Knowledge-Based Systems 148 (2018) 85–99

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

HAR-SI: A novel hybrid article recommendation approach integrating with social information in scientific social network

Gang Wang[a,*], XiRan He[a], Carolyne Isigi Ishuga[b]

[a] School of Management, Hefei University of Technology, Hefei, Anhui, PR China
[b] Department of Management Science, Kenyatta University, Nairobi, Kenya

ARTICLE INFO

Article history:
Received 13 June 2017
Revised 9 February 2018
Accepted 13 February 2018
Available online 22 February 2018

Keywords:
Article recommendation
Scientific social network
Hybrid approach
Content-based filtering
Collaborative filtering

ABSTRACT

With the rapid development of information technology, scientific social network, i.e. SSN, have become the fastest and most convenient way for researchers to communicate with each other. However, a lot of published articles are shared via SSN every day which make it difficult for researchers to find highly valuable articles. To solve above information overload problem, a novel hybrid approach integrating with social information, i.e., HAR-SI, is proposed for the article recommendation in SSN. Unlike the traditional CBF and CF recommendation approaches, the social tag information and social friend information, which have been proved playing a significant role on recommendation in many domains, are effectively integrated into these two approaches separately to improve the accuracy in SSN. Prediction results made by the improved CBF and CF separately are combined with a hybrid method. In order to verify the effectiveness of the proposed HAR-SI, a real-life dataset from CiteULike was employed. Experimental results show that the proposed approach provides higher quality recommendations than the baseline methods, thus providing a more effective manner to recommend articles in SSN.

1. Introduction

The rapid development of information technologies, especially Web 2.0 technology, has changed Internet users from being passive, consumption-driven to being active, production-driven [1]. A variety of platforms resulted from the Web 2.0 have lifted the barrier of adding information to the Internet and enabled collaborative content creation and modification among users. Among them, the social network, as a typical application of Web 2.0, has become one of the fastest growing online information interaction communities in recent years. With the rapid expansion of social network, the research domain has also appeared a kind of social network with its own characteristics, here referred to as scientific social network, i.e., SSN, such as ResearchGate, CiteULike, Academia.edu, ScholarMate and so on. The SSN softens research boundaries, strengthens social networking research, and allows researchers easily to find, use, and share research articles. These activities are of great benefits to researchers because they can keep up-to-date the current trends in their research fields. However, the rapid increases in the rate at which new articles are published and the ease of sharing them in these SSN platforms has led to information overload problem. This makes it difficult for researchers to find the most interesting scientific articles in their research field. From this point of view, building recommendation systems that can reduce irrelevant content and provide researchers with more pertinent articles has become an important tool in helping researchers to relieve the burden of time wasted on irrelevant articles.

Recently, article recommendation as a research field has attracted many researchers. Different article recommendation approaches to automatically find articles from an overwhelming number of available options have been developed [1–6]. Generally, the present article recommendation approaches can be classified into three types: Content Based Filtering (CBF), Collaborative Filtering (CF), and hybrid approach [4]. The CBF approach has roots in digital libraries to assist knowledge worker with helpful documents on the Internet due to its nature of content oriented [7]. It analyzes the content of an article based on a researcher's past behaviors, such as the browsing behaviors on articles, to extract personal preferences, and then recommends articles that are similar to those in which the researcher has shown interest in the past [2,8]. For example, Basu et al. proposed a technical article recommendation approach, in which the multiple information sources were combined to represent the preference of the researchers and the content feature of the articles [9]. Martin et al. employed semi-structured data, such as authors, journals, and keywords, to assess research article similarity and further employed this similarity to support article recommendation [10]. However, the CBF approach

* Corresponding author.
E-mail address: wgedison@hfut.edu.cn (G. Wang).

https://doi.org/10.1016/j.knosys.2018.02.024
0950-7051/

sometimes fails to represent the articles accurately because it suffers from the limitation of content feature extraction. Besides that, the lack of recommended diversity, the new researcher, and the ignorance of opinions of other researchers are all problems arising in the CBF approach.

In order to overcome the above mentioned problems, another recommendation approach, i.e., CF, has attracted more attention in the last decades [3,11]. As its widely application in E-commerce systems, such as Amazon, Netflix, and Alibaba, there have also emerged a large number of explorations in applying of this approach in scientific article recommendation. For instance, McNee et al. used the CF by analyzing the citation graph of articles, and building different rating matrices [12]. Boger and Bosch applied the traditional CF approach on recommending scientific articles and further found that the user-based approach was better than the item-based approach because of the data distribution factors [13]. Zhang et al. propose a cross-domain recommender system with consistent information transfer (CIT). In which domain adaptation techniques are used to map and adjust those user and item latent groups in both the source domain and the target domain to ensure the knowledge extracted from the source domain is consistent with the target domain during the transfer learning process [14]. When the CF approach is being used to recommend an article, it considers the likes and dislikes information of a researcher. For example, in SSN, given a target researcher's ratings for several articles and a database of other researchers' ratings, the CF predicts how the target researcher would rate an article which he has not read. The key idea of CF is that a target researcher prefers those articles that like-minded researchers preferred. The CF predicts on assumption that if researchers agree on the quality of some articles, they are likely to agree on the quality of other unknown articles. However, in spite of the success and popularity, the CF approach also encounters several limitations that make it hardly to achieve expected performance, including the data sparsity, and the cold start problem [15]. Of them, data sparsity is the major problem which is resulted from the fact that researchers do not willing to invest effort and time in rating most articles in real world, hence the user-item matrix is sparse.

To alleviate the respective disadvantages of the CBF approach and the CF approach mentioned previously and leverage the advantages of them at the same time, in recent years, a hybrid recommendation approach, in which the CBF and CF are combined together, has been proposed to recommend scientific articles [4,16,17]. For instance, Hwang developed a hybrid article recommendation method that combined the co-authorship network-based with the content-based techniques and switched between them on the basis of the content coherence of an article profile [18]. In [19], a researcher's reading records and research proposal were analyzed to model his background knowledge and target knowledge by LDA, then based on the constructed concept map, the well-matched articles were discovered. Unfortunately, these approaches fail to fully use the abundant social information that naturally existed in the SSN and have been proved to be effective in improving the accuracy of recommendation in other domains [20]. For example, SSN allows researchers freely assign articles with tags. These social tags describe the contents of an article, thus, the latent features of the paper are obviously affected by the tags with which it is labeled. Meanwhile, they are also provided by researchers based on their own understanding independently, so they can more likely to reflect each researcher's own concerns. Therefore, when building profiles to express users' preference and papers' features in CBF, considering these tag information can not only provide additional textual information for the feature extraction of articles but also make the preference extraction of researchers more characterized. On the other hand, with the deepening of communication between researchers in SSN, a lot of friend relationships have been established by researchers under the similar interests, hobbies, and research field independently. Just like the strength that the social tag information performed in the procedure of profiling, this social friend information also plays an important role in article recommendation due to the fact that people tend to be more willing to choose articles which have been chosen by their friends in real world [21]. That is to say, the preference of a specific researcher should be similar to that of his friends to some extent. Nevertheless, despite these benefits of social information, i.e., tag, and friend, in the existing setting of article recommending in SSN, there has been little exploration of how to mine the social information in SSN to overcome the problems, such as data sparsity involved in traditional CBF and CF article recommendation approaches to improve the performance.

Therefore, to addresses these specific problems above, and further improve the performance of article recommendation, a novel hybrid article recommendation approach integrating with social information in scientific social network, i.e., HAR-SI, is proposed in this research. First of all, the social tag information is incorporated into the standard CBF approach as an additional textual information to build more personalized researcher and article profiles. Since these social tags are the description of an article's main contents provided by researchers themselves subjectively, they can express not only the features of articles but also the concerns of researchers. Secondly, the social friend information is incorporated into the standard CF approach. This is because in reality, people always turn to their friends for recommendations, and their tastes and characters can be easily influenced by the friends they keep. Thirdly, to minimize the respective limitation of these two approaches, the improved CBF approach and the improved CF approach are combined by building a hybrid approach, in which the predictions made by CBF and CF are firstly kept separated for the purpose of allowing us to benefit from the advances made by either CBF or CF. Then with a weight to balance the importance of these two improved approaches, the final personalized recommendation lists are presented. The proposed novel HAR-SI approach is evaluated through the comprehensive experiments using real-life data from CiteULike, a leading scientific social network. The experimental results show that the proposed HAR-SI approach achieves 0.1593 on precision when recommend 10 articles and 0.3110 on recall when recommend 50 articles. Moreover, comparing with the baseline methods, the proposed HAR-SI can outperform on both precision and recall at an improvement of more than 10% thus giving an effective manner to recommend articles on SSN.

The main contributions of this article can be summarized as:

(1) An in-depth study on article recommendation issues in SSN, a kind of social platform that specializes in serving scientific researchers. In recent years, due to the generation of numerous articles, SSN is facing serious information overload problem. There is an urgent need for a way to help users discover valuable content quickly and accurately. However, no cure-all solution is yet available. Therefore, based on the analyses of SSN's unique characteristics, this paper conducts the research on the recommendation for the articles which are the most concerned by users.
(2) An application for social information to overcome the problems contained in traditional article recommendation approaches. As explained above, in SSN, we naturally have much social information, which has been proved playing a significant role in performing recommendations in many domains. Nevertheless, despite those benefits, in the existing setting of article recommendation in SSN, there has been little exploration of how to mine the social information existed in SSN. Motivated by the need to attempt this, in this paper, the social information is integrated into both the traditional

Table 1
Selected previous studies in article recommendation approaches.

Study	Year	Recommendation approach	Use social information	Application domain
Nallapati et al. [22]	2008	Content based filtering	Not	Online paper search
Chandrasekaran et al. [8]	2008	Content based filtering	Not	Digital library
Vellino et al. [26]	2010	Collaborative filtering	Not	Digital library
Lee and Brusilovsky [27]	2010	Collaborative filtering	Social group information	Social network
Nascimento et al. [21]	2011	Content based filtering	Not	Digital library
Kim et al. [33]	2011	Hybrid approach	Community information	Social network
Martin et al. [13]	2013	Content based filtering	Not	Online paper search
Lai et al. [28]	2013	Collaborative filtering	Social trust information	Digital library
Sun et al. [37]	2014	Hybrid approach	Social tag and neighbor information	Social network
Zhao et al. [19]	2016	Hybrid approach	Not	Digital library

CBF and the traditional CF approach as an additional supplement to achieve superior recommendations.

