sig web

Towards Higher Relevance and Serendipity in Scholarly Paper Recommendation

Kazunari Sugiyama and Min-Yen Kan
School of Computing,
National University of Singapore

Finding relevant scholarly papers is an important task for researchers. Such a literature search involves identifying drawbacks in existing works and proposing new approaches that address them. However, the growing number of scientific published papers results in information overload even for simple searches, such that researchers have difficulty in finding papers relevant to their interests. Recommendation systems can help address this problem to find relevant papers efficiently. In this article, we summarize our work on scholarly paper recommendation from both a relevance and serendipitous perspectives. Experimental results on a publicly-available scholarly paper recommendation dataset show that our proposed approaches provide promising recommendations for researchers, outperforming the state-of-the-art with statistical significance.

1. INTRODUCTION

According to *Nature*,[1] an unprecedented number of scholarly papers were published in 2012, which continues to this day. While these trends certainly mark the advent of the knowledge era, the sheer volume of scholarly knowledge creates a problem of overabundance. Commonly known as "information overload", it manifests itself when researchers find an overwhelming number of matches to their queries, but where the majority of the results are largely irrelevant to their latent information needs.

Work in recommendation systems is one promising approach to address this information overload. In digital library studies, this approach has been employed to obtain and refine search results to provide scholarly papers relevant to each user's interests [Torres et al. 2004; Gori and Pucci 2006; Yang et al. 2009; Nascimento et al. 2011] as well as citations relevant to each paper [McNee et al. 2002; Strohman et al. 2007; He et al. 2010; He et al. 2011; Lu et al. 2011; Huang et al. 2012; Caragea et al. 2013; Livne et al. 2014; Tang et al. 2014]. However, these approaches do not fully leverage the user's context, largely relying on the idea of session-as-context. To address this problem, we model a searcher's context in the form of a profile by capturing previous research interests embodied in their past publications to provide more relevant recommendations.

[1] Originally pinned from Nature (`http://www.nature.com/news/366-days-2012-in-review-1.12042`) with data from Thomson Reuters/Essential Science Indicators.

Aside from the quality of relevance, serendipitous recommendations are also important, and can be more highly valued than strongly similar relevant recommendations. For example, junior researchers need to broaden their range of research interests to acquire knowledge. Senior researchers have expertise in their own fields, but may seek to apply their knowledge towards other areas or import ideas familiar to other areas to their own. Such cross-pollination work is a hallmark of productive interdisciplinary exchange. Thus, serendipitous recommendation allows both type of researchers to derive benefits. Our work addresses this task by a process analogous to asking colleagues for advice or recommendation on what they find interesting. The preferences gathered from other users (dissimilar users and co-auhors) are used in the construction of the target researcher's user profile.

Furthermore, to achieve better recommendation accuracy, it is important to generate feature vector of candidate papers to recommend as well as user profile construction described above. Focusing on this point, we further mine additional signals from the full text of scholarly papers and their citation network – using (1) potentially cited papers and (2) their fragments. Here, *citation papers* are defined as papers that explicitly cite previous work and often contain a summary of its salient points. Such citation papers may be viewed as an endorsement of the cited paper, and they may help model the target paper more accurately. In addition, *fragments* are sections of a paper such as the abstract, introduction, conclusion, and so on.

In this article, we summarize our recent works on scholalrly paper recommendation [Sugiyama and Kan 2010; 2011; 2013; 2014], addressing higher relevancy and serendipity.

2. PROPOSED METHOD

We propose recommending papers based on an individual's (recent) research interests as modeled by a profile derived from their publication list. We hypothesize that this will result in high recommendation accuracy, as we believe that a user's research interests are reflected in their prior publications.

We first construct each researcher's profile using their list of previous publications, and then recommend papers by comparing the profiles with feature vectors constructed from candidate papers. Our approach is novel because it directly deals with each user's research interest using their publication history. A key aspect of our approach is that we include contextual evidence about each paper in the form of its *neighboring papers*: the papers that cite the target paper (we term these *citation* papers) and papers referenced by the target paper (*reference papers*). Another desirable property of our approach is that it is domain-independent – that is, our approach can be applied various kinds of documents that have links to other documents such as Web pages, patents, news paper articles as well as scholarly papers. Its simple requirement is that contextual information from such neighboring publications needs to be accessible.

Figure 1 shows an overview of our approach. (1) We first construct a user profile $\boldsymbol{P}_u$ from a researcher's list of published papers; (2) then construct feature vectors $\boldsymbol{G}^{p_j}$ $(j = 1, \ldots, N)$ for candidate papers to recommend; (3) compute cosine similarity $Sim(\boldsymbol{P}_u, \boldsymbol{G}^{p_j})$ between $\boldsymbol{P}_u$ and $\boldsymbol{G}^{p_j}$ and recommend papers with high similarity. In the following, we describe how to construct the user profile $\boldsymbol{P}_u$ and feature vectors $\boldsymbol{G}^{p_j}$ used in the first two steps.

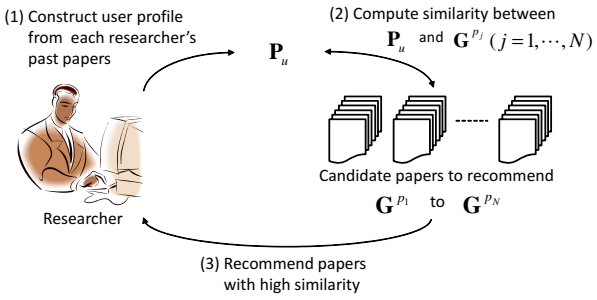


Fig. 1. System overview.

2.1 User Profile Construction for Relevant Recommendation

We model researchers as shown in Figure 2. Basically, researchers have multiple past publications, where their past publications may have attracted citations. Each paper also has reference papers.

