Mining Social Media Websites to Recommend Groups of Interest

Jianshan Sun
School of Management
Hefei University of Technology, Anhui, China
Email: sunjs9413@hfut.edu.cn

Dong Xu
School of Management
Hefei University of Technology, Anhui, China
Email:xudong@mail.hfut.edu.cn

Yuanchun Jiang
School of Management
Hefei University of Technology, Anhui, China
Email: ycjiang@hfut.edu.cn

Yezheng Liu
School of Management
Hefei University of Technology, Anhui, China
Email: liuyezheng@hfut.edu.cn

Abstract. Social media websites, such as YouTube and Flicker, are gaining increasing popularity nowadays. Online group functions are supported by these websites to enable users to collectively share their rich experience and information. However, the explosive growth of groups makes it increasingly difficult for users to find relevant ones that they are really interested in. This research proposes a novel approach to recommend interest groups to online users by leveraging semantic content and social connections involved in social media data. Semantic group recommendations and social group recommendations are aggregated via data fusion techniques. Two real social media websites are considered and experiments are conducted. The evaluation results exhibit that the proposed method is more effective than the baseline methods.

Keywords: Social media; group recommendation; semantic content; social connections

1. INTRODUCTION

Social media websites (e.g., YouTube and Flicker) are increasingly attracting people's attention nowadays. Recent years have witnessed a rapid convergence of online content sharing network websites. We observe that large amount of content can be generated and diffused by users in these social media websites. Due to the dynamical behavior of users in social media websites and the great volume of content generated by users, it imposes great challenges for traditional recommendations to provide personalized content to users. In the social media websites, the delivery of online content and information among users determines the popularity of the site. There are several ways by which such content and information can be shared among individuals. One of the most popular information sharing methods involves the formation of online groups that enable users to collectively share their content and rich experience with a group of people. More and more modern social Web sites such as Facebook, Flickr, CiteULike and Last.fm have supported group functions to involve users in sharing items and exchange insights. However, the explosive growth of online groups creates new challenges for researchers to help users locate relevant interest groups to join. It is a critical issue for online users to find relevant groups that they are really interested in. Users are flooded

in the sea of too much information and are struggling to make good decision, which refers to information overload problem (Aljukhadar et al., 2012). Manually browsing or searching the huge number of groups is very time-consuming and difficult. Thus, it is increasingly important to leverage social media data to recommend appropriate interest groups.

To alleviate the information overload imposed on online users and to facilitate group participation, we focus on recommending interest groups by mining social media websites. Previous group recommendation works are relatively limited when they are compared to item recommendation task. Existing works (Chen et al., 2008; Kim and El Saddik, 2013; Vasuki et al., 2010; Zheng et al., 2010) had limited exploration on social media data and they were lack of semantic analysis and social network analysis. In (Vasuki et al., 2010), only friendship information and membership information were considered to generate group recommendations while other useful metadata was ignored. Although content information and connection information were both used in Chen et al's work (2008), deeper content semantics analysis and various online connections were not exploited. Some recent work empirically verified the contributions of tagging information to improve recommendation performance. However, there is still room for improvement by leveraging rich social media data.

To address above issues, a hybrid recommendation approach, which is called the semantic-social fused group recommendation approach, has been proposed to recommend relevant interest groups in social media websites. The semantic content analysis and social network analysis are integrated via data fusion model to recommend highly semantic relevant and socially endorsed groups. The proposed approach is evaluated through a comprehensive experiment using CiteULike dataset and Last.fm dataset. The results show that the proposed approach outperforms the baseline methods in terms of recommendation accuracy.

The main contribution of this paper can be summarized as follows:

(1) We propose a semantic social group recommendation framework to recommend interest groups for online users, which leverages semantic content and online connections to improve recommendation performance.

(2) Mapreduce framework has been employed to support large scale similarity computation in social media contexts and data fusion techniques has been investigated in recommendation context and their effectiveness has been evaluated.

(3) To evaluate the performance of the proposed group recommendation framework, we conduct comprehensive experiments in two real datasets from social media websites and the results demonstrate the effectiveness of the proposed method and framework.

