Int J Digit Libr (2015) 16:91–109
DOI 10.1007/s00799-014-0122-2

A comprehensive evaluation of scholarly paper recommendation using potential citation papers

Kazunari Sugiyama · Min-Yen Kan

Received: 1 November 2013 / Revised: 5 July 2014 / Accepted: 5 July 2014 / Published online: 10 August 2014

Abstract To help generate relevant suggestions for researchers, recommendation systems have started to leverage the latent interests in the publication profiles of the researchers themselves. While using such a publication citation network has been shown to enhance performance, the network is often sparse, making recommendation difficult. To alleviate this sparsity, in our former work, we identified "potential citation papers" through the use of collaborative filtering. Also, as different logical sections of a paper have different significance, as a secondary contribution, we investigated which sections of papers can be leveraged to represent papers effectively. While this initial approach works well for researchers vested in a single discipline, it generates poor predictions for scientists who work on several different topics in the discipline (hereafter, "intra-disciplinary"). We thus extend our previous work in this paper by proposing an adaptive neighbor selection method to overcome this problem in our imputation-based collaborative filtering framework. On a publicly-available scholarly paper recommendation dataset, we show that recommendation accuracy significantly outperforms state-of-the-art recommendation baselines as measured by nDCG and MRR, when using our adaptive neighbor selection method. While recommendation performance is enhanced for all researchers, improvements are more marked for intra-disciplinary researchers, showing that our method does address the targeted audience.

Keywords Digital library · Information retrieval · Recommendation · Citation analysis · Collaborative filtering

1 Introduction

Newly discovered knowledge is now largely captured in digital form and archived throughout the world. Archival materials are also being digitized and are increasingly becoming more accessible online. The modern researcher has unprecedented level of access to the sum total of human knowledge. While certainly advantageous, this creates a problem of over abundance, commonly known as "information overload": where researchers find an overwhelming number of matches to their search queries, but for which the majority are largely irrelevant to their latent information needs.

Work in recommendation systems is one promising approach to address the information overload. In digital library studies, this approach has been employed to obtain and refine search results to satisfy each user's information needs [5,15,18,29,34]. However, these approaches do not fully leverage the user's context, largely relying on the idea of session-as-context. This legacy is ported from research in Web search, where session click-through data are used to form the context. To address this problem, in our previous work, we observed that the scholarly context allows us to leverage the role of the searcher-as-author [25]. We modeled a searcher's context in the form of a profile by capturing previous research interests embodied in their past publications, and showed elevated success at scholarly paper recommendation. Our approach in [25] also took advantage of the explicit citation network of publications as a source of knowledge to

This is an extended version of our paper, "Exploiting Potential Citation Papers in Scholarly Paper Recommendation" published in proceedings of the 13th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2013), pages 153–162.

K. Sugiyama (✉) · M.-Y. Kan
Computing 1, 13 Computing Drive, Singapore 117417, Singapore
e-mail: sugiyama@comp.nus.edu.sg; zakugus@gmail.com

M.-Y. Kan
e-mail: kanmy@comp.nus.edu.sg

improve recommendation accuracy. The contents of papers that cite an author's papers as well as the contents of the works referenced in the papers provide supplementary evidence used in modeling the author's research interests.

Following [25], we also proposed two extensions that further mine additional signals from the full text and citation network—using (1) potentially cited papers and (2) their fragments [27]. Citation papers are 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 parts of a paper such as abstract, introduction, conclusion, and so on.

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 hypothesize that we can model the target papers to recommend more accurately to achieve better recommendation performance. We applied collaborative filtering (CF) to find such potential citation papers. While CF is often used to recommend items to users directly, we applied CF to discover potential citation papers that help in representing target papers to recommend.

In [27], we found that imputation-based CF is more effective than CF with binary or similarity values in the discovery of potential citation papers. However, we also observed that if the topic of the target paper is intra-disciplinary, our proposed approach may perform erratically. Our analysis shows that the imputation approach discovers "skewed" potential citation papers. In this journal paper, we overcome this problem through our proposed adaptive selection of neighborhoods, further improving imputation-based CF (see (A3) in Sect. 3.2).

Through a series of experiments on a scholarly paper recommendation dataset, we show that proper modeling of potential citation papers—as well as properly representing papers with both their full text and assigning more weight to the conclusion—improve recommendation accuracy significantly ($p < 0.05$ or better) as judged by both mean reciprocal rank (MRR) and normalized discounted cumulative gain (nDCG). We also show that our approach can outperform state-of-the-art scholarly paper recommendation systems.

This paper is organized as follows: In Sect. 2, we review related work on scholarly paper recommendation for each user, citation recommendation for each paper, and link discovery. In Sect. 3, we detail our approach to find potential citation papers and present our new extension that addresses intra-disciplinary work. In Sect. 4, we present our publicly available dataset and experimental results obtained by our proposed approach and dissect the evaluation results in detail. Finally, we conclude the paper with a summary and directions for future work in Sect. 5.

2 Related work

As the field of recommendation systems is large, we focus our literature review on systems for scholarly paper recommendation for each user and citation recommendation for each paper. In addition, as finding potential papers can be viewed as a type of link discovery, we also briefly review on content link detection.

2.1 Scholarly paper recommendation relevant to each user's interests

With respect to scholarly paper recommendation, Torres et al. [29] proposed a method for recommending research papers by combining CF and content-based filtering (CBF). However, a single final ranking obtainable by merging the output from both CF and CBF is purposefully not done, as the authors claim that pure recommendation algorithms are not designed to receive input from another recommendation algorithm. Gori and Pucci [5] devised a PageRank-based method for recommending research papers. But in their approach, a user must prepare an initial set of relevant articles to obtain better recommendations, and the damping factor d that affects the score of PageRank [21] is not optimized. Yang et al. [34] presented a scholarly paper recommendation system using a ranking-oriented CF. Although their system overcomes the cold-start problem by utilizing implicit behaviors extracted from a user's access logs, the predefined settings for parameters used to select effective data are not justified nor investigated in detail. In recent work, Nascimento et al. [18] developed a scholarly paper recommendation system, in which they use the title to construct user profiles, and the title and abstract to generate feature vectors of candidate papers to recommend. However, we feel that such a small span of text does not effectively represent a user's interest and candidate papers. Actually, we observe that abstract is not effective in constructing feature vectors of candidate papers to recommend [27].

Scholarly paper recommendation studies are also emerging in data mining. Wang and Blei [31] proposed collaborative topic regression model which combines ideas from CF and content analysis based on probabilistic topic modeling. They used the abstract and title of the paper to model a user and characterize candidate papers to recommend, which occasionally results in irrelevant recommendations, similar to [18]. El-Arini and Guestrin [4] proposed a method for discovering a small set of scholarly papers that are relevant to a query yet diversified. They defined "influence" to capture the transfer of ideas as individual concepts among papers in the query. Their approach then returns papers related to these concepts. However, users need to prepare trusted papers in advance to discover relevant and diverse papers.

While the works described above recommend papers relevant to each user's interest, we addressed serendipitous scholarly paper recommendation [26].

2.2 Citation recommendation relevant to each paper

Researchers can benefit from a citation recommendation system because searching for relevant papers to cite is a laborious task. We can classify this field into collaborative filtering (CF)-based, content-based and translation model-based approach.

With respect to CF-based approaches, McNee et al. [15] proposed an approach to recommending citations. Their approach applied collaborative filtering (CF) to social networks to create a graph formed by the citations between research papers. This data can be mapped into a framework of CF and used to overcome the cold-start problem. To solve the problems in [15], Caragea et al. [3] employed SVD to provide better citation recommendation by assuming that an author of a paper possesses some background knowledge. To represent the author's background knowledge, however, users need to prepare initial set of citations relevant to the authors research topics.

With respect to content-based approaches, Strohman et al. [28] experimented with a citation recommendation system where the relevance between two documents is measured by a linear combination of text features and citation graph features. They concluded that the similarity between query and candidate documents, and the Katz distance [13] between the query and candidate documents expanded by their citations are the two most important features in this type of task. He et al. [7] developed a citation recommendation system based on a non-parametric probabilistic model. Their system requires a user to prepare query manuscript without a bibliography that indicates locations where citations are needed, resulting in additional burden for the user. In their subsequent work, they solved this problem by automatically analyzing the query manuscript to suggest locations where citations are needed [6].

Translation models are used originally to translate a text in one language to another language. In the citation recommendation, the citation contexts and the content of papers demonstrate different language properties, such that modeling the problem of citation recommendation task can be sufficiently modeled by translation models. Focusing on this point, Lu et al. [14] introduced the translation model into citation recommendation. They observed that translation models work better when they use the abstract as compared to the full text as document content for constructing the translation model. Following [14], Huang et al. [9] also employed translation models to recommend citations. They first define "descriptive language" and "reference language," which denote citation words in the paper before the reference section and references where each referenced paper is considered as a "word," respectively. However, their approach needs to construct a dictionary.

Patent documents, like scholarly papers, are also associated with citation links. Motivated by the insight that patent citations offer unique and important information about the value of cited patents to citing patents, Oh et al. [20] integrated patent citation information with patent bibliographic information to construct a heterogeneous patent citation–bibliographic information network, achieving effective patent citation recommendation by extracting promising features from the network.

2.3 Content link detection

Content link detection aims to discover similar content across different input and make such links explicit. In Wikipedia link detection, Milne and Witten [17] created explanatory links to all documents using supervised machine learning. They observed that decision tree generator gives better results than other learning techniques. West et al. [32] addressed the same task using unsupervised learning through principal component analysis. Following these studies, Kaptein et al. [12] proposed finding links from Wikipedia pages to external Web pages by using a language modeling approach. In story link detection, Nomoto [19] proposed a two-tier model of similarity, at both the document and collection levels. His similarity model adapted the idea of relevance feedback to link detection, where stories are measured for similarity not merely based on the document, but on a collection of relevant documents. Finally, by combining two algorithms proposed in [17] and [32], West et al. [33] created a hybrid algorithm that suggests topics to authors of text documents.

3 Proposed method

Our work tackles the core problem of matching users to candidate papers. Unlike existing scholarly paper recommendation systems which focused on user profile construction [5,18,25], our work leverages the scholarly papers more effectively, through the modeling of potential citation papers and their fragments, and enhancing the citation network with automatically identified potential citation papers. Unlike citation recommendation that provide relevant citations for each paper, we apply collaborative filtering to discover potential citation papers that help model target papers to recommend. And unlike previous work in content link detection which mainly focused on finding an effective learning framework, we focus on how to best use the scholarly corpora available to us.

