International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 584-593 www.ijaem.net ISSN: 2395-5252

Review of Structural Analysis Approach on Heterogeneous Information Networks Mining

1.Santosh Kumar Jha, 2. Dr.Shyam Krishna Singh
[1] *Research Scholar, Dept. of Computer Science,Magadh University, Bodh-Gaya, Bihar*
[2]*Retd. Professor, Dept. of Mathematics,Magadh University, Bodh-Gaya, Bihar.*

Submitted: 05-11-2021 Revised: 17-11-2021 Accepted: 20-11-2021

ABSTRACT: Objects and data in the real world are of varying nature in their types, nature, complexity and hence heterogeneous forming semi-structured information networks. Mostly, network science researches are focussed on homogeneous networks without characterising various types of objects and links in the networks. Our view relates with interconnected, multityped data as heterogeneous information networks. In this paper we focus on the study of how to manage the rich semantic meaning of structural types of objects and various links in the networks and present a structural analysis approach on mining semi-structured, multityped heterogeneous information networks. Also, a set of methods that can mine useful knowledge form heterogeneous information networks is summarized and pointed out some favourable research direction in the area of interest.

KEYWORDS: Bibliographic Information Networks, Link Analysis, Clustering, Ranking

I. INTRODUCTION

Rapid development in the fields of information technology and Internet technology, we are living in the age of information overload from the era of lack of information. It is the time of interconnected world. Thus, large volume of data, objects, individual agents, groups, or components are interconnected or interacting with each other and forming numerous, large, and sophisticated interconnected networks. Such interconnected networks are called information networks, without loss of generality. Some examples of information networks are social networks, the World Wide Web, research publication networks, biological networks, highway networks, public health systems, electrical power grids, and so on. Clearly, information networks are universal and form a critical component of modern information infrastructure. Nowadays, the analysis of information networks or their special kinds, such as social networks and the Web has gained enormously varied attentions from researchers in computer science, social science, physics, economics, biology, and so on, with exciting discoveries and successful applications across all the disciplines.

We propose to model real-world systems from different applications as semi-structured heterogeneous information networks by structuring objects and their interactions into different types, and investigate the principles and methodologies for systematically mining such networks. Different from many existing network models that view interconnected data as homogeneous graphs or networks, our semi-structured heterogeneous information network model leverages the rich semantics of typed nodes and links in a network and uncovers surprisingly rich knowledge from the network.Examples are, in a bibliographic database like DBLP[1] and PubMed[2], papers are linked together via authors, venues and terms. Similarly, in Flickr[3], as a social website, photos are linked together via users, groups, tags and comments. Various kinds of knowledge can be derived from such an information network view, such as discovery of clusters and hierarchies, ranking, topic analysis, classification, similarity search, and relationship prediction. These functions facilitate the generation of new knowledge in ubiquitous online databases and other online or offline systems in almost every industry. For example, different research areas and ranks for authors and conferences can be discovered by such analysis in a bibliographic database which will be useful for the users to better understand the data and obtain valuable knowledge.

In this article we present an overview of the techniques developed for information network analysis in recent years. The motivation and related concepts are briefly introduced in Section 2. The

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 584-593 www.ijaem.net ISSN: 2395-5252

major mining tasks and techniques are presented in Section 3, and more advanced topics are in Section 4. In Section 5, we propose several research directions along the line of mining heterogeneous information networks. Finally, Section 6 concludesFinally, Section 6 concludes our study.

Heterogeneous Information Networks Mining

Several current research on network science, social and information networks are usually influenced to have homogeneous, in which nodes are objects of the same entity type (e.g., person) and links are relationships from the same relation type (e.g., friendship). Interesting results have been generated from such studies with numerous influential applications, such as the community detection method and well-known PageRank algorithms [1] .

In fact, most real world networks are heterogeneous, where nodes and relations are of different types.

