Association Analytics for Network Connectivity in a Bibliographic and Expertise Dataset

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

Boanerges Aleman-Meza, Sheron L. Decker, Delroy Cameron, I. Budak Arpinar

Large Scale Distributed Information Systems (LSDIS) Lab
Computer Science Department, University of Georgia
Athens, GA 30602-7404, USA

Keywords
Semantic Web, Ontologies, RDF, DBLP, Bibliography Measures, Centrality, Collaboration Strength, Expert Finder.

1. 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 SwetoDblp ontology of DBLP data (lsdis.cs.uga.edu/projects/semdis/swetodblp/), and Andreas Harth's DBLP dataset in RDF (sw.deri.org/~aharth/2004/07/dblp/). Various studies have used DBLP data 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 datasets. In this paper, we describe a study of network connectivity in bibliography data. Our work expands upon earlier studies that have used subsets of DBLP data for analysis of collaborations in the field databases (Elmacioglu, 2005; Nascimento, 2003). The dataset we use includes not only data of publications in database field but also of research areas such as Artificial Intelligence, Web and Semantic Web.

Additionally, we describe how publication metadata from DBLP can be used for the creation of a dataset of topics and terms in Computer Science. Metadata of publications was used as a starting point for Web extraction of keywords. Due to the large number of keywords and terms that appear in abstracts of publications, it is possible to exploit this information for finding the most common terms. Based on this, we were able to identify potential terms that could be used in building a taxonomy of Computer Science topics. The main benefit is that these topics can be suggested to human and therefore the time required to build a taxonomy can be shortened. Additionally, the suggested terms come from and reflect the data itself (i.e., computer science publications). The identified terms can be analyzed to determine which are appearing only in the last few years. This can be the means to recognize possible new topics in Computer Science research. After keywords and terms are extracted, the relationships from terms to their respective publications can be seen as an indication that all authors of a paper have knowledge on the topics of the paper. Thus, if we look at an author in particular, it is possible to determine the topics on which s/he has expertise/knowledge. This is the basis to identify researchers that have high expertise on certain topics (according to the extracted data). We perform a study to validate this measure of expertise against lists of recognized researchers based on available lists of ACM fellows and IEEE fellows. We describe our choices for implementation in respect to RDF database used. We argue that this type of study can be done with existing Semantic Web technologies that are able to handle large datasets. We 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 terms 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.

2. 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 evolves and the characteristics of the social or collaborative interactions among authors. Many techniques for analysis of bibliography data have their roots or are related to social networks analysis, which focuses on the analysis of patterns of relationships among people, organizations, states and such social entities.

A quite common analysis in networks is that of determining whether the small-world phenomenon exists. The intuition comes from the "six degrees of separation" experiment by Milgram (1967). Many networks where humans participate exhibit a small-world phenomenon. Bibliography data is no exception. It could be said that a network (where humans participate) that does not exhibit such phenomenon might require revising whether the data has been correctly extracted. Thus, we verify in our analysis that the collaboration network indeed exhibits a social network phenomenon.

Related efforts in the literature that have addressed analysis of publications include analysis of publication venues in the field of Databases (Elmacioglu, 2005; Nascimento, 2003). In the area of Semantic Web, Golbeck (2006) addressed analysis of co-authorship, citation, and affiliation for authors of ISWC conferences. Their focus was more on visualization as compared as our approach involving analysis and highlighting the benefits of using semantics. Similarly, analysis of communities in Semantic Web has taken place by querying a search engine with names of researchers and research topics to determine associations between people and concepts (Staab, 2005).

We exploit the value of relating keywords and terms to authors in publications for the purpose of determining areas of expertise of researchers. Al-Sudani (2006) described this idea intended for finding knowledgeable personnel in certain areas of interest. However, they used a much smaller dataset of publications. In fact, they point out that data collection/extraction is a time-consuming task. We believe that our approach circumvents such problem by using the metadata itself of publications for selecting URLs that contain keywords and terms metadata to be extracted. Some manual work has to be done, in our case, for the creation of a web-scrapper for a specific web source such as ACM Digital Library. The advantage is that once such metadata is extracted, it can be safely assumed that it is not going to change. That is, the keywords of a published article will always remain unchanged.

The benefits of using semantics for expressing expertise or areas of interest for persons have been highlighted in a variety of scenarios and applications (Aleman-Meza, 2007). In fact, the ExpertFinder Initiative intends to identify use cases, challenges, techniques, etc. for semantics-based representation, retrieval, and processing of expertise data (rdfweb.org/topic/ExpertFinder). There is a close relationship between determining topics of papers to the use of such topics in determining expertise of authors. In fact, the topics of papers can be used, together with their date, to find trends in research areas (Decker, 2007; Tho, 2003).

3. ANALYTICS IN THE BIOBLIOGRAPHY DATASET

We selected several of the techniques for network analysis that were part of earlier studies of the Databases community (Elmacioglu, 2005; Nascimento, 2003). However, instead of simply replicating their work with an

updated dataset, we aim at demonstrating that further insight is possible by using RDF-encoded data. The data we use comes or is derived from DBLP. Where indicated, we used a subset of DBLP publications in the areas of Artificial Intelligence, Databases, Data Mining, Information Retrieval, Web and Semantic Web. We will refer to this subset as DBLP-subset.

3.1 Statistics about Authors

Centrality. There are known methods to identify participants in a network that are highly connected to the rest. The *closeness centrality* measure identifies how close an author is, on average, to all other authors in the network. Authors with low *closeness* values are connected to many authors within short path distances. Hence, it could be said that their 'influence' in the network is high. We computed centrality as the average of the shortest path that an author has to each author. Table 1 lists the top 10 *central* authors from the largest connected component in DBLP-subset. The first column lists authors computed by simply taking their name as they appear in DBLP.

Table 1. Top 10 *centrality* authors in DBLP-subset.

