Evaluating and Ranking Patents Using Weighted Citations

Sooyoung Oh
Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802
sooh@cse.psu.edu

Zhen Lei
Energy and Mineral Engineering
The Pennsylvania State University
University Park, PA 16802
zlei@psu.edu

Prasenjit Mitra John Yen
Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
{pmitra, jyen}@ist.psu.edu

ABSTRACT
Citation counts have been widely used in a digital library for purposes such as ranking scientific publications and evaluating patents. This paper demonstrates that distinguishing different types of citations could rank better for these purposes. We differentiate patent citations along two dimensions (assignees and technologies) into four types, and propose a weighted citation approach for assessing and ranking patents. We investigate five weight learning methods and compare their performance. Our weighted citation method performs consistently better than simple citation counts, in terms of rank correlations with patent renewal status. The estimated weights on different citations are consistent with economic insights on patent citations. Our study points to an interesting and promising research line on patent citation and network analysis that has not been explored.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms
Algorithms

Keywords
patent ranking, patent citation, weighted citation, patent renewal.

1. INTRODUCTION
Evaluating and ranking patents, such as identifying influential patents in a field or assessing the value of a patent or a patent portfolio, are important yet challenging problems. Patent citation counts, i.e. the number of citations that a patent receives, have been one of the most important and widely used indicator for patent value and importance [1,4,5,6]. Analogous to journal article citations that can help identify particularly seminal discoveries, patent citations suggest that the cited patents the relatively important precursor that defines the state of the art. The broader and more important the shoulder, the more likely it is to be cited. Furthermore, patent citations are referenced more parsimoniously than journal article citations as they define the scope of the citing patents and thus have significant legal and economic implications.

Previous studies have focused on the total number of patent citations a patent receives, paying little attention to a distinction among citing patents, a distinction that could have important implications for the value of the cited patent. Specifically, citing patents (and patent citations) can be distinguished among two dimensions: (1) whether a citing patent is owned by the same assignee (owner) as the cited patent (i.e. self-citations); and (2) whether a citing patent is the same technology class as the citied one. Patent self-citations suggest that a firm invests in further developing an innovation disclosed in the cited patent and thus presumably signify the economic value of the cited patent. With regard to technology classes, if a patent receives citations from other technology domains, it is likely that the patent is more valuable as it has impacts on subsequent innovations in other classes [4].

In this study, we classify patent citations along the two dimensions (assignees and technologies) into four groups: self-cited in the same class, self-cited from different class, other-cited in the same class, and other-cited from different class. We show that these four types of citations have different implications for patent value, and therefore, weighted patent citations are a better indicator for patent evaluation and patent ranking than a simple citation count that puts equal weights on different types of citations.

More specifically, we use patent renewal status as a reference to evaluate our weighted citation based measurement for patent value. U.S. patents are required to be renewed at the 4^{th}, 8^{th} and 12^{th} year after patent grant by paying maintenance fees, and it is no surprise that patent renewal status are indicative of patent value. We assume that the ranking of patents based on our weighted citations method correlates closely with the ranking pattern based on patent renewal status, and we measure this correlation by the Spearman's rank correlation coefficient [11].

We first apply an unconstrained nonlinear optimization method to estimate optimal weights on the four types of patent citations, by maximizing the Spearman's rank correlation coefficient. However, this numerical optimization approach is quite an expensive operation which sometimes even does not guarantee convergence. We then investigate other simpler regression-based approaches for estimation of weights and compare their performance with the performance of the optimal weights obtained from the nonlinear optimization approach, as an effort to find good methods for learning the weights.

According to our experimental results, our new weighted citation approach consistently shows a better performance, in terms of rank correlations with patent renewal status, than the simple unweighted citation approach, with around 20% improvement. The estimated weights for the four types of patent citations also shed interesting insights on economic implications of patent citations.

