Relevance Search on Signed Heterogeneous Information Network Based on Meta-path Factorization

Min Zhu[1], Tianchen Zhu[1], Zhaohui Peng[1(✉)], Guang Yang[1], Yang Xu[1], Senzhang Wang[2], Xiangwei Wang[3], and Xiaoguang Hong[1]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
zm13181@163.com, ztc1319@sina.cn, {pzh,hxg}@sdu.edu.cn, loggyt@yeah.net, zzmylq@gmail.com

[2] School of Computer Science and Engineering, Beihang University, Beijing, China
szwang@buaa.edu.cn

[3] State Grid Shandong Electric Power Company, Jinan, China
shandongwangxw@163.com

Abstract. Relevance search is a primitive operation in heterogeneous information networks, where the task is to measure the relatedness of objects with different types. Due to the semantics implied by network links, conventional research on relevance search is often based on meta-path in heterogeneous information networks. However, existing approaches mainly focus on studying non-signed information networks, without considering the polarity of the links in the network. In reality, there are many signed heterogeneous networks that the links can be either positive (such as trust, preference, friendship, etc.) or negative (such as distrust, dislike, opposition, etc.). It is challenging to utilize the semantic information of the two kinds of links in meta-paths and integrate them in a unified way to measure relevance.

In this paper, a relevance search measure called SignSim is proposed, which can measure the relatedness of objects in signed heterogeneous information networks based on signed meta-path factorization. SignSim firstly defines the atomic meta-paths and gives the computing paradigm of similarity between objects with the same type based on atomic meta-paths, with collaborative filtering using positive and negative user preferences. Then, on basis of the combination of different atomic meta-paths, SignSim can measure the relatedness between objects with different types based on multi-length signed meta-paths. Experimental results on real-world dataset verify the effectiveness of our proposed approach.

Keywords: Relevance search · Signed heterogeneous information network · Meta-path factorization

1 Introduction

Heterogeneous information networks are logical networks involving multiple typed objects and multiple typed links denoting different relationships, such as bibliographic networks, social media networks, and the knowledge network encoded in Wikipedia [1, 19]. In many heterogeneous information networks, the links could have positive or

J. Li and Y. Sun (Eds.): WAIM 2015, LNCS 9098, pp. 181–192, 2015.
DOI: 10.1007/978-3-319-21042-1_15

negative polarity denoting the positive or negative views and opinions of people. For instance, in Epinions network, consumers rate the products expressing viewpoints of likes or dislikes; in Slashdot Zoo, users can tag others as friends or foes. Such activities constitute the meaningful signed heterogeneous information networks, where links can be positive (“like”, “trust”) or negative (“dislike”, “distrust”).

In recent years, relevance or similarity search on heterogeneous information networks has attracted remarkable research interest [6, 11, 12, 13, 14]. By considering different linkage paths in the network, one could derive various relevance semantics, therefore, conventional work on relevance search usually take advantage of the various meta-paths which may contain different relevance semantics between objects. A meta-path is a path consisting of a sequence of relationships defined between different types of objects, namely the structural paths at the meta level. However, conventional research mainly focuses on non-signed information network, without considering the polarity of the links in the network.

The relevance could be computed simply based on meta-path in non-signed information networks. However, it is challenging to define the semantic information of meta-paths with negative links in a signed heterogeneous information network, especially when several different negative links exist simultaneously in one path, making the semantic more ambiguous. Directly utilizing previous meta-path-based methods to relevance search in signed heterogeneous networks might get undesirable or even totally opposite results. Therefore, it is challenging to model the meta-paths with both negative and positive links for relevant search in signed heterogeneous information networks.

In this paper, we propose a novel relevance search approach called SignSim to measure the relatedness of objects with different types in signed heterogeneous information network based on signed meta-path factorization. SignSim firstly defines the atomic meta-paths, based on which the computing paradigm of the similarity between objects with the same type, with collaborative filtering using positive and negative user preferences. Then, on basis of the combination of different atomic meta-paths, SignSim can measure the relatedness between objects with different types based on multi-length signed meta-paths.

