HAR: Hub, Authority and Relevance Scores in Multi-Relational Data for Query Search

Xutao Li* Michael K. Ng† Yunming Ye‡

Abstract

In this paper, we propose a framework HAR to study the hub and authority scores of objects, and the relevance scores of relations in multi-relational data for query search. The basic idea of our framework is to consider a random walk in multi-relational data, and study in such random walk, limiting probabilities of relations for relevance scores, and of objects for hub scores and authority scores. The main contribution of this paper is to (i) propose a framework (HAR) that can compute the hub, authority and relevance scores by solving limiting probabilities arising from multi-relational data, and can incorporate input query vectors to handle query-specific search; (ii) show existence and uniqueness of such limiting probabilities so that they can be used for query search effectively; and (iii) develop an iterative algorithm to solve a set of tensor (multivariate polynomial) equations to obtain such probabilities. Extensive experimental results on TREC and DBLP data sets suggest that the proposed method is very effective in obtaining relevant results to the querying inputs. In the comparison, we find that the performance of HAR is better than those of HITS, SALSA and TOPHITS.

1 Introduction

PageRank [18] and HITS [11] are two significant link analysis algorithms for determining the importance of objects in a graph. The PageRank scores are given by the entries of the principal eigenvector of a Markov matrix of objects transition probabilities across the entire graph. The PageRank score depends only on the topology of the graph and do not depend on the query. Topic/query-sensitive PageRank is also proposed and developed in [9]. On the other hand, HITS [11] first evaluates a focused subgraph, computes the principal singular vectors of the adjacency matrix of the focused subgraph, and then obtain authorities and hubs scores for answering the query. The HITS score is query-specific in that it computes the authority scores of the objects after a focused subgraph is used. In these two approaches, a single relation type is focused and studied.

There are many information retrieval and data mining applications where multiple relation types are involved. In such applications, objects have interactions with the others based on different relations. For example, researchers (or papers) cite the other researchers (or the other papers) based on different concepts/keywords [17], proteins interact with the other proteins under different conditions, webpages link to the other webpages via different semantic meanings [12], resource description format (RDF) resources connect to the other RDF resources through different RDF predicates [6]. We refer to the data of this type as multi-relational data. One natural way to represent the multi-relational data is using tensors, also known as multi-dimensional arrays. In Figure 1, we show an example of multi-relational data with five objects and three relations and its corresponding tensor representation. In the figure, an $5 \times 5 \times 3$ three-dimensional array is used, where (i, j, k) entry is nonzero if the jth object connects to the ith object with the kth relation.

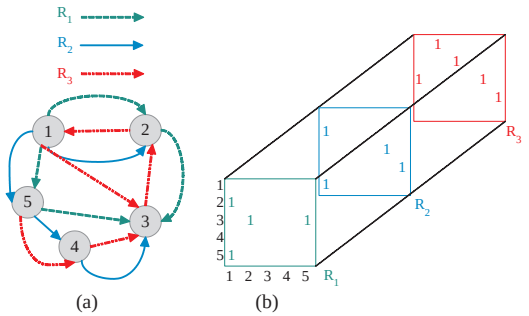

Figure 1: (a) An example of a multi-relational data in a graph representation and (b) the corresponding tensor representation.

In this paper, we propose a framework to study the hub and authority scores of objects in multi-relational

*Department of Computer Science, Shenzhen Graduate School, Harbin Institute of Technology. Email:xutaolee08@gmail.com

†Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong. Email: mng@math.hkbu.edu.hk

‡Department of Computer Science, Shenzhen Graduate School, Harbin Institute of Technology. Email:yeyunming@hit.edu.cn

data for query search. Besides hub and authority scores for each object, we assign a relevance score for each relation to indicate its importance in multi-relational data. In our proposal, these three scores have the following mutually-reinforcing relationships. (i) An object that points to many objects with high authority scores through relations of high relevance scores, receives a high hub score. (ii) An object that is pointed by many objects with high hub scores through relations of high relevance scores, receives a high authority score. (iii) A relation that is connected in between objects with high hub scores and high authority scores, receives a high relevance score.

Our idea is to compute hub and authority scores of objects and relevance scores of relations by considering a random walk in a multi-relational data/tensor, and studying the limiting probabilities arriving objects as hubs or as authorities, and using relations respectively. More specifically, we construct transition probability tensors for objects as hubs or as authorities and for relations, and then set up a set of tensor (multivariate polynomial) equations following the mutually-reinforcing relationship among hubs, authorities and relations described in the above mechanisms (i), (ii) and (iii). We obtain hub and authority scores for objects and relevance scores for relations by solving tensor equations. In order to handle query-specific search, we incorporate an input hub, authority or relation vector into the tensor equations to answer such query.

The main contribution of this paper can be summarized as follows. (i) We propose a framework (HAR) that can compute the hub, authority and relevance scores by solving limiting probabilities arising from multi-relational data, and can incorporate input query vectors to handle query-specific search. (ii) We show existence and uniqueness of such limiting probabilities so that they can be used for query search effectively. (iii) We develop an iterative algorithm to solve a set of tensor (multivariate polynomial) equations to obtain such probabilities. Extensive experimental results on TREC and DBLP data sets suggest that the proposed method is very effective in obtaining relevant results to the querying inputs. In the comparison, we find that the performance of HAR is better than those of HITS, SALSA and TOPHITS.

The rest of the paper is organized as follows. In Section 2, we review some related work. In Section 3, we describe notations in this paper and some preliminary knowledge. In Section 4, we present the proposed framework. In Section 5, we analyze the proposed methodology. In Section 6, we show and discuss the experimental results for real-world data sets. In Section 7, we give some concluding remarks and mention some future research work.

2 Related Work

Improvement of HITS can be found by considering probabilistic latent sematic indexing [3], using a weighted in-degree analysis [15] and utilizing document cluster information and language models [14]. The main challenge of these approaches is required to set suitable weights in the edges in order to obtain effective ranking results. However, the weights are not known in general. The novelty of this paper is to provide a framework to determine the weights automatically. In [5], a co-HITS algorithm was proposed to incorporate the bipartite graph with the content information from both sides as well as the constraints of relevance. The score propagation is the mutual reinforcement to boost co-linked entities on the graph only. We see from (4.10) in Section 4 that our method involves mutual reinforcement among all the hubs, authorities and relations.

In the literature, tensor factorization is a generalized approach for analyzing multi-way data [22, 16, 23, 24]. The query search problem in multi-relational data has been also studied recently. The main idea is to approximate a tensor by a low-rank decomposition and make use of the decomposition vectors to handle query search. Sun et al. [25] applied a 3-way Tucker decomposition [27] to the analysis of user, query-term and webpage data in order to personalize web search. Rendle el al. proposed a tensor factorization model to exploit the ternary relationships in tagging data and personalize the tag recommender [19]. Kolda et al. [13, 12] proposed TOPHITS by adding a third dimension to form an adjacency tensor that incorporates anchor text information, and employ a Three-way Parallel Factors (PARAFAC) decomposition [2, 8] to compute the singular vectors for query processing. Acar et al. [1] used various tensor decompositions of user, keyword and time data to separate different streams of conservations in chatroom data. In these methods, we need to select the number of decompositions (low-rank approximation) in the tensor factorization. The number of decompositions may not be known in advance. On the other hand, the computation of tensor factorization may not be unique as there are several numerical methods (e.g., the alternating least squares procedure) used to compute such factorization and the factorization results depend on the initial guess. There is no detailed algorithmic and mathematical analysis for the convergence of the method. Also the computational cost may be expensive for very large tensors [4].

