A Robust Model for Paper-Reviewer Assignment

Xiang Liu
New York University
Brooklyn, NY
xl493@nyu.edu

Torsten Suel
New York University
Brooklyn, NY
suel@poly.edu

Nasir Memon
New York University
Brooklyn, NY
memon@poly.edu

ABSTRACT

Automatic expert assignment is a common problem encountered in both industry and academia. For example, for conference program chairs and journal editors, in order to collect "good" judgments for a paper, it is necessary for them to assign the paper to the most appropriate reviewers. Choosing appropriate reviewers of course includes a number of considerations such as expertise and authority, but also diversity and avoiding conflicts. In this paper, we explore the expert retrieval problem and implement an automatic paper-reviewer recommendation system that considers aspects of expertise, authority, and diversity. In particular, a graph is first constructed on the possible reviewers and the query paper, incorporating expertise and authority information. Then a Random Walk with Restart (RWR) [1] model is employed on the graph with a sparsity constraint, incorporating diversity information. Extensive experiments on two reviewer recommendation benchmark datasets show that the proposed method obtains performance gains over state-of-the-art reviewer recommendation systems in terms of expertise, authority, diversity, and, most importantly, relevance as judged by human experts.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.3.3 [**Information Search and Retrieval**]: Relevance Feedback

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Review Assignment; Expert Retrieval; Information Propagation; Topic Model; Random Walk; Diversity; Ranking

RecSys '14, October 06 - 10 2014, San Jose or vicinity, CA, USA

http://dx.doi.org/10.1145/2645710.2645749

1. INTRODUCTION

The task of expert recommendation and assignment is a common problem in both industry and academia. Consider a job recruiting process where, in order to decide whether an applicant should receive an on-site interview, HR has to make an evaluation of an applicant's educational record, previous work experience, personal skills, and fit with job requirements. For the task of planning a workshop and deciding who should be invited, expertise, communication skills, and diverse background might be preferred. In general, for each expert recommendation task, several aspects are usually jointly considered to make the final decision. Accordingly, an automatic expert recommendation system should take into account multiple criteria for a specific task.

In this paper, we focus on the problem of paper-reviewer recommendation, which has been widely studied [2, 8, 3, 4]. For conference program chairs and journal editors, a good review assignment should satisfy several criteria, e.g., reviewers' authority, expertise, diversity, availability, conflict, etc. However, most existing methods often focus on one aspect, expertise (topic coverage), in the design of their approach, and attempt to maximize their definitions of expertise. In this work, we take three criteria as the main design objectives, including authority, whether the reviewer has a good recognition in the larger scientific community; expertise, whether the reviewer is a specialist in the specific domain related to the paper; and diversity, whether the selected reviewers have diverse research interests and background. The final goal of our work is to assign papers to reviewers that are considered highly qualified to perform the review. To achieve that goal, we are first targeting two intermediate goals: expertise, and authority. While expertise can be approximated by a variety of methods, e.g., text similarity, topic similarity, etc., authority can be explored by graph-based propagation. We will show that by maximizing expertise and authority, we can in fact achieve a better assignment as judged by human experts. Finally, we show how to also achieve the third objective, diversity, without significant decreases in expertise, authority, or quality as viewed by human experts.

Given a query paper and a candidate reviewer pool with N reviewers (shown in Figure 1), we first construct a graph with $N + 1$ nodes representing the candidate reviewers and the query paper. A standard topic model is then applied to the query paper and the published papers of the candidate reviewers, so that the link between a candidate reviewer and the query paper can be built and measured by the topic model-based similarity. Thus it ensures a strong connection between the query paper and a reviewer if he or she has strong background knowledge in the specific area (e.g., has published

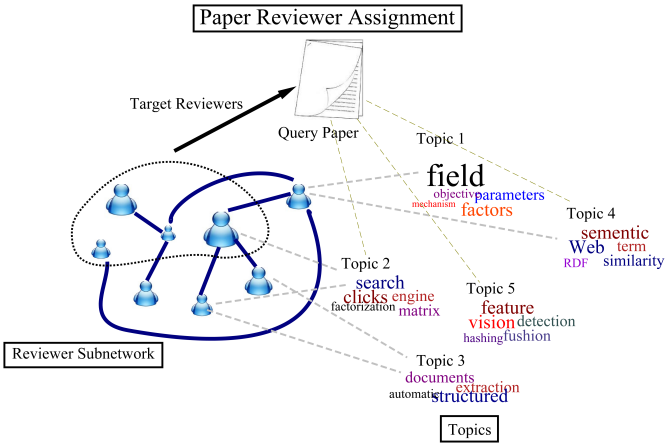


Figure 1: The proposed graph contains the query paper, candidate reviewers, and topics. Given a query paper, the recommendation algorithm selects the nodes with most expertise and propagates the query to the other nodes in the graph. After the propagation process, each node is assigned with a utility score, which is used to measure both the relevance between the node, and the query, and the authority of the node in the graph. Finally, a set of nodes with highest scores are selected as target reviewers.

a lot of papers on the same topic as the query paper). To achieve the authority objective, academic co-authorship is considered as the social link representation among the N reviewers (the Reviewer Subnetwork in Figure 1), so that by using graph propagation, the authority information can be incorporated. To better integrate the expertise and authority scores for each candidate reviewer, we propose a Random Walk with Restart (RWR) model. With such a model, there is some probability to jump back to the start node at each step. The hope is that the RWR will achieve a better balance between expertise and authority in the propagation process.

