Knowledge-Based Systems 154 (2018) 32–42

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A semantic-rich similarity measure in heterogeneous information networks

Yu Zhou[a], Jianbin Huang[*,a], He Li[a,b], Heli Sun[c], Yan Peng[d], Yueshen Xu[a]

[a] School of Software, Xidian University, Xi'an, Shaanxi, China
[b] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[c] Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, China
[d] College of Information Science and Engineering, Wuchang Shouyi University, Wuchang, Wuhan, China

ARTICLE INFO

Keywords:
Heterogeneous information network
Similarity
Meta path
Meta structure
Stratified meta structure

ABSTRACT

Most of the existing similarity metrics in heterogeneous information networks depend on the pre-specified meta-path or meta-structure. This dependency may cause them to be sensitive to different meta-paths or meta-structures. In this paper, we propose a stratified meta-structure-based similarity measure named SMSS in heterogeneous information networks. The stratified meta-structure can be constructed automatically and capture rich semantics.Then, we define the commuting matrix of the stratified meta-structure by virtue of the commuting matrices of meta-paths and meta-structures. As a result, the SMSS is defined by virtue of this commuting matrix. Experimental evaluations show that the existing metrics are sensitive to different meta-paths or meta-structures and that the proposed SMSS outperforms the state-of-the-art metrics in terms of ranking and clustering.

1. Introduction

Information network analysis attracts the attention of many researchers in the field of data mining because many real systems, e.g., bibliographic information database and biological systems, can be modeled as information networks. These kinds of networks are composed of multi-typed and interconnected objects and they are usually called heterogeneous information networks (HIN).

A fundamental problem in HINs is that of measuring the similarities between objects using structural and semantic information. All the off-the-shelf similarities in HIN are based on the pre-specified meta-path such as PathSim [1] and biased path constrained random walk (BPCRW) [2,3]. According to the literature [4], meta-paths can only capture biased and relatively simple semantics. Therefore, the authors proposed a more complex structure called meta-structure, and defined the meta-structure based similarity using the compressed-ETree, called the biased structure constrained subgraph expansion (BSCSE). However, the meta-structure must be specified in advance, as well.

It is truly quite difficult for users to specify meta-paths or meta-structures. For example, there are ten object types (Gene, Gene Ontology, Tissue, Chemical Compound, Side Effect, Substructure, Chemical Ontology, Pathway, Disease, Gene Family) and eleven link types in a complete biological information network [5,6]. Obviously, users hardly know how to choose appropriate meta-paths or meta-structures. In addition, different meta-paths and meta-structures have different effects on similarities between objects. This situation makes it more difficult for users to select appropriate meta-paths or meta-structures.

To alleviate the users' burden, we propose an automatically constructed schematic structure called the stratified meta-structure (SMS). This structure does not need to be specified in advance, and it combines many meta-paths and meta-structures. This combination ensures that (1) users need not worry about how to choose the meta-path or meta-structure; (2) rich semantics can be captured. We are inspired by the tree-walk proposed in [7]. The structure of a tree-walk is constructed by repetitively visiting nodes in the input graph. This idea can be employed here. As a result, we devise the stratified meta-structure, which is essentially a directed acyclic graph consisting of the object types with different layer labels. The SMS can be automatically constructed via repetitively visiting the object types in the network schema. In the process of construction, we discover that the SMS consists of many basic substructures and recurrent substructures (see Section 4.2). These basic substructures and recurrent substructures essentially represent specific relations. The SMS as a composite structure is therefore a composite relation. This is why the SMS can capture rich semantics.

