symmetry

Article

Group User Profile Modeling Based on Neural Word Embeddings in Social Networks

Jianxing Zheng [1,*], **Deyu Li** [1,2] **and Arun Kumar Sangaiah** [3]

[1] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China; lidysxu@163.com
[2] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China
[3] School of Computing Science and Engineering, VIT University, Vellore 632014, India; arunkumarsangaiah@gmail.com
* Correspondence: jxzheng@sxu.edu.cn; Tel.: +86-351-701-0566

Received: 4 September 2018; Accepted: 25 September 2018; Published: 26 September 2018

Abstract: How to find a user's interest from similar users a fundamental research problems in socialized recommender systems. Despite significant advances, there exists diversity loss for the majority of recommender systems. With this paper, for expanding the user's interest, we overcome this challenge by using representative and diverse similar users from followees. First, we model a personal user profile vector via word2vec and term frequency mechanisms. According to user profiles and their follow relationships, we compute content interaction similarity and follow interaction similarity. Second, by combining two kinds of interaction similarity, we calculate the social similarity and discover a diverse group with coverage and dissimilarity. The users in a diverse group can distinguish each other and cover the whole followees, which can model a group user profile (GUP). Then, by tracking the changes of followee set, we heuristically adjust the number of diverse group users and construct an adaptive GUP. Finally, we conduct experiments on Sina Weibo datasets for recommendation, and the experimental results demonstrate that the proposed GUP outperforms conventional approaches for diverse recommendation.

Keywords: micro-blog; word embeddings; diverse recommendation; group user profile

1. Introduction

Over the past years, millions of people have expressed and share their views on the social networking platform [1]. Especially in the micro-blog social network scenario, when people post instant messages for expressing their activities, their fans are willing to repost interesting messages and comment on related posts. These interaction behaviors generate a huge amount of abundant data. In these interaction behaviors, people's historical experience, cognitive consciousness and psychological emotion can reflect the characteristics of individuals; and social environment, business background can affect users' personalized behaviors. Combining cognitive experience with socialized attributes can mine users' potential interest tendency and behavior motivations. Recommender systems provide users with personalized service based on people's perceived experience and socialized behaviors. Therefore, how to acquire personalized rich interests is still a challenging problem in most recommender systems [2,3].

User profiles (UP) reflects the preferences of users, which is one of the main techniques in socialized recommender systems [4–8]. By modeling users profiles based on one's context (location, time, expertise, etc.), Benouaret et al. [9] designed a hybrid recommender system for mobile devices to improve the visitors' museum experience. With a groupware application (synchronous or asynchronous), Duque et al. [10] clustered similar users' model to obtain groups of users for the automatic generation of user interaction models.

Accurate user profiles can efficiently improve the quality of recommendation system. In social networks, we can model UPs with similar users to expand user's interests. Actually, micro-blogs are short and lack useful information for UP construction, which precludes discovering similar users. Meanwhile, sparse data derived from short texts can cause cold start problem for new target users. Existing solutions aim to discover latent similar users by modeling semantic representation for users and subjects. As a distributed representation of words, word embedding presents a vector representation method through a neural probabilistic language model, which can be applied in the fields of sentiment analysis, text classification, emotion classification, topic modeling, recommender systems and so on [11–15]. The low-dimensional feature vector is useful when finding semantics-related topics for user profile modeling. Thus, we can capture latent similar users in terms of user profile embedding to solve cold start recommendation problem.

In micro-blog social networks, most recommender systems adopt collaborative filtering or hybrid strategies to provide users with numerous valuable information from similar users [16–18]. However, existing works based on similar users often supply repeated items, which lack interest diversity. Indeed, the repeated items may cause users become bored with the recommendation results. The follow relationship in social network reflects users' preferences, which is helpful to mine user interests. More, based on different contents from different followees, we can get different diverse topics. However, for a special user, one's followee set may be large and each followee may post or repost lots of micro-blogs. As a result, it is time consuming to find interest from all similar followee users. Diverse interest mining from representative users can be used to decision analysis, vote, e-commerce, web search engines, socialized recommender systems and museum visitors' quality of experience and other fields [4,16,18–21]. Thus, it is reasonable that the followees could be replaced by a dissimilar and representative group to GUP for recommendation [7,10,19]. The GUP can efficiently provide diverse interests for target users than conventional UPs.

In this work, we introduce a novel GUP modeling method based on followees' representative users. The goal of our mechanism is to generate a robust GUP, which can make relevant and diverse interest recommendation. Concretely, considering users' interest subjects from micro-blogs, personal user profile vector is modeled by weighted subject embeddings. Then, a cosine metric on user profile vector is developed to measure the content interaction similarity, and a followee-in-common method is used to compute the follow interaction similarity. By combining content interaction similarity with follow interaction similarity, the social similarity is measured. Finally, based on the coverage degree of social similarity, the representative group of followees is discovered and aids to model a GUP for recommendation. When the scale of followees is changing, we can heuristically track the change of group to construct an adaptive GUP instead of recalculating new representative users. The method aims to automatically extract followees' diverse interests for target users.

The contribution of our paper are the following.

1. We model users' semantic relationships by considering user embeddings and subject embeddings, which are useful to find similar users with common semantic interests.
2. Based on the coverage ability of social similarity, representative and diverse group is selected to construct GUP. The diverse group can enrich the diversified interests for target users.
3. By adjusting the size of radius for representative group, GUP is updated automatically to match the changes of followees. The adaptive GUP can match the possible interest changes due to followees changes.

The remainder of this paper is organized as follows. In Section 2, we briefly discuss the works related to UP. In Section 3, we present a recommendation overview based on GUP. In Section 4, we introduce a method to build word embedding-based user profile. The adaptive GUP is modeled in

Section 5. In Section 6, we demonstrate the experimental results and discussions. Finally, Section 7 concludes the paper and addresses our future work.

2. Related Works

Research on representative group selection and diverse recommendation is relevant to the task of paper. In this section, we mainly discuss related work of user profile modeling, similar relationship discovery and recommendation approaches.

2.1. User Profile Modeling

In social recommender systems, modeling the UP from the raw micro-blog documents is usually challenging. Most of recommender systems developed UPs with individual content analysis or similar users' contents [22–24] to mine users' preferences. By classifying the visiting style of visitors, Zancanaro et al. [20] employed two unsupervised learning techniques for analzying museum visitors' behavior patterns. Using the keyword vector space from one's comments and articles, Meguebli et al. [25] exploited a user interest profile and article profile to recommend a list of articles. To achieve a personalized search, Du et al. [26] identified a user's likes and dislikes in terms of a method of multi-level user-profiling model. Based on user behavior modeling and user satisfaction capturing, Tsiropoulou et al. [27] adopted optimal and heuristic strategies to design a social recommendation and personalization approach, which can recommend a set of exhibits to the visitors. Considering basic information attributes, synthetic attributes and probability distribution attributes, Yang et al. [16] designed a framework named UMT (User-profile Modeling based on Transactional data) for modeling group profiles, which is suitable for personalization of transaction-based commercial application systems. Using social annotations, Mezghani et al. [28] contributed on characteristics of social user and techniques about modeling tag-based user profile and updating corresponding profiles for recommendation. Based on attribute types, attribute representations, and profiling methods, Shamri et al. [29] explored user-profiling approaches to make demographic recommendation. In addition, some researches use semantics model to detect user preferences. Lin et al. [30] tracked text of users and learned a semi-supervised latent dirichlet allocation (LDA) model to characterize a group of topics for modeling UP. Boratto et al. [31] tried to model users' interests and detect segments of users in terms of a vector representation of the words. By capturing the sequential effect, geographical influence, temporal cyclic effect and semantic effect into the embedding learning of his/her checked-in points of interests, Xie et al. [32] developed a novel time-decay method to dynamically compute users' latest preferences.

In this paper, considering the ontology structure, term frequency-inverse document frequency (TF-IDF) mechanism and word embedding, we model user profile vector to describe users' interests.

2.2. Similar Relationship Discovery

In social networks, socialized links and latent semantic relationships are used for finding similar users for recommendation. To address the sparsity problem of user–item interaction, Zhang et al. [17] utilized heterogeneous network embedding method to extract items' structural representations for network nodes and their relationships, which can improve the quality of recommender systems. By exploiting the links between users, Cantador et al. [33] extracted relations among similar users and identified semantic communities of interest, which are formed by a group of users with common interests. Considering both the similarity and the trustworthiness among users, De et al. [34] proposed a quantitative measure of group compactness for understanding the dynamics of group formation and evolution. According to the mature link relationships in the online social network, considering the similarity between user profiles, co-occurrence of user names and interaction behaviors, Xiong et al. [35] investigated a probabilistic graphical model to predict an accurate similarity of social relationship strengths and make recommendations. In addition, some authors presented semantics to model similar relationships between users. Li et al. [36] learned the distributional word representation model by mixed word embeddings to explore semantic relations for users. By incorporating various relations among users, tags and resource, Zhu et al. [37] proposed a novel heterogeneous hypergraph embedding framework for document recommendation. Based on user and item interaction graph, Dai et al. [38] proposed a novel deep coevolutionary network model to capture complex mutual relationship between users and items, and their evolution. Wen et al. [39] proposed a novel recommendation method in social networks based on user link network embedding mechanism.

Although existing user relationship studies have achieved the remarkable achievements, socialized recommendation fully relies on similar users, which can induce the repetitive recommendation results and ignore the diversity of users' interest needs.

2.3. Recommendation Approaches

Recommender systems involve three strategies: content-based recommendation, collaborative filtering (CF) method and hybrid method. Hybrid approach mainly performs recommendation by taking the hybrid between the content-based and CF method, which has been successfully applied into many recommender systems [40]. By collecting recent information on users' activities, Bok et al. [7] adopted the interests of other similar users in a group to make a group recommendation. Kefalas et al. [16] investigated two novel unified models by combining features from the collaborative filtering and content-based methods, which can provide review and Point-of-Interest recommendations. Applying the association rules to discover similar users, Kardan et al. [41] proposed a hybrid recommendation method to identify the user similarity neighborhood. In terms of topic enhanced word embeddings, Li et al. [42] used learning to rank algorithm to incorporate various features for recommending hashtags. Based on the historical interactions of the user on Web, Bai et al. [43] leveraged usage information to study the problem of news personalization, which showed a good ranked result list. By extracting implicit and reliable social information from user feedbacks, Zhang et al. [44] incorporated the top-k semantic friends information into matrix factorization (MF) and Bayesian personalized ranking (BPR) to solve the data sparsity problem. In terms of meta-path based random walk strategy, Shi et al. [45] designed node sequences for network embedding, which can mine latent structure features of users and items for heterogeneous information network based recommendation.

