Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach*

Yizhou Sun† Jiawei Han† Xifeng Yan‡ Philip S. Yu§
† University of Illinois at Urbana-Champaign, Urbana, IL
‡ University of California at Santa Barbara, Santa Barbara, CA
§ University of Illinois at Chicago, Chicago, IL
†{sun22, hanj}@illinois.edu ‡xyan@cs.ucsb.edu §psyu@cs.uic.edu

ABSTRACT

Most objects and data in the real world are interconnected, forming complex, heterogeneous but often semi-structured information networks. However, most people consider a database merely as a data repository that supports data storage and retrieval rather than one or a set of heterogeneous information networks that contain rich, inter-related, multi-typed data and information. Most network science researchers only study homogeneous networks, without distinguishing the different types of objects and links in the networks. In this tutorial, we view database and other interconnected data as heterogeneous information networks, and study how to leverage the rich semantic meaning of types of objects and links in the networks. We systematically introduce the technologies that can effectively and efficiently mine useful knowledge from such information networks.

1. INTRODUCTION

People usually treat a database as a data repository that stores a large set of data and supports indexing, retrieval, updating and 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. One important fact that most previous research has not paid much attention is that objects in (relational) databases are inter-related and linked (*e.g.*, via foreign keys, *etc.*) across multiple relations, 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 and retrieval functions provided in database systems. Moreover, in such information networks, objects and links are from different types. By leveraging the rich semantic meanings of the typed objects and links, much more interesting knowledge can be mined than using homogeneous network analysis approaches.

*The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), NSF IIS-1017362, MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC, and U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

Proceedings of the VLDB Endowment, Vol. 5, No. 12

2. TUTORIAL OUTLINE

This tutorial presents a comprehensive overview of the techniques developed for database-oriented information network analysis in recent years. We briefly introduce the outline of the tutorial in this section.

2.1 Database as a Heterogeneous Information Network

We first introduce what is a heterogeneous information network and how to view databases as information networks, and point out the major principles of mining such networks.

2.2 Mining Heterogeneous Information Networks

Clustering, classification and ranking are basic mining functions for information networks. We introduce several studies that address these tasks in heterogeneous information networks by distinguishing different types of links.

Ranking-based clustering in heterogeneous information networks. For link-based clustering of heterogeneous information networks, we need to explore links across heterogeneous types of data. Recent studies develop a ranking-based clustering approach (*e.g.*, RankClus [8] and NetClus [10]) that generates both clustering and ranking results 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.

Classification of heterogeneous information networks. 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 linked together 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 [3] and RankClass [2].

2.3 Meta-Path-Based Similarity Search and Mining

We then introduce a systematic approach for dealing with general heterogeneous information networks with a specified network schema, by using meta-path-based methodologies. Under this

framework, similarity search and other mining tasks such as relationship prediction can be addressed.

Meta-path-based similarity search in heterogeneous information networks. Similarity search plays an important role in the analysis of networks. By considering different linkage paths (*i.e.*, meta-path) in a network, one can derive various semantics on similarity in a heterogeneous information network. A meta-path based similarity measure, PathSim, is introduced in [7], which aims at finding peer objects in the network. PathSim turns out to be more meaningful in many scenarios compared with random-walk based similarity measures.

Meta-path-based relationship prediction in heterogeneous information networks. Heterogeneous information network brings interactions among multiple types of objects and hence the possibility of predicting relationships across heterogeneous typed objects. By systematically designing meta-path-based topological features and measures in the network, supervised models can be used to learn the best weights associated with different topological features in deciding the relationship prediction building [5, 6].

2.4 Relation Strength-Aware Mining

The heterogeneity of relations between object types leads to different mining results that can be chosen by users. Moreover, the strength of each relation should be automatically learned and used for better mining.

Relation strength-aware clustering via attributes selection. By specifying a set of attributes, the strengths of different relations in heterogeneous information networks can be automatically learned to help the clustering of objects [4].

Integrating user-guided clustering with meta-path selection. Different meta-paths in a heterogeneous information network represent different relations with different semantic meanings. User guidance in the form of a small set of training examples for some object types can indicate their preference on the results of clustering. Then the preferred meta-path or weighted meta-path combinations can be learned to reach better consistency between mining results and the training examples [9].

2.5 Advanced Topics

Methods for mining heterogeneous information networks can often help data cleaning, data integration, trustworthiness analysis, and role discovery, which in turn help construction of high quality information networks. More advanced operators such as OLAP is also necessary for better exploring the networks.

Role discovery in information networks. An information network contains abundant knowledge about relationships among objects. Unfortunately, such knowledge, such as advisor-advisee relationships among researchers in a bibliographic network, is often hidden. [12] successfully mines advisor-advisee hidden roles in the DBLP database with high accuracy. Such mechanism can be further developed to discover hierarchical relationships among objects under different kinds of user-provided constraints or rules.

Trustworthiness analysis in information networks. A major challenge for data integration is to derive the most complete and accurate integrated records from diverse and sometimes conflicting sources. The *truth finding* problem is to decide which piece of information being merged is most likely to be true. By constructing an information network that links multiple information providers with multiple versions of the stated facts for each entity to be resolved, novel network analysis methods, such as TruthFinder [13] and LTM [14], can be developed to resolve the conflicting source problem effectively.

Online analytical processing of heterogeneous information networks. The power of online analytical processing (OLAP) has been shown in multidimensional data analysis. However, the extension of OLAP to analysis of heterogeneous information network is nontrivial. We will introduce some preliminary studies on this issue, such as [11, 1, 15].

2.6 Research frontiers

Viewing database as an information network and studying systematically the methods for mining database-oriented heterogeneous information networks is a promising frontier in database and data mining research. There are still many challenging research issues. Here we illustrate only a few.

- Discovery and mining of hidden information networks
- Diffusion analysis in heterogeneous information networks, and
- Ontology and structure discovery in heterogeneous information networks

3. REFERENCES

[1] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu. Graph OLAP: Towards online analytical processing on graphs. In *ICDM '08*, Pisa, Italy, Dec. 2008.

[2] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *KDD '11*, San Diego, CA, Aug. 2011.

[3] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao. Graph regularized transductive classification on heterogeneous information networks. In *ECMLPKDD '10*, Barcelona, Spain, Sept. 2010.

[4] Y. Sun, C. C. Aggarwal, and J. Han. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. In *VLDB '12/PVLDB*, Istanbul, Turkey, Aug. 2012.

[5] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *ASONAM '11*, Kaohsiung, Taiwan, July 2011.

[6] Y. Sun, J. Han, C. C. Aggarwal, and N. Chawla. When will it happen? relationship prediction in heterogeneous information networks. In *WSDM '12*, Seattle, WA, Feb. 2012.

[7] 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 *VLDB '11/PVLDB*, Seattle, WA, Aug. 2011.

[8] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In *EDBT '09*, Saint-Petersburg, Russia, Mar. 2009.

[9] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *KDD '12*, Beijing, China, Aug. 2012.

[10] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD '09*, Paris, France, June 2009.

[11] Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. In *SIGMOD '08*, pages 567–580, Vancouver, BC, Canada, June 2008.

[12] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *KDD '10*, Washington D.C., July 2010.

[13] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the Web. *IEEE Trans. Knowledge and Data Engineering*, 20:796–808, 2008.

[14] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conflicting sources for data integration. In *VLDB '12/PVLDB*, Istanbul, Turkey, Aug. 2012.

[15] P. Zhao, X. Li, D. Xin, and J. Han. Graph cube: On warehousing and OLAP multidimensional networks. In *SIGMOD '11*, Athens, Greece, June 2011.