Data Science and Advanced Analytics (DSAA), 2014 International Conference on
Year: 2014, Pages: 361 - 366, DOI: 10.1109/DSAA.2014.7058097
IEEE Conference Publications

Hadoop based Deep Packet Inspection System for Traffic Analysis of E-Business Websites

Jiangtao Luo, Yan Liang, Wei Gao, Junchao Yang
Electronic Information and Networking Research Institute,
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
Email: Luojt@cqupt.edu.cn

***Abstract*—Internet traffic is experiencing an explosive growth, and online shopping is one of the significant drivers. However, alert network operators, unwilling to be dumb pipes, are making every effort to mine mass traffic with the help of *Deep Packet Inspection* (DPI) which is regarded as a big challenge especially for massive data when traditional methods and programming model are utilized. Hadoop provides an alternative approach with its strength in distributed storage and parallel computing. In this paper, a Hadoop based DPI system was reported, which was integrated with a web crawler. The system architecture and MapReduce models of packet analysis, web URL restoration were presented. As an example, live web traffic visiting the Tmall, the leading e-shopping giant in China, was specifically investigated using this system. Popularity of product, category and brand for a certain period was evaluated from page views of product. The detailed information of products was provided by the product information base built by the web crawler. This work explored the methodology of using Hadoop in DPI and presented valuable guidelines to develop such a system, which can be further used in analyzing other services and mining the value of network traffic by network operators.**

I. Introduction

Internet traffic is experiencing an explosive growth, and online shopping is one of the significant drivers. The global B2C (Business-to-Customer) E-Commerce was reported to grow at the rate of 20% annually, and one billion persons were expected to make purchases online this year [1]. As the center of emerging market, China contributed a great deal to the rapid development. In 2008–2012, Chinese online shopping market size ascended from RMB 128.2 billion to RMB 1,303.0 billion at the CAGR (Compound Annual Growth Rate) of 78.6% [2].

But for network operators, the huge mass of traffic means a big burden on network maintenance, and more important, it is a big challenge to monitor the traffic and find the deep value out of the mass data in order to avoid being dumb pipes. In this case, traffic classification or identification is not enough. For example, it is not much helpful and interesting to the industry just identifying application type with general traffic statistics for an e-business website. The detailed behaviour of consumers at least their favourite product brands were what they expected to know, which may bring potential commercial benefits for other industries or local economy. In such cases, technologies of *deep packet inspection* (DPI) are required.

It is a challenging task to learn the details of a product (title, brand and etc.) through DPI of network traffic, especially for a large volume of data. On one hand, the product details cannot be directly told from the Uniform Resource Locator (URL) in HTTP request since each product is always represented by a digital code. we have to find out the corresponding association between the codes and product information, such as its category hierarchy, product brand and title. On the other hand, a large quantity of data with orders of Perabytes or Exabytes obsoletes traditional data mining methods and calls for distributed and parallel approaches.

Hadoop, consisting of a distributed file system (HDFS) and a parallel programming model (MapReduce), is regarded as the most popular solution for massive data mining [3]. In this paper, the possibility of implementing DPI on a Hadoop cluster was explored and an architecture of a Hadoop based Internet traffic DPI system was presented. Some key issues including MapReduce implementation of packet analysis, input file size selection and etc. were addressed.

The contributions can be outlined as follows. First, a hadoop based DPI system architecture integrated with a web crawler was firstly proposed. Second, a methodology for web anatomy was presented; third, MapReduce model of HTTP URL restoration was designed and implemented. Finally, for verification, live traffic was captured and analyzed using the system. Specifically, traffic towards Tmall (tmall. com), one of the leading e-business websites in China, was in detail investigated. The popularity of product, subcategories and brand on the Tmall was retrieved with the help of parsing the pages fetched by a web crawler.

The remainder of the paper was organized as follows. Section II reviewed related work on traffic classification and analysis using MapReduce. Section III presented an architecture of DPI system based on Hadoop and integrated with a web crawler. The methods of website anatomy of Tmall, URL restoration and product identification through DPI were in detail investigated in this section. Then, section IV described the experiment setup and demonstrated the analysis results. Finally, section V concluded this paper and prospected for the future work.

II. Related Work

Hadoop/MapReduce has emerged in traffic classification and analysis of network measurements, but little used in DPI. Huang [4] proposed a cloud-based traffic classification service

TABLE III
BRANDS RANKING OF WOMEN'S WEAR.

Brand Name (id)	PVs	PV percentage
HSTYLE (brand8598007)	79	7.77%
Bershka (brand5764853)	73	7.18%
ELF SACK (brand10518561)	37	3.64%
LE'TEEN (brand4057780)	28	2.75%
Shinena (brand82745224)	20	1.97%
Other 288 brands	780	76.70%
total	1017	100%

subcategory. During the observation period, totally 48 subcategories under "women' wear" were observed. And the Top 3 sub-categories were "woolen sweater" (cat50025784), "cotton dress" (cat50047357) and "jeans" (cat50025227), which contributed nearly 26% to the total PVs, according to Table II.

