Chat Analysis to Understand Students Using Text Mining

Yao Leiyue and Xiong Jianying

Department of Software Research,
JiangXi Bluesky University,
Nanchang China 330098
ylyyly2001@163.com, special8212@sohu.com

Abstract. Network communication has been the main method in people communication. The goal is to help educators grasping the students' social and psychological status and personal characteristics by mining conversation contents of a student group. In this paper, the text mining is used to find hot topics in chat groups, personal language behavior, personal motivation, and members of the emotions involved in the overall performance trends and the personal emotional expression trend. By using word frequency analysis, co-occurrence word frequency analysis, and lexical emotional similarity analysis, experimental results show that the method can quickly and effectively grasp the characteristics of subjects for students to develop strategies provide the basis for education.

Keywords: Chat analysis, Text mining, word orientation, Student behavior.

1 Introduction

People interactive each other all the life, and with the development of information technology, to communicate through the network has become very popular, and is the mainstream of communication. Network communication includes receiving and sending e-mails, distance learning, online searching, cooperating in collaborative environments by chatting, or browsing the tremendous amount of online information. Currently, there are several chat tools available on the internet, and the emergence of such tools has realized the communication among the internet users from all over the world. Some of them are very popular such as ICQ, hotmail. In China, most Internet users prefer to use Tencent's QQ, it has also become an important tool for office and entertainment [1]. QQ chat text is conductively to a certain degree of monitoring, such as to aid in crime detection or even crime prevention. The children and young teenagers chats their thoughts and desires in a relatively short period of time, and expressed a certain literary form. In particular, university students is a popular, for a critical stage of psychological development, the analysis of the conversation text of them is conducive to help educators understanding the social psychology, personality status, hot issues among them, and make best education management for students [2].

The motivation is that the current chat content analysis techniques are basically manual [3], which is difficult, costly, and time consuming. A text mining techniques for the problem of automatic monitoring of QQ chat group is presents in this paper. By use text mining, these research including group membership, individual characteristics, hot issues, and emotions recognition among the group could been done.

D. Zeng (Ed.): Advances in Control and Communication, LNEE 137, pp. 235–243.
springerlink.com

The paper is organized as follows; section 2 describes the text mining and an overview of the chat room or group chat content monitoring problem. Section 3 gives an overview of our study and our experimental results. The experimental evaluation is discussed in section 4. Finally, section 5 concludes this paper and discusses possible extensions.

2 Related Work

Text mining is a rich semantic analysis process of text to understand the content and meaning it contains. It has become an important field of data mining. Existing text mining techniques rely on more structured, formal corpuses containing research papers, abstracts, reports, etc. Online chatting differs in a number of ways from everyday face-to-face conversation, both qualitatively and quantitatively. Text chat is a dynamic text, the article inputted by user each time is not as complete as would not necessarily follow strict grammatical structure, is a context-sensitive dynamic text. The user may express a specific theme by few words. The topics may be totally irrelevant, and chat text is formed by a short text. Approaches toward understanding the dynamics of chat conversation are limited, and as usage grows the need for automated analysis increase. From this point of view, it becomes necessary to analyse these conversations and to understand the characteristic of the speakers.

1) Summarizing conversation topic: In text mining applications, determination of conversation topic is one of the important study areas. Most of the studies are based on classification of news texts, and others' researches related to the determination of text writer's characteristics.

2) Understanding the user behavior: Radford examine the communication and information seeking preferences of the Internet users. They also compare traditional libraries and the Internet as the means for an information repository and emphasize the fact that the internet is starting to become an alternative for text based communication.

3) Investigation of chat user attributes: Gender variations was examined in Web logs using logistic regression techniques. However, the authors can not find any conclusive results binding the users genders and web writings. In their work, herring and danet examine several aspects of the language use in the Internet. They assert that gender is reflected in online discourse in every language they studied.

4) Understanding social and semantic interactions: The social and semantic relationships extracted from chat conversations can lead to a better understanding of human relations and interaction. Social clusters and conversational content are understood automatically.

5) Others research: Monitoring chat room conversations, extracting interesting information, authorship attribution and so on.

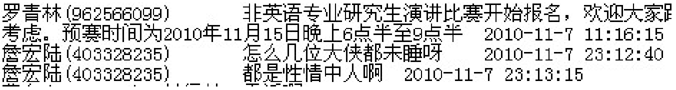
3 Methold of Chat Mining

3.1 Goal of Mining

We can obtain lots of information about the group and each people in the group from the conversation. Members of the group also have different responsiveness to conversation content. By using the text mining, group membership and individual characteristics can be analyzed by speech frequency and word frequency; the hot issues can be conducted by high frequency words and occurrence words; and understand emotions and personal feelings using the text orientation recognition.

