Discovering Web Workload Characteristics through Cluster Analysis

Fengbin Li, Katerina Goševa-Popstojanova, and Arun Ross
Lane Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV 26506-6109
{fengbinl, katerina, ross}@csee.wvu.edu

Abstract

In this paper we present clustering analysis of session-based Web workloads of eight Web servers using the intra-session characteristics (i.e., number of requests per session, session length in time, and bytes transferred per session) as variables. We use K-means algorithm and the Mahalanobis distance, and analyze the heavy-tailed behavior of intra-session characteristics and their correlations for each cluster. Our results show that clustering provides an efficient way to classify tens or hundreds thousands of sessions into several coherent classes that efficiently describe Web workloads. These classes reveal phenomena that cannot be observed when studying the workload as a whole.

1. Introduction

World Wide Web, the largest distributed system ever built, has become part of the fabric of our society. Its tremendous growth has changed the way of life and also brought huge challenges to Web site designers, content producers, and maintainers. The realistic and accurate characterization of Web workloads is the first essential step in the areas such as performance analysis and prediction, capacity planning, and admission control.

Understanding the nature and characteristics of Web workloads is fundamental to the goal of improving Web performance. In our earlier work [6] we introduced several inter-session (i.e., number of sessions per user, number of sessions initiated per day and per hour) and intra-session (i.e., number of requests per session, session length in time, and bytes transferred per session) characteristics which collectively describe session-based Web workload. Furthermore, we identified the invariants of the Web workload that applied through the eleven Web servers considered. Our results showed that session-based workload and reliability are better indicators of the users perception of the Web quality than the request-based metrics. In [7] we presented more detailed and rigorous statistical analysis of Web workloads based on empirical data extracted from the actual logs of four Web servers. Our results showed that all considered Web servers have long-range dependant request arrival process and Web session arrival process. The results also showed that under the Pareto model, intra-session characteristics for some data sets exhibit heavy-tailed behavior.

During our work in [7], we found that for most Web servers, session-based workloads showed several similar interesting patterns. First, the majority of sessions tended to have short session length in time, a few requests and a small number of bytes transferred. Then, very long sessions in time units tended to have more requests and bytes transferred. However, there were long sessions with large amount of bytes transferred, but a few requests. These sessions most likely represent users who were downloading files, images, or videos.

In this paper we present clustering analysis of session-based Web workloads of eight Web servers using the intra-session characteristics as variables. Our results show that clustering provides an efficient way to classify tens or hundreds thousands of sessions into several coherent classes that efficiently describe Web server workloads.

The paper is organized as follows. First, we discuss the related work in section 2. Then, we introduce the background on K-means clustering in section 3. In section 4, we present the clustering results on data extracted from eight Web servers. Finally, in section 5 we summarize the main observations and present the concluding remarks.

2. Related work

In order to improve the Web performance and better serve the user needs, a solid understanding of user sessions is essential. A session is defined as a sequence of requests from the same user during a single visit to the Web site. Till now considerable amount of research work has been focused on characterizing Web user sessions for different purposes such as capacity planning and finding user navigational patterns. Arlitt in [2] presented a detailed characterization of user sessions of the 1998 World Cup Web site. The author studied the effect of a wide range of session timeout values on numerous user session characteristics and showed

Sixth IEEE International Symposium on Network Computing and Applications (NCA 2007)
0-7695-2922-4/07 $25.00 © 2007 IEEE

how these characteristics can be utilized in improving Web server performance. However, the author did not use cluster analysis method for the characterization of Web workload.

In [3], clustering analysis was used to describe classes of users and robots for the purpose of capacity planning. The sessions were partitioned based on their resource demands and performance impact on the system. Each session was described by the fractions of cacheable, non-cacheable, and search requests. An Euclidean distance metric was used to compute the distance between two sessions. The authors used K-means clustering algorithm and chose the optimal number of clusters according to β_{cv}, the ratio of the coefficient of variation of the intra-cluster distance to the coefficient of variation of the inter-cluster distance. Again, the analysis was based on a single e-commerce site.

Menascé et. al in [8] used clustering to characterize the workloads of two e-commerce sites in order to find groups of customers that exhibit similar navigational patterns. Each request of a customer session was associated to a state (e,g, home page, browse, search, select, add to cart, and pay). If a session has n states, then it was described as a tuple $S = (C, W)$, where $C = [c_{i,j}]$ is an $n \times n$ matrix of transition counts between states i and j of the session, and $W = [w_{i,j}]$ is an $n \times n$ matrix of accumulated server-side think times between states i and j of the session. The authors used K-means clustering algorithm with Euclidian distance. The selection of the optimal number of clusters was based on four indexes.

In [10] the authors used clustering to discover different learning behaviors in a given on-line course. The similarity between two Web sessions was computed by finding the best matching between the two sequences using dynamic programming techniques. They used ROCK, CHAMELEON and TURN on their categorical data [10].

