Proc. of the Intl. Conf. on Advances in Computing and Communication – ICACC 2013

ISBN: 978-981-07-6260-5

A Survey on Internet Traffic Classification

Afzal Hussain Shahid
Department of CSE
National Institute of Technology
Patna, India

M. P. Singh
Department of CSE
National Institute of Technology
Patna, India

Prabhat Kumar
Department of IT
National Institute of Technology
Patna, India

***Abstract*— Internet traffic classification is remained to be important issue for years due to three major reasons. First, the fact that we cannot limit the number of users connected and obviously there would be continual increase in the number of users connected to the Internet with time. Second, increase in user access speed (due to the development of highly efficient hardware and software technology). Third, there are applications that require more network resources. This survey is done to find out the main problems in the field of Internet traffic classification and to acquire the knowledge about different classification techniques that have significant Impact on the scientific and research community.**

Keywords— Application Identification, Traffic Measurement, Classification

I. Introduction

Classification of Internet traffic are useful for various network management activities, such as bandwidth optimization, capacity planning and provisioning, fault diagnosis, traffic engineering, application performance, anomaly detection and pricing.

Broadband connection for Internet users is continually growing particularly those based on Asymmetric Digital Subscriber Line (ADSL) and cable Television infrastructure and technologies. These availabilities have opened the door for new ways of resource usage for small organizations and home users. Since the broadband Internet connection is always available as well as it has increased its quality of service, users are more inclined to use a broad range of services available in the current Internet, such as Voice over Internetworking Protocol , Internet banking and peer-to-peer (P2P) systems for resource sharing, particularly audio and video files. In other words, the increased capacity and availability has led to a complex behavior for a typical user, very different from a dial-up user. Therefore, Internet Service Providers should pay more attention to this more complex behavior. Furthermore, the current trend of moving phone calls from the public switched telephone network (PSTN) to the Internet via voice over IP (VoIP) applications represents a threat to telephony companies and its effects could not be completely understood [11].

Internet Traffic has been classified by different approaches viz. Port-based, Packet payload based but these methods are now become inefficient due to number of reasons discussed in section II. Recently the Machine Learning (ML) techniques (a subset of Artificial Intelligence discipline) are able to better classify the Internet Traffic. Several machine learning algorithms such as Naïve Bayes, RBF, C4.5, Bayes Net , MLP etc are found to be giving better accuracy.

Artificial Neural Networks (one of the machine learning technique) have been successfully used in a number of applications due to two major benefits. First, Neural Networks (NNs) derives computing power through massively parallel distributed structure. Second, generalization, that is production of reasonable output from inputs not encountered during training process. Also the capacity to handle non-linearity, quick adaptability to system dynamics and fault tolerant are very important properties and capabilities of NNs. They can be trained to efficiently recognize patterns of information in the presence of noise and non-linearity and classify information using those patterns. These properties can be exploited to use artificial neural networks in the actively researched field of Internet traffic classification.

Rest of the paper is organized as follows. Section II presents State-of-the-Art in Internet Traffic Classification. Section III presents related work. Section IV presents comparison of Traffic classification methods. Section V presents critical view. Section VI presents some open research challenges still to be answered in the field of Traffic classification. Section VII presents conclusion.

II. State-of-the-Art in Internet Traffic Classification

This section describes the limitation of the conventional port and payload based approach, definition of packet and flow based classification, and defines completeness and accuracy.

A. Limitation of Port based approach

There are basically three reasons why Port based approach is now inefficient in many cases. These are described below.

1. There are applications that have no (Internet Assigned Number Authority) IANA registered ports, but these applications uses ports already registered to other applications. Also ports can be randomly selected, or can be

Proc. of the Intl. Conf. on Advances in Computing and Communication – ICACC 2013

ISBN: 978-981-07-6260-5

defined by user. For example the ports of applications such as Napster and Kazaa are not registered with IANA official list. Also some applications allocates the ports dynamically e.g. The Real Video streamer allows the dynamic negotiation of the server port used for the data transfer.

2. The application designers and users can use well-known ports assigned to other applications to hide their traffic.

3. Encryption at the IP layer may also make obscure the TCP or UDP header, making it impossible to know the actual port numbers.

B. *Limitation of Payload based approach*

Payload-based approach scans packet content to identify byte strings associated with an application. This approach compares packet content (payload) to a set of stored signatures.