Our representations of the user profile are based on the foundation model where a paper represented as a feature vector. For each paper p in a researcher's publication list, we transform p into a feature vector $\boldsymbol{f}^p$ as follows:

$$\boldsymbol{f}^p = (w^p_{t_1}, w^p_{t_2}, \cdots, w^p_{t_m}), \tag{1}$$

where m is the number of distinct terms in the paper, and t_k $(k = 1, 2, \cdots, m)$ denotes each term. Using term frequency (TF),[2] we also define each element $w^p_{t_k}$ of $\boldsymbol{f}^p$ in Equation (1) as follows:

$$w^p_{t_k} = \frac{tf(t_k, p)}{\sum_{s=1}^{m} tf(t_s, p)},$$

where $tf(t_k, p)$ is the frequency of term t_k in a paper p. In a key step, we modify the assigned weights for terms to account for the influence of the papers in the citation network neighborhood. Papers that cite the target paper (termed *citation papers*) as well as those that the target paper references (termed *reference papers*) influence the original $\boldsymbol{f}^p$ weighting.

For each researcher, characterized as having n past papers p_i $(i = 1, \cdots, n)$, the individual feature vectors for each paper have an enlarged context accounting for possible citation and reference papers (corresponding to the additional second and third terms below, respectively):

$$\boldsymbol{F}^{p_i} = \boldsymbol{f}^{p_i} + \sum_{x=1}^{k} W^{p_{i_{cit_x}} \to p_i} \boldsymbol{f}^{p_{i_{cit_x}}} + \sum_{y=1}^{l} W^{p_i \to p_{i_{ref_y}}} \boldsymbol{f}^{p_{i_{ref_y}}} \tag{2}$$

where p_i $(i = 1, 2, \cdots, n)$, $p_{i_{cit_x}}$ $(x = 1, 2, \cdots, k)$, and $p_{i_{ref_y}}$ $(y = 1, 2, \cdots, l)$ denote a user's published papers, citation papers and reference papers, respectively. In addition,

[2] Note that we prefer TF over standard TF-IDF in the construction process, as the limited size of a researcher's publication list often does not allow reliable estimates of IDF.

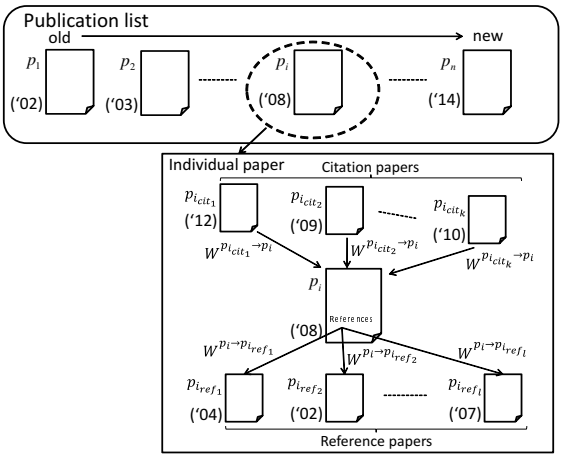


Fig. 2. Publication lists by researchers and relation between citation or reference papers and one of published papers p_i.

$W^{p_{i_{cit_x}} \to p_i}$ and $W^{p_i \to p_{i_{ref_y}}}$ denote weights defined by cosine similarity between $\boldsymbol{f}^{p_{i_{cit_x}}}$ and $\boldsymbol{f}^{p_i}$, and cosine similarity between $\boldsymbol{f}^{p_i}$ and $\boldsymbol{f}^{p_{i_{ref_y}}}$, respectively.

Secondly, as research interests of researchers change over time, the user profile construction process must model this. We capture this by using a tunable *forgetting factor* that assigns less weight to papers published further in the past. The user profile $\boldsymbol{P}_u^{rel}$ for the general case is thus defined as:

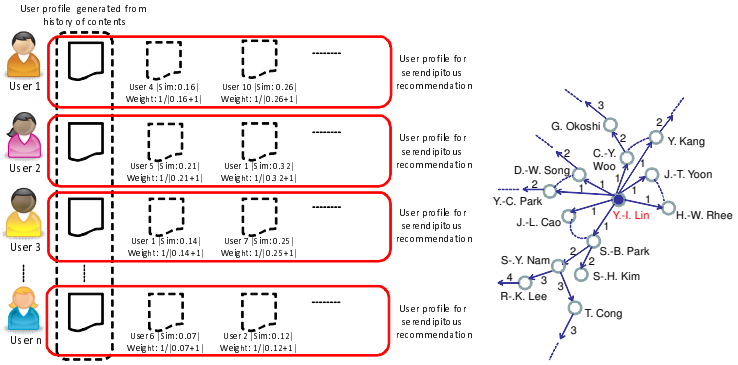
$$\begin{aligned} \boldsymbol{P}_u^{rel} &= e^{-\gamma \cdot d_{n \to 1}} \boldsymbol{F}^{p_1} + e^{-\gamma \cdot d_{n \to 2}} \boldsymbol{F}^{p_2} + \cdots + e^{-\gamma \cdot d_{n \to n-1}} \boldsymbol{F}^{p_{n-1}} + \boldsymbol{F}^{p_n} \\ &= \sum_{z=1}^{n} e^{-\gamma \cdot d_{n \to z}} \boldsymbol{F}^{p_z}, \end{aligned} \tag{3}$$

where $e^{-\gamma \cdot d_{n \to z}}$ denotes the forgetting factor between $[0, 1]$ assigned to paper $p_{n \to z}$ computed on the basis of the most recent paper p_n. Here, γ is the forgetting coefficient $(0 \leq \gamma \leq 1)$ and $d_{n \to z}$ is the difference between the published year of the most recent paper p_n and the previously published work p_z.

