Association for Information Systems
AIS Electronic Library (AISeL)

SAIS 2019 Proceedings | Southern (SAIS)

3-22-2019

Predicting Patent Value: A Data Mining Approach

Xiaoyun He
Auburn University at Montgomery, xhe@aum.edu

Feng Zhang
Pennsylvania State University - Abington, fzz34@psu.edu

PREDICTING PATENT VALUE: A DATA MINING APPROACH

Xiaoyun He
Auburn University at Montgomery
xhe@aum.edu

Feng Zhang
The Pennsylvania State University Abington
fzz34@psu.edu

ABSTRACT

Patents have long been recognized as a rich source of data for studying innovation, technical changes, and value creation. Patent data includes citations to previous patents, and patent citations allow one to create an indicator of patent value. Identifying valuable patents in a timely manner is essential for effectively harnessing the business value of inventions in the increasingly competitive global market. However, the existing methods of evaluating patent value suffer the issues of timeliness and accuracy. In this paper, we propose a data mining approach that utilizes the structural properties of patent citations networks to predict the value of patents while aiming to improve timeliness and accuracy.

Keywords

Patent value, citation network, data mining, prediction, timeliness and accuracy

INTRODUCTION

Over the last few decades, patents have been widely recognized as a rich and potentially fertile source of data for the study of innovation, technological evolution, and economic growth (e.g., Griliches, 1998; Trajtenberg, 1990; Hall, Jaffe, and Trajtenberg, 2001). According to USPTO (United States Patent and Trademark Office), for a patent to be granted, the innovation must be new, non-trivial, and useful. Patent data include citations to previous patents and to the scientific literature. Patent citations serve an important legal function, since they define the scope of the property rights awarded by the patent (Jaffe et al., 1993). For example, if patent B cites patent A (i.e., patent A is cited by patent B), it implies that patent A represents a piece of previously existing knowledge upon which patent B builds. In this case, patent A becomes part of patent B's backward citations, while patent B is one of patent A's forward citations. In other words, backward citations of a patent are what it cites. Citations received by a patent are defined as forward citations (Henderson et al., 1998; Trajtenberg, 1990). These citations opens up a great possibility of tracing multiple linkages between many factors, such as inventions, inventors, firms, locations, etc. For instance, patent citations allow one to investigate spillovers and to create indicators of the "importance" of individual patents, and thereby introducing a way of capturing the tremendous heterogeneity in the "value" of patents (Hall, Jaffe, and Trajtenberg, 2001).

Research has shown that both backward and forward citations are associated with patent value (Gambardella et al., 2008). Indeed, the citations among the patents naturally form a citations network. Specifically, each patent is representeded by a node in a citations network, and an edge or link between two nodes represents the the citation relationship between the corresponding patents. The words networks and graphs are often used interchangeably. Given a patent and a certain time window, all of its forward citations and forward citations among these forward citations within that time window form a forward citation network of this particular patent. Similarly, when only backward citations are considered, a backward citation network can be constructed for the patent. Incorporating both forward citations and backward citations form a comprehensive citation network of the patent.

Many concepts and theories have been proposed to deepen the understanding of the network-structured data and be used to solve many problems of practical interest represented by networks (Cormen et al., 2009). For example, the analysis of network structure can reveal "important" nodes and community. Often, structural properties such as density, average degree, degree centralization, connectedness, diameter, breadth, closure, etc. are examined as statistics across many networks (Ghoshal, 2009). Along these lines, it would be interesting to look into whether the data derived from the structural properties of patent citations networks helps to discover the potential values of patents in a timely and accurate manner.

Examining patent value has attracted significant attention from researchers across such disciplines as legal, economics, and management (e.g., Lanjouw and Schankerman, 2001; Hall et al., 2005; Gambardella et al., 2008). Some studies use revenue generated from a new technology as a proxy for patent value (e.g., Collins and Smith, 2006; Shapiro, 2006). However, this

class of methods suffers an ex post bias by including only commercialized technologies so that it has little strategic implications for identifying valuable technologies for commercialization in a timely manner. Other studies focus on social and technological impact of a technology by examining patent forward citations, i.e., the number of future technological inventions built upon a patent (e.g., Hall et al., 2005; Harhoff et al. 2003). However, as a long observation window is required for citations to accumulate, patent forward citation methods prevent value immediate estimation once an invention is created. Even after a certain period of accumulation, subsequent citations would continue to build up, which leads to the inaccuracy of an earlier estimation, namely a truncation problem (Barberá-Tomás et al., 2011; Hall et al., 2005).

