Towards an Effective and Unbiased Ranking of Scientific Literature through Mutual Reinforcement

Xiaorui Jiang
Key Lab of Intell. Inf. Processing
Inst. of Comput. Technol., CAS
Beijing, 100190, China
xiaoruijiang@kg.ict.ac.cn

Xiaoping Sun
Key Lab of Intell. Inf. Processing
Inst. of Comput. Technol., CAS
Beijing, 100190, China
sunxp@kg.ict.ac.cn

Hai Zhuge*
Key Lab of Intell. Inf. Processing
Inst. of Comput. Technol., CAS
Beijing, 100190, China
zhuge@ict.ac.cn

ABSTRACT
It is important to help researchers find valuable scientific papers from a large literature collection containing information of authors, papers and venues. Graph-based algorithms have been proposed to rank papers based on networks formed by citation and co-author relationships. This paper proposes a new graph-based ranking framework MutualRank that integrates mutual reinforcement relationships among networks of papers, researchers and venues to achieve a more synthetic, accurate and fair ranking result than previous graph-based methods. MutualRank leverages the network structure information among papers, authors, and their venues available from a literature collection dataset and sets up a unified mutual reinforcement model that involves both intra- and inter-network information for ranking papers, authors and venues simultaneously. To evaluate, we collect a set of recommended papers from websites of graduate-level computational linguistics courses of 15 top universities as the benchmark and apply different methods to estimate paper importance. The results show that MutualRank greatly outperforms the competitors including PageRank, HITS and CoRank in ranking papers as well as researchers. The experimental results also demonstrate that venues ranked by MutualRank are reasonable.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval models.

General Terms
Algorithms, Experimentation.

Keywords
Mutual reinforcement, iterative ranking, time distortion

1. INTRODUCTION
Researchers and students new in a particular area often feel at a loss when facing an ocean of papers published year after year. Researchers often need to answer the following questions. "Which classical papers in a particular area are the most influential and valuable to read for building consolidated background knowledge? Which papers and researchers are worth being paid attention to so as to catch up recent advance? And which venue is the most suitable for one to publish so as to maximize personal capacity?" Of all the three questions, the former two are more difficult to answer because of the huge numbers of papers and researchers. For example, many digital libraries like DBLP[1] and CiteSeer[2] contain paper metadata such as authors and venues, but it is hard for people to manually extract a series of important papers or authors for a given area.

To rank authors, papers and journals, citation count information has long been used, e.g., [6] [9] and [15] for journal ranking, h-index [7] for researcher ranking and citation count for paper ranking. But only using citation count based metrics for evaluation is still a controversial issue as most of these methods do not consider the network structure of literature information available. Recently, graph-based methods have been developed for literature ranking [3] [5] [13] [19] [23-24] [27]. They model a literature collection as a network and apply iterative computation over the adjacency matrix of the network to achieve a converged ranking vector for objects, just as PageRank over Web pages [2]. Most of these works focus on only one type of network, which limits their effectiveness in ranking objects. Recent works begin to consider multiple networks for ranking [4] [18] [21] [25-26]. For example, one of the most relevant research CoRank combines the citation network and the corresponding co-authorship network to achieve better ranking results for both authors and documents [26].

However, current research still does not fully leverage the available information of a general literature dataset which contains papers, authors and venues. All of the previous works on ranking papers and authors only utilize the paper and/or author metadata. But venue information is also important for ranking because we often assign reasonable importance values to new papers in prestigious venues with few citations and to young scientists with few collaborators. Besides, previous works assign only one type of importance values to each type of objects. But in many cases, different types of objects may, in nature, play different roles in a network and thus have different kinds of importance values. In the meanwhile, how to evaluate different ranking methods is still a problem. There lacks a common standard for ranking performance evaluation. Citation counts and future downloads are two predominant metrics widely adopted in previous research. However, we

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

* Corresponding author

[1] http://www.informatik.uni-trier.de/~ley/db.
[2] http://citeseer.ist.psu.edu.

argue that it is researcher who is the most appropriate for judging the relevance of papers and importance of authors. Thus any reasonable ranking algorithm should return many papers that are highly recommended by domain experts when user needs are taken into consideration.

To overcome these limitations, we propose a new ranking framework, MutualRank, which employs mutual reinforcement relationships across networks of papers, researchers and venues to achieve a more synthetic and reasonable ranking of papers, authors and venues simultaneously. The intuition of mutual reinforcement is that a higher value of x will lead to a higher value of y and vice versa. For instance, HITS [10] can be viewed as a mutual reinforcement process between authority and hub values of nodes. Mutual reinforcement is a ubiquitous phenomenon in real-world networks and has already been successfully studied and used in several works, e.g., the preferential attaching model depicts the features of complex networks in physical world [1]. It is also widely found that high degree nodes in social networks tend to link to other high degree networks [16]. In real applications, mutual reinforcement can be used for finding specific patterns. Jensen et al. use reinforcement relationships to mine significant places: places visited by authoritative users are significant and users visiting significant places gain authorities [8].

Compared to viewing ranking iteration as a random walk process, it is more instructive to take it as a reinforcement process. In MutualRank, three basic networks are constructed to capture the intra-network influences between papers, authors and venues respectively. Moreover, mutual reinforcement information across these networks are defined as follows: a paper which is written by an important researcher and published in a prestigious venue should be ranked high; a researcher who publishes highly ranked papers on prestigious venues gains personal importance; and a venue where important researchers publish highly ranked papers will be prestigious. Based on inter-network mutual reinforcement relationships, we integrate all these networks into one unified ranking framework to find important papers, authors and venues simultaneously. A synthetic transition matrix is then built for the iterative computation of ranks of different objects.

2. DATASETS AND BASIC DATA MODEL

2.1 Datasets

This paper uses the ACL (Association of Computational Linguistics) Anthology Network (AAN)[3] [20] dataset for evaluating ranking algorithms (the latest version till March 2011). AAN consists of all the metadata of ACL Anthology, a collection of papers published in journals and conferences hosted by the ACL. The dataset contains 18041 papers published in 273 venues and authored by 14386 researchers. From the AAN dataset we can obtain the authors, publication venue and reference list of a paper.

For joint conferences such as ACL-COLING 2006, we treat the corresponding published papers as being published in more than one venue. For example, the paper "P06-1001" belongs to both ACL'2006 and COLING'2006. If a conference accepts papers other than long, regular or full papers, we construct a new venue X-Companion for those papers in this conference, where X stands for the conference name. For example, there are two venues for the annual conference of ACL: ACL for full papers and ACL-Companion for short and student papers.

2.2 Scientific Literature Ranking Benchmark

For ranking papers. To evaluate the results of paper ranking, we collect papers from the reading lists of graduate-level courses in natural language processing or computational linguistics of top universities, and record the number of times each paper is recommended as the basic benchmark dataset *BenchP*. We discard those reading lists focusing only on a limited number of subareas of computational linguistics and only include those with an extensive coverage of the whole area. If a paper in the AAN dataset is on k universities' reading lists, i.e., receives k recommendations, its importance is k. If a paper receives recommendations from more universities, it is regarded as more relevant and more important in the area of computational linguistics.

In 218 papers collected from 15 universities, 181 are recommended once, 28 are recommended twice, 6 papers receive 3 recommendations, and the rest 3 papers occur 4, 7 and 8 times respectively on the reading lists. We believe the established benchmark well reflects the consensus of the research community because: (1) the benchmark papers are all recommended by prestigious researchers from world famous universities including MIT, CMU, Stanford, Cornell etc.; (2) although most papers has only one recommendation, the fact of winning recommendation is still an indicator of influence to a certain degree; (3) the benchmark set is consistent with the reference section of many survey papers and standard text books such as "Speech and Language Processing".

For ranking researchers. We construct two different benchmark collections for researchers. The first benchmark *BenchR1* is constructed from *BenchP*. If an author has at least 2 papers in *BenchP*, the author is put in *BenchR1*. There are in total 46 researchers in *BenchR1*. Four researchers are each recommended 16, 13, 8 and 7 times. Five occur on the reading lists 6 times. Six researchers win 5 and another six researchers get 4 recommendations. Each of the remaining 25 researchers gets 3 recommendations. The second benchmark collection for researchers *BenchR2* consists of the first 100 top-cited authors of AAN.[4]

2.3 Basic Data Model

We build three basic types of networks from the AAN dataset: Paper Influence Network (PIN), Research Influence Network (RIN) and Venue Influence Network (VIN). Figure 1 gives an example of the network construction.