To achieve the diversity objective, the N candidate reviewers are first clustered into groups based on their research topic distributions. We then select only a small set of individual nodes from each cluster so that an unbiased selection is achieved from diverse research areas. We call such strategy a sparsity constraint for each cluster. Finally, together with the sparsity constraint, an RWR process employed on the $N+1$ nodes will provide us a stable probability of each node. In our recommender system, this probability is considered as the ranking score for a reviewer under the criteria of expertise, authority and diversity.

We formulate the RWR and the group sparsity as a uniformed optimization framework. An efficient gradient descent-based method is proposed to solve the minimization objective. Extensive experiments confirm the effectiveness of the proposed method as compared to the state-of-the-arts.

The rest of this paper is organized as follows. In Section 2 we discuss related work. We explain our graph construction in Section 3. Then we describe our problem formulation and optimization approach for paper-reviewer assignment in detail in Section 4. In Section 5, we present our experimental results. Finally, we provide concluding remarks in Section 6.

2. RELATED WORK

There has been a lot of research on both reviewer assignment and graph propagation. In this section, we summarize the related work on these two topics.

2.1 Paper-Reviewer Assignment

Some recent work has focused on utilizing information retrieval and machine learning techniques to solve the problem of paper-reviewer assignment. For example, the widely used Toronto recommender system [5] addressed the assignment as a minimum cost network flow using some novel metrics. It also considered reviewers' bids, which expressed their interests or disinterests in specific papers, as available feedbacks. Hettich *et al.* [7] used TF-IDF to exploit the suitability between manuscripts and reviewers. Mimno and McCallum [8] applied a topic model to measure a reviewer's expertise. Charlin *et al.* [5] utilized LDA model, linear regression and collaborative filtering to determine reviewer assignments. Tang *et al.* [4] assumed that every reviewer had an expertise level, which was already known. Then they defined some specific matching criteria to optimize the reviewer arrangement procedure. Rodriguez *et al.* [9] built a co-authorship graph with the references of a submitted paper as starting points to suggest reviewers. Conry *et al.* [6] first studied the preference of reviewers for specific papers as available feedbacks. A linear programming-based optimization formulation was then used to solve the reviewer assignment problem.

Most existing papers have focused on improving the relevance between the query and experts. Expertise was often considered as the main criterion in these methods while diversity and authority were often ignored. Moreover, some of the previous work needed a labeling of the research interests and expertise levels of candidate reviewers as prior knowledge for better assignment. In comparison, our work incorporates three criteria at the same time, while no label information is needed. By collecting human judgments on relevance between candidate reviewers and the query paper, and evaluating our results on this ground truth data, we show the effectiveness of our model.

2.2 Graph Propagation

In network science, many algorithms have been proposed to determine the importance of the nodes in a network. Such well-known query-dependent ranking algorithms include HITS [11], Topic-Sensitive PageRank [12], and personalized PageRank [13].

Most of these models were based on random walks on the network structure. Random walk algorithms, which follow the trajectory of a random walker that takes successive random steps, have received a lot of attention. In this paper, we use a modified random walk model, RWR [1, 10], into which we can easily integrate both the expertise matching score between a candidate reviewer and the query paper, and the authority of the candidate reviewer together to make better recommendation. Different from the RWR model, we formulate our problem as an optimization framework that integrates RWR and a sparsity constraint together to obtain a stable probability for each node. That represents a balance of expertise, authority, and diversity.

3. GRAPH CONSTRUCTION

In this section, we introduce the procedure to construct a graph among the query paper and the reviewers, and explain how to measure the relations among the graph nodes.

For each query paper and all candidate reviewers, a graph can be constructed as follows: Let $\mathcal{R} = \{r_1, ..., r_N\}$ denote the candidate reviewers, and let $\mathcal{Q} = \{q_1, ..., q_m\}$ denote the query paper to be assigned reviewers. Here we consider the reviewers and the query paper as nodes in the graph, and then establish the edges and assign the associated weights.

3.1 Reviewer-Reviewer Connection

For reviewer r_i and reviewer r_j, we first search their previous publication and co-author lists using Microsoft Academic Search system[15]. An edge is established if and only if the two reviewers have co-authored at least one paper. The edge weight is set as the number of papers they have co-authored. The intuition behind this is that if a reviewer is well connected, e.g., has many co-authorship connections with others, he or she would be considered as having higher authority. As a result, during the process of connection construction between reviewer and reviewer, we incorporate the first type of criteria, the authority.

3.2 Reviewer-Query Connection

In order to establish edges between reviewers and the query paper, for each candidate reviewer we first crawl all of her previously published papers using Microsoft Academic Search. Suppose the list of published papers associated with reviewer r_i is denoted by $p_i = \{p_i^1, ..., p_i^j, ..., p_i^{m_{p_i}}\}$, where m_{p_i} is the number of publications and p_i^j is the jth paper of reviewer r_i. Then the edge weight between query paper q and reviewer r_i can be estimated by the similarity between q and the set of papers p_i published by reviewer r_i.