Different from these methods, we compute limiting probabilities of tensors for hub and authority scores of objects and relevance scores of relations to handle the

query search problem. We will show that such probabilities can be unique and computed very efficiently. In [17], we have developed a MultiRank algorithm for coranking both objects and relations together in a multi-relational data by considering undirected edges in multiple relations only. In this paper, we consider directed edges in multiple relations, and analyze both hub and authority scores in the new setting. Recently, Harth and Kinsella [7] used the power method to compute scoring vectors for subjects, objects, predicates and context relationships in semantic web in a fourth order tensor. However, the mathematical analysis of such scoring vectors is not given. We remark that we can extend our framework to cover a higher-order tensor e.g., subject-object-predicate-context relationships in semantic web, and study the theoretical properties of scoring vectors as limiting probabilities, see the concluding remarks in Section 7.

3 Preliminary

In this section, we describe notations and present some preliminary knowledge on tensors. As we analyze objects under multiple relations and also consider interaction between relations based on objects, we make use of rectangular tensors to represent them.

Let $\mathbb{R}$ be the real field. We call $\mathcal{T} = (t_{i_1,i_2,j_1})$ where $t_{i_1,i_2,j_1} \in \mathbb{R}$, for $i_k = 1, \cdots, m$, $k = 1, 2$ and $j_1 = 1, \cdots, n$, a real $(2,1)$th order $(m \times n)$-dimensional rectangular tensor. In this setting, we refer (i_1, i_2) to be the indices for objects and j_1 to be the indices for relations. For instance, five objects ($m = 5$) and three relations ($n = 3$) are used in the example in Figure 1. When there is a link from the i_1th object to the i_2th object when the j_1th relation is used, we set $t_{i_1,i_2,j_1} = 1$, otherwise $t_{i_1,i_2,j_1} = 0$. In addition, $\mathcal{T}$ is called nonnegative if $t_{i_1,i_2,j_1} \geq 0$.

Let $\mathbf{u}$, $\mathbf{v}$ be vectors of length m and $\mathbf{w}$ be a vector of length n. Let $[\mathcal{T}\mathbf{uw}]_1$ and $[\mathcal{T}\mathbf{uw}]_2$ be vectors in $\mathbb{R}^m$ such that

$$([\mathcal{T}\mathbf{uw}]_1)_{i_1} = \sum_{i_2=1}^{m}\sum_{j_1=1}^{n} t_{i_1,i_2,j_1} u_{i_2} w_{j_1},\ i_1 = 1, 2, \cdots, m,$$

or

$$([\mathcal{T}\mathbf{uw}]_2)_{i_2} = \sum_{i_1=1}^{m}\sum_{j_1=1}^{n} t_{i_1,i_2,j_1} u_{i_1} w_{j_1},\ i_2 = 1, 2, \cdots, m.$$

Similarly, $[\mathcal{T}\mathbf{uv}]_3$ is a vector in $\mathbb{R}^n$ such that

$$([\mathcal{T}\mathbf{uv}]_3)_{j_1} = \sum_{i_1=1}^{m}\sum_{i_2=1}^{m} t_{i_1,i_2,j_1} u_{i_1} v_{i_2},\ j_1 = 1, 2, \cdots, n.$$

As we consider a random walk in a nonnegative rectangular tensor arising from multi-relational data, and study the likelihood that we will arrive at any particular object as a hub or as an authority, and use at any particular relation, we construct three transition probability tensors $\mathcal{H} = (h_{i_1,i_2,j_1})$, $\mathcal{A} = (a_{i_1,i_2,j_1})$ and $\mathcal{R} = (r_{i_1,i_2,j_1})$ with respect to hubs, authorities and relations by normalizing the entry of $\mathcal{T}$ as follows:

$$\begin{aligned}
h_{i_1,i_2,j_1} &= \frac{t_{i_1,i_2,j_1}}{\sum_{i_1=1}^{m} t_{i_1,i_2,j_1}}, \quad i_1 = 1, 2, \cdots, m,\\
a_{i_1,i_2,j_1} &= \frac{t_{i_1,i_2,j_1}}{\sum_{i_2=1}^{m} t_{i_1,i_2,j_1}}, \quad i_2 = 1, 2, \cdots, m,\\
r_{i_1,i_2,j_1} &= \frac{t_{i_1,i_2,j_1}}{\sum_{j_1=1}^{n} t_{i_1,i_2,j_1}}, \quad j_1 = 1, 2, \cdots, n.
\end{aligned}$$

These numbers gives the estimates of the following conditional probabilities:

$$\begin{aligned}
h_{i_1,i_2,j_1} &= \text{Prob}[X_t = i_1 | Y_t = i_2, Z_t = j_1]\\
a_{i_1,i_2,j_1} &= \text{Prob}[Y_t = i_2 | X_t = i_1, Z_t = j_1]\\
r_{i_1,i_2,j_1} &= \text{Prob}[Z_t = j_1 | Y_t = i_2, X_t = i_1]
\end{aligned}$$

where X_t, Y_t and Z_t are random variables referring to visit at any particular object as a hub and as an authority, and to use at any particular relation at the time t, respectively. Here the time t refers to the time step in the random walk. h_{i_1,i_2,j_1} (or a_{i_1,i_2,j_1}) can be interpreted as the probability of visiting the i_1th (or i_2th) object as a hub (or as an authority) by given that the i_2th (or i_1th) object as an authority (or as a hub) is currently visited and the j_1th relation is used, and r_{i_1,i_2,j_1} can be interpreted as the probability of using the j_1th relation given that the i_2th object as an authority is visited from the i_1th object as a hub. We remark that the construction of a_{i_1,i_2,j_1} is related to the transpose of h_{i_1,i_2,j_1}. This is similar to the construction of SALSA algorithm to incorporate the link structure among objects for the role of hub and authority in the single relation data [15].

We note that if t_{i_1,i_2,j_1} is equal to 0 for all $1 \leq i_1 \leq m$, this is called the dangling node [18], and the values of h_{i_1,i_2,j_1} can be set to $1/m$. The same construction is for $a_{i_1,i_2,j}$. Similarly, if t_{i_1,i_2,j_1} is equal to 0 for all $1 \leq j_1 \leq n$, then the values of r_{i_1,i_2,j_1} can be set to $1/n$. With the above construction, we have

$$0 \leq h_{i_1,i_2,j_1} \leq 1,\ 0 \leq a_{i_1,i_2,j_1} \leq 1,\ 0 \leq r_{i_1,i_2,j_1} \leq 1,$$
$$\sum_{i_1=1}^{m} h_{i_1,i_2,j_1} = 1,\ \sum_{i_2=1}^{m} a_{i_1,i_2,j_1} = 1,\ \sum_{j_1=1}^{n} r_{i_1,i_2,j_1} = 1.$$

We call $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$ transition probability tensors which is similar to transition probability matrices in Markov chain [21]. In addition, it is necessary for us to know the connectivity among the objects and the relations within a tensor.

DEFINITION 3.1. *A $(2,1)$th order nonnegative rectangular tensor $\mathcal{T}$ is called irreducible if $(t_{i_1,i_2,j})$ (m-by-m matrices) for $j = 1, 2, \cdots, n$ are irreducible. If $\mathcal{T}$ is not irreducible, then we call $\mathcal{T}$ reducible.*

When $\mathcal{T}$ is irreducible, any two objects in multi-relational data can be connected via some relations. As we would like to determine the importance of both objects and relations simultaneously in multi-relational data, irreducibility is a reasonable assumption that we will use in the following discussion. It is clear that when $\mathcal{T}$ is irreducible, the corresponding tensors $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$ are also irreducible.

