HRank: A Path Based Ranking Method in Heterogeneous Information Network

Yitong Li[1], Chuan Shi[1,⋆], Philip S. Yu[2], and Qing Chen[3]

[1] Beijing University of Posts and Telecommunications, Beijing, China 100876
[2] University of Illinois at Chicago, IL, USA
[3] China Mobile Communications Corporation, Beijing, China

Abstract. Recently, there is a surge of interests on heterogeneous information network analysis. Although evaluating the importance of objects has been well studied in homogeneous networks, it is not yet exploited in heterogeneous networks. In this paper, we study the ranking problem in heterogeneous networks and propose the HRank method to evaluate the importance of multiple types of objects and meta paths. A constrained meta path is proposed to subtly capture the rich semantics in heterogeneous networks. Since the importance of objects depends upon the meta paths in heterogeneous networks, HRank develops a path based random walk process. Furthermore, HRank can simultaneously determine the importance of objects and meta paths through applying the tensor analysis. Experiments on three real datasets show that HRank can effectively evaluate the importance of objects and paths together. Moreover, the constrained meta path shows its potential on mining subtle semantics by obtaining more accurate ranking results.

Keywords: Heterogeneous information network, Rank, Random walk, Tensor analysis.

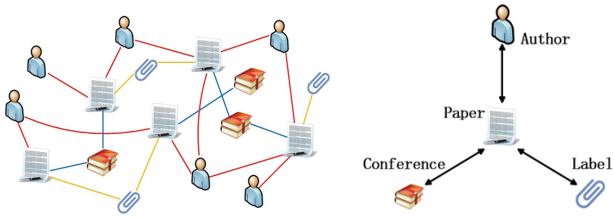
1 Introduction

It is an important research problem to evaluate object importance or popularity, which can be used in many data mining tasks. Many methods have been developed to evaluate object importance, such as PageRank [9], HITS [1], and SimRank [3]. In these literatures, objects ranking is done in a homogeneous network in which objects or relations are same-typed. However, in many real network data, there are many different types of objects and relations, which can be organized as heterogeneous network. Formally, Heterogeneous Information Networks (HIN) are the logical networks involving multiple types of objects as well as multiple types of links denoting different relations [2]. It is clear that heterogeneous information networks are ubiquitous and form a critical component of modern information infrastructure [2].

Fig. 1(a) shows a HIN example in bibliographic data and Fig. 1(b) illustrates its network schema. In this example, it contains four types of objects: papers (P), authors (A), labels (L, categories of papers) and conferences (C), and links

⋆ Corresponding author.

F. Li et al. (Eds.): WAIM 2014, LNCS 8485, pp. 553–565, 2014.



(a) Heterogeneous network (b) Network schema

Fig. 1. A heterogeneous information network example on bibliographic data. (a) shows heterogeneous objects and their relations. (b) shows the network schema.

connecting them. The link types are defined by the relations between two object types. For example, links exist between authors and papers denoting the writing or written-by relations. In this network, several interesting, yet seldom exploited, ranking problems can be proposed.

- One may pay attention to the importance of multiple types of objects simultaneously, and ask the following questions:
  Q. 1.1 Which are the most influential authors and reputable conferences?
  Q. 1.2 Which are the most influential authors and reputable conferences in data mining field?
- Furthermore, one may wonder which factor mostly affects the importance of objects. So he may ask the questions like this:
  Q. 2 Who are the most influential authors and which factor makes the author most influential?

The ranking analysis in HIN faces the following research challenges. (1) There are different types of objects and links in HIN. If we simply treat all objects equally and apply the random walk as PageRank does in homogeneous network, the ranking result will mix different types of objects together. (2) Different types of objects and links in heterogeneous networks carry different semantic meanings. The random walk along different meta paths has different semantics, which may lead to different ranking results. Here the meta path [4] means a sequence of relations between object types. So a desirable ranking method in HIN should be path-dependent. The study of related work can be seen on [10], which is the extension of this paper.

