

Classification of web-based email traffic in Thailand

Sirikarn Pukkawanna^{*}, Vasaka Visoottiviseth^{*}, Panita Pongpaibool[†]

^{*}Department of Computer Science, Mahidol University, Rama 6 Rd., Bangkok 10400
E-mail: g4836585@student.mahidol.ac.th, ccvvs@mahidol.ac.th

[†]NECTEC, 112 Phahol Yothin Rd., Klong Luang, Pathumthani 12120
E-mail: panita@nectec.or.th

Abstract— Traditional works in traffic classification usually measure usage of mail applications by monitoring only SMTP, IMAP, and POP3 traffic. The shortcoming of such measurement is that it does not take into account web-based email usage (Webmail) since the Webmail traffic is usually classified collectively as HTTP or web traffic. The simple way to identify Webmail traffic is mapping source or destination IP address with URLs of Webmail providers (e.g. Hotmail, Yahoo!, and Gmail), is neither flexible nor accurate. The URL mapping technique cannot detect some related Webmail traffic such as advertising banners, pictures, and news, which are requested from other servers. In this paper we propose a technique to detect Webmail traffic from regular HTTP traffic by matching unique Webmail keywords in HTTP payload, in combination with TCP flow analysis. The significance of our method is that it can identify Webmail traffic missed by using the URL mapping alone, and can identify all packets associated with a TCP flow in both sending and receiving directions.

I. INTRODUCTION

Web based e-mail (Webmail) is one of the most popular Internet applications that allows users to access their mailboxes from anywhere in the world. It provides an alternative to using an email client such as Microsoft Outlook, Mozilla Thunderbird, or Eudora. Many universities and organizations use such software to provide students and staffs with web-based access to their email accounts and services. Many content providers also offer free Webmail service for their customers. The most notable Webmail providers are Hotmail, Yahoo!, and Gmail [12].

Webmail is a web application that acts as a gateway between a web server and a mail server. Users connect to the Webmail engine within the web server via the HTTP protocol. Then Webmail communicates with the e-mail sub-system via the IMAP protocol in order to retrieve or send email messages. The Webmail engine then creates an HTML response, and sends it back to the client. Fig.1 shows the Webmail architecture. Because content data of e-mail is usually embedded in HTML pages, Webmail traffic often appears as HTTP traffic, i.e., using TCP port 80. Thus, it is difficult to distinguish Webmail traffic from other HTTP traffic.

A simple method to distinguish Webmail traffic from other HTTP traffic is to map source or destination IP address with URLs of Webmail providers (e.g. Hotmail, Yahoo!, and Gmail). This method requires knowledge of URLs of

Webmail providers. It is neither flexible nor accurate because the Webmail server usually redirects HTTP requests of e-mail contents and embedded objects, such as advertisement banners, flashes, and news, to other non-mail servers.

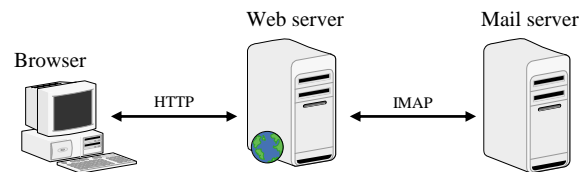


Fig. 1. Webmail traffic architecture.

In this paper, we propose a method for classifying Webmail traffic from HTTP traffic by matching unique Webmail keywords in HTTP payload, in combination with TCP flow analysis. Our algorithm can identify Webmail traffic that cannot be detected by the URL matching technique alone. In addition, we can identify all packets associated with a TCP flow in both directions (i.e. packets with the same source and destination IP addresses, same source and destination ports, and same protocol).

The rest of this paper is organized as follows. Section II outlines previous studies in the area of Webmail detection. Section III describes the traffic data used and characteristics of each location where traffic is captured from. Section IV describes in detail our proposed algorithm and specific keywords for Webmail traffic identification. Section V describes our experiments and results. Finally, conclusion and future work are discussed in Section VI.

