Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

Neighborhood-user profiling based on perception relationship in the micro-blog scenario

Jianxing Zheng[a], Bofeng Zhang[a,*], Xiaodong Yue[a], Guobing Zou[a], Jianhua Ma[b], Keyuan Jiang[c]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China
[b] Faculty of Computer and Information Sciences, Hosei University, Japan
[c] Department of Computer Information Technology & Graphics, Purdue University Calumet, IN, USA

ARTICLE INFO

Article history:
Received 7 September 2014
Received in revised form
25 June 2015
Accepted 26 June 2015
Available online 3 July 2015

Keywords:
Micro-blog
Ontology
Resource perception relationship
Follow perception relationship
Neighborhood-user profile

ABSTRACT

In the micro-blog scenario, personal user profiling relying on content is limited for recommending desired diverse subjects due to its shortcomings of short text, often leading to a poor recall. Currently, many methods only utilized the personal knowledge from each individual user to represent user profile without considering the neighborhood information. However, resource information related to neighboring friends play an important role in improving the performance of recommender systems. In this paper, we present the personalized expanded user profiling for micro-blog subject recommendation via ontology semantics structure. Next, taking into account diffusion ability of followee friends, we discuss resource perception relationship (RPR) and follow perception relationship (FPR). Finally, we discuss how, by adjusting the importance of RPR and FPR, the neighborhood is selected to construct neighborhood-user profile, which can mine new relevant subjects for target user. Our experimental results demonstrate the effectiveness of our neighborhood-user profiling in comparison to the existing collaborative filtering and personal user profile recommendation approaches on Sina micro-blog platform datasets.

1. Introduction

Nowadays, as a newly emerged communication tool and public medium on the Internet, micro-blog spreads popular hot topics from one user to millions of individuals just in a few minutes, which allows the user to receive desired information anytime and anywhere. Meanwhile, searching for personalized interests and feelings posted to the multi-source information platforms, such as micro-blog systems like Twitter, social network sites including Facebook and LinkedIn, and personal homepages and blogs as well as many others [1,2], is an interesting yet challenging task. Especially, in the micro-blog system platforms, people repost a lot of short messages about their daily activities and feelings so as to maintain latest interests or friendship.

Many researchers have successfully tested feasibility of applications in many areas including interesting topics [3] and micro-blog environments [4]. However, personal User Profile (UP) is a customized model of interest representing and reasoning for a user, which is implicitly contained and generated from one's behaviors, browsing contents, or feedbacks [5–7]. That is, how to fulfill personalized activities and information requirements with one's micro-blog user profile is an important yet challenging issue. Very little research, however, has been done on this issue.

In the micro-blog scenario, each micro-blog is short and lacks sufficient information for user profile construction. As is expected, a user profile is not only generated from individual short messages, but also profits from existing interactions of friends [8–10]. With more than 215 million users and more than 175 million postings per day in 2012, Twitter is one of the most prominent micro-blog services on the web [11]. In particular, most of users are used to forwarding tweets for communication, instead of directly posting. Hence, followees make an important role in the propagation and spread of personalized interests. Traditional user profiles capture personal interests over one's own knowledge [12–15], which are not holistic for discovering diverse information. In this case, items and products in user profiles could not reflect currently concerned subjects and socially propagative topics thoroughly.

In many scenarios, traditional collaborative filtering (CF) strategy provides users a lot of valuable information on the basis of mutual understanding and knowing. In social communities such as

[*] Corresponding author.
E-mail addresses: jxzheng185@gmail.com (J. Zheng), bfzhang@shu.edu.cn (B. Zhang), yswantfly@shu.edu.cn (X. Yue), guobingzou@gmail.com (G. Zou), jianhua@hosei.ac.jp (J. Ma), jiang@purduecal.edu (K. Jiang).

http://dx.doi.org/10.1016/j.websem.2015.06.005
1570-8268/© 2015 Elsevier B.V. All rights reserved.

Facebook, LinkedIn and Twitter, the solution of CF is challenging. First, sparse data derived from short text is insufficient to capture enough similar users to recommend desired items, which hurt both the precision and recall of recommender systems. Second, pluralistic society makes people generate diverse interests. Not only are users restricted to daily monotonous interesting item, but they may be interested in diverse items posted by their friends [16]. A vast amount of diverse data enables similarity between users is small, which also leads the capture of similar users is hard. Meanwhile, a small amount of similar users is difficult to discover the items of the high correlation, which cannot effectively be applied into CF strategy [9]. Actually, when we follow what the followees have written, we can reflect our interests in a tracking way; and when we glimpse other followees or communities, we can realize where the interests come from. Follow friends' knowledge is a kind of effective collective wisdom, which could extend personal interest to other latent but relevant subjects. Additionally, follow relation is a new back-to-back linkage, which can supply the target user diverse interests from collaborative users [9,17]. It is reasonable that these follow friends contain a group of intimate interest users, named as neighborhood. Therefore, neighborhood with sufficient knowledge could help an individual user build the Neighborhood User Profile (NUP), addressing the problem of information shortage in representing personal interest.

In this work, we, using Sina micro-blog data source, constructed novel neighborhood user profile based on the collective knowledge. First, taking into account roles of followee friends in the interest propagation of the target user, we investigated the follow perception relationship and resource perception relationship. Furthermore, by adjusting the importance of two kinds of relationships, we discovered the neighborhood of a user. Lastly, the NUP relying on neighbor interests is proposed. In addition, the proposed NUP is evaluated by comparing against the existing personal UP and CF recommendation methods through experiments on large amounts of data from the Sina micro-blog platform.

Our experimental results show that the proposed NUP approach outperforms other methods in both precision and recall but with a relatively higher time complexity. By analyzing the expanded interests by NUP, we have observed that the recommendations based on NUP can accelerate the diffusion of the user interest, especially some semantically related interest between friends. We introduce the idea of neighborhood to solve the problem of acquiring behavior interest of social users. In particular, with the consideration of both the roles of followees' friends and resource perception relationship equally, the selected neighborhood could expand semantic interest efficiently. When the neighborhood only includes oneself, the NUP becomes a conventional individual user profile. However, the zooming size of neighborhood is an important issue for interest supplement related to social networks and social Webs, which needs to leverage the adaptive diversification fusion algorithm for zooming-in and zooming-out of the neighborhood.

The remaining part of this paper is organized as follows. Section 2 briefly discusses the works related to user profile. In Section 3, an overview of our recommendation framework based on neighborhood user profile is presented. In Section 4, we introduce personal interest acquiring method. Concepts of neighborhood and detailed descriptions of interest extending in neighborhood user profile are presented in Section 5. In Section 6, we demonstrate the application of our system as well as our experiment results along with discussions on strength and limitations. Finally, we conclude in Section 7 directions of our future work.

2. A brief review of UP works

In the scenario of user profile construction, how to convert the raw micro-blog documents into user's interesting subjects is usually challenging. To exactly recommend appropriate products to the user, many researchers have published their works in discovering demonstrated ways to build user profiles [18–20,12–14,10,21,22]. In this section, we will briefly review some popular works related to user profile.

2.1. Content-based UP

Content-based user profile focuses on document content analysis to classify the categorization of browsing historical records for deriving the hot interest and meaningful subjects for a target user. Many researches use individual information from the current search session or personal information to construct UP [23]. Via the key word set from one's comments and article, Meguebli et al. [13] built a user profile and article profile for each user, and computed the similarity between the article profile and user profile to sort the recommendation article list. By taking full advantage of informal and unstructured labeled data in Tweets, Lim [8] proposed a LDA-based Twitter Opinion Topic Model (TOTM) to aggregate or summarize opinions of a product, which can discover target specific opinion words and improve opinion prediction. Instead of employing a human-generated ontology, Harvey [14] proposed a novel latent topic models to describe both the clicked URLs and the interests of users from click log data. Considering the characteristics of short text messages expressing user's opinions and interests, Esparza [24] described users and products from the terms used in abbreviated and highly personalized commentary and studied how Twitter-like short-form messages could be leveraged as a source of indexing and retrieval information. Lin [15] utilized a semi-supervised variant of LDA that accounts for both text and metadata to characterize version features into a set of latent topics for exploring UP modeling in the app domain. By adopting features including user interest match, content-dependent user relationship and user influence, Wang [9] proposed a machine learned ranking function to find a new interest group of users for newly twitter information diffusion. Although these works have obtained the remarkable achievements, the existing user profiling methods are not fit for micro-blog users. In the micro-blog scenario, traditional UP methods capture too sparse interests to permit robust personalization and only recommend limited subjects for user's diverse needs. In this work we focus on the roles of social relationships between users for user profiling as they provide a richer source of information about the user's sufficient interests and preferences.

2.2. Semantics-based UP

Semantics-based user profile focuses on researching the semantic linkages of blogs or articles to discover a user's interest. Based on clustering the keywords from the electronic academic publications in online service, Tang [10] focuses on extending scientific subject ontology to refine user interest profiling. Varga [25] introduced a new semantic graph, called category metagraph, to extract a more fine grained categorization of concepts to provide a set of novel semantic features from short text messages. As the cosine similarity and TF–IDF weighting scheme for terms occurring in news messages are used in most user profiles, Hogenboom [26] extended semantics based weighting techniques, Bing–SF–IDF+, by considering the synset semantic relationships and by employing named entity similarities using Bing page counts, to perform better of F1 than TF–IDF and SF–IDF methods. For the hierarchical semantic structure embedded in the query and the document, Huang [27] used a deep neural network (DNN) to rank a set of documents for a given query and proposed a series of Deep Structured Semantic Models (DSSM) for Web search. Tao [21] constructed user profiles based on the ontology with world knowledge and user local knowledge, and utilized semantic specificity of concept in each ontology layer to mine user's

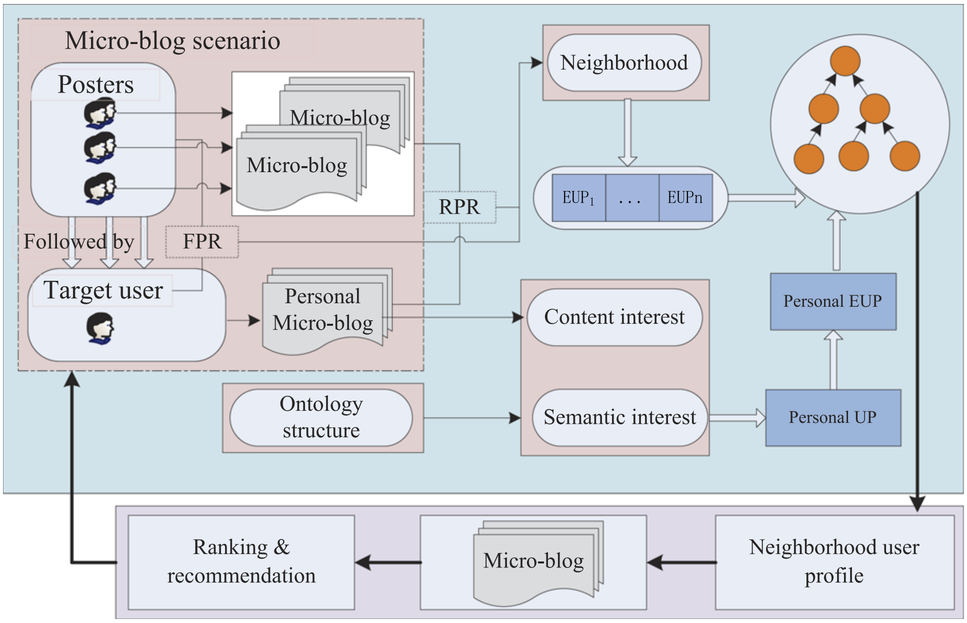

Fig. 1. Overview framework of recommendation based on NUP.

semantic interest. Cantador [22] clustered those personal semantic ontology user profiles from the tastes and preferences expressed by users to discover specific types of Community. Particularly, all these works indulge in semantics research in terms of vertical categories for personal user profile, and they did not take into consideration the semantic relationship of horizontal categories.

