TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/10 pp329-338
Volume 18, Number 4, August 2013

Meta-Path-Based Search and Mining in Heterogeneous Information Networks

Yizhou Sun* and Jiawei Han

Abstract: Information networks that can be extracted from many domains are widely studied recently. Different functions for mining these networks are proposed and developed, such as ranking, community detection, and link prediction. Most existing network studies are on homogeneous networks, where nodes and links are assumed from one single type. In reality, however, heterogeneous information networks can better model the real-world systems, which are typically semi-structured and typed, following a network schema. In order to mine these heterogeneous information networks directly, we propose to explore the meta structure of the information network, i.e., the network schema. The concepts of meta-paths are proposed to systematically capture numerous semantic relationships across multiple types of objects, which are defined as a path over the graph of network schema. Meta-paths can provide guidance for search and mining of the network and help analyze and understand the semantic meaning of the objects and relations in the network. Under this framework, similarity search and other mining tasks such as relationship prediction and clustering can be addressed by systematic exploration of the network meta structure. Moreover, with user's guidance or feedback, we can select the best meta-path or their weighted combination for a specific mining task.

Key words: heterogeneous information network; meta-path; similarity search; relationship prediction; user-guided clustering

1 Introduction

Real-world physical and abstract data objects are interconnected, forming gigantic and interconnected networks. By structuring these data objects and interactions between these objects into multiple types, such networks become semi-structured heterogeneous information networks. Most real-world applications that handle big data, including interconnected social media and social networks, scientific, engineering, or medical information systems, online e-commerce systems, and most database systems, can be structured into heterogeneous information networks.

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 semi-structured and typed, that is, nodes and links are structured by a set of types, forming a *network schema*. For example, in a bibliographic database like DBLP (http://www.informatik.uni-trier.de/∼ley/db/) and PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), papers are linked together via authors, venues, and terms, and in Flickr (http://www.flickr.com/), a social website, photos are linked together via

• Yizhou Sun is with the College of Computer and Information Science, Northeastern University, Boston, MA 02115, USA. E-mail: yzsun@ccs.neu.edu.
• Jiawei Han is with the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. E-mail: hanj@cs.uiuc.edu.
∗ To whom correspondence should be addressed.
Manuscript received: 2013-07-17; accepted: 2013-07-17

users, groups, tags, and comments. Different kinds of knowledge can be derived from such an information network view, such as discovery of clusters and hierarchies[1-3], ranking[1,3,4], topic analysis[5,6], classification[7,8], similarity search[9,10], and relationship prediction[11,12]. 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 users to better understand the data and obtain valuable knowledge.

Most of the current network studies are based on homogeneous networks. In order to apply homogeneous information network-based methods into heterogeneous information networks, we have to either project the heterogeneous networks into homogeneous ones, or simply ignore the type information associated with nodes and links. Unfortunately, both ways will cause severe information loss. Therefore, there is a need to provide mining methodologies directly on heterogeneous information networks, by utilizing the semantic meaning of heterogeneous nodes and links. As objects are connected via different semantic meanings in a heterogeneous information network, we propose to fully use the network schema of the heterogeneous information network. The network schema provides a meta structure of the information network, and it provides guidance of search and mining of the network and help analyze and understand the semantic meaning of the objects and relations in the network. More specifically, a meta-path-based approach is proposed. Meta-path is a path defined on the network schema, which is a relation sequence between two object types and defines a new or existing relation between objects.

In this article, we introduce meta-path-based approaches on three types of mining tasks on heterogeneous information networks, namely, similarity search, relationship prediction, and clustering. At the end of the article, we discuss some research frontiers along this direction.

2 Heterogeneous Information Network and Meta-Path

An information network represents an abstraction of the real world, focusing on the *objects* and the *interactions* between the objects. It turns out that this level of abstraction has great power in not only representing and storing the essential information about the real-world, but also in providing a useful tool to mine knowledge from it, by exploring the power of links. Formally, we define an information network as follows.

Definition 1 Information network An *information network* is defined as a directed graph $G = (\mathcal{V}, \mathcal{E})$ with an object type mapping function $\tau : \mathcal{V} \rightarrow \mathcal{A}$ and a link type mapping function $\phi : \mathcal{E} \rightarrow \mathcal{R}$, where each object $v \in \mathcal{V}$ belongs to one particular object type $\tau(v) \in \mathcal{A}$, each link $e \in \mathcal{E}$ belongs to a particular relation $\phi(e) \in \mathcal{R}$, and if two links belong to the same relation type, the two links share the same starting object type as well as the ending object type.

