Mining User Interests from Information Sharing Behaviors in Social Media

Tingting Wang[1], Hongyan Liu[2,*], Jun He[1,*], and Xiaoyong Du[1]

[1] Key Labs of Data Engineering and Knowledge Engineering, Ministry of Education, China
School of Information, Renmin University of China, 100872, China
{wtt526,hejun,duyong}@ruc.edu.cn
[2] Department of Management Science and Engineering, Tsinghua University, 100084, China
hyliu@tsinghua.edu.cn

Abstract. Mining user interests and preference plays an important role for many applications such as information retrieval and recommender systems. This paper intends to study how to infer interests for new users and inactive users from social media. Although some recently proposed methods can mine user interests efficiently, these works cannot make full use of relationship between users in their social network. In this paper, we propose a novel approach to infer interests of new users or inactive users based on social connections between users. A random-walk based mutual reinforcement model combining both text and link information is developed in the approach. More importantly, we compare the contribution of different social connections such as "follow", "retweet", "mention", and "comment" to interest sharing. Experiments conducted on real dataset show that our method is effective and outperforms existing algorithms, and different social connections have different impacts on mining user interests.

Keywords: Interest inferring, Social networks, Information sharing behaviors.

1 Introduction

With the advance of web technology, many micro-blogging websites are emerging rapidly. This kind of websites allows users not only to publish their views but also to share interests online. Twitter is one of the most famous micro-blogging services [14, 15], while in China Tencent Weibo is one of the largest micro-blogging websites, and it provides the similar social-networking functionality with Twitter.

A user registered in Tencent provides his profile information such as gender and date of birth, and publishes messages from time to time. Different people have different publishing frequency. Some are very active and some are very inactive. Besides, unlike other social network services that require users to send a request to make friends, another important part is that every user is allowed to "follow" the others without seeking any permission. In this case, the user who initiated this behavior is called "follower", while the one who is followed is called "friend". After a "follow"

* Corresponding authors.

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 85–98, 2013.

relationship is built, follower can read the messages published by his friends. Furthermore, follower can retweet a message or leave a comment to it if he or she is interested in the content. Unlike users registered in Twitter, users in Tencent can append some words when he retweets the message from others. Besides, a user is allowed to mention someone else when he publishes a message. To infer user interests, we use all of the above-mentioned information. For convenience, we call user's profile information and messages posted ***text*** information and social connection relationships regarding *follow*, *retweet*, *comment* and *mention* ***link*** information.

As we know, the number of micro-blogging users increased rapidly. A statistic from a newspaper shows that, the number of Tencent Weibo users has reached 310 million and the number of active users has reached 50 million up to November 2011. Furthermore, each day there are thousands of enthusiastic new users adding into the existing billions of actively engaged users. Although the large number of Tencent Weibo users benefit each other by information sharing, too much information results in information overload problem, which needs systems such as searching and recommender system to solve. Thus, it's really important to capture user interests and then provide personalized results and meet user's needs individually according to one's interests.

There are many good ways [1, 2, 3] to infer user's interests, however, most of them are not proposed for the social network. In these approaches user interest models are built according to the behavior history during web search, such as click-through web pages. For users in social network, the mainly behaviors are communication with the other users. Based on this difference, recently there are some methods proposed for social network. These algorithms mine user's interests through user contents such as micro-blog messages and user-generated tags. However, in social network there are some users preferring to read messages from other active users rather than publish some information about themselves. In this case, user-generated tags are not always representative of all of his interests, and only a small part of users provide tags. Especially for a user who just registered or the user who is not very active, both user-generated tags and micro-blog messages are difficult to get. Another kind of existing works inferred user interests from social neighbors to solve the problem of inactive users. However, this solution focuses on the three-degree ego network of a user and uses the information in a deterministic way. After inferring interests for one target user, its interests will not affect the target user's neighbors, which is unreasonable. In this paper, we emphasize mutual reinforcement between users through a process similar to random walk [16, 18].