Centrality using name		**Centrality using *same-as* information**	
Score	**Author Name**	**Score**	**Author Name**
4.0578	Gio WiederHold	3.9859	Gio WiederHold
4.1527	Richard T. Snodgrass	4.0517	Umeshwar Dayal
4.1900	Umeshwar Dayal	4.0616	Richard T. Snodgrass
4.2020	Philip A. Bernstein	4.0825	Elisa Bertino
4.2025	Elisa Bertino	4.1028	Christos Faloutsos
4.2087	Christos Faloutsos	4.1335	Philip A. Bernstein
4.2232	Kenneth A. Ross	4.1431	Christian S. Jensen
4.2299	Hector Garcia-Molina	4.1487	Jiawei Han
4.2340	David Maier	4.1535	Kenneth A. Ross
4.2427	Christian S. Jensen	4.1605	Erich J. Neuhold

It has been noted that DBLP does not have unique ID for authors (Elmacioglu, 2005). However, it could be said that the name of an author plays the role of a primary key. For the cases when different persons have the same name, a numerical value is appended in the name to differentiate the two entries in DBLP. For the cases when the same person is referred to in two (or more) forms, then such names (i.e., aliases) are related explicitly, we refer to these as 'same-as'. Common reasons for people having two names are the use of a shortened name (e.g., Tim and Timothy) and changes due to addition of hyphened name or middle initial. There are very few entries in DBLP data for authors with more than one name – probably due to the difficulty of detecting such ambiguities automatically. However, it is quite important to make use of information stating that two names refer to the same person. Otherwise, the publications count of an author that has two names would be incorrect. Similarly, co-authorship measures would miss out due to incorrectly counting the right number of co-authors. We compared results obtained when 'same-as' information is used in computing centrality scores of authors. Table 2 lists a couple of examples of authors that appear in DBLP-subset with more than one name. Each name appears with its own centrality score. It is noticeable how much of a change exists on the computed centrality score in the case of Alon Y. Halevy when both of his names spellings are considered. In the case of Timothy W. Finin, his centrality score is also smaller but his position among all computed centrality scores moves from 94 to 101. This happens because the positions of authors computed using same-as information affect not only authors that have more than one name, but also the scores of other authors in the network. This is quite evident in the second column in Table 1, which lists authors when their centrality score is computed using same-as information. It is interesting that the effect of using same-as information is such that the top *centrality* authors differ in both columns.

Table 2. Examples of improved centrality score by considering the 'same-as' information available

Using 'same-as' information		**Without 'same-as' information**	
Name of researcher	Centrality score	Names of researcher in the dataset	Centrality score
Alon Y. Halevy (37)	4.2707	Alon Y. Levy (51)	4.4026
		Alon Y. Halevy (111)	4.5498
Timothy W. Finin (101)	4.4051	Timothy W. Finin (94)	4.5123
		Tim Finin (1430)	5.0747

Collaborators Distribution. The distribution of number of collaborators per author, shown in Figure 1, clearly exhibits the power-law tail. This indicates that a large number of authors have a small number of collaborators (up to around 10). A much smaller number of authors have around 100 collaborators. A small number of authors have many publications (e.g., over 150). They would be the most likely authors to have many collaborators. Hence, the distribution of collaborators per authors indicates that the data exhibits a small-world phenomenon.

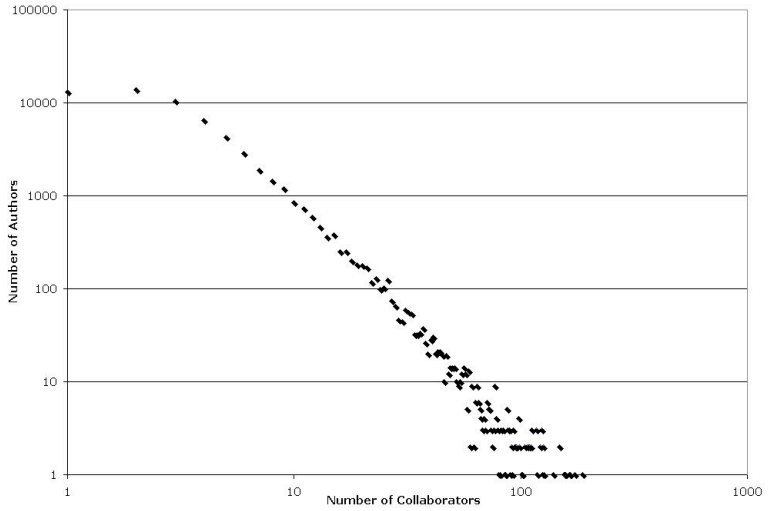

Figure 1. Distribution of collaborators per author

Collaboration Strength. Lastly, we measured the collaboration strength among authors. This method helps on identifying pairs of authors that collaborate frequently. For example, is expected that pairs of researchers with highest collaboration strength are those who work at the same organization for a long time and collaborate frequently. Instead of simply finding authors with highest frequent co-authors, we use a method that takes into account the number of authors in a paper as well as the number of papers that two people co-authored (Newman, 2001). The method adds a weight of $1/(n-1)$ to the collaboration strength of two authors for each paper they co-authored together (n is the number of authors in a paper). This measure captures quite well the collaboration among authors where a publication has very few authors. The assumption is that their collaboration strength is higher than in the case of publications with a large number of co-authors. Given that the computed collaboration strength for any two co-authors is symmetric, we show in Table 3 the highest ten pairs of collaborating researchers in the DBLP-subset. As expected, most of the highest collaborating researchers work/worked at the same organization. Only a few highest-collaborating researchers do not work at the same place.

Table 3. Highest ten pairs of collaborating researchers in DBLP-subset.

