Hitotsubashi University
Institute of Innovation Research

Claim Length as a Value Predictor of a Patent

Yoshimi Okada
Yusuke Naito
Sadao Nagaoka

IIR Working Paper WP#16-04

Revised in May 2016

Institute of Innovation Research
Hitotsubashi University
Tokyo, Japan
http://www.iir.hit-u.ac.jp

Claim length as a value predictor of a patent

May 2016

Yoshimi Okada[a]*, Yusuke Naito[a], Sadao Nagaoka[b]

[a] Hitotsubashi University, Institute of Innovation Research, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan

[b] Tokyo Keizai University, 1-7-34 Minamicho, Kokubunji, Tokyo 185-0021, Japan

Abstract

The claim of a patent defines the scope of patent right and provides crucial information on patent value. However, most empirical research uses only the number of claims as an indicator of patent value. We show that the breadth of a claim of Japanese patents, measured by the inverse of claim length, has significant explanatory power for patent value measured by applicant forward citations. Indeed, the explanatory power of claim breadth is comparable with that of the number of claims. The predictive power of claim breadth is stable for all quantiles in the discrete technology area, while it is far more significant for top-ranked patents in the complex technology area.

Keywords: Claim length, patent value, quantile regression, discrete technology, complex technology, applicant forward citation

JEL classification: O34

1. Introduction

The scope of patent rights is described by "claims," which should therefore provide crucial information on patent value. However, the existing empirical research predominantly uses only the number of claims to assess patent value. For example, Lanjouw and Schankerman (2004) use the number of claims in addition to forward and backward citations and family size as value indicators.

The higher the number of elements limiting the scope of the patent right, the longer the claim. We expect that the claim length is negatively correlated with the breadth of the patent's scope, and therefore, the patent value. Moreover, the predictive power of claim length is likely to differ between the complex technology and discrete technology areas,[1] since the value of a pioneer patent with a broad claim is likely to depend heavily to what extent complementary patents will emerge in the complex technology area (but not in the discrete technology area). However, except for Jansen (2009), to the best of our knowledge, there are no systematic studies on claim length with regard to patent value. Jansen (2009) examined the relationship between patent value and claims by investigating around 2700 European patents. He concluded that the length of independent claims[2] was not a significant predictor of patent value, contrary to practitioners' views.

This study aims to uncover how the length of an independent claim predicts patent value and how such predictive power differs between the complex technology and discrete technology areas.

2. Data construction

Generally, the first independent claim conveys the broadest inventive concept. We therefore focused on this claim. Because the Japanese language does not use spaces between words and it is hard to count the number of words, we used the number of characters instead of the number of words for the metric of claim length. Hereinafter, "claim length" denotes the number of characters in the first claim. We measured the breadth of the claim by the inverse of claim length and referred to it as "claim breadth."

We prepared our data sets from Japanese patent databases purchased from Artificial Life Laboratory, Inc., which covers all text data of patent publications as well as applicant citation data. In order to identify self- and non-self-citations, we utilized the dictionary of all major Japanese company names and the connection table for patent application provided by the National Institute of Science and Technology Policy (NISTEP) and considered only those patents filed by the applicants identified by the NISTEP database. Further, we used PATSTAT (2014 Autumn, the European Patent Office) to obtain US patent family data.

Since the relationship between claim length and patent value may differ significantly between product and process patents, we developed a program to divide patents into categories of "product" and "process."[3] Since product inventions accounted for more than 80% of patents,[4] we focused on product patents.

We restricted our assessment to patents filed between January 1991 and August 2002. The end limit was set to eliminate the influence of patent law change relating to disclosure of prior arts. The beginning limit was set

---

1 Cohen et al. (2000), for example, show how appropriability conditions differ between these two areas.
2 Claims that define all the essential components of the invention by itself are called "independent claims."
3 The randomly chosen 150 samples showed no errors.
4 The rates of product invention by technology are as follows: Total, 81.4%; Complex, 82.8%; Discrete, 78.3%; Computers & Communications, 85.8%; E & E, 79.1%; Mechanical, 84.5%; Chemical, 65.5%; Drugs and Medical, 90.9%; and Others, 83.4%.

because of the availability of text data. For simplicity, we eliminated the divisional applications from our data sets.

