An Automatic Scheme to Categorize User Sessions in Modern HTTP Traffic

Xiaozhu Lin
Computer and Information Management Center
Tsinghua University
Beijing, 100084, China
Email: linxz02@mails.thu.edu.cn

Lin Quan
Department of Computer Science
Tsinghua University
Beijing, 100084, China

Haiyan Wu
Computer and Information Management Center
Tsinghua University
Beijing, 100084, China

***Abstract*—The characterization of HTTP traffic is crucial for performance evaluation and server design. In this paper, we analyze massive web traces generated by various busy servers in recent years, trying to find the new features of modern HTTP traffic and user behaviors. Comparing the conclusions of earlier studies with our results, we have spotted considerable unconventional ingredients in modern HTTP traffic that could hardly be described by previous models. We also propose an innovative scheme to automatically categorize these various ingredients in modern traffic. The novel aspects of our work are: (1)It reveals the sophisticated composition of modern HTTP traffic with solid evidence, (2)It provides an automatic method to analyze the composition of modern HTTP traffic and (3)It promises a powerful manner to evaluate the possible performance implication of modern HTTP traffic on existing web servers. We hope this work would help researchers and designers to better understand new features of HTTP workloads and therefore make corresponding adaptations in design practice.**

I. Introduction

To understand how HTTP workloads behave is a prerequisite to design reliable and efficient web servers. However, to understand key characteristics in traffic is far more difficult than finding an analytical description for the aggregate level metrics such as throughput. In fact, HTTP traffic is the outcome of many intricately overlapped client 'sessions' which are temporal and logically related, making aggregate level metrics incapable to be approached by analytical models.

Moreover, on the session level, analytical modeling might also be hard. As we will show later, since user behaviors within a single session are affected by various factors such as networking condition and HTML page contents, user sessions might exhibit totally different characteristics in different web servers. Even many previous work well investigated representative behavioral patterns of human users, sessions in modern HTTP traffic are not necessarily from human users.

The two obstacles above largely challenge the analysis and reproduction of HTTP traffic. To overcome them, we have to understand what had happened with recent HTTP traffic, and carefully validate analytical traffic models proposed before. With sufficient analysis methods proposed by previous researchers, we go through complete web traces in the last a few years from four typical web applications in Tsinghua University. These web servers have high click-rates from wide range of users. We think their traces could hopefully reflect the characteristics of modern web traffics.

With thorough analysis of real traces, we find solid evidence showing that the modern HTTP traffic is a synthetic flow from various Internet applications including browsers. Thus, sessions, as the cells of traffic, should be categorized into different types according to their origins and performance implications. We propose a series of well-designed criteria to automatically categorize sessions, and validate its results through strict tests.

The main contributions of the paper are: (1) to extend previous HTTP traffic characterization work with modern networking conditions (2) to propose an innovative and effective scheme to categorize HTTP sessions automatically and (3) to provide the fundamentals for a session-pool-based traffic generator which will be highly tunable.

II. Related Work

Since 1990s, the efforts to characterize web traffic never stopped. [1] proposed a series of metrics to describe the aggregate traffic, most of which were heavily cited by subsequent researchers. Almost at the same time, the self-similarity in WWW traffic had been discovered [2]. The existence of self-similarity in HTTP traffic had been widely validated by later researchers and was used to evaluate various traffic generators such as [3]. [4] used massive TCP traces to analyze the aggregate characteristics in modern HTTP traffic. In recent years, new aggregate level models kept emerging. [5] used fractional Gaussian noise and fractional Brownian motion[6] to simulate self-similar processes.

Many researchers managed to decompose aggregate traffic flow into individual sessions and therefore construct model for sessions. Catledge and Pitkow [7] first investigated user sessions in web traffic at client side. [8][9] were also among the first to focus on session-level features. [10] focused on every 'click' instead of object, and employed IP address information to separate sessions and approximated session parameters by best distributions. The widely accepted definition of session was proposed formally by M. Arlitt [11]. [12] studied web traffic from three levels: aggregate level, intra-session level and inter-session level. Their thorough analysis on every level

largely helped researchers to understand the composition of HTTP traffic. Z. Liu *et al.* [13] used a stochastic marked point process to construct session-level model and focused on the elements in HTTP traffic. This model was employed to generate realistic traffic with much less parameters.

