Retrieval of Patent Documents from Heterogeneous Sources using Ontologies and Similarity Analysis

Siddharth Taduri, Gloria T. Lau, Kincho H. Law
Engineering Informatics Group
Stanford University
(staduri, glau, law)@stanford.edu

Jay P. Kesan
School of Law
University of Illinois Urbana-Champaign
kesan@illinois.edu

***Abstract*—In the past few years, there has been an explosive growth in scientific and legal information related to the patent system. Patents and related documents are siloed into multiple heterogeneous sources. Retrieving relevant information from diverse sources is a non-trivial task and poses many technical challenges. Among the challenges is the issue of terminological inconsistencies that are used in the documents. We tackle the terminological inconsistency issue by exploring domain knowledge through the use of ontology standards. Furthermore, we take advantage of cross-references and structural dependencies between the information sources to enhance terminological comparison. In this paper, we present a similarity analysis methodology which combines knowledge from two distinct sources – (1) domain ontologies and (2) ontologies which describe the information sources to assist a user in identifying relevant documents across several information sources simultaneously. Specifically, we explore the use of a rule-based system to infer relationships between documents based on pre-defined heuristics. We present our results through a use case in the bio-patent domain with a collection of 1150 patents and 30 court cases.**

Keywords-Ontology, Patent, Court cases, Information Retrieval, Knowledgebase

I. INTRODUCTION

In recent years, large volumes of information have been made available online. However, information pertaining to a subject is distributed across many sources maintained by autonomous entities, resulting in a very diverse, heterogeneous and unstructured collection of information. The heterogeneity is observed at multiple levels – structural, syntactic, semantic and system [1]. To integrate information from multiple diverse sources with various levels of heterogeneity is a non-trivial task.

In this research, we work with the U.S. patent system in which patents and relevant information are siloed into several sources; they include issued patents, patent applications, court cases, patent file wrappers, regulations and laws, and engineering and scientific publications. Currently, relevant documents are gathered by independently searching various sources, followed by manually correlating the information. For example, a startup company wanting to patent their new satellite communication technology needs to extensively search the patent database, court litigations and relevant scientific publications for any previous work in the field which may cause their application to be rejected. An infringement lawyer will need to search all the above sources including relevant laws and regulations. In both cases, multiple unstructured and semi-structured information sources are involved. Due to the sheer volume of information, finding relevant documents is a very tedious task. The motivation for working on this problem comes from the fact that the information contained within the patent system is very valuable for a wide range of users with different professional training and backgrounds and resources, from academicians to lawyers and from small inventors to large corporations.

When dealing with specific technical domains (for example, telecommunications, biotechnology, music etc.), there is an increased usage of domain terminology in the patent documents. However, terms are often inconsistently used in their various forms such as synonyms, hypernyms, and hyponyms etc. which cause the use of general language comparison during search to be inefficient. Domain ontologies provide a rich source of knowledge allowing applications to understand the semantics of the domain and handle terminological inconsistencies. For information retrieval, the use of domain knowledge can enhance the quality of results obtained [13, 14]. In this paper, we present our methodology based on documents in the biomedical domain. The recent advancements in biomedical ontologies have led to several ontology standards being extensively used in information systems [11, 12]. We use BioPortal, an online resource with over 250 bio-ontologies, as our source for comprehensive domain knowledge [15]. Specifically, we focus on a use case – erythropoietin – to illustrate our knowledge-driven approach to information retrieval from heterogeneous information sources.

Ontologies have been increasingly adopted in fields such as knowledge representation, information integration and information retrieval (IR) and extraction [2, 4]. We classify ontologies as our knowledge sources into two categories – (1) ontologies which provide a consistent representation for the information sources themselves; and (2) ontologies which provide consistent domain specific terminology and semantics. Creating a single unified ontology to tackle all forms of heterogeneity is not practical or feasible [10]. Integrating several types of ontologies are necessary in order

to improve IR. Ontology alignment, merging and mapping have been widely used to relate ontologies and facilitate inter-operability [5]. These techniques work very well for ontologies which describe similar information. In order to integrate information from two very different knowledge sources, however, the integration needs to be carried out at the application level [6].

The information sources in the patent system are heavily cross-referenced providing important information that one can reason along. We developed a patent system ontology that provides a consistent structure to the documents and captures the cross-referenced information. Since we are dealing with both a diverse set of information sources and users, it is challenging to determine the exact context of the query. In this paper, our goal is to provide a methodology that approaches the problem by integrating metadata information, cross-referenced information and semantics from domain ontologies to provide a reliable similarity analysis between documents spanning the various information sources in the patent system. Specifically, we implement a rule-based system to reason along the metadata and cross-referenced properties of patent system ontology to infer document similarity. At the application level, we tackle the terminological inconsistencies in the documents using domain ontologies. We hope this similarity analysis approach will not only yield higher quality results, but also facilitate evaluating the relations defined in the patent system ontology through the cross-referenced information.

This paper is organized as follows: Section II discusses previous work related to the problem stated in this discussion. A detailed explanation of our use case, the U.S. patent system and the bio-ontologies is provided in Section III. Section IV describes our methodology and the results are presented in Section V.