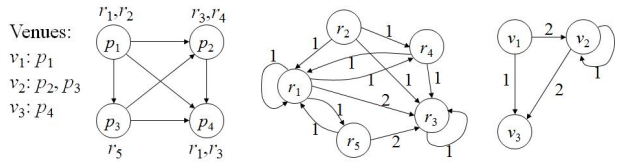
Network of papers. We construct the PIN based on the citation relationship. For each paper p_i in AAN, we add a corresponding node i to PIN. If paper p_i cites another paper p_j, we add an edge (i, j) to PIN. PIN is un-weighted.

Network of researchers. We use citation relationships to build the weighted directed RIN. For each researcher r_i in AAN, we add a node i to RIN. If r_i has a paper that cites another paper authored by r_j, we add an edge (i, j) to RIN. The weight on RIN edges is assigned as follows. Let

$$Cite(p_k, p_l) = \begin{cases} 1 & \text{paper } p_k \text{ cites } p_l \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

[3] http:// clair.si.umich.edu/clair/anthology/

[4] http://clair.si.umich.edu/clair/anthology/rankings.cgi



(a) Metadata and PIN (b) RIN (c) VIN

Figure 1. Construction of RIN and VIN from PIN.

and $P(r_i)$ be the set of papers written by researcher r_i. The weight of edge (i, j) in RIN is set as

$$w_{RIN}(i,j) = \sum_{p_k \in P(r_i), p_l \in P(r_j)} Cite(p_k, p_l). \quad (2)$$

The other type of researcher network is the Researcher Collaboration Network (RCN). RCN is a weighted undirected network. If two researchers r_i and r_j have coauthored at least one paper, there is an edge (i, j) in RCN. The weight of edge (i, j) in RCN is set as

$$w_{RCN}(i,j) = \left| P(r_i) \cap P(r_j) \right|. \quad (3)$$

Different from CoRank, we use RIN rather than RCN for ranking researchers. The motivation behind RCN is that the collaboration with an important researcher on some papers will raise the importance of researcher him/herself. However, we argue that if papers of a researcher are highly cited, we are confident that he/she is influential because his/her ideas may inspire a lot of other researchers in their own work. In Section 4, we will show that performances are slightly improved by substituting RIN for RCN. It should be noted that our framework is flexible enough for incorporating different ways of modeling researcher network.

Network of venues. We construct the VIN to incorporate venue information. VIN is a weighted directed network such that if a paper p_k published on venue v_i has another paper p_l published on venue v_j on its reference list, then there is an edge (i, j) in VIN. Let $P(v_i)$ be the set of papers published in v_i. Then the weight of edge (i, j) in VIN is set as follows. Let $P(v_i)$ be the set of papers published in v_i. Then the weight of edge (i, j) in VIN is set as

$$w_{VIN}(i,j) = \sum_{p_k \in P(v_i), p_l \in P(v_j)} Cite(p_k, p_l). \quad (4)$$

3. MUTUALRANK MODEL

Before delving into the details of MutualRank, we first make a preliminary study of the limitations of two classic ranking algorithms, PageRank and HITS, on ranking papers and discuss the causes of these limitations. Typically, similar phenomena appear in simply applying these algorithms on RIN (resp. VIN) for ranking researchers (resp. venues). The results obtained in this section well motivate the proposed MutualRank method.

3.1 Studies on PageRank and HITS

3.1.1 Results of PageRank

Table 1 lists the PageRank results that match *BenchP*. We only show the results from the top 100 answers returned by each algorithm. The matched ranking result of a paper is given in the format "o_p:*pname*[c_p]", where o_p is the ranked order, c_p is the recommendation count in *BenchP*, and *pname* is the paper id used in ACL anthology. For example, "4:J90-2002[1]" means that the paper J90-2002 is ranked 4th by PageRank and is recommended once in *BenchP*. "7:J93-2003[7]" is ranked 7th by PageRank and is recommended by 7 times in *BenchP*. Relevant papers published in the 1990s and 1980s are represented in boldface and italic respectively in Table 1

The Problem of Time Distortion. In PageRank, paper ranks are transferred from one paper to the other through the citing relationships. Thus, papers published in earlier years will inevitably gain higher ranks than new papers. This phenomenon is clearly reflected by Table 1. In all the 19 results, only 3 are published after year 2000 (P02-1040, A00-2018 and J02-3001). This contradicts to our experience that researchers not only read old classic papers but also pay much attention to recent papers to keep up with the fast-changing research fronts. We call this problem as *time distortion*.

Table 1. Relevant Papers in top-100 PageRank Results

4:J90-2002[1]	29:P02-1040[2]	**70:J93-2006[1]**
7:J93-2003[7]	**33:P97-1003[3]**	75:J02-3001[2]
8:J86-3001[1]	40:A00-2018[1]	**84:P98-2127[1]**
12:J96-1002[2]	**49:J95-4004[1]**	**88:P97-1023[1]**
13:J92-4003[1]	**54:J96-2004[1]**	96:P02-1053[1]
18:J88-1003[1]	**60:P95-1026[4]**	
25:W96-0213[2]	**65:C96-2141[1]**	

3.1.2 Results of HITS

We use a randomized version of HITS (RHITS) from [17] in this study because RHITS typically returns the same results with the HITS by Kleinberg [10] and is more easy to formulate and analyze in our context. RHITS can be formalized as follows. Let $\mathbf{P}$ be the adjacency matrix corresponding to PIN and $\hat{\mathbf{P}}$ is the corresponding transition matrix obtained by normalizing each row of $\mathbf{P}$. Similar to PageRank, random jump (aka. teleportation) is incorporated by rewriting the transition matrix $\hat{\mathbf{P}}$ into $\overline{\mathbf{P}}$ as follows:

$$\overline{\mathbf{P}} = \lambda(\hat{\mathbf{P}} + \mathbf{d}_r \frac{\mathbf{e}}{n_P}) + (1-\lambda)\frac{\mathbf{e}\mathbf{e}^T}{n_P}, \quad (5)$$

where n_P is the number of papers in PIN, $\mathbf{d_r}$ is an n_P-dimensional vector indicating which row of $\hat{\mathbf{P}}$ is zero, $\mathbf{e}$ is an n_P-dimensional identity vector and (1-λ) is the dangling factor controlling the probability in which the rank of a paper is evenly distributed onto every paper in the network. Similarly, $\mathbf{P^T}$ is rewritten into $\overline{\mathbf{P}^T}$.

Let **paut** and **phub** be the authority and hub vectors of papers in PIN. We formulate RHITS using a similar way as SALSA [12] which models the ranking process as two independent random walks as in Eq. (6–7).

$$\mathbf{paut}^{(t)} = \overline{\mathbf{P}}^T \mathbf{phub}^{(t-1)} \quad (6)$$

$$\mathbf{phub}^{(t)} = \overline{\mathbf{P}^T} \mathbf{paut}^{(t-1)} \quad (7)$$

Shown in Table 2, RHITS overcomes the problem of PageRank to some extent. (1) RHITS returns more relevant papers than PageRank. (2) 20 out of total 27 returned papers are in year after 2000. The second point is due to the mutual reinforcement nature of HITS. Unlike PageRank where ranks flow uni-directionally from new papers to old papers (forward flow), RHITS has the mechanism to let the ranks of old papers flow back to new papers

(backward flow). Therefore, new papers have much higher probabilities for being authoritative in RHITS than in PageRank. But Table 2 also shows that RHITS overestimate the values of new papers. Firstly, no papers in the 1980s are returned by HITS although many pioneering work in computational linguistics are done during this period. Secondly, it misses many important papers in the 1990s, the decade of statistical natural language processing which greatly boost the prosperity of the area and incubates many applications and new research problems in the new century. However, RHITS only returns 7 papers in this period.

In conclusion, RHITS performs better than PageRank in ranking on PIN because RHITS assigns two different kinds of paper ranks and the two types of ranks mutually reinforce each other to alleviate time distortion. Thus, mutual reinforcement may play an important role in improving ranking precision. However, the fact that RHITS may overestimate new papers is in fact another type of time distortion. It means that using PIN only for ranking papers may inevitably suffer from time distortion. All these findings lead us to develop a new ranking method.