3.2 Data Gathering

The conversation data were gathered from chat mediums by using QQ messenger log files. These data were subjected to pretreatment, and were prepared for data mining. In the pre-treatment of the data, basic steps of the data mining were taken into consideration. The formatted conversation which includes name, id, content, time is as follow:



```
罗青林(962566099)      非英语专业研究生演讲比赛开始报名，欢迎大家踊
考虑。预赛时间为2010年11月15日晚上6点半至9点半  2010-11-7 11:16:15
詹宏陆(403328235)      怎么几位大侠都未睡呀    2010-11-7 23:12:40
詹宏陆(403328235)      都是性情中人啊  2010-11-7 23:13:15
```

Fig. 1. Formatted conversation text

3.3 Mining Process

1) Exclude formatted data: the nickname, QQ number, speeches are formatted data, the conversation content was extracted, and the feature of text processing is required before this part of the data extracted. Text chat is a form of informal text, so there will be some sign language, network language. The definition of some commonly used languages in Table 1 and Table 2, and some other face figures as fig 3 defined in the QQ file can be read directly.

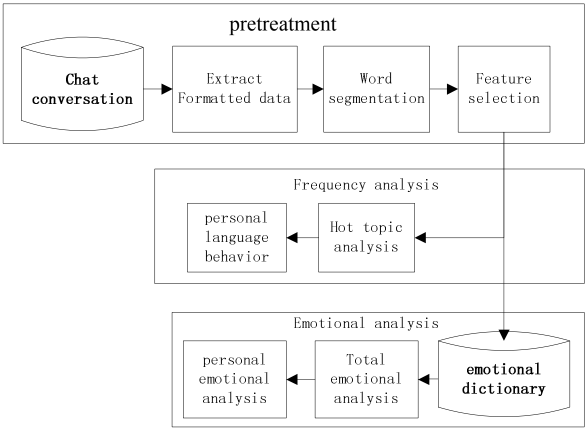


Fig. 2. Mining Process Flow

Fig. 3. QQ face

Table 1. Definition of network slang

slang	meaning
886, 88	goodbye
Lp	wife
Lg	husband
MMD, NND, TMD	curse
…	…

Table 2. Definition of signs

Signs	meaning
☺,:),:)),:D, :d	langhing
:-O	surprised
@>>--->---	rose
?_?	what
…	…

In order to speed up the efficiency of word segmentation, stop words can be filtered firstly. Segmentation algorithm is Reverse Maximum Method (RMM) which is a simple and effective method for Chinese word segmentation based on dictionary. Then filter characteristic words using defined nonsense words. Then conversation content is collected by speaker, and the storage data structure is as fig 3, and name is nickname, id is QQ account number, conFeature is collection of feature word list.

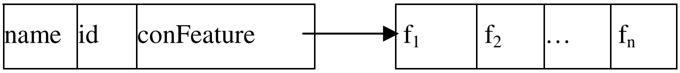


Fig. 3. Data structure

Algorithm described as follows:

```
WHILE(!endof(textFile))
    {
    read(textFile) /*read one conversation record*/
    extract(Name,Id,Time) /*extract the format data*/
    extract(Content) /*extract the unformatted data*/
    ConText=filter(Stoplist) /*filter by stoplist words*/
    segmentation(ConText)
    f=filter(InsignificanceList)
    addto(f,ConFeature)
    }
```

2) Analysis using word frequency: TF is word frequency, F is feature set, the hot words can be found by TF rank. When the TF beyond a threshold, it can be used as a hot word, is also means high frequency words. The hot words list can be defined as $T(t_1,t_2\ldots t_n)$. N(t,f) define as the occurrences of the hot words and other words, of which $t \in T$, $f \in F$. then select the high occurrences value to produce hot topic sentence.

3) Emotional recognition: Emotional orientation of words includes words polarity, intensity and context model analysis. The polarity and strength of words can be defined in dictionary.

