Mobile Netw Appl (2017) 22:228–239
DOI 10.1007/s11036-016-0790-9

The Recommendation System of Micro-Blog Topic Based on User Clustering

Shunxiang Zhang[1] · **Shiyao Zhang**[1] · **Neil Y. Yen**[2] · **Guangli Zhu**[1]

Published online: 27 December 2016

Abstract As a type of crowdsensing media, micro-blog has become an important crowdsensing place for a lot of real-time information dissemination and discussion. With the increasing of micro-blog users, there are more and more new topics emerging on this kind of platform, which has made the users difficult in finding out their own interesting topics. To solve this problem, this paper proposes a micro-blog topic recommendation system which can give corresponding suggestions/strategies for users. Firstly, the user relationship (i.e., a user adds a follow hyperlink to another user) in micro-blog can be effectively analyzed and saved to the user graph. In addition, an algorithm of computing user authority (which is similar to the idea of PageRank) is proposed to catch influential users based on the built user graph. Secondly, Topic Feature Graph (TFG) and User Micro-blog Feature Graph (UMFG) are respectively constructed based on the micro-blog text corpus of a topic and the micro-blog texts followed by a given user. Based on TFG and UMFG, User Topic Feature Vector (UTFV) and User Topic Feature Matrix (UTFM) can be achieved. After that, users' similarity is calculated based on the User Topic Feature Vector and User Topic Feature Matrix to realize the users clustering by the help of the hierarchical clustering algorithm. Incorporating topic heat degree and user authority, the recommendation algorithm is presented to realize Micro-blog topic personalized recommendation within user clustering set. Experiments show that our proposed recommendation system has a good accuracy which is up to 50.2%.

Keywords Recommendation system · Micro-blog topic · User semantic view · User topic feature vector · User clustering

1 Introduction

With the increasing popularity of the micro-blog, there are a total of 500 million registers in *Sina* micro-blog with the daily actively users participating in for 47 million. And the average number of followed of 1.18 million users has reached 469 [1]. Faced with large amount of information, it will be a problem when choosing the interesting topics in an effective way.

To solve this problem, this paper proposes a recommendation system of micro-blog topic based on user clustering. In order to better carry out the system, the following aspects should be taken into consideration. First of all, pay more attention to the users who are influential with their topics because they can offer some topics that can enjoy popularity. Secondly, it will be a problem to select the topics that can attract the users. Thirdly, it is about to gather the users who share similar interests. Fourthly, it is necessary to ensure the topics selected can be popular and acceptable by different kinds of users.

✉ Shunxiang Zhang
sxzhang@aust.edu.cn

Shiyao Zhang
yao__ing@163.com

Neil Y. Yen
neilyyen@u-aizu.ac.jp

Guangli Zhu
glzhu@aust.edu.cn

[1] Anhui University of Science & Technology, Huainan, Anhui 232001, China

[2] University of Aizu, Aizuwakamatsu, Fukushima 9658580, Japan

According to the above four aspects, the framework of the recommendation system of micro-blog topic based on user clustering is proposed, which is shown in Fig. 1. From Fig. 1, the proposed recommendation system is concerned with the following four layers.

(1) **The layer of user data processing.** First, the user graph is built according to the user relationship saved in user data set. Then, an algorithm of computing user authority, similar to PageRank, is proposed to catch influential users based on the built user graph.
(2) **The layer of the extraction of topic feature.** First, Topic Feature Graph(TFG) and User Micro-blog Feature Graph (UMFG) are respectively constructed by the technology of word of segmentation. Further, the User Topic Feature Vector(UTFV) and User Topic Feature Matrix(UTFM) are built by help of User Topic Feature Matrix Generation Algorithm.
(3) **The layer of user clustering.** Based on the User Topic Feature Matrix(UTFM), user similarity between any two users is calculated by the help of classical clustering algorithm known as hierarchical agglomerative clustering. Thus, all the micro-blog users with same feature are gathered and formed user clustering set.
(4) **The layer of Micro-Blog topic recommendation.** In this layer, topic heat degree is defined and computed. Incorporating topic heat degree and user authority computed in data processing layer, the recommendation algorithm is presented to realize Micro-blog topic personalized recommendation within user clustering set.

The rest of this paper is organized as follows: Section 2 introduces related works. Section 3 introduces the method of building micro-blog user graph. Section 4 presents the algorithm of building user micro-blog semantic view. Section 5 presents the clustering algorithm of Micro-blog user. The algorithm of micro-blog topic recommendation is presented in Section 6. Section 7 presents the experimental analysis from the efficiency of the proposed recommendation system. Conclusions are given in Section 8.

2 Related work

In this section, we briefly review some of the research papers related to our work. Papers about keywords extraction are show in the first part of this section. Works related to recommendation are represents in the second part of this section. Other technical used in out paper is described in the third part of this section.

2.1 The extraction of keywords from text

Keyword extraction is the fundament of building topic representation model. Many scholars have made contribution to this field. Li et al. proposed a new multi-strategy keyword extraction method based on TF/IDF and a specification of keywords depended on analyzing linguistic characteristics of news documents [2]. Litvak et al. proposed supervised and unsupervised graph-based approaches for the cross-lingual keyword extraction, which was used in extractive summarization of text documents [3]. Christian et al. presented three statistical methods to improve keyword extraction that went beyond the use of TF-IDF. Unlike the classical TF-IDF measure, they took the relations and context of words into account by using the so called co-occurrence distribution. This led to an improvement over TF-IDF based ranking [4]. Chen has improved the efficiency of keyword extraction in a set

