Text Similarity Computation Model for Identifying Rumor Based on Bayesian Network in Microblog

Chengcheng Li, Fengming Liu, and Pu Li
Business School, Shandong Normal University, China

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

Keywords: *Microblog Rumor, Similarity, Bayesian Network.*

Received May 8, 2018; accepted March 31, 2019
https://doi.org/10.34028/iajit/17/5/6

1. Introduction

Microblog is a new media platform based on the social network of user interaction to share, disseminate and exchange 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 [2, 4, 6, 11, 18, 25]. 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 [36]. Among these, the typical example is the spread of microblog rumors [8, 24, 33] 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.

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 [32, 35, 38] We take into account not only the rumor texts have similar characters, but also rumor 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 from 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 risen with the gradual growth and generalization of social networks. Therefore, here are many scholars begin to analyze behavior similarity of microblog users [27] so that better achieve users' forecast in some aspects [12, 30].

At present, developed countries are very active in the research of microblog and elaborate a lot of academic views. However, most views are based on a certain scientific basis. By studying the phenomenon of mutual concern, mutual comment, mutual forwarding, this paper analyzes deep reasons and discusses the internal factors. Also, foreign scholars have in-depth studies in the study of word of mouth marketing, raising funds, computer technology and interactive expansion of public relations of Twitter. Saini [31] presents a psychological study and analysis of the behavior of chatters in online chatting environments. There are a large number of Sina microblog users in China and many successful celebrities are using Sina microblog for social networking. Compared with Sina microblog, Phoenix and Tencent microblog begin to gradually develop in the past few years. On the basis of such a large user groups, many scholars in China conduct studies for microblog text, and the research on the spread forms of microblog rumors is developed in this period.

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 [37]. 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 [39], 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 [5, 15, 20, 21, 22, 23] 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.

To enhance the ability of users to distinguish events and reduce the damage of microblog rumors, this paper conducts a study on the similarity of microblog rumors, which strives to reduce or break the wantonly spreading of microblog rumors. In the process of social networking on the microblog, the text is the main form of microblog social and users are the foundation that both have their similar characters. Therefore, in this paper, we combine the text information of microblog with the user information, and propose an integrated model based on the Bayesian network to calculate the similarity of microblog according to the unique characters of microblog and other social networks. This method uses the Bayesian network to model the microblog text and microblog users respectively. Then, through the Bayesian network, the Bayesian reasoning and calculation are carried out. Finally, the two models are integrated and the measuring method of microblog rumors similarity is proposed, and then fully prove this method that is effectiveness and feasibility through comparing it with another similarity algorithm.

The contributions of this paper are:

- This paper provides a new flexible and unified method of calculating similarity. The microblog text and user are effectively combined to identify the rumor of text and predict the characters of users more accurately and effectively, and we further integrate the two models into a model. Through the experimental verification, the proposed algorithm is superior to the existing method in the real microblog data set, and which can improve similarity algorithm accuracy of identifying rumor from the massive microblog data.
- This paper introduces the Bayesian network into the microblog social network. The combination of probability theory and graph theory effectively provides a natural and intuitive way to reason and predict the uncertainty of microblog rumors and users, Then we can put forward the similarity calculation method, and microblog rumors similarity research provides the necessary technical support.
- 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.

In the future work, we will continue to consider the other characteristics of the microblog information to conduct in-depth research, such as the propagation characteristics, semantic features, image features, video features, etc., of the microblog information, thus enriching the combined feature similarity algorithm, which is more accurate. It reflects the complex relationship and potential relevance of rumor information to further improve the accuracy of identifying microblog rumor.

2. Preliminaries

In this section, we introduce related concepts and necessary theories.

2.1. Microblog Rumor

Microblog rumor, as the name implies that the network disseminates the unreal data or information, which will cover up the real situation. These words in the process of rapid dissemination will mislead the audience's ability to distinguish and hurt individuals, groups and society [29].

In microblog, we consider that not only the rumor texts have similar characteristics, but also the rumor producers have similar characteristics. Based on this, this paper believes that the concept of microblog rumor can be viewed from both the text and the user. From the text point of view, microblog rumor refers to the offensive and purposeful discourse without the fact through the microblog, but it has a higher degree of similarity with real information. From the user point of view, it refers to the microblog users as a rumor spread of the center, a single user to publish false information and users and their similar users between the reproduction of the text and diffusion of the degree of participation. The higher the similarity between the text and the rumored, and the higher the similarity between the user behavior and the rumor producer, the greater the probability of the rumor.

2.2. Bayesian Network

Bayesian network is a probabilistic network, which is based on mathematical model of probability reasoning. Bayesian network based on probability reasoning is proposed to solve the problem of uncertainty and incompleteness. The network has a strong mathematical basis and the image of intuitive semantics, which are widely used in the expert system, decision support, pattern recognition, machine learning and data mining and other fields. The Bayesian network is a directed acyclic graph, consisting of representing the nodes variables and connecting them to the edges. The nodes represent the random variables. The directional edges between the nodes represent the interrelationships between the nodes (from the parent node to their child nodes). The relationship between the nodes is expressed by the conditional probability, and no parent nodes are used to express the information with the prior probability. Bayesian network not only has a strong modeling function, but also has a perfect reasoning mechanism, through the effective integration of prior knowledge and the current observations to complete a variety of inquiries [19].