(3) A novel hybrid approach integrating with the social information for the article recommendation in SSN, called HAR-SI. This approach builds a hybrid approach with a balanced weight that minimize the respective limitation of these two approaches and benefit from the advances made by either CBF or CF at the same time. In addition, with an adjustable weight parameter, for other SSN other than experimental dataset CiteULike, our methods HAR-SI can achieve its best performance on that specific SSN through the adjustment of that weight parameter. To the best of our knowledge, this type of approach has rarely been addressed in the state-of-the-art article recommendation research in SSN.

(4) Comprehensive experiments were conducted using a real-life dataset, i.e., CiteULike, to evaluate the performance of the proposed HAR-SI as well as the impacts of the integrated social information. Experimental results were presented and compared to show the effectiveness of the proposed approach.

The rest of the article is organized as follows. In Section 2, we survey the related work about CBF, CF, and hybrid approach for article recommendation. The details of the HAR-SI approach are introduced in Section 3, while Section 4 presents the dataset, metrics and methodology used in the experiments. The results are analyzed in Section 5. We draw the conclusions in Section 6 and discuss future research.

2. Literature review

Our work is related to three main research threads: (1) content based filtering, (2) collaborative filtering, and (3) hybrid approach for article recommendation. We review the pertinent existing literature in these three fields. Table 1 presents the selected previous studies in article recommendation.

2.1. Content based filtering approach for article recommendation

The CBF approach tries to retrieve useful information which is usually textual data from articles that the researchers have shown interest in the past to build article feature profiles and researcher preference profiles. And then, relying on the researcher's historical behaviors together with these profiles, a list of new articles that are similar to the researcher preference are recommended [4]. Case studies of typical CBF recommendation methods recently can be given by Nallapati et al., in which the text and citation relationship were jointly modeled under the framework of topic model [22]. They introduced a Pair-Link-LDA model which models the presence or absence of a link between pairwise articles but does not scale to large digital libraries. They also introduced another model called Link-PLSA-LDA to model citations as a sample from a probability distribution associated with a topic. For the effective performance of the CBF approach, personal preferences are required to be identified from researchers and article contents [2]. Hong et al. designed an expansive personalized research paper recommendation system and implemented a user-profile-based algorithm for extracting keyword by keyword extraction and keyword inference [23]. These researcher profiles could be constructed through either explicit declaration by researchers or observation of researchers' implicit interaction. For example, the CiteSeer, as a successful digital library system, is designed to find out the most relevant research articles for online researchers based on the explicitly stated individual profiles of researchers [24]. On the contrast, in [7], the historical browsing behaviors of researchers were used to build the researcher profiles by analyzing documents previously read by the researcher. Then a semantic-expansion approach was adopted to recommend articles. PVA learned a researcher profile implicitly without researcher intervention [25]. The researcher profile is represented as a keyword vector in the form of a hierarchical category structure. Besides, recommendation systems in [8] was based on similarities between researcher profiles and the concepts of each article in the background data. In [21], a scholarly article recommendation system was developed by Nascimento et al., in which they used the title to construct researcher profiles, and the title and abstract to generate feature vectors of candidate articles to recommend.

However, as shown in Table 1, most of the above profile representation techniques extremely emphasized on the content of an article, thus encountering problems of the content feature extraction for the researcher profiling and article profiling, the lack of recommended diversity, the new researcher, and the ignorance of opinions of other researchers before recommending articles, which highly limit the satisfaction in CBF recommendation approach.

2.2. Collaborative filtering approach for article recommendation

With the prevalence of social bookmarking and social networking websites, the CF approach has attracted more attentions and is currently the most widely used technique for article recommendation. It recommends articles for researcher based on the relationships between other like-minded individuals or articles that have similar preferences or features as the target researcher. Bogers tried and evaluated three variants of user-based CF algorithms on article recommendation [13]. Experimental results showed that, owing to tag contribution, BM25-boosted CF achieved better performance over the other two CF methods i.e., Classic CF and Neighbor-weighted CF. To overcome the data sparsity problem, Vellino proposed a hybrid multi-dimensional approach where usage-based and citation-based methods for recommending research articles were combined [26]. The similar application also presented in [1], which provided an interpretable latent structure with the CF

and probabilistic topic modeling for researchers and articles, thus was able to make recommendations for both existing and newly published articles. Besides collective relations involved in user-item pairs, some attempts have been made to analyze and incorporate social relations into CF methods. Lee and Brusilovsky conducted article recommendation considering the self-defined social contacts while the group trust were mined together with the personal trust into the traditional CF techniques to recommend articles [27]. In [28], a personal trust model was proposed which adaptively combines the rating-based trust model and explicit trust metric to trust values. Then the trust values derived from their trust models were regarded as weightings in the CF method to identify the trustworthy recommenders for predicting article ratings.

As shown in Table 1, it is clear from the studies in the literature that the CF approach, as a popular and successful approach in e-commerce domain, can also be used in article recommendation and can provide recommendations of articles that are unable be considered by the CBF for its inner limitations [29–31]. However, the data sparsity and the cold start problems are both problems arising in the CBF for article recommendation that cannot be ignored.

2.3. Hybrid approach for article recommendation

Hybrid recommendation approaches combine the CBF with the CF approach in order to achieve the synergy between them [16]. The main assumption for this approach can be stated that "combining the approaches can provide more accurate recommendation than a single approach and disadvantages of each approach can be overcome by the other approaches" [32]. Recommendation systems in [33] described a two-stage approach. First, it adapted a CBF approach by representing books with keyword features to find neighbors, then it generated a CF-based recommendation list and removed irrelevant books from the CF-based list using the keyword preferences of individual members. Kodakateri reversed the stages [34]. It executed CBF approach based on the result trained through the ACM Computing Classification System, instead of combining CF and CBF approaches. In [18], another hybrid approach was proposed by switching between the coauthorship network-based and content-based techniques on the basis of the content coherence of a task profile. Apart from researching on the different ways about combining the CBF with the CF, the social information has also been considered to help generate the better recommendations. Among them, Yuan et al. introduced a hybrid recommendation approach, in which the rich content information of social media was explored to find the potential negative examples from the missing researcher-article pairs and further help researchers find their interesting articles [35]. Similarly, Weng and Chang recommended articles by taking into account other influential researchers in the community network environment combining the ontology to construct researcher profiles, and made use of researcher profile ontology as the basis to infer the preference of researchers [36]. Besides these studies, Sun et al. considered three different types of online social connection between researchers to find like-minded neighbors while combining with the CF approach to recommend scientific article [37].

The above-mentioned studies have combined CBF and CF approaches to exploit the benefits of each other and lessen their disadvantages. As shown in Table 1, a lot of experiments with different combining strategies have been conducted in their studies, which have proved its effectiveness of hybrid approach. However, in SSN with its natural characteristic of large social information, such as tags and friends, which has been applied in other domains and proved to significantly enhance the recommendation quality, few studies have explored using the hybrid approach integrated with this social information for article recommendation [38–40]. In the paper of our previous work [41], the article recommendation in SSN was formalized an One-Class Collaborative Filtering (OCCF) problem, then for the extreme data imbalance and data sparsity of OCCF problem, the social information was firstly applied to extract negative instance, then used to conduct an unified probability matrix factorization with the negative instance added data. Hence, as an extended research, having considering all the above-mentioned problems, in this research, a novel hybrid HAR-SI approach is proposed, in which the social information was integrated into both the traditional CBF and CF approaches separately, then the prediction results made by these two improved CBF and CF were combined with a hybrid method to minimize the respective limitation of them and leverage the advantages made by either CBF or CF at the same time.

3. A hybrid approach for article recommendation with social information in SSN

Scientific social network is the mostly research-centered community platform with high degree of article sharing and communicating. On these platforms, researchers accumulate articles when they wants to execute a research task. Researchers can also give tags to articles they have read and build friend relations with other researchers who are under the similar interests, hobbies or research fields with them. Having considered these nature characters of SSN, in this research, a novel hybrid HAR-SI approach is proposed to help researchers find useful articles for their research. The rationale behind the proposed approach is that an individual's interaction activity on the SSN platform may partially reflect and can further be used to satisfy his needs. Fig. 1 shows the overview of the proposed novel hybrid approach. The framework of the proposed HAR-SI contains four primary stages: data acquisition, improved CBF model, improved CF model, and hybrid model. The improved CBF model and the improved CF model are employed to firstly make predictions integrating with social tag and friend information separately. Then the results obtained from these two models are combined by the hybrid model with an adjustable parameter to further enhance the performance of recommendation. Our hybrid approach recommends articles from the perspective of researcher's information needs.

3.1. Data acquisition

The web crawler was used to search the web and retrieved all articles, researchers and friend relations in a certain SSN platform. The web crawler obtains a full technical datasheet for each article together with its browsed histories by researchers. Each specific characteristic of an article is a field that the individual recommendation system needs. Besides that, all tags which were labeled by researchers to summarize the article content from their own perspective and friend relations that were shown on the page were also collected.

The collected data is analyzed and preprocessed to generate researcher and article vectors. Words that appear in nearly all articles in the article vectors are cleaned up because they are irrelevant [42]. The cleaning process is mostly referred to as removal of stop words. These words are, for instance, prepositions such as "by", "for", "in", and others. They are removed from the collected data using Lucene, a full-text search engine toolkit. The goal in this stage is to help remove unwanted strings and leaving vocabulary that can adequately represent a particular article. The result of data pre-processing is used as experimental data.

