Mining Scholarly Semantic Networks from the Web

Mizuki Oka, Yutaka Matsuo
The University of Tokyo, Japan
mizuki.oka@gmail.com, matsuo@biz-model.t.u-tokyo.ac.jp

Abstract

With the increased usage of the Web and its availability of data, various scholarly information is now available on the Web. Extraction, aggregation, and visualization of such information is crucial for using our collective scholarly knowledge and expertise. Semantic network technology is prominent in structuring such knowledge. Our research project is designed to construct a scholarly semantic network from publicly available data that will be useful for researchers, business people in industry, and public servants in governmental organizations who are expected to make decisions appropriately. In this paper, we propose a practical architecture to construct a scholarly semantic network that integrates different scholarly entities such as researchers, papers, and keywords by taking into consideration ontological and web mining perspectives. We also provide an overview of our social network extraction system, which is used as an underpinning to build the architecture. The system, called POLYPHONET*, employs several advanced web mining and semantic technologies to extract relations of researchers, to detect groups of researchers, and to obtain keywords for a researcher. The public installation of the system at several academic conferences provides evidence of the system's usability and potential to facilitate the discovery of scholarly knowledge.*

Keywords— **scholarly semantic network, web mining, semantic technology, social network**

1 Introduction

In this era of information explosion, advanced usage of information related to the web is vital. To collect, store, and integrate information related to the web, semantic technology, as well as web mining technology, is necessary. By integrating a vast amount of information semantically, we can build an advanced type of knowledge base by web mining [18].

Our research, in particular, is intended to examine the potential for structuring scholarly knowledge through constructing semantic networks using web mining and semantic technologies [3]. We would like to provide an overview of the entire research domain, the trend of the domain, and also a detailed view of each research area. Social network analysis on researchers, conferences, and journals serves an important role to provide a macroscopic view of scholarly knowledge. More detailed text analysis and temporal data analysis also play an important role in providing a microscopic view. Visualization of the mined scholarly knowledge is another essential factor. Because the data must be processed and because their results might have many attribute types, not to mention that they might be large, it would be difficult to gain a comprehensive understanding at a glance. Intuitive visualization and easy navigation complements provision of a global view of our collective knowledge and wisdom.

Semantic network extraction and analysis have recently attracted the attention of many researchers in various research areas such as Semantic Web [1], social network analysis [23], and information visualization. A semantic network is a collection of entities that are linked by a set of relations [24]. Vastly numerous studies have been conducted using social network analysis [23]. Constructing semantic networks provides us with useful information such as calculating trustworthiness of a person [7] and detecting relevance and relations among different people [15]. In parallel efforts with semantic network extraction and analysis, visualization of social networks has been studied [5]. Studies have provided examples of the ways in which spatial position, color, size, and shape can all be used to encode information. With the widespread use of the internet, recent studies have applied such approaches to visualize and analyze domains such as e-mail communication [4] and co-authorship networks in scientific publications [16].

A scholarly semantic network can be constructed by recognizing relations among scholarly entities (e.g., papers, authors, keywords). Usually, classic databases of papers such as MEDLINE [1], PubMed [2], DBLP [3], and CiteSeer [4] merely enable access of scholarly knowledge through the Web. If we want to make greater use of web information such as web pages on new projects and new funds, press releases, user-generated knowledge on blogs

[1] http://medline.cos.com/.
[2] http://www.ncbi.nlm.nih.gov/PubMed/.
[3] http://www.informatik.uni-trier.de/ ley/db/.
[4] http://citeseer.ist.psu.edu/.

and Wikipedia [5], each entity and relation must be defined in a more formal way. Usage of the web enables us to aggregate such various kinds of entities and relationships, but it might also require visualization techniques that are suitable for exploration. Once the scholarly semantic network is obtained and visualized accordingly, it becomes useful in various ways by facilitating location of experts and authorities, and detection of topics among research papers and conferences.

In this paper, we provide an overview of our research project related to web mining: especially extraction, aggregation, and visualization for social network mining. We then describe the architecture used for structuring scholarly knowledge by web mining to improve the decision-making of researchers and other people who are interested in research, e.g. funding organizations. The contribution of our research is summarized as follows.