In this paper, we study the ranking problem in HIN and propose a novel ranking method, HRank, to evaluate the importance of multiple types of objects and meta paths in HIN. For *Q. 1*, a meta path based random walk model is proposed to evaluate the importance of single or multiple types of objects. Although meta path has been widely used to capture the semantics in HIN [6,4], it coarsely depicts object relations. By employing the meta path, we can only answer the *Q. 1.1*. In order to overcome the shortcoming existing in meta path, we propose the *constrained meta path* concept, which can effectively describe subtle semantics. The constrained meta path sets constraint conditions on meta path. Adopting the constrained meta path, we can further answer the *Q. 1.2*. Moreover, in HIN, based on different paths, the objects have different ranking values. The comprehensive importance of objects should consider all kinds of factors (the factors

can be embodied by constrained meta paths), which have different contribution to the importance of objects. In order to evaluate the importance of objects and meta paths simultaneously (i.e., answer *Q. 2*), we further propose a co-ranking method which organizes the relation matrices of objects on different constrained meta paths as a tensor. A random walk process is designed on this tensor to co-rank importance of objects and paths simultaneously. That is, random walkers surf in the tensor, where the stationary visiting probability of objects and meta paths is considered as the HRank score of objects and paths.

2 Preliminary

In this section, we describe notations used in this paper and present some preliminary knowledge.

In heterogeneous information networks, there are multiple object types and relation types. We use the network schema $S = (\mathcal{A}, \mathcal{R})$ to depict the object types and relations existing among object types, where $\mathcal{A} = \{A\}$ is a set of object types and $\mathcal{R} = \{R\}$ is a set of relations. A relation R existing from type S to type T is denoted as $S \xrightarrow{R} T$. Fig. 1(b) shows a network schema of bibliographic information network.

Different from homogeneous networks, two objects in a heterogeneous network can be connected via different paths and these paths have different meanings. These paths are called meta paths which can be defined as follows.

Definition 1 ***Meta path*** *[4]. A meta path $\mathcal{P}$ is a path defined on a schema $S = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$ (abbreviated as $A_1 A_2 \ldots A_{l+1}$), which defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$ between type A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations.*

It is obvious that semantics underneath these paths are different. The "*Author-Paper-Author*" (APA) path means authors collaborating on the same papers, while the "*Author-Paper-Conference-Paper-Author*" ($APCPA$) path means the authors' papers publishing on the same conferences. Based on different meta paths, there are different relation networks, which may result in different importance of objects. However, meta path fails to capture some subtle semantics. Taking Fig. 1(b) as an example, the APA cannot reveal the co-author relations in Data Mining (DM) field. In order to overcome the shortcomings in meta path, we propose the concept of constrained meta path, defined as follows.

Definition 2 ***Constrained meta path.*** *A constrained meta path is a meta path based on a certain constraint which is denoted as $\mathcal{CP} = \mathcal{P}|\mathcal{C}$. $\mathcal{P} = (A_1 A_2 \ldots A_l)$ is a meta path, while $\mathcal{C}$ represents the constraint on the objects in the meta path.*

Note that the $\mathcal{C}$ can be one or multiple constraint conditions on objects. Taking Fig. 1(b) as an example, the constrained meta path $APA|P.L =$ "DM" represents the co-author relations of authors in data mining field through constraining the label of papers with DM. Similarly, the constrained meta path $APCPA|P.L =$ "DM"&&$C =$ "$CIKM$" represents the co-author relations of

authors in CIKM conference and the papers of authors are in data mining field. Obviously, compared to meta path, the constrained meta path conveys richer semantics by subdividing meta paths under distinct conditions.

For a relation $A \xrightarrow{R} B$, we can obtain its constrained transition probability matrix as follows.

Definition 3 ***Constrained transition probability matrix.*** *W_{AB} is an adjacent matrix between type A and B on relation $A \xrightarrow{R} B$. U_{AB} is the normalized matrix of W_{AB} along the row vector, which is the transition probability matrix of $A \xrightarrow{R} B$. Suppose there is a constraint $\mathcal{C}$ on object type A. The constrained transition probability matrix U'_{AB} of constrained relation $R|\mathcal{C}$ is $U'_{AB} = M_{\mathcal{C}} U_{AB}$, where $M_{\mathcal{C}}$ is the constraint matrix generated by the constraint condition $\mathcal{C}$.*

The constraint matrix $M_{\mathcal{C}}$ is usually a diagonal matrix whose dimension is the number of objects in object type A. The element in the diagonal is 1 if the corresponding object satisfies the constraint, else the element is 0. Similarly, we can confine the constraint on object type B or both types.

Given a network following a network schema $S = (\mathcal{A}, \mathcal{R})$, we can define the constrained meta path based reachable probability matrix as follows.