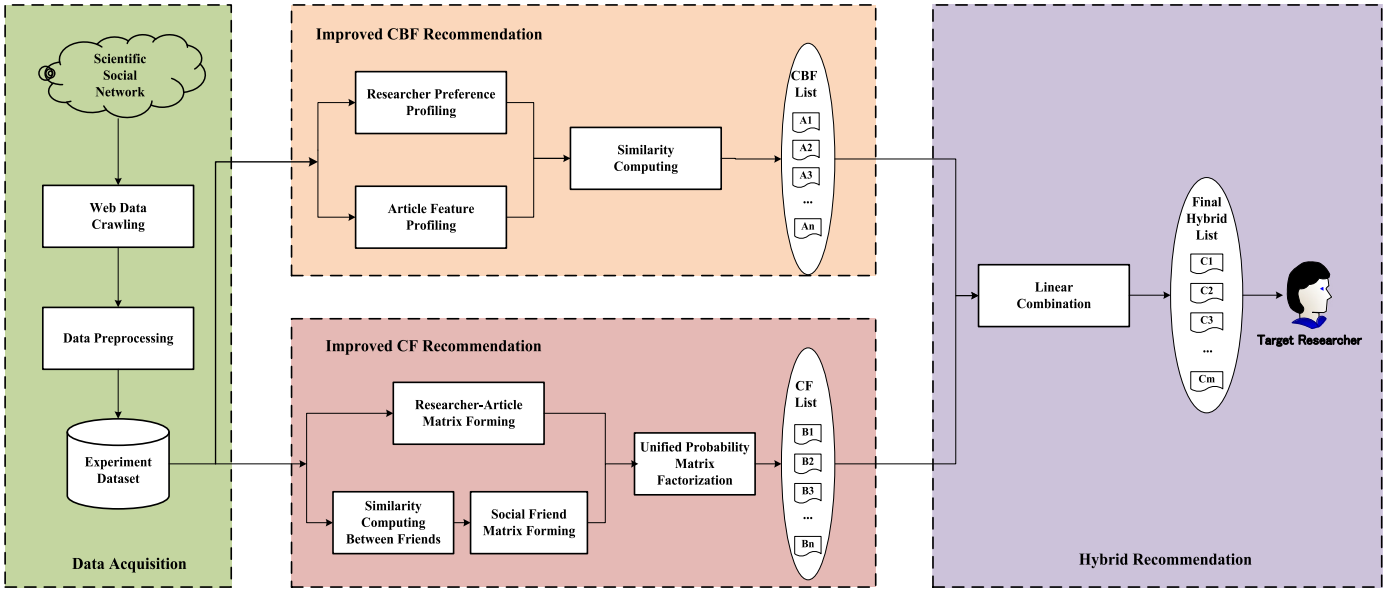

Fig. 1. Overview of HAR-SI for article recommendation in SSN.

3.2. The improved CBF model with social tag information

The CBF approach recommends a set of articles that are the most similar to the articles already read by the researcher. After analyzing researchers' activities in the SSN platform and retrieving the required experimental data, the next step is to build article profiles and researcher profiles, then the similarities between the article profiles and researcher profiles are calculated. Later, based on the assuming of CBF that the higher the similarity value between a candidate article and a target researcher, the more likely that candidate article is interesting to that researcher, the predictions of articles are given according to the similarity scores. It is important to note that in SSN, researchers are provided to give tags to articles they have read. This available tag information was also retrieved in this research along with other information of articles in the creation of article profiles and researcher profiles before making recommendations.

On the other hand, to identify articles similar to what a researcher likes, the Information Retrieval (IR) approach, which begins with a user query and searches for relevant documents from a large database, is used. Based on their query styles, document representation, and matching results, the IR techniques can further be classified into four basic models: boolean model, probabilistic model, vector space model, and language model [43–45]. The boolean model which provides boolean operators, such as AND, OR, and NOT, to increase search flexibility is initially used. However, in the view of this model, it does not consider any form of relevance ranking of the retrieved document set. The probabilistic model and vector space model have following been proposed in response. The probabilistic model uses keywords to estimate the probability that a document matches a query, while the vector space model utilizes arrays of keywords to represent queries and uses their similarities to make recommendations [46–48]. However, despite their better retrieval performance than that of the traditional boolean model, these models may also encounter problems that lead to inefficient searches [46]. For example, a user query may contain only two or three terms, which is inadequate for locating a highly relevant document. Given the problems above, a language modeling, i.e., LM, which has already been shown to perform well, is chosen to be used in the proposed hybrid approach as basic CBF approach for promising results [49,50]. Because LM is based on the assumption that a document is generated by some kind of underlying probabilistic model. It primarily based on the terms that occur in it, and then compare the query to that model rather than to the actual document to estimate the relevance of a query to a document. Considering these advantages, it is reasonable to infer that the LM can also perform well in our approach.

In sum, the main tasks of the improved CBF approach include: article feature profiling, user preference profiling, similarity computing and prediction. First of all, the article profiles are built based on the title and all tags labeled by researchers. It contains a set of features that describe and characterize a particular article. An article profile is obtained by analyzing the title and tags given to that specific article. Since each words contained in the article profiles has a weight which denotes the frequency of the word to a specific article, the weight of the word is determined. For example, $articleProfile(v_j)$ is a profile of an article v_j, and it contains the description of this article. $articleProfile(v_j)$ can therefore be defined as a vector of weights $(w_{v_j1}, w_{v_j2},...,w_{v_jt},...,w_{v_jl}), 1 \le t \le l$, where each weight w_{v_jt} represents the importance of a word k_t to the article v_j.

Then, in a similar way, the researcher profiles are built with the tags he has labeled and titles of all articles read by the researcher. Therefore, the researcher profile contains articles preferred by a specific researcher. Likewise, $researcherProfile(u_i)$ can be defined as a vector of weights $(w_{u_i1}, w_{u_i2},...,w_{u_is}...,w_{u_ir}), 1 \le s \le r$, where each weight w_{u_is} represents the importance of a word k_s to the researcher u_i.

After that, LM method is used to compute the profiles similarity between the target researcher and articles unread by the researcher. The weight for each term common between the query and article is determined. The LM for articles is estimated, and then the articles are ranked by the probability of the query with an estimated LM. The method assumes that a query q is "generated" by a probabilistic model based on an article v_j. When given a query $q = w_1w_2w_3 \dots w_t \dots w_l$ where w_t represents a word in the query. Probability that an article generates the observed q is given by:

$$P(q|v_j) = \prod_{w_t \in q} P(w_t|v_j) \tag{1}$$

In order to solve the sparsity problem, LM approach uses a data smoothing method. That is, adjusting the probability of maximum likelihood estimation so as to reduce the probability of having zero and non-zero values, thus improving the overall accuracy of the model. A model $p_t(w|v_j)$ used for "seen" words that occur in the articles v_j, and a model $p_u(w|v_j)$ for "unseen" words. The probability of a query q can then be written in terms of these models as follows:

$$P(w_t|v_j) = \frac{c(w_t, v_j)}{|v_j|} \tag{2}$$

where $c(w_t, v_j)$ is the number of times a word appears in an article v_j, while $|v_j|$ is the length of an article. With the profiles already constructed, similarity between a set of words from researcher vector u_i and article vector v_j is obtained. The cosine similarity measure was used as follows:

$$sim(u_i, v_j) = \frac{\sum_t u_i(t) \times v_j(t)}{\sqrt{\sum_t u_i(t)^2} \times \sqrt{\sum_t v_j(t)^2}} \tag{3}$$

In this formula, u_i is a researcher model vector, and v_j is one of the articles vectors unread by that researcher u_i. The similarity values $sim(u_i, v_j)$ range between 0 and 1, the nearer the values to 1, the higher the similarity between them. This calculation process is repeated for all candidate articles and target researcher, then the articles are sorted based on these calculated results.

Finally, after all similarity values, $sim(u_i, v_j)$, are assigned according to similarity scores, only Top N articles are considered. In the case of having the same similarity, they are randomly selected. Based on the above analysis, the predictions of this approach are made according to the similarity between the researcher preference profiles and article feature profiles. These articles that have a high degree of similarity to whatever the target researcher prefers are considered. This approach is able to select articles that are of similar interest to the researchers in the past. Fig. 2 shows the algorithm of the improved CBF recommendation approach.

3.3. The improved CF model with social friend information

Unlike the CBF approach, the CF approach predicts articles to researchers based on their previously rated articles. It does not require article's content information and can discover interesting associations that the CBF methods cannot. However, on the one hand, as occurring in the other domain, such as E-commerce, the application of CF in article recommendation also encounters the problems of data sparsity and cold start. On the other hand, considering the SSN, as a special form of social network, exists a large number of social friend information created by researchers independently, which has been proved to significantly relief the effects of these limitations and thus improved the recommendation performance [51]. Therefore, based on the fact that the users' favors can easily affected by their friends, the latent characteristic matrix of them is reasonable to be similar with his friends. And the higher the similarity between the user and his friend, the closer the latent characteristic matrix of them [51,52]. In this research, the friend relations information is utilized into the standard CF approach to solve the problem of low recommend accuracy resulting from data sparsity. The improved CF is implemented in the following three stages: researcher-article matrix forming, social friend matrix forming, and unified probabilistic matrix factorization conducting.

First of all, the researcher-article matrix is formed with the browsing histories of all researchers. In order to learn the latent characteristics of the researchers and articles, the Probabilistic Matrix Factorization (PMF) is employed to factorize the researcher-article matrix. The idea of researcher-article matrix factorization is to derive a high-quality D-dimensional feature representation U of researchers, and V of articles based on analyzing the researcher-article matrix $R^{M\times N}$. The researcher-article matrix is a collection of personal interest about a given researcher, which is the basis of CF approach. A researcher can be modeled depending on the amount of information stored in his researcher-article vector [53]. The matrix contains sets of articles by researchers, whereby each row represents ratings of a researcher and each column represents ratings of an article [54]. Given data consisting of M researchers and N articles, there exists $M \times N$ researcher-article matrix. An entry of this matrix represents the rating that a researcher gives to an article. Let $R_{i,j}$ represent the rating of user u_i for item v_j which contains binary data of "null" and "1" in this research, where the "null" represents the article that is unread by the researcher while the "1" represents that researcher has read on that specific article. And $U \in R^{D\times M}$ and $V \in R^{D\times N}$ are latent researcher and article feature matrices, with column vectors U_i and V_j representing the D-dimensional researcher-specific and article-specific latent feature vectors of researcher u_i and article v_j respectively.