1. Privacy policies and laws may prevent access to or archiving of packet content.
2. Encryption will simply make it impossible to get the actual stored signature in the payload.
3. Too much computationally expensive particularly on high-bandwidth links.

C. *Packet based and Flow based classification*

Traffic classification can be done by two ways packet or flow data. In packet based, packet level characteristics or application signature are used to classify captured packets. In flow based there are three ways of flow classification. First, some classifiers detect the application by ports used. Second, predefined flow characteristics are used e.g. connection patterns. Third, machine learning techniques are used.

D. *Traffic Classification Metrics*

Completeness - It is defined as the ratio of the number of flows (bytes) classified over the total number of flows (bytes).
Accuracy – It is the ratio between correct detection and total detection count.

III. Related Work

The various works that are done to classify Internet Traffic are discussed in this section.

Zhou et al. [1] proposes an approach based on feed forward neural network for Internet traffic classification and compared their approach with Naïve Bayes classifier. Naïve Bayes classifier is a probabilistic classifier based on Bayesian theorem and makes assumption that all feature properties are independent and subjected to Gaussian distribution. Feed forward neural network approach eliminates the disadvantages of port-based or payload-based classification methods. After the extensive experimentation and comparison it has been found out that, combined with a fast correlation-based feature selection filter, better performance and more accurate classification results is obtained using neural network method compared to Naïve Bayes method. The input to the neural network is the various statistical feature of the flow. Performance limitation of Naïve Bayes is due to its inappropriate inherent assumptions.

Trivedi et al. [2] uses statistical information (0nly packet size) and does not involve reading any packet headers to determine the application. Their approach is classifying IP traffic into different application types on the basis of packets attributes that can be obtained at OSI layer 3 (IP layer). They do not classify the traffic on per packet basis; rather the traffic flows are classified (that is classifying the traffic over a time frame). The classification with artificial neural networks is done using a conventional feed-forward back propagation network with three layers. The number of nodes in the hidden layer is empirically selected such that the performance function, i.e. the mean square error in case of feed-forward networks, is minimized. They have compared their method with a similar classification done using the classical statistical method of clustering. It is observed that both approaches give highly accurate result, however artificial neural networks giving slightly better results.

Auld et al. [3] presents a Bayesian trained neural network and uses a supervised machine learning to train a classifier. They attained a significantly higher degree of classification accuracy using less information than previous methods. This paper shows that information able to be derived from a traffic flow can allow the classification with an accuracy exceeding that of (standard) port-based mechanism.

Yuan et al. [4] proposes a machine learning method based on SVM (supporting vector machine) for accurate Internet traffic classification. This method classifies the Internet traffic into broad application categories according to the network flow parameters obtained from the packet headers. An optimized feature set is obtained via multiple classifier selection methods. Experimental results using traffic from campus backbone show that an accuracy of 99.42% is achieved with the regular biased training and testing samples. An accuracy of 97.17% is achieved when un-biased training and testing samples are used with the same feature set. Furthermore, as all the feature parameters are computable from the packet headers, the proposed method is also applicable to encrypted network traffic.

Bernaille et al. [5] proposes a technique that relies on the observation of the first five packets of a TCP connection to

Proc. of the Intl. Conf. on Advances in Computing and Communication – ICACC 2013

ISBN: 978-981-07-6260-5

identify the application. This result opens a range of new possibilities for online traffic classification. They use unsupervised clustering to detect a set of flows that share a common behavior.

Karagiannis et al. [6] (a flow based classification) presents a fundamentally different approach to classifying traffic flows according to the applications that generate them. In contrast to previous methods, their approach is based on observing and identifying patterns of host behavior at the transport layer (called connection pattern).Connection patterns are described by graphs, where nodes represent IP address and port pairs and edges represent flow between source and destination nodes. They analyze these patterns at three levels "(i) the social, (ii) the functional and (iii) the application level. This multilevel approach of looking at traffic flow is the most important contribution of this paper. Furthermore, their approach has two important features. First, it operates in the dark, having (a) no access to packet payload, (b) no knowledge of port numbers and (c) no additional information other than what current flow collectors provide. These restrictions respect privacy, technological and practical constraints. Second, it can be tuned to balance the accuracy of the classification versus the number of successfully classified traffic flows".