Definition 4 ***Constrained meta path based reachable probability matrix.*** *For a constrained meta path $\mathcal{CP} = (A_1 A_2 \cdots A_{l+1}|\mathcal{C})$, the constrained meta path based reachable probability matrix is defined as $PM_{\mathcal{CP}} = U'_{A_1 A_2} U'_{A_2 A_3} \cdots U'_{A_l A_{l+1}}$. $PM_{\mathcal{CP}}(i, j)$ represents the probability of object $i \in A_1$ reaching object $j \in A_{l+1}$ under the constrained meta path $\mathcal{CP}$.*

When there is a constraint on the objects, we only consider the objects that satisfy the constraint. For simplicity, we use the reachable probability matrix and the $M_{\mathcal{P}}$ to represent the constrained meta path based reachable probability matrix in the following section.

3 The HRank Method

In order to answer the two ranking problems proposed in Section 1, we design two versions of HRank, respectively.

3.1 Ranking Based on Constrained Meta Paths

For the question *Q. 1*, we propose the HRank-CMP method based on a constrained meta path $\mathcal{P} = (A_1 A_2 \ldots A_l|\mathcal{C})$.

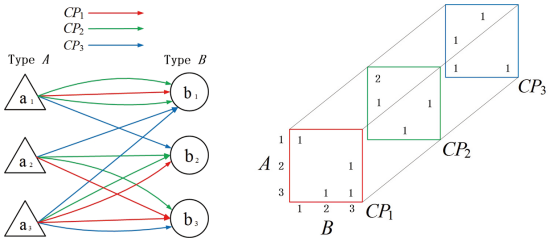
HRank-CMP is based on a random walk process that random walkers wander between A_1 and A_l along the path. The ranks of A_1 and A_l can be seen as the visiting probability of walkers, which are defined as follows:

$$\begin{aligned} R(A_l|\mathcal{P}^{-1}) &= \alpha R(A_1|\mathcal{P}) M_{\mathcal{P}} + (1-\alpha) E_{A_l} \\ R(A_1|\mathcal{P}) &= \alpha R(A_l|\mathcal{P}^{-1}) M_{\mathcal{P}^{-1}} + (1-\alpha) E_{A_1} \end{aligned} \tag{1}$$

where $M_{\mathcal{P}}$ and $M_{\mathcal{P}^{-1}}$ are the reachable probability matrix of path $\mathcal{P}$ and $\mathcal{P}^{-1}$. E_{A_1} and E_{A_l} are the restart probability of A_1 and A_l. Note that the path $\mathcal{P}$ is either symmetric ($\mathcal{P} = \mathcal{P}^{-1}$) or asymmetric ($\mathcal{P} \neq \mathcal{P}^{-1}$).

3.2 Co-ranking for Objects and Relations in HIN

There are many constrained meta paths in heterogeneous networks. It is an important issue to automatically determine the importance of paths [4,5], since it is usually hard for us to identify which relation is more important in real applications. To solve this problem (i.e., *Q. 2*), we propose the HRank-CO to co-rank the importance of objects and relations. The basic idea is based on an intuition that important objects are connected to many other objects through a number of important relations and important relations connect many important objects. So we organize the multiple relation networks with a tensor and a random walk process is designed on this tensor. The method not only can comprehensively evaluate the importance of objects by considering all constrained meta paths, but also can rank the contribution of different constrained meta paths.



(a) Multiple relations (b) Tenser representation

Fig. 2. An example of multi-relations of objects generated by multiple paths. (a) is the graph representation. (b) is the corresponding tensor representation.

In Fig. 2(a), we show an example of multiple relations among objects. There are three objects of type A, three objects of type B and three types of relations among them. These relations are generated by three constrained meta paths with type A as the source type and type B as the target type. To describe the multiple relations among objects, we use the representation of tensor which is a multidimensional array. We call $X = (x_{i,j,k})$ a 3rd order tensor, where $x_{i,j,k} \in R$, for $i = 1, \cdots, m$, $j = 1, \cdots, l$, $k = 1, \cdots, n$. m and n are the number of objects of type A and type B, respectively, and there are l types of relations among them. $x_{i,j,k}$ represents the times that object i is related to object k through the jth constrained meta path. For example, Fig. 2(b) is a three-way array, where each two dimensional slice represents an adjacency matrix for a single relation. So the data can be represented as a tensor of size $3 \times 3 \times 3$. In the multi-relational network, we define the transition probability tensor to present the transition probability among objects and relations.

