Design and Implementation of Scientific Article Recommendation on the Basis of Common Author Relation and Historical Preferences: A Review

Nitishkumar A. Dubey[1], Dr.P.M Jawandhiya[2].

M.E.CSE PLIT [1] *,PRINCIPAL PLIT* [2]
Email:nitishdubey5@gmail .com [1] *, principal@plit.ac.*

Abstract- Scientific article recommender systems are playing an increasingly important role for researchers in retrieving scientific articles of interest in the coming era of big scholarly data. Most existing studies have designed unified methods for all target researchers and hence the same algorithms are run to generate recommendations for all researchers no matter which situations they are in. However, different researchers may have their own features and there might be corresponding methods for them resulting in better recommendations. In this paper, we propose a novel recommendation method which incorporates information on common author relations between articles (i.e., two articles with the same author(s)). The rationale underlying our method is that researchers often search articles published by the same author(s). Since not all researchers have such author-based search patterns, we present two features, which are defined based on information about pairwise articles with common author relations and frequently appeared authors, to determine target researchers for recommendation. Extensive experiments we performed on a real-world dataset demonstrate that the defined features are effective to determine relevant target researchers and the proposed method generates more accurate recommendations for relevant researchers when compared to a Baseline method.

1. INTRODUCTION

With the rapid emergence of big scholarly data tremendous growth of knowledge is now largely captured in digital form and archived all over the world. Archival materials are also currently being digitized and provided online to people for free or by paying a fee. Such situation creates the commonly known information overload problem especially in academia while bringing a significant advantage that allows people to easily access more knowledge. For example, a researcher in academia needs to find articles of interest to read for generating a research idea or citing an article related to the article he is writing, an author needs to submit his manuscript to a certain journal of which the topic is relevant to the manuscript, an editor needs to assign a manuscript to a reviewer who is an expert in the domain which the manuscript belongs to, or a researcher in a domain needs to collaborative with another researcher in another domain. These academic activities involve in an overwhelming number of articles, journals, reviewers, and researchers.

Therefore, it is quite difficult for researchers to locate relevant articles, journals, reviewers, and researchers for the aforementioned purposes. Academic recommender systems aim to solve the information overload problem in big scholarly data such as finding relevant research paper, relevant publication venue, etc. Fig. 1 shows the corresponding recommendation tasks in above-mentioned scenarios, including (i) article recommendation [1], [2], [3], [4] for suggesting relevant articles to a researcher or an article for the purposes of reading or citation, (ii) reviewer recommendation [5], [6] for assigning a manuscript to the most appropriate reviewers (e.g., an expert in the same domain), (iii) venue recommendation [7], [8] for suggesting a topic-relevant conference or journal to publish a new article, and (iv) collaboration recommendation [9], [10] for suggesting new partners to execute joint research (e.g., exploring cross-domain solution). There exist some interesting studies on these recommendation tasks. Gori and Pucci [1] built a citation relation graph and employee randomwalk algorithmto compute rankingscores of each possible citation at all. [6] assigned relevant weights to various factors which affect the expertise of the reviewer to create a fuzzy set and then

compute the expertise. Yang and Davison [7] extracted features related to writing-style information for computing similarity between articles and then applied traditional collaborative filtering to recommend a venue for submission. Xia et al. [10] considered three academic factors (i.e., co-author order, collaboration time, and number of collaboration) to define link importance, and then employed a random walk algorithm to compute rankings of potential collaborators. In this paper, we focus on article-researcher recommendation, i.e., studying how to find articles of interest for target researchers in the context of big scholarly data. In the print age, researchers found articles of interest with the help of library catalogs. In recent years, web search tools employed by scientific digital libraries like IEEEXplore, and literature search engines like Google Scholar, can retrieve a list of relevant articles in diverse technological fields using keyword-based queries. However, these search tools have several drawbacks as follows: (i) It is not enough to describe searchers' needs depending on only several limited keywords; (ii) The obtained results are the same for all searchers if only the keywords are the same; (iii) It is not feasible to search articles when a searcher has no ideas of what they are looking for. Article-researcher recommender systems aim to automatically suggest personalized articles of potential interest for different targets ,there by overcoming the problems stated above. Existing studies [3], [11] generally compute the content similarity between articles to find articles which are similar to the target's articles of interest, or compute the similarity between the target's profile and an new article's content to find matches. However, content extraction is not such simple because an article includes too many words. In this paper, we extract only author information to build relations between articles, i.e., common author relations. Then, these relations and researchers' historical preferences are used together to build a heterogenious graph for

article ranking. The rationale of incorporating common author relations is that, the continuous development of internet technology enables researchers to easily build personal websites and share publications with others, which makes it more convenient to search articles published by the same author(s) for researchers who have such a search preference based on authors (we call it author-based search pattern). In addition, most studies ignore the fact that there exist different recommendation methods suitable for different targets. Therefore, we define features to find relevant target researchers who have author-based search patterns by analyzing information on common author relations existing in a researcher's historical preferences. In summary, we propose a novel Common Author relation-based recommendation method (CARE) for specific target researchers with author-based search patterns. Our main contributions in this work can be outlined as follows: •We present two features including the ratio of pairwise articles with common author relations and the ratio of the most frequently appeared author, to help determine relevant researchers with author-based search patterns. • We propose a novel recommendation method, which incorporates common author relations between articles to help generate better recommendations for relevant target researchers. • We conduct relevant experiments using a real-world dataset Cite ULike to evaluate the impact soft he defined features and the performance of the proposed method. In addition, other two features have also been defined and proved to be not effective for determining suitable targets. The rest of the paper is organized as follows. Section 2 reviews related work on article recommendation. Section 3 presents our problem definition. Section 4 introduces the details of our proposed method. Section 5 describes our experimental setup and discusses our results in detail. Section 6 finally concludes the paper.

2 RELATED WORK

Recommender systems aim to automatically suggest items of potential interest to users. As well-known effective tools

for solving information overload problems, recommender systems have been successfully applied in multiple domains including traffic [12], movies [13], [14], music [15], [16], news [17], e-commerce [18], [19], e-learning [20], [21], and so on. As aforementioned, with the rapid development of

information technology and ever-growing amounts of scholarly data, it is becoming increasingly popular and challenging to apply recommendation techniques in academia. In this section, we focus on reviewing related work on article recommendation.

