PatentsSearcher: A Novel Portal to Search and Explore Patents

Vagelis Hristidis
School of Computing and Information Sciences
Florida International University
Miami, FL
vagelis@cis.fiu.edu

Eduardo Ruiz
School of Computing and Information Sciences
Florida International University
Miami, FL
edjruiz@cis.fiu.edu

Alejandro Hernández
School of Computing and Information Sciences
Florida International University
Miami, FL
ahern066@fiu.edu

Fernando Farfán
Computer Science and Engineering Department
University of Michigan
Ann Arbor, MI
ffarfan@eecs.umich.edu

Ramakrishna Varadarajan
Computer Sciences
University of Wisconsin
Madison, WI
ramkris@cs.wisc.edu

ABSTRACT

There is an abundance of systems today to search for relevant patents, ranging from free ones like Google Patents (google.com/patents) to subscription ones like Delphion (delphion.com). After studying many existing systems, we found that they all apply general-purpose Information Retrieval (IR) techniques to rank patents. We argue that the quality of search can be significantly improved by exploiting the domain semantics: E.g., patents are organized into classes and subclasses, and have links to external publication and to other patents. Also patents' text is organized into various sections and uses specific legal wording.

We present the *PatentsSearcher* system, available at PatentsSearcher.com, whose key contribution is to leverage the domain semantics to improve the quality of discovery and ranking. PatentsSearcher also offers other novel functionalities to help users locate and navigate relevant and important patents or applications.

Keywords

Search, patents, ranking, user interface.

1. INTRODUCTION

According to the World Intellectual Property Organization [9] about 1,680,000 patents were filed in 2005 and there was an annual increase of about 7% in this number. The cost of filing patents, defining claims and defending a claim of infringement is also increasing with time, making the process too expensive for small companies or universities, and too time consuming for large companies (due to many rounds of refinement). A key reason of this increased cost, as claimed by patent attorneys contacted by the authors and other resources (e.g., [8]), is the cost to find relevant patents. [8] estimates patent search cost to $1,500 per patent filing.

There are many systems to search for relevant patents, ranging from free ones like Google Patents[1] to subscription ones like Delphion[2]. In our opinion, there are three major factors that define the value of a patents search system:

(a) *Coverage:* what patent and non-patent (e.g., legal state information) databases are included, from what countries or unions, and how up-to-date the data is.

(b) *Discovery and ranking:* how well does the system locate relevant patents or applications for a query, and how these are ranked.

(c) *Other features and analytics:* For instance, Delphion provides an interface to visually view the references among the patents of a result, or cluster the patents.

Our key contribution is in the area of "discovery and ranking". To the best of our knowledge, current patent search systems rank the relevant patents either by date (e.g., United States Patent and Trademark Office[3]) or using well-studied Information Retrieval (IR) techniques [5]. In particular, they rely on free (e.g., Lucene[4]) or commercial IR software. Further, some systems also use standard topic-extraction techniques like Latent Semantic Analysis [1] to achieve dimensionality reduction which intuitively means that relevant terms of the query terms are also searched. Such generic software is designed to search and rank any collection of documents, without any specific optimization based on the domain of the document corpus.

However, patents have some special characteristics and semantics, which deem generic search techniques suboptimal:

1. Patents are organized into classes and subclasses.
2. Patents have links to other patents and external publications.

[1]http://www.google.com/patents
[2]http://www.delphion.com
[3]http://www.uspto.gov
[4]http://lucene.apache.org

PaIR 2010 Toronto, Canada.

3. Patents' text is organized into various sections (abstract, claims, description and images).
4. Patents use specific legal wording in the claims section. Further, patent claims have references to other claims, that is, claims can be viewed as a graph.