- We present an overview of our five-year project undertaken during 2003–2007 at JSAI [6] academic conferences on web mining for scholarly knowledge structuring.
- We propose a practical architecture to construct a scholarly semantic network that integrates different scholarly entities by taking into consideration ontological and web mining perspectives.
- Our system might be used for locating experts and authorities, detecting topics among research papers, detecting relations among conferences and topics, and providing an overview and trends of the research domain.

The remainder of the paper is organized as follows. The following section briefly provides an overview of our social network mining project with the technical description of social network extraction and the implementation of interactive system called POLYPHONET. Section 3 presents an architecture for structuring scholarly knowledge as an expansion of the project. Section 4 describes related work and Section 5 concludes this paper.

2 Social Network Mining among Researchers

As an underpinning towards the realization of scholarly knowledge semantic network, we have developed social network mining techniques for researchers using the web [13, 25, 14]. It can recognize various kinds of collaborative relations among researchers. The *co-occurrence* of researcher's names on the Web is used as evidence of two researchers' collaboration using a general search engine. Text analysis is further conducted to verify the relation.

[5] http://www.wikipedia.org/ .

[6] http://www.ai-gakkai.or.jp/jsai/english.html.

Table 1: Error rate of relation types, precision and recall.

class	error rate	precision	recall
Co-author	4.1%	91.8% (90/98)	97.8% (90/92)
Lab	25.7%	70.9% (73/103)	86.9% (73/84)
Proj	5.8%	74.4% (67/90)	91.8% (67/73)
Conf	11.2%	89.7% (87/97)	67.4% (87/129)

The system called POLYPHONET has been used at several conferences and workshops in Japan to provide the overview of a research domain and to facilitate collaboration among researchers. The public installation of the system has provided evidence of the system's usability and potential for scholarly knowledge discovery. Because of space limitations, we provide only a brief overview of the project. We encourage the reader to visit the website for UbiComp2005 [7] and for JSAI2005 [8]. Please refer to [14] for additional technical details.

2.1 Methods to Extract Social Network from the Web

A social network is extracted through two steps. First we set nodes; then we add edges using a search engine. For example, assume that we intend to measure the strength of relations between two names, Yutaka Matsuo and Mizuki Oka, using a search engine. The number of hits estimates the strength of their relation by co-occurrence of their two names. We add an edge between the two corresponding nodes if the strength of relations is greater than a certain threshold.

Several indices can measure co-occurrence [12]: matching coefficient, mutual information, Dice coefficient, Jaccard coefficient, overlap coefficient, and cosine. Depending on the co-occurrence measure that is used, the resultant social network varies. Through comparison of the indices with a co-authorship relation, among them we conclude that the overlap coefficient is best for our purposes [13, 25].

Not only the strength of the tie, but also the type of relation is detected in POLYPHONET. Inference of the class of relationship is thereby reduced to a text categorization problem that can be addressed using a machine-learning approach. We first fetch the top several pages retrieved by the "X and Y" queries. Then we extract features from the contents of each page to classify pages into classes of relations. Especially, relations of four kinds are selected: Co-author, Lab (members of the same laboratory or research institute), Proj (members of the same project or committee), and Conf (participants in the same conference or workshop). Table 1 shows error rates of five-fold cross validation. Although the error rate for Lab is high, others have about a 10% error rate or less. Precision and recall are measured by manually labeling 200 additional Web pages.

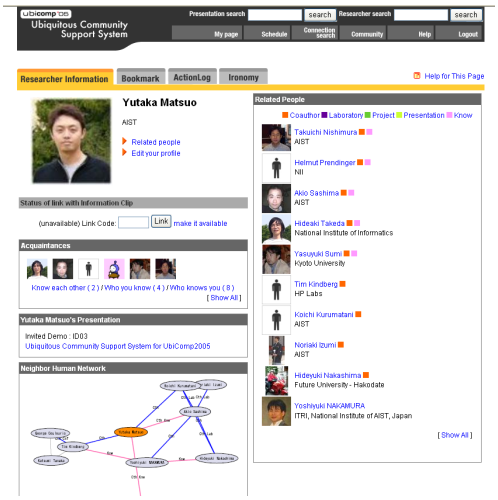

Figure 1: My page on POLYPHONET.

