Information Sciences 479 (2019) 55–75

Contents lists available at ScienceDirect

Information Sciences

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

Personalized recommendation based on hierarchical interest overlapping community

Jianxing Zheng[a,*], Suge Wang[a,b], Deyu Li[a,b], Bofeng Zhang[c]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, Shanxi 030006, China
[c] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Article history:
Received 7 January 2018
Revised 21 November 2018
Accepted 25 November 2018
Available online 26 November 2018

Keywords:
Ontology user profile
Multi-granularity similarity
Hierarchical interest overlapping community
Personalized recommendation

ABSTRACT

Ontology user profiles describe users' structural semantic interests. Studying similar relationships between user profiles is crucial to detecting interest overlapping communities. The novel view assumes that hierarchical interests of user profiles can generate multiple similarity relations, which is conducive to forming interest clusters. In this research, we develop a hierarchical interest overlapping community (HIOC) detection method and present a personalized recommendation model. First, content interest closeness and semantic interest closeness between user profiles are computed to measure multi-granularity subject similarity of users. Then, using the multi-granularity subject similarity and follow similarity of users, a heterogeneous hypergraph is constructed to represent an interest network. By application of the interest density peaks mechanism, the HIOC detection method is adopted for identifying communities of interest. Further, personalized interest prediction is implemented by consideration of the memberships of a user in a community and a subject distributed in a community. Finally, we verify the performance of the HIOC detection algorithm on several real networks and validate the effectiveness of the proposed recommendation approach. The experimental results illustrate that the proposed approach outperforms classical recommendation methods in precision and recall.

1. Introduction

Recently, various social media websites, such as Twitter, Tencent, and Sina Weibo, have become important information platforms for the provision of popular services [45]. However, the characteristics of short text in the micro-blog scenario renders the capture of sufficient preferences difficult, thereby reducing the quality of information services. While browsing content, people can interact with others and satisfy their needs in terms of numerous populations' relationships. These potential relationships reflect users' interest tendency and tastes. Therefore, online recommender systems have been studying users' relationships to predict personal preferences and improve users' reputations.

The quality of recommender systems highly relies on recommendation algorithms, which are mainly classified into three groups, namely, content-based filtering, collaborative filtering (CF), and hybrid approaches [1,4,14]. Although recent strategies produce positive effects, most recommender systems still encounter several challenges, namely, 1) cold start problem

* Corresponding author.
E-mail addresses: jxzhengsxu@163.com (J. Zheng), wsg@sxu.edu.cn (S. Wang), lidysxu@163.com (D. Li), bfzhang@shu.edu.cn (B. Zhang).

https://doi.org/10.1016/j.ins.2018.11.054
0020-0255/

occurring constantly for new users or new items; 2) disability in recommending multi-granularity interests due to disregard for relationships of structured user preferences; 3) insufficient diversity caused by the interest overlap derived from similar friends. In social recommender systems, calculation of user similarity and discovery of effective similar users are critical to addressing the above-mentioned questions.

Generally, users' interests are multi-dimensional and multi-granular. For example, some users may be passionate about coarse-grained interests such as "sports," whereas some other users prefer fine-grained subjects such as "basketball" and "football." For computing user similarity, conventional methods focus on the text distance or ontology concept distance, which ignores the interest structure and multi-layer semantic structure relationships [14]. In the case of two users who are interested in "sports," one of whom prefers "basketball" and the other wants "football," their similarity considers not only the similarity of the subject "sports" but also the semantic closeness of their structures. Comprehensive structured similarity calculation considers multi-grain levels of subjects, which reflect semantic structural relationships between users. The similarity can be used to find potentially intimate similar users and generate accurate recommendations. To our best knowledge, most recommender systems rarely provide interesting products from the perspective of multi-grain subjects. In this study, the similarity method models multi-granularity semantic interest relationships among users to find similar friends, thereby solving the cold start problem.

The social media information of users is useful in the development of socialized interest, which can form special interest communities [9]. Some studies have found that user profiles derived from communities can enhance the efficiency of recommender systems [29]. Various types of interests can model hierarchical multi-granularity semantic interest similarity and identify hierarchical interest communities. In an overlapping community, a user may belong to more than one community [9]. The overlapping communities that the user is involved in can supply relevant and diverse preferences. Various communities have different interest distributions. That is, in social networks, we can utilize hierarchical interest overlapping communities (HIOCs) to strengthen and expand socialized interests for users. Therefore, HIOC detection is considerably important in discovering similar users. The resource in overlapping communities solves the data sparsity and diversity problems in personalized recommendation.

We propose a novel recommendation method that is based on HIOC detection, which can effectively recommend relevant and diverse interests for target users. The ontology user profile is modeled for computing the multi-granularity subject similarity between users, which considers content interest closeness and semantic structural interest closeness. Then, in terms of multi-granularity subject similarity, we construct a heterogeneous hypergraph and compute the interest density of each node. On the basis of the interest density of nodes, we select community cores and perform HIOC detection to find communities of interest. By merging communities with substantial overlap, we analyze the membership of a user in communities and a subject distributed in communities to predict interest. Finally, we conduct experiments on Sina Weibo and Tencent Weibo datasets to verify the effectiveness of the proposed HIOC recommendation approach. The main contributions of this paper are threefold. (1) We design a multi-granularity similarity calculation method that describes comprehensive similar relationships between users. The similarity distinguishes semantic interests of users from structured aspects of subjects. (2) We develop the HIOC detection method in terms of interest density-fitness of a node. On this basis, users with similar interest distribution are clustered in the same community, thus potentially improving the accuracy efficiency of recommender systems. (3) We utilize the user and subject memberships in their overlapping communities to implement a novel recommendation model. The recommendation strategy based on memberships is conductive to expanding diverse interests.

The remainder of this paper is organized as follows. Section 2 briefly reviews the relevant literature. Section 3 explains the HIOC recommendation framework. Section 4 presents the calculation of the multi-granularity subject similarity of users. According to the multi-granularity similarity, the details of the proposed HIOC detection method and novel recommendation approach are elaborated in Section 5. Section 6 presents the computational experiments that compare the proposed technique with other recommendation methods. Finally, Section 7 concludes the paper.

2. Related works

2.1. Overlapping community detection

Overlapping community detection can describe the variability and diversity of user interests, which is helpful for satisfying users' information demands. Some research studies have focused on non-overlapping or disjointed communities. By introducing the modularity function of global optimization, Newman proposed the fast Newman algorithm to improve the quality of community detection [27]. However, Lancichinetti and Fortunato showed that the modularity function suffers from a resolution limit and extreme degradation under overlapping community conditions [18]. By considering the overlapping and the hierarchy of communities, Lancichinetti et al. proposed a local fitness maximization method for community detection [19]. According to the number of distinct w-edge paths, Long [24] defined edge intensity for selecting skeleton edges. Referring to core nodes connected by skeleton edges, each margin node is divided into the nearest community through the calculation of belongingness coefficients. Through spectral analysis of line graphs, Gui et al. [10] designed an overlapping and hierarchical community structure that completes a balance between overlap and hierarchy. Given the complexity of users' interests and needs, communities allow for easy presentation and modeling of user interest. With use of a novel density-based clustering method, Xie et al. incorporated multi-faceted relations in social media, proposed a mechanism of augmented folksonomy graph (AFG), and then discovered latent user communities from AFG [43]. By defining one's struc-

tural properties in a social network, Tang proposed a probabilistic topic model called the Role-Conformity Model (RCM), for modeling user behaviors [47]. Lee et al. [20] investigated a pervasively community in which nearly each node belongs to multiple communities. Cantador [7] proposed the automatic identification of communities of interest from personal ontology interest user profiles. Huang et al. [12] proposed an overlapping community detection method for heterogeneous user model social network; the method mainly considers the conductance in unweighted and undirected networks.

Most of these works have attempted to acquire users' interests from similar users in common communities to match user needs, neglecting the roles of interest distribution in different communities. In the present study, we detected multi-granularity interest overlapping communities and studied interest prediction by analyzing the membership relationship of users and subjects in each community.

2.2. Semantic similarity calculation

Research on semantic similarity has been concerned for decades, and the existing models can be divided into two types. The first comprises text-based metric models [13], which verify the similarity between users by the recurrence of users' interests. The second type involves structure-based metric models, which estimate the similarity between concepts in ontology. Text-based metrics can use contextual features to compute semantic similarity with the assumption that interest subjects shared with similar concepts have similar meanings [13]. Many research studies have been conducted extensively over past years, producing models such as Kullback–Leibler divergence, information radius, Manhattan norm [37], and the bag-of-words model [22]. Works have mainly focused on information content dimensionality reduction. Latent semantic analysis [25] has been applied for a large number of applications with various forms. Another popular non-hierarchical approach is the page-count-based metric, which uses the word association co-occurrence frequency in documents to compute the similarity of concepts, such as Jaccard and Dice coefficients [26], mutual information between words [11] and Google-based distance [8]. With the taxonomic relations of ontological concepts, hierarchical relations can be used to compute semantic similarity. The main idea of the method is that the shorter the distance between two concepts, the more similar the concepts [39]. Many works have adopted this hypothesis to compute the similarity or distance between two user profiles. In exploring research the semantic relationship among relevance words, ontology categories have been utilized to identify the difference of two users and verify the semantic relationship of interests [44]. Although proposed methods based on the vertical distance of category concept trees reinforce semantic links in terms of inheritance relationships, they do not handle more complex interest structure relations from the perspective of different granular subjects. Therefore, a feasible approach must be implemented for discovering the semantic structure relationships of hierarchical concepts from the perspective of multi-granularity.

2.3. Recommendation approaches

Many recommender systems utilize user profiles and similar users for recommending products. Content-based approaches model a user profile by reusing the user's previous products [2,40], whereas CF methods adopt the general tastes of similar user profiles to generate interest predictions for the target user [41]. By considering recent information from users' activities, Bok et al. [6] utilized the interests of similar users in a group to produce a collaborative recommendation. Kardan et al. [15] identified a similar neighborhood of a user with a hybrid recommendation method. By adopting the well-established pleasantnessCarousalCdominance paradigm, Bernab-Moreno [5] investigated the emotional profiling of locations from daily communicational activity to reveal people's emotional states. Likewise, the fuzzy linguistic approach can be used to model recommender systems [35,36,46]. On the basis of a fuzzy linguistic approach, Serrano-Guerrero et al. [36] proposed a recommender system that adopts the concept of competence to suggest personalized activities for each student, and designed a fuzzy linguistic label recommender system to provide new researchers useful resources from digital libraries [35].

Certain recent studies have focused on investigating document contexts or social networks to produce recommendations. By considering contextual information in documents, Kim et al. [16] integrated a convolutional neural network model into probabilistic matrix factorization to propose a robust document context-aware hybrid method. Considering the structural information of networks, Park and Kim et al. [30] modeled truster and trustee roles and introduced an optimized top-k ranking recommendation method. Liang et al. [3] proposed a semantic reasoning mechanism over ontologies that finds semantically related items to match users' actual interests. By exploiting knowledge from related networks, Yang et al. [48] made full use of collective wisdom from social user profiles and proposed an adaptive metric learning framework that maintains the integration of network topology and community structure. As shown by the above studies, most recommendations suppose that populations' interests are fully mixed and uniform; that is, users have the same chance to accept the recommendation result. By contrast, users may enjoy varied granular products due to different granular interests. This study thus investigates multi-granular interest overlapping communities to generate recommendations and satisfy users' diverse personal needs.

