A Survey of Entity Similarity Measures on Heterogeneous Information Network

ZHU Zichen
Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
zczhu2@cs.hku.hk

MA Chenhao
Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
chma2@cs.hku.hk

ABSTRACT

According to recent studies, *heterogeneous information networks (HINs)* consisting of multiple types of entities and relations has shown its power in many disciplines, such as, computer science, social science, physics and so on. More and more researchers have noticed the importance of HIN analysis and many novel data mining tasks have been exploited in such networks, such as similarity search, clustering, and classification. Among those tasks, similarity measure on HINs, which is mainly to evaluate the similarity of entities, is the basis of many data mining tasks, such as clustering and classification. In this paper, we provide a survey of similarity measures on heterogeneous information networks. We will introduce basic concepts of heterogeneous information networks analysis, examine the recent developments on similarity measures and make a general evaluation of those similarity metrics.

1. INTRODUCTION

Recently, researchers use heterogeneous information networks (HINs) to model real world relationships in many applications, especially for real systems containing multi-typed interacting components. For instance, in bibliographic database, like DBLP[1] [5], papers are connected together via authors, venues and terms; and in Instagram[2], photos or videos are linked together via users, locations, hashtags and comments. Compared to widely-used homogeneous information networks [1, 2], which are extracted from real interacting systems by simply ignoring the heterogeneity of objects and links or only considering one type of relations among one type of objects, the heterogeneous information network can effectively fuse more information and contain rich and specific semantics in nodes and links [4].

Because of HIN's property of rich semantics and information, since the concept of heterogeneous information network and meta path proposed in 2009 [7] and 2011 [6], respectively, more and more researchers have noticed the importance of heterogeneouse information network analysis and many novel data mining tasks have been developed in such networks, such as similarity search [3, 6], clustering [8]. In other words, HIN analysis has become a hot topic rapidly in the fields of data mining, database and information retrieval, involving similarity measure, clustering, classification, link prediction, ranking, recommendation and information fusion on HINs [4].

Among those analysis tasks, similarity measure is the fundamental problem of network analysis, because most high level tasks, such as clustering and classification, need to evaluate the similarity of objects or relations. What's more, most of the state-of-the-art homogeneous network similarity methods, which generally assume the networks don't carry semantics, do not generalize well in HINs, due to HINs' rich semantics. Thus, some similarity metrics based on meta-path, which represents semantics in HINs, have been designed to evaluate the similarity between entities or relations in HINs, such as PathSim [6] and RelSim [10]. Maybe some figure of HIN and an example...

In this paper, we attempts to clearly introduce basic concepts in heterogeneous network analysis and make a comprehensive investigation on contemporary research developments of similarity metrics on HINs. Then, we tentatively provide a general evaluation over some of these state-of-the-art algorithms to give some reference on similarity measure choosing in high level network analysis tasks, such as clustering and classification.

The following part is organized as follows. Section 2 introduces the basic concepts and examples about HIN. Section 3 presents recent designed similarity measures on HIN. Experiments and evaluation are conducted in Section 4. Finally, Section 5 summarizes and concludes this paper.

2. BASIC CONCEPTS AND DEFINITIONS

In this section, we introduce some basic concepts about HIN and give some HIN examples. We first define the information network and heterogeneous information network.

An information network represents an abstraction of the real world, focusing on the entities and the relations among these entities.

[1]http://dblp.uni-trier.de
[2]http://instagram.com

COMP 9102 Surveys 2017, CS Dept, HKU

Definition 1. **Information Network** [5, 9]. An information network is defined as a directed graph $G = (V, E)$ with an entity type mapping function $\varphi : V \to \mathcal{A}$ and a relation type mapping function $\psi : E \to \mathcal{R}$. Each entity $v \in V$ belongs to one particular entity type in the entity type set $\mathcal{A} : \varphi(v) \in \mathcal{A}$, and each relation $e \in E$ belongs to a particular relation type in the relation type set $\mathcal{R} : \psi(e) \in \mathcal{R}$. If two relations belongs to the same relation type, the two links share the same starting entity type as well as the ending entity type.

