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] [12], 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 [6, 7], 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 [11].

Because of HIN's property of rich semantics and information, since the concept of heterogeneous information network and meta path proposed in 2009 [14] and 2011 [13], 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 [10, 13], clustering [15]. 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 [11].

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 [13] and RelSim [18].

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

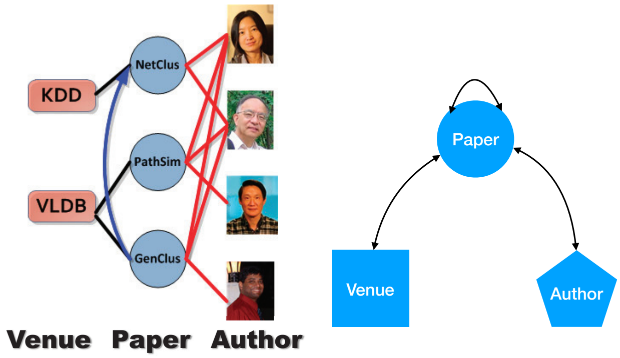
In this section, we introduce some basic concepts about HIN and give some HIN examples. We first define the information network and heterogeneous information network.

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

Definition 1. **Information Network** [12, 16]. An infor-

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

COMP 9102 Surveys 2017, CS Dept, HKU



(a) Network instance (b) Network schema

Figure 1: An example of heterogeneous information network on bibliographic data.

mation 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.

Example 1. Figure 1a shows a HIN example on bibliographic data [12]. A bibliographic information network, such as the bibliographic network shown in Figure 1a, is a typical HIN containing three types of entities: papers, venues and authors. For each paper, it has links to a set of authors, and a venue, and these relations belong to a set of relation types.

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** [12, 16]. 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.

Example 2. Figure 1a demonstrates the real entities and their relations on bibliographic data. Figure 1b shows the corresponding network schema of the HIN shown in Figure 1a. There are three different entity types: papers (P), authors (A) and venues (V). There are relations connecting different types of objects. The link types are defined by the relations between two entity types. For instance, links existing between authors and papers denote the writing or writen-by relations.

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

Definition 4. **Meta-path** [13]. 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 3. As the example shown in Figure 1, authors can be connected via different meta-paths, such as (A, P, A), i.e., "Author-Paper-Venue" path and (A, P, V, P, A), i.e., "Author-Paper-Venue-Paper-Author" path. It is obvious that semantics underneath these paths are different. The (A, P, A) path means authors collabrating on the same papers, while (A, P, V, P, A) path means authors publishing papers on the same venue.

The commuting matrix is defined in [13] to compute the frequencies of all the paths related to a meta-path.

Definition 5. **Commuting matrix**. Given a network $G = (V, E)$ and its network schema T_G, a commuting matrix $M_{\mathcal{P}}$ for a meta-path $\mathcal{P} = (A_1, A_2, \cdots A_{l+1})$ is defined as $M_{\mathcal{P}} = W_{A_1A_2}W_{A_2A_3} \cdots W_{A_lA_{l+1}}$, where $W_{A_iA_j}$ is the adjacency matrix between types A_i and A_j. $M_{\mathcal{P}}(i, j)$ represents the number of path instances between entities v_i and v_j, where $\varphi(v_i) = A_1$ and $\varphi(v_j) = A_{l+1}$, under mata-path $\mathcal{P}$.

3. SIMILARITY MEASURES

Similarity measure is to evaluate the similarity of entities. It is the basis of many data mining tasks, such as classification, clustering, and recommendation system. Similarity measure has been well studied on different kinds of data types for a long time. These studies can be roughly categorized into two types: **feature based approaches** and **link based approaches**. The feature based approaches measure the similarity of entities based on their feature/attribute values, such as cosine similarity, Jaccard similarity and Euclidean distance. The link based approaches measure the similarity of entities based on their link structures in a network. For instance, Personalized PageRank [5] evaluates the probability starting from a source object to a target object by randomly walking with restart, and SimRank[4] evaluates the similarity of two objects by their neighbors' similarities.

