Mining Knowledge from Data: An Information Network Analysis Approach

Jiawei Han †1, Yizhou Sun †2, Xifeng Yan *3, Philip S. Yu ‡4

† *Department of Computer Science, Univ. of Illinois at Urbana-Champaign, Illinois, USA*
1 hanj@illinois.edu, 2 sun22@illinois.edu

* *Department of Computer Science, Univ. of California at Santa Barbara, California, USA*
3 xyan@cs.ucsb.edu

‡ *Department of Computer Science, Univ. of of Illinois at Chicago, Illinois, USA*
4 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, many database researchers consider a *database* merely as a *data repository* that supports storage and retrieval rather than an information-rich, inter-related and multi-typed *information network* that supports comprehensive data analysis; whereas many network researchers focus on *homogeneous networks*. Departing from both, we view interconnected, semi-structured datasets as *heterogeneous, information-rich networks* and study how to uncover hidden knowledge in such networks. For example, a university database can be viewed as a heterogeneous information network, where objects of multiple types, such as students, professors, courses, departments, and multiple typed relationships, such as teach and advise are intertwined together, providing abundant information.**

In this tutorial, we present an organized picture on mining heterogeneous information networks and introduce a set of interesting, effective and scalable network mining methods. The topics to be covered include (i) database as an information network, (ii) mining information networks: clustering, classification, ranking, similarity search, and meta path-guided analysis, (iii) construction of quality, informative networks by data mining, (iv) trend and evolution analysis in heterogeneous information networks, and (v) research frontiers. We show that heterogeneous information networks are informative, and link analysis on such networks is powerful at uncovering critical knowledge hidden in large semi-structured datasets. Finally, we also present a few promising research directions.

I. Introduction

People usually treat a database as a data repository that stores a large set of data and facilitates indexing, updating, query processing, and transaction management. However, entities and objects in databases are not isolated records; they contain rich, inter-related semantic information that should be systematically explored. One important fact that most previous research has not paid enough attention is that objects in relational or semi-structured databases are inter-related and linked across multiple relations (*e.g.*, via foreign keys) or other structures, forming gigantic information networks. Information network analysis methods can be systematically developed for in-depth network-oriented data mining, which is far beyond the scope of traditional search and retrieval functions provided in database systems.

Example 1. Databases as information networks. In a bibliographic database, such as DBLP[1] and PubMed[2], papers are linked together via authors, venues and terms, and in a social media website, such as Flickr[3], photos are linked together via users, groups, tags and comments. A database therefore contains rich, inter-related, multi-typed data and information, forming a gigantic, interconnected, heterogeneous information network. Much knowledge can be mined from such a network by clustering, ranking, classification, role discovery, topic and ontology analysis, and so on. These new functions would be extremely useful considering the ubiquitous online databases in almost every industry. For example, clusters of research areas and ranks for authors and conferences can be discovered by such analysis in a bibliographic database, which will be very useful for better understanding and usage of the data stored in databases. ■

This tutorial presents a comprehensive overview of the techniques developed for database-oriented information network analysis in recent years. It will cover the following key issues.

- Database as an information network: A data analyst's view
- Mining information networks: Clustering, classification, ranking, similarity search, and meta path-guided analysis
- Construction of informative networks by data mining: Data cleaning, role discovery, trustworthiness analysis, and ontology discovery
- Evolution analysis, prediction, and diffusion analysis in heterogeneous information networks, and
- Research frontiers in database-oriented information network analysis.

This short paper is organized in a similar structure as the themes to be covered in the tutorial. Following a brief discussion on how a database can be viewed as a heterogeneous information network and why mining becomes interesting when it is so organized, Section 2 presents recent research

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

[2] http://www.ncbi.nlm.nih.gov/pubmed/

[3] http://www.flickr.com/

progress on mining heterogeneous information networks. Section 3 discusses how data mining methods may impact the construction of clean, intelligent, and informative networks. Section 4 presents methods for evolution analysis, prediction, and diffusion analysis in heterogeneous information networks. Finally, section 5 summarizes the results and points out some promising research frontiers.

II. Mining heterogeneous information networks

There are many studies on the analysis of homogeneous networks, such as network measures (*e.g.*, density, connectivity, centrality, *etc.* [1], [2]), statistical behavior study (*e.g.*, the small world phenomenon and the power-law distribution of degrees [3], [4]), and modeling of trend and dynamic evolution of networks [5], [4], [6], [7]. These themes will not be covered in depth since our focus is on mining heterogeneous information networks. Moreover, other recent work on homogeneous networks, such as clustering (*e.g.*, spectral clustering [8] and SCAN [9]), ranking (*e.g.*, PageRank [10] and HITS [11]), and similarity search (*e.g.*, SimRank [12] and Personalized PageRank [13]) will be introduced in our comparative study of heterogeneous networks.

