Information Processing and Management (2002) *38*(4), 473-490.

Analysis of large data logs: an application of Poisson sampling on excite web queries

H. Cenk Ozmutlu & Seda Ozmutlu

Department of Industrial Engineering

Uludag University

Gorukle Kampusu, Bursa 16059

Turkey

Amanda Spink

School of Information Sciences and Technology

The Pennsylvania State University

511 Rider I Building, 120 S. Burrowes Street

University Park, PA 16801

USA

Abstract

Search engines are the gateway for users to retrieve information from the Web. There is a crucial need for tools that allow effective analysis of search engine queries to provide a greater understanding of Web users' information seeking behavior. The objective of the study is to develop an effective strategy for the selection of samples from large-scale data sets. Millions of queries are submitted to Web search engines daily and new sampling techniques are required to bring these databases to a manageable size, while preserving the statistically representative characteristics of the entire data set. This paper reports results from a study using data logs from the Excite Web search engine. We use Poisson sampling to develop a sampling strategy, and show how sample sets selected by Poisson sampling statistically effectively represent the characteristics of the entire dataset. In addition, this paper discusses the use of Poisson sampling in continuous monitoring of stochastic processes, such as Web site dynamics.

Author Keywords: Poisson sampling; Large-scale in depth data analysis; Web user modeling; Search engine queries; Data mining

1. Introduction

One of the most frequent ways to retrieve information from the Web is via queries submitted to a search engine. Hence, query analysis is one important way to understand Web users' behavior. In this paper, we develop a methodology for sampling large data logs of user query sessions collected by Web search engines. Analyzing Web search engine query logs is problematic due to the large amount of data available. Everyday, millions of user query sessions are submitted to Web search engines. In databases of this size, most software packages have problems handling even the simplest of tasks, such as sorting, etc. The dynamics and characteristics of Web user query sessions are yet to be analyzed completely and many studies on Web user query sessions require context-wise interpretation of data that requires manual analysis (Spink, Jansen, & Ozmultu, 2000). The sheer enormity of Web data sets can render analysis difficult.

A major challenge for Web researchers is to select a sample from a data set that will effectively represent the whole data set. Previous studies on the dynamics of the user query sessions either have not used a specific strategy to select their samples, or the entire dataset is used (Spink, Bateman, & Jansen, 1999; Spink et al., 2000; Tomaiuolo & Packer, 1996). The problems of sampling strategy are faced in other research fields such as analysis of Web page dynamics and updates (Brewington & Cybenko, 2000). There is a need for a robust sampling strategy that can be applied to analyze large-scale Web data logs and other research fields.

This paper first compares the characteristics of systematic sampling and Poisson sampling. We then propose a random sampling strategy based on Poisson sampling, to tackle the task of choosing a sample of manageable size that is statistically representative of an entire data set. Poisson sampling algorithms select sample points from a certain dataset by skipping a random number of observations that is distributed according to a Poisson process. Poisson sampling is appropriate due to many properties. The most important property of Poisson sampling is that the query arrival process should not necessarily conform to the Poisson process that enables unbiased sampling for any stochastic arrival process (Wolff, 1982). An effective sampling strategy will enable the analysis of larger data sets. Moreover, a Poisson based sample selection methodology will increase the probability that a sample represents the entire data set, while preserving the statistical characteristics of the entire data set.

The following section summarizes the characteristics of Poisson sampling and systematic sampling.

2. Characteristics of systematic sampling and Poisson sampling

2.1. Systematic sampling

During the Web query analysis it is common to apply fixed interval sampling (selecting every *n*th query) that is also referred to as *systematic sampling*. During the sample selection process, every data point should have a chance of selection. Therefore, a random sampling strategy is required for assigning a probability for selection. However, the most critical issue is to determine the random stochastic process to use for random selection, where the probability of being selected is the same for all data points. Systematic sampling can provide sufficient data samples if there is no pattern in the data. However, the characteristics of the Web user query sessions differ according to their arrival times. The arrival times of sample points play critical role in the representative power of the sample data set. For example, considering hourly distribution of arrivals within a day, the number of samples taken in different hours should be proportional to the ratio of the mean of number of arrivals in hours of a day. Systematic sampling will provide proportional sampling if *n* (number of arrivals skipped before the next sample) is small enough so that there will be multiple samples at each hour of the day.