Definition 5 ***Transition probability tensor.*** *In a multi-relational network, X is the tensor representing the network. F is the normalized tensor of X along the column vector. R is the normalized tensor of X along the tube vector. T is the normalized tensor of X along the row vector. F, R, and T are called the transition probability tensor which can be denoted as follows:*

$$
\begin{aligned}
f_{i,j,k} &= \frac{x_{i,j,k}}{\sum_{i=1}^{m} x_{i,j,k}} \qquad i = 1, 2, \ldots, m \\
r_{i,j,k} &= \frac{x_{i,j,k}}{\sum_{j=1}^{l} x_{i,j,k}} \qquad j = 1, 2, \ldots, l \\
t_{i,j,k} &= \frac{x_{i,j,k}}{\sum_{k=1}^{n} x_{i,j,k}} \qquad k = 1, 2, \ldots, n
\end{aligned} \tag{2}
$$

$f_{i,j,k}$ can be interpreted as the probability of object i (of type A) being the visiting object when relation j is used and the current object being visited is object k (of type B), $r_{i,j,k}$ represents the probability of using relation j given that object k is visited from object i, and $t_{i,j,k}$ can be interpreted as the probability of object k being visited, given that object i is currently the visiting object and relation j is used. The meaning of these three tensors can be defined formally as follows:

$$
\begin{aligned}
f_{i,j,k} &= Prob(X_t = i | Y_t = j, Z_t = k) \\
r_{i,j,k} &= Prob(Y_t = j | X_t = i, Z_t = k) \\
t_{i,j,k} &= Prob(Z_t = k | X_t = i, Y_t = j)
\end{aligned} \tag{3}
$$

in which X_t, Z_t and Y_t are three random variables representing visiting at certain object of type A or type B and using certain relation respectively at the time t.

Now, we define the stationary distributions of objects and relations as follows

$$
x = (x_1, x_2, \cdots, x_m)^T, y = (y_1, y_2, \cdots, y_l)^T, z = (z_1, z_2, \cdots, z_n)^T \tag{4}
$$

in which

$$
x_i = \lim_{t \to \infty} Prob(X_t = i), y_j = \lim_{t \to \infty} Prob(Y_t = j), z_k = \lim_{t \to \infty} Prob(Z_t = k). \tag{5}
$$

From the above equations, we can get:

$$
\begin{aligned}
Prob(X_t = i) &= \sum_{j=1}^{l} \sum_{k=1}^{n} f_{i,j,k} \times Prob(Y_t = j, Z_t = k) \\
Prob(Y_t = j) &= \sum_{i=1}^{m} \sum_{k=1}^{n} r_{i,j,k} \times Prob(X_t = i, Z_t = k) \\
Prob(Z_t = k) &= \sum_{i=1}^{m} \sum_{j=1}^{l} t_{i,j,k} \times Prob(X_t = i, Y_t = j)
\end{aligned} \tag{6}
$$

where $Prob(Y_t = j, Z_t = k)$ is the joint probability distribution of Y_t and Z_t, $Prob(X_t = i, Z_t = k)$ is the joint probability distribution of X_t and Z_t, and $Prob(X_t = i, Y_t = j)$ is the joint probability distribution of X_t and Y_t. To obtain x_i, y_j and z_k, we assume that X_t, Y_t and Z_t are all independent from each other which can be denoted as below:

$$
\begin{aligned}
Prob(X_t = i, Y_t = j) &= Prob(X_t = i) Prob(Y_t = j) \\
Prob(X_t = i, Z_t = k) &= Prob(X_t = i) Prob(Z_t = k) \\
Prob(Y_t = j, Z_t = k) &= Prob(Y_t = j) Prob(Z_t = k)
\end{aligned} \tag{7}
$$

Consequently, combining the equations with the assumptions above, we get

$$\begin{aligned} x_i &= \sum_{j=1}^{l}\sum_{k=1}^{n} f_{i,j,k} y_j z_k, \qquad i = 1, 2, \ldots, m, \\ y_j &= \sum_{i=1}^{m}\sum_{k=1}^{n} r_{i,j,k} x_i z_k, \qquad j = 1, 2, \ldots, l, \\ z_k &= \sum_{i=1}^{m}\sum_{j=1}^{l} t_{i,j,k} x_i y_j, \qquad k = 1, 2, \ldots, n. \end{aligned} \tag{8}$$

The equations above can be written in a tensor format:

$$x = Fyz, \qquad y = Rxz, \qquad z = Txy \tag{9}$$

with $\sum_{i=1}^{m} x_i = 1$, $\sum_{j=1}^{l} y_j = 1$, and $\sum_{k=1}^{n} z_k = 1$.

