The Design and Implementation of the Information Processing Module of Micro-Blog Public Opinion Analysis System

Dezhi An1,2, Guangli Wu1,2, Jun Lu1,2, Shengcai Zhang1,2, Yan Li1,2
{adz6199@gsli.edu.cn}

1 School of Information Engineering, Gansu Institute of Political Science and Law Lanzhou 730070, China
2 Key Laboratory of Evidence Science of Gansu Province Gansu Institute of Political Science and Law, Lanzhou 730070, China

Abstract. The work of this paper is to design and develop the information processing module for a micro-blog public sentiment information analysis system. First of all, this paper introduces the overall framework design of micro-blog public opinion analysis system. It describes the information collection, indexing and segmentation module and related open source software and technology related to the system. The system is analyzed by means of micro-blog keyword to carry out, the use of its potential semantic analysis (LSA) also made a corresponding introduction. The last part mainly introduces the information processing module, gives the overall design of the structure, and real-time statistics, custom statistics, the same word merge, the analysis of the impact of microblogs and other function of the detailed design and coding.

Keywords: Micro-blog; public opinion analysis; latent semantic analysis; information processing; middle layer design

1 Introduction

Micro-blog is following the forum and blog, a new online exchange platform. Compared with the forum and blog, micro Bo the biggest characteristic is the length of each micro Bo has a limit, on twitter, the limits of 140 characters of English, and in Sina Weibo, length limit is 140 characters. This short text can be allowing publishers to quickly complete a micro Bo, also may let the reader in 5 seconds is read to the end of a micro Bo, so that we can in a very short period of time forward again. Micro-blog's main way of communication is by focusing on specific users to get their updates in real time, this concern is a one-way, no need to be concerned about people's confirmation. The unique nature of these micro-blog, micro-blog can quickly and efficiently determine the information spread throughout the micro-blog network. However, micro Bo door registration abuse low, without forcing the user information confirmation and microblogging audit mechanism, in micro Bo made Yao, framing, row slander and fraud have occurred, to the masses and society caused great losses. How to extract valuable content from the micro-blog information, and the negative, negative, false information in a timely manner and control, has become a hot topic of public opinion analysis

of the outstanding issues. Good solutions can not only supervise network behavior, purify network environment [1], but also in obtaining more useful information -- such as the mobile phone data for behavioral and information communication, human social science research with the burgeoning development of micro Bo, a good solution can not only purify the network environment, to enhance the user experience, the healthy development of the network public opinion, the construction of a harmonious society, the modernization of our country, have a important significance.

In order to help government agencies timely grasp the microblogging public opinion hotspot, accelerate for social emergency response speed, reduce the unnecessary labor and manpower cost, the use of information technology tools to effectively solve hot microblogging information discovery, tracking, analysis of information sources and sensitive topic the warning, development and construction of a custom-built by the government of microblog analysis system, meet the future years the government to monitor the microblog information needs, and deployment of the system are integrated into the "great political and ideological work platform, optimize the allocation of resources, rich platform function of the" great political and ideological work ".

2 Overview of Related Technologies and Theories

Vector space model is proposed by Salton et al in the late sixties of the 20th century, is a simple, effective and efficient text representation model and its theoretical basis is the algebra. Vector space model the user's query requirements and database document information expressed by search terms in a vector space, by calculating the distance between the vectors to determine the degree of similarity between a document and a query. Then, according to the similarity of the query results. The key of the vector space model is the selection of the feature vector and the calculation of the weight of the feature vector [2]. The two part is the key point of the vector space model. According to the "Bayesian hypothesis", it is assumed that the words or phrases that constitute the document are independent of each other in determining the function of the document class, and then use the set of words or phrases that appear in the document to replace the document. In other words, the category that belongs to an article has nothing to do with the order or position of the document in the document, but with the frequency of certain words or phrases that appear in the document. Obviously this will lose a lot of information about the content of the article, but this assumption can make the text representation and processing, and can achieve better results in the text classification. As mentioned above, if will constitute the text of all kinds of semantic units (such as characters, words and phrases) collectively become word for word the words appear in the text of the frequency, then a document contains various words phrase information to the correct classification.

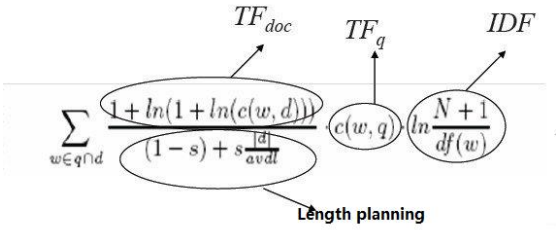

Fig. 1. Sketch map of VSM vector space model

2.1 Naive Bayesian Classification

Bayesian classification is a statistical classification methods, it is based on Bayes theorem can used to predict class membership probability, given text belongs to a particular class of probability and classification according to the prediction results of the sample to the class with the highest probability to be. Naive Bayesian classification model training process is actually a statistical every feature in all kinds of process, theoretically, speak of Bayesian classification error rate minimum. The test results, naive Bayesian algorithm in large data sets demonstrated rare speed and accuracy.