2.1 Article-Article Recommendation

Article-article recommendation, i.e., citation recommendation, includes global citation recommendation [1], [22], [23], [24], [25], [26], [27] and local citation recommendation [2], [28], [29], [30], [31]. Global citation recommendation aims to recommend a list of citations for a given query article. Strohman et al. [22] linearly combined text features and citation graph features to measure the relevance between articles. They conducted relevant experiments with their proposed citation recommender system and concluded that similarity between bibliographies of articles and Katz distance are the most important features. Gori and Pucci [1] used citation relations between articles to built a citation graph and applied a random walk algorithm in the graph to compute ranking scores of each article as a reference of a target article. Bethard and Jurafsky [23] incorporated a wide variety of features (including author impact, author citation habits, citation count, and publication ages) to build a retrieval model for literature search. After a training process, the model took abstract of an article as input to produce relevant reference lists. Nallapati et al. [24] jointly modeled the text and citation relationship under a framework of topic model. They introduced a model Pair-Link-LDA which models the presence or absence of a link between pairwise articles and does not scale to large digital libraries. They also introduced another model called Link-PLSA-LDA which models citations as a sample from a probability distribution associated with a topic. Meng et al. [25] incorporated various types of information including content, authorship, and collaboration network to build a unified graph-based model for personal global citation recommendation. Ren et al. [26] proposed to cluster article citations into interest groups to determine the significance of different structural relevance features for each group while deriving an article's relative authority within each group. Liu et al. [27] employed the pseudo relevance feedback (PRF) algorithm to determine important nodes like authors and venues on a heterogeneous bibliographic graph. Then, a random walk algorithm was run to compute the ranking scores of an article.

On the other hand, local citation recommendation aims to recommend citations for a given specific context such as a sentence in a paper. Tang and Zhang [2] formally defined the problem of topic-based citation recommendation and proposed to model article contents and citation relationships using a two-layer restricted Boltzmann machine. For a given context, they calculated the probability of each article being the reference based on the model. Lu et al. [28] proposed to recommend citations using a translation model which is originally used in

translating text in one language to another. They assumed that the languages used in citation contexts and article's content are different and translated one word in context to one word in citation. Based on the probability of translating one word to another, relevant articles were recommended to a citation context. Huang et al. [29] regarded an article as new 'words' in another language and employed a translation model for estimating the probability of citing an article given a citation context. Tang et al. [30] proposed a cross-language context-aware citation recommendation method for the purpose of recommending English citations for a given context of the place where a citation should be made in a Chinese article. Huang et al. [31] proposed a novel neural probabilistic model which jointly learns the semantic representations of citation contexts and cited articles and then estimated the probability of citing an article by a neural network.

2.2 Article-Researcher Recommendation

Article-researcher recommendation is our focus in this paper. Most existing studies compute similarities among researchers and articles based on articles' contents [3], [11], [32], [33], [34], [35], [36], [37], [38] or tags in social tagging systems [3], [37], [38], [39] and then apply traditional collaborative filtering to generate recommendations. Sugiyama and Kan [11] examined the effect of modeling a researcher's past works in scientific article recommendation to the researcher. A researcher's profile was derived from his past works and other works which are the references or citations of those works. Apart from previous explicit citations, Sugiyama and Kan [32] additionally took into account implicit citations. Potential citation articles were discovered using collaborative filtering and then combined with previous information to enhance the profiles of candidate articles and researchers. Finally, the two types of profiles were compared to compute similarity as their cosine measure. Wang and Blei [33] proposed a collaborative topic regression model for article recommendation, where each user was represented with interest's distribution and each article was described using content-based item topic distribution. Nascimento et al. [34] proposed a framework to generate structured search queries and obtained candidate articles using existing web information sources. Then, they computed the contentbased similarity for ranking candidate articles. Jiang et al. [35] employed a concept-based topic model to compute the problem-based similarity and solution-based similarity between a known article of interest and an unknown article. Tian and Jing [36] employed LDA (Latent Dirichlet Allocation) model to obtain each article's

representation based on content and computed the similarity between articles for determining their associations. Sun et al. [3] exploited semantic content and heterogeneous connections (i.e., social connection, behavioral connection, and semantic connection) to build two kinds of profiles of researchers and then computed researcher-article similarities and researcherresearcher similarities. Using a Social-Union method [40], the score of a target researcher on an article is defined by that of his nearest neighbors. Finally, it is fused with the results obtained based on researcher-article similarity computation to compute the final ranking scores. Pera and Ng [37], [38] proposed a personalized recommender for scientific articles, called PReSA. Given a target publication *P* which a user *U* expressed interest in, PReSA computes the three similarities between *P* and each candidate publication *CP* which is from the ones in the personal libraries of *U*'s connections:

a) tag similarity, b) title similarity, and c) abstract similarity. Then, these similarities and popularity of publications were fused to calculate the ranking score

of *CP* by employing a weight linear combination strategy. Finally, top-N publications were recommended to *U*. Xia et al. [39] built a active participant's profile and each group's profile based on tags annotated by participants. Then, their similarity was computed to recommend the active participant's to other participants in groups with higher similarities.

In this paper, we utilize only information on articles' authors to build common author relations between articles. Compared to analysis on content and tags, our work is simpler and more time-saving, because the number of terms in content and tags is enormous and there exist lots of irrelevant terms. In addition, these studies do not take into account specific target researchers suitable for their recommendation methods. We assume that, since our proposed method (CARE) incorporates common author relations, only a part of researchers can be selected as targets for highquality recommendation. Accordingly, we define features to determine such researchers.

3   PROBLEM STATEMENT

In academic social tagging websites such as CiteULike, each of the registered users is generally a researcher. When a researcher is

