A digital forensic model based on data mining

Peng Cheng [1, a], Hui Qu [2, b]

[1]Training Department, Engineering College of CAPF, Xi'an 710086, China

[2]Faculty of Science, Engineering College of CAPF, Xi'an 710086, China

[a]cpskull@163.com, [b]qh3798150@hotmail.com

Keywords: A digital forensic model; Data mining; Reliability.

Abstract. The paper designed a digital forensic model and Data mining is applied in the data analysis of the model. In order to meet rapid forensics needs in the era of computing era and to deal with the effectiveness, usefulness, depth issues and real-time and reliability problems according to the research of digital forensics at home and abroad, and the procedure of digital forensics.

Introduction

The development of technologies of computer and Internet brings us a great convenience. However, it also brings us unexpected negative impact. New types of crimes that are both aided by computer and aimed at computer are rising up. Computer crime has become an urgent problem for police and law enforcement agencies throughout the world. In this circumstance, digital forensics is emerging. Usually, the amount of original data, which is collected from so many sources and in different file formats, is massive. So, effective methods are needed to solve these numerous electronic data. Data mining is the very effective method. It can extract interest information from numerous electronic data.

The Classification of Data Mining

Data mining is a process as well as a knowledge discovery in database, namely, it is from the large, incomplete, noisy, fuzzy and random data to extract implicit information that people do not know in advance, but it is potentially useful information and it is a non- trivial process for knowledge. Data mining is originated from many disciplines, including database, artificial intelligence, statistics, machine learning, etc. Among them, the most important three fields are database, machine learning and statistics. These different historical influences made the different scholars hold different views on the function of data mining.

Data mining is related to many disciplines and methods, therefore, there are data a variety of classification methods for data mining. According to the task of data mining, it can be divided into classification or warning model discovery, data glamorization, clustering, association rules discovery, sequence pattern discovery and dependency relation or the dependent model discovery, exception discovery and trend discovery, etc; according to the object of data mining, it can be including the relational database, object-oriented database, spatial database, temporal database, text database, multimedia database heterogeneous database, heritage database and Web, etc; according to the method, it can be divided into the machine learning method, statistical method, neural network method and database method. While machine learning methods can be divided into inductive learning methods (decision trees, induction of rules, etc.) based on the case study, active learning, and genetic algorithms. Statistical analysis methods can be divided into regression (multivariate regression and aggressiveness regression, etc.), discriminant analysis (Bayesian discriminating, Fischer discriminant, non-parametric discriminant, etc.), cluster analysis (hierarchical clustering, clustering segmentation, etc), exploratory analysis (principal component analysis, correlation analysis, etc.) and so on. The artificial neural network method can be divided into feed forward neural networks (BP algorithm), self-organizing neural network (self-organizing feature map, competitive learning, etc.) The database method mainly includes the multidimensional data analysis, attribute-oriented induction method, etc.

Digital forensics. Digital forensics is a branch of forensic disciplines, which includes all of the work process of obtaining evidence from electronic devices and analysis of them. Digital forensics have been derived from the term of computer forensics, and its oriented object is electronic devices but not just the computer, which is due to law enforcement authorities find in today's environment that people use a variety of electronic devices. So the crime is not just computer-related equipment, particularly in the current popularity of intelligent terminals, the source object and criminal offenses are increasing. During the study of digital forensics, many experts has described its concept, but what now widely recognized by everyone, is still Fretsaws definition for digital forensics:" Digital forensics is a process of saving, mobile phones, validation, identification, analysis, interpretation, archiving and reasoning to the data of digital devices with the derived and proven methods, which ultimately facilitates that the investigators reconstruct the chain of evidence and criminal procedure.

There are many current network forensics methods, and following is a brief introduction of three commonly used network forensics methods:

IDS forensics methods. In 2000, Peter Stephenson proposed that intrusion detection systems can be used in forensics environment. Of course, the idea is not only respected by him, other scholars also very keen on this attempt: the evidence collection and system protection are combined to get more information about the evidence. At the same time this approach can solve real-time and continuity problems of acquirement of evidence. Meanwhile, Ado Payer completes real-time forensics model in his research work based on intrusion detection and proposed the ideas that "make sure to save the data in its original format," and "reconstruct of events scene, provide an analysis of the cases of a good start".

Network monitoring forensics method. Network monitoring method refers to the deployment of a network monitoring system in the network, typically formed by several modules, such as monitoring node, data storage centers and data analysis center. When the abnormal situation happens in the network, the monitoring node will report these abnormalities to data center and stores these abnormal flows, and gradually the original evidence base is formed. Chen Zhen gives a unified threat-based and traffic detection methods network forensics approach in the paper. Meanwhile, the approach has higher efficiency and better results after verification of experiment in terms of fighting the distributed attacks.

(SVM) forensics method. Support vector machine is one of the classic feature classification methods. It is mainly used in the feature selection of data packet in the intrusion detection, independent feature filtration and identification to abnormal behavior. Because of its efficiency and accuracy in the nonlinear classification issues, it is the one of the best methods in the intrusion detection and network forensics.

The design and implementation of digital forensics analysis. This paper presents a distributed network traffic forensics model based on Hardtop. According to the function, the system is divided into traffic acquisition layer, traffic analysis layer and storage layer of evidence. The following gives a detailed description for each layer, and proposes an improved algorithm based on past distributions algorithm of type SVM, and verifies and evaluates its effectiveness.