Since topic models such as LDA [16, 17] have been successfully applied in document analysis, we directly utilize LDA for the paper-to-paper similarity measurement. First, all of the published papers from all candidate reviewers and the query papers are collected and used as the input corpus of the LDA model. Each paper is considered as one document in the corpus. Suppose the total number of topics is T. After LDA topic model analysis, the published paper p_i^j associated with reviewer r_i will have a topic distribution denoted by $\theta_{p_i^j} \in \mathbb{R}^T$, and the query paper q will have a topic distribution denoted by $\theta_q \in \mathbb{R}^T$ [4]. Then cosine similarity is calculated between paper p_i^j and paper q in terms of the topic distribution representation. Finally, the edge weight between reviewer r_i and the query paper node q is estimated by max pooling [14] among all the similarities between reviewer r_i's published papers p_i and the query paper q.

Now we explain the rationale behind this construction. The basic assumption is that if one reviewer has published a paper in the same topic as the query paper, she should be considered as an expert candidate to review that paper. By involving the LDA model, we construct the connection between reviewers and query paper in terms of topic distribution representation. Accordingly the expertise criterion is taken into account during this graph construction process.

4. PROPAGATION OVER THE GRAPH

In this section, we present our proposed propagation method over the constructed graph. We first introduce the formulation based on RWR with a certain sparsity constraint, and then introduce an efficient gradient descent-based method to solve the objective.

4.1 Problem Formulation

4.1.1 Notation and Definition

Let $G = (V, E)$ denote a graph. The nodes set is $V = \{r_1, ..., r_N, q\}$ where r_i denotes a reviewer and q denotes the query paper. The edge set is $E = \{e_{ij} | 0 < i, j \leq N\}$ where e_{ij} denotes the edge between node v_i and v_j. Given the initial query paper node q, our goal is to propagate the initial query information through the entire graph and predict the query-reviewer relevance score for each node. Let $\mathbf{f} = [f_1, ..., f_{N+1}]^T$ denote the predicted score vector for all the nodes. Let $\mathbf{y} = [y_1, ..., y_{N+1}]^T$ denote the initial query vector where $y_i = 1$ if $v_i = q$ and $y_i = 0$ otherwise.

4.1.2 Modeling Expertise and Authority

In this work, we directly apply a Markov random walk process [21] on the graph. We use Q'_{ij} to denote the transition probability from node v_i to node v_j, which can be calculated in the following way:

$$Q'_{ij} = \begin{cases} \frac{exp(e_{ij})}{\sum_j exp(e_{ij})}, & \text{if } V_i = q \\ & V_j \neq q \\ \frac{exp(e_{ij})}{\sum_j exp(e_{ij})}, & \text{if } V_i \neq q \\ & V_j \neq q \\ 0, & otherwise \end{cases} \quad (1)$$

where the edge weight of e_{ij} is as defined in Section 3.

For a query, we would like to assign higher weight to nodes that are both well connected and relevant to the query. Thus we use an RWR process. Specifically, starting at a reviewer node, the random walker has two choices at each step: either moving to a neighbor, or jumping back to the origin node (query node) with probability η. Then the transition matrix $\mathbf{Q}$ is further revised as:

$$Q_{ij} = \begin{cases} (1-\eta)Q'_{ij}, & \text{if } V_i \neq q \text{ and } V_j \neq q \\ \eta, & \text{if } Vi \neq q \\ & V_j = q \end{cases} \quad (2)$$

According to the Markov random walk process, in order to obtain the stationary distribution vector $\mathbf{f}$ we need to solve the following eigenvector equation:

$$\mathbf{f} = \mathbf{Q}^T \mathbf{f} \quad (3)$$

We solve this by minimizing $||\mathbf{f} - \mathbf{Q}^T\mathbf{f}||_2^2$, such that two nodes connected by strong edges will have similar scores after propagation. Note that expertise and authority have already been incorporated during the construction process of the transition matrix $\mathbf{Q}$. Using the random walk process, the initial scores are propagated to the entire graph. After obtaining the stationary distribution of the random walk process, the probability score of each reviewer node provides us with a reviewer ranking that considers both expertise and authority.

4.1.3 Modeling Diversity

To achieve the diversity criterion, we first cluster candidate reviewers into groups according to their publication topic

distributions, to make sure that each cluster of reviewers has similar research interests. Then an ℓ_1-norm is applied within each group so that only a small set of individual nodes will get non-zero utility scores and thus be selected from each cluster by minimizing the ℓ_1-norm. The accumulation of the ℓ_1-norm across all the groups can be defined as follows:

$$\sum_{g=1}^{G} ||\mathbf{f}_g||_1 \tag{4}$$

where G denotes the number of groups, and $\mathbf{f}_g$ denotes the predicted score vector for all the nodes in group g. This sparsity constraint within each group is intuitive since we are trying to balance the group utility scores by not selecting all reviewers from the same research area. It is easy to see that the summation of $\mathbf{f}_g$ across all groups is equivalent to applying the ℓ_1-norm on $\mathbf{f}$.

4.1.4 RWR with Sparsity over the Graph

Finally, another term $||\mathbf{f}-\mathbf{y}||_2^2$ is added into the fomula, where the minimization on it will enforce the RWR process not go too far from the initial query.

Now we introduce our formulation which considers expertise, authority, and diversity at the same time as follows:

$$F(\mathbf{f}) \;=\; \frac{1}{2}||\mathbf{f} - \mathbf{Q}^T\mathbf{f}||_2^2 + \frac{\lambda}{2}||\mathbf{f} - \mathbf{y}||_2^2 + \gamma||\mathbf{f}||_1 \tag{5}$$

where $\lambda, \gamma > 0$ are two trade-off parameters, which can be tuned trough cross-validation.