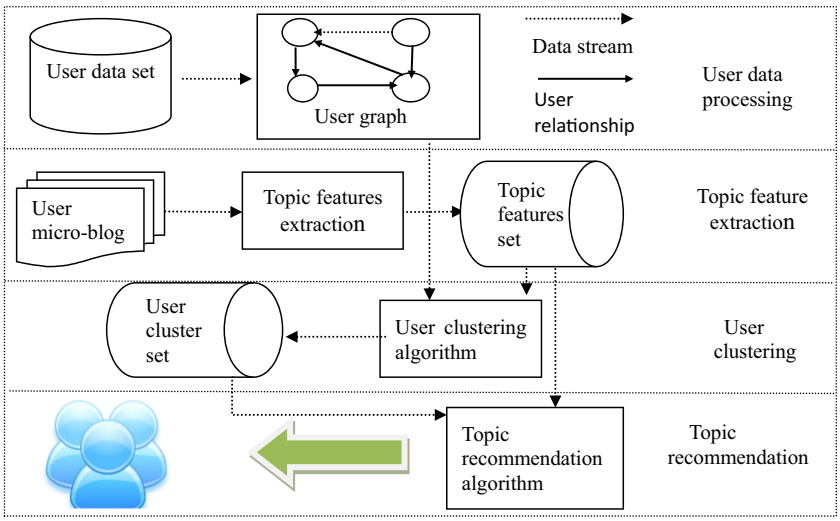

Fig. 1 The framework of micro-blog recommendation system

of relevant Chinese documents by using a new PAT-tree based approach [5]. Jiao et al. proposed a method of keyword extraction based on N-gram and word co-occurrence statistical analysis [6]. Zhang K et al. proposed the significance of keyword extraction with global context information and "local context information" on the basis of Support Vector Machines [7]. X Luo et al. a proposed a new method for semantic representation, which reduces the complexity of document analysis in an e-science Knowledge Grid effectively [8]. Xiangfeng Luo et al. proposed method considers a news event at a given time point as a system composed of different keywords. Based on keywords processing and other technical, Semantic Uncertainty is proposed to estimate the evolution potential of news events [9]. Junyu Xuan et al. propose an innovative graph topic model (GTM) to discovery latent topics among unstructured and structured data can be represented as graphs, such as documents and images [10]. Junyu Xuan et al. proposed IAT model to learn the hidden topics, authors' interests on these topics and the number of topics simultaneously [11]. Xiangfeng Luo et al. discuss two types of temporal features observed from the real time webpages covering an event. Based on that, an improved hidden Markov model is developed to predict the state transition of web events [12].

2.2 Recommendation system

A number of scholars conduct researches on the recommendation system. The work conducted by them can divide into the following categories:

First, on the basis of the behavior of users (e.g. comments or forwards), user interesting is got. Then, they use a model or a formula to recommend something to user by the help of user interesting. Reference [13] proposes a micro-blog system on the user interested in micro-blog's real-time recommendation based on the theme of the LDA model to infer the distribution of micro-blog's theme and the user's interest orientation. Reference [14] develops the use of a system to recommend the user by analyzing his or her interests, and using Conceptual Fuzzy Sets to expand a query. A system called Tweeter is proposed according to tweet published by users' fans and their friends, user relation in Twitter. What's more, the TF-IDF of Lucene is utilized to compute the weights of keywords [15]. Reference [16] proposed a recommendation algorithm that combines the contents of users' comments, reduce feature dimension of user recommends, calculate the similarity between the products, and then combine with the users scoring weights to get a comprehensive similarity, and finally recommend the products to the users may be interested in.

Second, based on the analysis of user historical information (e.g. micro-blog), user interesting is got. Then, they use clustering or other method to find similar interesting user to recommend something to them [17]. In the reference [18], a mobile micro-blog information recommendation method based on topic (model correlated) is proposed, and a visual mobile information recommendation system is designed based on this method. The reference [19] built a system in the forms of the named entity and core items by the help of Natural Language Processing tools. Reference [20] proposes a prediction algorithm to estimate the parameters of the probabilistic model, and use MapReduce to deal with large scale data. Reference [21] proposes a recommendation algorithm GCCR based on two stages clustering, which combines the graph summary method and the content-based similarity algorithm to achieve the purpose of the topic recommendation based on the users' interests.

Third, based on the handle of web resource, they use some new method/model to reorganize loose web resource and recommend user's interested resource to user. The reference [22] presents an Association Learning Model (ALM) which can efficiently provide association learning of Web resources in breadth or depth for learners. The reference [23] presents an application using C-ALN to organize Web services, which shows that C-ALN is an effective and efficient tool for building semantic link on the resources of Web services. The reference [24] develops a SP-based Webpage recommendation system which can significantly capture the different levels of the semantic uncertainties of Web events and it can be applied to Webpage recommendations. The reference [25] develops a SP-based Webpage recommendation system which can significantly capture the different levels of the semantic uncertainties of Web events and it can be applied to Webpage recommendations. The reference [26] proposed a method which can automatically generate the context and does not need any prior knowledge such as ontology or a hierarchical knowledge base. The reference [27] presented a hybrid image recommender system, which combines collaborative filtering (social techniques) with content-based techniques, leaving the user the liberty to give these processes a personal weight, improving the performance of existing systems to create a mobile social networks recommender with a high degree of adaptation to any kind of user. Reference [17] proposes a cloud-assisted approach for enriching end-user Quality of Experience (QoE) of drug recommendation, by modeling and representing the relationship of the user, symptom and medicine via tensor decomposition.

2.3 Other technical

Web ranking is one of the key technologies of search engine, the PR value got by PageRank Algorithm can reflect the influence of a web page, Li Z Y et al. make a summary of

PageRank Algorithm and analyzes the basic ideas and technical characteristics of various improved algorithms, aiming at the problems such as the theme drift, the emphasis on old web pages, and so on [28]. Meng F R et al. proposes information entropy based algorithm for efficient sub graph matching. Experiments show that the proposed method has a higher efficiency of inquiries. And in the long-tailed degree distributions of dataset, the effect is more apparent [29]. Pirasteh P et al. introduce new weighting schemes that allow us to consider new features in finding similarities between users These weighting schemes, first, transform symmetric similarity to asymmetric similarity by considering the number of ratings given by users on non-common items. Second, they take into account the habit effects of users are regarded on rating items by measuring the proximity of the number of repetitions for each rate on common rated items. The Experiments results show that adding weighted schemes to traditional similarity measures significantly improve the results obtained from traditional similarity measures [30].