II. PREVIOUS WORK

Recently, Internet traffic classification and measurement has been a subject of interests. However, there is a limited of number of publications that report the detection and measurement of Webmail traffic. Many traffic classification works [7, 8, 9] classify mail traffic based on mail protocols, such as IMAP, POP3, and SMTP. These works do not take into account Webmail usage. Webmail traffic is often classified and reported as HTTP traffic.

Previous studies of Webmail detection [6] observe only IP addresses of source and destination in the IP header, and map the IP addresses with a list of URLs of popular Webmail providers.

TABLE I
TYPE SIZE FOR PAPERS CHARACTERISTICS OF COLLECTED TRAFFIC BASED ON LOCATION

Location	Start date	Finish date	Duration	LAN Speed	Hosts	Packets	Bytes
Internet café	2006-06-09	2006-07-01	552hrs	100Mbps	22	238M	10 ⁶ GB
Mahidol University	2006-06-22	2006-07-01	240hrs	100Mbps	21	12M	172GB
NECTEC	2006-06-14	2006-07-01	432hrs	100Mbps	12	2M	1GB

Our approach differs from the previous work in four ways:

- We study characteristics of traffic from popular Webmail services, namely Hotmail, Yahoo!, and Gmail. We record unique keywords that appear in the packet's payload that can identify Webmail traffic of each provider.
- We observe characteristics and behaviors of Webmail traffic of each provider based on Webmail usage, such as logging-in, reading messages, and composing messages. We study factors that affect performance of different Webmail services.
- We combine the previous method of URL matching, with our approach of keyword matching and TCP flow analysis.
- We study the size of payload that is sufficient to detecting Webmail traffic.

III. TRAFFIC DATA

The traffic data analyzed in this paper is collected from three locations in Thailand, namely a shared-use Internet café, a university computer cluster, and a research office. We use tcpdump [5] to capture the first 96 bytes of each packet, which includes Ethernet, IP and TCP/UDP headers and some payload.

Table I lists characteristics of traffic captured from each location in this work.



Fig. 2. Users at Internet café

A. Internet Cafe

The Hub Coffee Internet café is a 24-hour shared-use Internet café, whose customers are mainly students and young adults. The Internet café offers 22 active hosts which are

connected to the Internet via an 8Mbps ADSL connection. We captured all packets going in to and out of ADSL routers 24 hours a day starting from June 9, 2006 until July 1, 2006. During the 23-day capture period, we see an average of 35 users per day. Fig. 2 shows a shared usage environment at the Hub Coffee Internet café.

B. Computer Cluster at Mahidol University

The graduate student's computer cluster at Mahidol University is a laboratory for computer science graduate students. This laboratory is connected to the university network via a 100 Mbps link. The University's connection to commodity Internet is 50 Mbps. We captured packets 24 hours a day starting from June 22, 2006 until July 1, 2006, during which period 21 hosts are active. Roughly 20 students visit the laboratory each day.

C. Research Office at National Electronics and Computer Technology Center (NECTEC)

An office within NECTEC that we captured the data from is a network research lab. Users in this office are engineers and researchers. The office connects to commodity Internet with an 8 Mbps connection. Since all hosts in this office are on the same broadcast LAN, we captured all packets broadcast to this LAN. The capture period is during June 14, 2006 to July 1, 2006, 24 hours a day from 12 observed hosts.

IV. PROPOSED TECHNIQUE FOR WEBMAIL DETECTION

In this Section, we propose a technique for Webmail detection based on keyword matching and TCP flow analysis. We discover unique keywords in the payload of each Webmail provider during three phases of Webmail usage, namely logging-in, reading messages, and composing messages.

A. Webmail Keywords

Our work is based on identifying specific Webmail keywords in the application-level user data. Table II shows lists of keywords in Webmail traffic of four Webmail providers (Gmail, Hotmail, Yahoo!, and Mahidol's Webmail) during three phases of usage, namely, logging-in, reading messages, and composing messages. Note that we attempt but fail to classify keywords in NECTEC's Webmail service. This is because NECTEC's Webmail uses Hypertext Transfer Protocol Secure (HTTPS), which encrypts user payload. This is a limitation of our purpose technique.