After the year 2000, some non-human ingredients had been spotted in HTTP traffic. Following the three-level concept from [12], Almeida [14] proposed several criteria to identify crawler activities from traces. These criteria served as the starting point of our categorization scheme. [15] tracked several famous crawlers in server traces and indicate their different patterns from human users.

These methods are enlightening, however, they shared some common flaws. They tried to describe all users, sessions or clicks using unified models without classifying them. Besides, most of the traces they discussed were too small or obsolete, making those models and parameters less accurate.

III. Server Traces Considered

In order to discover new trends in recent HTTP traffic we collect raw web traces since 2003 until now from four typical web servers.

The first trace set ("PORTAL") is from the official site of Tsinghua University. The server is always busy serving requests from all over the world, for both education purposes and non-education purposes. From 2005 until now the total size is around 100 GB.

The second trace set ("LEARN") is taken from the main online education server. The server has been widely used by all students and teachers (almost 40,000 people) to publish course materials or to submit homeworks. It also provides limited spaces to store personal pages. From 2003 the total size of the trace set is 217 GB.

The third set ("VSPACE") is from a web storage system. Students and teachers can upload, download and share their interested files. From 2004 we have totally 20 GB traces.

The fourth set ("XK") records every request in the web course-selection system. The traffic is highly bursty and concurrency could reach as high as 5000, which resembles modern E-commerce websites. The trace of one three-day period starting from September 19th 2007 is 27 GB.

Besides, we also take other's open web trace for comparison. EPA-HTTP[1] is a collection of trace in August 29th 1995 from Research Triangle Park. This 4.4 MB trace has also been investigated by other researchers.

IV. Analysis of Modern Web Traces

A. Motivation of Categorization

In order to determine the composition of modern web traffic, we analyze the ingredients on session level, based on the idea that sessions are the elements of web traffic[13]. Here by ingredients, we mean the different types of sessions overlapped together to form the synthetic HTTP traffic. We use the definition of HTTP session in [11], and set the inter-session timeout to 600 seconds. Note that we do not track IP address information to combine different sessions together, because (1) within a period longer than 600 seconds, IP address assignment could change and (2) for traffic from one specific user, autocorrelation in time scales larger than 600 seconds is weak.

Previous research[14] claims that one human session often consists of more than two requests. Therefore those extremely short sessions that are frequently observed in our traces are highly interested.

In practice, these 'short' sessions, as we understand before, may come from gentle crawlers which visit the server occasionally at long intervals and never fetch embedded objects (objects that are not main body of web pages, such as image banners). Short sessions can also be generated by Peer-to-Server-Peer (P2SP) clients. P2SP is a new feature in many P2P clients like Xunlei[2] (a.k.a. Thunder), Gigaget, Flashget and Bitcomet. Besides making use of traditional peers, they also heavily utilize HTTP/FTP resources to maximize the downloading speed automatically. We find that since April 2004 (according to their websites, it is time when P2SP tools emerge), all our web servers except XK have been harassed by P2SP traffic to different extents.

The behavioral patterns of P2SP sessions are quite distinct: (1) Almost all such sessions are requesting isolated URLs pointing to popular files such as .mp3, .rm, .wmv, etc. (2) Since P2SP clients usually request different segments of the target object from multiple sites, they will leave many 'partial response' logs in server traces. (3) Individual P2SP sessions request much fewer objects than human sessions.

Failed P2SP or crawler sessions and successful ones are basically different in that a successful session will take as much as 100 times of bandwidth (response size larger than 30KB) as of a failed session do (failure page about 300–500B), therefore exerting different influences on server performance.

