Leveraging Content and Connections for Scientific Article Recommendation in Social Computing Contexts

Jianshan Sun\textsuperscript{1,2, \ast}, Jian Ma\textsuperscript{2}, Zhiying Liu\textsuperscript{1} and Yajun Miao\textsuperscript{3}

\textsuperscript{1}School of Management, University of Science and Technology of China, Hefei, Anhui, PR China
\textsuperscript{2}Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
\textsuperscript{3}School of Public Affairs, University of Science and Technology of China, Hefei, Anhui, PR China

\ast Corresponding author: sunjs9413@gmail.com

Rapid proliferation of information technologies has generated a great volume of information that makes scientific information searching more challenging. Personalized recommendation is a widely used technique to help researchers find relevant information. Researchers involved in a social computing context generate abundant content and form heterogeneous connections. Existing article recommendation techniques fail to perform a deep analysis of this information. This research proposes a novel approach to recommend scientific articles to researchers by leveraging content and connections. In this approach, we first analyze the semantic content of the article by keyword similarity calculation and then extract online users' connections to support article voting and finally employ a two-stage recommendation process to suggest relevant articles. The proposed method has been implemented in ScholarMate (www.scholarmate.com), an online research social network platform. Two experiments are conducted and the evaluation results indicate that the proposed method is more effective than the baseline methods.

Keywords: article recommendation; social computing application; semantic content expansion; research social network

Received 15 December 2012; revised 24 June 2013
Handling editor: Guandong Xu

1. INTRODUCTION

The rapid proliferation of information technology, especially web 2.0, has led to a tremendous growth of online content. The presence of massive amounts of information has posed significant challenges for new information discovery especially in academia. Furthermore, online access to scholarly articles has become a common trend of scientific information searching. PubMed Central is a free full-text archive of scholarly articles in the biomedical field and currently holds over 1.7 million articles. But these articles represent only a tiny fraction of the estimated 50 million articles available on the web \cite{1}. In such an environment, there is an urgent requirement for an efficient and effective information acquisition technology which helps researchers to quickly find the relevant articles of interest.

The evolution of researchers’ information acquisition ways with the development of information technologies is shown in Fig. 1. As depicted it, traditionally, researchers in the print age found relevant articles by using library catalogs. This was ineffective due to the limited number of journal articles and books available in the library. Recently, online access services such as Google Scholar\textsuperscript{1} and Web of Science database\textsuperscript{2} have been able to provide a powerful search tool to find relevant articles. Digital libraries make academic resources online and provide a variety of convenient functions for effective information retrieval. For instance, keyword-based searching is a convenient and efficient approach to retrieve information. However, forming queries for finding new research articles can be difficult when researchers have no clear idea of what they are looking for.

\textsuperscript{1}scholar.google.com.
\textsuperscript{2}portal.isiknowledge.com.
Nowadays, with the spread of social network services such as CiteULike, ResearchGate, and ScholarMate, researchers can easily promote their own publications and bookmark others’ publications. Social network services provide a common platform for researchers to share research outcomes and communicate with each other. Recommendation technologies have been implemented in those social network sites to automatically suggest useful information to users based on their profiles. The accuracy of this profile-based recommendation is crucial to meet users’ satisfaction. In a social computing context, computing systems are interweaving with users' social behaviors. Users interact with each other (such as making friends, joining groups, etc.) and collectively generate online content (i.e. tags and reviews). Social computing applications ease information searching and generate more freedom to online users while bringing new challenges to information overload problems.

Article recommendation is a hot research topic and has been studied in different contexts. Collaborative filtering (CF) approaches have been adopted to recommend interesting articles, but they can hardly achieve the expected performance when the number of users is small and the number of articles is large. Content-based (CB) methods in article recommendation make use of the information based on researchers’ profiles and they perform better than CF approaches. Furthermore, keywords generated through a researcher’s query often have semantic ambiguities and such semantic relationships may also exist in the documents. Therefore, traditional keyword-based approaches suffer from a mismatch problem, i.e. the relevant documents which cannot exactly match the researcher’s query are discarded. In this paper, a keyword similarity matrix is used to deal with the mismatch problem.

To leverage the advantages of content-based (CB) approaches and collaborative filtering (CF) approaches and alleviate disadvantages of them, a hybrid recommendation approach, which is called the semantic-social aggregation approach, has been proposed to recommend relevant scientific articles in social computing contexts. The semantic content analysis and social connection analysis are integrated via a two-stage recommendation model to recommend highly relevant and socially endorsed articles. A prototype system for article recommendation has been implemented as an application service in ScholarMate. The proposed approach is evaluated through an offline experiment using the CiteULike dataset and an online user study in ScholarMate. The results show that the proposed approach outperforms the baseline methods in terms of recommendation accuracy.

There are three main contributions in this paper. First, the relevance of articles is analyzed from a novel and comprehensive perspective. Semantic content and online connections are deeply mined and aggregated to improve recommendation quality. Secondly, two experiments have been conducted to evaluate how the proposed semantic-social aggregation approach works in terms of improving the recommendation performance. Thirdly, a recommender system is developed to assist article recommendation in a social network environment. It supports content sharing and potential researcher collaboration services in ScholarMate.

