Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 4, April 2018, pp. 691-698
DOI: 10.23940/ijpe.18.04.p12.691698

A Measuring Method for User Similarity based on Interest Topic

Yang Bai[a,b,c,*], Guishi Deng[b], Liying Zhang[d,e], and Yi Wang[a]

[a]*School of System Engineering, Eastern Liaoning University, Dandong, 118003, China*
[b]*Institute of Systems Engineering, Dalian University of Technology, Dalian, 116024, China*
[c]*Department of Computer Science, The University of Texas at Dallas, Richardson, 75080, USA*
[d]*School of Information, Liaoning University, Shenyang, 110036, China*
[e]*Information Center, Liaoning University of Traditional Chinese Medicine, Shenyang, 110032, China*

Abstract

A key problem in user relationship analysis is the identification and representation of user interest. The basis to tackle this issue is user similarity measures. In social tagging system, users collaboratively create and manage tags to annotate and categorize content for searching and recommending. Due to the contribution to reflect users' opinions and interests, tags are metadata for user similarity measures. However, there are some issues about it such as data sparseness, the user none-distinguished interest areas and relatively little consider about user influence. This article argues a similarity measure method that based on user's interest topic division. First, we construct tag clustering and divide the user community according to user interest areas. Second, we improve user similarity measurement model using social network analysis (SNA) and PageRank. Finally, the validity of the improved method about user similarity calculation is verified using del.icio.us data set. Experimental results show that the improved method gets the highest $P@N$ and sorting accuracy compared with the traditional tag-based user similarity.

Keywords: social tagging system; user similarity; SNA; tag clustering; user community

(Submitted on December 22, 2017; Revised on January 30, 2018; Accepted on March 8, 2018)

1. Introduction

User similarity is the main research problem of collaborative filtering algorithm whose performance directly affects the quality of recommender. However, limitations like parse data and lack of data cause the user similarity accuracy. With the vigorous development of user generated content (UGC) network, personalities are easy to display. Meanwhile, the problem of user interest preference, another drawback of traditional user similarity, is becoming more and more significant. The success of social tagging system like Flickr, del.icio.us and Technorati has shown that tagging is a great collaboration tool. In social tagging system, tags have the ability of representing user interest and resource content characterization, which provides a new way to improve user similarity accuracy.

We present a measuring method for user similarity based on divided interest topic. Specifically, the main contributions of this paper are summarized as follows. We propose user different-interest community division based on tag clustering in AHC method. There are some advantages, such as it can alleviate data sparsity, reduce computational complexity and avoid interest preference of users due to each user community owes a same interest topic. In addition, we establish user influence model and introduce it into user similarity measurement, which avoids the disadvantages of the traditional recommendation method, such as ignoring the relationship between users and emphasizing the similarity.

The remaining of this paper is organized as follows: Section 2 concludes related work. Then, section 3 presents a method of user similarity measurement. Next, section 4 contains a detailed experimental study and discusses various extensions. Finally, section 5 concludes this paper and outlines some possible future works.

* Corresponding author.
E-mail address: by1997@163.com

2. Related Work

In recent years, tags have been widely used in user similarity studies. Tso-sutter [13] separately calculated the tag similarity and resource similarity in a personalized recommender. Liang [9] indicated that the noise problem is due to the freedom of the tag behavior. Therefore, he puts forward to use the 'popular' tag to measure user similarity and improve the recommender accuracy. Kim [6] used tags to filter user preferences and develop an algorithm to solve sparse data and 'cold start' user. In these studies, the similar neighbors of the target user were obtained by combining the users' rates. Compared to scoring data, the tag is more semantic features to reflect the user's real interest. Therefore, some studies use relevance among user, tag, and resource to calculate the user similarity, regardless of users' rating for resources. Hung [5] recommended based on the implicit score by calculating the similarity considering the number of the same users and tags between users. Vig [14] conducted a user study exploring the roles of tag relevance and tag preference in promoting effective tagsplanations.