Our proposed method is based on the homophily in social network. The phenomenon of homophily means that individuals with similar characteristics tend to associate with each other. Although homophily has been validated in different social networks such as twitter and student homepages [13, 17, 19], it isn't studied deeply in Tencent Weibo. As Tencent Weibo is mainly for Chinese users, and the culture and some information sharing mechanisms are different, it is necessary to study which kind of social connections reflect homophily and which one contributes more to interest sharing. To do this, before developing the interest inferring method, we conducted some statistical tests to study the relationship between social connections and similarity of

user interests (Detail is explained in Section 3.1). Existing work has studied semantics of the follow and retweet relationship in twitter [20]. In this paper, we studied more relationships such as comment, besides follow, retweet and mention.

According to this study, a conclusion that users with communication behaviors share more interests than those without can be made. Based on this finding, we propose a novel approach to infer user interests and we develop an algorithm to implement the approach. In this algorithm, first a directed graph to indicate potential interest propagation among users is constructed. And then text information is used to generate initial interests and link information is utilized to show how users affect each other in interests. Then, a mutual reinforcement process based on a random walk model is conducted to infer interests for new users and inactive users.

Our work offers two contributions.

First, we studied the relationship between social connections such as follow, retweet, comment and mention and common interests between users. Statistical study shows that different social behaviors have different influence on the interest similarity between users. We find that follow and retweet mean more strong connections for users in common interests than mention and comment.

Second, we propose a novel approach combining users' text information and link information (information about social connection) to infer interests for inactive users. In this approach, the mutual reinforcement between users is emphasized by a random walk model. Experiments show that this approach can improve the accuracy by up to 21.4%. Especially for inactive users, this approach can address the shortcomings of too little information.

The rest of the paper is organized as follows. In section 2, problem definition is given. The proposed algorithm is introduced in section 3. Experiment setup and results are described in section 4. In section 5, related work is discussed. Finally, conclusions are drawn in section 6.

2 Problem Definition

Let U be the set of users registered in a social media website such as Tencent WeiBo. Each user has a unique ID assigned by the system. In this paper, interest is defined as a pair of keyword and its weight about this keyword. For active users, keywords can be extracted from the text information of the user. Weight shows the favorite degree of the user to this keyword. The larger the weight is, the more the user likes the interest. One user can have one or more interests. Thus, a vector of pairs of keyword and weight is used to express the interests of users as shown in the definition below:

Definition 1. The interests of a user are expressed by a set of pairs of keyword and weight.

$$\{keyword_1{:}weight_1;\ keyword_2{:}weight_2;\ \ldots\ \ldots;\ keyword_l{:}\ weight_l\}$$

Example of interest information is shown in Table 1, where each integer represents a keyword, followed by weight. Further, the interests of a user a can be expressed by a vector I_a=<$weight_1$, $weight_2$, …, $weight_N$>, where N is the size of the union of all of user's interest keywords.

Table 1. Interest information

UserID	Interest Vector
10001	<101:0.4; 102:0.3; ·····
·····	·····
10005	<101:0.4; 103:0.3; ·····

Table 2. Follow relationship

UserID	Interest Vector
10001	<101:0.4; 102:0.3; ·····
·····	·····
10005	<101:0.4; 103:0.3; ·····

Table 3. Link information between users

User *a*	User *b*	RTnum	MEnum	CMnum
10001	10002	10	2	5
.....	.....	.....	.....	.....
10004	10007	3	0	0

The other kind of information is the link information between every pair of users. For users registered in Tencent Weibo, the basic behavior information between two users is the "follow" relationship. Besides, there are several other behaviors between two users, including "retweet (publishing other user's message)", "mention (mentioning other users when publish a message)", "comment (having a comment on other user's message)" and so on. These behaviors create links between users, which will be introduced in details in the next section. The *follow* relationship information is showed in Table 2. And the other behavior information, such as "retweet", "mention" and "comment" is given in Table 3, which shows the numbers of times of these different behaviors happened from user *a* to user *b*.

According to link information, a directed behavior graph $G(V, E)$ can be constructed to show the relationships among users. V is the node set which contains all the registered users. E is the edge set. Suppose a and b are two registered users. If user a has any action of follow, retweet, mention and comment to user b, an edge (a, b) is formed from user a to user b.