4 The Proposed Framework

Given three transition probability tensors $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$, we study the following conditional probabilities:

$$\begin{aligned}(4.1)\qquad &\text{Prob}[X_t = i_1] \\ &= \sum_{i_2=1}^{m}\sum_{j_1=1}^{n} h_{i_1,i_2,j_1} \times \text{Prob}[Y_t = i_2, Z_t = j_1]\end{aligned}$$

$$\begin{aligned}(4.2)\qquad &\text{Prob}[Y_t = i_2] \\ &= \sum_{i_1=1}^{m}\sum_{j_1=1}^{n} a_{i_1,i_2,j_1} \times \text{Prob}[X_t = i_1, Z_t = j_1]\end{aligned}$$

$$\begin{aligned}(4.3)\qquad &\text{Prob}[Z_t = j_1] \\ &= \sum_{i_1=1}^{m}\sum_{i_2=1}^{m} r_{i_1,i_2,j_1} \times \text{Prob}[X_t = i_1, Y_t = i_2]\end{aligned}$$

where $\text{Prob}[Y_t = i_2, Z_t = j_1]$ is the joint probability distribution of Y_t and Z_t, $\text{Prob}[X_t = i_1, Z_t = j_1]$ is the joint probability distribution of X_t and Z_t, and $\text{Prob}[X_t = i_1, Y_t = i_2]$ is the joint probability distribution of X_t and Y_t. In our approach, we consider the limiting probability distributions of objects as hubs and authorities, and relations, i.e., we are interested in hub and authority scores of objects

$$\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m]^T, \quad \bar{\mathbf{y}} = [\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_m]^T$$

and relevance scores of relations given by

$$\bar{\mathbf{z}} = [\bar{z}_1, \bar{z}_2, \cdots, \bar{z}_n]^T$$

respectively, with

$$\bar{x}_{i_k} = \lim_{t\to\infty} \text{Prob}[X_t = i_k], \quad \bar{y}_{j_k} = \lim_{t\to\infty} \text{Prob}[Y_t = i_k]$$

for $1 \le i_k \le m$, and

$$\bar{z}_{j_k} = \lim_{t\to\infty} \text{Prob}[Z_t = j_k]$$

for $1 \le j_k \le n$. In order to obtain $\bar{x}_{i_1}$, $\bar{y}_{i_1}$ and $\bar{z}_{j_1}$, we assume that the limiting joint probability distribution can be approximated by the individual limiting product distributions, i.e., the solutions are in tensor-product forms:

$$\begin{aligned}(4.4)\quad &\lim_{t\to\infty} \text{Prob}[Y_t = i_2, Z_t = j_1] &&= \bar{y}_{i_2}\bar{z}_{j_1} \\ (4.5)\quad &\lim_{t\to\infty} \text{Prob}[X_t = i_1, Z_t = j_1] &&= \bar{x}_{i_1}\bar{z}_{j_1} \\ (4.6)\quad &\lim_{t\to\infty} \text{Prob}[X_t = i_1, Y_t = i_2] &&= \bar{x}_{i_1}\bar{y}_{i_2}\end{aligned}$$

The tensor-product form solution has been considered in [7, 17]. Therefore, by making t goes to infinity, (4.1), (4.2) and (4.3) becomes

$$(4.7)\quad \bar{x}_{i_1} = \sum_{i_2=1}^{m}\sum_{j_1=1}^{n} h_{i_1,i_2,j_1}\bar{y}_{i_2}\bar{z}_{j_1}, \quad i_1 = 1, 2, \cdots, m,$$

$$(4.8)\quad \bar{y}_{i_2} = \sum_{i_1=1}^{m}\sum_{j_1=1}^{n} a_{i_1,i_2,j_1}\bar{x}_{i_2}\bar{z}_{j_1}, \quad i_2 = 1, 2, \cdots, m,$$

$$(4.9)\quad \bar{z}_{j_1} = \sum_{i_1=1}^{m}\sum_{i_2=1}^{m} r_{i_1,i_2,j_1}\bar{x}_{i_1}\bar{y}_{i_2}, \quad j_1 = 1, 2, \cdots, n.$$

We see from (4.7), (4.8) and (4.9) that the hub (or authority) score of an object is defined implicitly and depends on the number and authority (or hub) metric of all objects that have multiple relations to (or from) this object, and also the relevance values of these multiple relations. Similarly, the relevance score of a relation is defined implicitly and depends on which the objects to be linked and their hub and authority scores of these objects. It is clear that an object that is linked with high relevance score of relations from (or to) many objects with high hub (or authority) scores, receives a high authority (or hub) score itself. Also a relation that is linked with objects with high hub and authority scores, receives a high relevance score itself. It is interesting to note from (4.7), (4.8) and (4.9) that under the tensor operation, we solve the following tensor (multivariate polynomial) equations:

$$(4.10)\qquad [\mathcal{H}\bar{\mathbf{y}}\bar{\mathbf{z}}]_1 = \bar{\mathbf{x}}, \quad [\mathcal{A}\bar{\mathbf{x}}\bar{\mathbf{z}}]_2 = \bar{\mathbf{y}}, \quad [\mathcal{R}\bar{\mathbf{x}}\bar{\mathbf{y}}]_3 = \bar{\mathbf{z}},$$

with

$$\sum_{i_1=1}^{m} \bar{x}_{i_1} = 1, \quad \sum_{i_2=1}^{m} \bar{y}_{i_2} = 1, \quad \sum_{j_1=1}^{n} \bar{z}_{j_1} = 1. \tag{4.11}$$

For simplicity, we will drop the bracket and the number ($[\cdot]_1$,$[\cdot]_2$ and $[\cdot]_3$) for the above tensor-product operations in the following discussion.

When we consider a single relation type, we can set $\bar{\mathbf{z}}$ to be a vector $\mathbf{l}/n$ of all ones in (4.10), and thus we obtain two matrix equations $\mathcal{H}\bar{\mathbf{y}}\mathbf{l}/n = \bar{\mathbf{x}}$ and $\mathcal{A}\bar{\mathbf{x}}\mathbf{l}/n = \bar{\mathbf{y}}$. We remark that $\mathcal{A}$ can be viewed as the transpose of $\mathcal{H}$. This is exactly the same as that we solve for the singular vectors to get the hub and authority scoring vectors in SALSA. As a summary, the proposed framework HAR is a generalization of SALSA to deal with multi-relational data.

4.1 Query Search for Objects and Relations To deal with query processing, we need to compute hub and authority scores of objects and relevance scores of relations with respect to a query input. Motivated by the idea of topic-sensitive PageRank [9] and random walk with restart [26], we consider this issue by assigning the desired limiting probability distributions towards a query input. More specifically, we modify the tensor equations in (4.10) as follows:

$$\begin{aligned} (1-\alpha)\mathcal{H}\bar{\mathbf{y}}\bar{\mathbf{z}} + \alpha\mathbf{o} &= \bar{\mathbf{x}}, \\ (1-\beta)\mathcal{A}\bar{\mathbf{x}}\bar{\mathbf{z}} + \beta\mathbf{o} &= \bar{\mathbf{y}}, \\ (1-\gamma)\mathcal{R}\bar{\mathbf{x}}\bar{\mathbf{y}} + \gamma\mathbf{r} &= \bar{\mathbf{z}}, \end{aligned} \tag{4.12}$$

with (4.11), where $\mathbf{o}$ and $\mathbf{r}$ are two assigned probability distributions that are constructed from a query input. Here ($\sum_{i=1}^{m}[\mathbf{o}]_i = 1$ and $\sum_{j=1}^{n}[\mathbf{r}]_j = 1$), and $0 \leq \alpha, \beta, \gamma < 1$, are three parameters for controlling the importance of the assigned probability distributions in query input to the resulting hub and authority scores of objects and the relevance scores of relations.

Given a query input, one simple way is to construct $\mathbf{o}$ and $\mathbf{r}$ by using uniform distributions on the interesting objects and relations respectively, or by normalizing the weights of the interesting objects and relations respectively. Note that a query input can be composed of either objects or relations, or composed of both. Moreover, if we would like to emphasize the objects or relations satisfying some requirements in the query search, we assign higher probabilities to these objects or relations.

