Research-Insight: Providing Insight on Research by Publication Network Analysis

Fangbo Tao, Xiao Yu, Kin Hou Lei, George Brova, Xiao Cheng
Jiawei Han, Rucha Kanade, Yizhou Sun, Chi Wang, Lidan Wang, Tim Weninger
Dept. of Computer Science, University of Illinois at Urbana-Champaign
{ftao2, xiaoyu1, klei2, brova2, cheng88, hanj, kanade2, sun22, chiwang1, lidan, weninge1}@illinois.edu

ABSTRACT

A database contains rich, inter-related, multi-typed data and information, forming one or a set of gigantic, interconnected, heterogeneous information networks. Much knowledge can be derived from such information networks if we systematically develop an effective and scalable database-oriented information network analysis technology. In this system demo, we take a computer science research publication network as an example, which is an information network derived from an integration of DBLP, other web-based information about researchers, and partially available citation data, and construct a Research-Insight system in order to demonstrate the power of database-oriented information network analysis. We show that nontrivial research insight can be obtained from such analysis, including (1) ranking, clustering, classification and similarity search of researchers, terms and venues for research subfields and themes, (2) recommending good researchers and good research papers to read or cite when conducting research on certain topics, (3) predicting potential collaborators for certain theme-oriented research, and (4) predicting advisor-advisee relationships and affiliation history based on historical research publications. Although some of these functions have been studied in recent research, effective and scalable realization of such functions in large networks still poses challenging research problems. Moreover, some function are our ongoing research tasks. By integrating these functionalities, Research-Insight may not only provide with us insightful recommendations in CS research but also help us gain insight on how to perform effective data mining in large databases.

Categories and Subject Descriptors

H.2.8 [**Information Systems Applications**]: Database Applications—*Data Mining*

Keywords

heterogeneous information network, recommendation system

SIGMOD'13, June 22–27, 2013, New York, New York, USA.

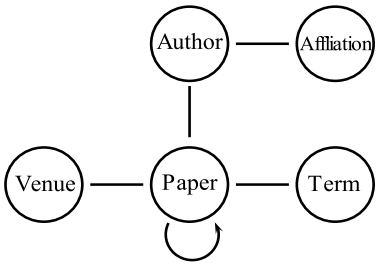

Figure 1: Academic Information Network Schema

1. INTRODUCTION

People usually treat a database as a data repository that stores large sets of data and supports indexing, retrieval, and various kinds of updating and complex query processing. However, entities/objects in databases are not isolated tuples; they contain rich, inter-related semantic information that can and should be systematically explored. Objects in databases are inter-related and linked (*e.g.*, via foreign keys, *etc.*) across multiple relations or entity sets, forming gigantic information networks. Information network analysis methods can be systematically developed for in-depth network-oriented data mining and analysis, which is far beyond the scope of traditional search functions provided in database systems.

To gain this insight, we propose the Research-Insight demonstration system that takes an information network and turns it into a useful information resource for effective clustering, classification, recommendation, and prediction by database-oriented information network analysis. We first briefly describe our data sets and outline the major functions being implemented in this demonstration system.

As a starting point for a general system, we begin with the well-known DBLP[1] dataset, which can be viewed as an information-rich computer science publication network. Recent studies show that such databases can be extended by mining the Web sites information within a dataset; in the computer science domain, the Web can be mined semi-automatically to recover and link affiliation information with author-records in DBLP [10]. Moreover, a good (though incomplete) set of citation information related to computer science publications is available from Arnetminer and Citeseer. By integrating such multiple sources of information, a Computer Science Research Information Network dataset (we call CSR-Net) can be constructed to facilitate a system-

[1]http://www.informatik.uni-trier.de/∼ley/db/

atic study of information network analysis methods by our Research-Insight system. In particular, the schema of this integrated information network consists of a set of interconnected node types, as shown in Figure 1. This particular schema implies that the network consists of a set of papers, which are linked by their authors, publication venues, and terms contained in the title. Our analysis functions illustrated as follows are built and tested on CSR-Net.

2. MAJOR FUNCTIONAL MODULES

Research-Insight consists of the following functional modules: (1) similarity search in heterogeneous networks; (2) ranking and clustering of research fields, authors, venues, and terms; (3) ranking and classification of research fields, authors, venues, and terms, (4) recommendation of research papers to read or to cite; (5) prediction of potential research collaborations; and (6) prediction of advisor-advisee relationships and historical affiliations. Here we introduce each of these functional modules with some illustrative examples.