There are three main problems in the above research. The tag data matrix is often sparse, which results in reducing similarity accuracy due to its own arbitrariness, ambiguity and other characteristics. Another problem is that user interest areas have not been distinguished. The reason is that user's interests are often multifaceted, belonging to different areas. For instance, user x is interested in both animated movies and war movies, while user y is interested in war movies. According to the traditional collaborative filtering algorithm, user x will recommend animated movies to similar user y. However, user y is not interested in such topics. It leads out a recommended bias. We found that comparing the similarity of two users is a prerequisite by analyzing the essential causes. It is reasonable to calculate two-user similarity in the field that both users are interested in. Recent research like Zong [8] presented a common co-occurrence group similarity based on measuring the semantic relevance between tags. The third problem is that there is relatively little consideration about user influence in measuring similarity. Studies often ignore the relationship among users and emphasize the similarity. In fact, user influence is the key characteristic of the importance of social network system, and is also with great value for information retrieval and recommendation. Therefore, it is necessary to consider the user influence index in the measuring of user similarity.

3. User Similarity Measure Method

3.1. User Community Division Based on Tag Clustering

The conceptual model of the social tagging system proposed by Marlow [10] discusses the relationship among users, resources and tags. Users assign tags to a specific resource. Tags are represented as typed edges connecting users and resources. Resources maybe also be connected to each other and users may be associated by a social network. We propose two definitions as follows:

- Definition 1. Social tagging system model is written as $D=(U,R,T,J)$. $U=\{u_1,u_2,...,u_n\}$ is a collection of users; $R=\{r_1,r_2,...,r_m\}$ is a collection of resources; $T=\{t_1,t_2,...,t_k\}$ is a collection of social tags; $A:\{a=(u_i,r_j,t_s)|u_i\in U,r_j\in R,t_s\in T\}$ is the connections among users, resources and tags.

- Definition 2. Social tagging network is written as $G=\{T,L,W\}$. Where T is the tag node set, $L=\{l_{ij}|i,j=1,2,...k\}$ is a set of relationship among tag nodes, and $W=\{w_{ij}|i,j=1,2,...k\}$ represents the relationship weights among tag nodes.

3.1.1. Tag Clustering

In social tagging system, users are free to apply any tag to a resource, often resulting in a large number of tags that are ambiguous, redundant, or idiosyncratic. Data clustering is a common technique for statistical data analysis, which provides a means to overcome this problem, and thus reducing noise. At present, clustering algorithms are as follows: partition clustering (e.i. k-means [7]), layer clustering (e.i. agglomerative hierarchical clustering [3], AHC) and density clustering [2], et al. We adopt hierarchical clustering because of its hierarchical structure that can well show the tag association. At the same time, in order to describe the agglomerative topic, we also introduce Co-word Analysis techniques. Co-word Analysis is a topic analysis technique proposed by Callon [1]. It analyzes the frequency of word co-occurrence, and gathers close words into a cluster. Pan [11] designed a social tag network based on co-occurring tags, and proposed a novel graph-based clustering algorithm in order to identify densely connected semantic communities.

Tag co-occurrence refers to when two tags appear (i.e. x, y) in a same type of resource at the same time. The frequency of two tags in the same type of resource is called co-occurrence frequency. The larger the number of co-occurrence tags is,

the stronger the semantic relationship among tags is. We introduce Jaccard Coefficient to construct the co-occurrence similarity coefficient as shown in Equation (1).

$$Sim(x,y)=\frac{|x\cap y|}{Max(x,y)} \tag{1}$$

where $|x\cap y|$ indicates the number of co-occurrence between tag x and tag y, and $Max(x,y)$ indicates the maximum number of resource tagged by x or y. Next, we define tag-clustering set as follows:

- Definition 3. Tag-clustering set is represented as $TC=\{T_1,T_2,...,T_r\}$. Where $T_r=\{t_j|\, j\in[1,k]\}$ is a tag clustering, and t_j is a tag element in a cluster.