2.2. Poisson sampling

The Poisson sampling process is a useful random sampling process as it includes the following properties:

Unbiased sampling: When Poisson sampling is applied, all instances of a stochastic arrival process would have equal chance of selection (Bilinkis & Mikelsons, 1992).

Proportional sampling: The characteristics of the arrival process for the Web inquiry sessions may change due to some factors such as time of the day, day of the week, etc. These factors may stay in effect for different time periods during the sampling process. Changes in the effective parameters create different stages of the stochastic process of the arrival process, such as morning stage, Christmas stage, etc. The advantage of the Poisson sampling process is that the duration length ratio of stages is captured in the ratio of number of observations taken during the stages, allowing proportional sampling of the different stages of the observed stochastic process (Wolff, 1982).

Comparability of heterogeneous Poisson sampling arrivals: If $\{N_1(t),t\geqslant 0\}$ and $\{N_2(t),t\geqslant 0\}$ are both Poisson processes with rates λ_1 and λ_2, respectively ($\lambda_1\neq\lambda_2$), then the averages of the samples obtained by $N_1(t)$ and $N_2(t)$ can be compared or combined depending on the time interval they have been applied (Wolff, 1982). This property allows utilization of different studies that apply Poisson sampling even if they use different parameters.

Flexibility on the stochastic arrival process from which the sample is selected: Poisson sampling can be applied to any kind of stochastic arrival process (Wolff, 1982). This is the most important property for the applicability of Poisson sampling on the Excite Web search query data log, since there is no information on the type of stochastic arrival process of the Web query sessions. In addition, the time stamp on the Excite transaction log is not sensitive enough to do an analysis on the interarrival times. Fortunately, the lack of knowledge on the stochastic arrival process of the Web query sessions does not affect the applicability Poisson sampling.

Poisson sampling can be applied in two different cases: continuous time sampling and discrete time sampling. For continuous time sampling, selection of the next sample point is comparatively easy. The random timing of the next sample is generated according to an exponential distribution with parameter λ (interarrival time of the next sample $x\sim\mathrm{Exp}(\lambda)$). The effects of the parameter λ will be discussed in the following section. The formulation for the random number generator for exponential distribution can be derived from the cumulative density function (cdf) of the exponential distribution, given in Eq. (1):

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y = \begin{cases} 1-\mathrm{e}^{r-\lambda x}, & x \geqslant 0, \\ 0, & x < 0. \end{cases} \tag{1}$$

If x is generated according to an exponential distribution, then the outcome of cdf, $F(x)$ for $x\geqslant 0$, has a Uniform (0,1) distribution. Since random variables $u\sim$ Uniform (0,1) are fairly easy to obtain, it is logical to use a formula where the interarrival time $x\sim\mathrm{Exp}(\lambda)$ can be obtained by a variable $u\sim$ Uniform (0,1). By calculating the analytical inverse of the exponential cdf in Eq. (1), we can develop the desired formula, which is stated in Eq. (2).