Highest Collaborating Researchers	Collaboration Strength
Amr-El Abbadi – Divyakant Agrawal	57.3
Didier Dubois – Henri Prade	42.1
Beng Chin Ooi – Kian-Lee Tan	28.5
Charu C. Aggarwal – Philip S. Yu	28.4
Dimitris Papadias – Yufei Tao	21.4
Ee-Peng Lim – Wee-Keong Ng	21.4
Katsushi Inoue – Itsuo Takanami	19.4
Paolo Atzeni – Riccardo Torlone	19.0
Rakesh Agrawal – Ramakrishnan Srikant	18.0
Nick J. Fiddian – W. Alex Gray	17.8

3.2 Statistics about Papers

Common statistics about papers include computing the number of papers per authors and number of papers per year. However, our intention is to demonstrate that other statistics can be computed with a dataset represented using Semantic Web techniques. Hence, we chose to determine the number of different affiliations per year. This requires authors of papers to have affiliation information. Data from DBLP alone does not contain such information. We used the SwetoDblp ontology, which is created from DBLP data and includes affiliation data for some of the authors. SwetoDblp extracts affiliation of authors based on their homepage (whenever possible). DBLP contains homepage information for little over 10K authors. SwetoDblp extracts affiliation information for 7K of them (as of June 2007). Figure 2 illustrates the number of papers per year together with the number of affiliations of authors per year in DBLP-subset.

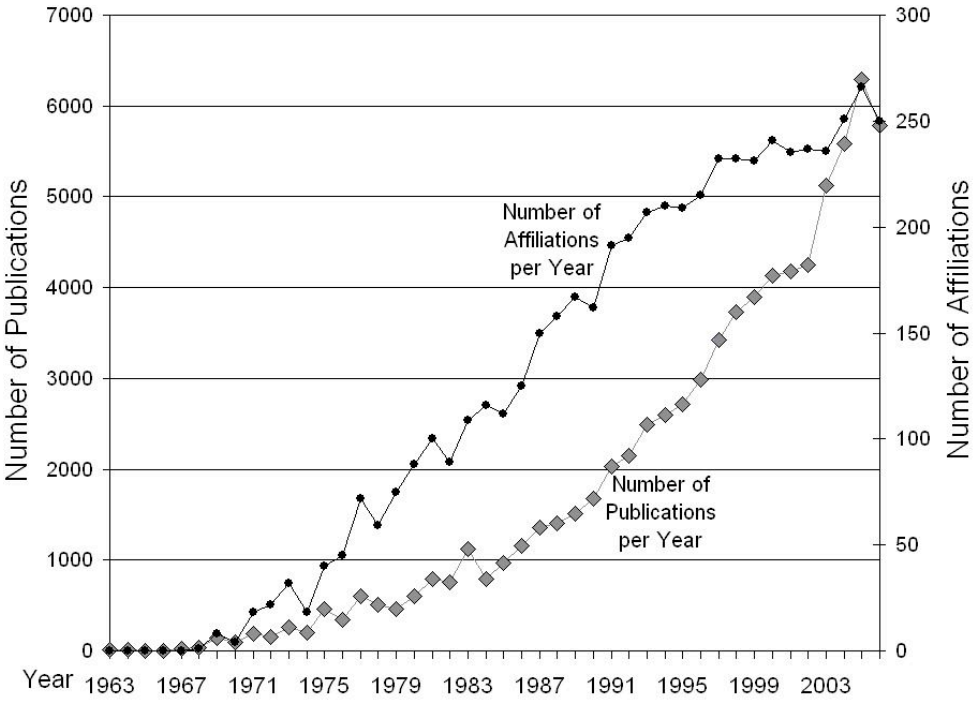

Figure 2. Number of Publications per Year and Number of Affiliations per Year

It can be seen that the number of affiliations (organizations) does not increase as quickly as the number of papers per year. Another interesting analysis that is possible is that of finding affiliations that appear only within the last couple of years. The intention is to find out which were the organizations that are relatively new into publishing research in Computer Science. Table 4 lists the URLs of such *new* organizations (using data of DBLP-subset). These results could be more accurate if the homepage (and affiliation) of authors were completely available and up to date. In fact, only about 1.8% of authors in DBLP have *homepage* information.

Table 4. Affiliations identified as only appearing within the last three years.

Relatively Newly Appearing Affiliations
www.curtin.edu.au, www.fudan.edu.cn, www.hartford.edu, www.isu.edu, www.nuaa.edu.cn, www.qc.edu, www.research.ibm.com/beijing, www.seu.edu.cn, www.uoguelph.ca, www.ustc.edu.cn, www.zjnu.edu.cn, www.zju.edu.cn

4. Linking Bibliography data to Expertise Data

Identification of researchers that have expertise on a specific research topic could be of great value. The value of extracting expertise data has been noted in other efforts (Hezinger, 2004; Mika, 2005). For example, the National Science Foundation (NSF) is an agency that funds projects in a wide arrange of research areas. They may wish to inquire on the productivity of the researchers that they have provided funding for with respect to specific research areas. Identifying experts on topics could help validate that their funding within certain areas has had a positive/productive impact in the research community. Moreover, a researcher on a new area could determine which publications are of importance for background information based on which researchers were identified to be experts in the area. To help facilitate the need of discovering such individuals, publications entities in a RDF dataset can be related to topics within a taxonomy. The authors of papers can then be implicitly related to topics for the purpose of identifying who is knowledgeable or has expertise on specific topics. The assumption is that authors of a paper have expertise on the topics of their papers.

In our previous work, we created a taxonomy of computer science topics, mostly in areas of Databases, Web and Semantic Web (Decker, 2007). The taxonomy was verified and adjusted based on the AKT ontology (Shadbolt, 2004) and CoMMa ontology (Gandon, 2001). The majority of topics within the taxonomy were based on brainstorming, discussion and feedback amongst several colleagues. However, we feel that the taxonomy was limited in regards to only the knowledge we pertain. Therefore, the decision was made to construct a taxonomy from scratch using data extraction methods from reliable computer science sources in order to obtain relevant terms from the data itself. Arguably, this would allow for the taxonomy to include past, present and emerging topics.