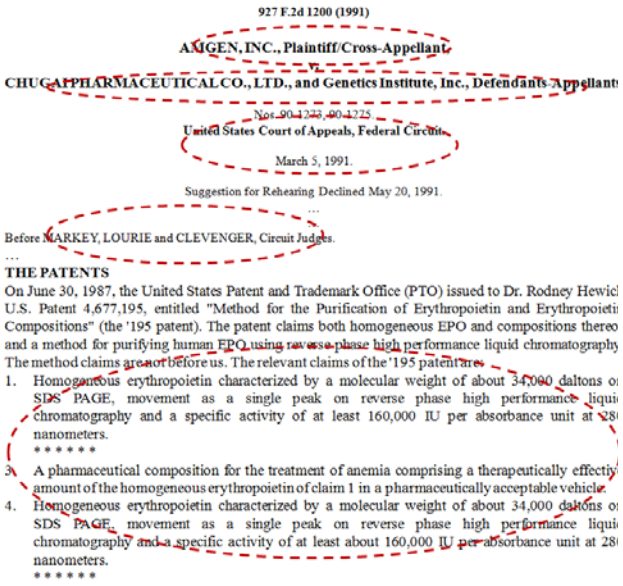
II. Background

IR has been an active field of study for quite some time. As more information is available online there is a constant push to develop methods that can accurately and efficiently retrieve relevant information. Document metadata such as citations, co-authorships etc. can provide very useful information which can be used in IR [30]. This leads to link analysis approaches where algorithms, such as PageRank, can be used to estimate the importance of a document. In the context of the web, these in-and-out links are hyperlinks connecting two web pages [29]. Citation based patent retrieval has been studied and shows an improvement in the search results [16]. Another method approaches patent classification and retrieval through technology categories [17]. Generally speaking, these approaches can scale to the entire patent corpus since they are not tied to any specific technical domain. However, due to the vast amount of information available, these general information retrieval approaches may not be the most efficient and accurate and research focus has shifted towards the use of terminologies and semantics [1]. When dealing with specific technical domains, one common problem is that the use of terminology is not consistent. A single concept can be represented in terms of its synonyms, hyponyms or even abbreviations. Several works have shown that the use of domain ontologies in IR facilitates semantic inter-operability and improves the quality of the results [11, 13, 14]. Another approach to handle terminological inconsistencies is Latent Semantic Indexing (LSI), where the various terms are mapped onto a single concept axis [28].

In the context of patent documents and regulations, it has been shown that methods that take into account structural aspects of documents improve the quality of IR [18, 19]. However, these methods deal with either a single document or a corpus with documents in similar domain. Since documents such as patent laws and regulations are not domain specific, semantic annotations from domain ontologies cannot directly be used to enhance retrieval. In such cases, the cross-referenced information and the metadata connecting other information sources become very useful. Thus, in addition to text-based similarity analysis, we explore a rule-based approach to reason along metadata and cross-referenced information in the patent system ontology to relate documents. Rule-based systems have traditionally been used mainly to develop expert systems such as question-answering and decision support systems [31]. In this work, rules add an additional layer of expressivity and allow us to reason over entities defined in the ontology [20]. While scalability is an important consideration for rule based systems and rule languages such as SWRL, newer reasoning algorithms and the use of DL-safe rules have shown a more computationally tractable performance [7].

III. Use Case

In this paper, we concentrate on a use case in the biomedical domain – erythropoietin. Erythropoietin is a hormone responsible for the production of red blood cells in the human body; the lack of which can cause serious diseases such as anemia. To establish a corpus for the experimental use case, we identified 5 core patents on the production of erythropoietin for treatment of related diseases. Following the backward and forward citations along these 5 core patents, we defined a set of 135 highly relevant patents as our gold standard. From the bio-ontologies in BioPortal, we extracted 43 concepts related to erythropoietin. Overall, our patent corpus consists of 1150 patents including the 135 relevant patents and the top 50-100 patents for each of the 43 concepts. Due to the high value of these patents, over the past decade there have been several court litigations involving other patents. Our corpus consists of around 30 court case documents mostly involving the 5 core patents. The 135 patents collectively cite over 3000 publications which are available in the PubMed database [22]. Furthermore, each patent is associated with a file wrapper; a collection of the original patent application and all communication between the applicant and USPTO regarding that patent application. However, since file wrappers are primarily available as image files, currently we are able to obtain the full-text for only one patent file wrapper (5,955,422). Our use case on erythropoietin clearly involves documents from diverse information sources. In addition, there are several bio-ontologies on BioPortal containing erythropoietin or a related concept to serve as domain



927 F.2d 1200 (1991)

AMGEN, INC., Plaintiff/Cross-Appellant,

v.

CHUGAI PHARMACEUTICAL CO., LTD., and Genetics Institute, Inc., Defendants-Appellants.

Nos. 90-1223, 90-1225.

United States Court of Appeals, Federal Circuit.

March 5, 1991.

Suggestion for Rehearing Declined May 20, 1991.

...

Before MARKEY, LOURIE and CLEVENGER, Circuit Judges.

...

THE PATENTS

On June 30, 1987, the United States Patent and Trademark Office (PTO) issued to Dr. Rodney Hewick U.S. Patent 4,677,195, entitled "Method for the Purification of Erythropoietin and Erythropoieti Compositions" (the '195 patent). The patent claims both homogeneous EPO and compositions thereo and a method for purifying human EPO using reverse-phase high performance liquid chromatography The method claims are not before us. The relevant claims of the '195 patent are:

1. Homogeneous erythropoietin characterized by a molecular weight of about 34,000 daltons o SDS PAGE, movement as a single peak on reverse phase high performance liqui chromatography and a specific activity of at least 160,000 IU per absorbance unit at 28 nanometers.

* * * * * *

3. A pharmaceutical composition for the treatment of anemia comprising a therapeutically effectiv amount of the homogeneous erythropoietin of claim 1 in a pharmaceutically acceptable vehicle.

