Advances in Computer Science Research, volume 80
International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018)

Research on the Similarity of Microblog Rumors Based on Bayesian Network

Chengcheng Li, Fengming Liu* and Pu Li
School of Management Science and Engineering, Shandong Normal University, Ji'nan 250014, China
*Corresponding author

***Abstract*—The research of text similarity, especially for rumors texts, which constructed the calculation model on the basis of known rumors and calculated its similarity. From which, people can recognize the rumors in advance, and improve their vigilance to effectively block and control rumors dissemination. Based on Bayesian network, the similarity calculation model of microblog rumors texts was built. At the same time, taking into account not only the rumors texts have similar characters, but also the rumors producers have similar characters, and therefore the similarity calculation model of rumors texts makers was constructed. Then, the similarity between the text and the user was integrated, and the microblog similarity calculation model was established. Finally, also experimentally studied the performance of the proposed model on the microblog rumors text and the user data set. The experimental results indicated that the similarity algorithm proposed in this paper can be used to identify the rumors of texts and predict the characters of users more accurately and effectively.**

Keywords—microblog rumors; similarity; bayesian network

I. INTRODUCTION

Microblog is a new media platform based on the social network of user interaction to share, disseminate and exchange the short and real-time data. Its features of originality, interaction, convenience and fragmentation will mainly focus on future development of microblog on the aspect of information construction, business model promotion, and other aspects [1-4]. In the microblog, users are not only foundation, but also main part that constitutes microblog framework combined with microblog text. The unique interaction of microblog among users build a complex and large social network, and its multi-level fission can make microblog content that is forwarded, continued to spread and amplified quickly by a lot of fans [5]. Among these, the typical example is the spread of microblog rumors [6-7] that causes a very bad impact not only on the microblog cyberspace, but also on people's daily life. So microblog rumors are widely explored and studied.

In order to scientifically and effectively manage massive amounts of data and user information, and refine microblog network space, researchers conduct a large number of analysis and experiment in the prediction of microblog rumors similarity [8-10]. We take into account not only the rumors texts have similar characters, but also rumors producers have similar features. Therefore, on the one hand, in the study of microblog text content, because microblog is short text with the features of short length and sparse feature words in its text, it results the previous calculation method of text similarity that cannot be directly applied in the detection of microblog text similarity. Thus, scholars begin to study the similarity of short text. On the other hand, in the exploration of microblog users, the number of users has been risen with the gradual growth and generalization of social networks. Therefore, here are many scholars begin to analyze behavior similarity of microblog users [11] so that better achieve users' forecast in some aspects [12-13].

For the calculation of the microblog content similarity, the traditional method is to transform the data and return the unified calculation to research based on extracted keyword and short text classification [14]. At present, the automatic extraction method based on semantic and conceptual terms is widely used. This method mainly uses the semantic dictionary to obtain the semantic knowledge among vocabularies and further to extract the text keywords. At the same time, for the calculation of users behavior similarity [15], researchers found that it has been widely used in enterprises microblog about customer service, and analyzed spending habits of potential customers who have similar behaviors through data mining; and search interests and hobbies information of users that further satisfy users' demands, which can increase the interaction between two parties [16-21]. However, in recent years, for the analysis and exploration of the text content of microblog rumors and the calculation of users behavior similarity is still in infant stage, and its related methods, indicators and verification are not enough integrity that needs to be further improved.

This paper adds the unique feature vector of microblog. In the study of text similarity, we consider the rhetorical devices, sentence features and sensitive words use. In the study of user similarity, we take into account the user's commenting behavior, forwarding behavior, @ behavior and other interactive behaviors. So from which we can get more comprehensive and targeted information of microblog text and user and improve the accuracy of the calculation method.

II. SIMILARITY MODELING AND CALCULATION METHOD OF MICROBLOG RUMORS BASED ON BAYESIAN NETWORK

A. Microblog Text Similarity Modeling Based on Bayesian Network

Due to the short length of microblog text, the characteristic words of composition text are less, the correlation between keywords is weaker and so on, and the processing of short text is becoming the mainstream of text processing. Through the

microblog rumors of the massive data study found that in the language of microblog rumors more popular, spoken language heavier, mostly using exaggeration, irony, citation and other rhetoric, rendering tension, rapid atmosphere; Phrases, sentences, affirmative sentences, exclamatory sentences, and syntactic style strong; in the use of words, the use of easy to stir up the group of emotional sensitive words [22]. Based on this, this paper constructs the Bayesian network model based on the similarity calculation method of short text keyword, and analyzes the characteristics of microblog rumor itself, adding text rhetorical devices, sentence features and sensitive words, then given the different weight coefficients, it can distinguish the importance degree of the various contribution degrees of eigenvectors in short text similarity calculations.