2.3. Relationship between UPs

Similar relationship between users has been studied over past years [17]. Group-based recommendation has been one of the most powerful techniques in collaborative filtering. By exploiting two levels of geographical neighborhood characteristics from location perspective, Liu [28] proposed a novel approach, namely Instance-Region Neighborhood Matrix Factorization (IRenMF) for location recommendation, which yields a more accurate modeling of users preferences on locations. In order to explore and model the structure correlations among users and items, Wang [29] designed a hierarchical group matrix factorization (HGMF) method for item recommendation. Aiming to reduce data sparsity and discover the latent characteristics of communities, Zhao [30] devised an LDA-based method for user recommendation in Twitter-style social networks.

Interactions among users in social networks also attracted a lot of attention [16]. By exploiting the "social" features of social folksonomies, Nocera [31,32] provided an "enhanced" user with recommendations of similar users and potentially interesting resources based on involved users and resources. Taking into account relations between friends in a large-scale real-world social network, Zhou [33] found that semantic web community evolves in a speckled fashion which is a highly distributed form. Based on identifying potential information flowing between nodes in a university email network, Kossinets [34] proposed a framework in order to analyze communication in networks. By analyzing the user-generated contents and opinions of their friends, Bao [35] explored a social probabilistic matrix factorization model to predict users' potential interests, which helps micro-blogging provide users with better personalized information services.

As described above, content-based approaches in the micro-blog scenario for recommendations have a low precision as tweet contents are typically short and noisy, while collaborative filtering approaches lead to a higher precision but data sparsity. Particularly, most mentioned researchers emphasize on recommending articles or products in terms of friends' roles in information diffusion while ignoring friends' important effects in the process of interest construction. In this paper, we consider roles of followee friends to find target user's interest and build NUP for the micro-blog recommendation.

3. Recommendation framework based on NUP

The NUP aims to discover a user's personal interest and discover his/her neighbor interests from followee friends and their similar resources. Illustrated in Fig. 1 is the overview of recommendation based on neighborhood user profile. As shown in Fig. 1, the personal user profile is obtained by integration of content interest and semantic interest. Then, personal interests can be accumulated in terms of semantics of the ontology category structure in order to form Expanded User Profile (EUP). Moreover, considering the social cognitive relationship between users, neighborhood can be discovered for stating target user's interest. Corresponding expanded user profile is modeled for each user in the neighborhood, as neighbor EUPs. Combining the personal EUP and EUPs from neighborhood, the intrinsic and latent interest of the target user is discovered and neighborhood user profile is modeled for effectively social recommendation.

The recommendation system mainly includes three steps. First, personal micro-blogs are used to acquire content interest. Additionally, based on the semantic specificity of the subject in ontology structure, personal UP is then constructed by considering the depth and width of the subject. Furthermore, using the subjects with the strongest semantic specificity, taking into account their content interest degrees in personal UP, personal EUP is generated by expanding the subjects and computing their interest degree in a vertical way at one level deep.

Second, similarity of users based on micro-blog resources is utilized to represent resource perception relationship (RPR). Simultaneously, similarity based on follow friend set is used to define follow perception relationship (FPR). By adjusting the relative weight of RPR and FPR, neighborhood of the target user is discovered. Then, EUP for each user in neighborhood is modeled so as to update target user's EUP by taking into account the relevance of subjects and attention degree of subjects. That is, close subjects in semantics from neighborhood are added to the latent interests of the target user for building neighborhood user profile.

Finally, the system recommends top k subjects based on the rank of interest degree in neighborhood user profile. This allows a chance to instantly receive common subjects derived from intimate friends and expand user's personal interests. The micro-blogs related to the above subjects are then pushed to the target user.

4. Personal interest extraction

In this section, we first model the personal content interest with TF–IDF weighting mechanism from micro-blogs. Then, we identify personal UP by considering the semantics of subjects and expand latent subjects in terms of ontology category structure.

4.1. Content interest acquisition

In the micro-blog scenario, there are numerous messages usually containing various opinions, which are difficult to be classified into appropriate categorizations. Many Web documents classification methods have been widely studied [36]. Here, we use message features (i.e., the term weighting scheme [37]) to analyze micro-blog texts. We use the data collected from Sina micro-blog platform to construct the dictionary and extract significant subjects and terms. In addition, the dictionary is updated regularly by crawling latest micro-blogs to ensure the accuracy of the corpus. Specifically, we consider personal local knowledge as a group of micro-blogs and each micro-blog m can be represented as a set of subject terms and corresponding weight, shown in Eq. (1).

$$m = \{(t_1, w_{1m}), (t_2, w_{2m}), \ldots, (t_p, w_{pm})\}. \tag{1}$$

The term frequency (TF) for term t in micro-blog m is calculated as:

$$tf_{tm} = \frac{freq_{tm}}{\max_l(freq_{lm})}, \tag{2}$$

where $freq_{tm}$ is the raw term frequency of t in micro-blog m and $\max_l(freq_{lm})$ is the frequency number of term l which has the maximum frequency in m. The inverse document frequency (IDF) for term t is defined as:

$$idf_t = \log\frac{N_m}{n_t}, \tag{3}$$

where N_m is the total number of micro-blogs and n_t is the number of micro-blogs that contains term t. Hence, the relative importance of term t to m can be formulated as:

$$w_{tm} = tf_{tm} * idf_t. \tag{4}$$

Furthermore, the similarity between two micro-blogs by means of the cosine measure can be defined as follows:

$$sim(m_i, m_j) = \frac{\overrightarrow{m_i} \bullet \overrightarrow{m_j}}{\|\overrightarrow{m_i}\| \cdot \|\overrightarrow{m_j}\|}. \tag{5}$$

With the similarity between micro-blogs, we obtain k-nearest neighbors for each micro-blog at adjusted threshold coefficient from the corpus training set. By considering the topic with the most number of neighboring samples, the target micro-blog can be divided into the appointed appropriate topic with *KNN* algorithm. At last, each micro-blog is partitioned into one of predefined topics such as Economy, IT, Sports and Culture, which are from categorization corpus of Sogou Lab Data. The length of the sample document vector was set 30 and the value of parameter k was set 15 to help discriminate the category of the micro-blog. After obtaining user's interest subjects, the interest degree of some subject in each category for each user can be modeled by the ontology category solution. According to the category structure, we can model user's interest tree and expand semantic related interest subjects. Fig. 2 shows part of categories structure related to the four topics, and we assume that there is a kind of part-of relation from a child node to the ancestor node.

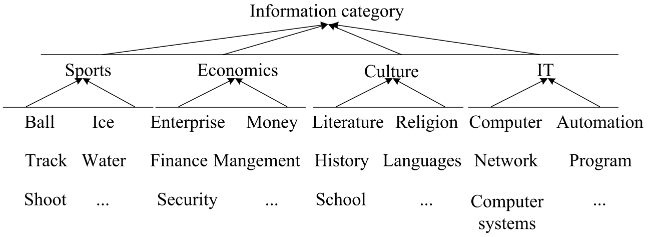

Fig. 2. Part of information categories on four topics.

As a powerful mechanism, ontology was proposed as a popular definition of an explicit specification of a conceptual model by Gruber [38]. Many researchers captured relational domain knowledge and developed personalized domain ontologies [39,40]. According to the Dewey decimal classification, King et al. [41] applied IntelliOnto in the field of distributed web information retrieval in order to improve the performance quality. Sieg et al. [42] used the Open Directory Project to create personalized ontologies for specifying user's preference and interest in web search. Downey et al. [43] developed Wikipedia to study user interest in queries. Also, we define domain ontology knowledge with Wikipedia categories to help individual users to discover semantic user profiles. Because ontology construction is beyond the scope of this paper, we focus primarily on the classification information of ontology. The primitive parts in Fig. 2 can be formally defined as follows.

Definition 1. Let 3-tuple $\Theta = (S, R, A)$ be an ontology, where S is the set of subjects; R is the set of relation types; A is the set of axioms that includes some rules and restrictions that constrain the attributes of subject object or relation.

Definition 2. Let R be a set of relations. S is a set of subjects. $\forall s_1, s_2 \in S, r \in R$, if $s_1 \xrightarrow{r} s_2$, then we define s_1 has a relation r to s_2.

Here, r is a kind of part-of relation from a child node to the ancestor node in Fig. 2.

According to the four topics and their subject classification information, we infer personal content interest degree of subject s based on all the micro-blogs involved in the topic C. Given all the micro-blogs involved in subject $s (s \in S)$, we can calculate content interest degree on subject s for user u as Eq. (6):

$$Cid_u(s) = \frac{\sum\limits_{m \in M_u^C} w_{sm} \times \eta(s, m)}{\sum\limits_{s_i \in S} \sum\limits_{m \in M_u^C} w_{s_i m} \times \eta(s_i, m)}, \tag{6}$$

where M_u^C represents the micro-blog set reposted by user u over topic C. $\eta(s, m) = 1$ if $s \in m$; otherwise $\eta(s, m) = 0$. Similarly, $\eta(s_i, m) = 1$ if $s_i \in m$; otherwise $\eta(s_i, m) = 0$.

