Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

Developing a Comprehensive Patent Related Information Retrieval Tool

Siddharth Taduri[1], Hang Yu[2], Gloria Lau[3], Kincho Law[4] and Jay Kesan[5]

Stanford University, Civil and Environmental Engineering, [1] staduri@stanford.edu, [3] glau@stanford.edu, [4] law@stanford.edu
University of Illinois Urbana-Champaign, College of Law, [2] hangyu@uiuc.edu, [5] kesan@illinois.edu

Received 09 August 2010; received in revised form 15 January 2011; accepted 25 January 2011

Abstract

In recent years, there has been a massive growth of regulatory and related information available online. This information is distributed across many different domains creating a problem for accessing and managing this data. This paper proposes a framework to access information across two such domains – patents and court cases. The framework is designed to boost the value of a set of patents based on information available in court cases by identifying and cross-referencing mutual information in the two domains. We test our framework by constructing a use case involving the hormone erythropoietin. A corpus of 1150 patents (including 135 closely related patents) and 30 court cases is gathered. Challenges associated with such integration and future plans are briefly discussed.

Keywords: Patents, Court cases, USPTO, Search, Information retrieval

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

1 Introduction

The administration of the government creates and enforces laws and regulations at various levels. At the top most level are the federal laws passed by Congress which focus on a wide range of areas, including science and technology. These laws are codified in the United States Code (U.S.C.). Broad power is given to administrative agencies, such as the Food and Drug Administration (FDA), the Federal Communications Commission (FCC) and the United States Patent and Trademark Office (USPTO), in order to create and enforce rules and regulations that then appear in the relevant chapters of the Code of Federal Regulations (C.F.R.). Huge amounts of information pertaining to science and technology is buried in this system and distributed across various incompatible and sometimes disconnected domains. These domains can be broadly classified into laws, regulations, the documents in the administrative agencies, the documents generated by the court system and other scientific and technological literature. Comprehensive regulatory knowledge on a particular topic is typically spread across several of these disparate domains. For example, a company working in the field of Global System for Mobile Communications (GSM) would likely need to know about existing patents, court litigations involving any of these patents, their competitors' work, and the relevant scientific literature. All of this information is available in different domains, namely (a) the administrative agency (USPTO in this case), (b) the federal court system, (c) the pertinent laws and regulations, and (d) the scientific literature. The task of retrieving information or knowledge relating to GSM requires thorough study of documents across all these domains. With the explosive regulatory growth and related information in the recent years, thorough study of such documents has become a very laborious task involving many hours of manual cross-referencing across different domains due to the lack of smart tools. There is a need for integrating such diverse sources of information and providing a common interface that has the ability to search and correlate information in various domains.

The recent years have seen a tremendous growth in research and developments in science and technology, and an emphasis in obtaining intellectual property protection for one's innovations. In 2009, around 485,312 patent applications were filed with the USPTO (Site 1). PubMed, a biomedical literature database, comprises of over 19 million records including MEDLINE citations. Searching for relevant information across these domains is a non-trivial task for two major reasons:

1. The domains are incompatible – The information in these domains is stored and expressed in different document formats, some of which are not computationally friendly.

2. The domains are highly distributed – The domains and the sub-domains are very widely distributed across many databases. For example, there are 94 federal judicial districts and 13 Courts of Appeal in the U.S., with data spanning across multiple silos of databases. Scientific literature is spread out even more widely with thousands of journals and conferences each having their own database.

Although tools exist to help users search across selected set of databases, little effort has been made to semantically correlate such diverse and heterogeneous documents beyond a keyword-based approach. The framework proposed in this paper attempts to provide such an integrated approach of retrieving relevant documents from across these different domains. We develop a use case in the biomedical area – erythropoietin, a hormone which regulates the production of red blood cells. To illustrate the proposed approach, we test the framework using two specific document databases, namely patent documents and court cases.

This paper is organized as follows – Section 2 discusses some common challenges associated with these domains and related work in the area. Section 3 introduces the use case and discusses the corpus of data. Section 4 presents the framework and Section 5 discusses tests and results based on the application of our framework to the use case study. Section 6 briefly states the continuing efforts and the future work.

2 Background and Related Work

In this section, we provide some background information on the two domains of interest – patents and federal court cases involving litigated patents. We briefly discuss the challenges associated with these domains. Currently available tools and other related work is also presented.

2.1 Patents

There are over 40 different patent issuing authorities in the world, such as the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO) and the Japan Patent Office (JPO). This results in a set of large and distributed databases. The USPTO alone has over 7 million issued patents (Site 1). In 2009, 485,312 patent applications were filed with the USPTO. An obvious challenge is integrating these databases together. Derwent World Patent Index (DWPI) is a manually indexed database of around 39 million patent documents from 41 different issuing authorities, created and maintained by Thomson Reuter's (Site 2). Several advanced online tools and systems exist for analyzing and searching patents (Sites 2 to 5). Due to the increasing number of patents, standard

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

document retrieval techniques tend to yield poor search results, even within a confined domain. As such, there is a need for improved search tools that are capable of more precise document retrieval.

2.2 Court Cases

Whether a patent or a set of patents have previously been challenged in court is a very important piece of information for various users. Patent examiners could use this information to deny or approve a patent application; patent applicants could re-write their claims so that they are distinguishable over other's works and so on. There are 94 US District Courts and 13 Courts of Appeals based on Jurisdiction. Each of these courts maintains a separate database. As in the case for patents, one of the key challenges is the integration of these databases. PACER (Public Access to Court Electronic Records) is one such system to access all case documentation (Site 8). However, PACER does not provide keyword-based search; as a result, users need to know the details such as the case number, the case class and other information. Another challenge is that PACER does not provide the documents in a text format; rather, search results of cases are rendered as images. Significant processing is required to be able to work with the text from these documents in image format. Other options for downloading court cases include LexisNexis and Westlaw (Sites 9 and 10).

2.3 Related Work

The claims of a patent define the invention in technical terms and the scope of protection sought by the inventor. Patent claims are considered the most important part of an issued patent. It is therefore necessary to process these claims into a more readable form and focus on the content of the claim. Various natural language techniques have been used to analyze patent claims [15], [16], [19]. In [15] is discussed a methodology to parse patent claims in order to set focus on the structure and key content of the claim. Another method to improve the readability of patent claims is proposed in [2]. These methods can be used in information retrieval engines that are claim-centric.

In [9], a cluster-based language model is developed for information retrieval tasks. The patent documents are clustered based on their International Patent Classification (IPC) codes. Another method makes use of patent citations as a basis of document retrieval [3]. Examiners are required to intensively search for prior art when examining a patent application. In [23], an entire patent application is converted into a query based on features and fields such as the title, abstract, primary claim etc. A bio-specific patent retrieval engine that makes use of domain specific annotations is described in [13]. Several other techniques are based on ontologies and semantic web approaches [2], [4]-[6], [10], [11], [17], [18], [21], [24]. Information retrieval methods for case laws are presented in [7] and [14].