In recent years, similarity measures on heterogeneous information networks begin to be noticed by more and more researchers. Apart from the structure similarity addressed by most homogeneous similarity metrics, similarity metrics on HIN also need to take the meta-path connecting these two objects into account. As we know, there are different meta-paths connecting two objects, and these meta paths contain different semantics meanings, which may lead to different similarities. So, the similarity measure on HIN is meta-path constraint [11]. Example? We present the recent state-of-the-art similarity metrics in the following part.

3.1 PathSim

PathSim [13] is the first meta-path based similarity measure to evaluate the similarity of same-typed entities based on symmetric meta-paths.

Definition 6. **PathSim**: Given a symmetric meta-path $\mathcal{P}$, PathSim between two entities u and v of the same entity type is:

$$\begin{aligned}\text{PathSim}(u, v) &= \frac{2 \times |\{p_{u \rightsquigarrow v} \in \mathcal{P}\}|}{|\{p_{u \rightsquigarrow u} \in \mathcal{P}\}| + |\{p_{v \rightsquigarrow v} \in \mathcal{P}\}|} \\ &= \frac{2 \cdot M_{\mathcal{P}}(u, v)}{M_{\mathcal{P}}(u, u) + M_{\mathcal{P}}(v, v)}\end{aligned} \tag{1}$$

PathSim is extended to ExPathSim for evaluting the asymmetric meta-path based and different-typed entities similarity in [20].

Definition 7.

$$\text{ExPathSim}(u, v) = \frac{2 \cdot M_{\mathcal{P}}(u, v)}{\sum_{w=1}^{N} M_{\mathcal{P}}(u, w) + \sum_{w=1}^{N} M_{\mathcal{P}^{-1}}(v, w)} \tag{2}$$

3.2 Distant Meta-Path Similarity

Distant meta-path similarity[17] is designed to evaluate text-based meta-path similarity between two distant (relatively isolated) entities. Here, distant entities means those two entities can not connected by the given meta-path.

Definition 8. **Distant meta-path similarity**. The distant meta-path similarity between an entity pair describes the proximity of the pair's neighborhood entities. Neighborhood entities are defined as the entities kinked via meta-path(s) to the pair. Let $\{M_{\mathcal{P}}(u, w)\}_{w=1}^{N}$ denotes the meta-path instantces between entity u and its neighborhood entities. The distant meta-path similarity between u and v is the decided by the proximity of $\{M_{\mathcal{P}}(u, w)\}_{w=1}^{N}$ and $\{M_{\mathcal{P}}(v, w)\}_{w=1}^{N}$. Entities u and v are called as distant neighbors to each other.

There are 53 similarity metrics, i.e., metrics to measure the similarity between $\{M_{\mathcal{P}}(u, w)\}_{w=1}^{N}$ and $\{M_{\mathcal{P}}(v, w)\}_{w=1}^{N}$, tested in [17] to find the best way to define a distant meta-path similarity. Experimental results in [17] show cosine similarity is consistent good for general use. Thus, we present cosine similarity based distant meta-path similarity here.

$$\begin{aligned}&\text{DistantSim}(u, v) \\ &= \frac{\sum_{m=1}^{M} \sum_{w=1}^{N} M_{\mathcal{P}_m}(u, w) M_{\mathcal{P}_m}(v, w)}{\sqrt{\sum_{m=1}^{M} \sum_{w=1}^{N} M_{\mathcal{P}_m}(u, w)^2} \sqrt{\sum_{m=1}^{M} \sum_{w=1}^{N} M_{\mathcal{P}_m}(v, w)^2}}\end{aligned} \tag{3}$$

3.3 HeteSim

The similarity of objects with different are needed in many applications, such as recommendation system [3] and medicine annotation analysis [8]. Thus, HeteSim [9] is proposed for evaluating the similarity of entities with different types.

Before giving the definition of HeteSim, we first introduce the decomposition of meta-path.