We started the *PatentsSearcher* project in an effort to address such domain semantics. A secondary goal of *PatentsSearcher* is to provide an intuitive user interface, while providing a suite of features that have been selected after discussing with patent attorneys. In particular, *PatentsSearcher* currently provides the following functionality, which is expanded every month:

1. Rank classes, inventors and assignees, in addition to patents. This is especially useful when patent attorneys do an exhaustive search of the most relevant classes for a topic. *PatentsSearcher* returns the most relevant classes using a complex ranking technique.
2. Export patent results into a spreadsheet.
3. Rank results by various attributes like date, relevance, and title.
4. Create a PDF file for one or more result patents.
5. Save and annotate a set of patents using the Patent Cart functionality.
6. Create and submit an expanded query to other data sources: *Google Scholar* and *Google Web Search*.
7. Search for similar patents to a specific patent.
8. Filter results by class, assignee, and inventor.
9. View and follow backward and forward patent reference links for a patent.
10. Query for very recent patents and application that have not been indexed yet by *PatentsSearcher*, but are retrieved on-the-fly from the USPTO web site.

The paper is organized as follows. Section 2 discusses the overall system architecture. Sections 3 and 4 present the details of the Searching and the User Interface Modules. Section 5 surveys the Related Work and Section 6 sketches our conclusions and future work.

2. ARCHITECTURE

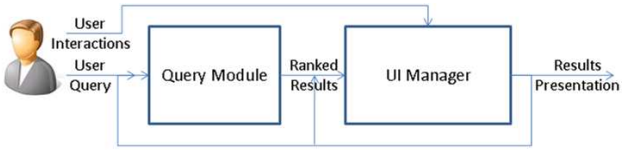

Figure 1: System Architecture.

This section presents the overall architecture of *PatentsSearcher*. As shown in Figure 1, there are two main modules. The Query Module inputs a query, which can be a simple keyword query or a query formulated using the Advanced Query Interface (described in Section 4). As detailed in Section 3, the Query Module discovers relevant patents and applications and ranks them by relevance and importance.

The results are displayed by the User Interface Manager Module, which allows the user to navigate the results by re-sorting, filtering, exploring, saving and annotating them. The UI Manager is presented in detail in Section 4.

Figure 1 shows the query-time components of the system. There is also a Data Collector module, which is not discussed in much detail in this paper. The Data Collector crawls the USPTO web site to retrieve all patents and applications; it then parses and indexes them. Further, the Data Collector crawls, stores and indexes other external data collections, related to patents. Currently, we crawl publications sites and the Web. These external sources are used to improve the precision and recall of the system as discussed in Section 3.

3. SEARCHING AND RANKING

Figure 2 shows the architecture of the Query Module, which inputs a user query and outputs a ranked list of patents or patent applications. The input query may be expressed using the Advanced Query Interface described in Section 4. We focus on the case where only a list of keywords is specified in the main text box and no advanced conditions are specified, since the advanced conditions play the role of filters.

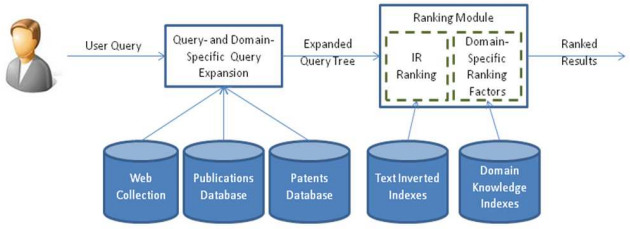

Figure 2: Query Module.

We follow disjunctive semantics, that is, not all keywords need to appear in the query.

Query- and Domain-Specific Query Expansion: The goal of this module is to input a list of keywords and output a set of synonymous or tightly relevant keywords or phrases. For instance, the query {*web ranking*} is expanded as {*(web, 1.0), (ranking, 1.0), (documents, 0.3), (pages, 0.25), (search,0.22)*}. For each relevant word, a weight is assigned denoting its semantic distance to the source keyword, *for the specific query.*

The synonyms (or relevant terms) and their weights are computed on-the-fly at query time for the context of the user query. The context of the query is defined using the result patents, relevant publications and Web documents. In particular, the query is submitted to these three data sources, as shown in Figure 2, which have been retrieved, parsed and indexed before queries arrive. Given the query context, defined as the top results from each of the three sources, we compute the most important words using an adaptation of the Rocchio work on pseudo-relevance feedback [4], where the different sources are weighted according to their relevance to the query. For instance, for some query, there may not be many relevant publications, but there are some relevant products, which are captured by the Web repository. In this case, the Web repository is more important than the publications repository in defining the query context. Figure 3 shows a screenshot of a query result page.