The rest of the paper is organized as follows. The details of semantic group recommendation framework are introduced in Section 2. Section 3 presents the design and methodology used in the experiments, while the results are analyzed in Section 4. Section 5 discusses conclusions and points out future research directions.

2. SEMANTIC-SOCIAL GROUP RECOMMENDATION MODEL

In this paper, we focus on recommending interest groups to individual users in order to address the following fundamental question: Given a particular user, which groups would be relevant to his/her interests? To answer this question, we mine the social media websites deeply and leverage content and connections to find interest groups for online users. Figure 1 depicts the architecture of the proposed group recommendation mechanism. There are three main modules developed to analyze the information from the social media websites. The objectives of the analysis modules included in the system are described as follows:

(1) The semantic content filtering module establishes the term vectors for group profiles and user profiles, computes the matching degree scores between them and recommends interest groups with semantic relevance. Moreover, the term similarity matrix is innovatively computed to support profile representations.

(2) The social aggregation filtering module analyzes heterogeneous user relations by examining online social activities, computes social aggregation scores from nearest neighbors and recommends interest groups with social influence.

(3)The group recommendation fusion module leverages a variety of data fusion strategies to combine two recommendation lists from previous steps and provides the final group recommendations to online users.

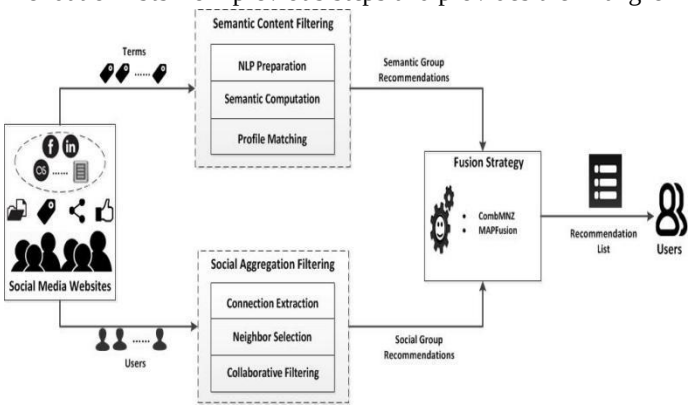

Figure 1: The overview of proposed semantic social group recommender system.

In social media websites, we leverage semantic expansion techniques and social mining methods to recommend most relevant interest groups for online users. The whole processes of the recommendation mechanism are detailed in the following subsections.

2.1 Semantic content filtering

2.1.1 NLP preparation

Since users' bookmarked items show their latent interests, we represent user profiles by analyzing the content and descriptions of items. Similarly, group profiles can also be represented by their member users' relevant items. Therefore, term vectors of all the items should be prepared. Using the classical NLP procedures (such as segmentation, stopping and stemming) (Pudota et al., 2010), the items and term features are represented by a Term-Item Matrix. The Term-Item (TI) matrix is a matrix to denote the association between terms and items, where is the number of terms and is the number of items. Traditionally, only terms in the title or description of items were considered However, this limited set of terms cannot capture the comprehensive content of the item since a low number of terms causes the TI matrix to be sparse leading to less accurate term correlation scores. In order to overcome the accuracy problem, rich tagging information is extracted and exploited. In the social media context, tagging information provides additional collective content descriptions. Some recent works have proved that tags were beneficial for document retrieval (Figueiredo et al., 2012; Hsu and Chen, 2011). In this paper, we use term frequency-inverse document frequency (TFIDF) measure (Chowdhury, 2010) to compute TI matrix. The TFIDF value of term t of item i, is defined as follows:

$$TFIDF_{t,i} = \frac{N_{t,i}}{\sum_{k \in V} N_{k,i}} * log \frac{|I|}{1 + |I_t|} \qquad (1)$$

Where $N_{t,i}$ is the occurrence count of term t in item i; V denotes the te-rm vocabulary; $|I|$ is the number of all the items and $|I_t|$ is the number of items which contains term t.