Ranking-based clustering in heterogeneous information networks. 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. Besides several interesting studies on clustering heterogeneous networks (*e.g.*, spectral clustering [14], LinkClus [15] and CrossClus [16]), recent studies develop a ranking-based clustering approach (*e.g.*, RankClus [17] and NetClus [18]) 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.

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 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 [19] and RankClass [20], for model construction in heterogeneous networks.

Similarity search in heterogeneous information networks. 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, where a meta-path is a structural path defined at the meta level (*i.e.*, relationships among object types). A new similarity measure, PathSim [21], is 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.

Meta-path guided analysis in heterogeneous information networks. Since different meta-paths in a heterogeneous information network may represent different semantic meanings, meta-path plays an important role in the analysis of heterogeneous information networks. User guidance in the form of a small set of training examples on some types of data in a heterogeneous network may effectively communicate with a network miner on what should be the user preference on the results of mining. Then the most preferred meta-path or weighted meta-path combinations can be selected based on the interaction between the mining parameters and the provided training examples to reach better consistency between mining results and the training examples. The essential role of meta-path will be demonstrated with multiple tasks at mining heterogeneous networks [22].

III. Mining for effective network construction

Database data can be used for construction of heterogeneous information networks; however, for effective knowledge discovery it is important to enhance such data by various data mining methods. Interestingly, methods for mining heterogeneous information networks can often help data cleaning, data integration, trustworthiness analysis, role discovery, and ontology discovery, which in turn help construction of high quality information networks.

Data cleaning in information networks. Before a database-derived information network can be used to mine interesting knowledge, data cleaning should first be applied on such data, since noise and inconsistency may exist in these real databases. For example, in bibliographic networks like DBLP, there are ambiguity and synonym problems for authors. Different authors may carry the same name whereas the same author may be presented with different names. Entity resolution methods have been developed by exploring data semantics. A novel algorithm, Distinct [23], has been developed to distinguish objects with identical names by link analysis.

Role discovery in information networks. Information network contains abundant knowledge about relationships among people or entities. Unfortunately, such knowledge, such as advisor-advisee relationships among researchers in a bibliographic network, is often hidden. Role discovery is to uncover such hidden relationships by information network analysis. For example, a time-constrained probabilistic factor graph model, which takes a research publication network as input and models the advisor-advisee relationship mining problem using a jointly likelihood objective function has been developed [24]. It 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, hence enhance information network construction.

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 [23] and LTM [25], can be developed to resolve the conflicting source problem effectively. Truth-finding will help data cleaning and data integration, hence improve the quality of information networks.

Ontology and structure discovery in heterogeneous information networks. Interconnected, multiple typed objects in a heterogeneous information network often provide critical information for generating high quality, fine-level concept hierarchies. For example, it is often difficult to identify researchers just based on their research collaboration networks. However, putting them in a network that links their publication, conferences, terms and research papers, their roles in the network becomes immediately clear. For example, NetClus [17] can help build concept hierarchies for bibliographic networks, and iTopicModel [26] can help build hierarchical topic models utilizing both text information and link information in a document network. By further integration of data cleaning, role discovery and trustworthiness analysis methods, ontology and structure discovery can be performed effectively using progressively enriched and refined heterogeneous networks.

IV. Trend and evolution analysis in heterogeneous information networks

There have been many studies on evolution and trend analysis in homogeneous networks. Comparatively, heterogeneous links capture more sensible information across multiple types of objects and thus facilitates such analysis more substantially.

Evolution analysis in heterogeneous information networks. Modeling co-evolution of multi-typed objects will capture richer semantics than modeling on single-typed objects alone. For example, studying co-evolution of authors, venues and terms in a bibliographic network can tell better the evolution of research areas than just examining co-author network or term network alone. Thus an important direction is how to model the co-evolution of multi-typed objects in the form of multi-typed cluster evolution in heterogeneous networks, such as EvoNetClus which builds a hierarchical Dirichlet process mixture model-based online model to study the real heterogeneous networks formed by DBLP and twitter [27].

Link and relationship prediction in heterogeneous information networks. One important application of network analysis is to predict links or interactions between objects in a network. Heterogeneous information network brings interactions among multiple types of objects and hence the possibility of predicting relationships across heterogeneous typed objects. For example, in a bibliographic network, there are multiple types of objects (*e.g.*, venues, topics, papers) 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 [22]. Moreover, by modeling the distribution of relationship building time between candidate objects with the use of the extracted topological features, one can also predict *when* a certain relationship will happen in the scenario of heterogeneous networks [28].

Diffusion analysis in heterogeneous information networks. In a well-connected online community, topics may spread ubiquitously among user-generated messages (*e.g.*, tweets) and documents. Together with this diffusion process is the evolution of topic content, where novel contents are introduced by documents which adopt the topic. Unlike explicit user behavior (*e.g.*, buying a TV), both the diffusion paths and the evolutionary process of a topic are implicit, making their discovery challenging. By modeling the task as a joint inference problem, considering textual documents, social influences, and topic evolution in a unified way, one can construct a probabilistic mixture model to track the evolution of an arbitrary topic and reveal the latent diffusion paths of that topic in a social community and achieve promising results [29].