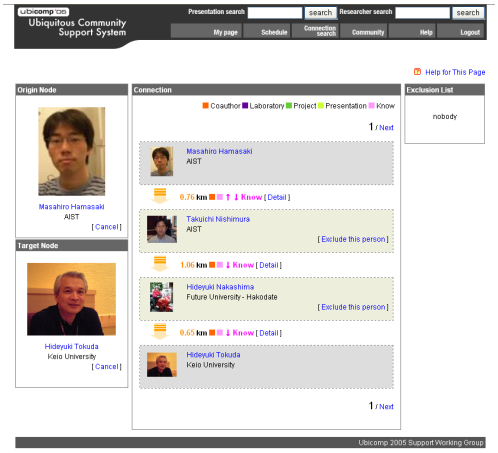

Figure 2: Shortest path from a person to a person on POLYPHONET.

2.2 POLYPHONET: Conference Support System

POLYPHONET [9] is a Web-based system for an academic community to facilitate communication and mutual understanding based on a social network extracted from the Web. We implement every module described above in POLYPHONET. The system has been used at JSAI[10] annual conferences successively for three years and at UbiComp2005[11] .

A social network of participants is displayed in POLYPHONET to illustrate a community overview. Retrieval tasks of various types are possible on the social network: researchers can be sought by name, affiliation, keyword, and research field; researchers related to a retrieved researcher are listed; and a search for the shortest path between two researchers can be made. Even more complicated retrievals are possible: e.g., a search for a researcher who is nearest to a user in the social network among researchers in a certain field. POLYPHONET is incorporated with a scheduling support system [8] and a location information display system [17] in the ubiquitous computing environment at the conference sites.

Figure 1 is a portal page that is tailored to an individual user, called *my page*. The user's presentations, bookmarks of the presentations, and registered acquaintances are shown along with the social network extracted from the Web. Figure 2 shows the shortest obtained path between two persons on a social network. Figure 3 portrays a screenshot that portrays at situation with three persons at an information kiosk; the social network including the three is displayed. More than 200 users used the system during each three-day conference, as shown in Table 2. Although POLYPHONET does not offer elaborate visualization of the extracted network and the interaction with users remains rather simple, comments from users were almost entirely positive; they enjoyed using the system. POLYPHONET presents evidence of potentials to facilitate scholarly knowledge discovery and provides an area for future work in exploring different kinds of visualization and their interaction with users.

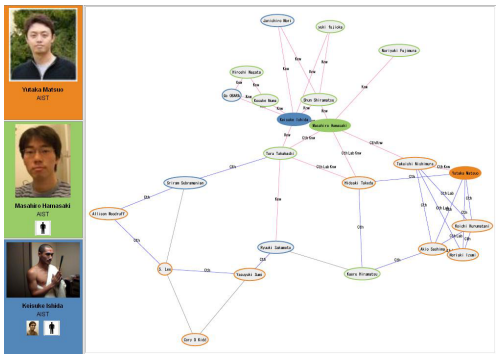

Figure 3: Social network among three persons on POLYPHONET.

Table 2: Numbers of participants at conferences.

	JSAI03	JSAI04	JSAI05	UbiComp05
#participants	558	639	about 600	about 500
#users	276	257	217	308

3 Scholarly Knowledge Mining

In this section, we describe an architecture to construct a scholarly semantic network from an ontological and web mining perspective as an extension of POLYPHONET. We consider not only researchers but also other scholarly entities and their relations based on web mining and semantic

[7] http://www.ubicomp-support.org/ubicomp2005/
[8] http://jsai-support-wg.org/polysuke2005/
[9] Polyphonet is a term that was coined using *polyphony* + *network*.
[10] http://jsai-support-wg.org/polysuke2005/.
[11] http://www.ubicomp-support.org/ubicomp2005/.

technology.

3.1 Architecture of Scholarly Web Mining

The web mining techniques we explained in the previous section mainly examine researchers specifically. Although a researcher plays an indispensable role in the research domain, entities of other kinds are also important: papers, journals, conferences, research topics, and so on.