Exploit Domain Knowledge in Results Ranking: The Ranking Module inputs the expanded query and executes the

Figure 3: Results for query "Web search ranking".

query on the patents database. The key novelty of this module is that in addition to a traditional Information Retrieval module used by most other patent search systems, the Ranking Module exploits a set of unique properties of the patents database. We give an overview of the key unique properties and the way we leverage them:

1. *Patents are organized into classes and subclasses:* We leverage this information to assign a degree of relevance to patents of the same class. That is, if a class has many results for a query, then this class becomes important for the query and hence its patents may be more relevant than the patents of other classes.
2. *Patents have links to external publication and to other patents:* For instance, a patent with many citations is better than a patent with few citations.
3. *Patents are organized into various sections (abstract, claims, description and images):* For instance, a word that appears in the abstract is better than one that appears in the description. Further, if both query keywords appear in the same section and in close proximity, it is better than if they appear in different sections.
4. *Patents use specific legal wording in the claims section. Further, claims have references to other claims, that is, claims can be viewed as a graph:* Hence if two query words appear in two linked claims it is better than if they appear on two disconnected claims.

Another challenge is to combine all the above ranking factors into a single score for each patent result. We have experimented with different combining functions using user surveys, and we have chosen to follow a weighted linear combination of all the factors, where the weights have been selected using user surveys.

Note that both the modules described above, the Query Expansion and Ranking, contribute in improving the precision and the recall of our system.

Example: Consider the query *Web search ranking*. At the time this paper was written, as shown in Figure 3, the second result of PatentsSearcher (*6356899: Method for interactively creating an information database including preferred information elements, such as preferred-authority, world wide web pages*) was ranked 120th in Google Patents[5], which is one of the most popular web patents search engines. Looking closer at this patent we can see that it is a critical patent in this area, cited by 76 other related patents. One of the reasons that most other patent search systems fail to recognize the relevance of this patent is that they rely on traditional IR methods, and this patent has very long title (high document length in IR) and only contains one of the three query keywords in the title.

On the other hand, the top result of a popular web patents search engine, *6073135: Connectivity server for locating linkage information between Web pages*, is mostly irrelevant to the query, since it tackles the problem of storing the links between pages and not searching or ranking the web.

4. USER INTERFACE MANAGER

This section presents the key functionality of PatentsSearcher, in addition to the actual ranking of the results. The UI Manager Module, shown in Figure 4, interacts with the Query Module and the user to facilitate the effective dissemination and navigation through the results.

[5]http://www.google.com/patents

Figure 4: UI Manager.

Advanced Query Interface: Like most patent search engines, our Advanced Query Interface, shown in Figure 5, allows specifying constraints on the Inventor, Assignee, Date, Class, and others. A unique feature is that, in addition to patents, the user may search for the most relevant Classes, Assignees or Inventors ([3] had also discussed searching for Classes). This feature was requested by patent attorneys we contacted. Finding the most relevant classes can be used to do an exhaustive class scan, whereas finding the most relevant inventors can be used to find area experts for litigations. Another novel feature of PatentsSearcher is the capability to view very recent, un-indexed patents or applications. We refresh our index approximately every month. In order not to miss very recent results, the user can check the "Possibly related patents that are too recent and not yet indexed by our system" checkbox in Advanced Query options. This shows, at the top of the results page, a list of patents or applications that contain all the query words, ordered by date, retrieved by directly querying the USPTO for the unindexed period. Further, the user may enable or disable the Query Expansion Module from the Advanced Query interface. For instance, if a user knows that the patents she is looking for contain some specific keywords, allowing query expansion may dilute the results.

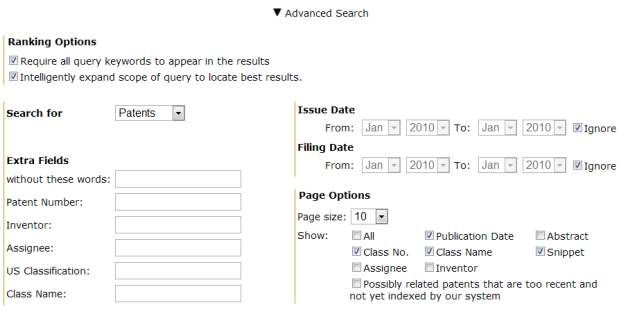

Figure 5: Advanced Query Interface.