For example, in a healthcare network, nodes can be patients, doctors, medical tests, diseases, medicines, hospitals, treatments, and so on. Assuming all the nodes as of the same type (e.g., homogeneous information networks) may lack important semantic information, whereas, treating every node as of a distinct type (e.g., labelled graph) may also lose valuable schema-level information. Therefore, it is important to know that patients are of the same kind, comparing with some other kinds, such as doctors or diseases. Thus, a typed, semi-structured heterogeneous network modelling may capture essential semantics of the real world.

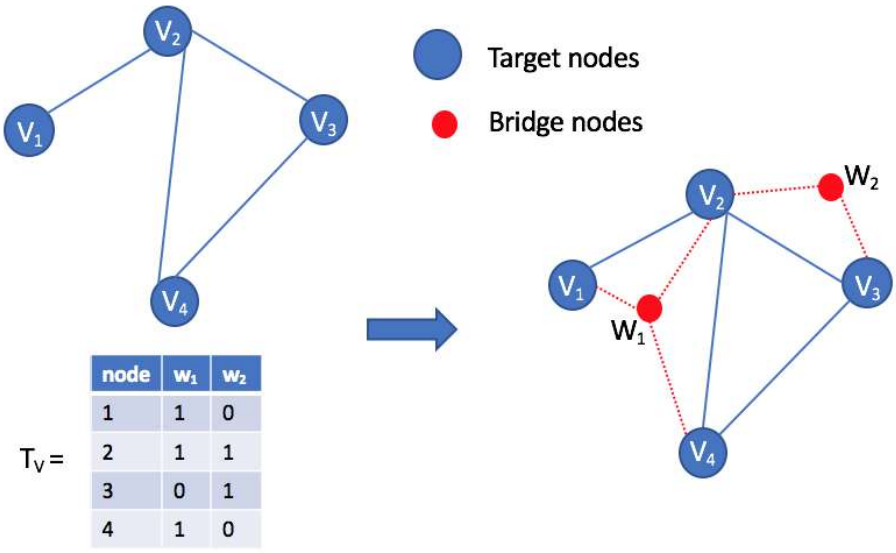


Figure 1: Converting homogeneous network in to heterogeneous network

II. EXPERIMENTATION

2.1 Definition

I. Information Network

Formally, a heterogeneous information network is defined as a directed graph G = (V, E) in which each vertex v ∈ V and each edge e ∈ E are associated with their type mapping functions τ (v) : V → TV and φ(e) : E → TE, respectively. TV and TE represent the sets of vertex and edge types, satisfying |TV | + |TE| > 2. If two edges have the same relation types, they share the same vertex types for both their source vertices and target vertices. If both |TV |=1 and |TE|=1, it is a homogeneous network with the same types of vertices and edges.

Here object types and relationship types in the network, which is different from traditional network definition. Note that, if a relation exists from type A to type B, denoted as A R B, the inverse relation R^{-1} holds naturally for B R^{-1} A. In this case of different types R and its inverse R^{-1} are usually not equal, unless the two types are the same and R is symmetric. When the types of objects |A| > 1 or the types of relations |R| > 1, the network is called heterogeneous information network. Otherwise, it is a homogeneous information network. Figure 2 shows the difference between the two networks.

