Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 17 (2014) 114 – 121

Procedia Technology

Conference on Electronics, Telecommunications and Computers – CETC 2013

Memory Management for Big Data Mining – Cache Hit Rate Estimation of LessFU

Kenichi Yoshida[a]

[a]*University of Tsukuba, Ostuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan*

Abstract

We have developed a network monitor which can find IP packets sent by Internet Virus from Internet backbone traffic. A data mining engine which can handle 10M transactions per second is the main component of the monitor. Although the data mining engine have to analyze over 200G byte data in theory, a memory management strategy named LessFU removes non-essential data to realize efficient processing. Our past experiments which use real Internet traffic shows the advantage of our approach. However, there exits no method to evaluate the cache hit rate of LessFU. Since the cache hit rate results in serious consequences on the data mining results, this paper proposes a method to estimate the cache hit rate of LessFU. The experimental results which show the advantage of the proposed method are also reported in this paper.

Peer-review under responsibility of ISEL – Instituto Superior de Engenharia de Lisboa, Lisbon, PORTUGAL.

Keywords: Big Data, Data Mining, Memory management, Zipf's law, Bloom Filter.

1. Introduction

"Big Data" and "Data Mining" are vogue buzz words. Although their definitions vary among researchers, we recognize the importance of the various data mining techniques for big data. Among such techniques, we are studying the memory management efficiency of data mining techniques since memory management efficiency is important in the analysis of big data.

Particularly, we have developed a network analyzer [1] that can find IP packets sent by undesirable applications such as Internet viruses and distributed denial of service (DDoS) software from Internet backbone traffic. A data mining engine that can handle 10 M transactions per second is the main component of the analyzer. Although data mining engines theoretically have to analyze over 200 GB of data, a memory management strategy named LessFU (Less Frequently Used) removes nonessential data to realize efficient processing. A spam filter, which we developed in a related study [2], also uses the same memory management strategy.

* Kenichi Yoshida Tel.: +81-3-3942-6896 ; fax: +81-3-3942-6829.
E-mail address: yoshida@gssm.otsuka.tsukuba.ac.jp

Peer-review under responsibility of ISEL – Instituto Superior de Engenharia de Lisboa, Lisbon, PORTUGAL.
doi:10.1016/j.protcy.2014.10.214

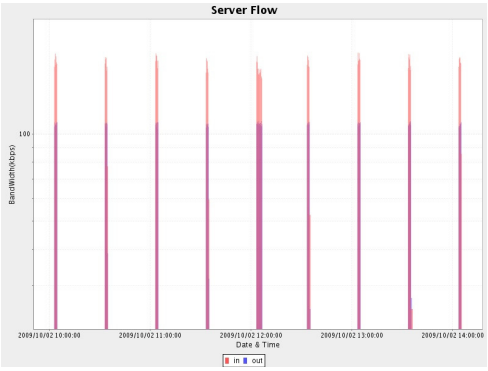

Fig. 1. Output of network analyzer

Our past experiments, which used real Internet traffic, show the advantage of our approach [1,2]. However, there exists no method to evaluate the cache hit rate of LessFU. Because the cache hit rate results in serious consequences on the data mining results, this paper proposes a method to estimate the cache hit rate of LessFU.

The rest of this paper is organized as follows. Section 2 summarizes the related studies. Section 3 proposes a mechanism to estimate the cache hit rate of LessFU. Section 4 reports on the experimental results, and Section 5 summarizes our findings.

2. Related Works

2.1. Data size of network analysis

Figure 1 shows an example of the outputs produced by our network analyzer. It shows that there exists a specific node that sends packets to more than 1000 destinations every 30 min. The node sends only a single UDP packet to each destination and receives a single UDP acknowledgment from each destination. After receiving the acknowledgments, the node does not send any packet for 30 min. Because this is a typical characteristic of the keep-alive behavior of Botnet, finding this type of strange packet flow is important to manage networks. [1]

Because the size of data to be analyzed is so huge, the finding of this type of hidden Botnets is difficult. For example, to find the flow shown in Figure 1, a network analyzer must analyze at least 216 GB of data. That is, a 10 Gbps network can send 10 M packets per second. To find nodes that send packets to various destinations, the network analyzer must analyze a 4 byte destination IP address, a 2 byte destination port number, a 4 byte source IP address, and a 2 byte source port number. Because the behavior of Botnets changes every 30 min, the network analyzer has to store the information of at least a 30 min period. Thus, the total data size becomes 216 Gbyte, i.e., 10 M transactions every 1800 s and 12 byte.

2.2. DRAM limitation

The difficulty associated with the network analysis is the required processing speed, i.e., 10 M transactions per second. To find packets sent by an Internet virus, DDoS software, and other undesirable Internet applications, a similar scale analysis is required. Because the analysis over 10 M transactions per second with 216 GB of data requires unfeasible CPU and memory resources, we developed LessFU [1], which removes nonessential data.

Note that the random I/O performance of DRAM systems is not very fast. For example, our network analyzer requires at least 15 data renewals, i.e., random memory read and write operations, per transaction. Thus, 10 M

[1] Botnets have the tendency to remain silent in order to hide their existence. They then become suddenly active for some specific purpose. Botnets used for DDoS attacks are a typical example of such Botnets. To prevent such undesirable use of a network, it is important to find this type of hidden Botnets.