An Information Gain Ratio based Discovery of User Similarity in *Sina* Blog Community

Wei Ren[†]
Computer and Information Science
Southwest University
Chongqing China
oicq@swu.edu.cn

Yepeng Qiu
College of Physical Education and
Health Science
Chongqing Normal University
Chongqing China
qiuyepeng@163.com

Xianghua Li
Computer and Information Science
Southwest University
Chongqing China
lxhgc@swu.edu.cn

ABSTRACT

Researchers have already found that the interaction data on Blogs bulletin board and social network are the reflection of human and human society. It can be used to build a user profile and tell the researchers what a network community is composed of. In previous studies we have noticed that users who are friends in real life tend to interact with each other with high frequency and the characteristics of their blogs shares more similarity, compared with that of strangers. It brings the idea that can we predict the behavior as well as personal value of a total new comer by the friends around him/her?

In this paper we proposed a method to classify the sub-community groups into even smaller and more alike ones by applying information gain ratio based decision tree to users' tags as well as behavior patterns. First we recognize the social network groups by community discover algorithm with *Sina* blog-oriented community definition and then further analyze the topology of the groups by frequency subtree mining to discover the real closely collected small groups. Then we collect the users' tags by text mining and assign every user a 10 words long tag set. Finally we use the information gain ratio decision tree to catalogue users into even smaller groups with greater similarity. Results show that our method has a higher accuracy.

CCS CONCEPTS

• Human-centered computing~ Social networks • Human-centered computing~ Blogs • Human-centered computing~ Empirical studies in collaborative and social computing

KEYWORDS

Information gain ratio, classification, social network users

ACM Reference format:

Wei Ren, Yepeng Qiu and Xianghua Li. 2018. An Information Gain Ratio based Discovery of User Similarity in Sina Blog Community. In *Proceedings of 2018 International Conference on Algorithms, Computing and Artificial Intelligence (ACAI'18). Sanya, China, 7 pages.* https://doi.org/10.1145/3302425.3302449

1 Introduction

"Birds of a feather flock together". We will finally find ourselves with people who are similar with ourselves. We make friends because we love the same movies, go out to the same restaurants, have same routines or support the same political leaders.

McPherson *etc.* [1] hold the belief that people's personal networks are homogeneous, that is, people choose their mate, friends based on similarity. Even the relationships in their work, information transfer, co-membership etc. are homogeneous. Homophily limits people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience. For instance, People of similar interests tend to make much more identical choices. Friends will go to the same place, share same routines and choose the same movie if they were together or not. The online social networks have been a part of people's everyday life [2][3][4]. Services like Twitter, Facebook, YouTube and *Sina* blog attracted tens of billions of users sharing and exchanging messages [5][6]. And have changed the way of living. We believe that the connections the user's online social network have truly reflect the inner deep of the user[7][8]: how he/she believe, how he/she sees the world, how he/she interprets others' behaviors and responds as well as the way he/she receives the outside messages. By observing one's close friends' behavior pattern, can we predict one's choices through the study of his/her close contacts?

Groups sharing similarities will tend to make the same choices or own common characteristics [9] [10]. The similarity study of social network users lies in three major benefits: First it enhances business marketing campaigns. More and more business companies are analyzing users' behavioral data streams in order to improve sales, improve advertising and target customer needs. Second it offers a better way to monitor public heathy. Similarity in users' routines may demonstrate the similarity in their physical condition and mental health. Finally, it provides government a

[†]Corresponding author: oicq@swu.edu.cn

ACAI '18, December 21–23, 2018, Sanya, China
© 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6625-0/18/12…$15.00
https://doi.org/10.1145/3302425.3302449

direct way to supervise the emerge as well as cognation of public opinion.