interested in an article, he will post it into his article library, read it extensively and then tag the article with one or more special keywords. As shown in Fig. 2, five
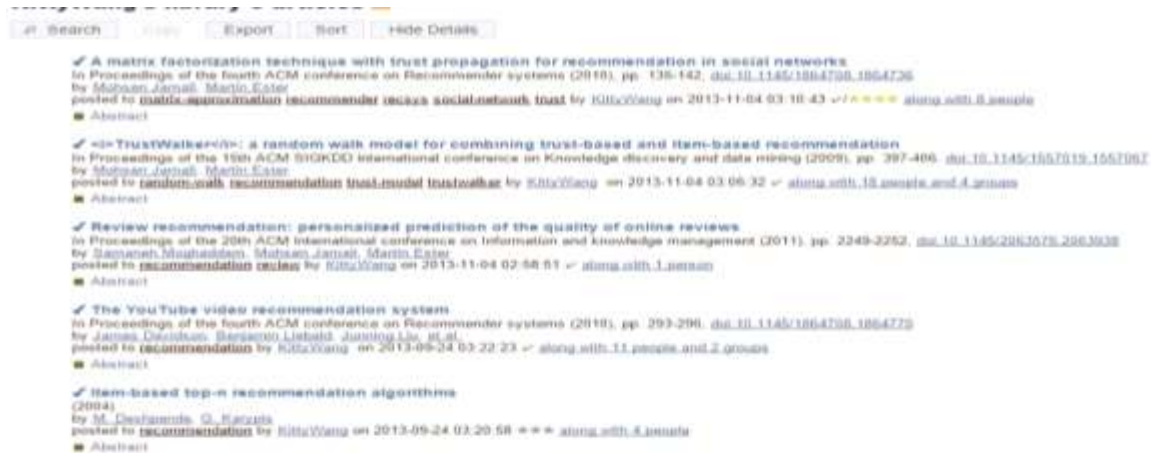
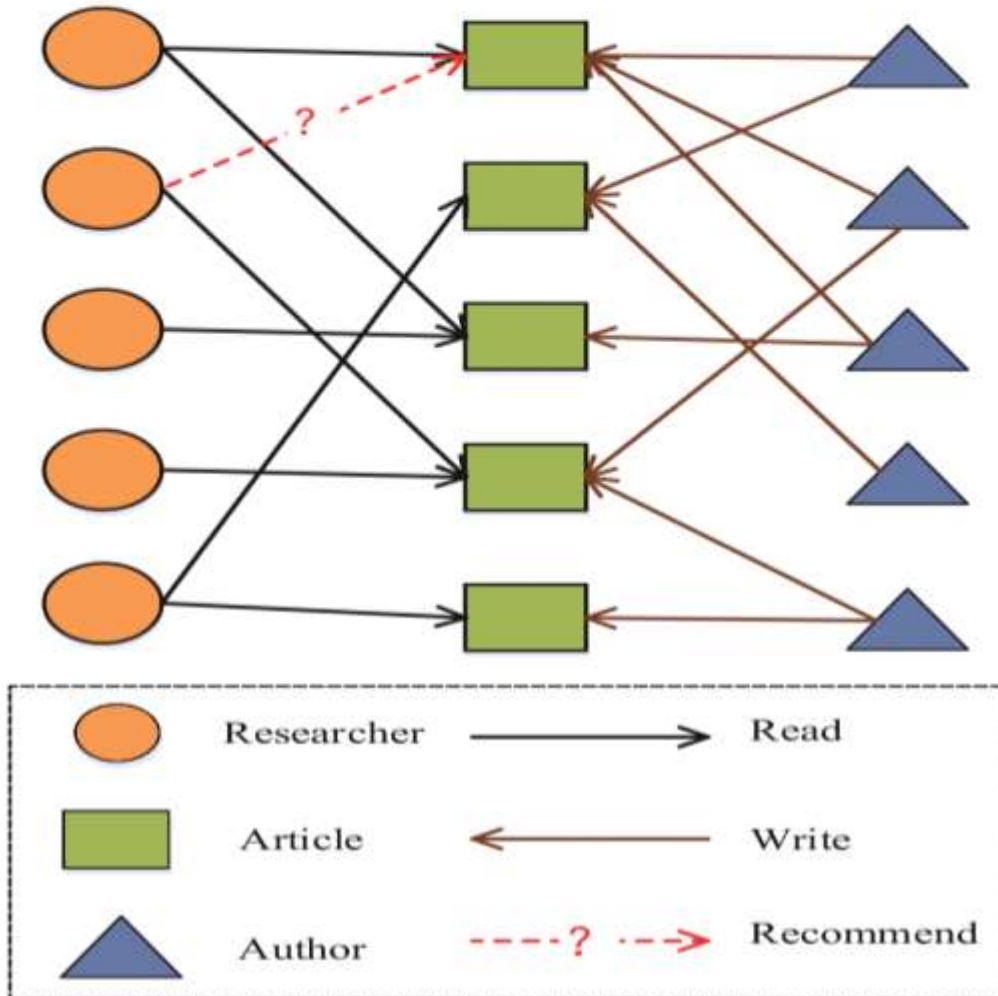Fig. 2. An example of article library.

Fig. 3. A example of recommendation scenario including three entities (researcher, article, and

author) and two relations (reading and writing)

articles have been included in KittyWang's library and each of them is given many different tags by the researcher. The researcher's historical preference is represented by the set of articles that interest him in his library. In this paper, our scientific article recommendation method aims to study how to automatically find the most possibly-preferred

articles which will be posted into a target researcher's library.

Fig. 3 shows our recommendation scenario. The scenario includes three objects: researchers, articles, and authors. Although there are possibly overlaps between researchers and authors, this situation is not taken into consideration
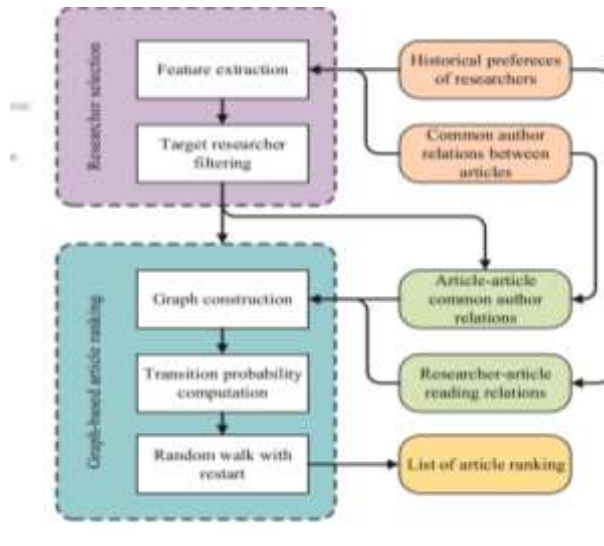

Fig. 4. The architecture of CARE

due to the facts that the CiteULike website does not provide enough information such as email address to determine each registered researcher's identity (i.e., whether he is an author of a certain article in article library) and the recommendation targets are the registered researchers rather than the authors. Additionally, there exist two kinds of links among the three objects. The first kind of link represents that a researcher, who is a registered user in CiteULike website, has read one or more articles he is interested in. The second kind of link represents that an article is written and published by one or more authors. Traditional collaborative filtering methods utilize the first kind of link to generate recommendations. The

rational underlying these methods is that, two researchers who are interested in the same articles are similar and then the tastes of similar researchers are used to predict those of target researchers. Generally, the second kind of link is ignored in these collaborative filtering methods. However, these additional information may influence recommendation quality due to the fact that researchers often focus the same authors' publications when they find these authors' work relevant to theirs. As a result, it is very necessary to incorporate the second kind of link to propose a novel recommendation method. This is the first problem we address in this paper.

4 DESIGN OF CARE

4.1 Overview

Our CARE method is inspired by two important facts: (i) researchers generally search articles written by the same authors; (ii) not all researchers have such an author-based search patterns. Fig. 4 shows the architecture of CARE, which mainly includes two components: (i)researcher selection module and (ii)

graph-based article ranking module. The first component is responsible for extracting relevant features from researchers' historical preferences and then selecting researchers with author-based search patterns as recommendation targets. The second component is responsible for incorporating common

author relations to build a graph and generating article ranking list through a graphbased random walk algorithm. In the domain of recommender systems, random walk-based ranking is a classical technique for recommendation. Based on the

4.2 Target researcher selection

For researchers who find articles of interest by searching article written by the same authors, in their online article libraries, possibly there are lots of articles which are mainly written by one or several authors. Therefore, we define two features which are relevant to common authors between any two articles to help determine target researchers:

- *FE*1, is the ratio of the total number of pairwise articles with common author relations to the total number of all pairwise articles for a researcher *FE*2, is the ratio of the occurrence number of the most frequently appeared author in articles to the total number of articles for a researcher.

For a researcher, when *FE*1 or *FE*2 is larger than a given threshold, this researcher will be considered to have an author-based search pattern and will be regarded as a target which is suitable for the next ranking component of our CARE method.