To mitigate the aforementioned issues in the existing methods, this study proposes a data mining approach that utilizes the structural properties of patent citations networks to predict the value of patents while improving the accuracy and timeliness. Specifically, we use the classification method to predict the patent value. We first obtain the data that measures the structural properties of patent citation networks, and then combine such measures with the information from the NBER (National Bureau of Economic Reserach) patent dataset to construct a training dataset, which is used to train the classification model. After the best model is selected, we use it to predict whether a patent is valuable or not. Using a selected patent data set, we conducted some initial testing of our proposed approach. The preliminary results show the promise of the approach.

RELATED LITERATURE

Timely identification of valuable technological inventions has strategic implications for organizations to maximize return on investments and overall social welfare (Zhang et al., 2018). As each patent contains highly detailed information on the invention itself and the technological area to which it belongs, patents increasingly reflect the inventive activity in the U.S. itself, but also around the world (Hall et al., 2001). Identifying a valuable invention has turned into identifying the patent value indicators. To date, the most popular and widely tested indicators of patent value are those based on patent citations (Zhang et al., 2018).

Patent citations suggest technological knowledge generation and diffusion between citing and cited patents. Both backward and forward citations are found to be associated with patent value (Gambardella, Harhoff, and Verspagen, 2008). Patent backward citations have been used to measure a technology's originality (Hall et al., 2001; Trajtenberg et al., 1997) and importance (Henderson et al., 1998; Lanjouw and Schankerman, 2001). Likewise, patent forward citations reveal a technology's generality (e.g., Henderson et al., 1998; Trajtenberg et al., 1997) and breadth (e.g., Hall et al., 2001, 2005; Henderson et al., 1998; Lanjouw and Schankerman, 2001; Sampat et al., 2003; Trajtenberg, 1990).

Due to the large variance in the technological and economic significance of individual patents, simple patent counts are considered as insufficient indicators of innovative value (Bessen, 2008; Gambardella et al., 2008). In other words, the explanatory power of citation counts is quite limited (Fischer and Leidinger, 2014). To overcome the deficiencies of simple citation counts, weighted counts of forward citations are introduced by incorporating subjective and arbitrary discount factors at different layers (Hall et al., 2005; Trajtenberg, 1990; Trajtenberg et al., 1997). He et al. (2008) propose a ranking-based metric to measure the value of patents, which not only takes into account of the citation counts, but also the importance ranking score of citing patents.

Yet, a truncation problem is often observed among indicators based on forward citations. Specifically, a long observation window is required for citations to accumulate, thus preventing accurate value estimation after an invention is created. Even after a certain period of accumulation, subsequent citations would continue to build up and thus affect the accuracy of an earlier estimation, namely a truncation problem (e.g., Barberá-Tomás et al., 2011; Hall et al., 2005). Other indicators suffer the same problem as well. For instance, a long enough period is required for a meaningful patent age measure and for foreign patent applications, renewal decisions, and oppositions/litigation to occur. Only if litigation about a patent happens, prosecution length can be calculated. Some legal scholars argue that opposition and litigation on valuable patents often happen soon after they are granted (Allison et al., 2003; Lanjouw and Schankerman, 2001). Yet, this measure is constrained by the fact that opposed or litigated patents might be valuable, but valuable patents are not always opposed or litigated. Fischer and Leidinger (2014) suggest that forward citations, family size and the number of claims that a patent has could serve as first indicators of patent value if at least five-year citation data are available. However, by examining the intertemporal distribution of patent returns, Giummo (2014) finds that highly valuable patents tend to receive significant returns at the later part of their term, suggesting the limitation of forward citations in assessing patent value regarding accuracy and timeliness.

PATENT CITATION NETWORKS

In this section, we conceptualize how a patent and the citations naturally form citation networks. Specifically, given a patent *p* that is applied in year t and a certain interval window Δ, its forward citation network records subsequent knowledge diffusion from patent *p* in Δ years (i.e., t+Δ). It has m layers, in which the 1st layer contains all the patents applied by the year of (t+Δ) that cite patent *p*. The 2nd layer contains all the patents applied by year t+ Δ that cite the 1st layer patents. The (m-1)th layer

contains all the patents applied by the year of (t+Δ) that cite the (m-2)th layer patents. Patent p's forward citation network boundary is set at mth layer that receives zero forward citation from any patent applied by the year of (t+Δ).

