MUI-ISIDA: An improved calculation method for user influence in social networks

Shulin Cheng*, Ziming Wang, Meng Qian#, Shan Yang, Xin Zheng
School of Computer and Information, Anqing Normal University, Anqing, China

ABSTRACT

With the popularization and development of the Internet, microblogs have become a mainstream social network platform. The evaluation of user influence has become a research hotspot. Most of the existing researches calculate influence by improving PageRank. But these researches ignored the relationship between users' interest theme similarity and information dissemination, and didn't have enough analysis about the interaction behaviors among users. Aiming at these problems, we proposed a new microblog user influence algorithm—MUI-ISIDA (microblog user influence based on Interest similarity and information dissemination ability) in this paper, which takes into account users' interest theme similarity and information dissemination ability. We verified the effectiveness of the proposed algorithm on Sina microblog dataset. The experimental results show that compared with PageRank and MR-UIRank, the proposed algorithm has achieved higher accuracy in user influence ranking.

Keywords: Microblog, user influence, PageRank, user interest, effective interaction

1. INTRODUCTION

In recent years, with the rapid development and popularization of the Internet, social network has played an increasingly important role in the disseminations of information, ideas, and influence, and it has become an important medium for users to obtain and exchange information. In China, microblog like Twitter has become one of the most extensive social network platforms due to its openness and content simplicity. Different behaviors among users make information spread rapidly, which form the microblog information dissemination network.

Influence refers to the ability of an individual to influence others, change their thoughts or behaviors[1]. While users receive and disseminate information, they are also in the process of influencing and being influenced. Users with different influences have different effects on the speed and scope of information dissemination. High-influence users in network can be used as seed nodes to maximize the spread of influence[2], and play active roles in many fields. Therefore, it's of practical significance to measure users' influences in social networks.

In order to achieve page ranking, Larry Page and Sergey Brin proposed the PageRank algorithm[3]. The algorithm has also been used to analyze microblog users' influences, but it has certain limitations. Therefore, many experts and scholars have proposed some new algorithms on the basis of PageRank. Inspired by previous studies, we proposed MUI-ISIDA algorithm, which is comprehensive and reasonable to calculate uses' influences by considering multiple factors, such as user interests, relationships between influence and theme[4], user information dissemination ability. The results are reasonably ranked and experiments have verified the effectiveness of the algorithm.

2. RELATED CONCEPTS AND METHODS

2.1 Microblog network

The 'following' relationships among users form the information dissemination network. The interactive behaviors among users, such as forwarding, commenting, etc., promoted the information spread again[5]. We selected the three behaviors of following, forwarding and commenting to construct the microblog network $G<V, E>$. V indicates the set of nodes and E represents the set of edges in the network. As shown in Figure 1, the nodes in the figure represent microblog users, and the edges represent certain relationships among users. If user A follows user B, forwards or comments on his/her microblog

* ChengshL@aqnu.edu.cn; qianmeng@aqnu.edu.cn

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 1226027 · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637483

posts, which implies that user *A* has interacted with user *B*. There is an edge from *A* to *B*, and the weight of the edge is the closeness of the connection.

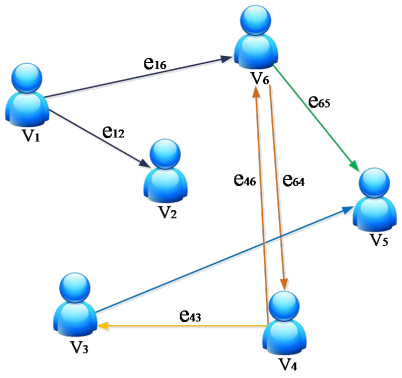

Figure 1. Microblog network.

