CLASSIFYING INTERNET APPLICATIONS ON EARLY STAGE IN 3G ENVIRONMENT

Taizhong Quan, Seongjin Lee, Youjip Won

Hanyang university, Korea
{coolquan| james| yjwon}@ece.hanyang.ac.kr

Abstract

Application identification plays a crucial role in network management. Among many different approaches introduced in the field, Bernaille et al. [1] introduces connection based clustering to identify applications. In this paper, first four packet sizes with directions are used as a feature set to cluster different applications in wireless environment, which is acclaimed to cluster with 93.7% of accuracy in [1]. However, our experiments not only show that it has low identification accuracy (89.3%), but also show clustering is not suitable for identification of wireless applications.

Keywords: Application classification, 3G, Machine learning, Clustering, network measurement

1 Introduction

Every year a variety of online applications are developed to enhance our lives. As the number of internet applications is rapidly growing, application identification becomes more and more significant in network management.

Using well-known TCP or UDP ports defined by International Assigned Number Authority (IANA) is a traditional and simple method to indentify internet applications. However, port-based approach is not very efficient. Because many popular applications, such as P2P, use ephemeral ports, and some applications such as MSN, VoIP, use the same port[8]. [4] suggests that only 50% applications can be identified by IANA. Another method is using payload based analysis. This approach exploits the payload of applications to find special signatures. However, it is complex, and it becomes infeasible when applications are encrypted.

In recently researches ([1], [2], [3], [6], [7], [8]), machine learning and the statistics of traffic features (such as inter-arrival, packet size distributions) are used to identify applications. Machine learning has two phases. Generally first phase is training phase. Where the traffic features from training trace are used to map different traffics to the desired classes. Selecting a efficient traffic features and a suitable cluster method is very important in this phase. Second phase is classifier phase. The identification system uses the desired classes to assign and label the online traffics. Bernaille et al. [1] introduce a methodology which combines port number and normal distribution model to successful identifying the unknown traffic problem and reduce the false-negative rate.

This paper follows the methodology introduced in [1] to classify wireless internet traffic, provided by SKT, Korean mobile service. We use five different groups of applications: VoD, Web, Upload, Download, Game.

The rest of this paper is organized as follows: we describe related work in section 2. The traces used in this paper are introduced in section 3. Four different feature sets are introduced and examined the performance with K-means clustering. And the clustering result and packet distribution are analyzing in section 5. Finally we conclude paper in Section 6.

2 Related work

Recently, many different methods have been introduced to solve the problem of application identification. Machine learning is one of the well known approaches. It generally has two phases – the training phase and the classifier phase. A good identification system requires “strong” features and a simple but powerful clustering method. An ideal machine has to be accurate and agile.

Bernaille et al. [1] presents an approach to identify applications using the direction and packet size of the first 4 data packets in each connection, where K-means, Gaussian and Hidden Markov Model are used in their system. Connection is defined as packets transmitted between a server and a client. Bidirectionally using TCP protocol, it initiates with three way handshake and ends with Fin control packet. They show that packet size and directions are powerful in clustering the applications. Their result shows that the identification accuracy is 93.7% with

clustering.

Williams et al. [2] carried out an empirical study of feature reduction algorithms. And they calculate the accuracy of these reduction algorithms with 5 different kinds of machine learning algorithms: Bayesian Network, C4.5 Decision, Naïve Bayes(NBD,NBK) and Naïve Bayes Tree. The result shows the accuracy using reduced subsets which is selected using their algorithms is almost as good as using the full feature set described in [3].

Moore and Zuev [3] used 248 flow features to classify different application types. Among these there were aggregate flow features such as median of bytes in packet and per-packet features like individual packet sizes. The result show that less than 20 efficient features can lead to an accuracy classification.

3 Description of data sets

The traffic traces used in this paper are taken from a SK Telecom's High-Speed Downlink Packet Access (HSDPA) Network. SKT is a major wireless telecommunication service provider in Korea. HSDPA is a mobile telephone data transmission protocol, which is also known as 3.5 Generation technology. Figure 1 gives abstract architecture of HSDPA Network of SK Telecom. Tapping device is placed before Access Control Router (ACR) device. ACR has mobility management and IP routing function for HSDPA services between remote access station (RAS) and core network. Mobile devices make connection to HSDPA services through RAS. Authentication, authorization, and accounting management server and application server are connected to ACR.

