Research on Data Mining Technology of Social Network Associated Information

Yanxin Jiang[1(✉)], Xian Mei[2], and Guanglu Sun[2]

[1] Heilongjiang University, Harbin 150080, Heilongjiang, China
jyxl977@sohu.com
[2] Harbin University of Science and Technology,
Harbin 150080, Heilongjiang, China

Abstract. With the popularization of Internet social networking service, the results of association data mining between friend dynamic, microblog and moments that user posting and giving feedback information, which have important influence on government planning, business management and personal affairs decision-making activities. This paper studies the data mining technology of social network related information, analyzes the text data in social network by using the finite state automata (DFSA) and word frequency - reverse file frequency (TF-IDF), and using tree algorithm to sort the data. The simulation results show that this method can realize the classification data mining of social network related information.

Keywords: Social network · Data mining · Ideological · Political education

1 Introduction

Social networking has evolved from the traditional World Wide Web. The main characteristic of the traditional World Wide Web is that the information of the graphic contents is organized and presented to the visitors through the links. Social networks pay more attention to the relationship between the creator and the reader of the graphic content. Usually, they can indicate the consistency of the content or the degree of concern by publishing the argument, reading, commenting or replying, praising, forwarding, and the like.

The six-degree theory of division proposed by Professor Stanley Milgram who is in Harvard University states that no two strangers will be separated by more than six people [1]. Through the analysis of social network related information can be found in the core content of social networks and the relationship G can be expressed as formula 1.

$$G = \langle V_c, V_u, E_f, E_r, E_c, E_a, E_t \rangle \tag{1}$$

Where Vc represents a content point set, Vu represents a user point set, Ef represents a side set of friend relationships between users, Er represents a side set of reading relationships, Ec represents a side set of comments or reply relations, Ea represents a point-like relationship side set, Et represents the edge set of the forwarding relationship.

S. Liu et al. (Eds.): eLEOT 2018, LNICST 243, pp. 17–26, 2018.
https://doi.org/10.1007/978-3-319-93719-9_3

This project mainly studies the influence of social network on the current college students 'groups and how to make use of the analysis of social network related data to understand the students' thoughts. On the basis of a large number of researches, the paper designs a university ideological and political support system based on social network associated data mining. The system uses DFSA and TF-IDF algorithms to record and analyze students 'social network data. Based on the decision tree theory, this system conducts an early warning assessment of the students' thoughts which release abnormally sensitive words. Then, the improved frequent set algorithm is used to find the people who are highly correlated with the problem Or the reason, and the key factor to solve the problem, prompt the mental educator to divert the pressure of the trainee, so as to put an early stop of the malignant event caused by the psychological reason in the stage of less damaging germination.

DFSA algorithm is an effective multi-pattern matching string algorithm. This algorithm constructs a finite automaton to convert the multi-pattern string matching problem to a relatively simple one-pattern string matching problem, and only needs to scan the body part once to check the matching of all the key words [2]. The system uses the DFSA multi-pattern string matching algorithm to compare sensitive words with the text of social network articles as well as to count a number of times and the importance of the expression of sensitive words in social network articles.

2 System Structure

The system solves the common problems of ideological education in colleges and universities by using social network-related information analysis. It consists of social network data collection and processing, decision-making early warning decision-making and auxiliary intervention and other three major modules, the specific structure shown in Fig. 1.

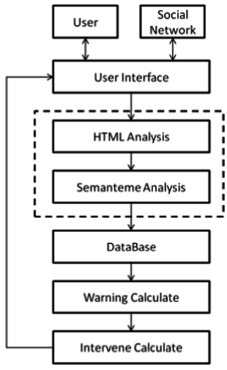

Fig. 1. Chart of system architecture.

2.1 Development Tools and Database

Due to the generally high security of social networks, traditional web crawler development tools often fail to log in and retrieve data. Therefore, the system adopts to introduce the Internet browser component, and completely simulates the live action to log in to the social network as a legitimate user, complies with the network application norms without violating the privacy of others, but only improves personal browsing speed and data sorting efficiency. However, in view of the current university computer hardware configuration level is generally not high, the database using Microsoft Office integrated Access database for the local database.

2.2 System Module Division and Function

As we all know, undergraduates are curious and exploratory. Majority of them don't have enough sufficient social experience, lack of correct judgment as well as control over things. What's worse, they are easily influenced by the outside world. [3] For instance, some students are addicted to the games that affects not only performance decline, but also their education and employment, which even lead to drop out of school; some students can't properly handle the relationship of classmates or roommates, then they do some over-excited behavior; others couldn't face the family, work and emotional problems properly, prone to psychological and emotional fluctuations, unable to extricate themselves, leading to suicide or self-mutilation [4]. And there is no doubt that we are in the era of "Internet +", the Internet has become an indispensable part of people's study and life. It is not uncommon for students on the campus even in the classroom to use their mobile phones frequently that deeply impact the students' learning concentration. Furthermore, it is also easy to give false guidance to public opinion. As has been stated above, it is the ideological and political work in colleges and universities that will face new problems and challenges.

