Exploiting Transitive Similarity and Temporal Dynamics for Similarity Search in Heterogeneous Information Networks

Jiazhen He[1,2], James Bailey[1,2], and Rui Zhang[1]

[1] Department of Computing and Information Systems, The University of Melbourne
[2] Victoria Research Laboratory, National ICT Australia

Abstract. Heterogeneous information networks have attracted much attention in recent years and a key challenge is to compute the similarity between two objects. In this paper, we study the problem of similarity search in heterogeneous information networks, and extend the meta path-based similarity measure *PathSim* by incorporating richer information, such as transitive similarity and temporal dynamics. Experiments on a large DBLP network show that our improved similarity measure is more effective at identifying similar authors in terms of their future collaborations.

Keywords: similarity search, heterogeneous network, meta path

1 Introduction

Heterogeneous information networks are ubiquitous in many real-world applications, such as bibliographic networks and healthcare networks. Different from homogeneous information networks (which only consider one type of object and link), heterogeneous information networks involve multiple types of objects and links. For example, heterogeneous bibliographic networks contain authors as well as other types of objects, such as papers, venues, and terms. In addition, heterogeneous information networks contain rich semantic information. For example, two objects can be connected through different links with different semantic meanings (i.e. two authors can be connected by co-authoring a paper or publishing different papers on a same venue). Such networks can more accurately model complex network data.

Heterogeneous information networks have been studied in many data mining tasks [6, 16, 15]. In this paper, we focus on the problem of similarity search in these networks. Similarity search aims to discover the most relevant objects with respect to a given query object. In heterogeneous information networks where multiple types of objects are available, we focus on identifying similar objects of the same type considering rich semantic information. For example, in a heterogeneous bibliographic network, given a query author, we can discover similar authors based on the diversified semantic meanings, such as co-author relationships and venues of publication.

Intuitively, two objects are similar if there many paths between them. A major challenge for similarity search in heterogeneous information networks is how to exploit the diversified semantic meanings under different paths. Existing similarity measures for *homogeneous information networks* cannot effectively capture such meanings since they treat all the paths between two objects equally without distinguishing the different semantic meanings. Some existing studies have recognised this problem and tackled similarity search in *heterogeneous information networks* based on the concept of meta paths[19, 14]. A *meta path* is a sequence of links between object types, which can capture a particular semantic meaning between its starting type and ending type. The meta path-based similarity measures treat the concrete paths following a given meta path equally. However, the impacts of the paths connected through different objects can vary. The challenge is how to model such impacts. In addition, heterogeneous information networks evolve over time, and contain rich temporal information. For example, the link between two objects is generally formed with a timestamp. The challenge is how to exploit this temporal information for similarity search.

In this paper, we extend the meta path-based similarity measure $PathSim$[14] by incorporating transitive similarity and temporal information. A meta path can be concatenated by multiple short meta paths. Given a meta path, we first decompose it into multiple short meta paths with the start type and end type of the same type. For example, meta path "*author-paper-author-paper-author* " ($APAPA$) describing two authors share same co-authors can be decomposed into two meta paths APA and APA. Then we add weights to the paths following a short meta path, according to the similarity between the two end objects of the short meta path, which is called transitive similarity. The transitive similarity between two objects can be obtained based on the different meta paths between them with different semantic meanings. The higher the transitive similarity between two objects, the more important the paths between them. For example, suppose two end authors x and y of $APAPA$ are connected through two common co-authors z_1 and z_2, if z_1 is more similar to x and y compared with z_2, the paths between x and z_1, and the ones between y and z_1 should be more important.

In addition, the paths between two objects are generally associated with temporal information, i.e., the building time. Intuitively, the recent paths should be more important than old ones. The paths are generally built as a result of an event. For example, the path "$Tom - P_1 - SIGKDD$" with building time 2012 following the meta path "*author-paper-venuer*" is built due to the event that Tom published paper P_1 in $SIGKDD$ in 2012. To differentiate the importance of different paths, we first decompose a meta path into multiple short meta paths with the maximum length that an event can affect, for example, meta path "*author-paper-venue-paper-author* " can be decomposed into "*author-paper-venuer*" and "*venue-paper-author*". Then we add weights to the paths following the short meta paths according to their building time.

On the other hand, evaluating a new similarity measure is difficult, since it is difficult to obtain ground truth. We approach this challenge by assuming that similar objects will exhibit their similarity by their future behaviour. For example, in the Flickr image network, similar images are more likely to share the same tags or be in the same categories in the future. In bibliographic networks, similar authors are more likely to have collaborations in the future. Under this assumption, we can obtain a ground truth to evaluate our extended similarity measure and compare it against existing methods.

The contributions of this paper are summarized as follows:

- We develop a new method that incorporates transitive similarity to capture the impacts of different paths between two objects given a meta path.
- We incorporate temporal information for similarity search in heterogeneous information networks, by assigning different weights for the paths with different building time.
- Experiments on DBLP network data demonstrate the effectiveness of our proposed methods.

The rest of the paper is organized as follows. Section 2 presents related work, then preliminary concepts and a problem definition are given in Section 3. Section 4 introudces our proposed methods, and Section 5 presents the experimental results. Finally, Section 6 concludes the paper.

