Research on Interest-Fusion Weibo User Similarity Computation (Postprint)

Authors: Huang Xianying, Yang Anzhi, Liu Xiaoyang, Liu Guangfeng

Date: 2018-11-29T00:00:00+00:00

Abstract

To address the problem of potentially excessive errors in traditional similarity calculations for microblog users based on post content and mutual friend counts, as well as the issues of high computational complexity and neglect of user interests in similarity computation models based on multi-source background information, this paper proposes a comprehensive similarity calculation method that integrates user interests and background information (BIBS). First, user interests are extracted from user tags; when tags are absent, user interest points are indirectly acquired by clustering influential users within the user' s following relationship network, thereby computing user interest similarity. Second, user background similarity is calculated based on demographic attributes such as gender, age, and location, enabling hierarchical mining of the most similar users. Finally, experimental analysis is conducted using Sina Weibo data. The results demonstrate that compared with the Microblog User Recommendation algorithm based on Multi-source Information Similarity (MISUR), the proposed method achieves improvements of 8.1%, 16.7%, and 13.6% in precision, recall, and F-score respectively, while requiring less computational time, thus validating the effectiveness and accuracy of the BIBS method.

Full Text

Preamble

Vol. 37 No. 1
Application Research of Computers (ChinaXiv Cooperative Journal)

Research on Similarity Computation of Microblog Users Combining User Interests

Huang Xianying, Yang Anzhi, Liu Xiaoyang, Liu Guangfeng
(School of Computer Science & Engineering, Chongqing University of Technol-

ogy, Chongqing 400054, China)

Abstract: Traditional methods for calculating microblog user similarity based on blog content and common friend counts suffer from excessive potential errors, while similarity computation models based on multi-source background information exhibit high computational complexity and neglect user interests. To address these issues, this paper proposes a comprehensive similarity calculation method (BIBS) that combines user interests and background information. First, user interests are extracted from user tags; when tags are missing, user interest points are indirectly obtained by clustering important users in the user' s follow relationship network, thereby calculating user interest similarity. Second, background similarity is computed based on user attributes such as gender, age, and location, hierarchically mining the most similar users. Finally, experimental analysis based on Sina Weibo data demonstrates that compared with the MISUR algorithm (user recommendation algorithm based on multi-source information similarity), the proposed method improves accuracy, recall rate, and F-measure by 8.1%, 16.7%, and 13.6% respectively while consuming less time, proving the effectiveness and accuracy of the BIBS method.

Keywords: microblog; interest; user clustering; similarity calculation

0 Introduction

With the advancement of information technology, online social networks have developed rapidly with increasing user participation. According to the 41st "Statistical Report on Internet Development in China" released by CNNIC, by December 2017, microblog users exceeded 310 million, with an annual growth rate of 16.4%. This massive user base creates information overload challenges when users search for information or establish interactive relationships on microblogs. Effectively helping users discover like-minded individuals among numerous nodes is crucial for both social network platforms and users themselves, and personalized recommendation represents an effective solution to this problem.

Traditional recommendation methods include collaborative filtering, content-based recommendation, and hybrid approaches, which have been applied in friend recommendations, news recommendations, music recommendations, and other domains. A critical component of personalized recommendation is similarity computation—such as user similarity and item similarity—which forms the foundation for generating recommendations. Most conventional recommendation algorithms build user interest models based on historical rating data to calculate user similarity and produce recommendations. With the rapid development of Web 2.0 and the popularity of online social networks like Twitter, Facebook, and Sina Weibo, traditional recommender systems have begun incorporating microblog users' background information and social behavior data to provide more relevant recommendations.

In recent years, numerous novel user similarity calculation methods have emerged in the microblog recommendation domain. For instance, Xu Zhiming et al. considered user background information, microblog text, and social information attributes to compute user similarity. Other researchers have combined user gender, age, and blog content information to propose comprehensive similarity calculation methods based on cosine distance. Yao Binxiu et al. integrated blog content, interaction information, and common follower counts to propose a microblog user recommendation algorithm based on multi-source information similarity. These methods construct feature vectors by comprehensively considering multiple aspects of user information and employ cosine distance to mine similar users. However, due to microblog length limitations, directly constructing user feature vectors and using cosine similarity is insufficient for measuring microblog user similarity and suffers from excessive potential errors and high computational complexity.