Definition 9. **Decomposition of meta-path**. An arbitrary meta-path $\mathcal{P} = (A_1, A_2, \cdots, A_{l+1})$ can be decomposed into two equal-length path $\mathcal{P}_L$ and $\mathcal{P}_R$, i.e., $\mathcal{P} = \mathcal{P}_L\mathcal{P}_R$, where $\mathcal{P}_L = (A_1, A_2, \cdots, A_{mid-1}, B)$ and $\mathcal{P}_R = (B, A_{mid+1}, \cdots, A_{l+1})$. If l is even, $B = A_{mid}$. Otherwise, B is the middle type entity E between the atomic relation $A_{\frac{l+1}{2}}A_{\frac{l+1}{2}+1}$. The new path becomes $\mathcal{P}' = (A_1, \cdots, E, \cdots, A_{l+1})$, so B is also the middle item of $\mathcal{P}'$.

Obviously, for a symmetric path $\mathcal{P} = \mathcal{P}_L\mathcal{P}_R$, $\mathcal{P}_R^{-1}$ is equal to $\mathcal{P}_L$. After tranforming the orginal meta-path, when its length is odd, the definition of HeteSim can be expressed:

Definition 10. **HeteSim**. Given a relevance path $\mathcal{P} = (A_1, A_2, \cdots, A_{l+1})$, the HeteSim score between two entities u and v ($u \in A_1, v \in A_{l+1}$) is:

$$\text{HeteSim}(u, v) = \frac{\sum_{w=1}^{N} M_{\mathcal{P}_L}(u, w) \cdot M_{\mathcal{P}_R^{-1}}(v, w)}{\sqrt{\sum_{w=1}^{N} M_{\mathcal{P}_m}(u, w)^2} \sqrt{\sum_{w=1}^{N} M_{\mathcal{P}_m}(v, w)^2}} \tag{4}$$

3.4 RelSim

RelSim [18] is a meta-path based relation similarity measure. It measures the similarity between two relation instances based on the latent semantic relation (LSR): two relation instances are more similar when sharing more important (heavily weighted) meta-paths.

Definition 11. **RelSim**. Given an LSR (latent semantic relation), denoted as $\{w_m, \mathcal{P}_m\}_{m=1}^{M}$, RelSim between two relation instances $r = \langle v^{(1)}, v^{(2)} \rangle$ and $r' = \langle v^{(1)'}, v^{(2)'} \rangle$ is defined as:

$$\text{RelSim}(r, r') = \frac{2 \times \sum_m w_m \min(x_m, x'_m)}{\sum_m w_m x_m + \sum m x_m x'_m} \tag{5}$$

where x_m is the number of path instances between $v^{(1)}$ and $v^{(2)}$ in relation r following meta-path $\mathcal{P}_m$, and x'_m is the number of path instances between $v^{(1)'}$ and $v^{(2)'}$ in relation r' following meta-path $\mathcal{P}_m$. We use a vector $\mathbf{x} = [x_1, \cdots, x_m, \cdots, x_M]$ to charachterize a relation instance r, and a vector $\mathbf{w} = [w_1, \cdots, w_m, \cdots, w_M]$ to denote the corresponding weights. M is the number of meta-paths.

3.5 SignSim

To handle meta-paths with negative links in a signed heterogeneous information network, SignSim[19] is proposed to measure the relatedness of objects with different types in signed HIN, based on signed meta-path factorization. To understand what the meta-path factorization is, we also need to know the definition of atomic meta-path. The two definitions are given as follows,

Definition 12. **Atomic Meta-path**. Atomic meta-path is a minimum part of a signed meta=path that can be used to compute the similarity between objects of the same type.

Definition 13. **Meta-path Factorization**. Meta-path factorization splits a signed meta-path into multiple atomic meta-paths, and then the relatedness between objects from different types can be computed based on the signed meta-paths.