3 Building micro-blog user graph

3.1 Basic conceptions

Definition 1: User Graph User Graph is a kind of network which reflects the complex relations among all users. Relationships between users are only in two situations. One is active following, the other is passive followed by others in the micro-blog. Relationship between any two users can be unidirectional. Namely, the user A follows the user B, user B does not follow the user A. Relationships between users can also be bidirectional, namely the user A follows the user B, user B follows user A at the same time. If users are regarded as the nodes in the graph, then the relations between users are the edge in this picture, and the edge in the graph are directed. In this way, the users and the relationships between them can construct a directed graph, we define the graph as user graph G. The user graph G can be described as follows:

$$G = \{(V_i, E_j)|i = 1, 2, 3\ldots k, j = 1, 2, 3\ldots n\} \quad (1)$$

Where V_i represents the node of the users. E_j represents the edge in the graph, that is, relationship between users. k denotes the number of the users, n represents the number of the relationship between users.

Definition 2: User Authority($\rho(U_i)$). User Authority is a kind of degree which can reflect a user's influence. For a micro-blog topic, each user's influence degree on the topic of micro-blog is different. For example, a micro-blog topic followed by user with several million fans is easy to become a hot topic, and a micro-blog topic followed by user with several fans is less likely to be a hot topic. On the searching engine, the PR value computed by PageRank algorithm can reflect the importance of a web page. This paper considers a micro-blog user as a web page, according to the PageRank algorithm calculating the value of the page(PR). User Authority ($\rho(U_i)$) is used to reflect users' influence. If the topic followed by a high authority user, the chance of obtaining other users' following is much greater than the common topics. User authority $\rho(U_i)$ can be described as follows:

$$\rho(U_i) = \{(U_i, \rho)|i = 1, 2, 3\ldots k\} \quad (2)$$

Where U_i represents a single user, k is the number of the users, ρ is the value of the PR for the single user.

3.2 User graph

Usually, a micro-blog displays user basic information such as the user id, nickname, location, gender, sexual orientation, relationship status, birthday, blood type. Some basic information can be chosen and defined as a 5-tuple for each user, namely $U = \{U_{id}, U_{name}, U_{address}, U_{sex}, U_{age}\}$ Where U_{id} is micro-blog ID of user, U_{name} is micro-blog nickname of user, U_{address} is user's location, U_{sex} is the sex of user , U_{age} is the age of the user. Each tuple is regarded as a node. The users are linked by the relations between any of users. The relationship is that if the user A follows the user B, then user A, B can be connected. The user A is out-degree, the user B is in-degree. Relationship between any two users can be unidirectional or bidirectional. Then a directed graph G = (*V*, *E*) is constructed by nodes and links. The vertex V is the set of all micro-blog users, while edge E is the link between the users. The specific situation is shown in Fig. 2.

The circle represents a single user in the Fig. 2, and the lines represent the relations. The user A points to the user B means A follows the user B. The user A, B, C, D, E and the relationships between them build a directed graph.

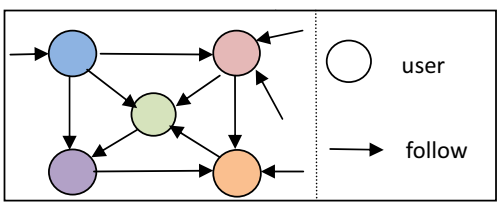

Fig. 2 A simple example of User Graph

3.3 The acquisition of the user authority

The relationships between users can be obtained according to the user graph G. The user authority ρ can be computed by the idea of PageRank algorithm based on these relationships.

The traditional PageRank algorithm is based on web link analysis for the search results of keywords [14]. It draws on the traditional citation analysis idea: when a web page A has a link to a web page B, B is considered as getting a contribution value from the A. The value depends on the importance of web page A itself. The more important of the web page A is, the obtained value of the web page B is higher. The calculation formula of PageRank is as follows:

$$PR(u) = (1-d) + d\sum_{v\in B(u)} \frac{PR(V)}{N_v} \tag{3}$$

Where, $B(u)$ represents the web page set that points to the web page u directly (u denotes the out-degree set; N_v indicates the number of the out-chain for web page v; $PR(v)/N_v$ refers to the value of the web page v which is average assigned to its out-chain web pages; d is the damping coefficient, whose general value is 0.85.

From the idea of PageRank algorithm, a micro-blog user is considered as a web page. The user B will get the contribution value from the user A when the user A follows the user B. The more people follow the user A, the more important the user A is, and the more contribution value user B gets. The value in the last iteration equals to the user's authority value.

The similar authority ρ calculation formula which we can get is as follows:

$$\mathrm{P}(u) = (1-d) + d\sum_{v\in W(u)} \frac{\rho(v)}{N_v} \tag{4}$$

Where $W(u)$ represents the user set that follow the user u directly; N_v denotes the number of the out-degree from the user v ; $\rho(v)/N_v$ indicates the authority value of the user v which is average assigned to the users who have followed him; d is the damping coefficient, the specific size obtained by experiments, tentatively scheduled for 0.85.

4 Building user topic feature matrix

In this section, topic feature graph and user micro-blog semantic view are defined to express micro-blog topic and user historical micro-blog. Then, the relationship between micro-blog topic and user are presented by user topic feature vectors. So that, user topic feature matrix UTFM is proposed that combines with the users and the topic features vectors. Besides, micro-Blog topic heat degree is represented to calculate topic heat degree.