After each sample point, a new uniform number *u* has to be generated to calculate the next exponentially distributed interarrival time using Eq. (2):

$$F^{-1}(u) = \begin{cases} -(1/\lambda) * \ln(1-u), & 0 \leqslant u \leqslant 1, \\ 0, & u < 0 \text{ or } u > 1. \end{cases} \tag{2}$$

In the other case of Poisson sampling, discrete time sampling is used where the stochastic process under observations has discrete arrivals. For discrete stochastic arrival processes, sampling is done by randomly generating a number $u\sim$ Uniform (0,1) and then find the corresponding *n*, the number of arrivals to skip before the next sample, using Poisson Process with parameter $\lambda>0,\{N(t),t\geqslant 0\}$. Note that the interarrival times of samples are distributed according to Poisson process, not the interarrival times of the process from where the samples are taken. The probability mass function of the Poisson process is given in Eq. (3):

$$F(x) = \frac{\lambda^k \exp(-\lambda)}{k!}, \quad \lambda > 0, \ k = 0, 1, \ldots \tag{3}$$

However, the analytical inverse of the Eq. (3) is not available.Therefore the following algorithm is used to generate the Poisson variate *n* (Mann, Schafer, & Singpurwalla, 1974):

Step 1. Set $j = 0$ and $y_j=u_0$, where $u_j\sim$ Uniform (0,1), j=0,1,…

Step 2. If $y_j\leqslant\exp(-\lambda)$, return $n=j$ and terminate.

Step 3. $j=j+1$, and $y_j=u_jy_j-1$

Goto Step 2.

As in the continuous sampling case, another random *n* is generated using the algorithm stated above.

Excite Web query sessions arrive according to a discrete stochastic process. Although, there is no available data study on the type of stochastic process that Web query sessions follow, the sampling strategy is not affected due to the fourth property of Poisson sampling. The data used in this study has time stamps for each query entry, however it is not sensitive enough to determine the stochastic arrival process. The smallest time unit of the time stamps was seconds, and on average, there were 31.8 arrivals in each second. One can argue that if the sampling time units are set in seconds, the arrival process can be considered as continuous time. Consequently, continuous time sampling becomes applicable. However, this discussion is not addressed in this study. To be on the safe side, we will apply discrete time Poisson sampling for the analysis of the data set.

In this paper we selected various samples that were created using Poisson sampling. These samples are compared to the whole data set to demonstrate the effectiveness of Poisson sampling. We also compare the Poisson sampling strategy to fixed interval (systematic) sampling.

3. Research goals

Our research goals in this study were:

1. To develop an effective sampling strategy for large Web data set analysis.

2. Compare Poisson and systematic sampling techniques for use with large Web data sets.

4. Research design

4.1. Excite data set

The data analyzed in this study consisted of a data set of 10,015 Excite Web queries (from 1000 users) selected from 1.7 million submitted on 20 December 1999. First, we assume that the 1000 session data is the entire data set for the purpose of demonstrating the effectiveness of Poisson sampling. By choosing a 1000 session data set as the entire data set, we subject Poisson sampling to stronger statistical tests. In other words, if we used the 1000 session data set as a sample and have applied the statistical tests accordingly (comparing one sample to a larger sample), the samples chosen from 1000 session data set by Poisson sampling would seem to be representative of the 1000 session data set with higher probability. The selected samples from 1000 session data log are used to estimate the characteristics of the 1000 session data log. In the Excite data log the entries are listed in the order they arrive that allows the identification of new sessions through a user ID and each query is given a time stamp in hours, minutes and seconds.