Although the above-mentioned studies can efficiently exploit users' interests for recommendation, the proposed approach takes a different stance to model diverse semantic preferences for target users. Differing from the existing works, a meaningful and representative group from followees is identified to generate a GUP for simulating target user's diverse interests.

3. Recommendation Overview Based on GUP

We design a group user profile by merging the interests of representative followees for a target user. The representative influential followees will be identified by considering the social similarity between user profiles. A snippet of the GUP recommendation as well as the generation process is presented in Figure 1.

Recommendation framework based on the GUP includes two stages. First, we find the representative group of followees for the target user. In particular, we applied the ontology structure, TF-IDF weight and word embeddings into the micro-blogs of users, which helps to generate the weighted subject embeddings. Next, we define the user profile vector and compute the content interaction similarity between users. Simultaneously, in the follow social network, the followees-in-common can measure the follow interaction similarity of two users. By the weighted sum of content interaction similarity and follow interaction similarity, the social similarity is computed. Based on the social similarity, the covered neighbors of each user are defined under a coverage threshold. Then, the representative group is discovered to simulate rich interests of the target user's followees.

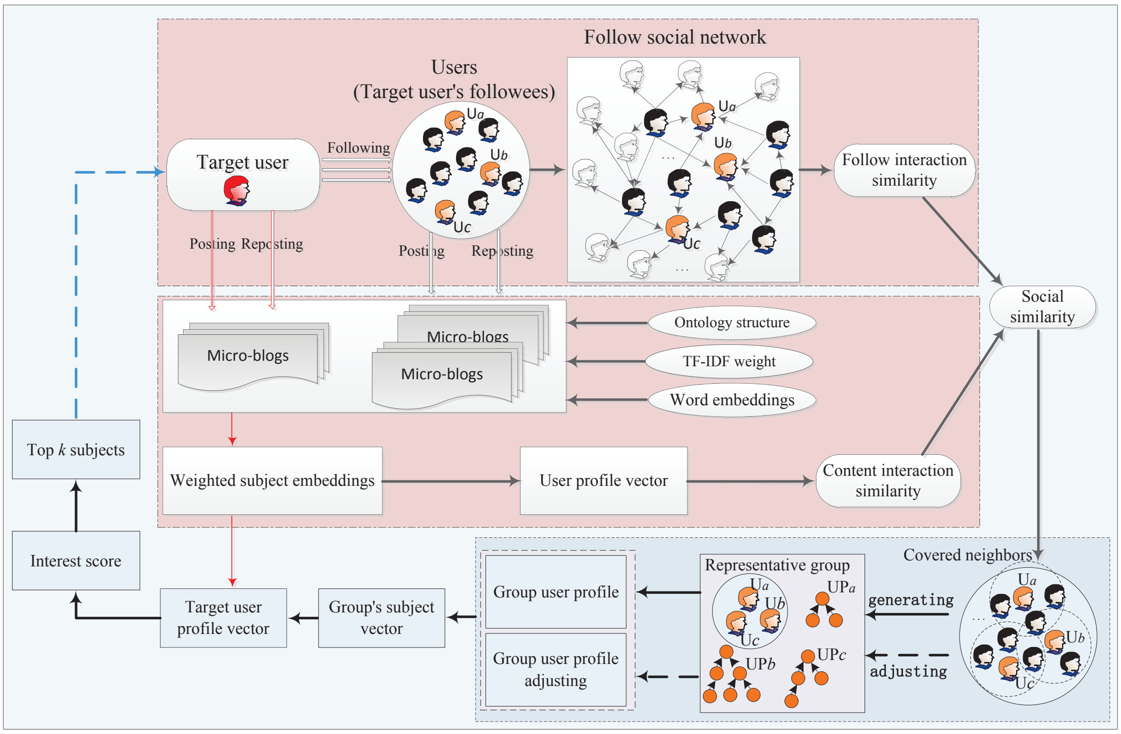

Figure 1. Recommendation overview based on the GUP.

Second, GUP is modeled and automatically adjusted with the change of zoom size of coverage threshold. We model group user profile based on UP (UPa, UPb, UPc) in the representative group in terms of collaborative strategy. Furthermore, by performing group's subject vector and target user profile vector jointly, the interest score of target user in the group's subject is computed. Specially, when the followees of target user change, we adjust the zoom size of coverage threshold and simultaneously update the representative group, which can assist to generate adapted group user profile. Based on the ranking of interest score, the top k subjects are selected and the relevant micro-blogs to these subjects are pushed to the target user to verify the comprehensiveness of the GUP.

4. Word Embedding-Based User Profile

In this work, we investigate the interest subjects of a user using word embeddings to form user profile vector. The method works in three steps (i) subject extraction; (ii) subject embeddings extraction and (iii) user profile representation. In the following, we will explain how each step works.

4.1. Subject Extraction

In social networks, based on the follows' resource corpus, we compute the TF-IDF weight of each term in micro-blogs. Hence, given a micro-blog m, we can define it as a vector of term and TF-IDF weight pairs in Equation (1).

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

Here, W_{pm} is the relative importance of term p in m. W_{pm} is a TF-IDF weight and defined as:

$$W_{pm} = \frac{freq_{pm}}{\max_l(freq_{lm})} \times \log \frac{N_m}{n_p} \tag{2}$$

$freq_{pm}$ is the frequency of term p in m, $\max_l(freq_{lm})$ is the maximum frequency of the term in m. N_m is the number of corpus micro-blogs. For each term p, n_p is the quantity of micro-blogs including p. The weight W_{pm} describes the importance of a term p in representing the message m.

In a micro-blog scenario, the noun entities play an important role in language understanding and generation, which efficiently reflects personal interest knowledge. However, all the noun entities on a corpus are disordered, and we match them with an ontology knowledge base to learn taxonomic structure. Figure 2 illustrates part of the categories related to five topics involved in Society, Sports, Economics, Culture and IT. The categorization corpus is from the Sogou Lab data, which describes a kind of inheritance relationship from an ancestor node to a child node.

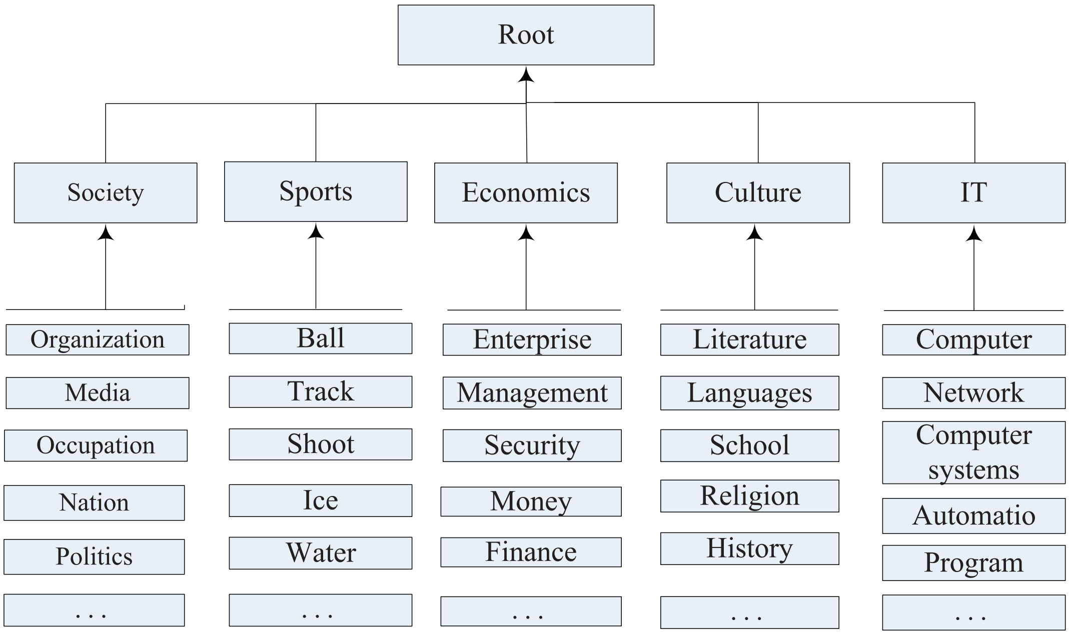

Figure 2. A sample of information categories for five topics.

According to the ontological category structure, for user u, we can extract the interest subjects appearing in personal micro-blogs as personal subject set PS_u. Furthermore, considering all the noun entities subjects, the accumulated weight of subject s is defined as:

$$W_u(s) = \frac{\sum\limits_{m \in M_u} W_{sm} \times \eta(s,m)}{\sum\limits_{s_i \in PS_u} \sum\limits_{m \in M_u} W_{s_i m} \times \eta(s_i,m)} \tag{3}$$

where M_u is a set of micro-blogs for user u. $\eta(s,m) = 1$ if $s \in m$; otherwise, $\eta(s,m) = 0$.

The weight $W_u(s)$ describes a user's interest in a particular subject absolutely. The larger the weight is, the more the user is interested in the subject. However, the subjects rely heavily on surface form of word co-occurrence features and are not fully robust when handling synonyms.

4.2. Subject Embeddings Extraction

For each subject $s \in PS_u$, we consider its neural word embeddings by using Google's word2vec [12], which is a popular tool to train word vectors from a large document corpus. Word2vec mainly utilizes the context to learn vector representation in order to find the latent syntactic and semantic word relationship of large corpus, which includes two training model: CBOW and Skip-gram [46]. The CBOW model is three-layer neutral networks to utilize the surrounding word of a micro-blog to predict the word representation while the Skip-gram model learns vector representation of word in the other way [47]. In our work, we adopt the Skip-gram model to learn the word representation. Figure 3 shows the illustration of Skip-gram model.