2.3. Similarity Research Status Quo

With the rise of short texts such as microblog text and short message, researchers' research on the similarity is mainly based on the text keywords and semantic concepts. Li *et al*. [17] proposed that the data that can be the concept and syntax and other information to calculate the short text, according to the similarity of language to test. Li *et al*. [13] analyzed the similarity calculation method of multiple characteristic sentences. Li *et al*. [16] extracted the domain keywords and used the probability theme model to classify the text, so as to calculate the similarity. Yang and Huang [34] calculated similarity by short text’s distance in the knowledge base and word library similarity algorithm.

At this stage, foreign scholars on the social network of user behavior similarity also elaborate a lot of academic views. Zhu and Liu [39] have proposed that by using human behavior data to analyze in-depth the user’s similarity and tap the potential needs of users to complete the recommended task. Qin *et al*. [28] proposed a search algorithm based on the similarity between the social network and the user's archival data, and further studied the evaluation effect between the user and the user in terms of the behavioral characteristic’s similarity. Hu *et al*. [9] formed a set of user attributes based on the collected data of personal characteristics and provided a new user preference similarity recommendation algorithm.

But so far, scholars have been based on previous studies in microblog rumors similarity in the area is still shallow. Most of the existing microblog rumor similarity calculation method is mostly text-based, while the traditional method ignores the user information, making the calculation is too simple and result in access to information is not comprehensive. In the microblog social network, a user to publish or reprint a microblog text, then the user's fans will judge whether the text is a rumor, and then come to make their decisions---reproduced or not reproduced. Thus from which we can effectively block rumor. Aiming at the existing algorithms, this paper proposed a flexible integrated model based on the similarity of microblog rumor---similarity algorithm based on Bayesian network, and we experimented on microblog text and user data sets. Specifically, we first divided the microblog rumor similarity research into microblog text and microblog user two different generative models for similarity analysis. Then we construct the Bayesian network and conduct similarity calculation according to the data information that has been captured to forecast the probability of rumor; finally, we integrate the two models into a unified model and put forward a microblog rumor similarity measurement method to calculate the similarity between detected microblog and microblog rumor. Among them, this paper by analyzing microblog rumor text similarity, we can more accurately determine whether the detected text is a rumor, and then help people make more reasonable and more effective decision-making. At the same time, by analyzing microblog rumor producer's behavior similarity, we can help people predict the probability that the user spreads rumors so that we can effectively block rumors. Based on the analysis of the similarity between text and user and the verification of experimental data, the algorithm based on Bayesian network similarity algorithm is better than that of single computing text similarity compared with other similarity algorithms, and it can effectively identify microblog rumor and improve the accuracy and stability of the forecast.

3. Similarity Modeling and Calculation Method of Microblog Rumors Based on Bayesian Network

3.1. 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 rumor of the massive data study found that in the language of microblog rumor 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 [1]. 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.

3.1.1. Related Parameters

Bayesian network microblog rumor text model structure as shown in the Figure 1, each node is subject to 0 ~ 1 distribution of random variables.

c for the microblog rumors samples set, a total of *m* articles; *C*={*c*1, *c*2,…,*c*3} is the model of the microblog sample space, where each element is a microblog rumor sample, then $\vec{c}=(c_1,c_2,...,c_m)$, among them $c_i \in \{0,1\}$. Each subset *u* in *C* is associated with a vector $\vec{c}$. Suppose $\vec{k}$ is a vector of *t*-dimensional space, $\vec{k}=(k_1,k_2,...,k_t)$, among them $k_i \in \{0,1\}$.

q is the eigenvector (text keyword, rhetoric, sentence feature, word use, etc.) of the microblog text's query node, $\vec{Q}=(w_{1q},w_{2q},...,w_{mq})$, which are the weights of the microblog rumors in the query text, respectively.

k for the keywords of the microblog sample, a total of *t*, $\vec{K}_i=(w_{1i},w_{2i},...,w_{mi})$, which are the weights of the microblog rumors in the text keywords, respectively.

d for the detected texts that are similar to the microblog rumor text, a total of *n* articles, $\vec{D}_j=(w_{1j},w_{2j},...,w_{mj})$, which are the weights of the microblog rumor samples in the texts that are similar to the query microblog texts, respectively.

3.1.2. Text Similarity Model Building

The model adopts a four-layer Bayesian network text retrieval structure, including query text node layer, microblog rumor information sample node layer, microblog sample keyword node layer and text node layer to be detected. The query text node is used to describe the query requirements, and the similarity between the query text and the microblog rumor sample is obtained by calculating the conditional probability between the nodes, and the microblog text is sorted accordingly, thereby calculating the probability that the query microblog text is a rumor.

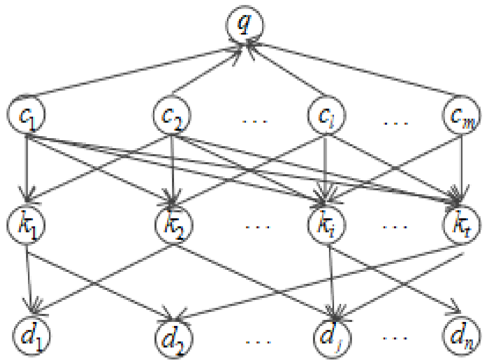

Figure 1. Microblog text modeling based on bayesian network.

