Scientific Ranking over Heterogeneous Academic Hypernetwork

Ronghua Liang[1] Xiaorui Jiang[2]*

[1] School of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China
[2] Department of Computer Science, Aston University, Birmingham, B4 7ET, UK

[1] rhliang@zjut.edu.cn [2] jiangx4@aston.as.uk

Abstract

Ranking is an important way of retrieving authoritative papers from a large scientific literature database. Current state-of-the-art exploits the flat structure of the heterogeneous academic network to achieve a better ranking of scientific articles, however, ignores the multinomial nature of the multi-dimensional relationships between different types of academic entities. This paper proposes a novel mutual ranking algorithm based on the multinomial heterogeneous academic hypernetwork, which serves as a generalized model of a scientific literature database. The proposed algorithm is demonstrated effective through extensive evaluation against well-known IR metrics on a well-established benchmarking environment based on the ACL Anthology Network.

Introduction

The tremendous scientific advancements, which result in a fast growing scientific literature database, have made automatic scientific ranking a much more important issue than ever before. Researchers need to know what papers are the must-read classics to gain a solid understanding of a scientific domain and, besides that, what papers are the most valuable for them to read to catch up with the current research fronts. However it is a great challenge for researchers to answer the above questions efficiently due to the explosive number of scientific publications. For example the ACL Anthology for the small subarea computational linguistics has indexed 18041 papers published by 14386 authors in 273 conferences or journals by 2011. It is so complex for humans to maneuver this dataset manually that there is an urgent need for effective automatic ranking.

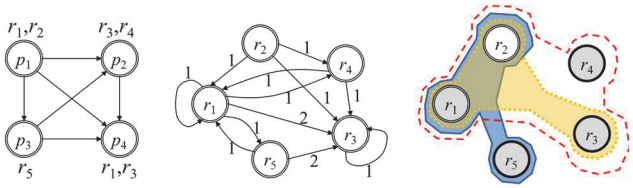
For the purpose automatic scientific ranking, a number of graph-based algorithms have been proposed. A recent trend in graph-based algorithms is to employ the structure of the heterogeneous academic network for an improved ranking (Zhou et al., 2007; Sayyadi and Getoor, 2009; Das et al., 2011; Yan et al., 2011; Wang et al., 2013). The most important assumption of these algorithm is that the authority of the papers and the importance of their authors (as well as publication venues) are mutually reinforced, that is a paper may be more authoritative if it is written by influential researchers and a researcher gains more importance from its papers that are regarded as influential.

Despite of their success in improving ranking effectiveness, graph-based algorithms all work on simple networks, either homogeneous or heterogeneous, which limits their abilities to model the real-life situation and thus their ranking performance. A toy example helps to understand this. For the scientific literature database in Figure 1(a), current works use simple network to model author[1] citations as in Figure 2(b). A paper citation results in several author citations, each of which links a citing author to a cited author in the author citation network. As a matter of fact, these author citation relations are just low-dimensional projections of a multinomial citation relation between the authors of the citing and cited papers. A more natural way is to model these multinomial citations using a *directed hypernetwork*[2] (Ducournao & Bretto, 2014). Figure 1(c) shows a subnetwork of the author citation hypernetwork. The red-dashed, blue-solid and yellow-dotted polygons represent three citation hyperedges. Double-lined and shaded vertices denote the tail and head vertices respectively. In such a hypernetwork, an author citation *hyperedge* links the set of authors of a citing paper to the set of authors of a cited paper. For example, the fact that p_1 cites p_4 results in the blue-solid hyperedge. Its tail vertices (i.e. citing authors) and head vertices (i.e. cited authors) are $\{r_1, r_2\}$ and $\{r_4, r_5\}$ respectively. Note that this is a case of author self-citation as r_1 acts as both a citing author and cited author.

We argue that using hypernetwork as a better model for the multinomial relationships between academic entities,

* Corresponding author: Xiaorui Jiang (jiangx4@aston.ac.uk)

[1] Author and researcher are used interchangeably throughout the paper.
[2] Although network and graph are typically used for the directed and undirected case respectively in discrete mathematics, this paper interchangeably uses hypernetwork and hypergraph.



(a) metadata (b) author citation network (c) three hyperedges

Figure 1. An exemplar Heterogeneous Academic Network.

we may get a more reasonable ranking result. As an illustration, applying PageRank to both networks in Figure 1(b) and Figure 1(c) results in the following researcher rankings: $\mathbf{r}_{1(b)}$ = [r_1:0.09, r_2:0.03, r_3:0.78, r_4:0.05, r_5:0.05] and $\mathbf{r}_{1(c)}$ = [r_1:0.33, r_2:0.03, r_3:0.41, r_4:0.11, r_5:0.11]. We can see that the importance of r_1 has been elevated by using author citation hypernetwork. This is reasonable because r_1 not only receives many citations but also citations from important researchers like r_3. We argue that PageRank on author citation hypernetwork returns a more reasonable ranking result than on simple author citation network.

Actually there also exist multinomial multidimensional relationships between different types of academic entities. For instance, a few researchers may have co-authored a number of papers. Current methods project these multimonial multidimensional relations to a number of simple binary relations which lead to information loss and thus poor ranking effectiveness. To overcome the above deficiencies, this paper proposes to use heterogeneous hypernetwork as a generalized model for the multinomial and multidimensional academic network, resulting in a *heterogeneous academic hypernetwork*, and proposes a better mutual ranking algorithm for scientific articles by employing the mutual reinforcement relationships between academic entities in the heteroegeneous academic hypernetwork.