We are also interested in overactive users, i.e. very long sessions. Robots, namely automatic programs or scripts, keep refreshing some specific pages or downloading certain files. They are different from P2SPs and crawlers in session length and cache performance. Robots could be excessive in many transaction-based sites including E-commerce and course-selection (XK). In addition, long-duration crawlers, possibly driven by search engines, routinely download various HTML pages from web sites. They are far more aggressive than short crawlers and have distinct performance implications, thus should be identified specifically. Besides, long crawlers are almost all successful.

B. Metrics Used to Categorize Sessions

We have introduced three important scores to evaluate individual sessions, they are: *Human Similarity*, *Diversity Factor* and *HTML Affinity*.

[1] http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html

[2] http://en.wikipedia.org/wiki/Xunlei

TABLE I
TYPE-BASED WEIGHTS USED TO CALCULATE HS

TYPE : S_i	$weight(S_i)$
mp3, wmv, rm, wma, exe, rar, zip, *etc.*	10
others	1

TABLE II
SCORES OF HTTP CODES USED TO CALCULATE HS

CODE : C_i	$score(C_i)$
4XX/5XX (Client/Server Error)	-20
3XX (Redirection)	-10
206 (Partial Content)	5
200 (OK)	10
others	0

Human Similarity is designed to estimate the similarity between specific session and human sessions, in the requested file types and the results of requests. It is defined as follows:

Definition 1: Let C_i and S_i denote HTTP status code and request object type for ith request within one session respectively, thus, for this session, Human Similarity (HS) is defined as,

$$HS = \sum_{i=1}^{N} score(C_i) weight(S_i) \quad (1 \le i \le N)$$

The mapping from file types to S_i is in Table I and C_i for HTTP codes are listed in Table II. In order to keep a close watch on objects preferred by P2SP sessions (.mp3, .wmv, etc.), we give them special weights in their corresponding S_i.

As a special rule, if a session begins with a request to an embedded object, we can almost assert that it is not from human users[14]. For such requests, $weight(S_i)$ is set to a very large value, such as $1.0e + 5$ to distinguish the session from others. Our experiment has proved this special rule is very effective to extract P2SP sessions.

We note that the distribution of HS is related to the object count of the session. Therefore, HS for short sessions and long sessions should be treated separately. For almost all short or long sessions, we always observe four potential peaks in the Probability Density Function (PDF) of HS (an example is shown in Figure 1). Two peaks appear near zero on both sides, and other two represent extremely high and low HS. This phenomenon motivates us to straightforwardly propose three thresholds T_1, T_2 and T_3 to separate the four peaks in PDF $f(n)$:

$$T_1 = -t_0 \quad T_3 = t_0 \quad T_2 = t$$

where

$$f(t) = min(f(n)) \quad (t_1 \le n \le t_2)$$
$$f(t_1) = max(f(n)) \quad (-t_0 \le n \le 0)$$
$$f(t_2) = max(f(n)) \quad (0 < n \le t_0)$$
$$t_0 = 50 \text{ (short sessions) or } 5000 \text{ (long sessions)}$$

Long sessions give us sufficient information to detect the fixed patterns. We propose *Diversity Factor* to measure how rigid the requesting behavior is:

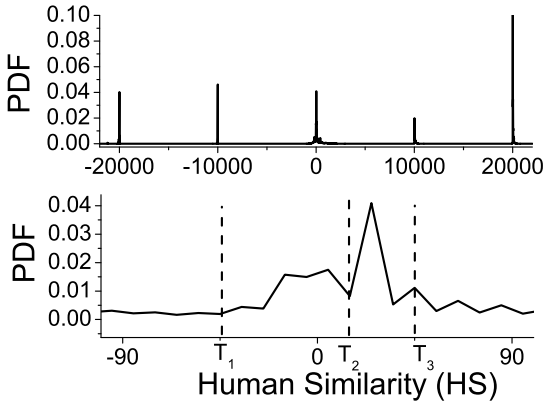

Fig. 1. The PDF of HS for all short sessions in a typical set of trace from LEARN. Upper: the overall distribution. Lower: the detailed distribution near zero.