Throughout the model, the first layer is the traffic acquisition layer, whose main goal is to detect the target network using network traffic monitoring tool, and to collect traffic data in real-time which is stored as formatted data so that its convenience of sampling, analysis and archive are improved. In this layer, how we can get real steady flow is very important, because the target traffic to obtain evidence is the first step to obtain evidence sources, which requires to obtain and store in standard way, and also to get legal recognition. The flow got by traffic acquisition layer is directly uploaded to the Hardtop HDFS (Hardtop Distributed File system (HDFS) for storage. As shown in fig 1.

Experimental results and analysis

Experiment on a platform carried out, and the number of training data set were 10,000, 20,000. Compare the result with the stand-alone operation SVM training, 30000 and 20000. The experimental data are shown in Table 1 below:

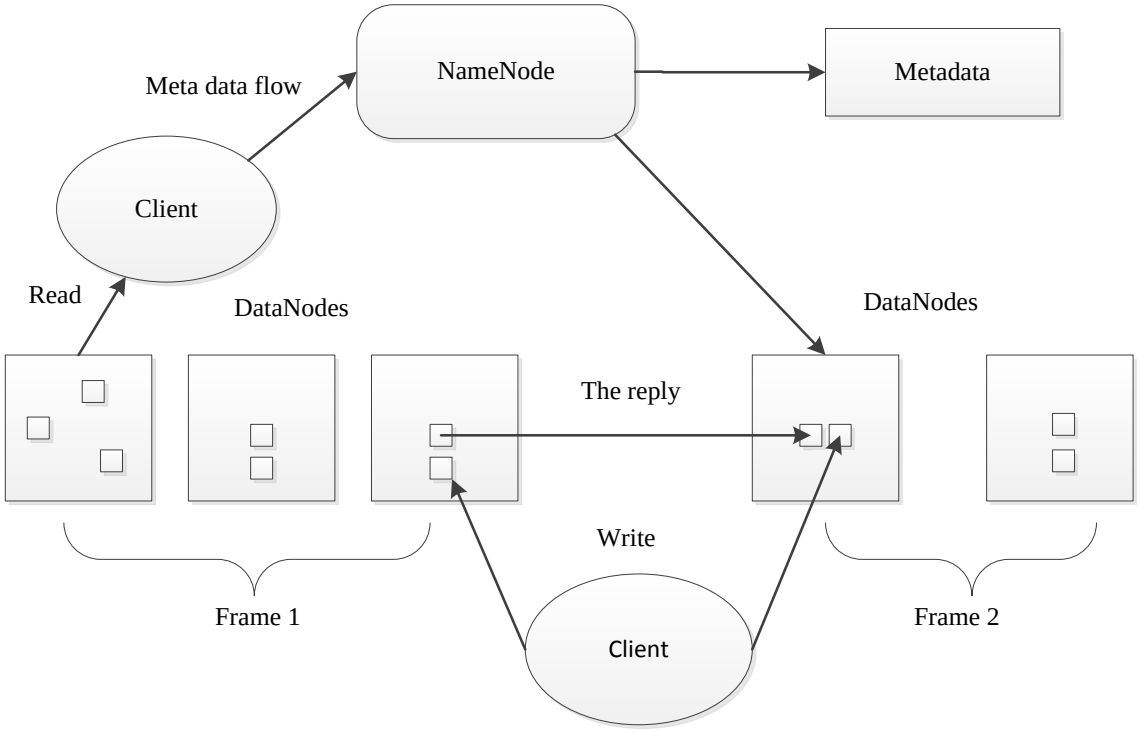

Fig 1. HPFS FRAMEWORK

Table 1 Experimental dates

The number of training data	10000	20000	30000	40000
The time on the experimental platform	9s	15s	22s	39s
Single SVM time-consuming	14s	44s	99s	176s
Platform accuracy rate	79.6%	81.1%	82.5%	84.6%
Stand-alone operation accuracy rate	83.2%	83.5%	85.4%	87.3%

After the understanding of the general process and the core idea of digital forensics and conduction of a more in-depth study, this article will use cloud computing technology in the digital forensics model. And from Hardtop version, distributed SVM algorithm is improved and each layer of cloud computing technology models have this good combination with the cloud computing technology. As an emerging technology, the ability of cloud computing to solve performance bottlenecks cannot be ignored. Meanwhile, the popularity of computer and its lower prices makes the cost of establishing a Hardtop cluster reduced. The algorithms with a way of improving the performance to make up

traditional machine learning is insufficient in the face of massive amounts of data, which is an effective way to renew vitality making these classic algorithms.

Conclusion

Digital forensics is an application-oriented topic, whose research has always revolved the needs of forensic work, so a good performance, high accuracy and stability are the three basic tasks needed attention. Combining data mining with digital forensics solves to find a new solution to many questions. The study results of this paper will provide new research methods and research perspective for electronic forensics researchers.

Reference

[1] Jie Ding . A model of intrusion detection forensics design [J]. Microcomputer development.vol.8. p:117-119.2004.

[2] Chang-yu Liang ,Distributed computer dynamic forensics model [J]. computer application.vol.6. p:1290-1293.2005.

[3] Carne M. An Historical Perspective of Digital Evidence: A Forensic Scientist' s View .International Journal of Digital Evidence. Springer Press. p: 5-7.2002.

[4] James R. Lyle .International Journal of Digital Evidence. NIST CFTT: Testing Disk Imaging Tools. Winter . Vol. 1 .2003.