2. RELATED WORK
The literature on patent citation and patent evaluation has been exclusively on simple patent citation counts. Lanjourw and Schankerman [7] found a positive statistical relationship between measurements such as patent citations, the number of claims, patent renewal and litigation. Bessen [1] used patent renewal data to assess the value of patents and found that citations received are

JCDL'12, June 10–14, 2012, Washington D.C., USA.

positively and significantly associated with renewal. Harhoff et al. [6] showed that patent citations are correlated with patent value obtained from a survey of German patent holders on the price at which they would have been willing to sell the patent right three years after filing. Hall et al. [5] found that a firm's market value is larger when its patents cite more of its older patents, suggesting that the firm has been doing accumulative and follow-on research in its specialized area. This study is somewhat related to our proposed distinction of patent citations between self-cited vs. other-cited, but in the context of firm evaluation rather than patent evaluation.

There is also a large literature on journal article citation analysis. In bibliometrics, citations received by a paper are a very popular measure of paper quality. Walker et al. [12] proposed a CiteRank algorithm that incorporates two parameters, the inverse of the average citation depth and a time constant that is biased toward more recent publications. Bollen et al. [2] used a weighted PageRank algorithm to measure journal prestige, and Radicchi et al. [9] proposed a weighted PageRank algorithm on a directed weighted author citation network for ranking scientists by considering the diffusion of their scientific credits.

3. WEIGHTED CITATIONS

We classify patent citations using information on patent assignees and patent technology classes. On the dimension of assignees, if a patent is cited by patents filed by the same assignee (e.g. patents owned by the same company), we classify these citations as *Self-cited*. *Other-cited* refers to citations made by patents filed by different assignees. On the dimension of technology domains, if a patent is cited by patents from the same technology domain, we classify these citations as *Same-class*; otherwise, as *Diff-class*. Therefore, each citation belongs to one of four types. Table 1 listed the types of patent citations and their distribution for patent citations to patents granted between 1981 and 2000.

Table 1. Type of Citations

Type of citations	By Assignee		By Technology Category		Ratio in the total citations
	Same assignee	Different assignee	Same category	Different category	
C_1. Self-cited in the Same-class	O		O		10.74%
C_2. Self-cited from the Diff-class	O			O	2.67%
C_3. Other-cited in the Same-class		O	O		66.38%
C_4. Other-cited from the Diff-class		O		O	20.21%

Our approach is to put different weights on each type of citations, to improve patent citation based measurement for patent evaluation and rankings. Let $Wcited_i$ be the weighted citation of a patent i and C_{ij} is the aggregated number of the type j citations received by a patent i. Then the weighted citation is defined as follows:

$$Wcited_i = \sum_{j=1}^{4} w_j C_{ij}$$

Patent Renewal Data

Patent holders have to pay maintenance fee to renew their patents. In the United States, maintenance fees on utility patents are due 3.5, 7.5, and 11.5 years from the date of the original patent grant. Table 2 lists the maintenance schedule and fees. For small entities, including for-profit companies with 500 or fewer employees, non-profit organizations or individual inventors, the fees are reduced by 50%.

Patent renewal status is a strong indicator for patent importance and value. When a patent owner pays a renewal fee, it implies that the patent is worth more than the fee required to keep it in force. Hence, we learn the weights on four types of patent citations by evaluating the correlation between a patent's renewal status and its weighted citations. In future work, we shall use the maintenance fees as another indicator to measure the value of patents.

Table 2. Maintenance Schedule and Fee

Renewal	Maintenance fee* (Large entity)	Maintenance fee* (Small entity)	Patents Expired	Ratio of Expired
No renewal	0	0	4th year	12.13%
4th-year renewal	$1,130	$565	8th year	20.45%
8th-year renewal	$2,850	$1,425	12th year	19.16%
12th-year renewal	$4,730	$2,365	20th year	48.26%

(*http://www.uspto.gov/web/offices/ac/qs/ope/fee092611.htm)

Spearman's Rank Correlation Coefficient

We use each patent's renewal status as the reference to learn the weights on different types of citations. We posit that the rank of patents by weighted citations is highly correlated with the rank by their renewal stages. Therefore, it seems appropriate to use Spearman's rank correlation coefficient [11] to measure the correlation between weighted citations and renewal stage. The Spearman's rank correlation coefficient is calculated using the Pearson correlation coefficient between the ranked variables. For a sample of size n, the n raw scores X_i, Y_i are converted to ranks x_i, y_i and the rank correlation coefficient, ρ, is computed from these:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Learning the Weights

