RANKING PATENTS FOR BETTER SEARCH CAPABILITIES

Mihai VLASE, Dan MUNTEANU

"Dunărea de Jos" University of Galati
Faculty of Computer Science
Department of Computers and Applied Informatics
111 Domnească Street, 800201-Galati, Romania
Phone/Fax: (+40) 236 460182; (+40) 236 461353
E-mail: mihai.vlase@ugal.ro, dan.munteanu@ugal.ro

Abstract: The continuous increasing number of patents worldwide makes the searching in these vast databases a real challenge. In the present paper we present a method, based on a computed rank that attempts to order the patents by a relevant order transforming a long list of hundreds of search results into a list ordered by relevance.

Keywords: database; search; rank; patents; citations.

1. INTRODUCTION

One of the most important steps that an inventor has to take in order to obtain a patent is searching for similar ideas. An inventor starts with an idea, then the first step that has to be taken in the long process from that idea to a registered patent is a preliminary search to find out if the idea is or not already available on the market or known to experts. The second time when an inventor involves searching in the patenting process is even more important and represents the prior-art search. If in preliminary search inventors can use any information available on the Internet (web pages, news pages, product information pages, virtual shopping sites, article, patents) to make a quick validation of the originality of his idea, in prior-art search inventors perform their searches more rigorously, mostly on articles or patent databases.

When an inventor proceeds to a search task, the inventor will start with the most popular patent database sites available worldwide. When a search is made in one of these sites, the number of results can be overwhelming. Most of these specialized sites offer fields filtering features to reduce the number of results. However, even when filters are used, the number of search results can be too large. Most of these specialized search sites list all these results ordered by application date and very few of them order the results by relevancy.

In the present paper we describe a method that calculates a rank for each patent from the database so the results list of a search is ordered by this rank. Thus, when a search has been made and a large number of results are displayed, we can order those results by this rank, so of the top of the list will be displayed the most relevant patents.

2. THE REFERENCE SYSTEM OF PATENTS

A patent document is structured in several sections. One of these sections consists of a list of references to older patents or other relevant documents or articles. These documents are obtained from a systematic and meticulous search and whose outcome depends of how strong will be the patent in case of infringement. Also, the active patents can be cited by the newer patents, thus creating a network of citations where newer patents cite older patents.

For example, patent EP0750986 cites sixteen patents: EP0064939, EP0453790, EP0559556, EP0648599, DE1957270, DE3638322, DE4205746, DE605994, DE8224870, FR397430, FR2561217, FR11969, GB1065028, GB1099069, GB2128953, US4618138. Same patent EP0750986 it is cited by other three newer patens: EP0851811, EP0960018, EP0977662.

This paper was recommended for publication by Cornelia Tudorie

Also, if we take one of the cited patents of the EP0750986, for instance EP0453790, this cites six other patents: EP0134526, DE1611379, DE3343811, DE3838078, GB2218953, US4819928 and it is cited by one: EP0888992.

In this way results a graph of citations where patents are nodes and referenced links. This graph is showed in figure 1.

3. APPROACH

The present study is based on a database from European Patent Office (EPO) which contains all accepted patents from 09 January 1980 until 29 December 2004. The total number of patents contained in this database is 707594.

The original data from EPO had to be processed from initial format (plain text), into a relational database, in our case MySql, in order to apply further data mining techniques.

In this study we start from the idea of improving the search capabilities by adding a rank to each patent and then display the search results ordered by this value.

3.1. Ranking Patents.

As we see in the previously chapter, patents contains a list of references. So we can make an analogy between patents references and web page references (links).

In this case we can apply the simple form of PageRank method to solve the problem of ranking.

PageRank is a web-page importance ranking algorithm developed by Page and Brin from the GOOGLE Inc. (USPTO (2001)). This method is related to the techniques of analyzing and assigning ranks to nodes in linked databases (Page *et al.*, 1998). The basic idea of this method is that it is based on citation (Lukach and Lukach, 2008).

According to the original publication of Page et al. (1998) the PageRank algorithm is defined as following:

Let u be a web page. Then let F_u be the set of pages u points to and B_u be the set of pages that point to u. Let $N_u = |F_u|$ be the number of links from u and let c be a factor used for normalization (so that the total rank of all web pages is constant) (Page *et al.*, 1998).

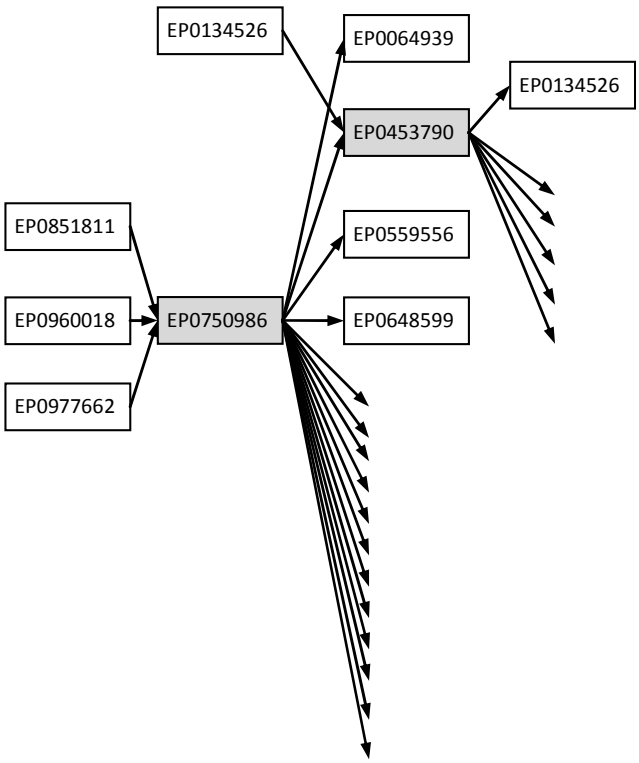

Fig.1. Graf of citations with full references for patents EP0750986 and EP0453790.

$$(1)\quad R(u) = c\sum_{v\in B_u}\frac{R(v)}{N_v}$$

The main idea of PageRank is that a page has high rank if the sum of the ranks of its back-links is high. This covers both the case when a page has many back-links and when a page has a few highly ranked backlinks (Page *et al.*, 1998).