2.1.2 Similarity computation

Since there may be certain semantic relationships in the item content, traditional vector space model (VSM) techniques generate term mismatch problem on account of ignoring term semantics (Quattrone et al., 2011; Sun et al., 2013). In this paper, we use semantic computing method to find semantics between terms. To compute pairwise similarity of all terms 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 term co-occurrence. However, these metrics will suffer from the problem of keyword's power law distribution in social items. Although the Latent Semantic Indexing (LSI) technique (Manning et al., 2008) 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 (Quattrone et al., 2011) relying on the mutual reinforcement principle. The method uses an iterative approach to compute similarities whereby the similarity between any two objects (terms or items) is computed based on the similarities already computed in the previous iteration. In detail, the similarity computation is performed as follows.

Initial Step,

$$st^0(t_m, t_n) = \theta_{mn}, \qquad si^0(i_m, i_n) = \theta_{mn} \qquad (2)$$

In p^{th} Step

$$st^p(t_m, t_n) = \frac{ST^p(t_m, t_n)}{\sqrt{ST^p(t_m, t_m)} \cdot \sqrt{ST^p(t_n, t_n)}} \qquad (3)$$

$$si^p(i_m, i_n) = \frac{SI^p(i_m, i_n)}{\sqrt{SI^p(i_m, i_m)} \cdot \sqrt{SI^p(i_n, i_n)}} \qquad (4)$$

Where:

$$ST^p(t_m, t_n) = \sum_{j,k=1}^{n_I} w_{mj} \cdot \varphi_{jk} \cdot si^{p-1}(i_j, i_k) \cdot w_{nk} \qquad (5)$$

$$SI^p(i_m, i_n) = \sum_{j,k=1}^{n_T} w_{jm} \cdot \varphi_{jk} \cdot st^{p-1}(t_j, t_k) \cdot w_{kn} \qquad (6)$$

In the initial step, term similarity $st^0(t_m, t_n)$ and the item similarity $si^0(i_m, i_n) = \theta_{mn}$ are defined. Each term (resp., item) is similar only to itself and it is dissimilar to all other terms (resp., item). At the p^{th} step, let $st^p(t_m, t_n)$ (resp., $si^p(i_m, i_n)$) be the term (resp., item) similarity between t_m and t_n (resp., i_m and i_n). In Equations (5-6), w_{mj} and w_{nk} are the entries in the Term-Item matrix while w_{jm} and w_{kn} are the entries in the Item-Term matrix; φ_{ij} is equal to 1 if i = j, otherwise it is equal to φ where φ is mutual reinforcement factor and $\varphi \in [0,1]$. The mutual reinforcement factoris guided to give higher relevance to terms that represented the very same items, (resp., to items represented by the very same terms). As operated in (Quattrone et al., 2011), the parameter can be learned from experiments. In this study, the best performance was achieved was set equal to 0.4. In this way, the term correlation matrix can be constructed and it is used to compute matching degree between two profiles as presented in the next section.

2.1.3 Profile matching

As the initial step, an expanded researcher profile is generated by adding more keywords which are similar to those in the original researcher profile. To make it less complicated we add three more terms to each term in the profile. Similar terms are identified based on the pre-computed term correlation matrix. Then, the enriched user profile is used to match with potential group profiles. The matching degree of terms between the extended user profile and group profile is calculated as follows:

$$MD\ (u,g) = \sum_{i=1}^{n_{ET}} w_{ui}\, w_{gi}\, sim_i \qquad (7)$$

where MD (u, g) denotes term matching degree of the user and group profile; n_{ET} is the number of distinct terms in extended user profile; w_{ui} represents the weight of term i in the extended researcher profile; w_{gi} represents the weight of term i in the group profile; sim_i indicates whether term i is an user profile term or expanded term, where sim_i=1, if it is the used keyword and sim_i in the term correlation matrix otherwise.

2.2 Social aggregation filtering

2.2.1 Connection extraction

There are three types of online connections between users in social media websites. The first one is explicit social linkages, such as friendships. We call these connections social connections, since they reflect the direct social interactions between online users. The second one is implicit relations derived from analyzing common social behaviors of users, such as bookmarking the same item and joining the same group. We call these relations behavioral connections, which share the similar meaning of neighbors in collaborative filtering settings. The third one is implicit relations calculated by the similarity of user profile. We call these relations semantic connections since they link people through the similarity of semantic profiles. Based on the graph representation of Figure 1, we can construct three different relationship matrices: User-User matrix, User-Object matrix and User-Term matrix to derive three types of connections respectively. For User-User matrix, various similarity measures (Liben-Nowell and Kleinberg, 2007) (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 (Symeonidis et al., 2011). The FriendTNS similarity measure is defined as follows:

$$sim_{soc}(u_i,u_j) = \begin{cases} 0 & if\ u_{ij}=0\ and\ u_{ji}=0 \\ \frac{1}{deg(u_i)+deg(u_j)-1}, & otherwise \end{cases} \qquad (8)$$