4.2. Personal UP based on content

4.2.1. Semantic coverage degree of subject

The approach of ontology semantic mining mainly focuses on hierarchical semantics relations. In our approach, we argue that semantic coverage degree (Scd) of a subject involves two aspects: vertical semantic coverage (Vsc) and horizontal semantic coverage (Hsc). Generally, a higher Vsc indicates that the subject is of high significance in specific search, while a higher Hsc states that the subject category is more ramose and feasible in wide search of subject.

For a specific subject in an ontology structure, Algorithm 1 gives the procedure which calculates its semantic coverage degree.

Algorithm 1 Semantic coverage degree calculation.

```
Input:
    ontology $\Theta = (S, R, A)$, a coefficient $1 < \lambda < 2$.
Output:
    $Scd(s)$ applied to ontology.
1:  get the root set $S_0$ of $S$ from $\Theta$, for $s \in S_0$, set $Scd(s) = 1$;
2:  remove $S_0$, get the new root nodes set $S'$ from $\Theta$;
3:  if ($S' == \varnothing$) then
4:      return;
5:  end if
6:  for each $s' \in S'$ do
7:      get the parent node $s_0$ of $s'$;
8:      if $s' \xrightarrow{r} s_0$ then
9:          compute vertical semantic coverage degree $Vsc(s') = \lambda \times Scd(s_0)$;
10:     end if
11:     get the sibling nodes set $S_{s'}$ of $s'$;
12:     compute horizontal semantic coverage degree $Hsc(s') = \log(1 + |S_{s'}|)$;
13:     compute semantic coverage degree $Scd(s') = Vsc(s') \times Hsc(s')$;
14: end for
15: $S_0 = S_0 \cup S'$, go to step 2.
```

For the semantic coverage degree of a subject in Algorithm 1, we consider two important assumptions about the subject density. First, as stated in [23], the subjects at upper levels toward the root are more abstract than those at lower levels toward the leaves. The lower level subjects will have more comprehensible concepts than upper level subjects in web search. Thus, the Vsc of a subject at a lower level should be greater than that of one at its upper level, which is reflected by the coefficient operation λ. Obviously, in step 9, the definition of Vsc states that semantics increases recursively. Second, a subject with more child nodes is more specific and understandable than that with fewer descendants for a disciplinary area, which displays more semantic coverage ability. That is, the Hsc of a subject should be larger than that of one with fewer descendants, which is represented by calculation of logarithmic operation log. In step 12, the assignment of Hsc reflects that the rise of semantics. Hence, we can design the semantic coverage degree of each subject with the logarithmic running time of the number of ontology tree nodes.

In above algorithm, the semantic coverage degree is becoming larger and larger with the increasing depth of subject, which makes remarkable discrimination for two detailed specified subjects in lower nodes. Actually, the semantics of lower specified subjects are closer than that of upper subjects. Thanks to this long tail, we normalize the Scd values to the range [0, 1] to distinguish semantics between upper nodes significantly by an exponential function. For a subject s, the normalized value $\overline{Scd}(s)$ of semantic coverage degree is given by the following formula:

$$\overline{Scd}(s) = \frac{1 - e^{-Scd(s)}}{1 + e^{-Scd(s)}}. \tag{7}$$

Analogously, when a user pays more attention to a subject with ample semantic coverage ability, one would have a strong semantics understanding ability to the subject. Based on the subjects the user is interested in, the semantic interest degree of s for user u can be calculated as Eq. (8):

$$Sid_u(s) = \frac{\sum\limits_{m \in M_u^C} \overline{Scd}(s) \times \eta(s, m)}{\sum\limits_{s_i \in S} \sum\limits_{m \in M_u^C} \overline{Scd}(s_i) \times \eta(s_i, m)}. \tag{8}$$

4.2.2. Personal UP

In this subsection, we infer a user's personal interest based on the content interest degree and semantic interest degree of a subject. The personal interest degree is derived by combining the roles of two factors as shown in Eq. (9):

$$I_u(s) = \frac{2 \times Cid_u(s) \times Sid_u(s)}{Cid_u(s) + Sid_u(s)}. \tag{9}$$

From Eq. (9), we can see that the target user only has reposted more number of micro-blogs for a specific subject, and the greater semantic coverage degree of the subject is, the more the user is interested in the subject.

Based on this idea, we can construct personal UP including the semantics by the following definition.

Definition 3. Let u be a target user, personal UP can be determined by two-tuples $\Theta_u = (S_u, I_u)$. $S_u \subseteq S$ is the set of subjects. I_u is a real number set. $\forall s \in S_u, \exists I_u(s) \in I_u, I_u(s) > 0$.

Personal UP is formed by a group of subjects that the target user is interested in and their corresponding interest degree of the subjects.

4.3. Personal EUP based on semantics

In the micro-blog scenario, personal interesting subjects of a user are quite few due to the limitations of the micro-blog short text. However, we can discover a user's supplementary interests by vertical semantics expansion in the ontology structure, which is called as personal EUP.

Here, we have an assumption that when a user is interested in a subject s, one should acquire tastes from all child subjects of s in terms of the ontology category structure. Therefore, according to the target user's existing interesting subjects, semantically related to subjects of personal UP could be added into S_u in order to generate the expanded subject set ES_u and expanded interest degree set EI_u.

Algorithm 2 describes the interest expansion of the personal UP.

As a new subject is added, the interest degree of the subject is also updated. Algorithm 2 shows specific interest expansion and interest degree update methods. Obviously, in steps 3 and 4, the extensible subject nodes are selected with time complexity of O(nlogn). In step 11, this algorithm shows us how to assign the interest degree of objects and add them into the original interest set. As an example, Fig. 3 shows the change process from personal UP to EUP. In Fig. 3, the green objects are ordinary subjects while the pink objects are interesting subjects for user u. We select black objects as the candidate set for available expanding subjects. This allows us to select their red child nodes in order to expand probable interesting subjects, which are added into personal UP forming as the EUP. Additionally, the interest degree of red nodes inherits that of their ancestor nodes because it is reasonable that the user is very interested in the added nodes as well as their parent nodes.

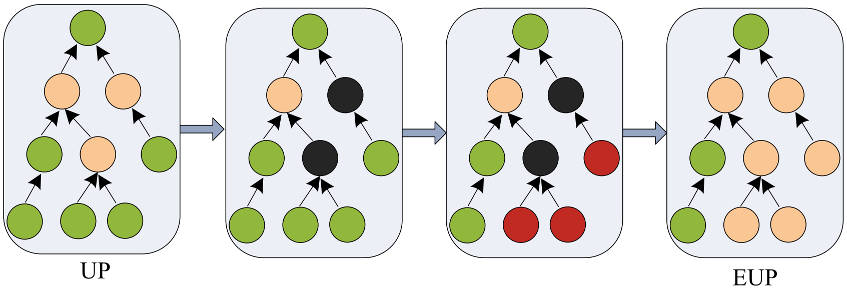

Fig. 3. Personal EUP construction based on semantics expansion.

Algorithm 2 Semantic interest expansion.

```
Input:
    ontology $\Theta = (S, R, A)$, personal UP $\Theta_u = (S_u, I_u)$, initial set $S_0 = \varnothing$.
Output:
    $E\Theta_u = (ES_u, EI_u)$ applied to ontology structure.
set $S_0 = S_u$, $ES_u = S_u$, $EI_u = I_u$;
for each $s \in S_0$ do
    while there exists $s_0 \in S_0$, it holds that $s \rightarrow \cdots \rightarrow s_0$ do
        $S_0 = S_0 - \{s_0\}$;
    end while
end for
for each $s \in S_0$, get the children node set $C_s$ of $s$ do
    if ($C_s == \varnothing$) then
        return;
    end if
    for each $s' \in C_S$ do
        compute interest degree of expanded subject $EI_u(s') = I_u(s)$;
        expand subject set $ES_u = ES_u \cup \{s'\}$;
        expand interest degree set $EI_u = EI_u \cup EI_u(s')$;
    end for
end for
```

5. Interest extraction based on NUP

In this section, we present a novel approach to discover neighborhood of the target user, and the NUP is modeled by integrating neighbor EUPs into a personal EUP. Afterwards, characteristics of NUP are analyzed.

5.1. RPR & FPR

In micro-blog platforms, UP ought to not only consider individual micro-blogs but also pay attention to behaviors from followee friends, which could enhance personal interest globally. In this subsection, we adopt resource relationship and follow relationship between users to discover the neighborhood of a user for expanding user's interests. In order to compute the neighborhood interest, we give preliminary definitions about neighborhood.

Definition 4. Let u be a target user. M_u represents the set of all the micro-blogs that u reposts.

Definition 5. Let u_i, u_j be two users. M_{u_i} and M_{u_j} correspond to micro-blog set of u_i and u_j, respectively. The RPR from u_i to u_j can be defined:

$$R^r_{ij} = \frac{\sum\limits_{m_i \in M_{u_i}} \sum\limits_{m_j \in M_{u_j}} sim(m_i, m_j)}{|M_{u_i}||M_{u_j}|}, \tag{10}$$

where $sim(m_i, m_j)$ is the similarity between m_i and m_j. $|M_{u_i}|$ and $|M_{u_j}|$ are cardinal number of micro-blog set M_{u_i} and M_{u_j}, respectively.

Followees are the best direct way to reflect useful interests of a user. In the micro-blog scenario, follow actions can depict perception relationship between users, which are formalized as follows.

Definition 6. Let u be a target user, friend set followed by u is denoted as $F_u = \{u_j | u \xrightarrow{f} u_j\}$.

Definition 7. Let u_i, u_j be two users. Followee friend sets are F_{u_i} and F_{u_j}, respectively. The FPR from u_i to u_j can be calculated by Eq. (11):

$$R^f_{ij} = \gamma \times f(u_i, u_j) + (1 - \gamma) \times \frac{|F_{u_i} \cap F_{u_j}|}{|F_{u_i}|}, \tag{11}$$

where $|F_{u_i}|$ and $|F_{u_j}|$ are cardinal number of followee friend set F_{u_i} and F_{u_j}, respectively. $f(u_i, u_j)$ represents the function of follow relationship from u_i to u_j, and $f(u_i, u_j) = 1$, if $u_i \xrightarrow{f} u_j$; otherwise $f(u_i, u_j) = 0$.