Extrapolated to patents, based on the citation system, we can say that a patent has a high rank if it is cited by a higher number of other patents, or if it is cited by a small number of highly ranked patents.

So, for the patents we can define the ranking as follow:

Let p be a patent. Then B_p let be all patents that cite patent p (cite by). Let T_p be all patents which are cited by the patent p (cite to). Let $N_p = |T_p|$ be the number of citations from patent p and c a constant.

Therefore the rank (weight) of patents p, $W(p)$, can be defined as:

$$(2)\quad W(p) = c \sum_{v \in B_p} \frac{W(v)}{N_p}$$

3.2. Results

As we mentioned before, the algorithm explained in previews section was tested on the EPO database which contains 707594 patents. The algorithm converged after 9, 10 steps. After 10 steps, the patent rank does not change significantly the order of patents.

We test the performance of the present algorithm by searching for a couple of common words (like mouse or screen) and then we order the patents results by rank. We note that in first positions are placed patents that are highly cited by other patents, which means that is most probably that we are also find this patents relevant for our own searches.

We further compare the search results ordered by the number of patents that cite each patent from result list and we note that we achieve approximately the same result.

The explanation for this behavior is that despite of more than 700000 patents the database is still small in data. The true value of the present algorithm can be reached if the database contains more patents (maybe a union of more patents databases).

The difference between search results ordered by count of citations and calculated rank is much more obviously when the list of search results contains a large number of patents with the same number or approximately same number of citations to them. In this case ordering the search result by the calculated rank, we retrieve more relevant patents on first positions.

For instance if we have a search result list with 20 patents with different number of patents that cite each of them, is very possible to have approximately same results if we order by rank or by number of citations. But if we have a search result list with 20 patents with the same number or approximately same number of citations to them, ordering by rank we retrieve a list of patents with most relevant patents placed on first positions.

So the patent search result ordered by rank is recommended instead of ordering by number of citations of each patent, when we retrieve a large number of patents with approximately same number of citations to each of them.

4. CONCLUSION

Even though there are many similarities between the relevancy of web sites and patents, there are many dissimilar aspects that should be considered for rank calculations for patents. The hierarchy of citations from the newer to older patents is one of these aspects.

The relevancy of a patent is also given by the type of search. It is possible that an inventor be interested in the latest technologies emerging on market so the inventor will consider relevant the newest patents. On the other hand, it is possible that an inventor may perform a prior-art search and the inventor will consider relevant older patents, patents that had the time to prove their value, being cited by other patents. And there are inventors who perform patent searches just to find information regarding the product that they are about to patent, so they are interested in all available patents. Often, corporations that apply for patents are interested in similar patents within their industry, in particular, patents applied for by their main competitors.

Using a relevancy algorithm yields considerably better results than simple ordering by year. Using our relevancy algorithm we retrieve a list ordered by this calculated rank, where in the first positions were patents applied by prestigious companies or that are highly cited by other patents.

The present algorithms leave open the possibility of improvement considering in the rank calculation other relevant patent properties.

It is possible to note a better performance if we increase the number of patents in database, for instance by adding patents from other databases as USTPO (United States Patent and Trademark Office) or WIPO (World Intellectual Property Organization).

Is also possible to improve the algorithm performance if we add in rang calculation other patent features as approved year, applicant name, or patent class. Each of this feature can add a small power to the calculated rank.

This approach let open further research for improving algorithm performance.

5. REFERENCES

Hagedoorn, J. and M. Cloodt (2002). Measuring innovative performance: is there an advantage in using multiple indicators? *Research Policy* **vol**.32 p. 1365–1379, Elsevier Science B.V,.

Lukach, R. and M. Lukach (2008). Ranking USPTO Patent Documents by Importance Using Random Surfer Method (PageRank). *Social Science Research Network*, Working Paper Series, Last revised: December 01 2008.

Page, Larry, Sergey Brin, R Motwani and T. Winograd (1998) The PageRank Citation Ranking: Bringing Order to the Web. *Stanford Digital Libraries Working Paper*.

Page, Lawrence (1998). Method for node ranking in a linked database, Stanford, CA, United States Patent 6,285,999. ,

Patil, A., A. Ambekar and M. K. Sundararajan (2007). Patent Analysis, *Computational Biology: Patent Analysis* - Technical Report.

Tseng, Yuen-Hsien, Chi-Jen Lin and Yu-I Lin (2007). Text Mining techniques for patent analysis, *Information Processing and Management*, **Vol** 43, No.5, Pages 1216-1247.

Vlase, M., Munteanu D., (2009) Patent relevancy on patent databases, *Networking in Education and Research*, RoEduNet International Conferince 8th Edition