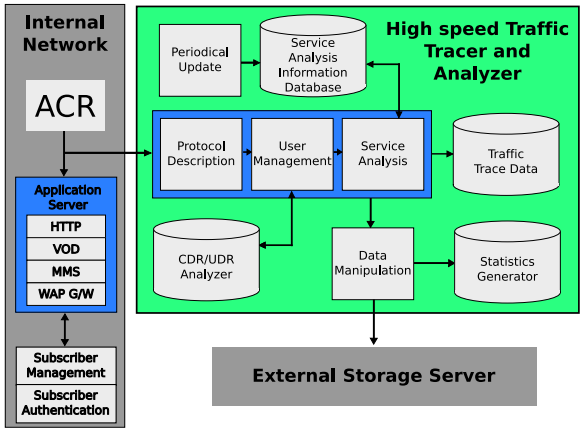

Figure 1. Architecture of HSDPA Network

Figure 2 shows HSDPA Traffic Tracer and Analyzer. High speed Tracer and Analyzer (HiTTA) has the capability to capture the traffic with 0.1 μ sec granularity. It is connected in between application servers and ACR. Application server provides services like HTTP, VoD, Multimedia Mobile Service (MMS), Wireless Application Protocol Gateway (WAP G/W), and etc. Subscriber management and authentication server systematically controls the subscribers in the internal network. Its role is not only to examine the Layer 3 and Layer 4 protocol, but also manages interaction of CDR/UDR Analyzer with User Management module to control and analyze Call Detail Records (CDR) and User Data Records (UDR) for accounting and billing system. It also updates Service Analysis Information Database with Service Analysis module. Purpose of the module is to monitor the user access behaviors and to obtain service usage statistics for future business models. Fourth, it creates general statistics and maintains traffic trace data for marketing purposes. Traffic Trace Data module in the HiTTA keeps the user generating packets temporarily. Only the data generated by the HiTTA is parsed and sent over to External Storage Server.

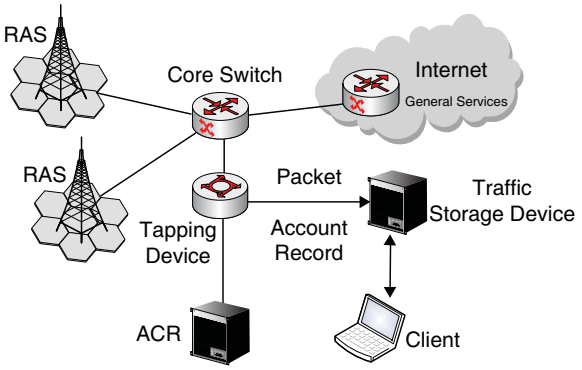

Figure 2. HSDPA Traffic Tracer and Analyzer

3.1 Packet trace

In this study, all the packet traces are collected during 2007.9 and 2007.11 from HSDPA network. They contain VoD, Game, Web, Upload, and Download 5 application types. The length of a trace is about 10 minutes. 1013 pcap traces are collected. All the traces contain all packets traversing the monitor during the measurement period. Table1 shows applications used in the paper.

3.2 Filtering Packet Traces

There are 12072 connections in the dataset, but only connections start after our traces are used. And the control packets are filtered out to capture actual data packets. After filtering, 6702 connections are left. There are 151 Download connections, 248 Game

Table 1 Description of application

Application Type	Description
Web	Naver,Daum,Empas
VoD	Daum UCC, YouTube
Downloads	Xtoc
Online Games	Koongpa, Maple Story
Uploads	Mail Uploads

Table 2 Elements of comparative experiments

	Experience 1	Experience 2	Experience 3	Experience 4
Feature set	IAT (inter arrival time)	Packet size	ITA+packet size	moudulous of packet size
Number of clusters	40	40	40	40
Number of packets	4	4	4	4
Training times	5	5	5	5
Number of training data	5	5	5	5
Number of connection	445	445	445	445

connections, 168 Upload connections, 89 VoD connections and 6046 Web connections.

4 Feature training

In Bernaille et al. [1], Moore and Zuev [3] packet size distribution of connections is an important flow feature in classification. Using K-means clustering number of experiments are conducted to find better feature set. There are two reasons to selecting K-means methods among the three models in [1]. First is that K-means is simple to realize and computationally efficient. Second the result in [1] shows that the performance of 3 different methods is very similar.