4.2 The Algorithm In this subsection, we present an efficient iterative algorithm to solve the tensor equations in (4.12) to obtain $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ for the hub and authority scores of objects and the relevance scores of relations. The HAR algorithm is summarized in Algorithm 1.

Algorithm 1 The HAR Algorithm

Input: Three tensors $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$, two initial probability distributions $\mathbf{y}_0$ and $\mathbf{z}_0$ with ($\sum_{i=1}^{m}[\mathbf{y}_0]_i = 1$ and $\sum_{j=1}^{n}[\mathbf{z}_0]_j = 1$), the assigned probability distributions of objects and/or relations $\mathbf{o}$ and $\mathbf{r}$ ($\sum_{i=1}^{m}[\mathbf{o}]_i = 1$ and $\sum_{j=1}^{n}[\mathbf{r}]_j = 1$), three weighting parameters $0 \leq \alpha, \beta, \gamma < 1$, and the tolerance ϵ

Output: Three limiting probability distributions $\bar{\mathbf{x}}$ (hub scores), $\bar{\mathbf{y}}$ (authority scores) and $\bar{\mathbf{z}}$ (relevance values)

Procedure:

1: Set $t = 1$;
2: Compute $\mathbf{x}_t = (1-\alpha)\mathcal{H}\mathbf{y}_{t-1}\mathbf{z}_{t-1} + \alpha\mathbf{o}$;
3: Compute $\mathbf{y}_t = (1-\beta)\mathcal{A}\mathbf{x}_t\mathbf{z}_{t-1} + \beta\mathbf{o}$;
4: Compute $\mathbf{z}_t = (1-\gamma)\mathcal{R}\mathbf{x}_t\mathbf{y}_t + \gamma\mathbf{r}$;
5: If $||\mathbf{x}_t - \mathbf{x}_{t-1}|| + ||\mathbf{y}_t - \mathbf{y}_{t-1}|| + ||\mathbf{z}_t - \mathbf{z}_{t-1}|| < \epsilon$, then stop, otherwise set $t = t + 1$ and goto Step 2.

In Algorithm 1, the HAR computations require several iterations, through the collection to adjust approximate hub and authority scores of objects and relevance scores of relations to more closely reflect their theoretical true values (underlying limiting probability distributions). The main computational cost of the HAR algorithm depends on the cost of performing tensor operations in Steps 2, 3 and 4. Assume that there are $O(N)$ nonzero entries in $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$, the cost of these tensor calculations are of $O(N)$ arithmetic operations.

5 Theoretical Analysis

In this section, we show existence and uniqueness of limiting probability distributions $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ so that it can be used in computing hub and authority scores for objects and relevance scores for relations very effectively. Based on these results, the convergence of HAR algorithm can be shown.

We let $\Omega_m = \{\mathbf{u} = (u_1, u_2, \cdots, u_m) \in R^m | u_i \geq 0, 1 \leq i \leq m, \sum_{i=1}^{m} u_i = 1\}$ and $\Omega_n = \{\mathbf{w} = (w_1, w_2, \cdots, w_n) \in R^n | w_j \geq 0, 1 \leq j \leq n, \sum_{j=1}^{n} w_j = 1\}$. We also set

$$\Omega = \{[\mathbf{x}, \mathbf{y}, \mathbf{z}] \in R^{2m+n} | \mathbf{x} \in \Omega_m, \mathbf{y} \in \Omega_m, \mathbf{z} \in \Omega_n\}.$$

We note that Ω_m, Ω_n and Ω are closed convex sets. We call $\mathbf{u}$ to be positive (denoted by $\mathbf{u} > 0$) if all its entries of u_i are positive.

It is easy to check that if $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are probability distributions, then the output $\mathcal{H}\mathbf{yz}$, $\mathcal{A}\mathbf{xz}$ and $\mathcal{R}\mathbf{xy}$ are also probability distributions (the correctness of Steps 2, 3 and 4 in Algorithm 1).

THEOREM 5.1. *Suppose $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$ are constructed in Section 3, $0 \leq \alpha, \beta, \gamma < 1$, and $\mathbf{o} \in \Omega_m$ and $\mathbf{r} \in \Omega_n$ are given. For any $\mathbf{x}, \mathbf{y} \in \Omega_m$ and $\mathbf{z} \in \Omega_n$, then $(1-\alpha)\mathcal{H}\mathbf{yz} + \alpha\mathbf{o}, (1-\alpha)\mathbf{Axz} + \alpha\mathbf{o} \in \Omega_m$ and $(1-\beta)\mathcal{R}\mathbf{xy} + \beta\mathbf{r} \in \Omega_n$.*

By using Theorem 5.1, we show the existence of positive solutions for the set of tensor equations in (4.12).

THEOREM 5.2. *Suppose $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$ are constructed in Section 3, $0 \leq \alpha$, β, $\gamma < 1$, and $\mathbf{o} \in \Omega_m$ and $\mathbf{r} \in \Omega_n$ are given. If $\mathcal{T}$ is irreducible, then there exist $\bar{\mathbf{x}} > 0$, $\bar{\mathbf{y}} > 0$ and $\bar{\mathbf{z}} > 0$ such that $(1-\alpha)\mathcal{H}\bar{\mathbf{y}}\bar{\mathbf{z}} + \alpha\mathbf{o} = \bar{\mathbf{x}}$, $(1-\beta)\mathcal{A}\bar{\mathbf{x}}\bar{\mathbf{z}} + \beta\mathbf{o} = \bar{\mathbf{y}}$, and $(1-\gamma)\mathcal{R}\bar{\mathbf{x}}\bar{\mathbf{y}} + \gamma\mathbf{r} = \bar{\mathbf{z}}$, with $\bar{\mathbf{x}}, \bar{\mathbf{y}} \in \Omega_m$ and $\bar{\mathbf{z}} \in \Omega_n$.*

Proof. The problem can be reduced to a fixed point problem as follows. We define the following mapping $T : \Omega \to \Omega$ as follows

$$\begin{aligned} T([\mathbf{x}, \mathbf{y}, \mathbf{z}]) &= [(1-\alpha)\mathcal{H}\mathbf{yz} + \alpha\mathbf{o}, \\ &\quad (1-\beta)\mathcal{A}\mathbf{xz} + \beta\mathbf{o}, \\ &\quad (1-\gamma)\mathcal{R}\mathbf{xy} + \gamma\mathbf{r}]. \end{aligned}$$

It is clear that T is well-defined (i.e., when $[\mathbf{x}, \mathbf{y}, \mathbf{z}] \in \Omega$, $T([\mathbf{x}, \mathbf{y}, \mathbf{z}]) \in \Omega$) and continuous. According to the Brouwer Fixed Point Theorem, there exists $[\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}}] \in \Omega$ such that $T([\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}}]) = [\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}}]$, i.e., $(1-\alpha)\mathcal{H}\bar{\mathbf{y}}\bar{\mathbf{z}} + \alpha\mathbf{o} = \bar{\mathbf{x}}$, $(1-\beta)\mathcal{A}\bar{\mathbf{x}}\bar{\mathbf{z}} + \beta\mathbf{o} = \bar{\mathbf{y}}$, and $(1-\gamma)\mathcal{R}\bar{\mathbf{x}}\bar{\mathbf{y}} + \gamma\mathbf{r} = \bar{\mathbf{z}}$.

Now suppose $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ are not positive, i.e., there exist some entries of $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ are zero. Let $I_1 = \{i_1 | \bar{x}_{i_1} = 0\}$, $I_2 = \{i_2 | \bar{y}_{i_2} = 0\}$ and $J_1 = \{j_1 | \bar{z}_{j_1} = 0\}$. Again I_1 and I_2 are proper subsets of $\{1, 2, \cdots, m\}$ and J is a proper subset of $\{1, 2, \cdots, n\}$. Let

$$\begin{aligned} \delta = \min\{&\min\{\bar{x}_{i_1} | i_1 \notin I_1\}, \\ &\min\{\bar{y}_{i_2} | i_2 \notin I_2\}, \\ &\min\{\bar{z}_{j_1} | j_1 \notin J_1\}\}. \end{aligned}$$

We must have $\delta > 0$.