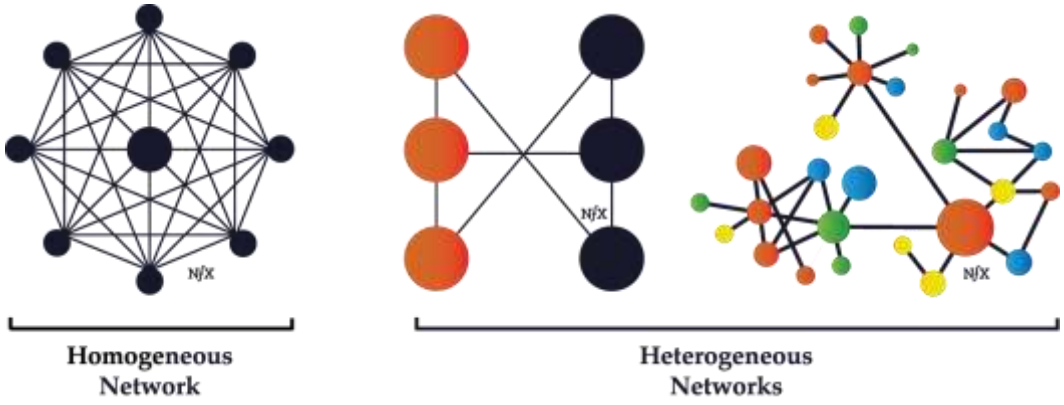
II. Network Schema

The network schema, denoted as TG = (A, R) is a meta template for a heterogeneous network G = (V, E) with the object type mapping τ : V → A and the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 584-593 www.ijaem.net ISSN: 2395-5252

link mapping φ : E → R, which is a directed graph defined over object types A, with edges as relations from R.

The network schema of a heterogeneous information network specifies type constraints on the sets of objects and relationships between the objects. These constraints make a heterogeneous information network semi-structured, guiding the exploration of the semantics of the network. An information network following a network schema is then called a network instance of the network schema. Heterogeneous information networks can be constructed from many interconnected, large-scale datasets, ranging from social, scientific, engineering to business applications. Here are a few examples of such networks.



[4]**Figure 2: Homogeneous vs. Heterogeneous Network**

Heterogeneous information networks can be constructed from many interconnected, large-scale datasets, ranging from social, scientific, engineering to business applications. Here are a few examples of such networks.

Bibliographic information network: A bibliographic information network, such as the computer science bibliographic information network derived from DBLP, is a typical heterogeneous network, containing objects in four types of entities: paper (P), venue (i.e., conference/journal) (V), author (A), and term (T). For each paper p ∈ P, it has links to a set of authors, a venue, and a set of terms, belonging to a set of link types. It may also contain citation information for some papers, that is, links to a set of papers cited by the paper and links from a set of papers citing the paper.

The network schema for a bibliographic network and an instance of such a network are shown in Fig. 3

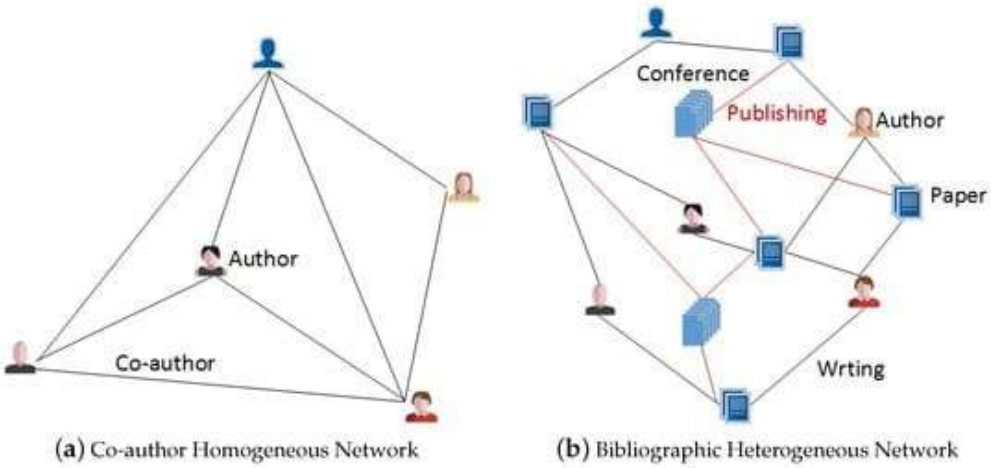


Figure 3: Bibliographic network instance (a) Homogeneous (b) Heterogeneous

B. Twitter information network: Twitter as a social media can also be considered as an information network, containing objects types such as user, tweet, hashtag and term, and relation (or link) types such as follow between users, post between users and tweets, reply between tweets,

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 584-593 www.ijaem.net ISSN: 2395-5252

use between tweets and terms, and contain between tweets and hashtags.