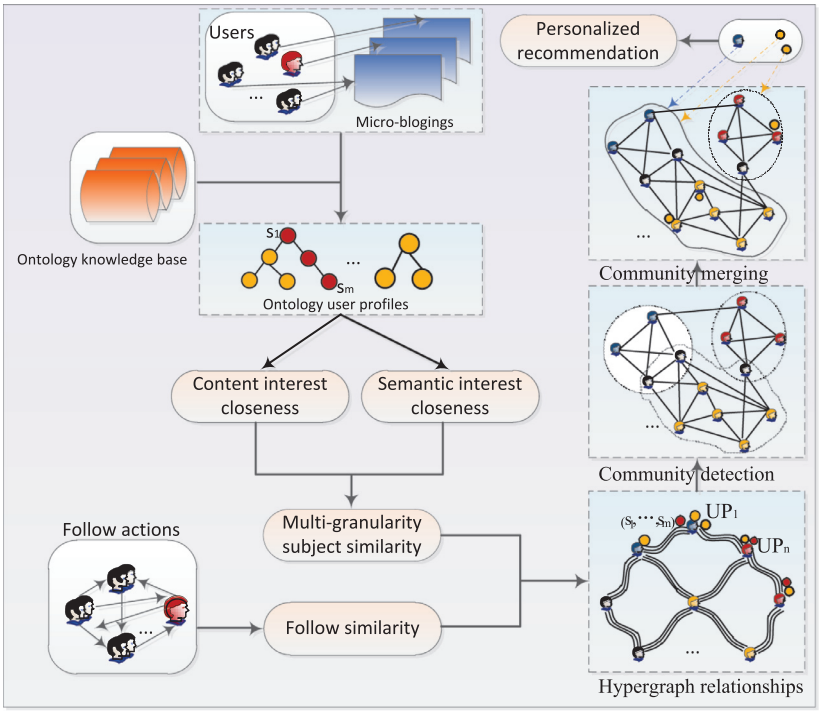

Fig. 1. Recommendation framework based on HIOC.

3. Recommendation system framework

We present a novel HIOC recommendation framework for terminal users. The framework uses the multi-granularity subject similarity of ontology user profiles, which considers content interest closeness and semantic interest closeness. Fig. 1 shows the detailed steps of the HIOC recommendation framework.

In the framework, we first create an ontology user profile for the HIOC recommendation system. From Sina Weibo websites, micro-blogs are crawled every day to extract important noun entities and compute the term frequencyCinverse document frequency (TF-IDF) weights. According to the ontology knowledge base, we model a user profile with the hierarchical interest subjects and their corresponding weights in terms of the TF-IDF mechanism. The hierarchical ontology user profile is modeled to reflect the affinity cap ability among subject categories.

Second, according to the interest weights and hierarchical interest tree structure, the content interest closeness and semantic interest closeness between different user profiles are computed. By adjustment of the relative importance of two factors, the weighted similarity at different levels of the structure can describe the multi-granularity subject similarity, which depicts the users' diverse similar relationships from different perspectives. Simultaneously, follow similarity between users is computed in terms of users' follow actions.

Then, with consideration for the follow similarity and multi-granularity subject similarity, a hypergraph based on the types of similar relationships is constructed to represent the multiple relations between user profiles. With the hyperlinks in the hypergraph, the HIOCs are detected by finding the link density peaks.

Lastly, according to the overlap proportion, the communities are merged to compute the memberships of a user and a subject in each community. On the basis of membership degrees in communities, the interest of the subject for the target user is predicted, and the top-k subjects are then pushed to the target user.

4. Multi-granularity similarity method

The ontology user profile can represent a user's omnifarious preferences. We implement a multi-granularity similarity model to identify interest communities for personalized recommendations. The multi-granularity similarity model involves two aspects: content-based interest closeness and semantics-based interest closeness.

4.1. Ontology user profile

In the micro-blog scenario, messages posted by a user usually contain different subjects, and we use subject features (that is, the term weighting scheme) to analyze the user's interests. In our work, stopword removal is processed to address

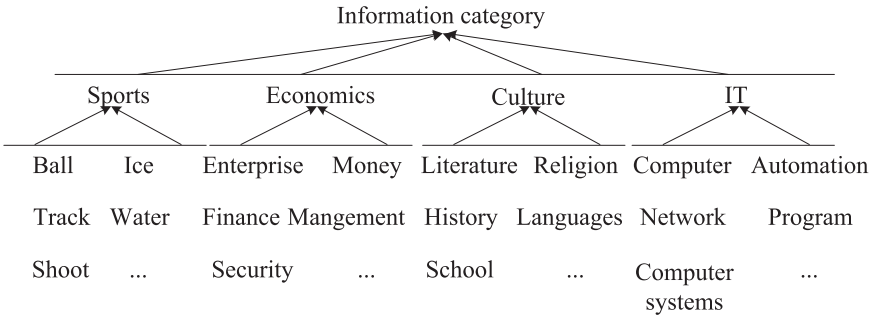

Fig. 2. Portion of the information categories on four topics.

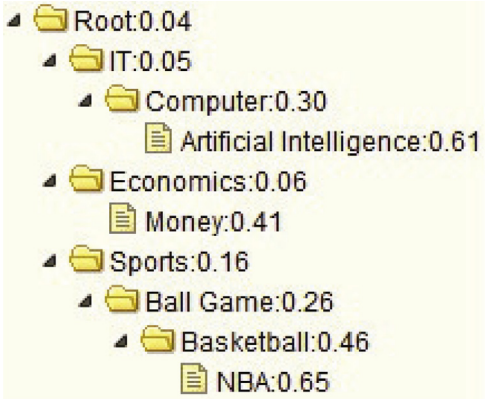

Fig. 3. Example of personalized user profile.

all the messages. By extracting subjects, each message is represented as $m = \{(t_1, W_{1m}), (t_2, W_{2m}), \ldots, (t_p, W_{pm})\}$, formed by a vector of attribute-value pairs.

Here, W_{tm} is the relative importance of term t in m. W_{tm} is computed by the TF-IDF term weighting scheme and defined as follows:

$$W_{tm} = \frac{freq_{tm}}{\max_l(freq_{lm})} \times \log \frac{N_m}{n_t}, \tag{1}$$

where $freq_{tm}$ is the raw term frequency of t in micro-blog m and $\max_l(freq_{lm})$ is the frequency of term l that has the maximum frequency in m. N_m is the total number of micro-blogs, and n_t is the number of micro-blogs that contain term t. The weight can describe the importance of a term in representing the message.

Furthermore, for a user, the micro-blogs posted by u can be defined as a set M_u. Considering all the micro-blogs involved in subject $s (s \in S)$, we can calculate the content interest degree of subject s for user u as Eq. (2).

$$Cid_u(s) = \frac{\sum_{m \in M_u} W_{sm} \times \eta(s, m)}{\sum_{s_i \in S} \sum_{m \in M_u} W_{s_i m} \times \eta(s_i, m)}, \tag{2}$$

where S represents the subject set over the ontology knowledge base category 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, $\eta(s_i, m) = 0$. The category C is involved in predefined topics, such as Sports, Economics, Culture and IT, which are from the categorization corpus of Sogou Lab Data. Fig. 2 shows part of the category structure related to the four topics.

On the basis of this analysis, we can model the interest tree of a user as a structured ontology user profile by content subjects and their interest degrees. Fig. 3 shows an example of a personalized user profile.

4.2. Closeness of ontology user profiles

In this part, a metric regarding interest degree is utilized to calculate content interest closeness, and a structured metric regarding the interest tree is used to compute the semantic interest closeness.

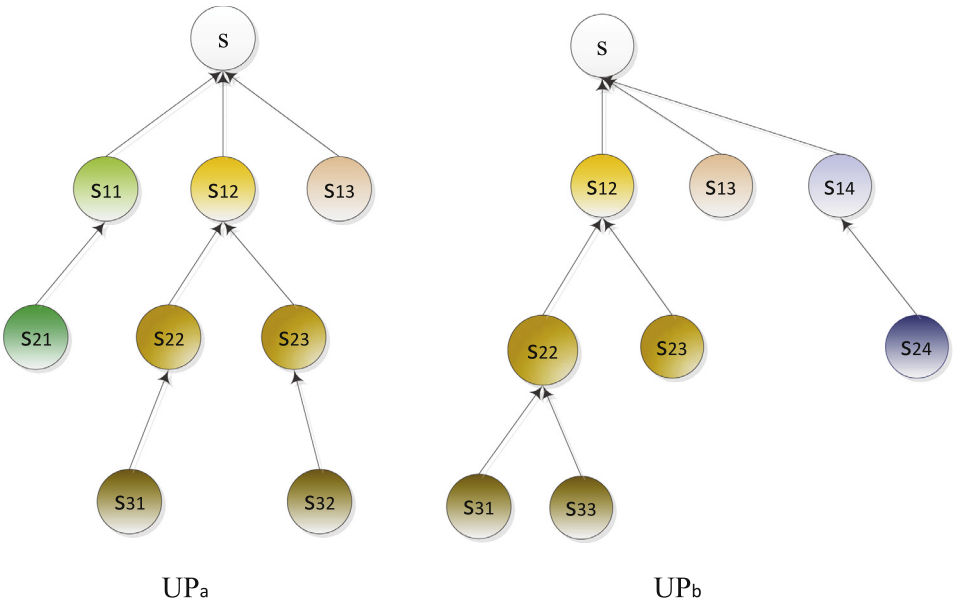

Fig. 4. Interest tree structure comparison for user profiles UP_a and UP_b.

4.2.1. Content interest closeness

If two users prefer a subject s to the same degree, then the similarity between the two user profiles should be large. We can model the content interest closeness of u_i and u_j as Eq. (3).

$$Csim_s(u_i, u_j) = 1 - \frac{|Cid_s(u_i) - Cid_s(u_j)|}{\max\{Cid_s(u_i), Cid_s(u_j)\}} \tag{3}$$

Eq. (3) depicts the content interest closeness from the aspect of the amount of interest. The larger and the closer the values of two users' content interest degrees, the bigger the content interest closeness.

4.2.2. Semantic interest closeness

According to the user profile interest tree, we can investigate the difference in structure between two user profiles to simulate the semantic interest closeness. Generally, a user considers the ancestors of the current subject node that he/she is interested in. For example, a user should be interested in "basketball" if he/she is immersed in news about the "NBA." Therefore, when we evaluate the structural similarity of the interest tree, we consider not only the current subject node but also its ancestors. Therefore, for a user, a change in the interest degree of a subject will influence a batch of nodes that comprises the parents or the ancestors of the current node. As expected, compared with the leaf node, the root node can represent the coarse interest field with poor specificity. The closer an ancestor of the current node is to the root, the less the influenced semantics. Based on the recursive spirit, the subjects in each level of the interest tree will be influenced by the leaf node. Specifically, there are three types of semantic scene similarity due to the node position in the interest tree, and these types are listed below. Fig. 4 shows an example of an interest tree structure comparison for user profiles UP_a and UP_b. In the figure, the node subjects s_{13} and s_{31} are the leaf node interests of UP_a and UP_b. Node s_{23} is a leaf node interest for UP_b but not for UP_a. In addition, the node subjects s, s_{12}, s_{22} are non-leaf node interests for UP_a and UP_b.