4. Homogeneous erythropoietin characterized by a molecular weight of about 34,000 daltons o SDS PAGE, movement as a single peak on reverse phase high performance liqui chromatography and a specific activity of at least about 160,000 IU per absorbance unit at 28 nanometers.

* * * * * *

Figure 1. Sample Court Case.

knowledge. Overall, the use case clearly represents the problem involving the use of ontologies for information retrieval across heterogeneous document sources. In the following sub-sections, we will describe the two information sources – the patent system as a whole; and the bio-ontologies as the domain knowledge.

A. Patent System

The U.S. patent system comprises of several diverse information sources including – (1) issued patents and patent applications; (2) court cases; (3) patent file wrappers; (4) scientific publications and (5) relevant chapters from the Code of Federal Regulations (C.F.R.) and the United States Code (U.S.C.). There are several challenges associated with the information sources. Some of these are discussed below:

1) Structure and Format:

Figure 1 shows an example of a court case. Relevant information is available in various parts of the unstructured document. In order to process the information, the document must first be conformed to a certain structure. Similarly, other documents also need to be re-structured. The documents are not only structurally different, but also come in different formats. While patent documents are available as plain HTML files from the U.S. Patent and Trademark Office (USPTO) database, file wrappers and court cases are image files. Hence, the documents need to be converted to a unified processable format.

2) Current tools and availability

The USPTO maintains a database for patents, applications, trademarks and copyrights [32]. The free-text database is searchable through a web interface. Public Access to Court Electronic Records (PACER) is an online resource for court cases and docket information [33]. However, PACER stores all documents as images. Hence, additional processing is required to extract text from these

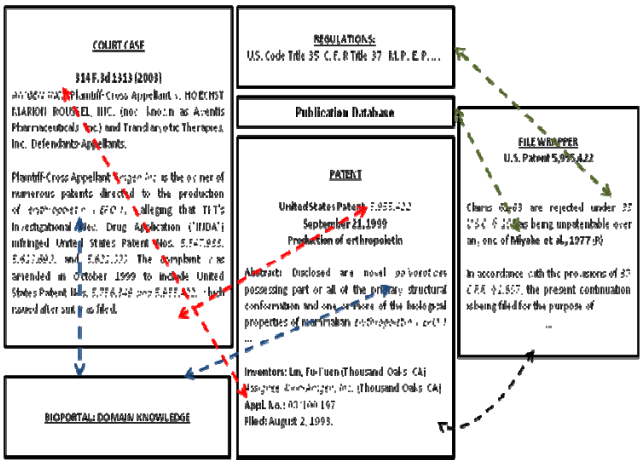


Figure 2. Cross-Referencing Between Information Sources.

patent documents. Alternative sources for IP related data are available from LexisNexis, Westlaw, HeinOnline, Thomson Reuters and Dialog LLC [9, 36, 37, 38]. PubMed is an example of a very comprehensive database for scientific publications. Entrez is a tool which provides web services to search the PubMed database. With such diverse sources of information, there is a lack of tools which integrate all these information and provide a unified search interface for all these documents. Currently, one needs to search independent databases and manually correlate the information.

3) Size

There are currently over 7 million U.S. patents and over 1,000 applications filed on an average per week [32]. There are 95 U.S. district courts and 13 Federal courts of appeals, each maintaining their own database. Furthermore, there are over 19 million citations for scientific publications in the PubMed database alone. However, only small subsets of all these documents are encompassed by a specific technical domain. General IR techniques do not cater to such specific domains.

With the challenges mentioned above, there is clearly a need for structured representation and information management to allow these information sources to interoperate. To address this issue, in our previous work, we have developed a patent system ontology, conceptualizing the patent system domain [23]. The ontology models the documents and has two important properties – (1) provides a consistent structure to the various documents; and (2) captures the cross-references between the domains which is important when developing relevancy measures. Figure 2 illustrates these references. The hierarchical classification of the patent system ontology is shown in Figure 3. Currently, the patent system ontology models the patent, court case and the file wrapper domains and is instantiated with the documents in the corpus. The knowledge base consists of around 50 classes, 35 properties and over 20,000 individuals. Although the topic is not within the scope of discussion, it is important to note that the information from physical documents is parsed accurately in order to realize the benefits of the ontology.

