ACM Transactions on the Web, v2n1, Feb 2008. Full 29-page copy at http://doi.acm.org/10.1145/1326561.1326568

Scalable Semantic Analytics on Social Networks for Addressing the Problem of Conflict of Interest Detection

BOANERGES ALEMAN-MEZA
University of Georgia
MEENAKSHI NAGARAJAN
Wright State University
LI DING
Stanford University
AMIT SHETH
Wright State University
I. BUDAK ARPINAR
University of Georgia
and
ANUPAM JOSHI and TIM FININ
University of Maryland, Baltimore County

In this article, we demonstrate the applicability of semantic techniques for detection of Conflict of Interest (COI). We explain the common challenges involved in building scalable Semantic Web applications, in particular those addressing connecting-the-dots problems. We describe in detail the challenges involved in two important aspects on building Semantic Web applications, namely, data acquisition and entity disambiguation (or reference reconciliation). We extend upon our previous

This research was supported by NSF-ITR Awards #IIS-0325464 and #0714441 titled 'SemDIS: Discovering Complex Relationships in the Semantic Web.'
This article is an extended version of our authors' paper "Semantic Analysis on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection," which appears in the Proceedings of the International World Wide Web Conference (WWW '06).
Authors' addresses: B. Aleman-Meza and I. B. Arpinar, LSDIS Lab, Department of Computer Science, University of Georgia, GA 30602; email: {boanerg, budak}@cs.uga.edu; M. Nagarajan and A. Sheth, Kno.e.sis Center, College of Engineering and Computer Science, Wright State University, OH 45435; email: {nagaran.5, amit.sheth}@wright.edu; L. Ding, Knowledge Systems, AI Lab, Department of Computer Science, Stanford University, CA 94305; email: ding@ksl.stanford.edu; A. Joshi and T. Finin, Department of Computer Science and Electrical Engineering, University of Maryland, MD 21250; email: {joshi, finin}@cs.umbc.edu.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, fax +1 (212) 869-0481, or permission@acm.org.
© 2008 ACM 1559-1131/2008/02-ART7 $5.00 DOI 10.1145/1326561.1326568 http://doi.acm.org/10.1145/1326561.1326568

ACM Transactions on the Web, v2n1, Feb 2008. Full 29-page copy at http://doi.acm.org/10.1145/1326561.1326568

work where we integrated the collaborative network of a subset of DBLP researchers with persons in a Friend-of-a-Friend social network (FOAF). Our method finds the connections between people, measures collaboration strength, and includes heuristics that use friendship/affiliation information to provide an estimate of potential COI in a peer-review scenario. Evaluations are presented by measuring what could have been the COI between accepted papers in various conference tracks and their respective program committee members. The experimental results demonstrate that scalability can be achieved by using a dataset of over 3 million entities (all bibliographic data from DBLP and a large collection of FOAF documents).

Categories and Subject Descriptors: H.4.m [**Information Systems Applications**]: Miscellaneous; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information Networks*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Semantic Web, social networks, conflict of interest, peer review process, semantic analytics, entity disambiguation, data fusion, semantic associations, ontologies, RDF, DBLP, swetoDblp

ACM Reference Format:

Aleman-Meza, B., Nagarajan, M., Ding, L., Sheth, A., Arpinar, I. B., Joshi, A., Finin, T. 2008. Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection. ACM Trans. Web, 2, 1, Article 7 (February 2008), 29 pages. DOI = 10.1145/1326561.1326568 http://doi.acm.org/10.1145/1326561.1326568

1. INTRODUCTION

Conflict of Interest (COI) is a situation where bias can exist or be perceived based on the relationships or connections of the participants involved either explicitly or implicitly. The connections between participants could come from various origins such as family ties, business (e.g., safeMinds.org) or friendship ties, and confidential information. Detecting COI is required to ensure "fair play" in many decision-making situations such as contract allocation, IPO (Initial Public Offerings) or company acquisitions, corporate law and peer-review of scientific research papers or proposals. Detection of COI is also critical where ethical and legal ramifications could be quite damaging to individuals or organizations. The underlying technical challenges are also related to the common *connecting-the-dots* applications that are found in a broad variety of fields, including regulatory compliance, intelligence and national security [Hollywood et al. 2004], and drug discovery [Laz et al. 2004].