Definition 2: Let N be the number of requests within one session, and $\{r_i\}, 1 \le i \le N$ be the sizes of requested objects within such session, we define $diversityfactor$ (DF) as the proportion of the count of non-zero unique object sizes to the count of non-zero objects:

$$DF = \frac{card_{1 \le i \le N}\{s | \exists s \in \mathbb{N}, s = r_i\}}{card_{1 \le i \le N}\{r_i | r_i \ne 0\}}$$

In the definition the notation $card\{\}$ refers the number of elements in the set. According to our repeated experiments, long sessions with $DF < 0.5$ are highly suspicious as from robots. It is also observed that, in PDF $f(n)$ when $0 < DF < 0.5$, there are always several obvious peaks. They may represent main types of long non-human sessions. In categorization we often want to include the biggest K peaks, therefore we propose an adaptive method to find the ideal threshold:

Let $f(n_1), f(n_2), ...f(n_K)$ be the K biggest peaks in $f(n)$ when $0 < n < 0.5$, and $k = max(n_1, n_2, ..., n_K)$, we have:

$$S = (a | f(a) < \beta f(k), \quad 0 < a < 0.5)$$
$$DFT = \begin{cases} sup\{S\} & S \ne \emptyset \\ 0.5 & S = \emptyset \end{cases}$$

where β is the peak decaying factor, usually 0.2 - 0.3 in practice.

The reason why we choose object size as the 'signature' of web pages, other than URLs, to determine DF is mainly because the pervasive deployment of dynamic pages: URL remains unchanged while clients navigate among different pages. We also find that Auto Correlation Function (ACF) of intervals between clicks or requests is not ideal to separate robots and human users, for interarrival times can be largely affected by networking conditions.

Among long sessions in all our traces, we can find massive ones continuously fetching HTMLs, JSPs or ASPs, but never get embedded objects. We believe that almost all these sessions are from crawlers[14]. Such crawlers usually have long durations and high DFs (because they fetch various pages).

Besides, since quite few modern web sites are composed of pure HTMLs, normal human users would not fetch a large portion of HTML pages. However, crawlers (possibly search engines) would ignore embedded objects in order to minimize throughput. To detect such pattern, we have introduced HTML Affinity (HA):

Definition 3: For a given session, let $\{J_i\}$ denote the set of requested objects within such session, HTML Affinity (HA) is defined as:

$$HA = \frac{card\{J_i | TYPE(J_i) \in \{\text{html,htm,jsp,asp}\}\}}{card\{J_i\}}$$

where $TYPE(J_i)$ represents the file type of object J_i.

Similarly, we use a threshold HAT when applying HA. In practical, 0.65 is a typical HAT value to detect suspicious sessions.

C. Automatic Session Categorization Scheme

Combing the traditional and new metrics mentioned above, we design an automatic session categorization scheme. As shown in Figure2, the scheme categorizes a specific session set as follows:

Firstly, the given session set is divided into short ones and long ones, according to the object count per session (*objcnt*). This is necessary because short sessions contain so limited information that it is impossible to apply DF and HA.

By applying HS value together with thresholds T_1, T_2 and T_3, short sessions will be further divided into four categories, i.e. short-p2sp-f, short-p2sp-s, crawler-f and crawler-s (the postfix -f stands for "failure" and -s stands for "success"). Short-p2sp-f and short-p2sp-s sessions have unusual HS values (extremely low or high, depending on whether they fail or succeed), due to the severe punishment to sessions starting with requests to hot type objects. The failed ones fall into the short-p2sp-f part, and the successful ones fall into the short-p2sp-s part. The rest of the sessions with HS in the scoring range(T_1, T_3) are almost all gentle crawlers. Even these crawlers request only one or two objects in each session, they are large in number and will take a considerable portion of the aggregate traffic (see Table III), thus could not be ignored in traffic analysis.

For long sessions ($objcnt > 2$), we combine HS, DF and HA to perform reliable division among them. The P2SP sessions (long-p2sp-s and long-p2sp-f), for they have extremely high or low HS values, are filtered out first. Next, sessions with DFs lower than the threshold are considered to be Robots. After that, we use HA to catch long-crawler sessions in the rest of the long sessions. Note that HA takes not only .htm(l), but also .jsp and .asp into consideration, which are all favored by modern crawlers.