2 Related Works

Researchers have done a lot of studies on users' classification and new users promotion methods. Agarwal *etc.* [11] introduced a way to identify familiar strangers by the contacts of a user. These familiar strangers refer to individuals who are not directly connected but exhibit some similarity (hobbies, careers, locations *etc.*). The blogger and citation network are used to showcase technical details and empirical results are collected [12]. Also, they employed contextual information and collective wisdom to aggregate similar bloggers in another research. Wang *etc.* [13] proposed to use collective wisdom from the crowd or tag networks and measure the similarity between pairs of tags to find the similar users. The proposed tag network approach achieves 108% and 27% relative improvements on the *BlogCatalog* dataset respectively.

Studies on other network system beside blog are also carried on how users of similar interests are connected. Li *etc.* [14] proposed a person representation method which uses a person's website to represent this person. The designed matching process takes person representation into consideration to allow the same representation to be used when composing the query. Ma *etc.* [15] proposed an approach for conquering the sparseness of behavior pattern space and discover similar mobile users with respect to their habits by leveraging behavior pattern mining. The experiments conducted on real data sets show that our approach outperforms three baselines in terms of the effectiveness of discovering similar mobile users with respect to their habits. Xiao *etc.* [16] models a user's GPS trajectories with a semantic location history then measure the similarity between that of different users' by using maximal travel match (MTM) algorithm. Kwak *etc.* [17] collected tags from six popular web services, and analyzed usage patterns, then connect users of similar interests based on user-labeled tags.
Some of the mentioned studies ignore the background and history context of the users' blog and others did not pay attention to the attributes other than users' tags, or based on a single collection of location history. The user features are mainly context based yet involved very a few other attributes. On the other hand, the works are focused mainly on the similarity of the users but did not pay much attention to the minor differences between closely interacted users by which may result major diversities in choices.

The work of this paper is based on *Sina* blogs and we pay attention to both user's personal descriptions and the contexts of their blogs. And we give more attention to the minor differences in similar users of a detected community. The community will be then further analyzed using an information gain ratio based decision method to catalogue users into smaller, more similar groups.

3 Information Gain Ratio Based Classification

The human behaviors are complicated and human decision is the result of multiple factors[18]. Thus, researches on similarity of online users should not be a single focus of one source but to take as many sources as possible into consideration.

We proposed an information gain ration based classification method on users' blogs. First we discover the community (closely interacted group) using *Sina* blog oriented community definition. The topology of studied network is semantic connections between blog subscribers and blogs been subscribed, as well as the "@" behavior between two users. Then an information gain ratio based decision process is carried using the keywords and pattern attributes to divide community users into even smaller groups with closest connections and highest similarity.

3.1 Semantic based Community Discovery in *Sina* Blog

In early researches [19][20], we have proposed a semantic online community discover algorithm on *Sina* blog network. The *Sina* blog is the national social network platform for message publishing and information interchanging. It is a real-time information network for sharing and discovering events around China. It enables users to send and read text-based messages. These messages are displayed on the author's profile page and delivered to the author's subscribers. Each user is a node in the network and the subscribe (out-degree) and being subscribed (in-degree) connections forms the topology of the network. Then we define the community as follows:
Suppose G is the topology of a social network, G_1, G_2, …, G_n are the subgraphs in network G, $G = G_1 \cup G_2 \cup ... G_n$, then G_i is a community in G, if $G_i \subset G$ follows:

(1) for any $v \subset G_i$, $d_{vj}(v) - d_{in}(v) = 0$ is the degree of node v in G_i , $d_{in}(v)$ is the sum of degree node v with other subgraphs except G_i . Then for other subgraphs G_j we have $d_{vj}(v) - d_{in}(v) \leq 0$;

(2) the average similarity of nodes in G_i is higher than the average similarity between subgraphs.

The modified definition simplified the find of community and can be adapted to networks of different scale. When v is an overlapped node, $d_{vj}(v) - d_{in}(v) = 0$.