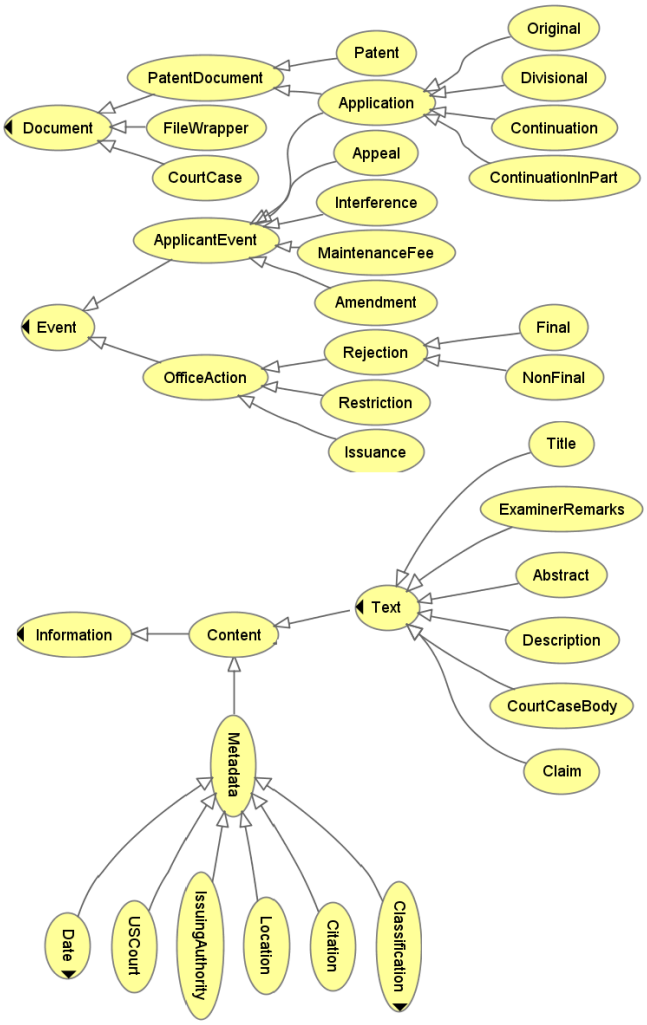


Figure 3. Class Hierarchy of the Patent System Ontology

B. *Bio-Ontologies*

Ontologies model terms and concepts within a particular domain (such as Gene Ontology and Medical Subject Headings) [15, 24]. Within the context of this paper, we refer domain ontologies as a set of existing and accepted terminological standards which facilitate semantic interoperability and consistent usage of the terminology. For example, consider two documents Doc1 containing the terms {"erythropoietin", "glycoprotein" …} and Doc2 containing the terms {"epo", "colony stimulating factor"…}. Docs 1 and 2 do not have any terms in common, thus, at first glance, there is no obvious similarity between them. A simple bag-of-words model will not identify any relationship between these documents. However, domain ontologies show that erythropoietin is synonymous to EPO and "colony stimulating factor" is a broader (parent) concept of EPO. Therefore, by comparing the semantics instead of the terms directly, a relevancy can be established between the documents. Given an initial query, we use the BioPortal REST web service to query the bio-ontologies for related concepts across all ontologies to expand the user query so that relevant and related documents that may not include exact terms and concepts can be retrieved.

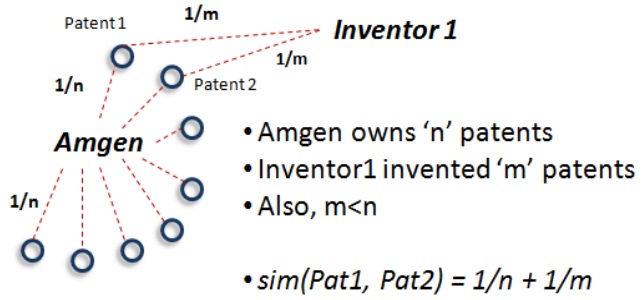


Figure 4. Weighing Rules to Rank Document Similarity

IV. METHODOLOGY

In this section we present our similarity analysis methodology which combines the use of domain knowledge with cross-referenced information to identify relevant documents.

A. *Developing Rule Based Similarity Measures*

Rules are declarative statements which operate over the entities defined in the knowledge base and are useful for inference. The Semantic Web Rule Language (SWRL), which combines OWL and RuleML, extends the expressivity of OWL [8, 20]. An inference engine or a reasoner executes the rules and infers new facts in the knowledge base. SWRL however comes at a price of decidability and computational complexity [7]. Since OWL is built on the semantics of RDF, we try to stay as close as possible to the boundary of RDF and OWL, and the use of DL-safe rules [34]. We use the Pellet reasoner and Jess inference engine to reason over the developed rules [25, 26]. The rules are developed based on similarity heuristics between documents. Examples of the heuristics for the rules are listed below:

- Two patent documents by the same inventor are potentially similar
- Two patents that appear in court litigation are potentially similar. Also, the court case is related to both the patents
- Two documents (patents, publications, court cases) etc. are similar if they occur in an *'interference*[1]*'* proceeding

The rules operate over the metadata and cross-referenced properties defined in the patent system ontology and infer pairs of similar documents. In order to differentiate between the inferences made by each rule, we define a property hasSimilarDocument_* for each rule, where * indicates the identifier for the rule. This allows us to apply several weighing schemes to the rules to distinguish between the more general and the more important specific rules. To illustrate, consider the example shown in Figure 4. Patents 1 and 2 are both owned by the same company 'Amgen' and invented by the same inventor. According to our rule base, these patents should be considered similar to one another according to at least two rules. However, intuitively, a large

[1] An interference is a proceeding where two parties claim the same invention. This information is contained in the patent file wrapper