To learn the weights on different types of patent citations, we first use an unconstrained nonlinear optimization method [8] to maximize the Spearman's correlation coefficient between weighted citations and patent renewal status, which gives us the optimal results if the rank correlation is our objective function. Let X be the patent citation data set and Y is the renewal status of patents. The row of X is patent citations that are classified into the four types shown in Table 1: self-cited in the same class, self-cited from different class, other-cited in the same class, and other-cited from different class. Each feature represents the number of citations received for each type. The renewal status of each patent is presented by ordered numbers such as 0 (expired at 4th-year), 1 (renewed at 4th-year but expired at 8th-year), 2 (renewed at 4th-year and 8th-year but expired at 12th-year), or 3 (full-term renewal). The nonlinear optimization will search for the optimal weights on different types of patent citations to maximize Spearman's correlation coefficient between patent weighted citations and patent renewal status, as described as follows:

$$\max_w \rho(Xw, Y)$$

However, it is expensive to use a numerical optimization method. Sometimes it does not even converge with an inappropriate initial condition. Hence, we turn to regression-based and simpler methods to learn the weights. First we use a linear regression to learn the weights, where the dependent variable is ordered patent renewal status (0,1,2,3).

Table 3. Rank Correlation Results of Patents Granted in 1990 with 10 Years Citations

Weight learning methods	C_1. Self-cited in the same-class	C_2. Self-cited from the diff-class	C_3. Other-cited in the same class	C_4. Other-cited from the diff-class	Constant	Spearman's rank correlation coefficient	Improvements
Not weighted	1	1	1	1	-	0.2037	(baseline)
1. Nonlinear optimization	2.1060 (14.79)	2.2610 (15.88)	0.2298 (1.61)	0.1424 (1.00)		0.2466	21.02%
2. Linear regression	0.0530 (4.57)	0.0593 (5.11)	0.0141 (1.22)	0.0116 (1.00)	1.7905	0.2393	17.46%
3. Exponential-scaled linear regression	0.3902 (4.41)	0.4648 (5.26)	0.1046 (1.18)	0.0884 (1.00)	10.1430	0.2392	17.42%
4. Linear regression on log-scaled features	0.3054 (4.82)	0.2460 (3.88)	0.1526 (2.41)	0.0634 (1.00)	1.5513	0.2453	20.38%
5. Nonlinear regression	0.9537 (8.63)	1.1064 (10.01)	0.1647 (1.49)	0.1105 (1.00)	4.2464	0.2451	20.31%

Table 4. Rank Correlation Results of Patents Granted from 1985 to 1989 with 10 Years Citations

Patents granted year (# of patents)	Not weighted (baseline)	1. Nonlinear optimization	2. Linear regression	3. Exponential-scaled linear regression	4. Linear regression on log-scaled features	5. Nonlinear regression
1985 (50,293)	0.1779	0.2169 (21.89%)	0.2049 (15.15%)	0.2038 (14.53%)	0.2151 (20.91%)	0.2149 (20.81%)
1986 (50,172)	0.1867	0.2257 (20.89%)	0.2186 (17.13%)	0.2180 (16.77%)	0.2239 (19.97%)	0.2245 (20.24%)
1987 (59,868)	0.1946	0.2309 (18.69%)	0.2241 (15.16%)	0.2237 (14.99%)	0.2294 (17.91%)	0.2295 (17.95%)
1988 (56,687)	0.1993	0.2377 (19.28%)	0.2075 (4.12%)	0.2072 (3.98%)	0.2335 (17.17%)	0.2364 (18.60%)
1989 (69,439)	0.2002	0.2430 (21.37%)	0.2347 (17.24%)	0.2341 (16.95%)	0.2410 (20.36%)	0.2419 (20.81%)

Table 5. Rank Correlation Results of Patents Granted in 1990 with Truncated Citations

