Detecting novelty in USPTO patent applications with neural networks

CS-433 Machine Learning (EPFL) - Project 2

Chun-Tso Tsai, Giacomo Orsi, Vittorio Rossi

***Abstract*—Despite providing a publicly accessible definition of novelty, the United State's Patents and Trademarks Office (USPTO) receives tens of thousands of non-compliant utility applications every year. This paper provides information regarding the non-linear correlations between a patent application's data and its novelty, and describes a model which predicts whether a patent will be approved.**

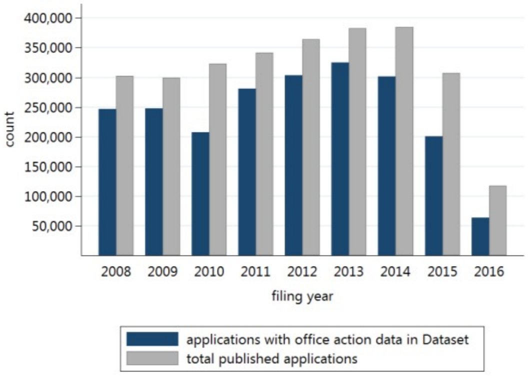

Figure 1. Comparison between the number of patent applications and the number of approved patents

I. Introduction

What is novelty? New inventions and ideas significantly shape the world we see today. A lot of research is carried out to understand the quintessence behind all these brilliant human creations. There exist different way to describe novelty. In this paper, we will use patents to assess novelty. It is evident that novelty have a close correlation with patents. In fact, a patent application can be approved only if it contains novelty and if it is not obvious. This analysis will aim to extract the idea of novelty by examining the decision issued by the United States patent office. This problem is called *novelty detection* in literature [1].

After the Obama administration's Open Government Agreement, the Public Patent Application Information Retrieval system (Public PAIR) data has become more readily available to researchers and other stakeholders [2]. Thanks to this, there are plentiful data available to analyse patents. On the contrary, it requires a lot of effort to select and organise the features related to novelty due to the gigantic amount of data.

By leveraging the data contained in the Patent Examinations, Office Actions [3], and Patent Claims [4] datasets, we build neural networks based on the *reconstruction-based* method of novelty detection [1]. This idea allows us to verify novelty by a machine classifier. Ideally, we want to analyse different aspects of applications, extracting information from an application's meta-data and claims with the help of our classifier model.

Due to the complexity of the patenting system, it is not unusual (Fig. 1) that inventors unknowingly retrace existing creations or that attorneys/agents in charge of the application on the inventors' behalf do not perform the due diligence in its entirety. These mistakes lead to the submission of non-compliant applications, which are rejected, and comprise the Office Actions dataset with their grounds for rejection.

Additionally, thanks to the advance of natural language processing techniques, researchers [5], [6] have unlocked previously unavailable features. Especially, we are able to extract the meaning behind the textual description of each patent as quantitative values. This allows us to integrate patent claims into consideration since they are critical to distinguish different patents.

By pairing inventor-side and USPTO-side data, it is possible to create an enriched profile associated with each application and infer meaningful remarks by observing how parts of this profile correlate to novelty. For this purpose, a variety of neural networks are used to learn non-linear relationships, and conclusions are obtained by varying and interpreting the set of input variables.

Section II explains the feature engineering process. Section III expands on the use of machine learning models, their selected structure, and the underlying reasoning. Section IV concludes by delineating the results obtained by the analysis, their interpretation, and possible alternative approaches.

II. Features

The incompleteness and heterogeneity [5] of the datasets provided by the USPTO pose significant challenges in distilling a subset to be used as valid features for the classification task. Not all patent applications published by the USPTO contain the list of inventors, attorneys, and the list of patent claims. Therefore, the analysis was restricted only to those patent applications that contain all the information mentioned above.

A. Data selection

Our analysis was carried out on patent applications that were revised in the period between 2010 and 2015. This was due to the incompleteness of the Office Action dataset.

As for the approved patents, we took all those for which the list of inventors, attorneys, and patent claims was available. As for rejected applications, we considered only those rejected for non-novelty (rejection code 102), for which the list of inventors, attorneys, and patent claims was available. Given that the USPTO can issue non-final rejections [3], inventors are allowed to modify the content of submitted applications, in substance or form, to meet the criteria of non-obviousness and novelty. Approved applications that were first rejected at an early stage for non-novelty were removed from the approved applications and retained only among those rejected.