2.2 PageRank

PageRank uses the link structures among web pages to calculate the qualities of web pages, and uses a directed graph to represent the relationships among web pages. Each node represents a web page. And each edge represents a link between web pages in the graph. It is calculated as below:

$$\text{PageRank}(p_i) = \frac{1-\alpha}{N} + \alpha \sum_{p_j \in M_{p_i}} \frac{Pagerank(p_j)}{L(p_j)} \tag{1}$$

where $M(p_i)$ stands for the set of all pages linked to p_i. $L(p_j)$ represents the number of all external links of the web page p_j. α is the damping factor, which represents the probability that the page is randomly accessed, and usually taking an empirical value of 0.85.

3. MAIN IDEA OF MUI-ISIDA

In this part, we took the interest theme similarities among users into account, and introduced the user information dissemination ability as a weighting factor in PageRank, finally constructed the MUI-ISIDA algorithm model.

3.1 Interest similarity

In microblog network, users often pay time and effort to publish microblog posts, so their historical original microblog posts can directly represent their own interests and hobbies[6]. We constructed the users' interest theme similarity model for calculating interest similarity. Since it is to identify topics that each user is interested in, rather than the topic of each microblog post is about. We integrated all original microblog posts published by the same user within a specified period of time into one document.

We chose the *LDA* model for topic mining. The *LDA* model, which treats each document as a 'bag of words', so each document emerges as a probability distribution over some topics[7]. We use this model and normalize it to get the document-topic distribution. And the result is represented by a matrix *DT*, which is denoted as *DT*.

$$DT = \begin{pmatrix} DT_{11} & DT_{12} & DT_{13} & \cdots & DT_{1j} \\ DT_{21} & DT_{22} & DT_{23} & \cdots & DT_{2j} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ DT_{k1} & DT_{k2} & DT_{k3} & \cdots & DT_{kj} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DT_{i1} & DT_{i2} & DT_{i3} & \cdots & DT_{ij} \end{pmatrix}$$

The interest similarities among users can be calculated by the users' topic distributions through *Pearson* correlation coefficient, and are described by equation (2).

$$Sim<i,k>=\frac{\sum_{j\in T}(DT_{ij}-\overline{DT_i})(DT_{kj}-\overline{DT_k})}{\sqrt{\sum_{j\in T}(DT_{ij}-\overline{DT_i})^2\sum_{j\in T}(DT_{kj}-\overline{DT_k})^2}} \tag{2}$$

where DT_{ij} stands for the probability of the word, which is assigned to the topic *j* in microblog posts of user *i*. $Sim<i,k>$ denotes the interest similarities among users. *T* represents the set of all topics in users' original microblog posts. $\overline{DT_i}$ denotes the mean interest degree of user *i* on these topics.

3.2 Information dissemination ability

In this part, we took the microblog quality coefficient and the assimilation effect coefficient as the evaluation index of the information dissemination ability.

3.2.1 Microblog quality coefficient. When a microblog post is forwarded or commented more frequently, it represents that the user who published it has higher influence. Users tend to share their own thoughts and participate in the discussion on the microblog posts they are interested in. The behaviors of forwarding or commenting embody the interests of the user are often closely related to the original microblog contents. Therefore, we can integrate these microblog posts and mine the latent topics to find the relevance to the original microblog contents. When the degree of relevance is higher, which implies that other users participated in the interaction with the original content more effectively, and the user's microblog contents are of higher quality.

Therefore, we integrated all original microblog posts by the same user into a document M_p, and integrated these thoughts related to the original contents into a document M_c. The topic probability distributions of the two documents $\theta_p(k)$ and $\theta_c(k)$ are generated through *LDA* model. *KL* distance[8] is usually used to measure the difference between the two probability distributions, which can be given by equation (3).

$$D(\theta_c \| \theta_p)=\sum \theta_c(k)\log\frac{\theta_c(k)}{\theta_p(k)} \tag{3}$$

The *JS* divergence[9] is a smoother and symmetrical probability distribution measurement based on the *KL* divergence. Inspired by the *JS* divergence, we transform the *KL* divergence equation and use equation (4) to symmetrize the *KL* divergence.