2.1 Similarity Search

Similarity search often plays an important role in the analysis of networks. However, it is challenging to define a good measure of similarity between objects in a heterogeneous information network. By considering different linkage paths in a network, one can derive various semantics on similarity. A meta-path based similarity measure is introduced in [6], where a meta-path is a structural path defined at the meta level (*i.e.*, relationships among object types). A new similarity measure, PathSim [6], was introduced for finding peer objects in the network (*e.g.*, find authors sharing similar research fields and with similar reputation), which turns out to be more meaningful in many scenarios compared with random-walk based similarity measures and is also efficient for top-k similarity search in heterogeneous networks. In Research-Insight, we will demonstrate this similarity function for finding top-k similar nodes for a given (query) node, and compare the PathSim measure with other measures, such as SimRank and Personalized-PageRank [4].

Example 1: Similarity Search in CSR-Net. Given an author (*e.g.*, Johannes Gehrke), find his/her top-k similar authors and explain why (by showing the corresponding meta-path and similarity measure). Do the same for a venue (*e.g.*, EDBT), a term (*e.g.*, SVM), and a research paper. Potential extensions include finding top-k most related heterogeneous typed objects (*e.g.*, given an author (*e.g.*, Christos Faloutsos), find his/her top-k most related venues and terms). ■

2.2 Ranking-Based Clustering

Most methods perform clustering based on attribute values of the data. However, for link-based clustering of heterogeneous information networks, we need to explore links across heterogeneous types of data. Our recent studies develop a ranking-based clustering approach, represented by RankClus [7] and NetClus [8], that generates interesting results for both clustering and ranking efficiently. This approach is based on the observation that ranking and clustering can mutually enhance each other because objects highly ranked in each cluster may contribute more towards unambiguous clustering, and objects more dedicated to a cluster will be more likely to be highly ranked in the same cluster.

Example 2: Rank-based clustering in CSR-Net. Given a subnetwork (*e.g.*, network formed in the fields of DB, DM, IR and ML) and a desired number of clusters (*e.g.*, 4), perform rank-based clustering and show top-k objects of each type (*i.e.*, authors, venue, terms) in each cluster. Do the same for a more restricted network (*e.g.*, DB) to find its m rank-based multi-object-typed clusters for a given m. ■

2.3 Ranking-Based Classification

Classification can also take advantage of links in heterogeneous information networks. Knowledge can be effectively propagated across a heterogeneous network because the nodes that are close to similar objects via similar links are likely to be similar. Moreover, following the idea of ranking-based clustering, one can explore ranking-based classification since objects highly ranked in a class are likely to play a more important role in classification. These ideas lead to effective algorithms, such as GNetMine [2] and RankClass [1], for model construction in heterogeneous networks.

Example 3: Rank-based classification in CSR-Net. Given a subnetwork (*e.g.*, network formed in the fields of DB, DM, IR and ML) and a set of label data (*e.g.*, authors and papers labeled by their fields), perform rank-based classification and show classification accuracy and top-k objects of each type (*i.e.*, authors, venue, terms) in each class. Do the same for a more restricted network (*e.g.*, AI only) to find its rank-based classes. ■

2.4 Literature Recommendation

Literature search is an essential task in scientific research. Traditional key-word-based search systems, such as Google Scholar and PubMed, answer keyword queries by ranking relevant documents based on document similarity between queries and papers. However, it is often more desirable to recommend literature based not only on keyword relevance but also on the reputation of authors and venues.

Example 4: Literature Recommendation. Suppose a researcher would like to find research papers on frequent pattern mining methods. However, there are thousands of publications on this topic but many of them may not even contain "frequent," "pattern," and "mining" in the title. Moreover, it is desirable to rank the retrieved papers based not only on keyword relevance but also on the reputations of authors and venues. Although Google Scholar may use citation count to rank papers, it will be hard to recommend newly published papers that have collected relatively few citation counts so far. Research-Insight will support this literature recommendation function based on our ongoing research. ■

Literature recommendation can be realized by integration of keyword based search with rank-based clustering and classification mechanisms. One can first preprocess the literature data to find a set of term clusters that represent a research theme, find and rank authors, venues, papers that are reputed on the theme (*e.g.*, using RankClus/RankClass [4] and citation counts). Then we can make recommendations based on meta-path-based feature space proposed in [11], freshness (*e.g.*, based on publication time), topic closeness, and influence of the papers (*e.g.*, based on the citation count). We will further explore ranking-based clustering on different types of entities (*e.g.*, authors, papers, affiliations) in heterogeneous information networks and combine both document similarity and network structural similarity to improve the quality of literature recommendation.

2.5 Collaboration Prediction

Recommending potential co-authors for a given author will help researchers explore new collaborations. Many studies treat co-author prediction as a link prediction problem in a homogeneous co-author network and prediction is usually made based on the common co-authors for a given pair of authors. However, coauthor relationships are often developed based on other factors, such as shared research themes, common publication venues, common affiliation, and so on. It is important to analyze different factors and predict new coauthor relationships by heterogeneous information network modeling and analysis.