2 Related Work

The key basis for similarity search is a similarity measure, which measures the similarity between two objects. Similarity measures for traditional data types have been widely studied, for example the Jaccard coefficient and cosine similarity. For graph data, a number of studies utilize link information to measure the similarity between two objects. Early similarity measures include co-citation[11] and co-coupling[7], which were developed for scientific papers. Other similarity measures based on random walks have also been developed, such as SimRank[4] and Personalized PageRank [5]. SimRank measures the similarity between two objects recursively, by averaging the similarity of their neighbours. Personalized PageRank measures the similarity between two objects by the probability of a random walk with restart starting from source object to target object.

The similarity measures defined in homogeneous networks ignore the different types of semantic information that is available under different paths in heterogeneous networks. There are several works on similarity search in heterogeneous information networks. In [14], a meta path framework was proposed for heterogeneous information networks, where a meta path corresponds to a sequence of links between the objects. Based on the framework, a similarity measure called *PathSim* was proposed, which aims to find similar objects with the same type. In [19], the similarity query ambiguity problem was studied, arising from the diversified semantic meanings in heterogeneous information networks. For a query object, users can provide example similar objects for the query as guidance for choosing related objects. Recently, relevance search in heterogeneous networks was studied in [10]. A relevance measure called *HeteSim*, was proposed to measure the relatedness of the objects in heterogeneous networks, either of the same or different type. Overall, these works are based on the meta path framework and can capture semantic information under a meta path. However, they do not differentiate the impacts of concrete paths given a meta path, which can affect the similarity between two objects.

Another line of work related to our problem is link prediction, as the similarity between two objects can be used to predict the existence of a link between them (i.e., friendships and co-authorship). In addition, since we evaluate the similarity measures considering the future behaviour between two similar objects, and such behaviour can be that a link will be formed between them in the future, our problem is similar to link prediction. However, we focus on developing similarity measures and the future information is only used for evaluation, while link prediction aims at developing methods to predict the existence of a link between two objects. The methods for link prediction can be directly using similarity measures[8] or more sophisticated such as using supervised learning[2].

There are several works on link prediction in heterogeneous information networks[12, 18, 13]. The most related work to our problem is co-author relationship prediction in heterogeneous networks. Sun et al.[12], considering heterogeneous meta path-based features, used a logistic regression-based co-author relationship prediction model, to predict future co-author relationships. Our similarity measure can actually serve as a heterogeneous feature for their link prediction model.

3 Preliminaries and Problem Statement

In this section, we briefly introduce concepts related to heterogeneous information networks and define the problem.

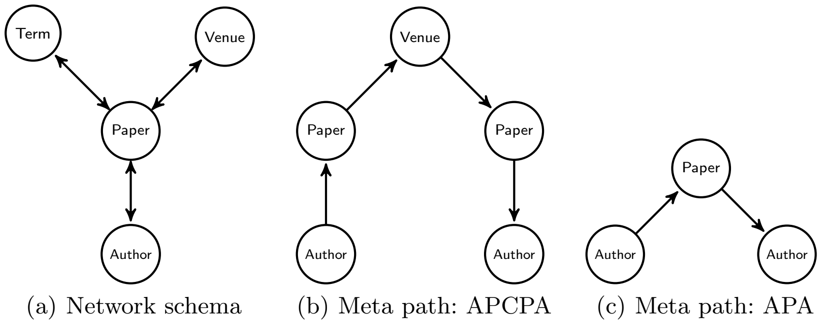
A ***Heterogeneous information network*** is defined as a graph $G = (V, E, \mathcal{T}, \mathcal{R})$ where V is a set of objects, E is a set of links, $\mathcal{T}$ is a set of object types and $\mathcal{R}$ is a set of link types between object types. Since a heterogeneous information network contains multiple types of objects and links, $|\mathcal{T}| > 1$ and $|\mathcal{R}| > 1$. Each object $v \in V$ is associated with a particular type $T_i \in \mathcal{T}$, and each link $e \in E$ is associated with a particular type $R_j \in \mathcal{R}$.

The concept of ***network schema***[14] has been proposed to describe the meta structure of a heterogeneous network for better understanding. It is a graph defined as $S_G = (\mathcal{T}, \mathcal{R})$ where each object is an object type and each link is a link type between object types.

For example, Fig. 1(a) shows the network schema for a bibliographic information network. There are four types of objects: papers (P), venues(conferences/ journals) (C), authors (A) and terms (T) which are the words appearing in the paper title. Also there are different links between the objects. For example, the links between authors and papers denote the writing or written-by relations.

A ***meta path*** $\mathcal{P}$ is a path defined over network schema, and is formalized as $T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} T_{l+1}$, which defines a composite relation between type T_1

and T_{l+1}. The length of $\mathcal{P}$ is the number of relations in it. The objects can be connected through different meta paths. Two examples of meta path are shown in Fig. 1(b) and Fig. 1(c). For simplicity, the meta path is denoted by the names of object types.



