Predicting Future Citations in Patents

Nolan Cooper
Undergraduate Student
Northen Illinois University
DeKalb, IL 60115
E-mail: Z1766022@students.niu.edu

Kethan Sai Yaram
Graduate Student
Northen Illinois University
DeKalb, IL 60115
Email: Z1840408@students.niu.edu

***Abstract*—In our research, we have applied machine learning to develop a predictive model for the citation of research papers in patents. By utilizing these predictions, researchers applying for patents can determine the practical applications for their work. We have used a Random Forest Classifier and a large set of training features to quickly predict patent citations across countries, fields of study, forms of social media, and levels of education. Data from the Altmetrics dataset provided a large corpus of material with which to train and test the model. This enhances the prediction capabilities of the model while making it generalizable to unseen data. In this study, we have managed to achieve a rather high accuracy and recall with our model and have designed it to be easily expanded upon in the future.**

Keywords—classifications, prediction, machine learning, patents, citations

I. Introduction

Patents are an excellent indicator of the practical applications for research and the current direction of scientific progress. Being cited in a patent shows that a paper covers relevant material and has potential for direct application in its field. Patent citations can help those applying for patents discover closely related patents or literature, as well as strengthen explanations of complex topics within their work; this improves their chances of receiving a patent. Patent citations can also be used from a competitive perspective for identifying others working in the same field. This can help someone who is applying for a patent to identify potential competitors while avoiding infringement issues. Overall, patent citations allow easy comparison of patent content, while also providing a metric for the relevance of the cited works. This benefits both those who are applying for patents and researchers who want to track the success of their work; however, researchers who are applying for patents can find difficulty tracking the success of their work. The goal of this research is to develop a predictive tool to ease this process using machine learning.

```
Model Performance
Average Error: 0.0120 degrees.
Accuracy = 98.80%.
[[2699490   11909]
 [  23978  264623]]
              precision    recall  f1-score   support

           0       0.99      1.00      0.99   2711399
           1       0.96      0.92      0.94    288601

   micro avg       0.99      0.99      0.99   3000000
   macro avg       0.97      0.96      0.96   3000000
weighted avg       0.99      0.99      0.99   3000000
```

Fig1. *Model Performance*

II. Related Work

There is a wealth of supporting work to look at and build on for our research project. Research has already been done to investigate connections between patent citations and quantitatively map technological trajectories, such as that of fuel cells. Being able to track knowledge flow quantitatively has been the subject of multiple papers, as it allows researchers to track the effectiveness of their work. The importance of this research can be seen through its relevance in scientific literature. Research has also shown that patent citation information can be used to home in on areas for further study based on the market effect of the patent. Companies use this