Although a very complicated system of interaction might pertain among such entities [21], we use a rather simple and practical architecture to represent the scholarly domain to fulfill the demand of system usage and the feasibility of web mining. In our model, the three most important entities are researchers, papers, and keywords. Papers are the main outcome of research activities Keywords are the most commonly used information to represent the contents of a paper. As a superclass of the three entities, we can define three other entities: organization, journal/conference, and research topics.

The six important entities we define in our research are as follows:

- Researcher: a subject who tries to conduct research activity
- Paper: an outcome of the research activity by researcher(s)
- Keyword: a representation of the contents of a paper
- Organization/Project (as a superclass of researchers): an organized subject which conduct research activities, which can be considered as a set of researchers.
- Journal/Conference (as a superclass of papers): an organized collection of papers as the outcome of broad research activities, which can be considered as a set of papers.
- Research topic (as a superclass of keywords): a set of keywords to represent what the paper is about, which is typically preprocessed using text processing.

Although the above categorization of entities might not be ontologically precise, in terms of web mining, we use this simple view to mine important information efficiently and to facilitate the integration of information.

By mining and recognizing the six entities and the relations among the entities, we can provide structured knowledge of various kinds. For example, the relevance of the research topic is measurable in relation to (i) keyword overlap, (ii) paper/conference overlap, and (iii) researcher/organization overlap. In other words, the two topics "stem cell" and "ontology" might be measured according to (i) how many relevant keywords the two topics share, (ii) how many papers or conferences contain these two topics, and (iii) how many researchers or organizations conduct research on both topics. This research topic map (or ontology of research topics) is useful for providing an overview of an interesting research topic, and even for providing the changing of trends of various research topics over time. This can assist, for instance, research funding agencies and others to make decisions appropriately.

Identifying relations among researchers and organizations or journals and conferences can provide us various useful information. The former relation is useful, for instance, for recruiting researchers, for locating researchers for collaboration, and for finding paper reviewers [22]. The latter is useful to provide conference maps to help researchers find where to submit papers and to assess the quality of conferences based on the key organizers of the conference [26].

3.2 Methods to Recognize Entities and Relations

To recognize three (or six) types of entities and relations among them, several approaches using web mining are possible. For our previous studies, we extracted relations among researchers and their relations to keywords. With the usage of the co-occurrence based approach and web mining technology in our studies, we can extract relations of various kinds among these entities from digital libraries as well as other information that is available on the Web.

3.2.1 Usage of Digital Library Database

Existing digital library databases such as CiteSeer and DBLP provide us relational information among researchers, papers, and conferences/journals. Extraction of relations from such information corresponds to conventional bibliography studies, which require no special web mining technology. Readily apparent information such as citations, co-authorships, and conferences and journals directly enable us to extract relations among these entities and run semantic analyses of the obtained network.

3.2.2 Usage of the Web

We can further consider potential extraction of relations of other kinds among entities from the Web; for example, co-reference of two papers on the same *Web* page (not on the same paper). These two papers might be two "to read" papers for a class in the university or for individuals, which implies that both papers similarly have good quality and are mutually relevant. This type of information is obtainable from collaborative bookmarking sites such as the del.icio.us [12] and the CiteULike [13].

The recognition of keywords and relation of keywords can be enhanced through several web mining techniques as

[12] http://del.icio.us/.

[13] http://www.citeulike.org/.

well. For example, we can measure the popularity of a keyword by simply using the web count produced by a general search engine. The usage of web hit counts to measure the relevance of two given words has been a common approach in web mining and natural language processing recently.

Another example of the usage of the Web is to recognize relations of terms (especially, named entities) using Wikipedia. Although the interest of researchers and that of common web users are not exactly the same, the recent growth and popularity of Wikipedia enables us to recognize the relation of two keywords that seem at first to be unrelated. The textual information, category information, and link information all contribute the recognition of two keywords.

3.3 Integration and Visualization of Scholarly Networks

Given a set of networks of relation among scholarly entities, how to integrate and visualize such various networks to obtain an overview is a challenging task. One way to integrate a set of networks is to process them one by one in a pipeline manner. Let us consider an example in which our goal is to extract the relation of authoritative papers with their related keywords given the networks of paper-paper relation, paper-keyword relation, and keyword-keyword relation. We can perform analyses of all three networks and detect authoritative entities on each network independently using a metric such as PageRank. We can then integrate them by connecting all the detected authoritative entities.

