Construction of Network Behavior Analysis System of Mobile User Based on Big Data Technology

Haiyan Luo[1,a], Tao Yang[1,b], Qiong Wu[1,c]

College of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang, 110866, China

[a]423233233@qq.com, [b]yangtaosx@yahoo.com.cn, [c]29038649@qq.com

Keywords: big data; mobile user; network behavior; behavior analysis

Abstract. The traditional user behavior analysis method lacks the support of big data technology and the user demands are not subdivided, which causes the mismatch of marketing and user demands. In view of the different behavior standards of mobile users, the personalized services have been accurately customized. At the same time, the network behavior analysis method of mobile users based on Hadoop, Hive, Zookeeper and other big data technologies has been proposed. Based on this method, the behavior analysis system of mobile users including the download program module of call bill log, the processing program module of call bill log, the classification program module of call bill log, the scheduler program module of task statistics and the storage program module of result statistics has been constructed. The results show that the system based on big data technology is of great significance to accurately position the user access demands, help the mobile operators to conduct the thematic analysis of big data and enhance the competitiveness of enterprises compared with the traditional methods.

Introduction

In recent years, with the decline of the smart phone price and the dropping of Internet advisory fees as well as the improvement of network coverage, the access threshold of mobile phone users have been lowered and the number of mobile phone users has been significantly increased. The network and the mobile phones that people often use in daily life will constantly produce large amounts of new data [1]. The rapid and explosive growth of data has brought great pressure to store, process and analyze the data, which is beyond the processing ability of traditional system. For the massive user behavior data, the stand-alone software or data warehouse can not effectively conduct the analysis of user behavior data [2]. However, the introduction of big data technology can not only meet the demands of the functions and the performance of system and reduce the cost of IT deployment, but also expand the application fields of data intelligent analysis. The big data technology has become the powerful tool of enterprises to enhance the competitiveness in the fast-changing and data explosion era.

User behavior analysis [3] refers to the process that the data in relevant sites is analyzed to find the principles of user to access the sites under the condition that page views of users are obtained. The traditional user behavior analysis lacks the behavior analysis means of big data support. The demands of users are not subdivided and the positioning is not accurate, which finally causes the mismatch of corporate marketing and user demands [4]. As the squeezing influence of OTT enterprises on operators is more and more evident, some services such as short messages, multimedia messages and videos are lost. The operators urgently need to build the blue ocean market of big data operation. Based on this, the behavior analysis method of mobile Internet users based on big data technology has been proposed. Finding an effective user behavior analysis program has become an important issue for mobile operators in the business development of mobile Internet.

Application of Big Data and Key Technologies on User Behavior Analysis

Hadoop is a distributed infrastructure organized by Apache open source [5-7] and it is an open-source computing platform. Hadoop can run the applications in clusters consisted by some cheap hardware. Through the stable and reliable interface provided by applications, the distributed system with high reliability and good scalability can be constructed, which determines that Hadoop is very good at processing the large-scale data [8]. In the behavior analysis system designed in this paper, Hapoop is mainly used to achieve the storage, the interaction and the statistics of data.

Hive [9] is a data warehouse based on Hapoop and the database file can be mapped to the database table. Meanwhile, it also has the simple SQL-like language query function, namely HQL. The corresponding statements can be parsed and converted to generate a series of MapReduce data tasks. Hive can be used to extract, convert and load large amount of data (ETL). In the behavior analysis system, MapReduce in Hadoop can be used through Hive so as to achieve the corresponding statistics of massive data.

ZooKeeper [10] is a subproject of Apache Hadoop project, namely the distributed coordination service of Hadoop. It is an open-source and powerful distributed synchronization service system that can be used to construct an ordinary distributed application. The whole ZooKeeper system is consisted by multiple server nodes and the clients can interact with ZooKeeper through ZooKeeper client library. Through ZooKeeper, the complex and error-prone critical services can be encapsulated. Then, the easily-used interface as well as the high-performance and stable-functional system can be provided for users.