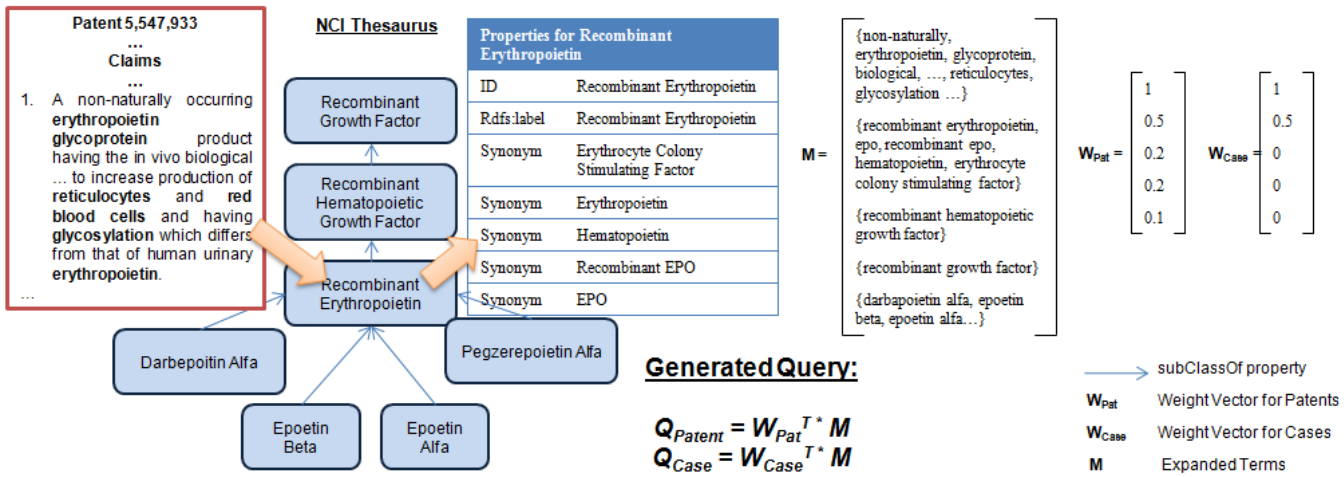


Figure 5. Example Usage of Bio-Ontologies

company such as Amgen is likely to own patents covering a broader range of topics than a single inventor would. If Amgen has '*n*' patents, then we will assume each link contributes a weight of *1/n*. Similarly, if Inventor1 contains '*m*' patents, then each link has a weight of *1/m*. Since *n>m*, the more general rules would be assigned a lesser weight. The resulting similarity score between the documents is a weighted sum of the number of rules that infer the two documents as similar:

$$Sim(A,B) = \sum_{i=1}^{\#\ of\ Rules} Wi * inference(i)$$

where W_i represents the importance of the rule and *inference(i)* = 1 if 'A hasSimilarDocument_i B' or 0 otherwise.. For illustration purpose, in this paper we simply give all rules equal weights and the score is equal to the number of rules that have concluded that the two documents are similar.

B. *Text-Based Semantic Similarity Measure*

As discussed in Section III.B, annotations from domain knowledge can help relate two similar documents which do not share any common terms. Many techniques employ ranking schemes where more general concepts (appeared as parents, grandparents etc. of a concept in the ontology hierarchy) are assigned lower weight than synonyms of the original query term [11, 13, 14]. However, in a corpus consisting documents from multiple information sources, some documents such as technical publications are completely written in technical language while other documents such as court cases tend to use minimal technical language. To handle the diverse amount of technical content in the documents, we cannot simply apply the ranking schemes or query expansion schemes to all types of documents alike. Therefore, we extend the above techniques by using domain terminology appropriately for each information source. As illustrated in Figure 5, we first retrieve all ontologies from BioPortal which contain the term 'erythropoietin'. The related terms are extracted and stored in M, where where $M = [\{Original\ Terms\}, \{Synonyms\}, \{Parents\}, \{Grandparents\}, \{Children\}]^T$. A different weight vector – W_{Pat} for patents and W_{Case} for court cases – is used to generate the new expanded query, which is of the form:

$$Q = W^T * M$$

A similar procedure is followed for the other bio-terms in the document. It is challenging to determine the exact context of the user query. For example, the term "recombinant erythropoietin" can be identified as two separate terms, or a single phrase. If expanded in the wrong context, this can lead to several spurious terms in the query which can drastically affect the quality of the search. We penalize the expanded terms by giving them a significantly lower weight than the original query terms. Hence, when performing similarity analysis, we take into account possible terminological inconsistencies and also limit the effect of spurious terms.

The documents are indexed using Apache Lucene, which is a text mining library commonly used in IR [27]. In our current implementation, the SWRL rules are encoded into the ontology and executed prior to indexing. Using a triple store such as Virtuoso to manage the data, we index the entire inferred OWL data to conform to a specific schema. Reasoning over large knowledge bases can be long process. Since structure and inferences from the patent system ontology are retained, the user is not required to wait through the long reasoning times when performing the search. Lucene employs a scoring scheme which is primarily based on tf-idf. The expanded query Q is executed on the Lucene index to retrieve a set of tf-idf ranked results. The final similarity score of a document is a linear combination of the scores from rule-based similarity and text-based similarity. Currently, a 60% weight is given to the rule-based inferences and a 40% weight is given to the similarity scores based on semantic annotations. Our future plan is to provide the user with the flexibility to choose weights for the two methods.

V. RESULTS

In this section, several examples are presented to illustrate the methodology used for similarity analysis of the documents in the patent system. The intention here is to