Although such a pipeline approach provides a means to integrate networks, it might cause errors to compound and allow decisions to be inconsistent because relation formation decisions are made independently from each other. Our project therefore seeks to provide a method that integrates such networks more seamlessly and that also provides information that is suitable for hierarchical visualization.

One means to offer such a seamlessly integrated network is to use a method called the Eigen Co-occurrence Matrix (ECM) method we proposed previously [20, 19]. The ECM method represents all relations among entities in a single matrix. In the example described above, the matrix will contain the relations of paper-paper, paper-keyword, and keyword-keyword in a single matrix. Through the application of principal component analysis (PCA) [10] to the matrix, it produces a network with principal features. Depending on the number of eigenvectors used to construct the network as a result of PCA, the resulting network can be represented in various ways from a microscopic view (with the usage of a larger number of eigenvectors) to a macroscopic view (with the usage of fewer eigenvectors). In other words, the ECM method enables us to integrate various networks seamlessly and enables us to inspect the results in a hierarchical manner. We plan to explore the potential of the method further through analyses of the actual data mined from the web.

4 Related Work

With the availability of digital libraries, numerous studies have been conducted to measure the quality and impact of scholarly entities such as authors, papers, and publication venues. Many such studies use citation-based metrics for ranking these entities from a digital library. A citation index enables the user to discover where and how often a particular article is cited in the literature, thereby providing an indication of the importance of the article (e.g. an impact factor). In the late 1990s, CiteSeer was introduced as an automatic citation indexing system. More recently, Google Scholar has become popular, which indexes not only papers and articles from various scholarly organizations, but also those available on the Web. Other services such as the DBLP, Thomson Scientific [14], and the ACM Digital Library [15] provide information related to coauthorships: listings of all papers by author, co-author, and co-author links are available for access.

The most widely adopted method to measure the quality of entities is to use Garfield's impact factor (IF) [6] which counts the average number of times the published papers are cited up to two years after publication. However, the IF has been criticized for its sole dependency on citation counts; many alternatives such as PageRank have been used to rank journals [2]. Our project is intended to benefit from the availability of the data from the Web and aggregated data not only from the digital libraries but also data mined from other sources such as blogs and Wikipedia.

The earliest work to benefit from the development of the Web dates back to work in the mid-1990s by Kauz and Selman. They developed a social network extraction system from the Web, called Referral Web [11]. Nevertheless, we share the basic concept with their work on the usage of the Web, the rapid development of the WWW, and the potential of semantic network technology. Our automatic extraction of relations among scholarly entities has much greater potential and demand now than when Referral Web was first developed. A recent work by Zhuang et al. underscores an instance of this development. They extract researchers' relations by mining the characteristics of Program committee members extracted from Call for Papers available on the Web to measure the quality of publication venues (e.g., conferences) [26]. Their method enables differentiation of conferences that have greater impacts from the remainder.

In terms of visualization of social networks, several social visualization projects have been proposed [9]. A no-

[14] http://scientific.thomson.com/

[15] http://portal.acm.org/

table work was reported by Heer and Boyd, who developed a system to visualize social network information collected from Friendster [16], a popular online social network service (SNS). Compared to studies of mining and visualizing network from the Web, these works are attempts to visualize social networks that are constructed by users collected at service providing sites. A system that visualizes social networks that have been mined automatically from the Web was proposed by P. Mika called Flink [15]. Flink extracts, aggregates and visualizes online social networks for a Semantic Web community. Social networks are obtained using analyses of Web pages, e-mail messages, and publications and self-created profiles (FOAF files). Our project therefore seeks how best to benefit from a long history of social network visualization as well as information visualization in general to provide a rich user interaction and exploration of the collected knowledge from the Web.

5 CONCLUSIONS

This paper provides an overview of our social network extraction system called POLYPHONET and presents an architecture for structuring scholarly knowledge using web mining and semantic technology. Usage of web mining technology enables the detection of relations among scholarly entities not only from conventionally available databases of papers but also from other information such as user-generated knowledge related to blogs and Wikipedia.