The detection of COI usually involves analysis of social network data, which is hard to obtain due to privacy concerns. We chose a special case of COI detection in the peer review process: detecting COI between computer science researchers. This case does not involve much of a privacy concern because researchers are open to expose their identity in published research (listing collaborators) and in their participation on the research community, for example, as reviewers or organizers of conferences. Hence, social and collaborative information is widely published via various media such as magazines, journals, and the Web. In particular, the advance of Web technologies has facilitated the access to social information not only homepages of persons and hyperlinks but also via many social networking sites.

ACM Transactions on the Web, v2n1, Feb 2008. Full 29-page copy at http://doi.acm.org/10.1145/1326561.1326568

Social networking websites attract more and more people to contribute and share information. For example, the LinkedIn social network comprises a large number of people from information technology areas, and it could be used to detect COI in situations such as IPO or company acquisitions. MySpace, Friendster, Orkut, and Hi5 contain data that could substantiate COI in situations of friendship or personal ties. The list keeps growing. Facebook was targeted to college students but it has expanded to include high-school students and now it is open to anyone. Club Nexus is an online community serving over 2000 Stanford undergraduate and graduate students [Adamic et al. 2003]. The creation of Yahoo! 360° and the acquisition of Dodgeball by Google are relatively recent examples where the importance of social network applications is evident not only considering the millions of users that some of them have but also due to the (even hundreds of) millions of dollars they are worth. Hence, it is not surprising that social network Web sites do not openly share their data. Other reasons for not doing so include privacy concerns. In some sites, the true identity of users is available only to their connections in the same network (e.g., Facebook, LinkedIn). Other sites such as LiveJournal publish the social connections of users openly, yet the true identity of users is (in most cases) hidden behind a nickname.

Although social network websites can provide data to detect COI, they are isolated even when their users might overlap a lot. That is, many people have accounts in more than one site. Moreover, much of the social information is still hosted in the distributed homepage-hyperlink style. Therefore, our case of demonstrating COI detection faces a big challenge: integration of different social networks. Meanwhile, our case also serves as a real-world showcase of Semantic Web technology. The Friend-of-a-Friend (FOAF) vocabulary can be used to publish information about persons, their relationships to workplaces and projects, and their social relations. We used a collection of FOAF documents from the Web where the *knows* relationship is explicitly stated. The aggregation of such FOAF documents by means of the knows relationship results in a social network. As a second network, we used the DBLP bibliography (dblp.uni-trier.de/), which provides collaboration network data by virtue of the explicit co-author relationships among authors. We made the assumption that this collaboration network represents an underlying social network. Although we anticipated significant challenges for the integration of the two networks, the effort needed in addressing this challenge surpassed our initial expectations. For example, DBLP has different entries that in the real world refer to the same person, such as the case of "Ed H. Chi" and "Ed Huai-hsin Chi." Thus, the need for entity disambiguation (also called entity resolution, or reference reconciliation) will likely continue to be a fundamental challenge in developing Semantic Web applications involving heterogeneous, real-world data. We believe that this integration effort of two social networks provides an example of how semantic technologies, such as FOAF, contribute to enhancing the Web.

This paper extends our previous work on semantic analytics on social networks [Aleman-Meza et al. 2006] where we demonstrated and explained the challenges of bringing together a semantic & semi-structured social network

ACM Transactions on the Web, v2n1, Feb 2008. Full 29-page copy at http://doi.acm.org/10.1145/1326561.1326568

(FOAF) with a social network extracted from the collaborative network in DBLP. We also introduced semantic analytics techniques to address the problem of COI detection and described our experiences in the context of a class of Semantic Web applications where COI was a simple yet representative application. In this article, our contributions go beyond those of the previous paper and can be summarized as follows.

—We verify scalability on bringing together a FOAF social network with the collaborative network in DBLP. We discuss the challenges in entity disambiguation to achieve integration of different social networks. Our evaluations demonstrate the need and feasibility of using large datasets (i.e., populated ontology with over 3 million entities).
—We improve upon our previous technique for COI detection by considering collaboration strength instead of basic co-authorship statistics. In addition, our new approach takes into account other relationships among people such as same-affiliation and co-editorship.
—We showcase the development process of creating scalable Semantic Web applications. Previously, we shed some light on what it takes to develop "connecting-the-dots" applications. Now we detail the challenges involved when, in addition, it is needed to use large-scale real-world datasets, in particular, social network data.