Where $u_{ij} \in \{0,1\}$ is the element of User-User matrix and if $u_{ij}=1$, a social link exists between u_i and u_j. $\deg(u_i)$ and $\deg(u_j)$ denote 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 User-Object matrix and 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:

$$sim_{beh}(u_i,u_j) = \frac{\sum_{vo\in O}(w_{u_i,o} * w_{u_j,o})}{\sqrt{\sum_{vo\in O}(w_{u_i,o})^2} * \sqrt{\sum_{vo\in O}(w_{u_j,o})^2}} \qquad (9)$$

$$sim_{sem}(u_i,u_j) = \frac{\sum_{vt\in T}(w_{u_i,t} * w_{u_j,t})}{\sqrt{\sum_{vt\in T}(w_{u_i,t})^2} * \sqrt{\sum_{vt\in T}(w_{u_j,t})^2}} \qquad (10)$$

Where $w_{u_i,i}$ denotes the social behavior of users to objects (items or groups) and it is often a binary value (0 and 1) for bookmarking (joining) or not. $w_{u_i,t}$ is the weight of terms in the user profile calculated in Section 2.1.1.

2.2.2 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 (Symeonidis et al., 2011) to combine three similarity scores from heterogeneous online connections. The aggregated user similarity scores are further employed to select nearest neighbors for recommendation. Social-Union method has three main steps: Normalization, Weighting and Aggregation. The formulas used in each steps are presented as follows:

Normalization Step,

$$sim_X(u_i,u_j) = \frac{sim_X(u_i,u_j)-\mu_X}{\sigma_X} \qquad (11)$$

Where X denotes types of online connections (social, behavioral, and semantic); μ_X denotes mean similarity value of X similarity matrix and σ_X denotes deviation of X similarity matrix.

Weighting Step,

$$dx = \frac{local\_x}{global\_x}\ , \qquad W_x = \frac{dx}{\sum_{x\in X} dx} \qquad (12)$$

Where local_x 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 ($\deg(u_i)/n$). global_x 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 number of users ($/n^2$).

Aggregation Step,

$$sim(u_i, u_j) = \sum_{x \in X} W_x \, sim_x(u_i, u_j) \quad (13)$$

Then, aggregated user similarity $\mathrm{sim}(u_i, u_j)$ is used to select nearest neighbors to support collaborative filtering process.

2.2.3 Collaborative filtering

The nearest neighbors with their corresponding similarity scores to the focal user are retrieved by the Social-Union method. Groups related to closest neighbors are selected and we assign voting score for those selected groups based on nearest neighbors' interest. The voting score of group g to user u is represented as $\mathrm{VS}(u, g)$ and it is determined by the following formula:

$$\mathrm{VS}(u, g) = \sum_{v \in \mathrm{Nr}(u)} \mathrm{sim}(u, v) * m(v, g) \quad (14)$$

Where $v \in \mathrm{Nr}(u)$ is a user in the user u's nearest neighbors set $\mathrm{Nr}(u)$. $\mathrm{sim}(u, v)$ denotes the aggregated similarity score between user u and user v. $m(v,g)$ denotes whether or not user v has a membership relation with group g and its value is set 1 or 0.

2.3 Group recommendation fusion

Group recommendation aims at recommending interest groups that are mostly semantic relevant and widely joined by similar users. The semantic matching degree calculated above is used to determine content-related groups. The social aggregation score is used to identity widely joined groups by connected users. The amalgamation of these two types of results is necessary to recommend most suitable ones. Therefore, we follow very popular data fusion methods to aggregate two types of results and to compute the final ranking score for the candidate groups. Data fusion has also been widely investigated in the information retrieval community. 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 (Fox and Shaw, 1994), and linear combination (Wu, 2012)). Ranking-based fusion methods require rank or position information to integrate different candidate ranking lists (such as Borda fusion (Aslam and Montague, 2001), Condorcet fusion (Montague and Aslam, 2002) and MAPFuse (Lillis et al., 2010)).