According to questionnaire survey to some college students, we have found that the vast students use social networking software to give vent to their emotions and express their stress and emotional inclination when they are under the influence of pressure or mood swings [5]. The dangers of misanthropy and misanthropy in college students often show up in their personal social software [6]. If Ideological and political workers can use the Internet to understand students' psychological and emotional changes, using data mining techniques to reflect the student thought in social software dynamic integrated data statistical analysis, they will find the problems in a timely manner so they can admonish and stop students, complete promptly warning education and undertake to the student psychological construction and channel. Only by doing so, can they urge college students better achieve academic record, reduce involved is not deep, personal injuries and property losses [7, 8].

According to users' requirements, the system is roughly divided into three functional modules. Firstly, the social network information acquisition and processing module is to log in the social network to obtain data and process the irrelevant information and query the frequency and effect of the sensitive words. Secondly, the

early warning data mining module is to deduce the classification rules of the data and form an early warning rule for the students who need to make early warning of ideological and political education. Last but not least, the function of auxiliary module Ideological education is to accept early warning education students strongly associated with friends or trigger students' psychological changes and any other main factors in order to carry out more targeted education or counseling.

3 Social Network Information Acquisition and Processing Module Design

3.1 Use DFSA Algorithm to Find the Key Words in the Main Body of Social Network

The system takes advantage of Microsoft Visual Basic's Internet Explorer component to log on to the social network as a normal user and retrieve the original Hypertext Transfer Protocol code from the social networking server. It is possible to legally obtain the important information of the network article body, the good friend ID, the good friend comment body and the like posted by the social friends by the analysis and filtering of the protocol code.

System uses the DFSA algorithm matching process to achieve the main pseudo-code described below:

```
Input: Social Network body content [] and its length n, jump function goto (), failure
function failure () and output function output ()
Output: Match the number of sensitive information
1. state = 0 ;
2. count = 0 ;
3. text_mentality = NULL ;
4. for( i=0 ; i<n ; i++){
5.   while( goto( state , content[i] ) == FAIL )
6.     state = failure( state ) ;
7.   state = goto( state , content[i] ) ;
8.   if( output( state )!=EMPTY )
9.     count = count + 1;
10. }
11. if(count>=1)
12.    text_mentality = "dispirited" ;
13. return count;
```

Using the DFSA algorithm, the text of a student's social network can be preprocessed, the keywords extracted and counted, and these text tags can be marked as the game related text, the text related to negative emotions, the text related to positive

emotions and criminal related body. What's more, it can pre-process multiple social networks for multiple students to facilitate early warning data mining.

3.2 Using TD-IDF Algorithm to Determine the Importance of the Key Vocabulary in the Social Network Text

However, in social network articles, the frequency of the mere statistics of sensitive words does not necessarily indicate whether the author's emphasis is on the semantic meaning of the sensitive words. Therefore, the system makes use of the TF-IDF algorithm to evaluate the importance of the word in the text and the corpus. In TF-IDF algorithm, TF represents the word frequency, that is, the number of occurrences of sensitive words in the document. IDF represents the inverse document frequency, meaning that if the document containing sensitive words is fewer, the short text's disguishability is better.

Take ti for an example, which is in a social network text and its importance in short text can be expressed as (2)

$$\mathrm{TF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

Wherein, the number of occurrences of this word in the text dj indicates the total number of occurrences of the word in the document set. For a particular sensitive term ti, the higher the frequency of occurrence in social articles, the lower appears in the entire document set, and the higher the TF-IDF value is, the greater the weight in the entire database.

4 Early Warning Data Mining Module Design

Decision tree classification method is one of the most widely used inductive inference algorithms. It is a method of approximating discrete-valued functions and it can obtain a tree-like representation [9]. In addition to, its algorithm is easy to understand and convert to binary or multi-branch classification rules with high classification efficiency.

The system adopts the popular ID3 algorithm in decision tree classification method to conduct early warning data mining. The ID3 algorithm can form an early warning decision tree by calculating the information gain by using the training class marked by the database class, and then it can use the established decision tree to analyze and predict the sample data.