The main contributions of this paper can be summarized as follow. (1)For the first time, to the best of our knowledge, we investigate the problem of relevance search in signed heterogeneous information networks. (2)A novel signed meta-path factorization based approach named SignSim for relevance search in signed heterogeneous information networks is proposed. To measure the relatedness between objects of different types, SignSim can effectively capture the positive and negative link semantics along the meta-paths. (3)Experiments on real-world dataset demonstrate the effectiveness of SignSim, by comparison with state-of-the-art approaches on non-signed heterogeneous information network.

The rest of the paper is organized as follows. We introduce the concepts and definitions in Section 2. The details of SignSim are given in Section 3. In section 4, extensive experiments on real data are performed to evaluate the effectiveness of our method. We discuss related work in Sections 5, and conclude the study in Section 6.

2 Problem Statement

In this section, we introduce some concepts and notations to help us state the studied problem.

Definition 1. (Signed Heterogeneous Information Network): A signed heterogeneous information network is an undirected graph G =(V, E) with an object type mapping function τ: V→A and link type mapping function φ: E→R and |A|>1, |R|>1. Each object v∈V belongs to a particular object type τ(v)∈A and each link e∈E belongs to a particular relation φ(e)∈R. For each link e∈E, it has polarity s(e)∈{Positive, Negative}.

As an example, a toy IMDB network is given in Fig.1. It is a typical signed heterogeneous information network, containing five types of objects: users (U), movies (M), genres (G), directors (D), and actors (A).

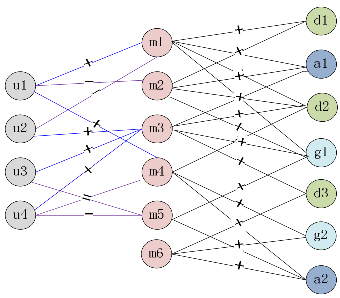

Fig. 1. A signed IMDB network

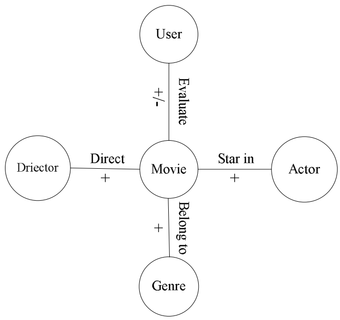

Fig. 2. The signed network schema of IMDB

Definition 2. (Signed Network Schema): The signed network schema, denoted as T_G=(A, R), is a meta-level template for a signed heterogeneous network G=(V, E) with the object type mapping τ: V→A and the link type mapping φ: E→R. In network schema, the nodes are the types of objects, and the edges are relationships between types, guiding the exploration of the semantics in the network.

For the IMDB network defined in Fig.1, the network schema is shown in Fig.2. The links between users and movies denote the rating or rated-by relationships with positive or negative polarities, while the links between movies and other types of entities only have the positive polarity.

Definition 3. (Signed Meta-path): The signed meta-path is a directed path defined on the signed network schema T_G = (A, R) of a graph G. A meta-path can be formally denoted as $\mathcal{P} = A_0 \xrightarrow{R_1} A_1 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_L$, which is a meta-level description of a path instance between two objects with type A_0 and type A_l respectively, usually abbreviated as $R_1R_2 \cdots R_L$ or $A_0A_1 \cdots A_L$.

Definition 4. (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 5. (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.

For instance, U (users)–>M (movies)–>U (users)–>M (movies) is one of the meta-path of the network schema shown in Fig.2. It denotes that two users have rated the same movie, and the second user also rated another movie. To get the relatedness between users and movies based on meta-path UMUM, SignSim uses meta-path factorization and splits the meta-path into UMU and UM. UMU is an atomic meta-path, based on which the similarity between users can be measured. UM is not an atomic meta-path, but based on UM the relevance between users and movies can be computed directly. Then SignSim combines the computing results and gets the final relevance.

Problem Statement. (Relevance Search on Signed Heterogeneous Network): In a signed heterogeneous information network G=(V, E), for a given node s $\in$ V and target type T, how to find the nodes in V of type T which are most relevant to s.

3 SignSim: A Relevance Measure

In this section, we elaborate a signed meta-path-based relevance framework and a novel relevance measure under this framework, SignSim. Taking Fig.1 for example, SignSim calculates the relatedness between users and movies based on multiple signed meta-paths.