In this case, due to differences of mutual follow relationships between users, the perception degree R^f_{ij} is different from R^f_{ji}, which is a kind of new cognitive relationship in social networks. As we know, if a user has followed another user, one would accept potential and latent preference from the interested individual easily. When they follow each other mutually and have not common friends, their perception relationship can be viewed as 1. Here, we only consider directed follow perception relationship, we set the weight parameter $\gamma = 0.5$ harmoniously to adjust the follow action weight from u_i to u_j.

Also, the increasing of γ can produce heavier weight for the follow action between users, which makes follow perception relationship increase easier. Hence, plenty of close friends can be selected for interest expanding.

Based on closeness relationships in follow action and resource content, the comprehensive perception degree from u_i to u_j can be calculated by Eq. (12):

$$R_{ij} = \alpha R^f_{ij} + (1 - \alpha) R^r_{ij}. \tag{12}$$

By using above perception relationship, neighborhood of a target user can be selected by predefined threshold δ, defined as below.

Definition 8. Let u_i be a target user. Given u_j, the perception relationship from u_i to u_j is R_{ij}. The neighborhood set of u_i can be defined as $N_{u_i} = \{u_j | R_{ij} \geq \delta\}$.

The value of α is used to weigh the relative importance between follow perception relationship and resource perception relationship in the process of selecting neighboring friends. Therefore, the variation of the parameter α would determine which perception relationship has greater influence in affecting the selection of neighborhood. Especially, when the value of α is increasing, the users followed tightly by the target user are selected; instead, the users close to the target user by the resource are chose with a lower value of α.

Note that for the desired interest of the target user, these neighbor users can reveal inevitable interesting subjects. Hence, we can extend personal EUP to NUP by taking neighbor interests into consideration.

5.2. NUP modeling

We design an interest extension method that extracts latent interesting subjects of a target user u from the users in neighborhood N_u, named as NUP. First, we define the NUP of u as below.

Definition 9. Let u be a target user. NUP of u can be defined as two-tuples $N\Theta_u = (NS_u, NI_u)$. $NS_u \subseteq S$ is the set of subjects. NI_u is a real number set. $\forall s \in NS_u, \exists NI_u(s) \in NI_u, NI_u(s) > 0$.

According to neighborhood set N_u, personal EUP for each user in N_u is modeled. Based on EUPs from N_u, some novel subjects for target user u are supplemental by the semantic relevance. Meanwhile, the interest degree of extended subject is updated with neighbor interest. Generally, interest selection mechanism of the herd effect can make users follow the interest degree by the will of the majority, and a leader has a strong control ability of interest. That is, the closer perception relationship between the target user and others is, the greater the impact on decision making is.

Definition 10. Let u_i be a target user. $E\Theta_{u_i} = (ES_{u_i}, EI_{u_i})$, $N\Theta_{u_i} = (NS_{u_i}, NI_{u_i})$ correspond to personal EUP and NUP. $\forall s' \in NS_{u_i}$, the interest degree of subject s' can be defined as:
if $s' \notin ES_{u_i}$,

$$NI_{u_i}(s') = \frac{\sum\limits_{u_j \in N_{u_i}(s')} R_{ij} \times EI_{u_j}(s')}{\sum\limits_{u_j \in N_{u_i}(s')} R_{ij}}; \tag{13}$$

if $s' \in ES_{u_i}$,

$$NI_{u_i}(s') = \begin{cases} EI_{u_i}(s') \quad \text{if } 0 < \dfrac{|N'_{u_i}(s')|}{|N_{u_i}(s')|} < 0.5 \\ \max\left\{ EI_{u_i}(s'), \dfrac{\sum\limits_{u_j \in N'_{u_i}(s')} R_{ij} \times EI_{u_j}(s')}{\sum\limits_{u_j \in N'_{u_i}(s')} R_{ij}} \right\} & \text{else.} \end{cases} \tag{14}$$

Where $N_{u_i}(s') = \{u | u \in N_{u_i}, EI_u(s') > 0\}$ is the set of users in N_{u_i} interested in subject s'; $N'_{u_i}(s') = \{u | u \in N_{u_i}(s'), EI_u(s') > EI_{u_i}(s')\}$ is the set of users from neighborhood, which are more interested in subject s' than target user. As one can see from Eqs. (13) and (14), neighborhood interest evaluates the interest effect of neighborhood for target user about a new subject in terms of collaborative wisdom; additionally, for an older subject, the attitude of target user is dominated by that of most of users in neighborhood. If more than half of users in neighborhood like subjects more than the target user, one may update the interest degree by neighbor collaborative strategy; otherwise, the target user insists on one's opinion. Based on the above strategy, Algorithm 3 gives an interest extension of personal EUP in terms of N_u.

Algorithm 3 Neighborhood interest expansion.

```
Input:
    ontology $\Theta = (S, R, A)$, personal EUP $E\Theta_u = (ES_u, EI_u)$,
    neighborhood set $N_u$, initial set $S_0 = \varnothing$.
Output:
    $N\Theta_u = (NS_u, NI_u)$ applied to ontology structure.
 1: set $S_0 = ES_u, NS_u = ES_u, NI_u = EI_u$;
 2: for each $u_j \in N_u$ do
 3:    for each $s' \in ES_{u_j}$ do
 4:       combine neighbor interest $NS_u = NS_u \cup \{s'\}$;
 5:    end for
 6: end for
 7: for each $s' \in NS_u$ do
 8:    while there exists $s \in S_0$, it holds that $s' \to \cdots \to s$ do
 9:       filter semantically related interest $S_0 = S_0 \cup \{s'\}$;
10:    end while
11: end for
12: generate neighbor subject set $NS_u = S_0$;
13: for each $s' \in NS_u$ do
14:    compute neighborhood interest degree $NI_u(s')$ by definition
       10;
15:    generate neighborhood interest degree set $NI_u = NI_u \cup$
       $NI_u(s')$;
16: end for
```

In Algorithm 3, the neighbor interest subjects from each user in neighborhood are selected by steps 3–5. In steps 8–10, the extensible subjects are filtered with the comparison against personal EUP. By the neighbor effects, the expanded subjects increase semantic focus and specificity of subjects in ES_u as much as possible. If a user in N_u is more specific interested in subject s than target user u, the corresponding children subjects of s should be merged for u, and the interest degree is assigned in terms of collaborative wisdom among users in neighborhood. The algorithm also belongs to the one with the time computational complexity of polynomial time. To illustrate the detailed steps of interest fusion, Fig. 4 shows the process of NUP construction. EUP_1 and EUP_n are extended UPs for users in N_u. The pink objects are their interesting subjects, which are supplied for EUP's expansion. Comparing against target user's subjects in EUP, the black objects are selected from neighbor EUP_1 and EUP_n for more semantic specificity. The crimson objects represent the subjects updated by most of users in neighborhood. Combining the origin EUP of u, the NUP is modeled.

As we expected, setting the value of δ, neighborhood relying on follow perception relationship is different from that of resource perception relationship. When the value of α is small, the neighborhood are dominated by resource perception relationship, which can supply redundant subjects. Otherwise, the neighborhood are dominated by follow perception relationship, which cannot provide sufficient interest to supplement.

Also, the number of neighborhood users may affect target user's interest supplement. When the value of δ is small, the fewer neighborhood users cannot supply enough subjects to add user's interest; meanwhile, diverse uncorrelated subjects are added to user if too many neighborhood users expand various interests.

5.3. Analysis of NUP

The NUP refers to the interesting subjects deriving from neighbor collective effects. Actually, the number of subjects in NUP is more than that of personal EUP and UP; simultaneously, the range of interest is wider than that of personal EUP and UP, which increases the selective chance of recommendation.

According to Algorithms 2 and 3, we reach the following conclusions.

Property 1. *Let u be a target user. $\Theta_u = (S_u, I_u)$, $E\Theta_u = (ES_u, EI_u)$, $N\Theta_u = (NS_u, NI_u)$ correspond to personal UP, EUP and NUP. We have $S_u \subseteq ES_u \subseteq NS_u$. $\forall s \in NS_u$, we also have $NI_u(s) \geq EI_u(s) \geq I_u(s)$.*

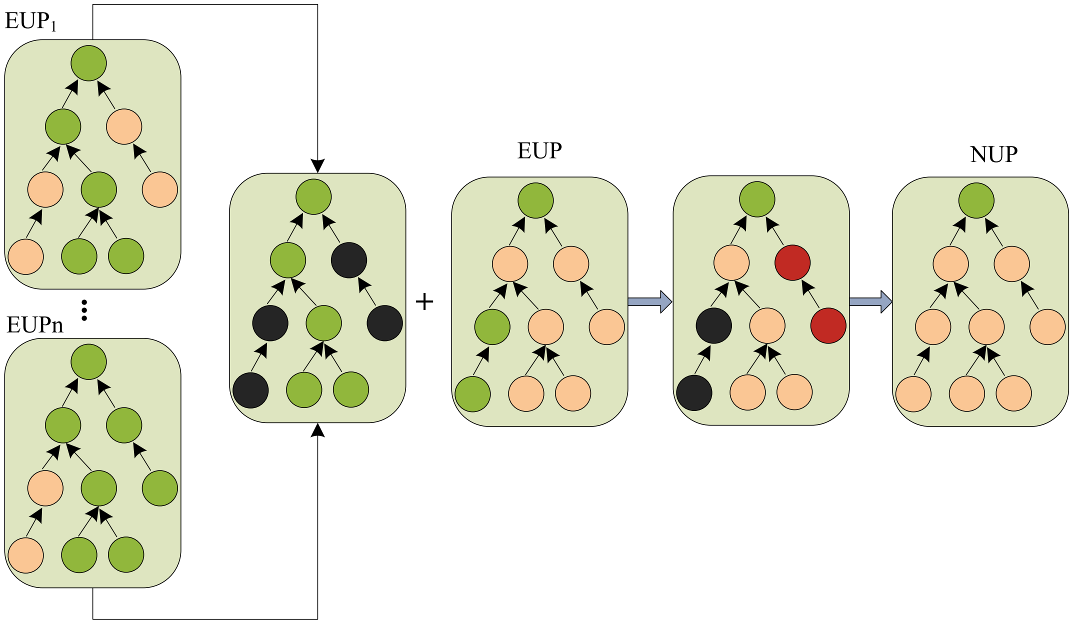

Fig. 4. NUP construction based on neighborhood.

This property states that neighborhood can increase user's absolute interest in a certain subject.