We can see from Table II that many keywords, such as "mail", "msg", "inbox" could appear in more than one Webmail providers. Therefore, in order to distinguish Webmail traffic of each provider, we must use these keywords in conjunction with other techniques.

TABLE II
WEBMAIL KEYWORDS

Provider	Log in	Read message	Compose message
Gmail	mail, setgmail, search=inbox, mail?auth, gmailchat, GMAIL_LOGIN	mail, mail?ik, mg	mail, msgbody, htmlcompose, name="to", name="cc", name="bcc"
Hotmail	hm, hmhome, curmbox, sbox, HMServer,	hm, HotMail, msg, getmsg, HMServer,	compose ,msn_hm, premail, msghdrid, newmail, encodeto, encodecc, encodebcc, UponSend, InReplyTo, sent_mail
Yahoo!	ym, b?P, mB, mail, Mail, crumb, ym/login, mailcommon, mailmain	box="Inbox", rb=Inbox, mB, MsgId, ym, ym/ShowFolder, ym/ShowLetter, bodyPart	ym/Compose, message, Message, sent, Attach, MailUser, ReplyTo
Mahidol Webmail	MailBox, mailbox, checkbox, mailbox=INBOX, mailbox=mail, message, INBOX, imapuser, IMAPServer, draft, sent-mail	message, message=Mail box, inbox, Inbox,contact, subject, Reply, draft, sent-mail	compose, name="message", name="to", name="cc", name="bcc", sent

B. Proposed Algorithm

In this paper we propose to apply a combination of methods, URL matching, keywords matching, and TCP flow analysis, to distinguish packets containing Webmail traffic from other HTTP packets. The flowchart of our proposed technique is illustrated in Fig. 3.

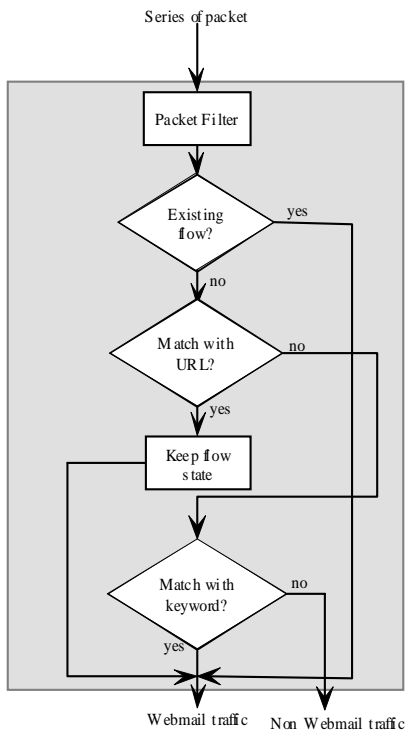


Fig. 3. Flowchart of the proposed technique for Webmail detection

The steps of our proposed technique for Webmail traffic classification are as follows. The Packet filter module filters out non-HTTP packets and sends HTTP packets to check with a list of existing flow records that have been classified as Webmail. If the HTTP packet belongs to existing Webmail flow record, the packet is counted as Webmail traffic. If the HTTP packet is not a part of an existing flow, we check source and destination IP addresses against the known URLs

of Webmail providers or URL keywords. For example, the keywords could be hotmail, gmail, webmail, or mailbox. A list of URL keywords is shown in Table III. Note that yahoo is not a valid URL keyword since it could refer to any non-mail services offered by Yahoo!.

TABLE III
URL KEYWORDS

URL keywords
mail, webmail, mailbox, hotmail, gmail, googlemail

If the source or destination IP address match with URLs or URL keywords, this HTTP packet is classified as Webmail. It is then forwarded to the Keep-flow-state module, which extracts the 5-tuple flow identification (protocol, source IP, destination IP, source port, and destination port) and records this flow into the Webmail flow record. On the other hand, if the URLs or URL keywords do not match with domain names of either source or destination, we look into packet payload to match keywords shown in Table II. If the keywords exist in this packet, it is then classified as Webmail traffic. Otherwise, we classify it as non-Webmail traffic.