Clustering was used in the previous research works to provide understanding of session-based Web workload for different purposes. However, using Euclidian distance to measure the similarity between sessions [2], [3], [8] does not take into account the correlation of the data, and using β_{cv} as a criterion to choose the optimal number of clusters [3], [8] is not accurate as we explain later in the paper. Furthermore clustering was applied on a single data set [3] or at most two data sets [8]. In this paper, we apply K-means clustering to empirical data collected from eight Web servers which allow us to generalize the observations. Different from the related work, we consider the number of requests per session, session length in time, and bytes transferred per session as variables, and use the Mahalanobis distance instead of Euclidian distance to measure the similarity between two sessions. In addition, we explore the heavy-tailed behavior and correlations between intra-session characteristics for the whole data sets, as well as for each cluster which has not been addressed in other research works so far.

3. Background on K-means clustering

In this paper, we use the K-means clustering algorithm [5] which works well on large data sets due to its linear time complexity. Its computational efficiency makes it a primary choice over other clustering methods, such as hierarchical methods, when large amount of data is concerned. The result of K-means clustering depends on the similarity measure, which defines the distance between the objects. Our earlier work [7] showed that the intra-session characteristics are correlated, (e.g., long sessions tend to have many of requests and large amount of bytes transferred). Therefore, we choose the Mahalanobis distance measure which corrects for interdependence of the variables. Furthermore, Mahalanobis distance is scale-invariant which benefits our analysis since the considered intra-session characteristics have different scales.

Let X and Y be two random variables of the same distribution with the covariance matrix Σ, then the Mahalanobis distance between $x \in X$ and $y \in Y$ is defined as

$$d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}.$$

Before conducting K-means clustering, the number of clusters k should be specified a priori. Different k will lead to different clustering results. To choose the best k is always a challenging task, especially when we do not have much information on the data to be analyzed. In [8] the authors used four indexes collectively to choose the best k: coefficient of variation for the intra-cluster distance (CV_{intra}), coefficient of variation for the inter-cluster distance (CV_{inter}), ratio between the intra- and inter-cluster variance (β_{var}) and ratio between the intra- and inter-cluster coefficient of variation (β_{cv}). The smaller the values of CV_{intra}, β_{var} and β_{cv}, the better the k is. However, it should be noticed that the coefficient of variation is the ratio of the standard deviation to the mean. In our data, some of the means of the intra- and inter-cluster distances are close to 0 which makes CV_{intra}, CV_{inter} and β_{cv} inaccurate indexes for deciding the number of clusters. Hence, in this paper we use β_{var} as a criterion for choosing the best k.

Let $C_1, C_2, \ldots, C_k$ be k disjointed clusters of a given dataset, and c_i and n_i be the centroid and size of C_i respectively, where $i = 1, 2, \ldots, k$. Then the variance of intra-cluster distance σ^2_{intra}, and the variance of inter-cluster distance σ^2_{inter} are computed as follows:

- Intra-cluster: $\sigma^2_{intra} = \frac{1}{k-1} \sum_{i=1}^{k} (d_i - \bar{d})^2, k > 1$, where $d_i = \frac{1}{n_i} \sum_{x \in C_i} d(x, c_i)$,

- Inter-cluster: $\sigma^2_{inter} = \frac{1}{\binom{k}{2}} \sum_{i=1}^{k} \sum_{j=i+1}^{k} (D_{i,j} - \bar{D})^2$, $i \neq j, k > 2$, where $D_{i,j} = d(c_i, c_j)$,

and $\bar{d}$ and $\bar{D}$ are the means of d_i and $D_{i,j}$ respectively. Then $\beta_{var} = \sigma^2_{intra} / \sigma^2_{inter}$. We select k for which β_{var}

Sixth IEEE International Symposium on Network Computing and Applications (NCA 2007)
0-7695-2922-4/07 $25.00 © 2007 IEEE

is the smallest, since the goal of clustering is to minimize the variance of the intra-cluster distance and to maximize the variance of the inter-cluster distance.

4. Clustering Results

The Web logs used in this paper were obtained from eight Web servers: six Web servers at the NASA Independent Verification and Validation Facility (NASA-Pub1, NASA-Pub2, NASA-Pub3, NASA-Pvt1, NASA-Pvt2, and NASA-Pvt3), Web server of the Lane Department of Computer Science and Electrical Engineering (CSEE), and university wide Web server at West Virginia University (WVU). Table 1 presents the time period, total number of requests, total number of sessions, and total kilobytes transferred for the Web servers analyzed in this paper. From Table 1 the total number of requests, total number of sessions, and total kilobytes transferred vary by four orders of magnitude on different servers.