4.1 Basic definition

Definition 3: Topic Feature Graph (TFG) Topic Feature Graph is type of graph which can reflect relationship between meaningful keywords in the topic. For each micro-blog topic, many high-frequency words will appear in discussions in the user micro-blog. These high-frequency words used a certain conditions are called feature words of micro-blog topic. Firstly, all micro-blogs related to a given topic are segmented by NLPIR. Then we extract meaningful words from segmented words as feature keywords. After that we calculate weight of each keyword based on TF-IDF. Afterwards the relationship between any two keywords will be mined according to association rule. Finally, each keyword regarded as a node and relationship between keywords regarded as edge will be built as a network TFG. TFG is defined as follows:

$$TFG = \left\{T_{tp}, E_{tp}, W_{tp}\right\} \tag{5}$$

Where each item of feature keyword set T_{tp} is the node of the TFG, each item of the relationship set E_{tp} is the edge of the TFG, each item of the weight set W_{tp} represents each weight on every edge.

Definition 4: User Micro-blog Semantic View (UMSV) Topic Feature Graph is type of graph which can reflect relationship between meaningful keywords in the user historical micro-blog. Firstly, all micro-blogs belonged to a user are segmented by NLPIR. Then we extract meaningful words from segmented words as feature keywords. After that, we calculate weight of each keyword based on TF-IDF. Afterwards the relationship between any two keywords will be mined according to association rule. Finally, each keyword regarded as a node and relationship between keywords regarded as edge will be built as a network UMSV for each user. UMSV is defined as follows:

$$\mathrm{UMSV} = \{T_x, E_x, W_x\} \tag{6}$$

Where each item of feature keyword set T_x represents the node of the UMSV, each item of the relationship set $E_x = \{e_{12} \ldots e_{ij} \ldots\}$ stands for the edge of the TFG, each item of the weight set $W_x = \{w_{12} \ldots w_{ij} \ldots\}$ represents each weight on each edge.

Definition 5: User Topic Feature Vector (UTFV) User Topic Feature Vector is a vector which can record user interesting in topics. All topics that a user is interested in will construct a vector. In this paper, UTFV is described as follows:

$$\mathrm{UTFV} = \{t_1, t_2, \cdots, t_m\}(i\geq 0, m\geq 0) \tag{7}$$

Where, t_i denotes the symbol that the topic is followed. When the topic is followed, the value of the t_i is 1. Otherwise, its value is 0. m is the number of the topics which are followed by users.

Definition 6: Micro-Blog Topic Heat Degree (φ). Micro-Blog Topic Heat Degree is a kind of degree which can embody heat degree of a topic. For a micro-blog topic, when the heat of a topic is high, from another perspective, when the micro-blog topic becomes hot, the probability of other users following it will be greatly increased. In this paper, we define a micro-blog topic heat as φ, the topic of micro-blog heat φ can be described as follows:

$$\varphi = R/100000 \tag{8}$$

Where, R denotes the current topic of search results returned by micro-blog search results, the search results of a topic can have very obvious reaction of the current heat of a topic. It impacts much for the future topic recommendation.

Definition 7: User Topic Feature Matrix (UTFM). User Topic Feature Matrix is a type of matrix which can incarnate users' interesting. According to definition 5, we get the user topic feature vector UTFV. Let n denotes the number of users. Obviously, n-topic feature vectors can constitute user topic matrix UTFM. The length of single user topic feature vector is m, the size of user topics eigenvectors UTFM is $n * m$, the topic of user feature vector can be described as follows:

$$\text{UTFM} = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1m} \\ t_{21} & t_{22} & \dots & t_{2m} \\ \dots & \dots & \dots & \dots \\ t_{n1} & t_{n2}{}^{n2} & \dots & t_{nm} \end{bmatrix} \tag{9}$$

Where, t_{ij} denotes the symbol that the topic is followed. When the topic is followed, the value of the t_{ij} is 1. Otherwise, its value is 0. m is the number of the topics that be followed by all users. n is the number of users.

Table 1 gives some examples of the topic matrix.

4.2 Acquisition of user topic feature matrix

For the micro-blogs from the same topic, we segment it and extract feature keywords from segmented words, then calculate the weight of each keyword based on TF-IDF, and then through association rule mining relationship between each keyword, build each topic as a network TFG. Finally, we will acquire a set of TFG. For the historical micro-blog from a single user, we do the same things on them as what we do on micro-blogs. Then user historical micro-blog will be built as a network UMSV. Finally we will acquire a set of UMSV.

Table 1. A simple example of user topic feature matrix

User/topic	Kobe retired	Curry MVP
User A	1	0
User B	0	0
User C	1	1

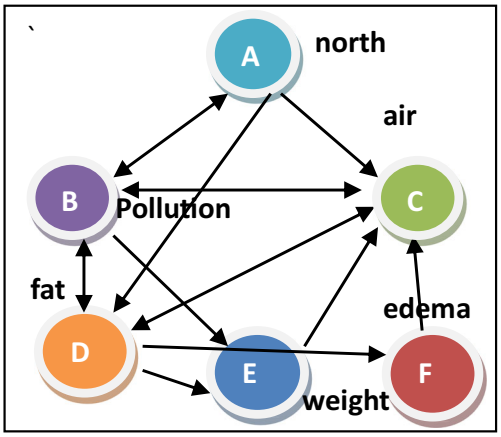

Fig. 3 Topic feature graph

Figure 3 shows an example of TFG. This example is a topic on the micro-blog named "Pollution can make people fat". Node A represents "north ", node b represents "pollution", node C,D,E,F respective represents "air", "far" ,"weight","edema". All the nodes are the keywords acquired by TF-IDF based on words above. Through the Association Rule Mining we get node A,B,D,E,F to connect Node C. Node A,C,D have a connection to Node B " pollution" and so on. The node and the relationship between any of them built an example of TFG.