Truncated year (# of patents)	Not weighted (baseline)	1. Nonlinear optimization	2. Linear regression	3. Exponential-scaled linear regression	4. Linear regression on log-scaled features	5. Nonlinear regression
1991 (17,507)	0.0718	0.0894 (24.47%)	0.0883 (23.00%)	0.0882 (22.90%)	0.0874 (21.72%)	0.0885 (23.21%)
1992 (35,509)	0.1031	0.1280 (24.11%)	0.1241 (20.31%)	0.1227 (18.97%)	0.1270 (23.17%)	0.1280 (24.09%)
1993 (46,298)	0.1243	0.1539 (23.80%)	0.1517 (22.09%)	0.1519 (22.19%)	0.1531 (23.18%)	0.1529 (23.02%)
1994 (52,845)	0.1464	0.1781 (21.68%)	0.1745 (19.17%)	0.1743 (19.08%)	0.1761 (20.29%)	0.1775 (21.22%)
1995 (56,867)	0.1640	0.1963 (19.72%)	0.1924 (17.35%)	0.1922 (17.24%)	0.1950 (18.91%)	0.1956 (19.31%)
1996 (59,566)	0.1757	0.2122 (20.74%)	0.2067 (17.61%)	0.2066 (17.55%)	0.2101 (19.56%)	0.2107 (19.90%)
1997 (61,485)	0.1821	0.2206 (21.16%)	0.2151 (18.14%)	0.2148 (17.94%)	0.2189 (20.19%)	0.2190 (20.27%)
1998 (63,259)	0.1914	0.2318 (21.10%)	0.2255 (17.80%)	0.2254 (17.76%)	0.2301 (20.20%)	0.2303 (20.27%)
1999 (64,548)	0.1976	0.2391 (21.04%)	0.2323 (17.58%)	0.2323 (17.56%)	0.2377 (20.30%)	0.2377 (20.34%)

Now, X also includes all 1's constant feature column.

$$Y = Xw$$

Given that the distribution of citations is more skewed than the distribution of ordered patent renewal status, we also try another linear regression where we use exponentials of Y as dependent variables, as specified below:

$$e^Y = Xw$$

We also try the nonlinear regression as follows:

$$Y = \log(Xw + 1)$$

Our last approach is a linear regression on logarithms of each type of citations, as shown below:

$$Y = w_0 + \sum_{i=1}^{4} w_i \log(x_i + 1)$$

4. EXPERIMENTS AND RESULTS

Patent Data Set

We use the U.S. utility patent citation data from the National Bureau of Economic Research (NBER) patent data project [10]. It includes information on patent citation between 1976 through 2006, patent assignees and technology categories. To classify technology domain for each patent, we use the HJT 6-class fields suggested by Hall, Jaffe, and Trajtenberg [4], which consist of six technology categories: Chemical, Computers and communications, Drugs and medical, Electrical and electronic, Mechanical, and Others. We also tested the HJT 36-class category which is a more refined system, and the results are similar.

For the patent renewal data, we use Patent Maintenance Fee event data set downloaded from Google USPTO Bulk download site [3]. This data set provides maintenance fee events from 1981

to present. We parse maintenance fee events to get the renewal stage for each patent.

Comparison of Learning Methods

To test our five weight learning methods, we investigate patent citations within a window of 10 years after patent grant for each of the patents granted in 1990. For evaluation, we use 10-fold cross validation and average the results from 10 10-fold cross validations.

The results in Table 3 show that using weighted citations improves the Spearman's rank correlation between patent citations and patent renewal status by more than 20%. The nonlinear optimization method leads to the optimal weights for the four types of patent citations and the highest rank correlation. Among the four regression methods, performance with the linear regression using log-scaled features and the nonlinear regression is quite close to the optimal performance.

Weights on Different Types of Citations

The weights on each type of citations, shown in Table 3, confirm that self-citations get much more weights than citations made by other assignees. While self-citations suggest that a firm has further developed innovations based on the cited patent, other-citations might indicate that competitors have entered the market and ruined the value of the cited patent. Thus we observe a stark difference between the weights for self-cited and for other-cited.

Moreover, when a self-citation is from different technology domains, it gets more weights than a self-citation from the same technology field, suggesting that, other things equal, a patent that get utilized in other domains within the firm is more valuable. However, if a patent becomes the shoulder for innovations and products in other domains by the competitors, then it is least valuable to the patent owner, as suggested by the weight on other-cited from different classes.

Results with Other Patent Sets

We confirm these results with other patent test sets. Table 4 presents the rank correlation results for five different sets of patents that were granted at different years. All of the five test sets show consistent results, including around 20% improvements in the performance of each of the five weight learning methods.