He et al. clustered users based on blog forwarding relationship networks and discovered that users in the same community share similar interests, demonstrating that social network relationships are primarily built on common interests. Incorporating user interests enables more accurate discovery of similar users within communities. Several studies have calculated user similarity based on interests, such as Huang Hongcheng et al., who investigated long-term and short-term interests for relationship prediction, and Chen Jie et al., who proposed a microblog recommendation method based on dynamic user interests. Interest-based recommendation has become increasingly popular. Xing et al. conducted in-depth research on user tags and found that verified users (VIP users) tend to add more tags than ordinary users, with experiments showing that tag information most effectively captures microblog user interests. Ma Huifang et al. also studied user-defined tags for recommendation purposes. Although tag-based recommendation is more accurate and effective, many ordinary microblog users lack custom tags. Alternative methods for extracting user interests include mining from user profiles and blog content, while Zhong Zhaoman et al. demonstrated that indirectly obtaining user interests through follow relationships is reasonable and effective—users follow celebrities because of interest in their domain, reflecting user interests in that field.

1.1 Traditional Calculation Methods

In traditional e-commerce services, personalized recommendation technology studies user preferences to recommend relevant products. User-based collaborative filtering algorithms, for example, calculate user similarity from rating matrices, identify the nearest neighbor set for target users, weight the nearest neighbors, and generate recommendations. These algorithms effectively utilize feedback from similar users to produce recommendations. With the rapid development of social networks, personalized recommendation technology has been applied in various ways in social networks, with numerous recommenda-

tion methods emerging in the microblog domain.

Earlier methods calculated microblog user similarity based on common neighbor counts, such as the Common Neighbors (CN) model and Jaccard similarity calculation model. The CN similarity model is defined as:

$$Sim_{CN}(A,B) = \frac{|CN(A) \cap CN(B)|}{|CN(A) \cup CN(B)|}$$

where $Sim_{CN}(A,B)$ represents the similarity between users A and B, $CN(A)$ represents user A' s friend set, and $CN(B)$ represents user B' s friend set. The more common friends users A and B share, the more similar they are. However, such algorithms suffer from poor recommendation accuracy. On one hand, they ignore information from microblog users themselves, such as preferences and age; on the other hand, unlike real-world friendships, social network users do not necessarily have strong connections with everyone on their friend list, resulting in low user satisfaction with recommendations based solely on common friend counts.

1.2 Multi-Source Information Fusion Methods

To address limitations of traditional recommendation algorithms, researchers began incorporating microblog users' background information to calculate similarity. Some studies examined user location, tag, and personal description information for background similarity calculation, while others combined gender, age, and geographic information. Building on existing research, this paper combines gender, age, and location information to compute microblog user background similarity.

These methods comprehensively consider multiple aspects of user information and use cosine distance to calculate similarity but still face challenges. For instance, blog posts exhibit high randomness, leading to excessive potential errors in similarity calculation. Additionally, constructing feature vectors from user background information and directly applying cosine similarity consumes significant resources in practical applications with relatively high computational complexity.

The MISUR algorithm (user recommendation algorithm based on multi-source information similarity) first preprocesses and segments user blog content, employs TF-IDF (term frequency–inverse document frequency) to obtain keyword tables, calculates blog content similarity using cosine distance, computes user interaction behavior similarity based on mutual interest in each other' s microblogs, and finally calculates social relationship similarity based on common followees and followers. The component formulas are defined as follows:

Blog Content Similarity:

$$Sim_{blog}(u,v) = \frac{\sum_{i \in r_u \cap r_v}(r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in r_u}(r_{ui} - \bar{r}_u)^2}\sqrt{\sum_{i \in r_v}(r_{vi} - \bar{r}_v)^2}}$$

where $Sim_{blog}(u,v)$ represents the blog content similarity between users u and v, and r_u, r_v represent the blog text vectors of users u and v.