We estimate the value of β_{var} for $k = 3, 4, \ldots, 20$, and choose the k with the smallest β_{var} to conduct the clustering analysis. As an illustration, Figure 1 plots β_{var} for different number of clusters k for NASA-Pub1 server. The plot shows an obvious decreasing tendency of β_{var}. From k=3 to k=9, β_{var} decreases dramatically. Starting from k=9, β_{var} shows a very slow decay. Therefore, we choose k=9 for the number of clusters. The shape of β_{var} as a function of k was similar for all Web servers. The values of k, chosen in a similar way, are given in the last column of Table 1. It should be noticed from Table 1 that the optimal number of clusters k is in the range of $4 \leq k \leq 13$ and is not depended on the workload. Thus, Web servers with higher workload not necessarily have higher k. For example, CSEE has the second largest number of sessions, but it has the smallest number of clusters. On the other hand, NASA-Pvt2 has the second lowest number of sessions, but it has the largest number of clusters.

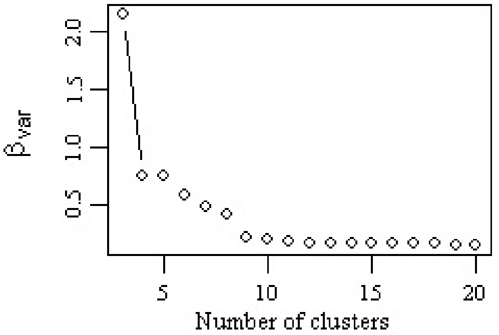

Figure 1. β_{var} for NASA-Pub1

After we choose the optimal number of clusters for each Web server, we apply K-means clustering to the empirical data. As an illustration, we present the detailed results for NASA-Pub1 server. In four weeks time period this server had 4,757 sessions, which K-means split into 9 clusters plotted in 3 dimensional space in Figure 2. For each cluster, Table 2 presents the percentage of sessions, the minimum, median and maximum of the number of requests per session, session length in seconds, and kilobytes transferred per session in that cluster. Note that throughout the paper the clusters are ordered decreasingly by their size, i.e., cluster 1 always denotes the largest cluster, cluster 2 denotes the second largest cluster, and so on.

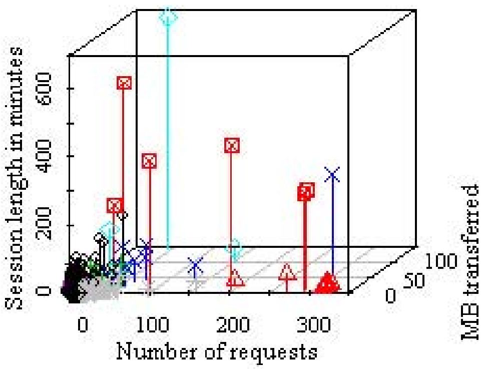

Figure 2. 3D plot of NASA-Pub1 clusters

As it can be seen from Table 2, cluster 1 contains the majority (70.53%) of sessions of NASA-Pub1. The sessions in cluster 1 have very small number of requests, session length in seconds, and kilobytes transferred compared with sessions in other clusters. Thus, the number of requests per session is at most 12. The sessions are very short, no more than 16 minutes, and there are no more than 2 megabytes transferred per session. These sessions represent users who usually go to the Web site, visit several pages, then leave the Web site.

Cluster 6 contains 0.65% of sessions of NASA-Pub1. These sessions have very large number of requests (at least 192 requests per session), moderate session length in time, and many kilobytes transferred. Although half of the sessions in cluster 6 are no longer than 3.23 minutes, they have at least 192 requests per session. This implies that for almost half of the sessions in cluster 6, there is at least one request per second. This phenomenon is due to the fact that 29 among 31 sessions in cluster 6 are due to robot visits.

Some other interesting clusters should also be noticed. For example, cluster 8 consists of 0.13% of sessions of NASA-Pub1. These sessions are long at least 4 hours. Among six sessions in this cluster, four are robots. Cluster 9 consists of 0.06% of sessions of NASA-Pub1, which have moderate number of requests, long session length in time, and very large amount of bytes transferred. These sessions are in the tail of distributions of session length in time and bytes transferred, but not in the tail of the distribution of the number of requests. These users were mainly downloading pdf, Power Point, and zip files from the Web server.

The percentages of sessions in each cluster for all eight

Data set	Time period	Requests	Sessions	KB transferred	k
NASA-Pub1	19-Sep-05 to 16-Oct-05	23,896	4,757	4,155,210	9
NASA-Pub2	19-Sep-05 to 16-Oct-05	131,052	14,331	3,193,722	10
NASA-Pub3	19-Sep-05 to 16-Oct-05	15,134	2,696	463,482	6
NASA-Pvt1	19-Sep-05 to 16-Oct-05	4,838	166	175,518	7
NASA-Pvt2	19-Sep-05 to 16-Oct-05	12,127	686	21,874	13
NASA-Pvt3	19-Sep-05 to 16-Oct-05	61,377	3,400	509,398	5
CSEE	07-Feb-05 to 13-Feb-05	323,219	46,607	13,141,292	4
WVU	02-Feb-04 to 08-Feb-04	14,343,952	184,847	39,824,799	9