DNSmasq is a very handy tool to configure DNS and DHCP. The behavior analysis system designed in this paper mainly uses the functions of domain name cache and custom domain name resolution. The domain name cache is mainly used in the crawling for Internet web contents. Through the domain name resolution and DNSmasq proxy resolution, the number of actual domain name resolution can be effectively reduced. The custom domain name resolution mainly provides the domain name resolution among the nodes of Hadoop.

HAProxy is a proxy software based on TCP and HTTP that can provide high availability and load balancing [11]. At the same time, it can also provide the fast and reliable proxy solution program based on TCP and HTTP. In the behavior analysis system, when the classifier is used for tagging through HTTP in LABEL program, HAProxy is used to receive the requests of label to send to each classifier so as to achieve the load of requests.

Design of User Behavior Analysis System Based on Big Data Technology

The overall design of user behavior analysis system is shown in Fig. 1. According to the overall demands of behavior analysis system, the system is mainly divided into the download program module of call bill log, the processing program module of call bill log, the classification program module of call bill log, the scheduler program module of task statistics and the storage program module of result statistics, etc.

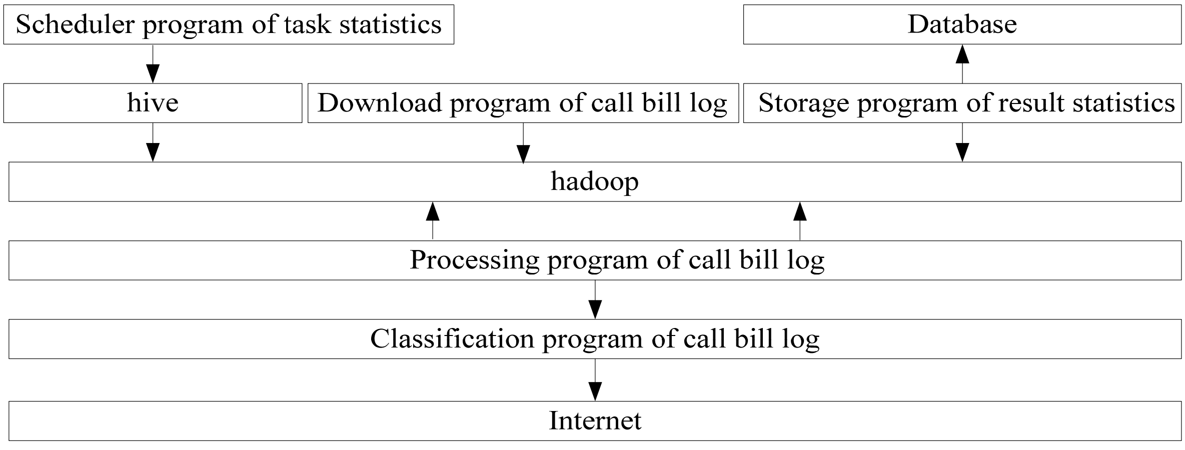
Scheduler program of task statistics
Database
hive
Download program of call bill log
Storage program of result statistics
hadoop
Processing program of call bill log
Classification program of call bill log
Internet

Fig. 1. Overall design of behavior analysis system

The behavior analysis system mainly analyzes the Internet pages. The URL information accessed by users can be obtained by logs and it is required to further analyze the text contents of pages. The download program module of call bill log mainly downloads the original bill from the specified

FTP server. After screening the records, it is necessary to extract the useful fields to generate the files with unified format that will be tagged after reunification processing. Meanwhile, the files should be stored in the specified directory of Hadoop. In addition to download GPRS logs, this module can also download the user traffic statistics and the user basic data set. The flowchart of the download program of call bill log is shown in Fig. 2.