Given a complex heterogeneous information network, it is necessary to provide its meta level (i.e., schema-level) description for better understanding the object types and link types in the network. Therefore, we propose the concept of network schema to describe the meta structure of a network.

Definition 2 Network schema The *network schema*, denoted as $T_G = (\mathcal{A}, \mathcal{R})$, is a meta template for a heterogeneous network $G = (\mathcal{V}, \mathcal{E})$ with the object type mapping $\tau : \mathcal{V} \rightarrow \mathcal{A}$ and the link mapping $\phi : \mathcal{E} \rightarrow \mathcal{R}$, which is a directed graph defined over object types $\mathcal{A}$, with edges as relations from $\mathcal{R}$.

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 are ubiquitous in real world, and we provide several examples in the following.

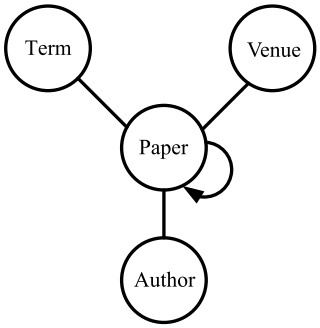
(1) **Bibliographic information network:** A bibliographic information network, such as the computer science bibliographic information network derived from DBLP, is a typical heterogeneous network, containing objects in four types of entities: *paper* (P), *venue* (i.e., conference/journal) (V), *author* (A), and *term* (T). For each paper, 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, these papers have 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. 1.

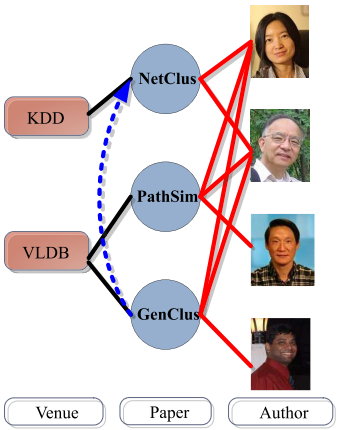
(2) **Twitter information network:** Twitter as a social media can also be considered as an information network, containing objects types such as *user, tweet, hashtag*, and *term*, and relation (or link) types such as *follow* between users, *post* between users and tweets, *reply* between tweets, *use*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

(3) **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 *comment* between comments and images.

(4) **Healthcare information network:** A healthcare system can be modeled 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.



(a) Schema of a bibliographic network



(b) A bibliographic network

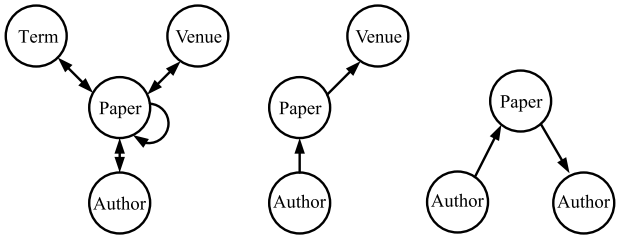
Fig. 1 A bibliographic network schema and a bibliographic network instance following the schema (only papers, venues, and authors are shown).

In heterogeneous information networks, objects can be connected via different types of relationships. In Ref. [9], we propose to use meta-path to systematically capture the relation type between two object types, which is formally defined as follows.

Definition 3 Meta-path A *meta-path* $\mathcal{P}$ is a path defined on the graph of network schema $T_G = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_l$ between types A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations.

For the bibliographic network schema shown in Fig. 2a, we list two examples of meta-paths in Figs. 2b and 2c, where an arrow explicitly shows the direction of a relation. We say a path $p = (a_1 a_2 \cdots a_{l+1})$ between a_1 and a_{l+1} in network G follows the meta-path $\mathcal{P}$, if $\forall i$, $a_i \in A_i$ and each link $e_i = \langle a_i a_{i+1} \rangle$ belongs to each relation R_i in $\mathcal{P}$. We call these paths as *path instances* of $\mathcal{P}$, denoted as $p \in \mathcal{P}$. The examples of path instances have been shown in Table 1, where we have listed the possible path instances between two authors and the meta-paths that these path instances belong to.