Related Work

Graph-based algorithms have recently been extensively applied to scientific ranking. Chen et al. (2007) was the first to apply Google's PageRank algorithm to the citation network to find the most prestigious papers. Later, Liu et al. (2008) proposed a PageRank-style personalized ranking algorithm for scientific publications by introducing time factors into the personalization vector for an acknowledgement of time recency. Almost at the same time, Walker et al. (2007) proposed the CiteRank algorithm which was derived in a different way where the rank of a paper is defined as the authority aggregated from other nodes through an authority flow downstream in the citation network as in Eq. (1) where α is the fraction of score flowing downward the citation network represented by $\mathbf{W}$. An important aspect of CiteRank is the setting of personalization towards each paper based on its publication time in Eq. (2) where t_{now} and t_i are respectively the current time and the publication time of the i-th paper in term of year or month, and τ is an adjustable decay factor.

$$\mathbf{s} = \mathbf{I}\cdot\mathbf{p} + (1-\alpha)\cdot\mathbf{W}\cdot\mathbf{p} + (1-\alpha)^2\cdot\mathbf{W}\cdot\mathbf{p} + \cdots \quad (1)$$

$$p_i = e^{-(t_{now}-t_i)/\tau} \quad (2)$$

Li et al. (2008) also considered a personalization strategy similar to CiteRank for favoring recent articles. Later, Yan et al. (2009) proposed to use the weighted PageRank to the coauthorship network and studied the effect of different damping factors. Radicchi et al. (2009) proposed SARA on author citation network, which works in quite a similar way to CiteRank on paper citation network. In their essence, CiteRank and SARA can be rewritten into an elegant matrix formulation which is almost equivalent to PageRank. All the above methods build and work on homogeneous network of only one type of academic entities.

Indeed, author importance and venue prestige were thought to have a significant impact on the assignment of paper authority, especially for those published recently without having attracted enough incoming citations. The above intuition was realized by recent efforts in heterogeneous graph based ranking of multi-type, multi-relational data. The seminal work CoRank proposes ranking framework based on a combination of intra-network (homogeneous) random walk in the citation and coauthorship network as well as the inter-network (heterogeneous) random walk between papers and its authors in the authorship network (Zhou et al., 2007). A large portion of later works followed or extended the similar idea of CoRank (Deng et al., 2009, Sayyadi and Getoor 2009; Yan et al., 2011; Ng et al., 2011; Jiang et al., 2012; Wang et al., 2013) while differed from CoRank in that they all employed only one-step random walk in both the homogeneous and heterogeneous networks. Deng et al. (2009) proposed the Co-HITS algorithm to bipartite graphs with one of its variant based on random walk works just in a similar way to the inter-network random walk in CoRank. The FutureRank algorithm (Sayyadi and Getoor 2009) was different from CoRank in that it did not rely on either the coauthorhisp or author citation network. The algorithms proposed by Yan et al. (2011) and Wang et al. (2013) also relied on the mutual reinforcement between papers and venues without considering homogeneous relationships between either authors or venues, while these information was used in Das et al. (2011).

Method

Heterogeneous Academic Hypernetwork

Most previous studies model the multinomial relationships between (the same type of) entities using undirected (homogeneneous) hypernetwork. To accommodate more gen-

eralized circumstances such as author citation network, this paper, inspired by Bellaachia and Al-Dhelaan (2015) and Ducournao and Bretto (2014), extends hypernetwork definition to the directed case as follows. A directed homogeneous hypernetwork is denoted as $\vec{\mathcal{H}} = (V, \vec{E})$ where (1) $V = V^+ \cup V^-$ denotes the set of vertices with V^+ and V^- called the set of tail and head vertices respectively; (2) $\vec{E} = \{\vec{e} \mid \vec{e} = (V^+(\vec{e}), V^-(\vec{e}))\}$ denotes the set of directed hyperedges with $V^+(\vec{e}) \subseteq V^+$ and $V^-(\vec{e}) \subseteq V^-$ being the sets of tail and head vertices that are incident to $\vec{e}$. Note that we allow $V^+(\vec{e}) \cap V^-(\vec{e}) \neq \varnothing$. $W_e(\vec{e})$ denotes the weight of hyperedge $\vec{e}$ and $\mathbf{W}_e = \mathbf{diag}(W(\vec{e}))$ is the diagonal matrix of hyperedge weights. As in Bellaachia and Al-Dhelaan (2015), we also allow vertices not only hyperedges to be weighted. Thus, a directed hypernetwork $\vec{\mathcal{H}}$ can be represented by two weighted adjacency matrices $\mathbf{W}_v^+$ and $\mathbf{W}_v^-$ where $W_v^+(v, \vec{e})$ and $W_v^-(v, \vec{e})$) denote the weight of the tail and head vertex of hyperedge $\vec{e}$ respectively. Correspondingly there are two Boolean adjacency matrices $\mathbf{H}_v^+$ and $\mathbf{H}_v^-$ .Superscripts "[P]" and "[R]" are used to distinguish between paper citation hypernetwork $\vec{\mathcal{H}}^{[P]} = (V^{[P]}, \vec{E}^{[P]})$ and author citation hypernetwork $\vec{\mathcal{H}}^{[R]} = (V^{[R]}, \vec{E}^{[P]})$.

