An Ontology-Based Interactive Tool to Search Documents in the U.S. Patent System

Siddharth Taduri, Gloria T. Lau
Stanford University
Civil and Environmental Eng.
Stanford University
Stanford, CA, USA
staduri, glau@stanford.edu

Kincho H. Law
Stanford University
Civil and Environmental Eng.
Stanford University
Stanford, CA, USA
law@stanford.edu

Hang Yu, Jay P. Kesan
College of Law
University of Illinois at Urbana-Champaign
IL, USA
hangyu, kesan@illinois.edu

ABSTRACT
The past few years have seen an explosive growth in scientific and regulatory documents related to the patent system. Relevant information is siloed into many heterogeneous and distributed information sources making it very challenging to retrieve information across multiple domains. In this demonstration, we present a tool that enables users to simultaneously search multiple information domains in the patent system. The presented tool is built upon the Patent System Ontology, which provides both a standardized representation of the patent system domain and required semantics to integrate the various information domains [1], [2]. The tool provides features such as integration with a biomedical knowledge base and recommendations for related articles. Future additions to the tool will provide features to analyze the data both statistically and visually to aid in research. We demonstrate how this tool can be helpful in expediting search and information retrieval in an intelligent and convenient way through a use case in the bio domain – erythropoietin.

Categories and Subject Descriptors
D.2.13 [**Software Engineering**]: Reusable Software – *Domain Engineering.*
H.3.4 [**Information Storage and Retrieval**]: System and Software – *Question-answering (fact retrieval) systems.*

General Terms
Design, Standardization.

Keywords
Search, Ontology, Patent, Court Cases, File Wrapper, Information Retrieval, Knowledgebase.

1. INTRODUCTION
In the recent years, there has been an explosive growth in the scientific and regulatory information available online. However, information pertaining to a particular subject is maintained by independent entities in the regulatory system, each enforcing different standards which results in a very heterogeneous set of documents segregated into information silos. Therefore, interested parties are required to simultaneously search multiple information silos in order to gather comprehensive information relating to a particular subject in the patent ecosystem. These information silos include (a) patents; (b) scientific publications; (c) court cases; (d) patent file wrappers and (e) relevant laws and regulations. Currently, search tools available in the market such as Thomson Reuter's Delphion and Web of Science provide a good starting point for trained professionals to perform Boolean searches within a single information silo. However, users of such tools belong to different backgrounds (for example – lawyers, startup companies, academicians etc.), who may have varying requirements. Hence, there is a need for a tool which provides a method to collectively search multiple information sources and also caters to the diverse set of users. To tackle the diversity of the documents, we have developed the Patent System Ontology, which standardizes the representation of the relevant document domains and provides the required semantics for the development of intelligent tools [1], [2]. In this demonstration, we present an interactive tool which enables users to simultaneously search and gather information across several diverse information sources in the patent system.

Our current implementation spans three information domains, namely issued patents, court cases and patent file wrappers. As we make progress with these documents, we intend to include other information sources such as scientific publications and regulations. The tool provides features such as integration with domain knowledge and with free text and search libraries such as Apache Lucene and Solr. These features are discussed in Section 3. We are currently focusing on the biomedical domain due to the vast resources available such as domain ontologies. Through a use case in the biomedical domain, erythropoietin, we demonstrate how this tool can be helpful in expediting search and information retrieval in an intelligent and convenient way. The presented tool will provide a general approach that can scale and be applied to other domains.

2. USE CASE
Erythropoietin is a hormone responsible for the production of red blood cells in the body through a process known as erythropoiesis. The use case involves several important patents which have been heavily litigated in the past two decades. In addition, the patents collectively cite many publications and each patent has a corresponding file wrapper. Hence, the use case provides us with a set of documents that are diverse and span several information sources. We identified five core patents

Dg.o'11, June 12–15, 2011, College Park, MD, USA.

related to the production of erythropoietin, namely U.S. 5,547,349, 5,618,698, 5,621,080, 5,756,349 and 5,955,422. Using the 5 core patents as a starting point, we built the corpus of documents following citations and keywords searches. Overall, the corpus consists of 1150 U.S. patents collected from the U.S. Patent and Trademark Office (USPTO) database. We collected around 30 U.S. federal and district court cases from the LexisNexis online database. The patents collectively cite over 3000 publications which can be found in PubMed, an online resource that contains over 20 million citations for biomedical literature. The file wrapper for U.S. patent 5,955,422 is also included in the corpus.

