SNAP Patent Citation Network: Citation Locality and Recommendations

Feras Ahmed*, Vincent Barila†, Hayley Bly‡ and Miguel Gavidia§ University of Miami

Email: *fahmed@students.law.miami.edu, †v.barila@umiami.edu, ‡h.bly@umiami.edu, §mgavidia@law.miami.edu

***Abstract*—We are proposing a model for recommending *citations* that may have prior art in cases where patents may have some overlap. With the observance of patent *citation* locality in various industries over a given time period the distribution can suggest the probability of related patents. With this collective analysis of the Stanford Network Analysis Project (SNAP) patent *citation* network of the National Bureau of Economic Research (NBER) patent data our goal is to discover any commonalities that may assist in the filing of new patents.**

I. Introduction

We will demonstrate a novel way of scoring the *citation* data on existing patents that have increased significantly over the years. Figure 1 is an overview of the patents awarded from 1975 to 1995. Utilizing Apache Hadoop as our main analysis tool for the model [1], we will index and rank individual patents based on the year and the industry and build an intelligent means to weight their locality and related patents.

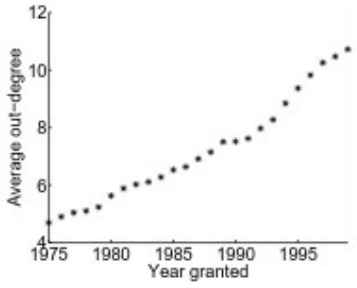

Fig. 1. **The average node out-degree over time of patents between 1975 and 1999.** [2]

II. Background and Significance

Patents and patent *citation* are among the most useful indicators in forecasting new technologies. The patent *citation* network describes the knowledge transfer process between technical fields, institutions, and countries. The NBER maintains the U.S. patent database. This database contains 37 years (January 1, 1963 to December 30,1999) of data. The data includes all the utility patents granted during that 37 year period, totalling 3,923,922 patents. The patent *citation* network is composed of all the citations made by patents granted between 1975, and 1999, totalling 16,522, 438 citations.

The significance of our research is in identifying trends among the major players, industries and sub-industries, and the relations between them. The more frequently a patent is cited by subsequent patents, the more diffused and available the technology can be said to be. Thus, patent *citation* data can be used to not only describe the current state of affairs in a particular industry or time, but also to predict reliably and accurately new and emerging technologies. This sort of predictive data is very valuable to not only attorneys and law firms, but also businesses and consulting firms.

III. Related Work

A survey of related work uncovers that patent citation analysis has been used for a number of purposes by both corporations and academics. Most closely related to our proposed objective, corporate researchers have concentrated their research on environmental scanning such as bibliometric/patent trend analysis [3] and market analysis to identify increasingly diversified needs of customers [4]. Similarly, academic researchers, have used citations analysis to assess research implementation and performance. [5], [6]. Researchers have also utilized patents as proxies for innovation in efforts to create and test hypothetical models for the process of technological development. For instance, Fleming [7] built a model of technological development in which innovation advances principally by a process of looking for new synergies between existing technological advances. Researchers

are also using patent citation analysis to explore ideas taken from present network science and social network studies to enlighten the structure of technological innovation. For instance, Weng et al. [8], tested the idea of structural equivalence, a key notion in the traditional hypothesis of social networks. Researchers have also measured the separation between patent citations as characterized by the shortest distance between points in the patent reference system. For instance, Lee et al. [9], examined a subset of the patent citation network system to explore the instance of electrical conducting polymer nanocomposites.

IV. Research Design and Methods

Using the *citation* network nodes to reference the NBER patent data [10] we will use MapReduce to create lookup information for specific industries and the given time the patent was filed so that we can analyze some generalities between the citations. Using k-Nearest Neighbors algorithm (k-NN) we intend to weight the total of citations in the indexed period by year and month to a single patent as a basis for our scoring mechanism. Clustering the data by these periods can give us temporal data that will be useful in determining locality. Figure 2 demonstrates two dendrograms representing the results of the hierarchical Ward clustering of patents of a particular subcategory between 18,833 patents on Jan. 1, 1994 (Graph A), and 25,624 patents on Dec. 31, 1999 (Graph B). Based on prior work on scoring recommendations [11], we will attempt to create a recommendation network for a given patent dependent on their citations and industry.