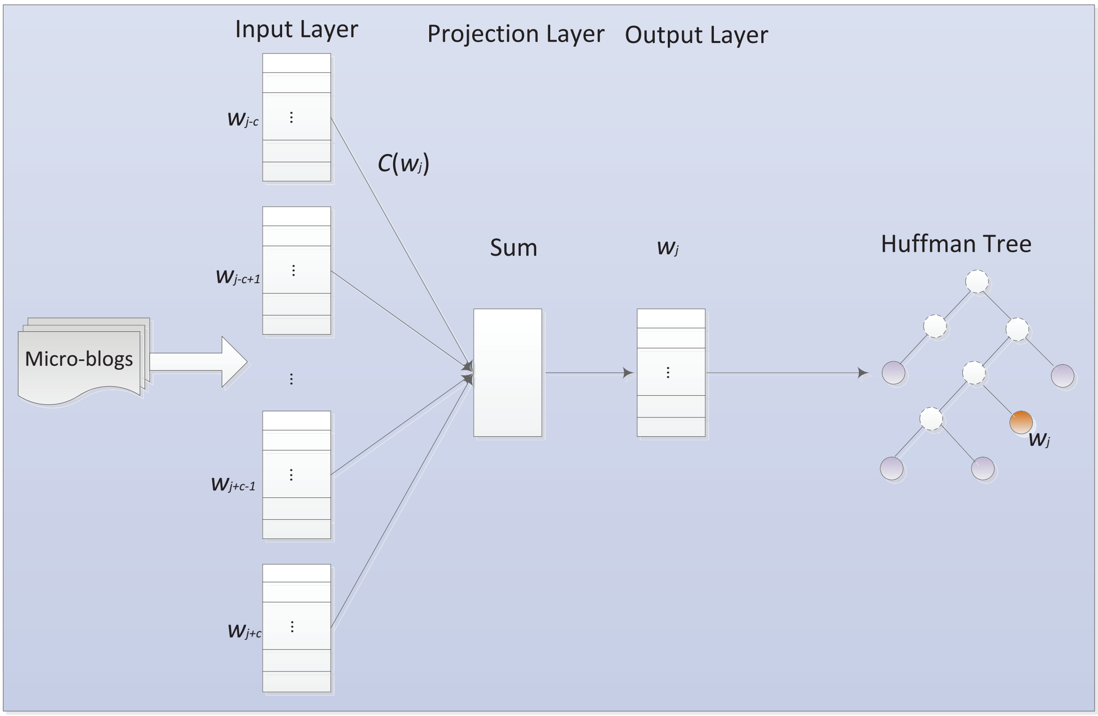

Figure 3. An illustration of Skip-gram.

According to the figure, given a sequence of training noun entities $w_1, w_2, ..., w_T$, the context of word w_j includes those of previous c words and behind c words, named as $C(w_j) = \{w_{j-c}, w_{j-c+1}, ..., w_{j+c-1}, w_{j+c}\}$. The Skip-gram learns the vector representation of each word w_j by maximizing the conditional probability $P(C(w_j)|w_j)$. Considering all the training nouns, the training goal of Skip-gram is maximizing the following average log probability:

$$L = \frac{1}{T}\sum_{j=1}^{T}\sum_{-c\leq i\leq c, i\neq 0}\log P(w_{j+i}|w_j) \tag{4}$$

where w_j is the center word and c is the size of the training context. For the Skip-gram model, we select hierarchical softmax method to speed-up training. At the output layer, it uses a binary tree representation with all the training vocabularies as the leaf nodes. For each word, there is a path from the root to that leaf with a corresponding weight in each edge. This structure promotes the words often occurring in the same context stay in near locations of tree.

4.3. User Profile Representation

The word vector $\overrightarrow{s}$ can represent the features of subject s for a user profile. However, users are immersed in expressing their interests by browsing content resource repetitively. So, we suppose that user's accumulated subject weight can affect user's interest distribution. That is, when the number of occurrences of the subject in personal micro-blogs increases, the user is also more interested in subject. Therefore, we can formally describe a personal user profile of user u as follows: $P\Theta_u = \{(W_u(s), \overrightarrow{s})|W_u(s) > 0\}$. A personal user profile is formed by a subject vector and accumulated subject weight pairs.

Using word2vec for subject representation, we represent a user profile vector by average weighted sum of all the subjects' vector:

$$\overrightarrow{P\Theta}_u = \frac{\sum\limits_{i=0}^{n} W_u(s_i)\overrightarrow{s}_i}{\sum\limits_{i=0}^{n} W_u(s_i)} \tag{5}$$

where n is the number of subjects for user u. The subject vector $\overrightarrow{s}$ used in the context of personal micro-blogs is utilized to model the user profile vector $\overrightarrow{P\Theta}_u$. In our paper, considering all the interest subjects, a collaborative weighted vector is performed to represent user. Then, user profile vector is employed to compute the content interaction similarity semantically.

5. Adaptive Group User Profile Modeling

Our aim is to build a group user profile automatically with the change of representative group of followees. The process is made up of three steps. In step 1, social similarity between users is computed. In step 2, group is discovered and assists the target user to generate GUP. In step 3, by changing the number of group users, GUP is adjusting to keep the probable change of interest.

5.1. Social Similarity

Social similarity mainly describes the closeness of users from aspect of socialized follow actions and interest resource, which includes follow interaction similarity and content interaction similarity. In micro-blog scenario, the more followees in common two users have, the more similar their interests may be. That is, when one person reposts a message from the common followee, another user would be also interested in the message and repost it. Here, we denote F_u as the followee set of user u. For two users u_i, u_j, based on common followees, their follow interaction similarity R_{ij}^{f} is defined as:

$$R_{ij}^{f} = \frac{|F_{u_i} \cap F_{u_j}|}{|F_{u_i} \cup F_{u_j}|} \tag{6}$$

In addition, user profile vector can be used to measure the content interaction similarity R_{ij}^{c} between users. Given two users' vector representation, we can compute R_{ij}^{c} as:

$$R_{ij}^{c} = \frac{\overrightarrow{P\Theta}_{u_i} \cdot \overrightarrow{P\Theta}_{u_j}}{||\overrightarrow{P\Theta}_{u_i}||||\overrightarrow{P\Theta}_{u_j}||} \tag{7}$$

According to the follow interaction and content interaction similarity, the social similarity between u_i and u_j can be calculated by Equation (8):

$$R_{ij} = \alpha R_{ij}^{c} + (1-\alpha)R_{ij}^{f} \tag{8}$$

Note that parameter α weights the relative importance of content interaction and follow interaction in the process of social similarity measuring. The social similarity evaluates the intimacy between users from aspect of socialized relation and semantic relation.

5.2. Group User Profile Constructing

Group user profile aims to identify a group of representative users from user's followees and provide rich interest for the target user. In traditional recommender systems, we mainly find k nearest similar users to enrich the interest of a user profile. However, it encounters limitations at aspect of interest diversity. In our study, our goal is to find a set of diverse representative users based on the dissimilar relationship of users. As stated in [48], the diverse subset, which considers not only the coverage ability of each object but also the difference among selected objects, can fully maintain the characteristics of the whole data set.

Given a set of followees F for the user, we want to select a subset G of these users such that all the users in G are dissimilar to each other and each other user out of G can be covered by a similar user in G. To simplify the presentation, $\forall u_i, u_j \in F$, we assume that the dissimilarity between user u_i and u_j can be defined as $d(u_i, u_j) = 1 - R_{ij}$. Given a threshold radius r, we consider the covered neighbors of u_i as $C_r(u_i) = \{u_j | d(u_j, u_i) \leq r\}$, which denotes the set of user objects that can be replaced by user u_i.

The covered neighbors reflect the influence of the user on the whole dataset. Generally, the bigger the $C_r(u_i)$ is, the larger social influence the user has. Based on this hypothesis, we can utilize a greedy heuristic algorithm to select the minimal diverse subset G from followee set F. Algorithm 1 describes the extraction method of a diverse group.

Algorithm 1 Diverse group extraction method.

```
Require: A set of objects $F$, radius $r$, $G = \oslash$ and $G^* = \oslash$.
Ensure: A diverse subset $G$ of $F$.
    Step 1. Select the object $u$ with the largest $|C_r(u)|$.
    Step 2. $G = G \cup \{u\}$.
    Step 3. While there exist $u_i \in F$ do
    Step 4. Set $G^* \leftarrow \oslash$;
    Step 5. Select the object(s) $u_i$ with the largest $|C_r(u_i)|$, and satisfy $d(u_p, u_i) > r$, $u_p \in G$;
    Step 6. $G^* = G^* \cup \{u_i\}$;
    Step 7. If $|G^*| \leq 1$ then
    Step 8. Go to step 12;
    Step 9. Else
    Step 10. Select the object $u_i$ from $G^*$ again that satisfies $\max_{i=1}^{|G^*|}\{\min\{d(u_p, u_i)\} | u_p \in G, u_i \in G^*\}$;
    Step 11. End if
    Step 12. $G = G \cup \{u_i\}$;
    Step 13. $F = F / \{u_i\}$;
    Step 14. For each $u_j \in C_r(u_i)$ and $u_j \in F$ do
    Step 15. If there do not exist $u_k \in F - G - C_r(u_i)$ that satisfies $d(u_k, u_j) \leq d(u_i, u_j)$ then
    Step 16. $F = F / \{u_j\}$;
    Step 17. End if
    Step 18. End for
    Step 19. End while
    Step 20. Return $G$.
```

In Algorithm 1, Step 5 ensures that the selected users have representativeness and Step 10 can maximize the difference among the candidate users. In the process of selecting the minimum group subset, all the followees are needed to compute their covered neighbors. Considering the number of followees n, the time complexity of Step 3 is $O(n)$. However, in Steps 14–18, for the covered neighbors, we need to decide whether to delete the object from the left followee set. The time complexity of deleting covered neighbors is $O(n \times |n - m|)$, where $|n - m|$ represents the cardinal number of the remaining followee set. Thus, the overall time complexity of GUP modeling is $O(n^2 \times |n - m|)$. Through several iterations of the operators, the group G is selected to predict rich preferences for the target user, as the users in group could reflect diverse characteristics of the global followees. In real social networks, as most of users have relatively few followees (usually no more than 100), the execution of the algorithm is not time-consuming.

For the group set $G(u)$, considering the subject vector representation, we can represent a group user profile as $G\Theta_u = \{(W_G(s), \overrightarrow{s}) | W_G(s) > 0\}$. The group user profile $G\Theta_u$ is composed of a set of group's collaborative weight $W_G(s)$ and subject vector pairs.