2. MOTIVATION AND BACKGROUND

This article intends to characterize the common engineering and research challenges of building large-scale practical Semantic Web applications rather than contribute to the theoretical aspects of Semantic Web. In fact, many of us in academia have seen multifaceted efforts towards realizing the Semantic Web vision. We believe that the success of this vision will be measured by how research in this field (i.e., theoretical) can contribute to increasing the deployment of Semantic Web applications [Lee 2005]. In particular, we refer to Semantic Web applications that have been built to solve commercial world problems [Miller 2005; Sheth 2005a; Sheth 2005b]. These include Semantic Search [Guha et al. 2003; Wasserman and Faust 1994], large scale annotation of Web pages [Dill et al. 2003], commercialized semantic annotation technology [Hammond et al. 2002] and applications for national security [Sheth et al. 2005]. The engineering process it takes to develop such applications is similar to what we present in this article. The development of a Semantic Web application typically involves the following multistep process.

1. *Obtaining High Quality Data*. Such data is often not available. Additionally, there might be many sites from which data is to be obtained. Thus, metadata extraction from multiple sources is often needed [Crescenzi et al. 2001; Laender et al. 2002; Sheth et al. 2002].
2. *Data Preparation*. Preparation typically follows the obtaining of data. Cleanup and evaluation of the quality of the data is part of data preparation.

ACM Transactions on the Web, v2n1, Feb 2008. Full 29-page copy at http://doi.acm.org/10.1145/1326561.1326568

3. *Entity Disambiguation*. This continues to be a key research aspect and often involves a demanding engineering effort. Identifying the right entity is essential for semantic annotation and data integration [Bergamaschi et al. 1999; Hassell et al. 2006].
4. *Metadata and Ontology Representation*. Depending on the application, it can be necessary to import or export data using standards such as RDF/RDFS and OWL. Addressing differences in modeling, representation and encodings can require significant effort.
5. *Querying and Inference Techniques*. These are needed as a foundation for more complex data processing and enabling semantic analytics and discovery [Anyanwu and Sheth 2003; Horrocks and Tessaris 2002; Karvounarakis et al. 2002; Sheth et al. 2002].
6. *Visualization*. The ranking and presentation of query or discovery results are very critical for the success of Semantic Web applications. Users should be able to understand how inference or discovery is justified by the data.
7. *Evaluation*. Often benchmarks or gold standards are not available to measure the success of Semantic Web applications. A frequently used method is comparing application output with results from human subjects.

These challenges are discussed throughout this article in the context of developing a large-scale application that addresses the problem of COI detection. Figure 1 illustrates the multistep process of building Semantic Web applications along with the steps involved in our approach for COI detection.

2.1 The Peer-Review Process

Throughout this article, we will focus on the peer-review process for scientific research papers. This process is commonly supported by semi-automated tools such as conference management systems. In a typical conference, (typically) one person designated as Program Committee (PC) Chair is in charge of the proper assignment of papers to be reviewed by PC members of the conference. Assigning papers to reviewers is one of the most challenging tasks for the Chair. State-of-the-art conference management systems support this task by relying on reviewers specifying their expertise and/or "bidding" on papers. These systems can then assign papers to reviewers and also allow the Chair to modify these assignments. A key task is to ensure that there are qualified reviewers for a paper. In addition, it is necessary to ensure that the reviewers will not have a-priori bias for or against the paper. These two requirements often conflict due to the trade-off between the two aspirations. Namely, a qualified reviewer is expected to be completely unbiased, yet s/he actually is a member of the same scientific community. Conference management systems can rely on the knowledge of the Chair about any particular strong social relationships that might point to possible COIs. However, due to the proliferation of interdisciplinary research, the Chair cannot be expected to keep up with the ever-changing landscape of collaborative relationships among researchers, let alone their personal

ACM Transactions on the Web, v2n1, Feb 2008. Full 29-page copy at http://doi.acm.org/10.1145/1326561.1326568