4.1 K-means clustering

K-means clustering is one of the simplest unsupervised learning algorithms. K-means clustering aims to partition n observations into k clusters. The main idea is to define a centroid for each cluster. Given a set $X = \{x_1, x_2, \cdots x_k\}$, X is the set of all connections used in this paper. Then K-means clustering partitions the X into k sets (k < n) clusters S. $S = \{s_1 s_2 \cdots s_k\}$, so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in X} \left\| x_j - \mu_i \right\| \qquad (1)$$

where μ_i is the mean of s_i.

K-means algorithm has 3 steps. First, centroids are randomly chosen for each cluster. Second, x_j $\left(x_j \in X\right)$ is assigned to the closest s_i $\left(s_i \in S\right)$. Third, new centroids are computed. And last two steps are iterated until all elements are clustered to a cenroid with minimum distance. Consult [5].for more detailed explanation on K-means.

4.2 Cluster quality metric

In order to measure the quality of clustering, the Normalized Mutual Information (NMI) ([1],[9]) is used .

X are an application connection sets and Y are cluster sets. In order to compute the quality of clustering, $NMI\left(X,Y\right)$ the mutual information between X and Y is compared as:

$$MI\left(X,Y\right) = \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{p_i p_j}\right) \qquad (2)$$

where p_{ij} is the probability that a connection in cluster j belongs to application i, p_i is application $i's$ probability in all the connections and p_j is cluster $j's$ probability in all the connections. $MI\left(X,Y\right)$ measures the shared information between X and Y . In order to normalize the value between 0 and 1:

$$NMI = \frac{MI}{\sqrt{H\left(X\right)H\left(Y\right)}} \qquad (3)$$

$H\left(X\right)$ and $H\left(X\right)$ are the entropy of X and Y, the NMI bounded between 0 and 1. When NMI is 1, it means that applications and clusters are one to one mapping.

4.3 Experiment

Choosing viable features is critical in the training phase. In [1], the packet sizes and packet directions of the first 4 packets in a connection are used. And the accuracy of their clustering system is 93.7%. In this section, several feature sets are compared to see how the features influence the clustering results.

In order to make comparison, four different experiments use the first 4 packet's feature sets to represent a connection. Using k-means clustering partition the connections into 40 clusters. All tests use different feature sets to represent a connection. E1(Experiment 1) uses first 4 IAT(Inter arrial time) and packet size. E2 uses first 4 packet size. E3 uses first 4 IAT. And E4 uses first 4 packet size without information of direction. Except E4 other 3 take packet directions into account. In each experiment, 5 different training data sets are clustered by K-means. Every training data set has 445 connections and consists of VoD, Web, Upload, Download and Game 5 different applications. To avoid bias in the clustering, 89 connections are randomly selected from each application. Because the NMI is influenced by the initial centroids, training phase

will be executed 5 times with the same training data, and getting the mean of NMI. In other words, 5 fold CV is used.

The x axis of Figure 3 means the number of training data. The y axis is the NMI. And every line represents the NMI of an experiment with 5 different training data. The NMI of E4 is lower than other 3. It shows the influence of direction of the packets in clustering. The NMI varies sharply even using the same feature sets, because the training data has effect on the clustering quality. Last but not the least is that the NMI is much lower than Compared to [1], where it says 0.7. Experiments show that the highest NMI is less than 0.65, although using the same connection feature set and the same clustering method. The mean NMI of E2 is only 0.6329 and the accuracy of classification is 89.3%.

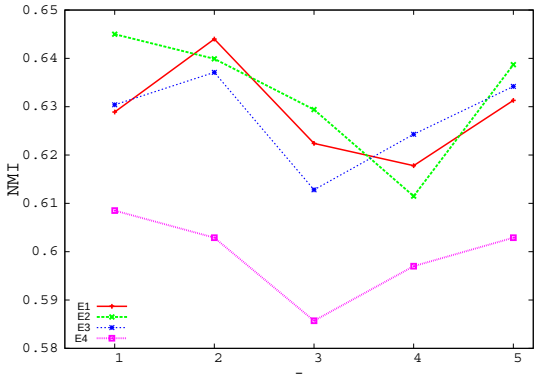

Figure 3. The NMI of Clustering result

5 Evaluation and analysis

This section analyzes the clustering results and the packet distributions.