Interaction Behavior Similarity:

$$Sim_{interaction}(u,v) = \frac{\sum_{r \in R_{uv}}(I_u(r) - \bar{I}_u)(I_v(r) - \bar{I}_v)}{\sqrt{\sum_{r \in R_{uv}}(I_u(r) - \bar{I}_u)^2}\sqrt{\sum_{r \in R_{uv}}(I_v(r) - \bar{I}_v)^2}}$$

where $Sim_{interaction}(u,v)$ represents the interaction behavior similarity between users u and v, $I_u(r)$ and $I_v(r)$ represent the interest levels of users u and v in microblog r, and $\bar{I}_u$ and $\bar{I}_v$ represent the average interest levels of users u and v across all interacted microblogs.

Social Relationship Similarity:

$$Sim_{social}(u,v) = w_1 \times Sim_{Following}(u,v) + w_2 \times Sim_{Follower}(u,v)$$

where $Sim_{social}(u,v)$ represents the social relationship similarity between users u and v, $Sim_{Following}(u,v)$ represents common followee similarity, $Sim_{Follower}(u,v)$ represents common follower similarity, and w_1 and w_2 represent the weights of each component, with $w_1 + w_2 = 1$.

Finally, the multi-source information similarity is computed by integrating these three components: blog content similarity, interaction similarity, and social relationship similarity.

1.3 User Interest Mining

Social network relationships are primarily built on common interests. To mine microblog user interests, Xing et al. conducted in-depth research on user tags and found that verified users (VIP users) tend to add significantly more tags than ordinary users. Experiments demonstrate that extracting microblog user interests from tag information is most effective. Other studies have investigated long-term and short-term interests, showing that tags represent long-term interests that are relatively stable.

Although tag information can accurately mine user interests and many celebrity VIPs willingly add more tags, ordinary users rarely define tags for themselves. Therefore, relying solely on tag information to mine microblog user interests has limitations. This paper analyzes the structure of microblog user follow

relationship networks and proposes a method to indirectly obtain user interests by leveraging important users in the follow relationship network, since many ordinary users lack custom tags representing their interests.

2.1 User Follow Network

In microblog social networks, when a user is interested in another user, they can follow that user and many other users of interest. Similarly, other users can follow them. The mutual following relationships among many users constitute a follow relationship network. In this network, some user nodes are followed by many other nodes due to their characteristics. These users are typically active community members with significant influence over other nodes and are called "opinion leaders." Compared to ordinary users, these important users in a user' s follow list better represent the user' s interest points.

Given the complexity of user follow relationship networks, the PageRank algorithm is employed to mine important users that can represent user interests. The follow relationship in social networks can be viewed as directed links. PageRank, proposed by Google' s founders for webpage ranking, is widely used in search engines. Page scores are obtained through iterative calculations, but issues like "rank leakage" and "rank sink" occur when pages have only incoming or outgoing links, producing unreasonable results. To address this, a random surfing model is introduced where each page can randomly access other pages. The final algorithm is represented as:

$$PageRank(x) = (1-d) + d \sum_{y \in L(x)} \frac{PageRank(y)}{N(y)}$$

where $PageRank(x)$ and $PageRank(y)$ represent the ranking scores of pages x and y, $L(x)$ represents the set of pages linking to x, $N(y)$ represents the total number of outgoing links from page y, and d is the damping coefficient representing the probability of a page being randomly accessed by other pages. This ensures pages receive a basic score even when not referenced, guaranteeing convergence.

When mining important users, $PageRank(x)$ and $PageRank(y)$ represent the importance of users x and y, $L(x)$ can represent the set of users followed by user x or interaction relationship sets, and $N(y)$ represents the number of friends in the corresponding set. Many studies have improved PageRank for mining important users in microblogs, such as Cao Jiuxin et al., who improved PageRank to mine opinion leaders.

2.2 Important User Mining

This paper uses the PageRank algorithm to mine important users in the follow network that best represent user interest points, indirectly obtaining user interests. Research shows that indirectly acquiring user interests through follow relationships is effective. If two users both follow celebrity Xie Na (with custom tags: "social idler," "host"), it indicates both users may be interested in hosting and can be predicted to potentially share interests in entertainment shows and variety programs. Therefore, when custom tags are scarce, mining user interests through follow relationships is more accurate than extracting interests from blog content.

