Research and Realization of College Students' Integrated Information Customization System Based on Hadoop

Liu Lijuan
Nanjing Normal University Taizhou College

Summary
The advent of large data age has provided the convenience of college students' comprehensive information acquisition. This paper studies and realizes college student comprehensive information customization system based on Hadoop. The system uses Htmlparser to complete the collection of information such as campus network、 campus office network and campus forum, Calls Bases classification algorithm with Mahout to realize the distribution of data mining algorithm, Stores the collected data with the distributed database Hbase and develops the client application based on the Android system, connects the mobile terminal with the distributed information processing platform, completes the successful push of the information. The system can provide college students with comprehensive information mobile phone customization services, facilitate students to access custom information at any time and improve the sharing of information services and targeted.

Key words:
Hadoop; comprehensive information; customized services; cloud services; data mining

1. Introduction

In recent years, the development of computer technology has promoted the construction of information technology in colleges and universities in depth, in order to achieve informational management, many colleges and universities have developed systems of different functions. For example: students visit the campus network can access the school news; visit the educational network can access the results elective information; visit the school network can obtain employment reward information; visit the campus forum can post communicate etc., the information of these systems is messy scattered and not easy to collect. When people access the various sites to obtain the required information through browser, there are data flow slow response poor targeting and other shortcomings.

Distributed storage system based on the cloud computing is the trend to achieve cheap storage [4], people integrate hardware resources through the cloud computing to provide a strong storage capacity and computing power, also achieve a unified management of data and improve the quality of service.

This paper focuses on putting the campus network information released by students employment information forums and other scattered data together, using mass storage capacity of Hadoop platform to solve problems of large data storage and query, on this basis , improving the efficiency of the query through algorithm based on data analysis、mining、classification, using mobile intelligent terminal for college students to provide comprehensive information services and improving the level of information services.

2. Hadoop platform [1]

Cloud computing is an Internet-based computing approach that uses new service delivery and schemas to provide users with the required hardware software platforms and network resources.

Hadoop is the cloud computing platform used widely, this paper uses Hadoop technology, Hadoop distributed file system HDFS and MapReduce [2] [5] as the core, to provide users with a variety of transparent services. Hadoop can help users easily integrate computer resources, make full use of computing and storage capabilities of the cluster, build a distributed computing platform and ultimately complete the massive data processing.

Hadoop cloud computing platform includes distributed parallel computing framework MapReduce distributed document system HDFS and distributed data Hbase [7], they complement each other, have their own strengths. The system can rent Ali cloud server to complete setup and configuration of Hadoop platform.

3. data collection

The system uses the Hadoop cloud platform to realize the background data storage processing and classification. The whole system can be divided into three layers: user presentation layer business logic layer [8] and data access layer.

The user presentation layer is responsible for satisfying the user's comprehensive customization function. The business logic layer is responsible for the distributed storage processing classification of data， in order to use for users.

Data access layer is mainly responsible for data collection, using htmlparser to crawl information of campus network、office network forum and other network. Html analytic

Manuscript received July 5, 2017
Manuscript revised July 20, 2017

library HtmlParser is writted by the Java language, can be used to transform or extract html, with the characteristics of efficient reliable, the process of information collection is shown in Fig 1.

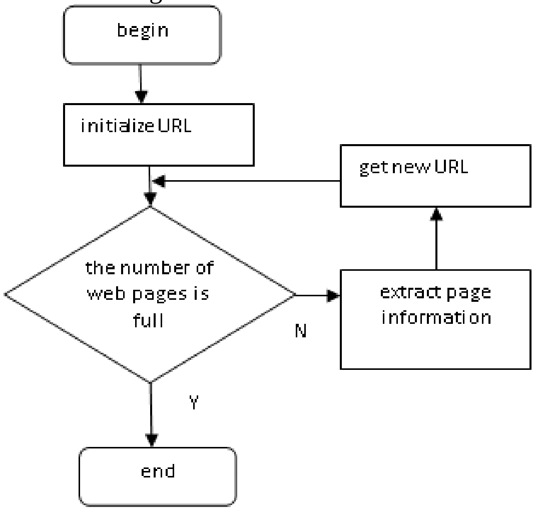

Fig 1 Htmlparser collects web page diagram

4. data analysis and processing

4.1 Split words、delete stop words

The collected site data need to be pre-processed and classified before they are stored into the database, pre-processing step contains several major links: split words delete stop words extract the characteristic words and classify data.