3.1.3. Microblog Text Content Similarity Calculation

$$\begin{aligned} P(d_j \mid q) &\sim \sum_{\forall \vec{c}} p(q \mid \vec{c}) p(d_j \mid \vec{c}) \\ &= (w_{1q}, w_{2q}, ..., w_{mq})(\sum_{\forall i} w_{ij} w_{1i}, \sum_{\forall i} w_{ij} w_{2i}, ..., \sum_{\forall i} w_{ij} w_{mi}) \end{aligned} \tag{1}$$

$$\begin{aligned} S_{content-j} &= \vec{Q} \cdot \vec{D} \\ &= \vec{Q} \cdot \sum_{\forall i} w_{ij} \vec{K}_i \\ &= (w_{1q}, w_{2q}, ..., w_{mq}) \cdot (\sum_{\forall i} w_{ij} w_{1i}, \sum_{\forall i} w_{ij} w_{2i}, ..., \sum_{\forall i} w_{ij} w_{mi}) \end{aligned} \tag{2}$$

Via Equations (1) and (2), we get 3

$$S_{content-j} \sim P(d_j \mid q) \tag{3}$$

In this model, $P(d_j \mid q)$ is used to calculate the similarity of the detected microblog text d_j and microblog keyword query *q*, $S_{content-j} \sim P(d_j \mid q)$. $P(d_j \mid q)$ reflects the query *q* to provide the coverage of the detected text *d*j. According to the structure of Figure 1, using the Bayesian formula, we can derive the following formula:

$$\begin{aligned} P(d_j \mid q) &\sim \sum_{\forall \vec{c}} p(q \mid \vec{c}) p(d_j \mid \vec{c}) \\ P(d_j \mid \vec{c}) &= \sum_{\forall k} p(d_j \mid \vec{k}) p(\vec{k} \mid \vec{c}) \\ P(d_j \mid q) &\sim \sum_{\forall \vec{c}} \{ p(q \mid \vec{c}) \sum_{\forall k} [p(d_j \mid \vec{k}) p(\vec{k} \mid \vec{c})] \} \end{aligned} \tag{4}$$

So we can retrieve the microblog rumor samples that are similar to the detected texts' key words in the microblog rumor library, then we take its average as text's key words similarity between query text and rumors sample, that is, $S_{content} = \frac{1}{n} \sum_{j=1}^{n} S_{content-j}$.

Similarly, we can calculate the similarity between microblog rumors and detected text in rhetoric, sentence features, sensitive words use, that is, $S_{rhetoric}$, $S_{sentence}$, S_{word} and give different weights respectively *A*, *B*, *C*, *D*. Finally, we can calculate the probability that detected microblog text is rumor, that is, P_{text}=*A*.$S_{content}$ + *B*.$S_{rhetoric}$,+*C*.$S_{sentence}$ + *D*. S_{word}.

3.2. 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.

3.2.1. Related Parameters

- $G = (V, E)$ is defined as the graph model of the microblog user.
- $V = \{u_i\}_{i=1}^{m}$ represents the user node set in the microblog social network model.
- $E \subseteq V \times V$ represents a set of links between users in microblog.
- $S = \{s_{ik}\}$ is defined as a set of similarity weights between users in microblog.

3.2.2. User Similarity Model Building

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 Figure 2.

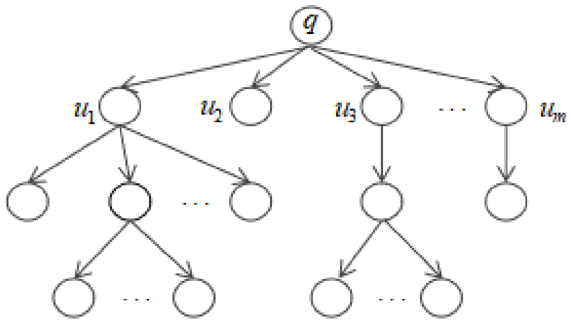

Figure 2. Microblog user modeling based on bayesian network.

3.2.3. Microblog User Behavior Similarity Calculation

Construct the conditional independence assumption of Bayesian network model.

$$P(X_i \mid \varphi_{xi}) = P(X_i \mid X_1, X_2, ..., X_{i-1}) = \prod_{t=1}^{i-1} P(X_i \mid X_t) \qquad (5)$$

Where *i*=1,2,…,*m*, $P(X_i \mid \varphi_{xi})$ represents the conditional probability distribution of the node u_i when its parent node takes the combined state value.

Calculate the prior probabilities of user nodes u_i in the case where the propagation probability is $q \in [0,1]$.

$$P(Q_{u_i} = q) = w_t / \sum_{t=0}^{1} w_t \qquad (6)$$

Where w_t indicates the number of rumors' probability *t* that the user has posted or reproduced.

To avoid the prior probability of the microblog user is 0, we introduce the Laplace estimator, then

$$P(Q_{u_i} = q) = (w_t + 1) / (\sum_{t=0}^{1} w_t + N) \qquad (7)$$

Calculate conditional probability that the similar user u_{f_k} propagates the rumor r_j is $q^{'}$ in the case where the query user propagates the rumor is *q*.

$$P(Q_{u_{f_k}} = q_k \mid Q_{u_i} = q) = g^{f_k}_{q_k,q} / \sum_{t=0}^{1} g^{f_k}_{t,q} \qquad (8)$$

Where u_{f_k} is the *k*-th similar user node of node u_i, $g^{f_k}_{q_k,q}$ represents the number that the similar user u_{f_k} propagates the same rumor's probability is $q^{'}$,in the case where the probability is *q* that the user u_i propagates rumors.