Even though the nonlinear optimization method and the nonlinear regression method show the best outcomes, but they are computationally expensive compared to linear regressions. It is interesting that the linear regression with log-scaled features shows promising results, which essentially takes the product (rather than the sum) of weight powered citations. We will explore this interesting finding more in future studies.

Truncated Citation Window

Table 5 shows the rank correlation results for patents granted in 1990 with truncated citations. It is a more difficult problem to estimate the value of patents with some shortage in citation information, for example, when a patent was granted not long ago. However, it might be more useful to predict a patent's importance as early as possible as the patent owner can decide more strategically what to do with the patent. At each truncation year, we conduct experiments for the subset of patents that had received at least one citation up to that year. The result shows that the weighted citation performs even better when patent citations are further truncated.

Table 6 shows the changes in weights for the four types of citations if citations are truncated. The test set is patents granted in 1990 and the weights are learned using the nonlinear optimization method. The weights in the table are relative to the weight on other-cited from different classes. The results show that the weights on self-citation continuously increased when the citation window becomes larger, whereas the changes in weights for other-cited are modest. One possible reason for why self-citations are much more sensitive than other-citations is that the proportion of self-citations among all citations is decreasing with more citation years, dropping from about 20% in a three year window to 10% in an eleven year window.

Table 6. The Ratio of Weights by Truncated Years

Citation Window	C_1. Self-cited in the same class	C_2. Self-cited from the diff-class	C_3. Other-cited in the same class	C_4. Other-cited from the diff-class
3-year	2.13 (17.46%)	2.31 (3.48%)	1.03 (62.44%)	1.00 (16.62%)
4-year	7.73 (15.69%)	8.42 (3.27%)	2.01 (63.66%)	1.00 (17.38%)
5-year	8.06 (14.38%)	9.14 (3.14%)	1.79 (64.46%)	1.00 (18.02%)
8-year	12.73 (11.54%)	12.84 (2.82%)	1.49 (65.73%)	1.00 (19.91%)
11-year	15.51 (9.71%)	15.02 (2.52%)	1.44 (66.43%)	1.00 (21.34%)

5. CONCLUSIONS

To our best knowledge, this is the first paper that distinguishes the different types of patent citations and proposes weighted citation based indicators for patent evaluation and ranking. We compare various methods for learning the weights for different citations. We find that the linear regression with log-scaled features leads to results as good as the optimal solutions obtained from the nonlinear optimization method. Our weighted citation approach shows consistent improvements on rank correlation with the renewal data. These results point to a promising approach to better ranking scientific literature and patents in digital libraries that has not been explored before.

6. REFERENCES

[1] Bessen, J. 2008. The value of us patents by owner and patent characteristics. *Research Policy*, 37(5), 932-945.

[2] Bollen, J., Rodriguez, M.A., and Van de Sompel, H. 2006. Journal status. *Scientometrics*, 69(3), 669-687.

[3] Google Patents: USPTO Bulk Downloads: Patent Maintenance Fees. http://www.google.com/googlebooks/uspto-patents-maintenance-fees.html

[4] Hall, B.H., Jaffe, A.B., and Trajtenberg, M. 2001. The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools. *NBER Working Paper*. No. 8498.

[5] Hall, B.H., Jaffe, A.B., and Trajtenberg, M. 2005. Market value and patent citations. *The RAND Journal of Economics*, 36(1), 16-38.

[6] Harhoff, D., Narin, F., Scherer, F.M., and Vopel, K. 1999. Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, 81(3), 511-515.

[7] Lanjouw, J.O. and Schankerman, M. 2004. Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495), 441-465.

[8] Nocedal, J. and Wright., S. J. 1999. *Numerical Optimization*. Springer-Verlag.

[9] Radicchi, F., Fortunato, S., Makines, B., and Vespignani, A. 2009. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*., 80(5).

[10] The New NBER Patent Data Project. https://sites.google.com/site/patentdataproject/

[11] Wackerly, D. D., Mendenhall, W., and Scheaffer, R. L. 2002. *Mathematical Statistics with Applications*. Duxbury Advanced Series.

[12] Walker, D., Xie, H., Yan, K.K., and Maslov, S. 2007. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 7, P06010.