2.3 User Interest Mining

The process involves first mining important users through PageRank, extracting their tags, constructing tag vectors, and clustering important users to obtain clustering results. The category vectors of clustering results are defined as $Cluster_1, Cluster_2, Cluster_3, \ldots, Cluster_N$, representing different clustering categories.

Next, the number of friends a user follows in each category is counted to construct the user' s interest vector, defined as:

$$Interest_A = (count_1, count_2, count_3, \ldots, count_N)$$

where $Interest_A$ represents user A' s interest vector, $count_1$ represents the number of users followed by user A in category 1, and $count_N$ represents the number of users followed in category N.

However, when a user follows many users in a particular category, it can interfere with cosine similarity results. Based on the TF-IDF concept, normalization is applied to obtain the user' s interest vector:

$$Interest_A = \left(\frac{count_1}{Num_1}, \frac{count_2}{Num_2}, \frac{count_3}{Num_3}, \ldots, \frac{count_N}{Num_N}\right)$$

where $Interest_A$ represents user A' s interest vector, and Num_1, Num_N represent the total number of users in category 1 and category N, respectively.

Finally, cosine distance is used to measure interest similarity between different users, with the calculation formula:

$$InterestSim(A, B) = \cos(Interest_A, Interest_B) = \frac{Interest_A \cdot Interest_B}{\|Interest_A\| \times \|Interest_B\|}$$

where $InterestSim(A, B)$ represents the interest similarity between users A and B, and $Interest_A$, $Interest_B$ represent the interest vectors of users A and B.

To illustrate the calculation process, an example is shown in [Figure 1: see original paper].

As shown in the figure, important users from the follow relationship network are clustered, resulting in categories 1, 2, and 3. User A follows users in categories 1 and 2, while user B follows users in categories 1, 2, and 3. Therefore, user A' s interest vector is $(1, 1, 0)$ and user B' s interest vector is $(1, 1, 1)$. After normalization, user A' s interest vector becomes $(\frac{1}{3}, \frac{1}{2}, 0)$ and user B' s interest vector becomes $(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$. The interest similarity between users A and B is then calculated using cosine distance.

By indirectly obtaining user interest points through important users in follow relationships, this method can mine user interests when custom tags are largely missing, enabling relevant recommendations. Compared with traditional algorithms that directly use user background information feature vectors to calculate similarity, this approach does not require computation for every user, resulting in relatively less time consumption and significantly lower complexity.

3.1 Interest Similarity

In social networks, users form different communities based on common interests, and users within the same community typically share similar interests. In traditional recommendation domains, researchers mine user interests from ratings of products and music. As microblogs become increasingly popular with expanding resources and data, mining user interests for recommendation becomes more important, though few methods leverage microblog user relationships for indirect interest mining. This paper mines user interests through important users in the social relationship network to calculate interest similarity.

3.2 Background Similarity

Many studies have comprehensively considered various background information. Some research calculated background similarity using user location, tag, and personal description information, while others combined gender, age, and geographic information. Building on existing studies, this paper computes microblog user background similarity based on gender, age, and location information.

Gender is often an important criterion for characterizing individuals. In the microblog domain, behavioral differences between genders are significant. Male users are generally more interested in sports, technology, and current affairs,

while female users tend to focus on beauty, entertainment, and weight loss. The gender attribute is defined as:

$$sex_U(A) = \begin{cases} 1, & \text{if user } A \text{ is male} \\ 0, & \text{if user } A \text{ is female} \end{cases}$$

where $sex_U(A)$ represents user A' s gender.

In social networks, users of different ages often exhibit significant differences. Users from different age groups typically have different experiences and concerns, resulting in lower similarity. Generally, the smaller the age difference and the lower the proportion of age difference relative to age, the closer the users'interests and the higher their similarity. The age attribute is defined as:

$$age_U(A) = \frac{age_A - age_{min}}{age_{max} - age_{min}}$$

where $age_U(A)$ represents the calculated age for user A, age_A represents user A' s actual age, age_{max} represents the maximum age value in the data, and age_{min} represents the minimum age value.