The rest of the paper is organized as follows. In Section 2, we review the related work on article recommendation. In Section 3, we propose the semantic-social aggregation approach for scientific article recommendation. The implemented system is presented in Section 4. Section 5 introduces evaluation metrics and experimental study. Finally, we conclude the paper and point out the future research directions.

2. RELATED WORK

The personalization techniques provide the ability to tailor content and services to individuals based on their preferences and tastes. Recommender systems, as the major tool of personalization, match the potential interesting content with users’ expectations. In academic contexts, such systems are often used to recommend scientific information of interest to researchers which guarantee remarkable reduction of burdens in information searching. In general, there are two fundamental approaches to conduct recommendations: CB approach and CF approach. The CB approach recommends items that are similar to those in which the user has shown interest in the past. The CF approach recommends items to the user based on other like-minded individuals who have similar preferences or tastes as the target user. The CB approach is typically applied to recommend products that have identifiable content, including websites, textual documents, news, and books. The main tasks of the CB approach include: extracting semantic
features from item content, building user interest profiles and matching [12].

This paper is about article recommendation in online scientific communities; thus we focus on the related work about scientific document recommendation. McNee’s work [13] was one of the pioneering and important references on article recommendation. In that work, the recommender system recommended research papers considering human computer interactions from a user-centric perspective.

CB filtering methods have been proposed to discover interesting articles. Recent studies [4, 5, 14] showed that the CB approach had better performance than the CF approach in recommending documents. These methods often required one to extract personal preferences to build user profiles from document content [15]. User profiles could be constructed through either explicit declaration by users or observation of users’ actions. CiteSeer [16] as a successful digital library system was designed to perform information-filtering and knowledge-discovery functions for online users. It used researchers’ personal profiles to track and recommend relevant research articles. Basu et al. [17] viewed the task of assigning technical papers to conference reviewers as a problem of recommending technical papers to researchers based on their interests and backgrounds. The different information sources about papers and reviewers were combined together for recommendation with the help of the word-based information representation language (WHIRL) system [18]. The CB approach used in [19] calculated keyword scores based on their locations and frequencies within the text. The recent study by Martin et al. [20] employed semi-structured data (authors, journals and keywords) to assess research paper similarity and this similarity could be further employed to support article recommendation. However, the above profile representation techniques ignored keyword semantics. Thus they failed to address synonym and polysemy problems which caused keyword mismatch problems. Regarding inadequate information, the semantic-expansion method [4] and concept-based method [21] were proposed to enhance user profiling and achieve highly relevant recommendations. Liang et al. [4] developed a spread activation model to expand the user interest profile with the goal of improving recommendation performance. Based on the existing Association for Computing Machinery (ACM) classification system, trees of concepts were constructed in [21] and similarities between user profiles and document profiles were computed using a tree-edit distance measure. Most of these methods had the problem of generalizability as they used existing dictionaries and taxonomies such as WordNet and ACM Taxonomy. Moreover, existing profile expansion methods led to a match irrelevance problem as they had no refinement mechanisms. In this paper, we construct a keyword similarity matrix for semantic expansion and employ rank aggregation techniques to deal with the mismatch problem and the match irrelevance problem identified in content recommendation applications.

With the prevalence of social bookmarking and social networking websites, CF methods have attracted more attention in recent years. Boger and Bosch [2] applied the traditional CF approach on recommending scientific publications in the CiteULike database and the experiments showed that a user-based approach was better than an item-based approach due to the data distribution factors. Furthermore, three variants of user-based CF article recommendation algorithms were proposed and evaluated [2]. Owing to tag contribution, BM25-boosted CF achieved better performance over the other two CF methods (Classic CF and Neighbor-weighted CF). To overcome the ‘data sparsity’ problem, Vellino proposed the usage-based and citation-based methods for recommending research articles [22]. Besides collective relations involved in user-item pairs, other social relations were analyzed and incorporated into CF methods for improving performance. Pera and Ng [23] conducted publication recommendation considering social contacts while self-defined social groups were mined together with traditional CF techniques to recommend bookmarked articles in [24]. It is clear from studies in the literature that social relations can be used to improve the recommendation quality. In this paper, we extensively analyze online users’ social activities to generate three types of user connections (social, behavioral and semantic) and a flexible integration method is proposed to aggregate these online connections.