Let C_1 and C_2 are two communities in G, $G' = C_1 + C_2$, A is the adjacency matrix of G' , n is the number of nodes in G' . Assume nodes' value in C_1 is 0, nodes' value in C_2 is 1, then G' has $Vector_{XG'} \in$ (0,1). Then for node v_i, the total connections $d_{out}(i)$ to other communities is

$$d_{out}(i) = (2x_i - 1)\left[x_i \sum_{j=1}^{n} A_{ij} - \sum_{j=1}^{n} (x_j A_{ij}) \right] \tag{1}$$

the inner connections within local community is

$$d_{in}(i) = (1-2x_i)\left[(1-x_i)\sum_{j=1}^{n} A_{ij} - \sum_{j=1}^{n}(x_j A_{ij})\right]$$

(2)

then the difference is

$$dif(i) = d_{out}(i) - d_{in}(i)$$

$$= (2x_i - 1)\left[x_i \sum_{j=1}^{n} A_{ij} - \sum_{j=1}^{n}(x_j A_{ij})\right] - (1-2x_i)\left[(1-x_i)\sum_{j=1}^{n} A_{ij} - \sum_{j=1}^{n}(x_j A_{ij})\right]$$

$$= (2x_i - 1)\left[\sum_{j=1}^{n} A_{ij} - 2\sum_{j=1}^{n}(x_j A_{ij})\right]$$

(3)

And we have G' node's vector *X* and unit vector *I*:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_i \\ . \\ x_n \end{bmatrix}, \qquad I = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix} \qquad (4)$$

In (4), $x_i \in \{0, 1\}$;

Then (3) is

$$dif(X) = (2X - I)\cdot(AI - 2AX)$$

(5)

Referring the definition of community, we have

$$dif(X) = (2X - I)\cdot(AI - 2AX) \le 0$$

(6)

$d_{if}(X)$ can be supplementary identification of community partition, $d_{if}(X)$ >0 may suggest the nodes are wrongly cataloged. The inner closeness of community is calculated by nodes' similarity. We experimented the *Sina* blog oriented definition on Dolphin social network and the result are demonstrated in Table 1.

Table 1: Communities Found in Dolphin Social Network

Community 1										
node	2	6	7	8	10	14	18	20	26	27
$d_{vj}(v) - d_{in}(v)$	-4	-4	-6	-1	-7	-8	-9	-2	-3	-3
node	28	32	33	42	49	55	57	58	61	
$d_{vj}(v) - d_{in}(v)$	-5	-1	-3	-5	-1	-7	-2	-7	-1	

Community 2										
node	1	3	4	5	9	11	12	13	15	16
$d_{vj}(v) - d_{in}(v)$	-6	-4	-3	-1	-6	-5	-1	-1	-12	-7
node	17	19	21	22	24	25	29	30	31	34
$d_{vj}(v) - d_{in}(v)$	-6	-7	-9	-6	-3	-6	-3	-9	-1	-10
node	35	36	37	38	39	40	41	43	44	45
$d_{vj}(v) - d_{in}(v)$	-5	-1	-5	-11	-8	-0	-6	-6	-7	-1
node	46	47	48	50	51	52	53	54	56	59
$d_{vj}(v) - d_{in}(v)$	-11	-2	-6	-2	-7	-10	-4	-2	-2	-1
node	60	62								
$d_{vj}(v) - d_{in}(v)$	-5	-3								

We discovered two communities in the dolphin society network of sixty-two dolphins. The network has 162 edges and the result is coherent.

3.2 User Characteristics Collection

The *Sina* blog has a 140 words limit and provide user with certification service[21]. Once you are certificated, the message you published will attract more subscribers. Then we choose a medium size of discovered community of 142 users with a higher percentage of certificated users (36.6%). Then we use crawler to collect 142 bloggers' profile data and their 3920 subscribers. The subscribers' profile data are collected too as well as the interaction with the community users.