To avoid the conditional probability of the microblog user is 0, we introduce Laplace estimator, then

$$P(Q_{u_{f_k}} = q_k \mid Q_{u_i} = q) = (g^{f_k}_{q_k,q} + 1) / (\sum_{t=0}^{1} g^{f_k}_{t,q} + N) \qquad (9)$$

Calculate the similarity weight between the user and the other user

First, calculate the similarity weight between two users.

For any two users' similarity weight, we consider two users at the same time concerned about the number of users, the value can be a better reflection of the microblog users' similarity. If the value of similarity reaches a certain threshold, then there is an edge between the two social network users, and the similarity weight between u_i and u_i is denoted by s_{ij}:

$$s_{ij} = \frac{N(u_i = 1, u_j = 1)}{N(u_i = 1) + N(u_j = 1) - N(u_i = 1, u_j = 1)} \tag{10}$$

Where N=(u_i=1,u_j=1) on behalf of the user u_i and u_j u_j at the same time concerned about the number of users. The denominator represents user u_i or user u_j concerned about the number of microblog users.

Assumed that ε is the given similarity threshold, that is, when the calculated similarity value $s_{ij} > \varepsilon$, then we consider that the microblog user u_i has a similar relationship with u_j.

Second, the calculation method of the similarity's direction.

In the microblog, we will see some interactive behavior as the measurement standard of similar relationship, such as comments behavior, reposts behavior, thumb-ups behavior, mentions (@) behavior. That is, by looking for the interactive behavior of the text to find similar users. If the user u_i has a high degree of attention to the user u_j, and they will have more interaction behaviors, the direction of the similarity relation is directed by the user u_i to the user u_j. Conversely, the user u_j points to the user u_i.

According to the Bayesian rule, we can obtain the posterior probability of the query user. For rumors r_j, we calculate the probability function that user u_i spreads rumor in the case of similar users to spread the rumor, as follows

$$P(Q_{u_i} = q \mid \Psi_{u_i}) = \frac{(\prod_{d \in Z_{u_i}} w_{qd} P(\Psi_d \mid Q_{u_i} = q)) P(Q_{u_i} = q)}{\sum_{t=0}^{1}((\prod_{d \in Z_{u_i}} w_{qd} P(\Psi_d \mid Q_{u_i} = t)) P(Q_{u_i} = t)} \tag{11}$$

Therefore, it is possible to calculate the prediction probability that the user u_i propagates the rumor r_j. Where Z_{ui} represents the set of nodes other than the query user u_i ; w_{qd} represents the attribute weight between the user u_i and the similar user d when the probability of propagating rumors is q; Ψ_d represents a combined state where other users of user d propagate rumors.

Finally, the system is sorted according to the size of the predicted probability, and we see the averaged probability as the probability that the microblog users reprint or publish rumors, that is,

$$P_{user} = \frac{1}{n}\sum_{j=1}^{n} p_j(Q_{u_i} = q \mid \Psi_{u_u}) \tag{12}$$

3.3. Microblog Rumors Similarity Algorithm Based on Bayesian Network (Integrated Model of Computing)

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 Figure 3.

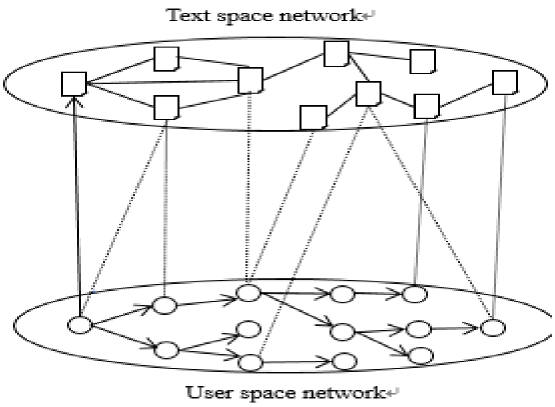

Figure 3. The microblog integrated model.

By using the causal probability reasoning of Bayesian network and calculating the first two sections' similarity, then we conduct respectively a weighted calculation for the prediction results of the two network models. Finally, we can get the integrated model calculation method proposed in this paper objectively:

$$S = \alpha \cdot P_{text} + \beta \cdot P_{user} \tag{13}$$

In this way, we can predict the similarity between detected text and microblog rumors sample in the microblog social network. In order to get the test results, we artificially set the "similar threshold" parameter. If the similarity between query microblog and rumors is greater than the similarity threshold, and we will think that the detected microblog text is similar to the rumor sample. That is, the detected microblog text is a rumor. Finally, through the experimental verification on the real data and the comparison of the evaluation index with other algorithms, we can get the validity and superiority of the proposed algorithm or some of the shortcomings.

4. 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" [26], which divides microblog rumors into several different categories, namely politics, economy, fraud, social life, common sense and other categories(Traffic, science and technology, education, international news, entertainment and gossip, medical and health, law and order). In the analysis, we deleted 92 too short (the text content of the index point symbol and the Chinese character adding up to a total of no more than 10 characters) microblog data,

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 randomly. 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 its 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.

