# Design and Implementation of User Behavior Analysis in Mobile Internet

Jing Fu[1, a], Zhizhong Zhang[1, 2, b] and Yuelong Chen[1, c]

[1]School of communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;

[2]Chongqing Chongyou Huice Communication Technology Company, Chongqing 401121, China.

[a]fj_anne@foxmail.com, [b]zhangzz@cqupt.edu.cn, [c]783804768@qq.com

## Abstract

In view of the low efficiency for existing user behavior analysis, designed and implemented the user behavior analysis system based on big data platform. Firstly, acquired data from the S1-U interface, decoded according to the interface protocol stack in decoding and synthesized the CDR(Call Detail Records) file; based on crawler information database that the web crawler climbed and CDR files, used DPI(Deep Packet Inspection) recognition technology to identify the users' business and flow; finally computed all kinds of users' business and traffic in the Hadoop big data platform, and found users' preference from the statistics. The system is verified by the current network data, and it can accurately identify the different behavior of users, and provide support for the operator's accurate marketing business.

## Keywords

Mobile internet, user behavior, DPI, big data.

## 1. Introduction

In recent years, the mobile Internet is developing rapidly. It plays an important role in people's daily necessities of life [1]. It can be said that the mobile Internet is changing or changed the way of people's life. The mobile Internet has two characteristics: one is the integration of mobile communication and the Internet, users can access the Internet at any time; the second is the large number of applications were designed with mobile Internet, combined the mobile terminal's mobility and portability to these applications, it can provide personalized service for users [2].

In order to provide comprehensive and high-quality personalized service for users, it is necessary for the operators to carry out a comprehensive and systematic study and analysis of the user's business and traffic. Through the user behavior analysis, we can better grasp the user's needs and support the mobile Internet service. Therefore, the user behavior analysis is not only a problem that the operators need to solve, but also is very useful for both the content providers and the users.

Different user behavior analysis system has selected different user behavior object and analysis means. Some systems analyze user logs and other related data, some systems analyze Web website content, and other systems analyze traffic contents, which focus on mining users' potential needs.

At present, there are a lot of researchers at home and abroad analyze user online behavior, and have made some achievements. Wan Fei, Zhao Xi put forward a model to analyze user behavior for Web log data, and evaluate the user behavior analysis model by an example [3]. Neelima G, Rodda S analyzes user behavior from a given log file, including data cleaning, user identification and session

identification for log files [4]. Chen K, Huazheng F U, Chen C proposed a hybrid n-gram feature based URL classification algorithm and DPI data classification method based on Doc2Vec model and text classification algorithm to classify user interests [5]. In this paper, a large data processing method is used to analyze the user behavior and improve the efficiency of the analysis.

## 2. LTE-A Network Architecture

LTE-A(Long Term Evolution-Advanced) is based on LTE (Long Term Evolution), it is at a higher stage of technology evolution and development, which not only meets and exceeds the needs of IMT-Advanced, but also maintains better backward compatibility with LTE network [6]. LTE-A follows the network architecture of the LTE system, its network architecture as shown in Fig. 1 [7]. LTE-A network has a flat architecture, which ensures low latency, low cost and low complexity requirements. It can also bring higher throughput to the system and provide better quality of service for users.

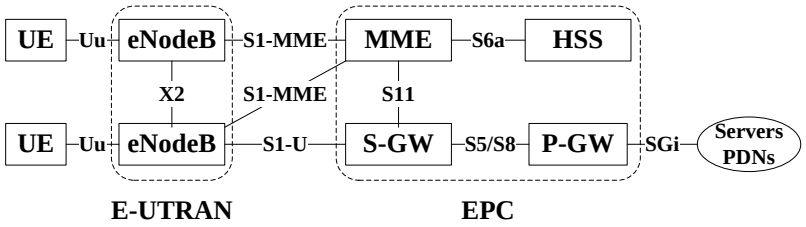


Fig. 1 LTE/LTE-A network architecture

The LTE-A system is composed of two parts about E-UTRAN (Evolved Universal Terrestrial Radio Access Network) and EPC (Evolved Package Core), the base station eNodeB consists of access network and core network incluedes S-GW (Serving GateWay), MME(Mobility Management Entity), P-GW (PDN GateWay) and many other network elements. The main interface and its functions of the LTE-A system network are shown in Table 1.