4.1 Taxonomy of Topics in Computer Science

The process of building a thorough taxonomy of topics in Computer Science is an arduous task. There are computer science classification systems readily available that could have been re-used in our approach. For instance, ACM's Computing Classification System (acm.org/class/1998/) provides a categorization of computer science related topics intended to reflect the current state of the field. It contains eleven primary research areas each including numerous subtopics. However, the system is comprised of a very "broad" four-level tree of topics that would not be very beneficial recognizing topics that are manifesting today. For example, a publication entitled "Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection" was classified with ACM's CCS with the primary topic 'Information Systems' because no other topics such as social networks, semantic analytics, or conflict of interest that were available. Therefore, we developed our own taxonomy of computer science topics that would help identify "newer" terms. Identification of newer terms is advantageous for the purpose of recognizing possible emerging trends that might be included in a taxonomy of topics.

In order to ensure that our taxonomy was comprised of the most relevant topics, we decided to use extracted data from DBLP. A number of publication venues (over 50 conference series and journals) were selected that include areas of Web, Databases, Semantic Web, and Artificial Intelligence. We selected papers in such publication venues for extracting data that will be used in creating a taxonomy of topics. The main aspect of our approach is how we retrieved a metadata of papers with the use of the electronic edition "ee" URL of individual papers (in DBLP). URLs having dx.doi.org/10.1016, doi.acm.org, or doi.ieeecomputerscociety.org were crawled to retrieve "keywords" and "abstracts" for the purpose of identifying a surplus of terms that are related to computer science. We experimented using metadata of keywords and abstracts separately. Using keywords alone brings limited data that does not have much added value from the research areas included in ACM's Computing Classification System. On the other hand, by incorporating terms extracted from the abstracts the method aided in identifying "newer" terms. The extraction of terms from abstracts was done using Yahoo! Term Extraction (developer.yahoo.com), which identifies phrases and terms from a given input text. We define newer terms as terms that have not appeared within publications before a certain year, in this case we selected the year 2005. Table 5 lists examples of terms that best illustrate newer terms identified with our approach. This was accomplished by determining which papers within our dataset labeled each term as keywords or included the term within its abstract and then retrieved the dates of those publications. A benefit of this approach is that it can keep up with changes in the field. In fact, Hepp (2007) pointed out the need for ontology engineering methods to quickly reflect domain changes to keep ontologies up to update. Our approach is based on whatever terms are contained in the data (keywords and abstracts) instead of creating a taxonomy on brainstorming or other methods with the intent to not limit our taxonomy with personal knowledge of topics in the field. Additionally, our method uses only the metadata and abstracts, without having to process the whole document content.

Table 5. Some of the identified terms appearing on year 2005 and afterwards.

Friendship, grid middleware, grid technology, phishing, protein structures, service oriented architecture (SOA), social network analysis, spam, wikipedia

Our approach retrieved more than 280 potential topics to consider for building the taxonomy. However, methods were used to narrow the results list because several of the terms and phrases were not relevant to computer science. As a means to retain the most common research topics accumulated, we kept a record of how many times each potential topic appeared. This allowed us to identify terms and phrases that were highly used as keywords and words within abstracts. Table 6 lists ten of the most frequently identified terms within the last ten years.

Table 6. Few of the top terms identified from URL extraction within last ten years.

Topic	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Algorithm(s)	87	99	111	89	219	222	381	418	608	71
Classifier(s)	0	7	1	2	33	30	47	80	94	5
Data Mining	12	10	20	13	46	62	88	104	184	8
Databases	13	17	19	19	28	32	43	53	63	6
Semantic Web	0	0	0	4	13	24	102	85	96	14
Semantics	19	16	26	22	28	24	90	75	86	11
Web Service(s)	0	0	0	0	4	2	67	82	69	1
XML	0	4	4	11	22	20	36	58	54	1

In identifying some of the most frequently identified terms in our approach, we were able to make three key observations pertaining to the results. First, we noticed that terms can be covered in a wide arrange of areas. Therefore, this may constitute for an extremely high volume count of a term compared to other terms. For example, the term *Algorithms* if a very broad term that is not only used as a reference to a research area but also as a means of defining or describing a particular method or technique. This is probably the reason why it appears so many times. Secondly, for a term such as *Databases*, which one would expect to appear more times than shown, we discovered that the total number of appearances is relatively small due to the large amount of synonyms used to represent this particular term. For example, data base, data bases, database management system, database management systems, and DBMS. Hence, if *Databases* is a topic in a taxonomy, then its synonyms should be added as alternate spellings of the term. Thirdly, we were able to identify broader terms, such as the term *Semantics*, which has been used in literature for several years. Although this term has been long used, we were able to detect related terms that have emerged within recent years, case in point being the term *Semantic Web*. This shows that the total number of appearances for these broader terms could be due to newer terms that are related to terms that have been used for a longer time.

The structure of our taxonomy was put together by determining how close topics are related. Our approach began by first retrieving all the URLs of the publications of each term from which the terms were included within. We then added each URL into a *set* for each term. Relationships among terms were identified using measures calculated from the intersection of the sets of two terms divided by the union of the sets. This would produce a measure ranging from 0 (which implies the two topics are not related) to 1. Pairs of terms with a value above 0.05 were considered to be related terms. The identification of relationships aids in building a tree-level organization of topics that can later turn into a taxonomy. Figure 3 illustrates examples of topics and their identified relations. Other approaches have done similar work in identifying relationships of topics. In the work by Mika (2005), research topics were identified specifically from the interests of researchers within a Semantic Web community. The associations between the topics were based on the number of researchers who have an interest in the given pair of topics. Our approach instead identifies computer science topics by means of crawling of the DBLP dataset and further data extraction; whereas in their work the topics were already known based on the supplied interests of researchers from FOAF. The work of Velardi (2007) is an example of research on taxonomy learning. In our work, we intend to demonstrate that the basic steps for suggesting terms in building a taxonomy can be achieved with off-the shelf tools such as Yahoo! term extraction.