By minimizing the objective function, we can obtain the ranking scores for all candidate reviewers as:

$$\begin{aligned} \min_{\mathbf{f}} \quad & F(\mathbf{f}), \\ \text{s.t.} \quad & \mathbf{f} \geq 0. \end{aligned} \tag{6}$$

We will show how to solve this based on a gradient descent method in the next section.

4.2 Optimization Procedure

The gradient of $\mathbf{f}$ in Equation (6) cannot be calculated due to the non-smoothness of the ℓ_1-norm regularizer. In this subsection, we show that by using the dual norm, the ℓ_l-norm term can be approached by a smoothing approximation. When the gradient of $\mathbf{f}$ is tractable, we will employ a gradient descent-based method for the optimization.

4.2.1 Smoothing Approximation

Note that the dual of the ℓ_1-norm is the ℓ_∞-norm. Similar to [21], based on Nesterov's smoothing approximation method [22], the $\|\mathbf{f}\|_1$ can be approximated by a smooth function as follows:

$$l_\mu(\mathbf{f}) = \max_{\|\mathbf{u}\|_\infty \leq 1} < \mathbf{u}, \mathbf{f} > -\frac{\mu}{2}\|\mathbf{u}\|_2^2 \tag{7}$$

where $<\cdot,\cdot>$ denotes the inner product operator and the optimal auxiliary variable $\mathbf{u}(\mathbf{f})$ can be defined as:

$$\mathbf{u}(\mathbf{f}) = S_\infty(\frac{\mathbf{f}}{\mu}) \tag{8}$$

where S_∞ is the projection operator which projects a value to the ℓ_∞-ball:

$$S_\infty(x) = \begin{cases} x, -1 \leq x \leq 1, \\ 1, x > 1, \\ -1, x < -1. \end{cases} \tag{9}$$

From the above approximation, the original formulation can be rewritten as the following smoothed objective function:

$$F_\mu(\mathbf{f}) \;=\; \frac{1}{2}||\mathbf{f} - \mathbf{Q}^T\mathbf{f}||_2^2 + \frac{\lambda}{2}||\mathbf{f} - \mathbf{y}||_2^2 + \gamma l_\mu(\mathbf{f}) \tag{10}$$

where the sparsity term in Equation (6) can be replaced by $l_0(\mathbf{f})$ with $\mu = 0$.

4.2.2 Optimization with Gradient Descent

The smooth objective function $F_\mu(\mathbf{f})$ is differentiable w.r.t. f_i as follows:

$$\begin{aligned} \partial_{f_i} F_\mu = (f_i - \sum_k f_k Q_{ki}) + \sum_{j \in \mathcal{N}_i} (f_j - \sum_k f_k Q_{kj})(-Q_{ij}) \\ + \lambda(f_i - y_i) + \gamma \nabla l_\mu(f_i) \end{aligned} \tag{11}$$

where $\mathcal{N}_i$ is the set of neighbors of node v_i. We summarize our optimization procedure in Algorithm 1.

Algorithm 1 Solving the Problem in Equation (6) by Gradient Descent Optimization

1: **Input:** $y \in \{0,1\}^{N+1}, \mathbf{f}^0 \in \mathbb{R}^n, \lambda, \gamma$
2: **Initialize:** Set $t = 0$, initialize $\mathbf{f}^t = 1$.
3: **repeat**
4: Employ Polack-Ribière conjugate gradient algorithm to estimate $\mathbf{f}^{t+1}$ based on $\partial_{f_i} F_\mu$.
5: Force the negative entries in $\mathbf{f}^{t+1}$ to 0.
6: $t = t + 1$.
7: **until** Convergence
8: **Output:** The optimized $\mathbf{f}^*$

In our experiments, we use the SLEP toolbox [23] to solve the objective function.

5. EXPERIMENTS

We have demonstrated the ability of our algorithms to incorporate expertise, authority, and diversity of candidate reviewers. In this section, we describe our experiments for evaluating the performance of the proposed method against the state-of-the-art methods. We first begin with a brief description of the two reviewer recommendation benchmark datasets and the evaluation metrics, and then introduce the baseline methods, followed by discussion of our experimental results.

5.1 Datasets

5.1.1 Multi-Aspect Review Assignment Dataset

The multi-aspect review assignment evaluation dataset is a benchmark dataset from UIUC [2]. It contains 73 papers accepted by SIGIR 2007, and 189 prospective reviewers who had published more than three papers from 1971 to 2006 in the main information retrieval conferences such as SIGIR, CIKM, and WWW. There is no label information between the 73 papers and the 189 reviewers. The dataset provides an extra expertise profile for each reviewer and each paper

to generate the pseudo-label between papers and reviewers. Specifically, 25 major topics based on the topic areas in the CFPs of ACM SIGIR in recent years were pre-defined by an information retrieval expert. For each paper in the set of 73 test papers, the expert provided a 25-dimensional label on that paper based on the defined topics. This could be considered as the expertise representation of that test paper. For the 189 reviewers, all of their publications were crawled, and through the same labeling procedure, each paper published by the reviewers also had a 25-dimensional expertise representation. By average pooling, it is then easy to achieve a similar expertise representation for each reviewer. Then the expertise matching score between each test paper and each reviewer could be measured by the distance of their pseudo-labels (e.g., cosine similarity based on the 25-dimensional expertise representation). The details are described in 5.2.2.