According to the social similarity in the group, we can compute the group's collaborative weight $W_G(s)$ as follows:

$$W_G(s) = \frac{\sum\limits_{u_j \in G(u)} R_{ij} \cdot W_{u_j}(s)}{\sum\limits_{u_j \in G(u)} R_{ij}} \tag{9}$$

Here, R_{ij} is the social similarity between target user u_i and user u_j in group. The group's weight measures the interest degree of subject from aspect of representative users in terms of collaborative strategy.

Then, considering all the subjects' vector representation in the group, we can describe group user profile vector of $G(u)$ as a average weighted sum:

$$\overrightarrow{G\Theta}_u = \frac{\sum\limits_{i=0}^{n} W_G(s_i)\overrightarrow{s}_i}{\sum\limits_{i=0}^{n} W_G(s_i)} \tag{10}$$

where n is the number of subjects interested in by users in group $G(u)$. The group user profile vector can be used to discover similar communities, which is not investigated elaborately here. For a specific subject, we can represent a group's weighted subject vector as $\overrightarrow{s}_G = W_G(s) \times \overrightarrow{s}$. The group's weighted subject vector can be applied to compute the similarity with the target user, which is helpful to the group recommendations.

Furthermore, for a new subject s out of the target user u but attached to the group, we can infer the user's interest score as $score = \overrightarrow{P\Theta}_u \times \overrightarrow{s}_G$. The subjects deriving from the representative group will be assigned an interest score and pushed to the target user according to ranked score.

5.3. Group User Profile Adjusting

Specially, the radius r can determine the scale of group, which may induce the change of interest subjects. Generally, a large r can define enough similar users to be covered, which generate a few novel group users for complementing the diverse interests. Contrarily, a small r can result in a small number of covered users, which probably generates too many dissimilar group users. Therefore, when the zoom size of radius r changes, we need to compute new representative group to adjust the GUP. As stated in [48], we adopt the greedy idea to get new representative group.

In micro-blog social networks, for a user, if the number of followees is small, we need to reduce the value of r and get a relative larger number of group users to keep the probable diverse interests. Let us consider a new $r' < r$, we need to generate a new set G' to model GUP. Algorithm 2 gives a group amplifying method.

In Algorithm 2, for a new radius $r' < r$, as the covered neighbors of user is decreasing, we will get a larger G' to maintain enough interests. Steps 2–4 try to select the outcast users which are not covered by the origin group G under new radius r'. Steps 5–21 aim to get some new members from the outcast users and add them into the G to generate G'. Based on the G', we can update the group user profile vector and make new recommendations. In the process of acquiring new group set G', Step 5 selects new members into the G'. The maximum number of new members is that of the whole followee set, n. Meanwhile, Steps 16–19 ensure that the objects covered by the new member are deleted. The time complexity is $O(n \times |n - m'|)$, where $|n - m'|$ is the cardinal number of remaining objects outside the neighbors under new radius r'. Therefore, the overall time complexity of group amplifying is $O(n^2 \times |n - m'|)$.

Algorithm 2 Group amplifying method.

```
Require: A set of objects $F$, radius $r$, group set $G$ and set $G^* = \oslash$.
Ensure: A new diverse amplifying subset $G'$.
    Step 1. $G' \leftarrow G$.
    Step 2. For all $u \in G$ do
    Step 3. $Temp\_G = Temp\_G \cup \{C_r(u) \backslash C_{r'}(u)\}$.
    Step 4. End for
    Step 5. While there exist $u_i \in Temp\_G$
    Step 6. $G^* \leftarrow \oslash$;
    Step 7. Select the object(s) $u_i$ with the largest $|C_{r'}(u_i)|$, and satisfy $d(u_p, u_i) > r'$, $u_p \in G'$;
    Step 8. $G^* = G^* \cup \{u_i\}$;
    Step 9. If $|G^*| \leq 1$ then
    Step 10. Go to step 14;
    Step 11. Else
    Step 12. Select the object $u_i$ from $G^*$ again that satisfies $\max_{i=1}^{|G^*|}\{\max\{d(u_p, u_i)\}|u_p \in G', u_i \in G^*\}$;
    Step 13. End if
    Step 14. $G' = G' \cup \{u_i\}$;
    Step 15. $Temp\_G = Temp\_G / \{u_i\}$;
    Step 16. For each $u_j \in C_{r'}(u_i)$ and $u_j \in Temp\_G$ do
    Step 17. If there do not exist $u_k \in Temp\_G - G' - C_{r'}(u_i)$ that satisfies $d(u_k, u_j) \leq d(u_i, u_j)$ then
    Step 18. $Temp\_G = Temp\_G / \{u_j\}$;
    Step 19. End if
    Step 20. End for
    Step 21. End while
    Step 22. Return $G'$.
```

In addition, when the number of followees is large, we will increase the value of r to get a few stable group and rhythmically adjust the interests. By setting a new radius $r' > r$, we can also get a relatively smaller number of group G''. Algorithm 3 describes a group reducing method.

In Algorithm 3, for a new radius $r' > r$, the covered neighbors of each user is increasing. So, the number of new group G'' may decrease. Specially, we divide the selection process of new group into two stages. First, due to the increscent coverage of users in origin group G, we need to select small number of candidate users from the origin G into new group set G'' under new radius r'; and then add the remaining covered users from origin G into set G'. Second, for the rest users in set G' out of the origin G, we also choose some new users and add them into new group set G''. Concretely, in Algorithm 3, Steps 1–18 try to eliminate those similar covered users into the remaining set G' under new radius r'. The time complexity is $O(n^2 \times |n - m'|)$, where n is the maximal number of users in origin group G and $|n - m'|$ is the cardinal number of those eliminated similar covered users under new radius r'. In the second half of algorithm, Steps 20–36 want to investigate some other representative users from the remaining user set G' into new group set G''. Similarly, the time complexity is $O(n'^2 \times |n' - m'|)$, where n' is the number of users in remaining set G'. Then, the overall time complexity of group reducing is $O(n^2 \times |n - m'| + n'^2 \times |n' - m'|)$. As $n' \leq n$, the time complexity is $O(n^2 \times |n - m'|$. The whole algorithm aims to get a new group G'' by avoiding recalculating the covered neighbors of each user.

Based on the above idea, we can adjust the value of radius r for different users to find appropriate group. Different group can induce different group user profiles, which can affect group's subject vector and corresponding recommendation results.

Algorithm 3 Group reducing method.

```
Require: A set of objects $F$, radius $r$, origin group set $G$, new group set $G'' = \oslash$ and set $G^*, G^{'} = \oslash$.
Ensure: A new diverse reducing subset $G''$.
    Step 1. While there exist $u_i \in G$ do
    Step 2. Set $G^* \leftarrow \oslash$;
    Step 3. Select the object(s) $u_i$ with the largest $|C_{r'}(u_i)|$, and satisfy $d(u_p, u_i) > r'$, $u_p \in G''$;
    Step 4. $G^* = G^* \cup \{u_i\}$;
    Step 5. If $|G^*| \leq 1$ then
    Step 6. Go to step 10;
    Step 7. Else
    Step 8. Select the object $u_i$ from $G^*$ again that satisfies $\max_{i=1}^{|G^*|}\{\max\{d(u_p, u_i)\}|u_p \in G'', u_i \in G^*\}$.
    Step 9. End if
    Step 10. $G'' = G'' \cup \{u_i\}$;
    Step 11. $G = G/\{u_i\}$;
    Step 12. For each $u_j \in C_{r'}(u_i)$ and $u_j \in G$ do
    Step 13. If there do not exist $u_k \in G - G'' - C_{r'}(u_i)$ that satisfies $d(u_k, u_j) \leq d(u_i, u_j)$ then
    Step 14. $G = G/\{u_j\}$;
    Step 15. $Temp\_G = Temp\_G \cup \{u_j\}$;
    Step 16. End if
    Step 17. End for
    Step 18. End while
    Step 19. $G^{'} = F/\{G^{''} \cup Temp\_G\}$;
    Step 20. While there exist $u_i \in G'$ do
    Step 21. $G^* \leftarrow \oslash$
    Step 22. Select the object(s) $u_i$ with the largest $|C_{r'}(u_i)|$, and satisfy $d(u_p, u_i) > r'$, $u_p \in G'$;
    Step 23. $G^* = G^* \cup \{u_i\}$;
    Step 24. If $|G^*| \leq 1$ then
    Step 25. Go to step 29;
    Step 26. Else
    Step 27. select the object $u_i$ from $G^*$ again, which satisfies $\max_{i=1}^{|G^*|}\{\max\{d(u_p, u_i)\}|u_p \in G', u_i \in G^*\}$;
    Step 28. End if
    Step 29. $G'' = G'' \cup \{u_i\}$;
    Step 30. $G' = G'/\{u_i\}$;
    Step 31. For each $u_j \in C_{r'}(u_i)$ and $u_j \in G'$ do
    Step 32. If there do not exist $u_k \in G' - G'' - C_{r'}(u_i)$ that satisfies $d(u_k, u_j) \leq d(u_i, u_j)$ then
    Step 33. $G' = G'/\{u_j\}$;
    Step 34. End if
    Step 35. End for
    Step 36. End while
    Step 37. Return $G''$.
```

6. Experiments and Discussions

In this section, we conduct experiments by making subject recommendations using the proposed GUP and some other strategies.

6.1. Recommendation Strategies

The proposed GUP aims to discover representative users and make collaborative recommendations. For the GUP recommendation, we first set the dimension of subject vectors and GUP vectors both as 100 to compute the interest score of subject. As the radius r determines the scale of group users in the process of GUP modeling, we perform greedy algorithms by setting different radius values belonging to range (0–1). By increasing the values of r with a step 0.2, we observe the optimal group radius for datasets. Base on diverse group, we model group user profile vector according to Equation (10), and compute the corresponding interest score of a new subject for a user. We can choose top k subjects according to the interest score of GUP and make micro-blog recommendations involved in these subjects for target users.