In this research we model group recommendation as a data fusion task. The CombSum, CombMNZ and MAPFuse aggregation method is applied to integrate existing ranking lists generated by applying semantic content filtering module and social aggregation filtering module consecutively. They are defined as follows:

$$\mathrm{Score}_{CombMNZ}(u, g) = \tau * (MD(u, g)_{norm} + \mathrm{VS}(u, g)_{norm}) \quad (15)$$

$$\mathrm{Score}_{MAPFuse}(u, g) = \mathrm{Map}_{MD} * MD(u, g)_{norm} + \mathrm{Map}_{VS} \mathrm{VS}(u, g)_{norm} \quad (16)$$

$$\mathrm{sim}_{\mathrm{norm}} = \frac{\mathrm{sim}_{\mathrm{org}} - \mathrm{sim}_{\mathrm{min}}}{\mathrm{sim}_{\mathrm{max}} - \mathrm{sim}_{\mathrm{min}}} \quad (17)$$

Before being used to calculate recommendation score, $\mathrm{MD}(u, g)$ and $\mathrm{VS}(u, g)$ should be processed through the normalization operation presented in Equation (17). τ is the count measure and if $\mathrm{MD}(u, g)_{\mathrm{norm}}$ and $\mathrm{VS}(u, g)_{\mathrm{norm}}$ are both more than zero, τ equals 2; if only one of $\mathrm{MD}(u, g)_{\mathrm{norm}}$ and $\mathrm{VS}(u, g)_{\mathrm{norm}}$ is more than zero, τ equals 1; if both of $\mathrm{MD}(u, g)_{\mathrm{norm}}$ and $\mathrm{VS}(u, g)_{\mathrm{norm}}$ are zero, τ equals 0. $\mathrm{Map}_{\mathrm{MD}}$ and $\mathrm{Map}_{\mathrm{VS}}$ are the map values (evaluation measure) calculated based on semantic content filtering and social aggregation filtering. Then, the final recommendations can be provided based on calculated fusion scores.

2.4 Large-scale similarity computation

Since the number of users and the number of groups are often huge in social media websites, previous research was lack of run-time efficiency to find relevant interest groups for users. We should refer to more intelligent computation tools for Large-scale Similarity Computation. With the rapid development of information techniques, MapReduce (Elsayed et al., 2008) is a popular framework for data-intensive parallel computation in shared-nothing clusters of machines, which includes two functions: map and reduce. The map function applies a user-defined function to each key-value pair in the input and generates a list of intermediate key-value pairs. These generated pairs are then sorted and grouped by the key and are further passed as inputs to the reduce function. The reduce function applies a second user-defined function to every intermediate key and all its associated values, and produces the final result. It has been successfully applied in many applications such as crawled document index, web access log analysis, and machine learning. So, in this research, MapReduce is suggested to improve the efficiency of group recommendation process in social media websites. Now we are calculating the user to group semantic similarity as an example. In order to compute sim(x,y) for all pairs of user profile and group profile in a batch mode, we first build an inverted index for all terms in the vocabulary. For each term t, there is a corresponding posting in the inverted

index Ind :

$$< (u_1, w_{t,u_1}), (u_2, w_{t,r_2}), \dots, (u_i, w_{t,r_i}) \dots (g_1, w_{t,g_1}), (g_2, w_{t,g_2}), \dots, (g_j, w_{t,g_j}) \dots >$$

Where u_i is a user's profile and g_j is a group profile. w_{t,r_i} and w_{t,r_i} are the corresponding weights. Then, we generate a mapper for each pair of user profile and group profile in each inverted index posting. Finally, all of the intermediate results calculated by these mappers are aggregated by the reducers. We summarize these steps in the following algorithm.