Besides the profile information available in the dataset, we have further constructed the co-authorship graph of the 189 prospective reviewers. Since this dataset was published in CIKM 2008, we call it the CIKM dataset for short.

5.1.2 NIPS Dataset

The second dataset was collected by Mimno and McCallum [8], who approximated the task of assigning reviewers to submitted papers by gathering expertise relevance judgments from humans experts. The dataset contains 148 papers accepted by NIPS 2006, and 364 reviewers. Several prominent researchers from the NIPS community were asked to provide a ground truth relevance judgment of a query paper and a proposed reviewer. The ground truth consists of 650 reviewer-paper relevance judgments from nine annotators using a four-level relevance scheme as follows: Very Relevant (score = 3), Relevant (score = 2), Slightly Relevant (score = 1) and Irrelevant (score = 0).

Since they labeled the ground truth according to the top 10 retrieved reviewers for their baselines, it is difficult to compare the performance of our proposed method with this based on the partial relevance judgments. We have further collected 766 more reviewer-paper labels from researchers in Machine Learning, following the same four-level relevance scheme. Moreover, we have crawled the publication lists of all 85537 co-authors of the 364 prospective reviewers and constructed the co-authorship graph.

5.2 Evaluation Metrics

In order to test the paper reviewer assignment performance, we define the following metrics to quantitatively evaluate the results.

5.2.1 Precision at Position k

Intuitively, for a given paper, it is desirable to retrieve the n reviewers with the highest relevance judgment scores. For the NIPS dataset with the ground truth, we apply precision at position k to measure the relevance of the top n results retrieved by a given query. Specifically, we measure the mean precision across all queries in terms of P@1, P@2, ..., P@10.

5.2.2 Expertise Matching Score

Unlike in the NIPS dataset, there is no ground truth in the CIKM dataset. Instead of using P@k to measure the assignment quality, we use the provided expertise profiles and apply the expertise matching score at position k as follows:

$$Expertise@k = \frac{\sum_{q=1}^{Q} \frac{\sum_{n=1}^{k} cosine(t_n, t_q)}{k}}{Q} \tag{12}$$

Here, $cosine(t_n, t_q)$ measures the expertise similarity between the n-th ranked reviewer R_n and the query paper q, while t_n and t_q represent the topic distribution of R_n and q, and Q represents the total number of query papers.

5.2.3 Authority

In addition to maximizing the paper reviewer relevance score, we also want to maximize the top n reviewers' authority. We use the h-index to measure the authority of each prospective reviewer.

$$Authority@k = \frac{\sum_{q=1}^{Q} \frac{\sum_{n=1}^{k} h\text{-}index(n,q)}{k}}{Q} \tag{13}$$

Here $h\text{-}index(n, q)$, the h-index of the n-th ranked reviewer of query paper q, represents the assignment quality in terms of authority, while Q represents the total number of query papers.

5.2.4 Diversity

We adopt two natural definitions of diversity. First, we use the Kullback Leibler (KL) divergence $KL(p||q)$, which represents the difference between two probability distributions p and q [24], to measure the dissimilarity between each pair of candidate reviewers based on their publication topic distributions, as mentioned in Section 3. The KL divergence is given by:

$$KL(p||q) = \sum_i p(i) log \frac{p(i)}{q(i)} \tag{14}$$

Similarly, we have the topic divergence measurement:

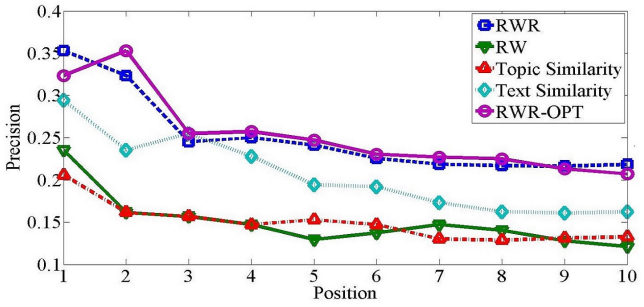
$$\begin{aligned} d_{topic} &= \frac{\sum_{q=1}^{Q} \sum_{i,j \in N_q} KL(i|j)}{Q} \\ &= \frac{\sum_{q=1}^{Q} \sum_{i,j \in N_q} \sum_t i(t) log \frac{i(t)}{j(t)}}{Q} \end{aligned} \tag{15}$$

Here N_q represents the retrieved reviewers of query q, $i(t)$ and $j(t)$ are the topic distribution of reviewers i and j over topic t, and Q is the total number of query papers.

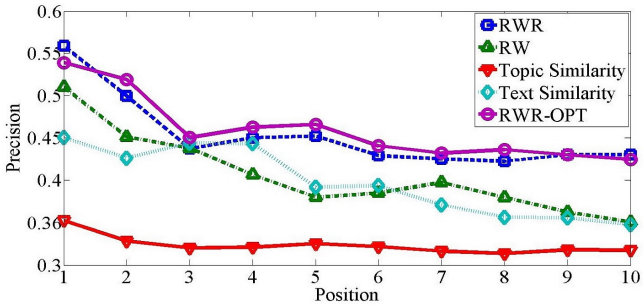
We also evaluate the diversity among retrieved reviewers by leveraging the notion of density from network science. The density of a graph is defined as the number of edges existing in the graph, divided by the maximal possible number of edges in the graph, as follows [18]:

$$d_{graph} = \frac{\sum_{u \in V} \sum_{v \in V, u \neq v} I[w(u,v) > 0]}{|V| \times (|V| - 1)} \tag{16}$$