Crawler is a powerful program to collect data from *Sina* blog. The working process of our data collection is as in Figure 1:

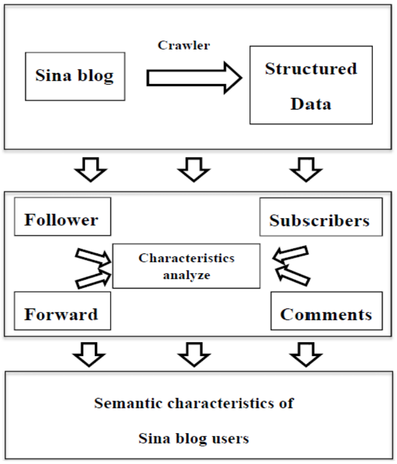

Figure 1: Process of User Characteristics Collection

The users' data are crawled from the Open Platform of *Sina* blog network[22]. The Application Programming Interface (API) and functions *Sina* blog provided users and researchers the ability to add function as well as resources without changing the system itself. So the crawling process will not change the topology of the network. The Table 2 are the crawled attributes.

Table 2 Information List Extracted by Crawler

	content	entity	usage
network topology	outlink list inlink list out-degree in-degree	follow fans user nodes	community discover Network topology study
user behaviors	number of posts post time number of forwards	user	user characteristics

The experiments server is 8GB memory and 4 kernels CPU, 2.33GHz. The crawling process lasts 15 hours. Figure 2 is a part of the crawled network:

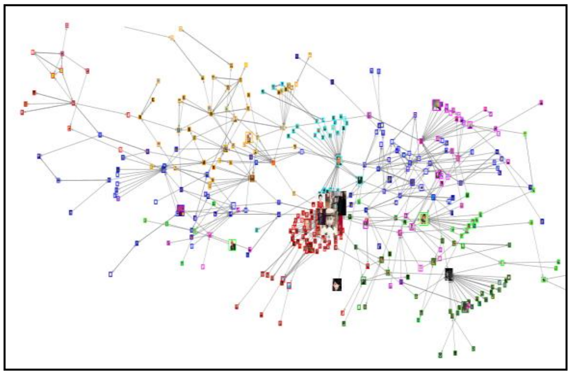

Figure 2: Part of Visualized *Sina* Network

The crawled network has 142 nodes and 8475 edges. The biggest strongly connected component has 94 nodes. The order is important in Chinese language. We will combine the user's self-statement tags and the post and comment he/she posted to build the characteristics set.

3.3 Semantic based Community Discovery in *Sina* Blog

There are classic algorithms such as ID3, C4.5, CART, Qusest, and Chaid. We design our algorithm by adding the semantic information gain ratio attribute to user characteristics and catalogue users into smaller and more similar groups. C4.5 uses divide-and-conquer strategy and the information gain ratio based characteristics will be analyzed on each level of the decision phase. The Information *In* is

$$In(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Where *S* is the set of *n* samples of tags (characteristics), S has *m* different categories C_i {*i* =1, 2,..., m); S_i is the number of sample in C_i, P_i is the probability any sample belongs to C_i.
Assume tag *T* has *m* values (a_1, a_2,..., a_m), *A* divide *S* into *m* subsets{s_1,s_2,...,s_m}; S_j is the valued sample of *S* in *T*, S_{ij} is the number of subsets of S_j which belongs to C_i. The entropy H is:

$$H(T, S) = \sum_{j=1}^{m} \frac{s_{1j} + s_{2j} + \ldots + s_{mj}}{s} In(s_{1j}, s_{2j}, \ldots, s_{mj})$$

Then Information gain on *T* is *G(T,S)* = $In(s_{1j}, s_{2j}, \ldots, s_{mj})$ -*H(T, S)*.

Changing of information is of crucial importance to decision tree generating. The changing is the gain of information. The more information gain the tag has, the more important that tag is. We use the gain ratio to evaluate the scale of information change. The higher the gain ratio, the more significant the tag will be, and the higher chance of changing the emerged tree would be. *GainRatio(S, T)* is evaluated by information gain *G (S, T)* and information split *SplitInfo(S, T)*:

$$GainRatio(Attri(Com),Q) = \frac{G(Attri(Com),Q)}{Semilarity(Attri(Com),Q)}$$

Where information split *SplitInfo(S, T)* is:

$$SplitInfo(S,T) = -\sum_{i=1}^{m} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$\{s_1, s_2, \ldots, s_m\}$ are the subsets *S* divided by tag *T*.
The tags (T_g) of individual user *v* are a set of keywords $T_g = \{t_1, t_2, \ldots, t_n\}$, and semantic. We use Semantic Similarity, Semilarity (S, T_g), to describe the semantic difference between users. The larger the Semilarity (S, T_g), the more alike the two users are.
Assume the sub-community *Com* has *n* users. *S(v)* is the set of semantic tags the user *v* has. Tag *Q* divides *Attri(Com) into k* subsets, $\{s_1, s_2, \ldots, s_k\}$*. And Attri(Com) is* the *n-1* sets users' tags of all community users except *v*. Then the similarity of tag Q of user v with *Attri(Com)* is:

$$Semilarity\left(Attri\left(Com\right),Q\right) = \sum_{i=1}^{k} \frac{|S_i|}{\left|Attri\left(Com\right)\right|} \log_2 \frac{|S_i|}{\left|Attri\left(Com\right)\right|}$$

Then information gain of v is *G(V,S)* = $In(s_{1j}, s_{2j}, \ldots, s_{mj})$ *-H(V, S).*

Then information gain ratio is:

$$GainRatio(Attri(Com),Q) = \frac{G(Attri(Com),Q)}{Semilarity(Attri(Com),Q)}$$

The nodes in social networks are semantic, the numerical evaluation along could not demonstrate the similarity as well as differences of users in the same community. Semantic based information gain ratio is suitable to evaluate the difference of characteristics. Then decision tree is generated as well.

4 Experiment and Results

4.1 Datasets

The whole process is start with the community discovery with *Sina* blog oriented community definition. Then users' tags as well as behavior patterns are added to calculate information. We choose CIPS-SIGHAN CLP 2010[23] simplified Chinese participle as training data. It developed by Chinese Academy of Sciences, and includes four fields, literature, computer science, medical science and finance, and has 200 thousand words in total. The training sets and the testing sets are collected from Internet and are suitable for our scenario. There is also a reference set in which the words are manually adjusted. Also we use ChnSentiCorp[24] to analyze sentimental words. The users' blogs we collected are 112837 posts. Then we use Sogou news [25] dataset and the blog posts as the training data.

Then we use ICTCLAS2015[26] for participle and parts of speech labeling.

The whole process is as in Figure 3:

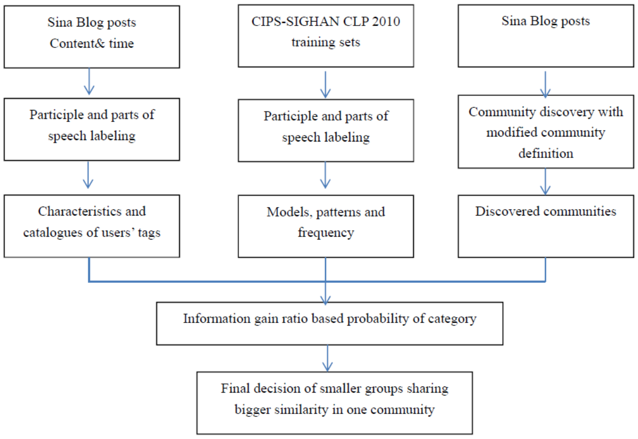
Sina Blog posts Content& time
CIPS-SIGHAN CLP 2010 training sets
Sina Blog posts
Participle and parts of speech labeling
Participle and parts of speech labeling
Community discovery with modified community definition
Characteristics and catalogues of users' tags
Models, patterns and frequency
Discovered communities
Information gain ratio based probability of category
Final decision of smaller groups sharing bigger similarity in one community

Figure 3: Process of Information Gain Ratio Based *Sina* Blog Users' Tag Classification

We set optimums on parameters that may vary during the process as in Table 3:

Table 3: Optimums of Parameters

	parameters	optimum
frequency model of words	minimum support (min_sup)	30
	maximum gap (max_gap)	1
	minimum difference threshold (min_diff)	0.65
information gain ratio based decision process	number of sub space of characteristics (*SS*)	10
	percentage of sampled sub space (α)	0.75
attributes of users	percentage of Minimum support (β)	0.015
	minimum confidence (min_con)	0.8
	minimum similarity (min_sim)	0.5

4.2 Results and Analysis

We compare our method with others and the results demonstrate ours has a better effect on smaller communities. First we compare different single classifiers with our method.

