Network Big Data Analysis and Mining Based on DPI Technology

Zongbao Ma*

School of Information Engineering, Xi'an University, No.168, Taibai south road, 710065, Xi'an, China

*Corresponding author e-mail: mzborbb@163.com

Abstract. In recent years, the rapid development of mobile data network, with 4G commercial wireless network has entered the data era, more and more big data impact on the network of mobile operators, network monitoring and network security need to be strengthened. Depth packet detection technology DPI. The purpose of this paper is to complete network data analysis by in-depth identification of key words in IP packets, so as to meet the current demand of operators for network analysis. In this paper, DPI technology is applied to the actual network, and the DPI data analysis system is constructed by combining the 2G and 3G networks of China mobile. The key technologies in DPI include character recognition, behavior pattern recognition and application layer gateway recognition. At the end of the experiment, the mass signaling data from GB interface and iu-ps interface were obtained to effectively identify various applications and data mining in the network. This paper expounds big data analysis and mining technology and its decision application research from different aspects, hoping to provide theoretical reference for experts and scholars studying the application of big data analysis and mining technology.

Keywords: DPI Technology, Network Big Data, Data Mining, Wireless Network Optimization

1. Introduction

In recent years, the network data business develops rapidly, and new businesses emerge one after another, such as P2P, VOIP, WEBTV, streaming media, video chat, online games and so on. As these new services are accepted by the majority of users, operators have more and more customers at the
same time, the huge growth of data traffic also impact the network's bottom traffic model and upper application model. A series of problems such as bandwidth management, information security, billing model, public opinion control and so on arise. Big data era, various industries in the process of development to produce large amounts of data, the analysis of the data is very important, to promote the development of transportation, lk sword, large data analysis technology to the data collected in various industries, stored data, scientific and effective to meet the need of transportation, the development of Ne analysis, improve the use efficiency of data analysis, improve the utilization rate of the data is to promote the development of industry needs. Data mining is the vast amounts of data how to extract useful data moving and towel wipes development in the field of industry process, the role of information technology and data mining, commonly used methods for modern large data analysis wipes, save for data analysis in specific jj industry often combined with some algorithm into the ij, such as deposit is usually combined with genetic algorithm is used in biological field, to improve the speed of data mining, data mining is scientific, accuracy, completeness, all play an important role, big data analysis is useful to data mining methods. Improve the role of information technology in data mining.

For the past decade, Swedish company Goceram Industrial Ventures AB of Molndal has been developing a medium-pressure injection molding technology for metal and ceramic powders. The aim is to provide a low-cost alternative to perfecting the high pressure injection molding technology, which also overcomes the degreasing and handling problems that hinder acceptance of the low pressure injection molding process. Robert Pompe and Jesper Brandt described in detail the equipment and processes involved in MEDPIMOULD™ technology and compared the cost and performance of low, medium and high pressure powder injection molding processes [1]. The rapid development of location capture technology promotes the generation of trajectory data, which can track the trajectory of moving objects. Trajectories are typically represented by a series of timestamps. A wide range of applications can benefit from trajectory data mining. Large-scale ballistic data presents unprecedented opportunities as well as enormous challenges. M. Tsukada reviewed various applications of trajectory data mining, such as trajectory discovery, position prediction, motion behavior analysis, etc. In addition, m. Tsukada also reviewed a large number of existing trajectory data mining techniques and discussed them in a trajectory data mining framework. The framework and survey results can be used as a reference for the design of future ballistic data mining solutions [2]. Prolonged QT interval drug-drug interactions (qt-ddis) can increase the risk of life-threatening arrhythmias. Despite regulatory guidelines for testing, these interactions are often discovered only after a drug has hit the market, and can go undetected for years. Using adverse event reporting, electronic health records (EHR), and laboratory experiments, Tal Lorberbaum's goal was to develop a data-driven pipeline to discover QT-DDIs. 1.8 million adverse event reports were mined for signals indicating qt-ddi.
hypothesized interactions were either refuted or confirmed using 1.6 million EHR results from 380,000 patients at the Tal Lorberbaum institute [3-4].