Our recent studies [3, 5] treat this problem as a problem of link and relationship prediction in heterogeneous information networks. To predict links or interactions between both homogeneous typed or heterogeneous typed objects (*e.g.*, predicting coauthor relationships or predicting whether an author will write papers on a theme) in a network, we model their interactions as a set of meta-paths in a heterogeneous information network. For example, in CSR-Net, there are multiple types of objects (e.g., venues, topics, papers, affiliations) and multiple types of links among these objects that may contribute to the co-author relation prediction. By systematically designing topological features and measures in the network, a supervised model can be used to learn the best weights associated with different topological features in deciding the co-author relationship, thus lead to high-quality coauthor relationship prediction [3]. Moreover, by incorporating data extracted with Web structure mining methods [10] and Web data from ACM Digital Library, author's affiliation information can be incorporated into this analysis (since affiliation can be an important factor that may influence authors' collaboration potential).

Example 6: Collaboration Prediction in CSR-Net. Given a research student, associated with his/her publication history and affiliation (*e.g.*, in the Database Group of Univ. of Michigan), one may get information about his likely advisor and group mates, as well as other professors and students in the same institution in a related discipline, as well as other researchers in the same field but different affiliations from CSR-Net. Such information will further help prediction of his/her top-k potential new research collaborators since it may use additional affiliation and research group information than those studied in [3]. Furthermore, for each recommended new potential co-author relationship, one can explain the reason why such a prediction is made, by displaying the weighted paths between the two potential collaborating researchers that are used for this prediction. ■

2.6 Prediction of Advisor-Advisee Relationship and Historical Affiliations

Researchers often collaborate on publications. In a research community, it is interesting to know at what period, whether two researchers ever have advisor-advisee relationship or whether they worked or are working in the same institution or in the same research group. Our study [9] shows that based on a set of training data and a small set of rules, such as "an advisor has more publications and a longer history than his/her advisee at the time of advising" and "once an advisee becomes advisor, s/he will not become advisee again", a time-constrained probabilistic factor graph can be constructed for quality advisor-advisee relationship prediction for the DBLP data. The extraction of

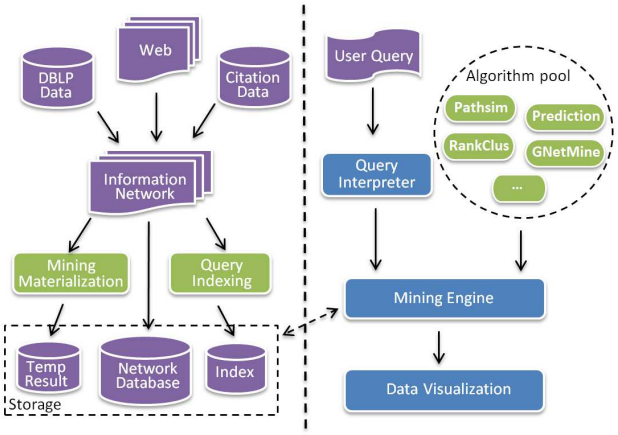

Figure 2: System Architecture of Research-Insight

partial affiliations from ACM Digital Library[2] and incorporation of Web structure mining methods [10] will further consolidate such discovered relationship. Notice that the Web data may miss most researcher's past affiliations, and our system will use collaboration and publication history to predict missing academic history based on his/her advisor's academic affiliation during the time s/he is being advised. Moreover, the prediction of advisor-advisee relationship between researchers in certain time period will help infer one's affiliation in that period since advisors are less likely change institutions than their advisees. Iterative constraint propagation will help uncover the hidden affiliation history for a good set of researchers. We will demonstrate the function of predicting advisor-advisee relationship and historical affiliations with CSR-Net.

Example 7: Prediction of Advisor-advisee relationships and Historical Affiliations. Given a researcher (*e.g.*, Jure Leskovec), our system will present his current institution (*e.g.*, Stanford, obtained from the Web), and the likely time (*e.g.*, since which year), as well as his/her historical affiliations(*e.g.*, in which year, as a Ph.D. student in which institution, following who as an advisee). Moreover, the system will report the associated reasoning paths and the likelihood of such a prediction to help readers understand the argument for such a prediction. ■

3. ABOUT THE SYSTEM AND THE DEMO

The Research-Insight system is designed with the architecture shown in Figure 2. It consists of the following modules: (1) generation of CSR-Net by integration of data from three input sources (*i.e.*, DBLP data, Web .edu data, and citation information data), (2) consolidation of CSR-Net, which generates consolidated CSR-Net by index construction, and pre-computation and materialization of some precomputed mining results, (3) mining query processing module, which processes user-queries by parsing the query, selecting and executing appropriate mining modules (*i.e.*, that mines on the consolidated CSR-Net to deriving mining results), and (4) result presentation by visualization and interpretation of the mining processes and results.