After constructing behavior graph $G(V, E)$, a directed graph $G'(V, E')$ called ***propagation*** graph is constructed to model how user interest propagates, in which a node is also on behalf of a user registered in the website. V is the same node set as that in graph G, and each edge (b, a) in E' corresponds to edge (a, b) in E. That is, if user a initiated an action such as "follow", "retweet", "comment" or "mention" to user b, there is an edge from the node b to node a. The direction of the edge is exactly opposite to the one in E. Because when an action is initiated from user a to user b, it reveals that user b's interests attract user a, and user b has some influence on user a about his interests. Thus, the interests should propagate from user b to user a. Besides, there is a weight assigned to this edge to indicate the influence on interests user b has on user a. And also an interest vector is assigned to each node according to the text information of users. However, first, not every user has this value, because some new registered users have little information published. And second, it's not easy to collect the interest information for every user especially for those inactive users who have little information. Thus, our mining task is to infer interests for these users in the network.

3 Interest Inferring Method

3.1 Hypothesis Tests

To infer user interests from his link relationship, several questions need to be answered to prove whether this approach is valid.

Questions 1, 2, 3 and **4:** Do users with "follower-friend", "retweet", "comment" or "mention" relationships in micro-blogging system in China have more similar interests than those without?

Besides these four questions directly related to the four kinds of link information we mentioned before, another factor indirectly related to the "follow" relationship is analyzed, that is the ratio of common friends of two users. In the next sections, this information is also discussed with the link information. Thus, another similar question is raised.

Question 5: Do users who have more common friends in Chinese micro-blogging system have more similar interests than those without?

To answer these questions, we give the definition of interest similarity of two users as follows:

Definition 2. Interest similarity of two users a and b can be measured by Equation 1.

$$ISim_{ab} = cos(v_a, v_b) \tag{1}$$

v_a and v_b are interest vectors of user a and b respectively, extracted from their text information.

Question 1 can be formalized as a two-sample t-test. Let u_{follow} be the mean interest similarity of the pairs of users with "follower-friend" relationship, while $u_{nofollow}$ the mean interest similarity of the pairs of users without. Let H_0 be the null hypothesis: $u_{follow}=u_{nofollow}$, and H_1 be the alternative hypothesis: $u_{follow}>u_{nofollow}$. Results show the null hypothesis is rejected at significant level $\alpha=0.01$ with a p-value of 3.14×10^{-5}. Question 2, 3, 4 and 5 are formalized as a two-sample t-test separately, too. Results show that the answers of Question 2, 3, and 4 are positive, and the null hypothesis is rejected at significant level $\alpha = 0.05$. To conduct a hypothesis test on Question 5, Equation 2 is used to measure the ratio of the common friends between two users.

$$cf_{ab} = \frac{|F_a \cap F_b|}{|F_a \cup F_b|} \tag{2}$$

F_a and F_b are the friend sets of user a and b separately. When selecting users who are used in hypothesis test, the users whose "common friends" measurement are larger than 0.8 are selected. Result shows that the null hypothesis is rejected at significant level $\alpha = 0.05$ with a p-value of 2×10^{-3}.

From these tests, we know that all the answers to these questions are positive, which shows that users who have these behaviors (follow, retweet, comment, mention and common friends) are more similar than users who don't. Based on this outcome, a novel approach to infer user interests is proposed in the next part.

3.2 Random Walk Based Inference Model

In this section, we focus on the problem of how to infer user interests after we construct the propagation graph. We will explain how to construct the graph later.

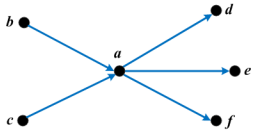

Fig. 1. User follow relationship

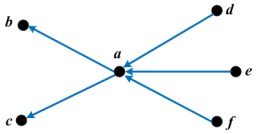

Fig. 2. User influence relationship on interests

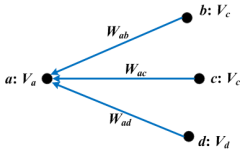

Fig. 3. One user and his in-degree neighbors

For one user in the social network, its local directed graph is shown in Fig. 1 to show its follow relationship.

And the directed graph to show its influence relationship in interests is shown in Fig. 2, right opposite to the direction in Fig. 1. In Fig. 1, user *a* follows user *d*, *e*, and *f*, and also is followed by user *b* and *c*. In Fig. 2, the direction of interest propagation is right opposite. Interests are propagated from user *d*, *e*, *f* to user *a*, while user *a* propagates interests to user *b* and *c*.