In this paper, the DPI data service analysis system was used to analyze the existing network of a mobile operator, and a large number of GB and iu-ps interface signaling data were collected. Through the analysis of the data, the problems in the network were actively discovered, the network performance was optimized, and the complaints were moved forward. At the same time, the complaints can be targeted to analyze the problem, shorten the problem discovery time, quickly solve the problem, improve the quality of the network. Through the system to provide a rich index system to establish a real-time omni-directional monitoring mechanism, prevent the occurrence of network problems, reduce the occurrence of complaints, complaints processing forward.

2. Proposed method

2.1. Classification of DPI technology

According to different recognition technologies, DPI technology can be divided into three categories: feature word recognition technology, application layer gateway recognition technology, and behavior model recognition technology. These three technologies are respectively used for different types of protocol recognition, and cannot be substituted for each other, but can be used in combination to achieve the optimal detection efficiency and flexibility[5].

(1) Feature word recognition technology

Different network applications generally use different transport protocols, and each protocol has its own special identification, like the only human "fingerprint". These "fingerprints" may be specific ports, specific strings, or specific Bit sequences. The recognition technology based on the characteristic word is to determine the application carried by the service by reading the fingerprint information in the datagram. Due to the different specific methods of data detection, the data recognition technology based on feature words is subdivided into three branch technologies, namely: fixed feature position matching, variable feature position matching and state feature word matching. Through a variety of protocols to constantly upgrade their fingerprint data, the recognition technology based on feature words can be very convenient for functional expansion, and realize the detection of new protocols[6].

(2) Application layer gateway recognition technology

The second is to use the application layer gateway recognition technology, this technology is mainly aimed at separating the control flow and business flow of business, and no. 7 signalling
relevant business, for example, it no characteristics of traffic flow, DPI by application layer gateway to identify its control flow, and according to the flow of control agreement to determine the specific application layer gateway for the interpretation of the business flow, so as to identify its business flow[7].

(3) Behavioral pattern recognition technology

Third class USES the behavior pattern recognition technology, operators through the network in all kinds of terminal is implementing the behavior study, on the basis of user behavior model, according to user behavior model, DPI technology further judgment according to the user has to implement the behavior of the users of ongoing and upcoming behavior, this kind of recognition technique is usually used cannot judge user behavior identification according to the agreement[8].

2.2. Network data method

(1) Data analysis method

In the era of big data, there are many types of data, large amount of data and difficult data analysis. In the era of Internet +, the development of enterprises produces more data, so we must adopt scientific data analysis methods, conduct scientific and accurate data analysis, improve the utilization rate of data, and give play to the role of data in the development of enterprises. In general, big data analysis is mainly analyzed from three aspects: data capacity, data rate and data format. There are many methods for big data analysis, and choosing appropriate data analysis methods based on specific data analysis types plays an important role in improving the efficiency of data analysis. It has an important application in the storage of data with a large capacity. Now the capacity of data generated by enterprises is relatively large, so the traditional data storage method is no longer suitable for modern data storage. Cloud storage is usually adopted for storage[9].

(2) Data mining methods

Data mining is a common way in data analysis, usually in the most useful data extraction, can not analyze the law of the data, can also analyze the law of the data, in the computer and other fields commonly used data mining to extract the data, give play to the significance of the data. There are many methods for data mining. Users choose appropriate mining methods according to the needs of data analysis. Sometimes nr is combined with a variety of data mining methods to give full play to the performance of each data mining method. Data mining mainly includes classification, homology analysis and Web data mining. Each type of data mining method has its own characteristics, users must choose the data mining method combined with the actual needs of data analysis[10].
3. Experiments

3.1. Experimental environment

This experiment through constructing the DPI of data analysis system, to obtain GB or IU interface - PS key agreement, decoding, synthesis, statistical analysis, take the initiative to find the problems existing in the network, optimize the network performance, will reach complaints, complaints occur at the same time can be targeted to analyze problems, shortening the time of problem found, quick to solve the problem. Figure 3-1 shows the network topology of the DPI analysis system, which is composed of three servers and one external storage. The existing network of a mobile company is utilized to collect the massive signaling data of GB/ iu-ps interface from the collection server. The GB decoding server and iu-ps decoding server are responsible for the decoding, call synthesis and statistical analysis of the data of GB/ iu-ps interface. The data volume of GB/ iu-ps interface is large and needs external storage. After analyzing the data through the server, a report is formed and extracted directly from the client.

