Analysis on US Patents

Vinayak Gaur, Ashish Solanki, Rakshit Makkar, Devashish Nayak
Department of Information Systems, California State University, Los Angeles

Abstract

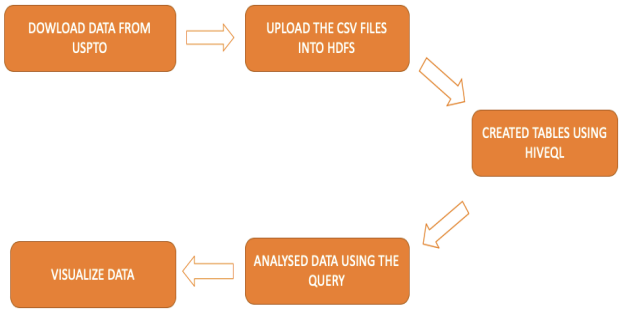
This project examines the patents assigned by USPTO in United States. The USPTO stands for United States Patent and Trademark Organization which assigns the patent not only in US but throughout the world in several fields. USPTO is releasing a series of datasets in formats convenient for our project, the database is roughly of 5 GB uncompressed with 6 million assignments and other transactions recorded during the 1970- 2017 period and affecting about 10 million patents or patent applications.
Throughout this project our team worked to answer several questions such as How patents are assigned? How USPTO categorized the patents? How much time would the take to assign a patent? and so on. To work on this large data set we have used Hadoop File system with capacity of 147 GB to store our data and Hive query language for data engineering. Other than this, we have also used the Microsoft Power BI for visualization of our data to make our analysis easier to analyze.

1. Introduction

Numerous studies have relied upon patent data as a valuable resource for analyzing technological innovation. While patent-related metrics have been used in published research since the 1950s (Schmookler 1954; 1957), the linking of patented inventions to the organizations that own them has proved particularly vexing for our research.
To have in depth analysis and to answer all of the questions we have used six different datasets and all of them describes the different thing such as the Assignment data set describes all of the details regarding the patents assigned since 1970 , assignor data tells about the details of the assignors including their contact details. The table assignee reflects the dates on which the patent is assigned and who assigned that and on what basis. Likewise, the Document_id and Documentid_admin tells the details about the patents such as the patents the date of reference, reel no , the title of the patent and the subject of the title and so on.
We have created a cluster in Oracle cloud, and used tools like Hadoop, Hive, Pig, Excel, Microsoft Power BI to perform analysis and draw insights into these two datasets.

2. Project Implementation

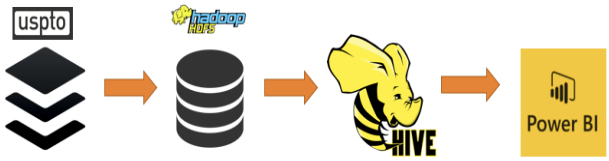
The figure below mentioned briefly describes the implementation of this project. To reach to our end visual we have been through various steps before that as following:
First of all we have downloaded the data from the USPTO website, the data downloaded is in the csv (comma separated version) form, and we directly uploaded that to the Hadoop file system by creating cluster in oracle cloud. After that we Have used several hive queries to process that data. Lastly, we downloaded the desired result into our local system and uploaded to that in Microsoft Power BI to visualize our data.



3. Hardware Specifications

- Cluster Type- Oracle Classic Compute Edition
- Cluster Version-2.7.1.2.4.2.0-258
- Number of Nods- 5 Nodes
- Memory Size- 150 GB
- Storage- 678 GB
- HDFS Capacity- 147 GB
- CPU Speed- 2.2 GHz
- Data Size – 12.2 GB

4. Technical Architecture



The technical architecture of this whole project would be first of all using the USPTO data base to download the appropriate data for our project after that we have upload that data into the HDFS and after wards in cluster that is created at Oracle cloud. Hive query language has been used to sorting and cleaning the data. For creating the visuals we have used the Microsoft Power BI

5. Data Description

As for the part of data description the whole data set is comprises of the six different file having data of different kind. Each of the data set has been described as follow:

5.1. Assignment Dataset

Basically, this dataset gives the full description of the patent. The assignment data file contains a single entry for each of the 6,328,178 transactions recorded at the USPTO between January 1970 and December 2014 (inclusive). While the earliest recording date in the data file is January 4, 1970, the number of transactions recorded in the initial years is negligible. It appears the data coverage in the Dataset is sufficient for time series analysis during 1981-2014. The assignment data file includes the recording date, a page count, and the correspondent name and address, typically reflecting the assignor's power of attorney or legal representative. The correspondent name and address fields consist of free-form text strings rather than distinct fields for street, city, state, etc.