According to the analysis above, we can design the following algorithm to co-rank the importance of objects and relations.

Algorithm 1. HRank-CO Algorithm

Input: Three tensors F, T and R, three initial probability distributions x_0, y_0 and z_0 and the tolerance ϵ.
Output: Three stationary probability distributions x, y and z.
Procedure:
Set $t = 1$;
repeat
Compute $x_t = Fy_{t-1}z_{t-1}$;
Compute $y_t = Rx_t z_{t-1}$;
Compute $z_t = Tx_t y_t$;
until $||x_t - x_{t-1}|| + ||y_t - y_{t-1}|| + ||z_t - z_{t-1}|| < \epsilon$

4 Experiments

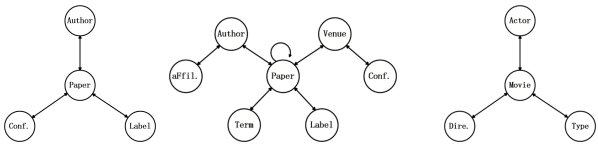
In this section, we do experiments to validate the effectiveness of two versions of HRank on three real datasets, respectively.

4.1 Datasets

We use three heterogeneous information networks for our experiments. They are summarized as follows:

DBLP Dataset [6,4]: The DBLP dataset is a sub-network collected from DBLP website [1] involving major conferences in two research areas: database (DB) and information retrieval (IR), which naturally form two labels. The dataset contains 9682 authors, 20 conferences and 22185 papers which are all labeled with one of the two research areas. The network schema is shown in Fig. 3(a).

[1] `http://www.informatik.uni-trier.de/~ley/db/`



(a) DBLP (b) ACM (c) Movie

Fig. 3. The network schema of three heterogeneous datasets. (a) DBLP bibliographic dataset. (b) ACM bibliographic dataset. (c) IMDB movie dataset.

ACM Dataset [6]: The ACM dataset was downloaded from ACM digital library [2] in June 2010. The ACM dataset comes from 14 representative computer science conferences. These conferences include 196 corresponding venue proceedings. The dataset has 12499 papers, 17431 authors, 1903 terms and 1804 author affiliations. The network also includes 73 labels of these papers in ACM category. The network schema of ACM dataset is shown in Fig. 3(b).

IMDB Dataset [7]: We crawled movie information from The Internet Movie Database [3] to construct the network. The related objects include movies, actors, directors and movie types, which are organized as a star schema shown in Fig. 3(c). Movie information includes 5324 actors, 1591 movies, 551 directors and 112 movie types.

4.2 Ranking of Heterogeneous Objects

Here, the experiments validate the effectiveness of HRank-CMP on constrained meta paths.

Experiment Study on Constrained Meta Paths. The experiments are done on the DBLP dataset. We evaluate the importance of authors and conferences simultaneously based on the meta path APC, which means authors publish papers on conferences. Two constrained meta paths ($APC|P.L =$ "DB" and $APC|P.L =$ "IR") are also included, which means authors publish DB(IR)-field papers on conferences. We employ HRank-CMP to rank the importance of authors and conferences based on these three paths. As the baseline methods, we use PageRank and the degree of authors and conferences (called Degree method). We directly run PageRank on the whole DBLP network by ignoring the heterogeneity of objects. Since the results of PageRank mix all types of objects, we select the author and conference type from the ranking list as the final results.

The top ten authors and conferences returned by these five methods are shown in Tables 1 and 2, respectively. As shown in Table 1, the ranking results of these methods on authors all are reasonable, however, the constrained meta paths can find the most influential authors in a certain field. For example, the top three

[2] `http://dl.acm.org/`
[3] `www.imdb.com/`

Table 1. Top ten authors of different methods on DBLP dataset. The number in the parenthesis of the fifth column means the rank of authors in the whole ranking list returned by PageRank.