2.2 User Profile Construction for Serendipitous Recommendation

We also investigated how serendipitous recommendation can be influenced through the modification of the user profile construction process. We do this by incorporating others' user profile weighting of papers into a target user's own user profile. We explain our two approaches in the following subsections.

2.2.1 *User Profile Construction Utilizing Dissimilar Users (DU).* Researchers whose interests differ from a target user may be promising candidates to generate interesting and surprising recommendations. For this reason, our dissimilar users (DU) approach utilizes profiles from users that are maximally different from our target user. We use the reciprocal of similarity between the target user and a candidate user to rank candidate users with re-



(a) Dissimilar users (b) Co-authors' network

Fig. 3. User profile construction with (a) dissimilar users and (b) co-authors' network for serendipitous recommendation.

spect to their dissimilarity. Figure 3 (a) shows the user profile construction with dissimilar users for serendipitous recommendation.

Suppose, for example, that User 1 is the target user and that the similarity between User 1 and User 4 is 0.16. In this case, the weight assigned to the profile of User 4 for User 1 is computed by taking the reciprocal of (the cosine similarity + k) : 1/(0.16+k). k is used to place a bounded limit on the dissimilarity value; in our work, we set $k = 1$, such that dissimilarity values range between $[\frac{1}{2}, 1]$.

We compute weights assigned to the other users' profiles for the target user in the same manner and combine them together with the original target user profile from the baseline. Let $\boldsymbol{P}_u^{srdp}$ be the modified user profile for user u for serendipitous recommendation. This scheme is formalized as follows:

$$\boldsymbol{P}_u^{srdp} = \boldsymbol{P}_u + \sum_{v=1}^{N_{du}} \left(\frac{1}{sim(\boldsymbol{P}_u, \boldsymbol{P}_v) + k} \times \boldsymbol{P}_v\right), \tag{4}$$

where $\boldsymbol{P}_u$ and $\boldsymbol{P}_v$ are the basic user profiles of user u and users v $(1, \ldots, N_{du}, v \neq u)$, who are the dissimilar users for user u. The fractional term is the weighting factor for the dissimilar user; as described above, it is essentially the reciprocal of the cosine similarity between u and v's profile.

While technically possible to involve all users aside from the target user u in the dissimilarity computation, it would be inefficient. A limited subset of dissimilar users would make the calculation tractable. We experiment with the following three methods for selecting dissimilar users:

(1) Selection by common title words (DU-title): The words in the title are good cues to find serendipitous papers because we often find papers serendipitous from titles in the proceedings or conference program. Thus, we employ titles of the paper to select dissimilar users.

(2) Selection by common references (DU-refs): We expect that references are also

good cues to find serendipitous papers. If different users refer to the same paper, user profile with different topics can be easily constructed. We expect such user profiles contribute to serendipitous recommendation.

(3) Selection by thresholding the cosine similarity (SIM-th): In this approach, we first set the threshold of similarity between users to select dissimilar users, and then construct user profile for serendipitous recommendation. Let th be the value of threshold. We select users whose similarity between a target user is less than th. The reason why we focus on "less than" is that our aim is to construct user profile with dissimilar users to recommend serendipitous papers.

2.2.2 *User Profile Construction Utilizing Co-author Network (CAN).* Researchers are often work collaboratively to achieve their goal; teaming up with others to do research and capitalize on each other's expertise. A trusted co-author often serves as a sounding board for ideas, lends inspiration and motivation, and importantly for us, can give a different and novel perspective on a research area.

In much the same way, our second approach modifies the construction of the user profile by utilizing a target user's co-author network. We modify the co-author network using the following two constraints, which give rise to networks similar to Figure 3 (b).

(c1) We place co-authors in the network at a "collaborative distance" with respect to their minimal transitive co-authorship distance to the target user u.

(c2) We ignore authoring relationships between other co-authors; in other words, we consider only the radial network centered on the target user.

In this (CAN) scheme, we define $\boldsymbol{P}_u^{srdp}$ as the user profile of user u for serendipitous recommendation as follows:

$$\begin{aligned}\boldsymbol{P}_u^{srdp} &= \boldsymbol{P}_u + \sum_{ca_{(1)}=1}^{N_{(1)}} w_{ca_{(1)}} \boldsymbol{P}_{ca_{(1)}} + \sum_{ca_{(2)}=1}^{N_{(2)}} w_{ca_{(2)}} \boldsymbol{P}_{ca_{(2)}} + \cdots \\ &= \boldsymbol{P}_u + \sum_{pl=1} \sum_{ca_{(pl)}=1}^{N_{(pl)}} w_{ca_{(pl)}} \boldsymbol{P}_{ca_{(pl)}}, \end{aligned} \tag{5}$$

where $\boldsymbol{P}_u$ and $\boldsymbol{P}_{ca_{(pl)}}$ are basic user profiles of user u and co-author $ca_{(pl)}$ that is separated from u with a path length of pl from the user u, respectively. Here, $w_{ca_{(pl)}}$ is the multiplicative coefficient used to integrate $\boldsymbol{P}_{ca_{(pl)}}$ with $\boldsymbol{P}_u$.

In constructing the modified user profile for recommendation, we explore the four methods to set $w_{ca_{(pl)}}$ (see [Sugiyama and Kan 2011] about the detailed definition of them):

W1. **Linear Combination (LC):** This weighting scheme simply combines user profile $\boldsymbol{P}_u$ of user u and user profile $\boldsymbol{P}_{ca_{(pl)}}$ of co-author $ca_{(pl)}$.

W2. **Reciprocal of Path Length (RCP-PL):** This weighting scheme assigns larger weights to closer co-authors and smaller weights to distant co-authors from the target user, using reciprocal weighting.