Then, the social friend matrix is built with the similarities between all researchers and their friends based on the researcher-article matrix. Given researcher-article matrix $R^{M\times N}$, similarity between a researcher u_i and his friend researcher u_k, $sim(u_i, u_k)$, can be obtained by using Jaccard measure as:

$$J_{i,k} = \frac{|R_{i,\cdot} \cap R_{k,\cdot}|}{|R_{i,\cdot} \cup R_{k,\cdot}|} \tag{4}$$

where $R_{i,\cdot}$ and $R_{k,\cdot}$ are the articles collection read by researcher u_i and u_k respectively. The results range from 0 to 1, and the bigger value indicates the more similar between the two researchers. Let $S^{M\times M} = \{S_{i,k}\}$ denotes the $M \times M$ matrix of researchers networks, which is also called the social friend matrix in this research. For a pair of vertices u_i and u_k, if they are friends, let $S_{i,k}$ denotes the similarity of the two researchers, and $S_{i,k} = 0$, otherwise.

Finally, integrate the social friend matrix $S^{M\times M}$ into researcher-article matrix $R^{M\times N}$ to conduct unified probabilistic matrix factorization. Getting the researchers' latent characteristic matrix U and articles' latent characteristic matrix V when the value of U^TV approximate the researcher-article matrix $R^{M\times N}$ as far as possible. Its probability graph model is shown in Fig. 3.

Where column vectors U_i and V_j representing the D-dimensional researcher-specific and article-specific latent feature vector of researcher u_i, and article v_j, respectively. $S_{i,k}$ indicates the Jaccard similarity between researcher u_i and u_k. $N(i)$ indicates the friends collection of u_i. Note that the solutions of U and V are not unique.

According to the definition above, the conditional distribution over the observed researcher-article matrix $R^{M\times N}$ is defined as follows:

$$P(R|U, V, \sigma_R^2) = \prod_{i=1}^{M}\prod_{j=1}^{N}\left[N(R_{i,j}|g(U_i^TV_j), \sigma_R^2)\right]^{I_{i,j}^R} \tag{5}$$

where $N(R_{i,j}|g(U_i^TV_j), \sigma_R^2)$ is the probability density function of the Gaussian distribution with mean $g(U_i^TV_j)$, and variance σ_R^2, and $I_{i,j}^R$ is the indicator function that is equal to 1 if researcher u_i read article v_j, and equal to 0 otherwise. The function $g(x)$ is the logistic function $g(x) = 1/(1 + exp(-x))$, which makes it possible to bound the range of $U_i^TV_j$ within the range [0, 1]. The zero-mean spherical Gaussian priors are also placed on user and item feature vectors:

$$P(U|S, \sigma_U^2, \sigma_S^2) = \prod_{i=1}^{M} N(U_i|0, \sigma_U^2 I) \times \prod_{i=1}^{M} N\left(U_i\middle|\sum_{k\in N(i)} S_{i,k}U_k, \sigma_S^2 I\right) \tag{6}$$

Input: observed preprocessed dataset with M researchers, N articles, researcher browsing records, title of articles and all tags given by researchers.

Output: similarity between researchers and unread articles of them

1: **For** j=1 to N **do**

2: Build article profile with its title and all tags labeled by researchers;

3: Create $articleProfile\left(v_j\right)$ as a vector of weights $\left(w_{v_j1}, w_{v_j2}, \ldots, w_{v_j1}\right)$ to represent the number of words occurring in the article profile;

4: **End for**

5: **For** i=1 to M **do**

6: Build researcher profile with title of articles he has read and all tags labeled by he;

7: Create $researcherProfile\left(u_i\right)$ as a vector of weights $\left(w_{u_i1}, w_{u_i2}, \ldots, w_{u_i1}\right)$ to represent the number of words occurring in the researcher profile;

8: **End for**

9: **For** i=1 to M **do**

10: **For** j=1 to N **do**

11: **If** article v_j has not read by researcher u_i

12: Compute $sim\left(u_i, v_j\right)$ using equation(3);

13: **End if**

14: **End for**

15: **End for**

16: **Return** $sim\left(u, v\right)$;

Fig. 2. The algorithm of improved CBF with social tag information.

$$P\left(V\middle|\sigma_V^2\right)=\prod_{j=1}^{N}N\left(V_j\middle|0,\sigma_V^2I\right) \tag{7}$$

Hence, through a Bayesian inference, we have

$$\begin{aligned}&P\left(U,V|R,S,\sigma_R^2,\sigma_U^2,\sigma_V^2,\sigma_s^2\right)P\left(R|U,V,\sigma_R^2\right)P\left(V|\sigma_V^2\right)P\left(U|S,\sigma_U^2,\sigma_s^2\right)\\&=\prod_{i=1}^{M}\prod_{j=1}^{N}\left[N\left(R_{i,j}\middle|g\left(U_i^TV_j\right),\sigma_R^2\right)\right]^{I_{i,j}^R}\times\prod_{j=1}^{N}N\left(V_j\middle|0,\sigma_V^2I\right)\\&\times\prod_{i=1}^{M}N\left(U_i\middle|0,\sigma_U^2I\right)\times\prod_{i=1}^{M}N\left(U_i\middle|\sum_{k\in N(i)}S_{i,k}U_k,\sigma_S^2I\right)\end{aligned} \tag{8}$$

This equation specifies the method on how to derive the researchers' latent feature space or researchers' characteristics based on the researcher-article matrix with considering the favors of the researchers' social friends. The log of the posterior distribution for the recommendations is given by:

$$\begin{aligned}&\ln P\left(U,V\middle|R,\mathrm{S},\sigma_U^2,\sigma_V^2,\sigma_R^2,\sigma_S^2\right)\\&=-\frac{1}{2\sigma_R^2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{i,j}^R\left(R_{i,j}-g\left(U_i^TV_j\right)\right)^2\\&-\frac{1}{2\sigma_S^2}\sum_{i=1}^{M}\left(U_i-\sum_{k\in N(i)}T_{i,k}U_k\right)^T\left(U_i-\sum_{k\in N(i)}T_{i,k}U_k\right)\\&-\frac{1}{2\sigma_U^2}\sum_{i=1}^{M}U_i^TU_i-\frac{1}{2\sigma_V^2}\sum_{j=1}^{N}V_j^TV_j-\sum_{i=1}^{M}\sum_{j=1}^{N}I_{i,j}^R\ln\sigma_R^2\\&-D\sum_{i=1}^{M}\ln\sigma_U^2-D\sum_{j=1}^{N}\ln\sigma_V^2+\mathrm{P}\end{aligned} \tag{9}$$

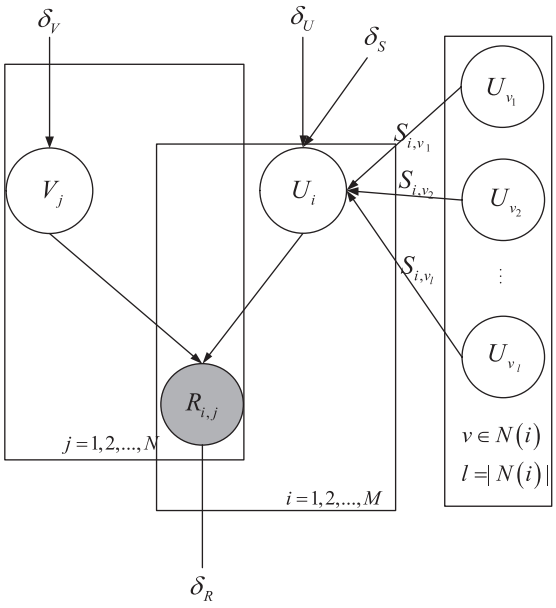

Fig. 3. Graph model for improved PMF with social friend information.

where P is a constant that does not depend on the parameters. Maximizing the log-posterior over two latent features with hyper parameters kept fixed is equivalent to minimizing the following sum-of-squared-errors objective functions with quadratic regularization terms:

$$\begin{aligned} E(U,V,R,S) &= \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} I_{i,j}^{R}\left(R_{i,j} - g\left(U_i^T V_j\right)\right)^2 \\ &+ \frac{\theta_S}{2}\sum_{i=1}^{M}\left(U_i - \sum_{k\in N(i)} T_{i,k}U_k\right)^T\left(U_i - \sum_{k\in N(i)} T_{i,k}U_k\right) \\ &+ \frac{\theta_U}{2}\sum_{i=1}^{M} U_i^T U_i + \frac{\theta_V}{2}\sum_{j=1}^{N} V_j^T V_j \end{aligned} \tag{10}$$

where $\theta_u = \frac{\sigma_R^2}{\sigma_u^2}, \theta_v = \frac{\sigma_R^2}{\sigma_v^2}, \theta_s = \frac{\sigma_R^2}{\sigma_s^2}$, which reflects the influence degree of each matrix on the objective function. A local minimum of the objective function given by equal (10) can be found by performing gradient descent in U_i, V_j,

$$\begin{aligned} \frac{\partial E}{\partial U_i} &= \sum_{j=1}^{N} I_{i,j}^{R}\left(g\left(U_i^T V_j\right) - R_{i,j}\right) g'\left(U_i^T V_j\right) V_j + \theta_U U_i \\ &+ \theta_S\left(U_i - \sum_{k\in N(i)} S_{i,k}U_k\right) - \theta_S \sum_{\{k|k\in N(i)\}} S_{i,k}\left(U_k - \sum_{w\in N(k)} S_{k,w}U_w\right) \end{aligned} \tag{11}$$

$$\frac{\partial E}{\partial V_j} = \sum_{i=1}^{M} I_{i,j}^{R}\left(g\left(U_i^T V_j\right) - R_{i,j}\right) g'\left(U_i^T V_j\right) U_i + \theta_V V_j \tag{12}$$

where $N(k)$ is the set that includes all the researchers who are friends of researcher u_k and $g'(x)$ is the derivative of logistic function $g'(x) = exp(x)/(1+exp(x))^2$.

