# Application of Patent Networks to Information Retrieval: A Preliminary Study

CS224W (Jure Leskovec): Final Project

12/07/2010

Siddharth Taduri
Civil and Environmental Engineering,
Stanford University, CA.

***Abstract*—Information pertaining to the patent system is scattered not only across the patent domain, but also across other scientific and regulatory domains. In recent years, there has been an explosive growth in scientific and regulatory documents related to the patent system. In this project, the use of networks and network analyses is applied to an Information Retrieval (IR) problem in the patent domain. First a patent citation network is developed and tested. Second, a text-based semantic network of patents is developed based on a use case "erythropoietin". An algorithm is proposed which attempts to identify the important patents in the network.**

***Keywords-patent, citation, semantic, network, information retrieval, search, similarity*.**

## I. INTRODUCTION

A patent is a document or set of rights, which gives the inventor of an invention, sole rights to make, use, sell or import the said invention. Patent documents hold high importance due to both their economic and technical value. Relevant information regarding a particular concept is spread across many diverse domains ranging from patent documents to scientific publications. For example, when a startup company is planning to launch a new product, they must take utmost care not to infringe existing patent documents. In addition, if they intend to patent their method, design or process, they need to search all available forms of printed publications which not only include patent documents, but also scientific publications, poster presentations etc. Similarly, a patent examiner examining a patent application must search all forms of related prior art before issuing the patent.

However, in recent years there has been an explosive growth in the amount of information made available online. Currently, the US patent database holds over 7 million issued patents and an average of 1000 new applications filed every single week. PubMed, which is a biomedical database of scientific literature holds over 15 million citations alone. When scaled to other technical databases such as IEEE, ACM etc. the number of documents is very large. This project attempts to explore the possibility of using network analyses, by creating networks of patent documents to enhance information retrieval.

First, a patent citation network is created, where the nodes represent patent documents and the edges represent citation links between them. Second, a text-based semantic network is created, where as in the previous network, the nodes represent patent documents and the edges represent a semantic relation between two patents. A third network is developed, which is a combination of the two earlier networks, with the aim of extracting the positive characteristics of both the networks. The generation of these various relations will help us study and compare various network properties to comment on how they affect the network configuration. The objective is to generate a complex network which potentially combines the positive effects of two distinct methods in order to make the information retrieval process efficient. An algorithm for identifying key patents is also proposed in Section III.4.

The project introduces the concept of multi-domain information retrieval, and subsequently the use of networks which span multiple information domains to aid in IR. For example, a network can be developed where edges can represent relations between documents of a similar type, ex. patents, or relations between two different types of documents, ex. patents and publications. Although the project does not provide any analysis on this network, the potential of such a network for multi-domain information retrieval is certainly an interesting future direction of this work.

Section II reviews some background material such as the databases used in this project, the use case and related work. Section III discusses the methodology employed to develop the graphs and algorithms. Section IV discusses the results obtained and Section V concludes with potential future directions.

## II. BACKGROUND

### *A. Introducing the Use Case*

The networks and analyses for this project will be demonstrated through a use case "erythropoietin". Erythropoietin is a hormone which controls erythropoiesis, which is the production of red blood cells in bone marrow. External preparation of erythropoietin has made possible the treatment of diseases such as anemia (the inability to produce sufficient red blood cells in the body). Synthetic production of this hormone has led to several commercially available drugs. Amgen Inc., an international biotechnology company, produced the first commercially available drug – Epogen and holds 5 patents for the production of erythropoietin – U.S.

5,547,933, U.S. 5,618,698, U.S. 5,621,080, U.S. 5,756,349 and U.S. 5,955,422, which have since been cited as well as challenged by many others. Following forward and backward citations of the five core patents, 135 closely related patents to the use case are identified. These 135 patents, including the 5 core patents will serve as an evaluation metric, or the ground truth to calculate precision and recall values where applicable.