Many recommendation systems in practical applications are based on location information, such as "people nearby" recommendations. Research based on location information has achieved good results, and location-based recommendations are increasingly popular. In microblogs, user location information includes province and city. The location attribute is defined as:

$$address_U(A) = (province_A, city_A)$$

where $address_U(A)$ represents user A' s location feature information, $province_A$ represents the province where user A is located, and $city_A$ represents the city. These need to be converted to corresponding numerical values for calculation.

Based on the analysis of various background information, a user background information vector is constructed by combining gender, age, and location:

$$BI_A = (sex_U(A), age_U(A), address_U(A))$$

where BI_A represents user A' s background feature vector, and $sex_U(A)$, $age_U(A)$, $address_U(A)$ represent user A' s gender, age, and location feature information, respectively.

The background similarity calculation formula is:

$$BI_{Sim}(A, B) = \cos(BI_A, BI_B) = \frac{BI_A \cdot BI_B}{\|BI_A\| \times \|BI_B\|}$$

where $BI_{Sim}(A, B)$ represents the background similarity between users A and B, and BI_A, BI_B represent the background feature vectors of users A and B.

3.3 Comprehensive Similarity Calculation

When calculating microblog user similarity, many studies have comprehensively considered user multi-source information. This paper systematically analyzes microblog user profiles, follow relationships, interaction relationships, and interest points, proposing a comprehensive model that combines interest similarity and background similarity to mine similar users in microblog social networks. The model structure is shown in [Figure 2: see original paper].

The steps for calculating user comprehensive similarity are: a) Data preprocessing b) Calculate interest similarity and mine the N users with most similar interests c) Calculate background similarity and mine users with similar backgrounds from the N interest-similar users d) Identify users with comprehensive similarity

The value of N relates to user scale and affects result accuracy. For example, when recommending the 10 most similar users for a target user, if N is too small, the algorithm may not guarantee finding the top 10 comprehensively similar users, but if N is too large, computational complexity increases. Therefore, N should be reasonably selected based on actual conditions.

Thus, when calculating user similarity, this method first mines interest-similar users, then calculates their background similarity, hierarchically identifying the most comprehensively similar users. This approach reduces the complexity of computing feature vectors for all users, improves algorithm performance, and ensures recommendation results are relevant to user interest points.

4.1 Data Acquisition

To verify the effectiveness of the proposed model for calculating microblog user similarity, the MicroblogPCU dataset from the UCI repository (https://archive.ics.uci.edu/ml/machine-learning-databases/00323/) was used for experiments. The dataset includes 59,191 users and 142,369 follow relationships. Among them, 782 users have detailed personal information including user ID, username, gender, account level, location information, tags, number of blogs, number of followees, and number of followers. Two hundred sixty-two users have custom tags, with a total of 1,441 tags. The experiment uses their follow relationships to construct the user network, with remaining users used for testing and model verification.

4.2 Results and Analysis

First, the user follow relationship network was analyzed to obtain a user interest diagram, shown in [Figure 3: see original paper]. The figure uses ForceAtlas layout for visualization, where each node represents a user and edges represent follow relationships between two users. Users followed by more users have larger nodes. The network shows clear categories of user interests, with most users having relatively concentrated interests. Some users follow multiple domains, resulting in broader interests. For users with multiple interests, traditional CN algorithms risk excessive potential errors in recommendations, as user interests are diverse and recommending similar users based solely on common friend counts yields poor results.

The PageRank algorithm was first applied to mine important users with many followers in the network. Their tags were segmented to construct tag vectors, and important users were clustered based on these tags. The key question is how many categories to cluster important users into. Clustering validity can be evaluated by measuring consistency with reference standards or assessing clustering quality under different numbers of clusters. This experiment uses the Calinski-Harabasz (CH) index to evaluate clustering results. The CH index judges clustering quality through the compactness of within-class dispersion matrices and the separation of between-class dispersion matrices:

$$CH(k) = \frac{tr(B_k)/(k-1)}{tr(W_k)/(n-k)}$$

where k represents the number of clustering results, $tr(B_k)$ represents the trace of the between-class dispersion matrix, and $tr(W_k)$ represents the trace of the within-class dispersion matrix. Larger CH values indicate tighter within-class elements and more dispersed between-class elements, reflecting better clustering effects.