CF and CB approaches had unique advantages and disadvantages. Researchers tried to combine both techniques and generated hybrid ones to improve performance [25]. The main assumption for hybrid methods was that fusing the algorithms could provide more accurate recommendations than a single algorithm and disadvantages of each algorithm could be overcome by other algorithms. Hwang and Chuang [12] combined article content and web usage information for literature recommendation in the digital library context. The experimental results demonstrated that content data and usage data were complements of each other and hybrid methods achieved more accurate recommendations. In [26], purely tagging data (user-tag-item assignments) were mined and a graph-based learning algorithm was proposed for document recommendation. In [27], tagging information and content metadata information were considered and different fusion strategies for different algorithms were proposed. They found that fusing methods indeed produced significant improvements in recommendation accuracy. Kim et al. [14] from a different perspective constructed a collaborative user profile to enhance the content recommendation process. Although the above-mentioned studies combined CF and CB approaches to exploit the benefits of each other and lessen their disadvantages, they largely ignored the semantic features of content, and heterogeneous user relations were lacking thorough study. Our proposed approach provides a systematic and comprehensive analysis of the features and characteristics in social computing contexts. We will leverage semantic content analysis and connection-empowered voting to recommend relevant scientific articles.

THE COMPUTER JOURNAL, 2013
3. SEMANTIC-SOCIAL AGGREGATION RECOMMENDATION APPROACH

3.1. Overview of the proposed approach

Computational applications now are pushing the boundaries of personal computing and beginning to facilitate collaboration and social interactions. Social computing is ‘an area of information technology concerned with the intersection of human and social studies connected by computer networks’ [28]. In this context, heterogeneous information is generated by online users. For example, there are three coexisting objects (user, tag and item) that are linked together in a social tagging system. Users have social relations between each other and items contain rich content information. To better explain our proposed approach, we can represent social computing data in a unified graph as follows (Fig. 2).

As shown in Fig. 2, social computing data can be represented by a heterogeneous three-layer weighted directed hyper-graph in which nodes of each layer represent users, tags and items, respectively. In the user space, hyper edges are social relations among users corresponding to affiliation groups in social networks. Other information about users, such as friendships and collaboration can also be added (marked in black edges). Interlayer links (marked in dotted edges) are formed based on the tags. Tags can be used to assemble users and to describe item content. In the item space, the links between items (marked in dashed edges) denote the similarity between them. Different types of item information (e.g. keywords and categories) can be used to compute item level similarity. In our context, researchers represent users and articles represent items. Therefore, we use researchers and users interchangeably in this article.

In this study, we propose a hybrid recommender system that employs semantic expansion to analyze online content and leverages user connections to support article voting. Figure 3 shows the main components and procedures of the proposed article recommendation approach. Our proposed semantic-social aggregation approach includes three modules: ‘Semantic Content Filtering’ module, ‘Article Social Voting’ module and ‘Two-Stage Recommendation’ module. The details related to each module are described as follows.

3.2. Semantic content filtering

This module proposes the semantic keyword weighting method to determine the content relevance of candidate articles. Direct keyword weighting techniques (such as TFIDF and BM25) are widely employed to find relevant documents to a particular query. However, as keywords in a user query often have semantic meanings and certain semantic relationships may exist in certain documents, these techniques may result in a mismatch problem. To solve this problem, this paper proposes a semantic expansion...
method to represent the profile of researchers and candidate articles and match them. This module initially uses Natural Language Processing (NLP) procedures to pre-process articles and constructs Keyword–Article (KA) matrix in which matrix elements represent weighted term frequencies. Then similarities of keywords are calculated and articles are matched with the researcher’s profile. Sub-processes of semantic content filtering are discussed in detail in the following subsections.

3.2.1. NLP preparation

Documents and additional content features should be prepared to support keyword similarity computation. Using the classical NLP procedures (such as segmentation, stopping and stemming) [29], the documents and keyword features are presented by a KA matrix. The KA matrix is an $n_k \times n_a$ matrix to denote the association between keywords and articles, where $n_k$ is the number of keywords and $n_a$ is the number of articles. Traditionally, only author-assigned keywords in articles were considered and the entry of such a matrix $K_{ka}$ is the binary token (1 or 0) indicating whether the $k$th keyword occurs in the $a$th article or not. However, this limited set of keywords cannot capture the comprehensive content of the article since a low number of assigned keywords causes the KA matrix to be sparse, leading to less accurate keyword correlation scores. In order to overcome the accuracy problem, rich content in the title and abstract of the article is extracted and exploited. Moreover, in the social computing context, tagging information provides additional collective content descriptions. Some recent works have proved that tags were beneficial for document retrieval [30, 31]. In this paper, we take the frequency of keywords in the title and abstract as well as social tags into consideration in order to handle the keyword sparsity problem. The elements in the KA matrix denote weighted frequency scores (FS). The FS can be calculated as

$$FS_{ka} = \lambda \ast (f_{tit} + \beta f_{key} + \gamma f_{abs}) + (1 - \lambda) \ast f_{tag},$$

(1)

where $f_{tit}$ is the frequency of keywords in the article title, $f_{key}$ is the frequency of keywords in the author-assigned keywords list and $f_{abs}$ is the frequency of keywords in the article abstract. By $f_{tag}$ we denote how many users have tagged the article; $\lambda \in [0, 1]$ is the weight parameter to control the importance between self-description keywords and social tagging keywords. Here, we set $\lambda = 0.5$ for simplicity, which gives equal weight to objective information and social information. $\alpha$, $\beta$ and $\gamma$, subject to the equation $\alpha + \beta + \gamma = 1$, are weights of $f_{tit}$, $f_{key}$ and $f_{abs}$, respectively. To determine the parameter value of $\alpha$, $\beta$ and $\gamma$, an initial test was conducted as follows. First, 10 queries were submitted and relevant scientific publications were retrieved. Secondly, three types of keyword weighting methods (namely, $f_{tit}$, $f_{key}$ and $f_{abs}$) were employed to return relevant documents. Thirdly, the number of relevant documents based on three methods was recorded respectively. Finally, the values over ten queries were aggregated and the relevance ratios of the three methods were calculated. As a result, we set $\alpha = 0.4$, $\beta = 0.4$ and $\gamma = 0.2$.