For each researcher, after getting all prediction ratings based on the latent characteristic matrix U and V, the top N articles which he has not read are selected. In the same way, for the articles with the same rating, they are selected randomly. This approach predicts how the target researcher would rate an article which he has not read according to their previously rated articles. Therefore, the higher the rating obtained by an article, the more likely the target researcher is interested in it. Fig. 4 shows the algorithm of the improved CF recommendation approach.

3.4. The hybrid approach for article recommendation in SSN

Although the CBF and CF are widely used to making article recommendations, they may not present the best recommendation results due to their individual limitations. Therefore, in this research, due to the fact that hybrid approach can provide more accurate recommendation than a single approach and the disadvantages of one approach can be overcome by the other approach, the two approaches which have been improved with social information are combined into a novel hybrid recommendation approach, i.e., HRA-SI. Data combination has also been widely investigated in the recommendation community. They were often divided into two categories: score-based and ranking-based [55,56]. Score-based combination methods require similarity information to conduct ranking list aggregation, such as CombSum, CombMNZ and weight combination [55,57]. Ranking-based combination methods require rank or position information to integrate different candidate ranking lists, such as Borda fusion, Condorcet fusion and MAP-Fuse [56,58,59].

In this research, the weight combination method is applied to integrate the existing prediction lists generated by the improved CBF approach the improved CF approach. To be more specific, the predictions resulting from the improved CBF and the improved CF are firstly produced separately with the purpose of allowing us to leverage the individual strengths of the both approaches, since there is no interdependency between them, then weightily combined with a set of adjustment parameters α from 0.1 to 0.9 as follows:

$$hybrid_{i,j}\_prediction = (1-\alpha)\mathrm{CBF}_{i,j}\_prediction + \alpha CF_{i,j}\_prediction \tag{13}$$

where the $hybrid_{i,j}\_prediction$ is a hybrid personal prediction for the target researcher u_i with respect to article v_j, $\mathrm{CBF}_{i,j}\_prediction$ is the prediction for the researcher u_i derived from the improved CBF approach and $CF_{i,j}\_prediction$ is the prediction for the researcher u_i derived from the improved CF approach.

In equal (13), the improved CBF's recommendations and the improved CF's recommendations are smoothed by the parameter α, which naturally fuses appropriate weight of these two approaches into the recommender systems. It is higher if the improved CF approach has a greater influence than the improved CBF approach. That is, the improved CF approach contributes more than the improved CBF approach to the recommended results of HAR-SI. Conversely, the improved CBF has a greater influence if the value of α is smaller. The parameter α controls how much impact the two approaches function on the final article recommendation results. Finally, only top M articles are recommended to a researcher based on the $hybrid_{i,j}\_prediction$.

4. Experimental design

This section presents the experiments of the proposed hybrid approach to evaluate the effectiveness of it. First two parts present the experimental datasets and evaluation metrics that were used for evaluating the proposed approach. Then the compared methods and experimental procedure are explained in details in the last part.

Input: observed researcher-article matrix R and social friend matrix S, number of iteration I, latent feature dimension D, regularization parameter θu, θv, θs and learning rate γ.

Output: researcher latent feature matrix U and article latent feature matrix V.

1: Initialize U and V randomly;

2: **For** iteration =1 to I **do**

3: **For** each <i,j>∈R **do**

4: Compute $\frac{\partial E}{\partial U_i}$ and $\frac{\partial E}{\partial V_j}$ using equation (11) and (12) with current U and V;

5: Update $U_i = U_i - \gamma \frac{\partial E}{\partial U_i}$;

6: Update $V_j = V_j - \alpha \frac{\partial E}{\partial V_j}$;

7: **End for**

8: **End for**

9: **Return** U and V;

Fig. 4. The algorithm of improved CF with social friend information.

4.1. Experimental dataset

In order to evaluate the proposed HAR-SI, the CiteUlike dataset was investigated in this research [1]. CiteULike is one of the leading SSN with an established social tagging and friend making system for researchers to find, storage, managing, and sharing academic articles online [60]. In the website, researchers can add scientific articles which they are interested in into their libraries and use tags to easily comment to any existing articles. Moreover, the CiteULike allows researchers with similar interests, hobbies and research field to establish friend relationship and communicate scientific research by creating research groups or join others' groups with specific research topics. Therefore, CiteULike dataset is suitable for experimental study of this research.

For the sake of obtaining required data, the CiteULike website was visited by using a crawler and all article content and social information, such as article title, tag, friend relation, and browsing history, were extracted from November 2005 to May 2013. Then data cleaning work were conducted and researchers with at least 15 article browsed records and articles which had been read by two or more researchers were chosen. Finally, a dataset of 2065 researchers, 85,542 articles with 231,604 researcher-articles pairs, and 46,391 tags was obtained. Table 2 is the description of the used dataset statistics.

Table 2
Statistics of filtered CiteULike dataset.

Description	Value
Number of researchers	2065
Number of articles	85,542
Number of researcher-articles pairs	231,604
Number of tags	46,391
Number of tag assignments	326,792
Number of groups	582
Number of group memberships	2261
Number of overall density (%)	0.131

4.2. Evaluation metrics

To compare the prediction accuracy of the proposed approach, precision and recall, which are widely used in recommendation systems to evaluate the recommendation quality, were employed in this research [48,61,62]. Precision is determined by the fraction of recommended articles that are actually interesting to a researcher, defined by Eq. (14). Recall is the fraction of interesting articles that have been accurately identified by the recommendation approach, illustrated by Eq. (15). The metrics are given as:

$$Precision@M = \frac{Number\ of\ correctly\ recommended\ articles\ in\ top\ M}{Number\ of\ recommended\ articles} \quad (14)$$

$$Recall@M = \frac{Number\ of\ correctly\ recommended\ articles\ in\ top\ M}{Number\ of\ collected\ articles} \quad (15)$$

where "Number of correctly recommended articles" refers to the number of articles included in both recommendation list derived from recommendation approach and the testing data. And "Number of recommended articles" is the size of a recommendation list while "Number of collected articles" is the size of a testing data. In this study, precision will be used in evaluating the proportion of researchers' interests towards the research article recommendation of our proposed HAR-SI. This proportion examines how correct the HAR-SI is in recommending articles of interest to researchers. On the other hand, recall estimates the ability that the HAR-SI approach can learn and also provide the right recommendation among the articles of interest for researchers, that is, it examines

the degree that the HAR-SI approach is able to satisfy the needs of researchers.

4.3. Experimental procedure

For the performance comparison in this research, some existing article recommendation methods in the literature were implemented. They are listed as follows:

(1) Term Frequency-Inverse Document Frequency, i.e., TF-IDF [63]: This approach uses TF-IDF value of keywords to represent researcher profiles and article profiles. Then article candidates are ranked by the calculated cosine similarity between researcher profiles vectors and article profiles vectors.
(2) Language Modeling, i.e., LM [64]: This approach converts queries collection of words, then arranges the words into sequences and solves data sparsity problem using smoothing data method.
(3) Personalized Research Paper Recommendation System, i.e., PRPRS [23]: This approach is one of the stat-of-the-art article recommendation approaches represented by Hong et al. We have reviewed it in Literature review.
(4) Singular Value Decomposition, i.e., SVD [65]: This approach uses the method of matrix singular values decomposition algorithm of matrix based only on user-item rating matrix to find a low-dimension model that maximizes the likelihood of observed ratings in recommendation systems.
(5) Probabilistic Matrix Factorization, i.e., PMF [66]: This is an approach for factorizing the user-item rating matrix using low-rank representations, and then utilizes them to make further predictions.
(6) Semantic-Social Aggregation Recommendation, i.e., SSAR [37]: This approach is one of the stat-of-the-art article recommendation approaches represented by Sun et al. which has been reviewed in Literature review.
(7) Common Author relation-based Recommendation, i.e., CARE [67]: This approach is proposed by Xia et al. [51], in which the common author relations between articles (i.e., two articles with the same author(s)) are used for the improvement of article recommendation.

Among these seven approaches, the first three represent baseline CBF approach, the following two represent CF approach and the last two represent hybrid approach. The collected data was used to construct researcher-article matrix and then divided into training dataset and testing dataset randomly. The 80% of researcher-article matrix was then used as the training dataset, while the remaining researcher-article preference matrix was used as the test dataset. Since each column in a matrix corresponds to articles, set of articles in test matrix was disjoint from those in the training matrix. Information in the training matrix together with content and social information was used to build the article profiles and researcher profiles, then different recommendation models were employ to make predictions for each researcher. To ensure experimental results were not sensitive to division of each dataset, the experiments were performed 10 times, each time we used different random splits. The parameter settings of our approach were $\alpha = 0.3$, $\theta_U = \theta_V = 0.001$, $\theta_S = 0.005$.

5. Result and discussions

5.1. Experimental results

In this section, the values for evaluation metrics were obtained and compared between the state-of-the-art approaches and our proposed approach on the CiteULike dataset. The detailed results are shown in Table 3. Table 3 gives the performance in terms of evaluation metric for Top @M ($M = 10, 20, 30, 40, 50$) obtained by the standard CBF, the improved CBF, the standard CF, the improved CF, the proposed HAR-SI and other state-of-the-art approaches. The best results are marked in bold.

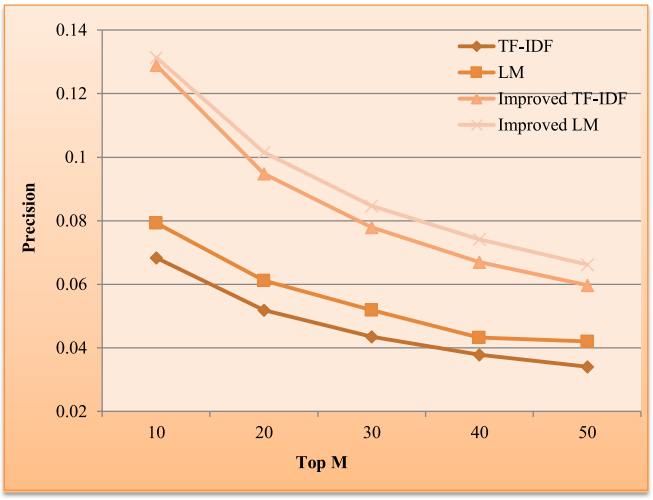

Fig. 5. CBF recommendation performance in terms of precision.