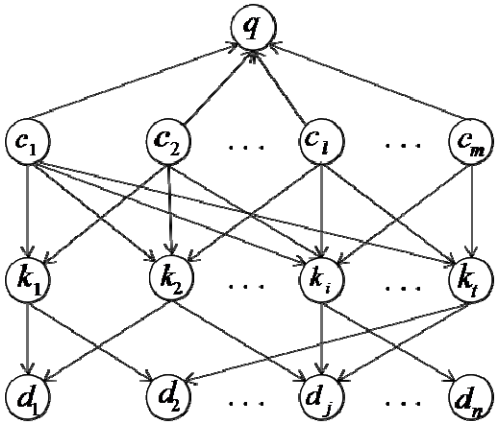

FIGURE I. MICROBLOG TEXT MODELING BASED ON BAYESIAN NETWORK

III. MICROBLOG USER SIMILARITY MODELING BASED ON BAYESIAN NETWORK

Bayesian network is through the directed acyclic graph to describe the probability of the relationship between the definition of microblog user set for the network node set. Each node represents a microblog user, between the node between the arc on behalf of the user similarity relationship. In order to form a directed acyclic graph between the user nodes, this paper establishes the query propagation tree by constructing the query propagation tree, and the query user node is the parent node of the tree. When the query user sends a query message, it will query its similar users, if the similar user spread the rumor, record the information, and then query the next similar user, if the user did not propagate the rumor, The similar user as a starting point, the downward expansion of the query, and so on, you can create a query rumor diffusion tree. It should be noted that when the number of layers of the query reaches the pre-specified value, the query is no longer extended downward. In addition, the query users can only receive a query, cannot be multiple inquiries, otherwise it will form a query storm.

The Bayesian network is constructed according to the query propagation tree, then the prior probability distribution of the nodes to be calculated and the conditional probability distribution among the nodes are calculated. And according to the original path return probability distribution information to the query user, query user according to the return information and Bayesian formula to predict the probability that the user publish or reprint rumors. Finally, this probability is compared with a given probability threshold, and if it is greater than the threshold, then the user will propagate the rumor. The query extension tree is shown in Fig. 2.

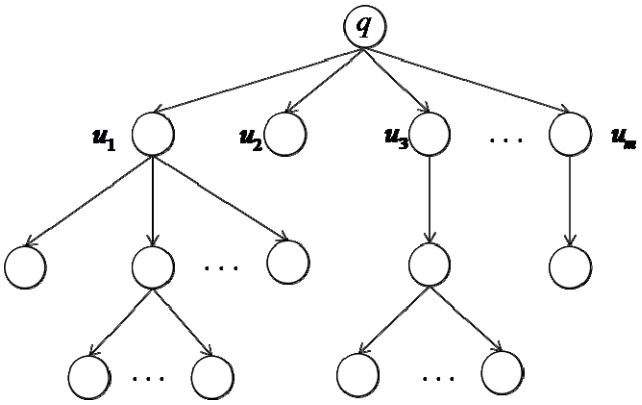

FIGURE II. MICROBLOG USER MODELING BASED ON BAYESIAN NETWORK

IV. MICROBLOG RUMORS SIMILARITY ALGORITHM BASED ON BAYESIAN NETWORK

In the first two sections, we focus on modeling the microblog text and the user, in microblog social network through constructing Bayesian network, we can predict the similarity between text and rumors sample (i.e. the probability that the detected text is the rumor) and the probability that the user of the detected text is the rumor producer. Now we propose an integrated model. That is, we take two models of the text and the user into one model, so that microblog text and its user information can be effectively embedded in the network to improve the accuracy and stability of the forecast. To achieve this goal, we have to define the two networks in the microblog social network, respectively, on behalf of the text space network and user space network. The integrated model is shown in Fig. 3.

