Student Innovation Challenge: Patent Volume Prediction Using Google Trends

Introduction

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. 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, number of employees in a particular year, etc. 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 Trends information with econometric models can help capture these microtrends. Google is the largest search engine in the world, and 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. 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. This project uses the seasonal autoregressive model with Google Trends in order to predict a given month's patent application volume.

Data Processing

The two sources of data used were the USPTO's Patent Examination Research (PatEx) Dataset and Google Trends. The PatEx dataset contains detailed information on each patent filed per line of the file. Preprocessing required counting the number of patents filed per month, and used data from 2006-2015. A dataframe was created with columns containing shifts of this data {1,2,...,12} months before. Google Trends provides query share data in a monthly format, so preprocessing required to use only 2006-2015 data. Both Trends and PatEx were normalized by subtracting the mean and dividing by standard deviation, allowing for comparison of coefficients to determine variable importance.

Abstract

This project aims to develop a method to forecast patent application volume using Google Trends information. One component of USPTO revenue is patent application filing fees therefore prediction of patent application volumes is important for budgeting purposes. This project utilized a mathematical model called a seasonal autoregressive model. as described in "Predicting the Present with Google Trends", Hal Varian et al. (2011). The model had parameters of previous patent filing volumes processed from the USPTO's Patent Examination Research Dataset (PatEx), as well as search query Google Trends data to fit a linear regression model. The model received a 5.62% error on training data and 6.57% error on testing data. This approach can hopefully be combined with other forecasting models currently used by the USPTO Office of the Chief Financial Officer to more accurately respond to real-time end user data via Google Trends .

Approach

The data was split into training and testing data, training data ranging from January 2006 to December 2013 and testing data ranging from January 2014 to November 2015. December 2015 was not included in the data due to the fact that it was a large outlier compared to the rest of the data. A randomized train test split was not used for the data because in a real world application, the model would be trained using all previous data and forecast on the next datapoint. Thus, having a set time interval for training and testing would more accurately reflect the model's real world usage.

The mathematical model used for this project was a seasonal autoregressive model below. SciPy's curve_fit was used to fit the function to the training data:

$$\bar{x}_t = a * x_{t-1} + b * x_{t-2} + \ldots + k * x_{t-11} + l * x_{t-12} + m * g_{1,t} + n g_{2,t} + \ldots + t * g_{8,t} + u * g_{9,t}$$

$\bar{x}_t$ is forecasted patent filing volume for month *t*

x_t is real patent filing volume in month *t*

$g_{n,t}$ is the query for Google Trend number *n* in month *t*

$\{a, b, c, \ldots, s, t, u\}$ are trainable parameters

Parameters:[-0.09893265, 0.0721132, 0.10281454, 0.17714237, 0.17040868, 0.1095406, 0.11416813, -0.07777901, 0.08863351, -0.31822784, -0.24101754, 0.63261328, 0.12249, 0.02644896, 0.21827038, -0.07727595, 0.15443251, 0.07632416, -0.09349702, 0.0066614, 0.06909871]

Results

Two metrics were used to measure the accuracy of the model - mean squared error and mean absolute percent error. These metrics were evaluated on the train and test data.

	Mean Squared Error	Mean Absolute Percent Error
Train	20791118.268	5.616 %
Test	21208389.918	6.574 %

Table 1: Results - Mean Squared Error and Percent Error of train and test data

Below is a plot showing the predicted volume against the actual volume for the test data. One important factor to notice is that the model is sensitive, meaning that it can capture microtends, or "spikes" in the data, very well.

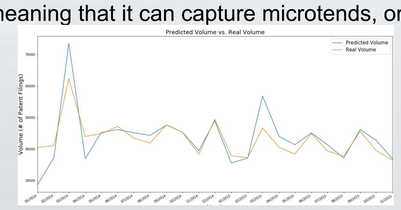

Figure 2: Predicted Volume vs. Real Volume

Results

Rank	Standardized Coefficient	Variable
1	0.632613	x_{t-12}
2	0.318228	x_{t-10}
3	0.241018	x_{t-11}
4	0.218270	"USPTO"
5	0.177142	x_{t-4}

Top Five Most Correlated Variables

Rank	Standardized Coefficient	Trend
4	0.218270	USPTO
6	0.154433	Patent Application Search
7	0.122490	Cost Of Patent
12	0.093497	PatentsView
15	0.077276	Patent Application Process

Top Five Most Correlated Trends

Discussion

This seasonal autoregressive model used for this project achieved optimal results. It received a 6.57% testing percent error on minimal data and was sensitive to changes in the data. The training data's mean squared error and percent error is slightly less than the test data's, indicating minor overfit.The model can be further improved by selecting higher quality Google Trends queries and using more granular data, resulting in a larger amount of data and possibly a higher accuracy model. Other machine learning or deep learning models can be explored such as Recurrent Neural Networks or Long Short-Term Memory networks, however these types of models usually require millions of data points. Because this model is optimal for finding microtrends, this model can be augmented with USPTO's current model to provide more accurate forecasting in the future.

Acknowledgements

This project was created as a part of the Student Innovation Challenge externship in the USPTO created by Mrs. Maria Conti. Special thanks to Mrs. Conti and others for mentorship.

Citations

Choi, H., & Varian, H. (10, April 2009). Predicting the Present with Google Trends. Retrieved August 16, 2018

Hingley, P., & Nicolas, M. (2006). *Forecasting Innovations: Methods for Predicting Numbers of Patent Filings*. Berlin: Springer.

Office of the Chief Financial Officer. (2018, August 6). USPTO Fee Schedule. Retrieved August 16, 2018

Code Repository: https://github.com/vishnurmurthy/uspto/

Python Packages Used: NumPy, SciPy, Pandas