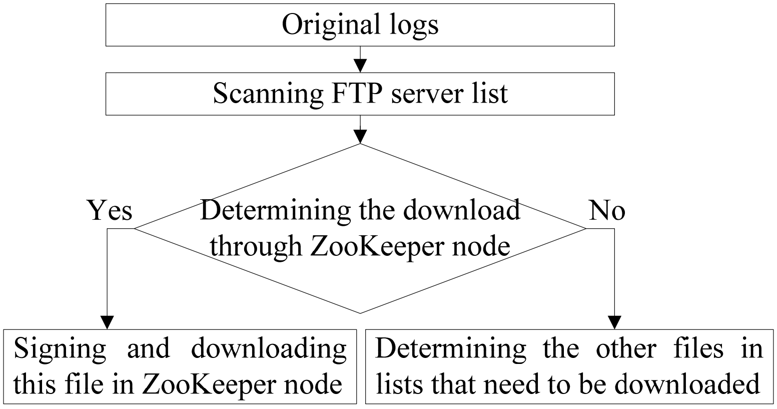
Original logs
Scanning FTP server list
Yes
Determining the download through ZooKeeper node
No
Signing and downloading this file in ZooKeeper node
Determining the other files in lists that need to be downloaded

Fig. 2. Flowchart of the download program of call bill log

In order to achieve the high availability, the parallel running of multiple instances can be realized through ZooKeeper so as to automatically share the file download tasks. The synergy method between the various instances can judge whether the node in ZooKeeper exists. If the node exists, it is indicated that other application instances have processed the file. If the node does not exist, it is required to create the corresponding file in HDFS. It should be continuously conducted if the file is successfully created. Otherwise, it will be proved that other application instances have been created.

The system employs Hadoop to store the daily user Internet information data and obtains the call bill files that will be processed from the distributed file system HDFS. Meanwhile, it can tag for the records every day. In order to achieve the high availability, the parallel running of multiple instances can be realized through the unity of HDFS and the download tasks can be automatically shared. After tagging the collected data, the download program of call bill log can output the information such as the numbers, the websites, the labels, the contents, the flow, the time-consuming and the keyword matching. It is required to adopt Map/Reduce technology to conduct the efficient massive data analysis. The Hadoop processing data flow is shown in Fig. 3.

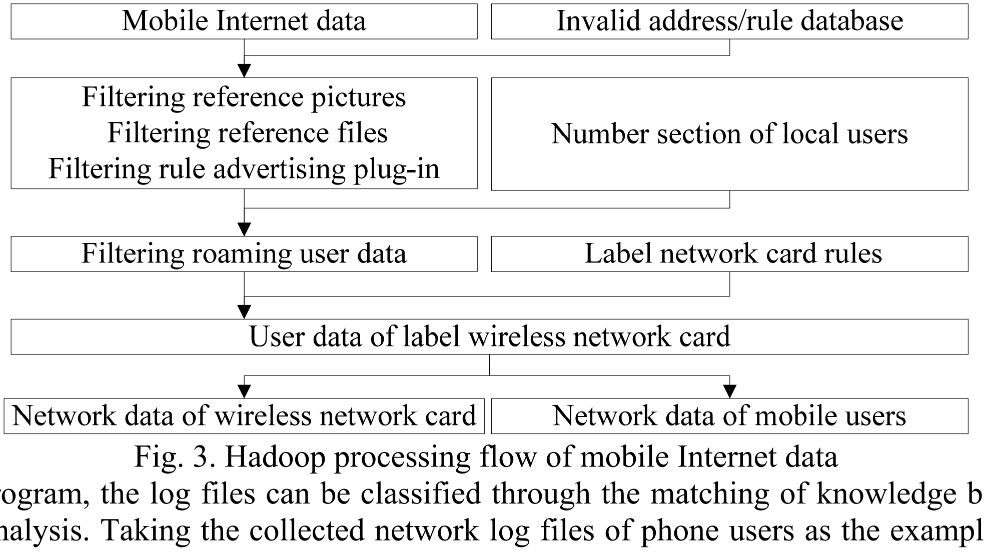
Mobile Internet data
Invalid address/rule database
Filtering reference pictures
Filtering reference files
Filtering rule advertising plug-in
Number section of local users
Filtering roaming user data
Label network card rules
User data of label wireless network card
Network data of wireless network card
Network data of mobile users

Fig. 3. Hadoop processing flow of mobile Internet data

In this program, the log files can be classified through the matching of knowledge base and the download analysis. Taking the collected network log files of phone users as the example, the main classification steps are:

(1) Classifying the popular websites;

(2) Crawling and classifying the website contents;

(3) Determining user’s preference according to the classification of user access;

(4) Analyzing other network behaviors of users, such as the period of surfing the Internet and the used business.