3.1 Baseline system [25]

Our method starts with our former scholarly paper recommendation system [25], and as such it is instructive to first describe our system and its basis. It consists of three steps:

Step 1: Construct a user profile $\boldsymbol{P}_{\text{user}}$ from a researcher's list of published papers;
Step 2: Compute feature vectors $\boldsymbol{F}^{p_j}$ $(j = 1, \ldots, t)$ for each of the papers in its scholarly paper knowledge base;
Step 3: Compute the cosine similarity $\text{Sim}(\boldsymbol{P}_{\text{user}}, \boldsymbol{F}^{p_j})$ between $\boldsymbol{P}_{\text{user}}$ and $\boldsymbol{F}^{p_j}$ $(j = 1, \ldots, t)$, and recommend papers with high similarity to the target user.

A candidate paper to recommend (p) is represented as a feature vector $\boldsymbol{f}^p$. We employ TF and TF-IDF [24] schemes in Steps 1 and 2, respectively. Both $\boldsymbol{P}_{\text{user}}$ and $\boldsymbol{F}^{p_j}$ are constructed as the combination of $\boldsymbol{f}^p$ as defined by Eq. (1). As such, our method views both user profiles and candidate papers to recommend as vectors of terms with specific, per-term tuned weights. As CBF relies on the item's content to provide its recommendations, it is important to represent an item's contents faithfully. A key innovative step in this approach was to model a target paper of interest based on not merely its own textual content but also 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 Fig. 1a).

When the text of such contextual papers is added to the original target paper weighted by cosine similarity to the target paper, recommendation accuracy was improved the most among other alternatives explored.

In [27], we further enhanced Step 2 above, to both **enlarge** what is meant by context through the discovery of potential citation papers (Fig. 1b), as well as **refine** its use in specific, well-linked parts of the contextual documents through the specific modeling of potential citation papers and their fragments.

To facilitate our continuing discussion, we show the original formula for Step 2 defined in [25] to compute the feature vector for each paper p:

$$\boldsymbol{F}^p = \boldsymbol{f}^p + \sum_{x=1}^{k} W^{p_{\text{cit}_x} \to p}\, \boldsymbol{f}^{p_{\text{cit}_x}} + \sum_{y=1}^{l} W^{p \to p_{\text{ref}_y}}\, \boldsymbol{f}^{p_{\text{ref}_y}}, \tag{1}$$

In Eq. (1), we define general weighting coefficients of the form $W^{u \to v}$ to denote the weight between target v and its source u. For these weights, we implement the cosine similarity $\text{sim}(\boldsymbol{f}^u, \boldsymbol{f}^v)$ between a feature vector representation of the two papers u and v.

By operationalizing this scheme into Eq. (1), 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. In addition, $W^{p_{\text{cit}_x} \to p}$ and $W^{p \to p_{\text{ref}_y}}$ are weights for the citation papers and weights for the reference papers, respectively.

Generally speaking, a target paper's feature vector comprises of three parts: words from its own body, words from papers that cite it, and words from papers that it refers to (first, second and third terms in Eq. 1, respectively).

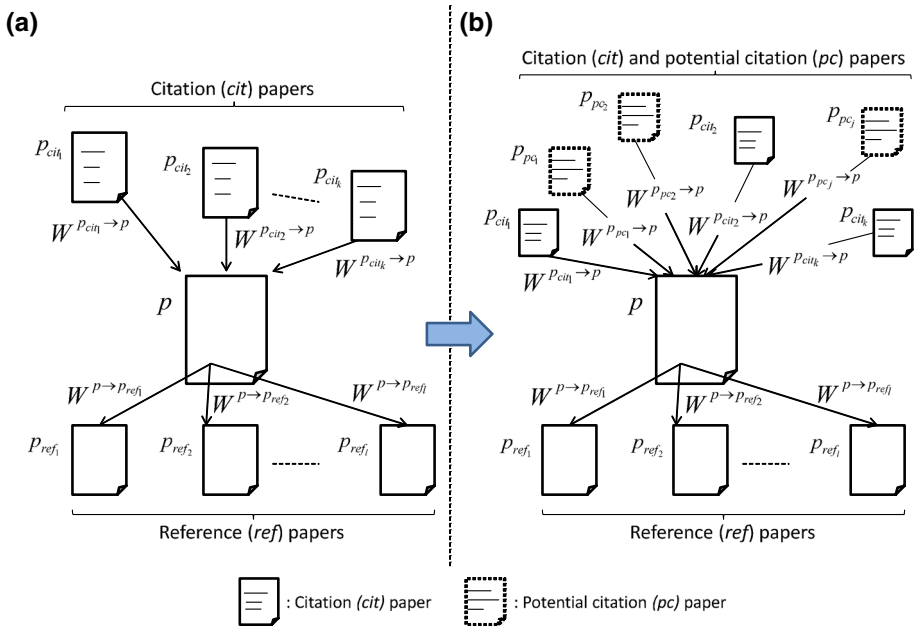


Fig. 1 Comparison of paper representations between our former works [25] and [27] (notations simplified from [25]). This article recaps how we leverage additional potential citation papers to enrich the description of a target paper. **a** Baseline system [25], **b** Enhanced system [27]

3.2 Leveraging potential citation papers

While a rich source of information, a citation network is subject to certain limitations that blunt its effectiveness in modeling target papers. We note that the citation network is constantly expanding; with every new publication, new citation links are added to older work. In studies depending solely on the citation network, cutting-edge work is marginalized as they do not have any citations yet; this is a kind of "cold-start problem" in scholarly recommendation systems that is analogous to the same problem in recommendation systems in general.

Also, because references and citations in a paper are static and never change, newer relevant papers to older ones have the "responsibility" of creating a citation link between them. The static nature of the citation network exacerbates missing and noisy citations.

Finally, the citation network is an artifact of the physical scholarly paper. In many cases, listing all relevant work would be infeasible, as the reference list may grow too long. Many venues have space limitations, ostensibly to help encourage authors to use their editorial powers to choose the most relevant references to include. However, this can also cause authors to prune potentially citable references from their bibliographies. We note that when authors save the space, the balance may be used to expand the description of their own approaches or experiments.

The above factors led us to believe that the observable, explicit citation network—while certainly of high-quality—is just "the tip of the iceberg"; where iceberg refers to the implicit set of relevant works for a target paper. We term papers in this implicit set potential citation (pc) papers. If we can predict these implicitly relevant papers, we obtain more content for representing a scholarly paper, which in turn, we hypothesized would improve recommendation performance.

In our approach, we discover such potential citation papers by applying collaborative filtering (CF). CF is usually used to recommend items directly to users. However, we employ it indirectly, by using it to discover potential citation papers, which are then used to represent papers to recommend. This discovery process is needed to better represent papers, which in turn enhances recommendation accuracy. Importantly, our use of CF operates on the paper–citation matrix, and is markedly different from its traditional one-step use in the user–item matrix; in contrast, we employ the citation network twice: both in directly representing target papers through citations and references as well as in finding potential citation papers. The details of our approach also break down into the discovery of potential citation papers [(A1) using CF and (A2) imputation-based CF] and (B) feature vector construction for target papers using the discovered potential citation papers [27]. However, in (A2), we observed that if the topic of the target paper is intra-disciplinary, our proposed approach tends to perform erratically. To overcome this problem, we extend (A2) in this journal version, proposing "(A3) discovery of potential citation papers with imputation-based CF using adaptive selection of neighborhoods" to address such intra-disciplinary research.

(A1) Discovery of potential citation (pc) **papers with CF**

We apply the neighborhood-based algorithm [8] in CF for use in discovering potential citation papers, by substituting papers for users and items for citations. At a high level, we can think of papers as actors that can recommend citations to each other, where CF lets papers that are more similar to a target paper (from a citation perspective) recommend citations with more weight. The algorithm has the following steps analogous to neighborhood-based CF:

A1.1: Weight all papers with respect to similarity to a target paper. As with the original CF algorithm, paper similarity is measured using the Pearson correlation coefficient between their citation vectors.
A1.2: Select n papers that have the highest similarity with the target paper. These papers form the *neighborhood* for the target.
A1.3: Compute a prediction from a weighted combination of the neighbor's values using a suitable similarity score.

In Step A1.1, the similarity between target paper p_{tgt} and other citation papers p_{cit_u} $(u = 1, \ldots, N)$, denoted as $S_{\text{tgt},u}$ is computed using the Pearson correlation coefficient:

$$S_{\text{tgt},u} = \frac{\sum_{i=1}^{N}(r_{\text{tgt},i} - \bar{r}_{\text{tgt}}) \times (r_{\text{cit}_u,i} - \bar{r}_{\text{cit}_u})}{\sqrt{\sum_{i=1}^{N}(r_{\text{tgt},i} - \bar{r}_{\text{tgt}})^2 \times \sum_{i=1}^{N}(r_{\text{cit}_u,i} - \bar{r}_{\text{cit}_u})^2}}, \quad (2)$$

where $r_{\text{tgt},i}$ is the score given to citation paper p_{cit_i} by paper p_{tgt}, and $\bar{r}_{\text{tgt}}$ is the mean score given by paper p_{tgt}, and N is the total number of papers in the dataset.

In Step A1.2, a subset of appropriate papers is 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 A1.3.

In Step A1.3, predictions are computed as the weighted average of deviations from the neighbor's mean, shown in Eq. (3):

$$p_{\text{tgt},i} = \bar{r}_{\text{tgt}} + \frac{\sum_{u=1}^{n}(r_{u,i} - \bar{r}_u) \times S_{\text{tgt},u}}{\sum_{u=1}^{n} S_{\text{tgt},u}}, \quad (3)$$

where $p_{\text{tgt},i}$ is the prediction for a target paper p_{tgt} for a citation paper p_{cit_i}. n is the number of papers in the neighborhood.

We explored two possible methods for calculating $S_{\text{tgt},u}$ and $p_{\text{tgt},i}$ above: a binary notion of citation (Fig. 2a), as well as a fine-grained similarity version of citation (Fig. 2b).