All of the methods help in patent retrieval and analysis. However, little effort has been made to integrate these diverse documents. One of our goals which is out of the scope of this paper is to develop a formal standardized representation of these diverse domains to facilitate the integration of the information they contain. We review the related work done in the field of information integration of diverse knowledge sources. Wache et al. provide a detailed comparison of many information integration approaches which suggest many ways in which an ontology can be developed [20]. When integrating diverse information sources, it may also be preferable to integrate only parts of the knowledge that are needed by the application instead of developing a single global ontology [12], [22]. The framework proposed in this paper aims to search more than a single domain of documents, and to correlate them. In particular, court cases and patents are analyzed in parallel. The framework utilizes information available in court cases to determine the importance of certain patents, and vice versa. It goes beyond a traditional keyword-based retrieval approach, and specific features of both court cases and patents are considered. Eventually, we plan to use the semantics provided by the integrated ontologies representing the domains, to perform the search methodology presented in this paper.

3 Use Case

Erythropoietin, or EPO in short, is a hormone whose primary known function is to regulate erythropoiesis, the production of red blood cells in the body. EPO is produced in the kidney and liver, and is also known as hematopoietin. Synthetic erythropoietin is used as an external stimulant in the treatment of diseases such as anemia. Anemia is the most common blood disorder in which the body is unable to produce enough red blood cells. Acute anemia could prove to be fatal to the living being, which is why research and development of erythropoietin is important.

Amgen Inc. was the first company to produce the synthetic form of erythropoietin called Epogen. Amgen holds five key patents related to the production of erythropoietin namely US Patents 5618698, 5621080, 5756349, 5955422 and 5547933. A total of 135 related patents are identified by following the inbound and outbound patent citations of these five core patents by Amgen. Several court litigations, involve these five patents and some others, dated back to the late 1980's. These litigations include some major companies in the domain such as Amgen Inc., Chugai Pharm., Hoechst Marion Roussell, Genetics Inc., and the like. Plenty of scientific literature is also available on this use case as over 3000 publications are cited amongst the 135 patents.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

Erythropoietin represents a good use case as its intellectual property rights have been heavily litigated. The amount of relevant patents and court cases is sufficient for us to build an illustrative corpus of our proposed framework. The subject area of EPO is non-trivial, with synonyms and non-standardized nomenclature which makes for an interesting use case where a traditional keyword-based approach is likely lacking.

3.1 Bio-ontologies

The field of bio-medicine is growing at a very fast rate. There is a constant introduction of new terms and it is getting increasingly hard to keep up with them. Domain specific ontologies are one initiative to produce a controlled vocabulary. These ontologies act as a knowledge base for tools and methods used especially by bio-informaticians and other IT professionals. An ontology is a formal representation of a domain in terms of its concepts, entities, relations and properties. Domain specific ontologies have been used as the backbone of many existing information retrieval systems. GoPubMed uses Gene Ontology (GO) and MEdical Sub Headings (MESH) as a driver for their information retrieval engine (Site 6).

Typically an ontology is created to meet the needs of a specific domain. Different ontologies hence look at a particular concept in different perspectives. For example, the gene ontology has three organizing principles – the concept as a cellular component, as a biological process or a molecular function. Figure 1 shows *erythropoietin receptor binding* as a part of the gene ontology. The properties of each of these concepts provide more information such as the synonyms, definitions and number of children concepts.

Some ontologies are very general, while some are very specific. When using ontologies to support the process of information retrieval, it is useful to have more than a single ontology which provides a broader and more comprehensive term base. BioPortal is an initiative started by the National Center for Biomedical Ontologies, which provides a common interface to search and download over 130 frequently updated ontologies (Site 7). In the following sections of this paper, we discuss how these ontologies have been used and how they could help improve the performance of our system.

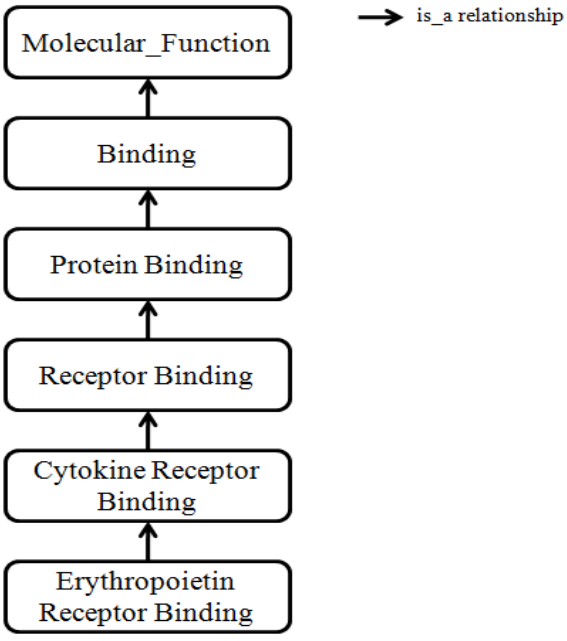

Figure 1: Erythropoietin as a concept in gene ontology (Site 7)

3.2 Databases

The corpus used for testing our framework consists of patent documents and court cases related to our use case of EPO. First, we build our corpus by gathering all potentially relevant documents. We start by identifying all relevant concepts to erythropoietin by searching for erythropoietin on BioPortal, which results in a set of around 11 ontologies. From these 11 ontologies, 43 concepts are extracted by following common relations such as parent, subclass etc. Only the children, parents, grandparents and synonyms of the starting concept erythropoietin are considered. For example, from the ontology shown in Figure 1, we extract the concepts *erythropoietin receptor binding*, *cytokine receptor binding* and *receptor binding*. Going beyond the grandparents results in very general concepts, such as the concept *protein binding*, which do not help the information retrieval process. For each of these 43 concepts, the top

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

50-100 matching patents are downloaded to form a database of 1150 patents. In addition, 135 patents are identified as being closely related to our use case by traversing the inbound and outbound citations for each of the five core patents as previously mentioned in this Section. These 135 patents are also included in the database of 1150 patents. A script is written to automatically download the documents from USPTO.

Court cases usually do not include the technical details of the invention. Hence, the technical terms used in the text are limited to the key concept and strongly related terms such as its synonyms. We observed that searching court cases for all 43 expanded concepts resulted in documents not related to our use case, driving our search in the wrong direction. We therefore focus on the concept erythropoietin and its synonyms such as EPO which resulted in 30 court case documents. These documents were downloaded manually by searching LexisNexis for litigations involving the keywords erythropoietin and its synonyms. The court cases span across almost 25 years from the late 1980s to 2009. They involve major companies such as Amgen, Hoechst Marion Roussel, Chugai Pharmaceutical, etc.

Patents and court cases are available in different formats and need to be re-formatted into a compatible format to ease computation. Various fields, such as inventor, assignee, and such, also need to be extracted and encapsulated in a machine understandable format. The choice for the common format is XML. The patent documents are available in HTML from the USPTO. These documents are very standard in their structure. A script is written to parse the HTML documents and extract relevant fields. Using the XML markups as shown in Figure 2, a new XML document for each of the 1150 patents is created. Similarly, court cases are also re-formatted into XML.

