Chapter VIII
Association Analytics for Network Connectivity in a Bibliographic and Expertise Dataset

Boanerges Aleman-Meza
University of Georgia, USA

Sheron L. Decker
University of Georgia, USA

Delroy Cameron
University of Georgia, USA

I. Budak Arpinar
University of Georgia, USA

ABSTRACT

This chapter highlights the benefits of semantics for analysis of the collaboration network in a bibliographic dataset. Metadata of publications was used for extracting keywords and terms, which can be the starting point towards building a taxonomy of topics. The aggregated effect of the topics over all publications of an author can be used to determine his/her areas of expertise. We also highlight the value of using a taxonomy of topics in searching experts on a given topic.
INTRODUCTION

Large-scale bibliography datasets are becoming increasingly available for use by Semantic Web applications. For example, DBLP is a high-quality bibliography of Computer Science literature. Its data is available in XML but it has also been made available in RDF as DR2Q-generated RDF data (Bizer, 2003), also in the SwetoDbp ontology of DBLP data (http://lsdis.cs.uga.edu/projects/semdis/swetodbp/), and Andreas Harth’s DBLP dataset in RDF (sw.deri.org/~aharth/2004/07/dblp/). DBLP data has been used to analyze co-authorship, collaborations, degrees of separation and other social network analysis measures. We claim that further and more detailed analysis is possible by using semantically marked-up data. In this paper, we describe a study of network connectivity in bibliography data. Our work expands upon earlier studies that have used sub-sets of DBLP data for analysis of collaborations in the field of databases (Elmacioglu & Lee, 2005; Nascimento et al., 2003). The dataset we use includes not only the data of publications in database field. It also includes data of publications in areas such as Artificial Intelligence, Web and Semantic Web.

Further analysis of bibliography data is possible when information of topics or research areas is available. Metadata of publications from DBLP can be used for the creation of a dataset of topics in Computer Science. In addition, keywords and terms that appear in abstracts of publications can be used for finding the most common topics or research areas. Based on this, we were able to identify potential terms to be used in building a taxonomy of Computer Science topics. The main benefit is that when these topics are suggested to human, the time required to build a taxonomy of topics could be shortened. Additionally, the suggested terms come from (and reflect) the domain in question (e.g., Computer Science). The identified terms can be analyzed to determine which ones appear only in the last few years. This can lead to the identification of possible new topics or emerging research trends. After the topics of a publication have been identified, it can be said that all authors of a paper have (at least some) knowledge on such topics. Thus, if we look at an author in particular, it is possible to determine the topics on which s/he has expertise/knowledge based on her/his publication track. This is the basis of our method to identify researchers that have high expertise on certain topics. We perform a study to validate this measure of expertise against well-known lists of recognized researchers (e.g., based on available lists of ACM fellows and IEEE fellows). We argue that this type of study can be done with existing Semantic Web technologies that are able to handle large datasets. We also describe the datasets used, which are freely available online.

In summary, the objectives of this chapter are to highlight the benefits of using semantics for analysis of the underlying collaboration network in a bibliography dataset. We describe how keywords and terms can be extracted and linked to metadata of publications. Then, we rely on the aggregated effect of terms/keywords of all publications of an author to determine his/her areas of expertise. We explain how analysis of terms and keywords of publications can help human to create a taxonomy of topics by identifying the most common terms as well as terms commonly occurring together. The use of topics to glean expertise of researchers is validated when top experts on certain topics compared quite well with researchers that have received awards such as ACM Fellows. In doing so, we highlight the value of using a taxonomy of topics to better match expertise of researchers.

BACKGROUND

Bibliography datasets have been used to measure how authors are connected, publication output, citations, etc. The motivation of such analysis typically is gaining insight of how a community
Related Content

An Evaluation of Ontology Based Domain Analysis for Model Driven Development
www.irma-international.org/article/an-evaluation-of-ontology-based-domain-analysis-for-model-driven-development/145855/

Data Mining, Validation, and Collaborative Knowledge Capture
www.irma-international.org/chapter/data-mining-validation-collaborative-knowledge/65692/

Context-Aware Presentation of Linked Data on Mobile
www.irma-international.org/article/context-aware-presentation-of-linked-data-on-mobile/129762/

Semi-Automatic Ontology Construction by Exploiting Functional Dependencies and Association Rules
www.irma-international.org/chapter/semi-automatic-ontology-construction-exploiting/76172/

Incremental Ontology Population and Enrichment through Semantic-based Text Mining: An Application for IT Audit Domain
www.irma-international.org/article/incremental-ontology-population-and-enrichment-through-semantic-based-text-mining/145230/