Table 1 LTE-A network main interface and its function

| Interface name | Connection network element | Interface function description |
|---|---|---|
| S1-U | eNodeB - SGW | Building a tunnel between GW and eNodeB devices, transmitting user data services--user plane data |
| S1-MME | eNodeB - MME | For the transmission of SM (session management) and MM (mobility management) information |
| X2 | eNodeB - eNodeB | Control and user plane information between base stations |
| S5 | SGW - PGW | Building a tunnel between GW devices, transmitting user plane data and control plane information |
| S8 | SGW - PGW | When roaming, the interface between the network PGW and the network SGW, the transmission of the control and user plane data |
| S11 | MME - SGW | Building a tunnel between MME and GW devices to transmit control plane data |
| SGi | PGW - External Internet | Building a tunnel to transmit user plane data |

The user plane interface S1-U is located between eNodeB and S-GW, and the S1-U interface protocol stack is shown in Fig. 2. The transmission network layer based on IP transmission, he GTP-U is above UDP/IP to transfer the user plane's PDU(Protocol Data Unit) between S-GW and eNodeB, including HTTP PDU and DNS PDU, so this article analyzes the users behavior is mainly through the analysis of the acquisition data from S1-U interface.

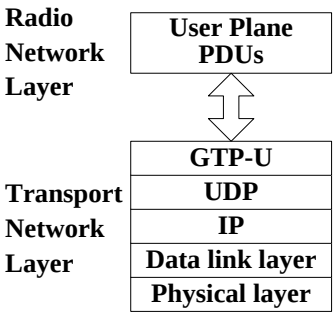


Fig. 2 S1-U interface protocol stack

## 3. Design and Implementation of User Behavior Analysis System

In order to better solve the user needs and expand business of the operators, this paper designed and built a user behavior analysis system based on the Hadoop big data platform [8,9]. The user behavior analysis system consists of three functional modules, such as acquisition and decoding, business identification and statistical application. The system architecture is shown in Fig. 3.

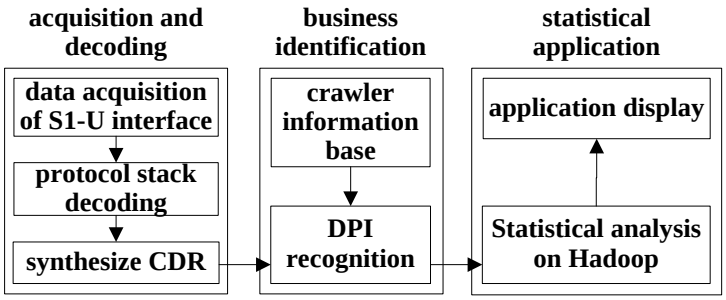


Fig. 3 System framework for user behavior analysis

### 3.1 Acquisition and Decoding

The data source of user behavior analysis is the S1-U interface data. It is acquired in real time by collecting card, then decode data by protocol stack, and finally synthesize CDR file for DPI module to identify business. The acquisition and decoding module is mainly divided into two parts: protocol stack decoding and CDR synthesis.

3.1.1 Protocol Stack Decoding

The decoding of the protocol stack mainly completes the decoding function of S1-U interface data, and its specific process is shown in Fig. 4. The protocol stack decoding is from the bottom to the upper. As for S1-U interface data,Decode IP, UDP, and GTP-U in turn. First determine whether the data is empty, if it is not empty, then recognize the protocol, using the corresponding protocol decode function to decode the data. After that, judge whether it has the upper data. If it has upper data, continue decoding until there is no data to decode. Finally, using the FillTreeBuf function to construct a tree structure, filling the decoding result.

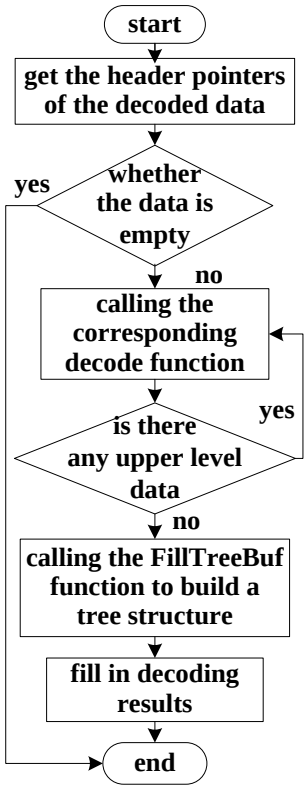


Fig. 4 Protocol stack decoding process

3.1.2 Synthesize CDR

The synthesis of CDR mainly synthesizes the decoding results of data, and records the correspondence between CDR and the original message in memory. The CDR data includes three types about the CDR of public data, the CDR of HTTP data, and the CDR of DNS data, the CDR of public data is defined as follows:

```
class CServiceBaseCDR : public CBaseCdr
{
  public:
   int64 uTcpConnID;
   ip_protocol bearPro;
   pro_type cdrPro;
   char chIMEI[16];
   uint8 uCdrType;
   uint32 uPTMSI;
   uint32 uMTMSI;
   char chMSISDN[24];
int32 nLac;
.....
};
```

