A Survey on Mining Heterogeneous Information Network

Jing Yan†
Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
jyan@cs.hku.hk

Xiaodong Li †
Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
xdli@cs.hku.hk

ABSTRACT
The world we are living in is connected. In the other words, most of our real-world applications, like interconnected social media and social networks, scientific, engineering, and medical information systems, online e-commerce systems, and most database systems, can be structured into an expression of Network. Multi-type, and Multi-interactions make the network more like a semi-structured Heterogeneous Information Network. Realized the necessity to analyze such network, recently there are a lot of related papers about the Heterogeneous Information Network and leverage network analysis approach to reveal the rich information in the network. These work are mainly focused on Clustering, Ranking, and some Model Measurement. In this paper, we provide a survey of heterogeneous information network analysis. We will introduce basic concepts of heterogeneous information network analysis, examine its developments on different data mining tasks, discuss some advanced topics, and point out some future research directions.

1. INTRODUCTION
The world is interconnected with multiple relations and multiple objects. Networks (or graphs) is able to model real world entities and their relationships by objects and links. There have been rich researches focusing on the analysis of homogeneous information network which is partially reflect the real world. To explore the rich semantic information among nodes, many works have been done in Homogeneous Information Network (in which all objects/links are of one single type), link mining and analysis [8] social network analysis [34] hypertext and web mining [7] network science [17], and graph mining [9].

Most of these work are very based on the assumption, the type of objects of the network is unique. However, most of the real-life application system contained more than one type. For example, facebook Open Graph 1, contained objects of different types, such as messages, groups, locations, posts etc. With the rapid development, there are complex knowledge base. For instance, Yago 2 is a knowledge base that captures information derived from Wikipedia, WordNet and GeoNames. Yago is a repository of information on more than 10 million objects (such as persons, organizations, cities, etc.) and it records more than 120 million facts about these entities. The limitation of previous research which based on the Homogeneous Information Network become less able to describe or model the systems properly. Therefore, many research turned to Heterogeneous Information Network which models systems more accurately. A Heterogeneous Information Network (HIN) is a network whose objects are of different types and whose links represent different kinds of relationships between objects. Figure one are two simple instances on Heterogeneous Information Network. Compared to widely-used homogeneous information network, the heterogeneous information network can effectively fuse more information and contain rich semantics in nodes and links, and thus it forms a new development of data mining.

In this paper, we are going to talk about interesting and popular topics, including ranking and clustering, meta-path based HIN research, and other new models on Heterogeneous Information Network. Also, we will introduce some interesting application based on Meta-path, like co-author prediction, personalized recommendation, query recommendation. Finally, we will conclude and give some research frontiers from our opinion.

1 https://developers.facebook.com/docs/sharing/opengraph
2 http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago
2. BASIC CONCEPTS

In this paper we are going to talk about the mining work on Heterogeneous Information Network. In this section, we are going to have a clear and complete definition of Heterogeneous Information Network.

Definition 1. Information Network, is a directed graph \( G = (V, E) \) with an object type mapping function \( \phi : V \rightarrow A \) and a link type mapping function \( \psi : E \rightarrow R \), where each object \( v \in V \) belongs to an object type \( \phi(v) \in A \), and each link \( e \in E \) belongs to a link type \( \psi(e) \in R \).

Definition 2. Heterogeneous Information Network (HIN), is a Complex Information Network where the node types \(|A| > 1\) and like relations \(|R| > 1\). Otherwise, the information network is a Homogeneous Information Network.

Definition 3. HIN Schema. Given an HIN \( G = (V, E) \) with mappings \( \phi : V \rightarrow A \) and \( \psi : E \rightarrow R \), its schema \( T_G \) is a directed graph defined over object types \( A \) and link types \( R \), i.e., \( T_G = (A, R) \).

Figure 1 illustrates an HIN, which is also a bibliography network. A paper object can link (or be linked) to its authors, a venue and its related topics. Note that multiple edges of distinct types between two objects may exist. Figure 2 illustrate the instance of the Network Schema in a Publication Bibliography there are for bi-directed relations among four different node types.

3. RANKING AND CLUSTERING

Ranking and clustering analysis are fundamental tasks in data mining [5]. Clustering is to partition a set of data objects into a set of clusters, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. Clustering is based on the features of objects, such as k-means [13]. Also, there are some work about clustering based on networked data (e.g., community detection [23]).