In addition to pointing out the meta-paths we are interested in, we also need to consider how to quantify the connection between two objects following a given meta-path. Typically, we can use the number of path



(a) Network schema (b) Meta-path: APV (c) Meta-path: APA

Fig. 2 Bibliographic network schema and meta-paths.

Table 1 Path instance vs. meta-path in heterogeneous information networks.

	Path instance	Meta-path
Connection Type I	Jim-P_1-Ann Mike-P_2-Ann Mike-P_3-Bob	Author-Paper-Author
Connection Type II	Jim-P_1-SIGMOD-P_2-Ann Mike-P_3-SIGMOD-P_2-Ann Mike-P_4-KDD-P_5-Bob	Author-Paper-Venue-Paper-Author

count, random walk-based measures, or PathSim[9] to quantify the meta-paths, more discussions of these measures can be found in Refs. [9,11,12]. Analogously, a meta-path-based measure in an information network corresponds to a feature or a feature type in a traditional data set, which can be used in many mining tasks.

In the next several sections, we demonstrate how meta-path-based approaches can be used in three very critical mining functions, i.e., similarity search, relationship prediction, and clustering.

3 Similarity Search

Similarity search in an information network aims to find the most similar or most proximate node for a given node. Links play a significant role in deciding the similarity between nodes, such as studied in personalized PageRank[13] and SimRank[14]. However, when coming to a heterogeneous information network, similarity measures can be defined according to different semantics. We then propose using meta-path to capture the different semantics of relationship type between two object types, and the meta-path-based similarity search framework is proposed accordingly.

3.1 Meta-path-based similarity search framework

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. For example, Table 2 shows that by using different meta-paths, one can find different author lists that are most similar to Christos Faloutsos, who is a very well-known data mining researcher at CMU. For example, by using *author-paper-author* meta-path, we can find Christos's students or close collaborators; by using *author-paper-venue-paper-author* meta-path, we can find other researchers with similar research area and reputation to Christos.

Table 2 Top-10 most similar authors to "Christos Faloutsos" under different meta-paths on the *full-DBLP* dataset.

(a) Path: APA		(b) Path: APVPA	
Rank	Author	Rank	Author
1	Christos Faloutsos	1	Christos Faloutsos
2	Spiros Papadimitriou	2	Jiawei Han
3	Jimeng Sun	3	Rakesh Agrawal
4	Jia-Yu Pan	4	Jian Pei
5	Agma J. M. Traina	5	Charu C. Aggarwal
6	Jure Leskovec	6	H. V. Jagadish
7	Caetano Traina Jr.	7	Raghu Ramakrishnan
8	Hanghang Tong	8	Nick Koudas
9	Deepayan Chakrabarti	9	Surajit Chaudhuri
10	Flip Korn	10	Divesh Srivastava

By quantifying the meta-path in different ways, we can further define similarity measures with different properties. A meta-path-based similarity measure, PathSim, is introduced in Ref. [9], for finding peer objects in the network, which generates better results, compared with random-walk-based similarity measures. Another measure, HeteSim, introduced in Ref. [15], computes relevance score between objects of different types.

3.2 PathSim: Finding similar peers

Although there have been several similarity measures, such as personalized PageRank and SimRank, they are biased to either highly visible objects or highly concentrated objects but cannot capture the semantics of peer similarity. For example, the path count and random walk-based similarity always favor objects with large degrees, and the pairwise random walk-based similarity favors concentrated objects where the majority of the links goes to a small portion of objects. However, in many scenarios, finding similar objects in networks is to *find similar peers*, such as finding similar authors based on their fields and reputation, finding similar actors based on their movie styles and productivity, and finding similar products based on their functions and popularity.

This motivated us to propose a new, meta-path-based similarity measure, called *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, we confine PathSim to symmetric meta-paths. It is easy to see that, *round trip meta-paths* in the form of $\mathcal{P} = (\mathcal{P}_l \mathcal{P}_l^{-1})$ are always symmetric.

Definition 4 PathSim: A meta-path-based similarity measure Given a symmetric meta-path $\mathcal{P}$, PathSim between two objects x and y of the same type is

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

where $p_{x \rightsquigarrow y}$ is a path instance between x and y, $p_{x \rightsquigarrow x}$ is that between x and x, and $p_{y \rightsquigarrow y}$ is that between y and y.