Based on the definition of information network, we derive the definitions of heterogeneous/homogeneous information network.

Definition 2. **Heterogeneous/homogeneous information network**. The information network is called heterogeneous information network if the type of entities $|\mathcal{A}| > 1$ or the types of relations $|\mathcal{R}| > 1$; otherwise, it is a homogeneous information network.

For better understanding the entity types and relation types in a complex heterogeneous information network, the network schema provides a high-level description of a given heterogeneous information network.

Definition 3. **Network schema** [5, 9]. The network schema, denoted as $T_G = (\mathcal{A}, \mathcal{R})$, is a meta template for an information network $G = (V, E)$ with the entity type mapping $\varphi : V \to \mathcal{A}$ and the relation type mappling $\psi : E \to \mathcal{R}$, which is a directed graph defined over entity types $\mathcal{A}$, with edges as relations from $\mathcal{R}$.

The network schema of a HIN specifies type constraints on the sets of entities and relationships among those entities. An information network following a network schema is called a **network instance** of the network schema. For a relation type R connecting entity type S to entity type T, i.e., $S \xrightarrow{R} T$, S and T are the **source entity type** and **target entity type** of relation type R, which can be denoted as $R.S$ and $R.T$, respectively. The inverse relation R^{-1} holds naturally for $T \xrightarrow{R^{-1}} S$. Generally, R is not equal to R^{-1}, unless R is symmetric.

Another important concept, meta-path, is proposed to systematically define relations between entities at the schema level.

Definition 4. **Meta-path** [6]. A meta-path $\mathcal{P}$ is a path defined on a schema $S = (\mathcal{A}, \mathcal{R})$, 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 entities $A_1, A_2, \cdots, A_{l+1}$, where $\circ$ denotes the composition operator on relations.

For simplicity, we can also use entity types to denote the meta-path if there are no multiple relation types between the same pair of entity types: $\mathcal{P} = (A_1 A_2 \cdots A_{l+1})$. We say a concrete path $p = (a_1 a_2 \cdots a_{l+1})$ between entities a_1 and a_{l+1} in network G is a **path instance** of the relevance path $\mathcal{P}$, if $\forall a_i, \varphi(a_i) = A_i$ and $\forall e_i = \langle a_i, a_{i+1} \rangle, \psi(e_i) = R_i$ in $\mathcal{P}$. It can be denoted as $p \in \mathcal{P}$. A meta-path $\mathcal{P}$ is a **symmetric path**, if the relation R defined by it is symmetric (i.e., $\mathcal{P}$ is equal to $\mathcal{P}^{-1}$), such as APA. Two meta-paths $\mathcal{P}_1 = (A_1 A_2 \cdots A_l)$ and $\mathcal{P}_2 = (B_1 B_2 \cdots B_k)$ are **concatenable** if and only if A_l is equal to B_1, and the concatenated path is writen as $\mathcal{P} = (\mathcal{P}_1 \mathcal{P}_2)$, which equals to $(A_1 A_2 \cdots A_l B_2 \cdots B_k)$.

Example 1. Example and figure about HIN, schema, meta-path

3. SIMILARITY MEASURES

4. REFERENCES

[1] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 393–402. ACM, 2010.

[2] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.

[3] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu. Relevance search in heterogeneous networks. In *Proceedings of the 15th International Conference on Extending Database Technology*, pages 180–191. ACM, 2012.

[4] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):17–37, 2017.

[5] Y. Sun and J. Han. Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter*, 14(2):20–28, 2013.

[6] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003, 2011.

[7] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM, 2009.

[8] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):11, 2013.

[9] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 797–806. ACM, 2009.

[10] 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. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 621–629. SIAM, 2016.