Results Presentation and Navigation: Once the ranked list of results is displayed, the user may re-sort them by Date, Title, and so on, as shown on the left pane of Figure 3. Further, the user may filter the results by Class, Inventor, and so on, by clicking the corresponding link at the end of a displayed result. PatentsSearcher also provides two useful export functionalities. First, the results can be exported to an Excel spreadsheet, where each result is a row and its attributes (number, title, inventor, and so on) are columns. Second, the current page of results can be exported to a single *PDF* file, which is the concatenation of the *PDF*s of the results of the page.

Patent Cart: Another useful feature we provide, after discussing with patent attorneys and patent searchers, is to save selected patents/applications to the Patent Cart. The Patent Cart is similar to the concept of Shopping Cart in E-commerce applications. Figure 6 shows the Patent Cart displayed side-by-side with the main search window. A patent/ application is added to the cart by clicking the shopping cart icon next to it. The user can then add some notes for each item in the cart, and export the information in the cart in Excel or *PDF* format.

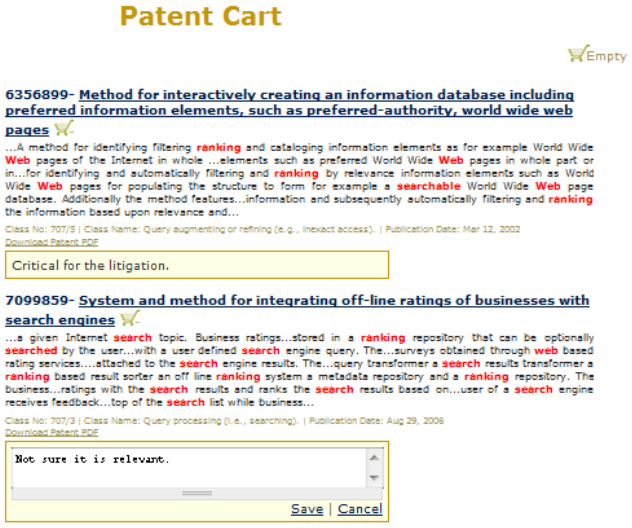

Figure 6: Patent Cart. Notice the annotations the user adds to the items in the cart.

Search External Sources: A key need in searching for prior art is to search non-patent data. Our system provides some basic functionality towards this direction. In particular, as show on the left pane of Figure 3, the user can submit a query to Google for general relevant data or to Google Scholar for relevant publications. For each of these two sources, we offer two variants: submitting the original query or the expanded query.

5. RELATED WORK

In this section we present a survey of related techniques and products to search patent databases.

Patent Indexes and Databases: Several organizations publish indexes of patents to aid the users to find the appropriate information.

For instance, the USPTO publishes the *Official Gazette for Patents* [7], a weekly report of the patents issued that week. It contains bibliographic text and a representative drawing from each patent. Similarly, the European Patent Office publishes the *European Patent Bulletin* [2], also issued weekly. This office also provide web search services. The *Derwent World Patents Index* (DWPI) [6], by Thomson Reuters, is a manually curated index of patents from all over the world. The index provides enhanced patent information including patent titles and abstracts written in English, using clear, consistent, industry-specific terms. DWPI contains over 41 million patent documents, with coverage from over 41 major patent issuing authorities worldwide.

Manually-generated reports: Several systems provide manually-conducted searches, in which humans do the search and return a report for a fee. Generally the user interface of these systems is an online form to specify the search terms.

For instance, *Legalzoom* (www.legalzoom.com) is an online service to create legal documents, which also offers a patent search service focused on *prior art patents*. A similar service is offered by *Questel* (www.questel.com/en/prodsand services/search_pat.htm), providing both basic and complex searches and charging on a per-case basis. *International Patent Search* (www.internationalpatentsearch.com) and *LexisNexis* (www.lexisnexis.com) also provide similar services. Note that the latter also provides a search engine, as described below.

Free online search sites: A large number of online search systems do not charge for the provided service, providing the user with a rich web-based user interface to perform the searches. These systems vary in the techniques involved to perform the search.

