Predicting Patent Citations to measure Economic Impact of Scholarly Research

Abdul Rahman Shaikh
z1841128@students.niu.edu
Northern Illinois University
DeKalb, IL

Hamed Alhoori
alhoori@niu.edu
Northern Illinois University
DeKalb, IL

ABSTRACT

A crucial goal of funding research and development has always been to advance economic development. On this basis, a considerable body of research undertaken with the purpose of determining what exactly constitutes economic impact and how to accurately measure that impact has been published. Numerous indicators have been used to measure economic impact, although no single indicator has been widely adapted. Based on patent data collected from Altmetrics we predict patent citations through various social media features using several classification models. Patents citing a research paper implies the potential it has for direct application in its field. These predictions can be utilized by researchers in determining the practical applications for their work when applying for patents.

KEYWORDS

Patents, Economic Impact, Social Media, Altmetrics

ACM Reference Format:
Abdul Rahman Shaikh and Hamed Alhoori. 2019. Predicting Patent Citations to measure Economic Impact of Scholarly Research. In *Proceedings of Urbana-Champaign '19: JCDL Joint Conference on Digital Libraries (Urbana-Champaign '19).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The main purpose of the patent system is to stimulate innovation in the market. Given that relevant data are readily available, patents and patent statistics are widely used by researchers to identify and explore areas of technical change and innovation simultaneously analyzing economic growth. Lv et al. [4] identified 161 converging technologies by performing cluster analysis on USPC class patents of five parties for 10 years from 2005 to 2015. Langinier and Moschini [1] found that protecting innovations through patents is a crucial task for technical improvements in the industry and patents produce innovations that stimulate economic growth. Patent citations have great potential in terms of providing a way to measure economic impact through indicators of patent quality. Squicciarini et al. [6] provided various indicators to measure patent quality. Specifically, they found that indicators such as patent family size, patent citations, patent renewals, and claims provide information pertinent to the technological and economic value of innovations. Further, they proposed a patent quality index based on four to six dimensions of patent quality. Based on an analysis of patents and their citations for the period of 1963 to 1995, Hall et al. [3] found that patent citations include valuable information regarding the market value of firms, RD, and patent counts.

Patent data and patent citations are very useful indicators of the inventions and the RD expenditures of a firm. Based on a survey of previous studies on patent statistics, Griliches [2] found a very strong relationship between patents and a firms RD expenditures in cross-sectional dimensions. Ribeiro et al. [5] examined 167,315 United States Patent and Trademark Office (USPTO) patents from 2009 and the papers cited in those patents and found that a global knowledge flow exists between universities and firms Van Raan [7] studied the background of scientific non-patent references (SNPR) and concluded that patents with a high economic value invention were cited highly. Patent quality can be assessed by multiple indicators of economic value such as patent claims, SNPRs, patent family size, patent renewal, and forward citations of patents.

Predicting amount of patent citations can be helpful in measuring the economic impact of research and in understanding how knowledge is commercialized. Our study built classifiers to predict the likelihood of a research article being cited in patents using social media features.

2 DATA COLLECTION

A random dataset was collected from Altmetrics.com which initially contained a million records. We only considered the articles published after 2010 since those records would have higher social media mentions. We performed data cleaning on the dataset to look for duplicate records and null values ending up with 784,665 data records. From these records 372,755 records were cited by patents and 411,910 records were not cited by patents. We removed few social media features such as Weibo, F1000, QA and Reddit since they mostly had null values for most records.

We used social media features since scholarly research is mostly being published or discussed on social media. We had a total of nine features of which one was the target variable and the other eight were predictors used to build classification models. We analyzed the target variable Patent citation and transformed the target variable into binary format such that if a record had a patent citation then the target variable would be 1 and if the patent citation is 0 then the target variable would be 0. The eight predictors were counts of scholarly articles mentions on news outlets, blogs, policy documents, Twitter, Facebook, Wikipedia, Google+ and Mendeley. Those