Figure 4 shows an example for UMSV, this example is part of User Micro-blog semantic. After segment words to user historical micro-blog, we get numbers of words. We will get keywords through TF-IDF. So node A "fat", node B

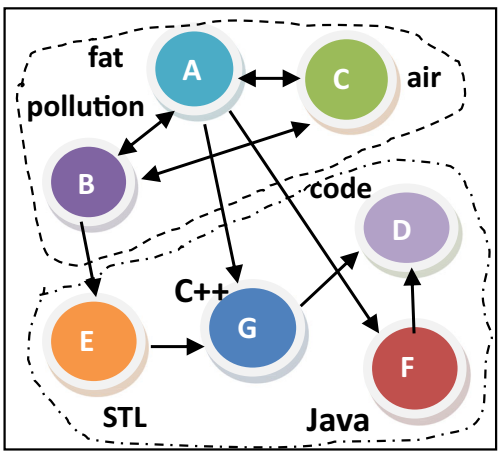

Fig. 4 User micro-blog semantic

"pollution", node C "air", node D "code" and so on are got. Through the he association rule mining we acquire the relationship between each keyword. From the Fig. 4,it is easy to see this example contains two topics, that are, "Air pollution can lead to fat" and "two main programming languages".

Based on TFG for each topic and UMSV for each user, this paper calculates matching degree for each TFG and each UMSV. When matching degree is greater than the threshold ∈, we believe that the user is interested in this topic.

Algorithm 1 User Topic Feature Matrix Generation Algorithm.

Algorithm 1 User Topic Feature Matrix Generation Algorithm

Input: topic feature graph set $TFG\_Set = \{TFG_1, TFG_2, \ldots, TFG_i, \ldots, TFG_m\}$, the number of topics m, the number of users n, the User Micro-blog semantic view list set $UFSV\_Set = \{UFSV_1, UFSV_2, \ldots, UFSV_i, \ldots, UFSV_n\}$

Output:User topic characteristic matrix UTFM

```
1：int UTFM[n][m]
2：i=0,j=0
3 ：while(i<n)
4：        while(j<m)
5:                UTFM[i][j] = 0
6：              j++;
7:          endwhile;
8：        i++;
9：endwhile
10: i=0;
11: int k
12：for UFSV in UFSV_Set
13：       for TFG in TFG_Set
14:           k = Match(TFG, UFMG);
15              if k>∈
16：              UTFM[i][j] = 1;
17              endif
18:         j++
19:       endfor
21：    i++;
22：endfor
23：end
```

In this algorithm, step 1 builds user topic characteristic matrix UTFM. Step 3 to step 9 are used to initialize the user topic characteristic matrix UTFM. Step 14 is used to acquire the marching degree for each UTFM and each UMTF. Steps 14 to 17 are used to realize matching. If the matching degree is greater than ϵ, then we set the value of this topic as 1, which means the user is interested in this topic. Steps 12 to 18 are used to find each user's followed topic and we can set its value as 1.

The complexity of this algorithm is mainly constituted by two double loops, the time complexity of first loop (steps 3 to 9) is O (n * m) and a time complexity of second loop (step 12 to 22) is O (n*m*m), the time complexity the algorithm is O (n*m*m). The algorithm is mainly used for the n*m two-dimensional matrix to storage UTFM, so the space complexity of the algorithm is O (n*m).

5 The clustering of micro-blog user

According to the topic of user characteristics matrix UTFM obtained above, we can cluster users to obtain user clusters of similar topics interest. Users within a user cluster may follow

the similar topic. Then the next step is to recommend the topic to users based on that.

Condensation hierarchical clustering algorithm is a bottom-up strategy. Since, first, each object is a cluster; second, merge those clusters until conditions are satisfied or only one cluster is remained. There are three types of hierarchical clustering, the single chain, the whole chain and the group average, this paper adopts the whole chain to calculate proximity. User Topic Feature Matrix (UTFM) is as input.

Cosine distance is used as distance between any two different users. Cosine distance is more to distinguish from the direction difference, and it is not sensitive to the absolute value. Thus, it is more content for calculating user interest similarity, and fixes the issues about user metrics are not uniform.

$$\mathrm{dist}\{A,B\} = \cos\frac{< A,B >}{\left|A\right| * \left|B\right|} \tag{10}$$

Then the user clustering algorithm based on condensation hierarchical clustering algorithm is as follows.

Algorithm 2: Micro-blog User clustering algorithm based on condensation hierarchical clustering.

Algorithm 2:Micro-blog User clustering algorithm based on condensation hierarchical clustering

Input:User topic characteristic matrix UTFM
Output:User clustering cluster C
1: int i=0;
2:if C=∅
3：for UTFV in UTFM
4：{ UTFV = UTFM[i]
6: { C_i =UTFV
7: { i++;
8: endfor
9: calculation proximity among each cluster in C, the proximity can form a Matrix M
10:else
11:repeat
12: selected max ($dist\ (U_i,\ U_j)$), $U_i = U_i \cup U_j,\ C = \{C_1, \ldots,\ C_{n-1}\}$.
13: update proximity matrix M,
14:until to reach the number of clusters α
15:end

In this algorithm, step 1 to 8 considers each user as a cluster. Step 10 calculates the similarity between each pair of user and the initial similarity matrix. Step 12 selects max distance between any two of users and merge selected user, then the number of cluster will reduce. Step 14 updates the similarity matrix until the number of clusters is reached α.

The algorithm uses adjacency matrix, which is assumed to be symmetric, and there are m users, so total space complex degree is O (m^2). The time complexity of Step 4 is O (m^2), step 5 of the time complexity is O (m) all the same, the total of the time complexity is O (m^3), if use an ordered list for store cluster, time complex degree can be reduced to O($m^2 logm$).