This paper also considers to employ the multinomial relationship between papers and authors to boost scientific ranking performance. We use $\mathcal{H}^{[PR]} = (V^{[PR]}, E^{[PR]})$ to denote the heterogeneous hypernetwork where (1) the superscript "[PR]" indicates $\mathcal{H}^{[PR]}$ is a heterogeneous hypernetwork capturing the multinomial relationships between Papers and Researchers; (2) $V^{[PR]} = V^{[P]} \cup V^{[R]}$ is the vertex set and $E^{[PR]} = \{e^{[PR]} \mid e^{[PR]} = (V^{[P]}(e^{[PR]}), V^{[R]}(e^{[PR]}))\}$ denotes the edge set. For an undirected heterogeneous hyperedge $e^{[PR]}$, $V^{[P]}(e^{[PR]})$ and $V^{[R]}(e^{[PR]})$ denote the subsets of papers and researchers that are incident to hyperedge $e^{[PR]}$ respectively. $\mathcal{H}^{[PR]}$ is represented by a $|V^{[P]}|$-by-$|E^{[PR]}|$ weighted adjacency matrix $\mathbf{W}_v^{[P/R]}$ and a $|V^{[R]}|$-by-$|E^{[PR]}|$ matrix $\mathbf{W}_v^{[R/P]}$, which indicate the weights of a paper and a researcher in a heterogeneous hyperedge $e^{[PR]}$ respectively, and two corresponding binary adjacency matrices $\mathbf{H}_e^{[P/R]}$ and $\mathbf{H}_e^{[R/P]}$. Each heterogeneous hyperedge $e^{[PR]}$ has a weight $W_e^{[PR]}(e^{[PR]})$ and a diagonal edge weight matrix $\mathbf{W}_e^{[PR]} = \mathbf{diag}(W_e^{[PR]}(e^{[PR]}))$.

Heterogeneous Academic Hypernetwork Ranking

Based on the above definitions of the heterogeneous academic hypernetworks, our scientific literature ranking algorithm is formulated based on the mutual reinforcement assumptions as follows. Each paper p_i is assigned two importance values, the authority $pa(i)$ and the $ph(i)$, and the importance of each researcher r_k is denoted as $ri(k)$. Thus for all the papers and researchers we have three ranking vectors **pa**, **ph**, and **ri**. Intuitively, one purpose of hub is for assessing a paper's "soundness", that is whether a paper covers enough authoritative related work. The other is for introducing a mechanism of backward flow (Jiang et al., 2013) from the cited to citing papers, otherwise old papers or dangling papers (without outgoing citations caused by dataset cutoff) will absorb the ranking values.

The idea of the ranking algorithm is explained using the assignment of paper authority vector as follows.

Computing the Paper Authority Vector

The computation is done at an iterative fashion. The authority $pa^{(t+1)}(i)$ of paper p_i at time $(t + 1)$ has the following three parts with the constraint $\alpha_{11} + \alpha_{12} + \alpha_{13} = 1$.

(1) α_{11} part is directly inherited from its historical value, that is $pa^{(t+1)}(i) \leftarrow \alpha_{11} \cdot pa^{(t)}(i)$.

(2) α_{12} part is transmitted from the hub values of its citations, that is $pa^{(t+1)}(i) \leftarrow \alpha_{12} \cdot \sum_{p_i \to p_j} ph^{(t)}(j)$ where $p_j \to p_i$ means a citation from paper p_j to paper p_i.

(3) The remaining α_{13} part is reinforced by its authors' importance values as in $pa^{(t+1)}(i) \leftarrow \alpha_{13} \cdot \sum_{r_k \in A(p_j)} ri^{(t)}(k)$ where $A(p_i)$ is the set of authors of paper p_i and $ri^{(t)}(k)$ is the k-th researcher's importance at time t.

First formulate part (2) of paper authority $pa^{(t+1)}(i)$. Because each citing paper p_j may be connected to p_i via different hyperedges, to transmit the hub value of p_j to p_i over $\vec{\mathcal{H}}^{[P]}$, we need to first select a directed hyperedge $\vec{e}^{[P]}$ going out of p_j according to the following probability, which is in essence the normalized weight of hyperedge $\vec{e}^{[P]}$,

$$\frac{W_e^{[P]}(\vec{e}^{[P]}) \times H^{[P+]}(j, \vec{e}^{[P]})}{\sum_{\tilde{e}^{[P]} \in V^{[P]}} W_e^{[P]}(\tilde{e}^{[P]}) \times H^{[P+]}(j, \tilde{e}^{[P]})}. \quad (3)$$

This means a portion of hub value (specified in Eq. (3)) is propagated from paper p_j to the head vertices incident to $\vec{e}^{[P]}$. Among this, the fraction that goes to p_i is determined by Eq. (4), which is in essence the normalized weight of p_i as a head vertex in $\vec{e}^{[P]}$.

$$\frac{W_v^{[P-]}(i, \vec{e}^{[P]})}{\sum_{\tilde{i} \in V^{[P]}} W_v^{[P-]}(\tilde{i}, \vec{e}^{[P]})} \quad (4)$$

Combining these two probabilities, we get the transition probability from a citing paper p_j to a cited paper p_i in the homogeneous hypergraph $\vec{\mathcal{H}}^{[P]}$ as follows,

$$\Pr^{[P]}(j, i) = \sum_{\vec{e}^{[P]} \in E^{[P]}} \frac{W_e^{[P]}(\vec{e}^{[P]}) \times H^{[P+]}(j, \vec{e}^{[P]})}{d^{[P+]}(j)} \times \frac{W_v^{[P-]}(i, \vec{e}^{[P]})}{\delta^{[P-]}(\vec{e}^{[P]})} \quad (5)$$

