2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress

Socialized User Modeling in Microblogging Scenarios for Interest Prediction

Mingqing Huang†, Bofeng Zhang†,*, Guobing Zou*, Chengwei Gu, Peiye Wu, Shulin Cheng, Zhu Zhang
School of Computer Engineering and Science
Shanghai University
Shanghai 200444, China
Emails: bfzhang@shu.edu.cn (Bofeng Zhang), gbzou@shu.edu.cn (Guobing Zou)

***Abstract*—For adapting functionality to individual users, systems need interest prediction information of these users. Social media provides opportunities to gather mass user data for this purpose. To effectively extract user interests, this paper proposes a hybrid user modeling approach which integrates isolated interests discovery via text analysis into the social relationship-based method by community detection for the interest prediction of microblog users. We conduct extensive experiments on large-scale microblogging datasets with 12,746 users. The experimental results demonstrate that our hybrid approach for socialized user modeling can significantly improve prediction accuracy of user interests, in comparison of the text-based method.**

***Index Terms*—user modeling; interest mining; community detection; LeaderRank sort; microblog platform.**

I. Introduction

Recently, due to the rapid development of Web 2.0 technology, many people post their opinions via online social networking services. In particular, microblogs as popular social media are responsible of suitable platforms for casual, real-time communication. Many microblog users state opinions from the perspectives of many facets, such as products, services, and brands. It is essential for enterprises to aim at improving the quality of their products and services based on customers interests. As a way of using user opinions for marketing, reputation analysis technologies have received extensive attentions over recent years [1]. Different from traditional marketing style based on questionnaire surveys, online opinion analysis possesses a lot of advantages, such as high volume, real-time feedback, and low cost.

User interest prediction is of vital importance for marketing analysis in recommender applications, since individual users interests vary with time. Considering many microblog users do not frequently express their views, in many cases it is a difficult task to draw sufficient opinions for interest prediction using the microblog content they posted, so that how to design a novel approach for user interest prediction has become a research challenge to be solved.

Several text-based methods have been suggested to extract user demographic information [2, 3]. However, on account of difficulties in achieving the efficiency and accuracy up to a level sufficient for practical application, there exist only few approaches for practical and large-scale interest analysis. Furthermore, user modeling performs an important role among all methods for users interests extraction and prediction.

The user model, also named user profile, as a mirror of a real user in the cyber world, is constructed to reveal the user demographic information and personal interests. There are many ways to present the interests of a user model. The Vector Space Model [4] has been widely applied to depict users, but it models users in a bloated style, has no uniform expression standard, and is unable to execute semantic extension. To solve these problems, the semantic ontology representation [5] is adopted in this study to describe users interests.

Although many methods have applied the user model to recommend information, they cannot accurately portray the user properties. The approach of recommending microblogs based upon user model [6] does not provide insight into the interaction relationships among users. The study on topics recommendation in microblogging scenarios based on neighborhood-user profile [7] simply binds semantic expansion and user social relationships. Twitter user profiling [8] employs social information only when the community demonstrates remarkable characteristics, and detects communities in the multi-hop network instead of the whole network of users. Consequently, it is extremely difficult to obtain the exact community partition.

To optimize the quality of user modeling, this paper proposes a hybrid user modeling approach combining the text-based method with the relationship-based method for more accurate interest prediction of microblog users. The predictions are estimated by analyzing the historic microblog contents and clustering of friends/followers, where the interaction relationship complements the microblog text for user modeling to address cold-start problems as well as sparse user profiles. In any clustered group of users, the individual influences are

†Contribute equally and share the first authorship.
*Corresponding authors.
This study is sponsored by the National Natural Science Foundation of China (61303096), the Shanghai Natural Science Foundation (13ZR1454600), the Specialized Research Fund for the Doctoral Program of Higher Education (20133108120029), and the Innovation Program of Shanghai Municipal Education Commission (14YZ017).

978-1-5090-4065-0/16 $31.00 © 2016 IEEE
DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.39

identified and regarded differently to reflect the different roles in society. In addition, the user susceptibilities to interests are taken into account in conformity with the objective situation.

The main characteristics and innovations of this study are threefold, which are listed as follows.

- The semantic interest extraction from microblogging history is performed via ontology base, which quantifies the interest degree on topic keywords in different granularities.
- We construct the user model by combining the microblog content with the interaction relationship of users for interest prediction, which can be exploited to effectively solve the data sparse problem.
- The individual influences and susceptibilities are estimated to accurately depict the interest impacts among different users through interactive relationships.

