BioNav: An Ontology-Based Framework to Discover Semantic Links in the Cloud of Linked Data

María-Esther Vidal[1], Louiqa Raschid[2], Natalia Márquez[1], Jean Carlo Rivera[1], and Edna Ruckhaus[1]

[1] Universidad Simón Bolívar, Caracas, Venezuela
{mvidal,nmarquez,jrivera,ruckhaus}@ldc.usb.ve
[2] University of Maryland
louiqa@umiacs.umd.edu

Abstract. We demonstrate BioNav, a system to efficiently discover potential novel associations between drugs and diseases by implementing Literature-Based Discovery techniques. BioNav exploits the wealth of the Cloud of Linked Data and combines the power of ontologies and existing ranking techniques, to support discovery requests. We discuss the formalization of a discovery request as a link-analysis and authority-based problem, and show that the top ranked target objects are in correspondence with the potential novel discoveries identified by existing approaches. We demonstrate how by exploiting properties of the ranking metrics, BioNav provides an efficient solution to the link discovery problem.

1 Introduction

Emerging infrastructures provide the basis for supporting on-line access to the wealth of scientific knowledge captured in the biomedical literature. The two largest interconnected bibliographic databases in biomedicine, PubMed and BIOISIS, illustrate the extremely large size of the scientific literature today. PubMed publishes at least 16 million references to journal articles, and BIOSIS more than 18 million of life science-related abstracts. On the other hand, a great number of ontologies and controlled vocabularies have become available under the umbrella of the Semantic Web and they have been used to annotate and describe the contents of existing Web available sources. For instance, MeSH, RxNorm, and GO are good examples of ontologies comprised of thousands of concepts and that are used to annotate publications and genes in the NCBI data sources.

Furthermore, in the context of the Linking Data project, a large number of diverse datasets that comprise the Cloud of Linked Data are available. The Cloud of Linked Data has had an exponential growth during the last years; in October 2007, datasets consisted of over two billion RDF triples, which were interlinked by over two million RDF links. By May 2009 this had grown to 4.2 billion of RDF triples interlinked by around 142 million of RDF links. At the time this paper was written, there were 13,112,409,691 triples in the Cloud of Linked Data; datasets can be about medical publications, airport data, drugs, diseases, clinical trials, etc. It is of particular interest, the portion of the Cloud that relates life science data such as diseases, traditional Chinese medicine, pharmaceutical companies, medical publications, genes and proteins, where concepts

L. Aroyo et al. (Eds.): ESWC 2010, Part II, LNCS 6089, pp. 441–445, 2010.

are derived from sites such as ClinicalTrials.gov, DrugBank, DailyMed, SIDER, TCM-GeneDIT, Diseasome and OMIM. To fully take advantage of the available data, and to be able to recognize novel discoveries, scientists will still have to navigate through the Cloud and compare, correlate and mine linked data; thus, they may have to spend countless hours to recognize relevant findings.

To provide support on the discovery task of potential novel associations between already published topics in existing bibliographic datasets, Literature-based discovery (LBD) techniques have been developed. LBD methods follow a disease-cure trajectory to guide the search in the space of implicit associations between scientific publications and their annotations or cites. Annotations correspond to concepts from controlled vocabularies or ontologies. LBD can perform Open or Closed discoveries, where a scientific problem is represented by a set of articles that discuss a particular topic A, and the goal is to prove the significance of the associations between A and some other topics C discussed in the set of articles reachable from the publications relevant to the topic A. Srinivasan et al [5] improved previous LBD techniques by recognizing that articles in PubMed have been curated and heavily annotated with controlled vocabulary terms from the MeSH (Medical Subject Heading) ontology. Srinivasan's algorithm considers that topics A, B and C are MeSH terms used to annotate or index PubMed publications. Thus, links from topic A to publications in the second layer are built by searching with topic A on PubMed. Links from publications in the second layer of the graph to MeSH terms in the third layer, are constructed by extracting the MeSH terms annotations from the publications and selecting the ones associated with the UMLS (Unified Medical Language System) semantic types: Gene or Genome; Enzyme; and Amino Acid, Peptide or Protein. The extracted MeSH terms or B set, are used to search on PubMed and a new set of publications is recovered. Again MeSH term annotations are extracted from these publications, and the terms of the UMLS types Disease or Syndrome and Neoplastic Process are retrieved to comprise the set of topics C. Figure 1 illustrates the LBD experiment reported by Srinivasan et al [5] where the MeSH term curcumin is associated with top-5 MeSH terms. Edges of the graph are labeled with weights that are computed by using an extension of the TF*IDF scores; these scores are used to rank the different paths and determine the potential novel associations.