We generated features containing information about the inventors and attorneys who prosecuted the patent applications by combining the research datasets mentioned in I. It is reasonable to assume that based on the professional history of inventors, the number of previously approved patents and the approval rate of applications may influence the outcome of new applications. The same is true for attorneys. More experienced attorneys will tend to produce higher-quality written patents.

For each application, we computed the number of attorneys and inventors and their respective number of applications. This data refers to the entire period 2010-2015.

The table I contains all the features that we extracted for each patent application.

B. Patent claims

Patent claims concisely and clearly formulate statements that explain what the invention consists of and the protection that is being claimed[1]. They are important references in the prosecution and litigation process, and they influence the quality of a patent application [4]. Therefore, patent claims can be used to distinguish one patent from another, thus potentially leading to ultimately understand the essence of each patent.

Despite the satisfying property that patent claims seem to characterize each patent well, there are still some difficulties to analyze the dataset containing the patent claims. The first challenge is the method to represent patent claims, as sentences, by quantitative data. Fortunately, the strict format and precise word choices for patent claims [2] ameliorates the performance to apply existing word embedding tools, resulting better performance compare to other NLP tasks [6]. This analysis uses *Doc2Vec* models to represent patent claims with word vectors. *Doc2Vec* model is a more generalized approach based on *Word2Vec* [7] and can represent arbitrary documents (long sentences) using a vector with a pre-assigned length. It was also shown to be a good way to represent patent claims in some precedent works [8], [9].

Within our targeted dataset, each application has 18 claims, on average. Combining that there are about 1.6 million applications in our targeted dataset, the amount of text is enormous. Therefore, a sample of 10% of the dataset was taken to train the *Doc2Vec* model. The dimension of the output vector was chosen to be 450, since USPTO classify patents into 450 classes.

Eventually, we computed the vector representation of the patent claims and the number of words, and used those as features for the classification task (studies shown that the number of words contained in the patent claims seems to be correlated with the approval rate [4])

III. Method

A naive way to compute the novelty of a patent application is to compare its vector representation with all the previously approved patents. This is the *distance-based* approach of novelty detection [1]. One can claim a new patent application is novel if there's no existing patents similar enough to the new one. Several studies already used this method [8], [9]. But the downside is that it is unfeasible to calculate word vectors for all the patents in the history, and performing a trasversal comparison would have time complexity of $O(N^2)$, with N being the number of approved patents in history.

According to the *reconstruction-based* method of novelty detection [1], we aim to detect novelty by training a classifier. Often, this approach is not feasible due to the lack of proper novelty labels on various applications. However, in our case, we took advantage of the non-novelty rejections issued by USPTO, so that we have labels to novel applications as `approved` and non-novel applications as `rejected` for non-novelty. The aim is to build a neural network which can classify whether

[1] Definition from *Swiss Federal Institute of Intellectual Property.*

[2] See USPTO Claim Drafting, 2019.

Table I
Features extracted from USPTO research datasets

Feature	Description
`application_number`	(will not be used for NN)
`approved`	`1` if approved, `0` if rejected for non-novelty
`examiner_art_unit`	Art unit of the examiner
`inventors_count`	Number of inventors
`inventors_avg_patents`	Average number of patents per inventors
`attorney_count`	Number of attorneys
`attorney_avg_patents`	Average number of patents per attorney

a patent application will be approved. That means that the model is able to identify whether a patent application contains novelty or not.

Two slightly different neural networks were trained, one with a single hidden layer, as previous literature in the field [10], [11] states that tabular data can be handled properly even by low-complexity models, and one with four hidden layers, as a comparison benchmark in each step of the analysis.

Training and testing were performed with an incremental number of features, starting only with subsets from the set described in Section II-A, then adding USPTO-side information, such as `examiner_art_unit` embedded with a one hot encoding, as proposed in [12]. As final addition, vector representation of the patent claims were added. In the first step, performance obtained by a *boosted trees* model was used as additional comparison metric. Attorney and inventor information required 4 input nodes, the encoded version of `examiner_art_unit` an additional 660, and claim-related features 451.

The dataset was stripped of dates, so the period of analysis was considered as a homogeneous set. From this whole, an 80-20 train-test split was used at all steps of the experiment.

IV. Results

A. Metrics

Contrary to what was hypothesized, both the neural networks and the tree model showed no correlation between input and output when fed with attorney and inventor-related features. The obtained model was comparable with a random classifier, with precision and recall valued at 0.5. Adding the embedded `examiner_art_unit` showed a significant increase in performance, improving accuracy to 64%. The last step of the analysis, which consisted of adding the word embeddings for all patent claims, did not improve the two neural network models.

