Creating An Innovation Index From Patent Data

Methodology Overview

Tiffany Schleeter, Ph.D.[1]; Daniel Byler, M.S.[1]; Vamsi Krishna[2]; Venkatesh Gangavarapu[2]

1 Deloitte Services, LP, Arlington, VA; 2 Deloitte Services LP, Hyderabad, India

I. Purpose

This methodology overview is a summary of how Deloitte designed and computed the innovation index for the study 'Patenting Innovation in the Oil and Gas Space'[1]. It is intended for statisticians and data scientists who wish to understand technical details of the calculations and/or replicate or update our results. While it contains some of the results of the overall study, the business implications of our findings are primarily detailed in the research study itself.

II. Background: The purpose of using patent data

Patent data[2] is very useful for analyzing a technology's growth, change, and impact. An individual patent granted by the United States Patent and Trademark Office (USPTO) contains a large amount of information used to examine the progression of knowledge acquisition. Some essential data points that are used to investigate patents are their assigned technology categories, patent assignee information, and citation records. Our patent analysis relies primarily on three discrete pieces of information from patent records[3]:

- Technology classification
- Assignee
- Patent citations

Technology Classifications

The Cooperative Patent Classification (CPC) scheme[4] is a detailed system of assigning a patent to descriptive categories based on what technology they primarily pertain to. Used in 45 patent offices, this system is an international standard and pertains to all granted US patents[5]; as of 2013 it officially replaced the previous patent classification system in the United States (the US Patent Classification System)[6],[7]. Granted patents in the U.S. have been classified under this scheme since 2010. Patents prior to 2010 have been programmatically re-assigned to maintain consistency across a longer time period.

The technology classification system is hierarchical; the first tier contains 8 categories, and the lowest sub-level is comprised of over 200,000 detailed categories. Deloitte's study used the third tier, which comprises almost 1,000 technology categories. Information gathered from a patent classification list allows us to quantify how much a technology field is evolving based on the number of patents they are assigned to annually; we can also identify the field(s) in which a patent holder is specializing in.

Assignee

The assignee of a patent is another valuable source of information that we obtain from this data. To obtain intellectual property, an applicant (or group of applicants) undergoes a lengthy and comprehensive process before they are able to receive patent ownership. Many patent assignees are corporations and research institutions but some patents are granted solely to individuals. Often times, these companies are associated with a particular market or industry.

Patent Citations

Patent citations are extremely detailed and are used heavily in our research; patents can glean techniques from previous patents, combine information from multiple patents, and take previous patent methodologies a step further to produce something worthy of a new patent. Whenever a patent builds on prior technologies, patent applicants have to cite all previous patents they have used as a resource. Knowing the relationships between patents from citation data is a foundation for identifying the relationships between technologies. The more patents demonstrating the same link between two classifications, the more those two technologies are connected to each other; understanding the trend of contributions among technologies, we have another way to evaluate innovation based on intellectual property.

III. Methodology: Patent Analysis Approach

Part 1: Acquiring and managing patent data

Deloitte's analysis for the purposes of the current study, cited above, focused on the 2 million+ patents that were granted from 2006-2015 (our time period of interest) in the United States. Jointly, the technological classifications mentioned before and the citations data make patent data ideal for network analysis[8],[9].

In the patent network, patent classifications are discrete, categorical variables that can be represented by nodes. Patent citations can be seen as quantifiable links between two patents or technologies. From there, various networks can be created with this data. First, we can consider all patents to be a universal network; furthermore, network theory can be applied to subsets of patents based on assignees, industries, or technologies of particular interest. This way we can understand the influence of a technology among others in a network.

Patent classifications represents a social network, where the size of the nodes, the distances between them, and the links among them appear random; however, in actuality, these characteristics are determined mathematically to be graphed this way. Using a package called *igraph*[10] in R Studio Version 99.879, 3 variables that have been gathered from the patent data are used as input variables:

1. The "source" name, s – the description of an object (technology) from the patent that is being cited.

2. The “target” name, t – the description of an object (technology) from the patent that cites other patents that is not the same as the source.

3. The target-source weight, ω – the total number of times a patent classified under “target” cites a patent classified under “source” in a year. In a subnetwork with P patents,

$$\omega = \sum_{i \leq P, s \neq t} (s \rightarrow t)_i$$

Depending on the size (number of objects) in the particular network, ω may be scaled. Frequently, this number will receive a threshold value determined by the user; the threshold is a weight minimum in which data falling below a certain value is considered negligible and does not get included in the resulting network graph.