5.2. Assignment_Conveyance

We made a reasonable attempt to assign all entries to the enumerated conveyance type categories, using key search terms such as "assignor's interest," "government interest," or "merger" for pattern matching. In a few cases, we can identify an employer assignment through a keyword search (for example, the conveyance text specifies "employment agreement"). In most instances however, the conveyance text associated with a (presumable) employer assignment is indistinguishable from that of an inter-firm assignment (or reassignment).

5.3. Assignee

The assignee file contains data captured for the assignee(s) for each rf_id in assignment. It includes the assignee's name and address (street, city, state/country, postal code). Data coverage of assignee street, city, and state/country improve for transactions recorded in the early 1990s to near 100 percent for post1996 recordings. Note that typically only the state field is populated for assignees with US addresses, whereas only the country field is populated for assignees with foreign addresses.

5.4. Assignor

the assignor data file contains data recorded for the assignor(s) for each rf_id in assignment. It includes only the assignor name as assignor addresses are not collected. A potential, though highly imperfect, substitute for assignor address may be the correspondence address in the assignment file. In addition to the assignor name, the assignor file contains date fields capturing the execution date (exec_dt; the date that the transaction actually took place) and the acknowledgement date (ack_dt; the date of signature of acknowledgement).

5.5. Document ID

The documentid data file contains identification data for the patent(s) and/or application(s) conveyed in each transaction. It contains the application number (appno_doc_num) of each property as well as other applicable identifiers (Pre-Grant Publication number, patent number), patent title, and relevant prosecution dates (application filing date, pre-grant publication date, patent grant date).

5.6. Documentid_Admin

In documentid_admin we attempted to identify such inaccuracies by matching application numbers (appno_doc_num) and patent numbers (grant_doc_num) to internal USPTO administrative data, writing the results to variables in documentid_admin. For each patent number in the assignment records, we extracted the corresponding application number from the administrative data (admin_appl_id_for_grant), and for each application number in the assignment records we extracted the administrative patent number (admin_pat_no_for_appno).

6. Visualizations

6.1. Average Estimated Time for a Patent

The first thing ever came to everyone's mind is how much time would it take to get a patent. So, we have taken this as our first question for analysis and for this we have used assignment table as this is the only table having the date on which the individual or the company have put their application in USPTO and the date on which the patent has been assigned to them. After the visualization a very impressive result came out as the estimated time to get a patent has dropped drastically since 1970's as the average time to get a patent in 1978 was almost 23 years and rose to 27 years by 1980, but dropped as nearer to 21st century, where the average estimated time to achieve a patent was no more than 7.5 years.

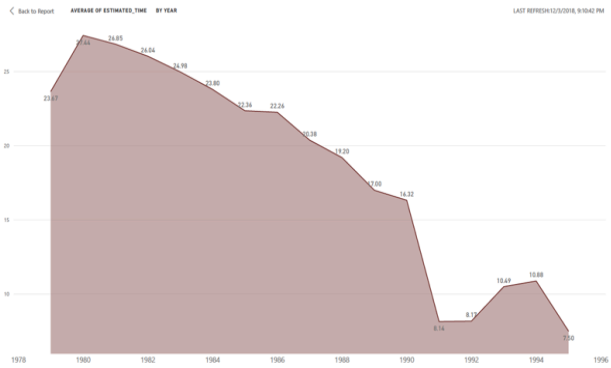


Fig 6.1. Average Estimated Time by year

6.2. Geographic Visualization

Moving forward the second thing we need to resolve in our analysis is the distribution of the patents throughout the world, as this help us to know which are the nations more actively innovating and gets the patents for their innovation. For this analysis we have used document_id table. As this visualization clearly shows that except the Russia and Asian countries the patents are distributed throughout the world, even though the Japan is the world leader in assigning the patents as by the end of 2017 the total number of 206103 patents are assigned to the Japan followed by the Germany and France with 69611 and 28380 patents.