3. FEATURES

Once the ontology is populated, it can be queried to infer and extract useful information. To query or visualize the ontology, an ontologist may use tools such as Protégé. However, the use of such tools can be cumbersome and complicated and involves a steep learning curve for using the tool itself. The tool provides a convenient interface for searching the Patent System Ontology. The important features of the tool are discussed in this Section.

3.1 Integration with BioPortal

There is a very inconsistent usage of terminologies in the biomedical community. Biomedical ontologies provide a rich knowledge base and a consistent representation of biomedical terminology. BioPortal is a source for over 250 bio ontologies. Given a set of keywords, this feature will use BioPortal's web services and automatically expand the query to include several related concepts such as synonyms, children and parents. These expanded terms are used to improve recall and precision of the search. The users can selectively filter ontologies and optionally provide weights to the retrieved concepts (see Figure 1). A sample of the results page is shown in Figure 3.

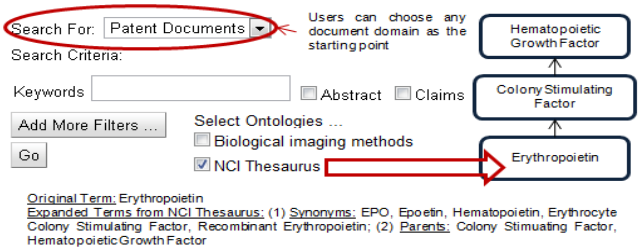

Figure 1. Search Interface

3.2 Cross-Referencing

Although the information sources are very diverse, they highly cross-reference one another. The cross-referenced information provides strong relevancy measures between documents from different information sources. The Patent System Ontology represents the cross-referenced information in the form of object properties. The tool reasons upon these properties to produce un-obvious results much faster than manual interpretation. As a result, users can traverse along these properties from one information domain to another to search for similar and related documents without having to deal with the underlying semantics of the ontology. For example, there is no connection between U.S. patents 5,955,422 and 4,879,272 that is obvious from the patent documents themselves. However, the file wrapper for patent 5,955,422 reveals that they are strongly related.

3.3 Full Text-Search and Faceted Search

Apache Lucene is a widely used text-indexing and searching library. The tool integrates the power of Lucene with the Patent System Ontology to allow fast and efficient text-based searches. Faceted searching allows users to explore the results by narrowing down to documents of their interest based on several dynamically generated filtering criteria. Solr is an Apache Lucene based search platform which provides libraries for full-text and faceted searching. However, when dealing with a diverse set of documents, the indexing schema can get complicated. We explore existing approaches such as LARQ (Lucene+ARQ) and SARQ (Solr+ARQ) to provide a methodology to integrate the Patent System Ontology with tools such as Solr and Lucene.

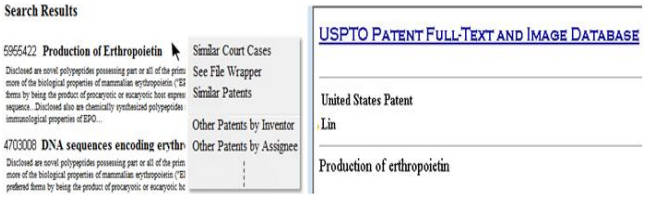

Figure 2. Search Results

4. CONCLUSION

In this demonstration, we present an interactive tool which integrates and searches several heterogeneous information sources in the patent system. The current version of the tool is built upon an instantiated OWL ontology for the patent system which consists of 54 classes, 41 properties and over 15,000 individuals from 1150 patent documents, 30 court cases and 1 file wrapper in the biomedical domain. In addition to the features explained in Section 3, the tool provides several search support features such as abstract snippets, direct links to USPTO database, highlighting a selected criterion and ability to create search profiles. Although we deal with documents from the biomedical domain, the tool provides an approach that can scale to other domains as well. The tool is designed to help expedite the research of a wide range of users including start-up companies, patent examiners, litigators, and will especially interest the DGO community.

5. 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.

6. REFERENCES

[1] S. Taduri, G. T. Lau, K. H. Law, Hang Yu and J. P. Kesan. 2011. An Ontology to Integrate Multiple Information Domains in the Patent System. *2011 IEEE International Symposium on Technology and Society (ISTAS)*, May 23-25, 2011, Accepted.

[2] S. Taduri, G. T. Lau, K. H. Law, Hang Yu and J. P. Kesan. 2011. Developing an Ontology for the U.S. Patent System. *12th Annual International Conference on Digital Government Research (dg.o 2011)*, June 12–15, 2011, Accepted.