The rest of this paper is structured as follows. In Subsection II-A, we illustrate the framework of our approach for user modeling. The isolated user modeling method and the community detection algorithm for the friends/followers network are introduced in Subsection II-B and II-C, respectively. The hybrid of the text-based method and the relationship-based method for socialized user modeling is described in depth in Subsection II-D. The experimental results of performance evaluation are displayed and analyzed in Section III. Finally, we conclude this paper by prospecting the future work in Section IV.

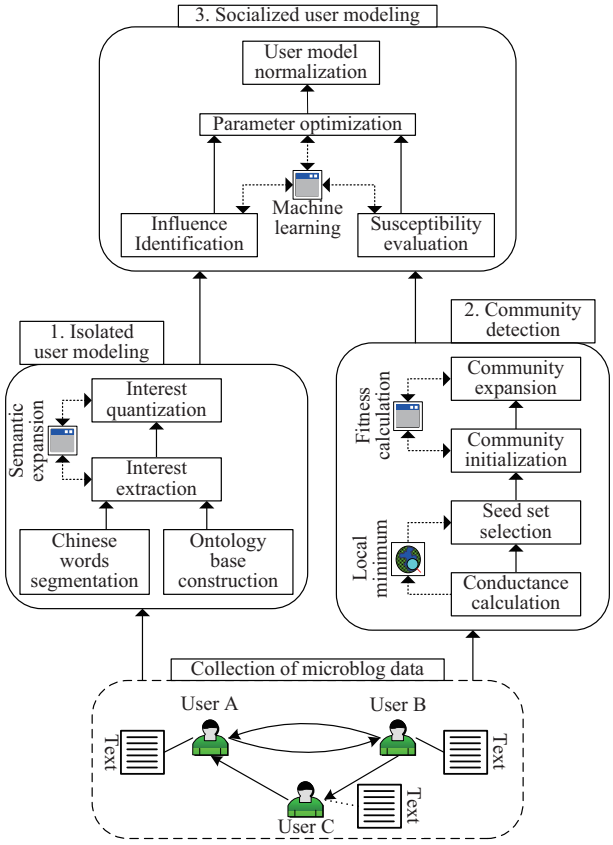

Fig. 1. The overall framework of the approach for user modeling.

II. Proposed Method

A. The Framework of Our Approach

In this study we focus on the Chinese microblogging as the research platform. However, the proposed hybrid approach with little modification can be applicable to any microblogging or social media in any other languages.

To achieve the goal of mining user interests, the process of user modeling consists of three crucial steps of isolated user modeling, community detection of friends/followers, and socialized user modeling. The detailed overall framework is illustrated in Fig. 1, where the collection of microblog data has been plugged as the foundation.

First, the text analysis is carried out on the microblog content via an ontology base for user interest extraction and degree quantization. Then, we detect overlapping communities using initial community set expansion and optimization in the directed network of friends/followers. Finally, the individual influence and susceptibility are evaluated to complement the text interest mining for hybrid user modeling through the parameter tuning.

B. Isolated User Modeling Based on Text Analysis

In this subsection, we first construct an ontology base for logic relationships management of concepts, and then mine and quantify user text interests, including semantic expansion and model representation.

1) *General ontology base construction*

Rather than the domain ontology, the general ontology is employed for keywords extraction, since the user microblogging history doesn't involve only a single field. Actually it covers a variety of areas in real-world microblogging applications.

Baidu Encyclopedia [9] is the most influential, comprehensive and attractive Chinese ontology, which was provided in an attempt to structure a Chinese information collection platform covering all domains of knowledge. In the January 2015 edition, there are 11 keyword channel categories in the first layer, 89 ones in the second layer, and 380,627 keywords in all four layers of *Baidu Encyclopedia*, as seen in Fig. 2. Because of these advantages, in this study *Baidu Encyclopedia* is adopted as the semantic ontology base for text interest extraction and analysis of large-scale user microbloggings.

2) *Interest extraction and quantization*

Through text analysis on the microblog contents of users, we quantify individual interests by interest degrees on specific topic keywords of the ontology base.