We use Fig. 5 as an example scenario for illustrating the computation process of the above two features. Fig 5(a) shows the writing relations between article and author for a target researcher *X*, where each edge links an article to one of all its authors and

4.3 Graph-based Article Ranking

4.3.1 Graph Construction

As aforementioned, in field of academic recommendation, there are many entities such as researchers, articles, conferences, journals, and so on. In this paper, we consider some of them and then design a method for recommending scientific articles. Scientific article recommender systems include a set of *n* researchers $R = \{R_1, \cdots, R_n\}$ and a set of *m* articles $A = \{A_1, \cdots, A_m\}$. Based on researchers' historical preferences, we can give the pairwise reading relations between researchers and articles, denoted as $W_{RA} = W_{RA}{}^n \times {}^m$ with $W_{RA}(i,j)$ and $W_{AR} = W_{AR}^{m \times n}$ with $W_{AR}(j,i)$, indicating whether a researcher R_i has read and expressed interest in an article A_j, as shown in Equation (1). As we consider undirected relations in our method, $W_{RA}(i,j)$ is equal to $W_{AR}(j,i)$. As stated in the previous section, we convert the writing relations between articles and authors into common author relations between articles. Then, we also give the pairwise relations between articles, denoted as $W_{AA} = W_{AA}{}^m \times {}^m$ with $W_{AA}(i,j)$ indicating whether there is/are common author(s) between the two article A_i and A_j, as shown

technique, many researchers [41] have successfully applied it to various recommendation scenarios. Next, we will introduce the two components in detail.

researcher *X* expressed interest to all articles in the figure. For example, article *A*1 is linked to its three authors *U*1, *U*3, and *U*4 through three different edges respectively. We consider two articles are related if they are linked to the same author(s) in Fig. 5(a). In this way, we convert the writing relation graph into a common author relation graph, as shown in Fig. 5(b), where two articles are linked to each other if they have common author(s). From Fig. 5(b), we can easily obtain the number of pairwise articles and it is equal to 4. For these articles, the number of all possible relations between articles is equal to $C_2^4 = 6$. As a result, *FE*1 is equal to 4/6 = 0.67. In addition, *U*1 is the most frequently occurred author and its occurrence number is 3. Then, *FE*2 is equal to 3/4 = 0.75. If the thresholds of *FE*1 and *FE*2 are set to 0.2 and 0.3 respectively, then 0.67 >0.2 and 0.75 >0.3. Therefore, the researcher *X* is a relevant target suitable for CARE method. There may be other features to determine relevant target researchers, but in this paper we consider the above two features and conduct experiments to verify their effectiveness in Section5.

in Equation (2). Likewise, we employ undirected common author relations. Therefore, $W_{AA}(i,j)$ is equal As stated in the previous section, we convert the writing relations between articles and authors into common author relations between articles. Then, we also give the pairwise relations between articles, denoted as WAA = Wm×m AA with WAA(i,j) indicating whether there is/arecommonauthor(s)betweenthetwoarticle Ai and Aj, as shown in Equation (2). Likewise, we employ undirected common author relations. Therefore, WAA(i,j) is equal to WAA(j,i). In addition, as there is no consideration on relations between researchers, we denote WRR = [0]. WRA(i,j) ={1 if Ri expressed interest in Aj 0 otherwise (1) WAA(i,j) =☐ ☐ ☐ ☐ ☐ 1 if there is/are common author(s) between articles Ai and Aj 0 otherwise (2) Based on the above two relation matrices, we con- struct a graph for applying a random walk-based article ranking algorithm, as shown in Fig. 6. Let G = (VR∪VA,ERA∪EAA), where ERA ⊆ VR×VA and EAA ⊆ VA ×VA. VR and VA indicate the set of researcher vertices and the

set of article vertices, respectively. ERA and EAA describe the set of reading relations between researchers andarticlesandthesetofcommonauthorrelationsbetween articles.Anedgelinkingare searcheri toanarticlej

4.3.2 Transition Probability Computation

A random walk in the graph is actually a transition from a vertex to another vertex. Therefore, we subsequently utilize the above three matrices to build a transition matrix, of which each element represents the transition probability

existsin the graph if WRA(i,j) or WAR(j,i) is equal to 1. Similarly, an edge linking an article i to another article j exists in the graph if WAA(i,j) or WAA(j,i) is equal to 1.

betweencorrespondingvertices(articletoarticle,articleto researcher, and researcher to researcher). The computation process is as follows. When a random walk starts with a
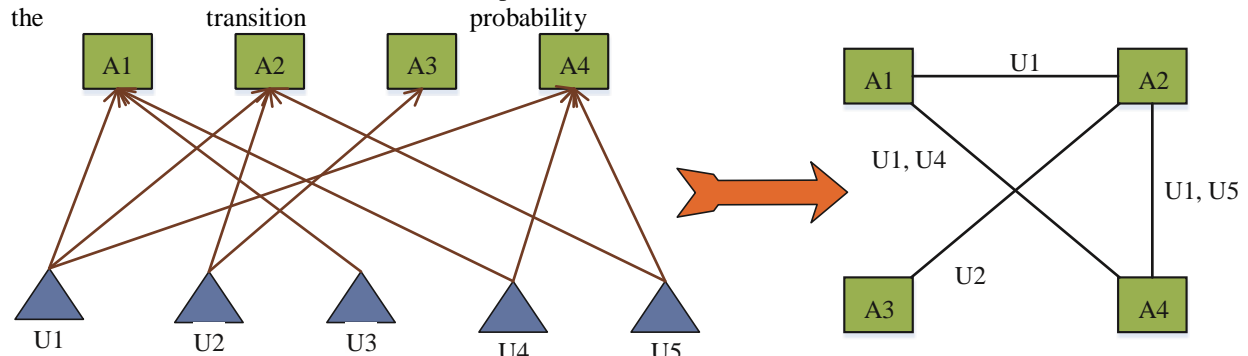
(a) Writing relations between articles and authors (b) Common author relations between the articles

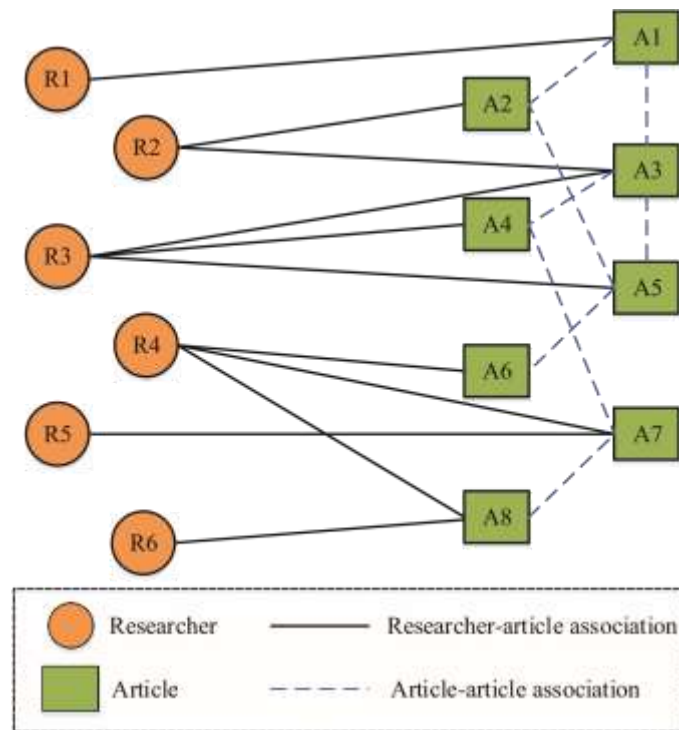Fig. 5. An example scenario for a researcher.