5.1 Clustering result

Table 3 shows the clustering result with 5 different training data. First column represents the different training data sets used in section 4. Table 3 shows the number of clusters which contains the application groups. For example, after training No. 1 training data, 7 clusters contain download connections. In table 3, almost half of the clusters contain Web and Game connections. It shows that Web and Game connections have various of types. Web connections are collected from 3 websites described in Table 1. Nowadays website provides more and more services. People can watch flash, download data, chat with others on the website. Different services lead to different connection feature types. It leads the Web connections widely distributed in the clusters.

Table 3 Cluster number of applications

	Download	Game	Upload	VoD	Web
1	7	14	3	6	20
2	7	17	3	6	18
3	3	14	3	7	27
4	2	20	1	6	20
5	4	20	3	5	22

Table 4 Connections group by the number of MTU

	Web	Upload	Download	VOD	Game
M_0	2	59	51	0	0
M_1	2818	0	0	85	17
M_2	331	106	0	2	0
M_3	71	0	0	0	2
M_4	22	0	99	0	0
M_5	29	0	1	0	0
Total	3273	168	151	87	19

5.2 Packet sizes distribution

Maximum Transmission Unit (MTU) in the data set is 1434 bytes. Note that, almost all the connections have 1434 bytes packet. Table 4 shows the separation of connections with the number of MTU in the first 20 packets of a connection. The separation rule is as following: $p_i \in$ P $(0 \le i \le 20)$, If p_i =MTU $(i \ge n)$ and $p_i \ne$ MTU $(i < n)$ Then P$\subset M_n$.

$P = \{p_1, p_2, \cdots p_{20}\}$, P is the set of packet sizes in a connection. p_i is the i th packet's size. In a connection, if the first n-1 packets are all smaller than MTU and from n to 20th packets are all MTU, the connection belongs to group M_n.

Table 4 shows the number of connections from M_0 to M_5 group. All Upload connections are grouped in M_0 and M_2. 99.3% of Download connections are in M_0 and M_4. 99.5% of VoD are in M_1 group. They all belong to one or two groups. But Web and Game have various of patterns. Web connections are distributed in all the groups from M_0 to M_5. Providing kinds of services lead the Web application have kinds of connection types. It adds the difficult to separate web connections with other Applications which provide similar services. For instance, In M_1, Web and VoD connections all provide flash services. First 20 packets of the connections in group M_1 all have MTU size, except the first packet. And the first packet also doesn't have distinct difference in size or direction. It is hard to separate connections in group M_1 which use packet size and direction only. In real clustering, it is the main issue to lead a low accuracy result. Although method described in [1] is simple and fast in identification, it is not suitable to identify

applications with similar services.

6 Conclusions

In this paper, we follow the methodology introduced in [1] to classify wireless internet applications provided by SKT Korean mobile service. A number of comparative experiments are carried out with four different feature sets using K-means clustering methodology. In our experiments, the NMI is lower than [1], and the accuracy of total classification is 89.3%. The result shows that clustering method is not suitable for identification of wireless applications.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. R0A-2007000201140-2008), and Seoul R&BD Program (KU080661).

References

[1] L. Bernaille, R. Teixeira, and K. Salamatian. Early Application Identification. In The 2nd ADETTI/ISCTE CoNEXT Conference, Lisboa, Portugal, December 2006.

[2] N. Williams, S. Zander, and G. Armitage. A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. SIGCOMM Computer Communication Review, October 2006.

[3] A. W. Moore, D. Zuev, Internet Traffic Classification Using Bayesian Analysis Techniques, in Proceedings of ACM SIGMETRICS, Banff, Canada, June 2005.

[4] J. R. Quinlan. C4.5: Program for Machine Learning. Morgan Kaufman, 1993.

[5] K. Alsabti, S. Ranka, and V. Singh. An efficient k-means clustering algorithm. In Proceedings of the First Workshop on High Performance Data Mining, Orlando, FL, March 1998.

[6] J. Erman, M. Arlitt, and A. Mahanti. Traffic Classification Using Clustering Algorithms. Proceedings of ACM SIGCOMM Minenet Workshop, Pisa, Italy, September 2006.

[7] F. Porikli. Trajectory distance metric using hidden markov model based representation. In IEEE European Conference on Computer Vision, PETS Workshop, 2004.

[8] W. Li and A. W. Moore. A Machine Learning Approach for Efficient Traffic Classification. In Proceedings of the IEEE MASCOTS, Oct.2007

[9] A. Strehl and J. Ghosh. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research (JMLR), 2002.