BioPortal is an online repository of over 150 bio-domain ontologies [6]. Based on an initial search for the keyword "erythropoietin", around 11 ontologies are identified. From these ontologies, the synonyms, parent, grandparent and children classes are extracted to generate a term base of 43 closely related concepts to erythropoietin. The top 50-100 patents for each of the 43 concepts are gathered from USPTO, resulting in a total of 1156 US Patents related to the use case erythropoietin. These 1156 patents will be used for the generation of the patent networks for part 1 and part 2 of the project. The documents on USPTO have a standard format and available as HTML files. This allows us to write an automatic parser which extracts selected information from the document such as citations, inventor names etc. The patents were gathered automatically through a script which queries the USPTO database and downloads the corresponding patent document.

The set of 135 patents alone reference a set of over 3000 publications. PubMed is a very comprehensive biomedical database available online. They provide tools and web services which allow us to automatically download publications such as ESearch and EFetch [5]. However, not all of the 3000 publications identified through the patents are available. Furthermore, the full-text of many publications is unavailable, and hence, generating a parallel publication citation network is a much harder task. Hence, part 2 will only do a preliminary study of the network where edges can represent relations between patent documents, and cross-over to represent relations with publications. For the network in part 2, each of these publications will be given a unique node ID so they can be identified through the analysis. Together, the networks are expected to have about 1156 nodes representing patents, and over 5000 nodes representing publications.

### B. *Related Work*

Several researches have made use of patent citations to analyze statistics, and also perform information retrieval in the patent domain [1], [2], [3], [8], [9]. In [1], a text based network is developed which lays the basis for the semantic network described in this project. In [2] and [3], they use various patent networks t perform statistical analysis and inferences about the domain inventors, assignees, classifications etc. An approach which makes use of patent classification to enhance patent retrieval has also been proposed [9].

The work presented in this paper attempts to extend the information retrieval techniques on the basis of the related literature by incorporating semantics into the patent network.

## III. METHODOLOGY

### A. *Generating the Networks*

#### 1) *Patent Citation Network*

Citations to prior art in any document provide a basis for the work presented. Any pair of documents related via a citation link can be considered to have a substantial relevance with respect to one another. In this network, the nodes represent patent documents and the edges represent a citation link between the documents. This graph is considered undirected ignoring the fact that no two documents can cite each other as prior art. In order to generate the network, the patent document is parsed to extract the prior art section containing the citation links to patent documents. The scope of the citations is limited to the 1156 documents in order to limit the expanse of citations which could increase exponentially otherwise. Hence, if a document cites another document which is not in the list of 1156 chosen ones, is discarded.

Citations are sometimes subject to local biases, self citations and lack of inventor knowledge or thorough examination [7]. Citations alone cannot express the semantic relation between two documents, and do not describe what it is exactly which related the documents. Moreover, the citation graph for the 1156 nodes mentioned here in the project has very few edges interconnecting them (see Section IV). To overcome some of these potential drawbacks, a text-based semantic network is developed as explained in Section III.2.

#### 2) *Text-based Semantic Network*

In [1], they generate a network based on the similarity of keywords between two documents. They consider the most frequently occurring words to compare the documents. However, this approach has many limitations. Many technical words are used interchangeably between one another in forms of synonyms, homonyms, hypernyms, abbreviations etc. When comparing two documents, one must go a level above standard vector space modeling techniques such as stop word filtering and word stemming, and consider all these variations. Another drawback of this approach is that using only the most frequent words of a document may not capture the semantics underlying these documents. If one is interested in the concept erythropoietin, a relation between two documents drawn due to the common occurrence of the concept GSM is irrelevant.

In order to address this issue, a more concept based approach is followed. From the ontologies made available by BioPortal, the synonyms of the concept erythropoietin are extracted to create a term base of 13 concepts. The documents are now modeled in vector space of the 13 concept terms, where the magnitude along each dimension represents the frequency of occurrence of the term in that document [10], [11]. The similarity score for a document pair is the weighted cosine similarity measure of the 13 term concept vector. By experimenting with various cut-off values, an edge can be drawn if a certain similarity measure is greater than the set cut-off. Note that the resulting edges are weighted edges, where the weight is the similarity score between the patent documents, with 1.0 indicating an exact match and 0.0 indicating no match. The change in the network configuration by manipulating the

cut-off value can be studied, and the ideal value for the cut-off can be determined experimentally.