Table 1. Summary of the raw data

	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9
perc_ses	70.53%	13.56%	11.37%	2.75%	0.67%	0.65%	0.27 %	0.13 %	0.06%
Req_min	1	1	2	10	1	192	13	33	5
Req_med	1	2	2	17	13	310	51	147	43
Req_max	12	22	54	152	56	324	292	292	147
Len_min (sec)	0	0	675	0	0	138	297	14,780	2,042
Len_med (sec)	0	0	1,082	42	602	194	2,986	20,316	6,734
Len_max (sec)	947	1,429	12,351	2,638	4,288	2,856	17,470	33,779	41,040
KB_min	0	888	0	7	5,103	6,613	23,391	2,820	82,483
KB_med	96	1,523	18	384	8,140	24,392	38,113	9,257	108,969
KB_max	2,125	5,412	24,312	10,566	21,022	29,566	68,415	64,443	148,461

Table 2. Summary of intra-cluster characteristics for clusters of NASA-Pub1

Web servers using the value for k in Table 1 are summarized in Table 3. There are no more than 5 sessions in each of the clusters 7, 8 and 9 of WVU dataset, so the percentages in Table 3 are given as 0.

The following important observations can be made from Table 3. For all servers, except NASA-Pvt1 and NASA-Pvt2, there is one very large cluster containing the majority of sessions (56.9%-93.27%). As in case of NASA-Pub1, the largest cluster consists of sessions with small number of requests, short session length in time, and small amount of bytes transferred. In general, K-means clustering does not always produce a large cluster which dominates the other clusters in size. The phenomenon that NASA-Pub1, NASA-Pub2, NASA-Pub3, NASA-Pvt3, CSEE and WVU all have a large, dominant cluster is due to the nature of data. NASA-Pvt1 and NASA-Pvt2 do not have a dominant cluster. Instead, they have three clusters with close sizes respectively. These two NASA servers are private sites used by certain domain of users. They are implemented to provide quick references and limited functionalities. The nature of these two private sites explains the lack of a dominating cluster present in case of all the other Web servers. It should be noticed that although NASA-Pvt3 is also a private site, it has higher workload than the two of the NASA public sites, NASA-Pub1 and NASA-Pub3. Furthermore, NASA-Pvt3 is used frequently by significant number of users to do scheduling and other daily tasks. Therefore, the clustering results for NASA-Pvt3 server resemble the results obtained for the public servers – the majority of sessions belong to one large cluster close to the origin in the 3 dimensional plot.

As it can be seen from our analysis, clustering provides an efficient way to classify tens or hundreds thousands of sessions into several coherent classes that efficiently describe Web server workloads. So, instead of looking at huge number of sessions, one can group the sessions into coherent clusters and study the characteristics of the clusters. For example, the CSEE Web server has 46,607 sessions, and the largest cluster (CL1) contains 93.27% of the sessions which is significantly higher percentage compared to the other servers. The minimum and median of the number of requests in CL1 are both 1, i.e., at least half of the sessions in CL1 have only one request. The median of session length is less than one second, and the median of KB transferred is 5. This is not surprising since whenever a user makes a request to the CSEE Web server, the request is redirected to another server (CEMR), and only in some cases, such as visiting homepages of faculty and students in the department, returns to the CSEE Web server. Another interesting cluster of CSEE is the smallest cluster CL4. The sessions in CL4 have 565MB to 699MB transferred . These users are mainly downloading image files (.IMG and .ISO), and the Debian Linux packages from the server.

The results presented in this section can be summarized as follows. (1) For most Web servers analyzed, majority of sessions (56.9%-93.27%) belong to a cluster characterized

Data set	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL10	CL11	CL12	CL13
NASA-Pub1	70.53	13.56	11.37	2.75	0.67	0.65	0.27	0.13	0.06				
NASA-Pub2	69.67	16.66	9.85	1.88	0.75	0.46	0.36	0.35	0.02	0.01			
NASA-Pub3	56.90	30.60	9.98	1.34	1.08	0.11							
NASA-Pvt1	34.34	28.92	18.67	6.63	4.82	4.82	1.81						
NASA-Pvt2	24.78	23.32	20.41	8.02	4.96	4.23	3.79	3.50	3.21	1.46	1.02	0.73	0.58
NASA-Pvt3	77.71	14.76	4.50	2.21	0.82								
CSEE	93.27	6.63	0.09	0.01									
WVU	58.21	32.91	7.67	0.76	0.42	0.02	0	0	0				

Table 3. Percentage of sessions in clusters

by a small number of requests, short session length in time units, and small amount of bytes transferred. (2) For NASA-Pvt1 and NASA-Pvt2 servers, there are three clusters with close sizes due to the nature of the data. (3) Different small size clusters characterize different Web workload phenomena, such as long sessions that have moderate number of requests and bytes transferred, or long sessions that have large number of requests and bytes transferred.