Table 4: Results Compared with Single Classifiers

	LR	DT	SVM
TF-IDF	84.19%	75.20%	80.85%
Word2Vec	83.57%	72.03%	81.32%
our	89.01%	79.83%	86.52%

approach			

TF-IDF and Word2Vec are open source tools.

Then we compare our method to classic algorithms such as Bagging, Boosting *etc.*.

Table 5: Results Compared with Classic Algorithms

	LR	DT	SVM
Bagging	87.92%	82.04%	85.96%
Boosting	85.38%	80.43%	84.89%
Random Subspace	88.46%	81.68%	89.14%
our approach	89.01%	79.83%	86.52%

From Table 4, TF-IDF achieved better results than Word2Vec, as the former one has taken into the consideration of word order, the length of sentences and intervals. And Word2Vec is mapping words into a fixed length vector, which may cause the lost or wrong transferred of the word's meaning. Also, the semantics we introduced helps with the higher accuracy. From Table 5 we can see the classic algorithms are better performed than the single classifiers. The reasons may lie in the sparse vectors the latter have. The single classifiers are poor performed dealing with the sparseness of the vectors

Our method outperformed the rest and gained higher accuracy in all the three aspects. It achieved 89.01% accuracy on LR. The reasons may be the users are in closely connected community. These communities are the results of our community discovery algorithm with modified community definition. Another reason is that we take user behavior pattern and semantics of both user interaction and words into consideration.

5 Conclusion

This paper proposed a new method of user similarity discover in *Sina* blog. The mining process includes three parts: the first is collecting *Sina* blog users' relations, posts of blogs and post time. The crawler is used through *Sina* provided API. The topology of the users' network, the semantic of users' tags and the patterns of user behaviors are collected without the verification of original network. Then we use a *Sina* blog oriented community definition to mining the communities. The process simplified the information gain ratio decision process and helps achieved better results. Finally we use the open datasets to train and to test proposed method; the overall experiments demonstrate our method achieves higher accuracy.

The Chinese language is studied less than the English, and the meaning of Chinese words are more diverse under the different pronunciations, all of which lead to the difficulty of Chinese language mining. For one aspect, if we could further mining the pattern of users behaviors with more diverse dataset, the results will be better. More methods and mechanisms are in need to improve the accuracy context mining..

REFERENCES

[1] M. McPherson, L. Smith-Lovin, J. M. Cook. Birds of a Feather: Homophily in Social Networks[J]. Annual Review of Sociology, 2001, 27(1): 415-444.

[2] S. Aybike and K. Resul, Using swarm intelligence algorithms to detect influential individuals for influence maximization in social networks, EXPERT SYSTEMS WITH APPLICATIONS, Vol.114, pp:224-236, 2018.

[3] Alp, ZZ and Oguducu, SG, Identifying topical influencers on twitter based on user behavior and network topology, KNOWLEDGE-BASED SYSTEMS, Vol:141, pp:211-221, 2018.

[4] Aral, S. Aral and D. Walker, Identifying Influential and Susceptible Members of Social Networks, SCIENCE, Vol:337, Iss:6092, pp:337-341,2012

[5] G. Suciu, C. Boscher, L. Prioux,*etc.*, Insights into Collaborative Platforms for Social Media Use Cases, STUDIES IN INFORMATICS AND CONTROL, Vol:26. Iss: 4, pp: 435-440, 2017.

[6] MG. Ozsoy, F. Polat, and R.Alhajj, Making recommendations by integrating information from multiple social networks, APPLIED INTELLIGENCE, Vol: 45. Iss: 4, pp: 1047-1065, 2016.