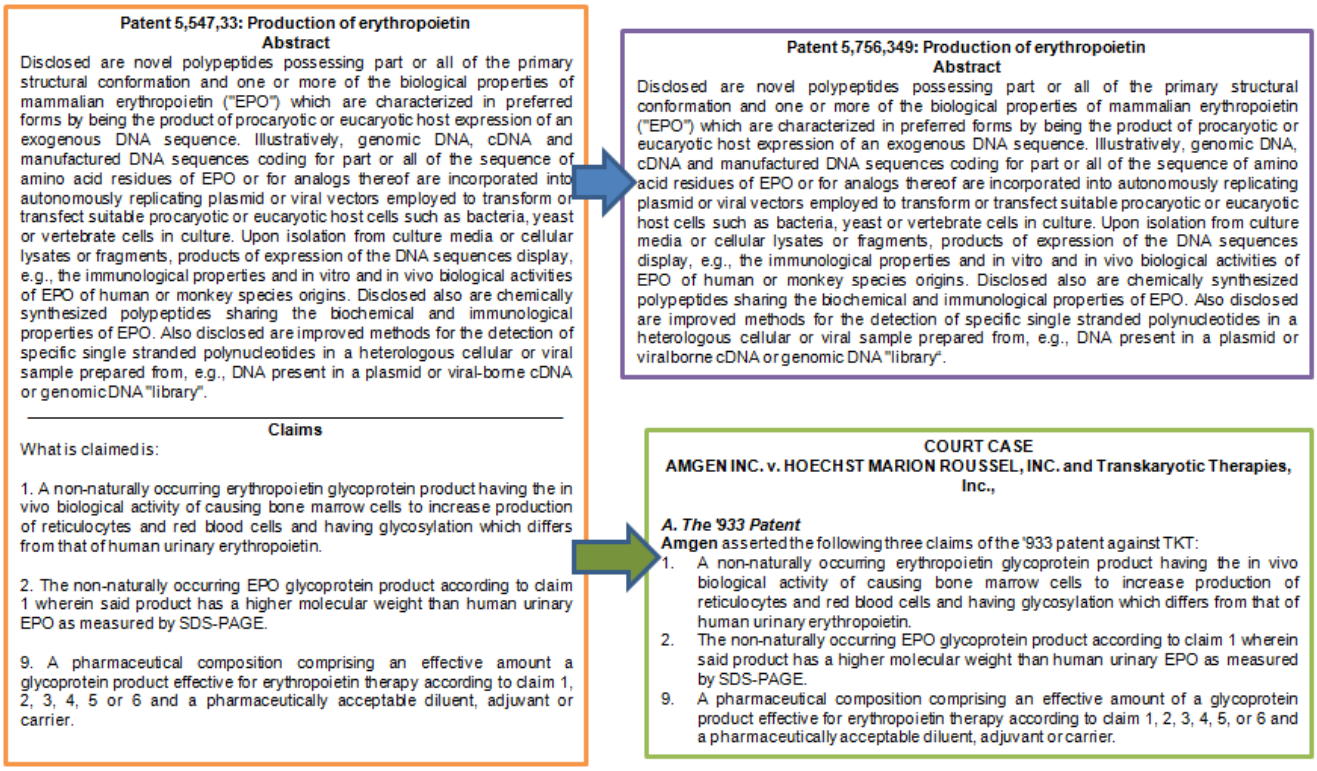


Figure 6. Structural Dependencies Between Documents.

show that the rule-based inferences and domain knowledge are both equally important to relate documents. We make an assumption that the user is able to identify at least one document from the corpus relevant to the search query [14]. The similarity analysis approach allows the user to quickly identify other relevant documents to the selected one.

Patent documents consist of both a technical section (Abstract and Description) and a legal section (Claims). Comparing these documents as a whole will yield low scores for relevant documents, thus making the similarity analysis inefficient. The relevancy between individual sections of the documents must be identified and compared in order to relate the documents. This is illustrated in our first example (see Figure 6), in which the top matches for patent 5,547,933 (Doc1) from both court case and patent repositories are shown. We observe that the text of court cases focus on the legal aspects of patent documents, i.e. the claims. We take advantage of the relevant sections identified in similarity analysis by comparing only the claims section of the patents with the body of the court case. For court cases, we limit the bio-annotation of the terms to only their synonyms and present our results as the best overlap (tf-idf score) of the expanded claim terms. Since the claims specifically use terms such as “erythropoietin” and “EPO”, this procedure will also identify other court cases which do not cite the claims exactly. Also, case documents make direct references to the parties and patents involved which provides additional information which can be used as relevancy metrics. These inferences are captured by the rule-engine component.

We perform several experiments when comparing patent documents. The patents are compared as a whole, and with respect to sub-sections such as the title, abstract, claims and description. Although the title is an important section of the document, it is often very short and does not provide sufficient information. On the other hand, the technical description can often run into several pages of information. Expanding the terms in the technical description can result in extremely large queries and also lead to imprecise results. In order to keep the generated query lengths within reasonable sizes, we compare patents by their abstracts which provide an average query length of ~200 terms after bio-annotation. The top match for Doc1 is shown in Figure 6, which is another core patent and thus verifies the procedure. Using the claims as a relevancy measure between patent documents and court cases has an additional advantage. For example, since the two patents are very similar to each other, if the user searches for similar documents to the court case only based on the abstracts, they would both obtain a high similarity score. The true relationship between Doc1 and the court case, i.e. the fact that Doc1 is directly involved in the court case is not identified. However, in addition to the rule-based inference, using the claims to relate the documents, Doc1 achieves a significantly higher score and is returned as a top match.