W3. **Reciprocal of Similarity (RCP-SIM):** The purpose of this work is to provide serendipitous recommendation for scholarly papers. We assign larger weight to dissimilar users under this scheme.

W4. **Product of W2 and W3 (RCP-PLSIM):** This final weighting scheme multiplicatively combines path length and cosine similarity, taking the product of W2 and W3 above.

2.3 Feature Vector Construction for Candidate Papers

Unlike the TF representation of papers used in the user profile, we employ TF-IDF [Salton and McGill 1983] for the calculation of the feature vector $\boldsymbol{g}^p$ of a candidate paper p to be considered for recommendation. Identical to Equation (1), we first define the feature vector $\boldsymbol{g}^p$ of p as follows:

$$\boldsymbol{g}^p = (w^p_{t_1}, w^p_{t_2}, \ldots, w^p_{t_m}), \tag{6}$$

where m is the number of distinct terms in the paper, and t_k $(k = 1, 2, \ldots, m)$ denotes each term. Using TF-IDF, each element $w^p_{t_k}$ of $\boldsymbol{g}^p$ in Equation (6) is defined as follows:

$$w^p_{t_k} = \frac{tf(t_k, p)}{\sum_{s=1}^{m} tf(t_s, p)} \cdot \log \frac{N}{df(t_k)},$$

where $tf(t_k, p)$ is the frequency of term t_k in the target candidate paper p, N is the total number of papers to recommend in the collection, and $df(t_k)$ is the number of papers in which term t_k appears. We favor TF-IDF here over pure TF – which we used for candidate papers – as the pool for candidate papers is usually much larger. In our experiments, as we describe later in Section 3.1, our candidate paper base consists of about ten thousand papers, making IDF more reliable and consistent. Critically, our dataset also contains clean citation information that allows us to construct correct citation and reference papers. Therefore, we also use this information to characterize a candidate paper better and obtain high recommendation accuracy: Let $\boldsymbol{G}^p$ be the feature vector for paper to recommend, this is denoted as follows:

$$\boldsymbol{G}^p = \boldsymbol{g}^p + \sum_{x=1}^{k} W^{p_{cit_x} \rightarrow p}\, \boldsymbol{g}^{p_{cit_x}} + \sum_{y=1}^{l} W^{p \rightarrow p_{ref_y}}\, \boldsymbol{g}^{p_{ref_y}}, \tag{7}$$

where p_{cit_x} $(x = 1, \ldots, k)$ and p_{ref_y} $(y = 1, \ldots, l)$ denote papers that cite p and papers that p refers to, respectively. As well as Equation (2), $W^{p_{cit_x} \rightarrow p}$ and $W^{p \rightarrow p_{ref_y}}$ denote weights defined by cosine similarity between $\boldsymbol{g}^{p_{cit_x}}$ and $\boldsymbol{g}^p$, and cosine similarity between $\boldsymbol{g}^p$ and $\boldsymbol{g}^{p_{ref_y}}$, respectively.

As shown in Figure 1, our approach basically employs content-based filtering (CBF), which relies on the item's content to provide its recommendations. Therefore, it is important to represent an item's contents faithfully. A key innovative step is to model a target paper of interest based on not merely its own textual content but also based on an appropriately weighted inclusion of the text from its context as defined by the neighborhood of scholarly works it referenced, as well as those works that cite it (see Figure 4 (a)). However, authors of papers also may not cite certain relevant papers in their publications, either purposefully (*e.g.*, to save space) or not (*e.g.*, were unaware of the specific relevant work). If we enhance the citation network with such potentially citable papers (hereafter, *pc*), we

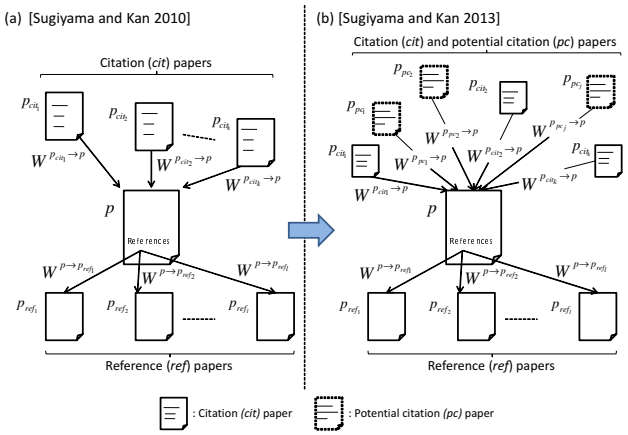


Fig. 4. Comparison of paper representations between our works.

hypothesize that we can model the target papers to recommend more accurately to achieve better recommendation performance.

We further enhance this step, to both **enlarge** what is meant by context through the discovery of potential citation papers (Figure 4(b)), as well as **refine** its use in specific, well-linked parts of the contextual documents by discovering potential citation papers with imputation-based collaborative filtering (pc-IMP) [Sugiyama and Kan 2013] and its adaptive selection of neighborhoods (pc-IMP (adp)) [Sugiyama and Kan 2014] on paper-citation matrix. In this article, we outline "pc-IMP (adp)" that employs clustering to adaptively select neighborhoods in collaborative filtering. "pc-IMP (adp)" consists of the following steps:

Step 1: Impute similarities between all papers, recording them into an intermediate imputed paper–citation matrix (Figure 5).

Step 2: For the target paper, find the n most similar clusters from the "(b) imputed matrix" in Figure 5:
- Generate clusters of papers by means of k nearest neighbor clustering [Jarvis and Patrick 1973], where the similarity between papers is measured using the Pearson correlation coefficient between the papers' citation vectors,
- Select n clusters that have the highest similarity with the target paper than the threshold (CL_{th}). These clusters form the *n-neighborhood* for the target paper. In Figure 6, C_1 and C_2 are determined to be the 2–neighborhood for p_1.