Part 2: Building the analysis

The software is capable of outputting various results from the input parameters. For our variables of interest, the resulting output of the network theory analysis returns[11],[12]:

1. **Closeness Centrality** measures network centrality of a node x by taking the total number of nodes N in the graph and dividing by the sum of the distances between x and some other node in the network y_i:

$$C(x) = \frac{N}{\sum_{y_i \in N-1} d(y_i, x)}$$

   Small values for $C(x)$ would indicate that x is closer to the center of the network, while high values for $C(x)$would indicate that x is distant from the centre of the network.

2. **Betweenness Centrality** measures network centrality of a node x as the total number of shortest paths from nodes m_i to nodes n_i that pass through x over the total number of shortest paths from all nodes m_i to all nodes n_i:

$$B(x) = \sum_{m_i \neq x \neq n_i}^{i \in N} \frac{\theta_{m_i n_i}(x)}{\theta_{m_i n_i}}$$

   Higher values for $B(x)$ indicates that x has a greater impact over the connection of objects in the network.

3. **Page Rank** is a value assigned to a node in the network based on a probability distribution of the likelihood that the node is randomly selected from the network. Page rank is not a simple calculation as it is computed by an iterative algorithm.

   To compute page rank, we imagine someone attempting to randomly traverse the network after starting on any given node. The value d is the damping factor, a parameter selected by the user, it represents the probability that someone randomly traversing the network will eventually stop at any given step.

To find the page rank of some node x_i in the network, consider the specific set of nodes that connect to it as M. The page rank of x_i is the sum of the page rank of all the nodes in M divided by the number of outbound links L of the node x_j in M that links to x_i.

$$PR(x_i) = \frac{1-d}{N} + d \sum_{x_j \in M(x_i)} \frac{PR(x_j)}{L(x_j)}$$

As the process is iterative, it can be seen that $PR(x_i)$ is determined by the page ranks of other nodes. Those page ranks are assumed to have been calculated at an earlier time step in the algorithm using the same formula therefore, the initial step is the first page rank calculation which is simply:

$$PR(x_0) = \frac{1}{N}$$

Once the page ranks of all the nodes in the network are known, the values can be interpreted as the importance value of that node, ordered largest to smallest.

The two measures of centrality determine where a node gets placed in the network based on their already known connection. Page rank, on the other hand, measures significance. Although one technology may connect with many others, indicating good centrality, the number of patents produced on it may be few, which means a low page rank. Conversely, few connections with high frequencies of contributions suggests greater prominence of a field or industry. Ultimately, the page rank of a technology corresponds to the size of its node in the graph. The target-source weight, ω, not only used for volume of citations, is also the measure of strength and direction of the link between the two referenced technologies in the graph.

Other researchers may choose to substitute other measures of centrality for page rank (including the ones mentioned).

However, this work uses page rank because not only does it measure importance quantitatively, but it assesses the quality of the nodes in the network. We also tested closeness centrality and eigenvector centrality and found the same overall result as when we used page rank. This makes us confident we are observing a real phenomena and not a fluke due to methodological choices.

IV. Apply to an Industry Subspace

The most efficient way to analyze patent activity in a particular subspace is to simplify the data in the whole patent universe first. A researcher could determine the subgroup in 3 different ways—listing the specific patents they want to examine (by the identifying number); knowing the names of specific companies, corporations, and subsidiaries as assignees; or selecting technologies at the third tier level which they want to investigate. Using these filters creates a subset of cited patents in the subspace that generate new values for ω. Once this is computed, the same centrality methods from above can be reused.

In our approach, Deloitte gathered patents owned by selected companies that were representative of the oil and gas industry. To assess this industry's contribution to innovation overall, Deloitte compared their growth in patents to the overall universe. A percent change in an industry's patent activity that is not the same to that of the universe indicates a faster or slower rate of contribution. More specifically, we can look into patent activity by technology to understand the direction an industry is moving to or away from. Again, the patent activity of a technology for a specific industry can be compared with the universe to gauge contribution to innovation.