Urbana-Champaign '19, June 2019, Urbana-Champaign, IL

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

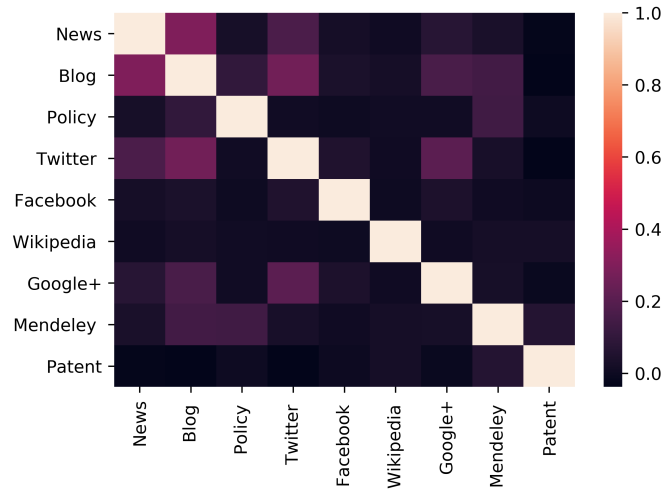

Figure 1: Correlation matrix of features

counts were considered before a patent was issued. Fig 1 shows the correlation between the features, we can notice that there is very low correlation between the features. We also included the citation count of research papers in the dataset to analyze the relationship between paper citations and patent citations.

3 METHODOLOGY AND RESULTS

We built four classification models using the processed dataset which are Logistic Regression (LR), Decision Tree (DT), Naive Bayes(NB), and Random Forest (RF). We evaluated the models based on their F1 score and accuracy as shown in Table 1. The table also shows the values of Precision and Recall for the models built on the dataset. After building classification models on the dataset, we observed that Random Forest performed the best in comparison to other models with an accuracy of 93.9% and an F1-score of 94.5.

	LR	DT	NB	RF
Accuracy	89.7%	92.6%	90.5%	93.9%
F1-score	90.3	93.0	90.4	94.5
Precision	90.2	92.6	90.7	94.2
Recall	90.4	93.4	90.1	94.8

Table 1: Accuracy and F1 based on several machine learning algorithms

From the dataset of 784,665 records, around 154,270 records had higher than 100 paper citations and among these records 124,267 records had an economic value or were cited by patents. Analyzing this dataset and the correlation matrix we found that papers which were highly cited by other papers were mostly cited by patents as well.

4 CONCLUSION AND FUTURE WORK

In this paper, we proposed a classification model which to predict whether a research article would have an economic impact by appearing in a patent using various social media features. We also found that papers which were highly cited by other papers were mostly cited by patents as well. In the future, we plan to use market values of patents to analyze the economic impact of a research article and create a framework which could help measure and predict economic impact of a given research article through patents.

REFERENCES

[1] GianCarlo Moschini Corinne Langinier. 2002. The Economics of Patents: An Overview. In *Intellectual Property Rights in Animal Breeding and Genetics* (1st. ed.), Max Frederick Rothschild and Scott Newman (Eds.). CABI publishing, Oxon, UK, Chapter 3, 31–50. https://doi.org/10.1079/9780851996417.0000

[2] Zvi Griliches. 1998. Patent Statistics as Economic Indicators: A Survey. In *RD and Productivity: The Econometric Evidence* (1st. ed.), Zvi Griliches (Ed.). University of Chicago Press, Chapter 13, 287 – 343. http://www.nber.org/chapters/c8351

[3] Bronwyn H. Hall, Adam B. Jaffe, and Manuel Trajtenberg. 2005. *Market Value and Patent Citations: A First Look.* Technical Report. National Bureau of Economic Research, Inc.

[4] Lucheng Lv, Tao Han, Yajuan Zhao, Xuezhao Wang, and Ping Zhao. 2018. Identification and Analysis of Converging Technology Based on Patent Co-Classification Relationship. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18).* ACM, New York, NY, USA, 363–364. https://doi.org/10.1145/3197026.3203903

[5] Leonardo Costa Ribeiro, Glenda Kruss, Gustavo Britto, Américo Tristão Bernardes, and Eduardo Motta E Albuquerque. 2014. A Methodology for Unveiling Global Innovation Networks: Patent Citations As Clues to Cross Border Knowledge Flows. *Scientometrics* 101, 1 (Oct. 2014), 61–83. https://doi.org/10.1007/s11192-014-1351-2

[6] Mariagrazia Squicciarini, Helene Dernis, and Chiara Criscuolo. 2013. Measuring Patent Quality: Indicators of Technological and Economic Value. *OECD* (2013). https://doi.org/10.1787/5k4522wkw1r8-en

[7] Anthony F.J. van Raan. 2017. Patent Citations Analysis and Its Value in Research Evaluation: A Review and a New Approach to Map Technology-relevant Research. *Journal of Data and Information Science* 2, 1 (2017).