In the second example (see Figure 7), we show two relevant patent documents to Doc1 – patent 4,677,195 (Doc2) and 4,999,291 (Doc3). Without the use of domain knowledge, the similarity score of Doc1 and Doc3 is very low and hence does not show up in the top results. The domain knowledge provides the relationship between “erythropoietin” and “colony stimulating factor” and thus improves the similarity score between the documents. After

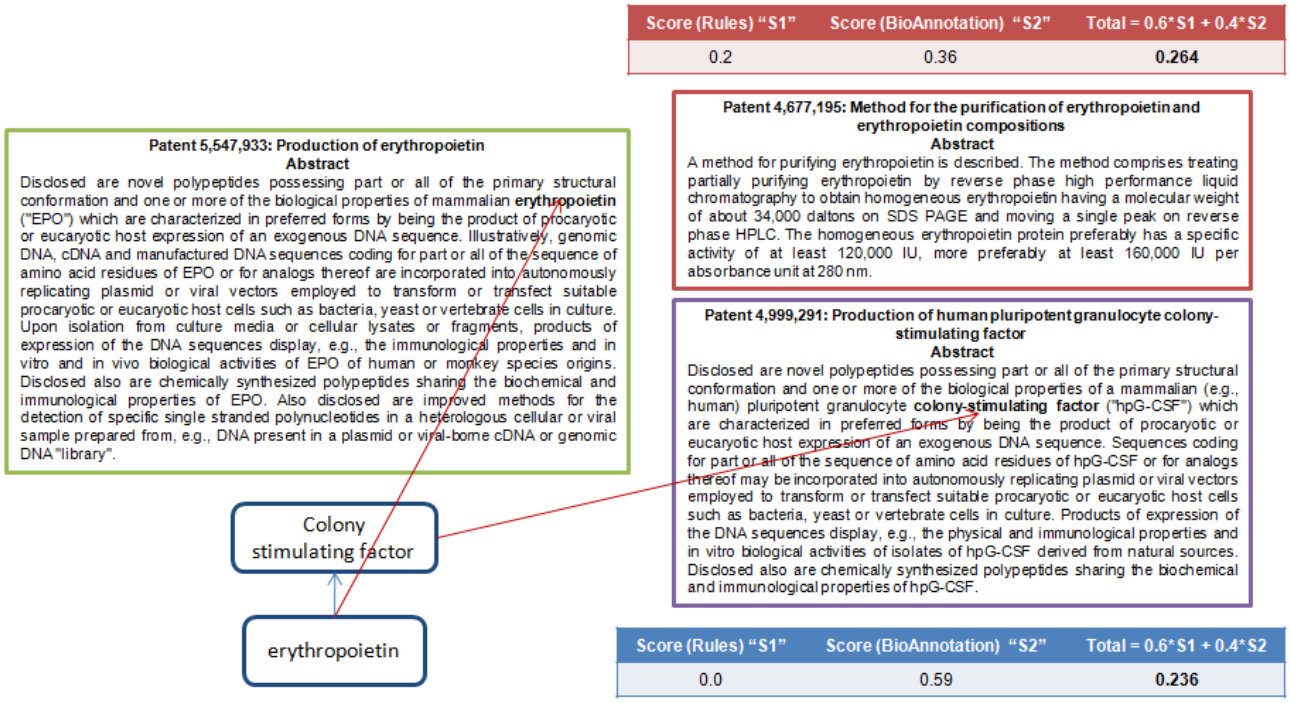


Figure 7. The Use of Bio-Ontologies and Rule-Based Inferences.

bio-annotation, we notice that the results based only on bio-annotation of the abstract of Doc1 give Doc3 a higher score than Doc2. However, Doc1 and Doc2 are very highly related and have been challenged together several times in the court. This relationship between the documents is not captured through the bio-annotations alone and could result in Doc2 not being identified by the user. The rule-based inferences identify this relationship and give the two documents a similarity score of 0.2 (i.e. 2 out of 10 rules infer that Doc1 and Doc2 are similar). The similarity score for Doc1 and Doc3 through rule-based inferences is zero. Hence, the resulting score after the linear combination of the two results gives Doc2 a higher score than Doc3 in the results. Through these examples, we show that the similarity analysis procedure must account for semantic similarity, metadata and cross-referenced information in order to achieve the best results. In the above example, although Doc1 and Doc2 get a low similarity score (0.2) through the rule-based component, the fact that they were challenged together in a court case could hold higher significance to some users. Hence, in future implementations, we propose to allow the modification of weighing scheme to suit the needs of individual users.

VI. Conclusion

In this paper, we present a similarity analysis methodology to retrieve relevant documents from multiple information sources in the patent system. The problem of gathering relevant information across multiple information sources in the patent system faces two major challenges – (1) inconsistent use of technical terminology; (2) diverse structures in the information sources. We tackle the first challenge through the use of domain ontologies which provide the semantics to handle inconsistencies and establish relationships between the terms. For a document chosen by the user, we generate an expanded query using the domain knowledge by annotating terms in the document with their synonyms, the hierarchical ontology concepts (parents, children and grandparents). We realize that the level of technical terminology used in different information sources vary significantly. Therefore, we apply the use of domain ontologies differently to each information source in order to improve the quality of the results. For example, we expand the query to synonyms and the hierarchical concepts when comparing patent documents, but limit the expansion to only synonyms when comparing court cases. Comparing the entire text of documents is both inefficient and yields low similarity scores. Using the patent system ontology, we make use of the structural dependencies between documents as a relevancy metric and compare only relevant sections with one another. Metadata and cross-referencing between the documents provide additional information which is not accounted for by the use of domain ontologies alone. We developed a rule-based system which operates on the metadata and cross-referenced information defined in the patent system ontology to infer similarity between the documents. The rules are developed on pre-defined similarity heuristics.

We present our results through a use case in the bio-domain, erythropoietin and a set of 1150 patents and 30 court cases. Our results show that the integration of the rule-based inferencing and domain knowledge leads to a strong similarity analysis approach that allows users to quickly identify a set of relevant documents based on the initial query. Currently, rule based systems face the issue of scalability when compared to traditional systems. OWL-RL,

a flavor of the OWL-2 language is primarily based on rule-based inferencing and is a potential candidate for our methodology [35]. In future, we plan to address this issue and provide a thorough analysis and an insight on potential improvements on the use of rule-based systems in IR. Presently, it is hard to define a gold standard to provide a formal analysis through measures such as precision and recall. We plan to interact with typical users of such an application to develop several scenarios to facilitate such a formal analysis.