Szabo et al. [7] proposes a combined method that includes the advantages of different approaches using the gained knowledge about the strengths and weaknesses of the existing approaches in order to provide a high level of classification completeness and accuracy. As a result, the ratio of the unclassified traffic becomes significantly lower. Further, the reliability of the classification improves, as the various methods validate the results of each other. The novel method is tested on several network traces, and it is shown that the proposed solution improves both the completeness and the accuracy of the traffic classification, when compared to existing methods.

Moore et al. [8] proposes a classification methodology that relies on the full packet payload. Their technique is not automated due to the fact that a particular Internet application could be new whose behavior is not yet known or particular application could satisfy more than one classification criterion.

John et al. [9] classifies the Internet Backbone Traffic based on connection pattern present at the transport layer. Instead of looking at individual packets or flows, sequences of flows to or from a specific endpoint are matched with a set of predefined heuristics. These heuristics typically don't require packet payload and could potentially even disregard port numbers. A complete traffic classification can be provided even for short 'snapshot' traces, including identification of attack and malicious traffic. The usefulness of the heuristics is finally shown on a large dataset of backbone traffic, where in the best case only 0.2% of the data is left unclassified.

Mohd et al. [10] uses machine learning algorithms for the classification of traffic and showed that random tree, IBI, IBK, random forest respectively are the top 4 highest accuracy in classifying flow based network traffic to their corresponding application among thirty algorithms with accuracy not less than 99.33%.

IV. Comparison Of Traffic Classification Methods

This section presents the comparison of well known recently proposed methods that performs well.

The evaluation of the Bayesian method [3] is summarized below in Table I. This method gives the best accuracy. This method needs to be trained with a data set that was previously classified then it is tested on a different data set. The authors investigate the accuracy of the approach but do not address completeness. The accuracy of the P2P traffic classification is lower than in case of other applications. This is due to the fact that their main characteristics are difficult to grab.

Application	Bayesian
WWW	99.27%
Mail	94.78%
Bulk Transfer(FTP)	82.25%
Services	63.68%
Database	86.91%
Multimedia	80.75%
P2P	36.45%

Table I. Accuracy of the Bayesian method

The method proposed by [5] also uses machine learning, but the algorithm reads only the first few packet headers in each connection. The accuracy of the classification methods is summed up in Table II. The authors use payload analysis as a reference. According to the results, the on-the-fly method works roughly as accurately as the Bayesian method even though it relies on significantly less and simpler input.

APPLICATION	ON THE FLY
HTTP	99%
HTTPS	81.8%
SMTP	84.4%
POP3	0%
POP3S	89.8%
BULK TRANSFER (FTP)	87%
NNTP	99.6%
KAZAA	95.24%
EDONKEY	84.2%
SSH	96.92%

Table II. Accuracy of ON THE FLY algorithm

The results of the BLINC method [6] are summarized in Table III. The authors [6] develop their own byte signatures and use them for validating the proposed connection pattern based classification method. The BLINC algorithm is able to classify the main application types and it performs better in terms of accuracy than completeness.

Application	Metric	BLINC
WWW	Completeness	69 - 97%
	Accuracy	98 – 100%
Mail	Completeness	78 – 95%

Proc. of the Intl. Conf. on Advances in Computing and Communication – ICACC 2013

ISBN: 978-981-07-6260-5

	Accuracy	85 – 99%
Bulk Transfer (FTP)	Completeness	95%
	Accuracy	98%
Chat	Completeness	68%
	Accuracy	98%
Network Management	Completeness	85 – 97%
	Accuracy	88 – 100%
P2P	Completeness	84 – 90%
	Accuracy	96 – 98%

Table III. Accuracy of BLINC method

All the tables (Table I, Table II, and Table III) above show that the algorithms perform well on the analyzed traces. But all proposed methods use heuristics therefore extra effort is needed to fit the methods to other traces.