- Leaf node subject

To calculate the structural similarity of a leaf node subject, we adopt symbolic data and then flag the state of interest. For a leaf node subject lying at the bottom of two user profiles, if the subject is of interest to two users, then the structural similarity is 1. Otherwise, the structural similarity is 0. The specific formula is shown in Eq. (4).

$$Ssim_s^1(u_i, u_j) = \begin{cases} 1 & if \quad Cid_s(u_i) > 0 \quad and \quad Cid_s(u_j) > 0 \\ 0 & else \end{cases}, \tag{4}$$

where the subject $s \in LS(u_i)$, $LS(u_j)$. $LS(u_i)$ and $LS(u_j)$ are the leaf node sets of users u_i and u_j. For example, in Fig. 4, the structural similarity of subject s_{13} for UP_a and UP_b is 1.

- Leaf node and non-leaf node subjects

Generally, the structures of two user profiles are different. A subject is frequently a leaf node of the first user profile and a non-leaf node of the second user profile. When the users are both interested in it, we define the structural similarity

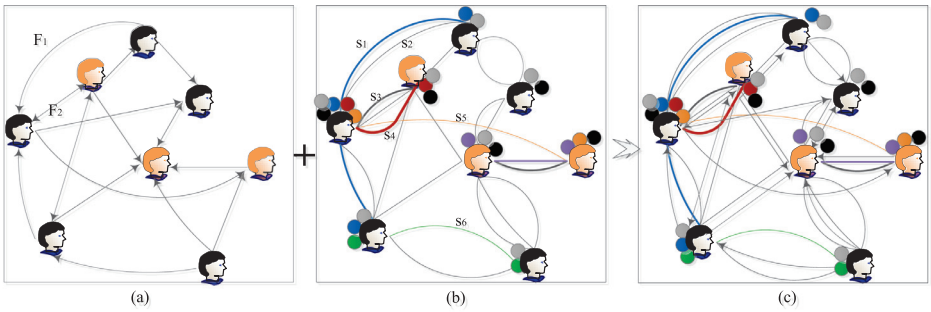

Fig. 5. Hypergraph based on follow similarity and subject similarity.

of the subject to be 0.5. Otherwise, the structural similarity of the subject is 0. The granularity level of the subject in the user profile tree represents the cognitive understanding ability of the user for the subject domain. For the same subject, the granularity of a leaf node can reflect that the user is a domain expert in the field, whereas that of a non-leaf node states that the user may know only certain things about the field. Hence, the different locations of the subject in different user profiles can discriminate the similarity of users, which can be formalized as Eq. (5).

$$Ssim_s^2(u_i, u_j) = \begin{cases} 0.5 & if \quad Cid_s(u_i) > 0 \quad and \quad Cid_s(u_j) > 0 \\ 0 & else \end{cases}, \tag{5}$$

where the subject $s \in LS(u_i)$, $NLS(u_j)$ or $s \in NLS(u_i)$, $LS(u_j)$. $LS(u_i)$ and $NLS(u_j)$ are the leaf node set of user u_i and non-leaf node set of u_j. For example, in Fig. 4, the structural similarity of subject s_{23} for UP_a and UP_b is 0.5.

- Non-leaf node subject

As interpreted above, when two users are interested in the same non-leaf node subject, the similarity will be affected by their child nodes. With a recursive characteristic, if $s \in NLS(u_i)$, $NLS(u_j)$, then we can define the structural similarity between the non-leaf nodes as in Eq. (6).

$$Ssim_s^3(u_i, u_j) = \frac{\sum_{t \in T_i^s \cap T_j^s} Snsim_t^3(u_i, u_j)}{|T_i^s \cup T_j^s|}, \tag{6}$$

where T_i^s, T_j^s is the set of child nodes of s in their corresponding user profile trees for users u_i, u_j, and t is a subject element in the set of $T_i^s \cap T_j^s$. For example, in Fig. 4, the structural similarity of subject s for UP_a and UP_b is 0.375.

4.2.3. Multi-granularity subject similarity

Finally, for a subject, given the similarity of three types of positions, the closeness between user profiles is measured by the weighted sum of semantic interest closeness and content interest closeness. As in Eq. (7), the specific similarity between u_i and u_j is formulated as follows:

$$sim_s(u_i, u_j) = \alpha Ssim_s(u_i, u_j) + (1 - \alpha) Csim_s(u_i, u_j), \tag{7}$$

where $Ssim_s(u_i, u_j)$ is one of the three types of structural similarity situations and the coefficient α reflects the relevant importance of determinants for the aspects of interest structure and interest degree.

From the perspective of the hierarchical structure of the ontology user profile, subjects in different levels represent various interest granules. In each granular level, the similarity between users can be calculated to describe the special closeness relation. In other words, for each subject, a similarity value can show subtle differences for two users at a certain semantic level. The similarities from all subjects in different granular levels can reflect the comprehensive similar relationship between users. Fig. 5(b) shows an example of multi-granularity subject similarity among users. In the figure, for each pair of users, edges that express their similarity on different subjects may be numerous.

5. Hierarchical interest overlapping community detection

In this section, we model a heterogeneous hypergraph in terms of various similar values between users and detect the interest overlapping communities on the basis of the density peaks of the hypergraph.

5.1. Heterogeneous hypergraph construction

For the personalized ontology user profile, the hierarchical subjects can depict the structural interests of a user. First, we compute the similarity between users in each subject. According to the ontology concepts, for the kth layer, we can

infer that users might be more similar in fine granular subjects than in coarse granular subjects. In other words, subjects in lower levels can reflect much more similarity than those in upper classes. Thus, we can take the semantic effect in terms of subject depth, $se_k = e^{-(layer-k)/\lambda}$, to differentiate the interest difference between users, where *layer* is the total number of ontology layers and $\lambda > 0$ is a real number that indicates the decay of the semantic effect for the aspect of structural depth. Given the semantic effect of subject s in the kth layer $se_k(s)$, we can differentiate the user-user resource similarity as follows:

$$sim_s^d(u_i, u_j) = se_k(s) \times sim_s(u_i, u_j). \tag{8}$$

The social follow action can reflect the interest tendency of users. Mutually following users can be more familiar with each other than one-way-following users. In the micro-blog scenario, let u_i and u_j be two users; their followee sets are F_{u_i} and F_{u_j}, respectively. The user-user follow similarity from u_i to u_j can be calculated by Eq. (9).

$$sim^f(u_i, u_j) = \frac{|F_{u_i} \cap F_{u_j}|}{|F_{u_i}|} \tag{9}$$

Given the differences in mutual follow actions, the similarity $sim^f(u_i, u_j)$ is different from $sim^f(u_j, u_i)$, which is a type of oriented interest cognitive mechanism in social networks.

We can combine the differentiated subject similarity $sim_s^d(u_i, u_j)$ in each layer and the oriented follow similarity $sim^f(u_i, u_j)$ between users and generate a heterogeneous hypergraph with a serious of vertexes and weighted hyperedges. Fig. 5 shows an example of a hypergraph that is based on subject similarity and follow similarity between users. In Fig. 5(a), F_1 and F_2 represent one-way and mutual followee relationships, respectively; in Fig. 5(b), the different colored edges represent multi-granular subject similar relations, such as on subjects s_1, s_2, s_3, s_4, s_5, s_6. By combining the follow similarity and subject similarity, a weighted heterogeneous hypergraph is constructed in Fig. 5(c).

The notations required for understanding the details of the hypergraph are as follows. We use $G(V, E, w)$ to denote a hypergraph, where V is the set of user vertexes, E is the set of hyperedges that represent a follow relation or subject relation, and $w : E \to R^+$ is the value of similarity. In addition, each user vertex $u \in V$ has a series of adhesive rounded subjects. For each subject s, the user's interest satisfies $Cid_u(s) > 0$.

As shown in Fig. 5(c), for a vertex $u \in V$, we can define the degree of user u as $d(u) = \sum_{e \in E} h(u, e)$. Here, $h(u, e) = 1$ if the vertex u or its adhesive subjects have an entry to the edge e; otherwise $h(u, e) = 0$. The entry to an edge simulates two types of similarity: follow similarity and subject similarity.

According to the follow similarity and subject similarity shared by two adjacent vertexes, the weighted unfamiliarity from user vertex v_i to user vertex v_j can be defined by Eq. (10).

$$d^*(v_i, v_j) = \sum_{v_i, v_j \in V, e \in E} d(v_i, v_j) h(v_i, e) h(v_j, e) \tag{10}$$

Here, the unfamiliarity $d(v_i, v_j) = 1 - w(v_i, v_j)$ is a type of mutual perception derived from the follow similarity $sim^f(u_i, u_j)$ or subject similarity $sim_s^d(u_i, u_j)$. A simple regularization is also needed for the weighted unfamiliarity $d^*(v_i, v_j)$, defined as follows:

$$\bar{d}(v_i, v_j) = \frac{e^{d^*(v_i, v_j)} - e^{-d^*(v_i, v_j)}}{e^{d^*(v_i, v_j)} + e^{-d^*(v_i, v_j)}}. \tag{11}$$

This regularization aims to address the imbalance of multi-granularity subject similarity, which can drive the value domain $\bar{d}$ into (0, 1). To control the value of $\bar{d}(v_i, v_j)$, we consider the similar relations of top 20 subjects for each user pairs.

As the follow similarity is oriented, the regularization unfamiliarity $\bar{d}(v_i, v_j)$ is an oriented perception difference that can differentiate two users' reciprocal attitudes asymmetrically. Considering all the subjects, we can observe the fine and synthetic differences from the aspect of fine-grained interests.

5.2. Overlapping community detection

In Fig. 5(c), the number of links between vertexes is not homogeneous and reflects the different interest link densities of users. The density of each vertex varies, showing different levels of activity in the social networks. A density peak clustering method was introduced for classifying objects into the local maximum of the density fields. The method has a base assumption that the cluster cores are found by a relatively high local link density and are at a relatively large distance from any point with higher local density. In the current research, we use the density peak idea in [33] to find the core of the hypergraph and then perform overlapping community detection.