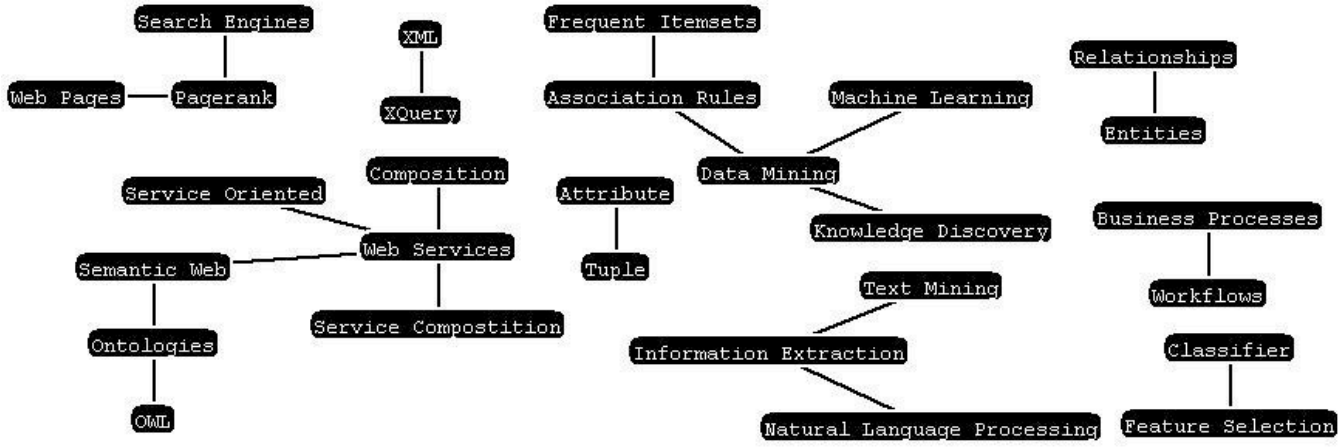

Figure 3. Snippet of Identified Relationships Among Terms

4.2 Measuring Expertise

A dataset abundant with publication data is important for accurately characterizing the knowledge areas and/or expertise of a researcher. The greater the number of publications that can be linked to various topics, the more accurately "Expertise Profiles" can be captured and represented for a researcher. We exploit this principle in order to obtain rich and accurate expertise profiles for each researcher. It is noteworthy to emphasize that the publication venue of a publication is also helpful in determining the commensurate weight of a publication. For example, a workshop publication might have lesser impact than an article in a high impact journal. Existing publication-impact data of various publication venues (conferences, journals, etc) thus play a role in determining the expertise of researchers.

In our previous work, we detailed an approach for measuring expertise in order to find relevant reviewers for Program Committee (PC) membership in a Peer-Review process (Cameron, 2007). Such expertise measure considered not only the number of publications in developing the profile of a researcher, but also the aggregated impact of the publications across a given area. The use of a taxonomy of topics provided the additional benefits of organizing topics in a hierarchy for querying and/or aggregation. An evaluation of such approach yielded plausible results through comparison with PCs of past conferences. A good percentage of the experts recommended by the system ranked highly in PCs of previous conferences. However, a thorough evaluation of methods that compute expertise of researchers is a challenging task due to the many factors in which expertise can be compared (many of which are of subjective nature). We expand the evaluation of measuring expertise by exploiting the availability of bibliography data. This can be achieved by comparing detected top *experts* by the system against awards where experts or influential individuals are determined or selected by humans.

4.2.1 Comparison with 'social' measures of technical achievements and recognitions

Recognitions of outstanding research accomplishments are an important aspect of many research communities. Professional associations such as the Association of Computing Machinery (ACM) and IEEE recognize distinguished fellows annually across a wide variety of areas in Computer Science. IEEE has a history of distinction of deciding with some unanimity those worthy of recognition across many areas of Engineering and Technology at large. We perceive these organizations as credible sources for validating experts. However, newer forms of recognition include sites such as Wikipedia (www.wikipedia.org). The content of Wikipedia is compiled from a large number of participants. However, the mechanisms in Wikipedia make it extremely difficult to create (and keep) a new Wikipedia entry for a person. That is, a Wikipedia page about a person can be created only if such person is arguably famous, has an important position, has important achievements, etc. Hence, we assume that Wikipedia entries about Computer Science researchers can be viewed as evidence of their important contributions. In fact, there are many Wikipedia pages for pioneers in Computer Science research. Table 7 shows our findings from the comparison of our SEMEF application and the Class of 2007 ACM Fellow Inductees, Wikipedia and IEEE Fellows. Other measures of researchers with high impact are based on their citation impact. For example, the h-index (Hirsch, 2005) could be used to validate experts determined by our method. However, it would require extensive manual work to determine the h-index of researchers by topic, mostly due to the fact that citation information is needed.

Table 7. Comparing experts identified by the system against recognized experts

Researcher	Rank without Taxonomy	Rank with Taxonomy	Award or Recognition	Topic	Contributions/Explanation of the Award or Recognition
Rakesh Agrawal	6	4	ACM Fellow	Data Mining	"... data mining"
Ming-Syan Chen	15	16	ACM Fellow	Data Mining	"... query processing and data mining"
Susan B. Davidson	11	12	ACM Fellow	XML	"... distributed databases, … semi-structured data …"
C. Lee Giles	78	19	ACM Fellow	Search	"... information processing and web analysis"
Jiawei Han	1	1	ACM Fellow	Data Mining	"... knowledge discovery and data mining"
Rudi Studer	20	33	Wikipedia Person	Ontologies	"... query processing and data mining"
Philip S. Yu	2	2	ACM Fellow	Data Mining	"... theory and practice of analytical performance modeling of database systems"
Amit P. Sheth	40	3	IEEE Fellow	Web	"... information integration and workflow management"

We make a couple of observations based on these results. First, we recognize the importance of using a taxonomy of topics for finding experts. For example, consider the case of researcher C. Lee Giles, he appears to have a significantly larger number of publications in the subtopics of research topic "Search" rather than the topic itself. His rank on this research topic increased almost 60 spots when including publications under the subtopics of Search within the taxonomy. A similar situation is also evident for researcher Amit P. Sheth. Many of his recent publications span subtopics of the "Web" topic (e.g., Semantic Web and Web Services). On the other hand, researchers Rudi Studer and Ming-Syan Chen had their expertise rank decreased when considering their publications in subtopics of the given topic. This alerts us that there may be in many cases other experts whose expertise in those subtopics surpasses them, while not true for the topic itself. In other cases still, including Phillip S. Yu, Susan B. Davidson and Jiawei Han, the inclusion of the taxonomy of topics does not affect their ranking in the topics listed in the table – an indication that the larger percentage of their publications are in data mining itself. We present the following conclusion based on these findings. The taxonomy of topics is important in determining expertise at finer levels of granularity. In many cases it identifies experts whose areas of expertise are at very specific levels, while not particularly broad in scope given a specific topics. In other cases, experts whose expertise is of greatest relevance are identified quite easily. Lastly, researchers whose expertise is distributed with some degree of consistency across the topic and subtopics of the given research area are also easily identifiable.