3.2.2. Keyword similarity computation

To compute pairwise similarity of all keywords within the dictionary, a variety of metrics (such as cosine similarity, Jaccard coefficient and Pearson correlation) have been proposed in the literature, which are often simply calculated based on keyword co-occurrence. However, these metrics will suffer from the problem of power law distribution of keywords in scientific articles. Although the Latent Semantic Indexing (LSI) technique [32] has been used in Information Retrieval to deal with the above problem, it has raised several concerns due to its computational cost and long parameter tuning time. In this work, we employ the novel keyword similarity method proposed in [33] relying on the mutual reinforcement principle. The method uses an iterative approach to compute similarities whereby the similarity between any two objects (tags or resources) is computed based on the similarities already computed in the previous iteration. In detail, the similarity computation is performed as follows.

Initial Step

$$sk^0(k_m, k_n) = \theta_{mn}, sa^0(a_m, a_n) = \theta_{mn}.$$  

(2)

In the $p$th step

$$sk^p(k_m, k_n) = \frac{SK^p(k_m, k_n)}{\sqrt{SK^p(k_m, k_m) \ast SK^p(k_n, k_n)}},$$

$$sa^p(a_m, a_n) = \frac{SA^p(a_m, a_n)}{\sqrt{SA^p(a_m, a_m) \ast SA^p(a_n, a_n)}},$$

(3)

(4)

where

$$SK^p(k_m, k_n) = \sum_{i,j=1}^{n_a} F_{sim} \cdot \phi_{ij} \cdot SA^{p-1}(a_i, a_j) \cdot FS_{jn},$$

$$SA^p(a_m, a_n) = \sum_{i,j=1}^{n_a} F_{sim} \cdot \phi_{ij} \cdot SK^{p-1}(k_i, k_j) \cdot FS_{in}.$$  

(5)

(6)

In the initial step, keyword similarity $sk^0(k_m, k_n)$ and the article similarity $sa^0(a_m, a_n)$ are defined. Each keyword (respectively, article) is similar only to itself and it is dissimilar to all other keywords (respectively, article). At the $p$th step, let $sk^p(k_m, k_n)$ (respectively, $sa^p(a_m, a_n)$) be the keyword (respectively, article) similarity between $k_m$ and $k_n$ (respectively, $a_m$ and $a_n$). In Equations (5)–(6), $\phi_{ij}$ is equal to 1 if $i = j$; otherwise it is equal to $\varphi$ and here $\varphi$ is the mutual reinforcement factor and $\varphi \in [0, 1]$. The mutual reinforcement factor $\varphi$ is guided to give higher relevance to keywords that represented the very same articles (respectively, to articles represented by the very same keywords). As operated in [33], the parameter $\varphi$ can be learned from experiments. In this study, the best performance was achieved when $\varphi$ was set equal to 0.4.
In this way, the keyword correlation matrix can be constructed and it is used to compute the matching degree between two profiles as presented in the next section.

3.2.3. Semantic profile matching
In this research, we apply a filtering strategy to efficiently recommend a selected list of articles from the pool of a large set of articles. Since exact keyword matching could generate inaccurate results, we employ semantic keyword matching to filter out irrelevant articles. As the initial step, an expanded researcher profile is generated by adding more keywords that are similar to those in the original researcher profile. To make it less complicated, we add five more keywords to each keyword in the profile. Similar keywords are identified based on the pre-computed keyword correlation matrix. Then, the enriched profile is used to match with potential article profiles. The matching degree of keywords between the extended researcher profile and article profile is calculated as follows:

$$\text{MD}_k(r, a) = \sum_{i=1}^{n_k} \text{FS}_{\text{kr}}(i) \cdot \text{FS}_{\text{ka}}(i) \cdot \text{sim}(i),$$

where $\text{MD}_k(r, a)$ denotes the keyword matching degree of the researcher and article profile; $n_k$ is the number of distinct keywords in the extended researcher profile; $\text{FS}_{\text{kr}}(i)$ represents the frequency score of keyword $i$ in the extended researcher profile; $\text{FS}_{\text{ka}}(i)$ represents the frequency score of keyword $i$ in the article profile; $\text{sim}(i)$ indicates whether keyword $i$ is an author-used keyword or expanded keyword, where $\text{sim}(i) = 1$, if it is the used keyword and $\text{sim}(i)$ in the keyword correlation matrix otherwise.