We divided the technological fields into six large categories: "Chemical (Chem.)," "Computers & Communications (C & C)," "Drugs and Medical (D & M)," "Electrical & Electronic (E & E)," "Mechanical (Mech.)," and "Others" (see the appendix Table A.1 for the International Patent Classification (IPC) correspondence table). "C & C," "E & E," and "Mech." belong to the complex technology area, and "Chem.," "D & M" and "Others," the discrete technology area.

We eliminated the patents involving chemical formulae, mathematical formulae, and/or tables in the first claim, utilizing "tag information" and its ilk; the crucial part of the patent right is provided by image data rather than text data in these patents. The amount of eliminated data was less than 5% in each field except for Chem.[5]

Figure 1 shows the distribution of the natural logarithm of claim breadth of product patents aggregated by all fields. Its shape is similar to that of a normal distribution.

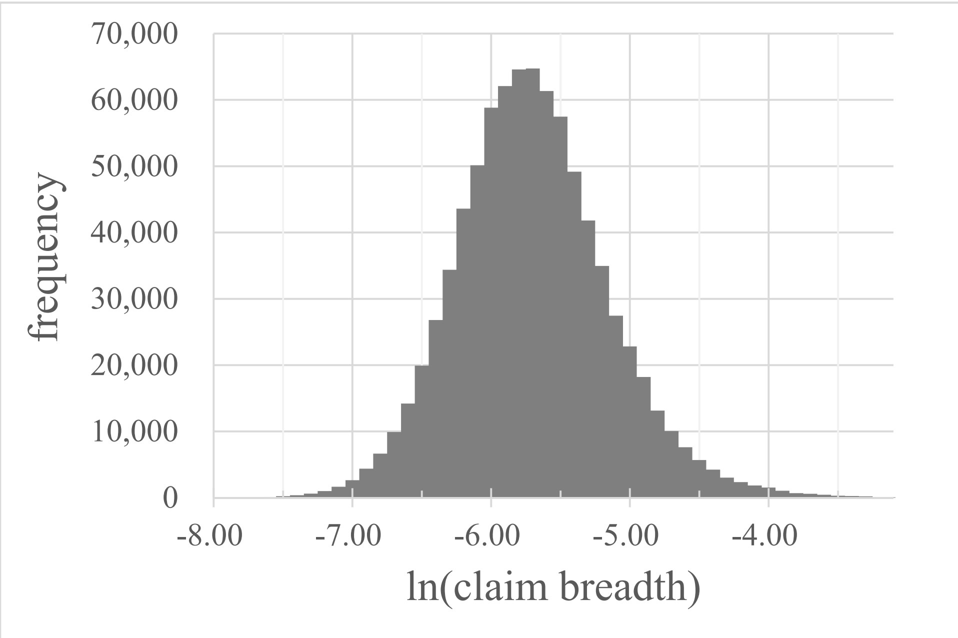

Figure 1 Distribution of the natural logarithm of claim breadth (-ln(claim length)) of product patents aggregated by all fields

3. Estimation model

We used ordinary least squares (OLS) as well as quantile regressions, which can accommodate the possibility of the value of a patent with broad claims having high variance. We used the number of forward citations as a value indicator. We assessed how claim breadth can be utilized to predict the number of forward citations, adding to the predictive power of the combination of all conventionally used major indicator variables, including the number of claims. Specifically, we used the following model, where Q_x is the *x* quantile or mean.

[5] The rates of the eliminated data are: all, 3.9%; Complex, 1.8%; Discrete, 9.0%; C & C, 0.4%; Mech., 1.0%; Others, 1.1%; E & E, 3.6%; D & M, 4.8%; and Chem., 25.9%.

$$Q_x(ln\_Num\_Forward\_Citation|explanatory\ variables) = \beta_0 ln\_claim\_breadth + \beta_1 ln\_N\_claims$$
$$+ \beta_2 ln\_N\_inventors + \beta_3 US1EP1\_dummy + \beta_4 US1EP0\_dummy + \beta_5 US0EP1\_dummy$$
$$+ \beta_6 ln\_N\_Bnonself\_Citn + \beta_7 ln\_N\_Bself\_Citn$$
$$+ \beta_8 ln\_N\_Bttl\_ExCitn + \beta_9 ln\_N\_Bttl\_CoCitn$$
$$+ \beta_{10} subseq\_appln\_dummy \ + \beta_{11} co\_ownship\_dummy$$
$$+ \beta_{year}\ effective\_filing\_year\_dummies + \beta_{tech}\ tehnology\_dummies$$
$$+ \beta_{year,tech}\ effective\_filing\_year\_dummies \times technology\_dummies + \varepsilon \qquad (1)$$

The meanings of the variables are shown in Table 1.