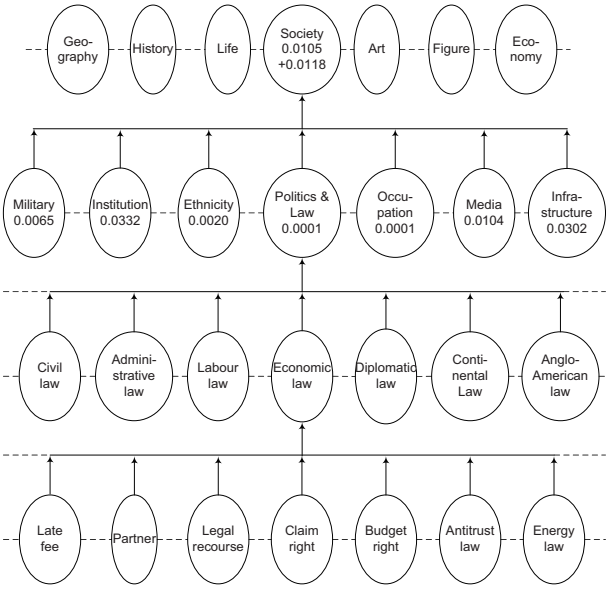

Fig. 2. The general ontology base of *Baidu Encyclopedia*.

The detailed process of quantifying the interest degree of an individual user consists of four steps as follows.

a) *User interest extraction*

The microblogs of each user during a predefined recent period of time (three months for an instance) are accumulated as a separate document. We perform the operation of Chinese words segmentation on each document that leads to a set of partitioned keywords. Furthermore, those keywords are queried and only those that exist in the general ontology base are extracted as user interests.

b) *Interest degree quantization*

Term Frequency - Inverse Document Frequency (*TF-IDF*) [10] is applied to build the user interest models. It is defined as below.

$$TF-IDF(D_i,t_j,D)=\frac{tf(t_j,D_i)}{\sum_{t_k\in D_i}tf(t_k,D_i)}\times idf(t_j,D). \tag{1}$$

Where $TF-IDF(D_i,t_j,D)$ stands for the weight value of keyword t_j to document D_i in document collection D; $tf(t_j,D_i)$ represents the number of times that item t_j appears in D_i; and $idf(t_j,D)$ is calculated as:

$$idf(t_j,D)=log(\frac{|D|}{|\{D_k\in D:t_j\in D_k\}|}). \tag{2}$$

The *TF-IDF* values of microblog keywords extracted from each document are calculated, and all *TF-IDF* values correlated with the same user are accumulated and transferred to a newly-built four-layer ontology base, which is referred as the initial user interest model of this user.

c) *Semantic value propagation*

To strengthen the interest degrees of a user model in its upper layers, the *One-way Upward Diffusion Model* is employed to perform semantic expansion and interest degree propagation. In other words, the interest degree of a user on keyword t_i in the fourth layer is multiplied by a coefficient α_i and added to its parent keyword. In the same way, the value propagation is then executed in the third and second layers. Herein α_i is the reciprocal of the number of siblings of keyword t_i. The diffusion model is calculated as below.

$$I'_j=I_j+\sum_{i=1}^{n^j_{child}}(\alpha_i I^j_i)=I_j+\sum_{i=1}^{n^j_{child}}(\frac{I^j_i}{n^j_{child}}). \tag{3}$$

Where I_j stands for the initial interest value on keyword t_j; I'_j is the updated one after the propagation is implemented; n^j_{child} represents the number of child nodes; and I^j_i denotes the interest degree on the *i*-th child keyword.

For example, as displayed in Fig. 2, the interest value of an initial user interest model on keyword "*society*" equals to 0.0105, and the interest values on all relevant seven child keywords after propagation sum up to 0.0825. As a result, this interest degree is then changed to $0.0105+(0.0825/7)=0.0223$ in the updated user interest model.

d) *Model representation with reduction*

The *Baidu Encyclopedia* ontology base consists of 380,627 keywords. As a compromise, the final user interest model in this study holds only top two layers from the ontology base, which amounts to 100 keywords for the representation of user interests.

Without loss of generality, the interest degrees in each layer are normalized. That is, the sum of each layer is equal to 1 after uniformly scaling on each keyword.

Taking a microblog user as an instance, the final user interest model is demonstrated in Fig. 3, which is constructed as a three-layer tree with topic keywords as nodes and corresponding interest degrees as annotations.

C. Community Detection in Microblog Platform

Consistent with previous work [11], directed edges are exploited to construct the social relationship network with arrows pointing from followers to friends. To efficiently mine communities from the social network, multiple nodes are first selected as seeds based on local minimum conductance. Then, we expand the initialized community to complete the overlapping community detection, which includes the membership degree calculation.