Many works has adapted the traditional methods on Homogeneous Information Network to Heterogeneous Information Network. GNetMine [18] was proposed to model the link structure in information networks with arbitrary network schema and arbitrary number of object/link types. Recently, Luo et al. proposed HetPathMine [21] to cluster with small labeled data on HIN through a novel meta path selection model, and Jacob et al. [12] proposed a method to label nodes of different types by computing a latent representation of nodes in a space where two connected nodes tend to have close latent representations. Some works also extend inductive classification that is to construct a decision function in the whole data space.

Ranking is an important data mining task in network analysis, which evaluates object importance or popularity based on some ranking functions. Many ranking methods have been proposed in homogeneous networks. For example, PageRank [24] evaluates the importance of objects through a random walk process, and HITS ranks objects using the authority and hub scores [16]. These approaches only consider the same type of objects in homogeneous networks. PupRank [?] aims at ranking popularity of web objects. They have considered the role difference of different web pages, and thus turn web pages into a heterogeneous network.

3.1 RankClus Algorithm

In Heterogeneous Information Network, since the complexity and the large-scale and raw networks, links between nodes are in a mess. It is often difficult to understand the network in that way. [29] proposed a clustering ranking based algorithm to solve this problem.

In [29], the author proposed a novel algorithm to do the ranking procedure. A novel clustering framework called RankClus is proposed that directly generates clusters integrated with ranking.

The Algorithm can be divided into four parts, As illustrated in Table 1 and Table 2.

<table>
<thead>
<tr>
<th>Table 1: RankClus Algorithm Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 0</td>
</tr>
<tr>
<td>Step 1</td>
</tr>
<tr>
<td>Step 2</td>
</tr>
<tr>
<td>Step3</td>
</tr>
<tr>
<td>Step4</td>
</tr>
</tbody>
</table>

3.2 Bi-type HIN Rankclus Case Study

To better understand the algorithm, we introduce a case study in this section. The case is based on a Bi-type network, which is a Conference-author network. Links can exist between Conference (X) and author (Y), Author (Y) and
Also, we define a link matrix $W_{xy}$. The information network can be denoted as $G = (X \cup Y, W_{xy})$.

**Step 0: Initialization** The initial clusters for target objects are generated, by assigning each target object with a cluster label from 1 to $K$ randomly.

**Step 1: Ranking**

[29] Based on current clusters, $K$ cluster-induced networks are generated accordingly, and the conditional rank distributions for types $Y$ and $X$ are calculated. In this step, we also need to judge whether any cluster is empty, which may be caused by the improper initialization of the algorithm. When some cluster is empty, the algorithm needs to restart in order to generate $K$ clusters.

For the ranking calculation, [29] introduced two major ranking methods: Simple Ranking and Authority Ranking.

**Simple Ranking**, which is proportional to degree counting for objects (E.g., in bi-typed, this is the number of publications of authors). Simple Ranking only considers only immediate neighborhood in the network. In this case, simple ranking of conferences and authors is based on the number of publications, which is proportional to the numbers of papers accepted by a conference or published by an author. The ranking score for $X$ and $Y$ is:

$$f_X(x) = \sum_{j=1}^{m} W_{XY}(x,j) \frac{r_Y(j)}{\sum_{i=1}^{n} W_{XY}(i,j)}$$

$$f_Y(y) = \sum_{i=1}^{n} W_{XY}(i,y) \frac{r_X(i)}{\sum_{j=1}^{m} W_{XY}(i,j)}$$

The time complexity of Simple Ranking is $O(|E|)$, where $|E|$ is the number of links.

**For Authority Ranking**, the ranking score is propagated by iterations using rules 2 and 3, or rules 2 and 3. Take the case as example. the authority ranking of $X$ and $Y$ turned out to be primary eigenvectors of some symmetric matrix. Also, this ranking method considers the impact from the overall network, which should rank better than simple ranking.

For authority ranking, the time complexity is $O(t|E|)$, where $t$ is the iteration number and $|E|$ is the number of links in the graph. Notice that, $|E| = O(d|V|^2)$ in a sparse network, where $|V|$ is the number of total objects in the network and $d$ is the average link per each object.

**Step 2: Generate New Measure Space**

A naive method on this is to map target object to a $K$-dimensional vector directly by considering a sub-network induced by it.

In this paper, the author proposed a mixture model method. In their method, they consider each target objects links are generated under a mixture distribution of ranking from each cluster.

To be more Specific, We can consider ranking as a distribution: $r(Y) \rightarrow p(Y)$. And this distribution could be considered as a mixture model over $K$ component distributions, which are attribute types conditional rank distributions on $K$ clusters. Based on this, [29] proposed a mixture model where each target object $x_i$ is mapped into a $K$-vector $\pi_{i,k}$. This function use the $\pi_{i,k}$ to denote the $x_i$’s coefficient for component $k$.