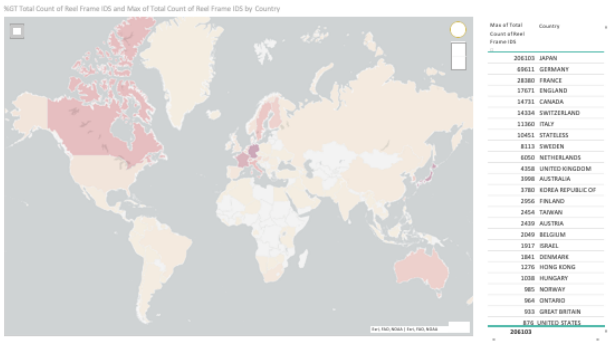


Fig 6.2. Geographical View of patents assigned throughout the world

6.3. Patents Based on Government Interest

The next big question we need to resolve is to find how the data has been categorized by the USPTO. So we came up with the answer that the USPTO categorized the patents on the basis of the interest of the people who appeal for the patent. The patents are basically categorized into the government interest, merger interest, security and so on. The first visualization is on the patents based on the government interest, as this will lead us to the point where we can find which nation's government is more active in applying new innovative ideas. As the visualization results into that the Canadian government have the 20% of the total patents assigned on government interest followed by Japanese and German Government.

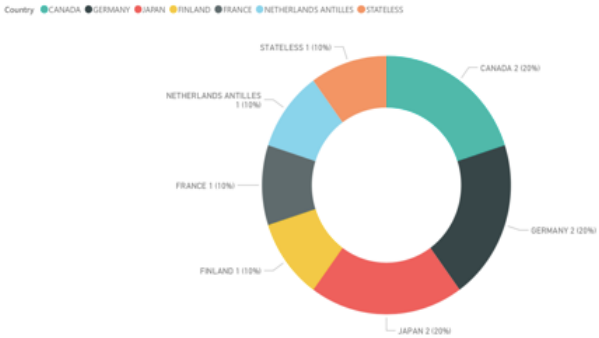


Fig 6.3. Patents Based on Government Interest

6.4. Patents Based on Security

Moving forward, the second category on which the patents are categorized are the patent based on Security Purpose. As large number of countries are indulge in applying latest innovative ways for security. From this visualization we can clearly see that the small countries are rapidly moving towards innovative ways to secure their boundaries as total number of 262 patents are till now has been allocated to them followed by Canada with 72 patents.

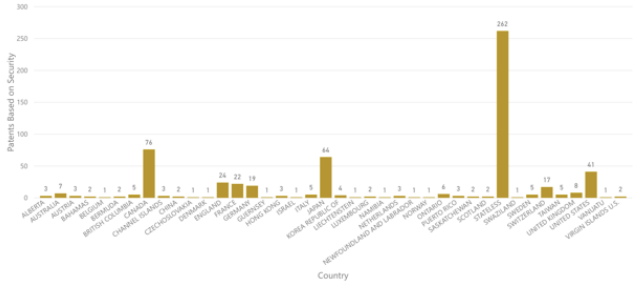


Fig. 6.4. Patents Based on Security

6.5. Patents Based on Merger

While exploring the data base we have came up to a new finding that some of the patents are assigned to the two or more merger of companies, As in this visual we still get a very unexpected result as the small countries like Japan and Italy are on the top positions while considering the patent based on merger with total number of 86 and 52 patents and after that nobody came out with the half of it till 2017.