1) *Multiple seeds selection*

Definition 1 (**Conductance**). *By extending the concept of conductance [12] to directed networks, we get the definition as:*

$$\phi(S)=\frac{cut(S)}{min(vol(S),vol(\bar{S}))}. \tag{4}$$

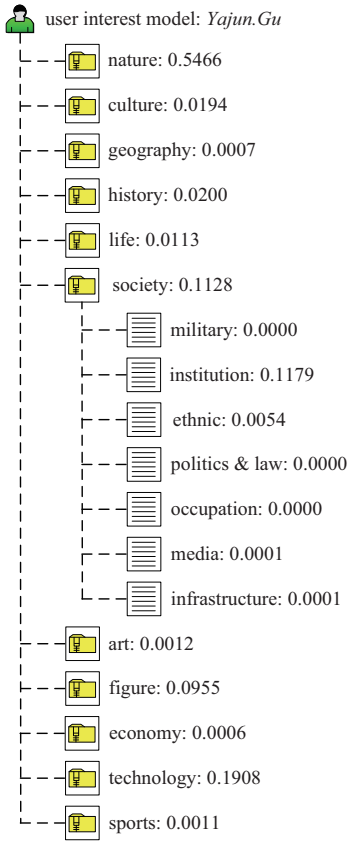

Fig. 3. A user interest model of the *Baidu Encyclopedia* top two-layer ontology base.

Where S denotes a connectivity node set in the network; $cut(S)$ is the number of edges pointing from nodes in S to other external ones; $vol(S)$ represents the sum of out-degrees of nodes in S; and $\bar{S}$ stands for the complementary of S, composed of nodes out of S. Since $vol(S) << vol(\bar{S})$ in large-scale social networks, we just need to calculate $vol(S)$.

By using the conductance, we simultaneously select multiple seeds for community detection as follows. For each follower node that does not belong to any detected community, if the node set consisting of this follower and all its friends has the conditional local minimum conductance, then this node is labeled as a community seed. The conditional local minimum conductance is determined by:

$$\phi(F(v)) \leq \phi(F(w)). \tag{5}$$

Where $F(v)$ represents the node set formed by node v and its friend nodes; w is every node except v in $F(v)$ which doesn't belong to any mined community.

2) *Overlapping community detection*

The initial community is extracted from the directed network by taking full advantage of the attribute of seed nodes, and then we expand it for overlapping community detection.

a) *Community initialization*

Definition 2 (**Community fitness**). *The fitness coefficient [13] of community C is calculated to express the cohesion degree by the function:*

$$f_C = \frac{W_{in}^C}{W_{in}^C + W_{out}^C}. \tag{6}$$

Where W_{in}^C and W_{out}^C denote the number of edges in community C and the number of edges pointing from nodes in C to other external ones, respectively.

Given a seed node, we identify the node set of its one-hop network as S_i, and then examine the items in S_i one by one. If the deletion of a node from S_i can increase the fitness value of S_i, we kick this node out of S_i. By doing so, the purified node set S'_i is returned as an initial community corresponding to the seed node.

b) *Community expansion*

Based on any incomplete initial community S_C, we perform the expansion process to add new nodes or delete those nodes with negative contributions to the community fitness that can optimize the detected community. It consists of six steps for each iteration.

(i) A loop test is executed over all neighbor nodes of S_C;
(ii) The neighbors with less than a threshold number [14], which increase the largest fitness, are added to S_C, yielding a larger community S'_C;
(iii) If there is no one neighbor that can lead to a larger fitness, the process stops and returns the current community as the final one;
(iv) The contribution of each node in S'_C to the community fitness is recalculated;
(v) The nodes with the largest negative values, which do not exceed another given threshold number, are excluded from S'_C, yielding a new community S''_C;
(vi) If step (v) occurs, the process jumps to step (iv), otherwise it repeats from step (i).

c) *Membership degree calculation*

Since some of the nodes can span several overlapping communities with different importance on interest preferences, their membership degrees are distributed unbalancedly.

Assume that C_p^i denotes the i-th community of node p, and N_p^i stands for the node set whose members are all included in C_p^i and adjacent to p. Under above parameter settings, the membership degree of p in C_p^i indicated by B_p^i is defined as:

$$A(p,q) = \frac{1}{m}(M_{pq} - \frac{d_p^{out} d_q^{in}}{m}). \tag{7}$$

$$B_p^i = \sum_{q \in N_p^i} \frac{1}{o_p o_q}(A(p,q) + A(q,p)). \tag{8}$$

Where m, M_{pq} and $A(p,q)$ are respectively the number of edges, the adjacent matrix element and the affinity degree pointing from node p to q in the social relationship network; d_p^{out} and d_q^{in} represent the out-degree of p and the in-degree of q, respectively; and o_p is the number of communities to which node p belongs.