(a) Network schema (b) Meta path: APCPA (c) Meta path: APA

Fig. 1. (a) A bibliographic network schema; (b) meta path "author-paper-venue-paper-author" (APCPA) describing authors publish papers in the same conferences; (c) meta path "author-paper-author" (APA) describing co-author relationship.

PathSim[14] is a meta path-based similarity measure, which aims at finding similar peer objects for a query object, such as finding similar authors in terms of research area and reputation. Given a symmetric meta path $\mathcal{P}$, *PathSim* computes the similarity between two objects x and y according to

$$s(x,y) = \frac{2 \times |\mathcal{P}_{x \rightsquigarrow y}|}{|\mathcal{P}_{x \rightsquigarrow x}| + |\mathcal{P}_{y \rightsquigarrow y}|} \tag{1}$$

where $\mathcal{P}_{x \rightsquigarrow y}$ is the set of paths between x and y following $\mathcal{P}$, $\mathcal{P}_{x \rightsquigarrow x}$ is that between x and x, and $\mathcal{P}_{y \rightsquigarrow y}$ is that between y and y. The intuition behind *PathSim* is that two similar peer objects should not only be strongly connected, but also share comparable visibility. Their connectivity is defined as the number of paths between them following $\mathcal{P}$, and the visibility is defined as the number of paths between themselves[14].

Given a symmetric meta path $\mathcal{P} = T_1 T_2 \cdots T_l$, *PathSim* similarity between two objects $x_i \in T_1$ and $x_j \in T_l$ with the same type $s(x_i, x_j)$, can be computed through the ***commuting matrix*** M, which is defined as $M = W_{T_1T_2}W_{T_2T_3}\cdots W_{T_{l-1}T_l}$, where $W_{T_iT_j}$ is the adjacency matrix between type T_i and type T_j. M_{ij} denotes the number of paths between object $x_i \in T_1$ and objects $y_j \in T_l$ following meta path $\mathcal{P}$, and $M_{ij} = |\mathcal{P}_{x_i \rightsquigarrow x_j}|$. Similarly, $M_{ii} = |\mathcal{P}_{x_i \rightsquigarrow x_i}|$ and $M_{jj} = |\mathcal{P}_{x_j \rightsquigarrow x_j}|$.

Problem Statement: The problem studied in this paper is as follows. Given a heterogeneous information network and a query object, the goal is to find the top-k objects with the same type and the highest similarity with respect to the query object.

4 Proposed Methods

In this section, we introduce our methods to extend *PathSim* by incorporating transitive similarity and temporal information.

4.1 Transitive Similarity

Given a meta path $\mathcal{P} = T_1 T_2 \cdots T_l$, where T_1 and T_l are the same type ($T_1 = T_l$), $\mathcal{T}_m$ is the set of intermediate types which are the same as T_1 and T_l, $\mathcal{T}_m = (T_{m1}, T_{m2}, \cdots, T_{md})$ where d is the cardinality of $\mathcal{T}_m$. Therefore, $\mathcal{P}$ can be concatenated by multiple meta paths $\mathcal{P}_i (i = 1, \cdots, d+1)$, which is shown in Eq.(2).

$$\mathcal{P} = \underbrace{T_1 \cdots T_{m1}}_{\mathcal{P}_1}\underbrace{\cdots T_{m2}}_{\mathcal{P}_2} \cdots \underbrace{T_{md} \cdots T_l}_{\mathcal{P}_{d+1}} \tag{2}$$

PathSim [14] treats all the paths between object $x \in T_1$ and $y \in T_l$ connected through different transitive objects $z \in T_{mh}$ equally. However, intuitively, we are more likely to trust the paths betweens the objects which are more similar to each other. We can put different weights on the paths following $\mathcal{P}_i$ considering the transitive similarity between the start type and the end type of $\mathcal{P}_i$. A simple way of obtaining the transitive similarity is to utilize *PathSim* over different meta paths with different semantic meanings. Therefore, for meta path $\mathcal{P}$, its commuting matrix can be computed as

$$M_{\mathcal{P}} = M^s_{\mathcal{P}_1} M^s_{\mathcal{P}_2} \cdots M^s_{\mathcal{P}_{d+1}} \tag{3}$$

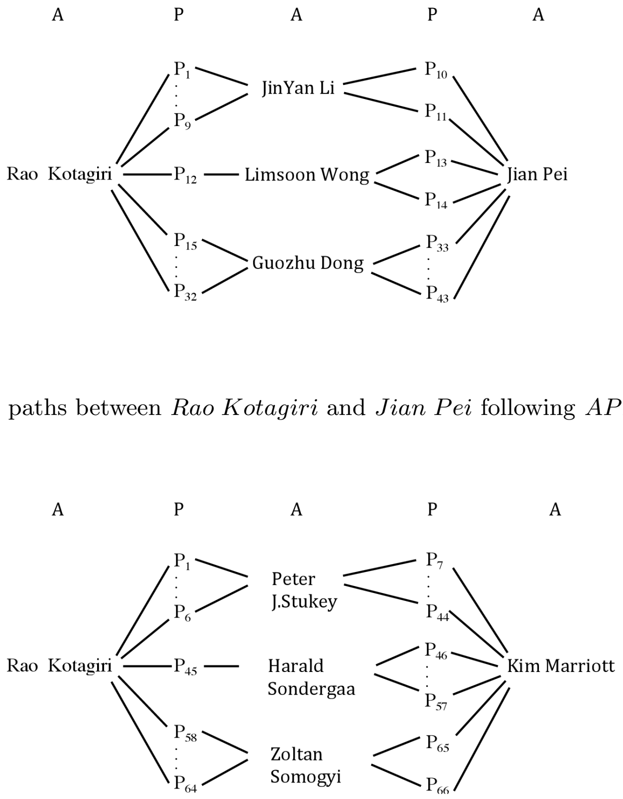
where $M^s_{\mathcal{P}_i}$ is the commuting matrix for meta path $\mathcal{P}_i$ with transitive similarity incorporated, and can be computed as

