A LITERATURE SURVEY ON AN EFFICIENT TRAFFIC MONITORING AND IDENTIFICATION BASED ON CATEGORIZATION IN INTERNET ACCESS

Submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY
in
Computer Science and Engineering

by

SEELAM GOYAL -18K91A05J8
Y SAIKIRANKUMAR REDDY -18K91A05P0
VADTHYAVATH RAMU -18K91A05M1
SAACHI JAISWAL -18K91A05J5

Under the guidance of

DR.S.A.KALAISELVAN
Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(ACCREDITED BY NBA AND NAAC WITH 'A' GRADE)
Medbowli, Meerpet, Saroornagar, Hyderabad-500097

CERTIFICATE

This is to certify that literature survey report entitled **A LITERATURE SURVEY ON AN EFFICIENT TRAFFIC MONITORING AND IDENTIFICATION BASED ON CATEGORIZATION IN INTERNET ACCESS.**, being submitted by Mr.SEELAM GOYAL bearing Hall Ticket Number: **18K91A05J8**, Mr.YELLANKI SAIKIRAN KUMAR REDDY bearing Hall Ticket Number: **18K91A05P0**,Mr.VADTHYAVATH RAMU bearing Hall Ticket Number: **18K91A0M1**,Ms.SAACHI JAISWAL bearing Hall Ticket Number: **18K91A05J5**, in partial fulfillment of requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering, to the TKR College of Engineering and Technology is a record of bonafide work carried out by him/her under my guidance and supervision.

Signature of the Guide
Dr.S.A.Kalaiselvan
Professor

Signature of the HoD
Dr.A.Suresh Rao
Professor

Place: Meerpet

Date:

TABLE OF CONTENTS

Chapter 1

LITERATURE REVIEW

1.1 Review of Literature

1.2 Evaluating visualisation approaches to detect abnormal activities in network traffic data

AUTHORS:Dong Hyun Jeong,Soo-Yeon Ji,Bong-Keun Jeong

In this paper, we examined various visualisation approaches used to identify abnormal network activities.Based on the literature review, we identified four common visualisation techniques: data filtration and transformation, pixel-based visualisation, graph representation, and coordinated multi- views.Data filtration is a simple technique that focuses on extracting portions of data or attributes with userspecified criteria.Data transformation utilises various approaches such as statistical analysis, machine learning, and data min- ing to extract transformed data.Pixel-based representation is a popular technique that supports the representation of data elements as coloured pixels.Graph representations is used to support the analysis of network traffic patterns and determine anomalous behaviours. Coordinated multi views enable users to analyse data from different perspectives by creating tightly integrated multiple views

1.3 Network Traffic Classification Techniques-A Review

AUTHORS:Yoga Durgadevi Goli, Dr. Ambika R

In this paper we discuss few techniques for network traffic classification. Three types of methods are available for network traffic classification: port- based, payload-based and machine learning-based.In port based techniques network traffic identification can be done based on the port numbers.In these techniques, many applications use the port number assigned by (IANA) Internet Assigned Numbers Authority.Unfortunately, these techniques suffers from the following disadvantages.As the applications are growing, they may use unpredictable port numbers.These techniques may not be suitable when some applications didn't registered their port numbers with IANA.In payload based technique, which is also

known as Deep Packet Inspection (DPI), the contents of the packets are observed by referring characteristic signatures of network applications in the traffic.Although these techniques provide accurate results, it has some disadvantages.Some of the limitations of port-based and payload- based techniques can be avoided by using Machine learningbased methods.Using machine learning techniques for traffic classification leads to reduce computational costs and identify encrypted traffic easily.Applying Machine Learning Techniques for Traffic classification involves a number of steps that need to be followed: First, the features need to be identified. These features are required to identify and differentiate unknown traffic.The next step is to train the Machine Language classifier by associating the set of features with known traffic classes.Then, a machine learning algorithm can be used to classify unknown traffic using previous training.