If analyze emotional tendency to all chatters by all words, social and psychological tendencies of the group can be understood. And if analyze emotional tendency to each chatter by his own words, the individual's social and psychological tendencies can be got. By comparing all individual's incline to the overall degree, the differences DIFF can be calculated as follows.

$$\text{DIFF}=\text{EMP}-\text{EMT} \tag{1}$$

$$EM=\sum_{k=1}^{t} X_k \tag{2}$$

EMP means individual's emotional orientation, and EMT means the overall emotional orientation which is calculated as the sum of all the characteristics of emotional words.

If DIFF is positive, then the individual holds a positive attitude on the issue than the whole, greater value, and the greater the individual differences.

If DIFF is negative, then the individual on the issue of negative attitudes held higher than the overall, greater value, the greater the individual differences.

If DIFF is equal to or close to 0, then the issue held by individuals is consistent with the general.

4 Results and Discussion

Total size of the conversations is 2.37MB. The conversation which included the chat content in 2 months has 41 talkers. 1839 conversation records were gathered from these mediums.

4.1 Basic Analysis

Speech frequency: From the speech frequency of the 49 talkers, only several people performs actively, most of people have little interest, the result illustrate that the cohesion of the group is lower.

Characteristic words: Through segmentation, the word frequency is greater than 3 times a 501. After the definition of the word meaningless word filters are 408 features, combined with HOWNET Chinese dictionaries, 182 nouns separated by a synonym for the word combination are features of 146 seeds as a hot topic the search words.

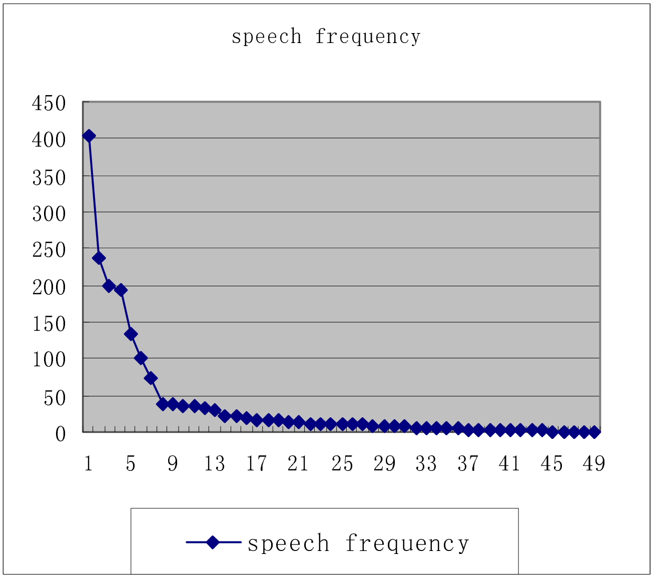


Fig. 5. Speech frequency distribution

4.2 Hot Topic

Select as term frequency greater than 10 hot words, a total of 41, through co-occurrence analysis, we found that the 41 nouns with a total of 106 key words are distributed scatter as follows:

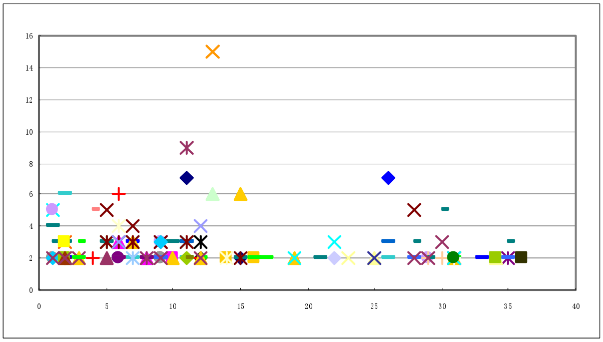


Fig. 6. Words Co-occurrence Distribution

If connecting the hot words with those words that have high Co-occurrence, the topic can be deduced. Such as teacher call the roll, paper issued, economics research, game theory research, Chinese politics status. For the group is made up by doctoral student, so the topics are in line with the doctoral student population characteristics.

4.3 Overall Emotion Analysis

The different emotions results are as Table 3. From the overall distribution data, the overall group community emotion lay particular stress on positive,(35.2% vs 13.21%). Neutral attitude very clear(51.59%).

Table 3. Tendency distribution

Emotion	records	distribution
positive	653	35.20%
neutral	957	51.59%
negative	245	13.21%

The different polarity level of positive emotions segment results are as Table 4.

Table 4. Positive emotion polarity distribution

polarity	records	distribution
Common(0~10)	500	26.95%
Middle(10~20)	116	6.25%
High(>20)	37	1.99%

The different polarity level of negative emotions segment results are as Table 5.