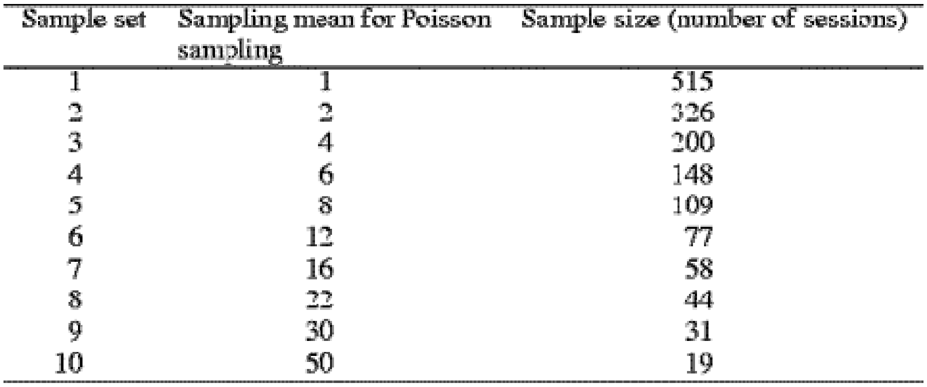
4.2. Data analysis

The data analysis was performed to demonstrate the properties of Poisson sampling. The data analysis consists of two parts: (1) demonstrates session statistics and compares the mean number of queries per session and the mean session durations of various samples chosen with Poisson sampling with the corresponding values from the main population, and (2) investigates the time-based arrival of sessions and queries.

4.2.1. Analysis for session statistics

10 different sample sets were generated from the entire data set of 10,015 queries using different means for Poisson sampling. The sampling mean is the average number of arrivals between two sample points. As the mean used for Poisson sampling increases, the number of observations skipped between the data units that contribute to the sample set increases, implying a decrease in sample size. Table 1 shows the means used for Poisson sampling and the sample size for each individual sample.

Table 1. Sampling mean for Poisson sampling and the sample size for number of queries per session and session duration analysis



Sample set	Sampling mean for Poisson sampling	Sample size (number of sessions)
1	1	515
2	2	326
3	4	200
4	6	148
5	8	109
6	12	77
7	16	58
8	22	44
9	30	31
10	50	19

4.2.1.1. Analysis for number of queries per session

It was important to determine if the number of queries per user obtained from the data sets was representative of the entire data set of 10,015 queries. Therefore, each individual sample set was compared with the entire data set. We used hypothesis testing at the 5% and 1%

significance levels to test the statistical significance of the difference in mean number of queries between each sample set and the entire data set. The hypotheses are as follows:

$$H_0 : \mu_i = \mu.$$
$$H_1 : \mu_i \neq \mu.$$

where μ is the mean number of queries obtained from the entire data set and μ_i is the mean number of queries obtained from sample set $i,i=1,\ldots,10$. Each sample was tested with respect to the entire data set, considering the mean from the entire data set as a constant. Hence, the tests are based on comparing a sample mean to a specific value. The test statistics for the first nine samples are z statistics since the sample size is over 30 and the test applied for the last sample is the t test since the sample size is less than 30.

4.2.1.2. Analysis for duration of sessions

Session durations have been frequently emphasized in studies on Web user dynamics (Spink et al., 2000). We analyzed the strength of the 10 samples in Table 1 in terms of session durations compared to the entire dataset. The analysis of session duration was performed as the analysis for number of queries per session. Hypothesis testing at the 5% and 1% significance levels was performed to test the 10 samples obtained using Poisson sampling.

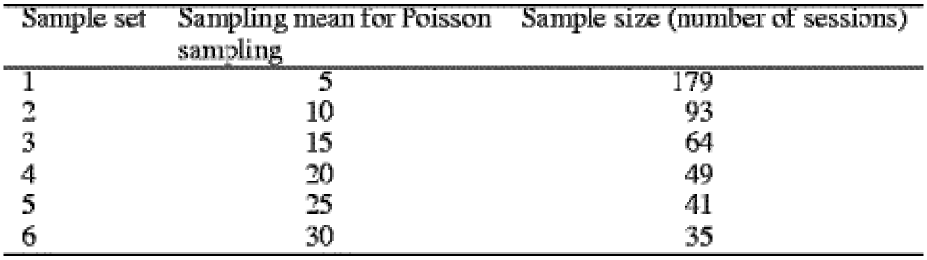
4.2.2. Analysis for time-based session and query arrivals

The second part of data analysis consists of data analysis for time-based session and query arrivals. Time-based analysis of sessions investigates the changes in session and query arrivals with respect to hours of the day.