3.3. Article social voting
As mentioned above, online connections of researchers can be extracted and mined to support article recommendation. CF techniques often employ like-minded neighbors to form recommendations. However, the emergence of social computing applications generates additional implicit or explicit social relationships for online users. For example, people make friends, add contacts and tend to trust others on the web. Furthermore, we can generate implicit connections by analyzing common bookmarking behavior or profile similarity. In this section, we present how diverse online connections are extracted and aggregated to vote for article recommendations.

3.3.1. Extraction of online connections
There are three types of online connections between researchers in social computing contexts. The first one is the social connection which reflects the direct social interactions between online users. The second one is the behavioral connection which shares the similar meaning of neighbors in CF settings. The third one is the semantic connection which links people through the similarity of semantic profiles. Based on the graph representation of Fig. 2, we can construct three different relationship matrices: User–User matrix, User–Item matrix and User–Keyword matrix to derive three types of connections. For the User–User matrix, various similarity measures [34] (i.e. Adamic and Adar index, FriendTNS, Jaccard Coefficient, Common Neighbors index, Random Walk with Restart (RWR), etc.) can be employed to analyze the node proximity in the network. In this paper, we choose the FriendTNS metric to calculate social connectivity in terms of its good performance in other related applications [35]. The FriendTNS similarity measure is defined as follows:

$$\text{sim}_{\text{soc}}(u_i, u_j) = \begin{cases} 0 & \text{if } u_{ij} = 0 \text{ and } u_{ji} = 0, \\ \frac{1}{\text{deg}(u_i) + \text{deg}(u_j) - 1} & \text{otherwise}, \end{cases}$$

where $u_{ij} \in \{0, 1\}$ is the element of the User–User matrix and if $u_{ij} = 1$, a social link exists between $u_i$ and $u_j$. We denote by $\text{deg}(u_i)$ and $\text{deg}(u_j)$ the degrees of nodes $u_i$ and $u_j$, respectively. For non-adjacent nodes $u_i$ and $u_j$, we multiply the similarity values between the intermediate nodes of the shortest path between $u_i$ and $u_j$. For the User–Item matrix and the User–Keyword matrix, we use cosine similarity to extract implicit behavioral connections and semantic connections. The user similarities of behavior connections and semantic connections are defined as follows:

$$\text{sim}_{\text{beh}}(u_i, u_j) = \frac{\sum_{i \in I} (W_{u_i, i} \ast W_{u_j, i})}{\sqrt{\sum_{i \in I} (W_{u_i, i})^2} \sqrt{\sum_{i \in I} (W_{u_j, i})^2}},$$

$$\text{sim}_{\text{sem}}(u_i, u_j) = \frac{\sum_{i \in K} (\text{FS}_{u_i, k} \ast \text{FS}_{u_j, k})}{\sqrt{\sum_{i \in K} (\text{FS}_{u_i, k})^2} \sqrt{\sum_{i \in K} (\text{FS}_{u_j, k})^2}},$$

where $W_{u_i, i}$ denotes the social behavior of users to items (binary value for bookmarking or integer value for rating) and $\text{FS}_{u_i, k}$ is the frequency score of keywords in the user profile calculated in Section 3.2.1.

3.3.2. Social-union for neighbor selection
After extracting three types of online connections, the overall similarity between two users can be calculated by aggregating the three similarity scores. In this paper, we apply the Social-Union method [35] to combine three similarity scores from heterogeneous online connections. The aggregated user similarity scores are further employed to select nearest neighbors for recommendation. The social-Union method has three main steps: Normalization, Weighting and Aggregation. The formulas used in each step are presented as follows:

**Normalization Step:**

$$\text{sim}_X(u_i, u_j) = \frac{\text{sim}_X(u_i, u_j) - \mu_X}{\sigma_X},$$
where $X$ denotes types of online connections (social, behavioral and semantic); $\mu_X$ denotes the mean similarity value of $X$ similarity matrix and $\sigma_X$ denotes deviation of $X$ similarity matrix.

Weighting Step:

$$\text{sim}_X(u_i, u_j) = \frac{\text{sim}_X(u_i, u_j) - \mu_X}{\sigma_X},$$  \hspace{1cm} (12)

$$dx = \frac{\text{local}_i}{\text{global}_i}, \quad W_x = \frac{dx}{\sum_{x \in X} dx},$$  \hspace{1cm} (13)

where $\text{local}_i$ is the local density of the selected user $u$ into the adjacency matrix, i.e. the number of non-zero values in its row divided by the number of users $(\text{deg}(u_i)/n)$, and $\text{global}_i$ is the global density of the adjacency matrix, i.e. the number of non-zero values in the full matrix divided by the square of the number of users $(n^2)$.

Aggregation Step:

$$\text{sim}(u_i, u_j) = \sum_{x \in X} W_x \text{sim}_x(u_i, u_j).$$  \hspace{1cm} (14)

Then, aggregated user similarity $\text{sim}(u_i, u_j)$ is used to select nearest neighbors to support article voting.