V. Preliminary Suppositions and Implications

There are inherent issues when trying to realize a traditional benchmarking [10] of the *citation* data. In particular, substantial temporal advancement in patenting rates and thus the numerical increase in citations Figure 1, as well as the unavoidable loss of some data due to the highly-populated nature of the base, make quite difficult the appropriate and meaningful direct use of the mere raw *citation* occurrence data received by different patents. There are two NBER-proposed approaches to rectify this matter. The first of which is analysis of *fixed-effects*, requiring the utilization of the average citation count for a group of patents relevant to the particular one in question for appropriate comparison; the second is to distinguish the multiple effects on citation rates via econometric estimation. We will be using *fixed-effects* in our model.

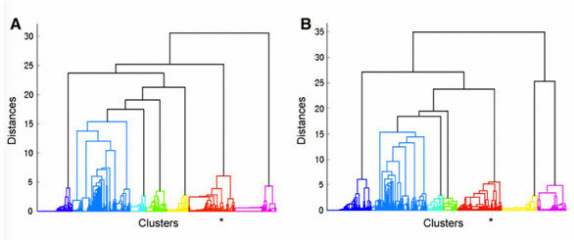

Fig. 2. **Temporal changes in the cluster structure of the patent system.** Dendrograms representing the results of the hierarchical Ward clustering of patents in subcategory 11, based on their *citation* vector similarity on Jan. 1, 1994 (18833 patents in graph a) and Dec. 31, 1999 (25624 in graph b). The x axis denotes a list of patents in subcategory 11, while the distances between them, as defined by the *citation* vector similarity, are dendrogram correspond to the 7 most widely separated clusters. While the overall structure is similar in 1994 and 1999, interesting structural changes emerged in this period. The cluster marked with the red color and asterisk approximately corresponds to the new class 442, which was established in 1997, but was clearly identifiable by our clustering algorithm as early as 1991. [12]

VI. Data Modeling

A. k-Nearest Neighbors Algorithm

k-NN is a non-parametric method used for classification and regression. Being non-parametric is useful as there is no underline assumption to the similarities of patents in the dataset. The learning is therefore lazy and approximations of locality and computation can be deferred until classification. As such we assign the closest examples in the space based off patent *fixed-effects* and distances between *sub-categories* as the weight. We can therefore forgo any training phase and test the data directly. Neighbors are then extracted from the set of patents via their parametric weight to contribute to the overall average distance between each. The neighbors are then elected based on a threshold of distance (i.e. $<.5\%$) and a maximum of n.

B. Data cleansing

In order to present how the model is created using the NBER patent data in conjunction with the SNAP *citation* data we have to prune it of erroneous entries. Further we need to ensure that the data accurately reflects one another. That is, if for each line in the SNAP *citation* data β, where P_1 is the *from* patent and P_2 is the *cited* patent, we verify if P_1 or P_2 exists in the NBER data δ. We then discard any patent that doesn't match. Next, we tabulate the sums of each matching patent in each column for verification from the existing numbers recorded in the

NBER data file. This is the *Number of citations made* P_1, and *Number of citations received* P_2. [10]

$$\text{Cites} \quad \leftarrow \quad \sum_{i=0}^{n} \begin{cases} P_{1i}, & \text{if } P_{1i} \in \delta \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$\text{Cited-by} \quad \leftarrow \quad \sum_{i=0}^{n} \begin{cases} P_{2i}, & \text{if } P_{2i} \in \delta \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

C. Similarity

Once the data has been cleaned, we then need to create a measurement of similarity in relation between a given patent and the *citations* that it has. This is what we label as a patents *similarity-index*. Thus, for each *citation* C we calculate the *Euclidean Distance* between the *sub-categories* denoted by α of patent P.