For the parameters in this function, the author proposed the EM [6] algorithm to estimate, which maximizes the log-likelihood given all the observations of links.
Step 3: Adjusting Cluster

In this step, the author adjusted the cluster. The cluster center is in a new measure space regarding to step 1 and step 2. The procedure for adjustment is, measure the distance by $1 - \cosinelsim(Y)$ and assign to the cluster with the nearest center.

Step 4: Repeat Step 1, 2, 3

Repeat Steps 1, 2 and 3 until clusters change only by a very small ratio or the number of iterations is bigger than a predefined value $\text{iterNum}$. In [29], they set $\varepsilon = 0$, and $\text{iterNum} = 20$.

Figure 4 illustrates the steps and results on the bi-typed information network. And Figure 5 illustrates the results of the RankClus based on the dataset from DBLP.

Figure 4: Steps of RankClus

<table>
<thead>
<tr>
<th>DB</th>
<th>Network</th>
<th>AI</th>
<th>Theory</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VLDB</td>
<td>INFOCOM</td>
<td>AAMAS</td>
<td>SODA</td>
</tr>
<tr>
<td>2</td>
<td>ICEDE</td>
<td>SIGMETRICS</td>
<td>IICAI</td>
<td>STOC</td>
</tr>
<tr>
<td>3</td>
<td>SIGMOD</td>
<td>ICNP</td>
<td>AAAI</td>
<td>FOCS</td>
</tr>
<tr>
<td>4</td>
<td>KDD</td>
<td>SIGCOMM</td>
<td>Agentia</td>
<td>ICALP</td>
</tr>
<tr>
<td>5</td>
<td>ECDM</td>
<td>MIRONCOM</td>
<td>AAM/IAI</td>
<td>CCC</td>
</tr>
<tr>
<td>6</td>
<td>EDRF</td>
<td>ICDCS</td>
<td>ECAI</td>
<td>SPAA</td>
</tr>
<tr>
<td>7</td>
<td>DASFAA</td>
<td>NETWORKING</td>
<td>RoboCap</td>
<td>PODC</td>
</tr>
<tr>
<td>8</td>
<td>FODS</td>
<td>MobILoc</td>
<td>IAT</td>
<td>CRYPTO</td>
</tr>
<tr>
<td>9</td>
<td>SSDM</td>
<td>ISOC</td>
<td>ICMAS</td>
<td>APPROX-RANDOM</td>
</tr>
<tr>
<td>10</td>
<td>SDM</td>
<td>SenSys</td>
<td>CP</td>
<td>EUROCRYPT</td>
</tr>
</tbody>
</table>

Figure 5: Results of RankClus based on DBLP

3.3 Summary

The RankClus method, which provides a novel framework for the ranking-based clustering on Heterogeneous Information Network has raised great influence. From then on, there are many publications, which have discussed about the clustering or ranking on Heterogeneous Information Network.

The Rankclus might be efficient to comparing to other algorithms while comparing the similarity, need to calculate pairwise similarity [14], there are some weaknesses for RankClus: (1) it has not demonstrated the ability to clustering on networks with arbitrary number of types; and (2) the clusters generated by RankClus only contain one type of objects. To solve this problem, in [7], the author proposed a NetClus, which can generate net-clusters comprised of objects from multiple types, given any star network [30]. Besides, many works, [20, 4], aimed at clustering objects from different types simultaneously. Given different cluster number needed for each type of objects, clusters for each type are generated by maximizing some objective function.

4. META-PATH BASED RESEARCH

In this section, we are going to introduce some works which are based on the meta-path. Meta-path is the concept first introduced in [3] and from then on is widely used in exploiting the Heterogeneous Information Network.

Definition 4. A meta path, denoted by $P$, is essentially a path defined on an HIN schema $T_G$, with the types of source object and target object on both ends of the path.

Figure 6: Instances of Meta-path in a bibliography network

In this section, we are going to talk about the similarity search, discover met-path and some interesting topics, including co-authorship prediction, personalized recommendation and query recommendation.

4.1 Meta-path based similarity Search

In a bibliographic network, a user may be interested in the (top-k) most similar authors for a given author, or the most similar venues for a given venue. Therefore, it is important to provide a similarity search function. Before, there are many works about the similarity search, like SimRank [14], and Personalized Page Rank [15]. However, Adoption of such measures to heterogeneous networks has significant drawbacks: Objects of different types and links carry different semantic meanings, and it does not make sense to mix them to measure the similarity without distinguishing their semantics. To solve these problems, [14] proposed a new measurement function: PathSim.