Table 2. Relevant Papers in top-100 RHITS Results

1:P02-1040[2]	29:A00-2018[1]	**58:W96-0213[2]**
4:J93-2003[7]	30:P03-1054[2]	59:P06-1096[1]
7:P05-1033[1]	33:P07-1019[1]	60:P05-1012[1]
9:J04-4002[2]	38:N04-1021[1]	75:D07-1090[2]
14:C96-2141[1]	40:P05-1074[2]	77:E06-1032[1]
24:J96-1002[2]	49:P05-1074[1]	79:P01-1030[2]
26:J90-2002[1]	51:D07-1091[1]	85:P06-1055[3]
27:N06-1014[1]	52:P05-1022[2]	**89:J92-4003[1]**
28:W06-1606[1]	**57:P97-1003[3]**	97:J04-2003[1]

3.2 Mutual reinforcement in the dataset

We develop the MutualRank framework for ranking papers, researchers and venues simultaneously. The MutalRank algorithm employs mutual reinforcement relationships between papers, researchers and venues to improve the precision of recommending valuable papers. In MutualRank, the importance measures of different objects are as follows:

- **paut(p_i):** Authority of a paper p_i;
- **psnd(p_i):** Soundness of a paper p_i;
- **rimp(r_j):** Importance of a researcher r_j;
- **vprs(v_k):** Prestige of a venue v_k.

Authority **paut**(p_i) depicts whether this paper is of high influence and promotes research advancements in its corresponding area. Soundness **psnd**(p_i) describes whether this paper grounds itself on a thorough study of major previous works. Authority and soundness correspond to the authority and hub values in the traditional HITS algorithm. We use soundness instead of hub to make the semantics of this importance measure clearer in our context. Low authority implies that this paper is somewhat less important or inspiring. High soundness indicates that this paper might be a good reference for background knowledge of a particular area.

For each paper p_i published in venue v_k and each researcher r_j authoring p_i, the mutual reinforcement relationships are as below:

(1) If a researcher r_j is important, a paper p_i of r_i will receive high authority and high soundness, i.e., **paut**(p_i) $\propto$ **rimp**(r_j) if $p_i \in P(r_j)$, and if a paper p_j is of high authority and high soundness, the researchers r_i authoring p_j receives high importance, i.e., **rimp**(r_j) $\propto$ **paut**(p_i) and **rimp**(r_j) $\propto$ **psnd**(p_i) if $p_i \in P(r_j)$.

(2) For a paper p_i published in venue v_k, p_i is of high rank if v_k is prestigious and v_k gains more prestige if p_i is ranked high, i.e., **paut**(p_i) $\propto$ **vprs**(v_k), **psnd**(p_i) $\propto$ **vprs**(v_k), **vprs**(v_k) $\propto$ **paut**(p_i), and **vprs**(v_k) $\propto$ **psnd**(p_i).

(3) Similarly, for a researcher r_j authoring some papers in venue v_k, we have **rimp**(r_j) $\propto$ **vprs**(v_k) and **vprs**(v_k) $\propto$ **rimp**(r_j).

To compute the ranks of objects we need to consider both intra-network relationships and inter-network relationships.

3.3 Intra-network ranking model

Intra-network ranking iteration is modeled by using PIN, RIN and VIN respectively. The classical PageRank is applied to RIN and VIN. To compensate the biases of PageRank and RHITS towards old and new papers respectively and improve ranking on PIN, we derive the First-Order HITS as follows.

In traditional HITS in Eq. (6)–(7), $\mathbf{paut}^{(t+1)}$ only relies on $\mathbf{psnd}^{(t)}$ and $\mathbf{psnd}^{(t+1)}$ only relies on $\mathbf{paut}^{(t)}$. We call this **Zero-Order HITS**, which means **paut** (resp. **psnd**) does not inherit its historical values directly. The name is borrowed from notations of Markovian processes. However, in **First-Order HITS**, **paut** (resp. **psnd**) also inherits a portion of its historical values directly (Eq. (8)–(9)). More specifically, $\mathbf{paut}^{(t+1)}$ inherits $\xi\mathbf{paut}^{(t)}$ from its nearest historical value, while the remaining comes from $\mathbf{psnd}^{(t)}$, i.e., $(1-\xi)\overline{\mathbf{P}}^{\mathrm{T}}\mathbf{psnd}^{(t)}$. Similarly, a ξ portion of $\mathbf{psnd}^{(t+1)}$ comes from $\xi\mathbf{psnd}^{(t)}$, while the remaining is reinforced by $\mathbf{paut}^{(t)}$, i.e., $(1-\xi)\overline{\mathbf{P}^{\mathrm{T}}}\mathbf{paut}^{(t)}$.

$$\mathbf{paut}^{(t+1)} = \xi\mathbf{paut}^{(t)} + (1-\xi)\overline{\mathbf{P}}^{\mathrm{T}}\mathbf{psnd}^{(t)} \quad (8)$$

$$\mathbf{psnd}^{(t+1)} = \xi\mathbf{psnd}^{(t)} + (1-\xi)\overline{\mathbf{P}^{\mathrm{T}}}\mathbf{paut}^{(t)} \quad (9)$$

Eq. (8) and (9) can be rewritten into the following matrix equation,

$$\mathbf{r}^{(t+1)} = \mathbf{M}_1^{\mathrm{T}}\mathbf{r}^{(t)}, \quad (10)$$

where $\mathbf{r} = [\mathbf{paut}^{\mathrm{T}}, \mathbf{psnd}^{\mathrm{T}}]^{\mathrm{T}}$ and

$$\mathbf{M}_1 = \begin{bmatrix} \xi\mathbf{\Lambda}_1 & (1-\xi)\overline{\mathbf{P}} \\ (1-\xi)\overline{\mathbf{P}^{\mathrm{T}}} & \xi\mathbf{\Lambda}_1 \end{bmatrix}$$ where $\mathbf{\Lambda_1}$ is n_{P}-dimensional identity matrix.

$\mathbf{M_1}$ is a transition matrix corresponding to a Markovian process. $\mathbf{M_1}$ is *irreducible* because there is only one communicating class in $\mathbf{M_1}$ as for each pair $<i, j>$, i and j reaches each other. $\mathbf{M_1}$ is *aperiodic* because there exists self-loops in the transition matrix. And $\mathbf{M_1}$ is clearly also a *stochastic* matrix. So according to Frobenius theorem [11], $\mathbf{r}$ calculated in Eq. (10) converges to the principle eigenvector of $\mathbf{M_1}$ with the principle eigenvalue being 1.

3.4 Inter-network ranking model

For incorporating the mutual reinforcement relationships between different networks, 3 undirected networks are constructed: Paper-Researcher Network (PRN), Paper-Venue Network (PVN) and Researcher-Venue Network (RVN). They are constructed as follows: If a researcher r_i has written a paper p_j published in venue v_k,

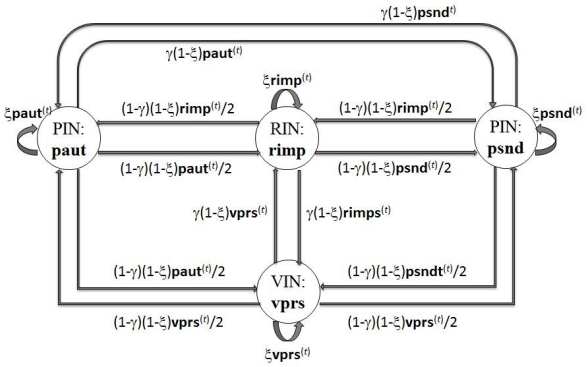


Figure 2. Mutual reinforcement relationships between paut, psnd, rimp and vprs.

we add an edge (i, j) to PRN, an edge (j, k) to PVN and an edge (i, k) to RVN. Then, we obtain two transition matrices for each of the above three networks as follows.