AHC algorithm is initialized with each original data point as a separate cluster. Then, two clusters are merged as maximum similarity every time until all ones are merged into a large class. The similarity and distance of clustering are the basis of hierarchical clustering. We use the above-mentioned co-occurrence tag coefficient, and adopt Single-Linkage (SL) method to calculate the distance between the new cluster and all the old ones in merging clusters iteratively. That means distance of SL is the maximum similarity between any object in one cluster and any one in the other cluster. The specific algorithm is as follows:

- Step 1. Initialize each tag in the input data set as a cluster.
- Step 2. Calculate the similarity coefficient of co-occurrence tag between a pair of tags, written $Sim(x,y)$, and construct initial similarity matrix.
- Step 3. According to the similarity matrix, two clusters with maximum similarity are merged so the new cluster contains all co-occurrence tag pairs of the old cluster.
- Step 4. Calculate the similarity between the new cluster and the current cluster, and update the similarity matrix.
- Step5. Repeat the above steps until only one cluster is left or the terminated conditions are met.

3.1.2. User Community Division

After clustering social tags together, we get different interest topics. According to it, we will divide the whole user space into communities. Thus, interests of users in the same community are similar. On the contrary, interests of users in the different community differ widely. We propose definition 4 and then definition 5 combined with symbols from definition 1, 3 and 4.

- Definition 4. User community collection is written as $UC=\{U_1,U_2,...,U_s\}$. U_s is a collection of users within a community.

- Definition 5. User community tag clustering mapping model represented with a quaternion $E=(UC,TC,B)$. Where $B:\{b=(u_i,t_j)|u_i\in U_s,t_j\in T_r\}$ refers to the relationship between users and tags, and $B\subset A$.

Followed by initial division for user community, the relevance strength between users and tags is to be discussed in the later section. The simple division method is that the users who use the tags in the same tag clustering are divided into a community. That is, if user u_i uses a tag in tag cluster T_j, the user u_i is divided into U_j corresponding T_j. However, because of the huge amount of data in the social tagging system, we refer to F-measure of the accuracy rate and the recall rate to indicate the relevance between user tag sets and the tag clustering for determining the division of the user community, as shown in Equation (2).

$$P=\frac{\sum_{t_j\in T_r}u_i(t_j)\cap T_r(t_j)}{\sum_{t_j\in T_r}T_r(t_j)},\quad R=\frac{\sum_{t_j\in T_r}u_i(t_j)\cap T_r(t_j)}{\sum_{t_j\in T_r}u_i(t_j)},\quad F(u_i,T_r)=\frac{(\alpha^2+1)P\times R}{\alpha^2\times P\times R} \tag{2}$$

where $F(u_i,T_r)$ indicates the relevance of user u_i and tag cluster T_r, α is the harmonic coefficient, $u_i(t_j)$ indicates the tag set of u_i, t_j is any one of the tags, $T_r(t_j)$ is the set of tags for T_r, and t_j is any one of the tags. Let the index threshold be θ. If $F(u_i,T_r)\geq\theta$, then u_i will be divided into U_r

3.2. User Similarity Measure Based on SNA

For each user within a well-divided user community, it is necessary to calculate the associated degree between each tag and each tag to establish a user-tag relevance matrix. Next, the user relationship matrix is to be established and the user influence introduced into the measuring model for user similarity is calculated.

3.2.1. User Relationship Matrix

As mentioned earlier, users were divided into interest-topic communities. In an interest-topic community, let user set be $U=\{u_i|i=1,2,...,m\}$, and tag collection be $T=\{t_j|j=1,2,...,n\}$. Because each user has its own tag set, according to the use of the tag features, we adopt VSM to represent each user's feature vector as shown in Equation (3).

$$u_i=\{(t_1,r_{i1}),(t_2,r_{i2}),...,(t_j,r_{ij}),...,(t_n,r_{in})\} \tag{3}$$

where r_{ij} indicates relevance between tag t_j and user u_i, which usually obtained from frequency that a user uses a tag or TF-IUF [15]. A user-tag relevance matrix is constructed for multiple users in a community. Using row u_i as a row vector. It is represented by a matrix R of $m\times n$ as shown in Equation (4). Element r_{ij} in the matrix R equals $Rel(u_i,t_j)$ (Equation 5) calculated through TF-IUF.