$$M^s_{\mathcal{P}_i} = M_{\mathcal{P}_i} \cdot S_{\mathcal{P}'} \tag{4}$$

where $M_{\mathcal{P}_i}$ denotes the commuting matrix of $\mathcal{P}_i$, with each element representing the number of paths between object $x \in T_s(\mathcal{P}_i)$ and object $y \in T_e(\mathcal{P}_i)$, where $T_s(\mathcal{P}_i)$ and $T_e(\mathcal{P}_i)$ represents the start type and the end type of $\mathcal{P}_i$ respectively. $S_{\mathcal{P}'}$ denotes a transitive similarity matrix computed on meta path $\mathcal{P}'$. $\mathcal{P}'$ can be different meta paths such that $T_s(\mathcal{P}') = T_e(\mathcal{P}') = T_s(\mathcal{P}) = T_e(\mathcal{P})$. $S_{\mathcal{P}'}$ allows us to incorporate different meta paths with different semantic meanings.

To better illustrate our method, we give an example in bibliographic networks. Fig. 2 shows the paths between *Rao Kotagiri*(*Rao*) and *Jian Pei*(*Jian*) following meta path *APAPA*, and the one between *Rao* and *Kim Marriott* (*Kim*) according to DBLP between 1990 and 2007. *Rao* and *Jian* (*Kim*) are not co-authors between 1990 and 2007. But they are connected through their common co-authors. Suppose *Rao* is the query author, the *PathSim* similarity between *Rao* and *Jian* according to Eq.(1) is,

$$\begin{aligned} s(Rao, Jian) &= \frac{2 \times |APAPA_{Rao \rightsquigarrow Jian}|}{|APAPA_{Rao \rightsquigarrow Rao}| + |APAPA_{Jian \rightsquigarrow Jian}|} \\ &= \frac{2 \times (9 \times 2 + 1 \times 2 + 18 \times 11)}{21280 + 15333} = 0.0119 \end{aligned}$$



(a) The paths between *Rao Kotagiri* and *Jian Pei* following $APAPA$

(b) The paths between *Rao Kotagiri* and *Kim Marriott* following $APAPA$

Fig. 2. Example of paths following $APAPA$ with *Rao Kotagiri* as the query author and two candidate authors

where the process of computation of $|APAPA_{Rao \rightsquigarrow Rao}| = 21280$ is not shown due to the space limitation, and the same for $Jian$ (15333). Similarly, $s(Rao, Kim)$ $= 0.0134$. However, according to our improved similarity measure,

$$\begin{aligned} s'(Rao, Jian) &= \frac{2 \times \sum_{c \in Co}(|APA_{Rao \rightsquigarrow c}| \times S_{Rao,c} + |APA_{c \rightsquigarrow Jian}| \times S_{c,Jian})}{19357.04 + 12594.43} \\ &= \frac{2 \times 3.59}{19357.04 + 12594.43} = 2.25E - 04 \end{aligned}$$

where c denotes a common co-author of Rao and $Jian$, $Co = \{JinYan\ Li, Limsoon\ Wong, Guozhu\ Dong\}$ denotes the set of common co-authors of Rao and $Jian$, $S_{Rao,c}$ denotes the transitive similarity between Rao and c (in this example, S

is computed based on APA), and similarly for $S_{c,Jian}$. The number of paths (weighted) between Rao and Rao (19357.04) is given directly due to the space limitation, and the same for $Jian$ (12594.43). Similarly, $s'(Rao, Kim) = 1.43E-04$. We assume that more similar authors are more likely to collaborate with the query author in future. In this example, based on the DBLP data between 2008 and 2013, $Jian$ has collaboration with Rao, while Kim does not. We can see that our improved similarity measure can rank $Jian$ higher compared with Kim.

4.2 Temporal Dynamics

Heterogenous information networks evolve over time, and also the similarity between two objects can change over time. We are more interested in finding similar objects now or even in the future. Intuitively, two objects are more similar if there are more recent connections between them. Instead of treating the paths given a single snapshot equally, we differentiate the impacts of paths formed at different timestamps. A simple way is to put different weights on the paths formed in different timestamps. Essentially, the older paths make less contribution to similarity than recent ones, and should be given lower weights.