ACKNOWLEDGMENT

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.

REFERENCES

[1] A. P. Sheth, "Changing focus on interoperability in information systems: From system, syntax, structure to semantics", In Interoperating, Geographic Information Systems, pp. 5–30, 1998.

[2] N. Guarino, "Formal ontology and information systems", 1998.

[3] 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 Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, pp. 108-117, 2001.

[4] A. G. Perez, O. Corocho and M. F. Lopez, "Ontological Engineering: with examples from the areas of Knowledge Management and e-Commerce and the Semantic Web", First Edition (Advanced Information and Knowledge Processing), Springer, 2004.

[5] J. Bruijn et al., "State-of-the-art Survey on Ontology Merging and Aligning", V1. SEKT-project report D4.2.1 (WP4), IST-2003-506826, 2003.

[6] P. Mitra, G. Wiederhold and S. Decker, "A Scalable Framework for Interoperation of Information Sources", Proceedings of the 1st International Semantic Web Working Symposium (SWWS `01), Stanford University, Stanford, CA, July 29-Aug 1, 2001.

[7] B. Parsia, E. Sirin, B. C. Grau, E. Ruckhaus and D. Hewlett, "Cautiously Approaching SWRL", Technical report, University of Maryland, 2005.

[8] Slides from Protégé Group. 2009.
http://protege.stanford.edu/conference/2009/slides/SWRL2009ProtegeConference.pdf

[9] LexisNexis. http://www.lexisnexis.com/

[10] S. Ray, "Interoperability Standards in the Semantic Web," Journal of Computing and Information Science in Engineering, ASME, vol. 2, March, 2002, pp. 65-69.

[11] A. Doms and M. Schroeder, "Gopubmed: exploring pubmed with the gene ontology", Nucleic Acids Research, vol. 33, July 2005, pp. 783–786.

[12] C. Jonquet, M. A. Musen and N. H. Shah, "A System for Ontology-Based Annotation of Biomedical Data", International Workshop on Data Integration in The Life Sciences 2008, DILS'08, Evry, France, Springer-Verlag, 5109, Lecture Notes in BioInformatics, pp. 144-152, 2008.

[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, vol. 30, pp.1066-1077, 2007.

[14] S. Taduri, Hang Yu, G. T. Lau, K. H. Law and J. P. Kesan, "Developing a Comprehensive Patent-Related Information Retrieval Tool", Journal of Theoretical and Applied Electronic Commerce Research, in press.

[15] BioPortal. Accessed on 05/12/2011. http://bioportal.bioontology.org

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

[17] 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, pp. 793-794, 2007.

[18] G. T. Lau, K. H. Law, and G. Wiederhold. "A Relatedness Analysis of Government Regulations using Domain Knowledge and Structural Organization", Information Retrieval, vol. 9, Sep 2006, pp. 657-680.

[19] 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(1), March 2008, pp. 21-23.

[20] I. Horrocks, P. F. Patel-Schneider and H. Boley et al., "SWRL: A semantic web rule language combining OWL and ruleML", W3C Member Submission, 21 May, 2004.

[21] OWL W3C Documentation. http://www.w3.org/TR/owl-ref/

[22] PubMed. http://www.ncbi.nlm.nih.gov/pubmed/

[23] S. Taduri, G. Lau, K. H. Law, H. Yu and J. P. Kesan, "Developing an Ontology for the U.S. Patent System", 12th Annual International Conference on Digital Government Research (dg.o 2011) Digital Government Innovation in Challenging Times , University of Maryland, College Park, MD, June 12–15, 2011 (Accepted).

[24] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, "Gene ontology: tool for the unification of biology. The gene ontology consortium", Nature genetics, vol. 25(1), May 2000, pp. 25–29.

[25] E. Sirin, B. Parsia, B. C.Grau, A. Kalyanpur, Y. Katz, "Pellet: A practical OWL-DL reasoner", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5 (2), Software Engineering and the Semantic Web, June 2007, pp. 51-53.

[26] E. Friedman-Hill, "Jess, the Rule Engine for the Java Platform", http://herzberg.ca.sandia.gov/jess/

[27] Apache Lucene. http://lucene.apache.org/

[28] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman, "Indexing by latent semantic analysis", J. Amer. Soc. Info. Sci., vol. 41, 1990, pp. 391-407.

[29] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report. Stanford InfoLab, 1999.

[30] C. L. Giles, K. D. Bollacker and S. Lawrence, "CiteSeer: an automatic citation indexing system", In Proceedings of the third ACM conference on Digital libraries (DL '98), ACM, New York, NY, USA, pp. 89-98, 1998.

[31] B. G. Buchanan, E. H. Shortliffe, "Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project", 1984.

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

[33] PACER. http://www.pacer.gov/

[34] B. Motik, U. Sattler and R. Studer, "Query Answering for OWL-DL with rules", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 3 (1), Rules Systems, July 2005, pp. 41-60.

[35] OWL-2 Profiles. 2009. http://www.w3.org/TR/owl2-profiles/

[36] HeinOnline IP Library. http://home.heinonline.org/

[37] WestLaw Website. http://www.westlaw.com/

[38] Thomson Innovation. http://www.thomsoninnovation.com/