Let **PR** (resp. **RP**) be the adjacency matrix from PIN (resp. RIN) to RIN (resp. PIN). We have

$$\mathbf{RP}(i,j) = \mathbf{PR}(j,i) = \begin{cases} 1 & \text{edge } (i,j) \text{ is in PRN} \\ 0 & \text{otherwise} \end{cases}. \quad (11)$$

Let **PV** and **VP** be the adjacency matrices coupling PIN and VIN together. We have

$$\mathbf{PV}(i,k) = \mathbf{VP}(k,i) = \begin{cases} 1 & \text{edge } (i,k) \text{ is in PVN} \\ 0 & \text{otherwise} \end{cases}. \quad (12)$$

Finally, **RV** and **VR** are the adjacency matrices between RIN and VIN and we have

$$\mathbf{RV}(j,k) = \mathbf{VR}(k,j) = \begin{cases} 1 & \text{edge } (j,k) \text{ is in RVN} \\ 0 & \text{otherwise} \end{cases}. \quad (13)$$

3.5 The unified iterative model

Now we have totally 9 adjacency matrices from the 6 networks which capture intra-network ranking iteration and inter-network mutual reinforcement. Similar to Eq. (5), we create the corresponding intra-network transition matrices $\overline{\mathbf{P}}$, $\overline{\mathbf{R}}$, $\overline{\mathbf{V}}$ and inter-network transition matrices $\overline{\mathbf{PR}}$, $\overline{\mathbf{RP}}$, $\overline{\mathbf{PV}}$, $\overline{\mathbf{VP}}$, $\overline{\mathbf{RV}}$ and $\overline{\mathbf{VR}}$. Thus, we can couple the First-Order HITS on PIN and PageRank on RIN and VIN into a unified framework as follows (see Figure 2 for illustration). Taking initial vectors of **paut, psnd, rimp** and **vprs** as inputs, the trilateral mutual reinforcement process updates these vectors at each iteration as follows.

1. Value of **paut** at time t reinforces values of **paut, psnd, rimp** and **vprs** at time t+1. $\xi\mathbf{paut}^{(t)}$ is reserved for $\mathbf{paut}^{(t+1)}$ in First-Order HITS on PIN. The remaining $(1\text{-}\xi)\mathbf{paut}^{(t)}$ is divided into three parts, one part is $\gamma(1\text{-}\xi)\mathbf{paut}^{(t)}$ and the other two parts are both $(1\text{-}\gamma)(1\text{-}\xi)\ \mathbf{paut}^{(t)}/2$ and $(1\text{-}\gamma)(1\text{-}\xi)\mathbf{paut}^{(t)}/2$. The first part $\gamma(1\text{-}\xi)\mathbf{paut}^{(t)}$ is used for reinforcing $\mathbf{psnd}^{(t+1)}$. The second and third parts are for $\mathbf{rimp}^{(t+1)}$ and for $\mathbf{vprs}^{(t+1)}$ respectively.
2. Value of **psnd** at time t reinforces values of **paut**, **psnd**, **rimp** and **vprs** at time t+1. Similarly, a portion $\xi\mathbf{psnd}^{(t)}$ is inherited by $\mathbf{psnd}^{(t+1)}$. The remaining $(1\text{-}\xi)\mathbf{psnd}^{(t)}$, divided into 3 parts $\gamma(1\text{-}\xi)\mathbf{psnd}^{(t)}$, $(1\text{-}\gamma)(1\text{-}\xi)\mathbf{pnd}^{(t)}/2$ and $(1\text{-}\gamma)(1\text{-}\xi)\mathbf{psnd}^{(t)}/2$, are for reinforcing $\mathbf{paut}^{(t+1)}$, $\mathbf{rimp}^{(t+1)}$ and $\mathbf{vprs}^{(t+1)}$ respectively.
3. For **rimp** at time t to reinforce **paut**, **psnd**, **rimp** and **vprs** at time t+1, a fraction $\xi\mathbf{rimp}^{(t)}$ is used in PageRank on RIN. The remaining $(1\text{-}\xi)\mathbf{rimp}^{(t)}$ is divided into 3 parts, $\gamma(1\text{-}\xi)\mathbf{rimp}^{(t)}$, $(1\text{-}\gamma)(1\text{-}\xi)\mathbf{rimp}^{(t)}/2$ and $(1\text{-}\gamma)(1\text{-}\xi)\mathbf{rimp}^{(t)}/2$. They are used to reinforce $\mathbf{vprs}^{(t+1)}$, $\mathbf{paut}^{(t+1)}$ and $\mathbf{psnd}^{(t+1)}$ respectively. $\mathbf{vprs}^{(t)}$ is processed in a way similar to $\mathbf{rimp}^{(t)}$.

To summarize, MutualRank is formulated in Eq. (14)–(17).

$$\begin{aligned}\mathbf{paut}^{(t+1)} &= \xi\mathbf{paut}^{(t)} + \gamma(1-\xi)\overline{\mathbf{P}}^{\mathrm{T}}\mathbf{psnd}^{(t)} \\ &+ \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{RP}}^{\mathrm{T}}\mathbf{rimp}^{(t)} + \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{VP}}^{\mathrm{T}}\mathbf{vprs}^{(t)}\end{aligned} \quad (14)$$

$$\begin{aligned}\mathbf{psnd}^{(t+1)} &= \gamma(1-\xi)\overline{\mathbf{P}}^{\mathrm{T}}\mathbf{paut}^{(t)} + \xi\mathbf{psnd}^{(t)} \\ &+ \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{RP}}^{\mathrm{T}}\mathbf{rimp}^{(t)} + \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{VP}}^{\mathrm{T}}\mathbf{vprs}^{(t)}\end{aligned} \quad (15)$$

$$\begin{aligned}\mathbf{rimp}^{(t+1)} &= \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{PR}}^{\mathrm{T}}\mathbf{paut}^{(t)} + \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{PR}}^{\mathrm{T}}\mathbf{psnd}^{(t)} \\ &+ \xi\overline{\mathbf{R}}^{\mathrm{T}}\mathbf{rimp}^{(t)} + \gamma(1-\xi)\overline{\mathbf{VR}}^{\mathrm{T}}\mathbf{vprs}^{(t)}\end{aligned} \quad (16)$$

$$\begin{aligned}\mathbf{vprs}^{(t+1)} &= \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{PV}}^{\mathrm{T}}\mathbf{paut}^{(t)} + \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{PV}}^{\mathrm{T}}\mathbf{psnd}^{(t)} \\ &+ \gamma(1-\xi)\overline{\mathbf{RV}}^{\mathrm{T}}\mathbf{rimp}^{(t)} + \xi\overline{\mathbf{V}}^{\mathrm{T}}\mathbf{vprs}^{(t)}\end{aligned} \quad (17)$$

Eq. (14)–(17) can be rephrased as follows:

$$\mathbf{r}^{(t+1)} = \mathbf{M}^{\mathrm{T}}\mathbf{r}^{(t)}, \quad (18)$$

where $\mathbf{r} = [\mathbf{paut}^{\mathrm{T}}, \mathbf{psnd}^{\mathrm{T}}, \mathbf{rimp}^{\mathrm{T}}, \mathbf{vprs}^{\mathrm{T}}]^{\mathrm{T}}$,

$$\mathbf{M} = \begin{bmatrix} \xi\mathbf{\Lambda}_1 & \gamma(1-\xi)\overline{\mathbf{P}}^{\mathrm{T}} & \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{PR}} & \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{PV}} \\ \gamma(1-\xi)\overline{\mathbf{P}} & \xi\mathbf{\Lambda}_1 & \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{PR}} & \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{PV}} \\ \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{RP}} & \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{RP}} & \xi\overline{\mathbf{R}} & \gamma(1-\xi)\overline{\mathbf{RV}} \\ \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{VP}} & \frac{(1-\gamma)}{2}(1-\xi)\overline{\mathbf{VP}} & \gamma(1-\xi)\overline{\mathbf{RV}} & \xi\overline{\mathbf{V}} \end{bmatrix} \quad (19)$$

and $\mathbf{\Lambda}_1$ is a diagonal matrix with each $\mathbf{\Lambda}_{1,i,i} = 1$ ($i = 1, \ldots, n_{\mathrm{P}}$).

The $(2n_{\mathrm{P}}+n_{\mathrm{R}}+n_{\mathrm{V}})\times(2n_{\mathrm{P}}+n_{\mathrm{R}}+n_{\mathrm{V}})$ dimensional matrix **M** is a transition matrix corresponding to a Markovian process. Similar to Eq. (10), it is easy to verify that **M** is an *irreducible*, *aperiodic* and *stochastic* matrix. So **r** converges to the principle eigenvector of **M** with principle eigenvalue of 1. Given initial values of **r**, MutualRank iteratively updates **r** until the convergence criterion $\|\mathbf{r}\|_1 \le$ ε is met for some small ε. The initial values of **paut**, **psnd, rimp** and **vprs** are set to $\mathbf{e}_{\mathbf{p}}/(2n_{\mathrm{P}}+n_{\mathrm{R}}+n_{\mathrm{V}})$, $\mathbf{e}_{\mathbf{p}}/(2n_{\mathrm{P}}+n_{\mathrm{R}}+n_{\mathrm{V}})$, $\mathbf{e}_{\mathbf{r}}/(2n_{\mathrm{P}}+n_{\mathrm{R}}+n_{\mathrm{V}})$, and $\mathbf{e}_{\mathbf{v}}/(2n_{\mathrm{P}}+n_{\mathrm{R}}+n_{\mathrm{V}})$ respectively, where $\mathbf{e}_{\mathbf{p}}$, $\mathbf{e}_{\mathbf{r}}$ and $\mathbf{e}_{\mathbf{v}}$ are n_{P}-, n_{R}- and n_{V}-dimensional unit vectors respectively.