After obtaining the SMS, the next step is to formalize its rich semantics. For meta-structures, the compressed-ETree is used to formalize the semantics. However, it cannot be used here because the SMS contains an infinite number of meta-structures. The semantics contained in meta-paths are usually formalized by their commuting matrices. In

* Corresponding author.
E-mail addresses: peterjone85@hotmail.com (Y. Zhou), jbhuang@xidian.edu.cn (J. Huang), heli@xidian.edu.cn (H. Li), hlsun@mail.xjtu.edu.cn (H. Sun), pey9076@hotmail.com (Y. Peng), ysxu@xidian.edu.cn (Y. Xu).

https://doi.org/10.1016/j.knosys.2018.05.010
Received 28 December 2017; Received in revised form 7 May 2018; Accepted 9 May 2018
Available online 09 May 2018
0950-7051/

essence, the meta-structures have the same nature as the meta-paths because they all have hierarchical structures. Therefore, we define commuting matrices of meta-structures by virtue of the Cartesian product in Section 3.2, and we further define the commuting matrix of the SMS by reasonably combining the infinite number of the commuting matrices of meta-structures. The proposed metric SMSS is defined by the commuting matrix of the SMS. Experimental evaluations suggest that 1) all the off-the-shelf metrics including PathSim, BPCRW and BSCSE are sensitive to different meta-paths or meta-structures and that 2) SMSS, on the whole, outperforms the baselines PathSim, BPCRW and BSCSE in terms of ranking quality and clustering quality.

The main contributions are summarized as follows. 1) We propose the stratified meta-structure with rich semantics, which can be constructed automatically, and we define a stratified meta-structure-based similarity measure SMSS by virtue of the commuting matrix of the SMS; 2) we define the commuting matrices of meta-structures by virtue of the Cartesian product, and we use them to compactly re-formulate the BSCSE; 3) we conduct experiments for evaluating the performance of the proposed metric SMSS. On the whole, the proposed metric outperforms the baselines in terms of ranking quality and clustering quality.

The rest of the paper is organized as follows. Section 2 introduces related works. Section 3 provides some preliminaries in HINs. Section 4 introduces the definition of SMSS. The experimental evaluations are introduced in Section 5. The conclusion is introduced in Section 6.

2. Related work

Sun et al. [8,9] and Sun and Han[10] proposed the definition of the HIN and studied ranking-based clustering in HINs. Shi et al. [11] gave a comprehensive summarization of research topics on HINs including similarity measure, clustering, link prediction, ranking, recommendation, and information fusion and classification. Article [12] proposed a novel meta-path based framework called HeteClass for transductive classification of target type objects. This framework can explore the network schema of the input HIN and incorporate the expert's knowledge to generate a collection of meta-paths. Articles [13,14] solved the team formation problem in the heterogeneous tripartite networks consisting of projects, experts and skills. Article [15] studied the social event organization approach in the heterogeneous networks consisting of users and events. Below, we summarize related works on similarity measures in information networks.

For similarity measures in homogeneous information networks, literature [16] proposed a general similarity measure SimRank combining the link information, which thought that two similar objects must relate to similar objects. Literature [17] evaluated the similarities of objects by a random-walk model with restart. Article [18] lists many state-of-the-art similarities in homogeneous information networks: (1) Local Approaches: e.g. Common Neighbors (CN), Adamic–Adar Index (AA), Resource Allocation Index (RA), Resource Allocation based on Common Neighbor Interactions (RA-CNI), Preferential Attachment Index (PA), Jaccard Index (JA), Salton Index (SA), Sorensen Index (SO), Hub Promoted Index (HPI), Hub Depressed Index (HDI), Local Leicht–Holme–Newman Index (LLHN), Individual Attraction index (IA), Mutual Information (MI), Local Naive Bayes (LNB), CAR-Based Indices (CAR), Functional Similarity Weight (FSW), Local Interacting Score (LIT); (2) Global Approaches: Negated Shortest Path (NSP), Katz Index (KI), Global Leicht–Holme–Newman Index (GLHN), Random Walks (RA), Random Walks with Restart (RWR), Flow Propagation (FP), Maximal Entropy Random Walk (MERW), Pseudo-inverse of the Laplacian Matrix (PLM), Average Commute Time (ACT), Random Forest Kernel Index (RFK), The Blondel index (BI); (3) Quasi-Local Approaches: Local Path Index (HPI), Local Random Walks (LRW), Superposed Random Walks (SRW), Third-Order Resource Allocation Based on Common Neighbor Interactions (ORA-CNI), FriendLink (FL), PropFlow Predictor (PFP).