Definition 1. (Local Density) Let $G(V, E, w)$ be a hypergraph. Let $V = \{v_1, v_2, \ldots, v_n\}$ be a set of vertex objects. $\forall v_i \in V$, the adjacent vertex set is $A(v_i)$. $A^+(v_i) = A(v_i) \cup \{v_i\}$. The local density of an object v_i is calculated according to Eqs. (12) and (13):

$$\rho(v_i) = \sum_{v_j \in A^+(v_i)} \chi(\bar{d}(v_i, v_j) - \varepsilon), \tag{12}$$

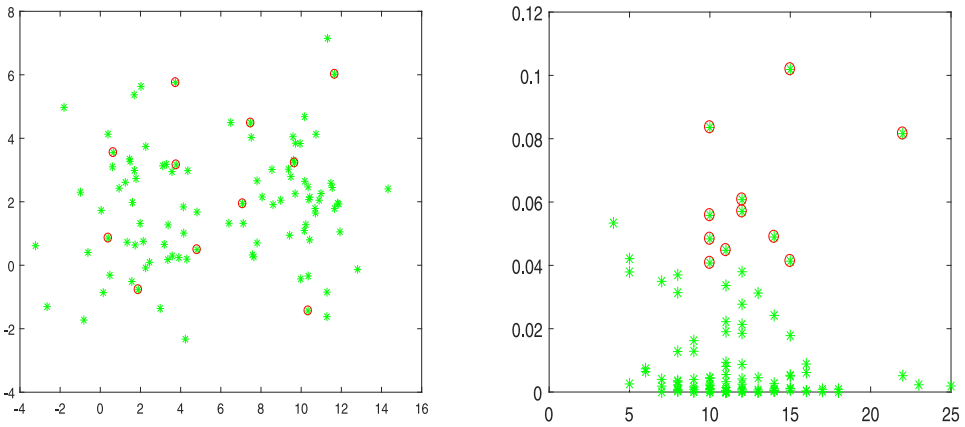

Fig. 6. Vertexes in a hypergraph and their density-decision graph. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\chi(x) = \begin{cases} 1 & if x < 0 \\ 0 & otherwise \end{cases}, \tag{13}$$

where ε is a predefined cutoff that controls the scale of the density, and the local density $\rho(v_i)$ depicts the number of objects closer than the cutoff ε to v_i.

Definition 2 (Density Unfamiliarity)**.** Let $G(V, E, w)$ be a hypergraph. Let $V = \{v_1, v_2, \ldots, v_n\}$ be a set of vertex objects. $\forall v_i \in V$, the adjacent vertex set is $A(v_i)$. $A^+(v_i) = A(v_i) \cup \{v_i\}$. The density unfamiliarity $\delta(v_i)$ of object v_i can be defined according to Eq. (14).

$$\delta(v_i) = \begin{cases} \max_{v_j \in A^+(v_i)} \bar{d}(v_i, v_j) & if \quad \rho(v_i) = \max_{v_k \in A^+(v_i)}(\rho(v_k)) \\ \min_{\forall v_j \in A^+(v_i): \rho(v_j) > \rho(v_i)} \bar{d}(v_i, v_j) & else \end{cases} \tag{14}$$

The value of $\delta(v_i)$ depicts the minimum unfamiliarity between the object v_i and any other objects with higher density.

Definition 3 (Graph Average Density)**.** Let $G(V, E, w)$ be a hypergraph. Let $V = \{v_1, v_2, \ldots, v_n\}$ be a set of vertex objects. $\forall v_i \in V$, the local density of v_i is $\rho(v_i)$. The graph average density $\rho(G)$ can be defined according to Eq. (15).

$$\rho(G) = \frac{\sum_{v_i \in V} \rho(v_i)}{|V|} \tag{15}$$

The value of $\rho(G)$ depicts the average interest link strength in the entire hypergraph.

In the hypergraph $G(V, E, w)$, according to the $\rho(v_i)$ and $\delta(v_i)$ of each vertex $v_i \in V$, we can obtain the decision graph generated by setting $\rho(v_i)$ as the x-axis and $\delta(v_i)$ as the y-axis. Fig. 6 shows an example of vertexes in a hypergraph and their density decision graph.

The decision graph shows some noticeable objects on the right with higher local density and density unfamiliarity than the rest and can thus be viewed as diverse density peak nodes. In terms of the local density of vertexes, $\rho(\cdot)$ and $\delta(\cdot)$ can help in the selection of several local density peak nodes. Furthermore, we can first initialize these density peak nodes as the initial community cores and then divide the hypergraph into several robust dense subgraphs. Algorithm 1 generates the initial community cores.

By considering the count of interest links and the strength of the interest links in the entire hypergraph, Algorithm 1 aims to mine the initialized cores of a hypergraph from the aspects of local density and density unfamiliarity. The initialized representative and diverse cores are conducive to improving the quality and speed of the community detection. In Fig. 6, the red objects are selected for comparison against the green objects as initialized cores.

According to the selected cores, we can find nodes with similar densities from the adjacent vertexes and control the scale of the community by fitness [19]. In our work, the communities are partitioned subgraphs comprising a subset of the vertexes in the hypergraph. The fitness of the community subgraph measures the contribution of the internal edges of nodes in the subgraph and the external edges with other nodes, which is calculated by Eq. (16) [19].

$$f(G) = \frac{deg_{in}^G}{(deg_{in}^G + deg_{out}^G)^\beta}, \tag{16}$$

where deg_{in}^G and deg_{out}^G are the total internal and external degrees, respectively, of the nodes of graph G, and β is a positive real-valued parameter that controls the size of the detected communities.

Algorithm 1 Density-based community core selecting.

Input:
Hypergraph $G(V, E, w)$, Initialized core set $C = \varnothing$.
Output:
Community core set C.

while $|V| > 0$ **do**
  for each $v_i \in V$, compute $\rho(v_i)$ and rank vertexes in a descending order $\rho(v_i)$ **do**
    Select node(s) v into temporary core set C^T from V with max $\rho(v)$;
    if $|C^T| > 1$ **then**
      Select one node v_i from C^T, which satisfy $\delta(v_i) \geq \max_{v_j \in C^T}\{\delta(v_j)\}$;
      C=$C \cup \{v_i\}$;
      $V = V/A^+(v_i)$;
    else
      C=$C \cup \{v\}$;
      $V = V/A^+(v)$;
    end if
  end for
end while
Return C.

By introducing the local density of a vertex and its similar interest link fitness contribution, we measure the importance of a vertex node. Therefore, for a graph G, we can measure the interest density-fitness contribution of G by Eq. (17).

$$f^{\rho}(G) = \frac{\rho_{in}^{G}}{(\rho_{in}^{G} + \rho_{out}^{G})^{\beta}} \tag{17}$$

Here, $\rho_{in}^{G} = \frac{\sum_{v_i \in G} \rho_{in}(v_i)}{|G|}$ is the average density link strength for the graph G, and $\rho_{in}(v_i) = \sum_{v_j \in G} \chi(\bar{d}(v_i, v_j) - \varepsilon)$ is the local density of an object v_i in G; analogously, considering the distance of object v_i to the other objects in $\bar{G}$, $\rho_{out}(v_i) = \sum_{v_j \in \bar{G}} \chi(\bar{d}(v_i, v_j) - \varepsilon)$ and $\rho_{out}^{G} = \frac{\sum_{v_i \in G} \rho_{out}(v_i)}{|G|}$.

Furthermore, the interest density-fitness contribution of a vertex p with respect to G can be computed by $f_G^{\rho}(p) = f_{G+\{p\}}^{\rho} - f_{G-\{p\}}^{\rho}$. Here, $f_{G+\{p\}}^{\rho}$ and $f_{G-\{p\}}^{\rho}$ indicate the density-fitness of graph G with node p inside and outside, respectively.

According to the interest density-fitness, the detailed steps of our approach for interest community detection of a graph are presented in Algorithm 2. We set $\beta = 1$ to have iterative operations for use in finding the scale of communities.

Algorithm 2 Community detection.

Input:
Core C.
Output:
Community $G(c)$.

A loop is performed over all adjacent vertexes $A(C)$;
Add the adjacent vertex vertex v_k from $A(C)$ into the core C, where $\rho(v_k) = \max_{v_j \in A(C)}\{\rho(v_j)\}$, generating a subgraph $G(C)$;
Calculate the density -fitness of each vertex of $G(C)$;
if $\exists p \in G(C)$, satisfy $f_{G(C)}^{\rho}(p) < 0$ **then**
  Delete p, yielding a new subgraph $G'(C)$;
end if
if Step 4 occurs **then**
  Repeat from Step 3;
else
  Repeat from Step 1 for subgraph $G'(C)$;
end if

The iterative process stops when the vertexes examined in Step 1 all have negative density-fitness values. Step 2 ensures selecting the tightly connected vertexes into C, which can adjust the interest density coordination of the community. Then, we output community $G(C)$ of the core C. If the detected communities fail to cover all the vertex nodes in hypergraph G,

then we need to find new cores and detect new communities for the remaining vertexes by implementing Algorithms 1 and 2 until all vertexes in G are contained in at least one community. By detecting the cover of a vertex in the hypergraph, the natural community of each node can be discovered. Algorithm 3 shows the process of overlapping community detection.

Algorithm 3 Overlapping community detection.

Input:
Hypergraph $G(V, E, w)$, community core set C.
Output:
Overlapping communities G'.
while $V \neq \varnothing$ **do**
  Find the core set C of V by algorithm 1;
  for each $c \in C$ **do**
    Detect the community $G(c)$ of core c by algorithm 2;
    $S = S \cup G(c)$;
    if $\exists p \in S$ and $p \in C$ **then**
      Delete p from C
    end if
  end for
  $V = V/S$;
end while
$G' = S$.

The algorithm stops when all vertexes have been assigned to at least one community. According to the above idea, for the community core set $C = \{c_1, c_2, \ldots, c_k\}$, we can obtain the detected overlapping communities as $G' = \{G_1, G_2, \ldots, G_k\}$. By the coverage search, the nodes of every community are either overlapping with other communities or not. That is, node v may belong to multiple communities, and two communities may have many overlapping vertexes. Therefore, a merging mechanism is needed for the combination of excessively intersected communities into a single community

5.3. Community merging

The overlap proportion is used to judge whether two communities can be combined or not, which has been adopted in many research studies [9]. We merge two small communities into a large community when they both have a high proportion of overlapping nodes. On the basis of this idea, the overlap proportion of communities G_i and G_j is measured as follows:

$$\gamma_{ij} = \frac{|G_i \cap G_j|}{\min(|G_i|, |G_j|)}. \tag{18}$$

Specially, the predefined threshold $\eta \in [0, 1]$ is used to determine the gate of combination of two communities. If $\gamma_{ij} > \eta$, then we merge the two communities as a new G. Algorithm 4 gives the procedure of overlapping community merging.

Algorithm 4 Overlapping community merging.

Input:
Communities $G' = \{G_1, G_2, \ldots, G_k\}$.
Output:
Community merging result G''
while $\exists G_i, G_j \in G'$ **do**
  if $\gamma_{ij} > \eta$ **then**
    $G_{new} = G_i \cup G_j$;
    $G' = G'/\{G_i, G_j\}$;
    $G' = G' \cup G_{new}$;
  end if
end while
Return merging results G' as G''

The algorithm starts merging from the two communities with the minimum number of members. By several iterations, it stops when neither community has a quality of overlapping nodes. After merging communities, users can still belong to more than one community. The merged communities include users with common/similar interests, thereby potentially aiding in the generation of specific personalized recommendations.

Table 1
Labeled network datasets for community detection.