For example, *Sumobrain*(www.sumobrain.com) provides a basic IR ranking, single, advanced and fielded search interfaces, and allows users to save searches and patents, and download as PDF documents. To the best of our knowledge, *Google Patents* (www.google.com/patents) only performs traditional IR analysis. All the keywords must be present for conjunctive queries. From an empirical analysis, it seems like the title and (probably) the abstract of the patent are viewed as the most important sections. The proximity between the keywords seems to also be considered. *PatentLens* (www.patentlens.net) provides a service similar to Google Patents. It also allows users to save patents and view them in a single page. *Patents.com* (www.patents.com) provides a basic search interface and a language for complex search, to search a corpus of U.S. and European patents. The search results are ordered by date. *FreePatentsOnline* (www.freepatentsonline.com) provides traditional IR techniques such as word stemming, proximity searching, relevancy ranking and search term weighting. The results can be ranked either by relevance or by date. Additional features include patent organization, annotation, sharing, and alerts. It also provides a functionality to search normalized chemical formulas. *PatentStorm* (www.patentstorm.us) US patents. Rank by IR relevance or by date. *WikiPatents* (www.wikipatents.com) searches patents from the U.S., Japan, Canada and several European nations. It only provides Boolean search on the title, abstract, and assignee, providing an unclear ranking. It also allows users to comment on a patents value or properties.

Subscription-based search sites: Several search sites provide access to their service with a monthly fee. Again, the IR methods and features of these services vary.

Delphion (www.delphion.com) has a tool to display the graph of citations among the patents. Citations are probably not used in searching or ranking. There is also a clustering tool. It can also work on DWPI. *WestLaw* uses Delphion in their integrated Patent Law Practitioner product (west.thomson.com/westlaw/practitioner/patent/demo.aspx). *Questel - Expert Searching* (www.questel.com/en/Prodsand services/Qweb.htm) offers a complex patent query language and a set of patent databases to choose from. The user can specify ranking attribute like date. Questel also offers other patent tools like building patent portfolios. Pantros IP's *ProSearch* (www.patentcafe.com/products/patent_search.asp) uses Latent Semantic Analysis for relevance ranking. As mentioned earlier, *LexisNexis* also provides a membership-based service, which searches patent applications from the U.S. and Europe, as well as patent abstracts from Japan, and Patent Cooperation Treaty patent applications.

6. CONCLUSIONS AND FUTURE WORK

The goal of *PatentsSearcher* is to incorporate cutting-edge research and domain knowledge into the patent search process. We continue to interact with patent attorneys, searchers and other stakeholders to adapt *PatentsSearcher* to their needs. A key direction is to make non-patents information available to the user in a more integrated way, instead of the current simple query spawning capability to external sources. Further, we plan to provide different types of searches like infringement, patent issuing, and landscape search, among others.

7. ACKNOWLEDGEMENTS

We would like to thank the National Science Foundation for supporting the basic research related to this project through grant # IIS-0811922, and the Kauffman Foundation for Entrepreneurship. We would like to thank attorneys Ruben Alcoba and Hannibal Travis for useful input on the patents domain. We also thank Jairo Quintana for crawling the classes hierarchy from USPTO.

8. REFERENCES

[1] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[2] European Patent Office. European Patent Bulletin. http://www.epo.org/patents/patent-information/european-patent-documents/european-patent-bulletin.html, accessed June 24 2010.

[3] L. Larkey. A patent search and classification system. In *Proceedings of the fourth ACM conference on Digital libraries*, page 187. ACM, 1999.

[4] J. Rocchio. Relevance feedback in information retrieval, The Smart retrieval system: Experiments in automatic document processing. *Salton ed*, pages 313–323, 1971.

[5] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Publishing Co., 1989.

[6] Thomson Reuters. Derwent World Patents Index. http://thomsonreuters.com/products_services/legal/legal_products/intellectual_property/DWPI, accessed June 24 2010.

[7] United States Patent and Trademark Office. Official Gazette for Patents. http://www.uspto.gov/news/og/patent_og/index.jsp, accessed June 24 2010.

[8] B. Wiens. *Understanding Patents*. http://www.benwiens.com/patents.html, 2010.

[9] World Intellectual Property Organization. World Intellectual Property Organization Statistics on Patents Report. http://www.wipo.int/ipstats/en/statistics/patents/, accessed June 23 2010.