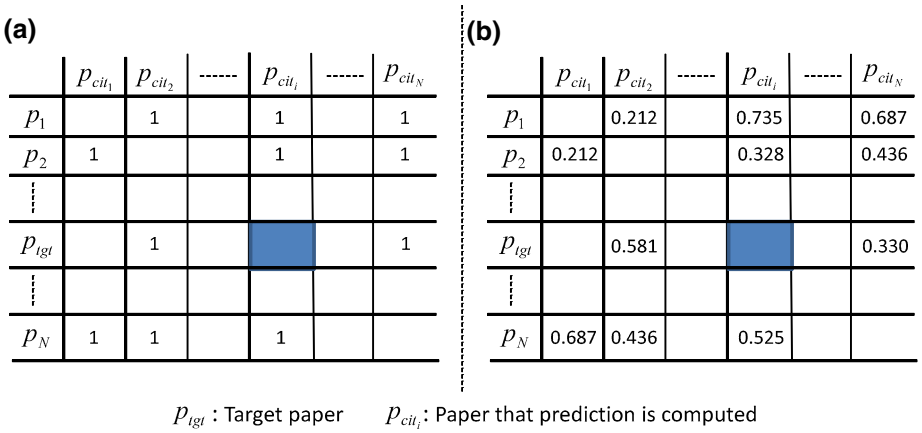


Fig. 2 Paper–citation matrix for our adapted collaborative filtering using **a** binary [pc-BIN] and **b** similarity [pc-SIM] weighting

The binary scheme is illustrated in Fig. 2a, which shows a paper–citation matrix with binary incidence values [pc-BIN]. Entries with a '1' indicate citations by the paper identified by the column to the target paper identified by the row (e.g., paper p_{tgt} is only cited in papers p_{cit_2} and p_{cit_N}).

It is generally agreed that citations have different functions. A key reference that acts as the foundation for the current work is likely more of a positive endorsement than a citation within a list of examples of applications of a particular model. We choose to use cosine similarity between papers as a simple means to model endorsement strength. Figure 2b shows a corresponding paper–citation matrix with similarity values [pc-SIM]. For example, in Fig. 2b, the similarity between the target paper p_{tgt} and p_{cit_2}, and the target paper p_{tgt} and p_{cit_N} is 0.581 and 0.330, respectively.

To be clear, in both models, multiple citations to the same target paper within a paper are not represented.

(A2) Discovery of potential citation (pc) papers with imputation-based CF

In Fig. 2, the matrices are sparse because each paper can only make a limited number of citations (see Sect. 4 about how sparse our dataset is). This affects the process of finding relevant potential papers. However, when the corpus of publications is large, we can utilize the fact that there are many other similar papers that potentially could have been cited but were not.

To leverage this opportunity and address sparseness, we employ imputation (hereafter, [pc-IMP]) as we can directly compute similarity between papers and citation papers, unlike the case of the user–item matrix based CF which requires manual ratings. This is a variant of [pc-SIM] and consists of three steps:

A2.1: Impute similarities between all papers, recording them into an intermediate imputed paper–citation matrix (Fig. 3).

A2.2: For the target paper, find the n most similar papers from the "(a) original matrix" in Fig. 3:

- Weight all papers with respect to similarity to the target paper (e.g., p_1). This similarity between papers is measured using the Pearson correlation coefficient between the papers' citation vectors,
- Select n papers that have the highest similarity with the target paper. These papers form the n-neighborhood for the target paper. In the left of Fig. 4, p_2, p_4, and p_5 are determined to be the 3-neighborhood for p_1.

A2.3: Compute a prediction from a weighted combination of the neighbor's similarity (Fig. 4, right). We use Fig. 3's "(b) intermediate imputed matrix" for the prediction calculation.

(A3) Discovery of potential citation (pc) papers with imputation-based CF using adaptive selection of neighborhoods

We found that the limitations in [pc-IMP] are that the imputation approach discovers "skewed" potential citation papers when the target paper is intra-disciplinary. In one instance, where the topic of a candidate paper concerned the "understanding mobile user's behavior patterns" that is equally embodied by mobile technology, user search behavior and clustering, [pc-IMP] discovers potential citation papers that only addressed mobile technology, and did not recommend any papers on behavior pattern mining. Our further analysis linked the cause of the skewed discovery of potential citation papers to the fact that the selected n-neighborhood of papers consists almost exclusively one specific topic, mobile technology.

In this journal paper, to overcome this problem and achieve balanced neighborhood selection, we introduce an enhancement of [pc-IMP] that employs clustering to adaptively select neighborhoods (hereafter, [pc-IMP (adp)]). [pc-IMP (adp)] consists of the following steps:

Fig. 3 Similarity imputation: **a** original matrix and **b** intermediate imputed matrix (imputed values are bolded)

(a)

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1		0.233			0.628
p_2	0.233		0.147		
p_3		0.147		0.265	
p_4			0.265		
p_5	0.628				

(b)

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1	1.000	0.233	**0.723**	**0.538**	0.628
p_2	0.233	1.000	0.147	**0.476**	**0.156**
p_3	**0.723**	0.147	1.000	0.265	**0.521**
p_4	**0.538**	**0.476**	0.265	1.000	**0.268**
p_5	0.628	**0.156**	**0.521**	**0.268**	1.000

Imputation

Fig. 4 Predictions computed for the target paper p_1 using corresponding neighbors, p_2, p_4, and p_5 with similarities from the intermediate imputed matrix

[Target paper p_1 and corresponding imputed similarities of neighborhood (p_2 , p_4 , and p_5) from "(b) imputed matrix" in Figure 3]

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1	1.000	0.233	?	?	0.628
p_2	0.233	1.000	0.147	**0.476**	**0.156**
⋮	⋮	⋮	⋮	⋮	⋮
p_4	**0.538**	**0.476**	0.265	1.000	**0.268**
p_5	0.628	**0.156**	**0.521**	**0.268**	1.000

Prediction of similarity for target paper p_1

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1	1.000	0.233	**0.682**	**0.453**	0.628

A3.1: Impute similarities between all papers, recording them into an intermediate imputed paper–citation matrix (Fig. 3).

A3.2: For the target paper, find the n most similar clusters from the "(b) imputed matrix" in Fig. 3:

- Generate clusters of papers by means of k nearest neighbor clustering [11], 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 Fig. 5, C_1, and C_2 are determined to be the 2-neighborhood for p_1.

A3.3: Compute a prediction from a weighted combination of the neighbor's values (Fig. 5b) using centroid vectors of clusters.

We review the two latter steps in more detail. In Step A3.2, the similarity between target paper p_{tgt} and centroid vectors of clusters g, is computed using the Pearson correlation coefficient similar to Eq. (2):

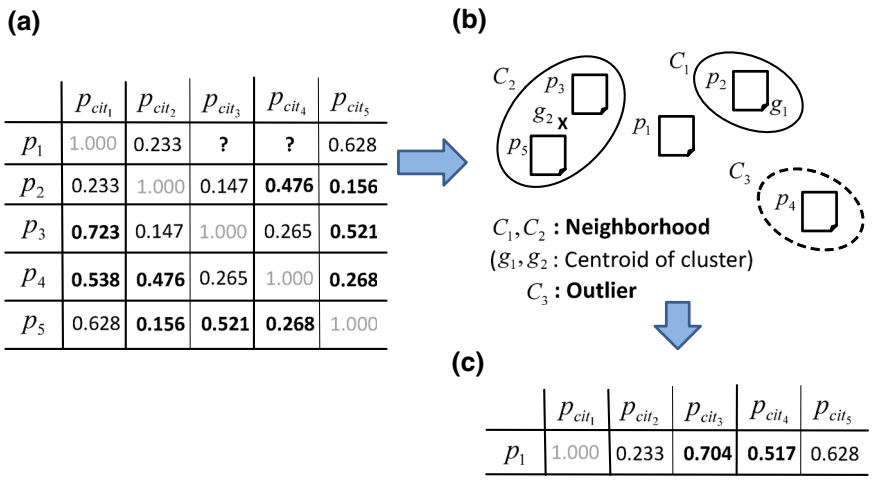
$$S_{\text{tgt},g} = \frac{\sum_{i=1}^{N}(r_{\text{tgt},i} - \bar{r}_{\text{tgt}}) \times (r_{g,i} - \bar{r}_g)}{\sqrt{\sum_{i=1}^{N}(r_{\text{tgt},i} - \bar{r}_{\text{tgt}})^2 \times \sum_{i=1}^{N}(r_{g,i} - \bar{r}_g)^2}}, \tag{4}$$

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 A3.3. In this step, the number of selected clusters may differ per target paper, hence our use of "adaptive." We expect that this method forms more relevant neighborhoods for certain target papers.

In Step A3.3, predictions are computed as the weighted average of deviations from the neighbor's mean, shown in Eq. (5):

$$p_{\text{tgt},i} = \bar{r}_{\text{tgt}} + \frac{\sum_{g=1}^{n}(r_{g,i} - \bar{r}_g) \times S_{\text{tgt},g}}{\sum_{g=1}^{n} S_{\text{tgt},g}}, \tag{5}$$

where $p_{\text{tgt},i}$ 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.



	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1	1.000	0.233	?	?	0.628
p_2	0.233	1.000	0.147	**0.476**	**0.156**
p_3	**0.723**	0.147	1.000	0.265	**0.521**
p_4	**0.538**	**0.476**	0.265	1.000	**0.268**
p_5	0.628	**0.156**	**0.521**	**0.268**	1.000

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1	1.000	0.233	**0.704**	**0.517**	0.628

Fig. 5 Predictions computed for the target paper p_1 using centroid vectors of corresponding clusters (neighbors), C_1 and C_2. **a** Target paper p_1 and intermediate imputed matrix, **b** target paper p_1 and generated clusters, **c** Prediction of similarity for target paper p_1

(B) Feature vector construction for target papers

With the discovery and weightage of our discovered potential papers, we can now build the feature vector for target papers. Let $\boldsymbol{F}^p$ be the feature vector for a paper to recommend p. We then define $\boldsymbol{F}^p$ as follows:

$$\boldsymbol{F}^p = \boldsymbol{f}^p + \sum_{x=1}^{j} W^{p_{\mathrm{pc}_x} \to p} \boldsymbol{f}^{p_{\mathrm{pc}_x}} + \sum_{y=1}^{k} W^{p_{\mathrm{cit}_y} \to p} \boldsymbol{f}^{p_{\mathrm{cit}_y}} + \sum_{z=1}^{l} W^{p \to p_{\mathrm{ref}_z}} \boldsymbol{f}^{p_{\mathrm{ref}_z}}, \quad (6)$$

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 weight for $W^{p_{\mathrm{pc}_x} \to p}$, $W^{p_{\mathrm{cit}_y} \to p}$, and $W^{p \to p_{\mathrm{ref}_z}}$ as it was found effective in our previous work [25].