This definition shows that given a meta-path $\mathcal{P}$, $s(x, y)$ is defined in terms of two parts: (1) their

connectivity defined by the number of paths between them following $\mathcal{P}$; and (2) the balance of their visibility, where the visibility of an object according $\mathcal{P}$ is defined as the number of path instances between the object itself following $\mathcal{P}$. Note that we do count multiple occurrences of a path instance as the weight of the path instance, which is the product of weights of all the links in the path instance.

Table 3 presents in three measures the results of finding top-5 similar authors for "Anhai Doan," who is an established young researcher in the database field, under the meta-path APVPA (based on their shared venues), in the database and information system (DBIS) area. P-PageRank returns the most similar authors as those published substantially in the area, that is, highly ranked authors; SimRank returns a set of authors that are concentrated on a small number of venues shared with Doan; whereas PathSim returns Patel, Deshpande, Yang, and Miller, who share very similar publication records and are also rising stars in the database field as Doan. Obviously, PathSim captures desired semantic similarity as peers in such networks.

3.3 User-guided similarity search

So far, we have seen that different meta-paths imply different similarity semantics. But how can we select the best meta-path or their combinations for a specific search task? We now introduce how meta-path can help user-guided similarity search.

As shown in Fig. 3, different users may prefer different similarity measures even for the same query entity. Given the DBLP network, similarity queries which share the same format can possess different semantic meanings. In Fig. 3, both Query 1 and Query 1′ aim to find authors having similar relationships with "Christos Faloutsos," however, if we use the same ranking function to answer both queries, the results might not be satisfactory. In Query 1, the hidden similarity semantic meaning is described by two examples "Jimeng Sun" and "Hanghang Tong." Judged with human knowledge, both authors were data

Table 3 Top-5 similar authors for "AnHai Doan" in the DBIS area.

Rank	P-PageRank	SimRank	PathSim
1	AnHai Doan	AnHai Doan	AnHai Doan
2	Philip S. Yu	Douglas W. Cornell	Jignesh M. Patel
3	Jiawei Han	Adam Silberstein	Amol Deshpande
4	Hector Garcia-Molina	Samuel DeFazio	Jun Yang
5	Gerhard Weikum	Curt Ellmann	Renée J. Miller

Query 1:	find author similar to "Christos Faloutsos" taking "Jimeng Sun", "Hanghang Tong" as examples
Answer:	Agma J. M. Traina, Spiros Papadimitriou, Jure Leskovec
Query 1′:	find author similar to "Christos Faloutsos" taking "Philip S. Yu", "Jiawei Han" as examples
Answer:	Hector Garcia-Molina, H. V. Jagadish, Divesh Srivastava
Query 2:	find movie similar to "the Dark Knight" taking "Batman Begins", "Batman" as examples
Answer:	Batman Returns, Batman Forever, Batman: Mask of the Phantasm
Query 2′:	find movie similar to "the Dark Knight" taking "Hancock", "The Curious Case of Benjamin Button" as examples
Answer:	Iron Man, Cloverfield, Indiana Jones and the Kingdom of the Crystal Skull

Fig. 3 Motivating example of user-guided similarity search.

mining researchers and students of Christos Faloutsos at Carnegie Mellon University (CMU). When a user issues Query 1, they might be looking for other data mining researchers from CMU who have regular collaboration relationship with Christos. However, in Query 1′ we can see from the user guidance that the query is likely looking for other highly reputed data mining researchers similar to Christos. In the IMDB (http://www.imdb.com/) dataset, Query 2 represents searches for movies about the same topic (i.e., Batman) while Query 2′ represents a search for movies produced around the same time and in the same genre.

In order to provide personalized similarity search that can satisfy users' different search intentions, we ask user to provide several examples together with the query entity, which is studied in Ref. [10]. The whole system can be divided into offline and online components. In the offline part, different meta-path-based ranking models are trained. In the online part, the intention of the query will be identified and the query will be dispatched to the corresponding ranking model. Finally the selected similarity ranking model will return all the other similar objects to the query entity, given the query with guidance as examples.

4 Relationship Prediction

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 for effective relationship prediction[11,12].