1.4 On Internet Traffic Classification: A Two-Phased Machine Learning Approach

AUTHORS:Taimur Bakhshi and Bogdan Ghita

Traffic classification serves as a fundamental requirement for network operators to differentiate and prioritise traffic for a number of purposes, from guaranteeing quality of service to anomaly detection and even profiling user resource requirements. To increase the flow classification accuracy, cascaded classification methodologies employing a combination of algorithms as well as semisupervised ML approaches have also been previously explored.ML tool used in traffic classification studies is Weka [37], incorporating a library (Java based) of supervised and unsupervised classifiers which can be readily implemented on test data set to evaluate the accuracy of the results from each methodology.Weka machine learning software suite (version 3.6.13) was employed to evaluate the eight most commonly utilised supervised machine learning algorithms in comparison with the proposed two-phased approach. The comparison evaluated (i) the classification accuracy of each algorithm and (ii) the computational overhead including the training and testing times to validate the results from each classification technique as well as (iii) provide perspectives on the scalability of our two-phased machine learning classifier.ML algorithms, unsupervised and supervised, were combined and the scheme used a probabilistic assignment during unsupervised cluster analysis to associated clusters with traffic labels.A high level overview

of the traffic classification scheme is shown (i) Preprocessing. Internet traffic is collected from end- user machines and marked with application labels accordingly. (ii) Cluster Analysis. Using unsupervised -means, flows belonging to individual applications are separately cluster analysed to extract unique subclasses per application, offering a finer granularity of the classification e.g., YouTube and Netflix flows would be classed as streaming and browsing. (iii) Classifier Training. Flows marked with their -means clusters, indicating the subclass they belong to, are afterwards fed to a C5.0 classifier for supervised training, leading to a decision tree. (iv) Evaluation. A separate data set is used for testing the accuracy of the algorithm.

1.5 A Review of Current Machine Learning Approaches for Anomaly Detection in Network Traffic

AUTHORS: Wasim A. Ali,Malika Bendechache,Mohammed Fadhel Aljunid

A computer network is a combination of many individual entities assembled together to provide complete and various communication services.Anomalies in these networks are also called abnormalities, outliers, or exceptions.Anomaly detection system is an automated security system used for monitoring, analysing, and detecting abnormal activities within a network or host.In this paper, we will investigate different types of attacks handled by using supervised, unsupervised and semi-supervised algorithms.Applying supervised techniques on the network data sets allows us to build a model, and the data instances can be labelled using a set of attributes. Many supervised algorithms are used to detect anomalies and intrusions in the network traffic and have proven effectiveness and efficiency, such as Support Vector Machine(SVM), Artificial Neural Network(ANN), Nearest Neighbour algorithm, Decision Trees, Knearest neighbour, Ensembles classifiers, and Naïve Bayes classifier.Unsupervised Network Detection Systems (NDS)are used to overcome the limitation of the supervised anomaly techniques system.There are many unsupervised algorithms used to cluster given data and detect anomalous/ abnormal activities in network traffic successfully like the K-means algorithm, Hidden Markov Model(HMM), Gaussian Mixture, Hierarchical clustering, and Neural Networks (NNs).Semi-supervised machine learning could be a combination of supervised and unsupervised machine learning approaches.

1.6 A Summary of Network Traffic Monitoring and Analysis Techniques

AUTHOR:Alisha Cecil

Network analysis is the process of capturing network traffic and inspecting it closely to determine what is happening on the network.Two Monitoring Techniques are discussed in this paper: Router Based and Non-Router Based.Router Based Monitoring Techniques are hard-coded into the routers and therefore offer little flexibility.The most commonly used monitoring techniques are Simple Network Monitoring Protocol (SNMP),Remote Monitoring (RMON) ,Netflow. Non-Router Based Monitoring Technique are limited in there abilities they do offer more flexibility than the router based techniques. These techniques are classified as either active or passive.Active monitoring [Active06] transmits probes into the network to collect measurements between at least two endpoints in the network.Passive monitoring [Curtis00] unlike active monitoring does not inject traffic into the network or modify the traffic that is already on the network.Although passive monitoring does not have the overhead that active monitoring has, it has its own set of downfalls.After reading this paper one can safely come to the conclusion that a combination of active and passive monitoring is better than using one or the other.Being able to monitor and analyse networks is vital in the job of Network Administrators.

1.7 Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms

Author: Muhammad Shafiq

Abstract-Network Traffic Classification is a central topic nowadays in the field of computer science. It is a very essential task for internet service providers (ISPs) to know which types of network applications flow in a network. Network Traffic Classification is the first step to analyze and identify different types of applications flowing in a network. Through this technique, internet service providers or network operators can manage the overall performance of a network. There are many methods traditional techniq ue to classify internet traffic like Port Based, Pay Load Based and Machine Learning Based technique. The most common technique used these days is Machine Learning (ML) technique. Which is used by many researchers and got very effective accuracy results. In this paper, we discuss network

traffic classification techniques step by step and real time internet data set is develop using network traffic capture tool, after that feature extraction tool is use to extract features from the capture traffic and then four machine learning classifiers Support Vector Machine, C4.5 decision tree, Naive Bays and Bayes Net classifiers are applied. Experimental analysis shows that C4.5 classifiers gives very good accuracy result as compare to other classifies.