Rank	APC	$APC\|P.L =$ "DB"	$APC\|P.L =$ "IR"	PageRank	Degree
1	Gerhard Weikum	Surajit Chaudhuri	W. Bruce Croft	W. Bruce Croft(23)	Philip S. Yu
2	Katsumi Tanaka	H. Garcia-Molina	Bert R. Boyce	Gerhard Weikum(24)	Gerhard Weikum
3	Philip S. Yu	H. V. Jagadish	Carol L. Barry	Philip S. Yu(25)	Divesh Srivastava
4	H. Garcia-Molina	Jeffrey F. Naughton	James Allan	Jiawei Han(26)	Jiawei Han
5	W. Bruce Croft	Michael Stonebraker	ChengXiang Zhai	H. Garcia-Molina(27)	H. Garcia-Molina
6	Jiawei Han	Divesh Srivastava	Mark Sanderson	Divesh Srivastava(28)	W. Bruce Croft
7	Divesh Srivastava	Gerhard Weikum	Maarten de Rijke	Surajit Chaudhuri(29)	Surajit Chaudhuri
8	Hans-Peter Kriegel	Jiawei Han	Katsumi Tanaka	H. V. Jagadish(30)	H. V. Jagadish
9	Divyakant Agrawal	Christos Faloutsos	Iadh Ounis	Jeffrey F. Naughton(31)	Jeffrey F. Naughton
10	Jeffrey Xu Yu	Philip S. Yu	Joemon M. Jose	Rakesh Agrawal(32)	Rakesh Agrawal

authors of $APC|P.L =$ "DB" are Surajit Chaudhuri, Hector Garcia-Molina and H. V. Jagadish, and all of them are very influential researchers in the database field. Similarly, as we can see in Table 2, HRank with constrained meta paths can clearly find the important conferences in DB and IR fields, while other methods mingle these conferences. For example, the most important conferences in the DB field are ICDE, VLDB and SIGMOD, while the most important conferences in the IR field are SIGIR, WWW and CIKM. Observing Tables 1 and 2, we can also find the mutual effect of authors and conferences.

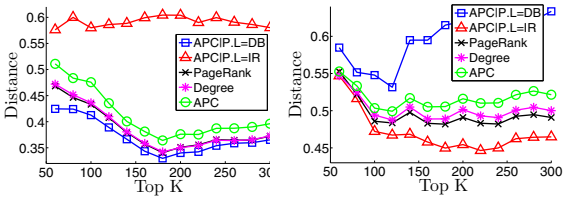
Table 2. Top ten conferences of different methods on DBLP dataset. The number in the parenthesis of the fifth column means the rank of conferences in the whole ranking list returned by PageRank.

Rank	APC	$APC\|P.L =$ "DB"	$APC\|P.L =$ "IR"	PageRank	Degree
1	CIKM	ICDE	SIGIR	ICDE(3)	ICDE
2	ICDE	VLDB	WWW	SIGIR(4)	SIGIR
3	WWW	SIGMOD	CIKM	VLDB(5)	VLDB
4	VLDB	PODS	JASIST	CIKM(6)	SIGMOD
5	SIGMOD	DASFAA	WISE	SIGMOD(7)	CIKM
6	SIGIR	EDBT	ECIR	JASIST(8)	JASIST
7	DASFAA	ICDT	APWeb	WWW(9)	WWW
8	JASIST	MDM	WSDM	DASFAA(10)	PODS
9	WISE	WebDB	JCIS	PODS(11)	DASFAA
10	EDBT	SSTD	IJKM	JCIS(12)	EDBT

Quantitative Comparison Experiments. Based on the results returned by five methods, we can obtain five candidate ranking lists of authors in DBLP dataset. To evaluate the results quantitatively, we use the author ranks from Microsoft Academic Search [4] as ground truth. Specifically, we crawled two standard ranking lists of authors in two academic fields: DB and IR. Then we use the *Distance* criterion [8] to compare the difference between our candidate ranking lists and the standard ranking lists. The criterion not only measures the number of mismatches between these two lists, but also considers the position of these mismatches. The smaller *Distance* means the smaller difference (i.e., better performance). Fig. 4 shows the differences of author ranking lists. We can observe that HRank with constrained meta paths achieve the best performances

[4] `http://academic.research.microsoft.com/`

on their corresponding field, while they have the worst performances on other fields. In addition, compared to that of PageRank and Degree, the mediocre performances of HRank with meta path APC further demonstrate the importance of constrained meta path to capture the subtle semantics contained in heterogeneous networks.