We first note that

$$(1-\alpha)\sum_{i_2=1}^{m}\sum_{j_1=1}^{n} h_{i_1,i_2,j_1}\bar{y}_{i_2}\bar{z}_{j_1} + \alpha o_{i_1} = \bar{x}_{i_1} = 0, \quad \forall i_1 \in I_1.$$

Let us consider the following quantity:

$$\begin{aligned} & (1-\alpha)\delta^2 \sum_{i_2 \notin I_2}\sum_{j_1 \notin J_1} h_{i_1,i_2,j_1} \\ \leq\ & (1-\alpha)\sum_{i_2 \notin I_2}\sum_{j_1 \notin J_1} h_{i_1,i_2,j_1}\bar{y}_{i_2}\bar{z}_{j_1} \\ \leq\ & (1-\alpha)\sum_{i_2=1}^{m}\sum_{j_1=1}^{n} h_{i_1,i_2,j_1}\bar{y}_{i_2}\bar{z}_{j_1} + \alpha o_{i_1} = 0, \forall i_1 \in I_1. \end{aligned}$$

Hence we have $h_{i_1,i_2,j_1} = 0$ for all $i_1 \in I_1$ and for all $i_2 \notin I_2$ for any fixed $j_1 \notin J_1$. Thus (h_{i_1,i_2,j_1}) $(j_1 \notin J_1)$ is a reducible matrix, and it implies that $\mathcal{H}$ is reducible. By using the similar argument and considering the other two equations $(1-\beta)\mathcal{A}\bar{\mathbf{x}}\bar{\mathbf{z}} + \beta\mathbf{o} = \bar{\mathbf{y}}$, and $(1-\gamma)\mathcal{R}\bar{\mathbf{x}}\bar{\mathbf{y}} + \gamma\mathbf{r} = \bar{\mathbf{z}}$, we can find that $\mathcal{A}$ and $\mathcal{R}$ are also reducible. According to these results, we obtain a contradiction. Hence $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ must be positive.

In [10], it has been given a general condition which guarantees the uniqueness of the fixed point in the Brouwer Fixed Point Theorem, namely, (i) 1 is not an eigenvalue of the Jacobian matrix of the mapping, and (ii) for each point in the boundary of the domain of the mapping, it is not a fixed point. In our case, we have shown in Theorem 5.2 that all the fixed points of T are positive when $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$ are irreducible, i.e., they do not lie on the boundary $\partial\Omega$ of Ω.

THEOREM 5.3. *Suppose $\mathcal{T}$ is irreducible, $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$ constructed in Section 3, $0 \leq \alpha, \beta, \gamma < 1$ and $\mathbf{o} \in \Omega_m$ and $\mathbf{r} \in \Omega_n$ are given. If 1 is not the eigenvalue of the Jacobian matrix of T, then the solution vectors $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ in Theorem 5.2 are unique.*

According to Theorem 5.3, if $\mathbf{x}_t = \mathbf{x}_{t-1}$, $\mathbf{y}_t = \mathbf{y}_{t-1}$ and $\mathbf{z}_t = \mathbf{z}_{t-1}$, in the HAR algorithm, then we obtain the unique solution vectors $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ for $(1-\alpha)\mathcal{H}\bar{\mathbf{y}}\bar{\mathbf{z}} + \alpha\mathbf{o} = \bar{\mathbf{x}}$, $(1-\beta)\mathcal{A}\bar{\mathbf{x}}\bar{\mathbf{z}} + \beta\mathbf{o} = \bar{\mathbf{y}}$, and $(1-\gamma)\mathcal{R}\bar{\mathbf{x}}\bar{\mathbf{y}} + \gamma\mathbf{r} = \bar{\mathbf{z}}$. When $\mathbf{x}_t \neq \mathbf{x}_{t-1}$, $\mathbf{y}_t \neq \mathbf{y}_{t-1}$ and $\mathbf{z}_t \neq \mathbf{z}_{t-1}$, there exist a subsequence $[\mathbf{x}_{t_s}, \mathbf{y}_{t_s}, \mathbf{z}_{t_s}]$ converges to $[\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}}]$ by using the fact that Ω is a compact space in R^{2m+n}. As we have shown that the solution vectors are unique, it implies that $[\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t]$ converges (up to a subsequence) to $[\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}}]$ which are the limiting probability vectors giving the hub and authority scores of objects and the relevance scores of relations respectively.

6 Experimental Results

In the section, we carry out two experiments to demonstrate the usefulness and effectiveness of the proposed method HAR. In the first experiment, we construct multi-relational data by considering links of different anchor texts among webpages from TREC Web data set. In the second experiment, our multi-relational data is constructed under the consideration of paper citations of different categories from DBLP data set. Both experiments concentrate on the query search tasks. For comparisons, we apply HITS, SALSA [15] and TOPHITS as well.

6.1 Evaluation metrics In this paper, we employ four evaluation metrics: the precision at position k (P@k), the normalized discounted cumulative gain at

position k (NDCG@k), the mean average precision (MAP) and the R-precision (R-prec).

1. P@k: Given a particular query q, we compute the precision at position k as follows:

$$P@k = \frac{\#\{\text{relevant documents in top } k \text{ results}\}}{k}$$

Since we have a set of queries in both experiments, we report the average P@k scores of these queries as final results.

2. NDCG@k: In order to emphasize the high-ranking relevant documents, the discounted cumulative gain DCG@k is defined as a measure to evaluate the effectiveness of search engine results. DCG@k discounts the contribution of low-ranking relevant documents and is calculated as

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(1+i)}$$

where $rel_i \in \{0, 1\}$ indicates whether the document ranked number i is relevant to the query topic or not. NDCG@k is a normalized version of this measure:

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

where IDCG@k refers to the ideal discounted cumulative gain at position k, which is computed by presenting an ideal ranking list. Similar to P@k, we report the average NDCG@k scores of all queries as final results.

3. MAP: Given a query, the average precision is calculated by averaging the precision scores at each position in the search results where a relevant document is found. MAP is then the mean of the average precision scores of all queries.

4. R-prec: Given a query, R-prec is the precision score after R documents are retrieved, i.e., R-prec=P@R, where R is the total number of relevant documents for such query. Again, we report the mean of R-prec scores of all queries as final results.

6.2 Experiment 1 In this experiment, we sample 100,000 webpages from .GOV Web collection in 2002 TREC. For query testing, we use the 50 topic distillation topics in TREC 2003 Web track as queries[1]. We preserve the webpages that are relevant to these query topics (516 webpages in total) when we sample the 100,000 webpages. Then we consider the links among them via different anchor texts. Given each anchor text, we preprocess it by eliminating the stop words and stemming. After that, we obtain 39,255 anchor terms in total, and 479,122 links with these anchor terms among the 100,000 webpages. We then use this data to construct a tensor and test our approach for query search task.

Tensor construction. We construct a tensor $\mathcal{T}$ based on links of webpages (objects) through different anchor temrs (relations). In this case, there are 100,000 objects and 39,255 relations. The tensor is constructed as follows: If the i_1th webpage links to the i_2th webpage via the j_1th anchor term, we set the entry t_{i_1,i_2,j_1} of $\mathcal{T}$ to be one. By considering all the links between these webpages, we construct the tensor $\mathcal{T}$. The size of $\mathcal{T}$ is $100,000 \times 100,000 \times 39,255$ and there are 479,122 nonzeros entries in $\mathcal{T}$. The percentage of nonzero entries is $1.22 \times 10^{-7}\%$. After constructing $\mathcal{T}$, we construct the transition probability tensors $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$. We note that only the nonzero entries and their locations are stored in the computational process. It is not necessary to store values $1/m$ or $1/n$ for the dangling nodes in $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$.