$$R=(r_{ij})_{m\times n},\quad r_{ij}=Rel(u_i,t_j), i=1,2,...,m; j=1,2,...,n \tag{4}$$

$$Rel(u_i,t_j)=TF(u_i,t_j)\times IUF(u_i,t_j)=\frac{w(u_i,t_j)}{\sum_{t_k\in T}w(u_i,t_k)}\times\lg\frac{\sum_{u_k\in U}\sum_{t_s\in T}w(u_k,t_s)}{\sum_{u_k\in U}w(u_k,t_j)} \tag{5}$$

where $TF(u_i,t_j)$ denotes the tag frequency, and $IUF(u_i,t_j)$ represents the inverse-tag-word frequency. If t_j is only used by u_i, t_j would be more influencing to u_i, and vice versa. $w(u_i,t_j)$ is the number of times that u_i uses t_j.

In fact, the closer the two users' behavior is, the more similar their interests are. We apply the concept from Hammock [12] to indicate strength of the relationship between the two users with similarity of tag usage behavior. We replace the same factor addition in matrix (transformation of 2-mode to 1-mode in SNA) with Vector Cosine Formula. The cosine distance u_{ij} between u_i and u_j is the element at the corresponding position of R_{U2U}. Combined with r_{ij} form Equation (3), user relationship matrix R_{U2U} is indicated as Equation (6).

$$R_{U2U}=(u_{ij})_{m\times m},\quad u_{ij}=\frac{\sum_{k=1}^{n}r_{ik}\times r_{jk}}{\sqrt{\sum_{k=1}^{n}r_{ik}^{\,2}}\times\sqrt{\sum_{k=1}^{n}r_{jk}^{\,2}}} \tag{6}$$

The usage of same tags represents similar users' interest, but which does not represent how much users' interest is. We measure the user interest with tag frequency. R_{U2U} denotes the similarity of different user interest, each $u_{ij}(i\neq j)$ denotes the TF-IUF similarity of the tags used by u_i and u_j. Especially, let $u_{ij}(i=j)$ in the diagonal equal to 0, and R_{U2U} be a symmetric matrix ($u_{ij}=u_{ji}$).

The user relationship matrix R_{U2U} is further standardized. We learn from R_{U2U} that $\sum_{j=1}^{m}u_{ij}$ is obtained from accumulating off-diagonal elements u_{ij} in row. Then, standardized user relationship matrix R_{U2U}' is combined with Equation (6) as shown in Equation (7).

$$R_{U2U}'=(s_{ij})_{m\times m},\quad s_{ij}=\begin{cases}\dfrac{u_{ij}}{\sum_{j=1}^{m}u_{ij}} & \sum_{j=1}^{m}s_{ij}=1, i,j=1,2,...,m, i\neq j\\ 0 & i=j\end{cases} \tag{7}$$

where s_{ij} indicates the degree of relevance between user u_i and u_j. s_{ij} and s_{ji} are not necessarily equal. R_{U2U}' not only reflects the user similarity, but also reflects the different ability to tag about users.

For example, R_{U2T} represents a user-tag relevance matrix consisting of five users and seven tags as shown in Table 1. It is converted to the user relationship matrix R_{U2U} as shown in Table 2. Then, it is standardized to the matrix R_{U2U}' as shown in Table 3.

3.2.2. User Influence Model

After obtaining the standardized user relation matrix R_{U2U}', we adopt UserRank method to analyze the relationship among users. In the following discussion, a user influence model is established.