A dataset of 3188 queries from 1064 users is another sample set selected from the 1.7 million Excite Web queries. The dataset of 10,015 queries is not used as the entire dataset in this portion of the study, since it only consists of a set of queries of a certain time portion of the day and do not provide enough variability in terms of the arrival time of query sessions. The 1064 sessions with 3188 queries were chosen using Poisson sampling from the 1.7 million queries. The 1.7 million queries were recorded throughout working hours of a day (9:00 a.m.–5:00 p.m.). Hence, a sample chosen from the 1.7 million queries through Poisson sampling will reflect the distribution of number of sessions and queries with respect to time. However, for analysis purposes and to keep the research methodology consistent with the first portion of the data analysis, the sample of 3188 queries will be used as the entire dataset for this portion of the data analysis. The 3188 queries were chosen with a sampling mean of 500 users for Poisson sampling.

Six different sample sets were generated from the entire data set of 3188 queries. The sample generation was done in the same fashion as applied in the first portion of the data analysis. Increasing mean for Poisson sampling implies a decrease in sample size. Table 2 shows the sampling means applied for Poisson sampling and the sample size for each individual sample.

Table 2. Sampling means for Poisson sampling and the sample size for the time-based session and query arrival analysis



Sample set	Sampling mean for Poisson sampling	Sample size (number of sessions)
1	5	179
2	10	93
3	15	64
4	20	49
5	25	41
6	30	35

4.2.2.1. Time-based analysis for session arrivals

The time-based analysis of session arrivals investigates the number of session arrivals to the search engine based on the hour of the day. It should be noted that a session does not necessarily end in the same hour that it begins.

The statistical testing applied in this portion of data analysis is slightly different than applied in the first section for number of queries per session and session duration. There are eight values for number of session arrivals for each sample set; each value corresponding to an hour of the workday. For example, for the entire dataset, we will calculate the session arrivals for 8:00–9:00 a.m. until 4:00–5:00 p.m., hence the eight values for session arrivals. Each sample will have a set of eight session arrival values. Since a set of values for different hours for each sample will be compared to a matching set of data for the entire dataset, paired *t*-tests will be used for hypothesis testing. In addition, since the total number of sessions is different for each sample and the entire dataset, the comparison should be based on the proportions of session arrivals at every hour to the total session arrivals, which we will call p_{ij} for sample set i=1,…,6 and j=1,…,8 (where j=1 for 9:00–10:00, etc.). The same proportions for the entire dataset will be referred as p_{ej} for j=1,…,8. Hence, the hypothesis testing of the difference in hourly number of session arrivals between the entire dataset and the six samples is performed using the proportion values.

The hypotheses are as follows:

$$H_0 : \mu_i = 0.$$
$$H_1 : \mu_i \neq 0.$$

where μ_i is the mean of differences between p_{ij} for sample set i, i=1,…,10;j=1,…,8 and p_{ej},j=1,…,8 for the entire dataset. The paired t-testing will be performed for 5% and 1% significance levels.

4.2.2.2. Time-based analysis for query arrivals

The number of query arrivals to the search engine based on the hour of the day was investigated. The analysis was conducted in the same way as in the time-based analysis of session arrivals, using paired t-testing The same kind of hypothesis testing used for time-based session arrivals was also used for testing time-based query arrivals. Hypothesis testing was performed at the 5% and 1% significance levels.

5. Results

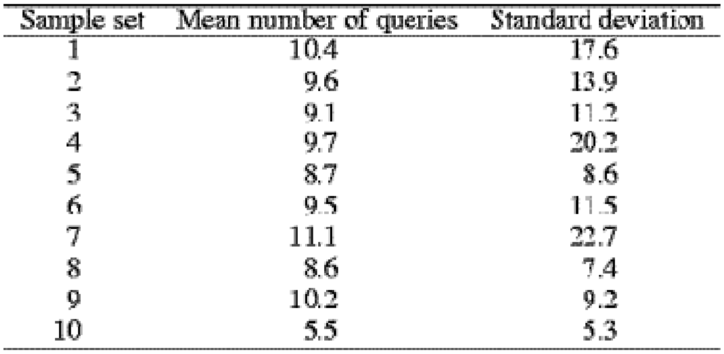
This paper extends preliminary findings reported in Ozmutlu, Spink, and Hurson (2001). The results of the study are reported in the same order as the data analysis methods presented, beginning with the results for session statistics (number of queries per session and session duration) and continuing with time-based session and query analyses.