Query settings. For all the comparison algorithms, we use the well-known BM25 model [20] to configure their query settings. The BM25 score of a document d with respect to a query q is computed by summing up the weight of those query terms that occur in the document:

$$W(d, q) = \sum_t w_t * q_t$$

the weight of the term t is computed as

$$w_t = \frac{(k_1 + 1) * tf_t}{k_1((1-b) + b\frac{dl}{avdl}) + tf_t} * log\frac{N - df_t + 0.5}{df_t + 0.5}$$

where tf_t is the frequency that the term t occurs in this document, df_t is the number of documents that contain the term t, dl is the length of this document, $avdl$ is the average length of documents, N is the total number of documents in the collection, k_1 and b are two parameters. We use $k_1 = 2.1$ and $b = 0.6$. Here we list several algorithms for comparison.

- For HITS and SALSA, we aggregate the multi-relational data into a simple graph, and then compute the BM25 scores of anchor terms with respect to the query topic and select 50 webpages that have the highest BM25 scores as a root set. Then we expand the root set based on aggregated

[1]The query topics can be found in http://trec.nist.gov/data/t12.web.html. We do not use the testing collection in TREC 2002 is because Hawking and Craswell point out that the testing collection in TREC 2002 is not a good relevance judgement for the topic distillation task [28].

graph to construct a subgraph for computing the hub and authority scores.

- For TOPHITS, we perform it on the tensor $\mathcal{T}$ and try the best 500-rank, 1000-rank and 1500-rank approximations respectively. The query vector is constructed by computing the BM25 score for each anchor term(i.e., relation) with respect to the query topic. Then the hub and authority scores are computed as in [12].
- In [29], Wu et al. combine the evidence from anchor terms and query-dependent link connections to handle topic distillation tasks, and show that this method produces state-of-the-art results. Therefore, we compare with this method as well. In this method, a weighted linear model is used to combine the BM25 scores of anchor terms with query-dependent indegree and out degree information. We refer to this method as BM25+DepInOut. We use the default weights reported in [29], i.e., the value 0.2 for the BM25 score, the value 0.1 for the dependent indegree information and the value 0.7 for the dependent outdegree information.
- For HAR, we use two query settings, i.e., query with relations only, and query with both relations and objects together. For query with relations, we set the entries of $\mathbf{r}$ to be the BM25 scores between the corresponding anchor terms and the query topic, and then normalize it. In this case, the parameters α, β and γ are set to be 0, 0 and 0.9 respectively. For query with both relations and objects, in addition to the construction of $\mathbf{r}$, we set the entries of $\mathbf{o}$ to be the BM25 scores between the content of the corresponding webpages and the query topic, and then normalize it. In this case, we set the parameters α, β and γ to be 0.5, 0.5 and 0.9 respectively. We set the stopping criterion ϵ to be 10^{-7} which is small enough for convergence.

Results and comparisons. Table 1 shows results of all the comparison algorithms. The results show that our proposed HAR outperforms the other algorithms significantly. We see that the results of HITS, SALSA and TOPHITS are not very good. For HITS and SALSA, this is because they can not differentiate multi-relations and they suffer from topic drift problems when constructing the subgraph from the aggregated graph to compute hub and authority scores. For TOPHITS, this is because it employs the decomposition results of the adjacency tensor $\mathcal{T}$ to compute hub and authority. However, the decomposition vectors suffer from two constraints: one is they are usually not unique and the other is negative entries may exist

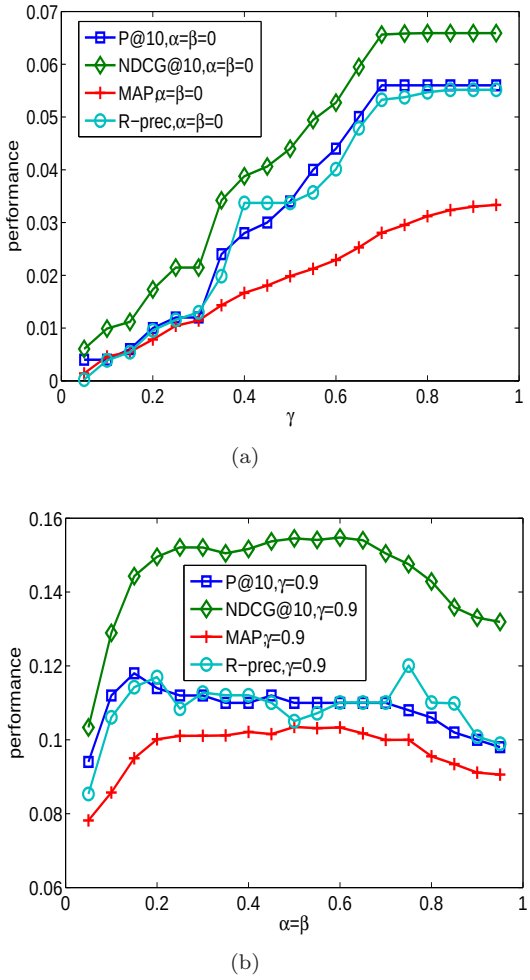

Figure 2: The parameter tuning test (a) tuning γ with $\alpha = \beta = 0$; (b) tuning α and β with $\gamma = 0.9$. For clearness, we do not show the curves evaluated with P@5, P@20, NDCG@5 and NDCG@20.

in these vectors. These two constraints may lead to non-unique or negative hub and authority scores for a particular query, which is unreasonable and hard to interpret. Different from TOPHITS, our method HAR employs the limiting probabilities to compute the hub and authority scores, which are unique and non-negative. Therefore it yields much better results than TOPHITS. Moreover, we see that HAR has slightly better performance than BM25+DepInOut when we query with relations only. When we consider query with both relations and objects, HAR has significant better performance than BM25+DepInOut. This is because it exploits all the useful information by combining the webpage content, the anchor terms and the link structure in a natural way.

	P@5	P@10	P@20	NDCG@5	NDCG@10	NDCG@20	MAP	R-prec
HITS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0041	0.0000
SALSA	0.0000	0.0160	0.0140	0.0000	0.0157	0.0203	0.0114	0.0084
TOPHITS (500-rank)	0.0040	0.0020	0.0010	0.0068	0.0044	0.0028	0.0008	0.0002
TOPHITS (1000-rank)	0.0080	0.0040	0.0020	0.0136	0.0088	0.0057	0.0016	0.0010
TOPHITS (1500-rank)	0.0080	0.0040	0.0030	0.0097	0.0063	0.0049	0.0011	0.0018
BM25+DepInOut	0.0400	0.0280	0.0180	0.0424	0.0419	0.0479	0.0370	0.0370
HAR (rel. query)	0.0480	0.0560	0.0410	0.0507	0.0659	0.0747	0.0330	0.0552
HAR (rel. and obj. query)	**0.1360**	**0.1100**	**0.0800**	**0.1398**	**0.1545**	**0.1765**	**0.1035**	**0.1051**

Table 1: The results of all comparison algorithms on TREC data set.

Parameters tuning study. Next we show how the performance changes with respect to the values of parameters α, β and γ. Figure 2(a) shows how the performance of HAR change when we tune the parameter γ. In this case, we consider query with relations only and set $\alpha = \beta = 0$. We see from this figure that the performance is increasing as we increase the value of γ, and it tends to be stable after $\gamma = 0.9$. Therefore we set γ to be 0.9 in all the experiments when we query with relations only. Figure 2(b) shows how the performance of HAR changes against the values of α and β. In this case, we consider query with both relations and objects. For simplicity, we set α and β to be equal, and then tune them. We see from this figure that the performance first increases then decreases when the values of α and β are increased. Therefore we set α and β to be 0.5 in all the experiments when we query with both relations and objects.