PathSim, that captures the subtlety of peer similarity. The intuition behind it is that two similar peer objects should not only be strongly connected, but also share comparable visibility. As the relation of peer should be symmetric, PathSim merely on the symmetric meta paths. It is easy to see that, round trip meta paths with the form $P = (P, P_P^{-1})$.

Definition 5. Path count(PC): the number of path instances $p$ between $x$ and $y$ following $P$: $s(x, y) = |\{p : p \in P\}|$.

Definition 6.
PathSim: A Meta path-based similarity measure.
Given a symmetric meta path $P$, PathSim between two objects of the same type $x$ and $y$ is:

$$s(x, y) = \frac{2 \times |\{p_{x \rightarrow y} : p_{x \rightarrow y} \in P\}|}{|\{p_{x \rightarrow x} : p_{x \rightarrow x} \in P\}| + |\{p_{y \rightarrow y} : p_{y \rightarrow y} \in P\}|}$$

From this definition, it is easy to find $s(x, y)$ is defined in terms of two parts: (1) their connectivity defined by the number of paths between them following $P$; (2) the balance of their visibility, where the visibility is defined as the number of path instances between themselves. Then the paper introduced the calculation of PathSim between any two objects of the same type given a certain meta path by the Commuting Matrix.

Definition 7.

Commuting matrix. Given a network $G = (V, E)$ and its network schema $T_G$, a commuting matrix $M$ for a meta path $P = (A_1A_2 \ldots A_l)$ is defined as $M = W_{A_1A_2}W_{A_2A_3} \ldots W_{A_{l-1}A_l}$, where $W_{A_iA_j}$ is the adjacency matrix between type $A_i$ and type $A_j$. $M(i,j)$ represents the number of path instances between object $x_i \in A_i$ and object $y_j \in A_j$ under meta path $P$.

Figure compares the similarity search among different functions.

While the definition of meta path-based similarity search is flexible to accommodate different queries, it requires expensive computations (matrix multiplications). It is time and space expensive to materialize all the possible meta paths. For example, in the DBLP network, the similarity matrix corresponding to a length-4 meta path, APCPA, for identifying similar authors publishing in common venues is a 710K 710K matrix. [14] provided the solution: Partially materialize commuting matrices for short length meta paths, and concatenate them online to get longer ones for a given query. However, we reserved our opinion on that method, and therefore we won’t discuss about the details in this paper (just refer to the paper for more details).

4.2 Link Prediction

Link Prediction is a very popular and hot topic in the social network. Early work mainly studies unsupervised methods [1, 19] and later on, supervised methods that are able to combine different features with different coefficients via training data sets are proposed by different studies [2, 32]. Leveraging the advantage of the rich content in the Heterogeneous Information Network, and then do the link prediction intuitively will have some interesting outcomes.

Case: Co-authorship Prediction

In this case, the target relation for prediction is co-authorship relation, which can be described using meta-path $A \rightarrow P \rightarrow A$. For the topological features, [28] study all the meta-path listed in Table 5.1 other than $A \rightarrow P \rightarrow A$ and all the measures listed in the last section. [28] introduced the relationship prediction model which models the probability of co-authorship between two authors as a function of topological features between them. Given the training pairs of authors, the author first extract the topological features for them, and then build the prediction model to learn the weights associated with these features.

In order to predict whether two authors are going to collaborate in a future interval, denoted as $y$, [28] used the logistic regression model as the prediction model. For each training pair of authors $<a_{i1}, a_{i2}>$, they let $x_i$ be the $(d+1)$-dimensional vector including constant 1 and $d$ topological features between them, and $y_i$ be the label of whether they will be co-authors in the future ($y_i = 1$ if they will be co-authors, and 0 otherwise), which follows Bernoulli distribution with probability $pi$ ($P(y_i = 1) = pi$). The probability $pi$ is modeled as follows:

$$p_i = \frac{e^{x_i\beta}}{e^{x_i\beta} + 1}$$

where $\beta$ is the $d + 1$ coefficient weights associated with the constant and each topological feature

4.3 Recommendation

Personalized Entity Recommendation

Most of the previous studies in personalized recommendation area only consider a single relationship type, such as friendships in a social network. In many scenarios, the entity recommendation problem exists in a heterogeneous information network environment. Recently, some researchers have begun to be aware of the importance of heterogeneous information for recommendations [27]. The comprehensive information and rich semantics of HIN make it promising to generate better recommendations.

Figure 8: IMDB movie recommendation

Figure 8 shows such an example in movie recommendation [35].