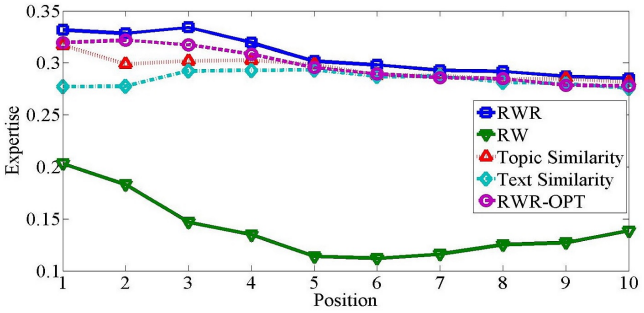
where $|V|$ is the number of nodes in graph G, $w(u, v)$ is the weight between node u and node v, and I is an indicator function. Given the top-n ranked reviewers of query paper q, we can construct a graph G_n with each node denoting one reviewer in N_q and each edge weight $w(u, v)$ defined as in Section 3.1. Then we use d_{graph} as an inverse measurement of diversity among the top-n reviewers.



(a) NIPS Dataset - Binary Judgment



(b) NIPS Dataset - Soft Judgement



(c) CIKM Dataset

Figure 2: The relevance performance for the NIPS dataset and the expertise matching performance for the CIKM dataset.

5.3 Experimental Results

In this section, we compare our proposed RWR method and RWR with sparsity constraint (RWR-OPT) with several baselines as follows: (1) Text Similarity: we first crawl all the published papers for a reviewer, and then use bag-of-words cosine similarity between the query paper and the reviewer's published papers to measure the relevance score between the query paper and the reviewer. (2) Topic Similarity: we first crawl all the published papers of a reviewer, and then use topic cosine similarity between the topic distribution of the query paper and the topic distribution of the reviewer's publication to measure the relevance score between the query paper and the reviewer. Both (1) and (2) are estimated by max pooling as introduced in Section 3.2. (3) Random Walk (RW): We only apply a random walk process on the graph instead of RWR. (4) State-of-art APT model as introduced in [8]. Since only p@5, p@10, ... , p@45 are provided in [8], we can only compare our results with theirs on p@5 and p@10.

5.3.1 *Precision Evaluation Results on NIPS Dataset*

Our first experiment shows the assignment performance of different models on the NIPS dataset. We evaluate each algorithm under two relevance settings. The first one is a binary relevance judgment, similar to [8], that only uses Very Relevant (score = 3) as relevant (label = 1); otherwise, label = 0. For the second setting, instead of binary judgment, we use a soft judgment as follows: Very Relevant (score = 3) as label = 1, Relevant (score = 2) as label = 0.67, Slightly Relevant (score = 1) as label = 0.33, and Irrelevant (score = 0) as label = 0.

The results for precision at position k for both relevance settings are shown in Tables 1 and 2, showing the performance (mean precision at position k) for all the methods. In Tables 1 and 2, the best result in each column is highlighted in bold. When a relevance value reaches the 5% level of test significance, it is denoted by one star, and when it reaches the 1% level, it is denoted by two stars. We see that our RWR method consistently achieves the best performance, with high significance level at several positions. Note that since there is less data at top positions, it is more difficult to get high-confidence numbers in these cases.

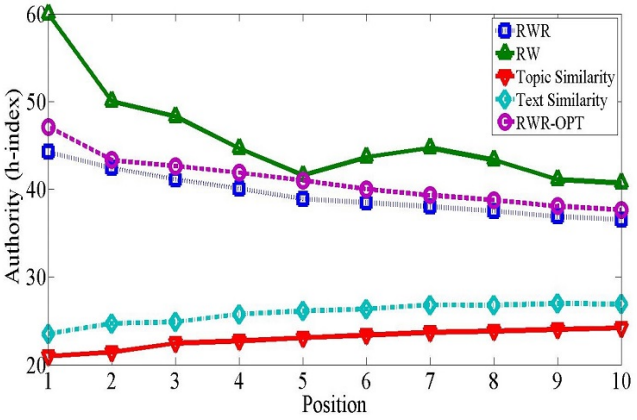
We further compare our proposed RWR-OPT method with the others and plot the performance comparisons in Figure 2 (a) and (b). The proposed RWR consistently beats all the other baselines (RW, text-similarity, and topic-similarity), which demonstrates its effectiveness in terms of relevance. In fact, RWR-OPT performs slightly better than the basic RWR algorithm in several cases. Thus, even with the added sparsity constraint, there is no reduction in relevance.

5.3.2 *Expertise Evaluation on the CIKM Dataset*

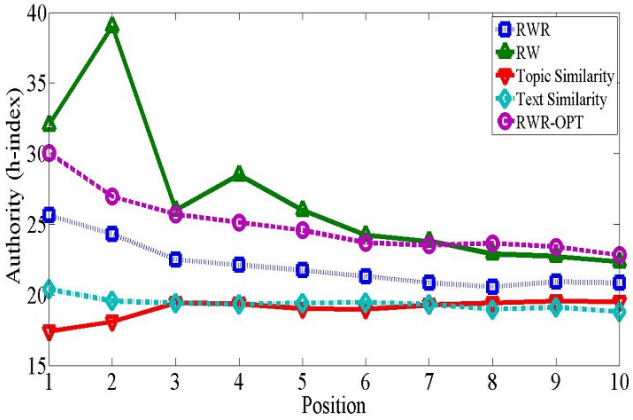
Since there is no human evaluation data in the CIKM dataset, we use the expertise matching scores described in Section 5.2.2 to evaluate the expertise performance of each model, shown in Figure 2 (c). From the experimental results, we see that the RWR and RWR-OPT models achieve the best performance.