Datasets	Vertexes	Edges	Communities
Polbooks	105	441	3
Polblogs	1490	16,718	2
Football	115	613	12
email-Eu-core	1005	25,571	42

5.4. Personalized recommendation

We consider the membership of a user in communities and the membership of a subject distributed in communities to produce an interest prediction.

After merging communities, if there are $n(n \geq 1)$ communities such as $\{G_1, \ldots, G_i, \ldots, G_n\}$, we can use a membership vector $p_u = (p_u^1, p_u^2, \ldots, p_u^n)$ to state the possibility of user u belonging to each community. The element of the vector can be viewed as the roles of all adjacent nodes of u playing in each community G_i. Given a user u, one's membership degree in the community G_i can be defined as Eq. (19).

$$p_u^i = \frac{\sum_{v \in A^+(u), v \in G_i} \bar{d}(u, v)}{\sum_{h \in G_i, g \in G_i} \bar{d}(h, g)} + \frac{|\{z | z \in A^+(u), z \in G_i\}|}{|G_i|} \tag{19}$$

The membership degree emphasizes the adhesion from the aspect of density proportion of interest linkage and quantity proportion of linkage users in community G_i. The membership vector can leverage the linkage roles of users in community G_i and other overlapping communities. In addition, for a new subject s, we can use the maximal interest of subject s in G_i to represent how a user may be interested in the subject with the maximum opportunity. Then, considering all the interests distributed in the associated communities in which subject s occurs, we can model the interest vector of subject s as $p_s = (p_s^1, p_s^2, \ldots, p_s^n)$, where $p_s^i = \max\{Cid_v^i(s) | v \in G_i\}$ represents the maximal interest degree of subject s in community G_i.

Considering that the membership vector of user u belongs to every community and the membership vector of interest subject s is distributed in all the communities, we can compute the interest degree of subject s for user u as follows:

$$Cid_u{}^G(s) = \overrightarrow{p_u} \cdot \overrightarrow{p_s}. \tag{20}$$

Naturally, on the basis of the overlapping communities that user u belongs to, we first aggregate their subjects of interest and compute their corresponding interest degree. Then, by ranking the interest degree of each subject, we select the top-k subjects as the recommendation list and allocate relevant micro-blogs to the target user.

6. Experiment evaluations

In this section, we conduct several experiments on real-world networks to test the performance of our proposed community detection algorithm and provide an experimental evaluation to show the effectiveness of our proposed recommendation mechanism.

6.1. Detection experimental results

6.1.1. Experiment design

In our study, several methods were selected for comparison with our approach in terms of community detection. These methods included the label propagation algorithm (LPA) [31]; clique percolation method CPM [33]; and LFM, which was proposed by Lancichinetti [19]. The experiments, over ten repetitions on the four labeled networks, were used to verify the average performance of community detection. The networks were downloaded from [17,21] and are listed in Table 1. In addition, we adopted two real Weibo networks to test the detection experiments. In the Sina Weibo platform, we crawled 13,722 posted micro-blogs and 5017 followee actions for 514 users from April 10, 2013 to April 29, 2013 to model the network named *SW* dataset, which was acquired from the *NLPIR* website (http://www.nlpir.org/). In the Tencent Weibo platform, 1296 users were used for collecting their reposted 76,176 micro-blogs, and 6809 follower-followee relationships among users were crawled in June 2015, which could help simulate all users' interest network and perform community detection; this was named the *TW* dataset. We divided the two datasets into two parts according to the timestamp. The earlier period was used for modeling the interest subjects, and the remaining part was adopted for testing. Details for the two micro-blog networks are shown in Table 2. We conducted our experiments on a computer with a 3.6 GHz Intel i7-4790MQ CPU and 16 GB RAM. The software platform was MathWorks MATLAB 2016b, which was run on 64-bit Windows 7.

6.1.2. Experiment metrics

We employ three metrics to evaluate the quality of the community detection approach.

Table 2
Micro-blog network datasets for community detection.

Datasets	Users	Followees' actions	Training micro-blogs	Testing micro-blogs
SW	514	5017	7424	6298
TW	1296	6809	36,794	39,382

Table 3
Experimental results on four labeled networks for six algorithms.

Datasets	Index	LPA	CPM	LFM	Long	Huang	HIOC
Polbooks	ARI	0.5888	0.0182	0.7083	0.0849	0.6745	**0.7867**
	NMI	0.4872	0.2725	0.5969	0.3558	0.5745	**0.6884**
Polblogs	ARI	0.4246		0.3924	**0.5992**	0.4356	0.4425
	NMI	**0.3276**		0.2579	NaN	0.2827	0.3147
Football	ARI	0.3943	0.6612	0.6573	0.6006	**0.7308**	**0.7308**
	NMI	0.7187	0.8061	0.8343	0.8307	**0.8421**	**0.8421**
Email-Eu-core	ARI	0.8983		0.8125	0.1373	0.8971	**0.8991**
	NMI	**0.9407**		0.8273	0.8332	0.9049	0.8635

(Normalized mutual information) Given a set V of N nodes and two partitions G_A, G_B, construct a confusion matrix **N**, where the rows correspond to the "real" communities G_A, and the columns correspond to the "detected" communities G_B. N_{ij} is the number of overlapping nodes between the real community i in G_A and the detected community j in G_B. $N_{i.}$ is the sum over row i of matrix **N**, and $N_{.j}$ is the sum over column j of matrix **N**. The normalized mutual information (*NMI*) [38] can be estimated by Eq. (21).

$$NMI = \frac{\sum_{i=1}^{G_A}\sum_{j=1}^{G_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{G_A} N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{G_B} N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \tag{21}$$

(Adjusted rand index) The adjusted rand index (ARI) [3] can be defined as follows:

$$ARI = \frac{\sum_{i=1}^{G_A}\sum_{j=1}^{G_B}\binom{N_{ij}}{2} - \left[\sum_{i=1}^{G_A}\binom{N_{i.}}{2}\sum_{j=1}^{G_B}\binom{N_{.j}}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_{i=1}^{G_A}\binom{N_{i.}}{2}+\sum_{j=1}^{G_B}\binom{N_{.j}}{2}\right] - \left[\sum_{i=1}^{G_A}\binom{N_{i.}}{2}\sum_{j=1}^{G_B}\binom{N_{.j}}{2}\right]/\binom{N}{2}}. \tag{22}$$

The larger the values of ARI and NMI, the better the detection result.

(Modularity) Newman's modularity [27,28] (Q_N) function is a widely known evaluation metric in the field of community detection, which is calculated as follows:

$$Q_N = \sum_{i=1}^{n}\left[\frac{L_i}{TL} - \left(\frac{D_i}{TL}\right)^2\right], \tag{23}$$

where n is the number of communities. L_i is the number of edges between vertexes within community i, D_i is the sum of the degrees of the vertexes in community i, and TL is the total number of edges of the network.

The more accurate the community detection result, the greater the value of modularity.

6.1.3. Experimental results

Labeled networks. We ran our HIOC algorithm in the four labeled networks for community accuracy detection and compared its performance with that of LPA, CPM, LFM, Long's method [24], and Huang's [12] method. For the HIOC algorithm, we set the parameter $\eta = 0.15$ to merge communities with high overlap proportion. In Long's method, we set the path parameter $P = 5$ to compute the edge density of networks and detect communities. In the experiment, the NMI and ARI values were used to judge the detection accuracy. Table 3 shows the experimental detection results of the six algorithms on four datasets. The results show a glaring difference between the HIOC method and the other algorithms in terms of NMI and ARI indexes. Specifically, on the Polbooks dataset, the ARI and NMI of the HIOC method are better than those of the other methods. On the Football, PolBlogs and Email-Eu-core datasets, the indexes of the HIOC method are close to the best results of the other algorithms. Similar to Long's and Huang's methods, our method can provide good performance. This finding is due to the fact that the HIOC algorithm first selects community cores according to the density of the network and then detects communities on the basis of the density-fitness contribution. In terms of density and distance for nodes in networks, we can generate a relatively appropriate number of cores and obtain stable communities with similar density. Therefore, although the HIOC method does not deliver the best performance on the ARI and NMI indexes for these datasets, the experimental results can illustrate the efficient usage of the proposed HIOC method for detecting the communities of these complicated networks.

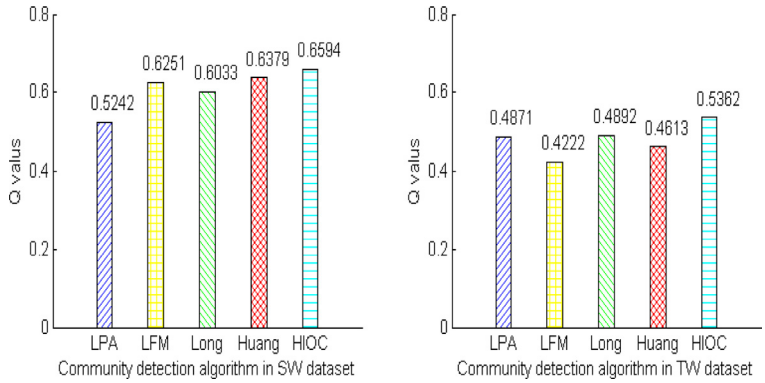

Fig. 7. Modularity values of HIOC and compared methods for two micro-blog social networks.

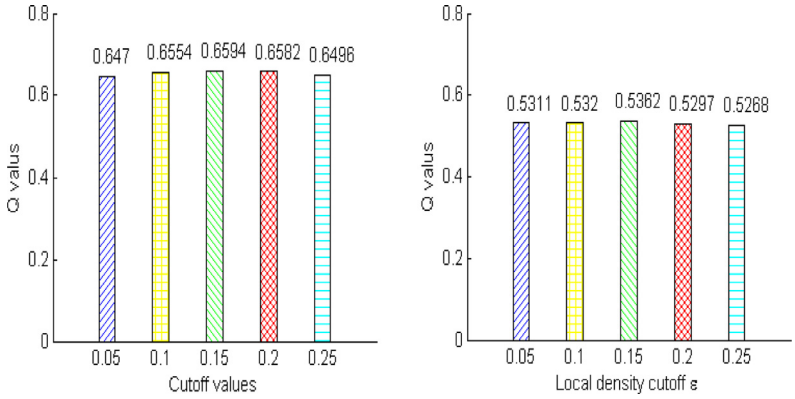

Fig. 8. Modularity values of HIOC method for two micro-blog social networks under different values of cutoff ε.

Unlabeled networks. We also applied the proposed HIOC algorithm to two real-world micro-blog social networks in Table 2 and compared the performance in terms of modularity Q_N. For capturing various types of subject similarity, we set $\lambda = 1.2$ to differentiate users from the aspect of semantic interests. For the *SW* dataset, according to the followees' actions and users' subject similarity, we selected 2996 nodes and 48,105 interest edges to construct the hypergraph. Similarly, the *TW* dataset contained 3324 nodes and 84,423 edges. Then, we set the cutoff ε=0.15 to control the scale of similar interest edges for a node and the interest density for a community. For the two unlabeled networks, we adopted the modularity Q_N values to observe the performance of HIOC method. Fig. 7 shows the experimental results of the HIOC algorithm and the LPA, LFM, Long, and Huang methods on two datasets. In the figure, we can see that the HIOC algorithm can perform better than the other algorithms, which indicates that the density peak-based core method can select a stable community structure. For example, for the *SW* dataset, the HIOC algorithm improves the modularity by 20.50%, 5.20%, 8.51%, and 3.26% compared with the LPA, LFM, Long, and Huang algorithms. In the *TW* dataset, users have rich activities or interest interactions with other users, and they would form a large number of interest communities. The Q_N for the *TW* dataset is smaller than that of the *SW* dataset.