The second major observation we make based on the results in Table 7 is the "Extent of Overlap" of the topics of expertise of the researchers identified by computer system and the actual explanation listed in their recognition (e.g., award) according to their appropriate areas of expertise. For example, the area of Data Mining produced based on our application produced close to 1,400 researchers with some expertise in the field. Many of the Fellows we show in the Table 7 are in the top 1% of those experts identified by the system. We feel that this overlap shows promise of the validity of using a computer system for identifying experts.

5. Experiments Setup

Most of the data used in this study comes or is derived from DBLP (as of June 2007). The analysis that uses topics of expertise was done using data of publications in the areas of Artificial Intelligence, Databases, Data Mining, Information Retrieval, Web and Semantic Web. The selection of publications on these areas was achieved selecting 28 journals and 112 conferences, symposiums and workshops. The publication venues selected is a superset of those used by Elmacioglu (2003). Selecting a subset of DBLP publications might seem a tedious task but it was relatively easy to do thanks to the naming convention that DBLP uses for BibTex entries of publications. For example, all publications in the World Wide Web Conference have "http://dblp.uni-trier.de/rec/bibtex/conf/www/" as prefix. The list of all prefixes used to create the subset we used, as well as other datasets mentioned here, are available online (http://lsdis.cs.uga.edu/~aleman/research/sa/).

The list of Computer Science authors that appear in Wikipedia was extracted in part from DBpedia (Auer, 2007) and from authors in DBLP having as homepage a Wikipedia entry. Most of the analysis was done with the SwetoDblp dataset in RDF containing DBLP data plus additions such as affiliations and organizations. We utilized BRAHMS system (Janik, 2005) for fast processing of the 919MB rdf/xml file. The DBLP-subset (file size of 100MB) was created from such file.

6. CONCLUSIONS

We presented some of the benefits of using semantics for analysis in a bibliographic dataset. For example, we found that the total number of universities affiliated with researchers is on the rise yet not at the same pace of publications from year to year. Centrality measures were also determined for researchers of publications included in our dataset. However, it was quite clear that there are benefits of using, if available, information of researchers that have more than one name or alias. Without using such 'same-as' information of researchers, the coputation of centrality values won't be correct. We were able to create a taxonomy of topics using metadata of keywords and terms from abstracts taking as starting point links of publications from DBLP. Our methods for extracting potential terms for the taxonomy were very effective in identifying topics that have been researched for many years and topics that are currently emerging. For example, terms that appear most frequently in the last few years include *phishing*, *spam*, and *wikipedia*. Then, using the terms related to papers, it is possible to determine areas of expertise of the authors. We used lists of ACM and IEEE fellows to compare with the experts determined using computer method. The areas for which such recognized researchers received their awards did match quite well with topics for which they are ranked very high in the computer method. In addition, we compared their rank with and without the use of a taxonomy of topics and found out that by using the taxonomy, the rank of the experts is a better match to what their expertise is. That is, they rank higher when the taxonomy is used. The current study and its evaluation show evidence of the promise for measuring experts on topics using a taxonomy-based approach but for future work we plan to do an analysis in more detail by considering multiple topics of expertise of a person such.

7. FUTURE RESEARCH DIRECTIONS

It is relatively straightforward to analyze bibliography data yet data about researchers also spans aspects such as social networks, events and blogs. Interlinking these aspects can provide options for analysis such as finding how certain communities interact. For example, which community is more active in the blogsphere? Or, which community has a denser social network, independent of its collaboration network. In addition, data quality issues remain, such as affiliation data. In our experiments, we found that special attention should be paid for organizations that have divisions that are referred or named in a variety of different ways. For example, it is useful to keep affiliation information of a researcher at IBM India Research Labs yet at the same time, a query or inference should take into account that such affiliation implies that it is part of IBM Corporation.

The compilation of metadata from papers based on its keywords and abstracts can be improved. In our work, we found that the information on some publishers' websites was somewhat difficult to extract. Thus, it is possible that the detected *new* terms might not have been new in reality. There are efforts by some publishers to make their information easier to access, such as by means of content feeds in XML. However, they rarely provide all relevant metadata items of a publication. The benefits of making available such information in machine processable formats can lead to better dissemination of the latest publications. Moreover, using richer metadata for determining topics on the field can lead to improved measures of the areas of expertise of researchers. A key aspect in this respect is to assign identifiers (e.g., URIs) for authors of papers towards solving ambiguity issues.

The measures of expertise of researchers could be somewhat controversial. However, this issue could be turned around so that authors themselves could help on improving the expertise data. For example, a researcher whose publications do not contain all metadata details might want to provide such data herself so that her expertise profile could be more complete. Citation-count is an important indicator of the impact of research. For example, the h-index (Hirsch, 2005) requires citation data to compute the h-number of a researcher. It might be

difficult to convince someone to provide machine processable details of the citations included in her papers. However, she might have a motivation to indicate which papers cite her papers because this type of data would directly impact her citation count. If a majority of researchers provide such information, existing measures of expertise or publications impact would have more practical value.

8. REFERENCES

Al-Sudani, S., Alhulou, R., Napoli, A., & Nauer, E. (2006). OntoBib: An Ontology-Based System for the Management of a Bibliography. Paper presented at the 17th European Conference on Artificial Intelligence, Riva del Garcia, Italy.

Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J. G., Mochol, M., Nixon, L. J. B., Polleres, A., & Zhdanova, A.V. (2007). Combining RDF Vocabularies for Expert Finding. Paper presented at the 4th European Semantic Web Conference. Innsbruck, Austria.

Auer, S. & Lehmann, J. (2007). What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. Paper presented at the Fourth European Semantic Web Conference. Innsbruck, Austria.

Bizer, C. (2003). D2R MRP - a Database to RDF Mappling Language. Paper presented at the Twelfth International World Wide Web Conference, Budapest, Hungary.

Cameron, D., Aleman-Meza, B., Decker, S., & Arpinar, I. B. (2007). SEMEF: A Taxonomy-Based Discovery of Experts, Expertise and Collaboration Networks. (Tech. Rep. No. 1114806563). University of Georgia, Computer Science Department.

Decker, S. L., Aleman-Meza, B., Cameron, D., & Arpinar, I. B. (2007). Detection of Bursty and Emerging Trends towards Identification of Researchers at the Early Stage of Trends. (Tech. Rep. No. 11148065665). University of Georgia, Computer Science Department.

Elmacioglu, E., & Lee, D. (2005). On Six Degrees of Separation in DBLP-DB and More. *SIGMOD Record, 34*(2), 33-40.

Gandon, F. (2001). Engineering an Ontology for a Multi-Agent Corporate Memory System. Paper presented at the Eighth International Symposium on the Management of Industrial and Corporate Knowledge. Université de Technologie de Compiègne, France.

Golbeck, J., Katz, Y., Krech, D., Mannes, A., Wang, T. D., & Hendler, J. (2006). PaperPuppy: Sniffin the Trail of Semantic Web Publications. Paper presented at the Fifth International Semantic Web Conference, Athens, Georgia, USA.

Hepp, M. (2007). Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies. *IEEE Internet Computing, 11*(1). 90-96.

Hezinger, M. & Lawrence, S. (2004). Extracting Knowledge from the World Wide Web, *PNAS 101*(supplement 1). 5186-5191

Hirsch, J. E. (2005). An Index to Quantify an Individual's Scientific Research Output. *PNAS 102*(46), 16569-16572.

Janik, M. & Kochut, K. (2005). BRAHMS: A WorkBench RDF Store and High Performance Memory System for Semantic Association Discovery. Paper presented at the Fourth International Semantic Web Conference. Galway, Ireland.

Mika, P. (2005). Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks. *Journal of Web Semantics, 3*. 211-223.

Milgram, S. (1967). The Small Wold Problem. *Psychology Today 2*, 60-70.

Nascimento, M. A., Sander, J., & Pound, J. (2003). Analysis of SIGMOD's Co-Authorship Graph, *SIGMOD Record, 32*(3), 8-10.

Newman, M. E. J. (2001). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, *Phys. Rev. E 64*. 016132

Shadbolt, N., Gibbins, N., Glaser, H., Harris, S., & Schraefel, M. m. c. (2004). CS AKTive Space, or How We Learned to Stop Worrying and Love the Semantic Web. *IEEE Intelligent Systems 19*(3), 41-47.

Staab, S., Domingos, P., Mika, P., Golbeck, J., Ding, L., Finin, T. W., Joshi, A., Nowak, A., & Vallacher, R. R. (2005). Social Networks Applied. *IEEE Intelligent Systems, 20*(1). 80-93.

Tho, Q. T., Hui, S. C., Fong, A. (2003). Web Mining for Identifying Research Trends. Paper presented at the 6th International Conference on Asian Digital Libraries. Kuala Lumpur, Malasya.

Velardi, P., Cucchiarelli, A., & Petit, M. (2007). A Taxonomy Learning Method and its Application to Characterize a Scientific Web Community. *IEEE Transactions on Knowledge and Data Engineering 19*(2). 180-191.

9. ADDITIONAL READING

Barabási, A.-L. (2002). Linked - The New Science of Networks. Perseus Publishing, Cambridge

Berkowitz, S.D. (1982). An Introduction to Structural Analysis: The Network Approach to Social Research. Butterworth, Toronto

Bojārs, U., & Breslin J. G. (2007). ResumeRDF: Expressing skill information on the Semantic Web. Paper presented at the 1st International ExpertFinder Workshop. Berlin, Germany.

Cameron, D., Aleman-Meza, B., & Arpinar, I.B. (2007). Collecting Expertise of Researchers for Finding Relevant Experts in a Peer-Review Setting. Paper presented at the 1st International ExpertFinder Workshop. Berlin, Germany.

Griffiths, T. L., & Steyvers, M. (2004). Finding Scientific Topics, *PNAS 101*(supplement 1). 5228-5235.

Iofciu, T., Diederich, J., Dolog, P., & Balke W. T. (2007). ExpertFOAF recommends experts. Paper presented at the 1st International ExpertFinder Workshop. Berlin, Germany.

Kleinberg, J. (2000). The Small-World Phenomenon: An Algorithm Perspective. Paper presented at the Thirty-second Annual ACM Symposium on Theory of Computing. Portland, Oregon, USA.

Li, L., Alderson, D., Doyle, J. C., Willinger, W. (2005). Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. *Internet Mathematics, 2*(4). 431-523.

Liu, B., & Chin, C. W. (2002). Searching people on the Web according to their interests. Poster presented at the Eleventh International World Wide Web Conference. Honolulu, Hawaii, USA.

Liu, P., & Dew, P. (2004). Using Semantic Web Technologies to Improve Expertise Matching within Academia, Paper presented at the 2nd International Conference on Knowledge Management. Graz, Austria.

Mane, K. K., & Börner, K. (2004). Mapping Topics and Topic Bursts in PNAS. *PNAS 101*(supplement 1). 5287-5290.

Miki, T., Nomura, S., & Ishida, T. (2005). Semantic Web Link Analysis to Discover Social Relationships in Academic Communities. Paper presented at the 2005 Symposium on Applications and the Internet.