4.1. Experimental Data

4.1.1. Text Data

The detected microblog text: 2012/8/10 8:39:53 Anxiously. The everyone to help turn: Chen Chen, two years old, was burned by just the pan of the oil, has been sent to the hospital. The doctor said that the situation is very dangerous, need to spend five hundred thousand, and need urgently good people to help donations. Every time you forward, a child can get 3 cents love support. More than one person to repost, more than a hope. Help this poor little boy.

- Tag: rumor
- Facts: This microblog picture of the boy named Xiao Huang, due to neglect of parents were burned, in 187 hospital burns department of Haikou in 2008.

Extract the text feature vector as follows,

- Content keywords: anxiously, donations, love support, help.
- Rhetoric: exaggeration.
- Sentence features: exclamation sentence, short sentences.
- Sensitive words: just the pan of the oil, very dangerous, need urgently.

4.1.2. User Data

- The number of comments: 46
- The number of reposts: 442
- The number of thumbs up: 0

The microblog rumor producer concerned about the number of followers with 21, the number of followees with 39. Compared with the similarity threshold, which has a similar relationship with 25 users, the 25 users concerned about the second users. Through extracting every user's recent 30 microblogs, at last we can get 750 microblogs. In order to reduce the complexity of the experiment, this paper only extracted two users' microblog quantity.

4.1.3. Parameter Setting

In order to guarantee the simplicity and validity of the similarity algorithm proposed in this paper, we take 1 as the weight of the eigenvector the keywords in the microblog text model. The value of the similarity threshold ε is 0.8 in the microblog user model. After several experiments to adjust the values of parameters α and β, the experimental results show that the experimental results when α and β are 0.6 and 0.4 respectively are the most significant, the F value is the largest and the accuracy is the highest. Therefore, in the combined similarity calculation formula proposed in this paper, α takes 0.4 and β takes 0.6.

4.2. Evaluation Indicators

We compared the three algorithms of similarity algorithm based on Bayesian network, similarity algorithm based on TF-IDF and similarity algorithm based on Vector Space Model (VSM). The three indexes of the accuracy rate *P*, recall rate *R* and *F*-measurement are used to detect the effectiveness of the algorithm in this experiment.

Accuracy rate *P*: The number of relevant microblog text correctly identify/the total number of identified text in the algorithm

Recall rate *R*: The number of related microblogs correctly identify/the total number of related texts in the model

$$F\text{-measurement}: \; F = \frac{2P \times R}{P + R} \tag{13}$$

4.3. Verifications

To meet the needs of verification, we cited data from Liu *et al* [26] "Statistical and Semantic Analysis of Rumors in Chinese Social Media" and transformed some microblog rumors into non-rumors. To strengthen the feasibility of this model, the experimental data will be randomly divided into a training data set and test data set to detect the two parts. So, we consider the political class, economy class, fraud class, social life class, common sense class microblog text a total of 3000, and according to various rumors in the rumors of the proportion of the allocation of randomly selected sets of text to be detected the number of various types of text. Among them, we select 200 fraudulent microblog texts as a training data set, including rumors and non-rumors related to the sample of 100. We select other texts as a test data set, among them are similar to the detected text with 1000, including rumors and non-rumors related to the text of the 500. As shown in Table 1.

Table 1. Microblog text classification table.

Microblog classification		**Fraud class**		**Political class**	**Economics class**	**society Life class**	**Common sense class**	**Total**
Training sample set	Rumors	200	100	0	0	0	0	200
	Non-rumors		100					
Testing data set	Rumors	1000	500	1000	400	1200	200	2800
	Non-rumors		500					
Total		1200		1000	400	1200	200	3000

4.4. Results and Analysis

4.4.1. Comparison of Experimental Results

To test the effectiveness of the microblog similarity algorithm based on Bayesian network, we compared the algorithm with the similarity algorithm based on TF-IDF [10] and the similarity algorithm based on the vector space model VSM [7, 14]. The experimental results are shown in Table 2.

Table 2. The experimental results of the three algorithms.

Sort the results according to the similarity of the size of the microblog texts	**The number of related texts has been correctly identified**			**The total number of identified text**	**The total number of related texts in the model**
	Similarity Algorithm Based on *BN*	**Similarity Algorithm Based on *TF-IDF***	**Similarity Algorithm Based on *VSM***		
100	96	71	68	100	1000
200	182	136	152	200	1000
300	258	171	237	300	1000
400	332	204	294	400	1000
500	382	242	368	500	1000
600	447	324	432	600	1000
700	511	413	497	700	1000
800	580	456	556	800	1000
900	639	504	612	900	1000
1000	701	556	672	1000	1000
1100	759	583	715	1100	1000
1200	822	612	768	1200	1000

4.4.2. Comparison of Evaluation Results

It can be seen from Table 3 that the accuracy rate of the microblog similarity algorithm based on Bayesian network can reach 70.1% in the top 1000 results, and the recall rate can reach 70.1%, and the *F*- measure can reach 70.1%, which shows that the proposed algorithm in this paper is still relatively satisfactory results. Compared with the similarity algorithm based on the TF-IDF, the average accuracy rate is improved by 22.9%, the average recall rate is increased by 6.7%, and the average *F*-measure value is increased by 14.5%. Compared with the similarity algorithm based on the vector space model VSM, the average accuracy rate is improved by 7%, the average recall rate is increased by 2.8%, and the average *F*- measure is increased by 4.2%.