V. EXPERIMENT AND RESULTS

In this Section, we show results of our classification, both overall data and results of each location. We demonstrate the performance of our method, described in Section IV to evaluate the success of our proposed technique. We present the effect of payload size for detecting Webmail traffic and also discuss different traffic characteristics of Webmail providers.

A. Overall Results

In Fig. 4, we present the results of traffic classification of all the traffic data from a shared-use Internet café, a university computer cluster, and a research office. Fig. 4(a) and 4(b) show by-packet and by-byte classification, respectively. The results of HTTP traffic classification are shown in Fig. 5. The popularity of Webmail providers based on percentage of number of packet and total amount in bytes is shown in Fig. 6(a) and 6(b).

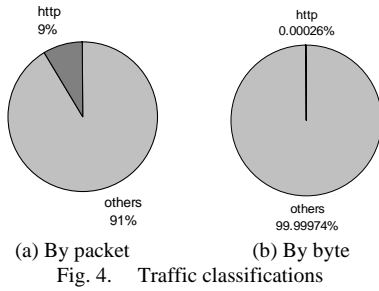


Fig. 4. Traffic classifications

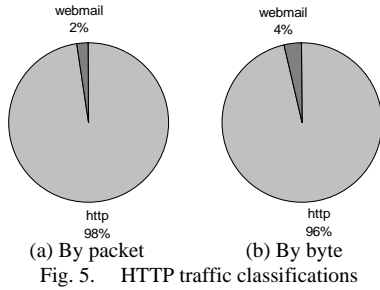


Fig. 5. HTTP traffic classifications

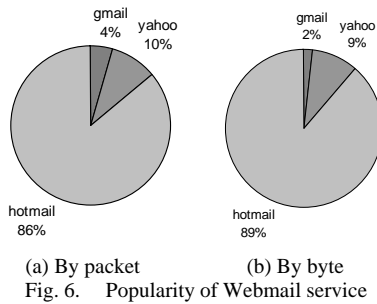


Fig. 6. Popularity of Webmail service

In the amount of traffic data, HTTP traffic represents 9% of all the packets. HTTP traffic comprised 2% of number of Webmail packet. Hotmail is large category, represents 86% of all Webmail packets. Yahoo! and Gmail represent 10% and 4% respectively, in the number of packets.

B. Results of Each Location

Fig. 7, 8 and 9 present the percentages of Webmail traffic in HTTP traffic at a shared-use Internet café, a university computer cluster, and a research office, respectively. We break down for each location into ratios based on total number of packets and bytes transferred.

From the traffic observed, Webmail application is used more in NECTEC than in other locations we collected data. Webmail comprised about 40% in the number of packets of NECTEC traffic, but only 1-2% of those from the Internet café and MU. This perhaps is due to the nature of office usage at NECTEC as oppose to temporary usage by students and users at the university cluster and at the Internet café.

C. Our Proposed Method vs. URL-Matching

In this Section, we present the performance of our approach. In Fig. 10-12, we compare the results between two methods, the URL-matching and the proposed method of Webmail classification at NECTEC, MU, and the Internet Café respectively. The results of Webmail detection of each location are classified by packet and by byte.