Table 1 Descriptive statistics for the data

Variable	Meaning	Complex: N=601,474		Discrete: N=231,858	
		Mean	Std. Dev.	Mean	Std. Dev
ln_Num_Forward_Citation	logarithm of “the number of forward citations + 1”	0.553	0.733	0.605	0.775
ln_claim_breadth	logarithm of the inverse of claim length	-5.78	0.506	-5.51	0.614
ln_N_claims	logarithm of the number of claims	1.18	0.818	1.05	0.763
ln_N_inventors	logarithm of the number of inventors	0.586	0.603	0.745	0.601
US1EP1_dummy	1 if both the corresponding US patent and the corresponding EP application exist; otherwise, 0	0.0933	0.291	0.0681	0.252
US1EP0_dummy	1 if there is a corresponding US patent but no corresponding EP application; otherwise, 0	0.143	0.350	0.046	0.209
US0EP1_dummy	1 if there is no corresponding US patent and there is an EP patent application; otherwise; 0	0.00816	0.0900	0.00995	0.0993
ln_N_Bnonself_Citn	logarithm of “the number of backward non-self-citations + 1”	0.272	0.489	0.345	0.580
ln_N_Bself_Citn	logarithm of “the number of backward self-citations + 1”	0.143	0.354	0.180	0.391
ln_N_Bttl_ExCitn	logarithm of “the number of “examiner citations + 1”	1.28	0.638	1.15	0.675
ln_N_Bttl_CoCitn	logarithm of “the number of backward citations cited both by the applicant and the patent examiner + 1”	0.101	0.268	0.124	0.300
subseq_appln_dummy	1 when subsequent divisional applications exist; otherwise; 0	0.0406	0.197	0.0414	0.199

co_ownship_dummy	1 if the patent is jointly held; 0 if the patent holder is single	0.0613	0.240	0.0705	0.256
effective_filing_year_dummy	earliest priority year of application				
technology_dummy	6 Categories				

4. Results and discussion

Table 2 shows the summary results for complex and discrete technology areas.

Table 2 Summary results of the regression

Dependent variable: *ln_Num_Forward_Citation*								
	complex				discrete			
VARIABLES	OLS	.99 quantile	.90 quantile	.70 quantile	OLS	.99 quantile	.90 quantile	.70 quantile
ln_claim_breadth	.0146***	.154***	.0406***	.00623***	.0544***	.0694***	.0815***	.0621***
	(.00192)	(.0111)	(.00437)	(.00197)	(.00283)	(.0150)	(.00598)	(.00308)
ln_N_claims	.0741***	.165***	.128***	.0648***	.0738***	.155***	.123***	.0727***
	(.00133)	(.00739)	(.00296)	(.00144)	(.00229)	(.0123)	(.00503)	(.00257)
ln_N_inventors	.0710***	.169***	.132***	.0664***	.0605***	.118***	.105***	.0552***
	(.00166)	(.00914)	(.00375)	(.00197)	(.00270)	(.0152)	(.00614)	(.00305)
US1EP1_dummy	.201***	.499***	.363***	.305***	.300***	.529***	.478***	.421***
	(.00388)	(.0233)	(.00933)	(.00482)	(.00814)	(.0431)	(.0162)	(.0100)
US1EP0_dummy	.115***	.266***	.205***	.167***	.160***	.270***	.298***	.204***
	(.00293)	(.0169)	(.00655)	(.00578)	(.00859)	(.0359)	(.0204)	(.0131)
US0EP1_dummy	.0878***	.137***	.152***	.135***	.123***	.252***	.188***	.211***
	(.0111)	(.0177)	(.0238)	(.0232)	(.0182)	(.0737)	(.0435)	(.0321)
ln_N_Bnonself_Citn	.145***	.419***	.263***	.177***	.136***	.298***	.224***	.175***
	(.00257)	(.0121)	(.00553)	(.00407)	(.00357)	(.0212)	(.00762)	(.0045)
ln_N_Bself_Citn	.0896***	.348***	.172***	.101***	.0781***	.130***	.117***	.0916***
	(.00370)	(.0178)	(.00737)	(.00485)	(.00499)	(.0274)	(.0103)	(.00534)
ln_N_Bttl_ExCitn	.0721***	.200***	.134***	.0593***	.0616***	.161***	.115***	.0555***
	(.00156)	(.00908)	(.00357)	(.00153)	(.00263)	(.0143)	(.00588)	(.00302)
ln_N_Bttl_CoCitn	-.0549***	-.240***	-.134***	-.0748***	-.0208***	-.160***	-.0531***	-.0178**
	(.00442)	(.0250)	(.00964)	(.00645)	(.00669)	(.0357)	(.0135)	(.00786)
subAppln_dummy	.340***	.751***	.577***	.467***	.345***	.853***	.611***	.452***
	(.00623)	(.0293)	(.0132)	(.0094)	(.0101)	(.0521)	(.0208)	(.0128)