From Fig. 2, the interests of user *a* can be collected from two aspects. One is the text information of user *a*. For a user who has published some messages, an interest vector can extracted from the messages he published, retweeted or commented. On the other hand, based on the finding in the previous section, some information sharing behaviors indicate the common interests between users. Thus, interest information can also be inferred from those users who have link relationship with them. The interests are propagated to the user in a certain probability, which is expressed by weight of the corresponding in-degree edge. We denote this probability by w_{ij} on edge (u_j, u_i). For example, the probability of the interests propagating from user *d* to user *a* is denoted as w_{ad}, as shown in Fig. 3. Combining the two resources of information, for user *a*, his interests according to this method is inferred by Equation 3.

$$I_a = \alpha \cdot \sum_{i \in U} w_{ai} \cdot I_i + v_a \tag{3}$$

In Equation 3, U is the set of all users in the network, v_a is the interest vector of user *a* extracted from the text information, and I_a is user *a*'s overall interest vector considering both text and link information, and α is the decaying factor of influence from user's in-degree neighbors. The lower α is, the less influence a user will be got from his friends, and vice versa.

According to this formula, the interests can be computed recursively, because users influence each other during the information sharing behaviors. Thus, a random walk process is utilized to implement the mutual reinforcement between users.

Suppose the whole propagation matrix is denoted as P. P is a $|U| \times |U|$ matrix, where each entry is equal to w_{ij}, as we described above. All users' interest vectors are collected into a $|U| \times N$ matrix, v, where N is the total number of keywords. Each row v_i of the matrix v is the interest vector extracted from text information of user i. Then the interests of all users in network can be computed.

The interest matrix of the users, denoted as I, where row j represents user j's interest vector. Matrix I can be calculated iteratively by Equation 4. I_t is a $|U| \times N$ matrix and represents the interests after t times of iterations, $t > 0$. Initially, $I_0 = v$.

$$I_t = \alpha \cdot P \cdot I_{t-1} + v \tag{4}$$

According to the property of Markov chain, convergence is guaranteed if P is stochastic. In the next sections, we introduce how to compute the weight of the propagation graph and make sure that P satisfies this requirement.

3.3 Generating Interest Vector from Text Information

There are several methods to produce initial interest vectors for users. Usually, user-generated tags can be considered as a way to express user interests. However, most people add few tags in the system, thus other information, such as messages one posted, can be utilized. All tweets posted by one user can be collected as a document for the user. These tweets include not only the tweets and comments published by the user himself and also those retweeted from others. And for all users in the website, a set of documents can be collected and used. In this paper, a typical method, *Latent Dirichlet Allocation* (*LDA*) [5, 12] model is applied to these documents, which is an unsupervised machine learning technique to identify latent topic from large document collection.

3.4 Assigning Weights to Edges

In the social media websites like Tencent Weibo micro-blogging system, users can communicate with each other by retweet, comment and mention behaviors. According to these different communications, five different factors are defined to compute the weight of edge (*b, a*) in propagation graph.

Based on Retweet
We measure the influence of user b to user a based on the amount of user b's tweets retweeted by user a. The more tweets retweeted by user a, the more influence user b has to user a, and the more common interests occurs between user a and user b. Let RT_{ab} be the number of tweet retweeted by user a from user b. The weight is measured by Equation 5.

$$w_{rt} = \frac{RT_{ab}}{\sum_{i \in U} RT_{ai}} \tag{5}$$

Based on Comment
The number of comments which user a gives to user b measures the degree user a shows interests on the tweets of user b. Let CM_{ab} be the number of comments user a gives to user b. Then the weight of the edge from user b to user a is calculated like by Equation 6.

$$w_{cm} = \frac{CM_{ab}}{\sum_{i \in U} CM_{ai}} \tag{6}$$

Based on Mention
Mention action is another communications between two users. To some extent, the frequency of this action can show the influence user b has to user a. let ME_{ab} be the number of "mention" actions user a gives to user b, then we measure the weight according to Equation 7.