The flow of the classification program of call bill log is shown in Fig. 4.

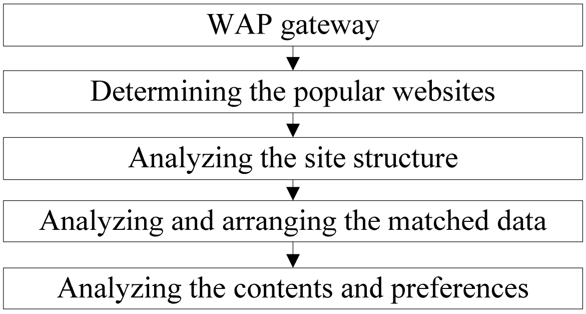
WAP gateway
Determining the popular websites
Analyzing the site structure
Analyzing and arranging the matched data
Analyzing the contents and preferences

Fig. 4. Classification flow of behavior analysis

This program uses the scripts that can regularly execute the configuration to conduct the statistical tasks and the data in previous day is processed every day. From the network log in previous day, the program will regularly scan the tasks and read all statistical tasks. After reading out the tasks, it is required to judge whether the forward task data of each task can be generated and the result partition of the tasks exists. If the conditions are met, the tasks should be submitted to the task queue of HADOOP. The program will control the number of submitted tasks. When the task successfully runs, the corresponding result partition will generate the data. When all the tasks are scanned in the next time, the task can be considered to be finished due to the generation of result partition data. The task statistics scheduler program is shown in Fig. 5.

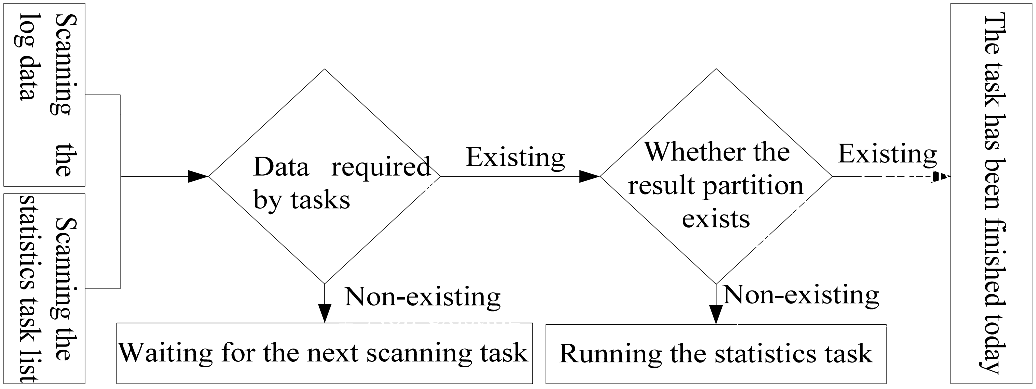
Scanning the log data
Scanning the statistics task list
Data required by tasks
Existing
Whether the result partition exists
Existing
The task has been finished today
Non-existing
Non-existing
Waiting for the next scanning task
Running the statistics task

Fig. 5. Flow of statistics scheduler program

This program can regularly store the tagging procedure in the processing result files on hdfs and it will be imported into hive base. Every day, it is started from processing the GPRS_LOG log generation files in previous day. This program will scan the tasks in hive lists every 10 minutes and all the statistics tasks in system will be read out. It is necessary to judge whether the forward task data of each task can be generated and the result partition of the tasks exists. When the conditions are met, the tasks will be submitted to the task queue of HADOOP. The flow of the storage program of statistics results is shown in Fig. 6.

The storage program of statistics results continuously scan the Hive result table to judge whether the data has been generated in result partition. If there is no data, it is required to wait for the next scanning. If the data is generated, the local data files will be generated. The scripts scan the local data files, when the data files are generated, the files should be renamed at first. Then, they should be added into the corresponding table of database. Otherwise, it is necessary to wait for the next scanning.