[7] R. Cerqueti, G. Ferraro, A. Iovanella, A new measure for community structures through indirect social connections, EXPERT SYSTEMS WITH APPLICATIONS, Vol:114, pp: 196-209, 2018

[8] A. Barrat, M. Barthelemy, A. Vespignani, Weighted evolving networks: Coupling topology and weight dynamics, PHYSICAL REVIEW LETTERS, Vol:92. Iss:22, 2004.

[9] R. Centeno, R. Hermoso, M. Fasli, On the inaccuracy of numerical ratings: dealing with biased opinions in social networks, INFORMATION SYSTEMS FRONTIERS, Vol:17. Iss:4, pp: 809-825, 2015.

[10] Palla et al., G. Uncovering the overlapping community structure of complex networks in nature and society. Nature , 2005.

[11] Agarwal, H. Liu, S. Murthy, et al. A Social Identity Approach to Identify Familiar Strangers in a Social Network[C]. International AAAI Conference on Weblogs and Social Media, San Jose, CA, USA, 2-9,2009..

[12] Agarwal, L. Huan, S. Subramanya, et al. Connecting Sparsely Distributed Similar Bloggers[C]. IEEE International Conference on Data Mining, Miami, FL, USA, 2009, 11-20

[13] X. Wang, H. Liu, W. Fan. Connecting users with similar interests via tag network inference[C]. ACM international conference on Information and knowledge management (CIKM), Glasgow, Scotland, UK, 2011, 1019-1024

[14] Q. Z. Li, Y. F. B. Wu. People search: Searching people sharing similar interests from the Web[J]. Journal of the American Society for Information Science and Technology, 2008, 59(1): 111-125

[15] H. Ma, H. Cao, Q. Yang, et al. A habit mining approach for discovering similar mobile users[C]. Proceedings of the international conference on World Wide Web (WWW), Lyon, France, 2012, 231-240

[16] X. Xiao, Y. Zheng, Q. Luo, et al. Finding similar users using category-based location history[C]. SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, California, 2010, 442-445

[17] H. Kwak, H.-Y. Shin, J.-I. Yoon, et al. Connecting Users with Similar Interests across Multiple Web Services[C]. International AAAI Conference on Weblogs and Social Media, San Jose, 2009, 246-249

[18] Wasserman, S., and Faust, K. 1994. Social Network Analysis. Cambridge University Press.

[19] W Ren, Y H Qiu, Micro-Blogging Based Network Growth Model of Semantic Link Network[J]. Applied Mechanics and Materials, 2014, 513-517:2211-2214.

[20] W. Ren, ZX. Huang, YH. Qiu, User Interaction Based Network Growth Model of Semantic Link Network, proceeding of 2010 Sixth International Conference on Semantics, Knowledge and Grids, pp: 66-73, 2011.

[21] Y. Wang, Brand crisis communication through social media A dialogue between brand competitors on Sina Weibo, CORPORATE COMMUNICATIONS, Vol: 21. Iss: 1, pp: 56-72, 2016.

[22] WJ. Li, YM. Feng, DJ. Li,*etc.*, Micro-Blog Topic Detection Method Based on BTM Topic Model and K-Means Clustering Algorithm, AUTOMATIC CONTROL AND COMPUTER SCIENCES, Vol: 50. Iss:4, pp: 271-277, 2016.

[23] Ma H , Cao H , Yang Q , et al. A habit mining approach for discovering similar mobile users[C]// International Conference on World Wide Web. ACM, 2012.

[24] H. Cao, T. Bao, Q. Yang, E. Chen, and J. Tian. An effective approach for mining mobile user habits. In Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM'10), pages 1677–1680, 2010.

[25] X. Xiao, Y. Zheng, Q. Luo, and X. Xie. Finding similar users using category-based location history. In GIS, pages 442–445, 2010.

[26] Newman, M. E. J. 2006. Modularity and community structure in networks. PNAS 103(23):8577–8582.