Step 3: Compute a prediction from a weighted combination of the neighbor's values (Figure 6 (b)) using centroid vectors of clusters.

We review the two latter steps in more detail. In Step 2, the similarity between target paper p_{tgt} and centroid vectors of clusters g, is computed using the Pearson correlation coefficient:

$$S_{tgt,g} = \frac{\sum_{i=1}^{N}(r_{tgt,i} - \bar{r}_{tgt}) \times (r_{g,i} - \bar{r}_g)}{\sqrt{\sum_{i=1}^{N}(r_{tgt,i} - \bar{r}_{tgt})^2 \times \sum_{i=1}^{N}(r_{g,i} - \bar{r}_g)^2}}, \quad (8)$$

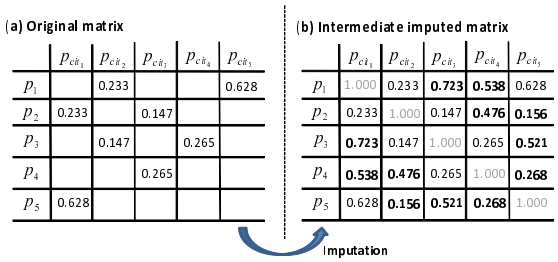


Fig. 5. Similarity imputation: (a) original matrix and (b) intermediate imputed matrix (imputed values are bolded).

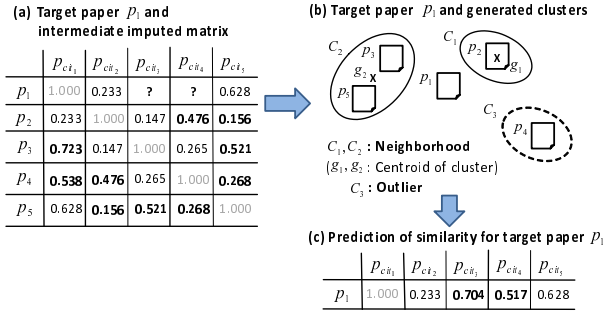


Fig. 6. Predictions computed for the target paper p_1 using centroid vectors of corresponding clusters (neighbors), C_1 and C_2.

where $r_{g,i}$ is the score given to citation paper p_{cit_i} by the centroid vectors of clusters g, and $\bar{r}_g$ is the mean score given by g. In addition, several clusters are chosen based on their similarity to the target paper, and a weighted aggregate of their scores is used to generate predictions for the target paper in Step 3. In this step, the number of selected clusters may differ per target paper, hence our use of the term "adaptive." We expect that this method forms more relevant neighborhoods for certain target papers.

In Step 3, predictions are computed as the weighted average of deviations from the neighbor's mean, shown in Equation (9):

$$p_{tgt,i} = \bar{r}_{tgt} + \frac{\sum_{g=1}^{n}(r_{g,i} - \bar{r}_g) \times S_{tgt,g}}{\sum_{g=1}^{n} S_{tgt,g}}, \tag{9}$$

where $p_{tgt,i}$ is the prediction for a target paper p_{tgt} for a citation paper p_{cit_i}. n is the number of centroid vectors of clusters in the neighborhood. According to the score of $p_{tgt,i}$, we select the top N_{pc} papers as potential citation papers.

With the discovery and weightage of our discovered potential citation papers, we can now modify the feature vector for target papers defined by Equation (7) as follows:

$$\boldsymbol{G}^p = \boldsymbol{g}^p + \sum_{x=1}^{j} W^{p_{pc_x} \to p} + \sum_{y=1}^{k} W^{p_{cit_x} \to p} \boldsymbol{g}^{p_{cit_y}} + \sum_{z=1}^{l} W^{p \to p_{ref_y}} \boldsymbol{g}^{p_{ref_z}}, \tag{10}$$

where p_{pc_x} $(x = 1, \ldots, j)$, p_{cit_y} $(y = 1, \ldots, k)$, and p_{ref_z} $(z = 1, \ldots, l)$ denote potential

citation papers, papers that cite p, and papers that p refers to, respectively. We employ cosine similarity weighting for $W^{p_{pc_x} \to p}$, $W^{p_{cit_y} \to p}$, and $W^{p \to p_{ref_z}}$ as it was found effective in our previous work [Sugiyama and Kan 2010].

In the above, we have artificially enriched the citation network to combat sparsity. We now also consider refining and improving the quality of information in the existing citation network. In this context, Abu-Jbara et al. [2013] analyzed "qualitative" aspect of citations in the ACL Anthology Network corpus [Radev et al. 2013] to identify the purpose of citing a paper and polarity of this citation.

Since citation sentences often present a clear representation of a target paper, we hypothesize that careful weighting of citation sentences improves recommendation accuracy. On the other hand, citation sentences are very small text fragments in citation papers. Larger text fragments of the (potential) citation papers may be more useful than using just single citation sentences. Thus, we also experiment with other larger fragments of the source paper: its abstract, introduction, and conclusion sections. In our experiments, we observe that the following equation, which assigns cosine similarity, featuring a tunable constant weight α $(0 \leq \alpha \leq 1)$, to feature vectors constructed from fragments in potential and explicit citation papers ($\boldsymbol{g}^{p_{pc_x}}_{(frg)}$ and $\boldsymbol{g}^{p_{cit_y}}_{(frg)}$, respectively), gives the best recommendation accuracy as shown in Table II:

$$\begin{aligned} \boldsymbol{G}^p = \quad & \alpha \quad \left(\sum_{x=1}^{j} W^{p_{pc_x} \to p}_{(frg)} \, \boldsymbol{g}^{p_{pc_x}}_{(frg)} \; + \; \sum_{y=1}^{k} W^{p_{cit_y} \to p}_{(frg)} \, \boldsymbol{g}^{p_{cit_y}}_{(frg)} \right) \\ & + \; (1-\alpha) \; \left(\boldsymbol{g}^p + \sum_{x=1}^{j} W^{p_{pc_x} \to p} \, \boldsymbol{g}^{p_{pc_x}} + \sum_{y=1}^{k} W^{p_{cit_y} \to p} \, \boldsymbol{g}^{p_{cit_y}} + \sum_{z=1}^{l} W^{p \to p_{ref_z}} \, \boldsymbol{g}^{p_{ref_z}} \right), \end{aligned} \tag{11}$$

where the first row are two added terms to Equation (10) that account for evidence from the fragments in potential and explicit citation papers, respectively. α represents the balance between the contribution from the full text and the fragments, and allows our model a bit more expressiveness in finding optimal parameters.

2.4 Recommendation of Papers

Using the user profile $\boldsymbol{P}^X_u$ $(X = rel, srdp)$ defined by Equations (3), (4) or (5) for researchers, and the feature vector for the candidate paper to recommend $\boldsymbol{G}^p$ as defined by Equation (11), our system computes the cosine similarity $sim(\boldsymbol{P}^X_u, \boldsymbol{G}^p)$ between $\boldsymbol{P}^X_u$ and $\boldsymbol{G}^p$:

$$sim(\boldsymbol{P}^X_u, \boldsymbol{G}^p) = \frac{\boldsymbol{P}^X_u \cdot \boldsymbol{G}^p}{|\boldsymbol{P}^X_u||\boldsymbol{G}^p|},$$

and ranks the set of candidate papers in order of decreasing similarity.

3. EXPERIMENTS

3.1 Experimental Data and Evaluation Measures

We use the publication lists of 50 researchers who have been engaged in various fields in computer science such as databases, embedded systems, graphics, information retrieval, networks, operating systems, programming languages, software engineering, security, user interface. The researchers also have publication lists in DBLP.[3] As DBLP lists many important venues in computer science, we assume here that a researcher's DBLP list is representative of their main interests.

We construct the user profile for each researcher using their respective publication list in DBLP. All 50 researchers' names are unambiguous with respect to the field of computer science studies.

The candidate papers to recommend is constructed from proceedings in the ACM Digital Library[4] (ACM DL). Among them, we collected 100,351 papers published in English, in conferences, symposiums, and workshops held more than three times. We also manually collected citation and reference papers for each paper. In collecting citation and reference papers, we used information on the "Cited By" tab attached in each paper in ACM DL (as of July, 2012), and those in the references section of each paper. Then, we construct feature vectors for these papers as described in Section 2.3. Stop words[5] were eliminated from each user's publication list and from the candidate papers to recommend. Stemming was performed using the Porter Stemmer[6] [Porter 1980]. We manually compiled the gold-standard results, by asking each researcher to mark papers relevant to their recent research interest. We performed 5–fold cross validation. In each fold, we divided these datasets into a training set (for parameter tuning) and a test set (for evaluation). Table I shows some statistics about our experimental data. We have made our entire dataset publicly available,[7] to encourage the community to work on this problem and to facilitate competitive benchmarking.

As in standard information retrieval (IR), top ranked documents are the most important, since users often scan just the first ranks. As such, we adopt ranked IR evaluation measures, specifically: (1) normalized discounted cumulative gain (nDCG) [Järvelin and Kekäläinen 2000], and (2) mean reciprocal rank (MRR) [Voorhees 1999].

3.2 Experimental Results and Discussion

In this section, we only show the experimental results after applying optimized parameters. Please refer to [Sugiyama and Kan 2010; 2011; 2013; 2014] about the detailed parameter tuning process and experimental results.

[3] http://www.informatik.uni-trier.de/~ley/db/
[4] http://dl.acm.org/
[5] ftp://ftp.cs.cornell.edu/pub/smart/english.stop
[6] http://www.tartarus.org/~martin/PorterStemmer/
[7] http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html

Table I. Some statistics on our scholarly paper dataset.

(a) Researchers

Number of researchers	50
Average number of DBLP papers	10.0
Average number of relevant papers in our dataset	75.4
Average number of citation papers	14.8 (max. 169)
Average number of reference papers	15.0 (max. 58)

(b) Candidate papers to recommend

Number of papers	100,351
Average number of citation papers	17.9 (max. 175)
Average number of reference papers	15.5 (max. 53)

3.2.1 *Experimental Results in Relevant Recommendation.* We compare our proposed approach with state-of-the-art scholarly paper recommendation systems [Nascimento et al. 2011; Wang and Blei 2011] and recent pseudo relevance feedback approach based on frequent term pattern mining [Algarni et al. 2010]. Table II shows recommendation accuracy obtained by applying the optimal parameters and selection of fragments to the test set.

We observe that our baseline system [Sugiyama and Kan 2010] outperforms others ([Nascimento et al. 2011], [Wang and Blei 2011], and [Algarni et al. 2010]) and that our approach outlined in this article ([*pc*-IMP (adp)] + [*frg*-TW]) gives the best recommendation accuracy.

In [Nascimento et al. 2011], user profiles are constructed from the title, and feature vectors of candidate papers are generated from the title and abstract, resulting in poor recommendation accuracy. This indicates that better representation of users and papers cannot be achieved by using short fragments such as title and abstract only. In [Wang and Blei 2011], a binary-valued user–paper matrix is applied to predict missing values to discover potential citation papers. These missing values are computed based on a probabilistic topic model generated from words in the abstract and title. We believe that these fragments are too short and uninformative, resulting in discovery of ineffective potential citation papers and irrelevant recommendation of scholarly papers. In light of these observations, we believe that our approach that uses full text and effective fragment (conclusion) in potential citation papers with appropriate tuning characterizes candidate papers better, resulting in more relevant recommendation. Algarni et al. [2010] gives the second highest recommendation accuracy among the other comparative approaches. This implies that implicit feedback based on frequent pattern mining is one of the effective methods for constructing user profiles.