Mochol, M., Jentzsch, A., Wache, H. (2007). Suitable Employees Wanted? Find them with Semantic Techniques. Paper presented at the Workshop on Making Semantics Work For Business. Vienna, Austria.

Mockus, A., & Herbsleb J. A. (2002). Expertise Browser: A Quantative Approach to Indentifying Expertise. Paper presented at the International Conference on Software Engineering. Orlando, Florida.

Ren, J., & Taylor, R. N. (2007). Automatic and Versatile Publications Ranking for Research Institutions and Scholars. *Communications of the ACM, 50*(6). 81-85.

Rodriguez, M. A., & Bollen, J. (2005). An Algorithm to Determine Peer-Reviewers. (Technical Report LA-UR-06-2261). Los Alamos National Laboratory.

Shadbolt, N., Hall, W., Berners-Lee, T. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems, 21*(3). 96-101.

Smeaton, A. F., Keogh, G., Gurrin, C., McDonald, K., & Sødring, T. (2003). Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century? *ACM SIGIR Forum, 36*(2). 39-43.

Song, X., Tseng, B. L., Lin, C.-Y., & Sun, M.-T. (2005). ExpertiseNet: Relational and Evolutionary Expert Modeling. Paper presented at the Tenth International Conference on User Modeling. Edinburgh, Scotland.

Takeda, H., Matsuzuka, T., & Taniguchi, Y. (2000). Discovery of Shared Topics Networks among People - A Simple Approach to Find Community Knowledge from WWW Bookmarks. Paper presented at the Sixth Pacific Rim International Conference on Artificial Intelligence. Melbourne, Australia.

The Yahoo! Research Team. (2006). Content, Metadata, and Behavioral Information: Directions for Yahoo! Research. *IEEE Data Engineering Bulletin, 31*(4). 10-18.

Thomas, C., & Sheth, A. P. (2006). On the Expressiveness of the Languages for the Semantic Web - Making a Case for 'A Little More'. In E. Sanchez (Editor), *Fuzzy Logic and the Semantic Web*. Elsevier.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. Cambridge University Press., Cambridge.

Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise Networks in Online Communities: Structure and Algorithms. Paper presented at the 16th International World Wide Web Conference. Banff, Canada.

Zhou, D., Ji, X., Zha, H., & Giles, C. L. (2006). Topic Evolution and Social Interactions: How Authors Effect Research. Paper presented at the 15th ACM International Conference on Information and Knowledge Management. Arlington, Virginia, USA.

Zhou, D., Manavoglu, E., Li, J., Giles, C. L., & Zha, H. (2006). Probabilistic Models for Discovering E-Communities. Paper presented at the 15th International World Wide Web Conference. Edinburgh, Scotland.

10. QUESTIONS FOR DISCUSSION

Beginner.

Give examples of keywords that (i) are strong indication that a paper is related to a very specific topic; (ii) are indication of a broad topic; and (ii) are not sufficient to determine that a paper is related to a topic.

Answer: (i) The keyword *PageRank* is strong indicator that a paper is related to the topic Search or Link Analysis. The keyword *XQuery* is a strong indicator that a paper is related to the topic XML.

(iii) The keywords *Data Mining*, *Semantic Web*, and *Databases* are examples that indicate that a paper is related to those (general) topics.

(iii) The keywords *algorithms*, and *evaluation* are not sufficient to determine that a paper is related to a topic.

Intermediate.

Provide an example of ten topics in computer science organized with ‘sub-topic’ relationships and including a synonym for five of the topic. Collect the answer from few participants and collaboratively try to ‘merge’ the topics of all participants. Discuss your findings.

Answer: Web Services (synonym: Web Service) with subtopic Semantic Web Services (synonym: Semantic Web Service). Web Service Composition (synonym: Service Composition), which is subtopic of Web Services. Semantic Web with subtopic Semantic Web Services. Semantic Search, which is subtopic of Semantic Web.

Search (synonym: Web Search) with subtopic Semantic Search. Intranet Search, which is subtopic of Search. Personalized Search, which is subtopic of Search. Discovery of Web Services (synonym: Web Services Discovery), which is subtopic of Web Services. Ontology Learning subtopic of Semantic Web.

Advanced.

Explain the differences of using a taxonomy of topics for finding experts on a given topic versus not using a taxonomy of topics.

Answer. There is a case where there is not difference on using or not a taxonomy of topics for finding experts. This is the case where a topic is a leaf node in the taxonomy. However, depending on the depth of the hierarchy, finding experts on a given topic will very likely produce better results when the taxonomy is used. This is because it is expected that the subtopics of the input topic will be included when finding matches for experts. For example, if the topic of interest is *Web Search* and the taxonomy contains the topic *Link Analysis* as a subtopic of *Web Search*, then the papers that are related to *Link Analysis* would be considered as a match for the input topic. However, it could be possible that most papers in a topic might always include the general topic in addition to more specific subtopics. In such cases, retrieval of experts using the general topic might not change much with or without the use of a taxonomy. The other side of the coin is that retrieval of experts on a more specific topic would indeed bring good results regardless of whether the general topic is related to the papers. However, the more specific the topic is, the closer it would be up to reaching leaves nodes. As before mentioned, this case would be a match in exactly same way whether the taxonomy is used or not. In summary, there are benefits when a taxonomy of topics is used yet these benefits are not that evident when the topics are near or at the leaf level of the taxonomy.

Practical Exercises.

Select one paper from the ACM Digital Library, one from the IEEE Digital Library and one from Springer. Make sure that all papers contain a list of keywords. Then, for every co-author in all papers, pick (if available) one paper for each and retrieve the keywords. Create a list of all keywords across all such papers, that is, three lists in total. Reorganize the lists by relating them to the authors whose papers are related to the keywords. If possible, determine the top experts on the keywords based on how many keywords are related to each author.

For every pair of collaborating researchers listed in Table 3, find out whether they collaborate often because they work at the same organization, or because they were in a advisor/advisee relationship and continue collaborating, or whether they are at different organizations yet collaborate frequently.