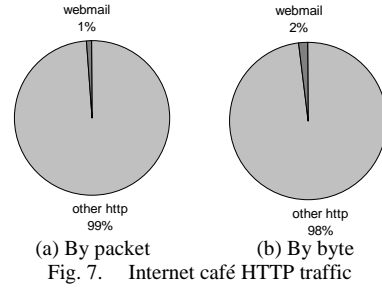


Fig. 7. Internet café HTTP traffic

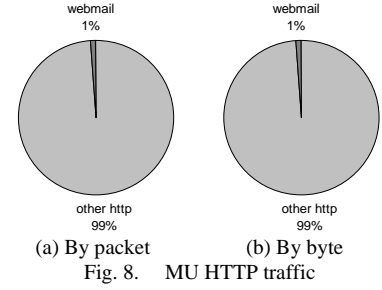


Fig. 8. MU HTTP traffic

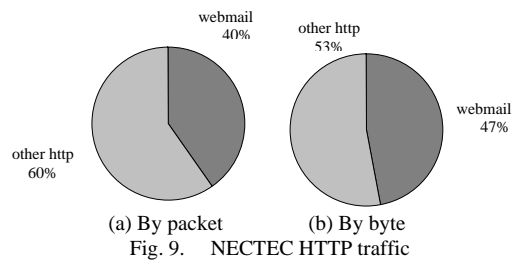


Fig. 9. NECTEC HTTP traffic

In all traces, our method can detect more Webmail traffic than using URL-only mapping. The improvement could be as high as 21 times in the case of Internet Café traffic classification by packet.

D. Effect of Payload Size

Next, we investigate the payload size that is sufficient to detect Webmail traffic. We analyze traffic data of NECTEC to find the effect of the size of user payload to the accuracy of our Webmail classification. Normally, each packet contains 20 bytes of TCP header, 20 bytes of IP header, and 14 bytes of Ethernet header. The size of user's payload we consider varies between one to 42 bytes. In other words, we applied the keyword matching to the first 55 to 96 bytes of a packet. Fig. 13 shows the effect of size of user payload on the number of packets classified as Webmail. This figure illustrates that keywords or strings are often found in the range of the sixtieth byte to the seventieth byte and also the eighty-fifth byte to the ninetieth byte of the captured packets. After the ninetieth byte, the number of packets that classified as Webmail does not change. Therefore, we can conclude that capturing the first 36 bytes of user payload (i.e., the first 90 bytes of a packet) is sufficient for detecting Webmail traffic.

E. Webmail Provider Characterizations

During the classification process, we have an opportunity to observe characteristics of several Webmail providers, such as Hotmail, Yahoo!, Gmail, and MU Webmail. We found that

for each e-mail session Hotmail and Yahoo! usually connect to a number of related servers. These servers are not necessarily mail servers, but servers for embedded objects, such as advertisement banners, flashes, and icons. On the other hand, we found that Gmail and MU Webmail connect to only one Webmail server during each e-mail session.

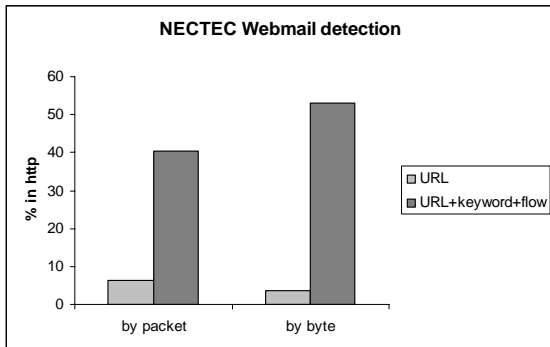


Fig. 10. NECTEC Webmail detection (%)

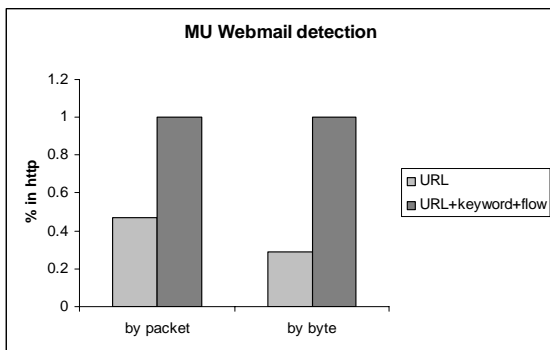


Fig. 11. MU Webmail detection (%)

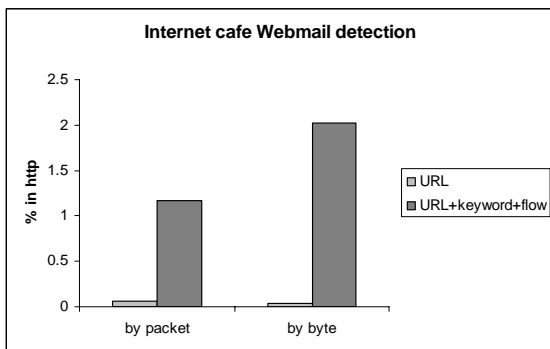


Fig. 12. Internet cafe Webmail detection (%)

To understand the effects and implications of the number of communicating servers, we measure number of packets per e-mail session and total amount of traffic per e-mail session. In our experiments, an e-mail session starts with a user’s log-in, and ends after a user reads five e-mail messages and writes five e-mail messages.