6.3 Experiment 2 In this experiment, we construct the multi-relational data with a data set crawled from DBLP. We crawled publication information of five conferences (SIGKDD, WWW, SIGIR, SIGMOD, CIKM) from DBLP[2]. Their publication periods are as follows: SIGKDD (1999-2010), WWW (2001-2010), SIGIR (2000-2010), SIGMOD (2000-2010) and CIKM (2000 and 2002-2009)[3]. Publication information includes title, authors, reference list, and classification categories associated with this publication[4]. There are in total 6848 publications, 10305 authors and 617 different categories in the data set. For query search task, we select 100 category concepts as query inputs to retrieve the relevant publications.

Tensor construction. We construct a tensor $\mathcal{T}$ based on citations of publications (objects) through different category concepts (relations). In this case there are 6848 objects and 617 relations. The tensor is constructed as follows: If the i_1th publication cites the i_2th publication and the i_2th publication has the j_1th category concept, then we set the entry t_{i_1,i_2,j_1} of $\mathcal{T}$ to be one, otherwise we set the entry t_{i_1,i_2,j_1} to be zero. By considering all the publications, t_{i_1,i_2,j_1} refers to the citation information from the i_1th publication to the i_2th publication via different category concept. The size of the tensor $\mathcal{T}$ is $6848 \times 6848 \times 617$ and there are 24901 nonzero entries. The percentage of the nonzero entries is $8.61 \times 10^{-5}\%$. After that, we generate transition probability tensors $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$. Again, it is not necessary to store values $1/m$ or $1/n$ for dangling nodes in $\mathcal{H}$, $\mathcal{A}$ and $\mathcal{R}$.

Query settings. Given the multi-relational data and 100 query concepts, we compare the performance of HITS, SALSA, TOPHITS and HAR. Since the query concept can exactly match the relations, we do not need to use BM25 function to compute the similarities between relations and the query concept.

For HITS and SALSA, we aggregate the multi-relational data as in experiment 1, and then select 50 publications that have the highest citations via the query concept as root set. Then we expand the root set based on the aggregated graph to construct a subgraph for computing hub and authority scores. For TOPHITS, we use 50-rank, 100-rank and 150-rank approximations respectively. The query vector is constructed by setting the entry that corresponds to the query concept to be one. For BM25+DepInOut, we use the same setting as in experiment 1. For HAR, we construct vector $\mathbf{r}$ by setting the entry that corresponds to the query concept to be one, and use parameters $\alpha = \beta = 0, \gamma = 0.9$. We set the stopping criterion ϵ to be 10^{-7} which is small enough for convergence.

Results and comparisons. Table 2 shows the results of HITS, SALSA, TOPHITS, BM25+DepInOut and our proposed HAR. Again, we see from this table

[2]http://www.informatik.uni-trier.de/ ley/db/

[3]Missing information for CIKM 2001 is due to fact that DBLP does not provide links to ACM Digital Library.

[4]For each publication, there are several strings indicating the classification categories of this publication, where each string provides the information from the most general concept to the most specific concept. For example, a string may be "H. information systems— H.3 information storage and retrieval— H.3.3 information search and retrieval". For each string, we choose the most specific concept as the classification category it indicates for the publication.

	P@5	P@10	P@20	NDCG@5	NDCG@10	NDCG@20	MAP	R-prec
HITS	0.2920	0.2260	0.1815	0.4122	0.3789	0.3792	0.2522	0.2751
SALSA	0.5240	0.4100	0.3105	0.6157	0.5606	0.5352	0.3462	0.3929
TOPHITS (50-rank)	0.1700	0.1360	0.1145	0.1958	0.1684	0.1557	0.0566	0.0617
TOPHITS (100-rank)	0.2080	0.1640	0.1340	0.2345	0.2012	0.1857	0.0646	0.0732
TOPHITS (150-rank)	0.2400	0.1920	0.1410	0.2649	0.2315	0.1998	0.0732	0.0765
BM25+DepInOut	0.0140	0.0170	0.0145	0.0118	0.0147	0.0138	0.0162	0.0109
HAR (rel. query)	**0.7280**	**0.5880**	**0.4155**	**0.8113**	**0.7472**	**0.6760**	**0.4731**	**0.4683**

Table 2: The results of all comparison algorithms on DBLP data set.

	P@5	P@10	P@20	NDCG@5	NDCG@10	NDCG@20	MAP	R-prec
HAR (rel. query)	0.2000	0.2000	0.2000	0.1312	0.1488	0.1647	0.1312	0.2075
HAR (rel. and obj. query)	**0.8000**	**0.6000**	**0.7500**	**0.8688**	**0.6995**	**0.7697**	**0.5422**	**0.6226**

Table 3: The results of HAR with two settings when we query "clustering" concept and "document" related papers. In this case, we judge a paper to be relevant if it has "clustering" concept and a "document" related title for evaluation.

HAR $\alpha = \beta = 0, \gamma = 0.9$, query by "clustering" concept only		
publication titles	target concept	target title
Agglomerative clustering of a search engine query log.	1	0
Information-theoretic co-clustering.	1	0
Clustering user queries of a search engine.	1	0
Random walks on the click graph.	1	0
Co-clustering documents and words using bipartite spectral graph partitioning.	1	1
Learning to cluster web search results.	1	0
Discovering evolutionary theme patterns from text: an exploration of temporal text mining.	1	0
Efficient clustering of high-dimensional data sets with application to reference matching.	1	0
Corpus structure language models and ad hoc information retrieval.	1	0
Document clustering based on non-negative matrix factorization.	1	1
An optimal and progressive algorithm for skyline queries.	1	0
Efficient similarity search and classification via rank aggregation.	1	0
Eliminating noisy information in web pages for data mining.	1	0
Fast and effective text mining using linear-time document clustering.	1	1
Evaluating strategies for similarity search on the web.	1	0
Using web structure for classifying and describing web pages.	1	0
Entropy-based subspace clustering for mining numerical data.	1	0
Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering.	1	0
Clustering by pattern similarity in large data sets.	1	0
Evaluation of hierarchical clustering algorithms for document datasets.	1	1
HAR $\alpha = \beta = 0.5, \gamma = 0.9$, query by "clustering" concept and also input the vector $\mathbf{o}$		
publication titles	target concept	target title
Document clustering based on non-negative matrix factorization.	1	1
Evaluation of hierarchical clustering algorithms for document datasets.	1	1
Co-clustering documents and words using bipartite spectral graph partitioning.	1	1
Fast and effective text mining using linear-time document clustering.	1	1
Corpus structure language models and ad hoc information retrieval.	1	0
Information-theoretic co-clustering.	1	0
Regularizing ad hoc retrieval scores.	1	0
Document clustering using word clusters via the information bottleneck method.	1	1
Document clustering with cluster refinement and model selection capabilities.	1	1
Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering.	1	0
As we may perceive: finding the boundaries of compound documents on the web.	1	1
An information-theoretic measure for document similarity.	1	1
A hierarchical monothetic document clustering algorithm for summarization and browsing search results.	1	1
On the merits of building categorization systems by supervised clustering.	1	0
A neighborhood-based approach for clustering of linked document collections.	1	1
Detecting similar documents using salient terms.	1	1
Document clustering with committees.	1	1
Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization.	1	1
Evolutionary document summarization for disaster management.	1	1
A matrix density based algorithm to hierarchically co-cluster documents and words.	1	1

Table 4: The top twenty results when querying by "clustering" concept and "document" related objects. The publications, appearing commonly in top twenty lists of two HAR results, are indicated with blue color. The target concept column shows whether the corresponding publication has the query concept according to its classification categories. The target title column shows whether the corresponding publication has a "document" related title.

that the performance of HAR is much better than those of the other comparison algorithms.