Finally we can figure out the behavioral patterns of real human user sessions: the requests contained in the session are enough; *HS* value is neither too large nor too small; requested objects are diverse so that DF is high enough; and HA is not very high. If one or more of these criteria are not satisfied, the session will be classified as non-human ingredients such as P2SP, robot, or crawler.

V. Experimental Results

A. Scheme Validation

In order to check if our scheme reaches its design goals, we apply it to various traces from PORTAL, VSPACE, LEARN and XK in various time periods, which have totally different aggregate characteristics. The categorization results are shown in Table III.

According to the protocol, crawlers should firstly fetch `/robots.txt`. We want to check if all such 'self-confessed' crawler sessions are correctly identified. Results show that, 1891 out of 521731 sessions try to fetch `/robots.txt` and they are all identified as crawler sessions successfully.

The client IP addresses recorded in raw traces also serve as 'labels' to validate our scheme. For instance, we know crawlers on PORTAL are almost outside of the campus and on-campus users are less likely to access PORTAL via P2SP tools. By checking IP addresses of non-human sessions we find only 2.8% non-human sessions are from on-campus users, in which 0.15% are crawler sessions, while on-campus users issue 10% sessions of all. We also check IP addresses of all crawler sessions and find that more than 90% sessions come from less than 10% class C subnets, which is also consistent with our knowledge regarding crawlers.

We also validate our scheme in a quantitative way. In order to study the characteristics of sessions before and after categorizations, we measure the corresponding session level metrics, and draw their CDFs (Cumulative Distribution Function).

We observe that, on a specific server (LEARN, PORTAL or VSPACE), for two trace sets, uncategorized sessions are highly different, which could be reflected by their overall CDFs. However, after categorization, metrics for one specific category from both sets show high consistency, even the time span is as long as days, weeks and even years and their intensities largely vary (shown in Figure 3). In Figure 4, the CDF curves suggest that human session characteristics perfectly matched each other, although the composition of traffic ingredients are totally different. The small discrepancies between CDFs in the right sub-figure might be ascribed to the insufficient crawler samples in 2004, and to the change in the sorts of crawlers.

B. Characteristics of Categories

Table III presents an overview of the proportions among all traffic ingredients. It contains a brief summary of the counts and percentages of sessions in traces analyzed above. From this table, we can clearly see the dramatic changes in HTTP traffic in recent 2 to 3 years. Generally speaking, at least before 2004, human users dominated web servers (as confirmed by results from EPA). However, nowadays the traffic composition is more complicated. One of the most remarkable ingredients is P2SP. The comparison between VSPACE 2004 and 2006 have shown that, in two years the absolute number of human sessions increased slightly, while their percentage has significantly dropped from 89% to 4%. This is mainly caused by massive failed P2SP sessions (rejected by server's authentication). Overall, P2SP sessions (including successful