ID3 algorithm is as follows:

```
Input：D：data set
C：classification attribute
Output：decision tree
1. create node N;
2. if(without other attribute in D)
3.   label N with most common value of C in D;
4. else if(all instances in T have the same value V of C)
5.   label N, "X.C=V with probability 1" ;
6. else {
7.  for each(attribute A in D)
8.    AM=the attribute of mini Avg Entropy(A,C,D);
9.  if (Avg Entropy(AM,C,D) is not substantially smaller than Entropy(C,D))
10.     label N with most common value of C in D;
11.   else{
12.     label N with AM;
13.     for( each value V of AM ){
14.       N1=ID3(subtable (D,A,V),C);
15.       if(N1!=null)
16.         arc from N to N1 labelled V;
17.       }
18.    }
19. }
20. return N;
```

5 Intervention Module Design

Once found that students who by reason of mood swings, or distracted with serious decline in academic performance or abnormal behavior trend, closely related to the need to immediately find classmates or relative personnel to understand the details such as making psychological counseling programs.

The Apriori algorithm is an efficient method to discover frequent itemsets in one-dimensional mode, which uses iterative methods of layer-by-layer search to explore larger frequent itemsets. Furthermore, the basic nature of this algorithm is that all non-empty sub-sets of frequent itemsets are frequent [10, 11]. Although the classic Apriori algorithm describes the steps and methods of discovering frequent itemsets more concisely, as a number of data items increases, the demand for system memory and CPU resources increases rapidly, and the system I/O load is huge as well [12].

Considering the time and space of ideological and political education in colleges and universities are limited, the system adopts the Apriori frequent set algorithm which is improved when the students need to look for students and social relations closely related to the early warning students, and excavates and sorts the network ID according to the degree of social concern [13].

That is to say, the first step is to mark all occurrences in the relevant social network ID as a set of elements. Next, the social network ID and social network ID of students who like the daily social network are marked as 2 sets of elements. Next, the number of occurrences of each 2 sets of items is recorded as p(i, j). Then, the number of times the social network ID and the social network ID of the student who responded to the message in the daily social network of a student is represented as r(i, j). Lastly, the two network ID's attention function can be expressed as (3)

$$a(i,j) = p(i,j)^{\wedge}(r(i,j)+1) \tag{3}$$

According to the attention function, the network ID is arranged in descending order as a reference for the close relationship with the trainee students. And give priority to the students or friends who are closest to the social network to try to understand the cause of the problem or the way to solve the problem. If the initial communication is not sufficient to acquire enough information to construct three sets of items, and try to make an appointment with two other members to find out the nature of the problem as soon as possible

6 System Implementation and Simulation

The ideological and political education work is a thousand times, the space limit cannot be listed in detail. Here is one of the contents of the academic warning part of the design of simulation experiments to test the practical effect of the system.

In the experiment, 32 students' average grade and data of the first six semesters were selected. Based on the recent seventh semester's online social data texts, DFSA algorithm and TF-IDF algorithm were used to determine the frequency and importance of using sensitive keywords, obtaining social network data Parameters as well as forming training data sheet for the seventh semester of academic assessment of early warning (Fig. 2).

Using the previous semester's social network data (see Fig. 3), the average grade (see Fig. 4), and the hanging data (see Fig. 5), the ID3 algorithm was used to construct the decision tree. We have the social network text-sensitive words, the average score,

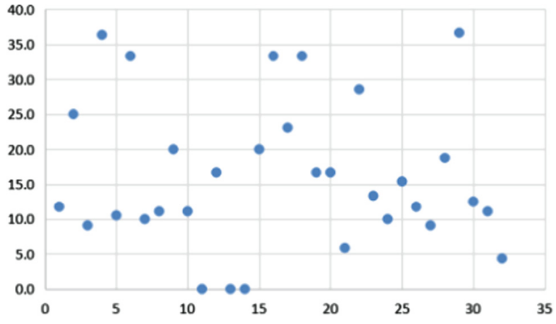

Fig. 2. Chart of social net data parameter distribution.

with or without history as a property. In other words, there are two types of social network text sensitive words, namely, more than 20% of social networking texts refer to sensitive words, less than 20% of social networking text Mention sensitive words. Also, the average grade is divided into three categories, a class of 65 points or less, a class of 65 points to 80 points, a class of 80 points or more. Besides, there is no history of hanging into two categories, with or without. The ID3 algorithm is used to mine the training samples to get the decision tree (see Fig. 6).

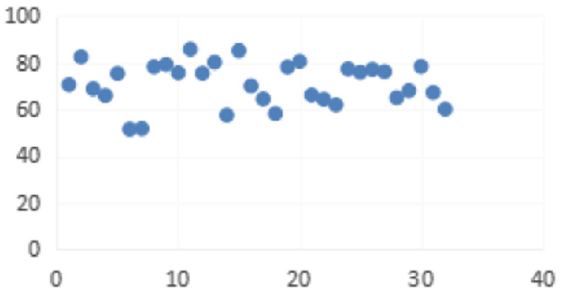

Fig. 3. Chart of average score distribution.