The main idea of SignSim can be expressed by the following three formulas.

$$P = P_1 P_2 \cdots P_i P_n \tag{6}$$

The first formula represents the decomposition of meta-path P, and $P_i (1 \leq i \leq n)$ is the atomic meta-path or redundant meta-path.

$$U_i = \bigcup_{o_{i-1} \in U_{i-1}} s(o_{i-1}, o_i | P_{i-1}) \; and \; U_1 = o_1 \tag{7}$$

In the second formula, U_i is the possible target node space of the node o_i, $s(o_{i-1}, o_i|P_{i-1})$ is the set of the object o_i positively related to object o_{i-1} based on the meta-path P_{i-1}.

$$\begin{aligned} &SignSim(o_1, o_n|P) \\ &= \sum_{o_2 \in U_2} \cdots \sum_{o_{n-1} \in U_{n-1}} sim(o_1, o_2|P_1) \cdots sim(o_{n-1}, o_n|P_{n-1}) \end{aligned} \tag{8}$$

In the last formula, $s(o_{i-1}, o_i|P_{i-1})$ is the relatedness between objects o_{i-1} and o_i based on meta-path P_{i-1}.

Next, an example about the relation between users and movies is given to show how to measure the similarity on atomic meta-path and redundant meta-path.

Atomic Meta-path:User-Movie-User. This meta-path contains two edges and both of them can be signed, which represents the user's rating toward a movie. The similarity between two users u_i and u_j can be obtained from the $n*m$ adjancy matrix $W_{UM} = [v_{ij}]_{n \times m}$ between users and movies. The m is the number of the network and n is the number of users. And we denote the row of the matrix with vector u^i in which $i (1 \leq i \leq n)$ is the row number. The similarity between the two users u_i and u_j is measured by cosine similarity based on vector space model. That is:

$$v_{ij} = \begin{cases} 1 & \text{user } u_i \text{ enjoys movie } m_j \\ -1 & \text{user } u_i \text{ doesn't enjoy movie } m_j \\ 0 & \text{user } u_i \text{ never see movie } m_j \end{cases} \tag{9}$$

$$sim(u_i, u_j|UMU) = cos(u^i, u^j) = \frac{u^i \cdot u^j}{|u^i| * |u^j|} \tag{10}$$

Redundant Meta-path:User-Movie. The redundant meta-path UM, containing a signed edge, describes the movie seen by the user. The relevance based on redundant meta-path can be computed directly without meta-path factorization. That is:

$$sim(u_i, m_j|UM) = \frac{v_{ij}}{|D_s(u_i|UM)| * |D_s(m_j)|UM|} \tag{11}$$

$$v_{ij} = \begin{cases} 1 & \text{user } u_i \text{ enjoys movie } m_j \\ -1 & \text{user } u_i \text{ doesn't enjoy movie } m_j \\ 0 & \text{user } u_i \text{ never see movie } m_j \end{cases} \tag{12}$$

$$D_s(u_i|UM) = \begin{cases} D_+(u_i|UM) & v_{ij} = 1 \\ D_-(u_i|UM) & v_{ij} = -1 \end{cases} \tag{13}$$

where $D_+(u_i|UM)$ is the neighbor set of node u_i based on positive edges of UM, and $D_-(u_i|UM)$ is the neighbor set of u_i based on negative edges of UM.

3.6 WsRel

WsRel [20] is a recently proposed relevance measure method, which can be applied in a weighted signed heterogeneous information network. It is devised to evaluate the relatedness of two objects with different types in a more complicated HIN. Specifically, it firstly transforms a signed network to a non-signed network, and then calculates relatedness of source object and target oobject based on various single meta-paths.

Sigmoid function $f(x) = \frac{1}{1+e^{-x}}$ is used to be the mapping function to implement the transformation from the origin signed network to a new non-signed network. After we obtain the new network, the definition of WsRel is given as follows:

Definition 14. **WsRel**. Given an meta-path $P = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, the relatedness of source object s and target object t following the meta-path P is:

$$\begin{aligned} &\mathrm{WsRel}(s, t|R_1R_2\cdots R_l) \\ &= \frac{1}{|O(s|R_1)|} \sum_{s' \in O(s|R_1)} w(s, s') \cdot WsRel(s', t|R_2R_3\cdots R_l) \end{aligned} \tag{14}$$

where $O(s|R_1)$ is the out-neighbors of s based on relation R_1, and $w(s, s')$ is reassigned weight value of the link between s and s'.