B. Results Interpretation

The first step of the evaluation reported no statistical patterns. This goes against the field's literature [10], [13], [14], in which inventors are used as a proxy for human resources. Such erroneous results could be mitigated by implementing a more rigorous data grouping approach, increasing the accuracy of informative content in data by only counting previous interactions with the USPTO. This facet is lost in the study due to the ripple effect of simplifications made during the data cleaning phase.

Adding art unit-related information brought the most significant improvement to the study. The related features showed an imbalance in different art unit approval rates, confirmed by testing the model on unseen data.

The vectorized claims, which were hypothesized to be the most information-rich feature, struggled to improve performance in the two models. This was most likely due to the shallowness of the neural networks proposed, and the authors encourage testing the features with deep structures to exploit their full potential.

V. Summary

This paper described a way to assess novelty by analyzing patent applications. Firstly, we analyzed and merged several research datasets provided by USPTO to build a meaningful subset to build a classification model. Secondly, we used *Word2Vec* to create a vector representation of the patent claims. Finally, we incrementally fed neural networks with distinct subsets of features to predict whether a patent application would be approved. This research is the starting point to analyze further some machine learning techniques that researchers can carry out on the dataset we built to improve the overall performance. In addition, the results, the insights, and the information we have found on patent-related datasets can be used for future research in this area.

Acknowledgements

The authors would like to thank Dr. Davide Bavato, postdoctoral researcher in the *Entrepreneurship and Technology Commercialization Chair*, EPFL, for his helpful suggestions and his support to carry out this project.

Source code

The source code of this project is published at this link: https://github.com/CS-433/ml-project-2-novae-2

References

[1] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "Review: A review of novelty detection," *Signal Process.*, vol. 99, p. 215–249, jun 2014. [Online]. Available: https://doi.org/10.1016/j.sigpro.2013.12.026

[2] S. J. Graham, A. C. Marco, and R. Miller, "The uspto patent examination research dataset: A window on the process of patent examination," *Georgia Tech Scheller College of Business Research Paper No. WP*, vol. 43, 2015. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2702637

[3] Q. Lu, A. Myers, and S. Beliveau, "Uspto patent prosecution research data: Unlocking office action traits," 2017. [Online]. Available: http://dx.doi.org/10.2139/ssrn.3024621

[4] A. C. Marco, J. D. Sarnoff, and C. deGrazia, "Patent claims and patent scope," *USPTO Economic Working Paper 2016-04*, 2016. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2844964

[5] L. Helmers, F. Horn, F. Biegler, T. Oppermann, and K.-R. Müller, "Automating the search for a patent's prior art with a full text similarity search," *PloS one*, vol. 14, no. 3, p. e0212103, 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0212103

[6] J. Risch and R. Krestel, "Domain-specific word embeddings for patent classification," *Data Technologies and Applications*, vol. 53, no. 1, pp. 108–122, Jan 2019. [Online]. Available: https://doi.org/10.1108/DTA-01-2019-0002

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[8] R. Whalen, A. Lungeanu, L. DeChurch, and N. Contractor, "Patent Similarity Data and Innovation Metrics," *Journal of Empirical Legal Studies*, vol. 17, no. 3, pp. 615–639, September 2020. [Online]. Available: https://ideas.repec.org/a/wly/empleg/v17y2020i3p615-639.html

[9] S. Feng, "The proximity of ideas: An analysis of patent text using machine learning," *PLOS ONE*, vol. 15, no. 7, pp. 1–19, 07 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0234880

[10] C. Lee, O. Kwon, M. Kim, and D. Kwon, "Early identification of emerging technologies: A machine learning approach using multiple patent indicators," *Technological Forecasting and Social Change*, vol. 127, pp. 291–303, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162517304778

[11] I. Kaastra and M. Boyd, "Designing a neural network for forecasting financial and economic time series," *Neurocomputing*, vol. 10, no. 3, pp. 215–236, 1996. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.8688&rep=rep1&type=pdf

[12] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," *arXiv preprint arXiv:1604.06737*, 2016. [Online]. Available: https://arxiv.org/pdf/1604.06737.pdf

[13] Z. Ma and Y. Lee, "Patent application and technological collaboration in inventive activities: 1980–2005," *Technovation*, vol. 28, no. 6, pp. 379–390, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166497207001071

[14] H. Ernst, "Patent information for strategic technology management," *World Patent Information*, vol. 25, no. 3, pp. 233–242, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0172219003000772

[15] G. Inc. (2017) Google patents public datasets: connecting public, paid, and private patent data. [Online]. Available: https://bit.ly/3eoL5Mg