Generally speaking, as shown in Table 3, although with increased recommended numbers M, the influence on different methods seems very alike with a trend of a decline on precision and an ascent on recall to a certain degree. This result is in line with the widely existence experimental results of the recommendation system, that is, the more the recommended number, the predict result is often the lower the precision, the higher the recall. Moreover, it is evident from Table 3 that no matter when recommend 10, 20, 30, 40 or 50 articles, the proposed HAR-SI can all outperform the best performed baseline method on both precision and recall at an improvement of more than 10%. And the best performance in terms of the precision is achieved by our proposed HAR-SI method to 0.1593 at a recommendation number of @10. For CBF approach, the improved LM approach with social tag information gets better performance and achieves its best result to 0.1314 at precision and 0.2549 at recall. For CF approach, the improved PMF approach with social friend information achieves better performance with its highest precision and recall to 0.0844 and 0.1785 respectively. Moreover, except recall@20, recall@30, recall@40 and recall@50 of LM approach, all approaches improved with social information in SSN achieves a higher performance of more than 50% compared with the standard approaches.

5.2. Discussions

5.2.1. Results analysis of content based filtering approach

As mentioned above, the CBF approach was compared in two forms: with title only, and title along with tags. Thus, in this section, the effect of the improved CBF approach with tags as well as title was evaluated by comparing its recommendation quality to those of the standard CBF approach merely with tags. The compared results are shown in Figs. 5 and 6. It can be easily observed from these two figures that no matter for LM or TF-IDF, our improved approaches with social tag information have made significant improvement on both evaluation metrics over the standard approaches with text information only. This result shows that social tags, as additional information, play a significant role in the article recommendation when determine which article is interested to a certain researcher since the improved approaches achieves a performance improvement of more than 57% relative to the standard approach in precision. It implies the natural advantage of so-

Table 3
Precision and recall results for proposed HAR-SI and other compared approaches.

Approach		Precision					Recall				
		@10	@20	@30	@40	@50	@10	@20	@30	@40	@50
Standard CBF	TF-IDF	0.0683	0.0519	0.0435	0.0379	0.0341	0.0701	0.1008	0.1224	0.1406	0.1537
	LM	0.0793	0.0612	0.0519	0.0433	0.0421	0.0752	0.1124	0.1373	0.1579	0.1752
Improved CBF	TF-IDF	0.1289	0.0948	0.0779	0.0670	0.0597	0.1113	0.1572	0.1881	0.2120	0.2316
	LM	**0.1314**	**0.1016**	**0.0847**	**0.0741**	**0.0662**	**0.1246**	**0.1683**	**0.2035**	**0.2317**	**0.2549**
Standard CF	SVD	0.0467	0.0355	0.0248	0.0233	0.0157	0.0496	0.0574	0.0675	0.0835	0.0946
	PMF	0.0543	0.0418	0.0363	0.0341	0.0302	0.0595	0.0729	0.0873	0.0991	0.1124
Improved CF	SVD	0.0741	0.0668	0.0579	0.0522	0.0483	0.0944	0.1047	0.1190	0.1329	0.1458
	PMF	**0.0844**	**0.0721**	**0.0689**	**0.0612**	**0.0574**	**0.1049**	**0.1254**	**0.1467**	**0.1641**	**0.1785**
PRPRS		0.0801	0.0729	0.0683	0.0599	0.0571	0.1032	0.1263	0.1458	0.1628	0.1783
SSAR		0.1355	0.0997	0.0915	0.0811	0.0753	0.1334	0.1825	0.2109	0.2476	0.2668
CARE		0.1416	0.1148	0.0987	0.0892	0.0814	0.1486	0.1994	0.2335	0.2650	0.2911
HAR-SI		**0.1593**	**0.1271**	**0.1105**	**0.0993**	**0.0910**	**0.1617**	**0.2185**	**0.2591**	**0.2829**	**0.3110**

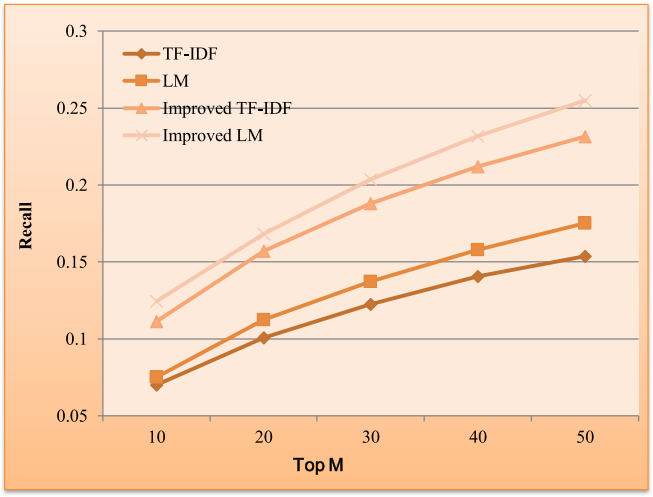

Fig. 6. CBF recommendation performance in terms of recall.

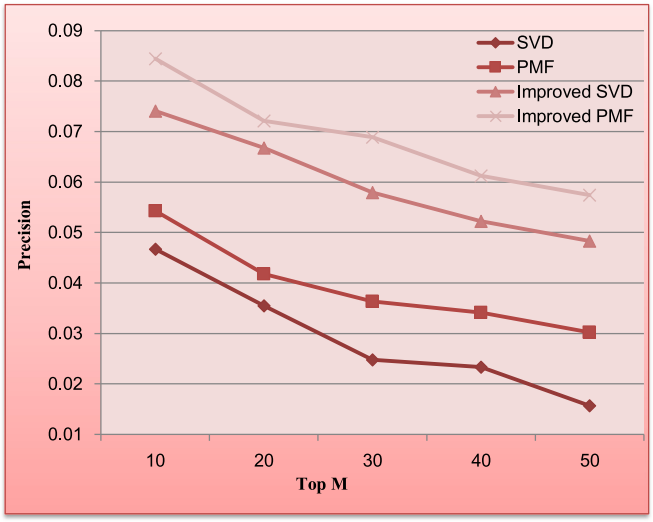

Fig. 7. CF recommendation performance in terms of precision.

cial tag information in characterization of researcher preference as well as article feature. Besides that, when further comparing the approaches in CBF, LM method can achieve higher performance, retrieve more useful articles in the recommendation list than TF-IDF method. It proved our inference illustrated in Section 3.2 that the LM can achieve outstanding performance due to its inner strengths, which is consistent with previous research [49,50]. Even more remarkable, although the standard LM approach performs better than the standard TF-IDF approach, after incorporating social information, the improved TF-IDF method surpasses and gains better results no matter on precision or recall than the standard LM method. This further proves the effectiveness of social information used in our proposed HAR-SI method for paper recommendation in SSN.

5.2.2. Results analysis of collaborative filtering approach

In this section, the effect of the improved CF approach was evaluated by comparing the standard CF methods, SVD method, and PMF method. Figs. 7 and 8 depict the performance improvement of our improved CF combined with social friend information against the standard CF, which does not use any additional information in the stage of probabilistic matrix factorization. It is clear that the proposed HAR-SI outperforms in terms of both the metrics to a great extent, which proves the effectiveness of social information in helping overcome the problems existed in CF, since it has gained a performance improvement of more than 50% over the best competitive approach of standard CF. This is owing to the fact that users are not independent and identically distributed in the real world. They not only have their own characteristics, but also can be easily influenced by their friends and prefer their friends' recommendations [38,39]. Thus, incorporating their social friend information can effectively enhance the article recommendation quality which has also been proved in our experiments. Besides that, similar to the above analysis in Section 5.2.1, the improved SVD method also achieves surpassing and gains better performance than the standard PMF method on both of the metrics.

5.2.3. Results analysis of hybrid recommendation approach

In this section, the results of our proposed novel hybrid article recommendation approach with different parameters setting were reported. Figs. 9 and 10 give the results of precision and recall under $\alpha = 0.3$ respectively. It is clear that the proposed HAR-SI significantly outperforms the other five comparison benchmark approaches in terms of both the two metrics, which proves the effectiveness of HAR-SI in overcoming the limitations of the individual approaches at the same time keeping their recommendation advantages. One the one hand, with the value of M increases, it is evident from the Figs. 9 and 10 that the precision is reducing significantly while the recall is rising to a certain degree.

Furtherly, according to the results presented in Figs. 9 and 10, the CBF approach has a higher impact on the hybrid recommendation approach, which also corresponds with the experimental conclusion proven by Mooney et al. [52] and Linden et al. [29]. This

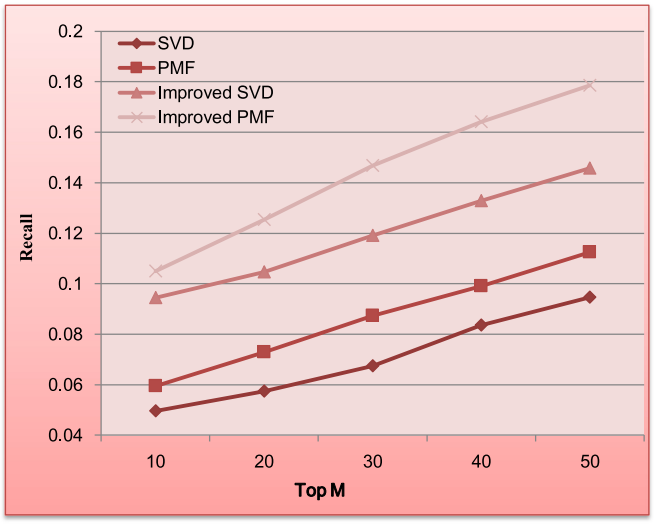

Fig. 8. CF recommendation performance in terms of recall.