The synthesis of CDR is first extracting the key information--Key from the decoding result message; judging whether the Key's corresponding of CDR record exists, if it exists, get the corresponding CDR of the Key from hash table, and update the attribute information of CDR; if it dose not exists, establishing the hash index and the new CDR record, and set the attribute information of CDR. Then, determine whether the current message is the CDR ending message. If not, the subsequent message

will improve the CDR content; if so, remove Key and output the complete CDR as CSV file for business identification module using.

### 3.2 Business Identification

Business identification module is mainly to identify data that users access to the Internet, including business types, such as video, shopping, reading and other specific contents. This module includes two parts: crawler information base and DPI recognition.

3.2.1 Crawler Information Base

In the crawler information base, there is some information that can be crawled by the web crawler, mainly includes the key information of the websites, such as the video name, the starring and other information. The specific process of crawling information by a web crawler is shown in Fig. 5.

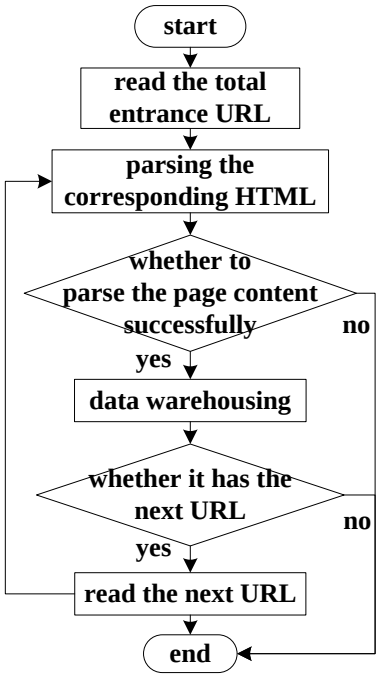


Fig. 5 Work flow of web crawler

To crawl information of the Youku video site as an example, first read the general entrance URL "http://list.youku.com/category/show/c_97.html", using the Jsoup to parse HTML content; judging whether the page is resolved or not, if successful, store the parsed contents in the database; then check whether there is next URL from the URL list, if there is, read the next URL, and analyze them, followed by the cycle until there is no presence of a URL, ending to crawl the information .

The key code for crawling information from a web crawler is as follows:

```
private void getDetail(String url)
{
Elements elements = null;
try
{
Document document = Jsoup.connect(url).get();
if (document != null)
{elements = document.getElementsByClass("p-base");}
}
if (elements.size() == 0)
{detail = number + "##" + id + "##" + jiShu + "##" + name + "##" + newTime + "##" + bigType + "##" + timeToMarket + "##" + actor + "##" + local + "##" + type;
```

```
writeToOracle(new VideoDetailInformation(detail));
new StringToFile(detail, "youku");}
else
{.....
}
}
```

3.2.2 DPI Recognition

The DPI identification process is shown in Fig. 6. First of all, loading the crawler information database, and read the CDR file.Secondly, according to the IP five tuple information, the source address of IP, the source port, the destination address of IP, the destination port and the protocol type to find out the business types. finally followed by Host, URL crawler information database, and other features to identify business, if it succeed, fill in the recognition statistics table, if it fail, end the recognition.