Table 1. R_{U2T}

	t_1	t_2	t_3	t_4	t_5	t_6	t_7
u_1	.1	0	.2	.5	0	0	.2
u_2	0	.1	.2	0	.3	0	0
u_3	.4	.2	0	.3	0	.6	0
u_4	.2	0	0	0	0	0	0
u_5	0	.1	0	.4	.3	.1	0

Table 2. R_{U2U}

	u_1	u_2	u_3	u_4	u_5
u_1	0	.1	.3	.1	.5
u_2	.2	0	.1	0	.7
u_3	.3	0	0	.3	.3
u_4	.3	0	.7	0	0
u_5	.4	.3	.3	0	0

Table 3. R_{U2U}'

	u_1	u_2	u_3	u_4	u_5
u_1	0	.1	.3	.1	.5
u_2	.2	0	.1	0	.7
u_3	.3	0	0	.3	.3
u_4	.3	0	.7	0	0
u_5	.4	.3	.3	0	0

A weighted directed network G is generated with R_{U2U}', which is the adjacency matrix and described as Definition 6:

- Definition 6: $G=\{U,E,W\}$ is user relationship network, where $U=\{u_i|i=1,2,...,m\}$ denotes a collection of user nodes, $E=\{e_{ij}|i,j=1,2,...m\}$ denotes a set of edges between user nodes, $W=\{s_{ij}|i,j=1,2,...m\}$ denotes a set of weights on the edges between user nodes. The relationship between u_i and u_j $R(u_i,u_j)$ is shown as Equation (8).

$$R(u_i,u_j)=\begin{cases} e_{ij} \text{ exists }, \text{ if } s_{ij}\neq 0 \\ e_{ij} \text{ not exists }, \text{ if } s_{ij}=0 \end{cases} \tag{8}$$

G denotes a social network composed of user relationship, whose link structure contains much important implicit information. It is beneficial to evaluate the influence of user nodes in the network. According to the PR algorithm, we calculate user influence index through analyzing the diffusion and attenuation mechanism. Introduced weight s_{ij} on edge e_{ij} into PageRank algorithm [4], improved UserRank (UR) model in Equation (9) is get to calculate the user's influence index. As seen from $\sum_{j=1}^{m} u_{ij}=1$ in Equation (7), we can get $OL(u_j)=1$ in Equation (9). Thus, Equation (9) is reduced to the Equation (10):

$$UR(u_i)=(1-\alpha)\times\lambda+\alpha\times\sum_{u_j\in IL(u_i)}\frac{PR(u_j)}{OL(u_j)}\times s_{ji} \tag{9}$$

$$UR(u_i)=(1-\alpha)\times\lambda+\alpha\times\sum_{u_j\in IL(u_i)}PR(u_j)\times s_{ji} \tag{10}$$

where the relevant parameters and terminated conditions of the algorithm are as follows. Let attenuation factor α=0.85, which is empirical value, and let λ obey uniform distribution as $\lambda=1/|U|$, where $|U|$ is the number of users in the community. The terminated condition that is the difference between the n-1th iteration and the nth iteration equals to 0.001.

3.2.3. Improved Measuring Model for User Similarity

Similarity measuring method is generally used Cosine-based similarity or Pearson correlation. From Equation (3) we can see that the relevance between user and the tag can be represented through row vector u_i of matrix R. We adopt Pearson coefficient to measure user similarity as shown in Equation (11).

$$Sim(u_i,u_j)=\frac{\sum_{k=1}^{n}(r_{ik}-\bar{r}_i)\times(r_{jk}-\bar{r}_j)}{\sqrt{\sum_{k=1}^{n}(r_{ik}-\bar{r}_i)^2}\times\sqrt{\sum_{k=1}^{n}(r_{jk}-\bar{r}_j)^2}} \tag{11}$$

Furthermore, we introduce UR model from Equation (10) into Equation (11) to get a measuring model for user similarity as shown in Equation (12).

$$Sim(u_i,u_j)'=UR(u_j)\frac{\sum_{k=1}^{n}(r_{ik}-\bar{r}_i)\times(r_{jk}-\bar{r}_j)}{\sqrt{\sum_{k=1}^{n}(r_{ik}-\bar{r}_i)^2}\times\sqrt{\sum_{k=1}^{n}(r_{jk}-\bar{r}_j)^2}} \tag{12}$$