Apache Lucene is a text mining library available under the Apache Software License. Lucene uses the standard vector space model to represent documents and provides libraries for tools such as stop word filtering and stemming (Site 11). At its core is a strong and effective scoring model which is based on a term frequency and inverse document frequency approach (tf-idf). Tf-idf maximizes the score of a term which is frequent in a single or a small set of documents and infrequent across the other documents in the database. We use Lucene to create a searchable index for our documents. Instead of indexing the entire document, the index is created field-by-field as shown in Figure 2 to provide a more granular control and filtering capabilities for our search.

```
<?xml version="1.0" encoding="UTF-8"?>
<PATENT>
        <title> Production of Erthropoietin </title>

        <assignee> Kirin-Amgen, Inc. </assignee>

        <inventor> Lin Fu-Kuen </inventor>

        <pat> 3033753 </pat>
        <pat> 3865801 </pat>
        <pat> 4237224 </pat>
        <pat> 4254095 </pat>
        ...

        <inPat> 7645733 </inPat>
        <inPat> 7629163 </inPat>
        <inPat> 7625756 </inPat>
        ...

        <pub> Nucleotide Sequence of a Bovine Clone Encoding the Agiogenic Protein,
        Basic Fibroblast Growth Factor </pub>
        ...

        <claim> A pharmaceutical composition comprising a therapeutically effective
        amount of human erythropoietin and a pharmaceutically acceptable diluent,
        adjuvant or carrier, wherein said erythropoietin is purified from mammalian cells
        grown in culture </claim>
        ...
</PATENT>
```

Figure 2: Sample of US patent 5955422 in XML format

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

4 Proposed Framework

In statistical terms, recall is the measure of completeness. It measures the number of relevant documents retrieved by a searcher divided by the number of all relevant documents. Recall is an important metric for patent research since a complete and comprehensive coverage of all relevant literature is often the deciding factor in IP litigations. However, it is difficult to attain a reasonable recall rate as literature exists in siloed databases using non-standardized vocabularies. In this context, recall is valued over precision, which measures the exactness of a searcher. Hence, the first step taken is to maximize recall by expanding the user query through domain specific ontologies. There is a lot of information shared between court cases and patents. Relevancy measures can be improved based on the shared information by correlating the documents. The next step is to rank and re-order the results centered on these relevancy measures. The framework proposed in this paper is a multi-step process, which effectively funnels down from a large set of results to a smaller set of highly relevant documents. The framework is shown in Figure 3. Our initial investigation focuses on the first three basic steps explained in Sections 4.1 through 4.3 below.

4.1 Step 1: Keyword Expansion

With the rapid expansion in the field of biotechnology, there is an increasing lack of standardized terminology. This makes it hard to search for documents based on a simple keyword based model. Bio-ontologies are domain specific knowledge bases which attempt to control the terminology and relate one concept to another. They define concepts, relations and properties parsing through which will provide more information, more keywords increasing our chances of retrieving relevant documents. As mentioned in Section 3.1, we use BioPortal's common interface to over 130 bio-ontologies to expand the user query.

When choosing which concepts to extract from bio-ontologies, it is important to consider the type of the document that is being dealt with. Court cases include fewer technical details when compared to patent documents. The extent of technical terminology used in the court cases is limited to the key concept and its synonyms as explained in Section 3.2. Expanding the query to concepts which are more distantly related to the key concept can lead to retrieval of court cases centered around irrelevant concepts. This is avoided by limiting the expansion of the key concept to its synonyms only. Patent documents on the other hand provide a detailed technical description of the invention and hence to obtain a broader coverage, the key concept is expanded beyond just its synonyms.

4.2 Step 2: Independently Search Databases

In this step, we use the expanded term base to search through the databases independently. In our proposed framework, patent documents are indexed based on various fields such as title, abstract, claims etc. We search for the occurrence of these new formed terms in various combinations of fields such as {title}, {title, abstract}, {abstract, claims}, etc. This is done in order to capture the importance of the document with respect to the term for all relevant fields. For example, if the term erythropoietin occurs in the title, the abstract and claims, it could imply that the document is strongly related to erythropoietin, whereas if erythropoietin only occurred once in the references cited, the document may not be as relevant. The goal of this stage is to maximize the recall to guarantee completeness.

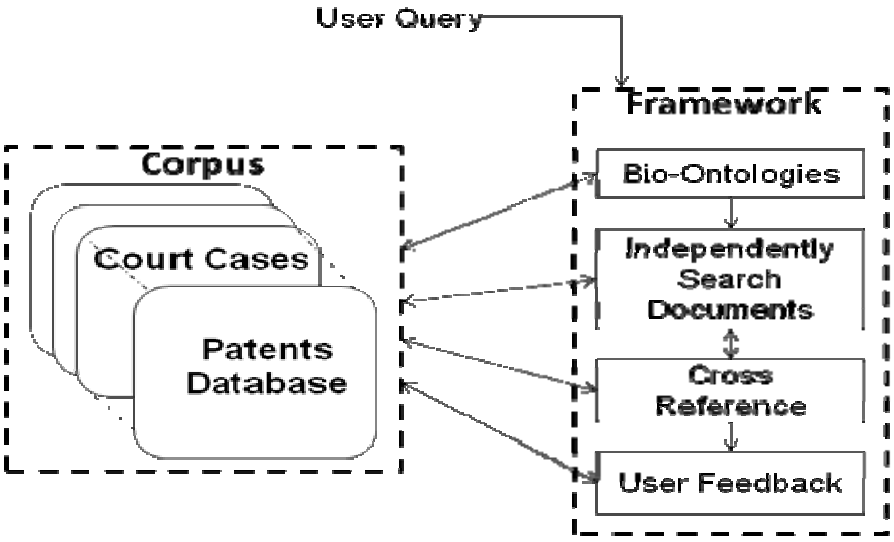

Figure 3: Proposed framework

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

4.3 Step 3: Cross-reference Domains

Each of these documents, patents and court cases, are created and maintained by completely independent agencies or organizations. Although they are available in incompatible formats across distributed databases, there is plenty of information which could inter-relate these documents. Since the goal of this framework is to return a set of related documents from various domains, this step is very crucial. Cross-referencing in our framework aims to extract key features or fields from these documents and improve relevancy measures by identifying the correlation between these key fields.

5 Methodology and Results

In this section, we present our methodology and results for the proposed framework described in Section 4.

5.1 Baseline Reference

In our use case, a patent attorney or patent agent will likely begin his/her research by using the keyword *erythropoietin* to search through the patent database. The result from this search will act as a baseline reference for the remaining tests. We will use the 135 patents identified by following inbound and outbound citations of the 5 core patents as the true positive to calculate precision and recall values. Our goal is to improve the precision and recall values with respect to the 135 patents. In addition to precision and recall computation by using the 135 patents as the gold standard, we are also interested in retrieving the five core patents in the result set since they are the most important patents in our use case. As a result, we will compute the average rank of these five core patents to further evaluate the effectiveness of our system.

The patent attorney or patent agent would also search the court case database for the keyword erythropoietin. As explained in Section 3.2, court cases are written for more general consumption and thus make use of fewer technical jargons. Hence, a search for erythropoietin alone is sufficient to retrieve all of the 30 court cases relevant to our use case from the database resulting in a recall of 100%.

Table 1 shows part of the baseline results. Notice that two documents out of the seven shown have a retrieval score of 0.0; in other words, they are not retrieved by the baseline search of *erythropoietin*. This is because some of the relevant patents do not make use of the concept erythropoietin, even if the patents are related to the general topic of erythropoietin. We have observed that there is a huge variance in the use of biotech terminologies in patents. The recall for this benchmarking search result is about 67%, which means only 90 documents out of the 135 are retrieved. The average rank of the five core patents in these set of results is found out to be 36.6.