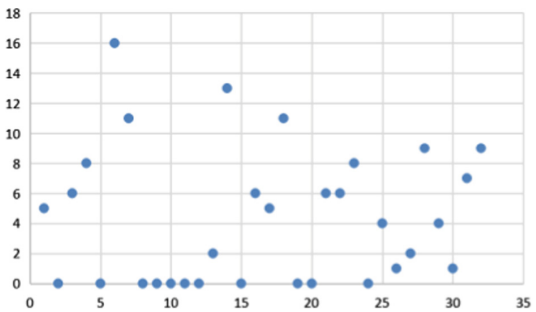

Fig. 4. Chart of fail the exam data distribution.

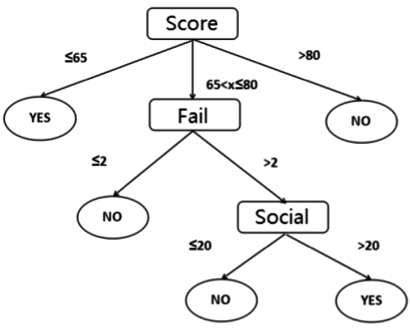

Fig. 5. Chart of early-warning decision tree.

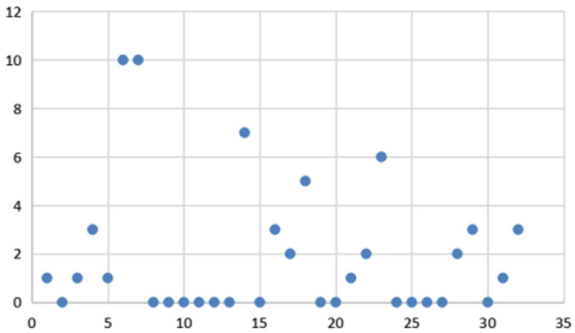

Fig. 6. Chart of early-warning data distribution

According to the results of the early warning, 12 students with the serial numbers of 4, 6, 7, 14, 16, 17, 18, 22, 23, 28, 29 and 32 were in a highly dangerous state and were in line with the actual results of the Seventh Semester and the distribution of hanging branches Fig. 6.

In combining with the frequent set algorithm, students who had close contact with early warning students active in social networks were found to have in-depth exchanges and we found that some of the major reasons of the drop in their academic performance were the fact that they recently participated in too many extracurricular practical activities. Thus, they couldn't arrange their time and energy very well. What mentioned above has accumulated experience for the follow-up ideological education guidance.

7 Conclusion

Designing and implementing a dynamic thinking system of college students is based on social network data mining technology. It uses DFSA and TF-IDF algorithms to analyze social network articles, ID3 algorithm deducing the early-warning decision tree and the improved Apriori algorithm finds frequent itemsets, so as to achieve early warning and timely disposal of college students' academic and ideological and political education.

References

1. Shi, Y., Yuan, Y.: Research on the mode of information dissemination based on social network. Libr. Trib. (06) 220–223 (2009)
2. Lu, D., Li, S., Xu, C.: DFSA Algorithm for Adaptive Frame Length Adjustment with CHI Tagging. J. Harbin Univ. Sci. Technol. (01), 56–60 (2015)
3. Yang, S., Wang, J., Dai, B., Li, X., Jiang, Y., Liu, Y.: Research status and prospect of user behavior in online social networks. Bull. Chin. Acad. Sci. (02), 200–215 (2015)
4. Yao, Q., Ma, H., Yan, H., Chen, Q.: Analysis of individual behavior of social network users from the perspective of psychology. Adv. Psychol. Sci. **22**(10), 1647–1659 (2014)

5. Huang, F., Peng, J., Ning, L.: An evolutionary model of social network views based on information entropy. Acta Phys. Sin. (16), 16–24 (2014)
6. Li, H., Zhou, Z.: Early warning system of university students' grade based on data mining. J. Daqing Pet. Inst. (04), 91–95 (2011)
7. Wu, K.: Machine learning based prediction system for student grading and research. J. Taiyuan Urban Vocat. Techn. Coll. (12), 178–180 (2016)
8. Wang, Y., Wang, P.: Study on construction of early warning system for college students. Shanghai Educ. Eval. Res. (03), 36–40 (2014)
9. Lu, D., Ling X.: DFSA algorithm for unequal long time slots in full subgroup. Technol. Meas. Control (09), 55–59 (2013)
10. Sun, J., Wang, X.: Adaptive fuzzy decision tree algorithm. Comput. Eng. Des. **34**(02), 649–653 (2013)
11. Li, Q., Zhou, X., Wang, L., Zhou, W.: Minimum combination method for mining maximal frequent sets. Appl. Res. Comput. **3**(03), 702–704 (2008)
12. Gao, C., Shen, D., Yu, G., Nie, T., Kou, Y.: A method for mining frequent sets based on uncertain data. Proc. Conf. Natl. Database Churches 82–87 (2008)
13. Chen, X.: A frequent mining of association rules with constraints. Comput. Eng. Appl. (02), 205–208 (2003)