In addition, [*pc*-IMP (adp)] can overcome the problem in [Sugiyama and Kan 2013], which tends to find only potential citation papers related to a single discipline when the topic of the target paper is intra-disciplinary. As shown in Figure 6, this approach selects neighborhoods as the centroid vector of clusters generated from citation papers. By employing this approach, some topics relevant to the target paper tends to be appropriately selected. This improved modeling provides better recommendations for the 15 intra-disciplinary researchers from our dataset of 50 researchers.

Table II. Recommendation accuracy obtained by applying optimal parameters and fragments to the test set. "**" and "*" denote the difference between the best results in the baseline system [Sugiyama and Kan 2010] (underlined scores) and the each result in [Sugiyama and Kan 2013] is significant for $p < 0.01$ and $p < 0.05$, respectively. "†" denotes the difference between the best results in [*pc*-IMP (adp)] + [*frg*-TW] and the best results in [Sugiyama and Kan 2013] (italic scores in [*pc*-IMP] + [*frg*-TW]) is significant for $p < 0.05$.

	nDCG@5	nDCG@10	MRR
***pc*-IMP** ($n = 4$, $N_{pc} = 6$) [Sugiyama and Kan 2013]			
***frg*-TW** ($\alpha = 0.4$, Full text + Conclusion)	***0.581*****	***0.577*****	***0.795****
***pc*-IMP (adp)** [Sugiyama and Kan 2014]			
(n:adaptive, $CL_{th} = 0.56$, $N_{pc} = 8$)			
***frg*-TW** ($\alpha = 0.4$, Full text + Conclusion)	**0.588**†	**0.598**†	**0.804**†
Baseline system [Sugiyama and Kan 2010]			
(Weight "SIM," $Th = 0.4, \gamma = 0.23, d = 3$)	0.527	0.482	0.752
Nascimento *et al.* [Nascimento et al. 2011]			
("Frequency of bi-gram" obtained from title and abstract)	0.335	0.311	0.437
Wang and Blei [Wang and Blei 2011]			
("In-matrix prediction" in collaborative topic regression)	0.396	0.374	0.498
Algarni *et al.* [Algarni et al. 2010]			
(4 times feedback)	0.460	0.433	0.630

Table III. Comparison of recommendation accuracy between our proposed approach and other approaches. "*" denotes that the difference between "MMR-Rafiei(+) in (DU)" and "our proposed approach (DU)" is significant for $p < 0.05$. "††" and "†" denote that the difference between "MMR-Rafiei(+) in (CAN)" and "our proposed approach (CAN)" is significant for $p < 0.01$ and $p < 0.05$, respectively.

	nDCG@10	MRR	nITN@10
Random	0.127	0.087	0.078
MMR-Rafiei(+) in (DU)	0.372	0.553	0.581
Our proposed approach (DU)	0.414*	0.612*	0.642*
(DU-title, $N_{du} = 16$)			
MMR-Rafiei(+) in (CAN)	0.353	0.544	0.568
Our proposed approach (CAN)	**0.426**†	**0.624**†	**0.656**††
(RCP-PLSIM, $pl = 3$)			

3.2.2 *Experimental Results in Serendipitous Recommendation.* We compare the effectiveness of our approaches with "random selection" and the well-known diversifications strategy of maximal marginal relevance (MMR) [Carbonell and Goldstein 1998], which combines query relevance with result novelty. However, one of the important factors in MMR is how to select an appropriate similarity measure between documents and query. In recent work, Rafiei et al. [2010] proposed using both reciprocal rank of the target document in search engine results and standard cosine similarity between the target documents as the similarity measure. Thus, we re-implement their approach and refer to it as "MMR-Rafiei(+)." Table III shows a comparison of the recommendation accuracy between our proposed approaches using their optimal settings and other approaches, "random selection" and "MMR-Rafiei(+)." According to Table III, we observe that the best recommendation accuracies in both of our approaches, (DU) and (CAN) outperform random selection approach and the variant of MMR [Rafiei et al. 2010]. Among the two of our proposed approaches, we find that user profile construction using co-author network (CAN) is most effective in achieving serendipitous recommendation.

4. CONCLUSION AND FUTURE WORK

We have summarized our recent works on scholalrly paper recommendation [Sugiyama and Kan 2010; 2011; 2013; 2014]. Our approach employs content-based filtering at its core, constructing user profiles and a feature vector for each candidate paper to recommend first, and then recommending papers with high similarity between them. In order to provide relevant and serendipitous paper recommendation, we examined two methods to construct the user profile by utilizing each researcher's published paper and its citation and reference papers and utilizing dissimilar users and co-author network, respectively. We also generate feature vectors of candidate papers to recommend by identifying potential citation papers with imputation-based collaborative filtering and its adaptive selection of neighborhoods and by using fragments in the citation and potential citation papers. Experimental results show that our proposed approaches outperform the state-of-the-art with statistical significance. In the course of our work, we have constructed a scholarly paper dataset, which we have made publicly available. Thus, in future work, we plan to analyze research trends using the dataset.

REFERENCES

ABU-JBARA, A., EZRA, J., AND RADEV, D. 2013. Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In *Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2013)*. 596–606.

ALGARNI, A., LI, Y., AND XU, Y. 2010. Selected New Training Documents to Update User Profile. In *Proc. of the 19th International Conference on Information and Knowledge Management (CIKM'10)*. 799–808.