(a) DB field (b) IR field

Fig. 4. The Distances between the candidate author ranking lists and the standard ranking lists on different fields on DBLP dataset

4.3 Co-ranking of Objects and Paths

Experiment Study on Co-ranking on Symmetric Constrained Meta Paths. In this experiment, we will validate the effectiveness of HRank-CO to rank objects and symmetric constrained meta paths simultaneously. The experiment is done on ACM dataset. First we construct a (2, 1)th order tensor X based on 73 constrained meta paths (i.e., $APA|P.L = L_j, j = 1 \cdots 73$). When the ith and the kth authors co-publish a paper together, of which the label is the jth label, we add one to the entries $x_{i,j,k}$ and $x_{k,j,i}$ of X. By considering all the publications, $x_{i,j,k}$ (or $x_{k,j,i}$) refers to the number of collaborations by the ith and the kth author under the jth paper label. In addition, we do not consider any self-collaboration, i.e., $x_{i,j,i} = 0$ for all $1 \leq i \leq 17431$ and $1 \leq j \leq 73$. The size of X is $17431 \times 73 \times 17431$ and the percentage of nonzero entries is $4.126 \times 10^{-4}\%$. In this dataset, we will evaluate the importance of authors through the co-author relations, meanwhile we will analyze the importance of paths.

Table 3. Top 10 authors and constrained meta paths (note that only the constraint (L_j) of the paths ($APA|P.L = L_j, j = 1 \ldots 73$) are shown in the table)

Rank	Authors	Constrained meta paths
1	Jiawei Han	H.3 (Information Storage and Retrieval)
2	Philip Yu	H.2 (Database Management)
3	Christos Faloutsos	C.2 (Computer-Communication Networks)
4	Ravi Kumar	I.2 (Artificial Intelligence)
5	Wei-Ying Ma	F.2 (Analysis of Algorithms and Problem Complexity)
6	Zheng Chen	D.4 (Operating Systems)
7	Hector Garcia-Molina	H.4 (Information Systems Applications)
8	Hans-Peter Kriegel	G.2 (Discrete Mathematics)
9	Gerhard Weikum	I.5 (Pattern Recognition)
10	D. R. Karger	H.5 (Information Interfaces and Presentation)

Table 3 shows the top ten authors (left) and paths (right) based on their HRank values. We can find that the top ten authors are all influential researchers

in DM/IR fields, which conforms to our common senses. Similarly, the most important paths are related to DM/IR fields, such as $APA|P.L =$ "$H.3$" (Information Storage and Retrieval) and $APA|P.L =$ "$H.2$" (Database Management). Although the conferences in ACM dataset are from multiple fields, there are more papers from the DM/DB fields, which makes the authors and paths in DM/DB fields ranked higher. We can also find that the influence of authors and paths can be promoted by each other. In order to observe this point more clearly, we show the number of co-authors of the top ten authors based on the top ten paths in Table 4. We can observe that there are more collaborations for top authors in influential fields. For example, although Zheng Chen (rank 6) has more number of co-authors than Jiawei Han (rank 1), the collaborations of Jiawei Han focus on ranked higher fields (i.e., H.3 and H.2), so Jiawei Han has higher HRank score. Similarly, the top paths contain many collaborations of influential authors.

Table 4. The number that the top ten authors collaborate with others via the top ten constrained meta paths (note that only the constraint (L_j) of the paths ($APA|P.L = L_j, j = 1 \dots 73$) are shown in the first row of the table).

Ranked A/CP	1 (H.3)	2 (H.2)	3 (C.2)	4 (I.2)	5 (F.2)	6 (D.4)	7 (H.4)	8 (G.2)	9 (I.5)	10 (H.5)
1 (Jiawei Han)	51	176	0	0	0	0	9	2	2	0
2 (Philip Yu)	51	94	0	0	9	0	3	0	13	0
3 (C. Faloutsos)	17	107	0	5	9	0	3	4	2	0
4 (Ravi Kumar)	73	27	0	3	13	0	18	5	0	0
5 (Wei-Ying Ma)	132	26	0	9	0	0	2	0	30	10
6 (Zheng Chen)	172	9	0	9	0	0	22	0	38	9
7 (H. Garcia-Molina)	23	65	3	0	0	0	1	0	0	4
8 (H. Kriegel)	19	28	5	0	0	0	6	0	7	4
9 (G. Weikum)	82	14	0	4	0	0	8	0	4	0
10 (D. R. Karger)	11	5	13	0	7	4	1	7	0	7

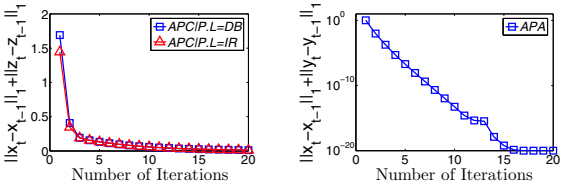
Experiment Study on Co-ranking on Asymmetric Constrained Meta Paths. The experiments on the Movie dataset aim to show the effectiveness of HRank-CO to rank heterogeneous objects and asymmetric constrained meta paths simultaneously. In this case, we construct a 3rd order tensor X based on the constrained meta paths $AMD|M.T = T_j, j = 1 \cdots 112$. That is, the tensor represents the actor-director collaboration relations on different types of movies. When the ith actor and the kth director cooperate in a movie of the jth type, we add one to the entries $x_{i,j,k}$ of X. By considering all the cooperations, $x_{i,j,k}$ refers to the number of collaborations by the ith actor and the kth director under the jth type of movie. The size of X is $5324 \times 112 \times 551$ and the percentage of nonzero entries is $7.827 \times 10^{-4}\%$.