The system uses FudanNLP to process data released by the user. FudanNLP is a development kit of Chinese natural language text processing based on machine learning developed by Fudan University in Shanghai. It is realized by java and is very convenient to call up.It can be used to achieve functions such as Chinese data segmentation syntax analysis and part of speech tagging etc., mainly relize word segmentation about the data which users published by calling edu.fudan.nlp.cn.tag.CWSTagger class, and filter stop words according to disable vocabularies.

4.2 extract the characteristic words

After Chinese words segmented, it’s need to extract the subject feature words, the system uses the TF-IDF [6] algorithm. The main idea of TF-IDF is that in an article, if a word or phrase appears to have a high frequency but is rarely present in other articles, the word or phrase has a good class distinction. TF-IDF algorithm mainly calculates TF_{ij}、 IDF_i、W_{ij} three parameters, as shown below:

$$TF_{ij}= n_{ij}/\Sigma kn_{k,j} \qquad (1)$$

$$IDF_i=\log(N/n_j) \qquad (2)$$

$$W_{ij}=TF_{ij}\times IDF_i \qquad (3)$$

In the expression (1), TF_{ij} refers to the frequency of the word k_i in a particular document d_j, n_{ij} refers to the number of times the word k_i appears in the document d_j, $\Sigma kn_{k,j}$ refers to the sum of the number of occurrences of all words in the document d_j. The IDF_i of the word k_i is calculated using the expression (2), where N is the total number of documents and n_j is the total number of documents containing the keyword k_i. The expression (3) can calculate the weight W_{ij} to measure the importance of the word k_i in the d_j document.

4.3 Data classification

When the feature word is obtained, the data is classified according to the characteristic word. The system is built in the cloud environment, Mahout is installed on the cloud platform and is an open source project of ASF, which uses Mahout training Bayes classifier to achieve distributed Naive Bayesian classification algorithm. Naive Bayesian algorithm is a classification algorithm widely used in data mining, using Mahout call Bayes algorithm for the user's information posted to split words、 extract keywords and classify, and finally the data will be processed to the user classification. The data classification process is shown in Fig 2.

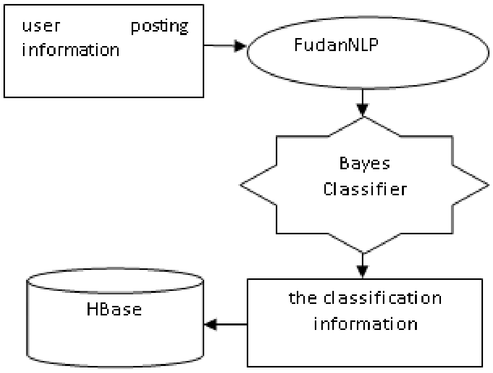

Fig 2 data classification process

5 .data storage

HBase is a geometric、distributed open source database. HBase uses Hadoop HDFS [3] as its file storage system, with the increase in the number of users using the system, the amount of data will be more and more, so choose HBase to store data collected.

Hbase uses tables to store data, each table consists of rows and columns. In the Hbase table, the element is uniquely determined by the row key column timestamp. So it’s need to hava a reasonable design of the row key when design a table, in order to quickly find data accessed. The data classified of the system are stored into the pre-designed Hbase table to storage.

6. Mobile terminal design

For college students, the use of mobile phones is very popular, they visit the network in addition to the form of the PC, the use of mobile phones is more frequently. therefore, developing a comprehensive information customization system for college students to facilitate access information, using the client based on Android system as a platform to access, it’s convenience for students to access network data at any time.

Android is based on the Linux kernel, mainly made up of three parts: the linux kernel layer,class library、the core component library layer and virtual machine、application framework layer. Android applications are developed with JAVA language and build web server with struts2 framework, in order to achieve mobile terminals、server-side communications. Web server architecture is shown in Fig 3.

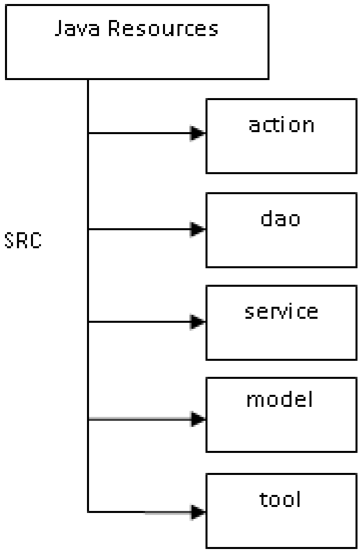

Fig 3 Web server architecture

7. System test

The system finally collected data from the campus network、school network、educational network、forum and other website, rented cloud platform, developed mobile terminals based on Android and achieved its connection with the cloud platform for college students to customize the comprehensive information. The final framework of the system is shown in Fig 4 below.