6 Micro-blog topic recommendation

Combining with the user topic feature matrix UTFM and user clustering result set, we can design micro-blog topic recommendation algorithm. The algorithm idea mainly considers two aspects, similar interest in a same user cluster and topic influence. In the first aspects, the users' topic interest is similar within the same cluster. And in the second aspect, user recommendation index filters the lower influence topic. Thus, this algorithm recommends the topic simultaneously meted two conditions. First, it is recommended by the algorithm that one user follows to other users who do not follow within a cluster. Second, this algorithm recommends a topic whose user recommendation index reaches a certain value.

6.1 Basic concepts

Definition 7: Topic Recommendation Index (λ) Topic Recommendation Index is an index determines recommended level for a topic. For a micro-blog topic, how to determine whether it is worth to be recommended to the user? In previous papers, the User Authority Degree and the Micro-Blog Topic Heat Degree was defined, then, Micro-blog topic recommended index is composed of User Authority and the Micro-Blog Topic Heat Degree. Higher micro-blog topic recommendation Index may bring to the rate increasing that users are interested in. Micro-blog topic recommendation index λ is described as the following form:

$$\lambda = a\rho + b\phi \tag{11}$$

Here a, b are parameters and experiment. According to statistics, among every 10 users, six users will be interested in to a topic that is hot and four users will be interested in a topic that is followed by famous star. So we defined that a values 0.4 and b values 0.6. ρ is the User Authority, ϕ is the Micro-Blog Topic Heat Degree. Defining a threshold χ, it will be recommended only if λ is greater than χ.

6.2 Micro-blog topic recommendation algorithm

The idea of micro-blog topic recommendation algorithm is that the user A and user B both belong to a cluster. Recommending the topic to user B that user A follows while user B does not. Because it belongs to the same user cluster, so the Micro-Blog Topic Heat Degree and User Authority are seen as two important variables. It can greatly improve the accuracy of the recommendation. From the perspective of micro blog, a high Micro-blog topic Heat Degree topic recommended by a user with a high User Authority (e.g., big V, a famous person in the country) is most possibly accepted by other users within a cluster. Recommendation algorithm is described in algorithm 3.

For each user cluster (corresponding to step 1 to 10), each user (correspond to step 2 to 8), each user follows (corresponding to step 3 to 7) topic, we calculated the Topic Recommendation Index, when it is greater than the threshold χ, and when the user follows symbol t = 1 (step 5), judging other users within user cluster whether they follow the topic (step 6) or not. If current topic is not followed by other users in the same cluster, then recommend it to other users. (user 6),

Assuming the number of user clusters is m , the number of users in a single user cluster is, the number of topics with a single user followed is p , the space of the algorithm requires the complexity of O ($m * n * p$), time complexity is also O ($m * n * p$).

Algorithm 3: Micro-blog topic recommendation algorithm.

Algorithm 3:Micro-blog topic recommendation algorithm

Input: User cluster set C-Set, $C\text{-}Set = \{C_1 \ldots C_m\}$

Output: User micro-blog topic recommendation

```
1: for the ith user cluster C_i in C-Set
2:   for the jth user u_j in C_i
3:     for topic t in u_j
4:       calculate Topic Recommendation Index λ of the topic
5:       if λ − χ >0 and t =1
6:           if current topic t not in other user in cluster ,recommend it to other
               users
7:       endif;
8:     endfor;
9:   endfor
10: endfor
11:end
```

7 Experimental analysis

In this section, we do two experiments to validate our work. In the first part of this section, we did a cluster experiment to see clustering effect. Then, in the second part of this section, we tested the effect of the recommendation system. The experimental environment is Dell XPS 13 laptop, MySql5.0 and Python2.7. And also other 10 personal laptop computers.

7.1 Cluster experiment

Clustering effect had a great influence on the accuracy of the final recommendation system. We selected 100 μ-blog users and 50 topic features, through the matching result of TFG and UMSV, accessed to the UTFM to cluster. The experiment is divided into the following steps:

(1) **Establishing the TFG:** We select 50 crawled topics and all their corresponding micro-blog texts are saved in a table/list of topic. Than TFG is built based on the topic table/list.

(2) **Establishing the UMSV:** We select 100 Micro-blog users as the experimental data, crawling their historical micro-blog in a table of Weibo. Than UMSV is built based on the table of Weibo.

(3) **Establishing the TFM**:TFM is built based on TFG and UMSV. Through the matching result of TFG and UMSV, we can get TFM.

(4) **Clustering**: Through the TFM acquired in the step 3, we can cluster user interested. The experiment result is showed in the Fig. 5.

We used silhouette coefficient to evaluate clustering effect. Silhouette coefficient is described as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{12}$$

For each datum i, a(i) is the average dissimilarity of i with all other data within the same cluster, b(i) is the lowest average dissimilarity of i to any other cluster.

From Fig. 5, the number of clusters is a significant factor of clustering result. The clustering results of different data sets are not the same. In our data set, when the number of cluster takes 5, the silhouette coefficient reaches highest. The silhouette coefficient is rising when the number of cluster range from 2 to 5. And the silhouette coefficient is declining when the number of cluster range from 5 to 8. When the number of clusters is 2, the silhouette coefficient is low, which shows that the clustering effect is not good. The reason for this is that the user's interest cluster is too large and it cannot display different user crowd interest accurately. When user clustering number is from 2 to5, the silhouette coefficient gradually increased to 5 to reach the highest. It shows when the number of clustering reaches 5, the interest similarity between any two of users in the cluster becomes highest. When the number of clusters is from 5 to 8, the silhouette coefficient is gradually reduced, which indicates that the clustering is too fine, and some similar interesting users are divided into different cluster.