Table 1: Baseline reference, max recall

5955422	2.316
6204247	0.000
6245740	0.025
6270989	0.317
6280977	0.036
6340742	8.010
6420339	0.000

Average Rank of the five core patents = **36.6**
Recall = **0.67**

5.2 Steps 1 and 2: Keyword Expansion and Searching Independent Databases

As shown in the previous section, recall in patent documents is relatively low when using a simplistic search strategy of keyword *erythropoietin*. One way to improve recall is to use the available ontologies to expand the query to include related concepts, such as the acronym *EPO*. However, one issue that arises is that some concepts are more general than others. We must carefully choose the terms used to search through the document corpus, as very general terms provide little or no information and tend to reduce precision as a result. For example, *protein* is a more general term than *erythropoietin*. A search for the keyword *erythropoietin* in the entire document on USPTO returns over 7000 documents, where as a search for a term such as protein returns more than a 150,000 documents. This implies that the size of the result set after the expansion of the query is larger than the baseline result set. In order to differentiate the strongly related terms from the more general ones, we calculate a weighted sum to be able to boost the weight for relevant terms and lessen the weight of more general or loosely related terms. To achieve this, we test four heuristic weighting functions. All four functions are devised to attenuate the weights as we branch away from the

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

original concept *erythropoietin* to more general concepts such as the parents and grandparents. The weights have a value between 0 and 1 according to the weighting functions shown in Figure 4. As suggested in Section 3.2, only parents, grandparents, children and synonyms of the starting concept *erythropoietin* are considered. We achieve a recall of 100% by expanding the query using relevant concepts from bio-portal ontologies. This means that all of the 135 documents were retrieved from the database of 1150. As such, even though we achieve 100% recall, the precision is on the lower side. Since the more general concepts are given a much lower weight, the weighting function also helps us improve the precision.

The 43 terms that are derived from the bio ontologies are used to search the title, abstract, claims, description and the entire document. As mentioned in Section 3.2, the documents are indexed using Lucene by the different fields which in our case are *title*, *abstract*, *claims*, *inventor* and so on. Lucene's score computation normalizes its document vector, which means that a higher score is assigned to a term occurring in a smaller field or with fewer words. Hence, if a keyword occurs in the title of a document, it is given a higher weight than when it is found in the description. In addition, we have observed that certain fields, such as title and abstract, tend to contain important information about the document. If a keyword occurs in the title, abstract and claim, it is more likely that the document under concern is of relevance when compared to a document where the keyword appears only in the description. Therefore, we also compute the score of various combinations of important fields such as {title and abstract}, {title, abstract and description} and so on. This is achieved by iteratively computing the scores for each combination of the fields in the patent index, for each of the 43 concepts. The score for each concept is the sum of the scores for each combination of the above fields. The final score of a patent document is a weighted sum of the scores for all the 43 keywords according to the selected weighting function from Figure 4. In general, for N concepts, this can be expressed as:

$$Final\ Score = \sum_{i=0}^{N} W_i * Score_i \quad (1)$$

Where W_i is the weight assigned to the concept from the weighting function, depending on whether it is a synonym, parent, grandparent or a child, and $Score_i$ is the accumulated score for that concept searched in the above mentioned combination of the fields – title, abstract, claims and description.

Table 2 shows the part of the results obtained from the use of bio-ontologies to expand our search. The patent numbers in colored cells are amongst the 135 relevant patents.

Table 2: Results after expansion with the bio-ontologies

Patent Number	Unweighted Score	Patent Number	Weighting Function-1	Patent Number	Weighting Function-2	Patent Number	Weighting Function-3	Patent Number	Weighting Function-4
7067477	25.249	**7067477**	24.337	5712370	22.463	5712370	21.548	6048971	19.279
7550433	25.135	5712370	23.825	**7067477**	21.784	5278065	20.388	5712370	18.935
6932968	25.034	5874224	23.739	5278065	21.586	**7067477**	20.369	**5955422**	18.444
5712370	24.364	**7550433**	23.582	4954437	20.787	4954437	20.114	5278065	17.682
5625035	24.360	**6696411**	23.378	**6696411**	20.593	**5955422**	19.598	4954437	16.900
6696411	24.350	5625035	23.142	6489293	20.584	6489293	19.415	**7067477**	16.839
5843726	23.619	5278065	22.414	6998124	20.436	**6696411**	19.112	6489293	15.996
6043211	23.159	5843726	22.293	5625035	20.257	6998124	18.646	**6696411**	15.697
5772992	23.067	6489293	22.257	**5955422**	19.934	**5441868**	18.366	**5441868**	15.221
5278065	22.940	6998124	21.645	6153190	19.329	5625035	17.996	6998124	14.936
6489293	22.934	4954437	21.512	**7550433**	19.175	**4667016**	17.950	5625035	14.564

Results are shown in Figures 5 and 6. Figure 5 shows the average rank of the five core patents with respect to the four functions. Function-1 gives a significantly high weight of 0.9, 0.8 and 0.7 as we branch away from the original concept. The average rank of the 5 core patents is about 74, which is greater than the baseline results of 36.6. We further attenuate the weights for the child, parent and grandparent in Functions 2, 3 and 4. The average rank of the five core patents is around 40, which is almost comparable to the baseline result of 36.6. The F-measure is a measure of a test's accuracy (Equation (2)). It combines the values of both precision and recall measures. It is defined as the harmonic mean of the precision and recall.

$$F - measure\ or\ F - value = 2 * \frac{precision * recall}{precision + recall} \quad (2)$$

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

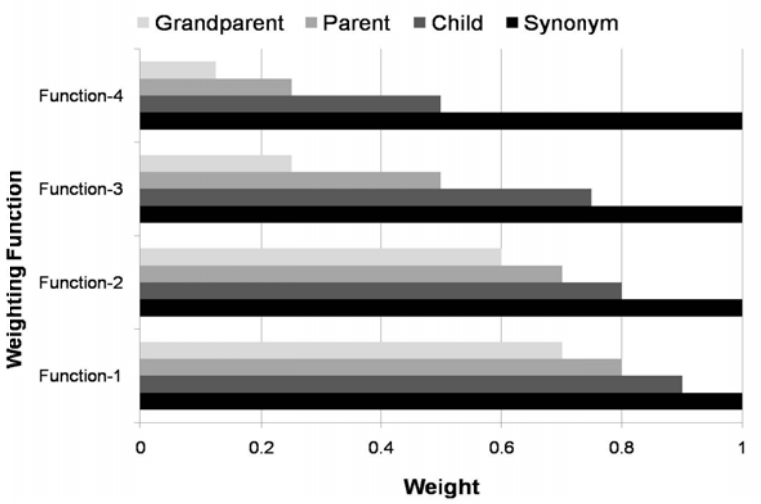

Figure 4: Four heuristic weighting functions

Figure 6 compares the F-value for each of the four functions with respect to a cut-off value. The cut-off value varied from 10 up to 250 in steps of 10. It represents the top results until the cut-off values. Function 3 gives us the best performance with a peak F-value of around 34%. The recall has improved by a significant margin from 67% to 100% when compared to the baseline results. Since Function 3 performs the best in both the average rank of the 5 core patents and the average F-measure, we will use this weighting function in the next section on cross-referencing patent documents and court cases.