Algorithm: Similarity computation via MapReduce
Input: Inverted index Ind
Process:
Initialize sim(x,y):
$sim(x,y) := 0, \forall\ x \in U, y \in G$
For all $t \in V$ Do
$p(t) := Ind(t)$
For all $x, y \in p(t)$ Do
Map: $map(key := x{:}y, v = x,y) \rightarrow < key := x{:}y, v' = w_{t,x} \cdot w_{t,y} >$
For all $x \in U, y \in G$ Do
Reduce: $sim(x,y) := \sum_{key=x:y} v'_{key}$
Output: sim(x,y) for all $x \in U, y \in G$

It's easy to see that the same algorithm can be employed for computing similarities for user paires.

3. EXPERIMENTAL DESIGN

3.1 Datasets

To evaluate the proposed semantic social group recommendation framework, we used two test datasets from social media websites. The first dataset was taken from CiteULike, which is a social tagging website where users can manage/share scholarly articles. In addition to tagging articles, users can create and join groups according to their research topics of interest. CiteULike offers daily dumps of their core database. We used the dump of May 27, 2013 as the basis for our experiments. A dump contains social tagging information and group membership information. It does not, however, contain other article metadata (such as title) information, so we crawl the article title ourselves from the CiteULike website using the article IDs. Since the dump only contain encrypted user IDs, we have no real user IDs and we cannot obtain user friendship relations. The second dataset was taken from Last.fm, which is a social music website where users can tag artists, tracks, and albums. It also allows users create and join groups based on common interests, music artists, and/or music genres. We used the published data from Schifanella et al.'s work (2010). This data set was crawled in the first half of 2009 and it contained social tagging information, item metadata information, user friendship information and group membership information which cover all the information our proposed method needed. In the original group data, many groups contained only one member and many users belonged to only one group. Therefore, we conducted the data cleaning work and removed single-member groups and ensured that items has been bookmarked by at least two users. Finally, we obtained 8741 users and 1764 groups with 12699 observed user-group pairs for CiteULike dataset and obtained 41615 users and 44191 groups with 725744 observed user-group pairs for Lastfm dataset. Table 1 is the description of the used data statistics.

Table 1: Statistics of filtered datasets used in experiments.

Dataset	Users	groups	Items	Memberships	Friendship	Bookmarks
CiteULike	8741	1764	112843	12699	/	308170
Lastfm	41615	44191	455075	725744	256446	3324955

3.2 Evaluation metrics

We treat group recommendation as a content retrieval system that recommends interest groups to online users. The evaluation metrics, Precision@K (P@K) and Mean Average Precision (MAP) (Croft et al., 2010) are employed to evaluate the recommendation accuracy of different methods. P@K measure only evaluates the ability to return overall relevant groups. However, MAP measure considers the rank information of relevant groups in the recommendation list. They are defined as follows.

$$P@K = \frac{N_{relevant}}{K} \tag{18}$$

$$MAP = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{1}{m_i} \sum_{j=1}^{N} P(R_{ij}) \tag{19}$$

where K is the number of recommended groups and in this setting, K is set to 10; $N_{relevant}$ is the number of relevant groups in the ranking list; $|U|$ denotes the number of user; m_i is the number of relevant groups to the user j; $P(R_{ij})$ represent the precision of recommended results from the top result until you get togroup k.

3.3 Experimental procedure

We evaluate our proposed algorithms in the Top N

recommendation evaluation framework. Similar to the experimentation in (Bogers and Van Den Bosch, 2011), We divide each data set into a training and test set by randomly selecting 10% of the users to be in our test set (871 users for CiteUlike and 4161 users for Lastfm). Final performance is evaluated on this 10% so-called active users by withholding 20% their joined groups. If an active user has less than five joined groups, we used one group affiliation in the test set. We optimize parameters involved in recommendation models on the training set using 10-fold cross-validation. The details of this evaluation setting can be referred to Boger and Bosch's work (2011). For the performance comparison of our method and existing methods, we implemented our method and existing group recommendation approaches in the literature. They were listed as follows:

1) Vector Space Model Method (abbreviated VSM): This method uses TF-IDF value of terms to represent user profile and group profile. Then group candidates are ranked by the calculated cosine similarity of user and group profile vectors.

2) Semantic Content Filtering Method (abbreviated SCF): This is our proposed semantic group recommendation method in Section 3.1. It is an enhanced VSM method.