$$w_{me} = \frac{ME_{ab}}{\sum_{i \in U} ME_{ai}} \tag{7}$$

Based on Follow Relationship

The "follow" relationship is the basis and most usual action in the social network. Mostly, user *a* will follow user *b* if he is interested in the tweets posted by user *b* or user *b* himself. Thus, this kind of relationship can reflect the relationship between two users and their interests. f_{ab} is used to show whether user *a* follow user *b*. According to this, the weight of the edge from user *b* to user *a* is measured by Equation 8, where f_{ab}=1 if user *a* follows user *b*, otherwise, f_{ab}=0.

$$w_f = \frac{f_{ab}}{\sum_{i \in U} f_{ai}} \tag{8}$$

Based on Intersection of Friends

According to the "follow" relationships of one user, a list of his friends can be got. For user *a* and user *b*, the larger the set of intersection of their friends, the more interests they share. Let F_a be the set of the friends of user *a*, and F_b be the set of the friends of user *b*. Then the influence user *b* has on user *a* can be calculated according to this Equation 2. And then the weight on the edge (*b*, *a*) in the propagation graph is measured by Equation 9.

$$w_{cf} = \frac{cf_{ab}}{\sum_{i \in U} cf_{ai}} \tag{9}$$

Considering this will generate a matrix with so many non-zero numbers, we neglect those values which are smaller than 0.1. Through this process, the non-zero values are reduced to 422380, which makes the matrix sparse and improves the efficiency of iteration.

If the denominators in Equations 5, 6, 7, 8, 10 are zeros, 1/|*U*| is used to replace the formula. Combining these five factors, a comprehensive computation formula is proposed in Equation 10.

$$w_{ab} = \frac{w_{rt} \cdot RT_{ab} + w_{cm} \cdot CM_{ab} + w_{me} \cdot ME_{ab} + w_f \cdot f_{ab} + w_{cf} \cdot cf_{ab}}{\sum_{i \in U} (w_{rt} \cdot RT_{ai} + w_{cm} \cdot CM_{ai} + w_{me} \cdot ME_{ai} + w_f \cdot f_{ai} + w_{cf} \cdot cf_{ai})} \tag{10}$$

After these factors are defined, the propagation matrix on interests can be identified. From these computation formulas, each row in the propagation matrix is sum up to 1, which makes the propagation matrix stochastic. This makes sure the iteration process will converge.

4 Experiment

4.1 Dataset

A large dataset collected from Tencent weibo is provided by the Tencent Company. To test the proposed method, a relatively small network is extracted by a *BFS* algorithm. After this extraction process, the total number of users in *U* is 5238. For every pair of users in *U*, the corresponding information is also extracted, including the follow relationship, the number of "retweet", "comment", and "mention" actions. Table 4 shows some information of this dataset *U*. The distribution of the followers for each user is shown in Fig. 4.

Basically, this distribution of followers per user follows a power-law distribution approximately. That is, most people have small number of followers, while only a small of users have a large number of followers, which proves that the experiment data extracted is reasonable and representative.

Table 4. Basic information of dataset *U*

items	value
# of users	5,238
# of users in training	4,190
# of users in test set	1,048
# of follow relation-	133,825

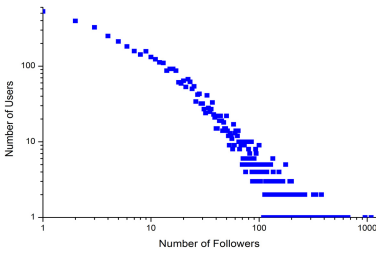

Fig. 4. Distribution of followers per user

For each user in the network, keywords information is extracted from messages posted. The number of distinct keywords for all 5238 users is 22376, and the average number of keywords which one user has is 29. A five-fold cross validation is conducted in this paper. We split these 5238 interest vectors into five parts. For every time, four parts of the vectors are used as training set, and the rest one is test set, whose interest vectors are regarded as truth. In the experiment, the text information of users in training set and the link information of users both in training set and test set are utilized to infer interests for users in test set.

