Journal Entry 0: A Novel Machine Learning Algorithm to Forecast Patent Filing Volume

Vishnu Murthy

September 7, 2018

1 Proposal

A large component of USPTO revenue is patent application filing fees. Basic filing fees range from $200-$300 depending on the type of patent and whether or not the filer is seeking a reissue [1]. Creating a model that can not only accurately predict the volume of patent application filings but also is sensitive to fluctuations in the market is very valuable, resulting in higher quality budget planning by the Office of the Chief Financial Officer and better preparation for large changes in volume.

Current models use an econometrics for prediction. The book, "Forecasting Innovations: Methods for Predicting Numbers of Patent Filings" by Hingley and Nicholas , summarizes the results of an external research programme that aimed to improve the forecasting methods in the European Patent Office. The models described in the book, which is similar to the models used in the USPTO, uses economic indicators such as Gross Domestic Product, Research & Development expenditures of various companies across a broad range of industries and number of employees in a particular year [2]. Although these models are helpful for forecasting patent filing volumes in the long-term, the economic models have difficulty in capturing short-term trends. Additionally, these indicators are relatively accurate at predicting patent volumes that come from amateurs or not from established companies that utilize patent attorneys.

Augmenting Google Patents information with econometric models can help capture these microtrends. Google is the largest search engine in the world, and Google Patents provides a search for existing patents, information being the number of searches per month on Google Patents and/or the number of searches particular search terms have (explained further below). Google's Chief Economist Hal Varian describes a method of using a seasonal autoregressive model on Google Trends data for predicting the present. The example model is a linear regression used for predicting car sales. The model uses the previous month's sales, the month in the previous year's sales, as well as Google Trend search queries in the first week of the current month to predict the current month's sales. [3] This project uses the seasonal autoregressive model with Google Patents in order to predict a given month's patent application volume.

There are three phases of this project:

1. Creating a baseline model with Google Trends: Google Trends provides trends data in real time. To receive data from Google Trends, search a Google query in the Google Trends website and download a .csv file with a time period and its normalized query share; "query share" meaning that similar search queries are grouped together, and normalized to a maximum of 100. Terms that a patent filer would search for, such as "Cost of Patent", "Patent Application Process", "Patent Application Search", "USPTO", "File for Patent", "Patent Filing Fees" can be used as parameters. Model:

$$\overline{x} = ax_{t-1} + bx_{t-1} + ... + kx_{t-11} + lx_{t-12} + mg_{1,t} + ng_{2,t} + ... + tg_{8,t} + ug_{9,t}$$

Figure 1: Model

where $\overline{x}$ is forecasted patent filing volume for month t, x_t is real patent filing volume in month t, $g_{n,t}$ is the normalized query share for Google Trend number n in month t, $a, b, c, ..., s, t, u$ are trainable parameters.

2. Using Google Patents data: Replace Google Trends data with Google Patents search data as below:

$$\overline{x} = ax_{t-1} + bx_{t-1} + ... + kx_{t-11} + lx_{t-12} + mg_{t-1} + ng_{t-2} + ... + tg_{t-11} + ug_{t-12}$$

Figure 2: Model

where $\overline{x}$ is forecasted patent filing volume for month t, x_t is real patent filing volume in month t, g_{t-n} is the number of searches the Google Patents interface has in month $t-n$, $a, b, c, ..., s, t, u$ are trainable parameters.

3. Forecast number of patents filed per category: Patent Offices use the Cooperative Patent Classification as a method of classifying patents into different departments (networking, systems, chemical engineering, etc) [4]. Phase 3 is not predict the raw number of patents filed, but predict the number of patents filed per category. This will be done by having models for each category, and getting the number of searches particular related search query have (for example, the materials science category would have queries "graphene", "nanoparticles", etc). Each category would have it's own model, and we would add time lags of 12 months for each of the queries.

2 Timeline

Date	Description
10/1	Complete Phase 1 - Google Trends information. Submit technical report
11/1	Get data on the number of searches Google Patents has per month
12/1	Filter out those search queries by IP address such that patent examiners are not counted in that figure
1/1	Complete Phase 2 - Google Patents information.
2/1	Get data on the number of searches particular search queries have per month
3/1	Complete Phase 3 - model for each CPC classifcation
4/1	Implement PatFT/AppFT (USPTO patent search interface) data with this model as well
6/1	Continue making improvements on model.

References

[1] O. of the Chief Financial Officer, Aug 2018.

[2] P. Hingley and M. Nicolas, *Forecasting innovations: methods for predicting numbers of patent filings*. Springer, 2006.

[3] H. Varian and H. Choi, "Predicting the present with google trends," Apr 2009.

[4] O. o. P. Classification, "Classification standards and development."