V. 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 of them.

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 not straightforward. One of the major challenges is how to discover concept hierarchies for entities based on both data objects and their interactions in a network. Another is how to systematically discover subnetworks, summarize a heterogeneous network, and provide multiple views at different granularity in a network. There are some preliminary studies on this issue, such as [30], [31], [32]. However, much work needs to be done to make OLAP heterogeneous networks a reality.

Discovery and mining of hidden information networks. Although a network can be huge, a user at a time could be only interested in a tiny portion of nodes, links, or sub-networks. Instead of directly mining the entire network, it is more fruitful to mine hidden networks “extracted” dynamically from some existing networks, based on user-specified constraints or expected node/link behaviors. For example, instead of mining an existing social network, it could be more fruitful to mine networks containing suspects and their associated links; or mine subgraphs with nontrivial nodes and high connectivity. How to discover of such hidden networks and how to mine knowledge (*e.g.*, clusters, behaviors, and anomalies) from such hidden but non-isolated networks (*i.e.*, still intertwined with the gigantic network in both network linkages and semantics) could be an interesting but challenging problem.

Acknowledgment

The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), U.S. National Science Foundation grants IIS-0905215, CNS-0931975, and IIS-1017362, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] M. Newman, *Networks: An Introduction*. Oxford Univ. Press, 2010.

[2] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge Univ. Press, 2010.

[3] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *Proc. ACM SIGCOMM'99 Conf. Applications, Technologies, Architectures, and Protocols for Computer Communication*, Cambridge, MA, Aug. 1999, pp. 251–262.

[4] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 2003.

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

[6] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proc. 2005 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'05)*, Chicago, IL, Aug. 2005, pp. 177–187.

[7] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The web as a graph: Measurements, models, and methods," in *Proc. Int. Conf. Computing and Combinatorics (COCOON'99)*, Tokyo, Japan, July 1999, pp. 1–17.

[8] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 2000.

[9] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "SCAN: A structural clustering algorithm for networks," in *Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'07)*, San Jose, CA, Aug. 2007.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," in *Technical Report*, Computer Science Department, Stanford University, 1998.

[11] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, pp. 604–632, 1999.

[12] G. Jeh and J. Widom, "SimRank: a measure of structural-context similarity," in *Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'02)*, Edmonton, Canada, July 2002, pp. 538–543.

[13] ——, "Scaling personalized web search," in *Proc. 2003 Int. World Wide Web Conf. (WWW'03)*, New York, NY, May 2004, pp. 271–279.

[14] B. Long, Z. M. Zhang, X. Wu, and P. S. Yu, "Spectral clustering for multi-type relational data," in *Proc. 2006 Int. Conf. Machine Learning (ICML'06)*, Pittsburgh, PA, June 2006, pp. 585–592.

[15] X. Yin, J. Han, and P. S. Yu, "LinkClus: Efficient clustering via heterogeneous semantic links," in *Proc. 2006 Int. Conf. on Very Large Data Bases (VLDB'06)*, Seoul, Korea, Sept. 2006.

[16] ——, "Crossclus: User-guided multi-relational clustering," *Data Mining and Knowledge Discovery*, vol. 15, pp. 321–348, 2007.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han, "A Bayesian approach to discovering truth from conflicting sources for data integration," in *submitted for publication*, 2012.

[26] Y. Sun, J. Han, J. Gao, and Y. Yu, "iTopicModel: Information network-integrated topic modeling," in *Proc. 2009 Int. Conf. Data Mining (ICDM'09)*, Miami, FL, Dec. 2009.

[27] Y. Sun, J. Tang, J. Han, M. Gupta, and B. Zhao, "Community evolution detection in dynamic heterogeneous information networks," in *Proc. 2010 KDD Workshop on Mining and Learning with Graphs (MLG'10)*, Washington D.C., July 2010.

[28] 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.

[29] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky, "The joint inference of topic diffusion and evolution in social communities," in *Proc. 2011 IEEE Int. Conf. on Data Mining (ICDM'11)*, Vancouver, Canada, Dec. 2011.

[30] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient aggregation for graph summarization," in *Proc. 2008 ACM SIGMOD Int. Conf. Management of Data (SIGMOD'08)*, Vancouver, BC, Canada, June 2008, pp. 567–580.

[31] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu, "Graph OLAP: Towards online analytical processing on graphs," in *Proc. 2008 Int. Conf. Data Mining (ICDM'08)*, Pisa, Italy, Dec. 2008.

[32] P. Zhao, X. Li, D. Xin, and J. Han, "Graph cube: On warehousing and OLAP multidimensional networks," in *Proc. of 2011 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'11)*, Athens, Greece, June 2011.