4.2 Performance Comparison

We conduct experiments based on five different behavior factors, "retweet", "mention", "comment", "follow" and "common friends". Correspondingly, we denote these methods as "*RT*", "*ME*", "*CM*", "*Follow*" and "*Common Friends* (or *CF*)". In the next tables, these abbreviations will be used to show the comparisons. For each factor, a separate experiment is conducted to compare which factor performs better. To combine all of these factors, five weights need to be determined. We extract a part of data from training set and compute five *NDCG* values according to different factors separately. Based on the *NDCG* values, we set the five weights. The larger the *NDCG* value is, the larger the weight of the factor. In this paper, we set these weights, w_{rt}, w_{cm}, w_{me}, w_f, and w_{cf} as 0.31, 0.25, 0.25, 0.32, and 0.34 respectively. The method combining the five factors is called "*combination*". When determining whether to stop the iterations, the sum of absolute errors of each entry of the result matrix is used. In our experiment, this value is set to 0.1. When the sum of the absolute errors is smaller than 0.1, the iterations will stop and the method get final results for each user in the test set.

Comparison against related algorithm is also conducted. The work in [4] is one of the classic related works in inferring interests. In this paper, user interests are inferred from his social connections, that is, his friends, friends' friends and 3-degree friends. This method is called "3*D-Friends*" here. For each user, a 3-degree ego network is constructed to infer the interests for inactive users. These results are compared in Table 5.

Table 5. Comparison of different methods

Methods	NDCG
RT	0.3120
CM	0.2533
ME	0.2576
Follow	0.3215
Common Friends (*CF*)	0.3360
Combination	0.3493
3*D-Friends*	0.2878

Table 6. Num. of edges in propagation graphs

Factor	# of edges
RT	112039
CM	25526
ME	27018
Follow	133825
Common Friends	422380
Combination	511676

From Table 5 we observe that the factor of common friends has more significant impact than the other four factors. The method based on follow relationship also works better, with *NDCG* only less than the one based on common friends. However, the methods based on "comment" and "mention" don't work very well. The reason why the performance of these two methods is not very well will be studied later. Besides, the method based on five comprehensive factors outperforms all the other methods. Our best method increases the quality of interest inferring than the existing method, 3*D-Friends*, by 21.4%.

The number of edges of the propagation graph based on each factor is shown in Table 6. For a graph that has 5238 nodes, the total number of edges of complete graph is 5238×5238=27436644. When the factor of "*Common Friends*" is considered, the graph is complete with a large number of edges. We reduce the number of edge constructed based on "*Common Friends*" to improve the efficiency. If the weights computed based on "*Common Friends*" factor is smaller than 0.1, the corresponding edge is removed. Through this process, the non-zero values in the propagation matrix are reduced to 422380, which make the matrix sparse and improve the efficiency of computation. Accordingly, the number of edges of the "*Combination*" method is not very large, too. After this edge removal step, the number of edges is reduced to 511676, which makes the propagation matrix sparse, too. For the other four factors, the number of edges is small, especially for the factors of "comment" and "mention". The propagation matrix based on "retweet" or "follow" has more non-zero values than those based on "comment" and "mention". This tells us that users prefer to follow the others and retweet the tweets more than comment or mention others. Basically, based on the comparison we can conclude that different user behaviors have different impact on user interests, which is same with the conclusion with Adamic and Adar [19], that is, some factors are better indicators of social connections than others.

We also compare the efficiency of our method with 3*D-Friends*, which is illustrated in Fig. 5. From this Figure, we know that the time spent is proportional to the number of the non-zero values in the propagation matrix. The efficiency of our method based on mention, comment, retweet or follow is better than the method 3*D-Friends*. Time spent on common friend graph and the combination graph is more than 3*D-Friends*, because these two graphs are much denser than the graph 3*D-Friends* uses.

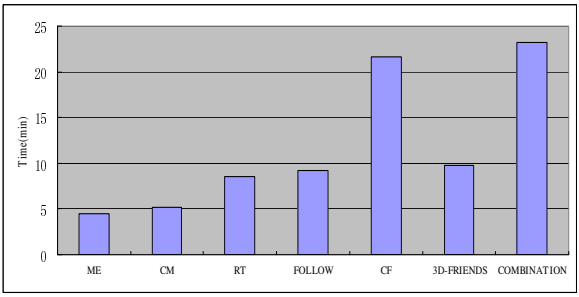

Fig. 5. Efficiency of different methods

4.3 Effect of Decay Factor

In this experiments described in the last section, we set the decaying factor α to 0.5. However, this decaying factor determines the important degree of the influence from a user's friends. The result of the algorithm will differ according to different decaying factor. Fig. 6 shows the changes of the results *NDCG* of all our six algorithms based on different decaying factor values.