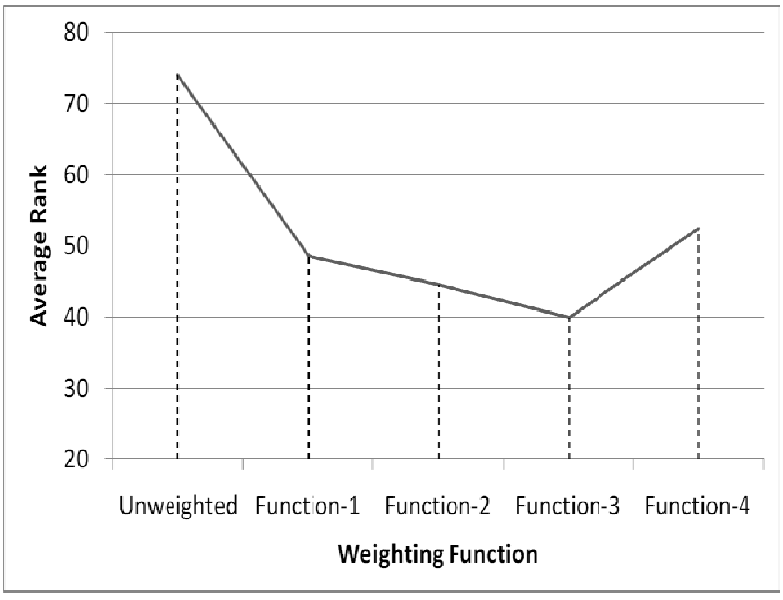

Figure 5: Average rank of the five core patents

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

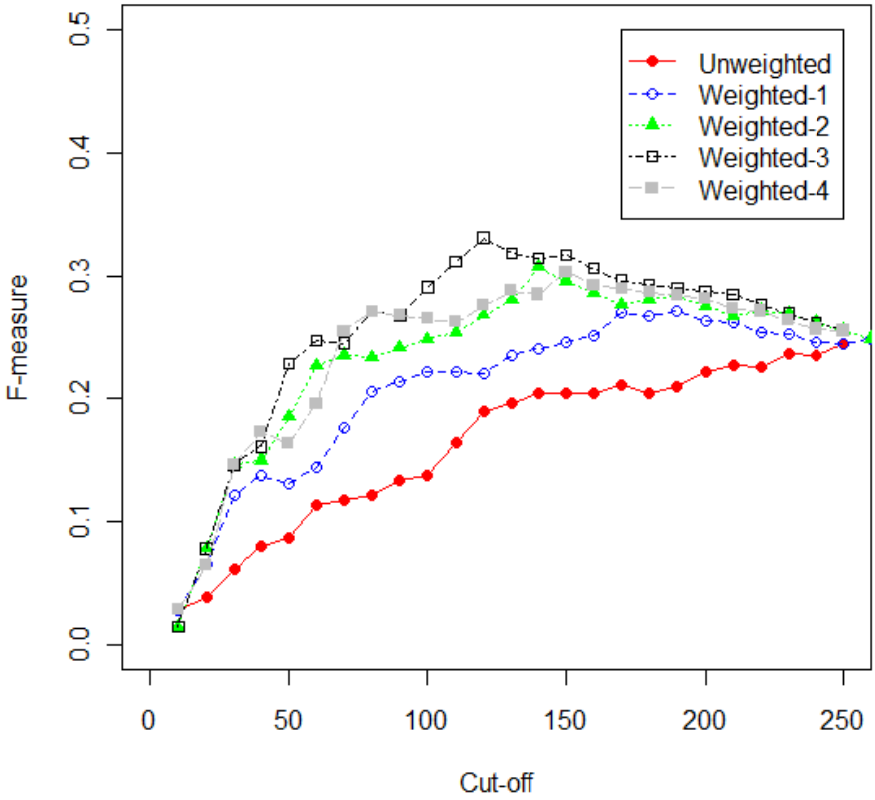

Figure 6: F-measures for the four weighting functions

5.3 Step 3: Cross-referencing Patents and Court Cases

A user researching a particular technology would be very interested in the patents related to that particular technology, especially those patents whose validity has been challenged in courts. A start up firm could be interested to know what litigations its competitors are involved in. Court cases provide such information for us to correlate patents that are litigated. Figure 7 illustrates the relevance between patents and court cases through matchup of different fields.

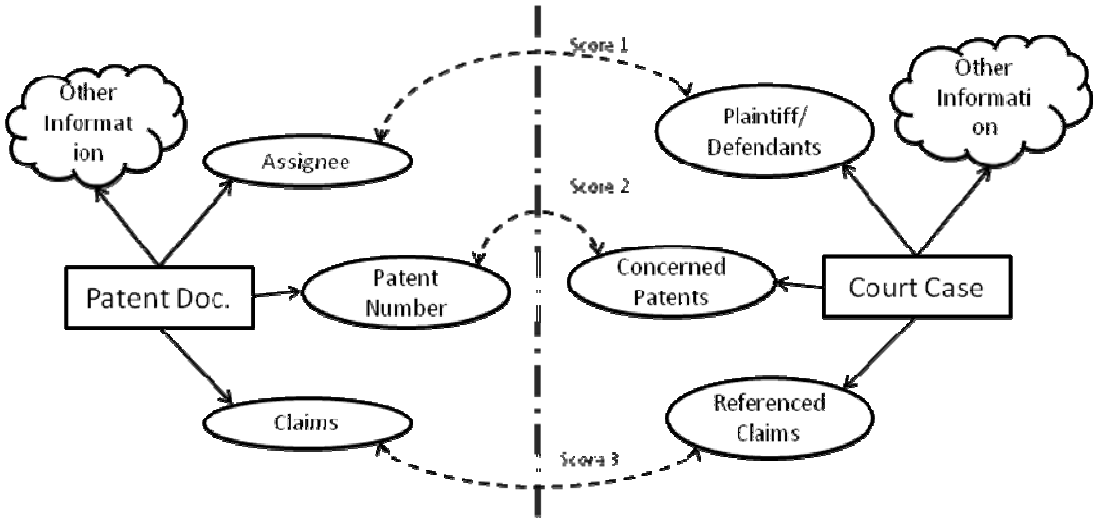

Figure 7: Co-Relating patent documents and court litigations

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

Court cases generally include the patent numbers of the patents being challenged. These patents can be directly looked up from the patent database; as such, score 2 in Figure 7 establishes the linkage between the two domains of patent and cases. In addition, we have observed that either the plaintiff or defendant or both are likely to have one or many patents. Score 1 in Figure 7 illustrates the linkage between a patent assignee and the plaintiff/defendant field in court cases. The last piece of metadata that is useful in correlating patents and court cases is the claims. The text of court cases usually state the claims under concern, which could be used as the third measure of relevance as shown as score 3 in Figure 7.

We would like to take advantage of these fields to boost the scores of the concerned patents, and the ones related to those. This section focuses on methods which cross-references these fields with the patent database to improve the precision and recall numbers. Pre-processing these documents is required in order to be able to easily capture the contents of these fields, such as claims and plaintiffs/defendants, as described in Section 3.2.