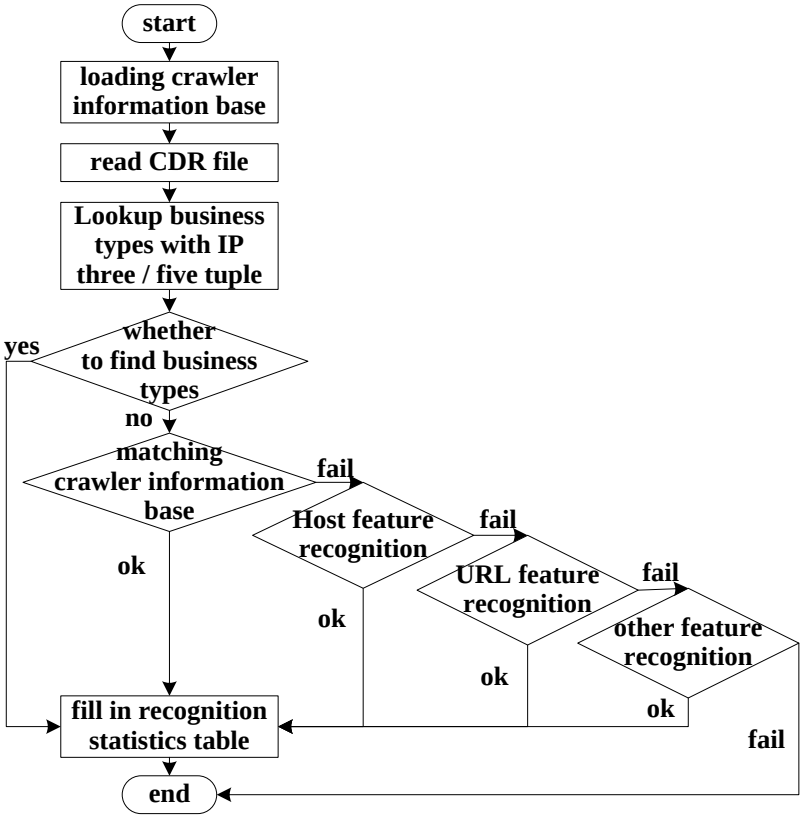


Fig. 6 DPI recognition process

The key code for DPI recognition implementation is as follows:

```
int32 CDPIDataFormat::Csv2DPI(const char *chCsv, int32 nCsvLen, CDpiFirst &firstDpi )
{
CParseCsvLine CsvLine((char *)chCsv, "$,#", 3, true);
CsvLine.GetSingleSection(53, firstDpi.chHost, sizeof(firstDpi.chHost));
if (firstDpi.chHost[0] == 0)
   ……
CsvLine.GetSingleSection(54, firstDpi.chUrl, sizeof(firstDpi.chUrl));
   ……
CsvLine.GetSingleSection(68, firstDpi.chPostInfo, sizeof(firstDpi.chPostInfo));
```

……

}

### 3.3 Statistical Application

The statistical application module includes two parts: statistical analysis and application display. Statistical analysis is complete functions such as the import, statistics, analysis of business recognition results data on the Hadoop platform. This part of the function is implemented by the hive component, using HQL (Hibernate Query Language) language [10]. The framework design is shown in Fig. 7.

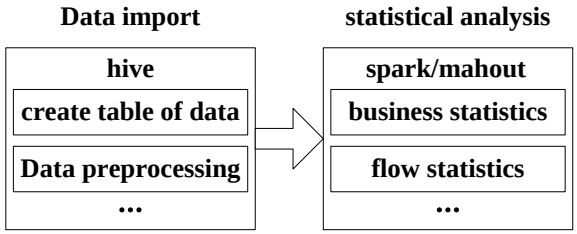


Fig. 7 Statistical analysis process

First, import the recognition statistics into hive and preprocess the data, then analyze the data based on different levels of time granularity through mahout/spark components, and form user behavior thematic analysis data. Application display is show the statistical analysis of the results, including the user specific business, the traffic generated by the business and so on.

The part of the code to create the table in hive is as follows:

```
create external table videoinfo(
ID_NO string,
IMSI_NO string,
websiteNum string,
updateTime string,
id string,
name string,
typeOne string,
typeTwo string,
protagonist string,
.....
)
row format delimited
fields terminated by ','
stored as textfile;
```

## 4. Verification Results of User Behavior Analysis System

The test verification of the user behavior analysis system is mainly divided into the verification of the crawler information base, the business identification and the statistical application. The information of crawler information base stored in the Oracle database. Take video as an example, the information stored in the database include the video site number(websiteNum), video identification(id), video name(name), the current time of output video information(updateTime), video classification(typeOne), specific video types(typeTwo) and others. Fig. 8 is the result of partial crawling video.