3.1 Framework

Due to the signed meta-path has positive or negative polarity, we cannot directly use the meta-path-based relevance measure to compute the relevance between objects from different types in the signed heterogeneous information network. Therefore, SignSim first splits the signed meta-path, getting computable atomic meta-paths, and then combines the computing results.

SignSim mainly has two points: (1) It splits the meta-path into multiple atomic meta-paths, and computes the similarity between objects of the same type based on the atomic meta-paths. (2) It determines the possibility space of the target nodes, and computes the relevance between objects of different types based on the combination of the similarity computed by various atomic meta-paths.

The signed meta-paths can be obtained by traversing on the network schema using BFS (breadth first search) [7]. Then the atomic meta-paths can be got by meta-path factorization. In meta-path factorization, we always find the first match to the atomic meta-path, and then repeat the above procedure on the remaining meta-path. We separate out the mismatch path and call them **redundant meta-path**. For example, from the IMDB network schema shown in Fig.2, we can get the meta-path $\mathcal{P} = U \xrightarrow{R_1} M \xrightarrow{R_2} A \xrightarrow{R_3} M \xrightarrow{R_4} A \xrightarrow{R_5} M$. It can be decomposed to redundant meta-path $\mathcal{P}_1 = U \xrightarrow{R_1} M$, atomic meta-path $\mathcal{P}_2 = M \xrightarrow{R_2} A \xrightarrow{R_3} M$ and $\mathcal{P}_3 = M \xrightarrow{R_4} A \xrightarrow{R_5} M$. The algorithm of meta-path factorization is shown in Algorithm 1.

Algorithm 1. Meta-path Factorization

Input: $\mathcal{P}$: the meta-path; **m**: the number of nodes of the meta-path

Output : $\mathcal{P}_i$: the atomic meta-path

```
j ← 1, k ← 2, k' ← 1
while k<=m do
  for i=k'; i<k; i++ do
    if the i-th node and the k-th node of meta-path are the same then
      k' ← k
      output path between two nodes as the atomic meta-path
    end if
  end for
  k ← k+1
end while
output all remaining paths as the redundant meta-path
```

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

$$\mathcal{P} = \mathcal{P}_1\mathcal{P}_2\ldots\mathcal{P}_i\ldots\mathcal{P}_n \tag{1}$$

Formula (1) represents the decomposition of meta-path $\mathcal{P}$, and $\mathcal{P}_i$ (1 <= i <= n) is the atomic meta-path or redundant meta-path.

$$U_i = \bigcup_{o_{i-1}\in U_{i-1}} s(o_{i-1}, o_i \mid \mathcal{P}_{i-1}) \ \ and \ \ U_1 = o_1 \tag{2}$$

In formula (2), U_i is the possible target node space of the node o_i, $s(o_{i-1}, o_i \mid \mathcal{P}_{i-1})$ is the set of the object o_i positively related to object o_{i-1} based on meta-path $\mathcal{P}_{i-1}$.

$$signsim(o_1, o_n \mid \mathcal{P}) = \sum_{o_2\in U_2}\sum_{o_3\in U_3}\ldots\sum_{o_{n-1}\in U_{n-1}} sim(o_1, o_2 \mid \mathcal{P}_1)\ldots sim\,(o_{n-2}, o_{n-1} \mid \mathcal{P}_{n-2}) sim\,(o_{n-1}, o_n \mid \mathcal{P}_{n-1}) \tag{3}$$

In formula (3), $sim\,(o_{i-1}, o_i \mid \mathcal{P}_{i-1})$ is the relatedness between objects o_{i-1} and o_i based on meta-path $\mathcal{P}_{i-1}$.

3.2 Atomic Meta-path-Based Similarity Measure

In this section, taking IMDB network as an example, we present how to construct atomic meta-paths and measure the similarity between objects of the same type. Based on Algorithm 1, we find out three kinds of atomic meta-paths.

Atomic Meta-path: UMU. The atomic meta-path UMU describes the two users seeing the same movie. It contains two edges and both of them are signed. The similarity between the two users is measured based on this atomic meta-path.