4.1 Heavy-tailed behavior of intra-session characteristics

In this section we explore the heavy-tailed behavior of the intra-session characteristics. For each Web server, we fit Pareto distribution to the whole dataset and to the clusters obtained from K-means clustering. We only fit Pareto to those clusters with more than 100 sessions to ensure the accuracy of the distribution fitting.

A random variable X, with cumulative distribution function F(x), is said to be heavy-tailed if $1 - F(x) = x^{-\alpha}L(x)$, where $L(x)$ is slowly varying as $x \to \infty$, i.e., $lim_{x\to\infty}L(\alpha x)/L(x) = 1$ for $\alpha > 0$ [9]. That is, regardless of the behavior for small values of the random variable, if the asymptotic shape of the distribution is hyperbolic, it is heavy-tailed. The simplest heavy-tailed distribution is Pareto, which is hyperbolic over its entire range. The classical Pareto distribution with shape parameter α (also known as tail index) and location parameter k has the cumulative distribution function $F(x) = 1 - (k/x)^{\alpha}$. If $1 < \alpha \leq 2$ the distribution has a finite mean and infinite variance. If $\alpha \leq 1$ the distribution has infinite mean and variance.

To estimate the tail index α of a Pareto distribution we use the log-log complementary distribution (LLCD) plots [1], [4]. These are plots of the complementary cumulative distribution function (CCDF) $P[X > x] = 1 - F(x)$ on log-log axes. If the distribution is Pareto, the tail of the LLCD plot will be approximately linear. Then, we use least-square regression to estimate the slope, which is equal to $-\alpha$.

Figure 3 shows the LLCD plots of bytes transferred per session for the whole WVU dataset and WVU cluster 5. The least square regression estimate of the heavy tail index for WVU whole dataset is $\alpha = 1.43$ with standard error $\sigma = 0.0003$. The coefficient of determination (R^2) is 0.997, which indicates a very good fit between the empirical and mathematical distributions. The heavy tail index for the cluster 5 is $\alpha = 0.75$, which implies that the distribution of bytes transferred in cluster 5 has infinite mean and variance.

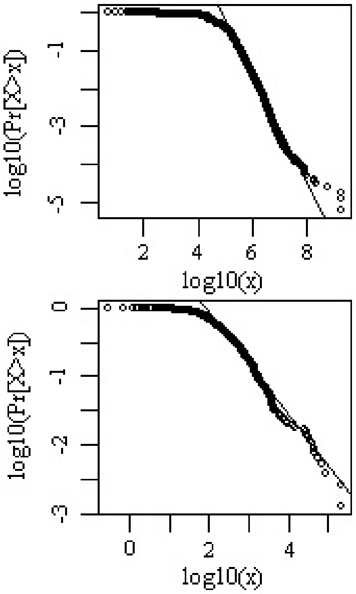

Figure 3. LLCD plot of bytes transferred for WVU, whole dataset (top) and cluster 5 (bottom)

Table 4 summarizes the α values for each intra-session characteristic for seven Web servers. We do not provide the values for NASA-Pvt1 due to its small data size. It should be noted that although some clusters have more than 100 sessions, large amount of sessions within these clusters have the same value of intra-session characteristics. For example, among 160 sessions in cluster 2 of NASA-Pvt2, 121 have the same number of requests, 14, per session. These cases were not considered in distribution fitting and are annotated with NA.

As it can be seen from Table 4, the number of request

			α_R	α_L	α_B
	whole	4,757	1.32	1.48	1.18
	CL1	70.53%	NA	11.08	4.25
	CL2	13.56%	2.24	2.39	7.68
NASA-	CL3	11.37%	1.65	2.96	1.58
Pub1	CL4	2.75%	2.78	2.11	1.74
	whole	14,331	1.76	1.97	1.10
	CL1	69.67%	2.75	8.61	1.48
	CL2	16.66%	7.16	3.56	2.58
	CL3	9.85%	4.02	3.16	1.45
NASA-	CL4	1.88%	12.42	2.52	1.95
Pub2	CL5	0.75%	1.84	2.74	3.36
	whole	2,696	1.64	1.25	0.73
	CL1	56.90%	1.58	26.16	1.78
NASA-	CL2	30.60%	1.64	1.95	0.86
Pub3	CL3	9.98%	5.02	1.30	1.68
	whole	686	2.51	3.35	2.82
	CL1	24.78%	3.85	2.93	2.33
NASA-	CL2	23.32%	NA	NA	NA
Pvt2	CL3	20.41%	NA	NA	NA
	whole	3,400	1.27	2.89	1.06
	CL1	77.71%	2.23	5.21	3.41
NASA-	CL2	14.76%	2.21	14.79	2.19
Pvt3	CL3	4.50%	2.90	2.92	2.52
	whole	46,607	1.71	2.16	0.88
	CL1	93.27%	2.08	4.56	0.97
CSEE	CL2	6.63%	1.98	2.61	0.99
	whole	184,847	2.17	1.86	1.43
	CL1	58.21%	2.84	16.91	1.99
	CL2	32.91%	8.22	6.41	1.88
	CL3	7.67%	5.08	2.25	2.14
	CL4	0.76%	2.19	2.40	1.91
WVU	CL5	0.42%	0.82	6.08	0.75