4. EXPERIMENTAL EVALUATION

4.1 Evaluation Metrics

In evaluation, we do not use the traditional IR metric *Precision-at-k* [14] because there exists a partial order between paper and researcher pairs. What's more, each benchmark paper has a recommendation count and our design of evaluation metric should

reflect this information which cannot be reflected by *Precision-at-k*. Meanwhile, it is also hard to employ *Normalized Discounted Cumulative Gain* (NDCG@*k*) in our context although NDCG@*k* is almost a de facto metric for ranking evaluation beyond binary relevance. This is because we are estimating the importance of papers in the whole computational linguistics area, and even for human, it is very difficult to judge the levels of relevance. The performance metric we use is called *Recommendation Intensity*. Let P be the list of top-k returned papers. For each paper p in P, the *recommendation intensity of p at k*, denoted as $RI(p)@k$, is defined as

$$RI(p)@k = \begin{cases} c_p / \log(c_p) + (1 - o_p / k) & p \in \text{BenchP} \\ 0 & p \notin \text{BenchP} \end{cases}. \quad (20)$$

where c_p is the number of recommendations of p in *BenchP* and o_p is the ranked order of p in P. Eq. (20) means that if a paper p has larger recommendation counts (with larger c_p) and is ranked higher (with smaller o_p), then its recommendation intensity $RI(p)@k$ is higher. $\log(c_p)$ is a controlling factor for preventing a few papers with extremely high recommendation counts from dominating the value of recommendation intensity of the whole result set. $(1-o_p/k)$ is to reflect the fact that a returned relevant result p (i.e. p is also in *BenchP*) is more useful if p is ranked higher in P. This is because users tend to look at only the first several items in the returned result list. Thus, the higher the rank of a relevant paper is, the more chances it has to be read by the user. *Recommendation intensity of P at k*, denoted as $RI(P)@k$, is defined as

$$RI(P)@k = \sum_{p \in P, o_p \le k} RI(p)@k \,. \quad (21)$$

The bigger $RI(P)@k$ is, the better a ranking method is. $RI(P)@k$ is an extension to *Precision-at-k*. In fact, if we ignore recommendation counts, $RI(P)@k$ degenerates to *Precision-at-k*. Recommendation intensity is also used for ranking researchers. Based on *BenchR1*, for each researcher r in the top-k result list R, *recommendation intensity of r at k* is defined as

$$RI_1(r)@k = \begin{cases} c_r / \log(c_r) + (1 - o_r / k) & r \in \textit{BenchR1} \\ 0 & r \notin \textit{BenchR1} \end{cases}, \quad (22)$$

where c_r is the recommendation count of r in *BenchR1* and o_r is the order of r in R. The *recommendation intensity of R at k* is

$$RI_1(R)@k = \sum_{r \in R, o_r \le k} RI_1(r)@k \,. \quad (23)$$

For evaluation using *BenchR2*, if a researcher r is the o_r-th one on the top-k result list R and the g_r-th in BenchR2, then

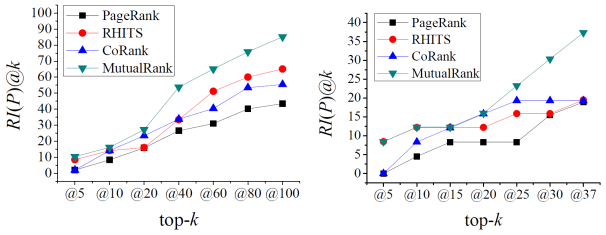
$$RI_2(r)@k = \begin{cases} 1 + (o_r - g_r) / k & r \in \textit{BenchR2} \\ 0 & r \notin \textit{BenchR2} \end{cases}, \quad (24)$$

and $RI_2(R)@k$ is defined as to $RI_1(R)$.

We compare MutualRank with PageRank, RHITS and CoRank using the above benchmark datasets and metrics. CoRank works on both PIN and RCN. CoRank is rephrased as follows,

$$\mathbf{r} = \begin{bmatrix} (1-\gamma)\overline{\mathbf{P}}^n & \gamma\overline{\mathbf{RP}}(\overline{\mathbf{PRRP}})^l \\ \gamma\overline{\mathbf{RP}}(\overline{\mathbf{PRRP}})^l & (1-\gamma)\overline{\mathbf{R}}^m \end{bmatrix}^{\mathrm{T}} \mathbf{r} \,, \quad (25)$$

where $\mathbf{r} = [\mathbf{prnk}^{\mathrm{T}}, \mathbf{rimp}^{\mathrm{T}}]^{\mathrm{T}}$, $\mathbf{prnk}^{\mathrm{T}}$ and $\mathbf{rimp}^{\mathrm{T}}$ are the rank vectors for papers and researchers respectively. The superscripts l is the number of times a random surfer does inter-network walk, and m and n denote the number of intra-network random walks.



(*a*) *#recommend* ≥ 1 (*b*) *#recommend* ≥ 2

Figure 3. Recommendation intensity curves of different methods on ranking papers.

4.2 Results for Ranking Papers

We first demonstrate how MutualRank compensate the problem of time distortion compared to PageRank and RHITS. Table 3 lists the relevant returned results of MutualRank under parameter settings that ξ = 0.5, γ = 0.5, λ = 0.85 and ε = 1e-5. Out of the top-100 results returned by MutualRank, 36 are verified as relevant according to *BenchP*. (1) Out of the 36 relevant results, 16 papers (in boldface) are published in the 1990s and 17 papers (in normal font) come from the 2000s which demonstrates that MutualRank not only accurately identifies the contributions of the last decade of the 20th century to the development of statistical natural language processing, but also assigns enough importance to the new century where many practical applications flourish, many new research problems emerge and the progresses in NLP are ever faster than before, armed with the promising tools developed and together with the boom of web science. To better catch up this, readers can refer to Table 5 and verify that all the top-15 papers published in the 1990s are indeed amongst the most influential papers through the years. (2) MutualRank returns extra papers of high recommendations that are lost by RHITS, e.g. P95-1026[4] and W02-1011[3] etc.

Table 3. Relevant Papers in top-100 MutualRank Results

2:J93-2003[7]	**24:P95-1026[4]**	63:J01-4004[1]
3:J86-3001[1]	25:P03-1054[2]	*66:J88-1003[1]*
4:P02-1040[2]	**26:P97-1003[3]**	68:J00-3003[1]
5:J96-1002[2]	**27:J95-2003[1]**	**72:H94-1028[1]**
8:J90-2002[1]	30:W02-1011[3]	*73:P87-1022[1]*
13:J92-4003[1]	33:W02-1001[2]	**76:C96-2141[1]**
15:J96-2004[1]	38:P02-1053[1]	79:P06-1055[3]
17:A00-2018[1]	43:P05-1022[2]	80:J01-2001[3]
18:J95-4004[1]	45:N03-1028[1]	82:P05-1012[1]
19:J02-3001[2]	**47:J98-1004[1]**	**86:J93-2006[1]**
21:P98-2127[1]	52:P05-1033[1]	**89:J94-2001[2]**
23:W96-0213[2]	58:J04-4002[2]	95:P02-1035[1]

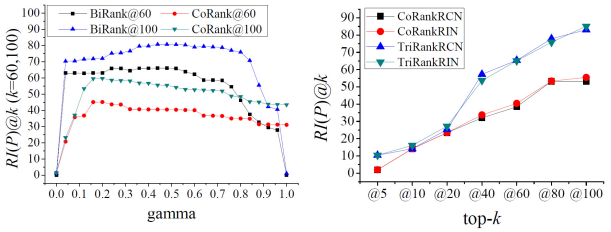
Note: ξ = 0.5, γ = 0.5, λ = 0.85 and ε = 1e-5.