We construct an $n*m$ adjacency matrix $W_{UM} = \left[v_{ij}\right]_{n\times m}$ between users and movies in the network, denoting the degree of user's preference to the movie. The m is the number of movies and n is the number of users. We denote the row of the matrix with vector u^i, in which i(1<=i<=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:

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

$$sim(u_i, u_j \mid UMU) = \cos(u^i, u^j) = \frac{u^i * u^j}{|u^i| * |u^j|} \tag{4}$$

If $sim(u_i, u_j \mid UMU) \geq \zeta$, the user u_i is considered as similar to the user u_j. ζ is a threshold parameter , which will be discussed in the experiments.

Atomic Meta-path: MUM. The atomic meta-path MUM describes the two movies seen by the same user. It contains two edges and both of them are signed. The similarity between two movies is measured based on the atomic meta-path.

We can measure the similarity between two movies based on the same adjacency matrix as presented above. We denote the row of matrix with vector m^i. The similarity measure between movies m_i and m_j is:

$$sim(m_i, m_j \mid MUM) = \cos(m^i, m^j) = \frac{m^i * m^j}{|m^i| * |m^j|} \tag{5}$$

Atomic Meta-path: MPM. The atomic meta-path MPM describes the two movies having common attributes (genres, directors, or actors). Here **P** represents **G** (genres), **D** (directors), and **A** (actors). It contains two edges and neither of them is signed. It is used to represent the relationship between movies and their various properties. The similarity between two movies can be measured by this atomic meta-path, just like formula (4) and (5).

3.3 Signed Meta-path-based Relevance Measure

For non-atomic meta-paths, we measure the relevance between objects by signed meta-path factorization. Different paths represent different semantics, thus path selection is essential to measure the relevance between objects from different types. The length of the meta-path may also affect the performance, and the previous works have shown that the accuracy will decrease as the length increases [6, 14]. In this section, we only study the meta-paths whose length is less than five.

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

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

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

$$D_s(u_i|UM) = \begin{cases} D_+(u_i|UM) & v_{ij}=1 \\ D_-(u_i|UM) & v_{ij}=-1 \end{cases} \tag{6}$$

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.

Meta-path: UMUM. The meta-path UMUM contains three edges and all of them are signed. For measuring the relevance between users and movies based on UMUM, we split it into one atomic meta-path UMU and one redundant meta-path UM, and compute the similarity or relevance respectively, then combine the computing results. Intuitively, the factorization means we find out the most similar users to the source user, and then we collect a set of favorite movies of these most similar users to the source user. The procedure can be formally depicted as follows:

$$U_2 = s(u,u_2 \mid UMU)$$
$$signsim(u,m \mid UMUM) = \sum_{u_2 \in U_2} sim(u,u_2 \mid UMU) sim(u_2,m \mid UM) \tag{7}$$

where $s(u,u_2 \mid UMU)$ is the set of users that u is similar to, and $sim(u_2,m \mid UM)$ is the relatedness between u_2 and m based on meta-path UM.

Meta-path: UMUMUM. This meta-path contains five edges and all of them are signed. For measuring the relevance between users and movies, we split the meta-path UMUMUM into UMU, UMU, and UM. Similarly, we first find out a set U_2 of most similar users with the source user based on meta-path UMU. Then, we find out a set U_3 of most similar users with users of set U_2 based on meta-path UMU. Finally, we collect a set of favorite movies of these users of set U_3. The procedure can be formally depicted as follows:

$$U_2 = s(u,u_2 \mid UMU)$$
$$U_3 = \bigcup_{u_2 \in U_2} s(u_2,u_3 \mid UMU)$$
$$signsim(u,m \mid UMUMUM) = \sum_{u_2 \in U_2} \sum_{u_3 \in U_3} sim(u,u_2 \mid UMU) sim(u_2,u_3 \mid UMU) sim(u_3,m \mid UM) \tag{8}$$

Meta-path: UMPM. This atomic meta-path contains three edges and one of them is a signed edge. For measuring the relevance between users and movies, we split this meta-path into one redundant meta-path and one atomic meta-path, which are UM and MPM.

Meta-path: UMPMPM. This meta-path contains five edges and one of them is signed. For measuring the relevance between users and movies, we could split it into one redundant meta-path and two atomic meta-paths, which are UM, MPM and MPM.

Meta-path: UMPMUM. This meta-path contains five edges and of which three are signed. For measuring the relevance between user and movie, we could split this meta-path into three meta-paths, which are UM, MPM, and MUM.