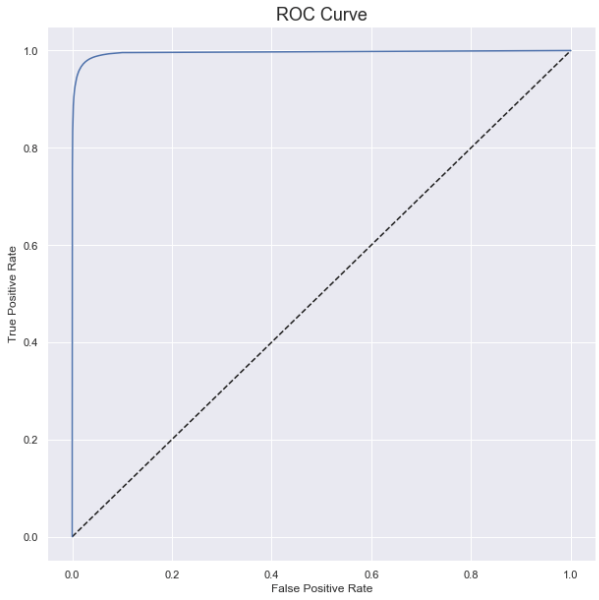

Fig. 2 *ROC Curve*

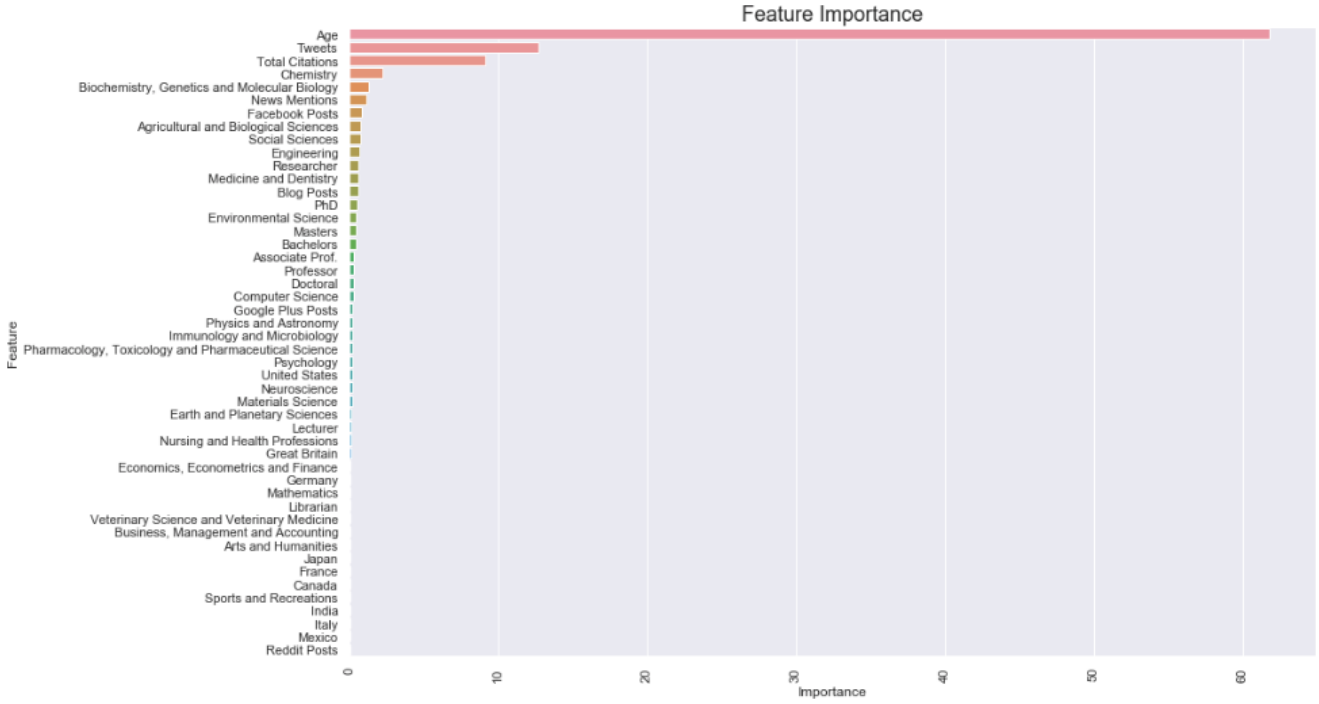

Fig. 3 *Feature Importance*

kind of data when considering mergers and acquisitions of other companies to determine the value of intellectual property. Even Google has created databases of patent citations to further understand their impact. Research into patent citation analysis has a strong place in multiple disciplines, ranging across market research, intellectual property valuation, and tracking of knowledge flow.

III. Dataset Characteristics

A. *Feature Selection*

The Altmetrics dataset is a dataset containing over 19.4 million records and hundreds of features, each record corresponding to a research paper. The features cover citation counts, associated field, social media mentions, countries where the paper is being read, and many more metrics. This provides an extremely large dataset to work with and many features to choose from. The data we are most interested in for our research is the patent citation count, which is sourced from IIFI CLAIMS and covers patents from nine international patent offices. Because the goal of our research was to determine whether a patent would be cited, the patent citation data was converted from a citation count into binary; 1 was given if the paper had been cited and 0 if it was not. Using all of the features available in the dataset would not be feasible for our research and would actually be a detriment to the classification process, so we made an initial feature selection to prevent this problem. By performing feature selection, we can reduce overfitting, increase accuracy, and decrease training time. We selected 48 features which were grouped into five different categories: (1) citation information, (2) popularity in media, (3) country of reader, (4) occupation of reader, and (5) paper topic. The full list of selected features can be seen in Fig. 3.