Figure 3 shows the results of ranking papers. For Figure 3(a), we compare the results of each competitor algorithm with *BenchP*. For Figure 3(b), we use a subset of papers in *BenchP* with at least two recommendations. To obtain the results in Figure 3, we set the parameters as follows, (i) for MutualRank ξ = 0.5, (ii) for

Table 5. Top-15 papers returned by MutualRank and their ranks by other competitors

MR	Title	AAN ID	Rank by #cite	CR	PR	HITS
1	Building A Large Annotated Corpus Of English: The Penn Treebank	J93-2004	1 (775)	5	6	17
2	The Mathematics Of Statistical Machine Translation: Parameter Estimation	J93-2003	2 (615)	6	7	4
3	Attention Intentions And The Structure Of Discourse	J86-3001	8 (343)	7	8	**1079**
4	Bleu: A Method For Automatic Evaluation Of Machine Translation	P02-1040	3 (591)	14	29	1
5	A Maximum Entropy Approach To Natural Language Processing	J96-1002	7 (344)	8	12	24
6	A Systematic Comparison Of Various Statistical Alignment Models	J03-1002	5 (473)	**58**	**71**	5
7	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text	A88-1019	16 (234)	1	1	**291**
8	A Statistical Approach To Machine Translation	J90-2002	29 (189)	3	4	26
9	Minimum Error Rate Training In Statistical Machine Translation	P03-1021	4 (475)	42	**64**	2
10	Moses: Open Source Toolkit for Statistical Machine Translation	P07-2045	10 (325)	**102**	**195**	6
11	Statistical Phrase-Based Translation	N03-1017	6 (436)	**60**	**83**	3
12	The Berkeley FrameNet Project	P98-1013	19 (220)	21	31	**267**
13	Class-Based N-Gram Models Of Natural Language	J92-4003	28 (193)	9	13	**89**
14	Accurate Methods For The Statistics Of Surprise And Coincidence	J93-1003	20 (218)	27	46	**103**
15	Assessing Agreement On Classification Tasks: The Kappa Statistic	J96-2004	32 (179)	30	54	**247**

Note: PR stands for PageRank.



(*a*) Parameters (*b*) Network models

Figure 4. Parameters and network models on ranking performance (BiRank: MutualRank without venue information; TriRank: MutualRank with venue information).

CoRank, $m = 2$, $n = 2$, $l = 1$ (parameters that obtain the best performances in the original paper [26]), and (iii) for both $\gamma = 0.5$. Results show that MutualRank has the best performance. Taking MutualRank and RHITS for example, there are 20 MutualRank-only relevant papers (those benchmark papers that are only recommended by MutualRank) on the top-100 result list and among them 5 are recommended more than twice. There are only 11 RHITS-only papers and only 2 of them are recommended more than twice. As MutualRank applies HITS-style ranking on PIN, the results verify that using different types of importance values for certain objects does improve ranking performance.

An interesting observation is that, although RHITS uses PIN only, it sometimes performs even better than CoRank, e.g. when $k \geq 40$ in Figure 3(a) and $k \leq 15$ in Figure 3(b). After investigating the results of both algorithms we find that RHITS returns more relevant results than CoRank and some RHITS-only papers have high recommendation counts in *BenchP*. A possible explanation is that CoRank uses PageRank for ranking in PIN which fails to reflect the soundness of the paper and thus still biases towards old papers to some extent. The time distortion of CoRank can also be seen from the CR column of Table 5. High rank papers are still mostly old papers and ranks of papers after 2000 are relatively lower. An extreme example is P07-2045 (shadowed in Table 5). Although it is ranked 10th by MutualRank and receives 325 citations in less than 4 years, it is only ranked 102th and 195th by CoRank and PageRank respectively.

To understand why MutualRank wins, we take a deep look at the relevant results. We see from Table 4 that, besides more relevant papers by MutualRank, most common (highly recommended) relevant results are ranked higher in MutualRank than in CoRank except for the papers in italic (Notice that these papers are published in earlier years). There are also 15 MutualRank-only results where 6 papers are recommended more than 2 times in *BenchP*.

Table 4. Shared Relevant Papers of MutualRank and CoRank

Paper id	#rec	Rank		Paper id	#rec	Rank	
		MR	CR			MR	CR
J93-2003	7	2	6	*W96-0213*	*2*	*23*	*15*
J86-3001	1	3	7	P95-1026	4	24	49
P02-1040	2	4	14	*P97-1003*	*3*	*26*	*23*
J96-1002	2	5	8	W02-1011	3	30	75
J90-2002	*1*	*8*	*3*	W02-1001	2	33	77
J92-4003	*1*	*13*	*9*	P02-1053	1	38	61
J96-2004	1	15	30	*J88-1003*	*1*	*66*	*19*
A00-2018	1	17	25	*H94-1028*	*1*	*72*	*50*
J95-4004	1	18	32	*C96-2141*	*1*	*76*	*64*
J02-3001	2	19	62	*J93-2006*	*1*	*86*	*67*
P98-2127	1	21	68				

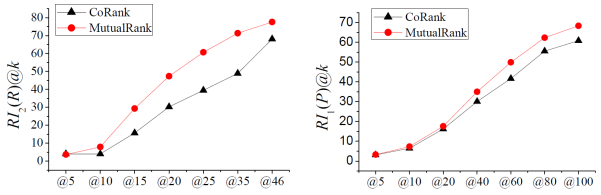
Note: MR stands for MutualRank and CR stands for CoRank.

Figure 4 shows the results of MutualRank under different different parameter settings and network models. We include CoRank for comparison with γ varying from 0 to 1 with parameter step of 0.02 (see Eq. (25)). To see how researcher and venue information helps improve ranking performance, we fix γ in MutualRank to 0.5 and vary ξ from 0 to 1 with parameter step of 0.02. For a fairer comparison, we do not use venue information. MutualRank thus degenerates to BiRank in Figure 4(a). In both MutualRank and CoRank, parameter γ governs how much information outside the PIN is incorporated for improving ranking. The larger γ is, the more researcher information is used. When $\gamma = 0$ CoRank degenerates to PageRank and MutualRank degenerates to First-Order

Table 7. Comparisons between different rankings of 8 well-known venues

	Venue Rank								Correlation				
	CL	**ACL**	**COLING**	**EMNLP**	**NAACL**	**EACL**	**HLT**	**IJCNLP**	**AM**	**MR**	**H1**	**H2**	**Avg_Correl**
AM	1	2	3	5	4	6	7	8	1.00	0.81	0.88	0.86	0.89
MR	**3**	1	2	4	6	7	***5***	8	0.81	1.00	**0.64**	0.83	0.82
H1	1	2	5	4	4	6	8	7	0.88	0.64	1.00	0.88	0.85
H2	2	1	4	3	6	5	7	8	0.86	0.83	0.88	1.00	0.89

Note: AM – ArnetMiner; MR – MutualRank; H – Human judgments; Avg_Correl – Average correlation with the other three rankings; CL – The Computational Liguistics journal



(a) *BenchR1* (b) *BenchR2*

Figure 5. Recommendation intensity curves of different methods on ranking researchers.

HITS. $\gamma = 1$ means that, loosely speaking, paper rank in both MutualRank and CoRank is the aggregated importance of its authors. As is expected, when γ increases from 0 to 1, ranking performances of both MutualRank and CoRank increase. This means adding metadata information such as research (or venue) influence network indeed improves ranking of papers. This performance gain peaks at some points (around $\gamma = 0.46$ for MutualRank and $\gamma = 0.20$ for CoRank) and turns to decrease when γ becomes too large. The results in Figure 4(a) show that using First-Order HITS instead of PageRank on PIN helps improve ranking performance a lot. Figure 4(b) shows the influence of network model on recommendation intensity, where both γ and ξ are fixed to 0.5. We compare two versions of MutualRank and two versions CoRank. On the whole, performances using RIN is slightly better RCN.

To get an intuitive understanding of MutualRank, we list the top-15 papers returned by MutualRank and its competitors in Table 5. We include the ranks by CoRank, PageRank, RHITS and #cite (citation count) for comparison. It is clear that the ranks given by MutualRank are quite reasonable. We also see how severely PageRank and RHITS bias towards old and new papers respectively (see the boldfaced items). The results of CoRank are in a sense close to PageRank. The IDs of papers are given so that interested readers can find them on the AAN website for more information.

4.3 Results for Ranking Researchers

For ranking researchers, we compare MutualRank to CoRank. The recommendation intensity curves in Figure 5 show that MutualRank outperforms CoRank by at most 53.8% under *BenchR1* and 19.7% under *BenchR2*.