4. EXPERIMENTS

We implemented ExPathSim, WsRel, SignSim, HeteSim in Python 3 to perform a general evaluation on similarity calculation on objects with different types.Note that SignSim originally does not fit weighted HIN and thus we made some improvement on the SignSim to make it support weighted HIN. The new-version SignSim will be introduced later in this section.

4.1 Dataset

We used hetrec2011-movielens[1] dataset for evaluation. It contains 2K users, 10K movies, 4K directors, 21K actors and 20 genres. According to the previous similar experiment in WsRel [20], we then chosen the rst 5 main actors from the original dataset according to the orders of actors in the cast. Also, we manually gave the relation from the movie to the actor, according to her ranking in he case. For example, if the actor a_i ranks 4th in the movie m_j, the weight of the relation from m_j to a_i is $6 - 4 = 2$. By this approach, the higher ranking means more weight of the relation.

At the same time, the rating scores range from 0.5 to 5.0 on the links between users and movies, where a higher score means a stronger preference. However, because of the existence of individual differences, if we directly dene that raw scores larger than some values, such as 3.0, means users like movies, it is not very reasonable. So we need to normalize the rating scores, such that it can be compared directly between different distributions of raw scores, which can also generate some links with negative weight.
The standard score z_u of user u's raw score r_u can be calculated by

$$z_u = \frac{r_u - \mu_u}{\sigma_u} \tag{15}$$

where μ_u is the mean of the rating scores of user u, σ_u is the standard deviation of the rating scores of user u. After applying the normalization, the weighted signed heterogeneous information network can be constructed on this dataset, which contains five types of objects: user, movie, dirctor, genre and actor.

4.2 Metrics

To evaluate these four approaches of similarity calculation, we randomly chose her 50% rating records and the remaining are testing data. And then we use the 50% rating records and the whole data on other-type objects to construct the heterogeneous information network. Given a user, we firstly use her rated movies in the remaining data as the candidate movies, and then according to different similarity calculation methods, we generate a list of K movies named R_u, which are top-K movies among all the candidate movies. If the movie in R_u is also in the top-K rated movies in the candidate movies. We call it a hit. Since here the recall is the same as the precision, we here only use standard precision as the metric to evaluate the results.

4.3 New-version SignSim(NSignSim)

To make SignSim applicable in weighted signed HIN, we made a little improvement in the origin method. Similarly, here we use the example meta-path UMU and UM to explain how to calculate NSignSim.

Atomic Meta-path:User-Movie-User. As we've introduced, the similarity between two users u_i and u_j can be obtained from the $n * m$ adjancy matrix $W_{UM} = [v_{ij}]_{n\times m}$ between users and movies. Note that v_{ij} doesn't mean the same thing as before. In NSignSim, v_{ij} represents the user u_i's rating on the movie m_j. The m is the number of the network and n is the number of users. And again, we denote the row of the matrix with vector u^i in which $i(1 \le i \le n)$ is the row number. The similarity between the two users u_i and u_j can be measured by cosine similarity or Euclidean distance. That is:

$$sim(u_i, u_j|UMU) = cos(u^i, u^j) = \frac{u^i \cdot u^j}{|u^i| * |u^j|} \tag{16}$$

$$sim(u_i, u_j|UMU) = \frac{1}{1 + ||u^i - u^j||_2} \tag{17}$$

where $||\overrightarrow{x}||_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2}$

Redundant Meta-path:User-Movie. The redundant meta-path UM, containing a signed and weighted edge, describes the movie seen by the user. Again, v_{ij} here means the user u_i's rating towards the movie m_j. We didn't change the revalance based on redundant meta-path. But there is a noteworthy thing for the following equation:

$$D_s(u_i|UM) = \begin{cases} D_+(u_i|UM) & v_{ij} > 0 \\ D_-(u_i|UM) & v_{ij} < 0 \end{cases} \tag{18}$$

where $D_+(u_i|UM)$ is the neighbor set of node u_i based on positive edges of UM, and $D_-(u_i|UM)$ is the neighbor set