3.3 Leveraging fragments in potential citation papers

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. As scholars, we often acknowledge the importance of others' previous work by citation, an explicit reference to previous work which is accompanied by a bibliographic reference to help others locate and trace the prior work. The in-text citation often clearly and succinctly describes a key point of the target paper of import to the current paper, as illustrated below:

Resnik (1999) addressed the issue of language identification for finding Web pages in the languages of interest.

Citation sentences have been used to build summaries of target papers [16,23] as well as for supporting scientific literature search [2]. However, to the best of our knowledge, they have not been used as an evidence source for recommendation.

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. We also examine the impact of using other short texts as evidence: keywords (1–10 words with the highest TF-IDF score), in place of citation sentences to model candidate papers to recommend.

We note that Mei and Zhai [16] proposed generating summaries using a paper and its citation context (hereafter, [CC]), rather than just using the bare citation sentence. Their approach fixed the citation context to two sentences before and after the citing sentence. For these reasons, we also explore varying the context, the number of sentences before and after the citing sentence, N_{cs} $(1 \leq N_{\mathrm{cs}} \leq 3)$.

Given these possible (potential) citation paper fragments, we tried the following two different schemes to weight the fragments' words in constructing the target candidate paper's feature vector $\boldsymbol{F}^p$.

1. [frg-SIM]: fragments with cosine similarity weighting

In this approach, we add an additional vector obtained from the fragment in the actual or potential citation paper.

This approach effectively allows the tunable weights to assign customized weights to the words that appear in the associated fragments, modifying the feature vector $\boldsymbol{F}^p$ to:

$$\begin{aligned}\boldsymbol{F}^p = &\sum_{x=1}^{j} W_{(\text{frg})}^{p_{\text{pc}_x} \to p} \boldsymbol{f}_{(\text{frg})}^{p_{\text{pc}_x}} + \sum_{y=1}^{k} W_{(\text{frg})}^{p_{\text{cit}_y} \to p} \boldsymbol{f}_{(\text{frg})}^{p_{\text{cit}_y}} \\ &+ \boldsymbol{f}^p + \sum_{x=1}^{j} W^{p_{\text{pc}_x} \to p} \boldsymbol{f}^{p_{\text{pc}_x}} \\ &+ \sum_{y=1}^{k} W^{p_{\text{cit}_y} \to p} \boldsymbol{f}^{p_{\text{cit}_y}} \\ &+ \sum_{z=1}^{l} W^{p \to p_{\text{ref}_z}} \boldsymbol{f}^{p_{\text{ref}_z}}, \end{aligned} \tag{7}$$

where the first row are two added terms to Eq. (6) that account for evidence from the fragments in potential and explicit citation papers, respectively. As in previous sections, we use cosine similarity as the weighting scheme for both coefficients.

2. [frg-TW]: [frg**-SIM] with tunable weight**

In this variation, we further augment the feature vector obtained from a fragment with tunable constant weight α ($0 \leq \alpha \leq 1$), which changes the feature vector calculation to:

$$\begin{aligned}\boldsymbol{F}^p = &\alpha \left(\sum_{x=1}^{j} W_{(\text{frg})}^{p_{\text{pc}_x} \to p} \boldsymbol{f}_{(\text{frg})}^{p_{\text{pc}_x}} + \sum_{y=1}^{k} W_{(\text{frg})}^{p_{\text{cit}_y} \to p} \boldsymbol{f}_{(\text{frg})}^{p_{\text{cit}_y}} \right) \\ &+ (1-\alpha) \Bigg(\boldsymbol{f}^p + \sum_{x=1}^{j} W^{p_{\text{pc}_x} \to p} \boldsymbol{f}^{p_{\text{pc}_x}} \\ &+ \sum_{y=1}^{k} W^{p_{\text{cit}_y} \to p} \boldsymbol{f}^{p_{\text{cit}_y}} \\ &+ \sum_{z=1}^{l} W^{p \to p_{\text{ref}_z}} \boldsymbol{f}^{p_{\text{ref}_z}} \Bigg), \end{aligned} \tag{8}$$

where α represents the balance between the contribution from the full text and the fragments, and allows our model a bit more expressiveness by finding optimal parameters.

To be clear, in both the [frg-SIM] and [frg-TW] methods, only the contribution of terms in a fragment is changed; i.e., Eqs. (7) and (8) only differ from Eq. (6) in the first row, dealing with the contribution of the fragments.

4 Experiments

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. Among them, 15 researchers conduct intra-disciplinary research (as subjectively assessed by the first author). The researchers also have publication lists in DBLP.[1] 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[2] (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, and those in the references section of each paper. Then, we construct feature vectors for these papers as described in Sect. 3. Stop words[3] were eliminated from each user's publication list and from the candidate papers to recommend. Stemming was performed using the Porter Stemmer[4] [22]. 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 1 shows some statistics about our experimental data. In the paper–citation matrix in Fig. 2, only 17.2 % of all cells are filled, demonstrating that the paper–citation matrix for our dataset is sparse. We have made our entire dataset publicly available,[5] to encourage the community to work on this problem and to facilitate competitive benchmarking.

4.1 Evaluation measures

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) [10], and (2) mean reciprocal rank (MRR) [30].

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

[2] http://dl.acm.org/.

[3] ftp://ftp.cs.cornell.edu/pub/smart/english.stop.

[4] http://www.tartarus.org/~martin/PorterStemmer/.

[5] http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html.

Table 1 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 (maximum 169)
Average number of reference papers	15.0 (maximum 58)
(b) Candidate papers to recommend	
Number of papers	100,351
Average number of citation papers	17.9 (maximum 175)
Average number of reference papers	15.5 (maximum 53)

(c) Intra-disciplinary researchers' research topics

Researcher	Research topics
R1	Creating and processing nursing documents, public vs. private work, human error, Bayesian networks
R3	Ajax, SQL, workflow apps, expert finding, enterprise search, search engine performance
R9	Distributed systems, network traffic analysis, protection from attacks, data mining
R13	User interaction, machine learning, text mining
R17	Processor, data mining, distributed system, fault tolerance, load balancing
R21	Real time applications, simulation, embedded systems, stream processing
R32	Aspect-oriented programming, software testing, mobile collaborative applications
R33	Mobile user browsing behavior, network monitoring, video streaming
R37	Transactional memory, work load, information flow control, privacy, data mining
R38	Authentication, protocol analysis, self-managing software patching, machine learning
R43	Code summarization, software readability, software documentation, machine learning
R44	Software maintenance, dataflow analysis, debugging, human factors
R45	Context-aware system, distributed applications, mobile network
R47	XML, user interface, workflow management
R49	Crowdsourcing, user behavior models, electronic markets, text mining

(1) Normalized discounted cumulative gain (nDCG)

nDCG is well-suited for the evaluation of recommendation systems, as it rewards relevant items in the top ranked results more heavily than those ranked lower. For a given user profile $\boldsymbol{P}_{\text{user}_i}$, the ranked results are examined top-down, where nDCG is computed as:

$$\text{nDCG}_i = Z_i \sum_{j=1}^{R} \frac{2^{r(j)} - 1}{\log(1+j)},$$

where Z_i is a normalization constant calculated so that a perfect ordering would obtain nDCG of 1; and each $r(j)$ is an integer relevance level (for our case, $r(j) = 1$ and $r(j) = 0$ for relevant and irrelevant recommendations, respectively) of the result returned at rank j ($j = 1, \ldots, R$). Then, nDCG_i is averaged over all our target researchers. As a typical recommendation system will just recommend a few items, we are only concerned about whether the top ranked results are relevant or not. Therefore, in this work, we use nDCG@R ($R = \{5, 10\}$) for evaluation where R is the number of top-R papers recommended by our proposed approaches.

(2) Mean reciprocal rank (MRR)

MRR indicates where in the ranking the first relevant item is returned by the system, averaged over all researchers. This measure provides insight in the ability of the system to return a relevant paper at the top of the ranking. Let r_i be the rank of the highest ranking relevant paper for a target researcher i, then MRR is just the reciprocal rank, averaged over all target researchers, N_{tr}:

$$\text{MRR} = \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \frac{1}{r_i}.$$

4.2 Experimental results

We first optimize our method's parameters using the training set, and then show experimental results after applying the optimal parameters to the test set. Since there are a few parameters to tune in our approach, we divide the tuning into two halves, where the first half (Phase 1) determines optimal parameters to discover potential citation papers, used in the two independent phases in the second half (Phases 2A and 2B) to leverage fragments.

For simplicity, we only show the best results obtained by using optimal tunable weight α in [pc-BIN] as the improvement compared with the baseline system (see Sect. 3.1) is marginal and we observe the same trends as [pc-SIM] and [pc-IMP]. In addition, in [CC] and "keywords," we only show the best result, namely, $N_{\text{cs}} = 1$ and five keywords, respectively. The remaining parameters ("Weight SIM," Th, γ, and d) that are inherited from the previous framework, are optimized here, following the methodology in [25].

We also compare our proposed approach with state-of-the-art scholarly paper recommendation systems [18,31] and recent pseudo relevance feedback approach based on frequent term pattern mining [1]. Nascimento et al.'s work [18] is a scholarly paper recommendation system based on content-based filtering which is the same approach as ours. In Wang and Blei's work [31], their experimental setting, "in-matrix prediction" to predict the score of paper–citation matrix is similar to ours. That is why we compare our approach with them. We apply their optimal settings to our experiments. For our implementation of [18], we construct the user profile using the title, and construct the feature vector of candidate papers to recommend using the bigram frequency extracted from the title and abstract. In [31], as described above, we apply their "in-matrix prediction" to predict the score of paper–citation matrix in Fig. 2 to discover potential citation papers. Finally, regarding the window size in [1], we set it to the number of each researcher's published papers in the past. We employ four times feedback as in [1].

Note that, in our own method, collaborative filtering is *indirectly used* to discover potential citation papers and expand citation network—our method's use of collaborative filtering is not a direct application, and thus we cannot compare our content-based filtering with collaborative filtering to recommend papers (i.e., we do not have user ratings for papers).