B. *Data Cleaning*

One downside to this dataset was the amount of processing needed to make it usable. The dataset took a considerable amount of time to process and clean due to its size, which was around 102GB. This difficulty was compounded by the format of the dataset, which was a directory structure composed of 196 folders containing a total of 380,518 specialized JSON files. Each file contains 50 lines, each line being an individual record stored in JSON format. This setup made it nearly impossible to load the dataset into memory all at once, so after completing our initial feature selection we loaded only the selected features, allowing the dataset to be kept in memory. This process, however, was lengthy and was being done every time the dataset was loaded for processing. An improved solution for handling the data was needed. We wrote a Python script that would load the necessary features from the dataset and put them into a pandas DataFrame which was then written out to a Comma Separated Variables (CSV) file. This CSV file was much smaller than the original dataset, weighing in at about 2GB, approximately a 98% reduction in size. The CSV file is also compressible to 276MB, making it much more portable for use on different machines. Because the data was stored as a single file, and not a collection of thousands like the original JSON format, it could be loaded much more quickly and with more control over the number of records. This cleaned and compressed dataset contained all 19.4 million records and selected features.

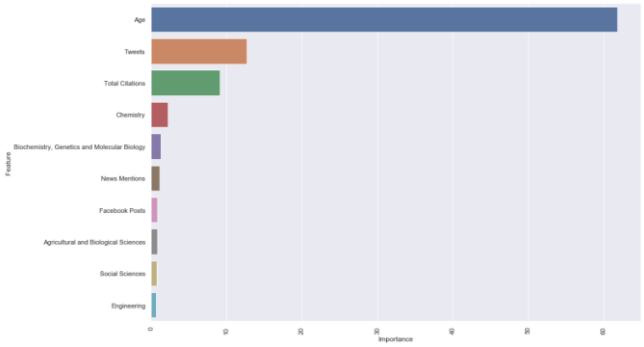

Fig. 4 *Top 10 Features by Importance*

TABLE I

Feature	Importance
Age	61.811013
Tweets	12.734072
Total Citations	9.136987
Chemistry	2.229020
Biochemistry, Genetics and Molecular Biology	1.324447
News Mentions	1.125326
Facebook Posts	0.819912
Agricultural and Biological Sciences	0.801383
Social Sciences	0.776511
Engineering	0.678347

IV. Methods

Our goal for our research was to predict whether a paper would be cited in a patent, so the binarized patent citation data was used as the target for this classification. Our tool of choice to generate this predictive model was a random forest classifier (RFC). The RFC was chosen due to its ability to quickly generate a model despite the large and sometimes sparse dataset. Other classifiers such as Support Vector Machines (SVM) and k-nearest neighbors (KNN) were examined. Downsides such as lower accuracy or longer training times made them poor candidates for our research. To implement our RFC, we used the scikit-learn RandomForestClassifier. The data was prepared using a train-test split of 70% training and 30% testing, and the data was stratified across the target feature to ensure a properly trained model. We evaluated the model by creating a function that displayed various metrics to measure the performance and ensure both accurate and useful predictions. These metrics consisted of average error, accuracy, precision, recall, F1 score, and a confusion matrix. Using this function, as well as other tools such as ROC Curves, the area under those curves (AUC), and cross validation, we set about tuning the model for maximum accuracy. Unfortunately, due to time and computing power constraints, we were unable to test and tune every hyperparameter of the RFC, so we focused on the ones which had the most impact on model performance. One of the most important hyperparameters was n_estimators, essentially the number of decision tress that would be created in the random forest. By default, this was set to 100 and we found, rather unsurprisingly that increasing this number improved model performance to a point. After fitting the model using a number of estimators ranging from 50 to 1000, we determined that a n_estimators value of 200 provided the most improvements with the least additional computation cost. Another important hyperparameter was max_features. This hyperparameter controls the number of features that the classifier considers when looking for the best split in trees. By default, this was set to "sqrt", meaning that the number of features considered for each split was the square root of the total number of features. After testing various configurations for this parameter, we found that setting it to "None", meaning that the model would consider all features for every split, greatly increased the model's performance. This did not come without increased computation times, but not enough to greatly delay our work. Two other hyperparameters were used for our model, setting n_jobs equal to -1 to run the model fitting processes on all available logical processors, and random_state equal to 42 to ensure consistent, repeatable results.