$$\text{Similarity-index} \quad \leftarrow \quad \sqrt{\textstyle\sum_{i=0}^{n}(P_{\alpha i} - C_{\alpha i})^2} \tag{3}$$

D. Fixed-effects

We use the NBER method of *fixed-effects* to minimize temporal differences between patents belonging to separate cohorts, see table I. That is the ratio of a patents *citations* C divided by the count of *citations* P in a cohort to the total amount of *citations* Z in all cohorts.

$$A \quad \leftarrow \quad \frac{\sum_{i=0}^{n} P_i}{\sum_{i=0}^{n} Z_i} \tag{4}$$

$$\text{Fixed-effect} \quad \leftarrow \quad \frac{\sum_{i=0}^{n} C_i}{A} \tag{5}$$

TABLE I
COHORT FIXED-EFFECT PERIODS

Period	Citations	A
1975	3458091	4.77
1985	3465102	4.76
1991	9593077	1.72

E. Distances

Combining our measurements to offset, or produce weight, we can now approximate the nearest neighbors for k-NN. This is done with the set of values obtained for the *similarity-index* σ, *fixed-effects* β and the subcategory ν using *Euclidean Distance* to gauge distance between two patents within the same cohort. This provides us with our k-NN map of neighbors for each patent in the entire dataset.

$$\text{Distance} \quad \leftarrow \quad \sqrt{\begin{matrix} (P_{1\sigma i} - P_{2\sigma j})^2 & + \\ (P_{1\beta i} - P_{2\beta j})^2 & + \\ (P_{1\nu i} - P_{2\nu j})^2 & \end{matrix}} \tag{6}$$

F. Scoring Neighbors

Finally, we set a threshold μ (i.e $<.5\%$) to compare the set of patents *distances*P_{2a} in the dataset to a given patents *similarity-index* σ and discard any that exceed μ. We then choose k patents from the new set whose distances are the closest to the given patents *similarity-index*.

$$P_{1k} \quad \leftarrow \quad \begin{cases} P_2, & \text{if } P_{2a} < \mu \text{ and } P_{2a} \approx P_{1\sigma} \\ \varnothing, & \text{otherwise} \end{cases} \tag{7}$$

VII. Implementation

In order to streamline the implementation of the model, we decided to use *Python*. The language is convenient as it allows almost immediate results with less lines of code. Algorithms 1, 2 and 3 sufficiently explain the prototyped environment.

A. Prototype

Algorithm 1 Euclidean Distance

procedure EUCLIDEANDISTANCE(A,B,$size$)
    set d = 0
    for each $i \in$ range($size$) **do**
        d += $(A[i] - B[i])^2$
    end for
    return $\sqrt{P}$
end procedure

Algorithm 2 Read Patent File

procedure PATENTREADER
    F = Open File
    P = Initialize Dictionary
    for each line $i \in F$ **do**
        let p be a new *Patent*
        Parse columns from i into p
        set cohort for p
        calculate *fixed-effects* for p
        calculate *similarity-index* for p
        insert p to P by patent number
    end for
    return P
end procedure

Algorithm 3 Get Neighbors

procedure GETNEIGHBORS(p, A, $size$)
    L = Initialize Array
    Extract C where $\forall k \in A_{cohort} = p_{cohort}$
    for each $i \in C$ **do**
        set t = $(p_\alpha, p_\beta, p_\sigma)$
        set u = $(C[j]_\alpha, C[j]_\beta, C[j]_\sigma)$
        append EuclideanDistance(t, u, 3) to L
    end for
    sort L ascending
    return L[0 .. $size$]
end procedure

VIII. Experimental Results

We were able to quickly find the model producing transient recommendations to patents within the category or adjacent sub-categories while observing the output of the *GetNeighbors* algorithm 3. However, due to the sheer size of patent citations, 16518952 in this case, our runtime is exponential on the order of $\Theta(n^2)$. Our last run was for about 6 days before crashing our test system, giving us an *Out of Memory* error. System specifications were on an Intel i7, with 12 Gigabytes of RAM. This is to say, there is not sufficient amount of time to retrieve a reasonable set of response data to gauge the effectiveness of the recommendation system.