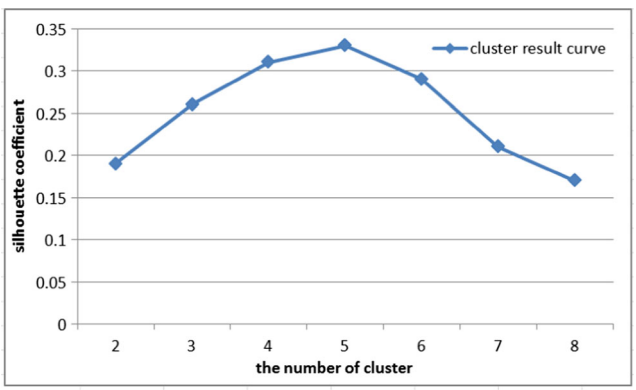

Fig. 5 The analysis of clustering result

7.2 Recommendation experiment

In this paper, the accuracy of the information retrieval system, $P = |C \cap R|/|R|$,is used to evaluate the effectiveness of the system. Wherein *R* is the number of topics recommended by the algorithm, *C* for topic number that users are interested in. In short, it is the ratio of the number of items that user interested and the total number of algorithm recommendation. The micro-blog recommendation system is different from other system, we must collect the users' feedback to judge the accuracy of the system recommended, and the experiment is divided into the following steps:

(1) Algorithm Operation

In the clustering experiment, we get the information about 100 μ-blog users. Then, we can crawl their relationship in the micro-blog and save them in the table usereation. After that we establishing the network diagram with table userelation, the PageRank computes the authority of each user with their historical micro-blog by extracting users' interest in the topic, and the establishment of user topic feature matrix TFM, the user clustering gain of 10 users were recommended by the topic.

(2) Collection of Statistical Results.

After Micro blog recommendation system has been operated, the recommended results can be concluded. We set the number of clustering is 5 and recommended to 10 users through the micro-blog automatic private letter by the help of algorithm. 10 users' feedback is shown in Table 2.

From Table 2, the number of topics recommended by user are different, but precision has been stable, which is near 0.502. To made a good recommendation effect, recommended index parameters *a* and *b* on the experimental results has played a key role

Table 2 User feedback form

user	Number of recommendation	Number of interest	accuracy
A	70	36	0.51
B	90	48	0.53
C	85	41	0.48
D	50	22	0.44
E	46	25	0.54
F	110	58	0.52
G	60	28	0.46
H	85	47	0.55
I	75	36	0.48
J	56	29	0.51

on the accuracy of recommendation. By adjusting these two factors, the algorithm can achieve better results.

(3) The Result with Different Clustering Number

It has been noted that the number of different clusters will lead to different clustering results. However, does the different number of clusters have an impact on the recommended result? The next experiment will illustrate this point. We also select the same number of 10 recommended users and 2–8 clusters. The experimental results are shown in Fig. 6.

In Fig. 6, the Y axis is the average of the accuracy of the 10 recommended user feedbacks. The X axis is the number of clusters. From the figure can be seen when the cluster number is from 2 to 5, the average accuracy of recommendation is gradually increased and at 5 to the highest. From 5 to 8, the recommended average accuracy decreases gradually. When the number of clusters is 2, the cluster is too large and the user's interest similarity is not high between any two of users in cluster, so, the result is not accurate. When the number of clusters increases from 2 to 5, the average accuracy is increasing and reaches highest at 5. This is due to the same interest users are suitable to gather together. When the number of user clustering is from 5 to 8, the accuracy rate is lower. Because the clustering is too fine, resulting in some of the same interest users did not gather in.

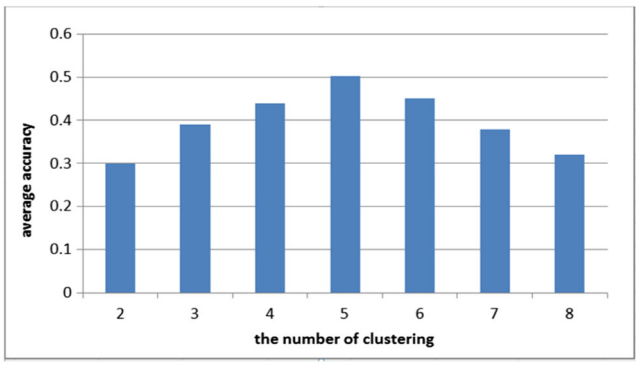

Fig. 6 The analysis of recommendation result

8 Conclusions

Micro-blog has become an important crowdsensing place for a lot of real-time information dissemination and discussion. At the same time, it brings the problem, overloading information, which is faced by micro-blog users (it is difficult for users can obtain their own interest to micro-blog topic with less time and effort). To solve this problem, we have designed a micro-blog recommendation system based on user clustering. Our contributions mainly include the following three aspects.

(1) The user graph of micro-blog has been constructed effectively. The user graph of micro-blog is a direct weighted graph whose nodes are micro-blog users and edges are the follow hyperlink from source user to destination user. And then an algorithm of computing user authority (which is similar to the idea of PageRank) is proposed to catch influential users based on the built user graph. If a user has followed another user, it can be considered as out-degree, and if a user that has been followed will be considered as in-degree.
(2) Topic Feature Graph (TFG) and User Micro-blog Feature Graph (UMFG) have been respectively constructed based on the micro-blog text corpus of a topic and the micro-blog texts followed by a given user. Further, User Topic Feature Vector (UTFV) and User Topic Feature Matrix (UTFM) can be achieved by matching UMFG with TFG. Segment words form is passed to the micro-blog users by extracting the feature of segmentation result, creating a user topic feature matrix.
(3) The algorithm of users clustering has been proposed for topic recommendation. The users' similarity among any two users is calculated based on the User Topic Feature Vector and User Topic Feature Matrix to realize by the help of the hierarchical clustering algorithm. The micro-blog topic recommendation index λ is introduced by combining the users' authority with the user topic feature matrix ρ and user clustering.

The experimental results show that the proposed recommendation system of micro-blog topic based on user clustering is accurate in some extend. It is effective and feasible in micro-blog topic recommendation for users.