3.2. Experimental steps

The signal acquisition scope of the system has covered the whole network of a city, and all BSC and RNC in the network have been connected to the DPI system. The data obtained through the system are as follows:

(1) Analyze the visits, process ranking and access file types of SP(direct provider of mobile Internet service content application service);

(2) Analyze the performance index of mobile self-owned service;

(3) Analyze the overall performance of MMS, success rate of sending and receiving, download rate and MMS customization business;

(4) Analyze customer complaints;

(5) Analyze the success rate of DNS resolution from the aspects of IP, domain name, terminal and business;

(6) Analyze the data flow from the terminal, user and network dimensions;

(7) Analyze user perception from the aspects of load bearing performance, terminal performance, business performance, user perception and cell performance;

(8) The terminal performance is analyzed from the aspects of terminal load bearing, user behavior
and locking network user.

4. Discussion

4.1. Data collection and performance analysis of dpi data analysis system

Through the analysis of the SP/HOST address of the mobile website, the top ones are mainly 10086.cn mobile website, mobile reading, mobile dream, mobile MM and mobile video. The number of visits generated by the mobile self-owned business was 1629,225, accounting for 1% of the total number of attempts and 0.86% of the total number of failures. The most influential ones were the mobile official website and mobile phone reading. The specific performance parameters are shown in Table 1.

| The name of the SP     | Number of attempted visits | The success rate | Downlink flow (MB) | Rate (Kbps) | Downlink mean time delay (ms) | Number of failures | Failure of |
|------------------------|-----------------------------|------------------|-------------------|-------------|-------------------------------|-------------------|------------|
| 10086.cn              | 632570                      | 80.48%           | 7801.8            | 2.35        | 1029.06                       | 123427            | 0.47%      |
| cmread.com            | 570968                      | 90.19%           | 4227.04           | 1.25        | 968.64                        | 57510             | 0.22%      |
| monternet.com         | 220118                      | 93.07%           | 1308.51           | 2.55        | 933.18                        | 14111             | 0.05%      |
| mmarket.com           | 109205                      | 82.58%           | 5371.75           | 14.45       | 915.21                        | 19510             | 0.07%      |
| cmvideo.cn            | 96364                       | 86.74%           | 690.77            | 3.19        | 801.42                        | 13344             | 0.05%      |

In the above table, in the traffic ranking of mobile self-owned business, mobile official website occupies the main position, which accounts for more than 40% of the traffic of mobile self-owned business, and the number of failures accounts for 54.16% of the total number of failures of mobile self-owned business. Second is cmread.com(mobile reading), which accounts for 35.05% of its own business and 25.23% of its total failures. From the data analysis of mobile business, the performance of each business is better.

4.2. Statistics of time distribution of each index

Statistics are made on the success rate of each SP, so as to obtain the influence of the access success rate of a single SP on the success rate of GET of the whole network. Through data analysis, qq.com has the greatest impact, with 5.76% of the total number of attempts and 17.58% of the total number of attempts. Followed by baidu.com and qlogo.cn, the number of failures accounted for 3.21% and 1.14% of the total number of attempts, accounting for 9.80% and 3.47% of the total number of attempts,
respectively. Finally, it is concluded that the failure of the TOP10 SP has an impact of 7.44% on the success rate of GET. According to the statistics of SP hourly index of the whole network of a city by DPI system, the peak of business flow appeared at 21 points, reaching 339GB. The number of attempts peaked at 20 at 31 million. As shown in Figure 1.