For similarity measures in heterogeneous information networks, Sun [1] proposed a meta-path based similarity measure in HINs, called PathSim. Lao and Cohen [2,3] studied the problem of measuring the entity similarity in labeled directed graphs, and defined a biased path constrained random walk (BPCRW) model, which can be applied to HINs. Huang et al. [4] proposed a similarity BSCSE, which can capture more complex semantics. Shi et al. [19] proposed a relevance measure HeteSim, which can be used to evaluate the relatedness of two objects with different types. For a user-specified meta-path, HeteSim is based on the pairwise random walk from its two endpoints to its center. Xiong et al. [20] studied the problem of finding the top-k similar object pairs by virtue of locality sensitive hashing. Zhu and Lglesias [21] proposed an integrated framework for the development, evaluation and application of semantic similarity for knowledge graphs which can be viewed as complicated heterogeneous information networks. This framework included many similarity tools and allowed users to compute semantic similarities. In the article [22], the authors studied the similarity search problem in social and knowledge networks and proposed a dual-perspective similarity metric called forward backward similarity.

3. Preliminaries

In this section, we introduce some important concepts related to HINs including network schema, meta-path and meta-structure.

3.1. HIN definition

As defined in article [23], an information network is essentially a directed graph $G = (V, E, \mathscr{A}, \mathscr{R})$. V and E denote the set of objects and links respectively, and $\mathscr{A}$ and $\mathscr{R}$ denote the set of object types and link types respectively. Map $\phi: V \to \mathscr{A}$ denotes the object type $\phi(v)$ of object $v \in V$. That is, each object in V belongs to a specific object type. Similarly, map $\psi: E \to \mathscr{R}$ represents the link type $\psi(e)$ of link $e \in E$, i.e., each link in E belongs to a specific link type. In essence, $\psi(e)$ contains some semantics because it is a relation between the source object type and the target object type. If two links belong to the same link type, they share the same starting object type as well as the ending object type. G is called a heterogeneous information network if $|\mathscr{A}| > 1$ or $|\mathscr{R}| > 1$. Otherwise, it is called a homogeneous information network.

Fig. 1 shows a toy bibliographic information network with four actual object types *Author* (A) in the shape of triangles, *Paper* (P) in the shape of circles, *Venue* (V) in the shape of pentagons and *Term* (T) in the shape of squares. The type P has six instances: P:PathSim [1], P:NetClus [8], P:HeteSim [19], P:HeProjI [23], P:GenClus [24], and P:PathSelClus [25]. Each paper has its author(s), a venue and its related terms. Hence,

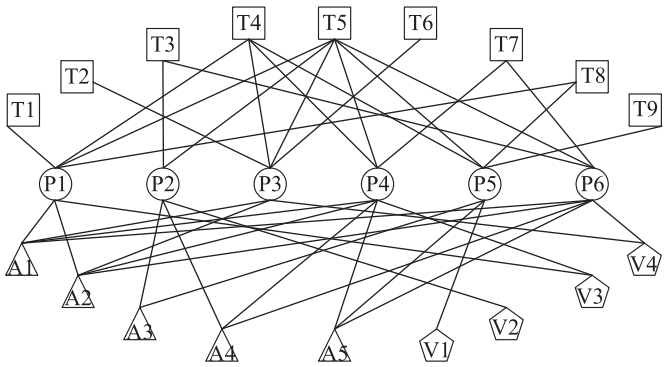

Fig. 1. A Toy Bibliographic Information Network. T1, T2, T3, T4, T5, T6, T7, T8 and T9 respectively stand for terms 'NetworkSchema', 'RelationStrength', 'Similarity', 'Clustering', 'HIN', 'Attribute', 'MetaPath', 'Ranking', 'NetworkSchema'. P1, P2, P3, P4, P5 and P6 respectively stand for papers 'NetClust', 'HeteSim', 'GenClus', 'PathSelClus', 'HeProjI', 'PathSim'. A1, A2, A3, A4 and A5 respectively stand for authors 'Yizhou Sun', 'Jiawei Han', 'Chuan Shi', 'Philip S. Yu'. 'Xifeng Yan'. V1, V2, V3 and V4 respectively stand for venues 'CIKM', 'TKDE', 'SIGKDD', 'VLDB'.