*3) Combined Network*

While it is largely considered that citations often suffer from local bias such as self-citations and limited domain knowledge, they still provide legitimate links. Two documents can be classified into completely different technological classes but may yet be related. While a concept based network may not detect this link, a citation based network might do so. The third network is a combination of the above two approaches, with a goal of combining the positive characteristics of the two approaches such that the positive aspects of both approaches are combined into one. Edges represent both citations as well as concept based similarity. The ideal cut-off value is chosen based on the analysis of the network in Section IV. Figure 1 illustrates a simple hypothetical case of these three networks. In the combined network, citation links are given a weight of 1.0, indicating a very high match.

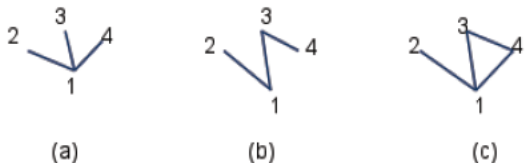


**Figure 1: (a) Citation Network; (b) Semantic Network; (c) Combined Network**

*4) Proposed IR algorithm*

A node with a high degree indicates some level of significance in the network. However, naively choosing these high degree nodes may not lead to the most accurate results. In the algorithm proposed in this project, at every iteration, the nodes with the least degree are deleted. Among the remaining nodes, a scoring function is applied to score the nodes:

score[node] = score[node] + degree[node]

Taking the example shown in Figure 2, simply going by the degree will give the same score to 10, 11, 12 and 13. However, the proposed algorithm will give 13 a higher weight than 10, 11 and 12 due to its position in the network. The algorithm also attempts to overcome imbalances such as when an older patent has higher citations, but a much more relevant newer patent has fewer citations due to the time it has been in the network.

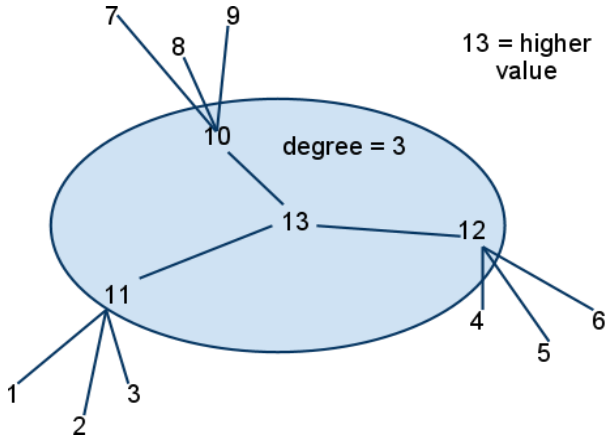


**Figure 2: Illustration of the Algorithm**

## IV. RESULTS

Six patent text based semantic networks were generated by specifying different cut-off values. Figure 3 shows the number of edges for each of these graphs. Notice that the citation graph has very few edges (~2033), amongst 1056 nodes, which is a very small number and probably not sufficient to model these documents. As expected, by decreasing the cut-off values while generating the semantic graphs, the number of edges significantly rises. The average path lengths also reduce with change in cut-off values indicating the nodes are now closer to each other (see Fig 4). The citation graph has a lot of individual components initially, which are either lone nodes, or a small set of nodes closely connected to each other. These components are soon deleted by the algorithm, however. The clustering coefficients also increase as the graphs become more denser as expected (see Fig 5).