Below, we use a simplified example to illustrate how the 2-layer forward citation network of a patent can be constructed. Suppose that patent *p* was applied in year 2000, which is used as the reference, and we adopt a 5-year investigation interval (i.e., Δ = 5). We construct a forward citation network G = (V, E) of patent *p* as follows. Each node in G corresponds to a unique patent. Assume that, at the 1st layer, there are two patents, namely p_1 and p_4, that were applied by the year of 2005 (i.e., 5 years after year 2000), cited patent *p*, then a link (or edge) between patent *p* and patent p_1 is added; and a link between patent *p* and patent p_4 is added as well. If, at the 2nd layer, there are two patents, namely p_2 and p_3, that were applied by the year of 2005, cited patent p_1, two separate links are added, which are between patent p_1 and patent p_2 as well as between patent p_1 and patent p_3. In addition, at the 2nd layer, if p_2 cited patent p_4, a link between patent p_4 and patent p_2 is added as well. Figure 1. shows the 2-layer forward citation network of patent *p*.

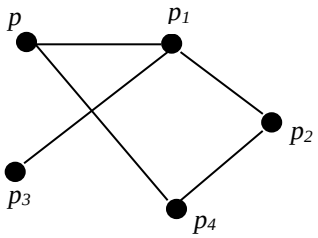

Figure 1. An Example: The Forward Citation Network of Patent *p*

The backward citation network of Patent *p* records the diffusion process of patented knowledge within the last Δ years (i.e., t-Δ) that resulted in patent *p*. The network contains n layers of backward citations. The 1st layer includes all the patents that patent *p* cites; the 2nd layer includes all patents that the 1st layer patents cite; the (n-1)th layer includes all patents that the (n-2)th layer patents cite. Patent *p*'s backward citation network boundary is set at nth layer when (n+1)th layer does not contain any patent applied within (t-Δ) years. Then, the construction of a comprehensive citation network of patent *p* is to incorporate its backward and forward citation networks, as well as the citations between them. Figures 2 illustrates a sample patent's comprehensive network, where Δ = 5.

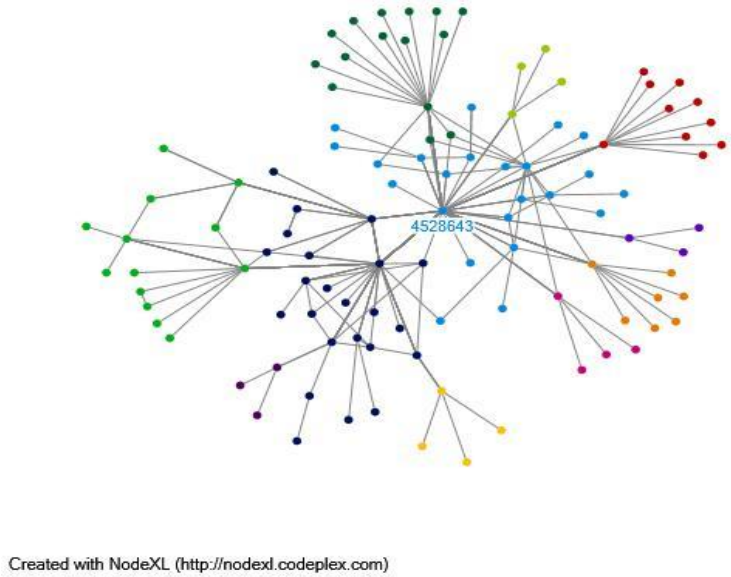

Figure 2. Comprehensive Citation Network of Sample Patent No. 4528643

PROPOSED APPROACH AND PRELIMINARY RESULTS

Networks can naturally model complicated structures and generic relationships among data objects in a wide variety of applications (Cormen et al., 2009). Such unique power offers us rich opportunity to analyse the network-structured data, and thus provide valuable insights. The fact that the relationships among the nodes that are reflected in the structure of a network

is often the focus of network analysis (e.g., Canning et al., 2018; Ghoshal, 2009). The analysis of network structure can reveal "important" nodes and community. Structural properties such as density, average degree, degree centralization, etc. are examined as statistics across many networks (Ghoshal, 2009). In this study, we are interested in utilizing the data derived from the structural properties of patent citation networks to predict the value of patents. We propose an approach that uses classification method to predict whether a patent is valuable or not. To do so, we follow the specific steps below.