Meta-path: UMUMPM. This meta-path contains five edges and of which three are signed. For measuring the relevance between user and movie, we could split it into three meta-paths, which are UMU, UM, and MPM.

4 Experiments

4.1 Experimental Setup

Dataset. In this section, we evaluate the effectiveness of the proposed SignSim approach with comparison to existing relevance search algorithms on the real dataset hetrec2011-movielens [18]. The dataset contains 2,113 users, 10,197 movies, 20 movie genres, 4,060 directors, and 11,019 actors. We extract the first two actors according to the actor ranking of movies. On average, there are 2.04 genres, 2 actors and 1 director per movie. By calculating the average ranking of movies given by users, a signed heterogeneous information network can be built on this dataset which contains five types of objects: user, movie, director, actor and genre.

Evaluation Metric. To evaluate our approach, we first sort all the rating records in the dataset according to the timestamp. For each user, the latest 30% movies he or she has rated are selected as testing data and the old ones are selected as training data. When conducting relevance search, we generate a list of *K* movies named *Ru* for each user *u*. If the testing movie appears in the result list, we call it a hit. The mean hit of movie rating can be calculated as follows:

$$MR_u = \frac{\sum_{m \in R_u} I(m \in T_u) r(u,m)}{\sum_{m \in R_u} I(m \in T_u)}$$

where I(·) is an indicator function, R_u is a set of top-*K* movies recommended to user *u*, T_u is the set of testing movies of user *u*, *r(u,m)* is the rating on movie *m* by user *u*. We use MR_u to evaluate the relevance search results. We select the following approaches as baselines:

MatrixCal: MatrixCal is the naïve algorithm to calculate relevance between users and movies. User's preference matrix was calculated to represent the preference to the movie features (such as directors, actors, genres). The feature matrix of movies was also obtained. Then the preference matrix and feature matrix were multiplied together.
HeteSim: HeteSim[14] can measure the relatedness of objects with the same or different types in heterogeneous networks. The relatedness of object pairs is defined based on the search path that connects two objects through following a sequence of node types.

4.2 Experimental Results

Case Study. In the following experiments, we set parameters ζ=0.1 and *K*=5 that ζ is the similarity threshold of objects in the same type and *K* is a parameter to control the number of movies recommended, and single out a user with ID 7815 as the source user.

Table 1 shows the top *K* search results of SignSim with varied meta-paths. We compare MR_u of SignSim with two baselines, MatrixCal and HeteSim, and show the results in Table 2. In both tables, the bold items mean the searched movies are hit, and the figures in brackets denote the score rated by the user.

As we can see in Table 1, the search results based on short meta-paths are much better than those based on long meta-paths, so the length of meta-path is critical to measure the relevance of objects. Table 2 shows the results with different algorithms. The results given by SignSim are better than the two baselines.

Table 1. The results of different signed meta-paths in SignSim

UMUM	UMPM	UMPMUM	UMUMUM
The Usual Suspects (5.0)	**Sicko (4.5)**	Red Dust	The Shawshank Redemption
The Shawshank Redemption	Heavenly Creatures	Ruby Cairo	**The Usual Suspects (5.0)**
American Beauty (4.5)	King Kong	The Bad and the Beautiful	**American Beauty (4.5)**
Shichinin no samurai (3.5)	JLG/JLG -autoportrait de décembre	Arabian Nights	**Shichinin no samurai (3.5)**
Trainspotting	The Endurance: Shackleton's Legendary Antarctic Expedition	Small Faces	Monty Python and the Holy Grail

Table 2. The results of different approaches

SignSim	MatrixCal	HeteSim
The Usual Suspects (5.0)	This Is Spinal Tap	**Sicko (4.5)**
The Shawshank Redemption	Finding Neverland	Star Wars: Episode I - The Phantom Menace
American Beauty (4.5)	Casanova	Super Size Me
Shichinin no samurai (3.5)	Recount	Unprecedented: The 2000 Presidential Election
Trainspotting	The Bourne Supremacy	Believers

Quantitative Comparison. In this section, we quantitatively compare the proposed SignSim with two baselines. Fig.3 shows the results of different algorithms. One can see that the average rating for all users calculated by SignSim is higher than that by HeteSim and MatrixCal.