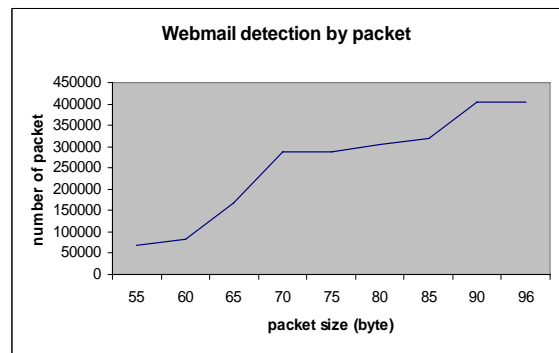


Fig. 13. Effect of payload size

In this experiment, all users have the same mailbox environment, for example, the same number of e-mails in their mailboxes, and read and write the same e-mail content. We repeat this experiment for each of the four Webmail services (Hotmail, Yahoo!, Gmail, and MU Webmail). The comparison of the average number of flows per email session, the average number of packets per flow, and the total number of packets per email session for each Webmail server are shown in Table IV.

TABLE IV
AVERAGE NUMBER OF FLOW PER E-MAIL SESSION, AVERAGE NUMBER OF PACKET PER FLOW, AND TOTAL NUMBER OF PACKET PER E-MAIL SESSION

Provider	# flow/session	# packet/flow	# packet/session
Hotmail	52	12	603
Yahoo!	78	7	522
Gmail	10	26	263
MU	88	5	434

Data in Table IV suggests that the number of communicating servers does not affect the number of TCP connections per e-mail session. That is, even though Gmail and MU Webmail only communicate with one server, their average numbers of flows per session are not necessarily smaller than that of Hotmail or Yahoo!

Moreover, we observe that on average Gmail has the smallest number of flows (10 flows per session) and the longest flows (26 packets per flow). This implies that Gmail uses a persistent connection or HTTP version 1.1 to send multiple objects in one TCP connection. At the same time, other Webmail services have more flows and shorter flows on average, which implies that they use a non-persistent connection or HTTP version 1.0.

The comparison of total amount of traffic per e-mail session in different Webmail services is shown in Fig. 14. From our experiment, Hotmail and Yahoo! require large number of total traffic per e-mail session. On the other hand, found that Gmail requires about half of number of packets and total traffic of Hotmail. The above data indicates that both Hotmail and Yahoo! provide e-mail usage with large bandwidth overhead and confirms that Gmail outperforms other Webmail services in terms of speed and efficiency. This is due to the unique feature of Gmail in which majority of its functionality is enacted client-side, i.e., on the browser rather than at the server, and is done with JavaScript [11].