Step 1: Given a set of patents, construct the citation networks for each of the patents. Recall that there are three types of citation networks: backward citation, forward citation, and comprehensive citation networks. For each type of these citation networks, the interval window Δ will be set as 5 years, 10 years, and 15 years respectively. For example, given a patent *p*, its forward citation networks will include 5-year forward citation network, 10-year forward citation network, and 15-year forward citation network.

Step 2: Run the network analysis tool to obtain the detailed measurement data on the structural properties of each of the citation networks. The tool we are currently using is the UCINET software (Borgatti et al., 2002). The structural properties that we measure include density, average degree, degree centralization, connectedness, diameter, breadth, closure, etc.

Step 3: Build the training data set. In this step, the structural properties from the step 2 are used as the predictors (or attributes). We need to add one target variable (or output variable) that is the variable being predicted in supervised learning. As our interest is to predict whether a patent is valuable or not, the variable named ValuablePatent is added as the target variable. If a particular patent is valuable, the value of variable ValuablePatent is set to 1; otherwise, it is set as 0. To build the training data, the actual value of ValuablePatent is based on the number of forward citations they received. The patents in the top 10% percentile are considered as valuable.

Step 4: Using the training data set from the previous step, we build and evaluate the classification models. There are a number of the classification techniques that can be used for our classification problem described above. Specifically, we start with SVM (Support Vector Machine) classifier, as it typically performs well for classification tasks and even small datasets (Han et al., 2011). Based on the evaluation, the best model is chosen.

Step 5: In this step, we apply the chosen model to the test data. That is, given a new patent, the chosen model is used to predict whether it is valuable or not.

To test our approach, we have conducted some initial experiments on a data set that we have collected so far. Specifically, we developed a C++ program to extract and construct the 5-year, 10-year, and 15-year forward citations networks of 70 patents respectively. The structural measures of these networks are used as the attributes of the datasets. Among these patents, about 30% of them are labeled as valuable, and the remainder as non-valuable. To estimate the overall accuracy of the SVM classification model, 10-fold cross-validation was used. That is, the complete dataset was randomly split into 10 mutually exclusive subsets of approximately equal size. The model was trained and tested 10 times. Each time, it was trained on all but one subset and then tested on the remaining single subset. The estimate of overall accuracy was then calculated by averaging the 10 individual accuracy measures.

The overall accuracy for the 5-year, 10-year, and 15-year forward citations networks was 81.8%, 82.3%, and 82.9% respectively. At this point, we were only able to obtain one set of the datasets with the three time intervals and the datasets are also relatively small. Thus, in terms of classification accuracy, we cannot draw any conclusion yet on whether the differences between the different time intervals are statistically significant. However, based on the closeness of the accuracies among the three time intervals, it seems to be a good indication that our proposed approach might be able to predict the value of patents in a short interval such as 5 years as accurately as the longer intervals such as 10-years or 15-years. That is, our approach could be promising in predicting whether a patent is valuable or not in a timely manner.

CONCLUSION AND FUTURE WORK

Identifying valuable patents in a timely and accurate way is essential for harnessing business value of new inventions. Yet, the existing methods suffer the issues of timeliness and accuracy. Aiming to mitigate such issues, this study proposes a data mining approach that utilizes the structural properties of patent citations networks to predict the value of patents. This study contributes to the literature in patent economics and research fields that would benefit from the timely and accurate prediction of valuable technological inventions. The proposed approach also has practical and commercial implications for businesses when quickly seizing the opportunities arising from innovations is critical for them to gain competitive advantage.

Our initial experiments show the promise of our proposed approach. However, the dataset that we have obtained so far is relatively small, and we have only tested our approach on the forward citation networks. Obtaining and preprocessing the data is a time consuming process. We plan to expand our dataset to include the patents from a variety of categories, and apply our

approach to all three types of citation networks. With each type of citation networks, we would like to experiment with different sets of attributes to find out the most relevant ones. We also intend to compare our approach with the existing approaches to demonstrate whether our approach can identify valuable patents faster and more accurate.

It is worth noting that patent value can be reflected in different dimensions, such as importance, originality, generality, complexity, and basicness (Sampat et al., 2003; Trajtenberg, 1990; Trajtenberg et al., 1997). As this study only represents the initial effort in developing a novel approach to identify valuable patents quickly and accurately, our primary focus is on patent value as a general concept. In this regard, future direction may include exploring the relationships between structural properties of citation networks and various dimensions of patent value.