As mentioned above, the length of meta-path can significantly affect the relevance search performance. In SignSim, the accuracy of the relevance decreases as the meta-path length increases. Fig.4 shows the effect of different length of meta-paths.

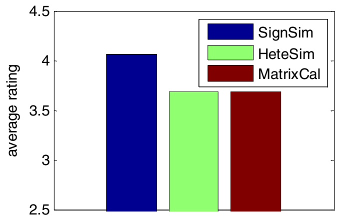

Fig. 3. The results of different algorithms

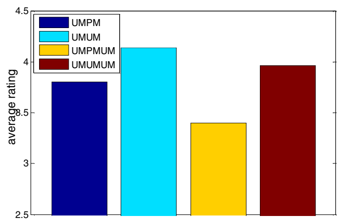

Fig. 4. The results of various meta-paths

Parameter Analysis. In this section, we will study the effect of parameter *K* and threshold ζ. *K* is used to control the amount of movies that we recommend to users and ζ is used to control the similarity between the same types of objects.

Fig.5 shows the results with various *K* values. One can see from Fig.5 that the performance shows a decrease trend with the increase of parameter *K*, which means a larger *K* can hurt the accuracy.

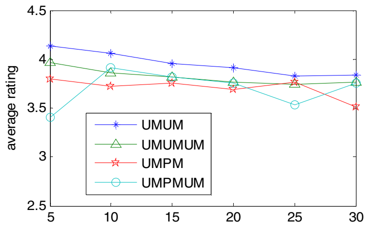

Fig. 5. The effect of parameter *K*

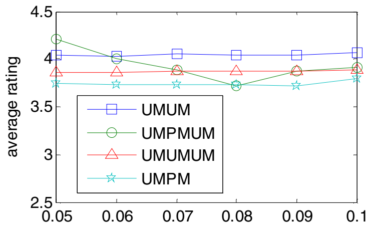

Fig. 6. The effect of parameter ζ

Fig.6 shows the performance of SignSim with the varied threshold parameters. One can see from Fig.6 that the performance of SignSim increases as the ζ increases for the meta-paths of UMUM, UMPM and UMUMUM, which means a higher similarity between the same type of objects helps to improve the final search accuracy, although the growth trend is not significant. As for UMPMUM before ζ=0.08, the performance decrease as the ζ increases and after ζ=0.08, the performance increase slowly. On the whole, SignSim can get a fairly better performance when ζ=0.06.

As to meta-path UMUMPM and UMPMPM, due to the space limit, we do not describe them in detail here. Their results are similar to those of other meta-paths.

5 Related Work

Relevance search on information network has gained wide attentions from researchers in link prediction [1, 7, 8, 9, 10], clustering [3], similarity query [6], text mining [2], etc. The most related work to relevance search is similarity search. Similarity calculation is used to measure the degree of similarity between two nodes, commonly used in data mining and natural language processing. Research on similarity search has made many significant achievements in recent years. Without considering the polarity of the edge in heterogeneous information network, i.e. links are all positive, similarity search broadly divided into two types: feature-based approach and link-based approach.

Feature-based measurement method is based on the characteristics of the object. Vector space model (VSM) is the most widely used similarity calculation model, in which objects are represented as vectors. However, feature-based approaches do not take the links between objects into account, so they cannot apply to network data.

Link-based approach is based on topological similarity of the object. SimRank [4] is a general algorithm determined only by the similarity of structural context, where two objects are similar if their neighbors are similar. Due to its relatively high computational complexity, most of the follow-up studies focus on how to improve the efficiency of the algorithm [17]. SCAN [2] is a structure clustering algorithm, which calculates the similarity between two objects through comparing their neighbor sets, while [6] proposed a meta-path-based similarity measure PathSim, which considers that objects are not only share similar visibility in the network but also strongly connected with each other. However, this approach just considers the objects with the same type. Different from the above study, HeteSim [14] is able to measure the correlation between two objects of different types based on arbitrary search path in heterogeneous information network. However, all of these methods don't consider the link polarity, i.e., their studies are built on a fundamental assumption that all edges are positive.