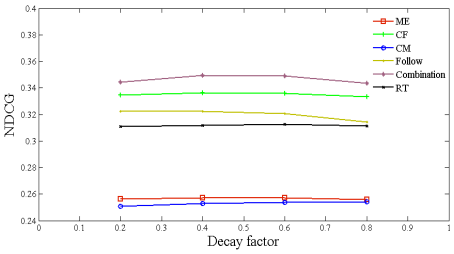

Fig. 6. Effect of α to *NDCG*

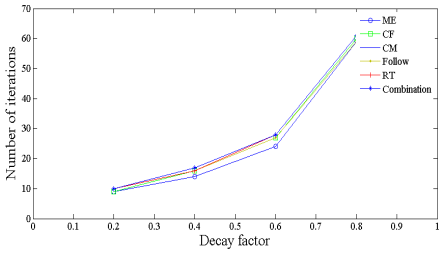

Fig. 7. Effect of α to number of iterations

From Fig. 6 we can observe that in most scenarios *NDCG* gets the best when α is between 0.4 to 0.6. However, the changes caused by different decaying factors are not very significant. No matter what value the decaying factor is, the basic trend among the results of the five methods remains similar.

The value of decaying factor has influence on not only the accuracy of the results, but also the efficiency of the algorithms. The number of iterations for different decaying factors is shown in Fig. 7. The larger the number of iterations is, the more time the corresponding method spends. From Fig. 7 we can see that the growth trends follow an exponential distribution. In our experiment, the decaying factor α is set to 0.5, and the number of iterations is about 21. When α is larger than 0.6, the run time increases a lot, and neighbor's influence becomes too heavy in the meantime.

5 Related Work

In this section, we briefly introduce related work. We category related work about user interest study into three groups: 1) based on user contents 2) based on user behavior 3) based on social cues.

Based on User Contents. Simply, explicit interests can be specified directly from users' profiles. In addition, other sources can also indicate users' interests. Some users prefer to use descriptive tags to express what they are interested in. Therefore, some researchers proposed approaches to find social interest based on user-generated tags [1, 9, 10]. However, these tags are not always representative, and some users don't like to add these tags to themselves. Although the tags extracted from user micro-blogs can replace the user-generated tags, it still cannot work well for new users and inactive users who have few micro-blogs.

Based on User Behaviors. Several algorithms have been proposed based on user behaviors during web search and browsing. Interests are captured from click through data or visited web pages [2, 6]. Qiu and Cho [7] focus on disambiguating the true intention of a query based on past click history. Kim and Chan [8] proposed a divisive hierarchical clustering algorithm to learn a user interest hierarchy from a set of web pages. These methods based on user behaviors, especially based on click-through history and web pages are mainly used to personalize user interests in the web search community. However, for users in social network, the click history is too sparse to be utilized to infer user interests. Most of users' behaviors are to share the messages from their friends.

Based on Social Cues. All of these above-mentioned methods use user individual information to infer interests. For the inactive users who have not many profiles and behaviors, these approaches cannot work well. Therefore, some researchers propose to infer user interests by leveraging social cues from other users. Similar with collaborative filtering systems, Glodberg et. al [3] proposed a method to mine the interests from the users who have similar opinions on a set of items. Their basic idea is that users who have similar behaviors will share similar interests. Similarly, White et.al [11] proposed a method to find a user's interests from other users that visit the same page as the user. In addition, another new approach is proposed by Wen and Lin [4]. It focuses on social cues from user's neighbors. In this work, for one target user, the neighbors in his three-degree ego network are considered. That is, for each user, a 3-degree ego network is constructed to infer interests. Relationships between users are built based on electronic communication data such email and instant messaging and Web2.0 social content such as social bookmarks and file sharing. The interests of active users in the network are extracted by LDA model from text information. Then inactive users' interests are computed based on their neighbors in a deterministic way, without considering user's mutual influence. Besides, Welch et al [20] demonstrate that in Twitter platform retweeting is a better indicator of topical interest than following behavior through the PageRank algorithm.