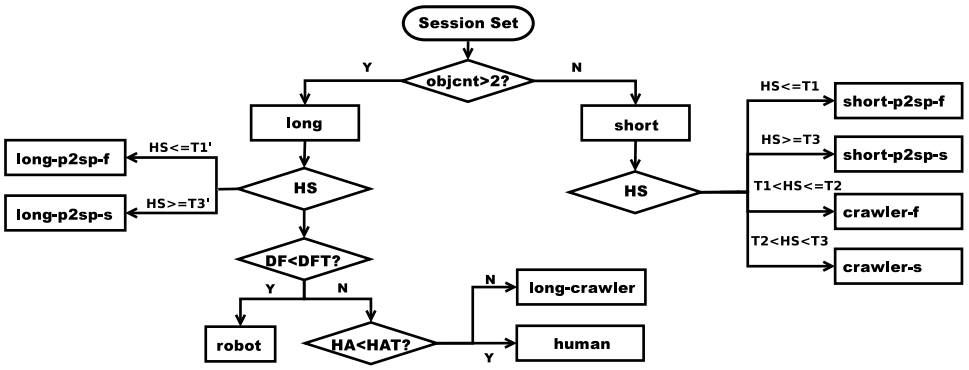

Fig. 2. Flowchart of the Categorization Scheme

TABLE III
THE STATISTICS OF THE COUNTS OF CATEGORIZED SESSIONS

Category	LEARN 2004/09/13 1 day	LEARN 2005/09/15 1 day	PORTAL 2007/09/15 4am 4 hours	PORTAL 2007/09/15 8am 4 hours	VSPACE 2004/03/21 1 week	VSPACE 2006/11/09 1 week	EPA 1995/08/29 1 day
human	10835(35.2%)	11719(22.6%)	867(16.7%)	5888(41.8%)	16705(89.1%)	17063(4.3%)	2586 (73.3%)
long-crawler	826(2.7%)	1621(3.1%)	484(9.3%)	640(4.5%)	411(2.2%)	483(0.1%)	120 (3.4%)
short-p2sp-s	7229(23.5%)	19851(38.3%)	520(10.0%)	2199(15.6%)	154(0.8%)	3996(1.0%)	128 (3.6%)
short-p2sp-f	4820(15.7%)	5382(10.4%)	122(2.3%)	359(2.5%)	199(1.1%)	166191(41.8%)	10 (0.3%)
long-p2sp-s	3281(10.7%)	6871(13.2%)	209(4.0%)	1227(8.7%)	198(1.1)	1262(0.3%)	73 (2.1%)
long-p2sp-f	1207(3.9%)	1130(2.2%)	65(1.2%)	177(1.3%)	323(1.7%)	206657(52.0%)	20 (0.6%)
robot	486(1.6%)	574(1.1%)	42(0.8%)	131(0.9%)	234(1.2%)	482(0.1%)	27 (0.8%)
crawler-s	1144(3.7%)	2608(5.0%)	2200(42.3%)	2761(19.6%)	475(2.5%)	1056(0.3%)	504 (14.3%)
crawler-f	925(3.0%)	2105(4.1%)	697(13.4%)	715(5.1%)	49(0.3%)	349(0.1%)	59 (1.7%)
Total	30753(100%)	51861(100%)	5206(100%)	14097(100%)	18748(100%)	397539(100%)	3527 (100%)

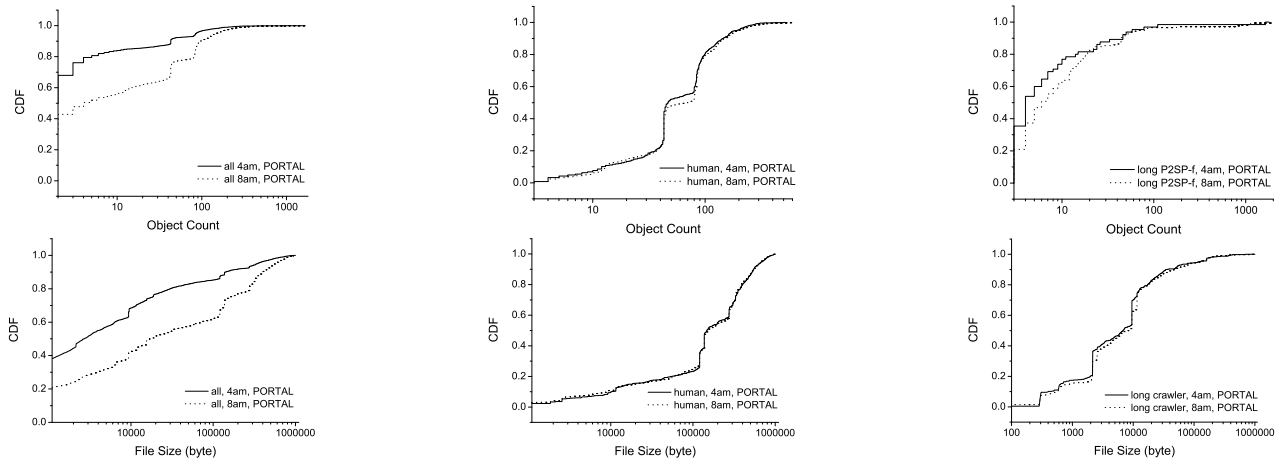

Fig. 3. Cumulative Distributions of Metrics for all (left), human (middle), long failed P2SP and long crawler (right) sessions, in two 4-hour periods of the same day from PORTAL. (upper: object count per session; lower: total requested object size per session)