C. Flickr information network: The photo sharing website Flickr can be viewed as an information network, containing a set of object types: image, user, tag, group, and comment, and a set of relation types, such as upload between users and images, contain between images and tags, belong to between images and groups, post between users and comments and comment between comments and images.

D. Healthcare information network: A healthcare system can be modelled as a healthcare information network, containing a set of object types, such as doctor, patient, disease, treatment, and device, and a set of relation types, such as used-for between treatments and diseases, have between patients and diseases, and visit between patients and doctors.

Diverse information can be associated with information networks. Attributes can be attached to the nodes or links in an information network. For example, location attributes, either categorical or numerical, are often associated with some users and tweets in a Twitter information network. Also, temporal information is often associated with nodes and links to reflect the dynamics of an information network. For example, in a bibliographic information network, new papers and authors emerge every year, as well as their associated links. Besides the structure information of information networks, such content information is also helpful or even critical in some tasks on mining information networks.

2.2 Need of Mining Heterogeneous Networks

A homogeneous information network is usually obtained by projection from a heterogeneous information network, but with significant information loss. For example, a co-author network can be obtained by projection on co-author information from a more complete heterogeneous bibliographic network. However, such projection will lose valuable information on what subjects and which papers the authors were collaborating on. Moreover, with rich heterogeneous information preserved in an original heterogeneous information network, many powerful and novel data mining functions need to be developed to explore the rich information hidden in the heterogeneous links across entities.

Based on our research into mining heterogeneous information networks, especially our studies on ranking-based clustering [2, 3], ranking-based classification [4, 5], meta-path-based similarity search [8], relationship prediction [9 10], and relation strength learning [6, 7], we believe there are a set of new principles that may guide systematic analysis of heterogeneous information networks. We summarize these principles as follows:

Information propagation across heterogeneous types of nodes and links: Similar to most of the network analytic studies, links should be used for information propagation in mining tasks. However, the new game is how to propagate information across heterogeneous types of nodes and links, in particular, how to compute ranking scores, similarity scores, and clusters, and how to make good use of class labels, across heterogeneous nodes and links. No matter how we work out new, delicate measures, definitions, and methodologies, a golden principle is that objects in the networks are interdependent, and knowledge can only be mined using the holistic information in a network.

Search and mining by exploring network meta structures: Different from homogeneous information networks where objects and links are being treated either as of the same type or as of un-typed nodes or links, heterogeneous information networks in our model are semistructured and typed, that is, nodes and links are structured by a set of types, forming a network schema. The network schema provides a meta structure of the information network. It provides guidance of search and mining of the network and helps to analyse and understand the semantic meaning of the objects and relations in the network. Meta-path-based similarity search and mining has demonstrated the usefulness and the power of exploring network meta structures.

User-guided exploration of information networks: In a heterogeneous information network, there often exist numerous semantic relationships across multiple types of objects, carrying subtly different semantic meanings. A certain weighted combination of relations or meta-paths may best fit a specific application for a particular user. Therefore, it is often desirable to automatically select the right relation (or meta-path) combinations with appropriate weights for a particular search or mining task based on user's guidance or feedback. User-guided or feedback-based network exploration is a useful strategy.

III. MAJOR TASKS AND TECHNIQUES

3.1 Clustering and Classification in Heterogeneous Information Networks Clustering, classification and ranking are basic mining functions for information networks. We

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 584-593 www.ijaem.net ISSN: 2395-5252

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 rankingbased clustering approach (e.g., RankClus [1] and NetClus [11]) 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 ranked high in the same cluster. It turns out that the accuracy of clustering results can be significantly enhanced compared with that either using projected homogeneous information networks or using only partial link information. Moreover, by integrating ranking and clustering, a cluster can be understood easily by reading the top-ranked objects in that 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, and different types of links have different level of strengths in determining this similarity. 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 [12] and RankClass [3]. It turns out that by distinguishing different types of links in a heterogeneous information network, classification accuracy can be significantly enhanced.