CF recommendation is a classic strategy that finds the most similar friends to recommend interests. Therefore, according to TF-IDF weight of subject in Equation (4), we can compute the similarity between users as follows:

$$sim(u_i, u_k) = \frac{\sum\limits_{s \in S_{u_i} \cap S_{u_k}} (W_{u_i}(s) - \overline{W_u})(W_{u_k}(s) - \overline{W_u})}{\sqrt{\sum\limits_{s \in S_{u_i} \cap S_{u_k}} (W_{u_i}(s) - \overline{W_u})^2} \sqrt{\sum\limits_{s \in S_{u_i} \cap S_{u_k}} (W_{u_k}(s) - \overline{W_u})^2}} \tag{11}$$

where $S_{u_i} \cap S_{u_k}$ is the set of subjects interested in by u_i and u_k; and $W_u(s)$, $\overline{W_u}$ are personal interest weight and average weight for subject s in traditional UP.

Furthermore, by selecting the same number of similar users into set G, the collaborative interest weight of subject s is shown in Equation (12).

$$CFW_{u_i}(s) = \frac{\sum\limits_{u_k \in G_{u_i}} W_{u_k}(s) \times sim(u_i, u_k)}{\sum\limits_{u_k \in G_{u_i}} |sim(u_i, u_k)|} \tag{12}$$

Then, top k subjects are used to mine relevant micro-blogs for the target user.

As a document-topic generation model, LDA method assumed that a document can be represented as a multinomial distribution over a set of T topics; and each topic is a multinomial distribution related to the set of vocabulary words [49]. In social networks, based on one's micro-blog contents, we can generate latent topic distribution. Given a topic $z_j \in T$, we can get topics with maxmized probability distribution for all micro-blogs, $\max\limits_i \{P(z_j|m_i)\}$. By ranking each topic's maxmized probability, we select top-k topics and push relevant words into the set of recommendation list. For the LDA method, we set the number of semantic topic as 100, hyper parameters $\alpha = \beta = 0.01$, and use Gibbs sampling to model micro-blogs' latent topic distribution.

6.2. Evaluation of Subject Recommendations

6.2.1. Data Description and Experimental Settings

Real datasets from the Sina micro-blog platform, such as the *NLPIR* dataset and *Application* dataset, are used to examine the performance of various recommendation mechanisms. For the *NLPIR* dataset, we collected data from 4 December 2011 to 23 December 2011 in the *NLPIR* website (http://www.nlpir.org/). Lastly, 144 users are selected for GUP modeling. The number of their followees is 7830. That is, averagely, each user has about 55 followees for mining representative group. In this dataset, most of followees rarely posted or reposted micro-blogs, so we utilized 1346 different activated followees who had posted or reposted micro-blogs for subject extraction. In addition, we collected 42,938 users who were followed by modeling users' followees for computing social similarity. For the *Application* dataset, we got the newest micro-blogs, followee friends, mutual responses and interactions from 10 April 2013 to 29 April 2013 in the Sina micro-blog platform (http://open.weibo.com). We selected 525 users for GUP modeling. The number of their followees is 39,017. For each user, the average number of followees is approximate to 75. Due to the repetition of the followees, we investigated 3514 different activated followees with micro-blogs for subject extraction. They posted or reposted 11,198 micro-blogs. Meanwhile, we count the number of followees' followees is 136,202 for calculating social similarity. The details for two datasets are shown in Table 1.

Table 1. The detail of $NLPIR$ and $Application$ dataset.

	Nlpir Dataset	**Application Dataset**
No. of users	144	525
No. of followees	7830	39,017
AVG followees per user	55	75
No. of activated followees	1346	3514
No. of users' micro-blogs	6375	8336
No. of followees' micro-blogs	2096	11,198
No. of testing micro-blogs	5869	11,435

In all experiments, we divide each dataset into two periods according to the timestamp. Specially, for the $NLPIR$ dataset, the data is from 4 December 2011 to 23 December 2011. So, we split it into two parts every ten days. That is, the data of the first half period is from 4 December 2011 to 13 December 2011. They are mainly used to model users' interests and construct GUP. The second half period is from 14 December 2011 to 23 December 2011, which is utilized to test the accuracy of recommendation result. Similarly, we divide the $Application$ dataset into two parts for our experiments. We compare the proposed GUP method with several existing methods, such as CF, LDA and personal UP method.

In the GUP experiment, we preprocessed the two dataset as follows. Given the micro-blogs corpus, the first step of preprocessing is word segmentation and stop words removing. After word segmentation, we mainly use a stop word dictionary to filter the conjunctions, adverbs, articles and prepositions. Then, we preserve the significant noun subjects and terms with Wikipedia ontology knowledge base to learn taxonomic structure. Next, we replace some irregular words, typographical errors and informal words by formal words in terms of an irregular word alternative list, which can form word vocabulary. By applying the TF-IDF mechanism, we can compute the accumulated weight of a word for target user. In terms of tensorflow tool, we implement the word in vocabulary list into a vector representation. For a user, considering all words' TF-IDF weight and corresponding vector, we can get one's user profile vector. we set the dimension of subject vectors and GUP vectors both as 100 to compute the interest score of subject. According to the rank of interest score, top-k subjects are provided for the target user.

The LDA method first models followees' hidden topic distribution according to all the posted or reposted micro-blogs. Then considering the top-k topic distribution of followees, we pushed the related subjects to the target user. In the process of modeling latent topic distribution, we set the hyper parameters $\alpha = \beta = 0.01$. For the CF recommendation, we utilize the method in Equation (11) to discover similar users and compute collaborative interest weight by Equation (12). Then, top-k subjects are selected for the target user. In addition, we adopted the content-based UP in Equation (3) for recommendation. Specially, for each kind of recommendation mechanism, we then push micro-blogs involved in recommended subjects to the target user.

6.2.2. Evaluation Metrics

Precision and recall are conventional properties that measure the degree to which generated recommendations accurately match the personal interests [33]. The metrics are shown as follows:

$$Precision@N = \frac{|S_T \cap S_R|}{|S_R|} \tag{13}$$

$$Recall@N = \frac{|S_T \cap S_R|}{|S_T|} \tag{14}$$

$$F1@N = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

where S_T is the actual subject set of testing messages and S_R is the subject set of recommendation list in the training messages relying on the GUP, CF, LDA and personal UP methods.

In addition, Discounted Cumulative Gain (DCG) is a popular recommendation accuracy metric that gives more importance to highly ranked subjects and the relevance of each subject [50]. The measure can be defined as:

$$DCG@N = \sum_{i=1}^{N} \frac{2^{rel(i)} - 1}{\log_2(i+1)} \tag{16}$$

where $rel(i) = 1$ if i-the subject is relevant to the actual subject set of testing messages; otherwise, $rel(i) = 0$. With the definition of $DCG@N$, the ideal $DCG@N$, denoted as $IDCG@N$, is defined:

$$IDCG@N = \sum_{i=1}^{C} \frac{1}{\log_2(i+1)} \tag{17}$$

where C is the number of relevant subjects in the recommendation list set S_R. Specifically, we adopt the normalized $NDCG@N = DCG@N/IDCG@N$ to measure the relevancy performance of the recommendation list.

In most case, users enjoy diverse and novel subjects while they pay attention to their interests. Hence, the diversity is also an important quality for evaluating the usefulness of recommendation subjects for each user, which is defined as:

$$DIV@N = \frac{2}{|S_R|(|S_R| - 1)} \sum_{s_i \in S_R|, j<i} d(s_i, s_j) \tag{18}$$

where $d(s_i, s_j)$ is the discrimination of two arbitrary subjects in the recommendation list S_R. Specifically, $d(s_i, s_j) = 1$, if $s_i \neq s_j$; otherwise, $d(s_i, s_j) = 0$.

Note that we average the metric scores of each user over the test set for every evaluation to measure the quality of recommendations.

6.2.3. Results and Evaluation

For the GUP, LDA, CF and UP methods, we make experiments with precision, recall, F1, IDCG and DIV metrics by varying the recommendation list size k. In the process of GUP modeling, the social similarity parameter α affects the selection of group, and the covered radius r also impacts the scale of group. So, we need to analyze the effect of parameters on the results. In the experiment, we get the best performance of GUP model at α = 0.5, r = 0.5 for the $NLPIR$ dataset. Similarly, for $Application$ dataset, the most effective result is achieved at α = 0.5, r = 0.7. Figures 4 and 5 show average performance obtained by several methods on $NLPIR$ and $Application$ datasets. For two datasets, on the precision indicators, the GUP strategy is superior overall to LDA, CF and UP methods. That is, the GUP recommendation method not only diversifies users' interests but also improves the relative interest score of recommended subjects. For example, in the $NLPIR$ dataset, the GUP approach achieves 0.5076 with a recommendation-list size of 5, higher than 0.4966 for LDA, 0.4792 for CF and 0.4087 for UP. This shows that more than 365 of the 720 subjects recommended by GUP are averagely received by 144 users. The phenomenon states that the representative group can cover most subjects of personal interests than an equal number of similar users. The representative group can contain diverse interests, which can be pushed to the target user.

The recall and F1 curves for GUP, LDA, CF and UP methods are given in Figures 4 and 5 (center, right). The performance of GUP is obviously higher than those of other methods, which can better express diverse semantics. As the size of the recommendation list increases, the recall of GUP improves. That is, the recommended subjects can cover personal interests wonderfully. In the figure, the recalls for LDA and CF methods become increasingly close as the number of recommended subjects increases. This is because semantic relevant subjects have already appeared, while the novel interests are constantly emerging in two methods at a large recommendation list size.

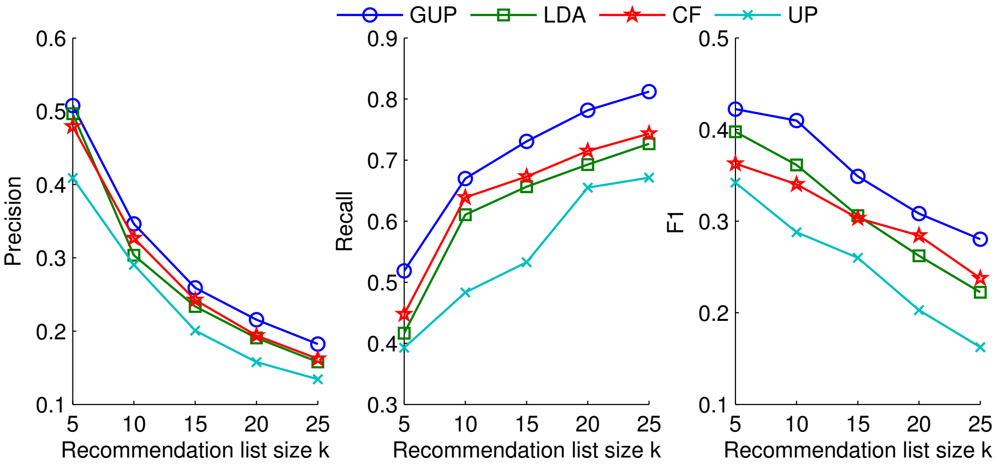

Figure 4. *Nlpir* dataset: precision, recall and F1 under different recommendation list size for 144 users ($\alpha = 0.5$, $r = 0.5$).