$$D_{KL}(\theta_c,\theta_p)=\frac{D(\theta_c \| \theta_p)+D(\theta_p \| \theta_c)}{2} \tag{4}$$

When $D_{KL}(\theta_c,\theta_p)$ gets greater value, which indicates that the similarity between M_p and M_c is smaller. There may exist some worthless communications or invalid forwarding in the user's microblog posts. Consequently, the quality coefficient should be appropriately reduced. So the effective factor δ is defined as:

$$\delta=e^{-D_{KL}(\theta_c,\theta_p)} \tag{5}$$

The microblog quality coefficient is expressed as follows:

$$Q_k=\frac{R_k+C_k}{N_k}\times\delta \tag{6}$$

where Q_k represents quality coefficient of user *k*. N_k stands for the total number of microblog posts published by user *k*. R_k, C_k denotes the number of microblog posts has been forwarded and commented of user *k* respectively. δ is the introduced effective coefficient.

3.2.2 Assimilation effect coefficient. Users can browse the contents of bloggers' microblog posts and they tend to forward or comment a microblog post they are interested in. Their interests and hobbies will also be imperceptibly affected during the process. The greater proportion of forwarding and commenting a user has, the user' s interests are more easily assimilated by others[10].

The user assimilation effect coefficient is shown in equation (7).

$$S_k = \frac{r_k + c_k}{n_k} \tag{7}$$

where r_k, c_k, and n_k indicate the number of forwarding, commenting and all microblog posts for user k, respectively. S_k stands for the assimilation effect coefficient of user k.

3.2.3 Information dissemination ability. The spread of information mainly through the forwarding and commenting among users. The larger the microblog quality coefficient and assimilation effect coefficient of users, the stronger the ability of users to control information transmission, which is conducive to information transmission in the microblog social network. Therefore, we defined the information dissemination ability of user k as W_k. The expression is shown as below:

$$W_k = Q_k \times S_k \tag{8}$$

3.3 MUI-ISIDA algorithm model

The basic idea of the MUI-ISIDA algorithm is to take the user's information dissemination ability as user's own weight. And according to the similarities among users, reasonably distribute the contribution of followers to the influence of bloggers. The more similar the users' interest theme, the greater the mutual influence among users. Therefore, the MUI-ISIDA value is computed as:

$$MUI-ISIDA(i) = \frac{1-\alpha}{N} + \alpha \sum_{j \in f(i)} MUI-ISIDA(j) \times \varphi(j,i) \times W_j \tag{9}$$

where $f(i)$ is the set of followers of user i. $\varphi(j,i)$ is the rate of MUI-ISIDA value that user j contributes to the user i, and the value is determined by the sum of the similarities among user j and the following users. The value is expressed as:

$$\varphi(j,i) = \frac{sim<j,i>}{\sum_{k=1}^{N} sim<j,k>}, \quad (k=1, 2, \cdots, N \quad k \in A(j)) \tag{10}$$

where $A(j)$ is the set of users followed by user j, k represents the k-th user in the set, and N denotes the total number of users followed by the same user. We defined the initial value of MUI-ISIDA as one and the MUI-ISIDA value could converge to a certain constant through repeated iterations. Consequently, the MUI-ISIDA value of all users can be obtained.

4. EXPERIMENT AND ANALYSIS

4.1 Data set

Sina microblog is a social network platform similar to Twitter. Until October 2020, monthly active users have reached 523 million. Therefore, we selected Sina microblog as the data source. We had obtained a batch of relevant users data. The total number of friends/followers records reaches 83,000. The dataset information is shown in Table 1.

Table 1. Data set information table.

Statistics	Data value
User counts	12746
Time	2015.8
Original microblog counts	441687
Forwarded microblog counts	66289
Comment counts	55784
Per capita microblog counts	39.85

4.2 Experimental results and analysis

To evaluate the validity of the ranking results, we listed the top-10 users of PageRank and MUI-ISIDA. The results are demonstrated in Tables 2 and 3.