3.2 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 interesting mining tasks, such as relationship prediction, can be addressed. Different from homogeneous information networks, two objects can be connected via different types of paths in a heterogeneous information network. For example, two authors can be connected via "author-paper-author" path, "authorpaper-venue-paper-author" path, and so on. Formally, these paths are called meta-paths, defined as follows:

Definition 3. (Meta-path) A meta-path P is a path defined on the graph of network schema TG = (A, R), and is denoted in the form of A1 $\xrightarrow{R1}$ A2 $\xrightarrow{R2}$ . . . $\xrightarrow{Rl}$ Al+1, which defines a composite relation R = R1 ◦ R2 ◦ . . . ◦ Rl between types A1 and Al+1, where ◦ denotes the composition operator on relations.

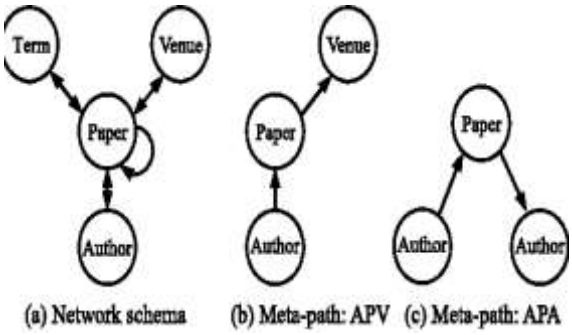


Figure 4: Bibliographic network schema and meta-paths.

For the bibliographic network schema shown in Figure 4 (a), we list two examples of meta-paths in Figure 4 (b) and (c), where an arrow explicitly shows the direction of a relation. We say a path p = (a1a2 . . . al+1) between a1 and al+1 in network G follows the meta-path P, if ∀i, ai∈ Ai and each link ei = haiai+1i belongs to each relation Ri in P. We call these paths as path instances of P, denoted as p ∈ P. Some path instance examples are shown in Table 1.

Through meta-paths, one can systematically specify how object types are connected in a network. Different meta-paths lead to different kinds of features. Multiple mining tasks can be explored under this framework.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 584-593 www.ijaem.net ISSN: 2395-5252

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.

Table 1: Path instances and their corresponding meta-paths in heterogeneous information networks.

	Connection Type I	Connection Type II
Path instance	Jim-P1-Ann Mike-P2-Ann Mike-P3-Bob	Jim-P1-SIGMOD-P2-Ann Mike-P3-SIGMOD-P2-Ann Mike-P4-KDD-P5-Bob
Meta-path	A(uthor)-P(aper)-A	A-P-V(enue)-P-A

IV. ADVANCED TOPICS

After the basic mining tasks discussed above, in this section, we introduce several advanced topics for mining information networks, which include role discovery, credibility analysis and co-evolution analysis, text mining in information networks, and OLAP in information networks. Many of these tasks can help better improve the quality of information networks, and others will help better understand the content rich information networks. More advanced operators such as OLAP is also necessary for better exploring the networks.

4.1 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. 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 [16]. It successfully mines advisor-advisee hidden roles in the DBLP database with high accuracy. Such mechanism can be further developed to discover hierarchical relationships [17] and ontology among objects under different kinds of user-provided constraints or rules.

4.2 Credibility Analysis in Information Networks

A major challenge for data integration is to derive the most complete and accurate integrated records from different 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 [18] and LTM [19], can be developed to resolve the conflicting source problem effectively. In [20], the authors propose to detect copying relationships among sources, which turns out to be critical in resolving conflicts among sources. Credibility analysis can help data cleaning and data integration, hence improving the quality of information networks.

4.3 Evolution Analysis in Heterogeneous Information Networks

Many current studies on network evolution are on homogeneous networks. However, in the real cases, different relationships exist in the heterogeneous network, and multityped relationships will co-evolve together. Modelling coevolution of multi-typed objects will capture richer semantics than modelling 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 [21].