4. Experiment and Analysis

Our experiments with real data from del.icio.us demonstrate the practical utility of our equations and the corresponding algorithms. Experimental dataset comes from hetrec2011-delicious-2k provided by Communitylens [16]. It includes 1867 users, 53388 tags, and 104,799 bookmarks with 69226 resource URLs. Specifically, two similarities methods (considering UR and none-considering UR) are compared with traditional tag-based user similarity. The content of our experiments include tag clustering, standard user similarity sequence, and user similarity comparison.

4.1. Tag Clustering Experiment

First, we obtain co-occurrence analysis with user-tag dataset, and get 106477 tag pairs. Any one of whose frequency of co-occurrence is more than twice even 65% of these tag pairs co-occur below 11 times. In order to facilitate testing, we filter out the co-occurrence tag pairs below 11 times and randomly tests 3000 pairs from the remaining tag co-occurrence data sets. Then, AHC tree is constructed starting with ranking tags in descending of co-occurrence frequency. The number of threshold cluster is set to 20. Thus, we acquire 20 tag clusters. A part of the results is shown in Figure 1. It can be seen that tags are relatively classed into the 20 clusters. We also found that a tag can be classed into multiple clusters in accordance with a reality of a resource with a variety of attributes, and a tag tagging multiple resources.

roots	nodes
design	design webdesign tools graphics informationdesign jobs faceb
book	books p2p california search historia photoshop book-rev
web	money 2010 html5 ajax html web software lego gem web_to
history	history chinese_history automobiles_history illustration (
culture	culture course diverse reference math reading thinking m
education	education education_games virtual_teenage blog youth_t
movies	movies media story_telling awarenessnetwork 21stcentur
music	music newmusic musicales jazz classic mp3 k-pop gratuit

Figure 1. Agglomeration hierarchical clustering results

4.2. Standard User Similarity Sequence

For each user in user community, let $u\in U$, $v\in U-\{u\}$. There are two steps in this section. We establish user-resource relevance matrix from dataset 'user-taggedbookmarks', then calculate the matrix element according to TF-IDF and introduce it into Equation (11). Thus, user-resource relevance $SimOnItem(u,v)$ is obtained. Next, we establish user-friend relevance matrix from dataset 'user-contacts'. The matrix element that introduce into Equation (11) is 1 if u and v are in friendship, otherwise is 0. Thus, user-resource relevance $SimOnContacts(u,v)$ is obtained. Standard User Similarity Set between u and v is obtained by combining the two relevance indexes into $Relation(u,v)$ as shown in Equation (13. According to the value of $Relation(u,v)$, v is ranked in descending order. Thus, we get an order-user-set based on resource and friendship. Standard User Similarity Sequence is written as $R(u)$.

$$Relation(u,v)=\alpha\times SimOnItem(u,v)+\beta\times SimOnContacts(u,v) \tag{13}$$

where α denotes resource relevance weight, β denotes relational relevance weight, and $\alpha+\beta=1$.

4.3. User Similarity Comparison

For each user in user community, let $u\in U$, $v\in U-\{u\}$. We calculate similarity between u and v according to user similarity algorithm to be tested, and then get the most similar ranked-user set which is called similar sequence $S(u)$ of u.

Next, we regard $S(u)$ as the object to be tested, and $R(u)$ as the standard. Measuring Model for User Similarity is evaluated through comparing the difference between $S(u)$ and $R(u)$.