Property 2. *Let u_i, u_j be two users. $E\Theta_{u_i} = (ES_{u_i}, EI_{u_i})$ and $E\Theta_{u_j} = (ES_{u_j}, EI_{u_j})$ correspond to personal EUPs for $\Theta_{u_i} = (S_{u_i}, I_{u_i})$ and $\Theta_{u_j} = (S_{u_j}, I_{u_j})$, respectively. If $S_{u_i} \subseteq S_{u_j}$, we have $ES_{u_i} \subseteq ES_{u_j}$.*

Property 3. *Let u_i, u_j be two users. N_{u_i} and N_{u_j} are neighborhood sets of u_i and u_j. $N\Theta_{u_i} = (NS_{u_i}, NI_{u_i})$ and $N\Theta_{u_j} = (NS_{u_j}, NI_{u_j})$ correspond to NUPs for $E\Theta_{u_i} = (ES_{u_i}, EI_{u_i})$ and $E\Theta_{u_j} = (ES_{u_j}, EI_{u_j})$, respectively. If $ES_{u_i} \subseteq ES_{u_j}$ and $N_{u_i} \subseteq N_{u_j}$, we have $NS_{u_i} \subseteq NS_{u_j}$.*

Property 3 describes a user with more neighbor followee friends will have stronger potential interest probability for other subjects. In addition, the proposed NUP could enhance the interest range of a user against CF method and personal UP. For example, Table 1 gives a useful comparison of NUP and other methods.

Example data (Table 1).

Example 1. An example of interesting subjects of NUP is given by Table 1.

This is a personal UP system of $\{u_1, u_2, u_3, u_4\}$. Here, $\{u_2, u_3, u_4\}$ is followee friends set of u_1. For u_1, we cannot discover new interesting subject set $\{s_4, s_5\}$ from historical set $\{s_1, s_2, s_3\}$. Therefore, the interest degree of $\{s_4, s_5\}$ for u_1 based on UP is represented as asterisks.

In addition, from the interest observations of $\{u_2, u_3, u_4\}$ on $\{s_1, s_2, s_3\}$, only u_4 can be used to predict the interest of $\{s_4, s_5\}$ according to the historical contents. Based on the collaborative idea, the interest degree of s_5 can be supplied with $I_{u_1}(s_5) = 0.5 \times \frac{0.64\times0.26}{\sqrt{(0.64-0.26)^2+(0-0.5)^2}} = 0.13$. The interest degree of s_4 is still not assigned.

However, considering the idea of NUP, suppose that $\alpha = 0.5$ and $F_{u_1} \cap F_{u_2} = \varnothing$, so there is $R_{12} = \alpha R^f_{12} + (1-\alpha)R^r_{12} = 0.5 \times 0.5 + 0.5 \times 0 = 0.25$. Similarly, $R_{13} = 0.25$; $R_{14} = 0.5 \times 0.5 + 0.5 \times 0.2650 = 0.38$. By setting $\delta = 0.2$, we can find $N_{u_1} = \{u_2, u_3, u_4\}$, so the interest degree of extended subjects $I_{u_1}(s_4) = \frac{0.25\times(0.44+0.45)}{0.25+0.25} = 0.45$, $I_{u_1}(s_5) = \frac{0.25\times0.58+0.38\times0.50}{0.25+0.38} = 0.53$.

6. Application and experiment evaluation

In this section, we present some recommendation applications using the proposed NUP, and evaluate the performance of the NUP by comparing with personal UP and CF recommendations. Simultaneously, we share some insights from our observations and analysis of the NUP-based subject recommendation.

6.1. Recommendation strategy

Using the NUP, top k subjects are selected to reflect the target user's future interest trend. Furthermore, we can recommend micro-blogs related to subjects for the target user. Particularly, the relevance of micro-blogs with the subject leverages traditional machine learning methods to classify the categories of micro-blogs, which is not investigated elaborately here.

In addition, the similarity between personal UPs considering content is computed to discover the most similar top k users for CF mechanism, as shown in Eq. (15):

$$sim(u_i, u_k) = \frac{\sum\limits_{s\in S_{u_i}\cap S_{u_k}} (I_{u_i}(s) - \overline{I_u})(I_{u_k}(s) - \overline{I_u})}{\sqrt{\sum\limits_{s\in S_{u_i}\cap S_{u_k}} (I_{u_i}(s) - \overline{I_u})^2}\sqrt{\sum\limits_{s\in S_{u_i}\cap S_{u_k}} (I_{u_k}(s) - \overline{I_u})^2}}. \tag{15}$$

where $S_{u_i} \cap S_{u_k}$ is the set of subject both interested in by user u_i, u_k; $I_u(s)$, $\overline{I_u}$ are the personal interest degree for subject s and average interest degree of user u in personal UP, respectively.

Based on similarity between users, the interest degree of subject s by the CF method is shown in Eq. (16):

$$Cfid_{u_i}(s) = \frac{\sum\limits_{u_k\in N_{u_i}} I_{u_k}(s) \times sim(u_i, u_k)}{\sum\limits_{u_k\in N_{u_i}} |sim(u_i, u_k)|}. \tag{16}$$

The top k subjects have been selected for target user to recommend relevant micro-blogs.

Table 1
The interests of NUP against UP and CF methods for user u_1.

	s_1	s_2	s_3	UP		CF		NUP	
				s_4	s_5	s_4	s_5	s_4	s_5
u_1	0.64			*	*	*	**0.13**	**0.45**	**0.53**
u_2		0.25		0.44		0.44		0.44	
u_3			0.36	0.45	0.58	0.45	0.58	0.45	0.58
u_4	0.26		0.50		0.50		0.50		0.50

Table 2
Statistics showing the number of users, followees, users' micro-blogs, followees' micro-blogs, test micro-blogs in each dataset.

	Nlpir	*Application* 1	*Application* 2
Users	85	75	338
Followees	1247	4013	5,467
Users' micro-blogs	4326	216	6,949
Followees' micro-blogs	2943	7863	179,890
Test micro-blogs	2641	666	6,952

6.2. Evaluation of recommending subjects

In this section, the effectiveness of NUP is interpreted by comparing the differences against the personal UP and CF recommendation. As is explained, the personal UP involves interesting subjects based on the content without taking into account semantics.

6.2.1. Data description and experiment design

In our experiment, we use *Nlpir* dataset and two *Application* datasets to create UPs involved in four domains. The *Nlpir* dataset is from *Nlpir* website (http://www.nlpir.org/), which is from Sina platform. The dataset was selected during a period from 4th of Dec. 2011 to 23rd of Dec. 2011, and 85 users and 1247 followee friends were used elaborately for NUP construction. For the *Nlpir* dataset, each user has more than 10 followees and 30 micro-blogs. In *Application* dataset 1, there are 4386 users in our database to crawl Sina micro-blog platform (http://open.weibo.com) for the newest micro-blogs, attended friends, mutual responses and interactions from the 10th of Apr. 2013 to the 29th of Apr. 2013. In order to get pure content information for NUPs, we only selected the primitive users including 75 test users and 4013 followee friends for our experiments and recommendation generation. For this dataset, each user has more than 50 followees and a few micro-blogs while their followees have large numbers of micro-blogs to supplement interest. In *Application* dataset 2, 10 thousand users are used to collect their reposting micro-blogs and their follow relationships from the 10th of Oct. 2013 to the 24th of Oct. 2013 in terms of Tencent Weibo platform. We select 338 users, 5467 followee friends and their 0.2 million micro-blogs to build neighborhood user profile. Each user has at least 30 followees and 40 micro-blogs to discover one's neighborhood friends. Statistics for all datasets are listed in Table 2. For all the datasets, we split user's micro-blogs set into two parts according to the time period, the earlier period was used for modeling user profile and latter period was used for testing.

In our research, Wikipedia was used to help understand categories structure of subjects (http://zh.wikipedia.org). Moreover, domain categories on four topics were utilized to discover the fundamental semantic coverage degree of the subjects, including sports, economics, culture, and IT. In the process of semantics transferring via categories information, we set coefficient $\lambda = 1.2$ to simulate the variation trend of semantic specificity of the subjects.

In order to evaluate the performance of our proposed recommendation mechanism, we used these performance evaluation metrics: precision and recall [22]. These metrics have been defined as follows:

$$Precision = \frac{|S_T \cap S_R|}{|S_R|}, \tag{17}$$

$$Recall = \frac{|S_T \cap S_R|}{|S_T|}, \tag{18}$$

where $S_T = \{s|I(s) > 0\}$ is the actual subject set where test micro-blogs involved and S_R is the recommendation subject set relying on the NUP, personal UP or CF method for each user.

6.2.2. Results and evaluation

This section presents the experimental recommendation results of several recommendation approaches. We conducted experiments and computed the precision and recall by varying the value of subject recommendation list size k from 4 to 12 with an increment of 2 for the first two datasets while the last one adopts the recommendation list size k from 4 to 20 with an increment of 4. As is known, the parameters α and δ will make a great influence on the performance of NUP. Figs. 5–7 (left) present precision results for all the datasets. From the figures, it is understood that there is a clear benefit for the NUP subject recommendation strategy than the CF and UP approaches. For example, in the case of *Nlpir* dataset with the recommendation list of size 4, we can see that the NUP approach enjoys a precision score of approximately 0.51, indicating that, on average, more than 170 of 340 recommended subjects by NUP are likely received by 85 users. However, for CF and personal UP methods, the precision values are 0.46 and 0.43, respectively. This represents that the subjects focused by most of followees is prone to be accepted by target user in the micro-blog scenario. The neighborhood can contain implicit interest of target user and make a prominent role in aspect of interest acquiring of target user. Additionally, Figs. 5–7 (left) shows that the precision results based on the personal UP and CF approaches are becoming gradually close when the list size is greater than 10. However, the precision values are still lower than that of NUP. As is known, in the micro-blog scenario, the similarity of users is small due to the shortage and impurity of text messages, and the performance of the CF method is certainly poor. Considering the FPR, the semantic relevant subjects from neighborhood are also recommended to the target user although micro-blog their contents are not consistent. Therefore, the precision of NUP outperforms those of other methods.

The recall results by three kinds of recommendation approaches are shown in Figs. 5–7 (right). Obviously, for all datasets, our NUP approach achieves a better performance than CF and UP while a small number of subjects is selected. Hence, there are enough subjects to match user's preferences in order to achieve high recall. For example, Fig. 7 shows significant recall performance which states that the expansion of the subjects. This is likely due to the results of comprehensiveness of neighborhood recommendation. Enough followees and supplement micro-blog content can supply adequate interests for the target user. However, for *Nlpir* dataset, the recall of NUP is more and more close to that of the CF method with the increasing recommendation list size. This is because the number of followee users in this dataset is relatively small, and the follow relationship between users is weak. The number of semantic relevant subjects expanded by NUP is small, which supply limited interesting subjects while most of recommended subjects are impure as k rises.