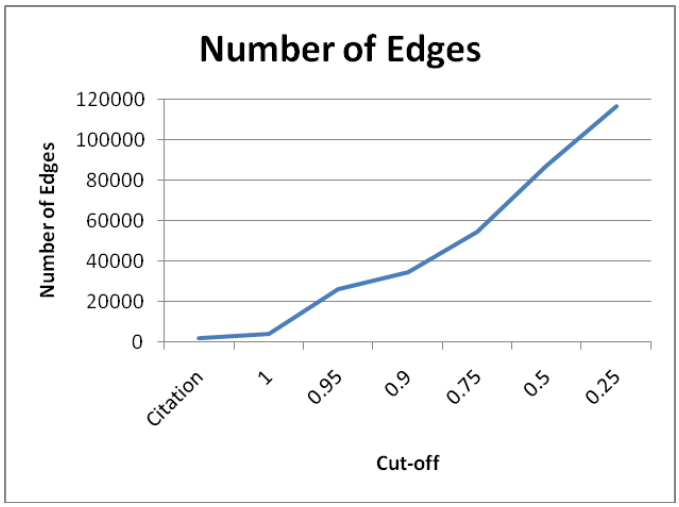


**Figure 3: Number of Edges for the Networks**

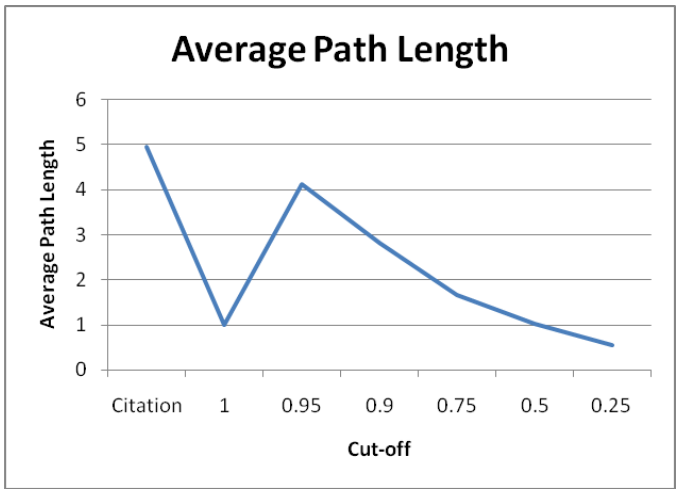


**Figure 4: Change in the Average Path Length**

Precision and recall are standard metrics to evaluate an IR algorithm. Figure 5 shows the graph of the precision and recall values of the proposed algorithm for each of the graphs that have been generated. The top 200 results were considered for the purpose of calculating the precision and recall statistics. We

notice that the best results are achieved for the citation network. This is rather surprising since the graph has much fewer edges when compared to the text based graphs. Since we ignore edges which do not fall completely among the selected 1156 patents, the remaning edges could possibly overlap the patents selected as ground truth as a coincidence. Also, since the ground truth is chosen by following citation links, this network probably performs very well.

Among the text based networks, the best precision and recall values are obtained for the graph with cut-off=0.5. By choosing this graph, we ask a question – will the combination of the two networks improve the results of the individual networks? The answer depends mainly on how many new patents the text-based semantic network has identified, which were not identified by the citation network. Upon testing this, no new patents were identified by the semantic network when compared to the citation network.


**Figure 5: Precision, Recall and Clustering for the Networks**

The two graphs were combined together to form a complex network, hoping to extract the positive characteristics from both the graphs (see Figure 6). The algorithm performes porrly on the combined network, when compared to both the original counterparts. To try and improve the results, an alternate scoring function was experimented. To the scoring function described in Section III.4, a measure of quality for every node is multiplied to give:

score[node] = score[node] + degree[node] * quality[node]

The chosen quality function in this case takes into account the weight of the edges incoming into a node. For example, a high degree node with weak edge weights gets a lower score than a node with lower degree but high quality links. The new scoring function improves the precision and recall values by a small fraction, indicating that alternate scoring schemes could be potentially explored. An interesting dimension can be added by considering annotated data such as classification, inventors, assignees etc. to act as predictors for the quality of a node (ex. highly cited class, inventor name observed for the first time etc.)