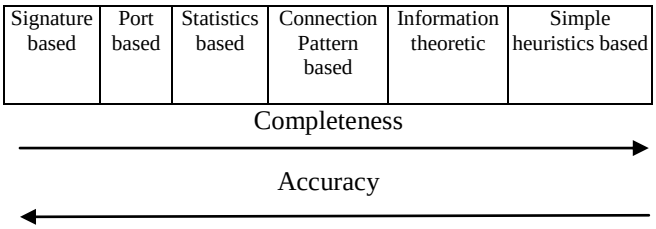
Uncertainty provided by the heuristics could be circumvented by running more algorithms in parallel and compare their result and final application classification decision is based on the result of the comparison, as proposed by [7]. This approach also has the advantage that mismatching classification results are recognized automatically. Also, such traffic may be dumped separately for further analysis and the knowledge gained can be incorporated into the algorithms.

Authors of [7] executed the different classification methods on different measurements one-by-one, and then the results were combined by their suggested decision mechanism. The accuracy and the completeness of the classification methods on different application types compared to their combined classification method. The accuracy and completeness is shown in Table IV.

Application	**Metric**	**PAYLOAD ANALYSIS**
Web	Accuracy	91%
	Completeness	134%
P2P	Accuracy	99%
	Completeness	61%
Chat	Accuracy	97%
	Completeness	76%
E-mail	Accuracy	97%
	Completeness	98%
File Transfer	Accuracy	99%
	Completeness	26%
Streaming	Accuracy	98%
	Completeness	3%
System	Accuracy	95%
	Completeness	75%
Tunneled	Accuracy	10%
	Completeness	120%

Table IV. Accuracy and Completeness of the Signature Based Analysis

The combined application – and the comparison run on the same input data – shows that some methods are stronger in accuracy and others provide more complete results. Strict protocol parsers (payload based approach) are the least complete while heuristic based approaches are the most complete. Regarding the accuracy, the order is reversed, thus signature based approaches (payload based approach) are more accurate than simple heuristic based approaches.



As a result, the traffic classification decision is a trade-off between the accuracy and completeness.

V. Critical View

This section presents the main problem of the various related works that needs to be circumvented in the future.

In [1] (feed forward neural network) the statistical features of different network flow classes may vary with time. Therefore, using data sets acquired at a later time to update the model of classification is expected to enhance the unbiasedness and robustness of the neural network in a long period. In [2] (feed-forward back propagation network) the algorithms used have to be refined to attain real-time classification and adaptation to changing network dynamics. In [3] (Baysian trained neural network) the authors have verified the classification scheme on a single network, and training and testing is done within that network. They made no claims about the stability of the composition of network traffic, or the difference of traffic between sites. Also the reduction of the feature set requires comment. In [4] (Support Vector Machine) authors require large number of training samples which could be reduced when classifying traffic in comparatively stable network. In [5] (Traffic classification on the fly) authors need to analyze the method on a much broader panel of traces. In [6] (BLINC) the method cannot identify specific application sub-types: This technique is capable of identifying the type of an application but may not be able to identify distinct applications. For instance, it can identify p2p flows, but it is unlikely that it can identify the specific p2p protocol (e.g., eMule versus Gnutella) with packet header information alone. In [7] Szabo et al. is giving good accuracy and completeness but accuracy cannot be granted as it is based on heuristics. In [8] Moore et al. uses full packet payload to classify traffic therefore requires much processing of traffic data and hence not suitable for real time classification. In [9] John et al. is again based on heuristics and hence accuracy cannot be granted. In [10] Mohd et al. have not considered the time factor in their work to find the four best machine learning algorithm for Bandwidth optimization.

VI. Some Open Research Challenges

The following are the challenges that are still to be addressed in the field of Internet Traffic Classification.

ISBN: 978-981-07-6260-5

1) From the research point of view, the problem is to find the minimum amount of data that needs to be measured in order to classify applications. However, storing the minimum amount of data may not be the best solution, since additional data may be needed to validate results.

2) Networks carry high traffic volume, which imposes a higher burden on traffic measurements. One way to deal with this problem is to apply sampling or other filtering techniques, e.g. measure the traffic of selected subscribers. It is not clear how much sampling can be used to keep a certain level of accuracy. It is also not clear how much information is lost given a certain sampling approach.

3) All Traffic classification methods leave parts of the traffic unclassified. Therefore a general goal is to reduce the amount of unclassified traffic.

4) Application classification methods use heuristics. As a consequence of that, their validation will always be problematic. Some heuristics are more reliable than others, e.g. a long byte signature is more accurate than a short one.

5) Updating heuristics is time consuming. This motivates the research of methods that can recognize new protocols or changes in protocol versions of the same application type automatically.