,

Where

$$d^{[P+]}(j) = \sum_{\vec{e}^{[P]} \in E^{[P]}} W_e^{[P]}(\tilde{e}^{[P]}) \times H^{[P+]}(j, \tilde{e}^{[P]}), \quad (6)$$

$$\delta^{[P-]}(\vec{e}^{[P]}) = \sum_{\tilde{i} \in V^{[P]}} W_v^{[P-]}(\tilde{i}, \vec{e}^{[P]}). \quad (7)$$

$d^{[P+]}(j)$ and $\delta^{[P-]}(\vec{e}^{[P]})$ are called the out-degree of paper p_j and in-degree of hyperedge $\vec{e}^{[P]}$ in $\vec{\mathcal{H}}^{[P]}$ respectively.

Based on Eqs. (5-7), part (2) of paper authority is calculated in Eq. (8).

$$pa^{(t+1)}(i) \leftarrow \alpha_{12} \sum_{p_j \in V^{[P]}} \Pr^{[P]}(j,i) \times ph^{(t)}(j)$$
$$= \alpha_{12} \sum_{p_j \in V^{[P]}} ph^{(t)}(j) \times \sum_{e^{[P]} \in E^{[P]}} \frac{W_e^{[P]}(\vec{e}^{[P]}) \times H^{[P+]}(j,\vec{e}^{[P]})}{d^{[P+]}(j)} \times \frac{W_v^{[P-]}(i,\vec{e}^{[P]})}{\delta^{[P-]}(\vec{e}^{[P]})} \quad (8)$$

A neater matrix-vector format of Eq. (8) is as follows

$$\mathbf{pa}^{(t+1)} \leftarrow \alpha_{12} \left(\mathcal{L}^{[P+]}\right)^{\mathrm{T}} \mathbf{ph}^{(t)}, \quad (9)$$

where $\mathcal{L}^{[P+]}$ is the normalized transition matrix for the paper citation hypernetwork as follows

$$\mathcal{L}^{[P+]} = \left(\mathbf{D}^{[P+]}\right)^{-1} \cdot \mathbf{H}_v^{[P+]} \cdot \mathbf{W}_e^{[P]} \cdot \left(\mathbf{\Delta}^{[P-]}\right)^{-1} \cdot \left(\mathbf{W}_v^{[P-]}\right)^{\mathrm{T}}, \quad (10)$$

where $\mathbf{D}^{[P+]} = \mathbf{diag}(d^{[P+]}(j))$ and $\mathbf{\Delta}^{[P-]} = \mathbf{diag}(\delta^{[P-]}(\vec{e}^{[P]}))$ are the diagonal matrices of the out-degrees of papers and the in-degrees of paper citation hyperedges respectively.

Part (3) of the authority of paper p_i is propagated from the importance of each of its authors r_k through the heterogeneous authorship hypergraph. Similarly this is also done by a two-step process. First, a portion of $ri(r_k)$ is propagated over $\mathcal{H}^{[PR]}$ via an undirected heterogeneous hyperedge $e^{[PR]}$ that is incident to r_k based on the normalized weight of $e^{[PR]}$. Then this amount of author importance is transmitted to the target paper p_i based on the normalized weight of p_i in $e^{[PR]}$. Thus the transition probability from author r_k to paper p_i over $\mathcal{H}^{[PR]}$ is

$$\Pr^{[R \to P]}(k,i) = \sum_{e^{[PR]} \in E^{[PR]}} \frac{W_e^{[PR]}(e^{[PR]}) \times H^{[R/P]}(k,e^{[PR]})}{d^{[R/P]}(k)} \times \frac{W_v^{[P/R]}(i,e^{[PR]})}{\delta^{[P/R]}(e^{[PR]})}, \quad (11)$$

Where

$$d^{[R/P]}(k) = \sum_{\tilde{e}^{[PR]} \in E^{[PR]}} W_e^{[PR]}(\tilde{e}^{[PR]}) \times H^{[R/P]}(k,\tilde{e}^{[PR]}), \quad (12)$$

$$\delta^{[P/R]}(e^{[PR]}) = \sum_{\tilde{i} \in V^{[P]}} W_v^{[P/R]}(\tilde{i},e^{[PR]}). \quad (13)$$

$d^{[R/P]}(k)$ denotes the *mode-R degree* of researcher r_k in $\mathcal{H}^{[PR]}$ while $\delta^{[P/R]}(e^{[PR]})$ denotes the *mode-P degree* of the heterogeneous hyperedge $e^{[PR]}$.

Based on Eq. (11), part (3) of paper authority is calculated by $pa^{(t+1)}(i) \leftarrow \alpha_{13} \sum_{k \in V^{[R]}} \Pr^{[R \to P]}(k,i) \times ri^{(t)}(j)$, and in matrix-vector format as follows

$$\mathbf{pa}^{(t+1)} \leftarrow \alpha_{13} \left(\mathcal{L}^{[R \to P]}\right)^{\mathrm{T}} \mathbf{ri}^{(t)}, \quad (14)$$

where $\mathcal{L}^{[R \to P]}$ is the normalized transition matrix from researchers to papers in the heterogeneous authorship hypergraph as follows

$$\mathcal{L}^{[R \to P]} = \left(\mathbf{D}^{[R/P]}\right)^{-1} \cdot \mathbf{H}_v^{[R/P]} \cdot \mathbf{W}_e^{[PR]} \cdot \left(\mathbf{\Delta}^{[P/R]}\right)^{-1} \cdot \left(\mathbf{W}_v^{[P/R]}\right)^{\mathrm{T}}, \quad (15)$$

where $\mathbf{D}^{[R/P]} = \mathbf{diag}(d^{[R/P]}(k))$ is the diagonal matrix of the out-degree of researchers and $\mathbf{\Delta}^{[R/P]} = \mathbf{diag}(\delta^{[R/P]}(e^{[PR]}))$ is the diagonal matrix of the in-degrees of papers in $\mathcal{H}^{[PR]}$.