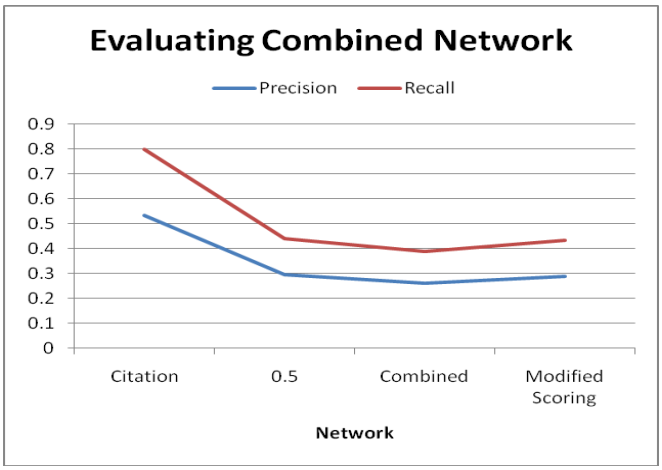


**Figure 6: Comparing the Combined Network and Applying Quality Function**

## V. CONCLUSION

This project explores the idea of generating semantic networks in addition to patent citation networks to enhance information retrieval in the patent domain. Through a use case erythropoietin, a set of 1156 patent documents is defined as the corpus using which three networks were generated – (1) citation network; (b) text-based semantic network; (c) combination of the two networks in (a) and (b). Upon these networks, the proposed information retrieval algorithm was executed and corresponding precision and recall values were noted.

While the semantic networks were outperformed by the citation based network, an alternative ground truth could prove otherwise. Additionally, one could follow the semantic relations from the bio-ontologies to extract more terms involving parents, children, grandparents so on. Observe that as we go away from a particular concept, we tend to become more generalized, or more specific, and hence a weighting function can be fit onto these concepts such that the synonyms have the highest weight, and the grandparents have the least weight. This will balance the effect of very general terms that are likely to occur in more documents and avoid unnecessary links.

Overall, this paper concludes that it is equally important to have a strong graph, as it is to have a good IR algorithm and other network analyses. Different users may have different expectations of a search results. The text-based networks are literally a semantic network representation of the patent domain with respect to the user query. The network edges change with respect to the query and can express the domain in a dynamic way rather than a static way such as citation networks, which will return the same set of results to all users irrespective of the subtle differences in their requirements.

### A. *Future Work*

A potential future direction of this project is to explore multi-domain networks. Let us consider an example of the publication domain and the patent domain. Patent documents cite both publications and patents in their prior art. These citations can cross-over from the patent domain to the

publication domain adding a new dimension to the publication domain. An interesting experiment will be to understand how publications are viewed in a commercial (patent) standpoint. Does adding this new dimension reveal new information about the publications which was previously not retrievable when only the publication domain was considered? This exploratory topic could potentially lay the foundation for multi-domain networks and their use for Information Retrieval.

## REFERENCES


[1] Byungun Yoon and Yongtae Park, "A text-mining-based patent network: Analytical tool for high-technology trend", The Journal of High Technology Management Research, Volume 15, Issue 1, February 2004, Pages 37-50

[2] Christian Sternitzke, Adam Bartkowski, Reinhard Schramm, "Visualizing patent statistics by means of social network analysis tools", World Patent Information, Volume 30, Issue 2, June 2008, Pages 115-131

[3] Li, X., et al. (2007a). Patent citation network in nanotechnology (1976–2004). Journal of Nanoparticle Research, 9, 337–352, 2007.

[4] USPTO. http://www.uspto.gov/

[5] PubMed Utilities. http://eutils.ncbi.nlm.nih.gov/

[6] BioPortal. http://bioportal.bioontology.org/

[7] Jaffe, Adam B., Trajtenberg, Manuel and Fogarty, Michael S., "The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees", NBER Working Paper No. W7631. Available at SSRN:http://ssrn.com/abstract=228106, 2000.

[8] Fujii, A., "Enhancing patent retrieval by citation analysis", In Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, New York, pp. 793-794, 2007.

[9] Kang, I., Na, S., Kim, J., and Lee, J., "Cluster-based patent retrieval", Information Processing and Management, pp. 1173-1182, vol. 43 (5). Sep. 2007.

[10] Baeza-Yates, R. and Ribeiro-Neto. B., *Modern Information Retrieval*, ACM Press, 1999.

[11] Manning, C.D., Raghavan, P. and Schütze, H., *An Introduction to Information Retrieval*, Cambridge University Press, 2009.