| websiteNum | id | updateTime | name | typeOne | typeTwo | local | protagonist |
|---|---|---|---|---|---|---|---|
| 6 | 1u657wp0v98un9p/a0022t6tsu3 | 2017/05/26 14:26:12 | 罗曼蒂克消亡史 | 电影 | 内地 动作 悬疑 院线 | 内地 | 葛优 章子怡 浅野忠信 钟欣潼 杜淳 倪大红 袁泉 |
| 6 | zm91ry6rctntum8/e0023xe9evh | 2017/05/26 14:26:23 | 熊出没·奇幻空间 | 电影 | 内地 动画 奇幻 冒险 | 内地 | |
| 6 | mqn4th6zsdepxd8/h0023ah3nzf | 2017/05/26 14:26:24 | 非凡任务 | 电影 | 内地 动作 院线 犯罪 | 内地 | 黄轩 段奕宏 郎月婷 祖峰 邢佳栋 王耀庆 |
| 6 | gm5raidogphgm60/q0023z0hyq8 | 2017/05/26 14:26:26 | 绑架者 | 电影 | 内地 动作 犯罪 院线 | 内地 | 白百何 黄立行 明道 |
| 6 | r18zt91aswyuk3w/p00234rvpit | 2017/05/26 14:26:26 | 合约男女 | 电影 | 内地 爱情 喜剧 院线 | 内地 | 郑秀文 张孝全 林雪 冯文娟 |
| 6 | c3qupkn17f0y0qa/o0022bfel6d | 2017/05/26 14:26:29 | 我们的十年 | 电影 | 内地 爱情 青春 院线 | 内地 | 赵丽颖 乔任梁 吴映洁 班嘉佳 范逸臣 冯铭朝 |
| 6 | vz8j5agokp0yugl/o00211t2bcb | 2017/05/26 14:26:30 | 绝地逃亡 | 电影 | 内地 动作 喜剧 院线 | 内地 | 成龙 范冰冰 约翰尼·诺克斯维尔 曾志伟 王敏德 |
| 6 | vz8j5agokp0yugl/v00213s81bn | 2017/05/26 14:26:30 | 绝地逃亡 | 电影 | 内地 动作 喜剧 院线 | 内地 | 成龙 范冰冰 约翰尼·诺克斯维尔 曾志伟 王敏德 |
| 6 | 1wa6zo6gmd93t02/a0022zblxjq | 2017/05/26 14:26:30 | 摆渡人 | 电影 | 内地 爱情 喜剧 院线 | 内地 | 梁朝伟 金城武 陈奕迅 Angelababy 张榕容 |
| 6 | 1wa6zo6gmd93t02/f0022inzefv | 2017/05/26 14:26:30 | 摆渡人 | 电影 | 内地 爱情 喜剧 院线 | 内地 | 梁朝伟 金城武 陈奕迅 Angelababy 张榕容 |
| 6 | xg95sxi4q7zc4uo/u0020m5jr73 | 2017/05/26 14:26:31 | 美人鱼 | 电影 | 内地 喜剧 爱情 科幻 院线 | 内地 | 邓超 罗志祥 张雨绮 林允 |
| 6 | yvt4p7vcd8xu184/u0023c02236 | 2017/05/26 14:26:31 | 欢乐喜剧人 | 电影 | 内地 喜剧 院线 | 内地 | 郭德纲 岳云鹏 罗温·艾金森 艾伦 张小斐 张泰维 |

Fig. 8 Crawler information base- video information

From the first data in Fig. 8, we can see that the number of crawling video sites is “6”, the video name is “history of the destruction of romance”, the type is “movie”, the specific types are “inland, action, suspense” and other information. This can prove the comprehensiveness of the information stored in the crawler information base.

DPI recognition is putting out the recognized results that stored as CSV file, this file mainly includes information such as HOST, URL, website name, DPI Key, specific business content and so on. Fig. 9 is the output result of partial DPI recognition.

| Host | URL | Sitetype | DPIKey | DPI |
|---|---|---|---|---|
| list.youku.com, | /show/id_z40efbfbd6208efbfbd75.1 | 优酷, | z40efbfbd6208 | 拆弹专家 香港 犯罪 悬疑 |
| list.youku.com, | /show/id_z5a6fefbfbd5b00efbfbd.1 | 优酷, | z5a6fefbfbd5b | 喜欢你 电影 内地 喜剧 爱 |
| www.hongxiu.com, | /book/7004586903966801, | 红袖, | 7004586903966 | 君上家的小妖花 仙侠奇缘 |
| video.tudou.com, | /v/XMjgyODI4MDY3Mg==.html?spm=a | 土豆视频, | XMjgyODI4MDY3 | 火影忍者 经典战役之卷, |
| www.xxsy.net, | /info/891650.html, | 潇湘书院, | 891650, | 都市之鬼女天师 悬疑 日玩 |
| www.iqiyi.com, | /lib/m_209031114.html?src=searc | 爱奇艺, | 209031114, | 漂洋过海来看你 电视剧 内 |

Fig. 9 DPI recognition results

As you can see from the first data in Fig. 9, the identified video ID is "z40efbfbd6208efbfbd75", and the video is called "bomb dismantling expert". Look up source code of Youku video web page, you can find the ID is "z40efbfbd6208efbfbd75", the video is the "bomb dismantling expert", it can be proved that DPI recognition can accurately identify the user behavior, and can meet the requirements of the system, and the recognition of the mainstream application accuracy is above 80%.