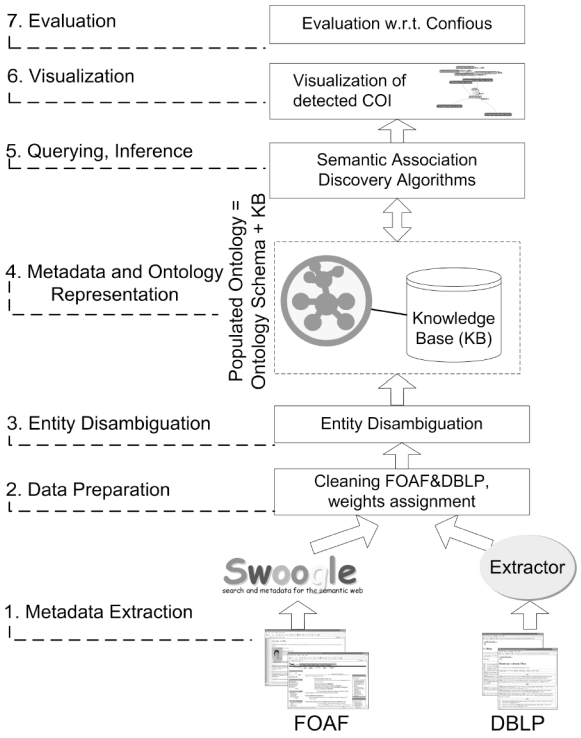

Fig. 1. Multistep process of Semantic Web applications.

relationships. Hence, conference management systems need to help the Chair with the detection of COIs.

Contemporary conference management systems support COI detection in different manners. EDAS (edas.info/doc/) checks for conflicts of interest based on declarations of possible conflicts by the PC members (e.g., while bidding for papers). Microsoft Research's CMT Tool (cmt.research.microsoft.com/cmt/) allows authors to indicate COI with reviewers. Confious (confious.com) automatically detects conflicts of interest based mainly on similar emails or coauthorship criteria. The similar email criterion tries to identify PC members and authors who are affiliated with the same organization based on the suffixes of the email addresses. The coauthorship criterion identifies users who have co-authored at least one paper in the past. However, Confious's relatively straightforward approach can miss out on COIs as exemplified by one recent case of a coauthor who now has a hyphened last name. On the other hand, this is a good example of how difficult COI detection might be. The approach presented in this article makes use of social relationships to detect COI other than just based on co-authorship. This is possible by combining DBLP data and FOAF data through entity resolution, which is the main improvement over Confious.

2.2 Online Social Networks

"A social network is a set of people (or organizations or other social entities) connected by a set of social relationships, such as friendship, co-working, or

ACM Transactions on the Web, v2n1, Feb 2008. Full 29-page copy at http://doi.acm.org/10.1145/1326561.1326568

information exchange" [Garton et al. 1997]. Social networks are receiving a lot of attention on the Web due to the increasing number of Web sites that allow users to post their personal information directly into online networked information spaces. The users of such Web sites form virtual or online communities that have become part of the modern society in many contexts such as social, educational, political and business.

The entity Person is the fundamental concept in online social networks. An entity can be identified by one or several of its properties, and different sources might use different set of properties. For example, a person can be identified by his/her name in an office, but will be identified by his/her policy number by an insurance company. Such heterogeneous contexts and entity identifiers necessitate entity disambiguation. A *link* is another important concept in social networks. Some sources directly provide links among person entities such as foaf:knows (where foaf refers to the FOAF namespace http://xmlns.com/foaf/0.1/). Other links, such as co-author, can be derived from metadata of publications.

Some of the online social networking sites provide machine readable personal information data using RDF/XML and FOAF vocabularies. Depending on the privacy policy of each Web site, the scope of published personal information ranges from nicknames and interests to sensitive information (e.g., date of birth). We acknowledge that there are privacy issues but a discussion on this topic is out of the scope of this article.

2.2.1 *Social Networks Analysis.* Social network analysis focuses on the analysis of patterns of relationships among people, organizations, states, etc. [Berkowitz 1982; Wasserman and Faust 1994; Wellman 1998]. Social network analysis has applications in analysis of networks of criminals [Xu and Chen 2003], visualization of co-citation relationships [Chen and Carr 1999] and of papers [Chen 1999], finding influential individuals [Nascimento et al. 2003; Smeaton et al. 2002], study of the evolution of co-authorship networks [Barabási 2002], etc. Our previous work in this respect demonstrated an ontological approach in integrating two social networks and using "semantic association" discovery techniques for identification of COI relationships [Aleman-Meza et al. 2006].

3. INTEGRATION OF TWO SOCIAL NETWORKS

In order to demonstrate our approach to the problem of COI detection, we bring together a semantic social network (FOAF) with a social network extracted from the underlying coauthorship network in DBLP. Here we describe these sources and explain the challenges involved with respect to entity disambiguation that have to be addressed to merge entities across (and within) these sources that in real-world refer to the same person.

3.1 Choosing Data Sources: FOAF and DBLP

We selected two representative online data sources for constructing two independent social networks and then we combined them into one social network