4.4 Online Analytical Processing of Heterogeneous Information Networks

The power of online analytical processing (OLAP) has been shown in multidimensional analysis of structured, relational data. Similarly, users may like to view a heterogeneous information network from different angles, in different dimension combinations, and at different levels of granularity. For example, in a bibliographic network, by specifying the object type as paper and link type as citation relation, and rolling up papers into research topics, we can immediately see the citation relationships between different research topics and figure out which research topic would be the driving force for others. However, the extension

of the concept of online analysis processing (OLAP) to multi-dimensional data analysis of heterogeneous information networks is non-trivial. Not only different applications may need different ontological structures and concept hierarchies to summarize information networks but also because multiple pieces of semantic information in heterogeneous networks are tangled, determined by multiple nodes and links. There are some preliminary studies on this issue, such as [22-24], but the large territories of online analytical processing of information networks are still waiting to be explored.

V. RESEARCH CHALLENGES

By viewing interconnected data as an information network and learning scientifically the methods for mining heterogeneous information networks is a promising factor in data mining research. There are still many challenging research issues. Here we illustrate only a few.

5.1 Constructing and Refining Heterogeneous Information Networks

Many studies on mining heterogeneous information networks assume that a heterogeneous information network to be investigated contains a well-defined network schema and a large set of relatively clean and unambiguous objects and links. However, in the real world, things are more complicated. A network extracted from a relational database may contain a well-defined schema which can be used to define the schema of its corresponding heterogeneous information network. Nevertheless, objects and links even in such a database-formed information network can still be noisy. For example, in the DBLP network, different authors may share the same name [25], that is, one node in a network may refer to multiple real-world entities; whereas in some other cases, different nodes in a network may refer to the same entity. Entity resolution will need to be integrated with network mining in order to merge and split objects or links and derive high quality results. Moreover, links in a network, roles of a node with respect to some other nodes may not be explicitly given.

5.2 Diffusion Analysis in Heterogeneous Information Networks

Diffusion analysis has been studied on homogeneous networks extensively, from the innovation diffusion analysis in social science [27] to obesity diffusion in health science [26]. However, in the real world, pieces of information or diseases are propagated in more complex ways, where different types of links may play different roles. For example, diseases could propagate among people, different kinds of animals and food, via different channels. Comments on a product may propagate among people, companies, and news agencies, via traditional news feeds, social media, reviews, and so on. It is highly desirable to study the issues on information diffusion in heterogeneous information networks in order to capture the spreading models that better represent the real world patterns.

5.3 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 subnetworks. 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 behaviours. 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 such hidden networks and how to mine knowledge (e.g., clusters, behaviours, 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.

5.4 Discovery of Application-Oriented Ontological Structures in Heterogeneous Information Networks

As shown in the studies on ranking-based clustering and ranking-based classification, interconnected, multiple typed objects in a heterogeneous information network often provide critical information for generating high quality, finelevel concept hierarchies. For example, it is often difficult to identify researchers just based on their research collaboration networks. However, putting them in a heterogeneous network that links researchers with their publication, conferences, terms and research papers, their roles in the network becomes evidently clear. Moreover, people may have different preferences over ontological structures at handling different kinds of tasks. For example, some people may be interested in the research area hierarchy in the DBLP network, whereas others may be interested in finding the author lineage hierarchy. How to incorporate user's guidance, and generate adaptable ontological structures to meet user's requirement

and expectation could be an interesting and useful topic to study.