6) New applications sometimes impose changes in network capacity and management (e.g., Voice over IP usage, Video on Demand, Internet Worms). Network operations cannot presently promptly react to traffic changes caused by new applications. Could there be a way around it.

7) Since broadband users tend to stay connected longer hours and sometimes even let their network access active just for connectivity (e.g., over-night downloads, VoIP applications waiting for a call), it is important to know the current typical traffic profile for these users. Is there a traffic profile trend towards the greater usage of certain applications? How do these applications behave and how will they behave with the growing number of users?

8) The literature exhibits a wide range of inconsistent terminology to describe approaches and metrics, making it difficult or impossible to compare studies or safely infer conclusions.

Any answers of these questions would greatly improve the way network managers operate their networks and adapt to changes.

VII. Conclusion

Many techniques have been proposed for Internet Traffic Classification but none of them could be final answer for Traffic Classification. Despite being the fastest and simplest, port-based approach is no longer relevant, since many applications, particularly those with a high network volume (e.g., P2P file sharing), bypass the rules and use known ports of other services. After that Payload-based approaches emerged, are very time-consuming, therefore cannot be utilized in high-speed links. Payload approach is considered to be reliable technique for Internet traffic classification, but it has privacy issues. Privacy policies and laws may prevent access to or archiving of packet content. It is easily circumvented by encryption, protocol obfuscation or encapsulation (e.g., tunneling traffic in HTTP), and prohibitively computationally expensive for general use on high-bandwidth links. Nowadays algorithms from the pattern recognition field using machine learning techniques have proven promising, especially in the face of obfuscated and encrypted traffic which rule out payload analysis. These systems learn from empirical data to automatically associate objects with corresponding classes.

There are still open challenges as to how well machine learning techniques can maintain their performance in the presence of packet loss, latency jitter, and packet fragmentation [12]. This Survey shows that all traffic classification methods have certain disadvantages.

References

[1] W. Zhou, L. Dong, L. Bic, M. Zhou, and L. Chen "Internet Traffic Classification Using Feed-forward Neural Network", IEEE ICCP 2011.

[2] C. Trivedi, Mo-Yuen Chow, A. Nilsson, and H. J. Trussell "Classification of Internet Traffic using Artificial Neural Networks". Publisher: North Carolina State University, Center for Advanced Computing and Communication, series/report no: TR-02/05, year: 2002.

[3] T. Auld, A. W. Moore, Member, IEEE, and S. F. Gull "Bayesian Neural Networks for Internet Traffic Classification", IEEE Trans. Neural Netw., Vol. 18, No. 1, Jan. 2007.

[4] R. Yuan, Z. Li, X. Guan, and Li Xu "An SVM-based machine learning method for accurate internet traffic classification" Springer Science + Business Media, LLC 2008.

[5] L. Bernaille, R. Teixeira, I. Akodjenou, A. Soule, and K. Salamatian "Traffic Classification On The Fly", ACM SIGCOMM Computer Communication Review, Volume 36, Number 2, April 2006, pp. 23-26.

[6] T. Karagiannis, K. Papagiannaki, and M. Faloutsos "BLINC: Multilevel Traffic Classification in the Dark", ACM SIGCOMM 2005, August/September 2005.

[7] G. Szabo, I. Szabo, and D. Orincsay, "Accurate Traffic Classification", WOWMoM 2007.

[8] A. W. Moore and D. Papagiannaki "Toward the accurate Identification of network applications", in poc. 6th passive active measurement. Workshop (PAM), mar 2005,vol. 3431, pp 41-54.

[9] W. John and S. Tafvelin "Heuristics to classify internet backbone traffic based on connection patterns" ICOIN, 2008.

[10] A. B. Mohd and S. M. Nor "Towards a Flow-based Internet Traffic Classification for Bandwidth Optimization" International Journal of Computer Science and Security (IJCSS), Volume (3) : Issue (2)

[11] A. Callado et al., "A Survey on Internet Traffic Identification and classification," IEEE Commun.Surveys & Tutorials, 2008.

[12] T. T. T. Nguyen and G. Armitage, "A Survey of Techniques for Internet Traffic Classification Using Machine Learning," IEEE Commun. Surveys & Tutorials, vol. 10, no. 4, 2008, pp. 56–76.