4.3.1. Evaluation index

We adopt evaluation indexes like $P@N$ and Rank Accuracy in the field of information retrieval to compare the differences between the two sequences $S(u)$ and $R(u)$, in order to evaluate user similarity model. $P@N$ is shown as Equation (14) :

$$P@N = \frac{1}{|U|}\sum_{u\in U}\frac{S_1^N(u)\cap R_1^N(u)}{N} \tag{14}$$

where the subset consisted of the first N elements of $S(u)$ and $R(u)$ is denoted as $S_1^N(u)$ and $R_1^N(u)$, $N = 5\times i, i = 1,2,...,10$. As the index $P@N$ does not evaluate the user's ranking, this article uses the Rank Accuracy to evaluate the user similarity algorithm as shown in Equation (15):

$$Accuracy = \frac{1}{|U|}\sum_{u\in U}\frac{1}{|R(u)|}\times\sum_{k=1}^{|F(u)|}\frac{1}{1+|S(u)\_rank_k - R(u)\_rank_k|} \tag{15}$$

where for each user k in $R(u)$, the ranking in which it appears in $R(u)$, $S(u)$ is recorded as $R(u)\_rank_k$, $S(u)\_rank_k$.

4.3.2. Experimental Setup

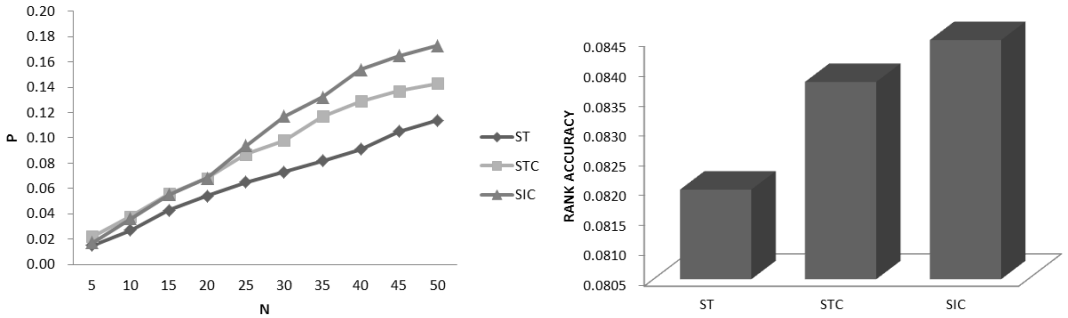
The relevance between resource relevance weight α and relational relevance weight β is obtained through Hierarchical Discriminant Matrix (HDM) method. HDM is in Table 4. Let resource relevance α = 0.167, and relevance weight β = 0.833.

Table 4. Standard similarity discrimination matrix

Standard similarity	User friend relevance	User resource relevance
User friend relevance	1	1/5
User resource relevance	1/5	1

4.3.3. Experimental Result

We calculate three algorithms: traditional similarity on tag [13] (ST), user similarity on tag clustering (STC), and similarity consider user-influence community (SIC). The experimental results of the similarity degree in P@N and Rank Accuracy are given respectively as shown in Figure 2(a) and (b). The experimental results show that SIC algorithm is the best. It can be seen that SIC algorithm significantly outperforms the other models. It shows that user communities can be generated according to tag-clustering method, which can help find appropriate neighborhoods for the target user. Moreover, it also reveals that user influence model can efficiently improve measuring model for user similarity.



(a) $P@N$ experimental results (b) Ranking accuracy test results

Figure 2. Comparison results of user similarity

5. Conclusions

The research method takes into account the diversity of user interests and uses tag clustering as the basis of user's interest topic. Introduced concepts of social network, it makes user similarity more adaptable to the current situation of online social network development. Moreover, due to the compact connection in a community and sparse connection between communities, it is helpful for alleviating sparsity effectively. In addition, by considering users' annotation behavior to analyze relevance between users and tags instead of users' subjective evaluation, the reliability and consistency of users

were improved. In this paper, we also found that the user influence is very important to measure users' similarity, so we proposed the improved similarity model considering user influence. It could increase each user's contribution to the similarity algorithm and would be helpful for deeply understanding the applicability of similarity. For future work, we plan to extend our tag-clustering algorithm to the semantic level for corresponding to meaningful topic domains. We also hope to propose a method for web services recommendation for further evaluate our work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 71372083). The authors thank the anonymous reviewers for their helpful suggestions.