5.5 Intelligent Querying and Semantic Search in Heterogeneous Information Networks

Given real-world data are interconnected, forming gigantic and complex heterogeneous information networks, it poses new challenges to query and search in such networks intelligently and efficiently. Given the enormous size and complexity of a large network, a user is often only interested in a small portion of the objects and links most relevant to the query. However, objects are connected and inter-dependent on each other, how to search effectively in a large network for a given user's query could be a challenge. Similarity search that returns the most similar objects to a queried object, as studied in this thesis [28] and its follow-up [29], will serve as a basic function for semantic search in heterogeneous networks. Such kind of similarity search may lead to useful applications, such as product search in ecommerce networks and patent search in patent networks. Search functions should be further enhanced and integrated with many other functions. For example, structural search [30], which tries to find semantically similar structures given a structural query, may be useful for finding pattern in an e-commerce network involving buyers, sellers, products, and their interactions. Also, a recommendation system may take advantage of heterogeneous information networks that link among products, customers and their properties to make improved recommendations. Querying and semantic search in heterogeneous information networks opens another interesting frontier on research related to mining heterogeneous information networks.

VI. CONCLUSIONS

Many database researchers consider a database merely as a data repository that supports storage and retrieval only. They do not focus on the aspect of information-rich, inter-related and multi-typed information network that supports comprehensive data analysis. Now a days, most objects and data in the real world are interconnected, forming complex, heterogeneous but often semi-structured information networks. Many network researchers focus on homogeneous networks. Keeping aside the both, we view interconnected, semi-structured datasets as heterogeneous, information-rich networks and study how to uncover hidden knowledge in these networks. In this article, we present a systematized concept on mining heterogeneous information networks and introducing a set of remarkable, effective and scalable network mining techniques. Also, we presented several encouraging research topics in this exciting direction which will definitely be helpful for new generation research scholars in this related area of interest.

REFERENCES

[1]. S. Brin and L. Page.The anatomy of a large-scale hypertextual web search engine. In Proc. 7th Int. World Wide Web Conf. (WWW'98), pages 107–117, Brisbane, Australia, April 1998.

[2]. 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.

[3]. 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.

[4]. 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.

[5]. 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.

[6]. Y. Sun, C. C. Aggarwal, and J. Han. Relation strengthaware clustering of heterogeneous information networks with incomplete attributes. PVLDB, 5:394–405, 2012.

[7]. 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 Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12), Beijing, China, Aug. 2012.

[8]. 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.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. N. A. Christakis and J. H. Fowler.The spread of obesity in a large social network over 32 years.The New England Journal of Medicine, 357(4):370–379, 2007.

[13]. H. Deng, J. Han, M. R. Lyu, and I. King.Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In Proceedings of the 12th ACM/IEEECS joint conference on Digital Libraries (JCDL'12), pages 71–80, 2012.

[14]. 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.

[15]. 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.

[16]. 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.

[17]. C. Wang, J. Han, Q. Li, X. Li, W.-P.Lin, and H. Ji. Learning hierarchical relationships among partially ordered objects with heterogeneous attributes and links. In Proc. 2012 SIAM Int. Conf. on Data Mining (SDM'12), Anaheim, CA, April 2012.

[18]. 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.

[19]. B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conflicting sources for data integration. In Proc. 2012 Int. Conf. Very Large Data Bases (VLDB'12), Istanbul, Turkey, Aug. 2012.

[20]. X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava.Global detection of complex copying relationships between sources. Proc. VLDB Endow., 3(1-2):1358–1369, Sept. 2010.

[21]. 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.

[22]. 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.

[23]. 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), pages 567–580, Vancouver, BC, Canada, June 2008.

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

[25]. X. Yin, J. Han, and P. S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. In Proc. 2007 Int. Conf. Data Engineering (ICDE'07), Istanbul, Turkey, April 2007.

[26]. N. A. Christakis and J. H. Fowler.The spread of obesity in a large social network over 32 years. The New England Journal of Medicine, 357(4):370–379, 2007.

[27]. E. M. Rogers. Diffusion of Innovations, 5th Edition. Free Press, 2003.

[28]. 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.

[29]. C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu.Relevance search in heterogeneous networks.In Proc. 2012 Int. Conf. on Extending Database Technology (EDBT'12), pages 180–191, Berlin, Germany, March 2012.

[30]. X. Yu, Y. Sun, P. Zhao, and J. Han.Query-driven discovery of semantically similar substructures in heterogeneous networks. In Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12), Beijing, China, Aug. 2012.