4.1 Case study 1: Co-authorship prediction

As a case study, the co-authorship prediction problem is examined in Ref. [11], which outputs the most significant meta-paths for predicting co-authorships, as shown in Table 4, and also provides better understanding why co-author relationships are built. Note that, predicting co-authors for a given author is an extremely difficult task, as there are too many candidate target authors (3-hop candidates are used in analysis), but the number of real new relationships are usually very small. Table 5 shows the top-5 predicted co-authors in time interval T_2 (2003-2009) using the $T_0 - T_1$ (topological features are collected in 1989-1995 and co-authorship building ground truths are collected in 1996-2002) training framework, for both the proposed hybrid topological features and the shared co-author feature. We can see that the result generated by heterogeneous features has a higher accuracy compared with the homogeneous one.

4.2 Case study 2: Author citation prediction with time

Traditional link prediction studies have been focused on asking **whether** a link will be built in the future, such as "whether two people will become friends?" However, in many applications, it may be more interesting to predict **when** the link will be built, such as "what is the probability that two authors will co-write a paper within 5 years," and "by when will a user in Netflix rent the movie *Avatar* with 80% probability?"

In Ref. [12], we studied the problem of predicting the *relationship building time* between two objects, such as, when two authors will cite the other for the first time in the future, based on the topological structure in a heterogeneous network, by investigating the meta-path-based relationships between authors in the DBLP network. First, we introduce the concepts of target relation and topological features for the problem encoded in meta-paths[9]. Then, a Generalized Linear Model (GLM)[16] based supervised framework is proposed to model the relationship building time. In this framework, the building times for relationships are treated as independent random variables conditional on their topological features, which can follow different types of distributions, such as Weibull distribution, and their expectation is modeled as a function of a linear predictor of the extracted topological features. We propose and compare models with different distribution assumptions for relationship building time, where the parameters for each model are learned separately.

To better understand the output of our model, we now show a case study of predicting when the citation relationship will be built for "Philip S. Yu" with other candidates. The model is trained by *GLM-weib* using a training interval of 9 years ($T_1^{\text{train}} = [2001, 2009]$), with the learned parameter $\lambda = 0.9331$, slightly less than 1, which means the citation relationship has a higher hazard happening at an earlier time. The ground truth of the citation building time, and the predicted median, mean, 25% quantile, and 75% quantile for several test pairs are shown in Table 6. It can be seen that the predicted median and confidence interval are very suggestive for predicting the true citation relationship building time. For those authors whose predicted being cited time is significantly different from the ground truth, in-depth studies may be needed. For example, David Maier is a prolific researcher in database area, and by intuition as well as suggested by the model, Philip should cite him. However, the ground truth says otherwise. Furthermore, this function can be used to recommend authors to any author in DBLP for citation purpose.

Table 4 Significance of meta-paths with *Normalized Path Count measure* for *HP3hop* dataset.

Meta-path	p-value	Significance level
A − P → P − A	0.0378	**
A − P ← P − A	0.0077	***
A − P − V − P − A	1.2974×10^{-174}	****
A − P − A − P − A	1.1484×10^{-126}	****
A − P − T − P − A	3.4867×10^{-51}	****
A − P → P → P − A	0.7459	
A − P ← P ← P − A	0.0647	*
A − P → P ← P − A	9.7641×10^{-11}	****
A − P ← P → P − A	0.0966	*

Notes: *, $p < 0.1$; **, $p < 0.05$; ***, $p < 0.01$; ****, $p < 0.001$

Table 5 Top-5 predicted co-authors for Jian Pei in 2003-2009.

Rank	Hybrid heterogeneous features	# of shared authors as feature
1	**Philip S. Yu**	**Philip S. Yu**
2	**Raymond T. Ng**	Ming-Syan Chen
3	Osmar R. Zaïane	Divesh Srivastava
4	**Ling Feng**	Kotagiri Ramamohanarao
5	**David Wai-Lok Cheung**	**Jeffrey Xu Yu**

Note: Bold font indicates true new co-authors of Jian Pei in the period of 2003-2009.

Table 6 Case studies of author citation relationship building time prediction.

a_i	a_j	Ground truth	Median	Mean	25% quant.	75% quant.
Philip S. Yu	Ling Liu	1	2.2386	3.4511	0.8549	4.7370
Philip S. Yu	Christian Jensen	3	2.7840	4.2919	1.0757	5.8911
Philip S. Yu	C. Lee Giles	0	8.3985	12.9474	3.2450	17.7717
Philip S. Yu	Stefano Ceri	0	0.5729	0.8833	0.2214	1.2124
Philip S. Yu	David Maier	9+	2.5675	3.9581	0.9920	5.4329
Philip S. Yu	Tong Zhang	9+	9.5371	14.7028	3.6849	20.1811
Philip S. Yu	Rudi Studer	9+	9.7752	15.0698	3.7769	20.6849