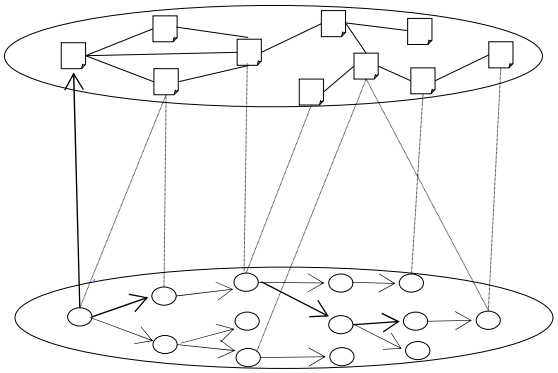

FIGURE III. THE MICROBLOG INTEGRATED MODEL

V. EXPERIMENTS

This paper chooses the data sets from Sina Weibo to start the experiment and verify the feasibility of this model. In the

process of collecting the experimental data, we refer to the data of "Statistical and Semantic Analysis of Rumors in Chinese Social Media"[23], which divides microblog rumors into several different categories, namely politics, economy, fraud, social life, common sense and other categories. In the analysis we delete too short text content or microblog picture.

In order to test and verify this model, we choose microblog rumors that are relatively more class - fraud class to verify. After that, in the fraudulent microblog rumors we extract a text to take pre-operation in a random way. Based on the analysis of the text similarity, we extract microblog's characteristic vector in the content, rhetorical devices, sentence features and sensitive words. Based on the analysis of user similarity, we find the user of the query text and study the publisher's attention to the number of users, fans, praise behavior, commenting behavior, reproduced behavior, etc., so that it's similar users in-depth study. In the process of verification, we have simplified the specific algorithm. The main purpose of simplification is to ensure the stability of the model, then we try to avoid the data to do too much of the changes. But the final test results are no substantive impact, so we see them as the experimental results of this study are credible.

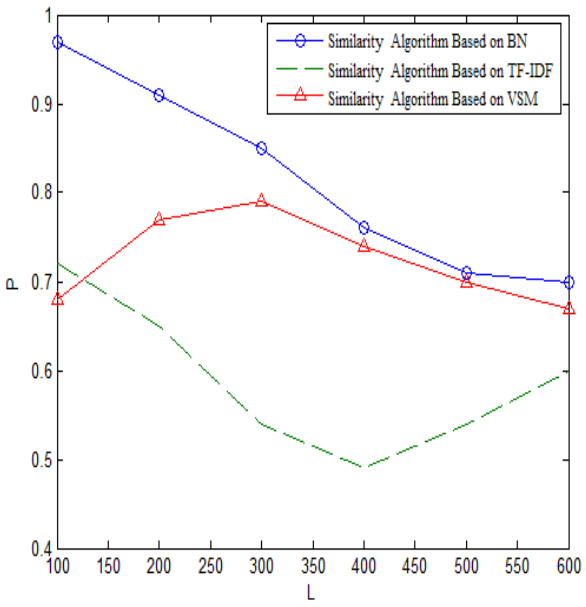

FIGURE IV. COMPARISON OF THE ACCURACY OF THE THREE ALGORITHMS

VI. CONCLUSION

This paper starts from the analysis of microblog rumors' similarity and uses Bayesian network to construct the model. First, in order to calculate the similarity of microblog social network rumors, we generate two models of the text and user into a unified model, so that we can improve the research accuracy of the microblog rumors similarity. Based on the study of the microblog rumor text's model, we calculate text similarity by extracting the eigenvector of the query microblog text. Based on the microblog user's model, we establish Bayesian network by using the probability information and the similarity relation between the users. Then by using Bayesian formula and the calculation, we can judge the probability that the rumor spreads. Finally, the experiment and analysis are carried out on the microblog rumor text and the user data set. The experimental results show that the similarity algorithm proposed in this paper can be used to identify the rumors of text and predict the characters of users more accurately and effectively.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (No. 61170038, 61472231, 71701115), the National Social Science Foundation of China (No. 14BTQ049), the Shandong Natural Science Foundation(ZR2017MF058), and Special project for Internet development of social science planning special program of Shandong province(17CHLJ23).

REFERENCES

[1] Ding Z Y, Jia Y, Zhou B. Survey of Data Mining for Microblogs. Journal of Computer Research and Development, 2014, 51(4):691-706.