Putting the above together we have

$$\mathbf{pa}^{(t+1)} = \alpha_{11}\mathbf{\Lambda}^{[P]}\mathbf{pa}^{(t)} + \alpha_{12}\left(\mathcal{L}^{[P+]}\right)^{\mathrm{T}} \mathbf{ph}^{(t)} + \alpha_{13}\left(\mathcal{L}^{[R \to P]}\right)^{\mathrm{T}} \mathbf{ri}^{(t)}, \quad (16)$$

where $\mathbf{\Lambda}^{[P]}$ is a n_P-dimensional identity matrix.

Putting the Remaining into a Unified Form

Paper hub is calculated similar to authority as in Eq. (17).

$$\mathbf{ph}^{(t+1)} = \alpha_{21}\left(\mathcal{L}^{[P-]}\right)^{\mathrm{T}} \mathbf{pa}^{(t)} + \alpha_{22}\mathbf{\Lambda}^{[P]}\mathbf{ph}^{(t)} + \alpha_{23}\left(\mathcal{L}^{[R \to P]}\right)^{\mathrm{T}} \mathbf{ri}^{(t)}, \quad (17)$$

where

$$\mathcal{L}^{[P-]} = \left(\mathbf{D}^{[P-]}\right)^{-1} \cdot \mathbf{H}^{[P-]} \cdot \mathbf{W}_e^{[P]} \cdot \left(\mathbf{\Delta}^{[P+]}\right)^{-1} \cdot \left(\mathbf{W}_v^{[P+]}\right)^{\mathrm{T}} \quad (18)$$

is the normalized transition matrix for the reverse paper citation hypernetwork.

Similarly research importance is computed by

$$\mathbf{ri}^{(t+1)} = \alpha_{31}\left(\mathcal{L}^{[P \to R]}\right)^{\mathrm{T}} \mathbf{pa}^{(t)} + \alpha_{32}\left(\mathcal{L}^{[P \to R]}\right)^{\mathrm{T}} \mathbf{ph}^{(t)} + \alpha_{33}\mathcal{L}^{[R]}\mathbf{ri}^{(t)}, \quad (19)$$

where

$$\mathcal{L}^{[P \to R]} = \left(\mathbf{D}^{[P/R]}\right)^{-1} \cdot \mathbf{H}_v^{[P/R]} \cdot \mathbf{W}_e^{[PR]} \cdot \left(\mathbf{\Delta}^{[R/P]}\right)^{-1} \cdot \left(\mathbf{W}_v^{[R/P]}\right)^{\mathrm{T}}, \quad (20)$$

$$\mathcal{L}^{[R]} = \left(\mathbf{D}^{[R+]}\right)^{-1} \cdot \mathbf{H}_v^{[R+]} \cdot \mathbf{W}_e^{[R]} \cdot \left(\mathbf{\Delta}^{[R-]}\right)^{-1} \cdot \left(\mathbf{W}_v^{[R-]}\right)^{\mathrm{T}}. \quad (21)$$

To put all the above together, we have the following unified iterative ranking equation

$$\begin{bmatrix} \mathbf{pa} \\ \mathbf{ph} \\ \mathbf{ri} \end{bmatrix}^{(t+1)} = \begin{bmatrix} \alpha_{11}\mathbf{\Lambda}^{[P]} & \alpha_{21}\mathcal{L}^{[P-]} & \alpha_{31}\mathcal{L}^{[P \to R]} \\ \alpha_{12}\mathcal{L}^{[P+]} & \alpha_{22}\mathbf{\Lambda}^{[P]} & \alpha_{32}\mathcal{L}^{[P \to R]} \\ \alpha_{13}\mathcal{L}^{[R \to P]} & \alpha_{23}\mathcal{L}^{[R \to P]} & \alpha_{33}\mathcal{L}^{[R]} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{pa} \\ \mathbf{ph} \\ \mathbf{ri} \end{bmatrix}^{(t)}. \quad (22)$$

Experiments

Experimental Setup

It is always difficult to put different ranking algorithms in a fair play because of lack of a common testing benchmark including dataset, gold-standard and evaluation metrics. Thus one of the contributions of this paper is to construct a comprehensive testing benchmark[3] for the design and evaluation of scientific ranking algorithms.

[3] Available at http://sites.google.com/site/xiaoruijiang/research.