In addition, we would like to show that the proposed method HAR can incorporate additional information to achieve different query purposes. For example, when we query "clustering" concept and "document" related papers, it may be difficult to set query input vectors for relations to express such concepts since there are no category concept labeled as "document". In this case, we construct the object vector $\mathbf{o}$ by setting the entries that correspond to the papers whose titles contain the word "document" to be ones and then normalizing it. After that, we apply the HAR algorithm by using "clustering" as relation query input and the object vector $\mathbf{o}$ in the search. Table 4 shows the top twenty papers when we apply HAR with two query settings. Table 3 shows the evaluation of these two results. We see that the results by using both relations and objects as input are more accurate than those by using "clustering" concept as input only.

In summary, the experimental results have shown that our proposed method HAR has better performance than HITS, SALSA and TOPHITS. It is shown that HAR is flexible to incorporate additional information in query search process. Theoretically, our approach has two advantages compared with TOPHITS: (i) HAR results in the unique limiting probability distributions of authorities, hubs and relevance scores, while TOPHITS employs tensor factorization results which may not be optimal; (ii) HAR is easier to interpret query results because the resulting limiting probabilities are all positive but tensor factorization may result in vectors with negative entries, which are hard to interpret.

7 Concluding Remarks

In this paper, we have proposed a framework HAR to determine hub and authority scores of objects and relevance scores of relations in multi-relational data for query search. Both experimental and theoretical results have shown the effectiveness of the proposed method. We have also demonstrated how to accomplish various query objectives using different settings of the proposed method. In the comparison, we find that the performance of HAR is better than those of HITS, SALSA and TOPHITS. We believe that our framework would be very useful in information retrieval tasks in practice. Here we point out several possible future research directions based on the proposed framework.

1. In our framework, we assume probability distributions satisfying (4.4), (4.5) and (4.6) to set up the tensor equations for hub, authority and relevance scores. It is interesting to study other forms of equations for this objective.

2. We set values for weighting parameters α, β and γ empirically in this paper. In order to make use the HAR algorithm more practically and effectively, these parameters can be learned via training data sets.

3. We consider in this paper the multi-relational data are represented as a $(2,1)$th order rectangular tensor. However, our framework is a general paradigm and it can be further extended to consider data with higher order tensors for potential applications. For example, we can consider the query search problem in semantic web using a $(1,1,1,1)$th order rectangular tensor to represent subject, object, predicate and context relationship as in [7]. By constructing four transition probability tensors $\mathcal{S}$, $\mathcal{O}$, $\mathcal{P}$ and $\mathcal{R}$ for subject, object, predicate and context relationship respectively. Based on the proposed framework, we expect to solve the following set of tensor equations:

$$\mathcal{S}\mathbf{opr} = \mathbf{s},\ \mathcal{O}\mathbf{spr} = \mathbf{o},\ \mathcal{P}\mathbf{sor} = \mathbf{p},\ \mathcal{R}\mathbf{sop} = \mathbf{r}.$$

In the HAR algorithm, we compute $\mathbf{s}_t = \mathcal{S}\mathbf{o}_{t-1}\mathbf{p}_{t-1}\mathbf{r}_{t-1}$, $\mathbf{o}_t = \mathcal{O}\mathbf{s}_t\mathbf{p}_{t-1}\mathbf{r}_{t-1}$, $\mathbf{p}_t = \mathcal{P}\mathbf{s}_t\mathbf{o}_t\mathbf{r}_{t-1}$, $\mathbf{r}_t = \mathcal{R}\mathbf{s}_t\mathbf{o}_t\mathbf{p}_t$ in the iterations. Similarly, assigned probability vectors can be incorporated into the above tensor equations for query input. By using the mathematical analysis in Section 5, we obtain the four limiting probability distributions corresponding to subject, object, predicate and context relationship for scoring vectors in query search.

8 Acknowledgement

X. Li's research was supported in part by NSFC under Grant no.61100190. Y. Ye's research was supported in part by NSFC under Grant no.61073195, Shenzhen Science and Technology Program under Grant no. CXB201005250024A and ZD201006100018A, and Natural Scientific Research Innovation Foundation in HIT under Grant no. HIT.NSFIR.2010128. M. Ng's research was supported in part by Centre for Mathematical Imaging and Vision, HKRGC grant and HKBU FRG grant.

References

[1] E. Acar, S. Camtepe, M. Krishnamoorthy, and B. Yener. Modeling and multiway analysis of chatroom tensors. *Intelligence and Security Informatics*, pages 256–268, 2005.

[2] J. Carroll and J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[3] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proc. of ICML*, pages 167–174, 2000.

[4] P. Comon, X. Luciani, and A. De Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, 23(7-8):393–405, 2009.

[5] H. Deng, M. Lyu, and I. King. A generalized Co-HITS algorithm and its application to bipartite graphs. In *Proc. of ACM SIGKDD*, pages 239–248. ACM, 2009.

[6] T. Franz, A. Schultz, S. Sizov, and S. Staab. Triplerank: Ranking semantic web data by tensor decomposition. *The Semantic Web-ISWC 2009*, pages 213–228, 2009.

[7] D. Galway and I. Park. TOPDIS: Tensor-based Ranking for Data Search and Navigation. 2009.

[8] R. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis. *UCLA working papers in phonetics*, 16(1):84, 1970.

[9] T. Haveliwala. Topic-sensitive PageRank. In *Proceedings of WWW*, 2002.

[10] R. Kellogg. Uniqueness in the Schauder fixed point theorem. *Proceedings of the American Mathematical Society*, 60(1):207–210, 1976.

[11] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[12] T. Kolda and B. Bader. The TOPHITS model for higher-order web link analysis. In *Workshop on Link Analysis, Counterterrorism and Security*, volume 7, pages 26–29, 2006.

[13] T. Kolda, B. Bader, and J. Kenny. Higher-order web link analysis using multilinear algebra. In *Data Mining, Fifth IEEE International Conference on*, page 8. IEEE, 2005.

[14] O. Kurland and L. Lee. Respect my authority!: HITS without hyperlinks, utilizing cluster-based language models. In *Proc.of ACM SIGIR*, pages 83–90. ACM, 2006.

[15] R. Lempel and S. Moran. SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, 19(2):131–160, 2001.

[16] Y. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher. MetaFac: community discovery via relational hypergraph factorization. In *Proceedings of the 15th ACM SIGKDD*, pages 527–536. ACM, 2009.

[17] M. Ng, X. Li, and Y. Ye. MultiRank: co-ranking scheme for objects and relations in multi-dimensional data. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1998.

[19] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 727–736. ACM, 2009.

[20] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241. Springer-Verlag New York, Inc., 1994.

[21] S. Ross. *Introduction to probability models*. Academic Pr, 2007.

[22] J. Sun, S. Papadimitriou, C. Lin, N. Cao, S. Liu, and W. Qian. Multivis: Content-based social network exploration through multi-way visual analysis. volume 9, pages 1063–1074, 2009.

[23] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *Proceedings of the 12th ACM SIGKDD*, pages 374–383. ACM, 2006.

[24] J. Sun, D. Tao, S. Papadimitriou, P. Yu, and C. Faloutsos. Incremental tensor analysis: Theory and applications. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(3):1–37, 2008.

[25] J. Sun, H. Zeng, H. Liu, Y. Lu, and Z. Chen. CubeSVD: a novel approach to personalized Web search. In *Proceedings of the 14th WWW*, pages 382–390. ACM, 2005.

[26] H. Tong, C. Faloutsos, and J. Pan. Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327–346, 2008.

[27] L. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[28] E. Voorhees, D. Harman, N. I. of Standards, and T. (US). *TREC: Experiment and evaluation in information retrieval*. MIT press USA, 2005.

[29] M. Wu, F. Scholer, and A. Turpin. Topic distillation with query-dependent link connections and page characteristics. *ACM Transactions on the Web (TWEB)*, 5(2):6, 2011.