CARAGEA, C., SILVESCU, A., MITRA, P., AND GILES, C. L. 2013. Can't See the Forest for the Trees? A Citation Recommendation System. In *Proc. of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL '13)*. 111–114.

CARBONELL, J. AND GOLDSTEIN, J. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. 335–336.

GORI, M. AND PUCCI, A. 2006. Research Paper Recommender Systems: A Random-Walk Based Approach. In *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006)*. 778–781.

HE, Q., KIFER, D., PEI, J., MITRA, P., AND GILES, C. L. 2011. Citation Recommendation without Author Supervision. In *Proc. of the 4th International Conference on Web Search and Data Mining (WSDM'11)*. 15–24.

HE, Q., PEI, J., KIFER, D., MITRA, P., AND GILES, C. L. 2010. Context-aware Citation Recommendation. In *Proc. of the 19th International World Wide Web Conference (WWW2010)*. 421–430.

HUANG, W., KATARIA, S., KARAGEA, C., MITRA, P., GILES, C. L., AND ROKACH, L. 2012. Recommending Citations: Translating Papers into References. In *Proc. of the 21st International Conference on Information and Knowledge Management (CIKM'12)*. 1910–1914.

JÄRVELIN, K. AND KEKÄLÄINEN, J. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*. 41–48.

JARVIS, R. A. AND PATRICK, E. A. 1973. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Transactions on Computers C22,* 11, pages 1025–1034.

LIVNE, A., GOKULADAS, V., TEEVAN, J., DUMAIS, S. T., AND ADAR, E. 2014. CiteSight: Supporting Contextual Citation Recommendation Using Differential Search. In *Proc. of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)*. 807–816.

LU, Y., HE, J., SHAN, D., AND YAN, H. 2011. Recommending Citations with Translation Model. In *Proc. of the 20th International Conference on Information and Knowledge Management (CIKM'11)*. 2017–2020.

MCNEE, S. M., ALBERT, I., COSLEY, D., P. GOPALKRISHNAN, S. L., RASHID, A. M., KONSTAN, J. S., AND RIEDL, J. 2002. On the Recommending of Citations for Research Papers. In *Proc. of the 2002 ACM Conference on Computer Supported Cooperative Work (CSCW '02)*. 116–125.

NASCIMENTO, C., LAENDER, A. H. F., DA SILVA, A. S., AND GONÇALVES, M. A. 2011. A Source Independent Framework for Research Paper Recommendation. In *Proc. of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011)*. 297–306.

PORTER, M. F. 1980. An Algorithm for Suffix Stripping. *Program 14,* 3, 130–137.

RADEV, D. R., MUTHUKRISHNAN, P., QAZVINIAN, V., AND ABU-JBARA, A. 2013. The ACL Anthology Network Corpus. *Language Resources and Evaluation 47,* 4, pages 919–944.

RAFIEI, D., BHARAT, K., AND SHUKLA, A. 2010. Diversifying Web Search Results. In *Proc. of the 19th International World Wide Web Conference (WWW2010)*. 781–790.

SALTON, G. AND MCGILL, M. J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

STROHMAN, T., CROFT, W. B., AND JENSEN, D. 2007. Recommending Citations for Academic Papers. In *Proc. of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*. 705–706.

SUGIYAMA, K. AND KAN, M.-Y. 2010. Scholarly Paper Recommendation via User's Recent Research Interests. In *Proc. of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL '10)*. 29–38.

SUGIYAMA, K. AND KAN, M.-Y. 2011. Serendipitous Recommendation for Scholarly Papers Considering Relations Among Researchers. In *Proc. of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. 307–310.

SUGIYAMA, K. AND KAN, M.-Y. 2013. Exploiting Potential Citation Papers in Scholarly Paper Recommendation. In *Proc. of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL '13)*. 153–162.

SUGIYAMA, K. AND KAN, M.-Y. 2014. A Comprehensive Evaluation of Scholarly Paper Recommendation Using Potential Citation Papers. *International Journal on Digital Libraries*, DOI: 10.1007/s00799–014–0122–2.

TANG, X., WAN, X., AND ZHANG, X. 2014. Cross-Language Context-Aware Citation Recommendation. In *Proc. of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)*. 817–826.

TORRES, R., MCNEE, S. M., ABEL, M., KONSTAN, J. A., AND RIEDL, J. 2004. Enhancing Digital Libraries with TechLens. In *Proc. of the 4th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)*. 228–236.

VOORHEES, E. M. 1999. The TREC-8 Question Answering Track Report. In *Proc. of the 8th Text REtrieval Conference (TREC-8)*. 77–82.

WANG, C. AND BLEI, D. M. 2011. Collaborative Topic Modeling for Recommending Scientific Articles. In *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*. 448–456.

YANG, D., WEI, B., WU, J., ZHANG, Y., AND ZHANG, L. 2009. CARES: A Ranking-Oriented CADAL Recommender System. In *Proc. of the 9th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2009)*. 203–211.

Kazunari Sugiyama is a researcher at School of Computing, National University of Singapore. His research interests include information retrieval, digital libraries (especially on recommendation systems), and natural language processing. He is a winner of the "Vannevar Bush Best Paper Award" at JCDL '13 for his joint work on scholarly paper recommendation, summarized in part here. He is a member of ACM, AAAI, and IEEE.

Min-Yen Kan (BS;MS;PhD Columbia Univ.; SACM, IEEE) is an associate professor at the National University of Singapore. He serves the School as an Assistant Dean of Undergraduate Studies. Min is a member of the executive committee of the Association of Computational Linguistics (ACL) and also helps to maintain the ACL's Anthology, the community's largest archive of published research. He is also an associate editor for the Springer "Information Retrieval" journal. His research interests include digital libraries and applied natural language processing. Specific projects include work in the areas of scientific discourse analysis, full-text literature mining, machine translation and applied text summarization.