1.8 Internet Traffic Identification Using Machine Learning

AUTHORS: Conference Papers

We apply an unsupervised machine learning ap- proach for Internet traffic identification and compare the results with that of a previously applied supervised machine learning approach. Our unsupervised approach uses an Expectation Max- imization (EM) based clustering algorithm and the supervised approach uses the Na ıve Bayes classifier. We find the unsupervised clustering technique has an accuracy up to 91outperform the supervised technique by up to 9that the unsupervised technique can be used to discover traffic from previously unknown applications and has the potential to become an excellent tool for exploring Internet traffic.

1.9 Machine Learning-Based Framework For Automatic Malware Detection Using Andriod Traffic Data

AUTHORS:UZOMA RITA ALO1,*, HENRY FRIDAY NWEKE2 , SYLVESTER I. ELE1, 3

One of the greatest challenges facing various organizations and institutions is information security. Attackers have devised means to steals mobile user identity by developing malware that might be inadvertently installed by users due to the open source nature of android operating system causing financial loses. Although various machine learning algorithms have been proposed recently for malware detection, it is challenging to detection malicious apps with single classification model. In this paper, we propose to detect malicious apps in android traffic using four (4) different machine learning algorithms. The proposed approach was evaluated on comprehensive and publicly available dataset. The result obtained shows that decision tree and tree based ensemble algorithms produced superior results when compared with support vector machine and logistic regression models. The results suggest the impact of multiple classification algorithms to improve the performance of malware detection

system. The finding can be utilized to guide security expert on the use of machine learning methods to detect malicious software.

1.10 A machine learning based framework for IoT device identification and abnormal traffic detection

AUTHORS : Ola Salman Imad H. Elhajj Ali Chehab Ayman Kayssi

Network security is a key challenge for the deployment of Internet of Things(IoT). New attacks have been developed to exploit the vulnerabilities of IoTdevices. Moreover, IoT immense scale will amplify traditional network attacks.Machine learning has been extensively applied for traffic classification andintrusion detection. In this paper, we propose a framework, specifically for IoTdevices identification and malicious traffic detection. Pushing the intelligence to the network edge, this framework extracts features per network flow to identify the source, the type of the generated traffic, and to detect network attacks. Differ- ent machine learning algorithms are compared with random forest, which gives the best results: Up to 94.5 accuracy for device-type identification, up to 93.5 accuracy for traffic-type classification, and up to 97 accuracy for abnormal traffic detection.

1.11 Performance evaluation of Internet traffic by network measurements

AUTHORS: Georgi P. Georgiev

A review of the main methods for measurement Internet traffic is made. Some of software platforms for network measurements are discussed. The reasons for the need of measurement and monitoring of traffic in IP-based networks are: optimization and network planning, quality assurance of services and detect security breaches. Internet traffic is heterogeneous and highly bursty. The trial network is a LAN, and serves two households. A measurement of the load on the network for a certain period with different reporting intervals is made. The change of network traffic also has been measured. It has been done a distribution by application layer protocols and by size of the packets. It is confirmed that the traffic is heterogeneous and highly bursty. The main protocols are UDP, from which we can conclude that the network is mainly used for transmission of multimedia. It is measured the size of the transmitted packets and it is found that the quantity of useful information transmitted is equal to the quantity of transmitted service information. Through software approximation is made

relating to the size of the package. With the expansion of modern IP-based networks, monitoring and measurement of traffic on them are becoming increasingly necessary.

1.12 In-depth Analysis of Internet Traffic

Author: Zoltán Móczár

One of the key components of today's Internet is the congestion control, which is implemented in TCP (Transmission Control Protocol). In the last decades, numerous versions of this protocol have been worked out to realize efficient congestion control in high-speed networks and lossy environments as well. Unfortunately, none of these approaches were entirely successful, and the researchers have shown that there is a need for a new paradigm. The Internet of the future requires inventing new network devices and solutions, which need careful design, but it is not possible to carry out without the deep knowledge of the traffic behavior. In addition, traffic analysis can also be used for many other purposes such as detecting distributed attacks, analyzing user activities and traffic-based accounting services.This thesis deals with the processing and analysis of traces originated from an ISP (Internet Service Provider). First of all, the literature background and the architecture of the network are reviewed, and then the results of the flow-level analysis are presented.The distribution functions and histograms of the flows as well as the relation between the most important traffic parameters are examined. In the following section flows are separated into groups based on three different dimensions (size, duration and rate). The characteristics of these categories and the connection between them are revealed. The features of the traffic generated by the applications and their impact on the aggregated traffic are also investigated.