5.3.1 Computing the Score

From all of the 30 court cases retrieved using the keyword *erythropoietin*, we extract the patent numbers that are cited and count their frequency of citation. From this we derive a probabilistic weight for each patent number. This can be considered a measure of importance, i.e., if a patent A is mentioned far more times than patent B in all the court cases put together, it is likely that patent A is more relevant to our use case than patent B. From all the court cases that are retrieved (including the false positives) the patent numbers of the patents involved in the litigation are extracted. Let *n* be the total number of times a particular patent, say US x, xxx, 345 has been cited. Let *N* be the total number of times all the patents have been cited. Then, *Weight = n/N.*

Table 3: Probabilistic score associated with each patent number

Patent Number	n	Weight = n/N
4,703,008	15	0.107
5,441,868	8	0.057
5,547,933	18	0.128
5,618,698	18	0.128
5,621,080	13	0.092
5,756,349	15	0.107
5,955,422	19	0.135

The total number of patent numbers retrieved from this set of court cases is 12. Table 3 shows the scores for some of the patents involved in the retrieved court cases. The five core patents are involved in most litigations and hence hold a high level of importance. These are shown in bold.

A similar approach can be followed to associate a weight to the plaintiff/defendant of court cases, which are directly related to the assignee field of patents. These scores are shown in Table 4. Notice again that Amgen Inc. is highly involved in these litigations.

Table 4: Probabilistic score associated with each patent number

Plaintiff/Defendant/Assignee	Count	Unweighted
Amgen Inc.	15	0.405
Chugai Pharmaceutical Co. Ltd.	3	0.081
Hoechst Marion Roussel Inc.	5	0.135

Assignees such as Amgen hold not only erythropoietin related patents, but many others. For example, around 743 patents are assigned to Amgen, only 144 of which have the occurrence of the term erythropoietin (Site 1). Using the plaintiff/defendant field from the court case, to boost the value of the patents by assignee could result in very general results, deviating from the topic of concern. Hence, we use the patents as weighed in Table 3 and find other patents which have a high relevance to them.

5.3.2 Document Similarity

As mentioned in Section 2.1, the number of issued patent documents is over 7 million in the US. Of these 7 million patents, there are an estimated 1.5 million patents in classes that broadly fall under the bio-medical domain (Site 1). A bag of words model comparing the entire patent would return a large set of results upon any bio keyword search, which becomes a laborious task to sort through the false positives. Hence, there is a need to be able to compare

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

documents by establishing similarity metrics in other information available in patent documents. We followed the inbound and outbound citations from the five core patents to identify 135 closely related patents to our use case. Others approaches to identify relevant patents involve meta information such as references, the inventors etc. The method discussed in this section establishes relevancy between a set of patent documents by computing the similarity on the basis of patent metadata namely the inventor, assignee, in and out patent citations, references to scientific publications and the claims. This method reorders the results obtained from Section 5.2 with respect to their similarity to the patents as mentioned in Table 3. Essentially, our system aims to produce a better ranking of results by pushing the more relevant patents onto the top of the result stack.

The claims of any patent document describe the scope and legal protection sought by the inventors for that invention. Two patents with fairly similar claims could be of equal interest to a user and hence establishing relevance with respect to these claims is critical. Citations are another good indicator of similarity between two patents. For instance, two patents referencing same or similar publications are likely to be relevant. A detailed analysis on patent citations is discussed in [1], [8] which suggest that patent citations can possibly suffer from local bias such as self-citations. Hence, although patent citations lead us to other relevant patents, they should not be the only indicator of similarity. The inventor and the assignee fields can be considered as a tuning factor in the similarity. Searching for patents only by inventors would most likely produce a small result set, but it gives some extra control to the user if one wishes to give a higher importance to this field.

The algorithm takes one patent as the starting point, and scores every other patent with respect to it. The selection of this patent is explained in Section 5.3.3. The score corresponding to the claims, references, assignee and the inventor are the cosine scores of similarity. The standard vector space model is used for this purpose. For computing the score of the patent citations, we take the ratio of the total number of matching citations in the two documents to the total number of citations in that document. Each of these fields are individually compared and scored for a patent document. The final score is a weighted sum of the scores indicating the similarity of the patents:

$$Final\ Score = \sum_{i=0}^{N} \mathrm{W_i} * \mathrm{Score_i} \tag{3}$$

Where *Field(i)* is one of the six fields – inventor, assignee, inbound citations, outbound citations, literature references and claims. *W(i)* is the corresponding weight attached to it. The weights are a value between 0 and 1 which can be adjusted to tune the results to the requirement. The results obtained can be considered a 6-dimensional representation of every patent document, with the magnitude on each dimension representing its similarity to the patents extracted from the court cases in that dimension.

5.3.3 Results and Analysis

We establish that the patent citations extracted from the court cases are important. For each of these patents, the document similarity method discussed in Section 5.3.2 is used to score the results from Section 5.2. The final score for a patent is the weighted sum of its similarity score for each of the patent numbers extracted from the court cases. Let patent A and patent B be the patents extracted from the court cases and *p1*, *p2* be their corresponding probabilistic weights from Table 3. Let the similarity score for a third patent C to patents A and B be *ScoreA* and *ScoreB* respectively. Then the final score of patent C will be:

$$ScoreFinal = p1 * ScoreA + p2 * ScoreB \tag{4}$$

We start by assigning a weight of 1 to each dimension of comparison, i.e., we include all fields with equal weight in the similarity method. The precision, recall and F-measure are shown in Figure 8. The precision and recall values drastically improve especially in the initial range. We achieved a high recall in step-2 of the framework as well, although now it reaches the value at a lower cut-off. All the 135 related patents are retrieved in the first 400 results. The F-value has improved from around 34% to a peak of around 63%. We do not completely rely on following backward citations and forward citations as we did to identify the 135 relevant patents, but we also consider other factors, such as claim, assignee, and etc, equally. From our preliminary result, establishing relevance or similarity with respect to these extra fields is shown to produce a better result set. All five core patents are now ranked in the top 7 with an average rank of *3.4*.

User intent in our system can be varied. The sheer amount of knowledge and the vast expansion in research areas of science and technology has made it hard to focus the search on a highly specialized topic. In other words, users have such diverse requirement that it is rather difficult to come up with a single similarity metric that would produce exactly what the user requires. In fact, precision and recall are measures that can vary per user's requirement. The document similarity method accounts for this diverse requirement by letting the user choose the basis of comparison; for example, all 6 dimensions with equal weight, or focusing on just a single dimension such as claims. Figure 9 shows the results for when claims are the only factor in calculating the similarity of the documents. The precision and recall values are inferior to the optimal result when all 6 dimensions are considered. However, we believe that it is important for our system to allow for the flexibility to compare only certain dimensions, such as claims, since patent searches are likely performed by a wide range of users with different intents.

Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 6 / ISSUE 2 / AUGUST 2011 / 1-16

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

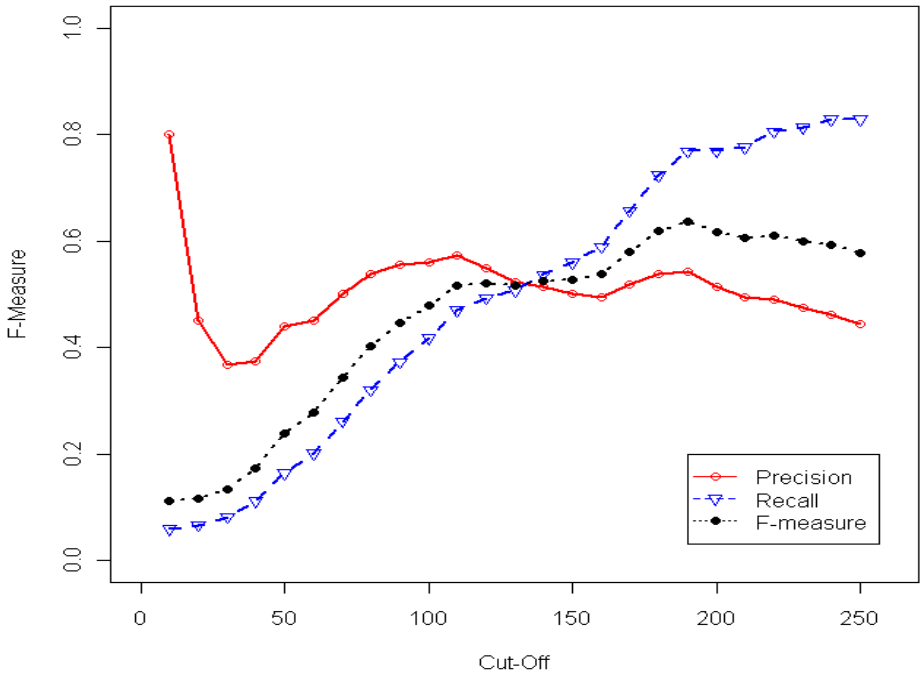

Figure 8: Comparing all the six fields of patent documents (with equal weights)

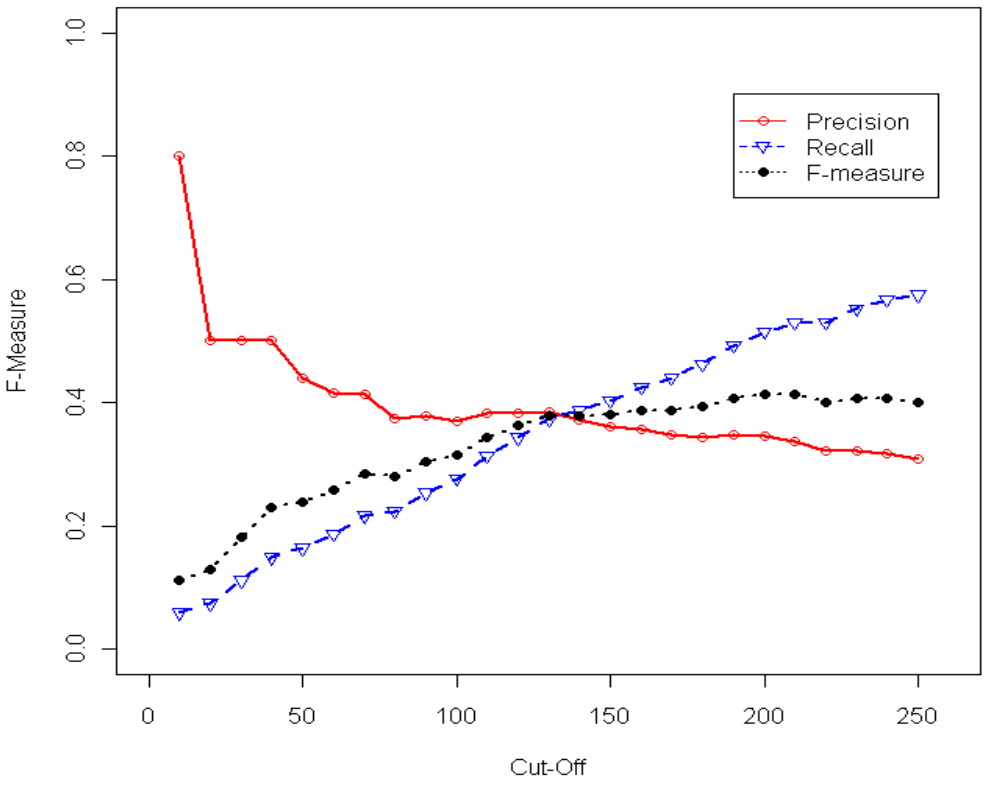

Figure 9: Considering only claims of patents

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

6 Conclusion

In this paper, we present a framework to help retrieval of relevant documents from two independent domains - court cases and patents. We demonstrate our results by constructing a use case in the field of biotech, erythropoietin. The test corpus consists of 1150 patents and 30 court cases relevant to the use case.

In the baseline approach, many relevant documents are not retrieved by simply searching for the keyword *erythropoietin*. The recall for this benchmarking search is 67%. Expanding the user query based on bio-ontologies helps improve recall to 100%. We then focus on establishing relevance between court cases and patent documents. We use the information available in the court cases to identify patents which have been challenged in court as they are likely to be important. Using the document similarity method described in Section 5.3.2, the patents are then re-ranked based on their similarity to the important patents identified from the court cases. The final results have a recall of 100% when compared to a recall of 67% achieved in the baseline test. In addition, the final results have a peak F-measure of 63% which is a 29% increase with respect to the results obtained in Section 5.2.

The methodology described in this paper heavily relies on multiple iterations of computation over the documents and deals with multiple databases. On a 2.5 GHz dual-core machine, running the entire process on the set of 1150 patent documents and 30 court cases takes a little over 15 minutes. The number of patent documents on USPTO and US court cases is very large when compared to the size of the corpus used to demonstrate the results of the methodology explained in this paper. Since our primary goal is to provide a framework which provides a valuable set of results, this paper does not study how this methodology will scale to larger databases. Although the results are very encouraging, the framework could prove to be less advantageous to a user whose search requirements are time critical, especially on a larger corpus. An obvious step to improve the performance would be to use faster and more resourceful hardware. However, this is an area which will have to be further studied for potential performance improvements. The issue of automatically retrieving court case documents from PACER or LexisNexis is however unsolved and an alternative will have to be worked out in order to expand the scope of the framework beyond the use case.

As mentioned in Section 1, information pertaining to science and technology is deeply buried under the regulatory system. In order to gather relevant information pertaining to a subject, one has to search many heterogeneous information domains which broadly include (1) agency documents such as the FDA and the USPTO; (2) scientific literature; (3) court litigations; and (4) relevant laws and regulations. The impact of this research falls on all the entities which create and manage this information, allowing them to communicate and share the information between them. Specifically the work presented in this paper deals with patent documents, court cases, scientific publications and patent file wrappers. The work presented in this paper impacts a wide variety of users ranging from smaller companies and individuals, lawyers to patent examiners amongst others. The ability to search and retrieve documents across multiple domains makes the process of gathering relevant information a much less daunting task.