Conclusion

The traditional mobile user behavior analysis lacks the support of big data technology and the rapid growth of data has been beyond the processing ability of traditional system. Therefore, the traditional method cannot effectively analyze the user behavior data. The introduction of big data technology can not only meet the requirements of the functions and the performance of system, but also expand the application fields of intelligent data analysis. The big data technology has become the powerful tool for enterprises to improve the competitiveness in data explosion era. In this paper, the applications of big data and other key technologies in network behavior system of mobile users have been analyzed and the network behavior analysis system of mobile users has been designed. Compared with the traditional method, the results show that the system based on big data technology is of great significance to deeply find the user's behavior data, provide personalized services for users and help the mobile operators to conduct the big data analysis as well as improve the service level of mobile operators.

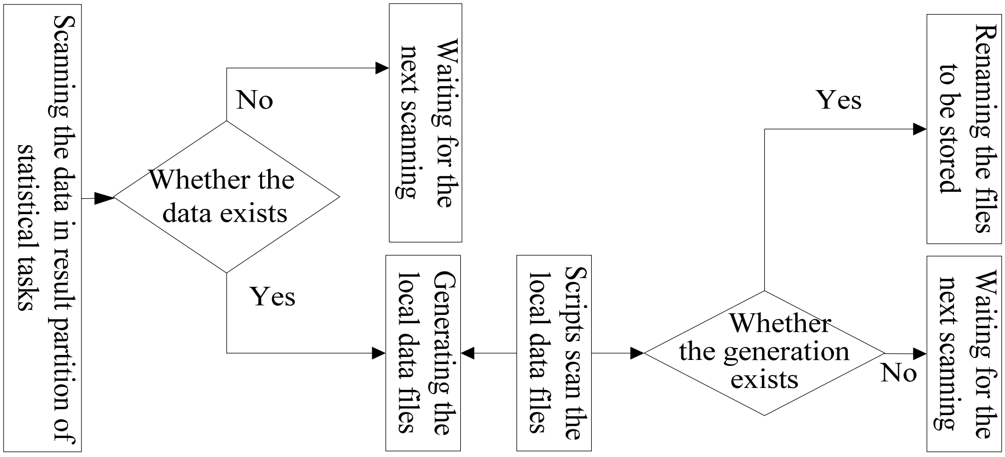
Scanning the data in result partition of statistical tasks
Whether the data exists
No
Yes
Waiting for the next scanning
Generating the local data files
Scripts scan the local data files
Whether the generation exists
Yes
No
Renaming the files to be stored
Waiting for the next scanning

Fig. 6. Flow of storage program of statistics results

References

[1] B. Zhang, "Overview on Big Data Management Technology Research," Computer Applications and Software, vol. 31, no. 11, pp. 1-5, 2014.

[2] L. Zhao, "Research of Massive Searching Logs Analysis System Based on Hive," Application Research of Computers, vol. 30, no. 11, pp. 3343-3345, 2013.

[3] A. Ma, "Design and realization of Accurate Selling System Based on User Behavior Analysis," Nanjing University of Posts and Telecommunications, 2013.

[4] P. Wang, "Key Technology of Web Users' Behaviors Analysis Based on Dynamic Behaviors Profile Database," Computer Technology and Development, vol. 19, no. 02, pp. 20-23, 2009.

[5] S. C. Chen, "User Behavior Modeling Method for Mobile Applications Based on Log Mining," Computer Science, vol. 41, no. 11, pp. 25-30, 2014.

[6] K. Wu, "Analysis of Mobile Internet User Behaviors," Beijing University of Posts and Telecommunications, 2013.

[7] Y. C. Sun, "MapReduce Designed to Optimize Computing Model Based on Hadoop Framework," Computer Science, vol. 41, no. 11A, pp. 333-336, 2014.

[8] Z. Zhu, "Research and Application of Massive Data Processing Model Based on Hadoop," Beijing University of Posts and Telecommunications, 2008.

[9] Y. Z. Liu, "Design and Implementation of Massive Web Log Analysis System Based on Hadoop/Hive," Dalian University of Technology, 2011.

[10] Y. F. Huang, "Design and Implement Distributed Coordination Framework Based on ZooKeeper," Zhejiang University, 2012.

[11] S. T. Zhou, "Research and Implementation of Persistent Connection Mutilplexing of TCP Based on Haproxy," South China University of Technology, 2011.