

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

María-Esther Vidal¹, Louiqa Raschid², Natalia Márquez¹,
Jean Carlo Rivera¹, and Edna Ruckhaus¹

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