co_ownship_dummy	-.0270***	-.0588**	-.0637***	-.0292***	-.00460	-.0534	.00222	-.0105*
	(.00393)	(.0293)	(.00883)	(.00412)	(.00625)	(.0381)	(.0133)	(.00598)
eff_filing_year	yes	yes	yes	yes	yes	yes	yes	yes
tech	yes	yes	yes	yes	yes	yes	yes	yes
eff_filing_year × *tech*	yes	yes	yes	yes	yes	yes	yes	yes
R-squared	.0686				.0972			
adjusted R-Squared	.0685				.0970			

Note. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

According to the results of the OLS estimation, the signs of the coefficients for claim breadth are positive and statistically highly significant in the two areas. Thus, claim breadth has a significant explanatory power for patent value, controlling for the existing indicators.[6] Its explanatory power in the complex technology area, however, is only one-third of that in the discrete technology area (0.015 vs. 0.054). Conversely, the values of the coefficients for number of claims are similar (0.074) and statistically significant in both areas. Although the explanatory power of claim breadth is smaller than that of the number of claims, it is comparable even if we consider the size of the standard deviations of claim breadth and number of claims (0.51 (0.61) and 0.82 (0.76) respectively in the complex (discrete) technology area).

The results of the quantile regressions reveal that the predictive power of claim breadth rises significantly with quantiles in the complex technology area (the highest (0.15) in the 0.99 quantile; negligible (0.006) for the 0.70 quantile), while it is stable for all quantiles in the discrete technology area. Conversely, the predictive power of the number of claims is relatively stable across all quantiles in both areas. Figure 2 compares the explanatory power of claim breadth relative to that of number of claims. It increases from 10% to 93% in the complex technology area (0.70 quantile to 0.99 quantile) and declines from around 85% to 45% in the discrete technology area. In the complex technology area, the value of a patent with a broad claim has high variance; the broader the inventive concept, the more frequently the patent will be ranked in the top 1% of commonly cited patents. The variance remains relatively constant in the discrete technology area.

[6] Jansen (2009) could not discover the predictive power of claim length possibly because his estimation was based on the mean effect of the indicators on the ranking of the value of a relatively small size of patents (2700), mostly in the complex technology area.