Phase 1: Parameter tuning to discover potential citation papers [TUNE:pc]

We first optimize parameters for finding the potential citation papers, namely the number of neighborhoods n and the number of potential citation papers N_{pc}. We optimize these parameters by using [pc-BIN], [pc-SIM], [pc-IMP], and [pc-IMP (adp)] approaches as described in Sect. 3.2. Table 2, Fig. 6a–c, d–f, and g–i show experimental results obtained by using [pc-BIN], [pc-SIM], [pc-IMP], and [pc-IMP (adp)], respectively.

Table 2 Tuning to address paper–citation matrix sparsity: Recommendation accuracy in [TUNE:pc] when modeling candidate papers using (potential) citation papers under [pc-BIN]

pc-BIN	nDCG@5	nDCG@10	MRR
$n = 2$, $N_{\text{pc}} = 5$	0.541	0.508	0.765
$\boldsymbol{n = 4}$, $\boldsymbol{N_{\text{pc}} = 5}$	**0.548**	**0.516**	**0.770**
$n = 8$, $N_{\text{pc}} = 5$	0.530	0.501	0.759
$n = 10$, $N_{\text{pc}} = 5$	0.526	0.498	0.757
Baseline [25] (Weight "SIM", Th $= 0.4$, $\gamma = 0.23$, $d = 3$)	0.521	0.489	0.750

From Table 2 and Fig. 6, we observe that the optimal parameters that give the best recommendation accuracy are ($n = 4$, $N_{\text{pc}} = 5$) in [pc-BIN] and [pc-SIM], ($n = 4$, $N_{\text{pc}} = 6$) in [pc-IMP], and ($CL_{\text{th}} = 0.56$, $N_{\text{pc}} = 8$) in [pc-IMP (adp)]. These n, N_{pc}, and CL_{th} values are held constant in Phase 2.

Phase 2A: Tuning fragments in frg-SIM [TUNE:frg-SIM]

After obtaining the optimized parameters n, N_{pc}, and CL_{th} to find potential citation papers, we further explore which fragments in the citation and potential citation papers give the best recommendation accuracy using Eq. (7). Table 3 shows the results.

Phase 2B: Tuning α and fragments in frg-TW [TUNE:frg-TW]

In the other second phase experiment, we optimize the weight for α in Eq. (8), and fragments in citation and potential citation papers that give the best recommendation accuracy. Table 4a and Fig. 7, Table 4b and Fig. 8 show recommendation accuracy obtained by using "only fragments" and "both full text and fragments" in citation and potential citation papers, respectively.

Finally, after applying the optimized parameters on the test set, we arrive at the final test recommendation accuracies. These results are shown in Table 5 and discussed in full later.

4.3 Discussion

In the experiments in [TUNE:pc], as shown in Fig. 6, for [pc-BIN], [pc-SIM], and [pc-IMP], when the number of neighbors n is too small ($n = 2, 3$) or too large ($n \geq 10$), we obtain poor predictions which result in selecting irrelevant potential citation papers. On the other hand, when the number of

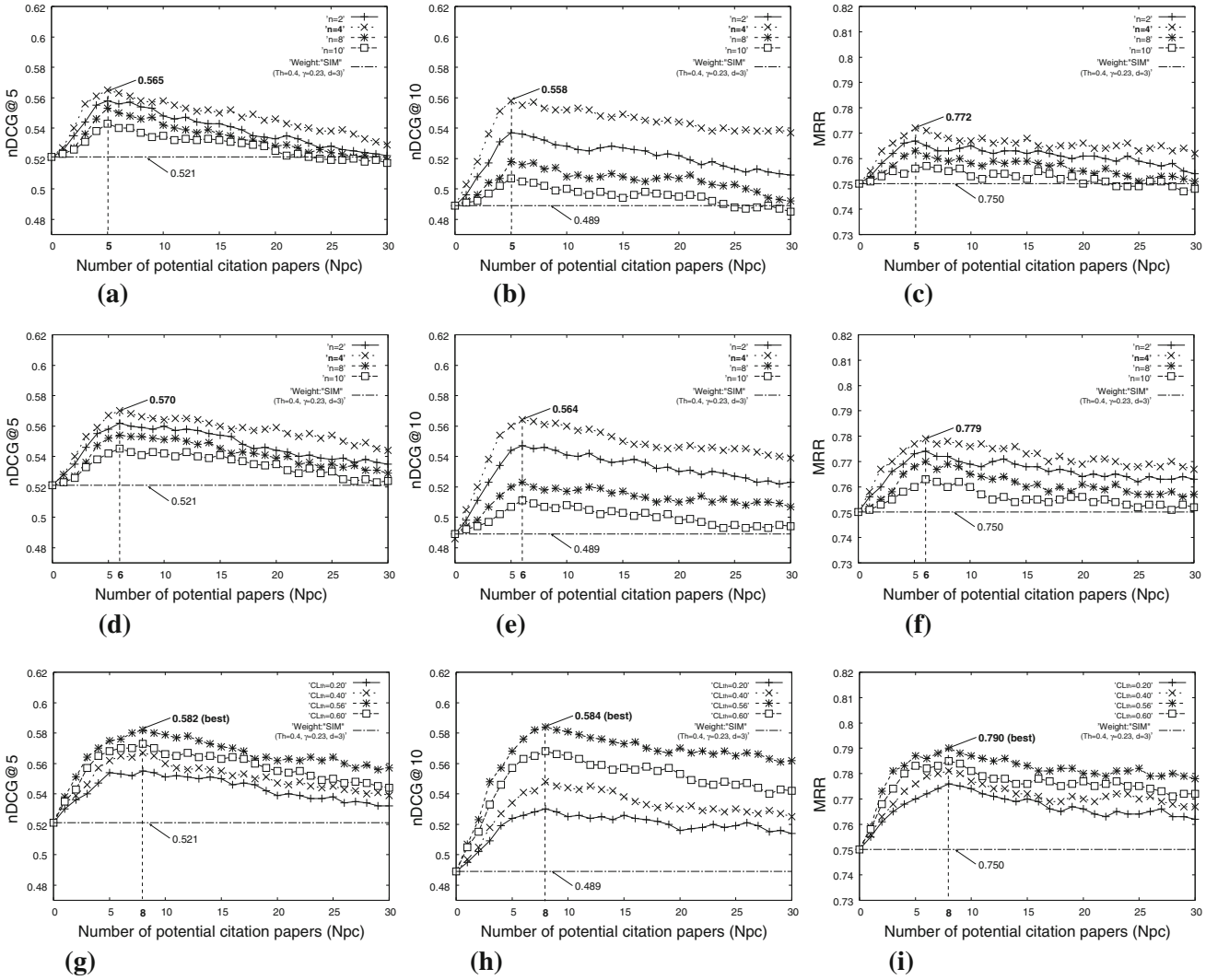


Fig. 6 Tuning to address paper–citation matrix sparsity: Recommendation accuracy in [TUNE:pc] when using a variable number of citation and potential citation papers in [pc-SIM] (**a**–**c**), [pc-IMP] (**d**–**f**), and [pc-IMP (adp)] (**g**–**i**). **a** nDCG@5 [pc-SIM], **b** nDCG@10 [pc-SIM], **c** MRR [pc-SIM], **d** nDCG@5 [pc-IMP], **e** nDCG@10 [pc-IMP], **f** MRR [pc-IMP], **g** nDCG@5 [pc-IMP (adp)], **h** nDCG@10 [pc-IMP (adp)], **i** MRR [pc-IMP (adp)]

neighbors n is 4, our method can select relevant potential citation papers (N_{pc}), resulting in higher recommendation accuracy. We also observe that N_{pc} remains stable—5 in [pc-BIN] (Table 2) and [pc-SIM] (Fig. 6a–c), and 6 in [pc-IMP] (Fig. 6d–f).

In [pc-IMP (adp)], according to Fig. 6g–i, when we set the threshold of similarity in clustering to a smaller value ($CL_{\text{th}} \leq 0.4$), the generated clusters do not form effective neighborhoods as much, resulting in selecting irrelevant potential citation papers. In addition, when we set the threshold of similarity in clustering to larger values ($CL_{\text{th}} \geq 0.6$), we observe that the results are improved compared with the smaller settings. But due to the higher threshold, it is difficult to merge a cluster into other clusters, generating many clusters with a single member. This tends to yield ineffective neighborhoods for obtaining good prediction results. Note that we only show the results obtained by $CL_{\text{th}} = 0.6$ in Fig. 6g–i as they are almost the same even if we set CL_{th} to much larger threshold. Finally, we observe that, in [pc-IMP (adp)], the optimal value of CL_{th} and N_{pc} are 0.56 and 8, respectively.

The above observation indicates that [pc-IMP] and [pc-IMP (adp)] find potential citation papers more effectively than [pc-BIN] and [pc-SIM]. We believe that its effectiveness is due to their use of the additional n-neighborhood of context papers in [pc-IMP] or generated clusters in [pc-IMP (adp)] used to impute the missing values for a target paper.

Table 3 Recommendation accuracy (nDCG@5, nDCG@10, and MRR) in [TUNE:frg-SIM] obtained by fragments in potential citation papers. The parameters n, N_{pc} and CL_{th} are optimized ones in [pc-BIN], [pc-SIM], [pc-IMP] and [pc-IMP (adp)