As described in Eq. (12), the cutoff ε can affect the local density of a user, which leads to variations in community detection results. By changing the value of the cutoff, we perform community detection using the HIOC method for two datasets and observe the different results shown in Fig. 8. For the *SW* dataset, the modularity value of the HIOC method first weakly increases and then weakly decreases as the value of the cutoff ε increases. A similar trend can be found in the modularity values of the *TW* dataset. As expected, a small ε can induce a small interest density, which produces many communities. A large ε can detect a small number of communities. In the figure, the change in the modularity values verifies the variations in community detection results under different cutoff ε values. Therefore, we conclude that neither too large nor too small a cutoff can achieve the best community detection results. In the figure, we can also see that the Q_N values of the HIOC method under different cutoffs are close, which verifies the stability of the algorithm.

Table 4
Confusion matrix for recommender systems.

$S_T \backslash S_R$	Positive	Negative
Positive	TP	FN
Negative	FP	TN

6.2. Recommendation experimental results

6.2.1. Experiment design

As described in Eq. (7), the proposed multi-granularity subject similarity distinguishes users from multiple levels of interest subjects. On the basis of the hypergraph constructed by user similarities, we developed community detections and performed community-based interest prediction.

To observe the effectiveness of the proposed similarity, we compared it with the Jaccard similarity [13] and adjusted cosine similarity [42]. We verified the recommendation performance of the HIOC algorithm according to several similarity methods.

According to the communities a user belongs to, we produced the top-k subject recommendations derived from the communities and pushed relevant micro-blogs to the target user by Eq. (20). In addition, we demonstrated our model by comparing it with personal UP, user-based CF (UBCF), and LDA.

For the personal UP recommendation, we first computed the content interest degree of a subject by the TF-IDF mechanism in Eqs. (1) and (2). Then, considering all the subjects that the user is interested in, we ranked one's subjects by interest degree and pushed the top-k subjects to the target user.

For the UBCF recommendation method, we first computed the similarity between users by adjusted cosine similarity [34,42], as shown in Eq. (24).

$$sim(u_i, u_k) = \frac{\sum\limits_{s \in S_{u_i} \cap S_{u_k}} (Cid_{u_i}(s) - \overline{Cid}_u)(Cid_{u_k}(s) - \overline{Cid}_u)}{\sqrt{\sum\limits_{s \in S_{u_i} \cap S_{u_k}} (Cid_{u_i}(s) - \overline{Cid}_u)^2} \sqrt{\sum\limits_{s \in S_{u_i} \cap S_{u_k}} (Cid_{u_k}(s) - \overline{Cid}_u)^2}}, \tag{24}$$

where $S_{u_i} \cap S_{u_k}$ is the set of subjects that u_i and u_k are interested in. $Cid_u(s)$ is the content interest of subject s for user u, and $\overline{Cid}_u$ is the average content interest weight of subject s from all UPs.

Then, on the basis of the similarity in Eq. (24), we could rank similar users to select collaborative users. By considering an equal number of similar users into set G_{u_i}, we could compute the UBCF interest degree as follows:

$$Cid_{u_i}^{CF}(s) = \frac{\sum\limits_{u_k \in G_{u_i}} Cid_{u_k}(s) \times sim(u_i, u_k)}{\sum\limits_{u_k \in G_{u_i}} |sim(u_i, u_k)|}. \tag{25}$$

According to the UBCF interest degree, we selected top-k subjects for target users.

The LDA topic model is a generative probabilistic graphical model for personalized topic mining [23]. Generally, the model generation process has two assumptions. First, given a set of T topics, each document can be viewed as a multinomial distribution over T topics. Second, based on a set of vocabulary words, each topic is also a multinomial distribution related to vocabulary words. The distribution is defined as $p(w|z)$ and $p(z|d)$, where z, w, d denote the latent topic, the word, and the document, respectively. A topic $z_i = j$ can be denoted as $P(z_i = j) = Multinomial(\theta_j^{(d_i)})$. A multinomial distribution related to the set of vocabulary words can be viewed as $P(w_i|z_i = j) = Multinomial(\varphi_{w_i}^{(j)})$, which depicts the meaning of the topic. Then, the dirichlet distributions of the document and the word can be defined as $\theta_j = Dirichlet(\alpha)$, $\varphi_i = Dirichlet(\beta)$.

In experiments, Gibbs sampling is used to train the latent topic distribution, and the hyper parameters are set $\alpha = \beta = 0.01$. According to personal users' all micro-blogs, we selected their related top-k topics. We selected the maximal related subject for each topic and recommendeded it to the target user.

6.2.2. Experiment metrics

For observing the classifying and misclassifying subjects, Table 4 shows the confusion matrices of recommender systems [32]. In this confusion matrix, TP is the true positive set, which signifies correctly classified positive examples; FN is the false negative set, which signifies incorrectly classified negative examples; FP is the false positive set, which signifies incorrectly classified positive examples; and TN is the true negative set, which signifies correctly classified negative examples. Therefore, $S_R = TP + FP$ is the set of subjects that is recommended by the system, and $S_T = TP + FN$ is the set of subjects that users are actually involved in.

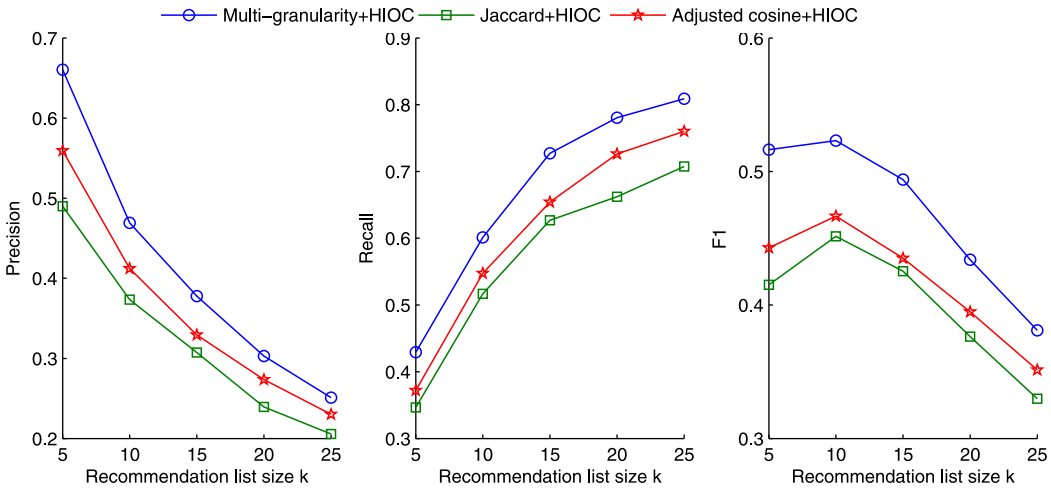

Fig. 9. Precision, recall, and F1 values based on the HIOC method under different similarity methods for the *SW* dataset.

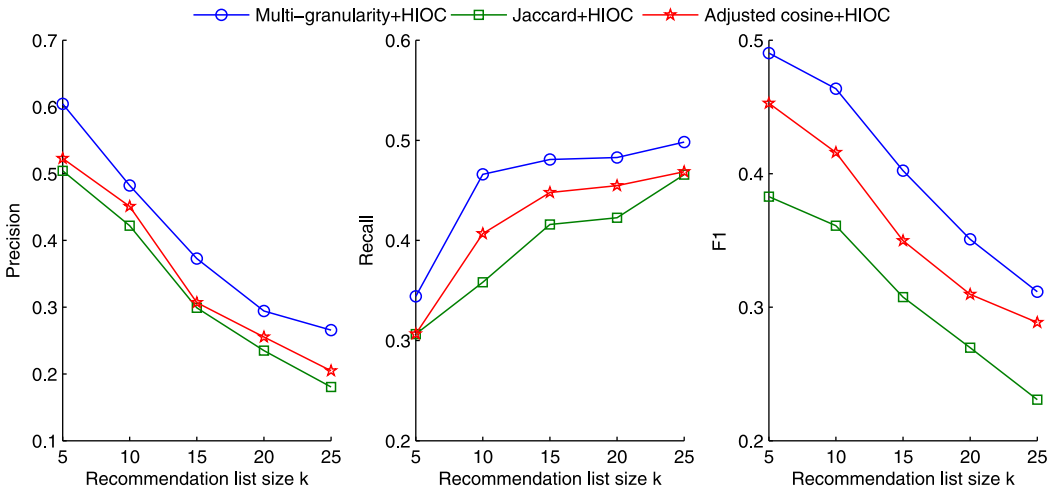

Fig. 10. Precision, recall, and F1 values based on the HIOC method under different similarity methods for the *TW* dataset.

To verify the performance of personalized recommendation, we used three popular evaluation measures: precision, recall, and F1. Precision, recall, and F1 are defined as follows:

$$Precision = \frac{|S_T \cap S_R|}{|S_R|}, \tag{26}$$

$$Recall = \frac{|S_T \cap S_R|}{|S_T|}, \tag{27}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{28}$$

6.2.3. Experiment results

To verify the effectiveness of the proposed similarity, Figs. 9 and 10 state the comparison results of the HIOC recommendation method under different similarity mechanisms. According to the figures, the HIOC recommendation based on the proposed similarity can perform better than the Jaccard and adjusted cosine similarities. Jaccard similarity considers the number of mutual interests but ignores the similarity derived from the interest degree of users. The adjusted cosine similarity attempts to find similar users from the angle difference between the whole vectors but also neglects the interest distance of each attribute dimension in the vector. The proposed similarity in Eqs. (3) and (7) considers the distance of interest degree and the semantic structures of user profiles, thereby finding more reasonable similar users than conventional

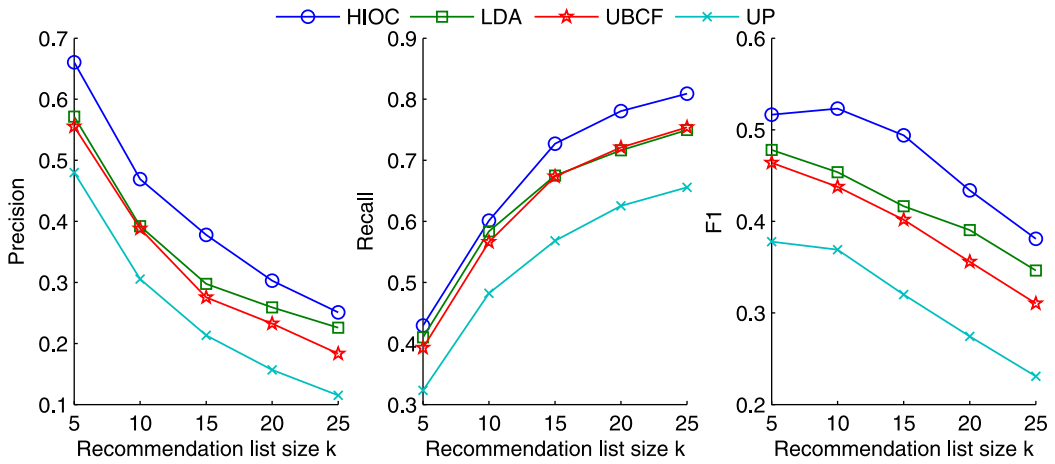

Fig. 11. Precision, recall, and F1 values based on the HIOC method and compared methods for the *SW* dataset ($\alpha = 0.5, \varepsilon = 0.15$).