Table 4. Fitting Pareto to data

per session of seven Web servers is reasonably well modeled by Pareto distribution with $1.27 \leq \alpha_R \leq 2.51$ for the whole dataset. Except for NASA-Pvt2 and WVU, the number of request is heavy-tailed with finite mean and infinite variance. The tail index of Pareto model for the largest cluster is in the range $1.58 \leq \alpha_R \leq 3.85$. The tail indexes for the largest clusters are greater than 2, except for the largest cluster of NASA-Pub3, which implies that the number of request is not heavy-tailed for most of the largest clusters. For clusters other than the largest one, some are heavy-tailed, some are not.

The session length in time is also well modeled by a Pareto distribution with $1.25 \leq \alpha_L \leq 3.35$ for the whole datasets. Although many servers have heavy-tailed session length in time units, none of the largest clusters is heavy-tailed ($2.93 \leq \alpha_L \leq 26.16$). For other clusters, some are heavy-tailed, some not.

As shown in Table 4, bytes transferred per session is the most heavy-tailed intra-session characteristic; the tail index for whole dataset is in the range $0.73 \leq \alpha_B \leq 2.82$. The tail index for the largest cluster is in the range $0.97 \leq \alpha_B \leq 4.25$. For other clusters, some are heavy-tailed, some not.

From the analysis of the heavy-tailed behavior of the intra-session characteristics, we draw the following interesting remarks. (1) Pareto model fits well for all intra-session characteristics for all Web servers. (2) Intra-session characteristics of the whole dataset are heavy-tailed with $\alpha \leq 2$ for most Web servers, i.e., they have infinite variances. (3) For most Web servers, the intra-session characteristics of the largest cluster are not heavy-tailed ($\alpha > 2$). (4) Even when an intra-session characteristic for the whole dataset is not heavy-tailed, it may be heavy-tailed for some cluster (e.g., the number of request of WVU). (5) Small size clusters might have huge influence on the heavy-tailedness of the whole dataset. For example, the session length of the whole dataset for WVU is heavy-tailed with $\alpha_L = 1.86$, but none of the clusters 1-5 are heavy-tailed. The remaining clusters 6-9 consist only of 0.03% of sessions, but all of them are in the 99.9% quantile of the distribution of session length, i.e., they are in the extreme tail of the distributions. The heavy-tailedness of the whole dataset is mostly due to these sessions in the smallest clusters.

4.2 Correlations between intra-session characteristics

In this section we analyze the correlations between intra-session characteristics for the eight Web servers and their clusters. Intuitively, one might expect long sessions to have large number of requests and large amount of bytes transferred for all Web servers. However, during our early work [7] we found that many long sessions that are in the tail of the distribution of session length are in the body of the distribution of the number of requests. Namely, many long sessions do not have very large number of requests. Clustering can be an efficient way to discover and group these kind of sessions into clusters and study their behaviors.

Figure 4 shows the pairwise correlations between the number of requests (R), session length (L) and bytes transferred (B) for the whole WVU data and its clusters. (Clusters with less than 20 sessions are not presented.)

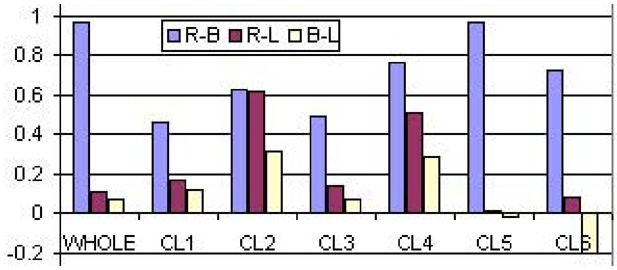

Figure 4. Correlations between intra-session characteristics for WVU

From Figure 4 the number of requests and bytes transferred are positively correlated for the whole WVU data and all the clusters. The correlation coefficient for the whole dataset is 0.97, which indicates a very strong linear relationship between these two intra-session characteristics. Most of the clusters have moderate to strong linear dependence between number of requests and bytes transferred ($0.46 \leq r_{R-B} \leq 0.97$). The correlation coefficient for the largest cluster (cluster 1) is 0.46, which indicates a moderate linear relationship. Figure 5 shows the scatter plots of bytes transferred vs. number of requests for the whole WVU dataset and the cluster 1. Notice that there are four points in the upper right corner of the scatter plot of the whole WVU dataset. These points are far away from the remaining 184,843 points and have huge impact on the correlation coefficient. When these four points are removed, the correlation coefficient for the rest of data decreases dramatically to 0.58. These four data points form a single cluster (cluster 8 not shown in Figure 4). From the bottom scatter plot, the correlation between the number of requests and bytes transferred for cluster 1 has a different direction and strength ($r_{R-B} = 0.46$) from the correlation for the whole dataset, which indicates that the correlation of the largest cluster does not necessarily guide the correlation of the whole dataset.