REFERENCES

1. Allison, J. R., Lemley, M. A., Moore, K. A., and Trunkey, R. D. (2003) Valuable Patents, *Boalt Working Papers in Public Law: UC Berkeley*.
2. Barberá-Tomás, D., Jiménez-Sáez, F., and Castelló-Molina, I. (2011) Mapping the importance of the real world: The validity of connectivity analysis of patent citations networks. *Research Policy*, 40, 3, 473-486.
3. Bessen, J. (2008) The value of US patents by owner and patent characteristics. *Research Policy*, 37, 5, 932-945.
4. Borgatti, S.P., Everett, M.G. and Freeman, L.C. (2002) Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.
5. Canning, J. P., Ingram, E. E., Nowak-Wolff, S., Ortiz, A. M., Ahmed, N. K., Rossi, R. A., and Soundarajan, S. (2018) Predicting Graph Categories from Structural Properties. *arXiv preprint arXiv:1805.02682*.
6. Collins, C. J., and Smith, K. G. (2006) Knowledge exchange and combination: The role of human resource practices in the performance of high-technology firms. *Academy of management journal*, 49, 3, 544-560.
7. Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009) *Introduction to algorithms*. MIT press.
8. Fischer, T., & Leidinger, J. (2014) Testing patent value indicators on directly observed patent value—An empirical analysis of Ocean Tomo patent auctions. *Research Policy*, 43, 3, 519-529.
9. Gambardella, A., Harhoff, D., and Verspagen, B. (2006) The Value of Patents, *1st Annual Conference of the EPIP Association: "Policy, Law and Economics of Intellectual Property"*. Munich, Germany.
10. Gambardella, A., Harhoff, D., and Verspagen, B. (2008) The value of European patents. *European Management Review*, 5, 2, 69-84.
11. Ghoshal, G. (2009) Structural and Dynamical Properties of Complex Networks. Dissertation, University of Michigan.
12. Giummo, J. (2014) An examination of the intertemporal returns of patented inventions. *Research Policy*, 43, 8, 1312-1319.
13. Griliches, Z. (1998) Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence*, University of Chicago Press, 287-343.
14. Hall, B. H., Jaffe, A., and Trajtenberg, M. (2001) The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). *National Bureau of Economic Research*.
15. Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005) Market value and patent citations. *RAND Journal of Economics*, 16-38.
16. Han, J., Pei, J., and Kamber, M. (2011) *Data mining: concepts and techniques*. Elsevier.
17. Harhoff, D., Scherer, F. M., and Vopel, K. (2003) Citations, family size, opposition and the value of patent rights. *Research policy*, 32, 8, 1343-1363.
18. He, X., Zhang, F., and Adam, N. (2008) Towards ranking the importance of patents. In IEEE Symposium on *Advanced Management of Information for Globalized Enterprises, AMIGE*, 1-5.
19. Henderson, R., Jaffe, A. B., and Trajtenberg, M. (1998) University as a source of commercial technology: a detailed analysis of university patenting, 1965-1988. *Review of Economics and Statisitics*, 80, 1, 119-127.
20. Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108, 3, 577-598.
21. Lanjouw, J. O. and Schankerman, M. (2001) Characteristics of patent litigation: a window on competition. *RAND Journal of Economics*, 129-151.

22. Sampat, B. N., Mowery, D. C., and Ziedonis, A. A. (2003) Changes in University Patent Quality after the Bayh-Dole Act: A Re-examination. *International Journal of Industrial Organization*, 21, 1371-1390.
23. Shapiro, A. R. (2006) Measuring innovation: beyond revenue from new products. *Research-Technology Management*, 49, 6, 42-51.
24. Trajtenberg, M. (1990) A Penny for Your Quotes: Patent Citations and the Value of Innovations, *RAND Journal of Economics*, 21, 1, 172-187.
25. Trajtenberg, M., Henderson, R., and Jaffe, A. B. (1997) University versus Corporate Patents: A Window on the Basicness of Invention. *Economics of Innovation and New Technology*, 5, 1, 19-50.
26. Zhang, F., Jiang, G., and He, X. (2018) Patent citations and value: through the lens of a social network approach. *International Journal of Management and Network Economics*, 4, 2, 115-143.