Fig. 6. An example graph for article ranking.

researcher vertex, the transition probability of moving to another researcher vertex is

$$T_{RR}(i,j) = 0 \qquad (3)$$

and the transition probability of moving to an article vertex

is

$$T_{RA}(i,j) = \frac{W_{RA}(i,j)}{\sum_k W_{RA}(i,k)} \quad (4)$$

Additionaly, when a random walk starts with an article vertex, the transition probability of moving to another article vertex is

$$T_{AA}(i,j) = \frac{W_{AA}(i,j)}{\sum_{k1} W_{AR}(i,k1) + \sum_{k2} W_{AA}(i,k2)} \quad (5)$$

and the transition probability of moving to a researcher vertex is

$$T_{AR}(i,j) = \frac{W_{AR}(i,j)}{\sum_{k1} W_{AR}(i,k1) + \sum_{k2} W_{AA}(i,k2)}$$

The transition probability matrix is

$$T = \left[\begin{array}{cc} T_{RR} & T_{RA} \\ T_{AR} & T_{AA} \end{array} \right] \quad (7)$$

Note that, in the above computation process of transition probability matrix, for each vertex, we assign equal values to all its neighbor vertices no matter what kind of vertex (researcher and article) the neighbor is. Specially, a vertex moves to any one of its neighbor vertices with the same probability 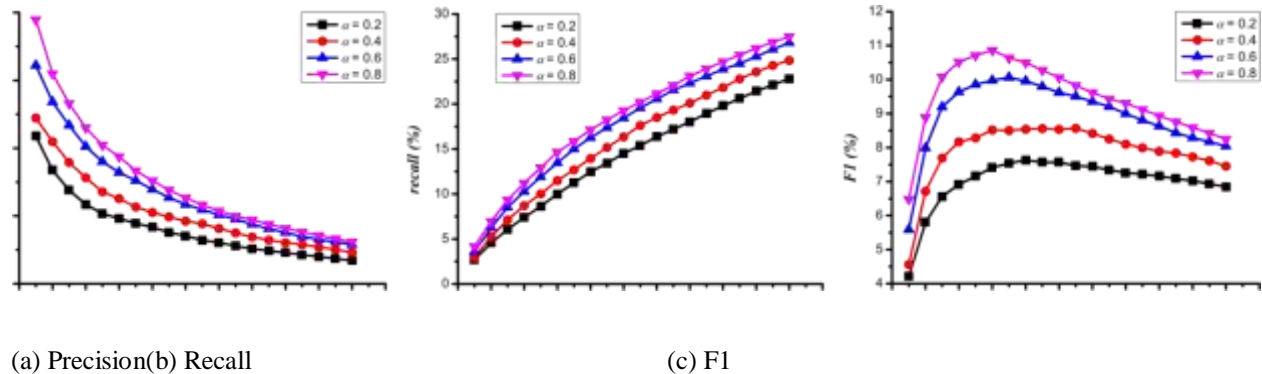even though these neighbors are different types of vertices. 4.3.3 Random Walk with Restart After obtaining the transition probability matrix, a random walk with restart method is employed to compute articles' rankings. Generally, the algorithm finds articles of interest based on the meta path: researcher-article-researcher-article. This means, are searcher is likely to be interested in an article which another researcher who has similar historical preferences expressed interest to. We incorporate common author relations between articles and then add another meta path: researcher-article-article. This means, a researcher is likely to be interested in an article which is similar to another article which another researcher has expressed interest to. Our algorithm considers the two meta paths. Starting from a source vertex v0 (target researcher), we perform random walk with restart in the graph built in previous section.After walking to any vertex vx, we continue then extrandom walk with probability α and walk to another vertex vy which links to vx with transition probability T(vx,vy). With probability 1−α, we return to source vertex v0. Algorithm 1 shows the process of graph-based article ranking. In this algorithm, a list of article rankings for target researcher are computed, and top-N articles which the researcher have not expressed interest in before, will be put in the recommendation list for a target

researcher.

(a) Precision(b) Recall   (c) F1

Fig. 7. Precision, recall, and F1 of CARE for different walking probability *α*.

Algorithm 1 Graph-based article ranking.

Input:

    Graph, *G*;

    Random walk probability, *α*;

    Target researcher vertex, v_0;

    Maximum step length of iteration, *maxStep*;

Transition probability matrix, T; **Output:**

Ranking scores of all article vertices, $ScoreArticle(1 : m)$; // m article vertices

1: Define ranking scores of all vertices, $ScoreAll(1 : n + m)$; // $n + m$ vertices

2: **for** each $v \in V_R \cup V_A$ **do**

3:      $ScoreAll(v) = 0$; //initial ranking scores are 0

4: **end for**
5: $ScoreAll(v_0) = 1$;
6: **for** $step = 0$; $step < maxStep$; $step + +$ **do**

7:**for** each $v \in V_R \cup V_A$ **do**

8:    $tmpScore(v) = 0$; //initial values are 0

9:    **end for**

10:      **for** each $v^x \in V^R \cup VA$ **do** $\times$ $\times$

11:    **for** each $v_y \in V_R$ V_A **do**

12:    $tmpScore(v_y) = \alpha$ $ScoreAll(v_x)$ $T(v_x, v_y) + tmpScore(v_y)$;
13:    **end for**
14:    **if** $v_x == v_0$ **then**
15:    $tmpScore(v_x) = tmpScore(v_x) + 1 - \alpha$;

16:    **end if**
17:     **end for**
18:      $ScoreAll = tmpScore$;
19:**end for**
20:$ScoreArticle(1 : m) = ScoreAll(n + 1 : n + m)$; // select ranking scores of article vertices 21: **return** $ScoreArticle(1 : m)$;

5   EXPERIMENTS

5.1   Dataset

CiteULike is a free web-based tool to help scientists, researchers, and academics store, organize, share, and discover links to academic research papers. With more than 3.5 million papers currently bookmarked and over 900000 visitors per month, CiteULike has grown to be one of the biggest and most popular social reference management

TABLE 1

Data statistics.