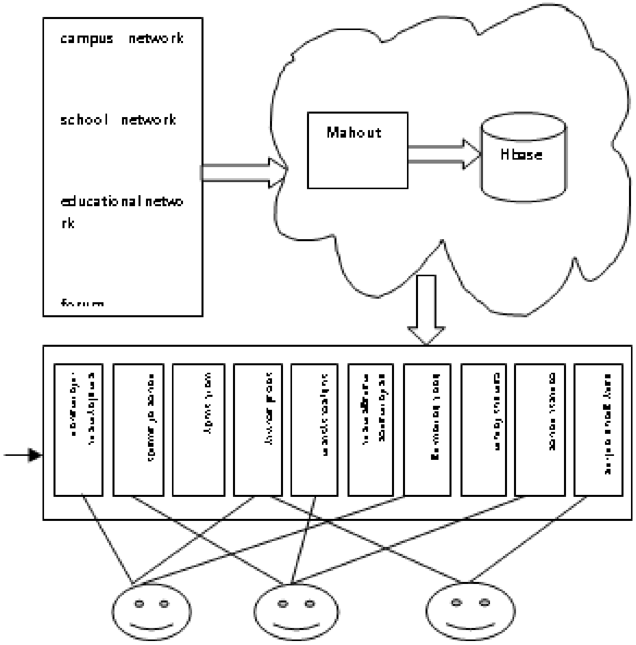

Fig4 The overall framework of the integrated information customization system

Students login system and set up custom information items that can freely choose different customized items depending on the content of each semester. Such as customing employment information, you can make network job search; customing award notification, you can participate in online awards of various awards, get notification、submit materials、query results; customing work-study, you can get work-study information, participate in the registration、receive the relevant information; customing community activities, you can join the various communities of the school、get the information about the activity; customing elective system, you can conduct selection process of the elective courses; customing performance management, you can find the results of each semester, the system will carry out academic early warning to poor students; customing book borrowing, you can book inquiries、 borrow and return etc.; customing campus forum, you can post、reply and exchange; customimg competition notice, you can be informed of a variety of competition information the college published; customing party group online, you can learn the party knowledge 、 watch online video 、 participate in party activities.

Through this system, students can easily customize different information according to their own needs, avoid logging on a variety of sites query one by one, missing information or causing the login process cumbersome, it’s more targeted、more in line with the needs of modern information technology development.

8. Conclusion

Large data era has come, a variety of information dissemination platforms are complex, it’s not conducive to the collection of information. This paper focuses on improving the comprehensive information management level of college students, studying and implementing the comprehensive information customization system of college students based on Hadoop. The system collected scattered data from campus network、school network、 educational network、 forum etc., used Mahout to train Bayes classifier for data mining, stored the classification results in the Hbase database, builded Hadoop cloud platform through rentaling cloud services, developed a mobile client using Android. After the test ,it shows that the system can meet the design requirements , provide personalized integrated information customization services for college students, promote the campus management informative and intelligent.

Acknowledgment

The author would like to express her cordial thanks to Professor Zhou Yan-huai for his valuable advice.

References

[1] Fu Wei. Design and implementation of Web log analysis platform based on Hadoop [D]. Beijing: Beijing University of Posts and Telecommunications, 2015.

[2] GAO Xian-wei . Study and design of context-based recommendation system based on Hadoop [D]. Shanxi: North University of China, 2015.

[3] Sun Ting. Research and implementation of microblogging recommendation system based on cloud computing [D]. Shandong: Shandong Normal University, 2015.

[4] Hu Rui, Chen Li-chun. Research and implementation of the micro-curriculum system based on Hadoop [J]. Software Development, 2016: 56-57.

[5] LIN Zhong-ming , LI Wen-jing . Study on Web user identification and news intelligence recommendation algorithm based on Hadoop [J]. Software Journal, 2016,15 (5): 27-29.

[6] Yang Hao , Zeng Xing-bin , He Jia-ming. Design and implementation of microblogging hot topic mining system based on Hadoop [J]. Technical program, 2016,2: 10-12.

[7] Jiang Yun-xia, Fu Qi. Research of the cloud teaching resources platform based on Hadoop [J]. Contemporary education theory and practice, 2016,8 (4): 111-113.

[8] Zhang Yong-hua. Design and implementation of the cloud teaching resource platform based on Hadoop [J]. Enterprise Technology Development, 2015,34 (16): 24-27.

Liu Lijuan received a master's degree from Nanjing University of Aeronautics and Astronautics, where she was an assistant in Nanjing Normal University Taizhou College (2008) and lecturer (2010). Her research interests include cloud computing、 big data theory and practical application.