Given a meta path $\mathcal{P} = T_1T_2 \cdots T_l$, its commuting matrix can be computed as

$$M_{\mathcal{P}} = M^t_{\mathcal{P}_1} M^t_{\mathcal{P}_2} \cdots M^t_{\mathcal{P}_g} \tag{5}$$

where $M^t_{\mathcal{P}_i}$ is the commuting matrix for meta path $\mathcal{P}_i$ with temporal information incorporated, and such that $\sum_{i=1}^{g} l(P_i) = l(\mathcal{P})$, where $l(\mathcal{P}_i)$ is the length of meta path $\mathcal{P}_i$. $\mathcal{P}_i$ is a meta path on which an event happens in a particular timestamp. For example, it can be APC in bibliographic networks which represents author publish paper in conference in a particular year. $M^t_{\mathcal{P}_i}$ can be computed as

$$M^t_{\mathcal{P}_i} = M_{\mathcal{P}_i} \cdot Y_{\mathcal{P}_i} \tag{6}$$

where $Y_{\mathcal{P}_i}$ is the temporal matrix on $\mathcal{P}_i$, with each element represents the weight of the path between object $x \in T_s(\mathcal{P}_i)$ and object $y \in T_e(\mathcal{P}_i)$. The weight can be assigned according to the timestamp of the path formed. Here, we define a function $f(t)$ of timestamp t to decide the weights,

$$f(t) = \alpha^{(t_1-t)}(t_0 \leq t \leq t_1) \tag{7}$$

where t_0 and t_1 represent the start time and end time of the data used for computing similarities. $\alpha(0 < \alpha < 1)$ can be varied. The path formed most recently in t_1 has the largest weight 1. The smaller α is, the more rapidly the weight of the less recent path drops. Different $f(t)$ can be defined. In this paper, we focus on the importance of incorporating temporal information instead of studying the impacts of different $f(t)$.

Based on the above proposed methods, we can improve $PathSim$ by incorporating transitive similarity and/or temporal dynamic, and find the top-k similar objects for a give query object based on our improved similarity measure.

5 Experiments

In this section, we compare the effectiveness of our improved similarity measure using the *PathSim* measure as a baseline.

5.1 Evaluation Measure

Assessing similarity is challenging since it is difficult to obtain ground truth providing a quantitative measure for the similarity between two objects. Most existing methods to evaluate the performance of similarity measures rely on user studies or on an reliable external measure of similarity. The study in [14] used case studies and manually labeled the results for a handful of queries, evaluating using domain knowledge based on these queries. In this paper, since we assume that similar objects will show similar behaviour in some way in the future, we can obtain ground truth to evaluate the similarity measure and provide a comprehensive experimental assessment using thousands of test queries.

We use NDCG (Discounted Normalised Cumulative Gain), a widely used measure in information retrieval [1][3], to evaluate the ranking performance. It rewards relevant objects in the top ranked results more heavily than those ranked lower. In particular, we use NDCG@n, which computes NDCG over the top n ranked objects, and which can be computed as

$$\begin{aligned} NDCG@n &= \frac{DCG@n}{IDCG@n} \\ DCG@n &= rel(1) + \sum_{i=2}^{n} \frac{rel(x_i)}{log_2(i)} \end{aligned} \tag{8}$$

where $IDCG@n$ denotes the Ideal DCG for a perfect ranking and $rel(x_i)$ denotes the relevance score for an object x_i at position i.

5.2 Experiment Setup

The DBLP dataset downloaded on 25th April 2013 is used in our experiments. The network schema of DBLP network is same as Fig. 1(a). The data from 1990 to 2007 (denoted as $T_{1990-2007}$) is used to compute similarity, while the data from 2008 to 2013(denoted as $T_{2008-2013}$) is used for evaluation. The number of authors, papers, conferences (including journals) and terms (after removing stopwords in paper titles) between 1990 and 2007 are shown in Table. 1.

Table 1. DBLP data between 1990 and 2007

Data	Author	Paper	Conference	Term
1990-2007	698,507	1,114,726	4,949	139,613

We focus on computing the similarity between two authors given a meta path between them. In particular, we use meta path $APAPA$ which implies two authors share the same co-authors. Given a query author q, the top n similar authors are returned with similarity computed based on the data in $T_{1990-2007}$. We assume that similar authors will exhibit their similarity by their future behaviour. For meta path $APAPA$, two similar authors might collaborate in the future ($T_{2008-2013}$). To easily capture such behaviour for evaluation, we only return the top n similar authors who have not collaborated with the query author in $T_{1990-2007}$. To evaluate the ranking performance, we need the relevance score $rel(x_i)$ for each returned similar author w.r.t. q. According to the number of co-authored publications between x_i and q in $T_{2008-2013}$, $rel(x_i)$ can be set as

$$rel(x_i) = \begin{cases} 0 & \text{if } N(q, x_i)\text{=}0 \\ \varphi\left(N(q, x_i)\right) & \text{if } N(q, x_i) \neq 0 \end{cases} \tag{9}$$