#paper	#author	#venue	#dangling	avg #citation (w/o dangling)
18041	14386	273	4187 (23.21%)	4.597/5.987

Table 1. Statistics of the Benchmarking Dataset

#rec.	2	3	4	5	6	7	8	9	10	
#GoldP	63	19	7	1	1	0	0	1	1	93
Grade	1	2	3	3	4	4	5	5	5	

Table 2. Statistics of the Gold Standard Sets

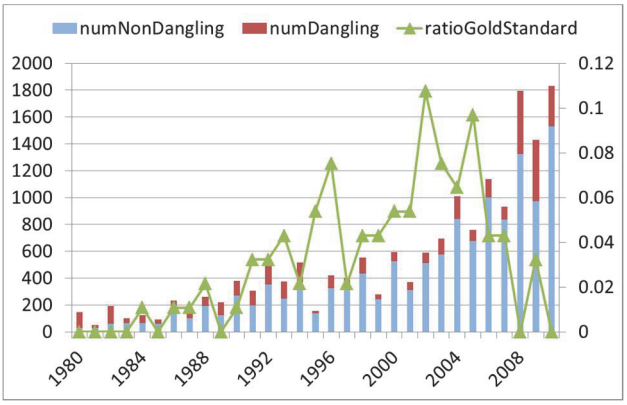


Figure 2. The Time Distributions for the AAN Benchmark Dataset and Gold Standard Papers.

Dataset Description

Previous studies evaluated their algorithms on different datasets, for example the arXiv hep-th dataset (Chen et al., 2007; Walker et al., 2007; Radicchi et al., 2009, the Information Science and Library Science publications (Ding et al., 2009; Yan et al., 2011), and different datasets of computer science publications (Zhou et al., 2007; Li et al., 2008; Sayyadi & Getoor, 2009; Das et al., 2011).

This paper uses ACL Anthology Network (AAN) as the benchmarking dataset (Radev et al., 2013). AAN contains the complete collection of computational linguistics articles published by the ACL (Association of Computational Linguistics). The 2011 release of AAN is used and its statistics is shown in Table 1. A distinguishing advantage of using AAN is that, by recording the full name of each author in such a self-contained field, the need for name disambiguation is minimized.

Note that in this paper, the citation hypernetwork $\vec{\mathcal{H}}^{[P]}$ is constructed by adding a hyperedge for each citing paper p with the citing paper alone as the tail vertex set and all the cited papers as the head vertex set. For the author citation hypernetwork $\vec{\mathcal{H}}^{[R]}$, we add a directed hyperedge for each pair of citing and cited papers. The authors of the citing and cited paper constitute the tail and head vertex sets of the hyperedge respectively. The heterogeneous hypernetwork $\mathcal{H}^{[PR]}$ is constructed by adding a hyperedge for each set of coauthors and the set of papers they collaborate on. For evaluation, this paper only considers the simple case where both hyperedges and vertices are unweighted and leaves weighting strategies as one of our future work.

Evaluation Methods

The gold standard set GoldP consist of 93 papers that are recommended at least twice by two well-known textbooks and/or the course reading lists of 15 world-famous universities, with the recommendation counts listed in Table 2. The non-uniform time distributions of both the AAN dataset and gold standard papers are illustrated in Figure 2.

As there are recommendation counts for the gold standard set of papers, we adopt two widely adopted graded relevance metrics for evaluating the ranking effectiveness. The first metric is the Normalized Discounted Cumulative Gain (NDCG; Järvelin and Kekäläinen, 2002). The second metric is Graded Average Precision (GAP; Robertson et al., 2010), a generalization of average precision to the multi-graded case. While NDCG is precision-oriented, GAP also considers recall by estimating the area under the non-interpolated graded precision-recall curve. For graded document ranking evaluation, let $G = [0, 1, \ldots, c]$ be the grade list for the top-K document list $R = \{d_k \mid 1 \le k \le K$ where the grade for each document d_k is $r_k \in G\}$, which is sorted in descending order of grade to $\tilde{R} = \{d_{[k]}\}$.

NDCG is defined as the Discounted Cumulative Gain (DCG) divided by its ideal value (IDCG), that is

$$NDCG@K = \frac{DCG@K}{IDCG@K} = \frac{\sum_{1\le i\le K} \tilde{r}_k}{\sum_{1\le [i]\le K} \tilde{r}_k}, \qquad (23)$$

where $\tilde{r}_k = \log d_k$ if $d_k > 1$ and $\tilde{r}_k = d_k$ otherwise.

GAP is based on such a hypothetic model that a user has a probability p_j to set a threshold at grade j, that is to treat grades 1, …, j as relevant. Thus GAP is formulated by the accumulated average precision contributed at all the K positions divided by the maximal accumulated average precisions contributed by the top-K documents, as in Eq. (24)

$$GAP@K = \frac{\sum_{k=1}^{K}\sum_{m=1}^{k} \delta(m,k)/k}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i} p_j}, \qquad (24)$$

where R_i is the number of relevant documents at grade i,

$$\delta(m,k) = \begin{cases} \sum_{j=1}^{\min(r_k, r_m)} p_j & r_m > 0 \\ 0 & \text{otherwise} \end{cases}. \qquad (25)$$

Experiment Results

Several representative ranking algorithms are evaluated in the same benchmarking environment, including PageRank (Brin and Page, 1998), the randomized version of HITS (Ng et al, 2001), CoRank (Zhou et al., 2007), FutureRank

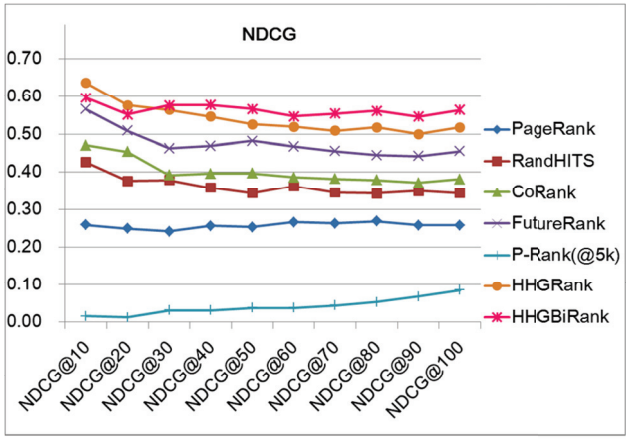


Figure 3. NDCG Curves for Top-100 Papers.