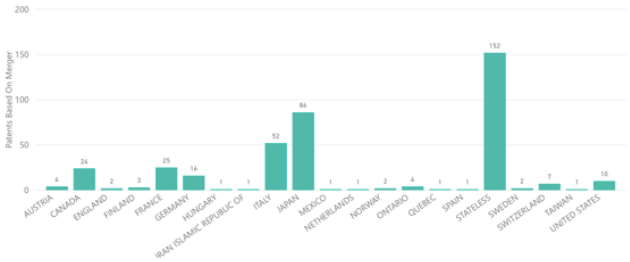
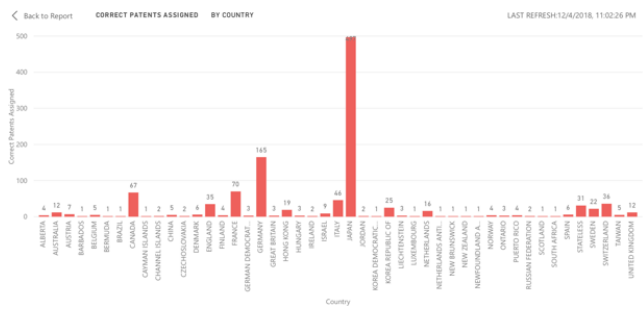


Fig. 6.5. Patent Based on Merger

6.6. Correct Patent Assigned

The next big task for our team was to separate the correct patent assigned from the total number of patents as the list comprises of correct and some incorrect patents, so to know the total number of right patents is as important as any other task. So after analysis e came out with the surprising result that the Japan is the world leader in innovation and in assigning patents with total number of 697 patents assigned till the end of year 2017 followed by Germany and Canada but no one came out with half of that number.



6.7. Dashboard

To show the relationship between the different categories on which the patents assigned throughout the world we have made a dashboard, which simply shows that the in three out of the four categories the smaller countries like Japan, Italy and Germany are able to get more patents as compared to other countries regardless of their size.

Fig. 6.7. Dashboard

6.8. Machine Learning and Artificial Intelligence related patents assigned

Furthermore, we have move towards more specific type of patents assigned, as the next question we try to resolve that is to find the number of patents assigned in the fields related to Machine Learning and Artificial Intelligence and we came up with the finding that almost 425 patents assigned on Neural Networking and 95 on Data Mining. As the fields like Machine Learning and Artificial are also emerging as high number of patents are assigned on them too such as 55 and 45 respectively.

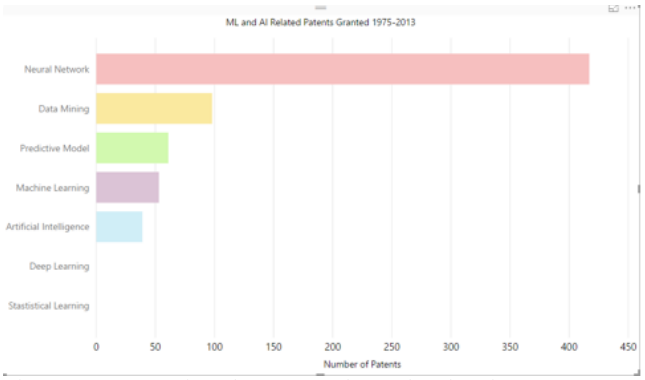


Fig. 6.8. Patent assigned on ML and AI related topics

6.9. Countries where Machine Learning and Artificial Intelligence Based Patents Invented

The continuing next step in this analysis is to find the distribution of patents related to Machine Learning and Artificial Intelligence throughout the world, for this we have used geographical map to display the distribution. So the visualization clearly shows that the most of the patents assigned were in the North America and followed by some of them in Western European Countries.

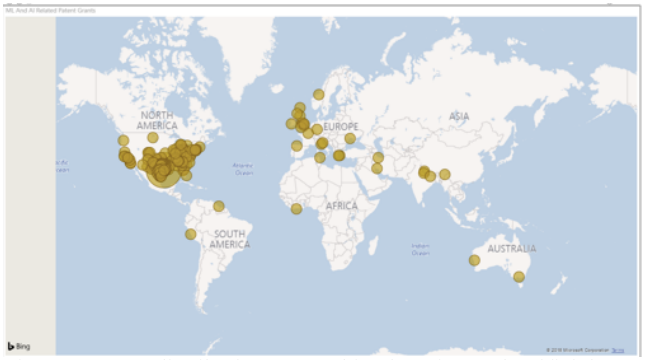


Fig. 6.9. Patent distribution on machine learning and artificial intelligence types of topics.