For the above model, the learned *top-4* most important topological features with the highest coefficients are:

(1) A − P − T − P − A, that is, if two authors are very similar in terms of writing similar topics, they tend to cite each other;

(2) A − P ← P → P − A, that is, if two authors are very similar in terms of being frequently co-cited by the common papers, they tend to cite each other;

(3) A − P − A − P → P − A, that is, an author tends to cite the authors that are frequently cited by her co-authors;

(4) A−P−T−P−A−P → P−A, that is, if two authors are similar in terms writing similar topics, they tend to cite the same authors.

These topological features provide insightful knowledge for people to understand the citation relationship building between authors.

5 User-Guided Clustering

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. The preferred meta-path or weighted meta-path combinations can be learned to reach better consistency between mining results and the training examples[2].

Example 1 Meta-path-based clustering A toy heterogeneous information network is shown in Fig. 4, which contains three types of objects: organization (O), author (A), and venue (V), and two types of links: the solid line represents the affiliation relation between author and organization, whereas the dashed one the publication relation between author and venue. Authors are then connected (indirectly) via different meta-paths. For example, AOA is a meta-path denoting a relation between authors via organizations

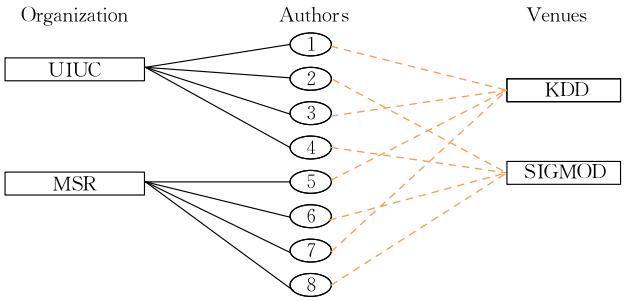


Fig. 4 A toy heterogeneous information network containing organizations, authors, and venues.

(i.e., colleagues), whereas AVA denotes a relation between authors via venues (i.e., publishing in the same venues). A question then arises: *which type of connections should we use to cluster the authors?*

Obviously, there is no unique answer to this question: different meta-paths lead to different author connection graphs, which may lead to different clustering results.

In Fig. 5a, authors are connected via organizations and form two clusters: $\{1, 2, 3, 4\}$ and $\{5, 6, 7, 8\}$; in Fig. 5b, authors are connected via venues and form two different clusters: $\{1, 3, 5, 7\}$ and $\{2, 4, 6, 8\}$; whereas in Fig. 5c, a connection graph combining both meta-paths generates 4 clusters: $\{1, 3\}, \{2, 4\}, \{5, 7\}$, and $\{6, 8\}$.

In Ref. [2], the PathSelClus algorithm is proposed to learn the importance of each meta-path as well as output the clustering results that are consistent with the user guidance. For example, to cluster authors into clusters in Example 1, a user may seed $\{1\}$ and $\{5\}$ for two

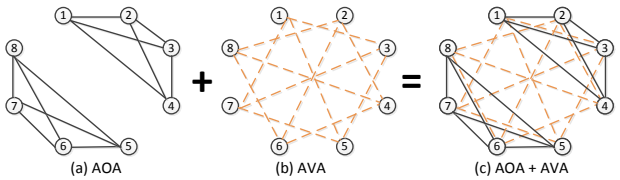


Fig. 5 Author connection graphs under different meta-paths.

clusters, which implies a selection of meta-path AOA; or seed $\{1\}, \{2\}, \{5\}$, and $\{6\}$ for four clusters, which implies a combination of both meta-paths AOA and AVA with about equal weight.

6 Research Frontiers

In this section, we discuss several research frontiers along the direction of meta-path-based mining on heterogeneous information networks.

6.1 Diffusion analysis in heterogeneous information networks

Diffusion analysis has been studied on homogeneous networks extensively, from the innovation diffusion analysis in social science[17] to obesity diffusion in health science[18]. 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, where meta-path can play an important role.

6.2 Recommendation in heterogeneous information networks

Recommendation function that recommends items to users is very useful in many real-world systems. Traditional recommendation algorithms utilize only the interaction between users and items as described in a user-item rating matrix or interaction matrix. However, a recommendation system may take advantage of heterogeneous information networks that link among products, customers, and their properties to make improved recommendations. For example, in Ref. [19], it has shown that by considering different recommendation factors that can be obtained via different meta-paths, the recommendation accuracy can be further enhanced.