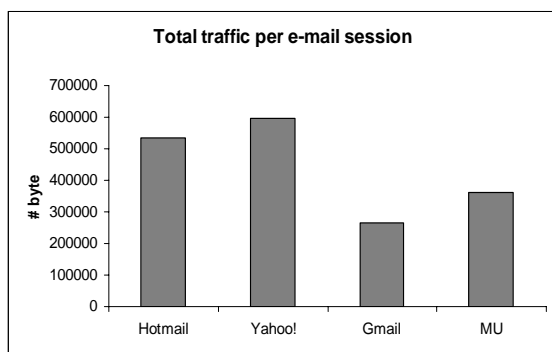


Fig. 14. The comparison of total traffic per e-mail session

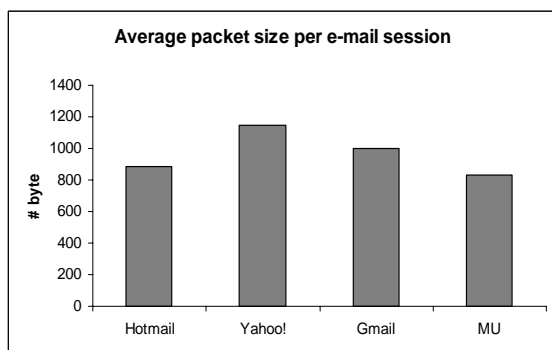


Fig. 15. The comparison of average packet size per e-mail session

We also observe average packet size per e-mail session of each Webmail provider. As shown in Fig. 15, Yahoo! transfers on average 1144 bytes of packet size per one e-mail session while Hotmail, Gmail and MU Webmail transfer on average 884, 1003 and 830 bytes to provide the same session. This implies that Yahoo! sends big objects such as advertisement banners. Moreover, from our analysis, we found that Yahoo! has the most of number of embedded objects in one e-mail session. This fact reveals from our data in Table IV that the number of flows per email session is large, and also from Fig. 14 that the total traffic in byte per email session is the largest among all of Webmail providers we have studied.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a Webmail detection approach and analyze the traffic of three locations in Thailand, namely a shared-use Internet café, a computer cluster, and a research office. Our method can identify Webmail traffic more than using URL-only mapping method. From the traffic analysis, users in NECTEC use Webmail service more than users at Mahidol University and at the Internet café. Also, we found that Hotmail has the highest percentage in number of packets and bytes usage. However, this does not mean that Hotmail is the most popular Webmail service in our trace, because both numbers of packets and bytes do not necessarily indicate popularity of Webmail usage.

From our analysis, we found out that the first 36 bytes of user payload are sufficient for detecting Webmail traffic. We

also found that Yahoo! transfers the most of numbers of embedded objects such as advertising banners, flashes, and icons which are factors that affect performance of different Webmail services.

In the future, we plan to perform additional classifications and measurements to gain more insights into Webmail traffic behaviors. We plan to develop an approach to classify advertising-related traffic from HTTP traffic and report ratios between content traffic and ad-related traffic. In addition, we plan to develop a more reliable method to measure popularity of each Webmail service. The number of byte and packet alone are not accurate indicators because some Webmail services send a lot of advertisement traffic. Finally, we plan to perform measurement and classification on other traffic disguised as HTTP (non-web traffic using port 80).

VII. ACKNOWLEDGEMENTS

We would like to thank the Hub Coffee Internet café, Mahidol University, and National Electronics and Computer Technology Center (NECTEC) in Thailand for allowing us to collect traffic data.

REFERENCES

- [1] Hotmail. <http://www.hotmail.com>
- [2] Google mail. <http://www.gmail.com>
- [3] Yahoo! mail. <http://www.mail.yahoo.com>
- [4] Mahidol webmail <http://webmail.mahidol.ac.th>
- [5] tcpdump. <http://www.tcpdump.org>
- [6] Bowei Du, Michael Demmer and Eric Brewer, "Analysis of WWW traffic in Cambodia and Ghana", in proceedings of the ACM World Wide Web Conference Committee (IW3C2), 2006.
- [7] Thomas Karagiannis, Konstantina Papagiannaki and Michalis Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark", in proceedings of the ACM SIGCOMM, 2005.
- [8] Laurent Bernaille, Renata Teixeira, Ismael Akodkenou, Augustin Soule and Kave Salamatian, "Traffic Classification on The Fly", in proceedings of the ACM SIGCOMM, 2005.
- [9] Thomas Karagiannis, Andre Broido, Michalis Faloutsos and Kc claffy, "Transport Layer Identification of P2P Traffic", in proceedings of the ACM Internet Measurement Conference, 2004.
- [10] Steve Martin, Anil Sewani, Blaine Nelson, Karl Chen and Anthony D. Joseph, "Analyzing Behavioral Features for Email Classification", Whitepaper 2006, <http://whitepapers.silicon.com/0,39024759,60178044p-39000575q,00.htm>
- [11] Demir Barlas, "How Gmail Works?", TechRepublic, December 2005, <http://whitepapers.techrepublic.com.com/abstract.aspx?docid=166323&promo=300111>
- [12] The popularity of Webmail provider <http://www.line56.com>