IX. Conclusion

Our goal was to create a model for recommending citations that may have prior relevance to a new patent in question in cases of potential overlap in between them. The patent data analyzed can be used not only to analyze the past or current state of affairs within an industry or time period, but also to predict emerging or significantly developing technologies. Referencing NBER patent data, we used MapReduce to create lookup information for specific lookup information for particular industries and the time of patent filing to analyze inter-citation generalities. Clustering data by period, we used the k-NN to weight total citations of a given month and year to a single patent as a scoring mechanism. We cross-checked individual data points from the SNAP citation network with information from the NBER records to ensure that our data was cleansed of erroneous entries and then implemented a similarity index to find potentially desirable relevant patents. Utilizing Python to process the algorithms of our proposed model, we ran our code on 16518952 citations and began to produce results that confirmed our expectations, but the immense size of our data collection resulted in a crashing emulator after multiple days of running. Unfortunately we do not have a favorable amount of data to discuss results in depth given the time remaining to complete the project, but we managed to pinpoint reasonable overlap between related patents and note connections between particular pairs of industries whose citation overlap increased as a given industry pair became more noteworthy with the emergence of new uses for such patents.

X. Future Work

We would like to extend this work into the Apache Hadoop space in the future. Hopefully, with utilities like *Hive* and/or *Impala* we may be able to create a

data structure that fits the proposed model so that we may query the system consistently and achieve a much more precise outcome. With a flattened model existent on HDFS it should be thus possible to devise different metrics that extend past recommendation systems. Further, with such a refined system in place we should be able to extrapolate additional data from the United States Patent and Trademark Office directly, post 1999, testing and training the model more extensively. This should bring the resultant outcomes more realistically closer to use in practice than what is actually in the NBER dataset.

REFERENCES

[1] V. Narayanan and M. Bhandarkar, "Modeling with hadoop," in *Proceedings of the 17th ACM SIGKDD International Conference Tutorials*, KDD '11 Tutorials, (New York, NY, USA), pp. 2:1–2:1, ACM, 2011.

[2] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187, ACM, 2005.

[3] J. P. Martino, "A review of selected recent advances in technological forecasting," *Technological Forecasting and Social Change*, vol. 70, no. 8, pp. 719–733, 2003.

[4] L. Fahey, W. R. King, and V. K. Narayanan, "Environmental scanning and forecasting in strategic planningthe state of the art," *Long Range Planning*, vol. 14, no. 1, pp. 32–39, 1981.

[5] E. Garfield and R. K. Merton, *Citation indexing: Its theory and application in science, technology, and humanities*, vol. 8. Wiley New York, 1979.

[6] H. F. Moed, *Citation analysis in research evaluation*, vol. 9. Springer Science & Business Media, 2006.

[7] L. Fleming, "Recombinant uncertainty in technological search," *Management science*, vol. 47, no. 1, pp. 117–132, 2001.

[8] C. S. Weng, W.-Y. Chen, H.-Y. Hsu, and S.-H. Chien, "To study the technological network by structural equivalence," *The Journal of High Technology Management Research*, vol. 21, no. 1, pp. 52–63, 2010.

[9] P.-C. Lee, H.-N. Su, and F.-S. Wu, "Quantitative mapping of patented technologythe case of electrical conducting polymer nanocomposite," *Technological Forecasting and Social Change*, vol. 77, no. 3, pp. 466–478, 2010.

[10] B. H. Hall, A. B. Jaffe, and M. Trajtenberg, "The nber patent citation data file: Lessons, insights and methodological tools," tech. rep., National Bureau of Economic Research, 2001.

[11] E. Tomppo and M. Halme, "Using coarse scale forest variables as ancillary information and weighting of variables in k-nn estimation: a genetic algorithm approach," *Remote Sensing of Environment*, vol. 92, no. 1, pp. 1–20, 2004.

[12] P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, and L. Zalányi, "Prediction of emerging technologies based on analysis of the us patent citation network," *Scientometrics*, vol. 95, no. 1, pp. 225–242, 2013.