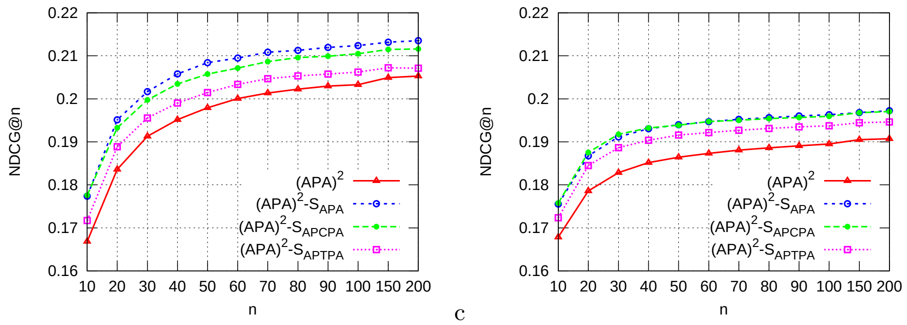
where $N(q, x_i)$ denotes the number of papers that q and x_i publish together in $T_{2008-2013}$. We use $\mathcal{C}$ to denote the set of all the candidate authors. The candidate authors are ranked in ascending order according to $N(q, x)(x \in \mathcal{C})$, and each candidate is assigned a ranking value according to its ranking position. For those who have same value of $N(q, x)$, the same ranking value will be assigned. $\varphi(\cdot)$ is a mapping function from $N(q, x_i)$ to the ranking value for x_i.

The query authors can be chosen from the set of authors who exist in $T_{1990-2007}$, and have new collaborations with authors exist in $T_{1990-2007}$ in future time interval $T_{2008-2013}$. We randomly select 3000 authors as query authors, and compute the averaged results over the 3000 authors. We compare our improved similarity measure with $PathSim$ using paired t-test with $p = 0.05$. This process is repeated 10 times, and the results reported in this paper are the averaged results over 10 runs. In addition, we show the effectiveness of our similarity measure on two sets of query authors, highly productive authors with more than 15 publications in $T_{1990-2007}$ (denoted as HP), and less productive authors with between 5 and 15 publications in $T_{1990-2007}$ (denoted as LP).

5.3 Experimental Results

Transitive Similarity Incorporated. In this group of experiments, we incorporate different kinds of transitive similarity into meta path $APAPA$. We compare our methods with the baseline method, $PathSim$ applied on $APAPA$. The results are shown in Fig.3, where $(APA)^2$ represents the baseline method, and $(APA)^2 - S_{APA}$, $(APA)^2 - S_{APCPA}$ and $(APA)^2 - S_{APTPA}$ represents our methods on $APAPA$ with incorporated transitive similarity based on APA, $APCPA$ and $APTPA$ respectively. All the results have statistical significance with p-value<0.05.

It can be seen from Fig.3 that after incorporating different similarity information, the performances of our methods are improved over all the varying n on both HP and LP queries. Basically, the similarity incorporated based on APA gives better performance compared with $APCPA$ and $APTPA$. In addition, the



(a) HP queries

(b) LP queries

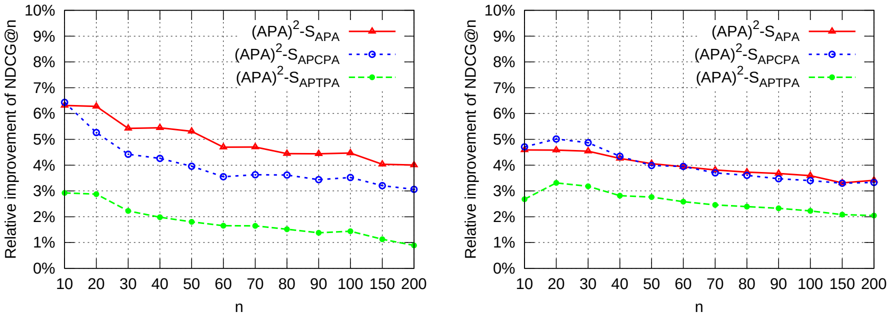
Fig. 3. NDCG@n of $(APA)^2$ denoting the baseline method ($PathSim$) on $APAPA$ and our methods $(APA)^2 - S_{APA}$, $(APA)^2 - S_{APCPA}$ and $(APA)^2 - S_{APTPA}$ denoting $APAPA$ with incorporated transitive similarity based on APA, $APCPA$ and $APTPA$ respectively, for (a)HP queries and (b)LP queries.

performances of all the similarity measures in terms of NDCG@N are low. The main reason is that ranking is generally difficult, especially in the case of similar authors in terms of future collaborators, and only using the raw similarity produced by the similarity measures. Actually, two authors can collaborate due to many external factors that cannot be captured using the similarity measures in this paper. Another reason is that for each run, among the 3000 queries, there are a number of queries with 0 for NDCG@n , which degrade the average results. Such queries do not have future collaborations with their 2-hop authors.

In addition, the overall performance of both the baseline method and our methods on LP queries is worse than that on HP queries. The reason is that for each run, among the 3000 queries, only about 1500 queries have new collaborations with their 2-hop authors for LP queries, while about 2200 for HP queries. Meanwhile, it indicates that HP authors are more likely to collaborate with their 2-hop authors compared with LP authors.