As mentioned in the above section, the value of parameter α can affect the perception relationship between users, which makes the neighborhood of a user varying so as to form different NUPs. Figs. 8–10 show the precision and recall results at different k values for different α values with $\delta = 0.1, \delta = 0.3, \delta = 0.5$ for

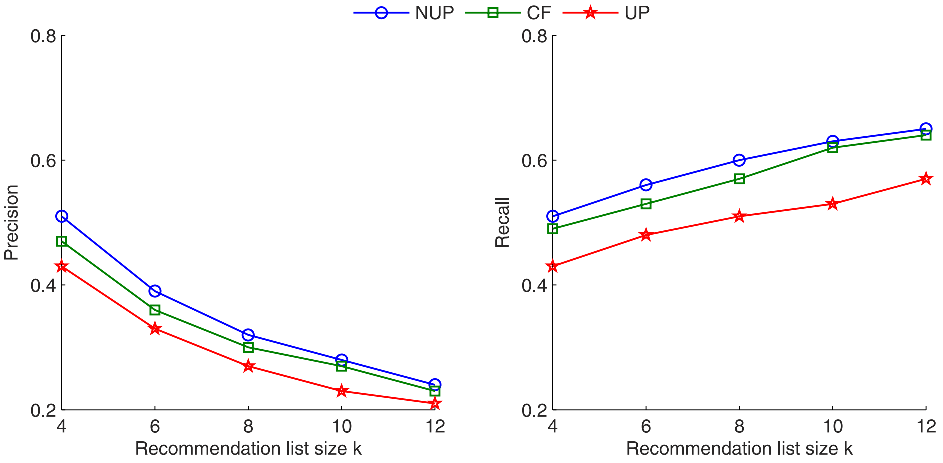

Fig. 5. *Nlpir* dataset: precision and recall based on NUP, CF, and UP recommendation approaches under different *k* for 85 users ($\alpha = 0.5, \delta = 0.1$).

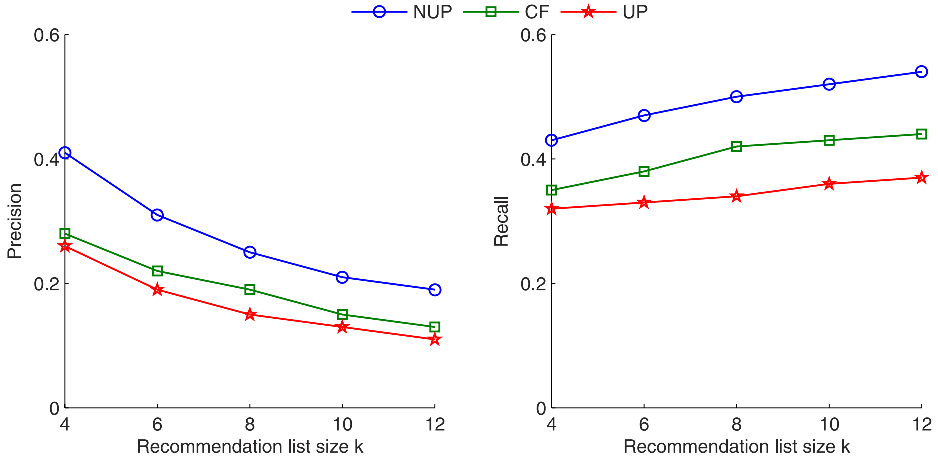

Fig. 6. *Application* dataset 1: precision and recall based on NUP, CF, and UP recommendation approaches under different *k* for 75 users ($\alpha = 0.5, \delta = 0.3$).

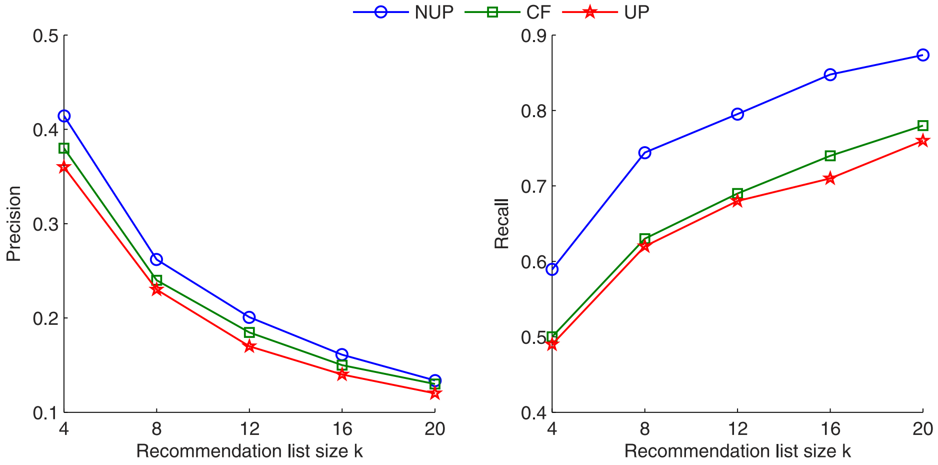

Fig. 7. *Application* dataset 2: precision and recall based on NUP, CF, and UP recommendation approaches under different *k* for 338 users ($\alpha = 0.5, \delta = 0.5$).

Nlpir, *Application* 1 and *Application* 2 datasets, respectively. Obviously, it can be seen that the same prominent changes happen at $\alpha = 0.5$ with precision first increasing and then decreasing for different recommendation lists. Interestingly, in all datasets, the recall is maximized at $\alpha = 0.5$ for most of result-lists of size, which describes that both considering the roles of follow perception relationship and resource perception relationship equivalently are beneficial for neighborhood selecting and NUP construction. Comparing Fig. 9 against Figs. 8–10, we find users with less reposting micro-blogs and more followees are likely to receive neighbor interest subjects. Specially, the curves change dramatically with parameter α varying from 0.3 to 0.5 for *Application* 1 dataset, which indicates that the roles of FPR and RPR change strongly in the process of NUP construction. We see that there is no prominent rising

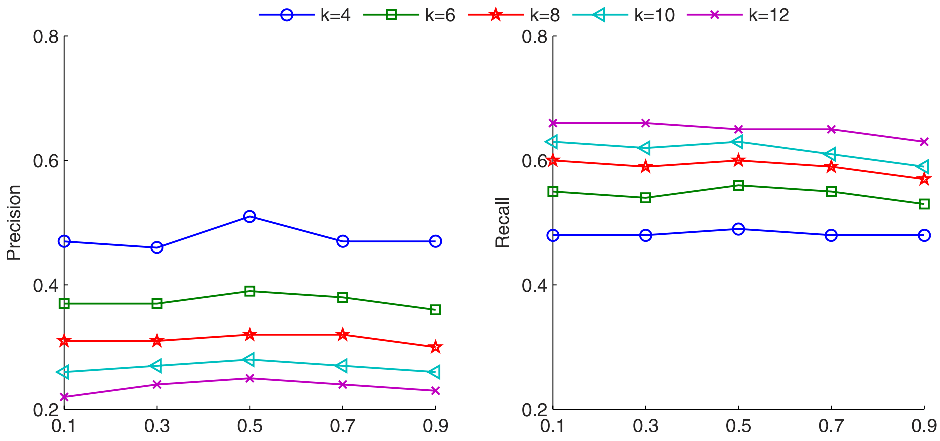

Fig. 8. *Nlpir* dataset: precision and recall based on NUP recommendation under different α with different values of k ($\delta = 0.1$).

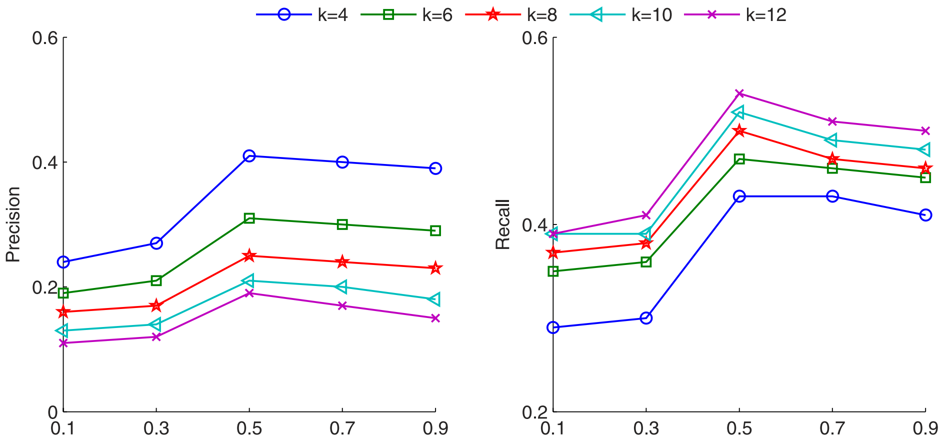

Fig. 9. *Application* dataset 1: precision and recall based on NUP recommendation under different α with different values of k ($\delta = 0.3$).

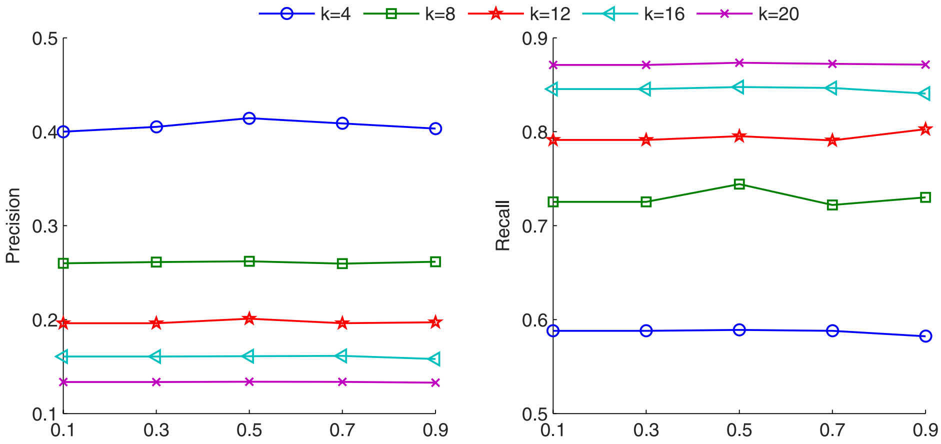

Fig. 10. *Application* dataset 2: precision and recall based on NUP recommendation under different α with different values of k ($\delta = 0.5$).

trend for the precision and recall when the value of α exceeds 0.5, implying that the role of FPR is dominated in the neighborhood selecting.