pc-BIN ($n = 4$, $N_{\mathrm{pc}} = 5$)	nDCG@5	nDCG@10	MRR
Abstract	0.521	0.468	0.746
Introduction	0.523	0.472	0.747
Conclusion	0.527	0.476	0.750
CC ($N_{\mathrm{cs}} = 1$)	0.518	0.472	0.738
5 keywords	0.515	0.467	0.727
Full text	0.548	0.514	0.768
Full text + abstract	0.552	0.519	0.772
Full text + introduction	0.555	0.525	0.774
Full text + conclusion	**0.560**	**0.530**	**0.775**
Full text + CC ($N_{\mathrm{cs}} = 1$)	0.552	0.524	0.771
Full text + 5 keywords	0.551	0.521	0.772
pc-SIM ($n = 4$, $N_{\mathrm{pc}} = 5$)	nDCG@5	nDCG@10	MRR
Abstract	0.525	0.478	0.747
Introduction	0.529	0.485	0.746
Conclusion	0.531	0.490	0.760
CC ($N_{\mathrm{cs}} = 1$)	0.522	0.475	0.740
5 keywords	0.519	0.470	0.735
Full text	0.563	0.558	0.771
Full text + abstract	0.562	0.557	0.774
Full text + introduction	0.565	0.562	0.773
Full text + conclusion	**0.571**	**0.568**	**0.777**
Full text + CC ($N_{\mathrm{cs}} = 1$)	0.562	0.554	0.773
Full text + 5 keywords	0.560	0.555	0.774
pc-IMP ($n = 4$, $N_{\mathrm{pc}} = 6$)	nDCG@5	nDCG@10	MRR
Abstract	0.535	0.498	0.757
Introduction	0.540	0.507	0.756
Conclusion	0.545	0.514	0.764
CC ($N_{\mathrm{cs}} = 1$)	0.536	0.512	0.754
5 keywords	0.533	0.508	0.745
Full text	0.568	0.562	0.778
Full text + abstract	0.571	0.566	0.788
Full text + introduction	0.569	0.563	0.787
Full text + conclusion	**0.576**	**0.572**	**0.790**
Full text + CC ($N_{\mathrm{cs}} = 1$)	0.570	0.564	0.789
Full text + 5 keywords	0.568	0.565	0.787
pc-IMP (adp) (n: adaptive, $CL_{\mathrm{th}} = 0.56$, $N_{\mathrm{pc}} = 8$)	nDCG@5	nDCG@10	MRR
Abstract	0.536	0.503	0.759
Introduction	0.545	0.512	0.761
Conclusion	0.556	0.520	0.772
CC ($N_{\mathrm{cs}} = 1$)	0.537	0.515	0.756
5 keywords	0.536	0.509	0.747
Full text	0.576	0.570	0.781
Full text + abstract	0.576	0.573	0.790
Full text + introduction	0.580	0.573	0.793
Full text + conclusion	**0.587**	**0.582**	**0.798**
Full text + CC ($N_{\mathrm{cs}} = 1$)	0.577	0.577	0.790
Full text + 5 keywords	0.574	0.573	0.788

Table 4 Tuning results: recommendation accuracy in [TUNE:frg-TW] obtained by modeling candidate papers to recommend using "(a) only fragments" and "(b) both full text and fragments" in citation and potential citation papers discovered by [pc-BIN]

(a) Only fragments pc-BIN ($n = 4$, $N_{pc} = 5$)	nDCG@5	nDCG@10	MRR
Abstract ($\alpha = 0.8$)	0.538	0.517	0.760
Introduction ($\alpha = 0.7$)	0.540	0.521	0.761
Conclusion ($\alpha = 0.7$)	**0.548**	**0.528**	**0.765**
CC ($N_{cs} = 1$) ($\alpha = 0.9$)	0.533	0.510	0.758
5 keywords ($\alpha = 0.9$)	0.529	0.507	0.756
Baseline system [25] (Weight "SIM", Th $= 0.4$, $\gamma = 0.23$, $d = 3$)	0.521	0.489	0.750
(b) Both full text and fragments pc-BIN ($n = 4$, $N_{pc} = 5$)	nDCG@5	nDCG@10	MRR
Full text + abstract ($\alpha = 0.5$)	0.549	0.522	0.768
Full text + introduction ($\alpha = 0.4$)	0.555	0.526	0.773
Full text + conclusion ($\alpha = 0.4$)	**0.560**	**0.533**	**0.779**
Full text + CC ($N_{cs} = 1$) ($\alpha = 0.6$)	0.545	0.518	0.767
Full text + 5 keywords ($\alpha = 0.6$)	0.542	0.515	0.765
Baseline [25] (Weight "SIM", Th $= 0.4$, $\gamma = 0.23$, $d = 3$)	0.521	0.489	0.750

According to the experimental results in [TUNE:frg-SIM] (Table 3), we obtain the best recommendation accuracy ("Full text + Conclusion": nDCG@5 of 0.587, nDCG@10 of 0.582, MRR of 0.798) in [pc-IMP (adp)]. We again observe that [pc-IMP (adp)] generally gives better results compared with [pc-BIN], [pc-SIM], and [pc-IMP]. Across the board, we see that fragments by themselves perform less well, but that they have a meaningful positive effect if used in conjunction with the full text. The length of the individual fragments may be an important consideration; as keywords and citation sentences are generally quite short. The conclusion may also serve as a factual summary of a paper, which omits introductory or motivating material common in abstracts and introduction fragments.

Table 4, Figs. 7 and 8 show experimental results in [TUNE:frg-TW]. In this approach, we obtain the best recommendation accuracy ("$\alpha = 0.4$, Full text + Conclusion": nDCG@5 of 0.589, nDCG@10 of 0.596, MRR of 0.798) when we employ [pc-IMP (adp)] (see Fig. 8g–i). When we employ "only fragments," according to Table 4a and Fig. 7, we observe that "Conclusion" gives better results than other fragments. Other fragments also yield better results when appropriately tuned. For example, "citation context (CC ($N_{cs} = 1$))" in nDCG@5 gives the recommendation accuracy of 0.533, 0.546, 0.552, and 0.555 at $\alpha = 0.9$ in [pc-BIN] (Table 4a), [pc-SIM] (Fig. 7a), [pc-IMP] (Fig. 7d), and [pc-IMP (adp)] (Fig. 7g), respectively. The same trends are also observed in nDCG@10 (Fig. 7b, e, h) and MRR (Fig. 7c, f, i). However, recommendation accuracy obtained by fragments "Abstract", "Introduction", "Citation Context (CC)", and "5 keywords" generally underperform "Conclusion". The same trend is also observed with nDCG@10 (Fig. 7b, e, h) and MRR (Fig. 7c, f, i).

On the other hand, when we employ "both full text and fragments" (Table 4b; Fig. 8), the recommendation accuracy generally outperforms that obtained using "only fragments." For example, "Full text + Conclusion" in nDCG@5 gives the best recommendation accuracy of 0.560, 0.575, 0.578, and 0.589 at $\alpha = 0.4$ in [pc-BIN] (Table 4b), [pc-SIM] (Fig. 8a), [pc-IMP] (Fig. 8d), and [pc-IMP (adp)] (Fig. 8g), respectively. As well as experimental results obtained using "only fragments" described above, other fragments do not give the better recommendation accuracy.

According to Table 4, and Figs. 7 and 8, we make a few observations concerning the value of α: from Table 4, we observe that the value of α that gives the best recommendation accuracy is different between the approach that uses "(a) only the fragment" ($\alpha = [0.7 - 0.9]$) and the approach that uses "(b) both full text and fragments" ($\alpha = [0.4 - 0.6]$). This indicates that in order to characterize candidate papers better, fragments (which contain relatively less text) need to be given larger weights compared to when they are used in conjunction with the full text. Interestingly, in Figs. 7 and 8, we see the same trends graphically, but further observe that fixing a particular α value leads to different fragments being more important: in Fig. 8, an α value of 0.6 yields "Citation Contexts (CC)" as most useful, 0.5 yields the "Abstract" as being most useful and 0.4 yields the "Conclusion" as best. This indicates that the tunable weight α is an important factor to set if optimal results are desired. We do note that the "Conclusion" fragment's performance is noticeably better and hence more stable than other fragments' performance levels, so we recommend this setting.

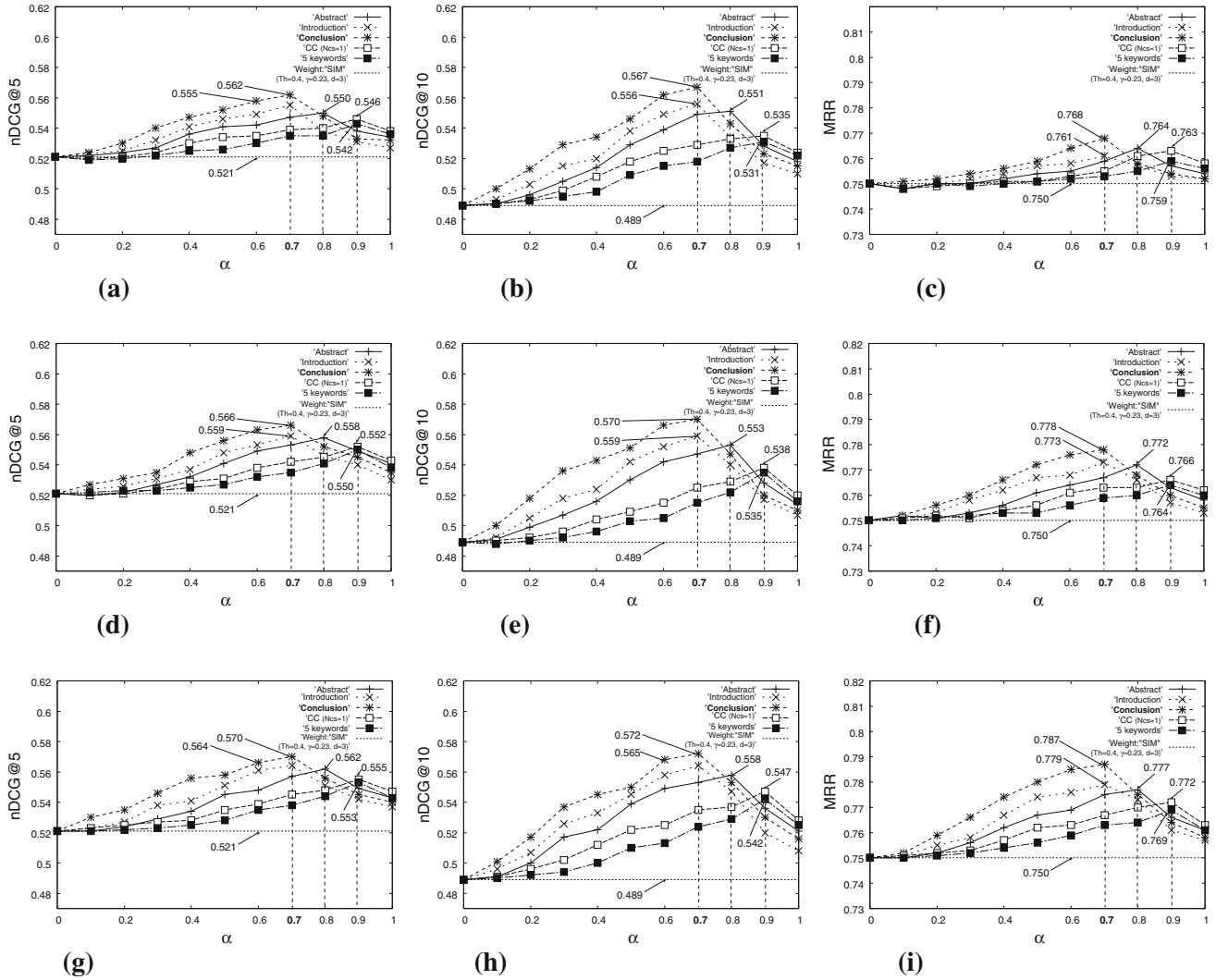


Fig. 7 Tuning (potential) citation papers' text weight: recommendation accuracy in [TUNE:frg-TW] obtained by modeling candidate papers using "only fragments" in (potential) citation papers discovered by [pc-SIM] (**a**–**c**), [pc-IMP] (**d**–**f**), and [pc-IMP (adp)] (**g**–**i**). **a** nDCG@5 [pc-SIM], **b** nDCG@10 [pc-SIM], **c** MRR [pc-SIM], **d** nDCG@5 [pc-IMP], **e** nDCG@10 [pc-IMP], **f** MRR [pc-IMP], **g** nDCG@5 [pc-IMP (adp)], **h** nDCG@10 [pc-IMP (adp)], **i** MRR [pc-IMP (adp)]