[2] Liao Y, Moshtaghi M, Han B. Mining Micro-Blogs : Opportunities and Challenges, Proceedings of Social Networks: Computational Aspects and Mining, Springer, 2011:1-28.

[3] Lei C C, Zhang A, Qi Q W, Su H M. Grid-based Location Microblog Data Fetching and Human Information Extraction. Science of Surveying and Mapping, 2017, (02):125-129.

[4] Cui J D, Du W Q, Guan Y, Luo W D. Research on the Evolution of Micro-blog User Information Personalized Recommendation Model Based on LDA. Information Science, 2017, (08):3-10.

[5] B D Davison, L Hong, D Yin. Structural Link Analysis and Prediction in Microblogs. Acm Conference on Information and Knowledge Management (CIKM2011). Glasgow, Scotland, UK, 2011(10):1163-1168.

[6] Lei H Z, Zhang J, Lan J L. The Diffusion Effect of "Weibo Community" Negative Information Based on the Rumor Spreading Model and Case Studies. Modern Information, 2015, 35(5):30-34.

[7] Li Y, Wang X Y, Zhang X G. Predicting Trending Messages and Diffusion Participants Based on Multi-types of Influences in Microblogs. Application Research of Computers, 2016, (10):2910-2913+2918.

[8] Sun Y F, Li S. Similarity-Based Community Detection in Social Network of Microblog. Journal of Computer Research and Development, 2014, 51(12):2797-2807.

[9] Yao B X, Ni J C, Yu P P, Li L L, Cao B. Micro blog user recommendation algorithm based on similarity of multi-source information. Journal of Computer Applications, 2017, (05):1382-1386.

[10] Zheng Z Y, Jia C Y, Wang Z F, Li D. Computing Research of User Similarity Based on Micro-blog. Computer Science, 2017, (02):262-266.

[11] Mepherson M, Smith-Lovin L, Cook J M. Birds of a feather: Homophily in Social Networks. Annual Review of Sociology, 2001: 415-444.

[12] Resnick P, Iacovou N, Suchak M. Grouplens: An Open Architecture for Collaborative Filtering of Netnews. Chapel Hill, North Carolina, United States: ACM, 1994.

[13] Lawrence R D, Almasi G S, Kotlyar V. Personalization of Supermarket Product Recommendations. Berlin Springer, 2001.

[14] Zhang B, Zhang Y, Gao K N. Combining Relation and Content Analysis for Social Tagging Recommendation. Journal of Software, 2012, 23 (3):476-488.

[15] Zhu X Q, Liu F M, Research on Behavior Model of Rumor Maker Based on System Dynamics, Complexity, Hindawi, Volume 2017, Article ID 5094218, 9 pages. https://doi.org/10.1155/2017/5094218.

[16] Li Q Q, Gu J F. Activity Driven Modelling of Online Social Network. Journal of Systems Engineering, 2015, 30 (1):9-15.

[17] Ding Y S, Liu F M, Tang B Y, Context-Sensitive Trust Computing in Distributed Environments", Knowledge-Based Systems, 2012(28) 105-114.

[18] Liu F M, Li X, Ding Y S, Zhao H F, Liu X Y, Ma Y H, Tang B Y, A Social Network-Based Trust-Aware Propagation Model for P2P Systems, Knowledge-Based Systems, 2013(41), 8–15.

[19] Liu F M, Wang L, Gao L, Li H X, Zhao H F, Sok Khim Men. A Web Service Trust Evaluation Model Based on Small-World Networks, Knowledge-Based Systems, 2014(57) 161-167.

[20] Liu F M, Wang L, Henric Johnson, Zhao H F, Analysis of Network Trust Dynamics Based on Evolutionary Game, Scientia Iranica, Transaction E: Industrial Engineering, 2015(22.6): 2548-2557.

[21] Liu F M, Zhu X Q, Hu Y X, Ren L H, Henric Johnson, A Cloud Theory-Based Trust Computing Model in Social Networks. Entropy 2017, 19(1), 11.

[22] Cheng A X, Xia C Q. On the Linguistic Features of Microblog Rumor. Southeast Communication, 2014, (11):98-100.

[23] Liu Z Y, Zhang L, Tu C C, Sun M S. Statistical and Semantic Analysis of Rumors in Chinese Social Media. Scientia Sinica (Informationis), 2015, (12):1536-1546.