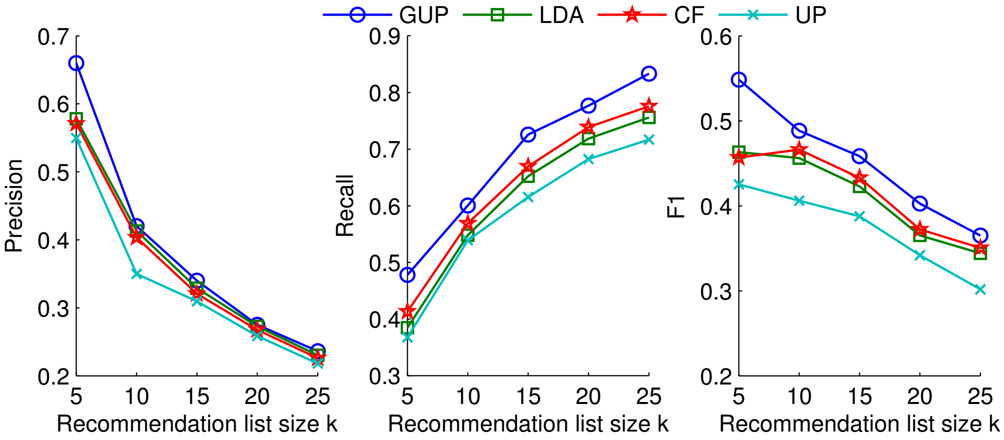

Figure 5. *Application* dataset: precision, recall and F1 under different recommendation list size for 525 users ($\alpha = 0.5$, $r = 0.7$).

In the process of GUP recommendation, as the number of diverse interests increases, the interest score of enhanced interests can also affect the ranking of recommendation list. Tables 2 and 3 measure influences of interest score by comparing the rank accuracy and the diversity of the recommendation list. According to the tables, GUP performs best among all compared methods. As k increases, the values of NDCG and DIV indicators in GUP method become significantly higher than those of other methods. This indicates that the interests in terms of GUP representation are highly ranked in the top of the list, which can both improve the performances of NDCG and DIV. For example, GUP outperforms all other models in terms of both NDCG@20 and DIV@20. On two datasets, NDCG and DIV values are basically stable, which indicates that the number of highly relevant diverse subjects is steadily increasing. It is clear that the interest score mechanism of GUP can gradually increase the diversity as the recommendation list size increases.

Table 2. *Nlpir* dataset: $NDCG$ and DIV under different k for 144 users.

	k = 5		k = 10		k = 15		k = 20		k = 25	
	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV
GUP	0.7076	0.7744	0.7304	0.7806	0.7439	0.7937	0.7235	0.7637	0.7583	0.7642
CF	0.5854	0.7152	0.5982	0.7155	0.6040	0.7012	0.6063	0.6993	0.6088	0.6913
LDA	0.6048	0.7134	0.6036	0.7146	0.5958	0.711	0.5939	0.7116	0.5929	0.7117
UP	0.5048	0.6634	0.5069	0.6646	0.5079	0.661	0.5081	0.6616	0.4980	0.6617

Table 3. *Application* dataset: $NDCG$ and DIV under different k for 525 users.

	k = 5		k = 10		k = 15		k = 20		k = 25	
	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV
GUP	0.8409	0.8158	0.8455	0.8074	0.8489	0.8252	0.8489	0.8174	0.8417	0.8121
CF	0.8147	0.7392	0.8151	0.7437	0.8125	0.7558	0.8121	0.761	0.8121	0.7652
LDA	0.827	0.7493	0.8271	0.7537	0.8246	0.766	0.8241	0.7612	0.8241	0.7555
UP	0.7483	0.7297	0.7687	0.7333	0.7747	0.7252	0.7725	0.7139	0.7732	0.6881

For the influence of social similarity, the value of parameter α can indirectly determine the number of covered neighbors, which induces the change of group. Figures 6 and 7 show precision, recall, and F1 at different recommendation lists under different α with fixed radius r for two datasets. These indicators obtain their maximal values at about α = 0.5 for different recommendation lists. Meanwhile, all the indicators are first increasing and then decreasing as the recommendation list size increases. This indicates that both the content interaction similarity and follow interaction similarity can influence the social similarity. The appropriate similarity can induce the best diverse group, especially when two factors have the same effect, which can generate the appropriate GUP. That is, when α = 0.5, most of recommendations can get relative balance for content interaction and follow interaction factors. Interestingly, the curves vary smoothly when α gradually increases but less than 0.5. Contrarily, when the value of α exceeds 0.5, the curves also change smoothly. This particularly reflects the importance of content interaction and follow interaction in the process of representative group selection.

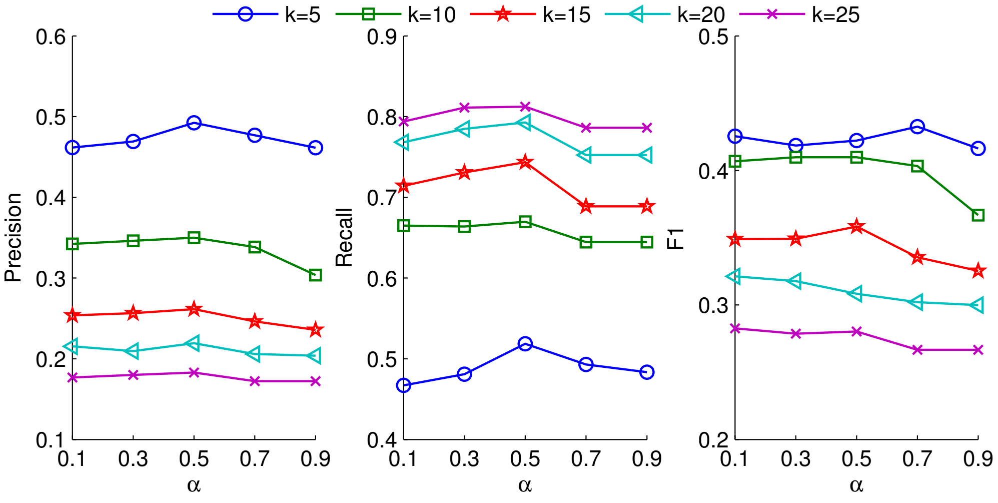

Figure 6. *Nlpir* dataset: precision, recall and F1 for GUP under different α and k (r = 0.5).

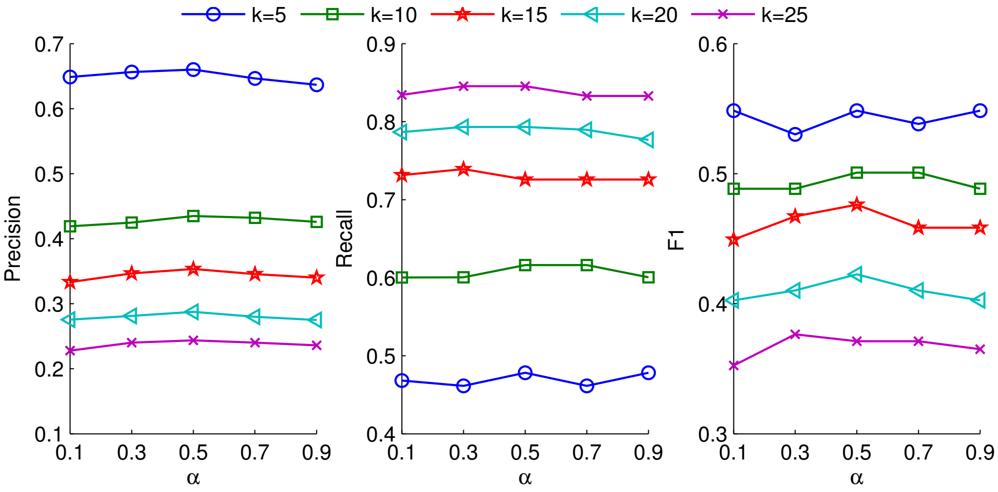

Figure 7. *Application* dataset: precision, recall and F1 for GUP under different α and k ($r = 0.7$).

Similarly, by changing the values of α and k, the results about NDCG and DIV based on GUP are shown in Tables 4 and 5. In the *Nlpir* dataset, the accuracy indicator (NDCG@20) of GUP first increases and then decreases when the α increases, which reaches its maximal value at $\alpha = 0.5$. For the novelty indicator DIV@20, a similar phenomenon can be achieved. In addition, we can see that NDCG slowly increases and DIV decreases with the value of k increasing. That is, based on the GUP, the recommended novel relevant subjects have large interest degree and are highly ranked. Meanwhile, the recommended irrelevant subjects rank lower, which are mostly the original interests of the target user. This observation suggests that the GUP can easily discover novel interests for the target user.

Table 4. *Nlpir* dataset: *NDCG* and *DIV* for GUP under different α and k ($r = 0.5$).

	***k* = 5**		***k* = 10**		***k* = 15**		***k* = 20**		***k* = 25**	
α	***NDCG***	***DIV***	***NDCG***	***DIV***	***NDCG***	***DIV***	***NDCG***	***DIV***	***NDCG***	***DIV***
0.1	0.6561	0.7889	0.6922	0.7944	0.7076	0.8148	0.6804	0.7622	0.6579	0.693
0.3	0.713	0.7811	0.7236	0.7857	0.7413	0.7942	0.7032	0.7374	0.6757	0.6988
0.5	0.7076	0.7744	0.7304	0.7806	0.7439	0.7937	0.7235	0.7637	0.7583	0.7642
0.7	0.7148	0.7721	0.7464	0.7782	0.7523	0.785	0.72	0.7528	0.6707	0.7051
0.9	0.7139	0.7695	0.7563	0.7725	0.7583	0.7763	0.72	0.761	0.6709	0.7064

Table 5. *Application* dataset: *NDCG* and *DIV* for GUP under different α and k ($r = 0.7$).