and failed sessions) takes a large portion in the aggregate traffic (about 30% – 40% or even more), and their behavioral patterns fluctuate much from day (very active) to night (less active). Unlike crawlers, P2SP traffic is similar to human in this sense. We ascribe this surprising feature to the fact that P2SP sessions are initiated by P2P clients such as Xunlei and Bitcomet, which are driven by human users.

The crawlers keep increasing in recent years and show special interests to very popular sites like PORTAL. We also find that their absolute number does not fluctuate much in different time periods. The percentages of crawler sessions are notable at night, but diluted by human sessions in daytime. Therefore, within a long period of time (weeks or months), the traffic from crawlers could be regarded as a kind of background noise in HTTP traffic with constant strength.

In our experiments, we have found that the percentage of robots is small (approximately 1%), and the average session arrival intervals are generally large (100 seconds in peak hours). We can infer that robot sessions are only a marginal part in various traffic ingredients nowadays.

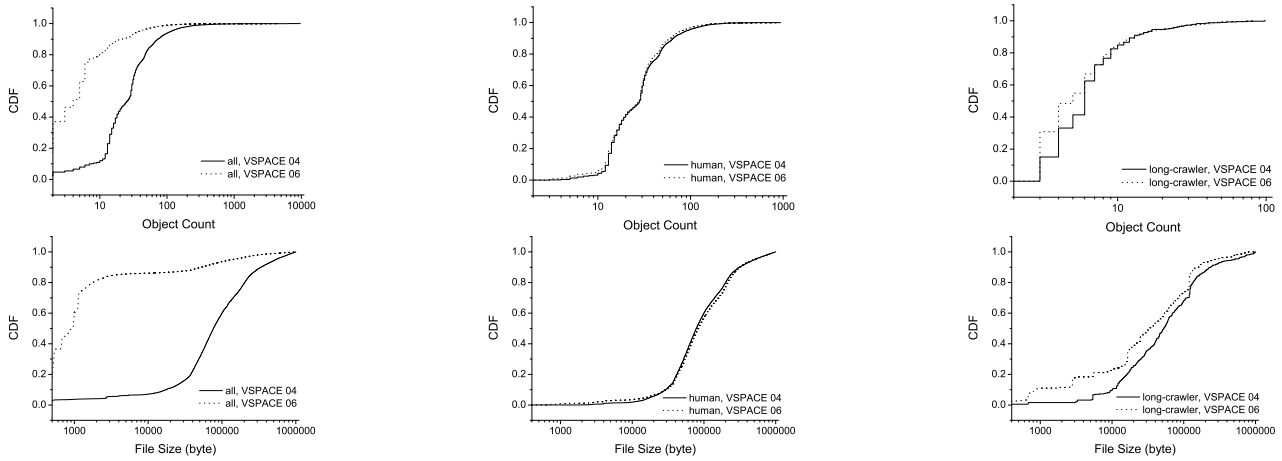

Fig. 4. Cumulative Distributions of Metrics for all (left), human (middle), long crawler (right) sessions, in two 1-week periods across three years from VSPACE. (upper: object count per session; lower: total requested object size per session)

VI. Conclusion and future works

In this paper we first analyze session characteristics of real HTTP traffic nowadays. We have found that the composition of HTTP traffic nowadays is more complicated, comparing with HTTP traffic earlier than 2004. We also argue that existing models could hardly describe the traffic on either session level or aggregate level. To address the challenge we designed an automatic scheme to classify sessions in HTTP traffic into several categories according to their features and sources. Results show that, for one specific server, sessions within the same category show highly consistent characteristics, even though such sessions might come from different time periods. On the other side, sessions from different categories are totally different in characteristics and performance implications.