Taking *BenchR2* for example, in the top-10 results returned by MutualRank, two matches *BenchR2*. They are $r_{(5)}$ = 'Della Pietra, Vincent J.' and $r_{(10)}$ = 'Manning, Christopher D.', but none of CoRank's top-10 results matches *BenchR2*. Moreover, compared to CoRank results, most MutualRank ranks are closer to their benchmark orders. For example, the top 4 researchers in *BenchR2* 'Och, Franz Josef', 'Ney, Hermann', 'Koehn, Philipp' and 'Marcu, Daniel' are ranked 24th, 19th, 50th and 16th by MutualRank and 45th, 11th, 65th, and 39th by CoRank respectively.

Table 6 shows the top-15 researchers by MutualRank with their benchmark and CoRank ranks. The columns "Rank in *BenchR1*" and "Rank in *BenchR2*" are in the form of o_r (*#rec*) and r (*#cite*), respectively, where o_r is the order of r in the benchmark, *#rec* and *#cite* are the number of recommendations in *BenchR1* and number of incoming citations in *BenchR2* respectively. Two lines in boldface are exact match of *BenchR2* by MutualRank. Exact match by *BenchR1* is italicized. The symbol '-- (--)' means that the returned researcher is not in *BenchR1*.

Table 6. Top-15 researchers returned by MutualRank

MR	**Researcher Name**	**CR**	**Rank in *BenchR1***	**Rank in *BenchR2***
1	Church, Kenneth Ward	1	-- (--)	19 (2667)
2	Marcus, Mitchell P.	9	-- (--)	12 (3411)
3	Pereira, Fernando	10	-- (--)	15 (3070)
4	Brill, Eric	26	22 (3)*	35 (1829)
5	***Della Pietra, Vincent J.***	***12***	***5 (6)***	***5 (3978)***
6	Joshi, Aravind K.	3	-- (--)	30 (2065)
7	McKeown, Kathleen R.	2	-- (--)	33 (1956)
8	Hovy, Eduard	8	-- (--)	24 (2327)
9	Grishman, Ralph	6	-- (--)	31 (2023)
10	Manning, Christopher D.	22	2 (13)	6 (3915)
11	**Mercer, Robert L.**	**4**	10 (5)	**11 (3453)**
12	Della Pietra, Stephen A.	15	10 (5)	10 (3549)
13	Collins, Michael John	21	10 (5)	7 (3909)
14	Johnson, Mark	13	5 (6)	17 (2908)
15	Yarowsky, David	17	22 (3)*	23 (2335)

4.4 Results for Ranking Venues

Although it is difficult to give a consolidated quantitative metric for verifying venue ranking, the results show that the top venues returned by MutualRank conform to our common knowledge on conference and journal reputations. Among all the 273 venues, the top 10 are $v_{(1)}$ = ACL, $v_{(2)}$ = COLING, $v_{(3)}$ = CL, $v_{(4)}$ = EMNLP, $v_{(5)}$ = HLT, $v_{(6)}$ = NAACL, $v_{(7)}$ = wDSANL, $v_{(8)}$ = ANLP, $v_{(9)}$ = CoNLL, and $v_{(10)}$ = EACL. wDSANL is a shorthand for 'Workshop on Speech and Natural Language' and CL stands for 'Journal of Computational Linguistics'. Following are some other reputed conferences including $v_{(11)}$ = LREC, $v_{(14)}$ = MU, $v_{(16)}$ = IJCNLP, $v_{(18)}$ = NLG, $v_{(19)}$ = wSMT, $v_{(20)}$ = VLC. MU is short for the 'Message Understanding Conference', NLG stands for 'International Conference on Natural Language Generation', wSMT stands for 'Workshop on Statistical Machine Translation' and VLC abbreviates 'Workshop on Very Large Corpora'.

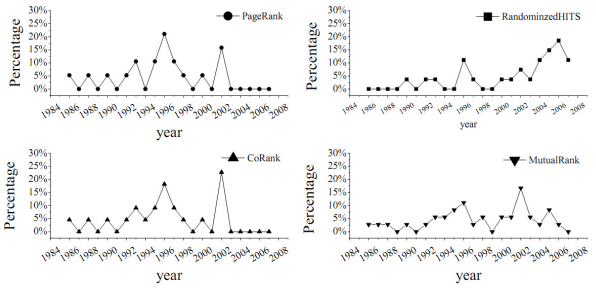
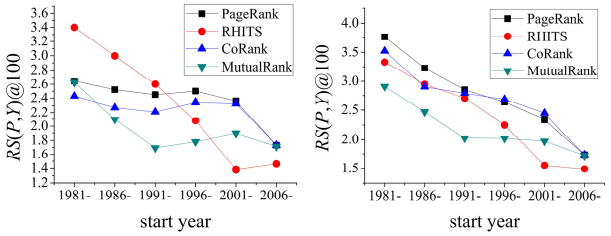


Figure 6. Distribution of relevant returned papers (k = 100)



(a) Curves of relevant papers (b) Curves of returned papers

Figure 7. Recommendation sensitivity curves of different methods on ranking papers (k = 100).

To make a qualitative investigation, we compare our results with the 2008 conference ranks given by ArnetMiner[5] [22] and 2 human evaluations (H1 and H2 in Table 7) of 8 well-known venues, and analyze the correlations between these four rankings. As Table 7 shows, the four rankings are highly correlated with each other. The only significant difference lies between MR and H1, where the correlation coefficient of 0.64 is not high enough to be consistent. With in-depth investigation of H1 and MR, we found the problem lies in the following points. Firstly, in MutualRank, a joint conference such as HLT-NAACL is split into two different venues. Such a way assigns extra prestige to less-prestigious conferences, that is HLT in this context. We are unable to distinguish between jointly held conferences using AAN metadata only. Secondly, the Journal of Computational Linguistics is by all means the most prestigious venue to NLP researchers. However, in MutualRank it is ranked lower than ACL and COLING because the latter two conferences have far more papers published every year and so their aggregated prestige is higher.

Very interestingly in MutualRank, ACL-Companion (12th), COLING-Companion (13th), HLT-Companion (15th) and NAACL-Companion (17th) all have high prestige too. This phenomenon reveals that, on one hand, researchers often focus on the most influential conferences and thus even the short or demo papers published in these conferences gain many citations, and on the other hand, it is more attractive to important researchers for publishing their short or demo papers on these prestigious conferences. Moreover, MutualRank returns some well-known workshops too, e.g., $v_{(21)}$ = wDAD (SIGDIAL Workshop on Discourse And Dialogue), $v_{(23)}$ = BioNLP (a series of workshops on Natural Language Processing in Biology and Biomedicine).

4.5 Time Distributions

Finally, we study the time distributions of returned papers of different ranking algorithms. Figure 6 shows the yearly distributions of the relevant papers in the top-100 papers returned by different algorithms. (1) From the chart we can see that two peaks around year 1996 and 2002 can be identified by PageRank, CoRank and MutualRank but the RHITS algorithm has an obvious bias towards new papers after 2000 year. Bias here means favoring certain years or year ranges. PageRank has a strong bias towards older papers and older papers receive citations more easily than new papers. CoRank and MutualRank are less biased in this sense. (2) Around each of these two peaks, MutualRank has a slowly-increasing and slowly-decreasing slope, which is much close to the real development process of a booming area while PageRank and CoRank has sharper busts at the two peaks. From the above, we argue that MutualRank not only captures the booming of statistical natural language in the 1990s but also shows enough respect of the flourishing development in the new century.

To make a deeper investigation of the problem of time distortion identified in Section3.1.1, we define a new metric *recommendation sensitivity* for ranking papers. Given a year range $Y = [y_s, y_e]$ and a paper list P returned by some ranking algorithm, for each year $y \in Y$, let $pub(y)@k$ and $pub(Y)@k$ be the percentages of papers in P that are published in year y and in year range Y respectively. Let $\overline{pub(y)@k} = \sum_{y \in Y} pub(y)@k \,/\, |Y|$. The *recommendation sensitivity of P at k*, denoted as $RS(P)@k$, is defined as

$$RS(Y)@k = \sum_{y \in Y} -dev(y)@k \bullet \log\left(dev(y)@k\right) , \qquad (26)$$

where $dev(y)@k = \left| pub(y)@k - \overline{pub(y)@k} \right|$. If $dev(y)@k = 0$, the corresponding *log*-element in Eq. (7) is set to 0. Recommendation sensitivity measures how biased the yearly distribution of returned papers during a year range is. Big recommendation sensitivity means that the yearly distribution of the returned papers has very sharp peaks, i.e. biasing towards certain periods. What's more, we can see how recommendation sensitivity changes during different periods by choosing different start years of the year ranges with fixed end year and drawing the recommendation sensitivity curves corresponding to the year ranges. Figure 7 shows the recommendation sensitivity curves (RSC) of different methods, where we fix the end year as 2010 and vary the start year from 1981 to 2006 with step = 5. Figure 7(a) shows the RSCs of relevant papers, i.e. the top-100 papers that match *BenchP*, while Figure 7(b) illustrates the RSCs of the top-100 papers returned by different algorithms. Parameters of different ranking algorithms are just set as in Figure 3. An unbiased and fair ranking algorithm should have both small recommendation sensitivity values in different year ranges and a flat RSC in the whole time period.