Table 1: Precision Rates of Four Methods under Different Meta-paths

Meta-path	ExPathSim	WsRel	NSignSim	HeteSim
U, M, D, M	0.4200	0.3600	0.4200	0.3200
U, M, A, M	0.2500	0.2200	0.2500	0.2000
U, M, G, M	0.3200	0.3600	0.3800	0.3400
U, M, U, M		0.2400	0.3400	

Table 2: Average Running Time (in minutes) of Four Methods under Different Meta-paths

Meta-path	ExPathSim	WsRel	NSignSim	HeteSim
U, M, D, M	0.0473	0.0010	0.1254	0.0037
U, M, A, M	0.3615	0.0051	0.2079	0.0224
U, M, G, M	22.2447	0.0017	0.1513	0.0729
U, M, U, M		0.1915	26.8311	

of u_i based on negative edges of UM. Actually we can also use $D_+(u_i|UM)$ represents the number of the users who has similar rating and $D_+(u_i|UM)$ represents the number of the movies that also has similar rating. In this experiment, to make it simpler, we only determine it by the weight's sign, just as the previous one did.

4.4 Experiment Results

We evaluate the efficience and effectiveness of HeteSim, WsRel, ExPathSim, NSignSim. We set $K = 10$. We run experiments on 10 users selected randomly under different meta-paths, then use the average running time and average precision rate to express the efficiency and effectiveness under this experiment set. Table 1 shows the precision rates of these 4 methods. From Table 1, we can find that NSignSim has best performance on this experiment. Table 2 contains the detail of these 4 methods' average running time, and result shows ExPathSim is the slowest method. WsRel and HeteSim have good efficiency according to this experiment.

5. CONCLUSION

In this survey, we introduce the basic definition about heterogeneous information network. Then, we provide detailed introduction about similarity measures on heterogeneous information network related to meta-path. Based on the similarity methods' properties and suitable ranges, we implemented four of these and run basic effectiveness and efficiency evaluation. But, our work still has some limitation: some meta-structure based similarity methods [2] are not included; due to the limit of dataset and time, we have not conducted other experiments on these methods, such as clustering and classification.

6. REFERENCES

[1] I. Cantador, P. L. Brusilovsky, and T. Kuflik. Second workshop on information heterogeneity and fusion in recommender systems (hetrec2011). 2011.

[2] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li. Meta structure: Computing relevance in large heterogeneous information networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1595–1604. ACM, 2016.

[3] M. Jamali and L. Lakshmanan. Heteromf: recommendation in heterogeneous information networks using context dependent factor models. In *Proceedings of the 22nd international conference on World Wide Web*, pages 643–654. ACM, 2013.

[4] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[5] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279. ACM, 2003.

[6] 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.

[7] 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.

[8] G. Palma, M.-E. Vidal, E. Haag, L. Raschid, and A. Thor. Measuring relatedness between scientific entities in annotation datasets. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 367. ACM, 2013.

[9] C. Shi, X. Kong, Y. Huang, S. Y. Philip, and B. Wu. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2479–2492, 2014.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] C. Wang, Y. Song, H. Li, Y. Sun, M. Zhang, and J. Han. Distant meta-path similarities for text-based heterogeneous information networks. 2017.

[18] 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.

[19] M. Zhu, T. Zhu, Z. Peng, G. Yang, Y. Xu, S. Wang, X. Wang, and X. Hong. Relevance search on signed heterogeneous information network based on meta-path factorization. In *International Conference on Web-Age Information Management*, pages 181–192. Springer, 2015.

[20] T. Zhu, Z. Peng, S. Wang, S. Y. Philip, and X. Hong. Measuring the relevance of different-typed objects in weighted signed heterogeneous information networks. In *Computer Supported Cooperative Work in Design (CSCWD), 2017 IEEE 21st International Conference on*, pages 556–561. IEEE, 2017.