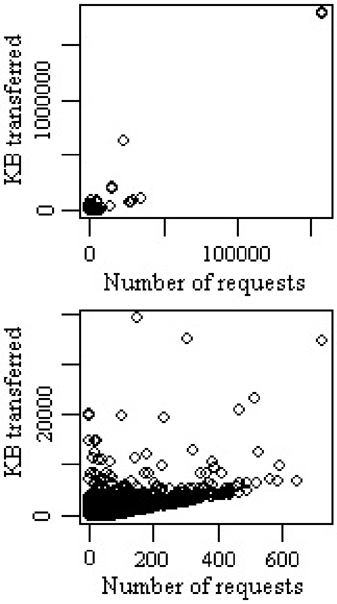

Figure 5. KB transferred vs. Number of requests for WVU (top) and WVU cluster 1 (bottom)

Compared to the number of requests and bytes transferred, the number of requests and session length have smaller linear dependence. Thus, the correlation coefficient for the whole dataset is 0.11 which indicates a very weak linear relationship. For the clusters the correlation coefficient is ranging from almost no linear dependence for cluster 5 ($r_{R-L} = 0.01$) to moderate dependence for cluster 2 ($r_{R-L} = 0.61$).

There is almost no linear relationship between the bytes transferred and session length for the whole WVU dataset ($r_{B-L} = 0.07$). The value of this correlation coefficients for the clusters is in the range $-0.2 \leq r_{B-L} \leq 0.31$. There is a small positive linear relationship between bytes transferred and session length for all the clusters, except clusters 5 and 6. The correlation coefficient for cluster 5 is -0.02, which indicates almost no linear dependence, while the correlation coefficient $r_{B-L} = -0.2$ for cluster 6 indicates a small negative linear relationship.

We repeat the same analysis on the other Web servers as well. The analysis shows that the correlation varies significantly across the Web servers and clusters. Figure 6 presents the correlation coefficients of intra-session characteristics for all Web servers and clusters.

In summary, we make the following remarks on the correlations. (1) The intra-session characteristics have positive pairwise correlations for the whole dataset as well as the largest clusters. (2) Out of total 45 clusters analyzed, only six clusters have negative pairwise correlations. (3) The correlation of the largest cluster does not guide the correlation of the whole dataset as illustrated on the example of WVU. (4) For six out of eight Web servers, the number of requests and bytes transferred have the highest correlation compared to the other pairwise correlation coefficients.

5. Concluding remarks

In this paper we have presented a detailed clustering analysis aimed at discovering the characteristics of session level Web workload based on data extracted from eight real Web servers. First, we applied K-means to session-based Web workloads using the intra-session characteristics (i.e., number of requests per session, session length in time, and bytes transferred per session) as variables. We used K-means algorithm which works well on large data sets due to its linear time complexity, and the Mahalanobis similarity measure which corrects for interdependence of the variables. Then, we studied the heavy-tailed behavior and the correlations of the intra-session characteristics for the whole datasets and the clusters obtained from K-means.

Since our analysis is based on data extracted from eight real Web servers, we are able to make the following generalized observations. (1) The number of clusters varies for different Web servers. In our data, the smallest k is 4 for CSEE and the largest k is 13 for NASA-Pvt2. (2) For most Web servers analyzed, majority of sessions (56.9%-93.27%) belong to a cluster characterized by a small number of requests, short session length in time units, and small amount of bytes transferred. This phenomenon is due to the nature of the data, not to the clustering method we applied. (3) Different small size clusters characterize different Web