There are some works [5, 13, 15, 16] discussing the trust prediction task in social networks which have both positive and negative links. The majority of those studies focus on homogeneous networks which have only one type of objects and one type of links. However, many networks in the real world, such as Epinions, Slashdot Zoo, and Wikipedia, are both signed and heterogeneous composed of multiple types of objects and multiple types of links. Therefore it is necessary to study new methods to address knowledge discovery in signed heterogeneous network.

6 Conclusion and Future Work

In the paper, we study the relevance search problem in signed heterogeneous information networks. A novel relevance search approach called SignSim is proposed. Based on signed meta-path factorization, SignSim is able to find the most relevant objects with various types to a given object in the signed heterogeneous information networks. SignSim first splits the signed meta-path into multiple atomic meta-paths. By utilizing the atomic meta-paths, SignSim next proposes to measure the similarity between objects of the same type. Finally, SignSim integrates various similarity results to measure the relatedness between objects from different types based on non-atomic signed meta-paths. Experimental results on a real-world dataset verify the effectiveness of the approach. In this paper, we only study the relevance based on individual signed meta-paths. Our future work will consider the combination of different signed meta-paths to get global relevance.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No.61303005, No.61170052), the Natural Science Foundation of Shandong Province of China (Grant No.ZR2013FQ009), and the Technology Project of State Grid (Grant No.2012GWK515).

References

1. Sun,Y., Han, J., Yan, X., Yu, P.: Mining knowledge from interconnected data: a heterogeneous information network analysis approach. In: Proceedings of the VLDB Endowment, pp. 2022–2023. ACM Press, New York (2012)

2. Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.J.: SCAN: a structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 824–833. ACM Press, New York (2007)
3. Lao, N., Cohen, W.: Fast query execution for retrieval models based on path constrained random walks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 881–888. ACM Press, New York (2010)
4. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543. ACM Press, New York (2002)
5. Ye, J., Cheng, H., Zhu, Z., Chen, M.: Predicting positive and negative links in signed social networks by transfer learning. In: WWW, Switzerland (2013)
6. Sun, Y., Han, J., Yan, X., Yu, P., Wu, T.: Pathsim: meta path-based Top-k similarity search in heterogeneous information networks. In: PVLDB, pp. 992–1003 (2011)
7. Sun, Y., Han, J., Aggarwal, C., Chawla, N.V.: When will it happen? : relationship prediction in heterogeneous information networks. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 663–672. ACM Press, New York (2012)
8. Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C., Han, J.: Co-author relationship prediction in heterogeneous information networks. In: Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 121–128. IEEE Computer Society Press, Washington (2011)
9. Sun, Y., Norick, B., Han, J., Yan, X., Yu, P.S., Yu, X.: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1348–1356. ACM Press, New York (2012)
10. Tong, H., Faloutsos, C., Pan, J.Y.: Fast random walk with restart and its applications. In: Proceedings of the Sixth International Conference on Data Mining, pp. 613–622. IEEE Computer Society Press, Washington (2006)
11. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol., 1019–1031 (2007). John Wiley & Sons, Inc. Press
12. Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 243–252. ACM Press, New York (2010)
13. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web, pp. 641–650. ACM Press, New York (2010)
14. Shi, C., Kong, X., Huang, Y., et al.: HeteSim: a general framework for relevance measure in heterogeneous networks. In: IEEE Transactions on Knowledge and Data Engineering, pp. 2479–2492. IEEE Computer Society, Washington (2014)
15. Symeonidis, P., Tiakas, E., Manolopoulos, Y.: Transitive node similarity for link prediction in social networks with positive and negative links. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp.183–190. ACM Press, New York (2010)
16. DuBois, T., Golbeck, J., Srinivasan, A.: Predicting trust and distrust in social networks. In: Agent and Multi-Agent Systems: Technologies and Applications, pp. 122–131. Springer Berlin Heidelberg Press, Heidelberg (2011)
17. Sun, L., Cheng, R., Li, X., Cheung, D.W., Han, J.: On Link-based similarity join. In: Proceedings of 37th International Conference on Very Large Data Bases (2011)
18. hetrec2011-movielens. http://grouplens.org/datasets/hetrec-2011/
19. Wang, S.Z., Yan, Z., Hu, X., Yu, P.S., Li, Z.J.: Burst time prediction in cascades. In: Proceedings of 29th AAAI Conference on Artificial Intelligence (2015)