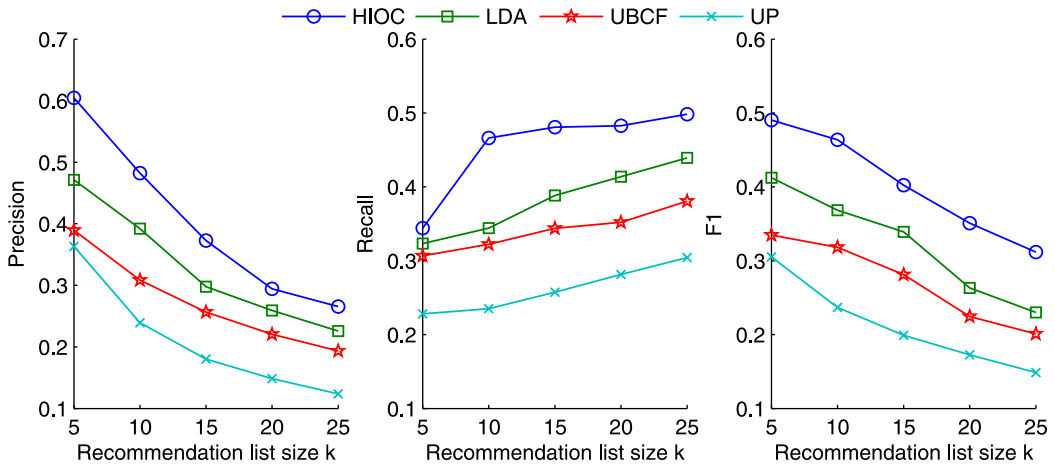

Fig. 12. Precision, recall, and F1 values based on the HIOC method and compared methods for the *TW* dataset ($\alpha = 0.9, \varepsilon = 0.15$).

methods. A large number of reasonable similar users provide rich interests for target users, which can expand users' preferences effectively. Hence, the proposed similarity method can be used to improve the diversification of results and solve the cold start problem.

Figs. 11 and 12 show the comparison results of the HIOC model and other recommendation approaches. As illustrated in the figures, the HIOC method can outperform all of the compared methods in terms of precision, recall, and F1. For the *SW* dataset, precision and recall have a faster rate of increase as the recommendation list size increases. Specifically, when k is 15, the mean precision of HIOC is 0.3778, which is higher than the 0.2977, 0.2754, and 0.2134 of LDA, UBCF, and UP, respectively; the mean recall of HIOC is 0.7272, which is higher than the 0.6751, 0.6734, and 0.5683 of LDA, UBCF, and UP, respectively. These findings are due to the division of the user into multiple communities by consideration of its fine-grained interests. Multiple communities can provide rich and accurate interests for the target user. Contrarily, the LDA and UBCF mainly take a single granularity of interest content to mine similar users. The similar users are not sufficient for the provision of comprehensive interests for personalized recommendation. Therefore, the HIOC structure can efficiently increase the recall of personalized recommendation.

On the *SW* and *TW* datasets, when the value of coefficient α varies, the multi-granularity similarity between users changes. Given a fixed value of cutoff ε, we could obtain different local density values according to varying similarity, thereby possibly generating various community detection results. Such results can assign different community friends for the target user, which affects the performance of recommendation. Figs. 13 and 14 show precision, recall, and F1 values based on HIOC method under different values of α for the two datasets. According to the figures, the precision and recall values can gradually increase with the importance of semantic interest closeness factor. In the *SW* dataset, by setting $\varepsilon = 0.15$, we could obtain the best performance at approximately $\alpha = 0.5$ in Fig. 13. For the *SW* dataset, the content interest closeness between users is large, and a certain amount of similar users can be selected in terms of cutoff ε, which can

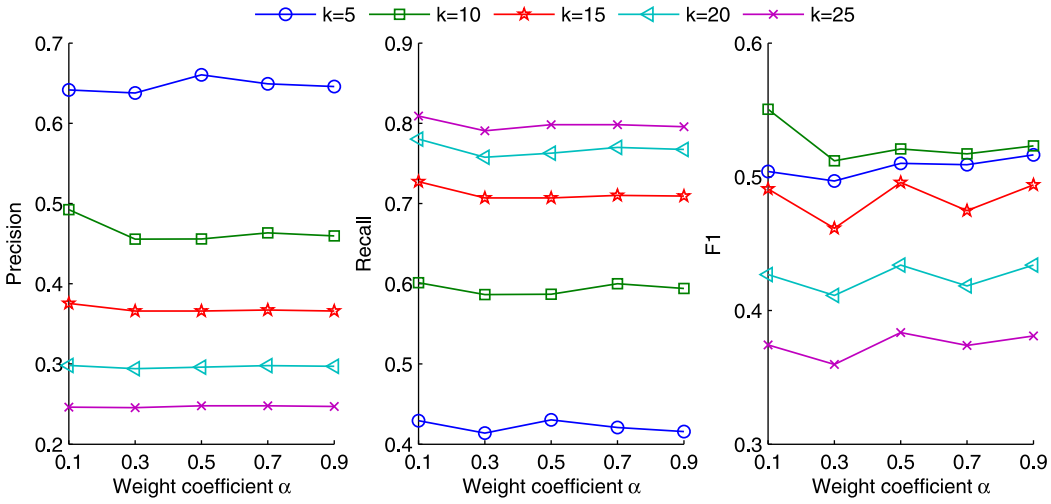

Fig. 13. Precision, recall, and F1 values based on HIOC method under different values of α for the *SW* dataset ($\varepsilon = 0.15$).

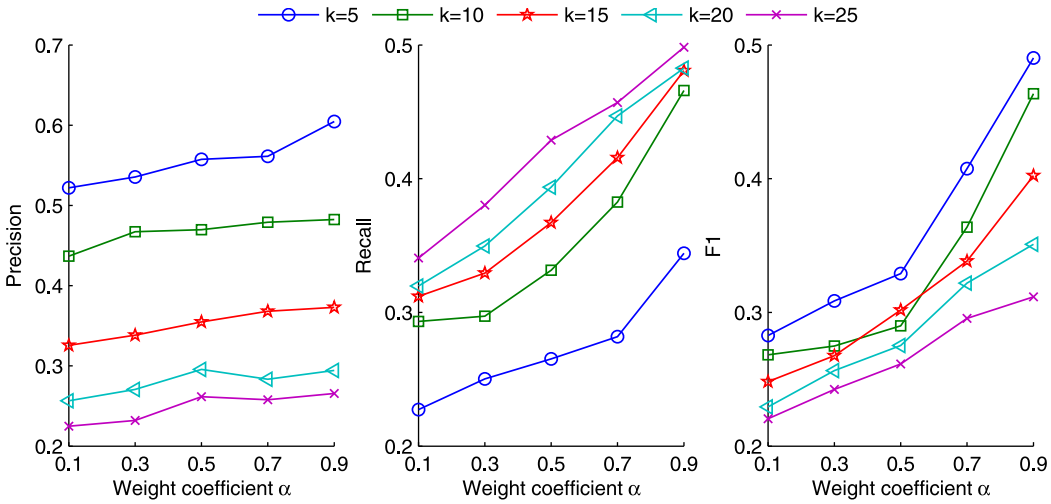

Fig. 14. Precision, recall, and F1 values based on HIOC method under different values of α for the *TW* dataset ($\varepsilon = 0.15$).

provide good recommendation service. That is, the semantic interest closeness based on user profile structures has a stable influence for the selection of similar users. In the *TW* dataset, the content interest closeness between users is small, and the structure-based semantic interest closeness can efficiently identify the implied similarity between users on certain topics, which could mine more similar relationship of users. Moreover, the community can find more potential similar users and consequently improve the quality of recommendation. In Fig. 14, with the weight of semantic interest closeness increasing, the similarity between users on a certain subject can become markedly large, and multi-granular similar relationships between users become rich, which can find fine-grained similar users with close interests. Hence, in Fig. 14, the HIOC method achieves the best performance at $\alpha = 0.9$ for the indexes of precision, recall, and F1.

Considering the local density of interest relationships among users, different values of the cutoff ε can generate various community detection results. As different communities can assign varied community friends for the target user, we utilize the different detection results to make various personalized recommendations. Figs. 15 and 16 show the precision, recall, and F1 based on the HIOC recommendation method under different values of ε for the two datasets. As depicted by the figures, for different recommendation list sizes, the precision, recall, and F1 of the HIOC method increase and then decrease with variations of the cutoff ε. This is due to the small number of communities induced by a large ε. Each community may own enough users to provide abundant interests. Superfluous subjects can lead to a decline in the rate of recall. By contrast, a small ε generates a large number of communities. Each community has relatively few users, which cannot supply rich interests for the target user.

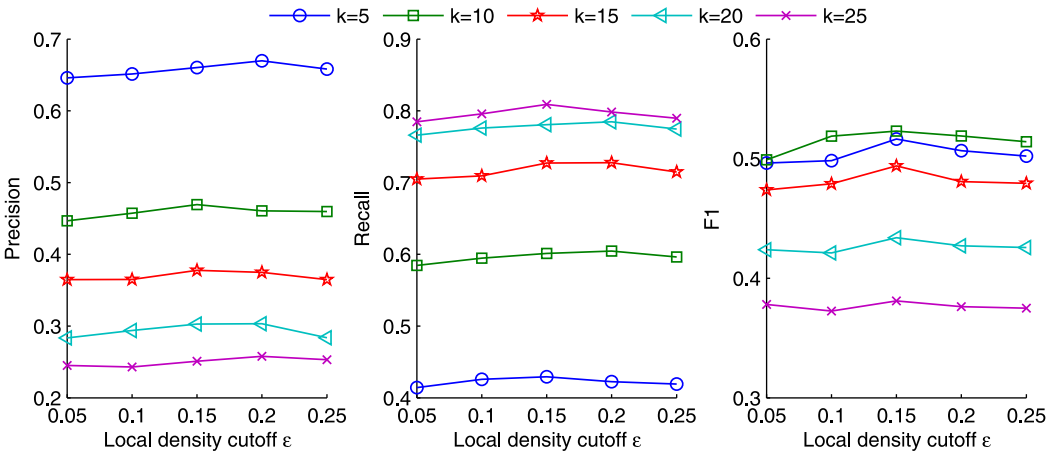

Fig. 15. Precision, recall, and F1 values based on the HIOC method under different cutoffs of ε for the *SW* dataset ($\alpha = 0.5$).