In [35], In order to take full advantage of the relationship heterogeneity in information networks, the author first introduced meta-path-based latent features to represent the connectivity between users and items along different types.
of paths (leveraging diffusion process along meta-paths, in figure 9). Then they defined recommendation models at both global and personalized levels and use Bayesian ranking optimization techniques to estimate the proposed models. Empirical studies show that our approaches outperform several widely employed or the state-of-the-art entity recommendation techniques.

Figure 9: User Preference Diffusion

Query Recommendation

Extracting entities in the query, and then using the entity’s shortest meta-path in a Knowledge Base (also can be seen as the Heterogeneous Information Network) to recommend the peer entities [10].

4.4 Meta Structure

While investigating the use of HIN information for relevance computation, most of work only utilize simple structure, such as path, to measure the similarity (PathSim [3]) between objects. [11] propose to use meta structure, which is a directed acyclic graph of object types (see Figure 10) with edge types connecting in between, to measure the proximity between objects. The strength of meta structure is that it can describe complex relationship between two HIN objects (e.g., two papers in DBLP share the same authors and topics).

Figure 10: Meta-path, and Meta-Structure

4.5 Summary

In this section, we studies researches and some interesting applications based on the meta-path. Meanwhile, we also introduced other model on HIN to better measure the relevance, or similarity.

5. FRONTIERS

In this section, we are going to discuss some of the frontiers based on our understanding and sense. Also, we will summarize some of the future-forwards from some of the cited papers.

5.1 Further Similarity Search

Dynamic and Reverse version Search

The current similarity search are mostly conducted on the PathSim method. On one hand, this method reveals the connections following a meta-path between peer nodes, on the other hand, the method might just reveals the importance to me from a single degree. For example, as the student of Reynold, till my graduation, I might only have few papers coauthored with him. If we utilized the SimRank to evaluate the similarity in a top-k based, did we reveal such? Table 2 illustrates the different results of Reynold’s PathSim in a time basis. This indicates we should also maintain a dynamic version and also the reverse version. By these, we may have further explorations.

Table 2: Reynold’s top-5 APA Similarity

<table>
<thead>
<tr>
<th>NO.</th>
<th>All-Year</th>
<th>Recent 5-Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ben Kao</td>
<td>Ben Kao</td>
</tr>
<tr>
<td>2</td>
<td>Sunil Prabhakar</td>
<td>David Wai-Lok Cheung</td>
</tr>
<tr>
<td>3</td>
<td>David Wai-Lok Cheung</td>
<td>Luyi Mo</td>
</tr>
<tr>
<td>4</td>
<td>Jinchuan Chen</td>
<td>Matthias Renz</td>
</tr>
<tr>
<td>5</td>
<td>Silviu Maniu</td>
<td>Xuan S. Yang</td>
</tr>
</tbody>
</table>

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 users query could be a challenge. Similarity search that returns the most similar objects to a queried object, as studied in this thesis [14] and its follow-up [26], 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 e-commerce networks and patent search in patent networks. Search functions should be further enhanced and integrated with many other functions. Querying and semantic search in heterogeneous information networks opens another interesting frontier on research related to mining heterogeneous information networks.

5.2 Refining Heterogeneous Information Networks

Many works on Heterogeneous Information Network assume that a HIN 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. Though the work [22], but the method is more applicable in the network like DBLP or is able to detect few paths in Wikipedia or Freebase. In [33] provides an framework in which the users only need to provide few or several instances, the system will automatically detect other meta-paths in the HIN. Since in section 7, we also pointed out that constructing other models are possible.
and effective in some cases, how to refine other models like meta-structure in HIN?

5.3 Heterogeneous Network Embedding

Recently, since the first release of paper on Network Embedding [31], many works start to explore the Embedding. Although similarity search in HINs has been studied previously, most existing approaches neither explore rich semantic information embedded in the network structures nor take users preference as a guidance. In [25], authors reexamine similarity search in HINs and propose a novel embedding-based framework. It models vertices as low-dimensional vectors to explore network structure-embedded similarity.

6. CONCLUSION

In this paper, we generally discussing the Heterogeneous Information Network which is a more accurate model for the real world containing rich semantic schema. We have discussed the basic mining methods like ranking and clustering, meta-path which is a typical schema on HIN based researches, and some interesting topics. We also examine the meta-path, and introduce some new models. Finally, we raise some interesting questions and frontier research directions in our standing.

7. REFERENCES


[33] C. Wang, Y. Sun, Y. Song, J. Han, Y. Song, L. Wang, and M. Zhang. Relsim: Relation similarity search in schema-rich heterogeneous information networks.