6.1 Future Work

USPTO file wrappers hold the entire application history of a patent. Key information regarding related prior art can be obtained from file wrappers by studying the office actions and amendments in the document. The difference in the initial claims as filed, and the claims as finally issued can reveal significant information about the invention. Other sets of documents such as scientific/technological publications and agency regulations are also sought to be incorporated into the framework. While cross-referencing information from such diverse domains is a difficult task, it must be noted that the framework will provide a much stronger reasoning ability with a view into all the documents. Hence, one of the primary directions of this research is to incorporate more document domains to build stronger relevancy measures.

User requirements are very diverse. The final step in this framework, as shown in Figure 3, aims at including the user relevancy feedback, and will help tune the results per the user's requirement. Our ultimate goal is to develop an ontology or multiple ontologies which will formally represent all relevant documents, including patents, court cases, file wrappers and scientific publications, as well as various relevancy measures as concepts, relations and properties. This ontology will be populated with instances of the actual documents to build and function as a knowledge base. The framework described in this paper will rely on the knowledge base as the backbone for its information. However, it must be noted that creating an ontology for such a diverse set of documents is a challenging task by itself.

Acknowledgments

This research is partially supported by NSF Grant Number 0811975 awarded to the University of Illinois at Urbana-Champaign and NSF Grant Number 0811460 to Stanford University. Any opinions and findings are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

Websites List

Site 1: USPTO
http://www.uspto.gov/

Site 2: Thomson Innovation
http://www.thomsoninnovation.com/ti/contentsets/patents/

Site 3: esp@cenet
http://www.espacenet.com/

Site 4: Patent Cafe
http://www.patentcafe.com/

Site 5: Dialog LLC
http://www.dialog.com/

Site 6: GoPubMed
http://www.gopubmed.org/

Site 7: BioPortal
http://bioportal.bioontology.org/

Site 8: PACER
http://www.pacer.gov/

Site 9: LexisNexis Academic
http://www.lexisnexis.com/hottopics/lnacademic/

Site 10: Westlaw
http://www.wipo.int/portal/index.html.en

Site 11: Lucene's Scoring Formula
http://lucene.apache.org/java/2_4_0/api/org/apache/lucene/search/Similarity.html

References

[1] J. Alcácer and M. Gittelman, Patent citations as a measure of knowledge flows: The influence of examiner citations, Review of Economics and Statistics, vol. 88, no. 4, pp. 774-779, 2006.
[2] J. Codina, E. Pianta, S. Vrochidis, and S. Papadopoulos, Integration of semantic, metadata and image search engines with a text search engine for patent retrieval, in Proceedings of the Workshop on Semantic Search at the 5th European Semantic Web Conference, June 2008, pp. 14-28.
[3] A. Fujii, 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, 2007, pp. 793-794.
[4] N. Ghoula, K. Khelif, and R. Dieng-Kuntz, Supporting patent mining by using ontology-based semantic annotations, in WI'07 Proceedings of the IEEE/WIC/ACM International Conference on Web intelligence, Washington, DC, November 2007, pp. 435-438.
[5] M. Giereth, S. Brügmann, A. Stäbler, M. Rotard, and T. Ertl, Application of semantic technologies for representing patent metadata, Lecture Notes in Informatics, vol. 94, pp. 297-305, 2006.
[6] M. Giereth, S. Koch, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, L. Serafini, and L. Wanner, A modular framework for ontology-based representation of patent information, in Proceedings of the 2007 Conference on Legal Knowledge and information Systems, 2007, pp. 49-58.
[7] P. Jackson, L. Al-Kofahi, C. Kreilick and B. Grom, Information extraction from case law and retrieval of prior cases by partial parsing and query generation, in Proceedings of the Seventh international Conference on information and Knowledge Management, Bethesda, Maryland, Nov 1998, pp. 60-67.
[8] A. B. Jaffe, M. Trajtenberg, and M.S. Fogarty. (2000, April) The meaning of patent citations: Report on the NBER/Case-Western reserve survey of patentees, NBER Working Paper No. W7631. [Online]. Available: http://www.nber.org/papers/w7631.pdf.
[9] I. Kang, S. Na, J. Kim, and J. Lee, Cluster-based patent retrieval, Information Processing and Management, vol. 43, no. 5, pp. 1173-1182, 2007.
[10] Y. Kitamura, M. Kashiwase, F. Masayoshi, and R. Mizoguchi, Deployment of an ontological framework of function design knowledge, Advanced Engineering Informatics, vol. 18, no. 2, pp. 115-127, 2004.
[11] L. T. McCarty, Deep semantic interpretations of legal texts, in Proceedings of the 11th international Conference on Artificial intelligence and Law, Stanford, California, June 2007, pp. 217-224.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762011000200002

[12] P. Mitra, G. Wiederhold, and J. Jannink, Semi-automatic integration of knowledge sources, presented at the 2nd International Conference on Information Fusion, Sunnyvale, CA, July 6-8, 1999.
[13] S. Mukherjea and B. Bamba, BioPatentMiner: an information retrieval system for biomedical patents, in Proceedings of the Thirtieth international Conference on Very Large Data Bases, 2007, pp. 1066-1077.
[14] E. L. Rissland and J. J. Daniels, A hybrid CBR-IR approach to legal information retrieval, in Proceedings of the 5th international Conference on Artificial intelligence and Law, Maryland, United States, May 1995, pp. 52-61.
[15] S. Sheremetyeva, Natural language analysis of patent claims, in Proceedings of the ACL Workshop on Patent Corpus Processing, Sapporo, 2003, pp. 66-73.
[16] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama, Patent claim processing for readability: structure analysis and term explanation, in Proceedings of the ACL-2003 Workshop on Patent Corpus Processing, Morristown, NJ, 2003, pp. 56-65.
[17] V. W. Soo, S. Y. Lin, S. Y. Yang, S. N. Lin, S. L. Cheng, A cooperative multi-agent platform for invention based on patent document analysis and ontology, Expert Systems with Applications, vol. 31, no. 4, pp. 766-775, 2006.
[18] R. M. Tong, C. A. Reid, G. J. Crowe, and P. R. Douglas, Conceptual legal document retrieval using the RUBRIC system, in Proceedings of the 1st international Conference on Artificial intelligence and Law, Boston, 1987, pp. 28-34.
[19] Yuen-Hsien Tseng, Chi-Jen Lin, Yu-I Lin, Text mining techniques for patent analysis, Information Processing & Management, vol. 43, no. 5, pp. 1216-1247, 2007.
[20] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, Ontology-based integration of information-a survey of existing approaches, in IJCAI–01 Workshop: Ontologies and Information Sharing, 2001, pp. 108-117.
[21] L. Wanner, R. Baeza-Yates, S. Brugmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, and V. Zervaki, Towards content-oriented patent document processing, World Patent Information, vol. 30, no. 1, pp. 21-23, 2008.
[22] G. Wiederhold and J. Jannink, Composing diverse ontologies, presented at 8th Working Conference on Database Semantics (DS-8), Rotorua, New Zealand, January, 1999.
[23] X. Xue and W. B. Croft, Automatic query generation for patent search, in Proceedings of the 18th ACM Conference on information and Knowledge Management, Hong Kong, China, November 2009, pp. 2037-2040.
[24] S. Y. Yang, S. Y. Lin, S. N. Lin, S. L. Cheng, and V. W. Soo, An Ontology-based multi-agent platform for patent knowledge management, International Journal of Electronic Business Management, vol. 3, no. 3, pp. 181-192, 2005.