In the future, based on our proposed recommendation system of micro-blog topic, we will do some related researches such as the tracking, recognition and credibility analysis of micro-blog topics.

Acknowledgement This paper is the extended version of the conference paper of MOBIMEDIA 2016.

This work was supported by the Natural Science Foundation of Anhui Province Universities (No. KJ2015A111), in part by the National Science and Technology Major Project under Grant 2013ZX01033002-003, in part by the National Science Foundation of China under Grant 61300202, and in part by the Science Foundation of Shanghai under Grant 13ZR1452900.

References

1. Mu FN (2013) Research on recommendation diversity for micro-blog users [D]. Harbin Institute of Technology, Harbin
2. Li J, Fan Q, Zhang K (2007) Keyword extraction based on tf/idf for Chinese news document. Wuhan Univ J Nat Sci 12(5):917–921
3. Litvak M, Last M (2008) Graph-based keyword extraction for single-document summarization[C] proceedings of the workshop on multi-source multilingual information extraction and summarization. Association for Computational Linguistics, Columbus, pp. 17–24
4. Wartena C, Brussee R, Slakhorst W (2010) Keyword extraction using word co-occurrence. In: Proceedings of the 2010 Workshops on Database and Expert Systems Applications. IEEE Computer Society, Bilbao, pp. 54–58
5. Chien LF (1989) PAT-tree-based keyword extraction for Chinese information retrieval. In: Machinery, ACM SIGIR Forum, Association for Computing, pp. 221–222
6. Jiao H, Liu Q, Jia HB (2007) Chinese keyword extraction based on N-gram and word co-occurrence. In: Computational Intelligence and Security Workshops, 2007 (CISW 2007), pp. 152–155
7. Zhang K, Xu H, Tang J et al (2006) Keyword extraction using support vector machine. Lect Notes Comput Sci 4333(016):85–96
8. Luo XF, Fang N et al (2008) Semantic representation of scientific documents for the e-science knowledge grid. Concur Comput: Pract Exp 20:839–862
9. Luo XF, Xuan JY, Zhang GQ et al (2016) Measuring the semantic uncertainty of news events for evolution potential estimation. ACM Trans Inf Syst 34(4):1–25
10. Xuan JY, Jie L, Zhang GQ, Luo XF (2015) Topic model for graph mining. IEEE Trans Cybernetics 45(12):2792–2803
11. Xuan JY, Lu J, Zhang GQ et al (2015) Infinite author topic model through mixed Gamma negative binomial processes. IEEE International Conference on Data Mining (ICDM 2015), Atlantic City, pp. 489–498
12. Luo XF, Xuan JY, Liu HM (2014) Web event state prediction model: combining prior knowledge with real time data. J Web Eng 13(5&6):507–524
13. Gao M, Jin CQ, Qian Q et al (2014) The real-time personalized recommendation for the micro-blog system [J]. J Comput Sci 04: 963–975
14. Sakaguchi T, Akaho Y, Takagi T et al (2010) Recommendations in twitter using conceptual fuzzy sets[C]. Fuzzy information processing society (NAFIPS), 2010 annual meeting of the north American. IEEE, Toronto, pp. 1–6
15. Hannon J, Bennett M, Smyth B (2010) Recommending twitter users to follow using content and collaborative filtering approaches[C]. ACM conference on recommender systems. ACM, Barcelona, pp. 199–206
16. Liu Y Personalized product recommendation system based on user comments[D]. Beijing University of Posts and Telecommunications
17. Zhang Y, Zhang D, Hassan MM et al (2015) CADRE: cloud-assisted drug recommendation Service for Online Pharmacies[J]. Mob Net Appl 20(3):348–355
18. Song SY, Li QD (2011) A method of information recommendation for mobile terminals based on mobile terminals [J]. Comput Therm Sci 38(11):137–139
19. Scott P, Jon W (2011) A feasibility study on extracting twitter users' interests using NLP tools for serendipitous connections[C]. Proceedings of the 3rd IEEE International Conference on Social Computing (SocialCom-2011), Boston, pp. 910–915
20. Kim Y, Shim K (2011) TWITOBI: a recommendation system for twitter using probabilistic modeling.[C]. 2013 I.E. 13th international conference on data mining. IEEE, Dallas, pp. 340–349
21. Kehan C, Panpan H, Wu J (2013) Recommendation algorithm based on user clustering for heterogeneous social networks [J]. Chinese J Comp 36(2):349–359
22. Zhang SX, Luo XF, Xuan JY, Chen X, Xu WM (2014) Discovering small-world in association link networks for association learning. World Wide Web 17(2):229–254
23. Luo X-F, Xu Z, Yu J et al (2011) Building association link network for semantic link on web resources. IEEE Trans Automation Sci Eng 8(3):482–494
24. Xu Z et al (2015) R2 Semantic based representing and organizing surveillance big data using video structural description technology. J Sys Software 102:217–225
25. Xuan JY, Luo XF, Lu J et al (2016) Uncertainty analysis for the keyword system of web events. IEEE Transactions on Systems, Man and Cybernetics: Systems 46(6):829–842
26. Xu Z et al (2014) Generating temporal semantic context of concepts using web search engines. J Netw Comput Appl 43:42–55
27. Sanchez F, Barrilero M, Uribe S et al (2012) Social and content hybrid image recommender system for mobile social networks[J]. Mob Net Appl 17(6):782–795
28. Li ZY, Yang W et al Summary of PageRank algorithm [J]. Comput Therm Sci 2011(B10):185–188
29. Meng FR, Zhang Q, Yan QY (2012) Information entropy based algorithm for efficient subgraph matching[J]. Appl Res Comp 29(11):4035–4037
30. Pirasteh P, Hwang D, Jung JE (2014) Weighted similarity schemes for high scalability in user-based collaborative filtering[J]. Mob Net Appl 20(4):497–507