Table 5 shows the top ten actors, directors and constrained meta paths (i.e., movie type). Basically, the results comply with our common senses. The top ten actors are well known, such as Eddie Murphy, Harrison Ford. Similarly, these directors are also famous in filmdom due to their works. These movie types obtained are the most popular movie subjects as well. In addition, we observe the mutual enhancements of the importance of objects and meta paths again. As we know, Eddie Murphy and Drew Barrymore (rank 1, 4 in actors) are famous

Table 5. Top 10 actors, directors and meta paths on IMDB dataset (note that only the constraint (T_j) of the paths ($AMD|M.T = T_j, j = 1 \ldots 112$) are shown in the table)

Rank	Actor	Director	Constrained meta path
1	Eddie Murphy	Tim Burton	Comedy
2	Harrison Ford	Zack Snyder	Drama
3	Bruce Willis	Marc Forster	Thriller
4	Drew Barrymore	David Fincher	Action
5	Nicole Kidman	Michael Bay	Adventure
6	Nicolas Cage	Ridley Scott	Romance
7	Hugh Jackman	Richard Donner	Crime
8	Robert De Niro	Steven Spielberg	Sci-Fi
9	Brad Pitt	Robert Zemeckis	Animation
10	Christopher Walken	Stephen Sommers	Fantasy

comedy and drama (rank 1, 2 in paths) actors. Higher ranked directors also prefer popular movie subjects.



(a) HRank-CMP (b) HRank-CO

Fig. 5. The difference between two successive calculated probability vectors against iterations based on the two versions of HRank

4.4 Convergence Experiments

In Fig. 5, we show the convergence of HRank on the previous experiments. The results illustrate that the two versions of HRank both quickly converge after no more than 20 iterations. In addition, we can also observe that HRank has different convergence speed in these two conditions. HRank-CMP almost converges on 9 iterations (see Fig. 5(a)). However, HRank-CO for co-ranking converges on 16 iterations (see Fig. 5(b)). We think it is reasonable, since it is more difficult to converge for more objects in HRank-CO. The time and space complexity is analyzed, and three fast computation strategies are designed to fasten the matrix multiplication process in [10].

5 Conclusions

In this paper, we first study the ranking problem in heterogeneous information network and propose the HRank method, which is a path based random walk method. In this method, we introduce the constrained meta path concept to capture the more subtle and refined semantics contained in HIN. In addition,

we further put forward a method to co-rank the paths and objects, since the paths effect the importance of objects. Experiments validate the effectiveness and efficiency of HRank on three real datasets.

Acknowledgments. This work is supported by the National Basic Research Program of China (2013CB329603). It is also supported by the National Natural Science Foundation of China (No. 61375058, 61074128, 71231002) and Ministry of Education of China and China Mobile Research Fund (MCM20123021).

References

1. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604–632 (1999)
2. Han, J.: Mining Heterogeneous Information Networks by Exploring the Power of Links. Discovery Science (2009)
3. Jeh, G., Widom, J.: Simrank: a Measure of Structural-Context Similarity. In: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543. ACM (2002)
4. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. In: VLDB, pp. 992–1003 (2011)
5. 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: KDD, pp. 1348–1356 (2012)
6. Shi, C., Kong, X., Yu, P.S., Xie, S., Wu, B.: Relevance Search in Heterogeneous Networks. In: 15th EDBT, pp. 180–191. ACM (2012)
7. Shi, C., Zhou, C., Kong, X., Yu, P.S., Liu, G., Wang, B.: HeteRecom: A Semantic-Based Recommendation System in Heterogeneous Networks. In: KDD, pp. 1552–1555 (2012)
8. Nie, Z., Zhang, Y., Wen, J.R., Ma, W.Y.: Object-level Ranking: Bringing Order to Web Objects. In: WWW, pp. 422–433 (2005)
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University Database Group (1998)
10. `http://arxiv.org/abs/1403.7315`