6.3 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 Ref. [9] and its follow-up[15], 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. For example, structural search[20], 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. Querying and semantic search in heterogeneous information networks opens another interesting frontier on research related to mining heterogeneous information networks, where meta-paths could be used to describe complicated constraints for these queries.

7 Conclusions

In this article, we have discussed the challenges raised by search and mining heterogeneous information networks. We point out that meta-path is a powerful tool to systematically define the relation between two object types and capture different semantic meanings of the connection between objects. Under the meta-path framework, we have introduced how similarity search, relationship prediction, and user-guided clustering can be studied. By utilizing the meta-path framework, we are not only able to produce higher accuracy in all these tasks, but also able to provide a better understanding towards which type of relations play an important role in these user specified tasks. Finally, we have introduced several research frontiers under this framework.

Acknowledgements

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-0905215,

CNS-09-31975, and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC.

References

[1] 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.

[2] 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.

[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] 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/IEEE-CS Joint Conference on Digital Libraries (JCDL'12)*, 2012, pp. 71-80.

[5] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin, Probabilistic topic models with biased propagation on heterogeneous information networks, in *Proc. 2011 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, San Diego, CA, USA, Aug. 2011.

[6] 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, USA, Dec. 2009.

[7] 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.

[8] 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.

[9] 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, USA, Aug. 2011.

[10] X. Yu, Y. Sun, B. Norick, T. Mao, and J. Han. User guided entity similarity search using meta-path selection in heterogeneous information networks, in *Proc. 2012 Int. Conf. on Information and Knowledge Management (CIKM'12)*, Maui, Hawaii, USA, Oct. 2012.

[11] 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, China, July 2011.

[12] 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, USA, Feb. 2012.

[13] G. Jeh and J. Widom, Scaling personalized web search, in *Proc. 2003 Int. World Wide Web Conf. (WWW'03)*, Budapest, Hungary, May 2003.

[14] G. Jeh and J. Widom, Simrank: A measure of structural-context similarity, in *Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'02)*, Edmonton, Canada, July 2002.

[15] 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)*, Berlin, Germany, March 2012, pp. 180-191.

[16] A. J. Dobson, *An Introduction to Generalized Linear Models, Second Edition*. Chapman & Hall/CRC, 2001.

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

[18] 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*, vol. 357, no. 4, pp. 370-379, 2007.

[19] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han, HeteRec: Entity recommendation in heterogeneous information networks with implicit user feedback, in *Proc. of 2013 ACM Int. Conf. Series on Recommendation Systems (RecSys'13)*, Hong Kong, China, Oct. 2013.

[20] 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.

Yizhou Sun received her PhD from UIUC in 2012. She is an assistant professor in the College of Computer and Information Science of Northeastern University. Her principal research interest is in mining information and social networks, and more generally in data mining, database systems, statistics, machine learning, information retrieval, and network science. She has over 40 publications in books, journals, and major conferences. Tutorials based on her thesis work on mining heterogeneous information networks have been given in several premier conferences, including EDBT'09, SIGMOD'10, KDD'10, ICDE'12, VLDB'12, and ASONAM'12. She received ACM KDD 2012 Best Student Paper Award.

Jiawei Han received his PhD degree from University of Wisconsin-Madison in 1985. He is bliss professor of Computer Science, University of Illinois at Urbana-Champaign. He has been researching into data mining, information network analysis, database systems, and data warehousing, with over 600 journal and conference publications. He has chaired or served on many program committees of international conferences, including PC co-chair for KDD, SDM, and ICDM conferences, and Americas Coordinator for VLDB conferences. He also served as the founding Editor-In-Chief of ACM Transactions on Knowledge Discovery from Data and is serving as the Director of Information Network Academic Research Center supported by U.S. Army Research Lab. He is a fellow of ACM and IEEE, and received 2004 ACM SIGKDD Innovations Award, 2005 IEEE Computer Society Technical Achievement Award, 2009 IEEE Computer Society Wallace McDowell Award, and 2011 Daniel C. Drucker Eminent Faculty Award at UIUC. His book "Data Mining: Concepts and Techniques" has been used popularly as a textbook worldwide.