Motivated by the success of categorization scheme, we propose to design a new type HTTP traffic generator working based on session-pools. It will rely on categorized sessions from specific servers instead of any assumptions of analytical model. Hopefully, our generator can synthesize highly realistic traffic, on both session level and aggregate level. Besides, the traffic generated will be completely tunable. Users could perform *what-if* test by tuning up or down certain kinds of sessions in the synthetic traffic, therefore to finer plan the system capacity and better predict server performance within specific context.

All data sets, figures and other analysis results, shown or not shown in the paper, are available upon request. In addition, we are also glad to provide our scripts and the full source code. We hope our category-based traffic analysis, could help researchers to improve web server design within modern conditions.

References

[1] Martin F. Arlitt and Carey L. Williamson. Web server workload characterization: The search for invariants. In *1996 ACM SIGMETRICS Conference*, pages 126–137. ACM, May 1996.

[2] Mark E. Crovella and Azer Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *Transactions on Networking*, 5(6):835–846, December 1997.

[3] Erich M. Nahum. Deconstructing specweb99. In *the Seventh International Workshop on Web Content Caching and Distribution (WCW)*. IEEE, August 2001.

[4] F. Donelson Smith, Felix Hernandez-Campos, Kevin Jeffay, and David Ott. What TCP/IP protocol headers can tell us about the web. In *SIGMETRICS/Performance*, pages 245–256, 2001.

[5] Cathy H. Xia, Zhen Liu, Mark S. Squillante, Li Zhang, and Naceur Malouch. Traffic modeling and performance analysis of commercial web sites. *SIGMETRICS Performance Evaluation Review, 2002.*, 30(3):32–34, December 2002.

[6] Cathy H. Xia, Zhen Liu, Mark S. Squillante, Li Zhang, and Naceur Malouch. Web traffic modeling at finer time scales and performance implications. *Performance Evaluation*, 61(8):181–201, 2004.

[7] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world wide web. *Computer Networks and ISDN Systems*, 26(6):1065–1073, December 1995.

[8] Daniel A. Menascé, Virgilio A. F. Almeida, Rodrigo Fonseca, and Marco A. Mendes. A methodology for workload characterization of e-commerce sites. In *EC '99: Proceedings of the 1st ACM conference on Electronic commerce*, pages 119–128, New York, NY, USA, 1999. ACM.

[9] Daniel A. Menascé, Virgilio A. F. Almeida, Rodrigo Fonseca, and Marco A. Mendes. Business-oriented resource management policies for e-commerce servers. *Performance Evaluation*, 42(2-3):223–239, 2000.

[10] Hyoung-Kee Choi and John O. Limb. A behavioral model of web traffic. In *ICNP*, page 327, Los Alamitos, CA, USA, 1999. IEEE Computer Society.

[11] Martin F. Arlitt. Characterizing web user sessions. *SIGMETRICS Performance Evaluation Review, 2000.*, 28(2):50–63, September 2000.

[12] Daniel Menascé, Virgílio Almeida, Rudolf Riedi, Flávia Ribeiro, Rodrigo Fonseca, and Jr. Wagner Meira. In search of invariants for e-business workloads. In *EC '00: Proceedings of the 2nd ACM conference on Electronic commerce*, pages 56–65, New York, NY, USA, 2000. ACM.

[13] Zhen Liu, Nicolas Niclausse, and César Jalpa-Villanueva. Traffic model and performance evaluation of web servers. *Performance Evaluation*, 46(2-3):77–100, October 2001.

[14] Virgílio Almeida, Daniel Menascé, Rudolf Riedi, Flávia Peligrinelli, Rodrigo Fonseca, and Wagner Meira Jr. Analyzing robot behavior in e-business sites. *SIGMETRICS Performance Evaluation Review, 2001.*, 29(1):338–339, June 2001.

[15] Marios Dikaiakos, Athena Stassopoulou, and Loizos Papageorgiou. Characterizing crawler behavior from web server access logs. In *The 4th International Conference on E-Commerce and Web Technologies (EC-Web 03)*, pages 369–378. Springer, September 2003.