Although Srinivasan's approach enhances previously LBD methods, this solution may be still costly. To provide an efficient solution, we propose the ontology-based

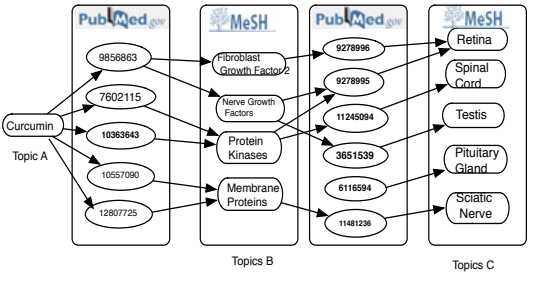

Fig. 1. Example of a Literature-Based Discovery experiment where the MeSH curcumin is related to the retinal disease

system BioNav that provides a framework to discover potential novel associations between drugs and diseases. In this demonstration, we will use the Srinivasan's experiment to show effectiveness and efficiency of the proposed discovery techniques. The paper is comprised of three additional sections: BioNav is presented in section 2, section 3 describes the use cases to be demonstrated, and conclusions are given in section 4.

2 The BioNav Architecture

In Figure 2 we present the BioNav architecture [6]. BioNav is comprised of four main components: a Catalog, a Query Optimizer, a Source Path Discovery component, and a Semantic Link Discovery Engine.

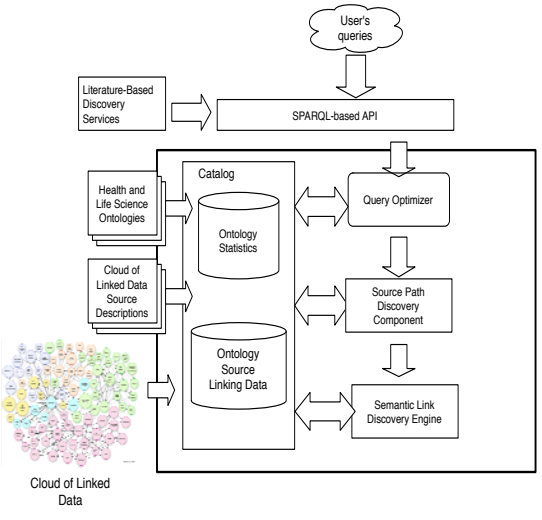

Fig. 2. The BioNav System

The Catalog maintains information about the sources and data that comprise the Cloud of Linked Data; also ontologies are used to describe the meaning of the data. Queries are expressed in terms of the ontology concepts and they are evaluated against the data stored in the Cloud. The result of evaluating a query corresponds to a list of MeSH terms that are semantically associated with terms in the query.

Once a discovery query is received, the parser checks if it is syntactically correct, and cost-based optimization techniques are performed to identify an efficient query execution. An execution plan corresponds to an ordering of the concepts referred in the query, and minimizes the cardinality of the intermediate facts that need to be computed to answer a query and the execution time [4].

Once the input query is optimized, it is rewritten in terms of the data sources that need to be accessed to evaluate the query; a graph-based meta-heuristic Best-First is used to enumerate the paths between data sources that need to be traversed to evaluate the query. Source paths are evaluated by traversing the sources in the order specified in the path.

The answer of the query will be comprised of paths in the Cloud of Linked Data that are locally stored in the BioNav catalog. Paths in the query answer will be computed by traversing the source paths. BioNav uses link-analysis and authority-flow ranking metrics to rank the discovered associations. BioNav ranking techniques assume that the data paths to be ranked comprise a layered graph $lgODG$=(V_{lg}, E_{lg}) of k layers, l_1, ..., l_k. Odd layers are composed of MeSH terms while even layers are sets of publications. An edge from a term b to a publication p indicates that p is retrieved by the PubMed search engine when b is the search term. An edge from a publication p to a term b represents that p is annotated with b. Each edge $e = (b, p)$(resp., $e = (p, b)$) between the layers l_i and l_{i+1} is annotated with the $TF \times IDF$ that represents how relevant is the term b in the collection of documents in l_{i+1}, or a document relevance regarding to a set of terms. The most important terms or publications correspond to the highly ranked Mesh terms or publications in the last layer of the $lgODG$. In this paper we focus on an extension of the Object Rank [1,2] and Path Count [3] metrics for layered graphs or *layered graph Weighted Path Count* (lgWP), which is defined as follows:

A ranking vector R of the target objects in the layered Open Discovery graph $lgODG$ of k layers is defined by a transition matrix A and an initial ranking vector R_{ini}:

$$R = A^{k-1}R_{ini} = (\prod_{l=1}^{k-1} A)\, R_{ini}$$

An entry $A[u, v]$ in the transition matrix A, where u and v are two data objects in $lgODG$, corresponds to $\alpha(u, v)$ or 0. The value of $\alpha(u, v)$ is the weight that represents how relevant is the object u for the object v. Nodes with high lgWP scores are linked by many nodes or linked by highly scored nodes.

$$A[u, v] = \begin{cases} \alpha(u, v) & \text{if } (u, v) \in E_{lg}, \\ 0 & \text{otherwise.} \end{cases}$$

To speed up the tasks of computing the lgWP, we build a Bayesian network with the knowledge encoded in the layered Open Discovery graph, and we perform a Direct Sampling algorithm to traverse the network and just visit the publications or MeSH terms that conduce to potential novel discoveries. This sampling approach has the ability to identify a large number of the potential novel discoveries, while a reduced number of nodes that need to be visited by at least one order of magnitude.

3 Demonstration of Use Cases

In this demonstration, we will show the different steps of the BioNav LBD process in several real-world experiments. The objective is to show the applicability and performance of BioNav for the top-5 diseases that can be treated with the MeSH terms curcumin and aloe. We will demonstrate the following scenarios:

– We show effectiveness by demonstrating the process to compute the potential novel associations between the substance curcumin and MeSH terms that correspond to diseases. We use a ranking metric lgWP to discriminate the potential novel discoveries. Weights in the edges of the graph represent the relevance of the MeSH terms

or publications, and they are computed by using an extension of the $TF \times IDF$ scores. We show that the top-5 MeSH terms in the last layer of the graph correspond to 80% of the top-5 potential novel diseases discovered by Srinivasan et al. [5].
- We show efficiency by computing the number of intermediate and target MeSH terms and publications that need to be visited to identify the novel associations. In BioNav, properties of lgWP are exploited with approximate methods to avoid traversing MeSH terms or publications that do not directly or indirectly link a potential novel discovery. We show that for the MeSH terms curcumin and aloe, the number of visited MeSH terms and publications can be reduced by at least one order of magnitude when these approximate methods are executed.
- Finally, we show the different steps of the open LBD process by selecting a MeSH term that corresponds to a drug or substance, and the potential novel associations of this term with MeSH terms that correspond to diseases. We discuss the impact of the proposed techniques in the Linked Data Cloud.

4 Conclusions

In this demonstration, we present BioNav, an ontology-based tool that supports the discovery of semantic associations between linked data. We demonstrate how link-analysis and authority-based ranking metrics can be used in conjunction with ontology annotations on linked data, to discover potential novel discoveries; also we demonstrate how the properties of the ranking metrics can be exploited to avoid traversing MeSH terms and publications that do not conduce to novel potential discoveries. We show real-world use cases that suggest BioNav is able to efficiently discover almost all the associations identified by state-of-the-art approaches.

References

1. Balmin, A., Hristidis, V., Papakonstantinou, Y.: Objectrank: Authority-based keyword search in databases. In: Proceedings VLDB, pp. 564–575 (2004)
2. Page, L., Brin, S., Motwani, R.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
3. Raschid, L., Wu, Y., Lee, W., Vidal, M., Tsaparas, P., Srinivasan, P., Sehgal, A.: Ranking target objects of navigational queries. In: WIDM, pp. 27–34 (2006)
4. Ruckhaus, E., Ruiz, E., Vidal, M.: Query evaluation and optimization in the semantic web. In: TPLP (2008)
5. Srinivasan, P., Libbus, b., Kumar, A.: Mining medline: Postulating a beneficial role for curcumin longa in retinal diseases. In: Hirschman, L., Pustejovsky, J. (eds.) LT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases, pp. 33–40 (2004)
6. Vidal, M.-E., Ruckhaus, E., Marquez, N.: BioNav: A System to Discover Semantic Web Associations in the Life Sciences. In: ESWC 2009-Poster Session (2009)