The attention of brand was evaluated by aggregating the PV counts of each brand. According to the results listed in Table III, products from total 293 brands were browsed during the observation period, and the top three brands of women's wear were HSTYLE (brand8598007), Bershaka (brand5764853) and ELF SACK (brand10518561) which in sum accounted for almost 18.6% of total product PVs.

V. CONCLUSION

A Hadoop based DPI system incorporated with a web crawler was proposed and verified in this paper. The system architecture and the methodology of website analysis, page fetching and parsing were presented. As the most critical part of the system, MapReduce models of packet analysis, URL restoration and as well as the product popularity assessment were thoroughly demonstrated. Simultaneously, some key issues of implementing such a system were addressed, like selection of input file size, page elements filtering, subpage crawling. For verification, web traffic of Tmall, Chinese leading e-commerce website, was analyzed using the proposed system. The popularity of product, category and brand during a certain period was investigated and presented.

In summary, this paper explored the possibility of integrating a network crawler with mass data processing capability of MapReduce in DPI of web traffic, which provided a set of guidelines on developing similar applications.

In future work, other websites will be analyzed; page views initiated by different users will be distinguished. The system will be fine optimized and extended to a larger scale so that it can facilitate traffic monitoring and marketing for network operators.

ACKNOWLEDGMENT

The authors would like to thank the joint financial aids from the *Program for Innovation Team Building* (2013), *Chongqing Municipal Engineering Research Center of Institutions of Higher Education*, and the *Chongqing Municipal Application and Development Planning Projects* (cstc2013yykfA40006). The authors also want to say thanks to the *Network Information Management Center of the CQUPT* for providing cloud platform and facilitating experimental data capturing.

REFERENCES

[1] Reportlinker.com, Global B2C E-commerce and online payment report 2013, New York, June 2013
[2] ResearchInChina, China B2C online shoping industry Report, 2013–2016, Oct 2013
[3] (2013, Dec.) Apache Hadoop, [online]. Available: http://hadoop.apache.org//
[4] Nen-FU Huang, Gin-Yuan Jai, Chi-Hao Chen, and Han-Chieh Chao, "On the Cloud-based Network Traffic Classificaiton and Applications Identification Service," in Proc. *2012 IEEE International Conference on Selected Topics in Mobile and Wireless Networking*, pp. 36–41, 2012.
[5] Yeonhee Lee, and Youngseok Lee, "Toward Scalable Internet Traffic Measurement and Analysis with Hadoop," *ACM SIGCOMM Computer Communication Review*, LNCS 6613, vol. 43, no. 1, pp. 6–13, Jan. 2013.
[6] (2013, Nov.) TcpDump & Libpcap, [online]. Available: http://www.tcpdump.org/
[7] Taghrid Samak, Daniel Gunter and Valerie Hendrix, "Scalable analysis of network measurements with Hadoop and Pig," in Proc. *2012 IEEE/IFIP 5th Workshop on Distributed Autonomous Network Management System (DANMS)*, pp. 1254–1259, 2012
[8] (2013, Dec.) Apache Pig, [online]. Available: http://pig.apache.org/
[9] (2013, Dec.) The R Project for Statistical Computating, [online]. Available: http://www.r-project.org/
[10] Jiangtao Luo and Qingchuan Li, "Packet domain monitoring system based on cloud storage", *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)*, vol. 24, no. 6, pp. 675–681, 2012.06
[11] T. Vieira, P. Soares, M. Machado, R. Assad, and V. Garcia, "Measuring distributed applications through MapReduce and traffic analysis," in Proc. *IEEE 18th International Conference on Parallel and Distributed Systems (ICPDS'2012)*, 2012
[12] (2013, Dec.) JXTA, the Language and Platform Independent Protocol for P2P Networking, [online]. Available: https://jxta.kenai.com/
[13] S. Kawano, T. Okugawa, T. Yamamoto, T. Motono, and Y. Takagi, "High-speed DPI method using multi-stage packet flow analyses," in Proc. *9th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT'2012)*, 2012
[14] G. Macia-Fernandez, Yong Wang, R. Rodriguez-Gomez, and A. Kuzmanovic, "ISP-Enabled Behavioral Ad Targeting without Deep Packet Inspection," in Proc. *INFOCOM 2010*, pp. 1–9, 2010
[15] (2013, Nov.) The Apache Nutch weibsite, [online]. Available: http://nutch.apache.org/
[16] Albetro Dainotti, and Antonio Pescape, "Issues and Future Directions in Traffic Classification," *IEEE Network*, Jan/Feb 2012, pp. 35–40, 2012.
[17] Thuy T. T. Nguyen, and G. Armitage, "A Survey of Techniques for Internet Traffic Classification using Machine learning," *IEEE Communications Survey & Tutorials*, vol. 10, no. 4, 4th Quarter, 2008.