Table 3. The evaluation results of the three algorithms.

Sort the results according to the similarity of the size of the microblog texts	**Similarity Algorithm Based on *BN***			**Similarity Algorithm Based on *TF-IDF***			**Similarity Algorithm Based on *VSM***		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
100	.960	.096	.175	.710	.071	.129	.680	.068	.124
200	.910	.182	.303	.680	.136	.227	.760	.152	.253
300	.860	.258	.397	.570	.171	.263	.790	.237	.365
400	.830	.332	.474	.510	.204	.291	.735	.294	.383
500	.764	.382	.509	.484	.242	.323	.736	.368	.491
600	.745	.447	.559	.540	.324	.405	.720	.432	.540
700	.730	.511	.601	.590	.413	.486	.710	.497	.585
800	.725	.580	.644	.570	.456	.507	.695	.556	.618
900	.710	.639	.673	.560	.504	.531	.680	.612	.644
1000	.701	.701	.701	.556	.556	.556	.672	.672	.672
1100	.690	.759	.723	.530	.583	.555	.650	.715	.681
1200	.685	.822	.747	.510	.612	.556	.640	.768	.698
Average value	.776	.476	.547	.568	.409	.402	.706	.448	.505

4.4.3. Results Analysis

Figures 4, 5, and 6 show the accuracy rate, recall rate, and *F*-measurement that we obtained under different similarity algorithms. Regardless of the evaluation index of the algorithm, the algorithm proposed in this paper is higher than the other two algorithms. In this case, with the number of detected texts increasing, the accuracy decreases, and the recall rate and *F*-measurement increases. This is mainly due to the number's increase of detected text, it increased the scope of the query, so resulting in increased recall rate. With continuous expansion of the inspection range, the error will increase and accuracy rate will decrease. It can be seen from Figure 7 that the accuracy rate of the similarity algorithm based on Bayesian network is obviously higher than that similarity algorithm based on *TF-IDF* and similarity algorithm based on VSM in the same recall rate.

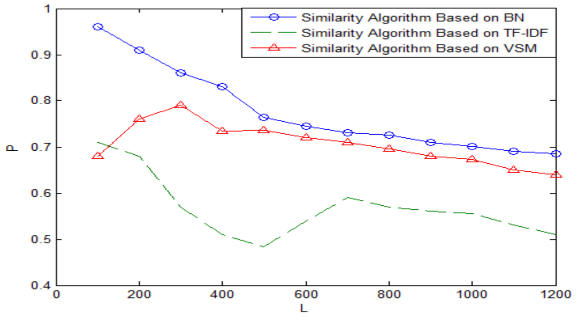

Figure 4. Comparison of the accuracy of the three algorithms.

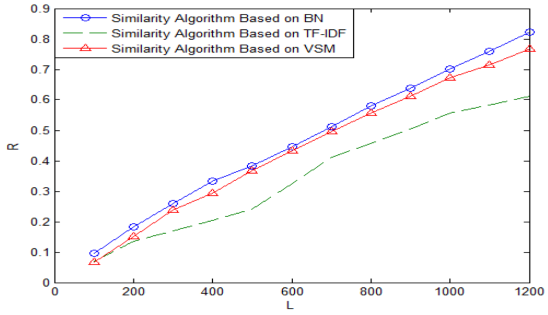

Figure 5. Comparison of the recall rate of the three algorithms.

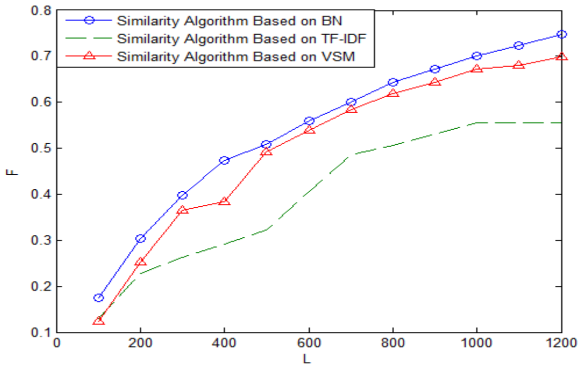

Figure 6. Comparison of the *F*-measurement of the three algorithms.

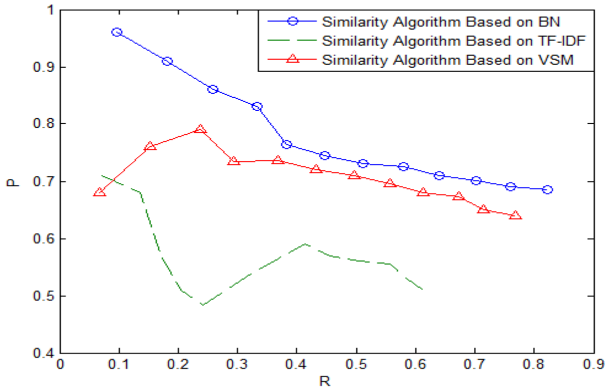

Figure 7. Comparison of the P / R of the three algorithms.

It can be seen from the above analysis that compared with the similarity algorithm based on TF-IDF and the similarity algorithm based on vector space model VSM, the similarity algorithm based on Bayesian network proposed in this paper has achieved good results in terms of accuracy rate *P*, recall rate *R* and *F*-measurement evaluation index, and verified that the algorithm in microblog rumours similarity is effective, which is very valuable in the research of microblog rumours

5. Summary

5.1. 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 rumor, we generate two models of the text and user into a unified model, so that we can improve the research accuracy of the microblog rumours 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 rumour text and the user data set. The experimental results show that the similarity algorithm proposed in this paper can be used to identify the rumor of texts and predict the characters of users more accurately and effectively.