Number of researchers	5550
Number of articles	15439
Number of researcher-article reading relations	200251
Sparsity of researcher-article reading relations	0.9977
Number of article-article common author relations	18646

Sparsity of article-article common author relations	0.9998

websites by helping users streamline their process of storing and managing academic references. Emamy and Cameron [42] have provided detail description on CiteULike. We used the version of CiteULike dataset collected by Wang et al. [33] in our experiments. This dataset includes all registered users' (researchers) historical preferences, i.e., articles in each user's library, and articles' contents. Note that there is no author information in the original dataset. We designed a web crawler to collect each article's author information from CiteULike website. Then, we compared pairwise articles' authors to determine their common author relations. To avoid the situation that some authors' names are the same, two articles are considered to be relevant if they have at least two same authors. Although there possibly exist his own article(s) in a researcher's library, this situation can be ignored due to the following facts: (i) CiteULike has not provided enough information on the registered researchers'

identities such as email address, so there is no way to determine whether an article in a researcher's article library belongs to him (i.e., its author); (ii) The registered researchers generally put other researchers' articles of interest into his library. The original dataset includes 5551 researchers and 16980 articles. We removed articles with less than 5 researchers who express interests in them. The distribution of the preprocessed dataset is shown in Table 1. Similar to most datasets for evaluating recommendation methods, this one has the characteristic that the researcher-article and article-article relations are very sparse, i.e. data sparsity. The sparsity indicates the ratio of the difference between numbers of all possible relations and existing common author relations to the number of all possible relations. Therefore, based on the spare data, if a novel recommendation method can be designed to improve recommendation quality, to some extent, the challenge of data sparsity will be solved.
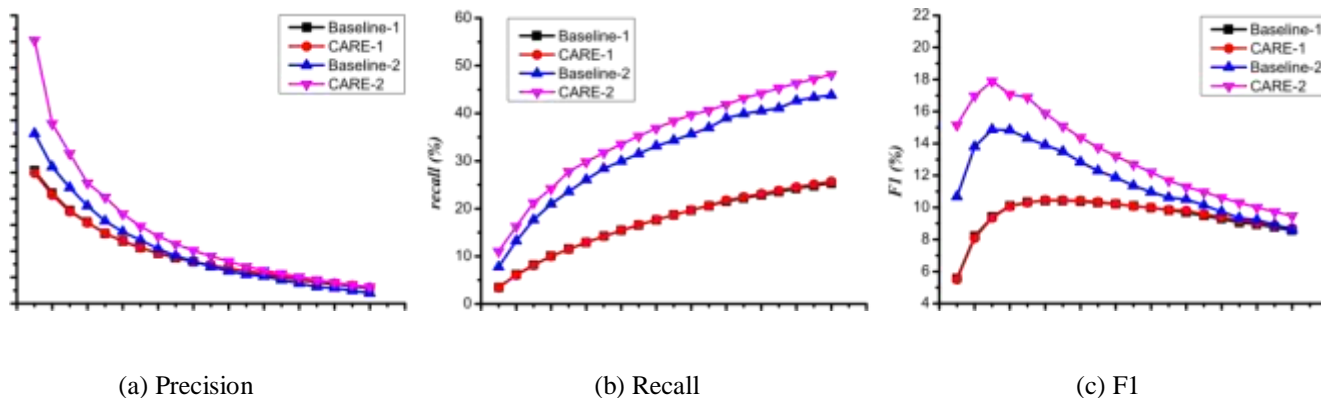

(a) Precision (b) Recall (c) F1

Fig. 8. Comparison of precision, recall, and F1 of Baseline and CARE for all researchers and relevant researchers.

5.2 Experimental Setup

To test our method's performance, the dataset is randomly divided into a training set (80%) and a test set (20%) using the following procedures. For researchers who have less than 5 articles in his library, we randomly select one article and the corresponding researcher into the test set. For others, we randomly select articles into the test set at the ratio of 20%. The training set is treated as known information used by our method for generating recommendations, while the test set is regarded as unknown information used for testing the

performance of recommendation results. To evaluate the recommendation quality of our proposed method, in our experiments, we employed three different metrics, namely, Precision, Recall, and F1, which have been widely used in the literatures [43], [44], [45] on the fields of recommender systems and information retrieval. Next, we give their definition information.

- *Precision*. Precision represents the probability that the selected articles appeared in the recommendation list which is shown as

$$Precision_i = \frac{N_{rt}^i}{N}$$
(8)

where *Precision$_i$* represents researcher R_i's precision, N_{rt}^i denotes the number of recommended articles that appeared in the researcher R_i's test set, and *N* represents the length of recommendation list. By averaging over all researchers' precisions, we can obtain the whole recommender systems' precision as

$$Precision = \frac{1}{n} \sum_{i=1}^{n} Precision_i$$
(9)

where *n* represents the number of researchers. Obviously, a higher precision means a higher recommendation accuracy.

Recall. Recall represents the probability that the recommended articles appeared in researcher's collected list shown as

$$Recall_i - \frac{N_{rt}^i}{N_t^i}$$
(10)

where *Recall$_i$* represents researcher R_i's recall and N_t^i is the number of articles collected by researcher R_i in the test set. Averaging over all individuals' recall, we can obtain the whole recommender systems' recall as

$$Recall = \frac{1}{n} \sum_{i=1}^{n} Recall_i$$
(11)

- *F1.* Generally speaking, for each researcher, recall is sensitive to *N* and a larger value of *N* generally gives a higher recall but a lower precision. F1, that assigns equal weight for precision and recall, is defined as

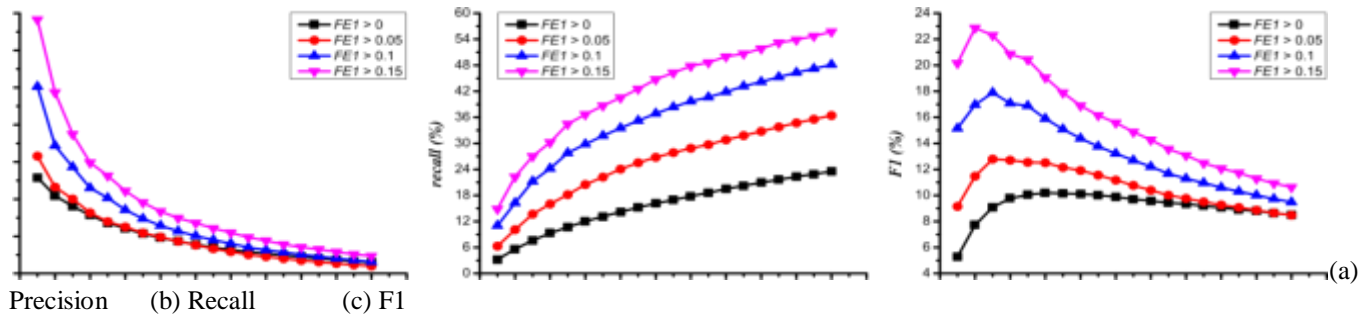$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$$
(12)

By averaging over all researchers' F1, we can also obtain the whole system's F1 as

$$F1 = \frac{1}{n} \sum_{i=1}^{n} F1_i$$
(13)

In order to demonstrate the effectiveness of our recommendation method, we compare CARE with the following method.

- **Baseline**: This is a random walk model with restart, which does not take into account common author relations between articles and does not differentiate between relevant researchers and irrelevant researchers.