1.13 Internet Traffic Classification Tool

AUTHORS : Muhammad Kamran Muhammad Arshad,Hamza Khalid

This research paper regards with the tool development for the internet traffic classification, discussing different types of data packets and different techniques to analyze data packets. For instance, it discusses DNS, VOIP, http/s, RTP and FTP types of data packets and throws light on port-based approach, payload-based approach, host-behavior based approach and flow feature-based approach to identify the Packer type and analyze the packets further. Moreover, literature is reviewed regarding Wireshark; an online free licensed project which

does the same work as the tool which this research regards with. Then the methodology is discussed after differentiating TCP Dump and WinDump. Furthermore, time bin is discussed in detail and specified for the tool which is under discussion in this paper. In addition, visuals and image references are also provided to clarify the complex concepts regarding the under-discussion tool. Taking into account, all of the mentioned and discussed factors, one can conclude that this research paper is extremely useful for those who want to explore the world of Computer Networks

1.14 Internet Traffic Classification with Federated Learning

AUTHORS:Hyunsu Mun and Youngseok Lee As Internet traffic classification is a typical problem for ISPs or mobile carriers, there have been a lot of studies based on statistical packet header information, deep packet inspection, or machine learning. Due to recent advances in end-to-end encryption and dynamic port policies, machine or deep learning has been an essential key to improve the accuracy of packet classification. In addition, ISPs or mobile carriers should carefully deal with the privacy issue while collecting user packets for accounting or security. The recent development of distributed machine learning, called federated learning, collaboratively carries out machine learning jobs on the clients without uploading data to a central server. Although federated learning provides an on-device learning framework towards user privacy protection, its feasibility and performance of Internet traffic classification have not been fully examined. In this paper, we propose a federated-learning traffic classification protocol (FLIC), which can achieve an accuracy comparable to centralized deep learning for Internet application identification without privacy leakage.

1.15 Internet Access Traffic Measurement and Analysis

AUTHORS: Steffen Gebert, Rastin Pries, Daniel Schlosser Klaus Heck The fast changing application types and their behavior require consecutive measurements of access networks. In this paper, we present the results of a 14-day measurement in an access network connecting 600 users with the Internet. Our application classification reveals a trend back to HTTP traffic, underlines the immense usage of flash videos, and unveils a participant of a Botnet. In addition, flow and user statistics are presented, which resulting traffic models can

be used for simulation and emulation of access networks.

1.16 An Efficient Internet Traffic Classification System Using Deep Learning for IoT

AUTHORS: Muhammad Basit Umair, Zeshan Iqbal, Muhammad Bilal, Tarik Adnan Almohamad, Jamel Nebhen, Raja Majid Mehmood

A network of devices connected to the internet and sharing a massive amount of data between each other and a central location. These IoT devices are connected to a network therefore prone to attacks. Various management tasks and network operations such as security, intrusion detection, Quality-of-Service provisioning, performance monitoring, resource provisioning, and traffic engineering require traffic classification. Due to the ineffectiveness of traditional classification schemes, such as port-based and payload-based methods, researchers proposed machine learning-based traffic classification systems based on shallow neural networks. Furthermore, machine learning-based models incline to misclassify internet traffic due to improper feature selection. In this research, an efficient multilayer deep learning based classification system is presented to overcome these challenges that can classify internet traffic. To examine the performance of the proposed technique, Moore-dataset is used for training the classifier. The proposed scheme takes the pre-processed data and extracts the flow features using a deep neural network (DNN). In particular, the maximum entropy classifier is used to classify the internet traffic. The experimental results show that the proposed hybrid deep learning algorithm is effective and achieved high accuracy for internet traffic classification, i.e., 99.23. Furthermore, the proposed algorithm achieved the highest accuracy compared to the support vector machine (SVM) based classification technique and k-nearest neighbours (KNNs) based classification technique.