	***k* = 5**		***k* = 10**		***k* = 15**		***k* = 20**		***k* = 25**	
α	***NDCG***	***DIV***	***NDCG***	***DIV***	***NDCG***	***DIV***	***NDCG***	***DIV***	***NDCG***	***DIV***
0.1	0.8246	0.8367	0.8294	0.84	0.8275	0.85	0.8254	0.8383	0.8242	0.82
0.3	0.8353	0.8233	0.8276	0.8337	0.8368	0.8378	0.8316	0.8363	0.8348	0.8278
0.5	0.8409	0.8158	0.8455	0.8074	0.8489	0.8252	0.8489	0.8174	0.8417	0.8121
0.7	0.8468	0.7984	0.8478	0.8021	0.8489	0.8174	0.8484	0.8158	0.8469	0.8021
0.9	0.8584	0.796	0.8658	0.8003	0.868	0.807	0.8625	0.807	0.864	0.8003

For the GUP modeling, the value of radius r can also influence the number of group users. Different groups result in different recommendation results. In Figures 8 and 9, we show precision, recall and F1 results under different radius settings. From the figures, the better results are obtained at $r = 0.5$ and $r = 0.7$, respectively. When the radius r is large, the covered users are more, which can lead to a small representative group. The small group may not provide abundant subjects to expand diverse interests. In contrast, a small radius r could deduce a larger number of representative users from followees. The superabundant representative users may generate disordered and unsystematic subjects, which are irrelevant to the interest of target user and reduce the quality of recommendations. Figures 8 and 9 verify that the appropriate group can achieve high recall and F1 at the optimum value of r.

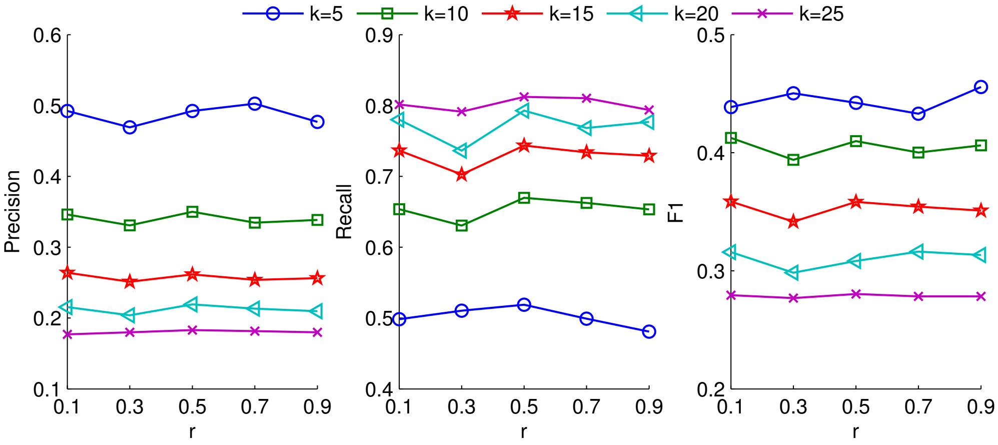

Figure 8. *Nlpir* dataset: precision, recall and F1 for GUP under different r and k (α = 0.5).

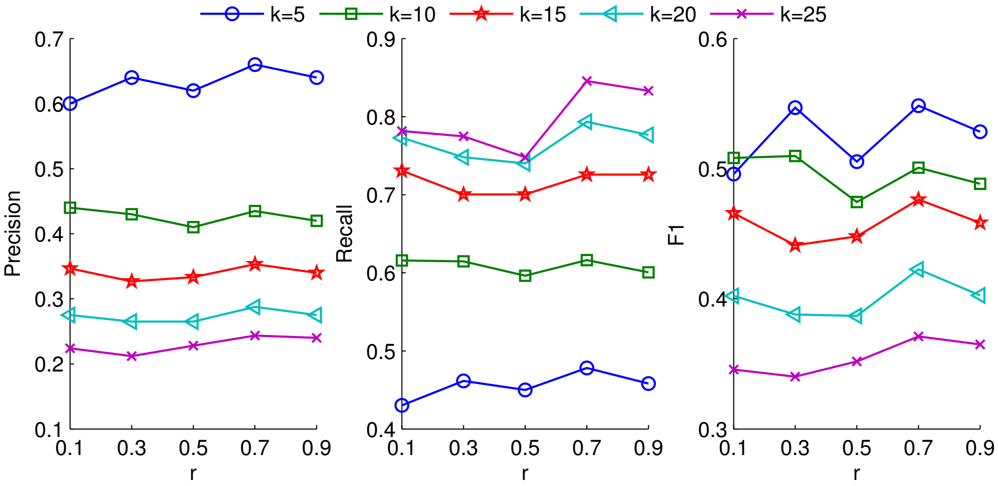

Figure 9. *Application* dataset: precision, recall and F1 for GUP under different r and k (α = 0.5).

To test the accuracy and diversity of GUP method under different radius r, Tables 6 and 7 show the NDCG and DIV results for GUP under different k for two datasets. In the *NLPIR* dataset, for each recommendation list size, the GUP performs best when r = 0.5, which is consistent with the phenomenon of precision. For the *Application* dataset, the radius r is 0.7. From the results, we can see that the number of group users can affect the number of recommended subjects. The appropriate group can provide rich novel subjects, which may be ranked in the top of the recommendation list, and promote the values of NDCG. Simultaneously, sufficient novel subjects can induce a large DIV. The values of DIV are maximal when r = 0.5 and r = 0.7 for two datasets.

Table 6. *Nlpir* dataset: *NDCG* and *DIV* for GUP under different r and k (α = 0.5).

	k = 5		k = 10		k = 15		k = 20		k = 25	
r	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV
0.1	0.6798	0.837	0.6522	0.7963	0.6976	0.8258	0.6838	0.7744	0.6579	0.8037
0.3	0.6688	0.7959	0.6601	0.7811	0.7134	0.8116	0.7196	0.7671	0.6757	0.7901
0.5	0.7076	0.7744	0.7304	0.7806	0.7439	0.7937	0.7235	0.7637	0.7583	0.7642
0.7	0.6931	0.7788	0.6957	0.7717	0.7578	0.7984	0.7308	0.7651	0.6707	0.7778
0.9	0.6957	0.7811	0.6977	0.7749	0.7448	0.7963	0.7391	0.7642	0.6709	0.7617

Table 7. *Application* dataset: *NDCG* and *DIV* for GUP under different r and k (α = 0.5).

	k = 5		k = 10		k = 15		k = 20		k = 25	
r	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV	$NDCG$	DIV
0.1	0.838	0.8035	0.8346	0.8123	0.8375	0.82	0.8275	0.8389	0.8221	0.81
0.3	0.8439	0.8233	0.835	0.8289	0.8368	0.8178	0.8368	0.8378	0.8359	0.8444
0.5	0.8331	0.8219	0.8407	0.8248	0.8433	0.8124	0.8433	0.8252	0.8392	0.8152
0.7	0.8409	0.8158	0.8455	0.8074	0.8489	0.8252	0.8489	0.8174	0.8417	0.8121
0.9	0.8477	0.7987	0.8564	0.7993	0.864	0.807	0.868	0.807	0.8552	0.8023

6.2.4. Discussions

Based on the above result analysis, we can see that the proposed GUP outperforms other conventional methods, especially for indexes of recall and DIV, but with a relatively high time complexity. In addition, there are some insights from our observations and analysis for the GUP recommendation. In Tables 2 and 3, the GUP not only can significantly enrich users' interests, but also keep related interests for the target user. In the process of GUP modeling, the sample group users can cover the whole followees' set as much as possible semantically. Meanwhile, the number of users in representative group is optimized minimum. Thus, in social networks, for users with many followees, there is a clear benefit that they can get required resources in terms of limited followee friends. Simultaneously, when followees the target user pays attention to changes, we need to select an adjusted group to match followees' interests. By setting different zoom size of neighborhood radius, we get different diverse group to obtain update interests.

Interestingly, the zoom radius can determine the scale of GUP. A large radius induces a small group, which can generate less additional preferences. On the contrary, a small radius leads a large group, which can expand rich interests for target user. Specially, when the value of radius is 1, the GUP will become original personal UP. Indeed, GUP provides a new diverse insight, which can maximize the profit in micro-blog advertisements, marketing, and recommender systems and visitors' quality of experience with small human resource cost.

7. Conclusions

In this paper, we have presented a novel GUP recommendation model based on representative groups. The users in a representative group are not only similar to each other, but also diverse. Concretely, the proposed work first models the subject vector and user profile vector from a semantic perspective. Then, by considering the follow interaction similarity and content interaction similarity between users, we have computed the social similarity and discovered a representative diverse group. Lastly, based on the diverse group, GUP is constructed and allows recommending subjects to target users.

Some other useful insights can be obtained from the experimental results, i.e., the proposed model outperforms the CF, LDA and personal UP methods at aspect of mining diverse interests. The main contribution of this work is identifying rich interests for target users in terms of a representative group.

Simultaneously, the heuristic strategy is efficiently used to adjust the scale of group for updating interests of group.

In micro-blog social networks, social interests relying on users' actions are multi-dimensional. In future, we intend focus on mining multi-dimensional semantic relations of subjects and make community detections based on semantic interests. Furthermore, considering representative users in a community, we can identify leader peers in the diffusion of interests.

Author Contributions: J.Z. conceived the idea of the paper and performed the experiments and write the paper; D.L. and A.K. Sangaiah reviews and revised the paper.

Funding: This research was funded by the National Natural Science Foundation of China (Nos. 61603229, 61672331 and 61573231) and Natural Science Foundation of Shanxi (No. 201601D202041).