In our study, we focus on leveraging the social network to infer user interests. The contents from active users are considered as initial interests. These interests are propagated through the social network, which is built according to the interest similarity between users. This approach has important differences from the above-mentioned existing work in two aspects. First, unlike the existing methods based on user content and behavior, our proposed approach works well especially for users who have no text information. Second, different from the methods based on social cues, we consider the social connections and emphasize the mutual reinforcement among users, instead of directly inferring from a couple of friends or similar users.

6 Conclusions and Future Work

In this paper, we propose a novel approach for user interests inferring especially for new users and inactive users who have few messages published. In this approach, a random walk on a propagation graph model is used to emphasize the mutual reinforcement between users. When constructing the propagation graph, both text information and link information of users are taken into account. Besides, we prove by statistical tests that information sharing behaviors such as follow, retweet, comment and mention are related to the common interests between users. And the experimental results conducted on real social network data set show that different kind of social connections have different influence to common interests. Experimental results demonstrate that our methods get a better performance not only in the quality but also in efficiency. In the future, we will utilize the approach to provide better results for recommender system in social network.

Acknowledgement. This work was supported by the 973 program of China under Grant No. 2012CB316205 and the NSF of China under Grant No. 71272029, 70890083, 71110107027 and 61033010.

References

1. Li, X., Guo, L., Zhao, Y.E.: Tag-based social interest discovery. In: Proc. of the 17th International Conference on World Wide Web, pp. 675–684. ACM, New York (2008)
2. Agichtein, E., Brill, E., Dumais, S.T.: Improving web search ranking by incorporating user behavior information. In: Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 19–26. ACM, New York (2006)
3. Glodberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Communications of the ACM-Special Issue on Information Filtering 35(12), 61–70 (1992)
4. Wen, Z., Lin, C.-Y.: On the quality of inferring interests from social neighbors. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2010, pp. 373–382. ACM, New York (2010)
5. Blei, D.M., Nq, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
6. Claypool, M., Brown, D., Le, P., Waseda, M.: Inferring user Interest. IEEE Internet Computing, 32–39 (2001)
7. Qiu, F., Cho, J.: Automatic identification of user interest for personalized search. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006, pp. 727–736. ACM, New York (2006)
8. Kim, H.R., Chan, P.K.: Learning implicit user interest hierarchy for context in personalization. In: Proceedings of the 8th International Conference on Intelligent User Interfaces, IUI 2003, pp. 101–108. ACM, New York (2003)
9. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.D.: Personalized recommendation in social tagging systems using hierarchical clustering. In: Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, pp. 259–266. ACM, New York (2008)

10. Stoyanovich, J., Amer-Yahia, S., Markow, C., Yu, C.: Leveraging tagging to model user interests in del.icio.us. In: AAAI Spring Symposium on Social Information Processing (2008)
11. White, R.W., Bailey, P., Chen, L.: Predicting user interests from contextual information. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, pp. 363–370. ACM, New York (2009)
12. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (eds.) Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum (2007)
13. Weng, J., Lim, E.-P., Jiang, J., He, Q.: TwitterRank: Finding topic-sensitive influential twitterers. In: Proceedings of the Third International ACM Conference on Web Search and Data Mining, pp. 261–270 (2010)
14. Micro-blogging, http://en.wikipedia.org/wiki/Micro-blogging
15. Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B., Magoulas, R.: Twitter and the micro-messaging revolution: Communication, connections, and immediacy-140 characters at a time. O'Reilly Report (November 2008)
16. Random Walk, http://en.wikipedia.org/wiki/Random_walk
17. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology 27(1), 415–444 (2001)
18. Tong, H., Faloutsos, C., Pan, J.-Y.: Fast random walk with restart and its applications. In: Proceeding of the 6th international Conference on Data Mining, pp. 613–622.
19. Adamic, L.A., Adar, E.: Friends and Neighbors on the Web. Social Networks 25(3), 211–230 (2003)
20. Welch, M.J., Schonfeld, U., He, D., Cho, J.: Topic semantics of twitter links. In: Proceedings of the Fourth International ACM Conference on Web Search and Data Mining (WSDM 2011), Hong Kong, China (2011)