Table 2. Top-10 users of PageRank.

Username	Followers counts	Interaction counts	Microblog counts	Microblog quality	PR value
Prometheus	804	67	35	1.91	4.1309
Liu HuiZ	625	16	6	2.67	3.4485
Zhang RuiJ	453	36	10	3.6	3.1501
Zhang W	455	14	28	0.5	3.1263
Ju Geg	881	5	27	0.19	3.0019
He BaoH	596	20	25	0.8	2.9966
Yang B	602	10	25	0.4	2.8629
EllieAga	928	7	2	3.5	2.8158
Xing XueP	390	17	31	0.55	2.7824
Shang LeiM	488	95	51	1.86	2.7631

Table 3. Top-10 Users of MUI-ISIDA.

Username	Followers counts	Interaction counts	Microblog counts	Microblog quality	MUI-ISIDA value
Shang LM	488	95	51	1.86	6.1642
Zhang PW	267	144	19	7.58	5.3367
Lun Zi	266	62	42	1.48	4.6187
Prometheus	804	67	35	1.91	3.7657
Wu YiT	452	48	20	2.4	3.5735
Li Yang	498	42	28	1.5	3.3868
Chen XiaoH	226	79	24	3.29	3.3493
Fang F	298	49	17	2.88	3.0384
Zhao XueM	617	37	28	1.32	2.9471
Zeng Gy	242	128	68	1.88	2.8498

4.2.1 Top-10 users of the two methods. As is shown in Tables 2 and 3, the interaction counts include the total number of the user's microblog posts have been forwarded and commented. The quality of microblog is defined as the ratio of the user's interaction counts to the number of microblog posts. It can be seen in Table 2 that some top users of PageRank have high follower counts, which demonstrates that the number of followers is directly related to user influence ranking. In the results of MUI-ISIDA, the top users seem relatively reasonable. More followers do not connote higher influence. Take 'Zhang PW' and 'Zeng Gy' as examples, although their followers counts are small, but other users interact with them frequently, so they have higher influences. That's to say, the interaction counts are directly related to user influence in MUI-ISIDA.

4.2.2 The accuracy of user influence ranking. In this part, we compared our algorithm with PageRank and the MR-UIRank[11], and analysed the experimental results. If a user's microblog posts are forwarded or commented more

frequently, it means that he has a higher influence. Therefore, the number of interactions can reflect the user's influence. In addition, users tend to browse high-quality microblog contents and interact with the information they are interested in. The higher quality of a user's microblog posts, the more frequent interaction will be attracted. So the quality of microblog posts surely reflect the user influence.

In Figures 2 and 3, the abscissa denotes the top k users of the three methods, the ordinate represents the hit rates of the corresponding results in microblog interaction rankings and microblog quality rankings. It can be seen from figures 2 and 3 that when the number of top k is 30, the hit rates of the MUI-ISIDA both increased by 23.3% compared with PageRank. The hit rates increased by 6.7% and 3.4% when compared with MR-UIRank respectively, good experimental results have been achieved. This is because the MUI-ISIDA algorithm considers multi-dimensional factors, and reasonably integrates into PageRank. In this paper, the hit rate is used as an indicator of the accuracy of judgment. The higher hit rate, the higher accuracy is. Experimental results show that the accuracy of MUI-ISIDA algorithm in influence calculation is better than PageRank and MR-UIRank.

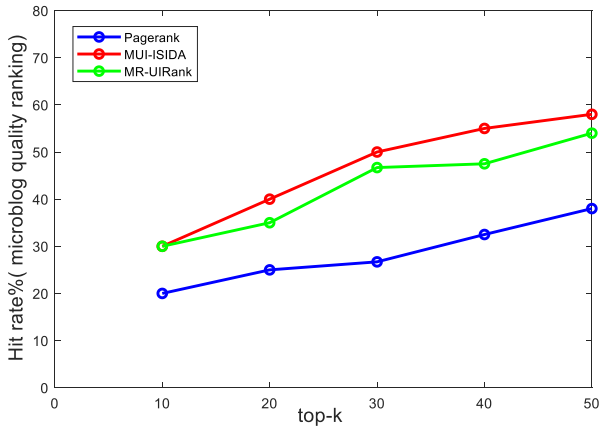

Figure 2. Microblog quality ranking.