Table 1: The mean precision performance for binary judgment on the NIPS dataset. The best result in each column is highlighted in bold. Star-annotated values indicate a significance level of 0.95, and two stars indicate a significance level of 0.99.

Method	P@1	P@2	P@3	P@4	P@5
RWR	**0.353**	**0.324**	0.245	**0.25**	**0.241***
RW	0.235	0.162	0.157	0.147	0.129
Topic-Sim	0.206	0.162	0.157	0.147	0.153
Text-Sim	0.294	0.235	**0.255**	0.228	0.194
APT200	-	-	-	-	0.2059
Method	P@6	P@7	P@8	P@9	P@10
RWR	**0.225**	**0.218**	**0.217***	**0.216****	**0.218****
RW	0.137	0.147	0.140	0.128	0.121
Topic-Sim	0.147	0.130	0.129	0.131	0.133
Text-Sim	0.192	0.173	0.162	0.161	0.162
APT200	-	-	-	-	0.1412

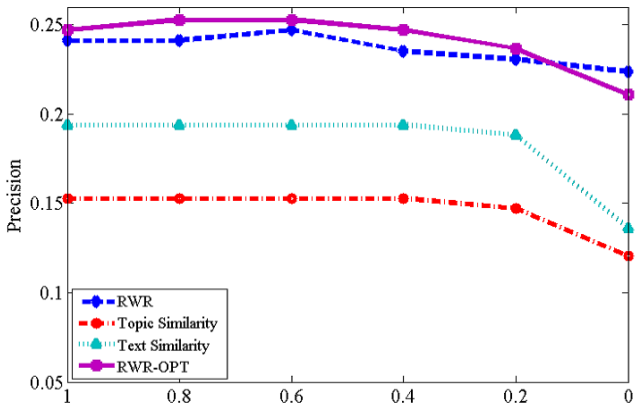


(a) NIPS Dataset

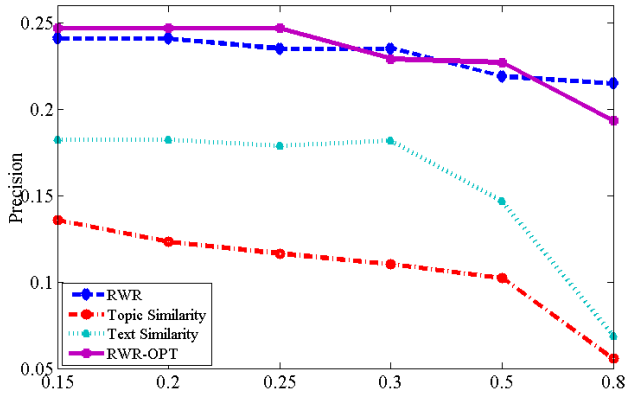


(b) CIKM Dataset

Figure 3: Average authority of top k reviewers retrieved, for NIPS and CIKM datasets.



(a) Graph Density d_{graph}



(b) Topic Divergence d_{topic}

Figure 4: The trade-off between relevance and diversity of a group of k (k=5) reviewers for NIPS dataset.

Table 2: The mean precision performance for soft judgment on the NIPS dataset. The best result in each column is highlighted in bold.

Method	P@1	P@2	P@3	P@4	P@5
RWR	**0.559**	**0.500**	0.438	**0.450**	**0.453**
RW	0.510	0.451	0.438	0.407	0.380
Topic-Sim	0.352	0.328	0.320	0.321	0.325
Text-Sim	0.451	0.426	**0.444**	0.443	0.392
Method	P@6	P@7	P@8	P@9	P@10
RWR	**0.429**	**0.425**	**0.423**	**0.430***	**0.430***
RW	0.385	0.398	0.380	0.363	0.351
Topic-Sim	0.322	0.316	0.314	0.318	0.318
Text-Sim	0.394	0.371	0.357	0.356	0.348

5.3.3 *Authority Evaluation*

Besides expertise, authority is another intermediate goal for us to achieve better reviewers as evaluated by humans. In this experiment, we evaluate the mean authority performance at position k. The authority performance of each model on the two benchmark datasets is shown in Figure 3. From the results, we can conclude that the RW method consistently beats all the other methods, which is to be expected since random walk processes are usually used for graph-based prestige measurement. We also find that the proposed RWR-OPT and RWR models perform better than the other expertise matching baselines (i.e., text similarity and topic similarity).

5.3.4 *Diversity Evaluation*

In this experiment, we measure how the average relevance for a group of k reviewers changes as the required diversify score for the same group increases. To explore the trade-off between relevance and diversity, we first define a set of diversity threshold scores T_i, and utilize a post-processing process to choose a group of k reviewers that has the highest accumulated ranking score while having at least diversity score T_i. Then the average relevance score for all the reviewers in the group is considered as the relevance score for the group. As T_i increases, more diversity is required.