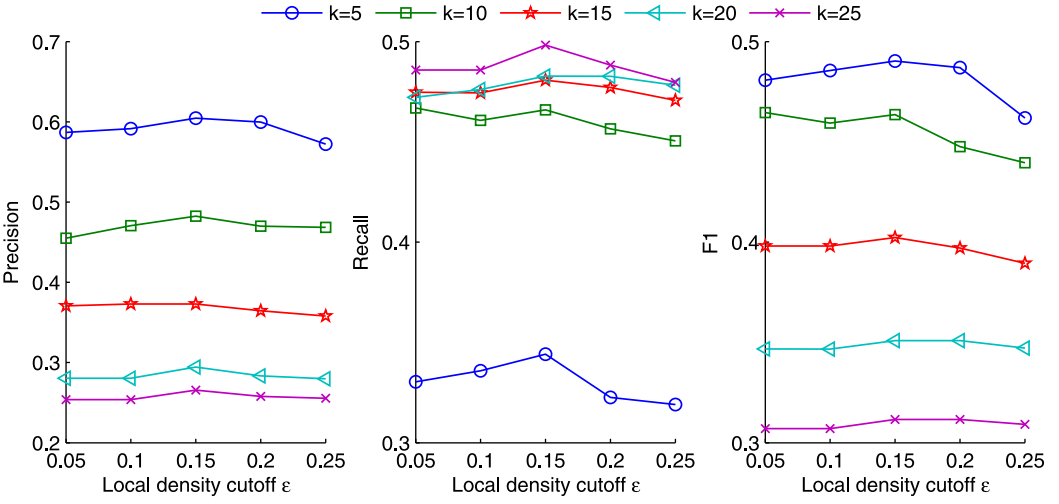

Fig. 16. Precision, recall, and F1 values based on the HIOC method under different values of ε for the *TW* dataset ($\alpha = 0.9$).

For the *SW* dataset, recall takes a highest value at $\varepsilon = 0.15$ under each recommendation list size. For the *TW* dataset, the best performance of recall is also achieved at $\varepsilon = 0.15$. The reason is that the appropriate cutoff can intuitively generate appropriate detected communities. The appropriate number of communities can distribute different interest subjects in various communities. As users in the same community are similar from the aspect of fine-granularity interest, we can push diverse subjects to the target user according to the communities that the target user belongs to. Hence, the effective values of ε can be adjusted for obtaining the appropriate communities for personalized recommendation.

7. Conclusions

In this study, we propose an HIOC-based recommendation model. Unlike previous similarity researchers, we utilize the interest tree structure of ontology user profiles to compute content interest closeness and semantic interest closeness between users. Considering hierarchical subjects, we combine multi-granular subject similarity and follow similarity of users to compute the interest linkage density of nodes and perform community detection. Then, by merging overlapping communities, we analyze the membership of a user in communities and that of a subject distributed in communities to complete personalized recommendation. The proposed scheme finds communities that have similar semantic interests and thus can provide multi-granular semantics-related subjects for target users. The evaluation results demonstrate that our mechanism shows improved performance in terms of precision, recall, and F1 measure compared with classical methods.

For future works, we will extend the proposed method to combination recommendation of different granular communities to satisfy users' diverse needs. In real-world networks, top leaders might differ in various granular communities. The process of finding top leaders in different granular interest communities is an interesting problem, and we also plan to study leaders' role migration in different granular interest communities. In addition, as users' interests shift over time, researching

the update of communities is a promising problem. Finally, we also aim to investigate the effect of community migration and evolution on multi-granular communities.

Acknowledgments

We would like to show our thanks to important anonymous reviewers for their insightful comments and wonderful suggestions that must lead to a much higher quality of our manuscript. This work was partially supported by the National Natural Science Foundation of China (nos. 61603229, 61632011, 61672331, 61573231, 61432011), and the Natural Science Foundation of Shanxi (201601D202041).

References

[1] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 734–749.
[2] A. Albadvi, M. Shahbazi, A hybrid recommendation technique based on product category attributes, Expert Syst. Appl. 36 (9) (2009) 11480–11488.
[3] L. Bai, X. Cheng, J. Liang, et al., Fast graph clustering with a new description model for community detection, Inf. Sci. 388 (2017) 37–47.
[4] X. Bai, B. Cambazoglu, F. Gullo, et al., Exploiting search history of users for news personalization, Inf. Sci. 385–386 (2017) 125–137.
[5] J. Bernab-Moreno, A. Tejeda-Lorente, C. Porcel, et al., Quantifying the emotional impact of events on locations with social media, Knowl. Based Syst. 146 (2018) 44–57.
[6] K. Bok, J. Lim, H. Yang, Social group recommendation based on dynamic profiles and collaborative filtering, Neurocomputing 209 (2016) 3–13.
[7] 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.
[8] R. Cilibrasi, P. Vitnyi, The google similarity distance, IEEE Trans. Knowl. Data Eng. 19 (3) (2004) 370–383.
[9] H. Feng, J. Tian, H. Wang, et al., Personalized recommendations based on time-weighted overlapping community detection, Inf. Manag. 52 (7) (2015) 789–800.
[10] C. Gui, R. Zhang, R. Hu, et al., Overlapping communities detection based on spectral analysis of line graphs, Phys. A Stat. Mech. Appl. 498 (2018) 50–65.
[11] P. Hanks, P. Hanks, Word association norms, mutual information, and lexicography, in: Meeting on Association for Computational Linguistics, 1989, pp. 76–83.
[12] M. Huang, G. Zou, B. Zhang, et al., Overlapping community detection in heterogeneous social networks via the user model, Inf. Sci. 432 (2018) 164–184.
[13] E. Iosif, A. Potamianos, Unsupervised semantic similarity computation between terms using web documents, IEEE Trans. Knowl. Data Eng. 22 (11) (2010) 1637–1647.
[14] Y. Jiang, J. Shang, Y. Liu, Maximizing customer satisfaction through an online recommendation system: a novel associative classification model, Decis. Support Syst. 48 (3) (2010) 470–479.
[15] A. Kardan, M. Ebrahimi, A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups, Inf. Sci. 219 (219) (2013) 93–110.
[16] D. Kim, C. Park, J. Oh, Deep hybrid recommender systems via exploiting document context and statistics of items, Inf. Sci. 417 (2017) 72–87.
[17] J. Kunegis, Konect2015, http://konect.uni-koblenz.de/networks/, 2015.
[18] A. Lancichinetti, S. Fortunato, Limits of modularity maximization in community detection, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 84 (2) (2011) 066122.
[19] A. Lancichinetti, S. Fortunato, J. Kertesz, Detecting the overlapping and hierarchical community structure of complex networks, New J. Phys. 11 (3) (2008) 19–44.
[20] C. Lee, F. Reid, A. Mcdaid, et al., Seeding for pervasively overlapping communities, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 83 (2) (2011) 066107.
[21] J. Leskovec, Snap2017, http://snap.stanford.edu/data/, 2017.
[22] D. Lewis, Naive (Bayes) at forty: the independence assumption in information retrieval, in: European Conference on Machine Learning, 1998, pp. 4–15.
[23] J. Lin, K. Sugiyama, M. Kan, 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, 2014, pp. 647–656.
[24] H. Long, Overlapping community detection with least replicas in complex networks, Inf. Sci. 453 (2018) 216–226.
[25] Z. Lu, H. Ip, Y. Peng, Contextual kernel and spectral methods for learning the semantics of images, IEEE Trans. Image Process. 20 (6) (2011) 1739–1750.
[26] P. Meo, A. Nocera, G. Terracina, et al., Recommendation of similar users, resources and social networks in a social internetworking scenario, Inf. Sci. 181 (7) (2011) 1285–1305.
[27] M. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 69 (6 Pt 2) (2004) 066133.
[28] M. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 69 (2 Pt 2) (2004) 026113.
[29] G. Paliouras, Discovery of web user communities and their role in personalization, User Model. User-adapt. Interact. 22 (1–2) (2012) 151–175.
[30] C. Park, D. Kim, J. Oh, Using user trust network to improve top-k recommendation, Inf. Sci. 374 (2016) 100–114.
[31] U. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 76 (3 Pt 2) (2007) 036106.
[32] X. Robin, N. Turck, A. Hainard, et al., Proc: an open-source package for r and s+ to analyze and compare roc curves, BMC Bioinform. 12 (1) (2011) 1–8.
[33] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492.
[34] B. Sarwar, G. Karypis, J. Konstan, Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 2001 International Conference on World Wide Web, 2001, pp. 285–295.
[35] J. Serrano-Guerrero, E. Herrera-Viedma, J. Olivas, et al., A google wave-based fuzzy recommender system to disseminate information in university digital libraries 2.0, Inf. Sci. 181 (9) (2011) 1503–1516.
[36] J. Serrano-Guerrero, F. Romero, J. Olivas, Hiperion: a fuzzy approach for recommending educational activities based on the acquisition of competences, Inf. Sci. 248 (6) (2013) 114–129.
[37] K. Siu, H. Meng, Semi-automatic acquisition of domain-specific semantic structures, Eurospeech (1999).
[38] X. Tang, T. Xu, X. Feng, et al., Learning community structures: global and local perspectives, Neurocomputing 239 (2017) 249–256.
[39] X. Tao, Y. Li, N. Zhong, A personalized ontology model for web information gathering, IEEE Trans. Knowl. Data Eng. 23 (4) (2010) 496–511.
[40] D. Vallet, I. Cantador, J. Jose, Personalizing web search with Folksonomy-based user and document profiles, in: European Conference on Advances in Information Retrieval, 2010, pp. 420–431.
[41] S. Wan, Y. Lan, J. Guo, et al., Informational friend recommendation in social media, in: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013, pp. 1045–1048.
[42] H. Wu, Y. Pei, B. Li, et al., Item recommendation in collaborative tagging systems via heuristic data fusion, Knowl. Based Syst. 75 (2015) 124–140.
[43] H. Xie, Q. Li, M. X.D., et al., Mining latent user community for tag-based and content-based search in social media, Comput. J. 57 (9) (2014) 1415–1430.
[44] G. Yin, Q. Sheng, Research on ontology-based measuring semantic similarity, in: International Conference on Internet Computing in Science and Engineering, 2008, pp. 250–253.
[45] Y.M. Li, Y. S., A diffusion mechanism for social advertising over microblogs, Decis. Support Syst. 54 (1) (2012) 9–22.

[46] C. Zhang, D. Li, J. Liang, Hesitant fuzzy linguistic rough set over two universes model and its applications, Int. J. Mach. Learn. Cybern. 9 (4) (2018) 577–588.
[47] J. Zhang, J. Tang, L. W.K., et al., Role-aware conformity influence modeling and analysis in social networks, in: Proceedings of 28th AAAI Conference on Artificial Intelligence, 2014, pp. 958–964.
[48] E. Zhong, W. Fan, Q. Yang, Adaptive User Distance Modeling in Social Media, 2014.