Since the absolute improvements can be misleading, we mainly report the relative improvements of NDCG@n (which is also used in studies in information retrieval[9, 17]) in the following experiments. The relative improvements of our methods over $PathSim$ on meta path $APAPA$ are given in Fig. 4. We can see that the relative improvements of our method with transitive similarity S_{APA} and S_{APCPA}, are more than 4% and 3% respectively over all the values of varying n on HP queries. Furthermore, the relative improvements for S_{APTPA} on HP queries is less than that on LP queries. The reason might be that HP authors are generally active in diverse research topics, which yields diverse terms.

Temporal Information Incorporated. In this group of experiments, we show the effectiveness of incorporating temporal information. We incorporate tempo-



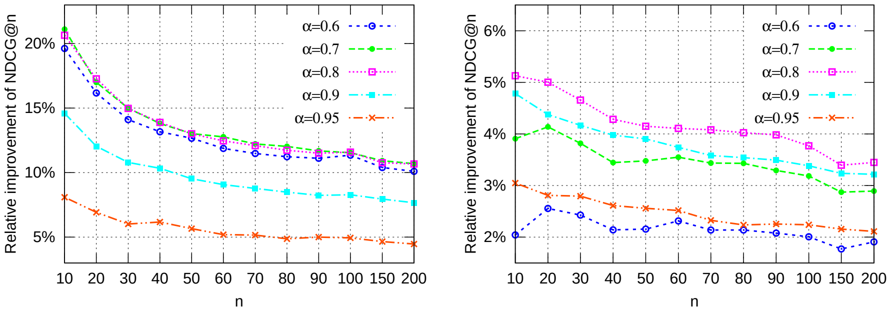
(a) HP queries

(b) LP queries

Fig. 4. Relative improvements of our methods $(APA)^2 - S_{APA}$, $(APA)^2 - S_{APCPA}$ and $(APA)^2 - S_{APTPA}$ over *PathSim* on *APAPA*

ral information into meta path *APAPA*, and use Eq.(7) to decide the weights of the paths following *APA*. Here, $t_0 = 1990$, $t_1 = 2007$.

First we study the impact of parameter α. Fig.5 shows the relative improvements of our method $(APA)^2\_T_{\alpha}$ with varying α over *PathSim* on *APAPA*, where $(APA)^2\_T_{\alpha}$ denotes incorporating the temporal information (with varying α) into *APAPA*. It can be seen that when $\alpha = 0.8$, our method can yield good performance on both *HP* and *LP* queries. In addition, the relative improvements on *HP* queries are much higher than *LP* queries. The reason might be that the links associated with *LP* authors are relatively sparse, and are formed in a relatively short time interval, which do not contain much diversified temporal information to be exploited.

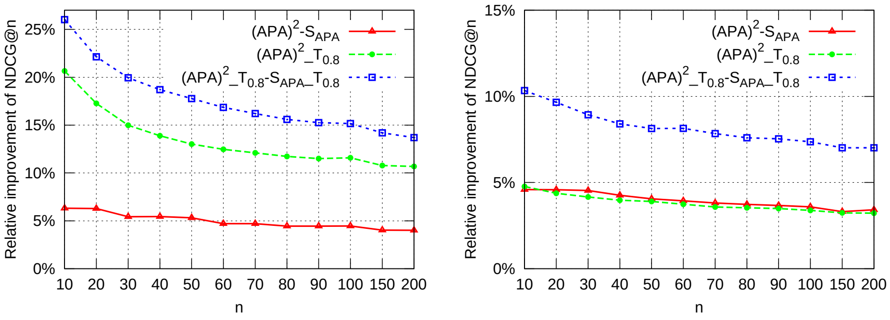


(a) HP queries

(b) LP queries

Fig. 5. Relative improvements of our method $(APA)^2\_T_{\alpha}$ denoting the temporal information (with varying α) incorporated to *APAPA* over *PathSim* on *APAPA*.

Furthermore, we compare the relative improvements over *PathSim* when incorporating temporal information and/or transitive similarity into $APAPA$. Fig. 6 shows the results when incorporating only transitive similarity ($(APA)^2\_S_{APA}$), only temporal information ($(APA)^2\_T_{0.8}$), and both of them ($APAPA\_T_{0.8} - S_{APA}\_T_{0.8}$) to $APAPA$.



(a) HP queries

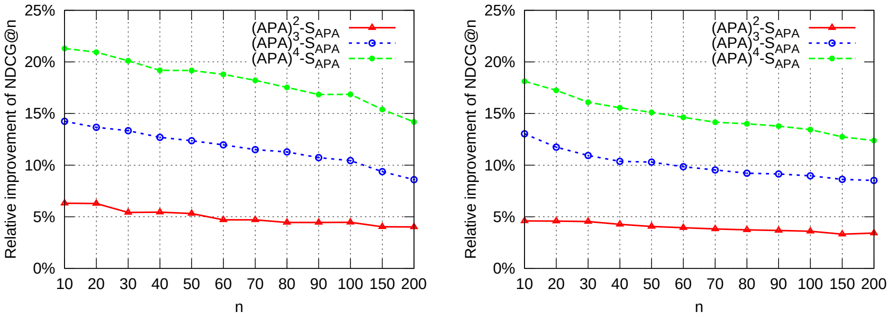
(b) LP queries

Fig. 6. Relative improvements of our method $(APA)^2\_S_{APA}$, $(APA)^2\_T_{0.8}$ and $APAPA\_T_{0.8} - S_{APA}\_T_{0.8}$ over *PathSim* on HP queries and LP queries.