6.10. Patents and Inventions by Countries

Afterwards we have broaden our view to list the countries leading in the patents and innovations in the technology field. The visualization clearly concluded that the North America and Canada are the world leaders in the field of technology as the largest number of patents are assigned to them followed by the developing countries like India and china. However the rest of the world has barely contributed to the inventions in this particular field.

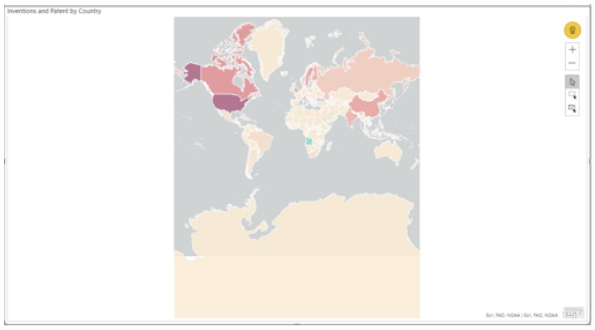


Fig. 6.10. Patents and Inventions by Countries

6.11. Number of Patents Owned by Market Giants.

The last analysis we have selected is to find the patents undertaken by the market giants in the field of Information Technology, the giants like Amazon, Facebook, Google and Samsung. All of them are the market leader in their respective businesses but when came up to the patents term the world biggest mobile manufacturer Samsung has the most number of patents with 36070 patents followed by Carrier company AT&T and Google Inc. with 17320 and 2384 patents.

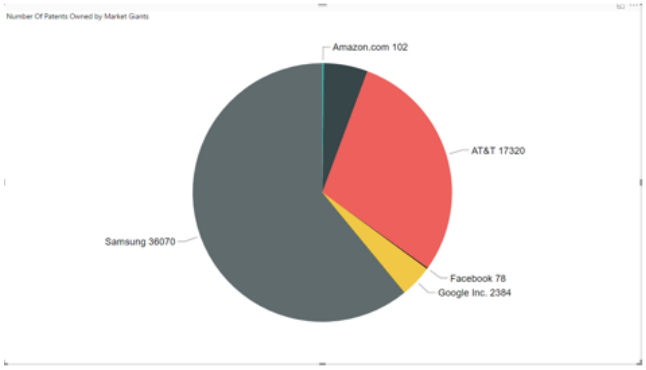


Fig. 6.11. Number of Patents owned by Market Giants.

6.12. Factors Behind Rejection of Patents

Since the beginning we have talked a lot about the patents that are assigned but we have yet to talk about the patents which are not assigned. What are the basis on which they are rejected? The USPTO factorize them into three basis such as

- Missing Application ID
- Incorrected ID
- Ambiguous

Missing application ID means that the ID was not there when the patent was taken into consideration, so that can't be assigned. Second factor is the incorrect ID, as the ID displayed might be incorrect or mismatched.

7. Limitations

Perhaps, there are couple of limitations that we have faced throughout the project.

- As data has more than 6 million rows which exceeds the limit of excel so we are not able clearly view the data in all of the rows.
- Not-having the in-depth details of the patents assigned, as none of the table shows the whole subject of the patent.
- Duplication of records- as there 6 different data sets we have to take care of the duplicate records.

8. Summary

- We have provided a comprehensive description and presented a set of stylized facts to help motivate future research.
- Despite some limitations in using these data the Dataset will open multiple avenues for original follow-on research.
- Possible areas include the markets for technology and innovation, the relationships of intangible assets to firm financing, and government interest sponsorship to innovation and commercial outcomes.
- We encourage researchers to learn from our efforts and these newly-released data in order to answer important questions related to these topics and others we do not have the prescience to predict.

9. References

Graham, Stuart J.h., et al. "Patent Transactions in the Marketplace: Lessons from the USPTO Patent Assignment Dataset." *SSRN Electronic Journal*, 2017, doi:10.2139/ssrn.2489153.

10. GitHub

https://github.com/asolank5/CIS5200-03- Group4