Table 5 shows recommendation accuracy obtained by applying the optimal parameters and selection of fragments to the test set. In the test set, we observe the same trends as in training: [pc-IMP (adp)] outperforms [pc-BIN], [pc-SIM] and [pc-IMP]. In particular, [pc-IMP (adp)] ($CL_{\text{th}} = 0.56$, $N_{\text{pc}} = 8$) with [frg-TW] ($\alpha = 0.4$, "Full text + Conclusion") gives the best recommendation accuracy, similar to the best results in the training settings. This shows that our dataset keeps a useful balance between the training and test sets.

Furthermore, we observe that our baseline system [25] outperforms others ([18], [31], and [1]) and our approach proposed in this paper ([pc-IMP (adp)] + [frg-TW]) gives the best recommendation accuracy.

In Nascimento et al. [18], 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's work [31], a binary-valued user–paper matrix—similar to Fig. 2a—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

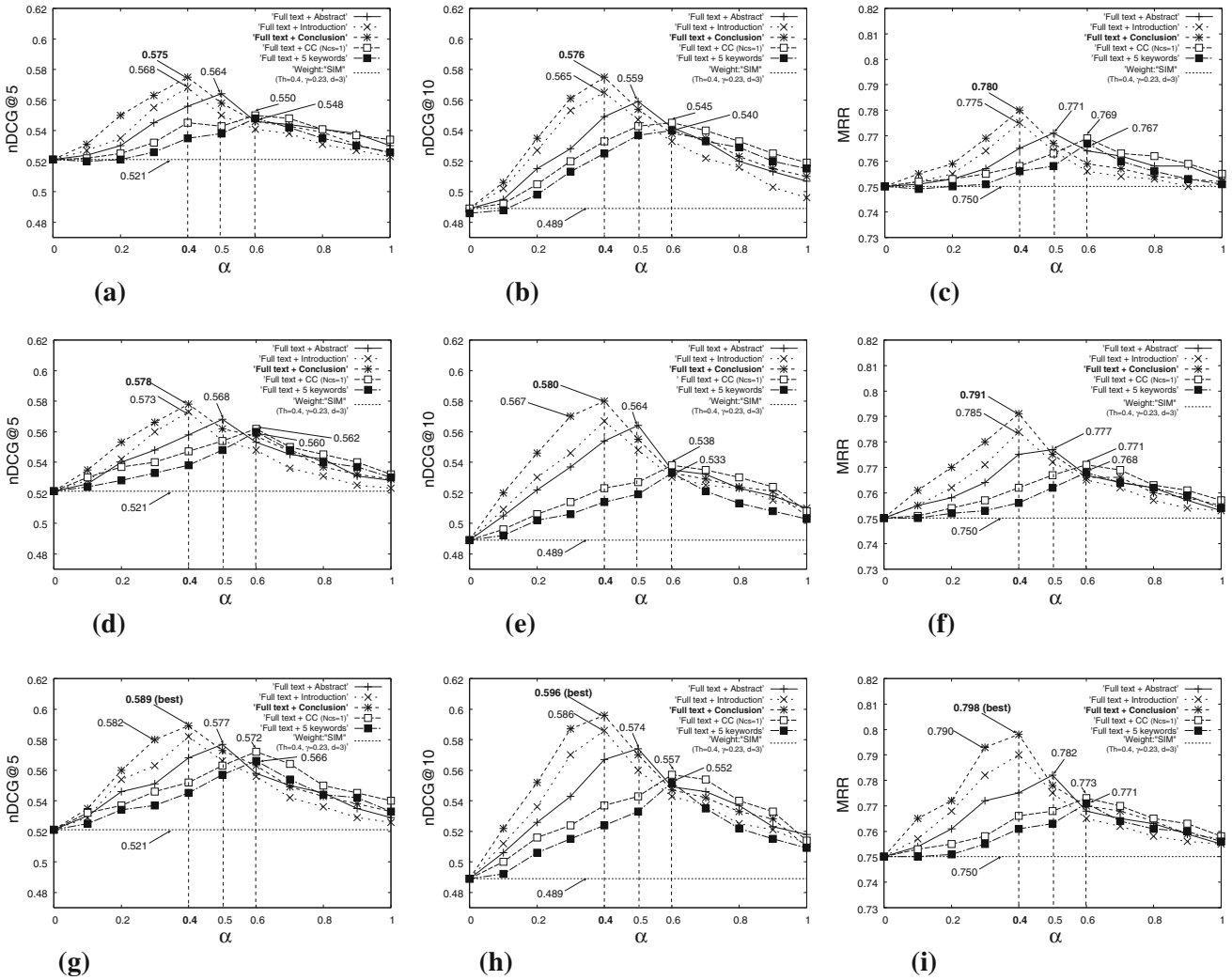


Fig. 8 Tuning (potential) citation papers' text weight: recommendation accuracy in [TUNE:frg-TW] obtained by modeling candidate papers using "both full text and fragments" in (potential) citation papers discovered by [pc-SIM] (**a**–**c**), [pc-IMP] (**d**–**f**), and [pc-IMP (adp)] (**g**–**i**). **a** nDCG@5 [pc-SIM], **b** nDCG@10 [pc-SIM], **c** MRR [pc-SIM], **d** nDCG@5 [pc-IMP], **e** nDCG@10 [pc-IMP], **f** MRR [pc-IMP], **g** nDCG@5 [pc-IMP (adp)], **h** nDCG@10 [pc-IMP (adp)], **i** MRR [pc-IMP (adp)]

more relevant recommendation. Algarni et al.'s approach [1] gives the second highest recommendation accuracy among the other comparative approaches. This implies that implicit feedback based on frequent pattern mining is one of effective methods in constructing a user profile.

Microscopic analyses

Diving into individual results, in [pc-IMP], we observe that one of the researchers who works on computer graphics, received relevant recommendations in our optimized system, where the first relevant recommendation was at the first rank. In the baseline system [25], he could not obtain any relevant recommendation in the top-10 results, and the first relevant result was ranked 63rd. Another researcher, who works on mobile computing, also was provided a relevant recommendation in the first rank, while his most relevant recommendation is ranked 52nd using the baseline.

Inspecting the topics of the papers identified as potential citations also helps to further understand the reach of our algorithm. We observed that our approach discovers relevant potential citation papers to characterize candidate papers to recommend. For example, when the topic of candidate paper to recommend is "access control of business documents based on role mining," which is often cited by security papers, our approach identifies papers about data mining in databases and papers about information theory related to security as potential citation papers. In another example, given the candidate paper topic of "real world gesture analysis", often cited in human–computer interaction, our imputation approach discovers potential citation papers whose

Table 5 Recommendation accuracy obtained by applying optimal parameters and fragments to the test set

	nDCG@5	nDCG@10	MRR
pc-BIN ($n = 4$, $N_{pc} = 5$)			
frg-SIM (full text + conclusion)	0.558*	0.536*	0.767*
frg-TW ($\alpha = 0.4$, full text + conclusion)	0.562*	0.541*	0.774*
pc-SIM ($n = 4$, $N_{pc} = 5$)			
frg-SIM (full text + conclusion)	0.569*	0.566**	0.772*
frg-TW($\alpha = 0.4$, full text + conclusion)	0.573**	0.571**	0.783*
pc-IMP ($n = 4$, $N_{pc} = 6$) [27]			
frg-SIM (full text + conclusion)	0.574**	0.570**	0.787*
frg-TW ($\alpha = 0.4$, full text + conclusion)—(A)	***0.581*****	***0.577*****	***0.795****
pc-IMP (adp) (n:adaptive, $CL_{th} = 0.56$, $N_{pc} = 8$)			
frg-SIM (full text + conclusion)	0.586†	0.588†	0.797†
frg-TW ($\alpha = 0.4$, full text + conclusion)—(B)	**0.588**†	**0.598**†	**0.804**†
Baseline system [25] (Weight "SIM", Th = 0.4, $\gamma = 0.23$, $d = 3$)	0.527	0.482	0.752
Nascimento et al. [18] ("Frequency of bi-gram" obtained from title and abstract)	0.335	0.311	0.437
Wang and Blei [31] ("in-matrix prediction" in collaborative topic regression)	0.396	0.374	0.498
Algarni et al. [1] (4 times feedback)	0.460	0.433	0.630

** and * denote the difference between the best results in the baseline system [25] (underlined scores) and the each result in [27] 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 [27] (italic scores in [pc-IMP] + [frg-TW]) is significant for $p < 0.05$

Table 6 Recommendation accuracy for intra-disciplinary researchers obtained by applying optimal parameters and fragments to the test set

	nDCG@5	nDCG@10	MRR
pc-IMP ($n = 4$, $N_{pc} = 6$) [27]			
frg-TW ($\alpha = 0.4$, full text + conclusion) [(A) in Table 5]	0.581	0.577	0.795
Intra-disciplinary researchers	*0.576*	*0.571*	*0.789*
pc-IMP (adp) (n:adaptive, $CL_{th} = 0.56$, $N_{pc} = 8$)			
frg-TW ($\alpha = 0.4$, full text + conclusion) [(B) in Table 5]	0.588	0.598	0.804
Intra-disciplinary researchers	**0.590*** (+0.014)	**0.604*** (+0.033)	**0.806*** (+0.017)

* denotes the difference between recommendation accuracy for intra-disciplinary researchers in this journal paper and [27] (italic scores) is significant for $p < 0.05$

topics are biomechanics, computer-based music conducting systems, and machine learning. These examples indicate how our approach can characterize papers better than the baseline. This is then reflected in more relevant recommendations and higher accuracy.