3) Graph Proximity Model Method (abbreviated GPM): This is a graph-based method leveraging Katz measure and it has achieved good performance for group recommendation (Kim and El Saddik, 2013).

4) Social Aggregation Filtering Method (abbreviated SAF): This is our proposed social group recommendation method in Section 3.2. It combines social connections, behavioral connections and semantic connections to recommend interest groups.

5) Semantic-Social Fusion Method (abbreviated SSF): This is our proposed fused group recommendation method which leverages semantic content and online connections to generate recommendations for the user. Since two data fusion strategies are employed in this study, we obtained two fused group recommendation methods: SSF_{Comb} and SSF_{Map}. .

Among these five methods, the first two represented content-based 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 are tested in the two datasets and the results will presented in the next section.

4. RESULTS ANALYSIS

In this section, we present the detailed comparison of results. The values for evaluation metrics are obtained and compared between the state-of-the-art methods and our proposed approach on the CiteULike dataset and Lastfm. The detailed results are shown in Table 2.

It can be easily observed from this table that our proposed SSF approaches achieve the best performance in terms of P@10 metric and MAP metric. For CiteULike dataset, collaborative filtering methods (GPM and SAF) achieves better performance than content-based methods (VSM and SCF) while the gains are not obvious. Among four baseline methods, GPM obtained highest P@10 scores and SAF obtained highest MAP scores. For Lastfm dataset, collaborative filtering methods achieves better performance than content-based methods and the gains are obvious. The reason for this difference between two datasets may be that groups in CiteUlike have more centralized semantic description while groups in Lastfm have dispersed semantic description. GPM obtained highest P@10 scores and highest MAP scores among baseline methods in Lastfm dataset. Although GPM has better performance than SCF, it often costs high time to compute proximity degree and shows inefficiency in the real applications. Our proposed methods obtain highest recommendation quality since they leverage advantages of content-based approaches and collaborative filtering approaches and alleviate disadvantages of them. method is good at improving P@10 measure, which indicates that it can recommend more relevant interest groups to online users. method does well in improving MAP measure, which indicates that it can rank relevant groups higher to online users.

Table 2: Comparison results of five methods on the two dataset.

		VSM	SCF	GPM	SAF		
CiteuLike	P@10	0.0865	0.0973	0.1025	0.1014	0.1247	0.1225
	MAP	0.0843	0.0965	0.1027	0.1036	0.1170	0.1208
Lastfm	P@10	0.0532	0.0521	0.0870	0.0854	0.1064	0.1042
	MAP	0.0501	0.0489	0.0881	0.0827	0.1012	0.1178

5. CONCLUSIONAND FUTURE WORK

As online groups in social media websites are coming into broad use as an important way of sharing experiences and information, locating interest groups has become a critical research issue. In this paper, we propose a semantic social fused group recommendation framework by leveraging semantic content analysis and social network mining. Profiles of users are built from two aspects: semantic content and heterogeneous connections. To overcome shortcomings of traditional content-based and collaborative filtering based methods, we rank the group candidates according to the fused recommendation score from the pre-computed matching degree score and social aggregation score. We also employ Mapreduce framework to support large scale similarity computation. Finally, the effectiveness of the proposed approach over baselines is verified in two real social media datasets.

There are several limitations in this research. First, we compute keyword similarity to expand user profile. We are aware that the use of domain ontology will greatly help to r esolve semantic ambiguity in keyword matching. Thus, in t he future, research domain ontology can be constructed to s upport extended profile matching. Second, this paper adopt s the CombMNZ technique as the rank aggregation method. Some complex data fusion techniques (Nandakumar et al., 2008) can also be considered.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (71490725, 91546114, 71371062, 71302064, 71501057) and Hefei University of Technology (JZ2014HGBZ0368).

REFERENCES

Aljukhadar,M.,Senecal,S.,Daoust,C.-E.,(2012).Using Recommendation Agents to Cope with Information Overload.*International Journal of Electronic Commerce* 17, 41-70.

Aslam, J.A., Montague, M., (2001) Models for metasearch, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, pp. 276-284.

Bogers,T.,Van Den Bosch,A., (2011) Fusing Recommendations for Social Bookmarking Web Sites. *International Journal of Electronic Commerce* 15, 31-72.