The Research-Insight system provides an easy-to-use web interface and beautiful data visualizations. We aggregate

[2]http://dl.acm.org/

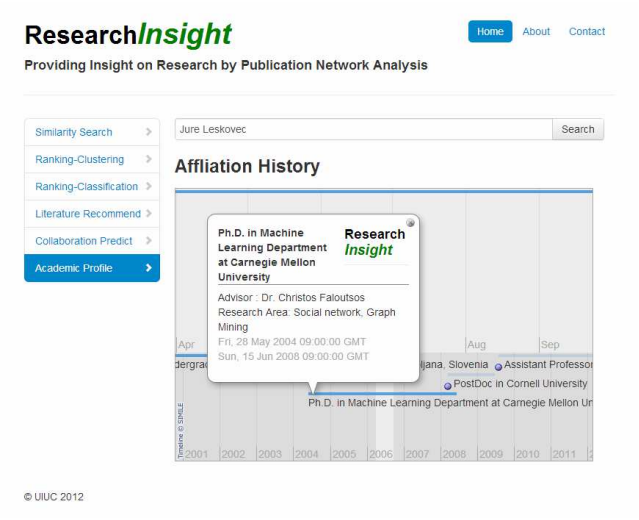

Figure 3: Illustration of advisor-advisee relationship and affiliation history on query: "Jure Leskovec"

all the functions proposed in Section 2 into two major web-based entry pages: (1) Author Search page, in which we show the affiliation history, collaboration prediction result and similar authors/conference results; and (2) Literature Search page, in which we show the literature recommendation results, clustering/classification result for literature keyword search. Figure 3 is a preliminary screen shot of an example result.

This proposed system is resulted from our multi-year research on mining heterogeneous information networks [4] and web structure mining [10], in which research publication networks are among the most popular examples and test data. Although many functional modules proposed here are studied in our recent publications, there are still many challenging problems to make such functions effective on various scenarios and scalable in large networks. Moreover, the proposed demo has introduced some new functionalities which need nontrivial extensions of our proposed methodology. We are conducting such research to ensure the proposed functionalities will be effectively realized, efficiently implemented, and being integrated into one coherent system.

The proposed demo is confined to computer science research publication networks. However, since most of the methods developed are not confined to computer science data sets, we expect the system should be extended and customized to multiple research publication networks, including ArXive, PubMed and other research publication networks. Since different scientific disciplines may have rather different culture, history, and research publication datasets, and researchers are likely have rather different search and mining needs, much research is still needed to extend such a Research-Insight system into other scientific and engineering disciplines.

A strong motivation of this study is to show that the massive data stored in databases are essentially information rich, heterogeneous information networks. Mining such data will generate useful knowledge and lead to deep insight on your data. We hope this demo may serve this purpose and promote the development of more powerful mining methods to tackle such big data.

4. REFERENCES

[1] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proc. 2011 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, San Diego, CA, Aug. 2011.

[2] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao. Graph regularized transductive classification on heterogeneous information networks. In *Proc. 2010 European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD'10)*, Barcelona, Spain, Sept. 2010.

[3] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proc. 2011 Int. Conf. Advances in Social Network Analysis and Mining (ASONAM'11)*, Kaohsiung, Taiwan, July 2011.

[4] Y. Sun and J. Han. *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers, 2012.

[5] Y. Sun, J. Han, C. C. Aggarwal, and N. Chawla. When will it happen? relationship prediction in heterogeneous information networks. In *Proc. 2012 ACM Int. Conf. on Web Search and Data Mining (WSDM'12)*, Seattle, WA, Feb. 2012.

[6] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proc. 2011 Int. Conf. Very Large Data Bases (VLDB'11)*, Seattle, WA, Aug. 2011.

[7] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In *Proc. 2009 Int. Conf. Extending Data Base Technology (EDBT'09)*, Saint-Petersburg, Russia, Mar. 2009.

[8] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proc. 2009 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, June 2009.

[9] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *Proc. 2010 ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'10)*, Washington D.C., July 2010.

[10] T. Weninger, F. Fumarola, C. Xide Lin, R. Barber, J. Han, and D. Malerba. Growing parallel paths for entity-page discovery. In *Proc. 2011 Int. World Wide Web Conf. (WWW'11)*, Hyderabad, India, Mar. 2011.

[11] X. Yu, Q. Gu, M. Zhou, and J. Han. Citation prediction in heterogeneous bibliographic networks. In *Proc. 2012 SIAM Int. Conf. on Data Mining (SDM'12)*, Anaheim, CA, April 2012.

Acknowledgments. Work supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA); NASA NNX08AC35A; U.S. NSF grants IIS-0905215, CNS-0931975, and IIS-1017362; U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265; and DHS-IDS MIAS Center at UIUC for Multimodal Information Access and Synthesis.