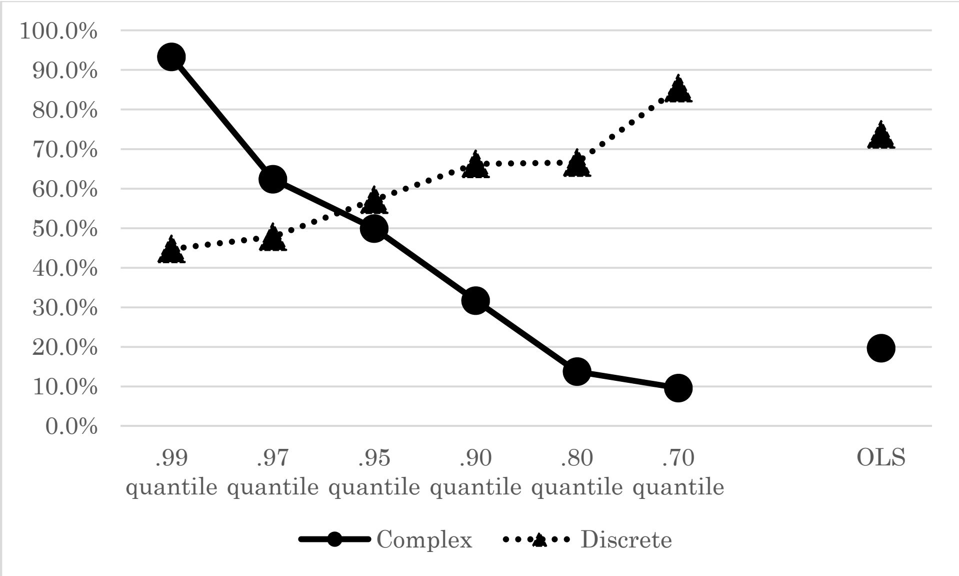

Figure 2 Relative value of the coefficient for claim breadth against that of number of claims

Note: The figure shows the values of (coefficient for claim breadth /coefficient for number of claims) estimated by quantile regression where the quantile is 0.99, 0.97, 0.95, 0.90, 0.80, and 0.70 as well as those estimated by OLS.

The differential effect of claim breadth between the two areas is consistent with the following interpretation. Since many patented technologies are combined to produce a product in the complex technology area, the value of a pioneering patent with a broad claim depends on the existence of many complementary technologies in this area. Thus, the value of a patent with a broad claim tends to have high variance in the complex technology area, since the availability of complementary technologies is uncertain, and such uncertainty rises with the scope of the patent claim. However, the value of a patent in the discrete technology area depends on the standalone value only; thus, a patent with a broad claim is not as uncertain.

5. Conclusions

The empirical results clearly show the considerable predictive power of claim breadth, which is measured by the inverse of claim length, for patent value. It is statistically significant in predicting the number of applicant forward citations, controlling for the major bibliographic indicators. Moreover, its predictive power is stable across all quantiles in the discrete technology area and rises significantly with quantiles in the complex technology area. These findings suggest that the value of a pioneering patent with few limitations on its claim is more uncertain in the complex technology area than in the discrete technology area, since the former depends on the existence of complementary technologies.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 26885031. We would like to thank Research Assistant Junko Morikawa for checking the samples.

References

Cohen, W.M., Nelson, R.R., Walsh, J.P., 2000. Protecting their intellectual assets: Appropriability conditions and why the US manufacturing firms patent (or not). NBER working paper 7552.

Jansen, W., 2009. Examining the relation between patent value and patent claims. Eindhoven: Eindhoven University of Technology.

Lanjouw, J.O., Schankerman, M.A., 2004. Patent quality and research productivity: Measuring innovation with multiple indicators. Econ. J.114, 441–465.

NISTEP, 2015, "Company name dictionary."

URL: http://www.nistep.go.jp/research/scisip/data-and-information-infrastructure (accessed 17.04.15).

NISTEP, 2015, "Connection table to other databases."

URL: http://www.nistep.go.jp/research/scisip/data-and-information-infrastructure (accessed 17.04.15).

Appendix

Table A.1 Technology sector classification

technology sector	IPC	Technology
Chemical	A01N, C01-C11(excludes C06),C21-C30	Nonorganic chemistry, Fertilizer, Organic chemistry, Pesticides, Organic molecule compounds, Dyes, Petroleum, Metallurgy, Coating metals
Computers & Communications	G04-G12, H03-H04	Clock, Controlling, Computer, Display, Information Storage, Instruments, Electronics circuit, Communication tech.
Drugs & Medical	A61-A63, C12-C14	Health and Amusement, Drugs, Biotechnology, Beer, Fermentation, Genetic Engineering
Electrical & Electronic	G01-G03, H01-H02,H05	Measurement, Optics, Photography, Electronics components, Semiconductor
Mechanical	B21-B32(excludes B31)-B44, B60-B68, F01-F04,F15-F17	Machine tools, Metal working, Casting, Grinding, Layered product, Printing, Transporting, Packing, Lifting, Engine, Pump, Engineering elements
Others	A01(excludes A01N), A21-A47, B01-B09, B31, B81,B82, C06, D01-D21, E01-E21, F21-F42, G21	Agriculture, Food stuffs, Personal and Domestic Articles, Separating, Mixing, Textile, Paper, Construction, Mining, Drilling, Lighting, Steam generation, Heating, Weapons, Blasting, Nuclear physics