5.2. Interventions

In this paper, we believe that the content characteristics and publishing user characteristics of rumors are different from non-rumors, and thus constitute the key elements for identifying rumors and non-rumors. Therefore, according to the research in this paper, we want to identify microblog rumors and control the spread of rumors. We have to conduct in-depth discussions from two aspects. On the one hand, the characteristics of the microblog text information can be analyzed, and the most representative features are extracted for calculation, thereby judging the probability that the microblog text content is a rumor. On the other hand, the behavior characteristics of the microblog users can be tracked, and the probability of the microblog users posting or reprinting a rumor is determined according to the behavior history records of the microblog users. Finally, a comprehensive measurement can more accurately determine the rumor information in the microblog information.

5.3. Prospect

In this paper, the similarity algorithm proposed in this paper provides a new way of thinking compared with the traditional algorithm. And it can be more effective and precise to predict the probability that microblog text is expected to be rumor from a wide range of microblog data. For the work of the research content in the future, this paper provides two directions to think and improve. On the one hand, based on the research direction of the microblog text content proposed in this paper, we can consider adding the image information, video image, link file and so on, and improve the similarity method and calculation method. On the other hand, according to the research of microblog social network user modelling proposed in this paper, we only takes into account the direct similarity between users, and do not consider the possible non-direct similarity between users. But the indirect similarity has a significant effect between users. It is hoped that the similarity discovery method of rumors in mass social network data will be studied in order to further improve the effect of text prediction.

Acknowlegement

This work was supported in part by the National Natural Science Foundation of China (No. 71701115), the Shandong Natural Science Foundation (ZR2017MF058), and Special Project for Internet Development of Social Science Planning Special Program of Shandong Province (17CHLJ23), Social Science Planning Program of Shandong Province(13BTQJ01).

References

[1] Cheng A. and Xia C., "On the Linguistic Features of Microblog Rumor," *Southeast Communication*, no. 11, pp. 98-100, 2014.

[2] Chengcheng L., An Z., Qingwen Q., and Huimin S., "Grid-based Location Microblog Data Fetching and Human Information Extraction," *Science of Surveying and Mapping*, vol. 42, no. 2, pp. 125-129, 2017.

[3] Chun Z., Xiao X., Zhu Z., "A Personalized Search Algorithm by Using Content-Based Filtering," *Journal of Software*, vol. 14, no. 5, pp. 999-1004, 2003.

[4] Dong C., Qiang D., Yang G., and Da L., "Research on the Evolution of Micro-blog User Information Personalized Recommendation Model Based on LDA," *Information Science*, no. 8, pp. 3-10, 2017.

[5] Ding Y., Liu F., and Tang B., "Context-Sensitive Trust Computing in Distributed Environments," *Knowledge-Based Systems*, no. 28, pp. 105-114, 2012.

[6] Fan B., Wang J., Ge Y., and Ma M., "Research on the Spatial Structure of Chengdu-Chongqing City Group Based on Weibo Sign-in Data and Its Inter-city Population Flow," *Journal of Earth Information Science*, vol. 21, no. 1, pp. 68-76, 2019.

[7] Guo Q., Li Y., and Tang Y., "Research on Text Similarity Calculation Based on VSM," *Application Research of Computers*, no. 11, pp. 3256-3258, 2008.

[8] Han Y. and Lu J., "A Summary of Weibo Proverbs Communication Model," *Network Security Technology and Applications*, no. 11, pp. 41-43, 2018.

[9] Hu M., Cai S., and Zhang Y., "Research on Personalized Recommendation-oriented Contextual User Profile," *Journal of Intelligence*, vol. 29, no. 10, pp. 157-162, 2010.

[10] Huang C., Yin H., and Hou Y., "A Text Similarity Measurement Method Combining Term Semantic Information and TF-IDF Method," *Chinese Journal of Computers*, vol. 34, no. 5, pp. 856-864, 2011.

[11] Hassan M. and Shoaib M., "Opinion within Opinion: Segmentation Approach for urdu Sentiment Analysis," *The International Arab Journal of Information Technology*, vol. 15, no. 1, pp. 21-28, 2018.

[12] Lawrence R., Almasi G., Kotlyar V., Viveros M., and Duri S., "Personalization of Supermarket Product Recommendations," *Data Mining and Knowledge Discovery*, vol. 5, pp. 11-32, 2001.

[13] Li F., Hou J., Zeng R., and Li C., "Research on Multi-Feature Sentence Similarity Computing Method with Word Embedding," *Journal of Frontiers of Computer Science and Technology*, vol. 11, no. 4, pp. 608-618, 2017.

[14] Li L., Zhu A., and Su T., "Research and Implementation of an Improved Vector Space Text Similarity Algorithm," *Computer Applications and Software*, vol. 29, no. 5, pp. 282-284, 2012.

[15] Li Q. and Gu J., "Activity Driven Modelling of Online Social Network," *Journal of Systems Engineering*, vol. 30, no. 1, pp. 9-15, 2015.