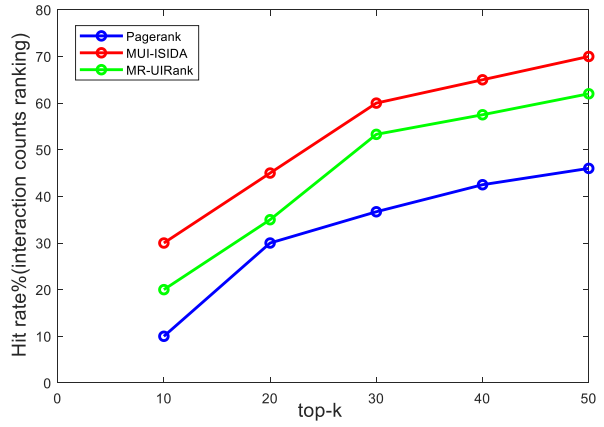

Figure 3. Interaction ranking.

5. CONCLUSIONS

We proposed a new algorithm to calculate users' influences, which fully integrated the interest theme similarity and information dissemination ability. The proposed algorithm improves the effectiveness and objectivity of user influence calculation to a certain extent. On the one hand, we analyzed users' original microblog posts to capture their interests, so as to distribute the influence reasonably. On the other hand, we quantified users' behaviors and verified the effectiveness

of them. Finally, we tested our algorithm on microblog datasets. According to the experimental results, our proposed algorithm has achieved a higher accuracy than other state-of-the-art algorithms.

ACKNOWLEDGEMENT

The work was supported by grants from the Nature Science Foundation of Anhui Province in China, No.2008085MF193 and No.1908085MF194, the University Synergy Innovation Program of Anhui Province, No. GXXT-2019-008, the Outstanding Young Talents Program of Anhui Province, Grant Number gxyqZD2018060, Provincial quality project of Anhui Province Education Department, No.2019jyxm0285. And we thank all the anonymous reviewers for their hard work and valuable comments.

REFERENCES

[1] Liu, Q., Xiang, B., Yuan, N. J., Chen, E. H., Xiong, H., Zheng, Y. and Yang, Y., ACM Transactions on Knowledge Discovery from Data, 11(3), 1-30 (2017).
[2] Wen, Z., Kveton, B., Valko, M. and Vaswani, S., Advances in Neural Information Processing Systems, 3022-3032 (2017).
[3] Page, L., Brin, S., Motwani, R. and Winograd, T., Stanford Digital Libraries Working Paper, 9(1), 1-14 (1998).
[4] Weng, J., Lim, E. P., Jiang, J. and He, Q., Proc. of the Third International Conf. on Web Search and Web Data Mining, New York, USA, (2010).
[5] Zhao, J., Gui, X. and Tian, F., IEEE Access, 5, 3008-3015 (2017).
[6] Liu, B. Y., Wang, C. R., Wang, C., Wang, J. W. and Huang, M., Journal of Software, 28(2), 246-261 (2017).
[7] Blei, D. M., Ng, A. Y. and Jordan, M. I., Journal of Machine Learning Research, 3, 993-1022 (2012).
[8] Cheng, S. and Wang, W., Information, 11(1), 4 (2020).
[9] Nguyen, H. V. and Vreeken, J., Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 173-189 (2015).
[10] Huang, M., Zhang, B., Zou, G., Gu, C. and Zhu, Z., Int. Conf. on Big Data Intelligence & Computing & Cyber Science & Technology Cong., 124-131 (2016).
[11] Sun, H. and Zuo, T., Journal of Chinese Computer Systems, 39(1), 42-47 (2018).