As described at (A3) in Sect. 3.2, a shortcoming of the [pc-IMP] method is that the imputation approach tends to find only potential citation papers relating to a single discipline when the topic of the target paper is intra-disciplinary. We designed the [pc-IMP (adp)] adaptive neighborhood approach to overcome this problem. As shown in Fig. 5, 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. In addition, as shown in Fig. 6, the number of potential citation papers (N_{pc}) that gives the best recommendation accuracy is 8, that is larger a little bit compared with the optimal one in [pc-IMP] ($N_{pc} = 6$). This indicates that more relevant potential citation papers can characterize the target paper, that is suitable for intra-disciplinary papers with some topics. In the same example of "understanding mobile user's behavior patterns" above, when we employ [pc-IMP (adp)], the details of eight potential citation papers are "mobile technology" (3 papers), "user search behavior" (3 papers), and "clustering" papers (2 papers). This indicates that the "understanding mobile user's behavior patterns" paper is more faithfully modeled by the identified potential citation papers.

This improved modeling provides better recommendations for intra-disciplinary researchers. As shown in

Table 1c, our dataset has 15 researchers, conducting intra-disciplinary research. In addition, as shown in Table 6, [pc-IMP (adp)] + [frg-TW] gives statistically significant recommendation accuracy compared with [pc-IMP] + [frg-TW] (nDCG@5 of 0.590, nDCG@10 of 0.604, and MRR of 0.806). Among the 15 researchers, the recommendation accuracies for the two researchers are significantly improved, which we review now in depth.

A researcher who works on software engineering (R43), focusing on software documentation and readability, receives relevant recommendations at the top as well as more relevant recommendations in the top-10 results when we employ [pc-IMP (adp)] + [frg-TW]. Our [pc-IMP] + [frg-TW] only identifies potential citation papers about software engineering only, the extended approach, [pc-IMP (adp)] + [frg-TW] can identify papers about text analysis, human factors, and machine learning as well as software engineering.

The other researcher, working on user behavior models in electronic commerce (R49), is also one of intra-disciplinary researchers whose recommendations were improved significantly. [pc-IMP] + [frg-TW] only identifies papers about econometrics and electronic commerce, whereas [pc-IMP (adp)] + [frg-TW] additionally identified papers about machine learning and user behavior mining.

These examples indicates the effectiveness of [pc-IMP (adp)] + [frg-TW] approach that can characterize intra-disciplinary papers much better, resulting in better recommendation for intra-disciplinary researchers.

Finally, [pc-IMP] + [frg-TW] is statistically significant compared with the baseline system [25], improving recommendation accuracy by almost 10 % when measured by nDCG@5 and nDCG@10 ($p < 0.01$), respectively, and by over 6 % when measured by MRR ($p < 0.05$). Additionally, we also observe that the difference between [pc-IMP (adp)] + [frg-TW] and [pc-IMP] + [frg-TW] is also statistically significant with improvement by 1.2 %, 3.6 %, and 0.9 % when measured by nDCG@5, nDCG@10, and MRR ($p < 0.05$ for each), respectively. This shows that relevant papers are recommended at the top for more researchers and more relevant papers are recommended in the top-10. We believe that our proposed approach is effective in characterizing candidate papers to recommend to obtain much higher recommendation accuracy.

5 Conclusion

We have explored two significant approaches to improve the state-of-the-art in scholarly paper recommendation. In particular, we examine (1) how to alleviate data sparsity using collaborative filtering to find potential citation papers, and (2) how to refine the use of citing papers in characterizing a target candidate paper using fragments in the citation and potential citation papers.

Our results show that, in the discovery of potential citation papers, imputation-based CF—especially when done using an adaptive selection of neighborhoods—is more effective than CF with binary or similarity values. Additionally, the potential citation paper approach, when appropriately tuned, can improve recommendation accuracy significantly. Especially, when we model candidate papers to recommend using "full text and conclusion" in both citation and potential citation papers, we achieve the best accuracy and outperform state-of-the-art scholarly paper recommendation systems.

The novel extension of our work here is to address intra-disciplinarity: where researchers in a discipline work on distinct subfields of a discipline. Through appropriate adaptive cluster selection, we overcome the limitation of the baseline model of assigning works to a single sub-discipline. Through both macro- and micro-level analyses, we demonstrate that our approach more faithfully models such researchers, improving overall recommendation quality. An important limitation to note is that our study is limited to a single discipline; in future work, it will be important to demonstrate—when an appropriate unencumbered scholarly paper and recommendation dataset is available—that these methods also hold for addressing multi-disciplinary scholars.

We believe that our approach can be applied more generally. The notion of enriching a network with potential items can be applied to any network that feature asymmetric directional links, such as social networks, patent documents and email dialogues. The notion of using potential citation papers can be applied wherever textual evidence is associated with the links, such as in patent documents and customer testimonials.

References

1. Algarni, A., Li, Y., Xu, Y.: Selected new training documents to update user profile. In: Proceedings of the 19th International Conference on Information and Knowledge Management (CIKM'10), pp. 799–808 (2010)
2. Bethard, S., Jurafsky, D.: Who should I cite? Learning literature search models from citation behavior. In: Proceedings of the 19th International Conference on Information and Knowledge Management (CIKM'10), pp. 609–618 (2010)
3. Caragea, C., Silvescu, A., Mitra, P., Giles, C.L.: Can't see the forest for the trees? A citation recommendation system. In: Proceedings of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL '13), pp. 111–114 (2013)
4. El-Arini, K., Guestrin, C.: Beyond keyword search: discovering relevant scientific literature. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11), pp. 439–447 (2011)
5. Gori, M., Pucci, A.: Research paper recommender systems: a random-walk based approach. In: Proceedings of the 2006

IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006), pp. 778–781 (2006)
6. He, Q., Kifer, D., Pei, J., Mitra, P., Giles, C.L.: Citation recommendation without author supervision. In: Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM'11), pp. 15–24 (2011)
7. He, Q., Pei, J., Kifer, D., Mitra, P., Giles, C.L.: Context-aware citation recommendation. In: Proceedings of the 19th International World Wide Web Conference (WWW2010), pp. 421–430 (2010)
8. Herlocker, J., Konstan, J., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), pp. 230–237 (1999)
9. Huang, W., Kataria, S., Karagea, C., Mitra, P., Giles, C.L., Rokach, L.: Recommending citations: translating papers into references. In: Proceedings of the 21st International Conference on Information and Knowledge Management (CIKM'12), pp. 1910–1914 (2012)
10. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), pp. 41–48 (2000)
11. Jarvis, R.A., Patrick, E.A.: Clustering using a similarity measure based on shared near neighbors. IEEE Trans. Comput. **C22**(11), 1025–1034 (1973)
12. Kaptein, R., Serdyukov, P., Kamps, J.: Linking wikipedia to the web. In: Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10), pp. 839–840 (2010)
13. Katz, L.: A new status index derived from sociometric analysis. Psychometrika **18**(1), 39–43 (1953)
14. Lu, Y., He, J., Shan, D., Yan, H.: Recommending citations with translation model. In: Proceedings of the 20th International Conference on Information and Knowledge Management (CIKM'11), pp. 2017–2020 (2011)
15. McNee, S.M., Albert, I., Cosley, D., P. Gopalkrishnan, S.L., Rashid, A.M., Konstan, J.S., Riedl, J.: On the recommending of citations for research papers. In: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work (CSCW '02), pp. 116–125 (2002)
16. Mei, Q., Zhai, C.: Generating impact-based summaries for scientific literature. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-08: HLT), pp. 816–824 (2008)
17. Milne, D., Witten, I.H.: Learning to link Wikipedia. In: Proceedings of the 17th International Conference on Information and Knowledge Management (CIKM'08), pp. 509–518 (2008)
18. Nascimento, C., Laender, A.H.F., da Silva, A.S., Gonçalves, M.A.: A source independent framework for research paper recommendation. In: Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011), pp. 297–306 (2011)
19. Nomoto, T.: Two-tier similarity model for story link detection. In: Proceedings of the 19th International Conference on Information and Knowledge Management (CIKM'10), pp. 789–798 (2010)
20. Oh, S., Lei, Z., Lee, W.C., Mitra, P., Yen, J.: CV-PCR: a context-guided value-driven framework for patent citation recommendation. In: Proceedings of the 22nd International Conference on Information and Knowledge Management (CIKM'13), pp. 2291–2296 (2013)
21. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. In: Technical Report, SIDL-WP-1999-0120, Stanford Digital Library Technologies Project (1998)
22. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
23. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling'08), pp. 689–696 (2008)
24. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, London (1983)
25. Sugiyama, K., Kan, M.Y.: Scholarly paper recommendation via user's recent research interests. In: Proceedings of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL '10), pp. 29–38 (2010)
26. Sugiyama, K., Kan, M.Y.: Serendipitous recommendation for scholarly papers considering relations among researchers. In: Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL '11), pp. 307–310 (2011)
27. Sugiyama, K., Kan, M.Y.: Exploiting potential citation papers in scholarly paper recommendation. In: Proceedings of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL '13), pp. 153–162 (2013)
28. Strohman, T., Croft, W. B., Jensen, D.: Recommending citations for academic papers. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), pp. 705–706 (2007)
29. Torres, R., McNee, S.M., Abel, M., Konstan, J.A., Riedl, J.: Enhancing digital libraries with TechLens. In: Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004), pp. 228–236 (2004)
30. Voorhees, E.M.: The TREC-8 question answering track report. In: Proceedings of the 8th Text REtrieval Conference (TREC-8), pp. 77–82 (1999)
31. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11), pp. 448–456 (2011)
32. West, R., Precup, D., Pineau, J.: Completing Wikipedia's hyperlink structure through dimensionality reduction. In: Proceedings of the 18th International Conference on Information and Knowledge Management (CIKM'09), pp. 1097–1106 (2009)
33. West, R., Precup, D., Pineau, J.: Automatically suggesting topics for augmenting text documents. In: Proceedings of the 19th International Conference on Information and Knowledge Management (CIKM'10), pp. 929–938 (2010)
34. Yang, D., Wei, B., Wu, J., Zhang, Y., Zhang, L.: CARES: A ranking-oriented CADAL recommender system. In: Proceedings of the 9th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2009), pp. 203–211 (2009)