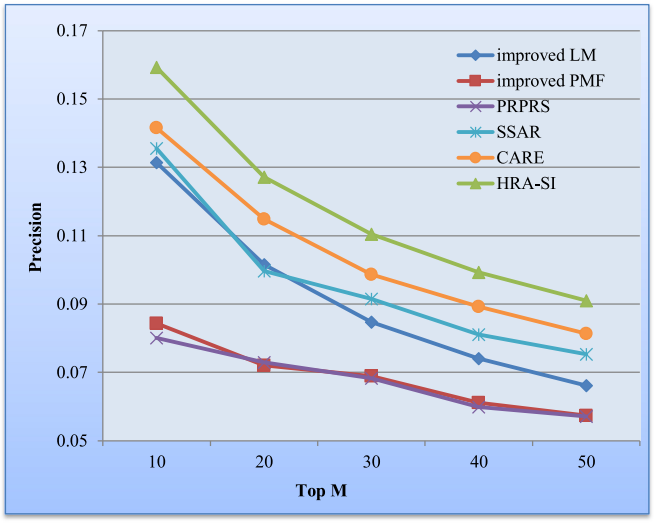

Fig. 9. Improved CBF, CF and proposed HAR-SI approach recommendation performance in terms of precision.

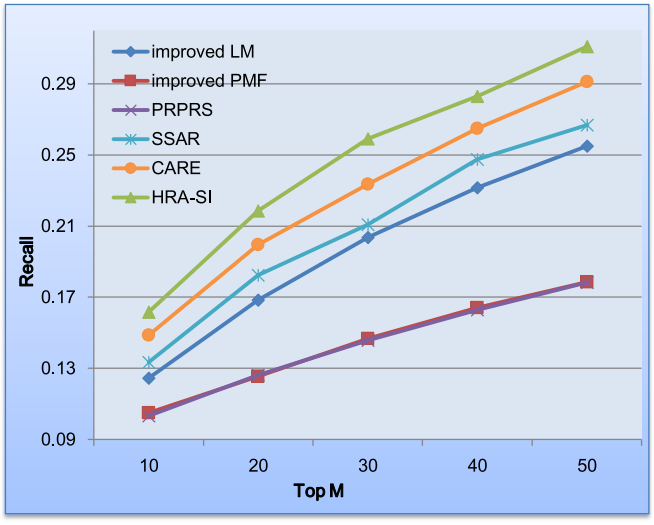

Fig. 10. Improved CBF, CF and proposed HAR-SI approach recommendation performance in terms of recall.

is that CBF is more effective in locating textual documents relevant to a topic, whereas CF is more popular with e-tailors that sell products such as in Amazon.com. Therefore, the CBF shown more competitive performance because in the recommendation of articles in SSN, the textual information plays a vital role. In addition, the tag information also improved the performance of CBF to a great extent. Please note that although compared with the CBF, the contribution of the CF is less prominent in the hybrid article recommendation approach. It is still essential to consider the impact of the CF on finding useful articles for the target researcher since it can provide recommendations of articles that are unable be considered by the CBF for its inner limitations. Thereby, giving a proper weight to the CBF and the CF approach in obtaining the hybrid recommendations of HAR-SI is necessary.

To combine the two prediction results of the improved CBF and the improved CF in HAR-SI, the adjustment parameter α is utilized to adjust the relative importance between them. In order to determine the optimal value for α, several experiments have been conducted to systematically adjust the values of α in an increment of 0.1 under the different recommended number M, as shown in Fig. 11. The value of the metrics for a particular Top-M was denoted as prec@M and rec@M. According to the experiment results, HAR-SI has the highest precision and recall when α is 0.3. This means that the relative importance for the improved CBF and the improved CF is 0.7 and 0.3, respectively. With the value of α increase, both of the two metrics of HAR-SI improve until reach their best performance when α is around 0.3. After that, an increase in value of α brings about a performance reduction of the HAR-SI recommendation instead.

Generally, it can be inferred that the social information available in SSN is significantly helpful in improving the performance of recommendation approaches. On the other hand, combining CBF approach and CF approach into a hybrid approach is very effective to improve recommendation accuracy for articles on SSN platforms.

6. Conclusions and future work

In this research, a novel hybrid approach integrating with social information in SSN was proposed for article recommendation. The social tag information was used into the CBF approach to build profiles for researchers and articles. At the same time, the social friend information was utilized into the CF approach to help conduct unified probabilistic matrix factorization. Subsequently, the results of the two improved approaches were combined with a weight of α, and a further application to article recommendation in SSN was made. The experimental results on real scientific social network dataset, i.e., CiteULike, proved that comparing with the baseline recommendation approaches, the proposed hybrid approach got the best recommendation results. With the better recommendation approaches, researchers will benefit from time saved on web browsing, increasing access to valuable information wherever needed and up to date developments in their field.

Several future research directions also exist for this research. Firstly, according to the results obtained in this research, it is clear that additional social information on the SSN can greatly improve recommendation accuracy. Thus, in the future, this work will be extended by considering incorporating other more available additional information in the social network, such as time factor and the quality of the scientific articles (citations, journal impact factor). Secondly, this research used a simple method of weighting as the hybrid approach. More complex ways of combining individual approaches can also be implemented to produce better results. Thirdly, in this research, only one type of SSN, i.e., CiteUlike dataset, was employed to verify the proposed hybrid approach. Other platforms such as ResearchGate, ScolarMate, and

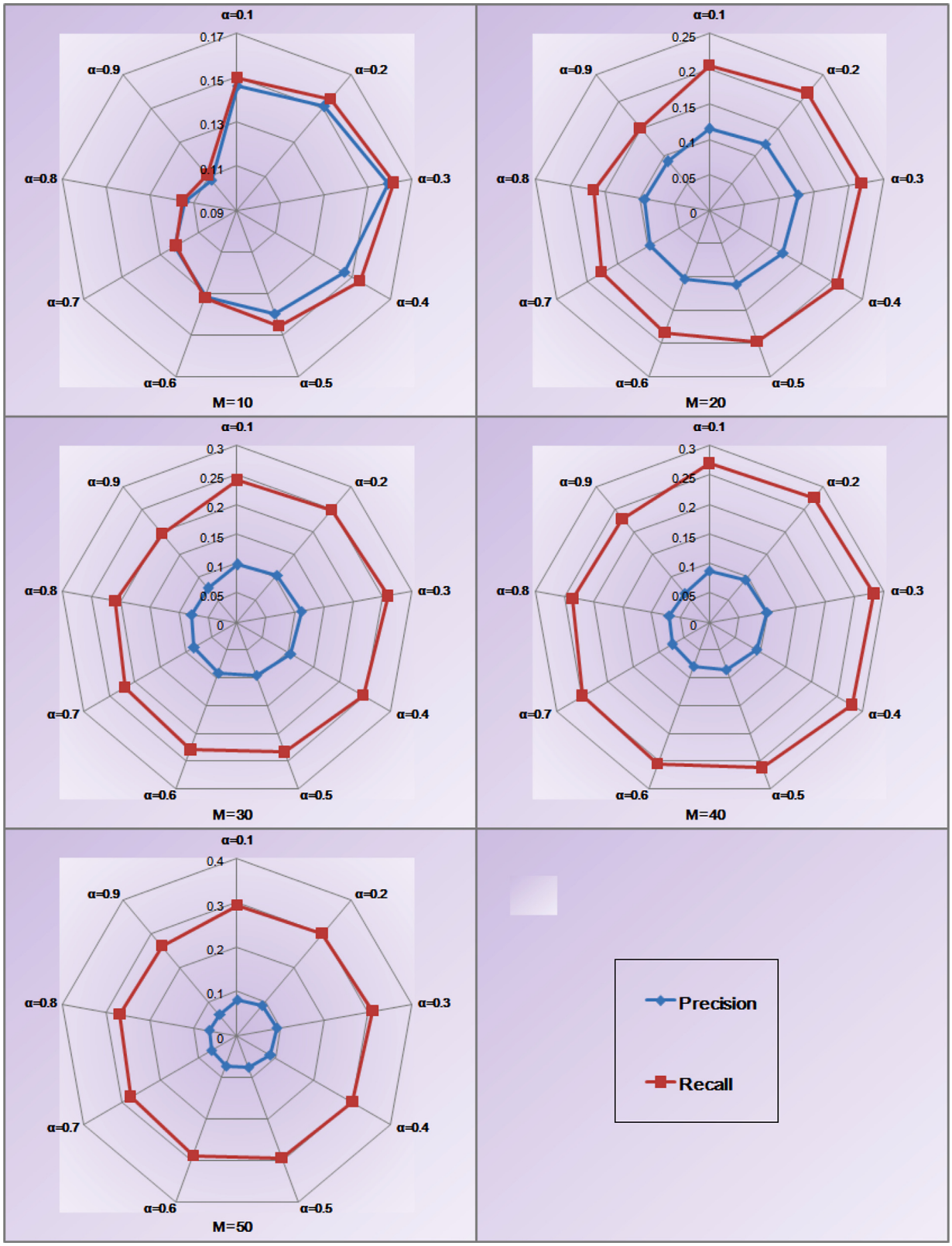

Fig. 11. The proposed HAR-SI approach recommendation performance in terms of precision and recall.

Academia.edu also have much additional information that might help enhance the recommendation performance. We will extend our approach to these other application platforms and further explore the combination of heterogeneous data sources to improve the recommendation applicability and generality. Finally, considering the characteristics of SSN that compared to other social networks with scoring, such as MovieLens, the data on SSN are mostly one class data which can only show the positive information. Thus, the next, we will also from the perspective of One Class Collabora-

tive Filtering (OCCF) to further explore the recommendation problems in SSN.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (91646111, 71471054), Anhui Provincial Natural Science Foundation (1608085MG150), Social Science Knowledge Popularization Foundation of Anhui Province (Y2016016), Training Program of Application of Scientific and Technological Achievement of HeFei University of Technology (JZ2017YYPY0235).