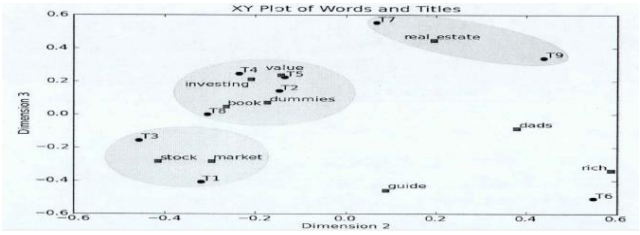

Fig. 2. Map of 2D coordinate system of words and documents

2.2 Lily Cloud Database Introduction

Lily is on the market first for social content applications and cloud data warehouse, it can also store content and attention data also has expansibility [3]. The software is based on the storage, indexing and search functions. The software allows developers can easily build a variable rich media content data warehouse and its application in a variety of large-scale data centralized type of Internet application, provide the function of media and resource management, archiving, document storage, and so on.

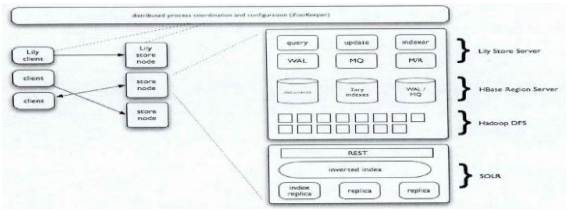

Fig. 3. Schematic diagram of Lily cloud database

3 System Function Module Design

In this paper, the micro-blog public opinion analysis system is modular design, each module through the document definition of interaction, with high cohesion, low characteristics. The system is divided into four modules, which are: information acquisition module, information processing module, information service module and front-end display module. And has an independent database for services. The main content of this paper is to design and implement the information processing module.

3.1 Information Processing Module

Information processing module in the whole system to play a connecting role, which real-time statistics by public opinion kernel analysis to obtain a statistical summary of the recent and for one hour a day, contain the information of geographical position and hot word statistics. Custom statistical default read lily to preserve historical hourly records for fast statistics, in addition to providing more comprehensive fine-grained statistics, the statistics will be to read the blog of the original data to statistical, high overhead for upper layer information service module generates a detailed report of the public opinion or for analysis as appropriate, the use of personnel [4].

3.2 System Interface Design

This section probably explains the connection between the information processing module and the other modules, which hides the details and functions of the information processing module. The location analysis and microblogging scoring system mainly for kernel services, in micro Bo public opinion analysis of kernel extraction data of sina microblogging, to every tweet for segmentation, geographic location analysis, and microblogging score, and statistics the frequency and location information, and maintain more new memory in the tree structure. Information service module calls for real-time statistics module and custom statistics module. The two modules through the micro Bo public opinion analysis system kernel corresponding analysis data acquisition, processing and after processing of the data to the information service module.

Because we build the micro-blog public opinion analysis system needs a higher real-time, every 60 seconds to extract data from micro-blog Sina and analysis, and the processed information is passed to the information service module for use. In addition, information processing and statistics module requirements implement certain filtering, geographical position filtration and keyword filtering, if using the traditional analysis method, Sina Weibo per hour million microblog analysis will take a lot of time, time can not be effectively guaranteed. This requires us to develop a new algorithm to effectively solve this contradiction. Methods to solve this contradiction is try not to read the data stored in the lily, and will require the use of data retained in memory, and every 60 seconds only need carries on the statistics to the new microblogging, without again before statistical micro blog data.

3.3 System Database Design

System parameters read and update: all the operating parameters of the system are stored in the Oracle database, using the JDBC connection database in this module to read and update the operation directly. Including: sensitive vocabulary to add, delete, modify, query function; filter rule table of new, delete, modify, medicine, the inquiry function; module setting the table to add, delete, modify, medicine, the inquiry function; the system log table query function. Micro-blog related information acquisition: the realization of this part is mainly divided into two, the use of micro-blog Sina API to obtain the latest information to get the latest micro-blog. In addition, the original data in the information acquisition module is realized, and the upper layer information service module can choose different information acquisition strategies according to their own needs. Including: access to the specified users and focus on users of the latest micro-blog released micro-blog; access to the specified user; acquiring the specified a number of users of micro-blog micro-blog acquisition of micro-blog original list; forwarding; access to the specified users of micro-blog; access to the specified user focus on the user's micro-blog two-way access to the specified number of comments forwarded; micro-blog; access to the specified the user's attention to the orientation of micro-blog; access to the specified user attention list; access to the specified users of common concern list; access to the specified user list details for fans; the relationship between two users; the current user access to the specified topic list [5]; micro-blog; get a list of users praise designated micro-blog.

4 Conclusion

In this paper, micro-blog public opinion analysis system of the other modules and system architecture as a deep and comprehensive analysis, introduced the theory and technology related to public opinion analysis system. Then, the requirement analysis is carried out for the bottom layer and the upper layer of the information processing module, and the design and implementation are completed on the basis of the information processing module. In the framework of the system design process, we are based on the specific circumstances, the different functions of the use and language, in order to ensure the efficiency of the system and coding efficiency. Some of the kernel related functions are divided into the information processing module to complete, reduce the core of the coupling, and the preparation of a detailed document for the future expansion of the function and the two development provides a convenient.

References

[1] Hong Ji Ouyang, Ge Meng, Zhao Qiang. Based on JDBC and design pattern database connection pool implementation method. Computer technology and development. No. 01. (2011)

[2] Li Honglei, Pakistan minje. The Internet public opinion involving measures of. China prosecutor. No. 17. (2010)

[3] Zhu Haodong, Zhong Yong. The combination of optimized document frequency and LSA feature selection method. Computer engineering and applications. No. 34. (2009)

[4] Cheng Chunrui, Liu Wanjun. High cohesion and low coupling software architecture for the construction of. Computer system applications. No. 07. (2009)

[5] Yao Yao, Liu Gongshen, Li Xiang. A text clustering algorithm based on vector space mode. Computer engineering. No. 18. (2008)