We use two definitions of diversity introduced in Section 5.2.4 to evaluate the trade-off between relevance and diversity, as shown in Figure 4. We see that from both (a) and (b) that each model has a relevance drop as the diversity increases. However, both the text-similarity and topic-similarity see a significant decrease when the diversity threshold is approaching the maximum diversity possible, while both RWR and RWR-OPT drop only slightly. Also we find that with the same precision score, RWR-OPT achieves higher diversity

than RWR, which proves the effectiveness of the sparsity constraint in our optimization formula.

6. CONCLUSION

As an expert retrieval problem, paper reviewer assignment is a labor-intensive task. To reduce the time required to manually assign submitted papers to suitable reviewers, many automatic review assignment systems have been introduced. The major disadvantage for existing work is that they are trying to conduct the matching according to expertise while omitting the other criteria. In this paper, we study how to rank candidate reviewers while balancing three objectives: authority, expertise and diversity. We propose a graph constructed on candidate reviewers and the query paper, and then an optimization framework with sparsity principle is introduced. We tested all the methods on two benchmark datasets. Experiment results show that the RWR outperforms text similarity and topic similarity baselines in both expertise and authority measurements, and the selected reviewers obtain higher diversity scores when we enforce group sparsity on the grouped reviewers.

7. REFERENCES

[1] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In IEEE International Conference on Data Mining (ICDM) 2006, pages 613-622.

[2] M. Karimzadehgan and C. Zhai. Constrained multi-aspect expertise matching for committee review assignment. In ACM International Conference on Information and Knowledge Management (CIKM) 2009, pages 1697-1700.

[3] F. Wang, B. Chen, and Z. Miao. A survey on reviewer assignment problem. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE) 2008, pages 718-727.

[4] W. Tang, J. Tang, and C. Tan. Expertise Matching via Constraint-Based Optimization. In IEEE/WIC/ACM International Conference on Web Intelligence (WI) and Intelligent Agent Technology (IAT) 2010, pages 34-41.

[5] L. Charlin, R. S. Zemel. The Toronto paper matching system: an automated paper-reviewer assignment system. In International Conference on Machine Learning (ICML) 2013, Workshop on Peer Reviewing and Publishing Models.

[6] D. Conry, Y. Koren, N. Ramakrishnan. Recommender Systems for the Conference Paper Assignment Problem. In ACM Recommender System Conference (RecSys) 2009, pages 357-360.

[7] S. Hettich, M. J. Pazzani. Mining for Proposal Reviewers: Lessons Learned at the National Science Foundation. In ACM SIGMOD Conference on Knowledge Discovery and Data Mining (KDD) 2006, pages 862-871.

[8] D. M. Mimno, A. McCallum. Expertise modeling for matching papers with reviewers. In ACM SIGMOD Conference on Knowledge Discovery and Data Mining (KDD) 2007, pages 500-509.

[9] M. A. Rodriguez, J. Bollen. An algorithm to determine peer-reviewers. In ACM International Conference on Information and Knowledge Management (CIKM) 2008, pages 319-328.

[10] Y. Fujiwara, M. Nakatsuji, M. Onizuka, and M. Kitsuregawa. Fast and Exact Top-k Search for Random Walk with Restart. In The Proceedings of the VLDB Endowment (PVLDB), 5(5):442-453, 2012.

[11] J. M. Kleinberg, Authoritative Sources In A Hyper-linked Environment. Journal of the ACM, 46(5):604-632, 1999.

[12] T. Haveliwala, Topic-sensitive PageRank. In International World Wide Web Conference (WWW) 2002, pages 517-526.

[13] P. Berkhin. A Survey on PageRank Computing. Internet Math. 2(1), 73-120, 2005.

[14] J. Yang, K. Yu, Y. Gong, T. Huang. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[15] Microsoft Academic Search. `http://academic.research.microsoft.com/`

[16] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research 2003, 3:993-1022.

[17] L. Li and N. Memon. Mining Groups of Common Interest: Discovering Topical Communities with Network Flows. In International Conference on Machine Learning and Data Mining (MLDM) 2013, pages 405-420.

[18] Q. Mei, J. Guo, and D. Radev. Divrank: the interplay of prestige and diversity in information networks. ACM SIGMOD Conference on Knowledge Discovery and Data Mining (KDD) 2010, pages 1009-1018.

[19] B. Wilson. The Machine Learning Dictionary. `http://www.cse.unsw.edu.au/~billw/mldict.html`

[20] T. Griffiths, M. Steyvers. Finding Scientific Topics. In Proceedings of the National Academy of Sciences, 2004, 101 (suppl. 1), 5228-5235.

[21] D. Liu, G. Ye, C. Chen, S. Yan, S. Chang. Hybrid Social Media Network. In ACM Multimedia, 2012.

[22] Y. Nesterov. Smooth minimization of non-smooth functions. Mathematical Programming, 2005.

[23] J. Liu, S. Ji, J. Ye. SLEP: Sparse learning with efficient projections. Software Tool. Arizona State University, 2009. `www.public.asu.edu/~jye02/Software/SLEP`

[24] L. AlSumait, D. Barbará, C. Domeniconi. On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In IEEE International Conference on Data Mining (ICDM), 2008.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[26] M.R. Morris, J. Teevan, and K. Panovich. A comparison of information seeking using search engines and social networks. In International Conference on Weblogs and Social Media (ICWSM) 2010, pages 291-294.

[27] D. Horowitz, S. D. Kamvar. Anatomy of a large-scale social search engine. In International World Wide Web Conference (WWW) 2010.