Statistical analysis is based on Hadoop big data platform for user traffic data analysis, including traffic and other specific information, and then classify users. Some users business statistical analysis results are shown in Fig. 10.

| IMSI | TYPE | CONTENT | FLOW | LABEL |
|---|---|---|---|---|
| 460026271273737 | 视频 | 电影 | 497420 | 一般偏好用户 |
| 460007070440083 | 视频 | 电视剧 | 33280 | 弱偏好用户 |
| 460026271519537 | 视频 | 电视剧 | 52500 | 弱偏好用户 |
| 460008232744466 | 视频 | 动漫 | 950720 | 强偏好用户 |
| 460026271222980 | 视频 | 电影 | 1250822 | 强偏好用户 |
| 460006191961542 | 视频 | 动漫 | 505824 | 一般偏好用户 |
| 460001064819130 | 视频 | 综艺 | 628006 | 一般偏好用户 |
| 460026271210433 | 视频 | 电影 | 82210 | 弱偏好用户 |

Fig. 10 Statistical analysis results

The user label is calculated for the number of visits, the use of flow and the use of time, which uses the weighted average algorithm. And then descending order of users, the first 30% users are labelled as strong preference users, 30% to 70% users are labelled as the general preference users, After 70%, the users are labeled as weak preference users.70% labels for weak user preferences.

## 5. Concluding Remarks

In this paper, achieving the mobile Internet user behavior analysis system based on protocol decoding, DPI recognition, hive and spark components of Hadoop big data platform. It can accurately identify the user behavior, including the analysis of users specific business , users flow and users preference, which may help operators achieve precision marketing. Further, we can use data mining algorithm, such as clustering algorithm, neural network to do deep mining of business identification data, and conduct multi-dimensional analysis of user behavior.

## Acknowledgements

This work is supported by Major National Science and Technology Special Project of China (2015ZX03001013), Ministry of Education - China Mobile Research Fund (Grant no. MCM20150508), Major Theme of Key Technology Innovation of Key Industries in Chongqing (cstc2017zdcy-zdzx0030), and Chongqing University Innovation Team (KJTD201312).

## References

[1] Wu J, Li W, Huang J, et al. Key techniques for Mobile Internet: a survey, Scientia Sinica Informationis, Vol. 45 (2015) No. 1, p.45-69.
[2] Matthew N.O. Sadiku , Adebowale E. Shadare , Sarhan M. Musa, Mobile Internet, International Journal of Engineering Research, Vol. 6 (2017) No. 9, p.429-430.
[3] Wan F, Zhao X, Liang X, et al. Search Behavior Study Based on the Mobile SearchLog, Journal of Chinese Information Processing, Vol. 28 (2014) No. 2, p.144-150.
[4] Neelima G, Rodda S. Predicting user behavior through sessions using the web log mining, International Conference on Advances in Human Machine Interaction. IEEE, (2016), p.1-5.
[5] Chen K, Huazheng F U, Chen C, et al. A real time approach to user interest classification using DPI, Telecommunications Science, Vol. 32 (2016) No. 12, p.109-115.
[6] Shih M J, Pang Y C, Lin G Y, et al. Performance Evaluation for Energy-Harvesting Machine-Type Communication in LTE-A System, International Journal of Wavelets Multiresolution & Information Processing, Vol. 11 (2015) No. 2, p.335-356.
[7] Lei L I, Zhang Z, Bing X I. Research and implementation of multi-protocol association scheme on Uu interface in LTE-Advanced network, Telecommunications Science, Vol. 32 (2016) No. 2, p.167-176.
[8] HU W Y, AI M, ZHOU G B, et al. Design and implementation of big data based signaling monitoring system, Video Engineering, Vol. 40 (2016) No. 1, p.95-101.
[9] LEI L, LI J W, GONG D P, et al. Study on data modeling and collection in OSS based on Hadoop, Telecommunications Science, Vol. 31 (2015) No. 1, p.128-138.
[10] Feng X, Xiyu W U, Zhao J, et al. Data warehouse of QAR based on Hive, Computer Engineering & Applications, Vol. 53 (2017) No. 11, p.90-94