The CH values for different numbers of clusters are shown in [Figure 4: see original paper], which displays results for clustering important users into 2 to 25 categories. While the CH value is better at $k = 5$, using this to build user interest vectors results in many frequently-followed users being in the same class, yielding poor accuracy. At $k = 10$, algorithm accuracy is relatively good. Therefore, important users are clustered into 10 classes to construct ordinary users' interest vectors, though the optimal number of clusters should be determined based on actual dataset conditions.

To verify algorithm effectiveness, precision rate, recall rate, and F-measure are used as evaluation metrics:

$$Precision = \frac{|r_u \cap N_u|}{|N_u|}$$

where r_u represents the set of friends recommended to user u, N_u represents the set of friends already followed by user u, and $Precision$ represents the ratio of correctly recommended similar users to total recommended users.

$$Recall = \frac{|r_u \cap N_u|}{|r_u|}$$

where $Recall$ represents the ratio of correctly recommended similar users to the number of friends already followed by the user.

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where $F\text{-}measure$ represents the harmonic mean of precision and recall, with larger values indicating more accurate results.

Comparison algorithms include the BIBS method (comprehensive interest and background similarity), BIS method (interest similarity only), MISUR algorithm from literature [8], and Common Neighbors (CN) algorithm. The experiment selected 131 users and their follow relationship data from the dataset to verify the accuracy, recall rate, and F-measure of the proposed algorithm and comparison algorithms.

Precision rate comparison results are shown in [Figure 5: see original paper]. When recommending 5 to 25 users, the BIBS method achieves the highest precision, demonstrating that obtaining user interests from relationship networks is effective and that combining interest and background information more accurately mines similar microblog users. Compared with the MISUR algorithm, precision improves by an average of 8.1%.

Recall rate comparison results are shown in [Figure 6: see original paper]. As the number of recommended users increases, all algorithms' recall rates rise. When recommending 25 users, literature [8]'s algorithm achieves the highest recall rate, but the BIBS algorithm demonstrates relatively better overall performance, with recall rate improving by an average of 16.7%.

F-measure comparison results are shown in [Figure 7: see original paper]. As a comprehensive evaluation metric, the BIBS algorithm proves most effective, with F-measure improving by an average of 13.6% compared with the MISUR algorithm. Overall, the proposed comprehensive user similarity method based on user interests and background information outperforms comparison algorithms.

The running time of each algorithm when recommending friends for 131 users is shown in [Figure 8: see original paper]. The proposed algorithm consumes the least time compared with the MISUR algorithm and CN algorithm. The MISUR algorithm extracts features from user blogs, constructs user feature vectors, builds background feature vectors, and measures similarity using cosine distance, requiring direct computation for all users and thus relatively longer

time. The CN algorithm similarly requires comparing common friend counts for all users. In contrast, the proposed algorithm first extracts user interests from follow relationships, finds interest-similar users, then calculates background similarity among these users to hierarchically identify the most similar users for recommendation, significantly reducing time consumption and demonstrating good practicality.

---

5 Conclusion

Existing research on microblog user similarity computation comprehensively considers user multi-source information, constructs user background feature vectors, and uses cosine distance to calculate similarity. These methods suffer from excessive potential errors, high computational complexity, and neglect user interest points. To address these issues, this paper proposes a comprehensive user similarity calculation method based on user interests and background information (BIBS).

The paper first reviews relevant research on microblog user similarity computation, then introduces a method for indirectly obtaining user interest points based on follow relationships. This method uses the PageRank algorithm to mine important users that best represent user interests from follow relationships, clusters these important users, and indirectly obtains user interest points. Combined with user background information (gender, age, and location), background similarity is calculated, and a comprehensive calculation model based on user interests and background information is proposed to hierarchically mine similar users in microblogs.

Experiments on Sina Weibo datasets verify the model' s effectiveness, demonstrating significant improvements in accuracy and performance with less time consumption. Future work will focus on optimizing and improving the algorithm by mining users' short-term interests from forwarding and comment relationship networks to further enhance algorithm accuracy in calculating microblog user similarity.