V. Classifier Performance

Using the methods described above, we measured the performance of our model. Our classifier performed strongly on all fronts, providing predictions with an accuracy of 98.81%. This is impressive but must be backed up by other metrics. These can be seen in Fig. 1, where our classifier achieved 99% precision and 100% recall for predicting papers that will not cited in patents, as well as 96% precision and 92% recall for papers that will be cited in patents. This results in F_1 scores of 99% and 94% respectively. These metrics in conjunction with an extremely low error of .0119 degrees and the confusion matrix point to extremely good predictive performance. This was initially unexpected due to the sparse data, which seems to have been cancelled out by the sheer volume of papers. Another important performance metric used to determine model performance was a ROC Curve, comparing the rate of true positives to the rate of false positives. Our model produced an exceptionally good ROC curve, with the AUC nearing 1 at a value of 0.9498032831366164. This was confirmed using a cross fold validation using 3 folds, returning AUCs of 0.99329288, 0.97491749, and 0.94669967, averaging to 0.97163668. Overall, this collection of performance metrics indicate that our model has extremely powerful predictive capabilities.

VI. Conclusion

In Figures 3, 4, and 5 the age of a research paper clearly has very high predictive importance when it comes to patent citations. This likely means a specific age group is regularly associated with increased patent citations, perhaps with research being generated specifically to be patented, or older papers having more chances over time to be cited. Tweets about a paper also contributed strongly to patent citations, although this does not mean the two features are linearly

correlated. It is possible the types of papers that are tweeted about are the exact opposite of papers that are regularly cited in patents. Further analysis of this trend could help to shed some light on the nature of this relationship. The total number of citations that a paper receives also plays a large role in the patent citations, likely showing that well known works are more likely to be referenced, almost like a snowball effect. As far as future work is concerned, preliminary tests show that this analysis could be done with a smaller data set; an overwhelming majority of the features have little effect on prediction accuracy. A reduction in features could drastically reduce computation times and facilitate more accurate results. Computation times could also be reduced by storing the fitted model as serialized file using Python's build in pickle library. These kinds of time saving efforts would leaving room for more detailed analysis for future researchers. Overall, patent citations allow easy comparison of patent content, while also providing a metric for the relevance of the cited works. This benefits both those who are applying for patents as well as researchers who want to track the success of their work.

ACKNOWLEDGMENT

We would like to thank Dr. Hamed Alhoori for his guidance in this research. His availability to answer any and all questions was extremely helpful throughout the project.

We would also like to thank Dr. Papka and the ddiLab for providing access to powerful computing resources, without which this project could not have been completed to such a degree of accuracy.

REFERENCES

[1] Collins, Peter & Wyatt, Suzanne, 1988. "Citations in patents to the basic research literature," Research Policy, Elsevier, vol. 17(2), pages 65-74, April.

[2] Sharma, P., & Tripathi, R. C. (2017). Patent citation: A technique for measuring the knowledge flow of information and innovation. World Patent Information, 51, 31-42.

[3] Kousha, K. and Thelwall, M. (2017), Patent citation analysis with Google. J Assn Inf Sci-Tec, 68: 48-61. doi:10.1002/asi.23608

[4] Li, R., Chambers, T., Ding, Y. , Zhang, G. and Meng, L. (2014), Patent Citation Analysis. J Assn Inf Sci-Tec, 65: 1007-1017. doi:10.1002/asi.23054

[5] Karki, M. M. S. (1997). Patent citation analysis: A policy analysis tool. World Patent Information, 19(4), 269-272.

[6] Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. Advances in Complex Systems, 10(01), 93-115.

[7] Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). National Bureau of Economic Research.

[8] Von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. Research Policy, 34(10), 1591-1607.

[9] Breitzman, A., & Thomas, P. (2002). Using patent citation analysis to target/value M&A candidates. Research-Technology Management, 45(5), 28-36.

[10] Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research, 15(1), 37-50.

[11] Thompson, P. (2006). Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations. The Review of Economics and Statistics, 88(2), 383-388.

[12] Valenzuela, M., Ha, V., & Etzioni, O. (2015, January). Identifying Meaningful Citations. In AAAI Workshop: Scholarly Big Data.

[13] Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014, September). Towards a stratified learning approach to predict future citation counts. In Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 351-360). IEEE Press.

[14] Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. Expert Systems with Applications, 34(3), 1804-1812.