Exploration Using Signatures in Annotation Graph Datasets

Louiqa Raschid
University of Maryland
louiqa@umiacs.umd.edu

Guillermo Palma
Universidad Simón Bolívar
gpalma@ldc.usb.ve

Maria-Esther Vidal
Universidad Simón Bolívar
mvidal@ldc.usb.ve

Andreas Thor
University of Leipzig
thor@informatik.uni-leipzig.de

Abstract

The widespread development and adoption of ontologies to capture semantic domain knowledge and the growth of annotation graph datasets has created many opportunities for large scale Linked Data analytics.

Ontologies are developed by domain experts to capture knowledge specific to some domain. The biomedical community has taken the lead in these activities. Every model organism database has genes and proteins that are widely annotated, e.g., with controlled vocabulary (CV) terms from the Gene Ontology (GO). The NCI Thesaurus (NCIt) version 12.05d has 93,788 terms and the LinkedCT dataset of clinical trial results circa September 2011 includes 142,207 drugs or interventions, 167,012 conditions or diseases, and 166,890 links to DBPedia, DrugBank and Diseasome. At the opposite end of the domain spectrum, the Financial Industry Business Ontology (FIBO) captures knowledge about the structure, properties and behavior of financial contracts.

LinkedData and the Linked Open Data cloud (LOD cloud) has also made available many annotation graph datasets where scientific concepts such as genes, drugs and diseases are marked up (annotated) with controlled vocabulary terms (CV terms) from ontologies. The challenge is to explore these rich and complex annotated datasets, together with the domain semantics captured within ontologies, to discover patterns of annotations across multiple concepts that may lead to potential discoveries. For genes, these patterns may involve cross-genome functional annotation, e.g., combining the GO functional annotations of two model organisms such as Arabidopsis thaliana (a plant) and C. elegans (a nematode or worm), to predict new gene functions or interactions.

Drug target prediction, with a goal of finding new targets for existing drugs, has received widespread media attention and has resulted in some notable successes, e.g., Viagra. Beyond drug target prediction, there are many applications where one may need to provide a comprehensive report of all known evidence about a pair or family of drugs, e.g., to make clinical policy recommendations. The New York Times reported on November 3, 2010 that Genentech began offering secret rebates to about 300 ophthalmologists in an apparent inducement to get them to use more `Ranibizumab` rather than the less expensive `Bevacizumab`. Both drugs are monoclonal antibodies (`_mab`) that target specific molecules or organisms. In 2008, `Bevacizumab` cost Medicare $20 million for about 480,000 injections, while `Ranibizumab` cost $537 million for 337,000 injections. Several studies have shown no superior effect of `Ranibizumab` over `Bevacizumab` for the treatment of macular degeneration.

Our research addresses the challenge of large scale Linked Data analytics of annotation graph datasets, using semantic knowledge from ontologies. We define an *Annotation Signature* between a pair of scientific concepts, e.g., a pair of drugs or a pair of genes. The annotation signature builds upon the shared annotations or shared CV terms between the pair of concepts. The signature further makes use of knowledge in the ontology to determine the ontological relatedness of the shared CV terms. The annotation signature is represented by N groups (clusters) of ontologically related shared CV terms. For example, the annotation signature for a (drug, drug) pair will be a set of N clusters, where each cluster includes a group of ontologically related disease terms.

We define the *Annotation Signature* problem of creating a many-to-many partitioning of the edges of a bipartite graph between two sets of annotations (Palma et al. 2013a; 2013b). We then summarize the challenges in exploiting domain specific semantic knowledge, including the ontology structure and relationship types between concepts. We show how we can tune the (ontologically related) similarity score between node pairs, and produce clusters of more closely related terms that are more useful to the domain scientist.

This research was partially supported by NSF award DBI1147144. We thank our collaborators: Eric Haag and Heven Sze, University of Maryland; Gilberto Fragoso and Sherri De Coronado, National Cancer Institute; Guoqian Jiang and Cui Tao, Mayo Clinic.

References

Palma, G.; Vidal, M.-E.; ; Raschid, L.; and Thor, A. 2013a. Annsigclustering: A semantic-driven clustering technique for annotated linked data. In *Proceedings of the LIS Workshop in conjunction with ISWC*.

Palma, G.; Vidal, M.-E.; Haag, E.; Raschid, L.; and Thor, A. 2013b. Measuring relatedness between scientific entities in annotation datasets. In *Proceedings of the ACM BCB Conference*.