3.3.3. Social voting for article recommendation

The nearest neighbors with their corresponding similarity scores to the focal user are retrieved by the Social-Union method. Articles related to closest neighbors are selected and we assign a voting score for those selected articles based on nearest neighbors’ interest. The voting score of article $i$ to user $u$ is represented as $\text{VS}(u, i)$ and it is determined by the following formula:

$$\text{VS}(u, i) = \sum_{v \in \text{Nr}(u)} \text{sim}(u, v) \ast r(v, i),$$  \hspace{1cm} (15)

where $v \in \text{Nr}(u)$ is a researcher in the user $u$’s nearest neighbors set $\text{Nr}(u)$, $\text{sim}(u, v)$ denotes the aggregated similarity score between user $u$ and user $v$, and $r(v, i)$ denotes the level of interest of user $u$ to article $i$.

3.4. Two-stage aggregated recommendation

Article recommendation aims at recommending articles that are mostly relevant and widely accepted by similar researchers. The semantic similarity measures calculated above are used to determine mostly related articles and filter out irrelevant ones. The social voting measure is used to identity widely used articles by related experts. The amalgamation of these two types of results is necessary to recommend most suitable articles. Therefore, we follow very popular data fusion methods to aggregate two types of results and to compute the final ranking score for the candidate articles. Data fusion has also been widely investigated in the information retrieval community. The fusion techniques can be useful for various tasks such as routing queries [36], expert search [37] and blog opinion retrieval [38]. They were often divided into two categories: score-based and ranking-based. Score-based fusion methods require similarity information to conduct ranking list aggregation (such as CombSum, CombMNZ [39] and linear combination [38]). Ranking-based fusion methods require rank or position information to integrate different candidate ranking lists (such as Borda fusion [40], Condorcet fusion [41] and MAPFuse [42]).

In this research, we model article recommendation as a data fusion task. The CombMNZ aggregation method is applied to integrate existing ranking lists generated by applying semantic similarity measures and social voting measures consecutively. The CombMNZ method regards not only the scores in each of the ranked list but also the number of supported evidence (non-zero scores in each list). Therefore an article positioned higher in the ranking list of generated recommendations is mostly relevant (in content semantics) to the researcher as well as socially recommended by the researcher’s online connections. The final recommendation score of the article $i$ to user $u$: $\text{RS}(u, i)$ can be fused as in equation (15).

$$\text{RS}(u, i) = \tau(\text{MD}(u, i), \text{VS}(u, i)) \ast \text{SUM}[\text{MD}(u, i), \text{VS}(u, i)],$$  \hspace{1cm} (16)

$$\text{sim}_{\text{norm}} = \frac{\text{sim}_{\text{org}} - \text{sim}_{\text{min}}}{\text{sim}_{\text{max}} - \text{sim}_{\text{min}}},$$  \hspace{1cm} (17)

where $\tau(\text{MD}(u, i), \text{VS}(u, i))$ is the count measure and if $\text{MD}(u, i)$ and $\text{VS}(u, i)$ are both more than zero, $\tau$ equals 2; if only one of $\text{MD}(u, i)$ and $\text{VS}(u, i)$ is more than zero, $\tau$ equals 1; if both of $\text{MD}(u, i)$ and $\text{VS}(u, i)$ are zero, $\tau$ equals 0. $\text{SUM}[a, b]$ is the function to calculate the sum score of $a$ and $b$. Before being used to calculate the recommendation score, $\text{MD}(u, i)$ and $\text{VS}(u, i)$ should be processed through the normalization operation presented in Equation (16).

4. SYSTEM IMPLEMENTATION

ScholarMate is an online professional social network community platform, particularly developed for academic researchers by the author’s team. It aims to foster a knowledge-sharing cyberspace for researchers to collect and share different kinds of resources (e.g. publications, research progress reports). Different from other online researcher communities which require researchers to input research outputs manually, it can automatically collect a particular researcher’s outputs from various sources, like CNKI,6 ISI and Scopus.7 On the ScholarMate, researchers can add other researchers into their contact list as friends. Besides, researchers with similar interests can collaborate via self-organized special interest group (SIG) functions.

6http://www.cnki.net/
They can share their professional works in terms of publications, projects, papers with other community members, and receive comments and suggestions. Through ScholarMate, researchers can benefit more from sharing and recommending scientific outputs. Information publisher (author of the article) can gain high reputation and become popular by spreading their outcomes. Information receiver (any researcher) can receive relevant and useful information to support their research. Therefore, ScholarMate provides a potential platform to test and evaluate the proposed approach.

The proposed approach is implemented as one of the application services in ScholarMate. The system provides the main interfaces to extract article-related data, including title, keywords and abstract. Once the system has gathered the required information, matching degrees between article profiles and researcher profiles are calculated. Also, the system extracts different relationships of online researchers to support social network-empowered recommendation. Figure 4 presents the interface of the article recommendation application with descriptive features (content features and social features). As shown in the right panel of Fig. 4, additional details about article quality are also provided by the system. The decision button (accept or reject) is displayed in the last column.