![Figure 1. Trend chart of SP traffic and number of attempts in the whole network](image)

It can be seen from the data in figure 1 that the success rate is not proportional to the traffic, and the lowest success rate of 22-point attempts is only 76.78%, but the traffic in this period is not the highest. From this, it can be concluded that busy data will affect the success rate of business access, but not necessarily. The average delay of the whole network remained around 1100ms, and the highest time period of the rate was 6.34kbps, which appeared at 23 points. Therefore, the rate trend was relatively stable. Secondly, the application of SP traffic in the top 20 is analyzed. Tencent business ranked the first in terms of traffic, with a daily downstream of 375.57gb and a success rate of 88.78%. Second, qpic.cn belongs to tencent weibo access, with 208.62gb of traffic. Again,baidu.com, which belongs to baidu website, has a success rate of 86.28% and generates a large amount of traffic. The daily traffic generated by baidu.com is 161GB, and the number of failures accounts for 9.8% of the total number of failures.

5. Conclusions

In this paper, according to the actual circumstance of city net now, using network resources to build
DPI data analysis system, through the analysis of the system data has the overall understanding to the municipal data network, and according to the results of focuses on network optimization, DPI technology as a new analytical technology on the depth of the network will be more and more accepted by the operator. With the high-speed development of Internet, more and more new business will also appear, the challenge for the network optimization will be more and more, operators need to constantly from daily optimization and gain more experience in maintenance, improve their own network, efficient service to provide users with higher quality, create the best products in a network for users to use. With the development and application of big data technology, there have been earth-shaking changes in various social fields such as finance, medical treatment, telecommunications and education, and a large amount of data will be formed every day. Because there are many uncertain factors in the society, the processing of data types will be more complicated. After the application of computer technology, there are still great limitations in the traditional processing methods and problem solving. Now with the help of data mining technology, big data problems can be effectively solved, the future will be the era of "data is king".

References

[1] R. Pompe, J. Brandt. Goceram's MEDPIMOULD technology offers cost-effective PIM production[J]. Metal Powder Report, 2016, 56(6):14-17.

[2] Zhenni Feng, Yanmin Zhu. A Survey on Trajectory Data Mining: Techniques and Applications[J]. IEEE Access, 2017, 4(9):2056-2067.

[3] Tal Lorberbaum, Kevin J. Sampson, Jeremy B. Chang. Coupling Data Mining and Laboratory Experiments to Discover Drug Interactions Causing QT Prolongation[J]. Journal of the American College of Cardiology, 2016, 68(16):1756-1764.

[4] T. H. Tuchscheerer, Dipl.-Phys. Dr.-Ing habil. J. Kuprianoff. Die Erkennbarkeit einer erfolgten Bestrahlung bei Lebensmitteln[J]. European Journal of Lipid Science & Technology, 2016, 67(2):120-124.

[5] Dr. P. A. Arsenev, Dipl.-Phys. I. Senff, A. V. Potemkin. The Search for Prospective Media for the Application as Active Laser Elements[J]. Crystal Research & Technology, 2015, 13(1):5-16.

[6] Dr. Michael Kleinkes, Dipl.-Phys.-Ing. Christian Schmidt…. Safety and Styling of Future Rear Lights[J]. Atzautotechnology, 2015, 114(1):48-51.

[7] M Tsukada, R Irie, Y Yonemochi. Adhesion force measurement of a DPI size pharmaceutical
[8] Min Zhao, Yu You, Yachao Ren. Formulation, characteristics and aerosolization performance of azithromycin DPI prepared by spray-drying[J]. Powder Technology, 2016, 187(3): 214-221.

[9] V. Krishnaiah, G. Narsimha, N. Subhash. Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review[J]. International Journal of Computer Applications, 2016, 136(2): 43-51.

[10] Qiming Sun, Libao Shi, Yixin Ni. An enhanced cascading failure model integrating data mining technique[J]. Protection & Control of Modern Power Systems, 2017, 2(1): 5.