To quantify the contribution of an industry overall, we created a variable we call the "innovation index" which is based on patent activity over time. We consider the proportion, p, of all patents belonging to a specified industry, denoted by b, with respect to the universe which consists of N technologies, $X = \{x_1, x_2, \cdots, x_N\}$. Proportion takes into consideration the change of technology influence in the universe from year to year so it is more effective than just volume. We calculate the industry measure of an individual technology by taking the product, π, of its universal page rank, PR, with p. Summarizing this for all N technologies, we find the value for the innovation index, I, for some given year, t:

$$I_{b,t} = \sum_{i=1}^{N} p_b(x_i)PR(x_i) = \sum_{i=1}^{N} \pi(x_i)$$

The greater the value of I, the more an industry is considered to be influential to innovation. Frequently, an industry will not have patents granted in all technology areas; in those cases, the value of π would be 0, obstructing a potential increase of the innovation index. This makes sense as fewer connections means weaker centrality, therefore having less significance. Because patent activity will change from year to year, I_t can be normalized by dividing it by the total page rank of the universe in t to weight I_b over time.

Many industries are characterized by select technologies. Citations to and from industry patents allow for other patent classifications to grow into or noticeably fade out of a network over time. The weight of these relationships and how they change can reveal information about knowledge flow evolves in a patent subspace. When identifying relationships between technologies, it is often simpler to visualize them rather than read numbers from a table. D3, a JavaScript library, can produce vibrant, user-friendly illustrations in a web browser showing relationships in images whose sizes correspond to volume.

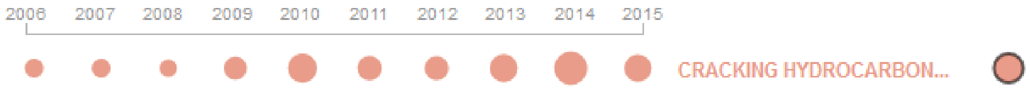

Figure 1: The relationship between "Fuels Not Otherwise Provided For" and "Cracking Hydrocarbon Oils from 2006-2015.

A dynamic visualization of universal network or subnetworks can also be created in D3[13] once the locations, sizes, and links are known. Network graphs should be generated for each year of data to observe noticeable changes over time. A rotation and translation can be applied to a graph of an earlier year to better align to its future counterpart.