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their helpful comments and suggestions. This research was funded by the National Natural Science Foundation of China (Nos. 61603229, 61672331 and 61573231) and Natural Science Foundation of Shanxi (No. 201601D202041). This work was partially supported by the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, Q.; Wu, J.; Zhang, Q.; Zhang, P.; Long, G.; Zhang, C. Dual influence embedded social recommendation. *World Wide Web* **2018**, *21*, 1–26. [CrossRef]
2. Ma, Y.; Zeng, Y.; Ren, X.; Zhong, N. User interests modeling based on multi-source personal information fusion and semantic reasoning. In Proceedings of the 7th International Conference on Active Media Technology, Lanzhou, China, 7–9 September 2011; pp. 195–205.
3. Ball, F.; Geyer-Schulz, A. How symmetric are real-world graphs? A large-scale study. *Symmetry* **2018**, *10*, 29. [CrossRef]
4. Abel, F.; Gao, Q.; Houben, G.J.; Tao, K. Analyzing user modeling on twitter for personalized news recommendations. In Proceedings of the International Conference on User Modeling, Adaption and Personalization, Girona, Spain, 11–15 July 2011; pp. 1–12.
5. Ikeda, K.; Hattori, G.; Ono, C.; Asoh, H.; Higashino, T. Twitter user profiling based on text and community mining for market analysis. *Knowl.-Based Syst.* **2013**, *51*, 35–47. [CrossRef]
6. Xie, H.; Li, X.; Wang, T.; Lau, R.Y.; Wong, T.L.; Chen, L.; Li, Q. Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy. *Inf. Process. Manag.* **2016**, *52*, 61–72. [CrossRef]
7. Bok, K.; Yang, H.; Yang, H.; Yoo, J. Social group recommendation based on dynamic profiles and collaborative filtering. *Neurocomputing* **2016**, *209*, 3–13. [CrossRef]
8. Tsiropoulou, E.E.; Thanou, A.; Papavassiliou, S. Modelling museum visitors' Quality of Experience. In Proceedings of the 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Thessaloniki, Greece, 20–21 October 2016; pp. 77–82.
9. Benouaret, I.; Lenne, D. Personalizing the museum experience through context-aware recommendations. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Kowloon, China, 9–12 October 2015; pp. 743–748.
10. Tîrnăucă, C.; Duque, R.; Montaña, J. User interaction modeling and profile extraction in interactive systems: A groupware application case study. *Sensors* **2017**, *17*, 1669. [CrossRef]
11. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
12. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *Comput. Lang. arXiv* **2013**, arXiv:1301.3781.
13. Xun, G.; Gopalakrishnan, V.; Ma, F.; Li, Y.; Gao, J.; Zhang, A. Topic discovery for short texts using word embeddings. In Proceedings of the 16th IEEE International Conference on Data Mining, Barcelona, Spain, 12–15 December 2016; pp. 1299–1304.

14. Hu, W.; Tsujii, J. A Latent Concept Topic Model for Robust Topic Inference Using Word Embeddings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 380–386.
15. Zhang, M. Efficient correlated topic modeling with topic embedding. In Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 225–233.
16. Yang, Y.; Marques, N.C. User group profile modeling based on user transactional data for personalized systems. In Proceedings of the 12th Portuguese Conference on Progress in Artificial Intelligence, Covilhã, Portugal, 5–8 December 2005; pp. 337–347.
17. Zhang, F.; Yuan, N.J.; Lian, D.; Xie, X.; Ma, W.Y. Collaborative knowledge base embedding for recommender systems. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 353–362.
18. Ullah, F.; Lee, S. Social content recommendation based on spatial-temporal aware diffusion modeling in social networks. *Symmetry* **2016**, *8*, 89. [CrossRef]
19. Li, Y.M.; Shiu, Y.L. A diffusion mechanism for social advertising over microblogs. *Decis. Support. Syst.* **2012**, *54*, 9–22. [CrossRef]
20. Zancanaro, M.; Kuflik, T.; Boger, Z.; Gorenbar, D.; Dan, G. Analyzing museum visitors' behavior patterns. In Proceedings of the 11th International Conference on User Modeling, Corfu, Greece, 25–29 June 2007; pp. 238–246.
21. Lykourentzou, I.; Naudet, Y.; Tobias, E.; Antoniou, A.; Lepouras, G.; Vassilakis, C. Improving museum visitors' Quality of Experience through intelligent recommendations: A visiting style-based approach. In Proceedings of the 9th International Conference on Intelligent Environments, Athens, Greece, 16–17 July 2013; pp. 507–518.
22. Zheng, J.; Zhang, B.; Zou, G. Multi-granularity recommendation based on ontology user model. In Proceedings of the 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, Beijing, China, 20–23 August 2013; pp. 2194–2199.
23. Xie, H.; Li, Q.; Mao, X.; Li, X.; Cai, Y.; Rao, Y. Community-aware user profile enrichment in folksonomy. *Neural. Netw.* **2014**, *58*, 111–121. [CrossRef] [PubMed]
24. Iglesias, J.A.; Angelov, P.; Ledezma, A.; Sanchis, A. Creating evolving user behavior profiles automatically. *IEEE. Trans. Knowl. Data Eng.* **2012**, *24*, 854–867. [CrossRef]
25. Meguebli, Y.; Kacimi, M.; Doan, B.L.; Popineau, F. Building rich user profiles for personalized news recommendation. In Proceedings of the 22nd Conference on User Modelling, Adaptation and Personalization Workshops, Aalborg, Denmark, 7–11 July 2014.
26. Du, Q.; Xie, H.R.; Cai, Y.; Leung, H.F.; Li, Q.; Min, H.Q.; Wang, F.L. Folksonomy-based personalized search by hybrid user profiles in multiple levels. *Neurocomputing* **2016**, *204*, 142–152. [CrossRef]
27. Tsiropoulou, E.E.; Thanou, A.; Papavassiliou, S. Quality of experience-based museum touring: A human in the loop approach. *Soc. Netw. Anal. Min.* **2017**, *7*, 1–13. [CrossRef]
28. Mezghani, M.; Zayani, C.A.; Amous, I.; Gargouri, F. A user profile modelling using social annotations: A survey. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 969–976.
29. Shamri, A.; Mohammad, Y.H. User profiling approaches for demographic recommender systems. *Knowl.-Based Syst.* **2016**, *100*, 175–187. [CrossRef]
30. Lin, J.; Sugiyama, K.; Kan, M.Y.; Chua, T.S. New and improved:modeling versions to improve app recommendation. In Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, Queensland, Australia, 6–11 July 2014; pp. 647–656.
31. Boratto, L.; Carta, S.; Fenu, G.; Saia, R. Using neural word embeddings to model user behavior and detect user segments. *Knowl.-Based Syst.* **2016**, *108*, 5–14. [CrossRef]
32. Xie, M.; Yin, H.; Wang, H.; Xu, F.; Chen, W.; Wang, S. Learning graph-based poi embedding for location-based recommendation. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; pp. 15–24.

33. Cantador, I.; Castells, P. Extracting multilayered Communities of Interest from semantic user profiles: Application to group modeling and hybrid recommendations. *Comput. Hum. Behav.* **2011**, *27*, 1321–1336. [CrossRef]
34. De, M.P.; Ferrara, E.; Rosaci, D.; Sarne, G.M. Trust and compactness in social network groups. *IEEE Trans. Cybern.* **2015**, *45*, 205–216.
35. Xiong, L.; Lei, Y.; Huang, W.; Huang, X.; Zhong, M. An estimation model for social relationship strength based on users' profiles, co-occurrence and interaction activities. *Neurocomputing* **2016**, *214*, 927–934. [CrossRef]
36. Li, J.; Li, J.; Fu, X.; Masud, M.A.; Huang, J.Z. Learning distributed word representation with multi-contextual mixed embedding. *Knowl.-Based Syst.* **2016**, *106*, 220–230. [CrossRef]
37. Zhu, Y.; Guan, Z.; Tan, S.; Liu, H.; Cai, D.; He, X. Heterogeneous hypergraph embedding for document recommendation. *Neurocomputing* **2016**, *216*, 150–162. [CrossRef]
38. Dai, H.; Wang, Y.; Trivedi, R.; Song, L. Deep coevolutionary network: embedding user and item features for recommendation. *arXiv* **2017**, arXiv:1609.03675.
39. Wen, Y.; Guo, L.; Chen, Z.; Ma, J. Network embedding based recommendation method in social networks. In Proceedings of the Companion of the the Web Conference, Lyon, France, 23–27 April 2018.
40. Le, H.S. Dealing with the new user cold-start problem in recommender systems: A comparative review. *Inf. Syst.* **2016**, *58*, 87–104.
41. Kardan, A.A.; Ebrahimi, M. A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups. *Inf. Sci.* **2013**, *219*, 93–110. [CrossRef]
42. Li, Q.; Shah, S.; Nourbakhsh, A.; Liu, X.; Fang, R. Hashtag recommendation based on topic enhanced embedding, tweet entity data and learning to rank. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; pp. 2085–2088.
43. Bai, X.; Cambazoglu, B.B.; Gullo, F.; Mantrach, A.; Silvestri, F. Exploiting search history of users for news personalization. *Inf. Sci.* **2017**, *385*, 125–137. [CrossRef]
44. Zhang, C.; Yu, L.; Wang, Y.; Shah, C.; Zhang, X. Collaborative user network embedding for social recommender systems. *Proc. SIAM Int. Conf. Data Min.* **2017**. [CrossRef]
45. Shi, C.; Hu, B.; Zhao, X.; Yu, P. Heterogeneous information network embedding for recommendation. *IEEE Trans. Cybern.* **2017**, *99*, 1–14. [CrossRef]
46. Giatsoglou, M.; Vozalis, M.G.; Diamantaras, K.; Vakali, A.; Sarigiannidis, G.; Chatzisavvas, K.C. Sentiment analysis leveraging emotions and word embeddings. *Expert Syst. Appl.* **2017**, *69*, 214–224. [CrossRef]
47. Enrquez, F.; Troyano, J.A.; Lpez-Solaz, T. An approach to the use of word embeddings in an opinion classification task. *Expert Syst. Appl.* **2016**, *66*, 1–6. [CrossRef]
48. Drosou, M.; Pitoura, E. Disc diversity: Result diversification based on dissimilarity and coverage. *Proc. VLDB Endow.* **2012**, *6*, 13–24. [CrossRef]
49. Wei, X.; Croft, W.B. LDA-based document models for ad-hocretrieval. In Proceedings of the 29th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 6–11 August 2006; pp. 178–185.
50. Wu, H.; Pei, Y.J.; Li, B.; Kang, Z.Z.; Liu, X.X.; Li, H. Item recommendation in collaborative tagging systems via heuristic data fusion. *Knowl.-Based Syst.* **2015**, *75*, 124–140. [CrossRef]