Without loss of generality, the membership degrees on all correlated communities are then normalized so as to sum the degree distribution on each overlapping node as 1.

D. Socialized User Modeling with Hybrid Method

The influence of others and susceptibility of themselves are assessed in the community, which are applied to estimate the complementarity of the social relationships to the microblog content for socialized user modeling.

1) *Influence identification*

Although LeaderRank [15] is analogous to PageRank, it is more stable and robust than PageRank in terms of disturbances and attacks. In addition, LeaderRank is a parameter-free method for practical use. Since these attractive advantages, LeaderRank is adopted to rank the user influence in this study.

Based on PageRank, LeaderRank introduces a ground node g, which connects two directed edges e_{gi} and e_{ig} to every node i in the original network. The rank score $r_j(t)$ of note j at discrete time t is determined by:

$$r_j(t) = \sum_{k=1}^{n+1} \frac{M_{kj}}{d_k^{out}} r_k(t-1). \tag{9}$$

Initially, we set $r_g(0) = 0$ for the ground node g and $r_k(0) = 1$ for every other node k. After reaching the steady state, the rank score of the ground node is equally assigned to all other nodes to conserve scores without in-degree edges. Therefore, the final score of node j is defined as:

$$r_j = r_j(t_\infty) + \frac{r_g(t_\infty)}{n}. \tag{10}$$

Where $r_j(t_\infty)$ denotes the rank score of node j under the stable condition.

2) *Susceptibility evaluation*

In the research of quantifying user influence on Twitter [16], the investigation on the influence and relative attributes of large-scale Twitter data verified a hypothesis that the largest diffusion cascades tend to be generated by users who have a large number of followers and have been influential in the past. From another perspective, the greater proportion of forwarding and commenting a user has, the more susceptible of assimilation it would be for a microblog user to others interests.

For simplicity sake, the forwarding and commenting are treated equally without any distinction in this study, and we get the susceptibility formula as follows:

$$s_p = \frac{n_p^f + n_p^c}{n_p^m}. \tag{11}$$

Where n_p^f, n_p^c and n_p^m denote the number of forwarding, commenting and all microblogs for user p within recent times of a designated interval, respectively.

3) *Hybrid user modeling*

The community interest model is estimated based upon all user interest models in a community by taking account of the differences in individual influences, and all the isolated user interest models are then updated by this community interest model.

a) *Community interest model*

Given a social community, every microblog user with different personal interests may impact one another to varying degrees. The community interest model is formed under the long-term effects of each other to depict the overall interests of the community. Based on this idea, we construct the community interest model with the differences among all members as:

$$I_C = \frac{\sum_{i \in C} r_i \times I_i}{\sum_{i \in C} r_i}. \tag{12}$$

Where I_i represents the interest model of user i by vector with constant length, and r_i stands for the influence score of the same user.

b) *Hybrid of text and relationship*

The hybrid method targets almost all users by making the best of the microblog platform, including both microblogs as text information and followers/friends as social relationship information. With the combination of these two factors, we construct the socialized user model by a mixing formula derived as:

$$I_p^s = \alpha \sum\nolimits_{i \in C_p} B_p^i I_{C_p^i} + (1-\alpha)(1-\beta s_p) I_p. \tag{13}$$

The interest values in each layer are normalized finally for generality. Where α and $\beta$are the influence coefficient expressing the influence degree of the community to the individual interest and the susceptibility coefficient implying the ability to become assimilated, respectively.

The parameters α and β can get the optimal values for socialized user modeling through machine learning. For simplicity, in this study we select the best one from multiple combination values as the ideal value.

III. Experimental Evaluation

We evaluate the performance of the proposed approach using the data, experimental environment and evaluation criterions introduced in the following subsections.

A. Collection of Microblog Data

Launched by Tencent Co. Ltd., Tencent microblogging is an online social networking service analogous to Twitter. By the winter of 2012, the number of registered members accounted for up to 540 million, which is still quickly increasing as time goes. Specifically, the number of daily active users has exceeded 100 million. Furthermore, Tencent microblogging offers a free and open API [17] for developers to gather the global user data. Given the significant advantages above, Tencent microblogging is chosen as the experimental platform for the performance evaluation.