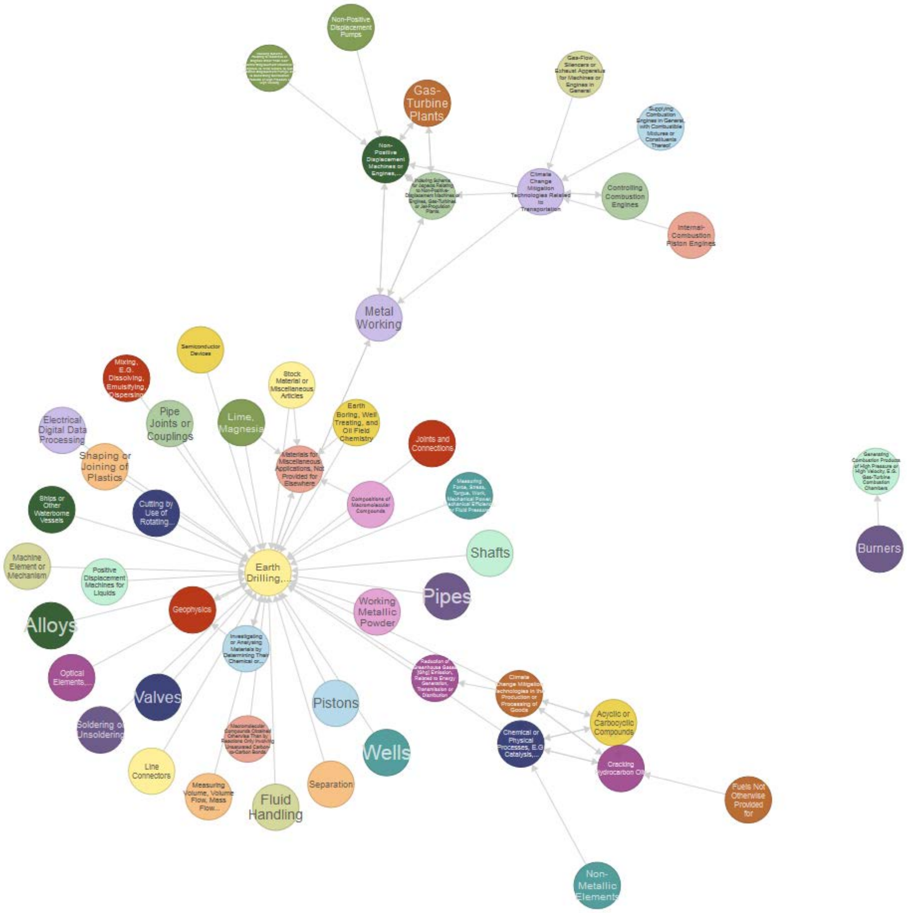
Non-Positive Displacement Pumps
Gas-Turbine Plants
Gas-Flow Silencers or Exhaust Apparatus for Machines or Engines in General
Supplying Combustion Engines in General, with Combustible Mixtures or Constituents Thereof
Non-Positive Displacement Machines or Engines,...
Indexing Scheme for Aspects Relating to Non-Positive-Displacement Machines or Engines, Gas-Turbines or Jet-Propulsion Plants
Climate Change Mitigation Technologies Related to Transportation
Controlling Combustion Engines
Internal-Combustion Piston Engines
Metal Working
Semiconductor Devices
Mixing, E.G. Dissolving, Emulsifying, Dispersing
Stock Material or Miscellaneous Articles
Electrical Digital Data Processing
Pipe Joints or Couplings
Lime, Magnesia
Earth Boring, Well Treating, and Oil Field Chemistry
Joints and Connections
Shaping or Joining of Plastics
Materials for Miscellaneous Applications, Not Provided for Elsewhere
Ships or Other Waterborne Vessels
Cutting by Use of Rotating...
Compositions of Macromolecular Compounds
Measuring Force, Stress, Torque, Work, Mechanical Power, Mechanical Efficiency, or Fluid Pressure
Generating Combustion Products of High Pressure or High Velocity, E.G. Gas-Turbine Combustion Chambers
Machine Element or Mechanism
Positive Displacement Machines for Liquids
Earth Drilling,...
Shafts
Burners
Geophysics
Pipes
Alloys
Working Metallic Powder
Investigating or Analysing Materials by Determining Their Chemical or...
Optical Elements,...
Valves
Reduction of Greenhouse Gases [Ghg] Emission, Related to Energy Generation, Transmission or Distribution
Pistons
Climate Change Mitigation Technologies in the Production or Processing of Goods
Acyclic or Carbocyclic Compounds
Soldering or Unsoldering
Macromolecular Compounds Obtained Otherwise Than by Reactions Only Involving Unsaturated Carbon-to-Carbon Bonds
Wells
Chemical or Physical Processes, E.G. Catalysis,...
Cracking Hydrocarbon Oils
Line Connectors
Measuring Volume, Volume Flow, Mass Flow...
Fluid Handling
Separation
Fuels Not Otherwise Provided for
Non-Metallic Elements

Figure 2: Example illustration of a patent sub-network in D3

References

[1] (Authors). 2017 February 14. Patenting Innovation in Oil and Gas. Retrieved from (url).

[2] Reed Tech, USPTO Data Sets. 2016. Retrieved from Reed Tech.

[3] Cooperative Patent Classification. United States Patent and Trademark Office. 2013.

[4] CPC scheme and CPC definitions. United States Patent and Trademark Office. 2013.

[5] "3rd EPO-USPTO CPC Annual Meeting with national offices classifying in CPC" (PDF). European Patent Office. 2016 February 26.

[6] European Patent Office. 2013 June 4. Europe and China agree to use same patent classification system (CPC) [Press Release].

[7] United States Patent and Trademark Office. 2013 June 5. USPTO and KIPO Announce Launch of Cooperative Patent Classification System Pilot [Press Release].

[8] Wasserman, Stanley, and K. Faust. 1994. Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press, Cambridge, UK.

[9] Èrdi, Péter et. al. 2013. Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics: Volume 95, Issue 1, Pages 225-242.

[10] R Core Team. 2015. R: Network Analysis and Visualization. R Foundation for Statistical Computing, Vienna, Austria. Package: 'igraph'.

[11] Borgatti, Stephen. 2005. Centrality and Network Flow. Social Networks: Volume 27, Issue 1, Pages 55-71.

[12] Page, Larry and S. Brin. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report, Stanford InfoLab.

[13] Bostock, Michael, J. Heer, and V. Ogienetsky. 2016. D3: Data-Driven Documents Version 4.2.0 [computer software]. Stanford, California, USA.