Using our interdisciplinary approach, which enhances the usage of web mining and semantic technology over various scholarly entities, we hope to contribute not only to methods of semantic network extraction and analysis but also to their visualization. We believe that structured scholarly knowledge can be useful for researchers, business people in industry, and public servants in governmental organization who must make appropriate decisions.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web, 2001. Scientific American.

[2] J. Bollen, M. A. Rodriguez, and H. V. de Sompel. Journal Status. *Scientometrics*, 69(3):669–687, 2006.

[3] P. J. Carrington, J. Scott, and S. Wasserman. *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005. ISBN: 0521600979.

[4] D. Fisher and P. Dourish. Social and temporal structures in everyday collaboration. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 551–558, New York, NY, USA, 2004. ACM.

[5] L. C. Freeman. Visualizing Social Networks. *Journal of Social Structure*, 1(1), 2000.

[6] E. Garfield. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, 122(3159):108–111, 1955.

[7] J. Golbeck and B. Parsia. Trust network-based filtering of aggregated claims. *International Journal of Metadata, Semantics and Ontologies*, 1(1):58–65, 2006.

[16] http://www.friendster.com

[8] M. Hamasaki, H. Takeda, I. Omukai, and R. Ichise. Scheduling Support System for Academic Conferences Based on Interpersonal Networks. In *Demonstration and Poster. Proceedings of Hypertext*, pages 50–51, 2004.

[9] J. Heer and D. Boyd. Vizster: Visualizing Online Social Networks. In *INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, page 5, Washington, DC, USA, 2005. IEEE Computer Society.

[10] I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, 2 edition, 2002. ISBN: 0387954422.

[11] H. Kautz, B. Selman, and M. Shah. Referral Web: combining social networks and collaborative filtering. *Commun. ACM*, 40(3):63–65, 1997.

[12] C. D. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[13] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. Finding social network for trust calcularion. In *ECAI 04: Proceedings of the 18th European Conference on Artificial Intelligence*, pages 510–514, 2004.

[14] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. POLYPHONET: An Advanced Social Network Extraction System from the Web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 397–406. ACM Press, 2006.

[15] P. Mika. Flink: Semantic Web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):211–223, 2005.

[16] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 5200–5205, 2004.

[17] T. Nishimura, Y. Nakamura, H. Itoh, and H. Nakashima. System Design of Event Space Information Support Utilizing CoBITs. In *ICDCSW'04: Proceedings of the 24th International Conference on Distributed Computing System Workshops*, pages 384–387, 2004.

[18] N. F. Noy. Semantic integration: a survey of ontology-based approaches. *SIGMOD Rec.*, 33(4):65–70, 2004.

[19] M. Oka, T. Koiso, and K. Kato. Extracting Features of Patients using the Eigen Co-occurrence Matrix Algorithm. In *Proceedings of the the 15th ECML/PKDD Workshop on Discovery Challenge*, pages 86–97, 2004.

[20] M. Oka, Y. Oyama, H. Abe, and K. Kato. Anomaly Detection Using Layered Networks Based on Eigen Co-occurrence Matrix. In *Proceedings of the Seventh International Symposium on Recent Advances in Intrusion Detection (RAID)*, pages 223–237, 2004.

[21] M. A. Rodriguez, J. Bollen, and H. V. de Sompel. A practical ontology for the large-scale modeling of scholarly artifacts and their usage. In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pages 278–287, New York, NY, USA, 2007. ACM.

[22] M. A. Rodringuez and J. Bollen. An Algorithm to Determine Peer-Reviewers. *ArXiv Computer Science e-prints*, May 2006.

[23] J. P. Scott. *Social Network Analysis: A Handbook*. Sage Publications Ltd, 2 edition, 2000. ISBN: 0761963391.

[24] M. Steyvers and J. B. Tenenbaum. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29(1):41–78, 2005.

[25] Y. M. H. Tomobe, K. Hashida, H. Nakajima, and M. Ishizuka. Social network extraction from the web information. *Journal of the Japanese Society for Artificial Intelligence*, 20(1E):46–56, 2005.

[26] Z. Zhuang, E. Elmacioglu, D. Lee, and C. L. Giles. Measuring conference quality by mining program committee characteristics. In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pages 225–234, New York, NY, USA, 2007. ACM.