References

1. M. Callon, J. P. Courtial, and F. Laville, "Co-word Analysis as a Tool for Describing the Network of Interactions Between Basic and Technological Research: the Case of Polymer Chemsitry", *Scientometrics*, Vol. 22 No. 1, pp. 155-205, 1991.
2. Y. X. Chen, R. Santamaria, A. Butz, and R. Theron, "TagClusters: Semantic Aggregation of Collaborative Tags Beyond TagClouds", in *International Symposium on Smart Graphics*, pp. 56-67, 2009.
3. J. Gemmell, A. Shepitsen, B. Mobasher, and R. Burke, "Personalizing Navigation in Folksonomies Using Hierarchical Tag Clustering", in *International Conference on Data Warehousing and Knowledge Discovery*, pp. 196-205, 2008.
4. T. H. Haveliwala, "Topic-sensitive PageRank", in *International Conference on World Wide Web*, pp. 517-526, 2002.
5. C. C. Hung, Y. C. Huang, Y. J. Hsu, and K. C. Wu, "Tag-Based User Profiling for Social Media Recommendation", *AAAI Workshop - Technical Report*, 2008.
6. H. N. Kim, A. T. Ji, I. Ha, and G. S. Jo, "Collaborative Filtering Based on Collaborative Tagging for Enhancing the Quality of Recommendation", *Electronic Commerce Research & Applications*, Vol. 9 No. 1, pp. 73-83, 2010.
7. S. S. Kumar, and H. H. Inbarani, "Web 2.0 Social Bookmark Selection for Tag Clustering", in *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pp. 510-516, 2013.
8. H. Z. Li, X. G. Hu, Y. J. Lin, H. E. Wei, and J. H. Pan, "A Social Tag Clustering Method Based on Common Co-occurrence Group Similarity", *Frontiers of Information Technology & Electronic Engineering*, Vol. 17 No. 2, pp. 122-134, 2016.
9. H. Liang, Y. Xu, Y. Li, and R. Nayak, "Collaborative Filtering Recommender Systems Based on Popular Tags", *Adcs Proceedings of the Fourteenth Australasian Document Computing Symposium*, 2009.
10. C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read", in *Hypertext 2006, Proceedings of the ACM Conference on Hypertext and Hypermedia, Odense, Denmark*, pp. 31-40, August 2006.
11. W. Pan, S. Chen, and Z. Feng, "Automatic Clustering of Social Tag Using Community Detection", *Applied Mathematics & Information Sciences*, Vol. 7 No. 2, pp. 675-681, 2013.
12. X. Su, and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques", *Advances in Artificial Intelligence*, Vol. 2009 No. 12, pp. 4, 2009.
13. K. H. L. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme, "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms", *Acm Symposium on Applied Computing*, pp. 1995-1999, 2008.
14. J. Vig, S. Sen, and J. Riedl, "Tagsplanations: Explaining Recommendations Using Tags", in *International Conference on Intelligent User Interfaces*, pp. 47-56, 2009.
15. S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring Folksonomy for Personalized Search", in *International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore*, pp. 155-162, July 2008.
16. "Hetrec2011-delicious-2k", Available at *https://grouplens.org/datasets/hetrec-2011/*, Last accessed on February 1, 2018.

Yang Bai graduated from the School Information, Liaoning University, for the degree of Bachelor, Master, and now is a Ph. D. student at Dalian University of Technology. She visited the University of Texas at Dallas in the U.S.A. as a visiting scholar from 2015 to 2016. Now she is an associate professor of school of System Engineering, Eastern Liaoning University, Dandong, China. Her current research interests include data mining, information service and intelligent business.

Guishi Deng is a professor, doctoral supervisor of Dalian University of Technology. His research interests include intelligent business and complex system analysis.

Liying Zhang is a master student in School Information, Liaoning University. Her research interest is data mining.

Yi Wang is an associate professor of school of System Engineering, Eastern Liaoning University. Her research interests include intelligent business and intelligent information processing.