1.17 A Survey of Network Traffic Monitoring and Analysis Tools

AUTHORS: Chakchai So-In

From hundreds to thousands of computers, hubs to switched networks, and Ethernet to either ATM or 10Gbps Ethernet, administrators need more sophisticated network traffic monitoring and analysis tools in order to deal with the increase. These tools are needed, not only to fix network problems on time, but also to prevent network failure, to detect inside and outside

threats, and make good decisions for network planning. This paper surveys all possible network traffic monitoring and analysis tools in non-profit and commercial areas. The tools are categorized in three categories based on data acquisition methods: network traffic flow from NetFlow-like network devices and SNMP, and local traffic flow by packet sniffer. The popular tools for each category and their main features and operating system compatibilities are discussed. The feature comparisons on each category are also made.

1.18 An Internet Traffic Analysis Method with MapReduc

AUTHORS: Youngseok Lee,Wonchul Kang, Hyeongu Son

Internet traffic measurement and analysis have been usually performed on a high performance server that collects and examines packet or flow traces. However, when we monitor a large volume of traffic data for detailed statistics, a longperiod or a large-scale network, it is not easy to handle Tera or Peta-byte traffic data with a single server. Common ways to reduce a large volume of continuously monitored traffic data are packet sampling or flow aggregation that results in coarse traffic statistics. As distributed parallel processing schemes have been recently developed due to the cloud computing platform and the cluster filesystem, they could be usefully applied to analyzing big traffic data. Thus, in this paper, we propose an Internet flow analysis method based on the MapReduce software framework of the cloud computing platform for a large-scale network. From the experiments with an open-source MapReduce system, Hadoop, we have verified that the MapReduce-based flow analysis method improves the flow statistics computation time by 72, when compared with the popular flow data processing tool, flow-tools, on a single host. In addition, we showed that MapReduce-based programs complete the flow analysis job against a single node failure.

1.19 Internet Traffic Measurement

AUTHORS: Carey Williamson

The Internet is simply a connection to these applications. They are shielded from the details of how the Internet works, through the information-hiding principles of the Internet protocolstack, which dictates how user-level data is transformed into network packets for transport across the network and put back together for delivery at the receiving application. For many networking researchers, however, the protocols themselves, rather than the

information they carry, are of interest. Using specialized network measurement hardware or software, these researchers collect information about network packet transmissions, including their timing structure and contents. With detailed packet-level measurements and some knowledge of the IP stack, they can use reverse engineering to gather significant information about both the application structure and user behavior,which can be applied to a variety of tasks like network troubleshooting, protocol debugging, workload characterization, and performance evaluation and improvement.

1.20 Internet Traffic Classification And Recognition With Statistical Approach And Intrusion Detection

AUTHORS :Conference paper

The machine learning (ML) technique is based on the labeled dataset. In this technique, a machine learning classifier is trained as input, and then using the trained sample prediction, unknown classes are classified. I first establish a reference standard performance for five classifiers (Random Forest, AdaBoost.M1, C 4.5, MLP, and SVM) using publicly available network traffic NSL KDD datasets. The features are calculated using the wrapper method and then the ML classifier is trained with these features with known traffic classes and creates the classifier model known as the Memorization process. This model is then used to classify unknown traffic known as testing or Generalization process. Five ML algorithms are used for IP traffic classification with mentioned datasets and also intrusions are detected for the same datasets. Finally, performance analysis is done with the help of several evaluation metrics.

1.21 Traffic Classification Prioritization

AUTHORS:Conference paper

Traffic classification is useful for traffic engineering and network security. Network administrators can use it to allocate, control and manage the network resources as per their requirements. Classification methods can be used to classify P2P traffic, encrypted traffic, web, streaming, download or any specific application. Our classification model classifies traffic into two classes, i.e., multimedia and download. We used supervised machine learning algorithms (Decision Tree and K-NN) to build the classification model. This model is trained using pre-labeled training instances and later used to classify the traffic in real-time. We use

packet level statistics (average packet size, average inter-arrival time, receiver's window size, flow duration etc.) as features for classification algorithms. Prioritization module ensures that once the flow is identified as multimedia it will get higher priority over the download flows. We used HTB (Hierarchical Token Bucket Filter) for this purpose. We have also developed heuristics that can automatically label the training data set with some manual inputs, i.e. labeling each flow in the data set as either multimedia or download. These heuristics look at URI of HTTP GET request and search for multimedia file formats in it, if found then it labels that flow as multimedia.