5.3 Impact of Walking Probability

As stated in Section 4.3, for a certain vertex, α represents the walking probability from the vertex to its neighbor vertices and $(1-\alpha)$ represents the walking probability from the vertex to the source vertex (target researcher). Different values of α may produce different impacts on recommendation quality. We conducted relevant experiments using our proposed CARE method for different values of α. Fig. 7 shows the comparison results of precision, recall, and F1 when α is equal to 0.2, 0.4, 0.6, and 0.8, respectively. As shown in these sub-figures, for a larger value of α, our CARE method achieves larger values of precision, recall, and F1. For example, when N is equal to 6, CARE (α is equal to 0.2) achieves the worst results (6% precision, 10% recall, and 7.5% F1), and CARE (α is equal to 0.8) achieves the best results (8.5% precision, 14% recall, and 11% F1). This indicates that different walking probabilities have different impacts on CARE method. However, because α is

Precision   (b) Recall   (c) F1

Fig. 9. Comparison of precision, recall, and F1 of CARE for different thresholds of FE1.
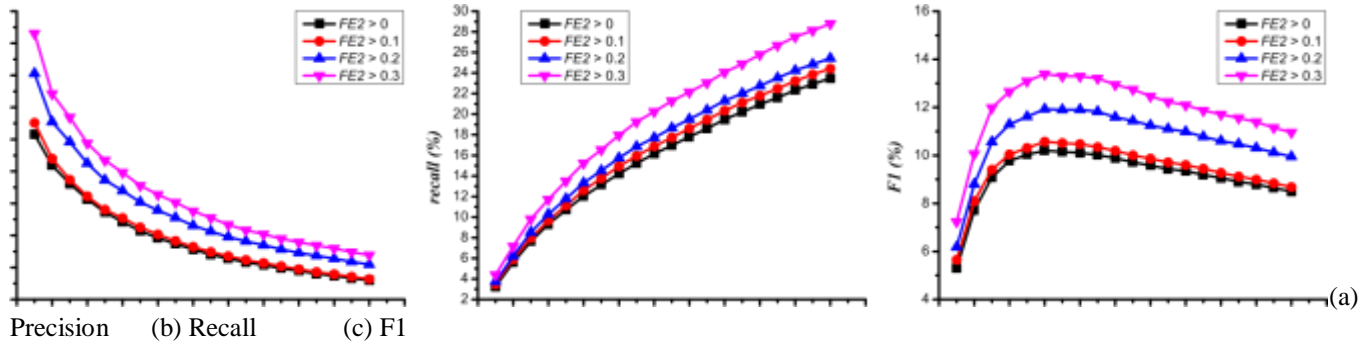
Precision   (b) Recall   (c) F1

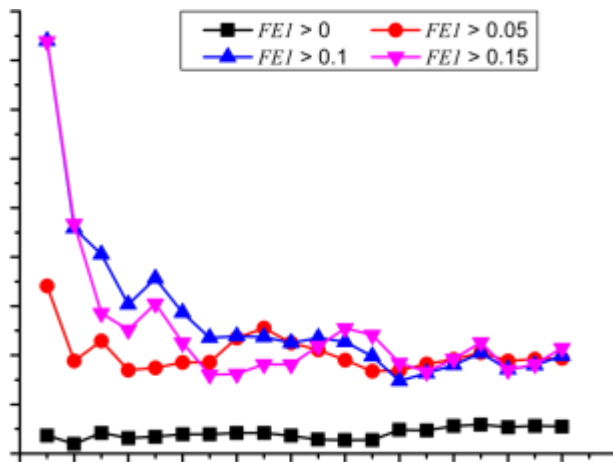Fig. 10. Comparison of precision, recall, and F1 of CARE for different thresholds of FE2

Fig. 11. Impact of FE1 on increase rate.

the common parameter for CARE and Baseline methods, it is enough to discuss their comparison results only if they employ the same value of α.
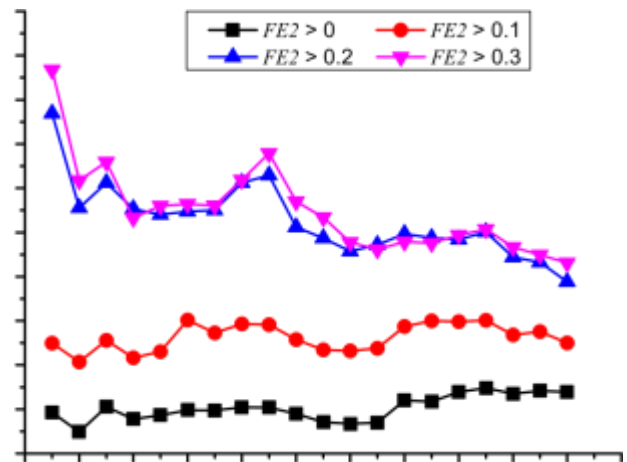

Fig. 12. Impact of FE2 on increase rate.

Therefore, we assign an empirical value of 0.8 to α for next experiments.

5.4 Comparison against Baseline Method

In this section, we conducted two groups of experiments: (i) compare the results of CARE and Baseline when all researchersaretargetsandthetwomethodsinthissituation are called CARE-1 and Baseline-1; (ii) compare the results of CARE and Baseline when a part of researchers are selected as relevant targets through the previously-defined features (FE1 > 0.1 and FE2 > 0.1) and the two methods under thissituationarecalledCARE-2andBaseline-2.Fig.8shows the result comparisons of the two groups of experiments. In the comparison of CARE-1

to Baseline-1, it can be obviously seen that, their precision, recall, and F1 are almost the same. Actually, the results of CARE-1 are a little worse than those of Baseline-1. However, we can see that, precision, recall, andF1ofCARE-2methodaresignificantlylargerthanthose of Baseline-2 method. It means that, CARE performs better than Baseline method for relevant researchers filtered using two features FE1 and FE2. This indicates that, incorporating common author relations is able to help generate accurate recommendations for relevant researchers rather than all researchers.

5.5 Impact of Researcher Features

In this section, we conducted relevant experiments to dis- cuss the impact of researcher's features previously defined in Section 4.2 on recommendation quality of CARE method. Fig. 9 shows the comparison results of CARE for different thresholds of FE1 (i.e., 0, 0.05, 0.1, and 0.15). We can see that, for a larger threshold of FE1, CARE achieves larger precision, recall, and F1. For example, for the length of recommendation list is equal to 2, when the threshold of FE1 is equal to 0, their results are 12% precision, 6% recall, and 7% F1 (i.e., the least values), and when the threshold of FE1 is equal to 0.15, their results are 23% precision, 22% recall, and 23% F1 (i.e., the largest values). This indicates that, a larger threshold of FE1 can help to find more relevant researchers with author-based search patterns and then our proposed CARE method generates better recommendations for these targets. In addition, Fig. 10 shows the comparison results of CARE for different thresholds of FE2 (i.e., 0, 0.1, 0.2, and 0.3). We can also see that, CARE with a larger threshold of FE2 performs better than that with a smallest threshold of FE2.Similarly,this demonstrates that,

FE2 is an effective feature for determining relevant target researchers who have author-based searcher patterns. In addition, we defined increase rate to represent CARE's improvement ratio to Baseline for different thresholds of FE1 and FE2 using Equation (14). Note that increase rate is the same for precision, recall, and F1. Fig. 11 shows the increase rate when the thresholds of FE1 are 0, 0.05, 0.1, and 0.15, respectively. It can be observed that, the increase rate is positive for these four situations. This demonstrates that, for the same researchers who are filtered using the threshold of FE1, CARE performs better than the Baseline method. Fig. 12 shows the increase rate when the thresholds of FE2 are 0.1, 0.2, 0.3, and 0.4. We can also see that, the increaserateispositiveforallsituations. Especially,whenthe threshold of FE2 is larger (e.g., 0.3 or 0.4), CARE performs much better than the Baseline method. These experiments further illustrate that these two features are useful to help find researchers with author-based search patterns and our CARE method is effective in terms of generating better recommendations for those targets.