From Figure 7(a), we see that RHITS is very time-sensitive. The sharp slope reflects the fact that RHITS heavily biases towards new papers. By contrast, PageRank and CoRank are relatively not that sensitive. However, the sharp slope from year 2001 to 2006 reveals that PageRank and CoRank both bias towards old papers. On the contrary, the smaller values and flatter slop of MutualRank RSC from year 1986 to 2010 demonstrate that MutualRank is not only less sensitive to changes in year range but also less biased

[5] http://v1.arnetminer.org/page/conference-ranking/html/NLP.html

towards papers of certain time periods. There are two more points about MuturalRank RSC worth note. (1) The sharp slope from 1981 to 1986 shows that MutualRank returns few papers which are too old. (2) The flat slope from year 2001 to 2006 reveals the fact that MutualRank does not omit some important recent papers as PageRank and CoRank do. To sum up, MutualRank is the most unbiased method for ranking papers. Similar observations can be found in Figure 7(b). The point reserving attention is that, unlike in Figure 7(a), PageRank and CoRank also have sharp slopes from year 1981 to 1986 and from year 1986 to 1991. This is understandable because in the early days fewer papers are published every year, so all the methods have fewer returned results in these years (Note that these returned papers are not necessarily relevant with respect to *BenchP*).

5. CONCLUSION

This paper proposes a new framework MutualRank towards an unbiased ranking of papers. MutualRank employs the mutual reinforcement relationships between papers, researchers and publication venues to fairly rank papers of certain time periods, as well as researchers and venues. It models the intra- and inter-network rankings in a unified way and computes the ranking vectors of papers, researchers and venues in an iterative fashion. Using the manually collected benchmark datasets for papers and researchers, we show through experiments that MutualRank outperforms state-of-the-art competitors including PageRank, Randomized HITS and CoRank in the following aspects: (1) MutualRank returns more relevant highly-ranked papers and researchers, and (2) results of MutualRank are more unbiased. The MutualRank rankings of venues are verified to be reasonable. We also present a detailed discussion on the problem of time distortion in ranking papers and propose methods for judging biases of ranking.

In the future work, we will study the convergence properties of the ranking algorithms, establish a better benchmark, and consider more indicators for evaluation such as awards for ranking papers and researchers and social factors. The idea of MutualRank can be also applied to other network resource ranking scenarios where multiple heterogeneous sub-networks are interwoven with different node types and edge semantics. For example, in a heterogeneous social network where friend network, email network, collaboration network and affiliation network co-exist, mutual reinforcement information can be modeled to help ranking nodes in a synthetic way.

6. ACKNOWLEDGEMENT

This work was partially supported by National Science Foundation of China (No.61075074 and No.61070183), Natural Science Foundation of Chongqing (No.cstc2012jjB40012), and the Key Discipline Fund of National 211 Project (Southwest University: NSKD11013). We thank Dr. Yang Liu for annotating venue rankings. Thanks also go to Dr. Jianmin Yao for venue ranking annotation and helpful discussions and comments.

7. REFERENCES

[1] Barabási, A.-L. and Albert, R. 1999. Emergence of Scaling in Random Networks. *Science*, 286: 509–512.

[2] Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.*, 30, 1-7, (Apr. 1998) 107–117.

[3] Chen, P., Xie, H., Maslov, S., and Redner, S. 2007. Finding scientific gems with Google's PageRank algorithm. *J. Informetrics*, 1, 1 (Jun. 2007), 8–15.

[4] Das, S., Mitra, P., and Lee Giles, C. 2011. Ranking Authors in Digital Libraries. In *Proc. JCDL'11*, 251–254.

[5] Ding, Y., Yan, E., Frazho, R., and Caverlee, J. 2009. PageRank for Ranking Authors in Co-citation Networks. *J. Am. Soc. Inf. Sci. Technol.*, 60, 11 (Jun. 2009), 2229–2243.

[6] Garfield, E. 1972. Citation analysis as a tool in journal evaluation. *Science*, 178, 60 (Nov. 1972), 471–479.

[7] Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci.*, 102, 46 (Nov. 2005), 16569–4.

[8] Jensen, C. S., Cao, X., and Cong, G. 2010. Mining Significant Semantic Locations from GPS Data. In *Proc. VLDB*, 3, 1–2 (Sep. 2010), 1009–1020.

[9] Katerattanakul, P., Han, B., and Hong, S. 2003. Objective quality ranking of computing journals. *Commun. ACM*, 46, 10 (Oct. 2003), 111–114.

[10] Kleinberg, J. M. 1999. Authoritative Sources in a Hyperlinked Environment. *J. ACM*. 46, 5 (Sep. 1999), 604–632.

[11] Lefebvre, M. 2006. *Applied Stochastic Processes*. Springer.

[12] Lempel, R., and Moran, S. 2001. SALSA: The Stochastic Approach for Link-Structure Analysis. *ACM Trans. Internet Tech.* 19, 2 (Apr. 2001), 131-169.

[13] Li, X., Liu, B., and Yu, P. 2008. Time Sensitive Ranking with Application to Publication Search. In *Proc. ICDM'08*, 893–898.

[14] Manning, C. D., Raghavan, R., and Schütze, H. 2008. *Introduction to Information retrieval.* Cambridge University Press.

[15] Nerur, S., Sikora, R., Mangalaraj, G., and Balijepally, V. 2005. Assessing the relative influence of journals in a citation network. *Commun. ACM*, 48, 11 (Nov. 2005), 71–74.

[16] Newman, M. E. J. 2002. Assortative Mixing in Networks. *Phys. Rev. Lett*., 89, 20: 208701–5.

[17] Ng, A. Y., Zheng, A. X., and Jordan, M. I. 2001. Stable Algorithms for Link Analysis. In *Proc. SIGIR'01*, 258–266.

[18] Ng, M. K., Li, X., and Ye, Y. 2011. MultiRank: co-ranking for objects and relations in multi-relational data. In *Proc. KDD'11*, 1217–1225.

[19] Radicchi, F., Fortunato, S., Markines, B., and Vespignani, A. 2009. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E*, 80, 5 (Nov. 2009), 056103–12.

[20] Radev, D. R., Muthukrishnan, P., and Qazvinian, V. 2009. The ACL Anthology Network. In *Proc. NLPIR4DL'09*, 54–61.

[21] Sayyadi, H., and Getoor, L. 2009. FutureRank: Ranking Scientific Articles by Predicting their Future PageRank. In *Proc. SDM'09*. 533–544.

[22] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. 2008. ArnetMiner: extracting academic social networks. In *Proc. KDD'08*. 990–998.

[23] Walker, D., Xie, H., Yan, K.-K., and Maslov, S. 2007. Ranking scientific publications using a model of network traffic. *J. Stat. Mech.*, 7 (Jun. 2007), 06010–9.

[24] Yan, E., and Ding, Y. 2009. Applying centrality measures to impact analysis: A coauthorship network analysis. *J. Am. Soc. Inf. Sci. Technol.* 60, 10 (Oct. 2009), 2107–2118.

[25] Yan, S., and Lee, D.-W. 2007. Toward Alternative Measures for Ranking Venues: A Case of Database Research Community. In *Proc. JCDL'07*, 235–244.

[26] Zhou, D., Orshanskiy, S. A., Zha, H., and Lee Giles, C. 2007. Co-Ranking Authors and Documents in a Heterogeneous Network. In *Proc. ICDM'07*, 739–744.

[27] Zhuge, H., and Zhang, J. 2010. Topological Centrality and Its e-Science Applications. *J. Am. Soc. Inf. Sci. Technol.*, 61, 9 (May 2010), 1824–1841.