In addition, from Figs. 8–10, we can see that the precision and recall under other values of α are both relatively worse than that of $\alpha = 0.5$. Therefore, we can predict that the best balance of precision and recall is achieved at $\alpha = 0.5$ for most of recommendation lists. This is because FPR is dominant in the process of neighborhood selecting when $\alpha \geq 0.5$, which leads to the fact that subjects recommended are too irrelevant. Contrarily, thoroughly considering RPR makes the number of neighborhood decrease because differences of the resource content give rise to a low similarity between users when $\alpha < 0.5$.

In Figs. 11–13, we compare the precision and recall results of NUP under different δ with different values of k for three datasets. It can be seen that the best performance is achieved with $\delta = 0.1$, $\delta = 0.3$ and $\delta = 0.5$, respectively. For example, the precision increases first and then decreases with the value of δ changing in *Nlpir* dataset, which reaches the maximal at $\delta = 0.1$, presenting 0.51, 0.39, 0.32, 0.28, 0.24 for list size 4, 6, 8, 10, 12, respectively. Interestingly, the precision values are also maximal as 0.41, 0.31, 0.25, 0.21 and 0.19 for list size 4, 6, 8, 10, 12 at $\delta = 0.3$ in *Application*

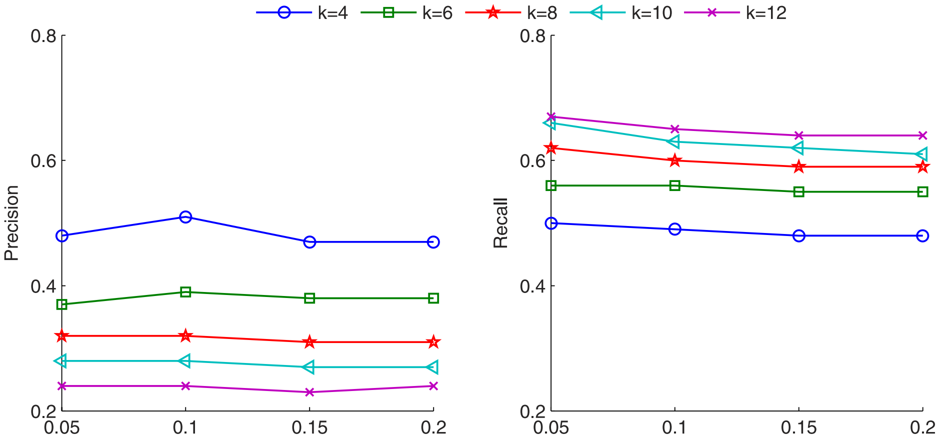

Fig. 11. *Nlpir* dataset: precision and recall based on NUP recommendation under different δ with different values of k ($\alpha = 0.5$).

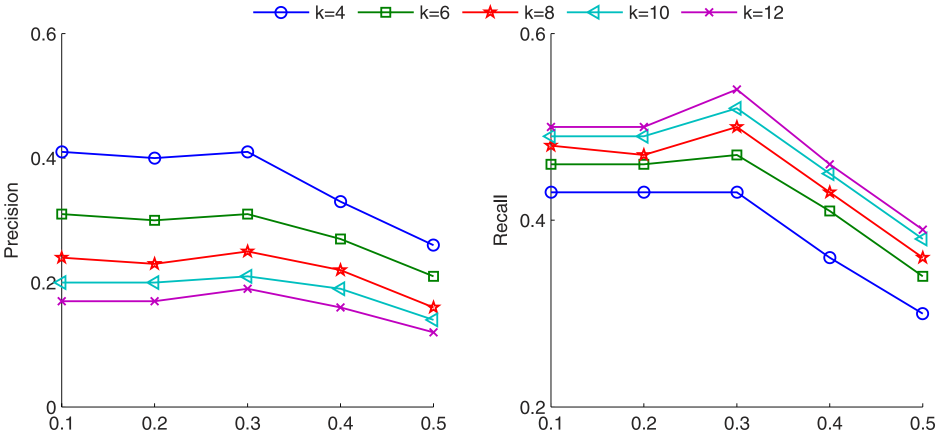

Fig. 12. *Application* dataset 1: precision and recall based on NUP recommendation under different δ with different values of k ($\alpha = 0.5$).

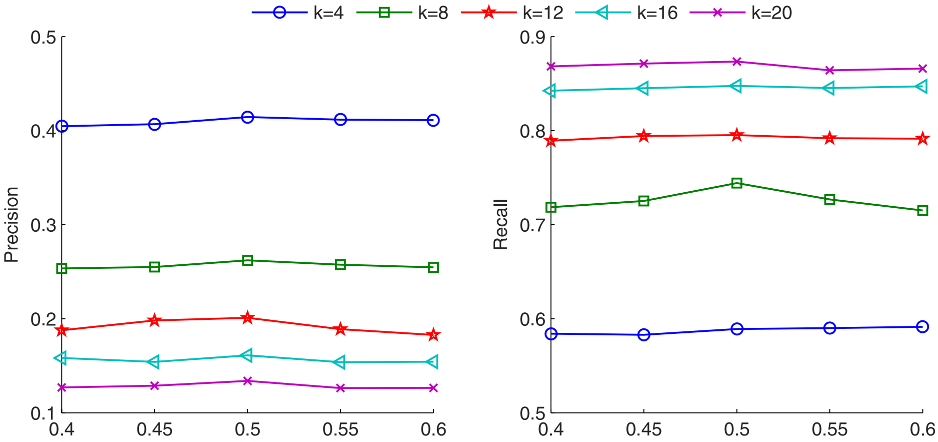

Fig. 13. *Application* dataset 2: precision and recall based on NUP recommendation under different δ with different values of k ($\alpha = 0.5$).

dataset 1. Also, the precision values present as 0.41, 0.26, 0.20, 0.16 and 0.13 for list size 4, 8, 12, 16, 20 at $\delta = 0.5$ in *Application* dataset 2. As we can observe, the mean number of neighborhood is small when the parameter δ has a relatively large value, and the number of subjects recommended is also few. Additionally, when the value of δ is too low, the mean number of neighborhood is big enough to supply abundant subjects so as to reduce the purity of recommendation performance. However, we note that a lot of follow relationships exist in *Application* dataset 1 and the perception relationship degree between users is large. Therefore, we set a larger parameter $\delta = 0.3$ to maintain the relatively stable number of neighborhood users to build NUP. In *Application* dataset 2, as selected users have a lot of followees and many reposting micro-blogs, the perception relationship between users is large enough to discover one's neighborhood, which leads to a gentle curve for the variation of performance in Fig. 13.

6.2.3. *Analysis of interest extension*

To analyze the expansion of interest subject, we adopt the AUC [51] metric (ROC) to verify the effectiveness of our method. For all users, their mean AUC value states that the average percent of non-relevant subjects of recommendation methods. Table 3 shows

Table 3
Mean AUC of NUP, CF and UP methods for 338 users.

	Percent of non-relevant subjects		
	NUP	CF	UP
Precision	0.27	0.30	0.34

Table 4
Precision and recall under different numbers of followee friends for *Application* dataset.

	The number of followee friends		
	(0–10)	(10–20)	(20–30)
Precision	0.39	0.48	0.38
Recall	0.41	0.38	0.39

Table 5
Precision under different numbers of UP, micro-blogs, test micro-blogs for *Application* dataset.

	No. of UP	No. of micro-blogs	No. of test micro-blogs
[0–0.3)	30	75	255
[0.3–0.6)	24	97	193
[0.6–1]	21	44	218

the mean AUC of NUP, CF and UP methods for *Application* dataset 2. In Table 3, we can see that the NUP method recommend less non-relevant subjects than other methods.

As described above, the recommendation performance of the proposed NUP is affected by the features of a user, including the social activity and interaction ability of a user. As is expected, when the number of friends for NUP is few, there are not enough appropriate subjects to help the target user obtain the desired information in order to improve the precision. However, large numbers of followee friends can present abundant subjects to precipitate the user lost the most popular topics and lower the recommendation performance. Table 4 shows the precision and recall statistics with different numbers of followee friends in *Application* dataset 1. In Table 4, the results show that most of users are prone to receiving their followees' information to obtain favorite subjects when the number of their followees lingers about 10–20.

Generally, if a user reposts micro-blogs frequently, one can get enough information without needing neighbor helps. In contrary, an isolated or silent user would prefer obtaining relevant information from neighbor friends. Table 5 shows variations of precision range under different number of UP, micro-blogs, test micro-blogs for 75 users in *Application* dataset 1. The precision intervals are divided in three range types [0–0.3), [0.3, 0.6), [0.6, 1]. From Table 5, we analyzed the differences of users under different precision ranges, and attempted to find which kind of users is fit for the NUP approach. As we can see, the numbers of users are 35, 24 and 16 under precision range [0–0.3), [0.3, 0.6) and [0.6, 1], respectively. Simultaneously, the number of micro-blogs for the NUP is gradually smaller with the precision increasing. For these users, neighborhood plays an important role in the process of interest acquiring. Hence, in Table 5, there is a basic phenomenon for NUPs that users have less micro-blogs in personal Ups generate better recommendation results with a high precision, which is consistent with what we expected.

To examine the variations of the number of users based on the NUP and personal UP in different subjects, we performed an experiment to construct the NUPs at $\alpha = 0.5, \delta = 0.3$. Fig. 14 shows distribution of the number of users in four subjects based on the NUP and personal UP, respectively. The results show that the number of users interested in different subjects is prominently changing. In general, the most interesting topics for the great majority of NUPs imply that the subjects gradually become popular in the propagation of social interest. For each user, that personal UP reveals that the existing partial individual interest. Actually, expanded subjects by neighborhood can be used to spread friends' interest to the target user, which has made the number of users interested in the subject larger and larger. In Fig. 14, the number of NUPs increases markedly comparing with that of the origin personal UP in each subject. For example, the number of users in IT and Culture rises up by 22 and 20, respectively. The ideal result is that all the users are interested in the extended subjects, which can be revealed by the effects of group interest. An application of this insight is to maximize the profit in micro-blog advertisements.