Table 5. Negative emotion polarity distribution

polarity	records	distribution
Common(-10~0)	188	10.13%
Middle(-20~-10)	48	2.59%
High(<-20)	3	0.16%

In further emotion classification, most of the word's polarity of positive and negative is common; and the high level words occupied very small. It hints that the sociality of this group is normal, during this period.

4.4 Difference between Individual and Overall Emotion

There is existed great difference between individual and overall in positive and neutral attitude from the individual emotion minus overall, as fig 7 shows the neutral and positive two lines fluctuate badly. It is means that the attitude towards some topics has little consistency. But to the negative attitude, as fig 7 shows the flat negative lines. It implicated that more people express negative towards topics has consistency.

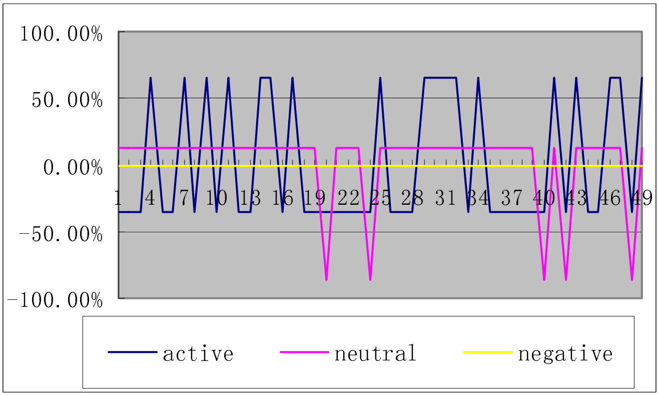


Fig. 7. Speech frequency distribution

5 Conclusion

Development of information technology makes network communication become one of the mainstream communications. In China, many point to point chat tools provide group chat, such as Tencent's QQ which is an important tool in office work and entertainment. Especially for students group, their thoughts and desires often express when chatting in a relatively short period of time, is a certain product of literary form of a campus cultural phenomenon. It is beneficial to formulate education management from the analysis of the text. In this paper, word frequency analysis, document analysis, and lexical emotional recognition are used in the mining of chat content, in order to understand personal language behavior, personal motivation, and members of the emotions involved in the overall performance trends and the personal emotional expression trend. Experimental results show that the method can quickly and effectively grasp the characteristics of subjects for students, and to develop strategies provide the basis for education.

References

1. Tian, F.: Hot QQ: a new social intercourse. China Youth Study (3), 16–19 (2003)
2. Li, X., Ma, L.: Psychological Perspective from QQ chat behavior of college students. Higher Science Education (6), 72–75 (2009)
3. Meehan, A., Manes, G., Davis, L., Hale, J.: Packet sniffing for automated chat room monitoring and evidence preservation. In: Proceeding of the 2001 IEEE, Workshop on Information Assurance and Security (June 2001)
4. Li, S.Y., He, W.: A study on the method of feature selection in chat text. Computer Science (5), 202–204 (2007)
5. Bengel, J., Gauch, S., Mittur, E., Vijayaraghavan, R.: ChatTrack: Chat Room Topic Detection Using Classification. In: Chen, H., Moore, R., Zeng, D.D., Leavitt, J. (eds.) ISI 2004. LNCS, vol. 3073, pp. 266–277. Springer, Heidelberg (2004)

6. Argamon, S., Sacic, M., Stein, S.: Style mining of electronic message for multiple authorship discrimination:First results. In: Proceddings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 475–480 (2003)
7. Dickey, M.H., Burnett, G., Chudoba, K.M., Kazmer, M.M.: Do you read me? Perspective making and perspective taking in chat communities. Journal of the Association for Information Systems 8(1), 47–70 (2007)
8. Walther, J.B., Bunz, U., Bazarova, N.N.: Rules of virtual groups. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences (2005)
9. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style and classification techniques. Journal of the American Society for Information Science and Technology 57(3), 378–393 (2006)
10. Aas, K., Eikvil, L.: Text categorisation: A survey, tech. rep., Norwegian Computing Center (June 1999)
11. Kerrand, D.S., Murthy, U.S.: Divergent and convergent idea generationin teams: A comparison of computer-mediated and face-to-face communication. Group Decision and Negotiation 13, 381–399 (2004)
12. Liao, X., Cao, D., Fang, B.: Research on Blog Opinion Retrival Based on Probabilistic Inference Model. Journal of Computer Research and Development 46(9), 1530–1536 (2009)