Chen,W.-Y., Zhang,D.,Chang,E.Y., (2008) Combinational collaborative filtering for personalized community recommendation,*Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM*, pp. 115-123.

Chowdhury, G., (2010) Introduction to modern information retrieval. *Facet publishing*.

Croft,W.B.,Metzler,D.,Strohman, T., (2010) Search engines: Information retrieval in practice. *Addison-Wesley*.

Elsayed,T.,Lin, J.,Oard,D.W., (2008) Pairwise document similarity in large collections with MapReduce, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics*, pp.265-268.

Figueiredo, F., Pinto, H., Belém, F., Almeida, J., Gonçalves, M.,Fernandes,D.,Moura,E.,(2012) Assessing the quality of textual features in social media. *Information Processing & Management*.

Fox, E., Shaw, J., (1994) Combination of multiple searches. *NIST SPECIAL PUBLICATION SP*, 243-243.

Hsu,M.-H.,Chen,H.-H., (2 0 1 1) Efficient and effective prediction of social tags to enhance web search. *Journal of the American Society for Information Science and Technology 62*, 1473-1487.

Kim,H.-N.,El Saddik,A.,(2013)Exploring social tagging for personalized community recommendations. *User Modeling and User-Adapted Interaction* 23, 249-285.

Lee,D.H.,Brusilovsky,P., (2010) Using self-defined group activities for improvingrecommendations in collaborative tagging systems, *Proceedings of the fourth ACM conference on Recommender systems. ACM*, pp. 221-224.

Lillis,D.,Zhang,L.,Toolan,F.,Collier,R.W.,Leonard,D., Dunnion,J.,(2010) Estimating probabilities for effective data fusion, *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM*, pp. 347-354.

Manning,C.D.,Raghavan,P.,Schütze,H., (2008) Introduction to information retrieval. *Cambridge University Press Cambridge*.

Montague,M.,Aslam,J.A., (2002) Condorcet fusion for improved retrieval,*Proceedings of the eleventh international conference on Information and knowledge management. ACM*, pp. 538-548.

Nandakumar, K.,Chen,Y.,Dass, S.C.,Jain, A.K., (2008) Likelihood ratio-based biometric score fusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30, 342-347.

Pudota,N.,Dattolo,A.,Baruzzo,A.,Ferrara,F.,Tasso,C., (2010) Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems* 25, 1158-1186.

Quattrone, G., Capra, L., De Meo, P., Ferrara, E., Ursino, D., (2011) Effective retrieval of resources in folksonomies using a new tag similarity measure, *Proceedings of the 20th ACM international conference*

on Information and knowledge management. ACM, pp. 545-550.

Schifanella,R.,Barrat,A.,Cattuto, C., Markines, B., Menczer, F., (2010) Folks in folksonomies: social link prediction from shared metadata, *Proceedings of the third ACM international conference on Web search and data mining. ACM*, pp. 271-280.

Sun, J.,Ma, J., Liu, Z.,Miao,Y.,(2013) Leveraging Content and Connections for Scientific Article Recommendation in Social Computing Contexts. *The Computer Journal*, bxt086.

Symeonidis, P.,Tiakas, E.,Manolopoulos,Y., (2011) Product recommendation and rating prediction based on multi-modal social networks, *Proceedings of the fifth ACM conference on Recommender systems. ACM*, pp. 61-68.

Vasuki,V.,Natarajan,N.,Lu,Z.,Dhillon,I.S.,(2010) Affiliation recommendation using auxiliary networks, *Proceedings of the fourth ACM conference on Recommender systems*. *ACM*, pp. 103-110.

Wu,S.,(2012)Applying the data fusion technique to blog opinion retrieval. *Expert Systems with Applications* 39, 1346-1353.

Zeng,W.,Zeng,A.,Shang,M.-S.,Zhang,Y.-C., (2013) Membership in social networks and the application in information filtering. *The European Physical Journal B* 86, 1-7.

Zheng,N.,Li,Q.,Liao,S.,Zhang,L. (2010) Which photo groups should I choose? *A comparative study of recommendation algorithms in Flickr. J. Inf. Sci.* 36, 733-750.