It is observed that every user has dozens of friends on average. If we select some users stochastically to construct a directed network of friends/followers, it cannot express accurately the social attributes of these users. Therefore, we begin from a specified user who is assigned as the initial node in the directed network, and all followers of this user are then added to the network. After that, all the followers of these new members are recruited in the same way. We repeat this process until the number of nodes satisfies the experimental demand. Besides, the individuals who have followers or friends up to more than 1000 are excluded, since they are referred to as stars or friend abusers.

We carry out the experiments with two batches of microblog data associated with the same 12,746 users, which are acquired in June and August 2015 for user interests prediction and verification, respectively. The total number of friends/followers records reaches 83,809 among these users in June.

B. Experimental Environment and Evaluation Criterions

1) *Experimental environment*

We execute the experiments on an ordinary computer with a four-core 2.33-GHz processor, 8GB RAM, and Windows Server 2008 R2 OS using the Java 1.7 64-bit.

2) *Evaluation criterions*

To validate the effectiveness of our proposed hybrid approach for user modeling, we apply the following two evaluation criterions in our experiments, including accuracy rate and Pearson correlation coefficient.

a) *Accuracy rate*

The accuracy of the interest distribution ratio on each topic keyword is a metric to clarify that the prediction ratio is not biased towards some particular interest regions. For marketing analysis, the distribution ratio of the user interests is important to perform, e.g. the follower interests analysis of a company. High accuracy means that the method returns most of the correct results. Given a specified user, the Accuracy rate of interest distribution ratio is defined as:

$$Accuracy = \sum_{t \in T} \frac{I_p^t \bigcap I_a^t}{|T| I_a^t}. \tag{14}$$

Where I_p^t and I_a^t represent the predicted and actual interest degree on topic keyword t in collection T, respectively; $|T|$ denotes the number of items in T. Here, we set $I_p^t \bigcap I_a^t = \min\left(I_p^t, I_a^t\right)$ to get the intersection set of continuous data.

b) *Pearson correlation coefficient*

Several metrics have been proposed in literature with the purpose of measuring the interest similarity between users, among which the Pearson correlation coefficient [18] gets wide acceptance. In this study, it defines the similarity between the predicted and actual interests of a specified user. It is expressed by:

$$sim\left(I_p, I_a\right) = \frac{\sum_{i=1}^{n} \sum_{j \in T_i} \left(I_p^j - \bar{I}_{p,T_i}\right)\left(I_a^j - \bar{I}_{a,T_i}\right)}{\sqrt{\sum_{i=1}^{n} \sum_{j \in T_i} \left(I_p^j - \bar{I}_{p,T_i}\right)^2 \sum_{i=1}^{n} \sum_{j \in T_i} \left(I_a^j - \bar{I}_{a,T_i}\right)^2}}. \tag{15}$$

Unlike the traditional formalization of Pearson correlation coefficient, herein we make differences among topic keywords since there are multiple layers in the ontology base. Where T_i is the keyword set of the *i*-th ontology layer; and $\bar{I}_{p,T_i}$ denotes the mean interest degree of the user on the corresponding layer.

C. Experimental Results

Out of all the 12,746 users, we pick out 1,340 active ones for the verification on interest prediction. On the contrary to the remainders who didn't update their blogs frequently, most of them wrote, forwarded or commented at least 15 microblogs during the two months interval.

Both theoretical analysis and experimental data manifest that the variations of the influence and susceptibility coefficient can significantly affect the accuracy of interest prediction. As illustrated in Fig. 4 and 5, the measurements in the above-mentioned metrics vary with the changes of α and β, where isolated user modeling approach is referenced as a comparison with socialized user modeling method.

From the experimental results, we find that the proposed hybrid method is significantly superior to the other approach in all two evaluation metrics. When the parameters are assigned as $\alpha = 0.5$ and $\beta = 0.2$, our approach gains the highest performance, where the effectiveness rates are increased by 24.7% and 8.5% in Accuracy rate and Pearson correlation coefficient, respectively.

As illustrated in the experimental results, the isolated user modeling approach maintains different constant values in two evaluation metrics, since this method is impervious to the influence and susceptibility coefficient. As for the socialized user modeling method, the measurement values increase first and then decrease, where the influence coefficient impacts more tremendous to the interest prediction accuracy than the susceptibility coefficient.