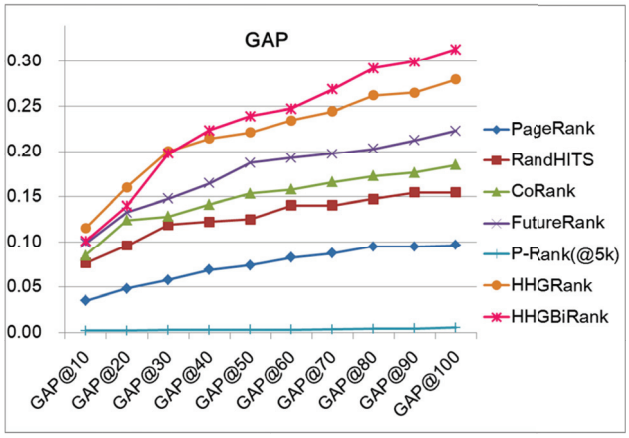


Figure 4. GAP Curves for Top-100 Papers.

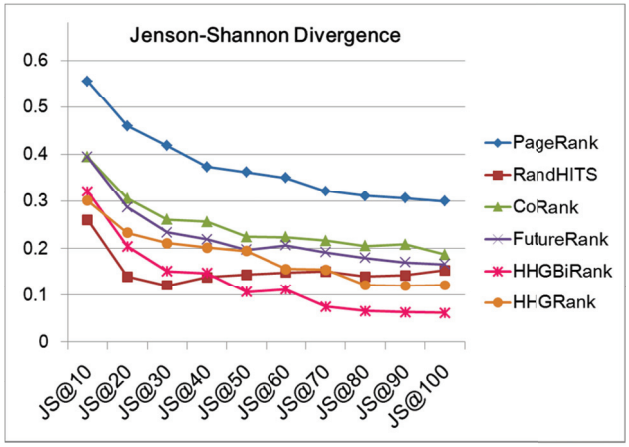


Figure 5. Jensen-Shannon Divergence of Top-100 Papers.

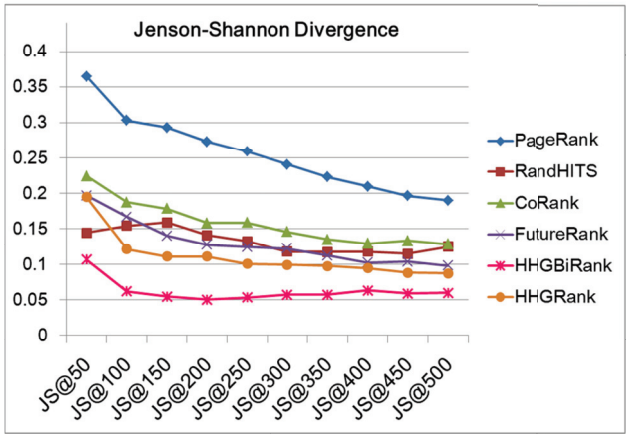


Figure 6. Jensen-Shannon Divergence of Top-500 Papers.

(Sayyadi and Getoor 2009), P-Rank (Yan et al., 2011). For CoRank, α and β ($= 1 - \alpha$) parts of all the ranking vectors come from intra- and inter-network iterations respectively. FutureRank also has an additional γ part of uniform teleportation for paper ranking such that $\alpha + \beta + \gamma = 1$. On the contrary, P-Rank adopts personalized PageRank for paper ranking with γ part of non-uniform (i.e. personalized) teleportation, where α and β ($= 1 - \alpha$) portions of the personalization vector come from author and venue (journal and/or conference etc.) rankings respectively. The algorithm proposed in this paper is denoted as HHGBiRank. However, to demonstrate the usefulness of distinguishing paper authority and hub, this section also consider a variant called HHGRank which applies PageRank-style ranking to paper citation hypernetwork. Note that although time factor is not considered in FutureRank for the purpose of a fair play, we will see how HHGBiRank improves ranking performance by resolving a related issue of time factor. We report the results under parameter settings that have been proved effective so that real performances of different competitors can be reflected. Thus we have $\alpha_{11} = \alpha_{12} = \alpha_{21} = \alpha_{22} = \alpha_{31} = \alpha_{32} = 0.3$ and $\alpha_{13} = \alpha_{23} = \alpha_{33} = 0.4$ for HHGBiRank; $m = 2$, $n = 2$, $l = 1$ (as in its original paper) and $\alpha = 0.3$ For CoRank; $\alpha = 0.3$ and $\beta = 0.45$ For FutureRank and P-Rank, $\alpha = 0.5$; and for all the algorithms the teleportation factor is set as $\gamma = 0.15$.