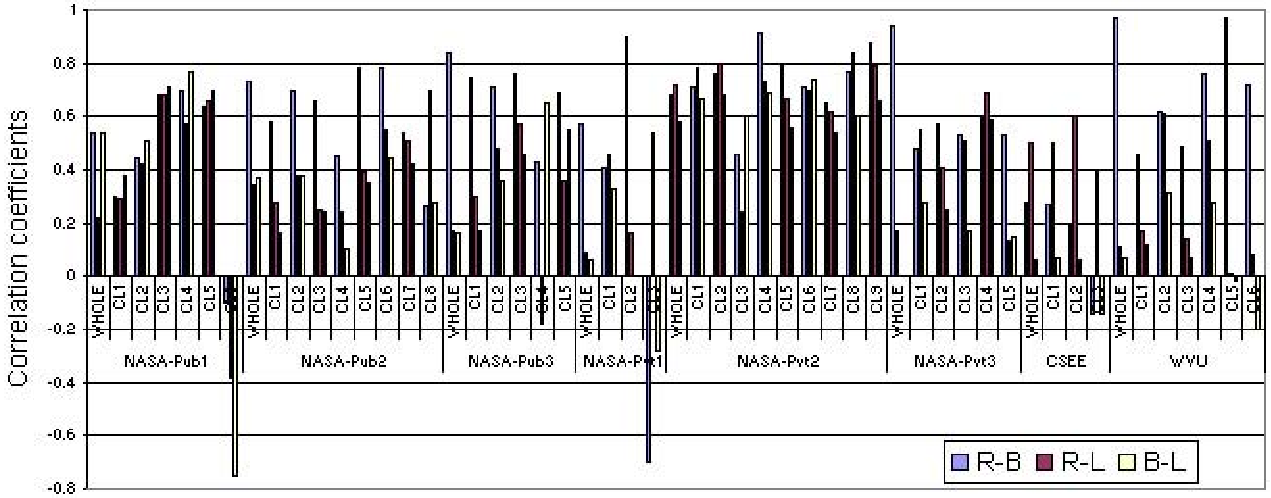

Figure 6. correlation coefficients of intra-session characteristics for all Web servers and clusters

workload phenomena, such as long sessions with moderate number of requests and bytes transferred, or long sessions with large number of requests and bytes transferred. (4) The intra-session characteristics for all Web servers are well modeled by Pareto distribution; they are heavy-tailed for most Web servers. Furthermore, the intra-session characteristics are positively correlated. (5) The intra-session characteristics for the largest clusters are very rarely heavy-tailed. For the largest clusters, the intra-session characteristics are positively correlated. The correlation of the largest cluster does not guide the correlation of the whole dataset. (6) Small size clusters may have huge influence on the heavy-tailedness of the whole dataset. For some small size clusters, the pairwise correlation between intra-session characteristics is negative.

In summary, the results presented in this paper show that clustering provides an efficient way to classify tens or hundreds thousands of sessions into several coherent classes that efficiently describe Web server workloads. Furthermore, clustering allows us to observe some phenomena that are not visible from the analysis of the whole data sets.

Acknowledgements

This work is funded by the National Science Foundation under CAREER grant CNS-0447715 and by the NASA OSMA SARP under grant managed through NASA IV&V Facility in Fairmont. The authors thank David Krovich and David Olsen of West Virginia University and Brian Kesecker, Roger Harris, Lisa Montgomery, and Ken McGill of NASA IV&V Facility for giving permisson to use the data and making the Web logs available.

References

[1] M. Arlitt and C. Williamson, “Internet Web Servers: Workload Characterization and Performance Implications”, *IEEE/ACM Trans. Netw.*, Vol.5, No.5, 1997, pp. 631-645.

[2] M. Arlitt, “Characterizing Web User Sessions,” *SIGMETRICS Perform. Eval. Rev.*, Vol. 28, No. 2, 2000, pp. 50-63.

[3] M. Arlitt, K. Diwakar, and Rolia Jerry, “Characterizing the scalability of a Large Web-based Shopping System,” *ACM Trans. Internet Technology*, Vol. 1, No. 1, 2001, pp. 44-69.

[4] M. E. Crovella and A. Bestavros, “Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes”, *IEEE/ACM Trans. Netw.*, Vol.5, No.6, 1997, pp. 835-846.

[5] B. Everitt, *Cluster Analysis*, Halsted Press, New York, 1980.

[6] K. Goševa-Popstojanova, A. Singh, S. Mazimdar and F. Li, “Empirical Characterization of Session-based Workload and Reliability for Web Servers,” *Empirical Software Engineering Journal*, Vol.11, No.1, 2006, pp. 71-117.

[7] K. Goševa-Popstojanova, F. Li, X. Wang, and A. Sangle, “A Contribution Towards Sovling the Web Workload Puzzle,” *Int’l Conf. Dependable Systems and Networks*, 2006, pp. 505-514.

[8] D. A. Menascé, V. A. F. Almeida, R. Fonseca, and M. A. Mendes, “A Methodology for Workload Characterization of E-commerce Sites,” *1st ACM conf. Electronic Commerce*, 1999, pp. 119-128.

[9] S. I. Resnick, “Heavy Tail Modeling of Teletraffic Data”, *The Annals of Statistics*, Vol.25, No.5,1997, pp. 1805-1849.

[10] W. Wang, and O. Zaiane, “Clustering Web Sessions by Sequence Alignment,” *13th Int’l Workshop on Database and Expert Systems Applications*, 2002, pp. 394-398.

Sixth IEEE International Symposium on Network Computing and Applications (NCA 2007)
0-7695-2922-4/07 $25.00 © 2007 IEEE