The time complexities of the isolated user modeling and community detection are $O(n)$ and $O(m)$, where n and m

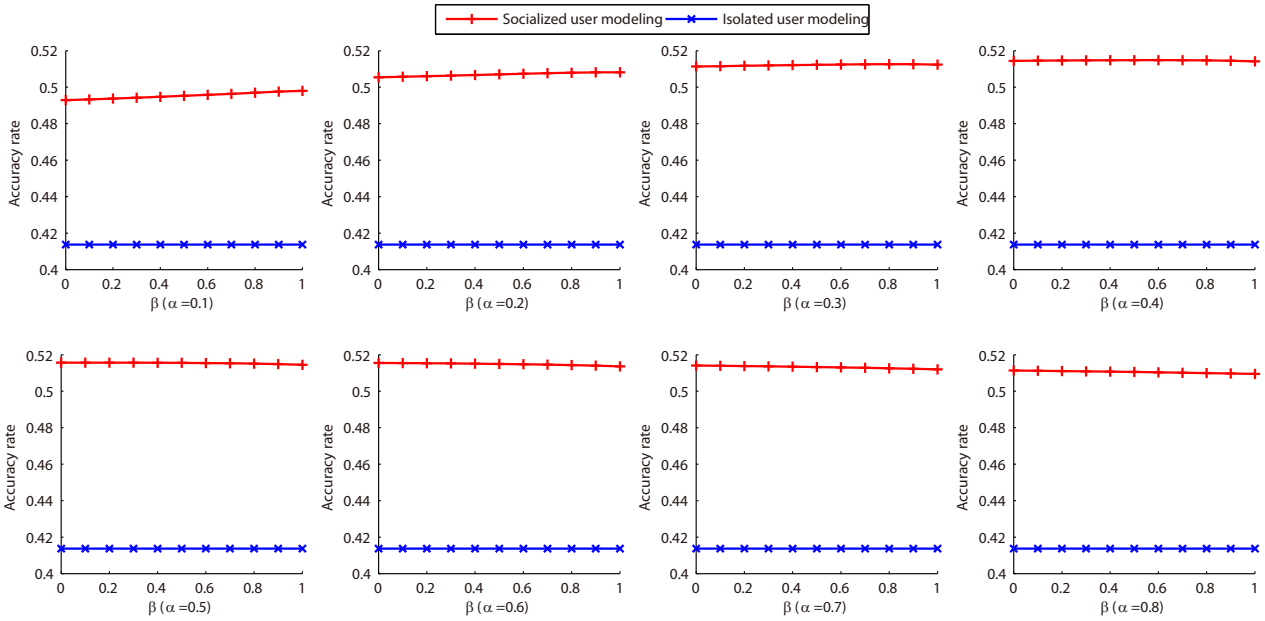

Fig. 4. Comparison of Accuracy rate of the socialized user modeling method with the isolated user modeling approach.