5. EXPERIMENTAL EVALUATION

In this section, we empirically evaluate the proposed approach and compare its performance against that of the benchmark algorithms. To demonstrate the prospects of our proposed method, we conduct two experimental studies, a traditional offline experiment with the CiteULike dataset and a real online user study in ScholarMate.

5.1. Data collection

For offline experiment evaluation, we obtained the users and their libraries of articles from the CiteULike website. CiteULike is one of the leading social tagging systems for managing and sharing bibliographic references. Besides bookmarking relevant articles, users of CiteULike can create research groups and also join others’ groups with specific research topics. In order to obtain required data, we visited the CiteULike website on 1 December 2011 using a crawler and extracted article content and social data for what we needed. Since we employed social connections to improve recommendation quality, we collected all group names that were shown on the page at the time of the visit. Then, the group collections, the group members and each member’s personal collection were collected. Each article of the collections contained a title and an abstract. (The other information about the articles, such as the authors, journals and editors, is not used in this paper.) Users’ social tagging information was also included in our dataset. We conducted the data cleaning work and removed single-member groups and users with insufficient collections ($n < 15$). Finally, we obtained a dataset of 1660 users and 70,032 articles with 206,289 observed user–item pairs. Table 1 is the description of the used data statistics.
TABLE 1. Statistics of filtered CiteULike dataset.

<table>
<thead>
<tr>
<th>Users</th>
<th>Articles</th>
<th>User–Item pairs</th>
<th>Tags</th>
<th>Tag assignments</th>
<th>Groups</th>
<th>Group memberships</th>
</tr>
</thead>
<tbody>
<tr>
<td>1660</td>
<td>70 032</td>
<td>206 289</td>
<td>60 642</td>
<td>476 462</td>
<td>469</td>
<td>1965</td>
</tr>
</tbody>
</table>

For the article content, we concatenated its title and abstract and also incorporated the attached social tags into the content. We removed stop words and used term frequency-inverse document frequency (TF-IDF) to choose the top 80% distinct words as the vocabulary.

For the online user study, registered users in ScholarMate were selected as subjects of this survey. Prior to their participation, recommendation lists (top 10 articles) from different methods were computed. Then, each subject assessed the randomized combined recommendation list (no more than 30 articles). The subjects were asked to accept or reject each recommended article. If the participant considered the recommended article relevant and useful, he/she could accept it, otherwise reject it. As a result, we could use the obtained feedback data to evaluate the effectiveness and accuracy of our proposed approach.

5.2. Experimental design and metrics

To evaluate the performance of the proposed method over existing article recommendation methods, we randomly divided the CiteULike dataset into training set and testing set. The users’ libraries of articles were split into a testing set with 10 items per user (16 600 articles in the testing set) and a training set with the remaining articles (53 432 articles in the training set). The training set together with content and social information was used to build the user profile and learn different recommendation methods. In the online user study, the experiment subjects were registered researchers from areas of Management Information Systems, Operations Research, Computer Science and so on. Hundred researchers were randomly selected subject to the condition that each of them should have at least five publications in ScholarMate. We made an online survey of their perceptions on the quality of recommendation results obtained by the proposed method and the other existing alternatives. After the experiments, we collected 74 valid responses. The response rate was 74% and it was acceptable. The responding researchers were from 12 distinct disciplines (we refer the reader to the National Natural Science Foundation of China discipline tree8), and among them 26% were with the title of professor, 59% with the title of associate professor and 15% with the title of lecturer.

For the performance comparison of our method and existing article recommendation approaches in the literature. They are listed as follows:

1. TF-IDF Vector-based Method (abbreviated TFIDF): This method uses TF-IDF value of keywords to represent user profile and article profile. Then article candidates are ranked by the calculated cosine similarity of user and article profile vectors.

2. Semantic-expansion Content Filtering Method (abbreviated SeCon): SeCon is the enhanced version of the TFIDF method by using the pre-computed keyword similarity matrix.

3. User-based Collaborative Filtering Method (abbreviated UCF): This method employs the user’s behavioral neighbors to predict the recommendation score of the article candidate.

4. Social Voting Recommendation Method (abbreviated SVR): This method combines behavioral connections and social connections to find nearest neighbors. Then, the recommendation score of the article candidate is determined by social votes from calculated nearest neighbors.

5. Semantic-Social Aggregation Recommendation Method (abbreviated SSAR): This is our proposed method. It leverages semantic content and online connections to generate recommendations for the user.

Among these five methods, the first two represented CB methods and the next two represented CF-based methods. The last one could be considered as a hybrid recommendation method. All of the five methods were tested in the offline experiment, while three selected methods (SeCon, SVR and SSAR) were tested in the online user study. We chose to test these three methods by using the survey due to their representativeness and recommendation tolerance of participants.