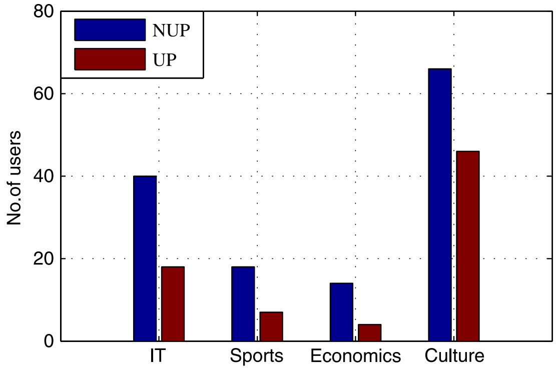

Fig. 14. Number of users about NUP and UP on different subjects for 75 users with $\alpha = 0.5, \delta = 0.3$.

7. Conclusions

This paper investigates how neighborhood of a user can be used to help build a novel NUP in the micro-blog scenario by addressing the drawbacks of existing UP approaches immersing in use of personal knowledge. First, personal UP is constructed based on integrating the content interest and semantic interest, and then the improved EUP is created. Second, taking into account the roles of RPR and FPR, neighborhood of a user is discovered. Lastly, we propose a NUP modeling approach based on neighbor interest and personal EUP.

In our experiments, the proposed NUP approach outperforms the personal UP and CF method, both in recommendation accuracy and coverage. We contribute to utilizing collective neighborhood information to mine personal potential interest to enhance the quality of recommending micro-blog subjects. In addition, the evaluation results show that FPR and RPR are both equally important for NUP construction, and the users rarely reposting micro-blog are fit for neighborhood helps. Interestingly, the threshold δ can determine the number of neighborhood users. Especially when the value of δ is 1, our NUP becomes the traditional personal UP.

As a short text form, micro-blogs from friends could be viewed as an important channel for mining social interest. However, linkages of users are various. Especially, social links based on user's action is multidimensional. In the future work, we plan to utilize our approach to discover the potential multidimensional interest community for cross-domain recommendation. In addition, the optimal number of neighborhood friends is necessary to research so as to reduce the time complexity of neighborhood selecting and decrease the repeatability of relevant interest. Hence, further study is required to investigate these effective approaches for improving diverse recommendation performance in the micro-blog scenario.

Acknowledgments

We would like to thank all of the anonymous reviewers for their insightful and constructive comments and useful suggestions that have significantly improved the quality of the manuscript. This work was partially supported by the National Natural Science Foundation of China (Nos. 61303096, 61103067), and the Natural Science Foundation of Shanghai (13ZR1454600).

References

[1] Y. Ma, Y. Zeng, X. Ren, N. Zhong, User interests modeling based on multi-source personal information fusion and semantic reasoning, Active Media Technol. 6890 (2011) 195–205.

[2] D. Rosaci, G.M.L. Sarné, Recommending multimedia web services in a multi-device environment, Inform. Syst. 38 (2) (2013) 198–212.

[3] K. Lerman, R. Ghosh, Information contagion: an empirical study of spread of news on dig and twitter social networks, in: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, Washington, USA, 2010.

[4] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, 2010, pp. 851-860.

[5] Y.W. Seo, B.T. Zhang, Learning user's preferences by analyzing web browsing behaviors, Artificial Intelligence 15 (6) (2001) 381–387.

[6] F. Carmagnola, F. Cena, O. Cortassa, C. Gena, I. Torre, Towards a tag-based user model: how can user model benefit from tags? User Model. 2007 (2007) 445–449.

[7] D. Rosaci, G.M.L. Sarné, Efficient personalization of e-learning activities using a multi-device decentralized recommender system, Comput. Intell. 26 (2) (2010) 121–141.

[8] K.W. Lim, W. Buntine, Twitter opinion topic model: extracting product opinions from tweets by leveraging hashtags and sentiment lexicon, in: Proceedings of the 23rd International Conference on Information and Knowledge Man-agement, CIKM2014, Shanghai, China, 2014, pp. 1319–1328.

[9] B. Wang, C. Wang, J.J. Bu, C. Chen, et al. Whom to mention: expand the diffusion of tweets by @ recommendation on micro-blogging systems, in: Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 2013, pp. 1331–1340.

[10] X. Tang, Q. Zeng, Keyword clustering for user interest profiling refinement within paper recommender systems, J. Syst. Softw. 85 (2012) 87–101.

[11] J. Lin, K. Sugiyama, M.Y. Kan, T.S. Chua, Addressing cold-start in app recommendation: latent user models constructed from twitter followers, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 2013, pp. 283–292.

[12] X. Liu, Y. Liu, K. Aberer, C. Miao, Personalized point-of-interest recommendation by mining users' preference transition, in: Proceedings of the 22nd International Conference on Information and Knowledge Management, CIKM2013, Burlingame, CA, USA, 2013, pp. 733–738.

[13] Y. Meguebli, M. Kacimi, B. Doan, F. Popineau, Building rich user profiles for personalized news recommendation, in: Proceedings of the 22nd Conference on User Modelling, Adaptation and Personalization, UMAP2014, Aalborg, Denmark, 2014.

[14] M. Harvey, F. Crestani, M. Carman, Building user profiles from topic models for personalised search, in: Proceedings of the 22nd International Conference on Information and Knowledge Management, CIKM2013, Burlingame, CA, USA, 2013, pp. 2309–2314.

[15] J. Lin, K. Sugiyama, M.Y. Kan, T.S. Chua, New and improved: modeling versions to improve app recommendation, in: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, Australia, 2014, pp. 647–656.

[16] P. De Meo, A. Nocera, D. Rosaci, Recommendation of reliable users, social networks and high-quality resources in a social internetworking system, AI Commun. 24 (1) (2011) 31–50.

[17] Y.M. Li, Y.L. Shiu, A diffusion mechanism for social advertising over microblogs, Decis. Support Syst. 54 (1) (2012) 9–22.

[18] J. Zheng, B. Zhang, G. Zhou, Multi-granularity recommendation based on ontology user model, in: Proceedings of the 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, Beijing, China, 2013, pp. 2194–2199.

[19] F. Tang, B. Zhang, J. Zheng, Friend recommendation based on the similarity of micro-blog user model, in: Proceedings of the 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, Beijing, China, 2013, pp. 2200–2204.

[20] J.A. Iglesias, P. Angelov, A. Ledezma, A. Sanchis, Creating evolving user behavior profiles automatically, IEEE Trans. Knowl. Data Eng. 24 (5) (2012) 854–867.

[21] X.H. Tao, Y.F. Li, N. Zhong, A personalized ontology model for web information gathering, IEEE Trans. Knowl. Data Eng. 23 (4) (2011) 496–511.

[22] I. Cantador, P. Castells, Extracting multilayered communities of interest from semantic user profiles: application to group modeling and hybrid recommendations, Comput. Human. Behav. 27 (4) (2011) 1321–1336.

[23] R.W. White, P. Bailey, L. Chen, Predicting user interests from contextual information, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, Massachusetts, 2009, pp. 363–370.

[24] S.G. Esparza, M.P.O. Mahony, B. Smyth, Mining the real-time web: a novel approach to product recommendation, Knowl.-Based Syst. 29 (2012) 3–11.

[25] A. Varga, A.E.C. Basave, M. Rowe, F. Ciravegna, Y. He, Linked knowledge sources for topic classification of microposts: a semantic graph-based approach, Web Semant. 26 (2014) 36–57.

[26] F. Hogenboom, M. Capelle, M. Moerland, et al. Bing-SF-IDF+: Semantics-driven news recommendation, in: Proceedings of the 23rd World Wide Web Conference, Seoul Korea, 2014.

[27] P.S. Huang, X.D. He, J.F. Gao, et al. Learning deep structured semantic models for web search using click through data, in: Proceedings of the 22nd International Conference on Information and Knowledge Management, CIKM2013, Burlingame, CA, USA, 2013.

[28] Y. Liu, W. Wei, A. Sun, et al. Exploiting geographical neighborhood characteristics for location Recommendation, in: Proceedings of the 23rd International Conference on Information and Knowledge Management, CIKM2014, Shanghai, China, 2014, pp. 739–748.

[29] X. Wang, W. Pan, C. Xu, HGMF: Hierarchical group matrix factorization for collaborative recommendation, in: Proceedings of the 23rd International Conference on Information and Knowledge Management, CIKM2014, Shanghai, China, 2014, pp. 769–778.

[30] G. Zhao, M.L. Lee, W. Hsu, Community-based user recommendation in Uni-directional social networks, in: Proceedings of the 22nd International Conference on Information and Knowledge Management, CIKM2013, Burlingame, CA, USA, 2013, pp. 189–198.

[31] A. Nocera, D. Ursino, An approach to providing a user of a "social folksonomy" with recommendations of similar users and potentially interesting resources, Knowl.-Based Syst. 24 (8) (2011) 1277–1296.

[32] P.D. Meo, E. Ferrara, D. Rosaci, G.M.L. Sarnè, Trust and compactness in social network groups, IEEE Trans. Cybernet. 45 (2) (2015) 205–216.

[33] L.N. Zhou, Li Ding, T. Finin, How is the semantic web evolving? a dynamic social network perspective, Comput. Human. Behav. 27 (4) (2011) 1294–1302.

[34] G. Kossinets, J. Kleinberg, D. Watts, The structure of information pathways in a social communication network, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, Las Vegas, NV, USA, 2008, pp. 435–443.

[35] H. Bao, Q. Li, S.S. Liao, S. Song, H. Gao, A new temporal and social PMF-based method to predict users' interests in micro-blogging, Decis. Support Syst. 55 (2013) 698–709.

[36] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[37] X. Quan, W. Liu, B. Qiu, Term weighting schemes for question categorization, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 1009–1021.

[38] T. Gruber, Translation approach to portable ontology specifications, Knowl. Acquis. 5 (2) (1993) 199–220.

[39] W.N. Borst, Construction of engineering ontologies for knowledge sharing and reuse (Ph.D. thesis), University of Twente, Enschede, 1997.

[40] L.A. Zadeh, Web intelligence and world knowledge the concept of Web IQ (WIQ), in: Proceedings of Annual Meeting of the North American Fuzzy Information Processing Society, 2004, pp. 1–3.

[41] J.D. King, Y. Li, X. Tao, R. Nayak, Mining world knowledge for analysis of search engine content, Web Intell. Agent Syst. 5 (3) (2007) 233–253.

[42] A. Sieg, B. Mobasher, R. Burke, Web search personalization with ontological user profiles, in: Proceedings of 16th ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, 2007, pp. 525–534.

[43] D. Downey, S. Dumais, D. Liebling, E. Horvitz, Understanding the relationship between searchers' queries and information goals, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, 2008, pp. 449–458.