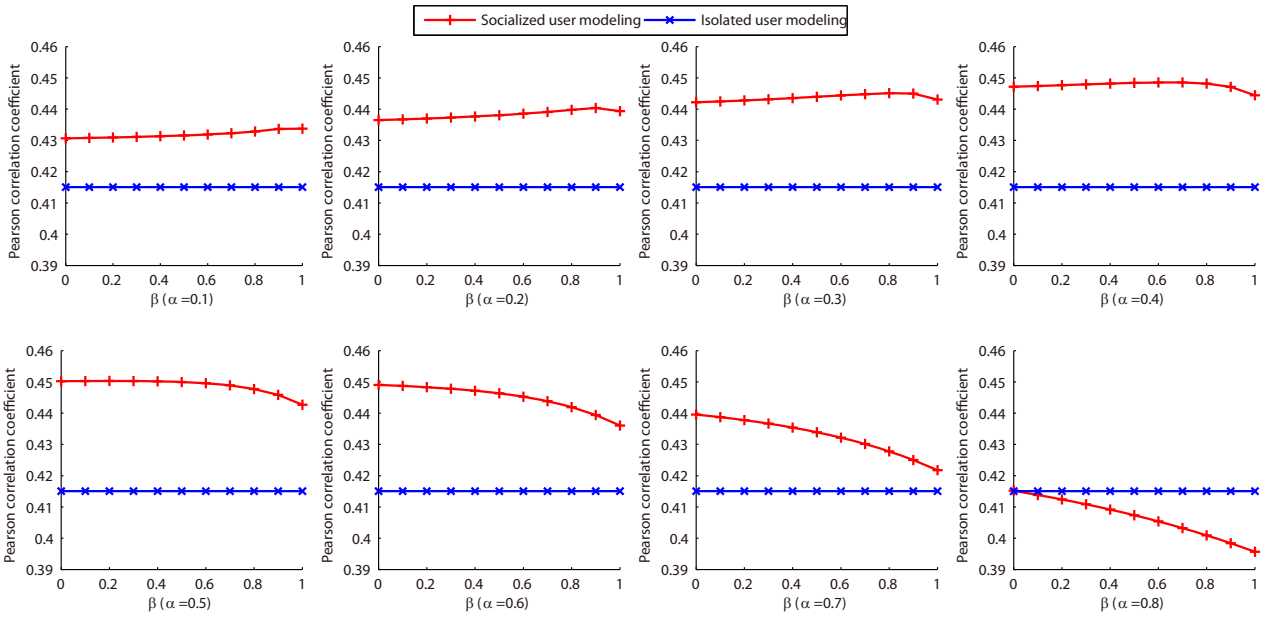

Fig. 5. Comparison of Pearson correlation coefficient of the socialized user modeling method with the isolated user modeling approach.

respectively denote the number of users and relationships of friends/followers, so the time complexity of the hybrid method is also $O(m)$, since $m > n$ in the overwhelming majority of cases. The processing time of the experiment on 12,746 users with multi-threading is only one day, which demonstrates convincingly that our method is suitable for practical application.

IV. Conclusions and Future Work

Accurate interest prediction on targeted users in cyberspace for market analysis has become an attractive research problem. It can highlight many applications, such as recommendation systems. In this paper, we propose a hybrid user modeling method which integrates isolated user modeling based on text analysis and community detection for social relationship discovery. The experimental results on large-scale datasets

demonstrate that our proposed method can significantly improve the prediction accuracy of user interests with low time consumption, can satisfy the requirement of actual application.

The current design of our method takes into account only the friends/followers as social relationships. In the future, we intend to improve it by incorporating more user interactions, including the forwarding and commenting. In addition, we will further validate the performance of our method in large-scale microblog datasets.

ACKNOWLEDGMENT

We would like to thank all the anonymous reviewers for their constructive suggestions and insightful comments that can significantly improve the quality of our manuscript.

REFERENCES

[1] J. Wiebe and E. Riloff, "Finding mutual benefit between subjectivity analysis and information extraction," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 175–191, 2011.

[2] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," *Communications of the ACM*, vol. 52, no. 2, pp. 119–123, 2009.

[3] D. D. Pham, G. B. Tran, and S. B. Pham, "Author profiling for vietnamese blogs," in *Proceedings of the International Conference on Asian Languages Processing*, 2009, pp. 190–194.

[4] J. Kamahara, T. Asakawa, S. Shimojo, and H. Miyahara, "A community-based recommendation system to reveal unexpected interests," in *Proceedings of the 11th International Multimedia Modelling Conference (MMM05)*, 2005, pp. 433–438.

[5] S. Calegari and G. Pasi, "Personal ontologies: Generation of user profiles based on the yago ontology," *Information processing and management*, vol. 49, no. 3, pp. 640–658, 2013.

[6] K. Mei, B. Zhang, J. Zheng, L. Zhang, and M. Wang, "Method of recommend microblogging based on user model," in *Proceedings of the 12th International Conference on Computer and Information Technology*, 2012, pp. 1056–1060.

[7] J. Zheng, B. Zhang, X. Yue, G. Zou, J. Ma, and et al, "Neighborhood-user profiling based on perception relationship in the micro-blog scenario," *Journal of Web Semantics*, vol. 34, pp. 13–26, 2015.

[8] K. Ikeda, G. Hattori, C. Ono, H. Asoh, and T. Higashino, "Twitter user profiling based on text and community mining for market analysis," *Knowledge-Based Systems*, vol. 51, pp. 35–47, 2013.

[9] "Baidu encyclopedia," *<http://baike.baidu.com>*.

[10] G. Salton and M. J. Mcgill, "Introduction to modern information retrieval," 1986.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," 1999.

[12] D. F. Gleich and C. Seshadhri, "Vertex neighborhoods, low conductance cuts, and good seeds for local community methods," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 597–605.

[13] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, 2009, article. 033015.

[14] M. Huang, G. Zou, B. Zhang, Y. Liu, Y. Gu, and etc., "Overlapping community detection in heterogeneous social network via user model."

[15] L. Lü, Y. C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PloS one*, vol. 6, no. 6, 2011, article. e21202.

[16] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the 4th ACM international conference on Web search and data mining*, 2011, pp. 65–74.

[17] "Tencent microblog platform," *<http://dev.t.qq.com>*.

[18] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise reduction in speech processing*. Springer Science and Business Media, 2009.