References

[1] C. Wang, D.M. Blei, Collaborative topic modeling for recommending scientific articles, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011.
[2] A.S. Vivacqua, J. Oliveira, J.M. De Souza, i-ProSE: inferring user profiles in a scientific context, Comput. J. 52 (7) (2009) 789–798.
[3] C.-N. Ziegler, S.M. McNee, J.A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th International Conference on World Wide Web, ACM, 2005.
[4] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 734–749.
[5] Q. Li, S.H. Myaeng, B.M. Kim, A probabilistic music recommender considering user opinions and audio features, Inf. Process. Manage. 43 (2) (2007) 473–487.
[6] S. Kangas, Collaborative Filtering and Recommendation Systems, VTT Information Technology, 2002.
[7] T.-P. Liang, Y.-F. Yang, D.-N. Chen, Y.-C. Ku, A semantic-expansion approach to personalized knowledge recommendation, Decis. Support Syst. 45 (3) (2008) 401–412.
[8] K. Chandrasekaran, S. Gauch, P. Lakkaraju, H.P. Luong, Concept-based document recommendations for citeseer authors, International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Springer, 2008.
[9] C. Basu, H. Hirsh, W.W. Cohen, C.G. Nevill-Manning, Technical paper recommendation: a study in combining multiple information sources, J. Artif. Intell. Res. 14 (2001) 231–252.
[10] G.H. Martín, S. Schockaert, C. Cornelis, H. Naessens, Using semi-structured data for assessing research paper similarity, Inf. Sci. 221 (2013) 245–261.
[11] W. Wang, G. Zhang, J. Lu, Member contribution-based group recommender system, Decis. Support Syst. 87 (2016) 80–93.
[12] S.M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S.K. Lam, A.M. Rashid, J.A. Konstan, J. Riedl, On the recommending of citations for research papers, in: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, ACM, 2002.
[13] T. Bogers, V.D.B. Antal, Recommending scientific articles using CiteULike, ACM Conference on Recommender Systems, Recsys 2008, Lausanne, Switzerland, October, 2008.
[14] Q. Zhang, D. Wu, J. Lu, F. Liu, G. Zhang, A cross-domain recommender system with consistent information transfer, Decis. Support Syst. 104 (2017) 49–63.
[15] M. Al-Hassan, H. Lu, J. Lu, A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system, Decis. Support Syst. 72 (2015) 97–109.
[16] G. Özbal, H. Karaman, F.N. Alpaslan, A content boosted collaborative filtering approach for movie recommendation based on local & global similarity and missing data prediction, Comput. J. 54 (9) (2010) 109–112.
[17] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 734–749.
[18] S.-Y. Hwang, C.-P. Wei, Y.-F. Liao, Coauthorship networks and academic literature recommendation, Electron. Commerce Res. Appl. 9 (4) (2010) 323–334.
[19] W. Zhao, R. Wu, H. Liu, Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target, Inf. Process. Manage. 52 (5) (2016) 976–988.
[20] M. Mao, J. Lu, G. Zhang, J. Zhang, Multirelational social recommendations via multigraph ranking, IEEE Trans. Cybern. 47 (12) (2017) 4049–4061.
[21] C. Nascimento, A.H. Laender, A.S. da Silva, M.A. Gonçalves, A source independent framework for research paper recommendation, in: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, ACM, 2011.
[22] R.M. Nallapati, A. Ahmed, E.P. Xing, 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, ACM, 2008.
[23] K. Hong, H. Jeon, C. Jeon, Personalized research paper recommendation system using keyword extraction based on userprofile, J. Converg. Inf. Technol. 8 (16) (2013) 106.
[24] K.D. Bollacker, S. Lawrence, C.L. Giles, Discovering relevant scientific literature on the web, IEEE Intell. Syst. Appl. 15 (2) (2000) 42–47.
[25] C.C. Chen, M.C. Chen, Y. Sun, PVA: A self-adaptive personal view agent, J. Intell. Inf. Syst. 18 (2-3) (2002) 173–194.
[26] A. Vellino, A comparison between usage-based and citation-based methods for recommending scholarly research articles, in: Proceedings of the American Society for Information Science and Technology, 47, 2010, pp. 1–2.
[27] D.H. Lee, P. Brusilovsky, Using self-defined group activities for improving recommendations in collaborative tagging systems, in: Proceedings of the Fourth ACM Conference on Recommender Systems, ACM, 2010.
[28] C.H. Lai, D.R. Liu, C.S. Lin, Novel personal and group-based trust models in collaborative filtering for document recommendation, Inf. Sci. 239 (4) (2013) 31–49.
[29] G. Linden, B. Smith, J. York, Amazon.com recommendations: item-to-item collaborative filtering., IEEE Internet Comput. 7 (1) (2003) 76–80.
[30] S.T. Park, D.M. Pennock, Applying collaborative filtering techniques to movie search for better ranking and browsing, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August, 2007.
[31] A.S. Das, M. Datar, A. Garg, S. Rajaram, Google news personalization: scalable online collaborative filtering, International Conference on World Wide Web, 2007.
[32] Van Setten, M., Supporting people in finding information: hybrid recommender systems and goal-based structuring. 2005.
[33] H.K. Kim, H.Y. Oh, J.C. Gu, J.K. Kim, Commenders: a recommendation procedure for online book communities, Electron. Commerce Res. Appl. 10 (5) (2011) 501–509.
[34] A. Kodakateri Pudhiyaveetil, S. Gauch, H. Luong, J. Eno, Conceptual recommender system for CiteSeerX, in: Proceedings of the Third ACM Conference on Recommender Systems, ACM, 2009.
[35] T. Yuan, J. Cheng, X. Zhang, Q. Liu, H. Lu, Enriching one-class collaborative filtering with content information from social media, Multimedia Syst. 22 (1) (2016) 51–62.
[36] S.-S. Weng, H.-L. Chang, Using ontology network analysis for research document recommendation, Expert Syst. Appl. 34 (3) (2008) 1857–1869.
[37] J. Sun, J. Ma, Z. Liu, Y. Miao, Leveraging content and connections for scientific article recommendation in social computing contexts, Comput. J. 57 (9) (2014) 1331–1342.
[38] H. Ma, I. King, M.R. Lyu, Learning to recommend with social trust ensemble, International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July, 2009.
[39] H. Ma, H. Yang, M.R. Lyu, I. King, SoRec: social recommendation using probabilistic matrix factorization, ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October, 2008.
[40] H. Ma, D. Zhou, C. Liu, M.R. Lyu, I. King, Recommender systems with social regularization, Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February, 2011.
[41] G. Wang, X. He, C.I. Ishuga, Social and content aware One-Class recommendation of papers in scientific social networks, PLoS One 12 (8) (2017).
[42] G. Leroy, H. Chen, J.D. Martinez, A shallow parser based on closed-class words to capture relations in biomedical text, J. Biomed. Inform. 36 (3) (2003) 145–158.
[43] C.D. Paice, A thesaural model of information retrieval, Inf. Process. Manage. 27 (5) (1991) 433–447.
[44] H.-C. Tu, J. Hsiang, An architecture and category knowledge for intelligent information retrieval agents, Decis. Support Syst. 28 (3) (2000) 255–268.
[45] J. Xu, Solving the Word Mismatch Problem Through Automatic Text Analysis, Citeseer, 1997.
[46] J. Xu, W.B. Croft, Query expansion using local and global document analysis, in: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1996.
[47] N. Fuhr, H. Hüther, Optimum probability estimation from empirical distributions, Inf. Process. Manage. 25 (5) (1989) 493–507.
[48] Salton, G. and M.J. McGill, Introduction to Modern Information Retrieval. 1986.
[49] L. Hong, B.D. Davison, Empirical study of topic modeling in twitter, in: Proceedings of the First Workshop on Social Media Analytics, ACM, 2010.
[50] X. Quan, G. Liu, Z. Lu, X. Ni, L. Wenyin, Short text similarity based on probabilistic topics, Knowl. Inf. Syst. 25 (3) (2010) 473–491.
[51] Z. Zheng, H. Ma, M.R. Lyu, I. King, Qos-aware web service recommendation by collaborative filtering, IEEE Trans. Serv. Comput. 4 (2) (2011) 140–152.
[52] R.J. Mooney, L. Roy, Content-based book recommending using learning for text categorization, in: Proceedings of the Fifth ACM Conference on Digital libraries, ACM, 2000.
[53] C. Froschl, User Modeling and User Profiling in Adaptive E-learning Systems Master Thesis, Graz, Austria, 2005.
[54] O. Kirmemis, A. Birturk, A content-based user model generation and optimization approach for movie recommendation, Workshop on ITWP, 2008.
[55] S. Wu, Applying the data fusion technique to blog opinion retrieval, Expert Syst. Appl. 39 (1) (2012) 1346–1353.
[56] M. Montague, J.A. Aslam, Condorcet fusion for improved retrieval, in: Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM, 2002.
[57] E.A. Fox, J.A. Shaw, Combination of Multiple Searches, NIST Special Publication SP, 1994 243-243.
[58] J.A. Aslam, M. Montague, Models for metasearch, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2001.

[59] D. Lillis, L. Zhang, F. Toolan, R.W. Collier, D. Leonard, J. Dunnion, Estimating probabilities for effective data fusion, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2010.
[60] K. Emamy, R. Cameron, Citeulike: a researcher's social bookmarking service, Ariadne 51 (51) (2007).
[61] Y. Cao, Y. Li, An intelligent fuzzy-based recommendation system for consumer electronic products, Expert Syst. Appl. 33 (1) (2007) 230–240.
[62] F.H. del Olmo, E. Gaudioso, Evaluation of recommender systems: a new approach, Expert Syst. Appl. 35 (3) (2008) 790–804.
[63] T. Joachims, in: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, Springer, US, 1997, pp. 143–151.
[64] S.F. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, Comput. Speech Lang. 13 (4) (2010) 359–393.
[65] S. Zhang, W. Wang, J. Ford, F. Makedon, J. Pearlman, Using singular value decomposition approximation for collaborative filtering, In: Seventh IEEE International Conference on, 2005. E-Commerce Technology, 2005. CEC, 2005, pp. 257–264.
[66] A. Mnih, R. Salakhutdinov, Probabilistic matrix factorization, International Conference on Machine Learning, 2012.
[67] F. Xia, H. Liu, I. Lee, L. Cao, Scientific article recommendation: exploiting common author relations and historical preferences, IEEE Trans. Big Data 2 (2) (2016) 101–112.