[16] Li X., Cao H., Ding C., and Huang L., "Short text Classification Based on HowNet and Domain Keyword Set Extension," *New Technology of Library and Information Service*, no. 2, pp. 31-38, 2015.

[17] Li X., Liu K., Ding C., and Liao X., "Title Information Classification Based on Hownet Semantics Feature Extension," *Library Journal*, no. 2, pp. 11-19, 2017.

[18] Liao H., Wang Y., and Guang P., "Topic Mining and Viewpoint Recognition of Different Communicators in Weibo Public Opinion Communication Cycle," *Library and Information Work*, vol. 62, no. 19, pp. 77-85, 2018.

[19] Ling H., "Research Overview on Bayesian Network," *Natural Sciences Edition*, vol. 23, no. 1, pp. 33-40, 2013.

[20] Liu F., Li X., Ding Y., Zhao H., Liu X., and Ma Y., Tang B., "A Social Network-Based Trust-Aware Propagation Model for P2P Systems,"

Knowledge-Based Systems, no. 41, pp. 8-15, 2013.

[21] Liu F., Wang L., Gao L., Li H., Zhao H., and Sok Khim Men., "A Web Service Trust Evaluation Model Based on Small-World Networks," *Knowledge-Based Systems*, no. 57, pp. 161-167, 2014.

[22] Liu F., Wang L., Johnson H., and Zhao H., "Analysis of Network Trust Dynamics Based on Evolutionary Game," *Transaction E: Industrial Engineering*, vol. 22, no. 6, pp. 2548-2557, 2015.

[23] Liu F., Zhu X., Hu Y., Ren L., and Johnson H., "A Cloud Theory-Based Trust Computing Model in Social Networks," *Entropy*, vol. 19, no. 1, pp. 1-11, 2017.

[24] Liu M., "Post-Truth Era Microblog Rumors Spread and Governance-Taking the "Henan Eye Cancer Girl Incident" as an Example," *New Media Research*, vol. 4, no. 22, pp. 21-23, 2018.

[25] Liu Y., Liang X., and Yang X., "Information Spreading Model of Weibo Network Based on Petri Net," Chinese Management Science, vol. 26, no. 12, pp. 158-167, 2018.

[26] Liu Z., Zhang L., Tu C., and Sun M., "Statistical and Semantic Analysis of Rumors in Chinese Social Media," *Scientia Sinica Informationis*, no. 12, pp. 1536-1546, 2015.

[27] Mepherson M., Smith-Lovin L., and Cook J., "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, pp. 415-444, 2001.

[28] Qin J., Wang W., Xiao C., Lu Y., Lin X., and Wang H., "Asymmetric Signature Schemes for Efficient Exact Edit Similarity Query Processing," *Acm Transactions on Database Systems*, vol. 38, no. 3, pp. 1-44, 2013.

[29] Ren Y., Wang Y., and Wang G., "Research on the Evolution Mechanism of Micro-blog Rumors," *Journal of Intelligence*, vol. 31, no. 5, pp. 50-54, 2012.

[30] Resnick P., Iacovou N., Suchak M., Bergstrom P., and Riedl J., "Grouplens: An Open Architecture for Collaborative Filtering of Netnews," *in Proceedings of ACM Conference on Computer Supported Cooperative Work*, North Carolina, pp. 175-186, 1994.

[31] Saini J., "Psychoanalysis of Online Behavior and Cyber Conduct of Chatters in Chat Rooms and Messenger Environments," *Advanced Networking and Applications*, vol. 6, no. 2, pp. 2214-2221, 2014.

[32] Sun Y. and Li S., "Similarity-Based Community Detection in Social Network of Microblog," *Journal of Computer Research and Development*, vol. 51, no. 12, pp. 2797-2807, 2014.

[33] Wang H. and Cai P., "Research on Weibo Proverbs Communication Network," *Library and Information Research*, vol. 11, no. 1, pp. 37-42+49, 2018.

[34] Yang Z. and Huang H., "Graph Based Word Sense Disambiguation Method Using Distance Between Words," *Journal of Software*, vol. 23, no. 4, pp. 776-785, 2012.

[35] Yao B., Ni J., Yu P., Li L., and Cao B., "Micro Blog User Recommendation Algorithm Based on Similarity of Multi-Source Information," *Journal of Computer Applications*, no. 5, pp. 1382-1386, 2017.

[36] Yin D., Hong L., and Davison B., "Structural Link Analysis and Prediction in Microblogs," *in Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, Glasgow, pp. 1163-1168, 2011.

[37] Zhang B., Zhang Y., and Gao K., "Combining Relation and Content Analysis for Social Tagging Recommendation," *Journal of Software*, vol. 23, no. 3, pp. 476-488, 2012.

[38] Zhi-yun Z., Chun-yuan J., Zhen-fei W., and Dun L., "Computing Research of User Similarity Based on Micro-blog," *Computer Science*, no. 2, pp. 262-266, 2017.

[39] Zhu X. and Liu F., "Research on Behavior Model of Rumor Maker Based on System Dynamics," *Complexity*, pp. 1-9, 2017.

Chengcheng Li is a graduate student studying in Shandong Normal University. Her research interests include rumor spreading and governing.

Fengming Liu is a professor of the school of business at Shandong Normal University. His research interests include trust and social computing, game theory, and network behaviors dynamics.

Pu Li is a graduate student studying in Shandong Normal University. Her research interests include rumor spreading an d governing.