We treat the personalized article recommendation as a content retrieval system that recommends to researchers a list of ten articles relevant to their profiles. The evaluation metrics, Precision@K (Prec@K), Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) [43] are employed to evaluate the recommendation accuracy of different methods. The Prec@K measure only evaluates the ability to return overall relevant articles. However, MAP and MRR measures consider the rank information of relevant articles in the recommendation list. They

The Prec@5 scores obtained by using the SeCon method and are defined as follows:

\[ P@K = \frac{N_{\text{relevant}}}{K}, \]  

\[ \text{MAP} = \frac{1}{|U|} \sum_{j=1}^{|U|} \frac{1}{m_j} \sum_{k=1}^{N} P(R_{jk}), \]  

\[ \text{MRR} = \frac{1}{|U|} \sum_{j=1}^{|U|} \frac{1}{\text{rank } F_j}. \]

where \( K \) is the number of recommended articles and, in this setting, \( K \) is set to 5 and 10; \( N_{\text{relevant}} \) is the number of relevant articles in the ranking list; \(|U|\) denotes the number of subjects; \( m_j \) is the number of relevant articles to researcher \( j \); \( P(R_{jk}) \) represent the precision of retrieved results from the top result until you get to article \( k \); and \( \text{rank } F_j \) is the rank (position) of the first relevant article to the user \( j \).

### 5.3. Results and discussion

In this section, we present the detailed comparison of results from the offline experiment and the online user study. The values for evaluation metrics are obtained and compared between the state-of-the-art methods and our proposed approach on the CiteULike dataset. The detailed results are shown in Table 2. It can be easily observed from this table that our SSAR approach significantly improves the recommendation quality and outperforms the other competitive algorithms in terms of various performance metrics. The results also show that the proposed method has a major advantage on article recommendation since it has gained a performance improvement of more than 30% over the best competitive algorithm. In general, the UCF method gets the worst performance due to the high sparsity of the user–item matrix in the CiteULike. CB methods obtain the comparative performance over the UCF due to the fact that content information plays an important role in article recommendation. The social voting approach incorporates new connection information into traditional CF methods and also has obtained good performance. Our hybrid approach aggregates the advantages from both sides since it leverages the semantic content of articles and social connections of online users.

To further verify the effectiveness of the proposed method, the user study results of three selected methods are shown in Fig. 5. The Prec@5 scores obtained by using the SeCon method and the SVR method are 0.22 and 0.12, respectively, which are at least 0.18 lower than that of our proposed method. The Prec@10 scores of SeCon, SVR and SSAR in the Fig. 5 are 0.19, 0.11 and 0.25, respectively. Our proposed method generates high-quality results than other baseline methods. These results show that our method can recommend more relevant articles than the other two methods. We further evaluate the rank performance of the three methods. MAP scores reflect the browsing efforts of the researchers before locating the relevant articles. The MAP scores of the three recommendation methods are also shown in Fig. 5. The MAP score of the proposed method SSAR, 0.64, is over 20% higher than the scores of SeCon and SVR methods, which are 0.44 and 0.25, respectively. The improvements on the MAP value clearly show that our method is more effective than the SeCon and SVR methods, which gives a higher ranking for the relevant articles in the recommendation list. The large improvements in MRR scores also demonstrate our method’s ability to rank relevant articles higher.

Moreover, we test the significance of the results of the proposed method over baseline methods by means of the paired t-tests. The significance result in Table 3 indicates that improvements of our approach in Prec@5, Prec@10 and MAP are all statistically significant.

### 6. Conclusion and Future Work

In this paper, a novel approach leveraging content and connections is proposed for the personalized article
recommendation in social computing contexts. Profiles of researchers are built from two aspects: semantic content and heterogeneous connections. To overcome shortcomings of traditional CB and CF-based methods, we rank the article candidates according to the aggregated recommendation score from the pre-computed matching degree score and social voting score. The effectiveness of the proposed approach over the baseline methods is verified in two experimental contexts. The proposed algorithm and designed recommender system have been incorporated into the existing research social network website to facilitate content sharing and potential collaboration. Also, the algorithm and system can be generalized to other personalization applications in digital libraries and social media websites.

There are several limitations in this research. First, we compute keyword similarity to expand the user profile. We are aware that the use of domain ontology will greatly help to resolve semantic ambiguity in keyword matching. Thus, in the future, research domain ontology can be constructed to support extended profile matching. Secondly, this paper adopts the CombMNZ technique as the rank aggregation method. Some complex data fusion techniques can also be considered such as Condorcet fusion [41] and other techniques that model score distribution [44]. Moreover, since ScholarMate provides a good platform to collect recommendation feedbacks, relevance feedback weighting methods can be applied in the future. Thirdly, it is better to consider the time factor and the quality of the scientific articles (citations, journal impact factor) when recommending articles to researchers. Therefore, we will extend this work by considering factors of time and quality in the future.

FUNDING

This research was partially supported by the General Research Fund of the Hong Kong Research Grant Council (CityU 119611), the National Natural Science Foundation of China (71171172, 71001103 and 71101042), the Specialized Research Fund for the Doctoral Program of Higher Education (2011011120014) and City University of Hong Kong (6000201).

REFERENCES

[22] Vellino, A. (2010) A Comparison between Usage-Based and Citation-Based Methods for Recommending Scholarly Research