Figure 3 and Figure 4 illustrate the NDCG curve and GAP curve of the top-100 papers returned by different algorithms respectively. Note that, in both figures, the results of P-Rank are calculated using top-5k returned papers, i.e. top-50, top-100, …, top-500 due to its poor performance. In Figure 4, we set the probabilities p_j of setting thresholds at each grade j equally to 0.20. Similar GAP curves are obtained if we use different thresholding probabilities and are omitted due to space limit. Consistent to previous studies, FutureRank and CoRank perform better than HITS and PageRank, which justifies employing the structure of heterogeneous academic networks. HHGRank improves ranking effectiveness further by 10% to 15% in term of NDCG and about 30% in term of GAP. This demonstrates the superiority of using heterogeneous academic hypernetworks as the more generalized model. We also see that HHG-

BiRank beats all the competitors in a general sense, almost 10% and 15% better than HHGRank in terms of NDCG and GP respectively, which justifies our assumption about the distinction between authority and hub.

As the time distribution for the gold standard papers is non-uniform, it is intuitive that an algorithm is better if the time distribution of its top-ranked papers is closer to that of the gold standard, for which Jenson-Shannon divergence (JS) is a promising evaluation metric. Figure 5 shows the JS curves for the top-100 papers returned by different algorithms. A smaller JS value for a certain k and a flatter JS curve mean a better algorithm. In this sense, PageRank, CoRank and FutureRank are the three worst algorithms. It also implies that there is an inverse relationship between NDCG or GAP and JS divergence. HHGRank is clearly better than the above three algorithms due to the powerful modeling capability of heterogeneous academic hypernetworks, and HHGBiRank performs even much better than HHGRank due to the distinction between authority and hub.

Although when k is small HITS has a smaller JS value than HHGRank ($k \leq 60$) and HHGBiRank ($k \leq 40$), it is not valid to say HITS is better than heterogeneous academic network based algorithms. Actually when we expand the ranked list to top-500 as in Figure 6, HHGRank is better than all the previous algorithms and HHGBiRank further enlarges the performance gain by a nontrivial margin.

Conclusion

This paper introduces a new scientific ranking algorithm based on the multinomial multidimensional relationships between different types of academic entities. Using heterogeneous academic hypernetwork as a generalized model, the algorithm formulates the mutual reinforcement relation between these academic entities. The proposed algorithm has been extensively evaluated and demonstrated effective using widely-adopted IR metrics in a common benchmarking environment with ACL Anthology Network as the dataset and a carefully built gold standard set of papers.

Acknowledgement

This paper is partially supported by National Science Foundation of China (61402412) and Zhejiang Province (LY14F020016).

References

Bellaachia, A., and Al-Dhelaan, M. 2015. Short text keyphrase extraction with hypergraphs. *Progress in Artificial Intelligence* 3(2): 73–87.

Bretto, A. 2013. Hypergraph theory: An introduction. Springer, Heidelberg.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30: 107–117.

Chen, P., Xie, H., Maslov, S., and Redner, S. 2007. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics* 1: 8–15.

Das, S., Mitra, P., and Lee Giles, C. 2011. Ranking Authors in Digital Libraries. In *Proc JCDL2011*.

Deng, H., Lyu, M. R., and King, I. 2009. A generalized Co-HITS algorithm and its application to bipartite graphs. In *Proc. KDD2009*.

Ding, Y., Yan, E., Frazho, R., and Caverlee, J. 2009. PageRank for Ranking Authors in Co-citation Networks. *Journal of the American Society for Information Science and Technology* 60(11): 2229–2243.

Ducournao, A., and Bretto, A. 2014. Random walks in directed hypergraphs and application to semi-supervised image segmentation. *Computer Vision and Image Understanding* 120: 91–102.

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20(4): 422–446.

Jiang, X., Sun, X., and Zhuge, H. 2012. Towards an effective and unbiased ranking of scientific literature through mutual reinforcement. In *Proc. CIKM2012*.

Jiang, X., Sun, X., and Zhuge, H. 2013. Graph-based algorithms for ranking researchers: not all swans are white! *Scientometrics* 96(13): 743–759.

King B., Jha R., and Radev D. R. 2014. Heterogeneous Networks and Their Applications: Scientometrics, Name Disambiguation, and Topic Modeling. *Transactions of Association for Computational Linguistics* 2: 1–14.

Li, X., Liu, B., and Yu, P. 2008. Time sensitive ranking with application to publication search. In *Proc. ICDM2008*.

Ng, A. Y., Zheng, A. X., and Jordan, M. I. 2001. Stable algorithms for link analysis. In *Proc. SIGIR2001*.

Radev, D. R., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. 2013. The ACL anthology network corpus. *Language Resources and Evaluation* 47(4): 919–944.

Radicchi, F., Fortunato, S., Markines, B., and Vespignani, A. 2009. Diffusion of scientific credits and the ranking of scientists. *Physics Review E* 80(5): 56103–56112.

Robertson S. E., Kanoulas E., and Yilmaz E. 2010. Extending average precision to graded relevance judgments. In *Proc. SIGIR2010.*

Sayyadi, H., and Getoor, L. 2009. FutureRank: Ranking Scientific Articles by Predicting their Future PageRank. In *Proc. SDM2009*.

Walker, D., Xie, H., Yan, K.-K., and Maslov, S. 2007. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment* 7.

Wang, Y., Tong, Y., and Zeng, M. 2013. Ranking scientific articles by exploiting citations, authors, journals, and time information. In *Proc. AAAI2013*.

Yan E., Ding Y., and Sugimoto C. R. 2011. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology* 62(3): 467–477.

Zhou, D., Orshanskiy, S. A., Zha, H., and Lee Giles, C. 2007. Co-Ranking Authors and Documents in a Heterogeneous Network. In *Proc. ICDM2007*.