It can be seen that there is little difference for the relative improvements of incorporating transitive similarity on HP queries and LP queries. But incorporating temporal information makes huge differences, and basically it works better for HP queries. In addition, the more information incorporated, the higher the performance is, which can be seen from Fig.6 that, $APAPA\_T_{0.8} - S_{APA}\_T_{0.8}$ achieves the best performance with relative improvements more than 15% on HP queries and more than 7% on LP queries.

Impacts on Different Length of Meta Path In this group of experiments, we check the impacts of transitive similarity on different length of meta path. Fig. 7 shows the relative improvements of incorporating transitive similarity (based on APA) into different length of meta path APA over *PathSim* applied on corresponding length of meta path APA, where $(APA)^4 - S_{APA}$ represents the relative improvements of incorporating transitive similarity (based on APA) into $(APA)^4$ over *PathSim* on $(APA)^4$, and similarly for $(APA)^3 - S_{APA}$ and $(APA)^2 - S_{APA}$.

It can be seen that the relative improvement on longer paths is much higher than shorter paths. This is because *PathSim* does not distinguish the importance of different paths given a meta path. When increasing the length of a meta path, *PathSim* will treat more remote (and possibly irrelevant) neighbours as similar, whilst our methods which take into account transitive similarity can alleviate this effect.



(a) HP queries

(b) LP queries

Fig. 7. Relative improvement on NDCG@n for different length of APA with transitive similarity based on APA incorporated

6 Conclusion and Future work

We have studied the problem of similarity search in heterogeneous information networks and we have proposed an improved meta path-based similarity measure which incorporates transitive similarity and temporal information. Experimental results show that our improved similarity measures outperforms the baseline existing method. We also found that using temporal information can provide greater gains on highly productive authors than less productive authors. Furthermore, using transitive similarity and temporal information simultaneously can produce the best performance. In future, we plan to consider in more detail other types of objects and networks.

References

1. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 19–26. ACM (2006)
2. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: Proceedings of the Sixth SIAM Data Mining Workshop on Link Analysis, Counter-terrorism and Security (2006)
3. Balasubramanian, N., Kumaran, G., Carvalho, V.R.: Predicting query performance on the web. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 785–786. ACM (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. KDD '02, ACM, Edmonton, Alberta, Canada (2002)
5. Jeh, G., Widom, J.: Scaling personalized web search. In: Proceedings of the 12th international conference on World Wide Web. pp. 271–279. WWW '03, ACM, Budapest, Hungary (2003)

6. Ji, M., Han, J., Danilevsky, M.: Ranking-based classification of heterogeneous information networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1298–1306. ACM (2011)
7. Kessler, M.M.: Bibliographic coupling between scientific papers. American documentation 14(1), 10–25 (1963)
8. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the twelfth international conference on Information and knowledge management. pp. 556–559. CIKM '03, ACM, New Orleans, LA, USA (2003)
9. Qin, T., Zhang, X.D., Wang, D.S., Liu, T.Y., Lai, W., Li, H.: Ranking with multiple hyperplanes. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 279–286. ACM (2007)
10. Shi, C., Kong, X., Yu, P.S., Xie, S., Wu, B.: Relevance search in heterogeneous networks. In: Proceedings of the 15th International Conference on Extending Database Technology. pp. 180–191. EDBT '12, ACM, Berlin, Germany (2012)
11. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science 24(4), 265–269 (1973)
12. Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C., Han, J.: Co-author relationship prediction in heterogeneous bibliographic networks. In: Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining. pp. 121–128. ASONAM '11, IEEE Computer Society, Washington, DC, USA (2011)
13. Sun, Y., Han, J., Aggarwal, C.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. WSDM '12, ACM, Seattle, Washington, USA (2012)
14. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment 4(11) (2011)
15. Sun, Y., Tang, J., Han, J., Gupta, M., Zhao, B.: Community evolution detection in dynamic heterogeneous information networks. In: Proceedings of the Eighth Workshop on Mining and Learning with Graphs. pp. 137–146. ACM (2010)
16. Sun, Y., Yu, Y., Han, J.: 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. pp. 797–806. ACM (2009)
17. Yeh, J.Y., Lin, J.Y., Ke, H.R., Yang, W.P.: Learning to rank for information retrieval using genetic programming. In: Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007) (2007)
18. Yu, X., Gu, Q., Zhou, M., Han, J.: Citation prediction in heterogeneous bibliographic networks. In: Proceedings of the Twelfth SIAM International Conference on Data Mining. pp. 1119–1130. Anaheim, California, USA (2012)
19. Yu, X., Sun, Y., Norick, B., Mao, T., Han, J.: User guided entity similarity search using meta-path selection in heterogeneous information networks. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 2025–2029. CIKM '12, ACM, Maui, Hawaii, USA (2012)