References

[1] Statistical Report on Internet Development in China [R]. Beijing: China Internet Network Information Center, 2018.

[2] Meng Xiangwu, Liu Shudong, Zhang Yujie, et al. Research on social recommender systems [J]. Journal of Software, 2015, 26(6): 1356-1372.

[3] Huang Zhenhua, Zhang Jiawen, Tian Chunqi, et al. Survey on learning-to-rank based recommendation algorithms [J]. Journal of Software, 2016, 27(3): 691-713.

[4] Wu Xiaokun, Huang Yongfeng. SigRA: a new similarity computation method in recommendation system [C]// Proc of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. 2017: 148-154.

[5] Xu Zhiming, Li Dong, Liu Ting, et al. Measuring similarity between microblog users and its application [J]. Chinese Journal of Computers, 2014, 37(1): 207-218.

[6] Zheng Zhiyun, Jia Chunyuan, Wang Zhengfei, et al. Computing research of user similarity based on micro-blog [J]. Computer Science, 2017, 44(2): 262-266.

[7] Duan Xulei, Zhang Yangsen, Sun Yizhuo. Research on sentence vector representation and similarity calculation method about microblog texts [J]. Computer Engineering, 2017, 43(5): 143-148.

[8] Yao Binxiu, Ni Jiancheng, Yu Pingping, et al. Microblog user recommendation algorithm based on similarity of multi-source information [J]. Journal of Computer Applications, 2017, 37(5): 1382-1386.

[9] Pandey N. Density based clustering for Cricket World Cup Tweets using cosine similarity and time parameter [C]// Proc of India Conference. IEEE, 2016: 1-6.

[10] He Yuan, Wang Cheng, Jiang Changjun. Mining coherent topics with pre-learned interest knowledge in Twitter [J]. IEEE Access, 2017, 5(99): 10515-10525.

[11] Huang Hongcheng, Lu Weijin, Hu Min, et al. User relationships prediction algorithm with interest similarity measurement [J]. Journal of Frontiers of Computer Science & Technology, 2017, 11(7): 1068-1079.

[12] Chen Jie, Liu Xuejun, Li Bin, et al. Personalized microblogging recommendation based on dynamic interests and social networking of users [J]. Acta Electronica Sinica, 2017, 45(4): 898-905.

[13] Jain A, Gupta A, Sharma N, et al. Mining application on analyzing users' interests from Twitter [C]// Proc of International Conference on Internet of Things and Connected Technologies. 2018: 1-8.

[14] Xing Qianli, Liu Lie, Liu Yiqun, et al. Study on user tags in Weibo [J]. Journal of Software, 2015, 26(7): 1626-1637.

[15] Ma Huifang, Jia Meihuizi, Zhang Di, et al. Microblog recommendation based on tag correlation and user social relation [J]. Acta Electronica Sinica, 2017, 45(1): 112-118.

[16] Zhong ZhaoMan, Guan Yan, Hu Yun, et al. Mining user interests on microblog based on profile and content [J]. Journal of Software, 2017, 28(2): 278-291.

[17] Rong Huigui, Huo Shengxu, Hu Chunhua, et al. User similarity-based collaborative filtering recommendation algorithm [J]. Journal on Communications,

2014(2): 16-24.

[18] Cao Jiuxin, Chen Gaojun, Wu Jianglin, et al. Multi-feature based opinion leader mining in social networks [J]. Acta Electronica Sinica, 2016, 44(4): 898-905.

[19] Zou Zhiqiang, Xie Xingyu, Chao Sha. Mining user behavior and similarity in location-based social networks [C]// Proc of International Symposium on Parallel Architectures. 2016: 167-171.

[20] Ding Yong, Liu Jing, Jiang Cuiqing, et al. A study of friends recommendation algorithm considering users' preference of making friends in the LBSN [J]. Systems Engineering-Theory & Practice, 2017, 37(11): 2975-2982.

[21] Zhou Kaile, Yang Shanglin, Ding Shuai, et al. On cluster validation [J]. Systems Engineering-Theory & Practice, 2014, 34(9): 2417-2431.

Note: Figure translations are in progress. See original paper for figures.