CONCLUSIONI

In this paper, a novel method that exploits information pertaining to common author relations and historical preferences has been proposed to recommend articles of interest forspecificresearcherswithauthor-basedsearchpatterns.In order to determine specific targets, we defined two features (i.e. FE1 and FE2) which are relevant to common author relations between articles. Then, the information on common authors relations was incorporated to build a graph-based article ranking algorithm for generating a recommendation list. The experimental results demonstrated that, for relevant targets determined by two features, our proposed method performs better

different recommendation methods which are suitable for different researchers can be

than the Baseline method and the two features have impacts on recommendation quality. In addition,wealsodefinedtwootherfeatures(FE3 and FE4 in Section 5.5) and they are proved to be ineffective for suitable targets selection through relevant experiments. Asourfuturework,weplantodefinenewfeaturesto explore which researchers have author-based search patterns. In addition, it is potential to incorporate additional social relations such as citation relationships to design a citation- based recommendation method. Then, relevant targets are determined by analyzing the information on citation relations between articles. Finally,

(a) Precision   (b) Recall   (c) F1

Fig. 13. Comparison of precision, recall, and F1 of CARE for different thresholds of FE3.
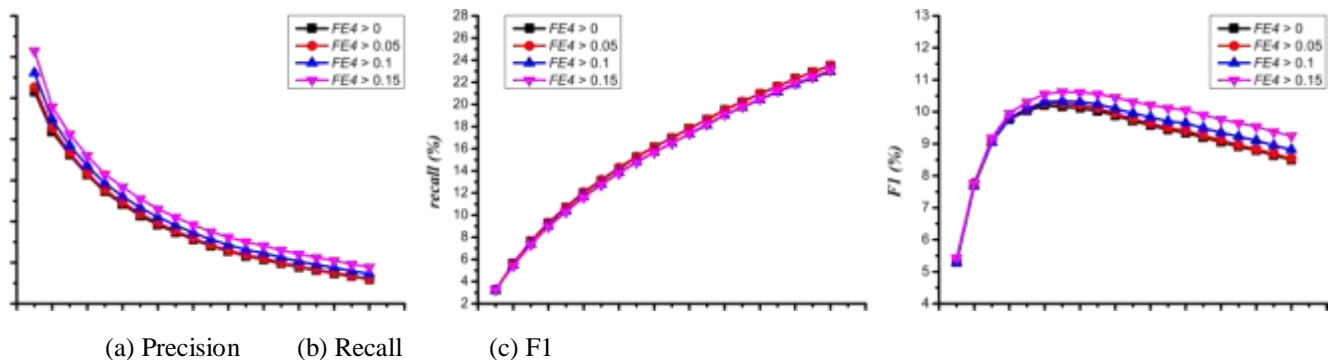

(a) Precision   (b) Recall   (c) F1

Fig. 14. Comparison of precision, recall, and F1 of CARE for different thresholds of FE4.

combined into a hybrid framework so that all researchers can obtain satisfactory recommendations.

REFERENCES

M. Gori and A. Pucci, "Research paper recommender systems: A random-walk based approach," in *2006 IEEE/WIC/ACM International Conference on Web Intelligence*, **2006, pp. 778–781.**

J. Tang and J. Zhang, "A discriminative approach to topic-based citation recommendation," in *PAKDD'09 Proceedings of the 13th Pacific-Asia Conferences on Advances in Knowledge Discovery and Data Mining*, **2009, pp. 572–579.**

J. Sun, J. Ma, Z. Liu, and Y. Miao, "Leveraging content and connections for scientific article recommendation in social computing contexts," *The Computer Journal*, **vol. 57, no. 9, pp. 1331–1342, 2014.**

H. Liu, Z. Yang, I. Lee, Z. Xu, S. Yu, and F. Xia, "Car: Incorporating filtered citation relations for scientific article recommendation," in *The 8th IEEE International Conference on Social Computing and Networking (SocialCom)*, Chengdu, China, Dec 2015.

T. Kolasa and D. Krol, **"A survey of algorithms for paper-reviewer assignment problem,"** *IETE Technical Review*, **vol. 28, no. 2, pp. 123–134, 2011.**

D. K. Tayal, P. Saxena, A. Sharma, G. Khanna, and S. Gupta, "New method for solving reviewer assignment problem using type-2 fuzzy sets and fuzzy functions," *Applied intelligence*, **vol. 40, no. 1, pp. 54–73, 2014.**

Z. Yang and B. D. Davison, "Venue recommendation: Submitting your paper with style," in *2012 11th International Conference on Machine Learning and Applications*, **vol. 1, 2012, pp. 681–686.**

E. Medvet, A. Bartoli, and G. Piccinin, "Publication venue recommendation based on paper abstract," in *Proceedings of the 26th IEEE International*

Conference on Tools with Artificial Intelligence, 2014, pp. 1004–1010.

J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1285–1293.

F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, "Mvcwalker: Random walk based most valuable collaborators recommendation exploiting academic factors," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 364–375, 2014.

S. Bethard and D. Jurafsky, "Who should i cite: learning literature search models from citation behavior," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010, pp. 609–618.

R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 542–550.

F. Meng, D. Gao, W. Li, X. Sun, and Y. Hou, "A unified graph model for personalized query-oriented reference paper recommendation," in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, 2013, pp. 1509–1512.

X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and

J. Han, "Cluscite: Effective citation recommendation by information network-based clustering," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 821–830.

X. Liu, Y. Yu, C. Guo, and Y. Sun, "Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation," in *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, 2014, pp. 121– 130.

Y. Lu, J. He, D. Shan, and H. Yan, "Recommending citations with translation model," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 2011, pp. 2017–2020.

W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and L. Rokach, "Recommending citations: translating papers into references," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, pp. 1910–1914.

P. Symeonidis, E. Tiakas, and Y. Manolopoulos, "Product recommendation and rating prediction based on multi-modal social networks," in *Proceedings of the 5th ACM Conference on Recommender Systems*, 2011, pp. 61–68.

M. Gori and A. Pucci, "Itemrank: A random-walk based scoring algorithm for recommender engines," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 2007, pp. 2766– 2771.

K. Emamy and R. Cameron, "Citeulike: a researcher's social bookmarking service," *Ariadne*, no. 51, p. 5, 2007.

J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5–53, 2004.

J. Bobadilla, F. Ortega, A. Hernando, and A. Gutierrez, "**Recom-´ mender systems survey,**" *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

D. A. Grossman and O. Frieder, *Information retrieval: Algorithms and heuristics*, 2012, vol. 15.