5.1. Results for session statistics

5.1.1. Number of queries per session

The mean number of queries obtained from the entire data set is 10.015 with a standard deviation of 15.76. The mean number of queries obtained from each sample and the standard deviation of the sample of queries are listed in Table 3.

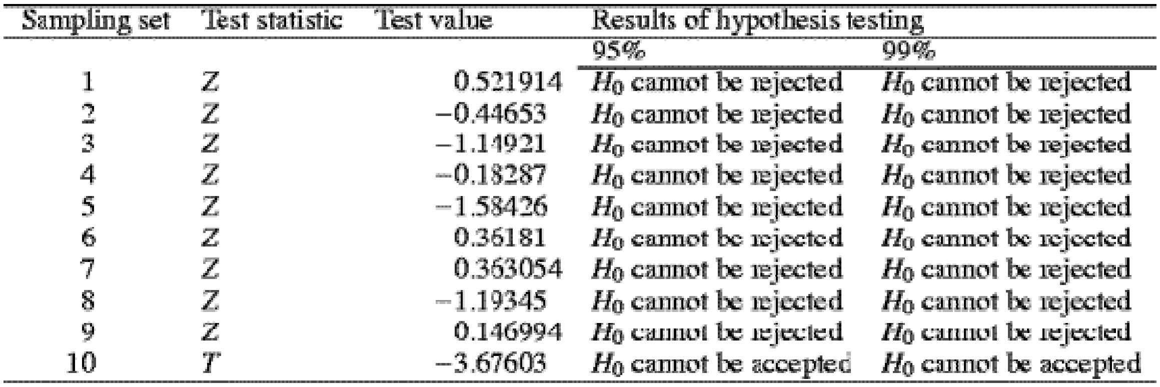
Table 3. Mean and standard deviation of number of queries for 10 different samples



Sample set	Mean number of queries	Standard deviation
1	10.4	17.6
2	9.6	13.9
3	9.1	11.2
4	9.7	20.2
5	8.7	8.6
6	9.5	11.5
7	11.1	22.7
8	8.6	7.4
9	10.2	9.2
10	5.5	5.3

The statistical tests are performed as described in the data analysis for number of queries per session. Table 4 displays the test statistics for the statistical significance of the difference between the mean number of queries from each sample and the mean number of queries from the entire data set, the test values for the relevant tests and the results of the hypothesis testing on the 5% and 1% significance levels.

Table 4. The hypothesis tests used for testing the statistical significance of difference in mean number of queries between entire data set and sample sets and the results of the hypothesis testing



Sampling set	Test statistic	Test value	Results of hypothesis testing	
			95%	99%
1	Z	0.521914	H_0 cannot be rejected	H_0 cannot be rejected
2	Z	−0.44653	H_0 cannot be rejected	H_0 cannot be rejected
3	Z	−1.14921	H_0 cannot be rejected	H_0 cannot be rejected
4	Z	−0.18287	H_0 cannot be rejected	H_0 cannot be rejected
5	Z	−1.58426	H_0 cannot be rejected	H_0 cannot be rejected
6	Z	0.36181	H_0 cannot be rejected	H_0 cannot be rejected
7	Z	0.363054	H_0 cannot be rejected	H_0 cannot be rejected
8	Z	−1.19345	H_0 cannot be rejected	H_0 cannot be rejected
9	Z	0.146994	H_0 cannot be rejected	H_0 cannot be rejected
10	T	−3.67603	H_0 cannot be accepted	H_0 cannot be accepted

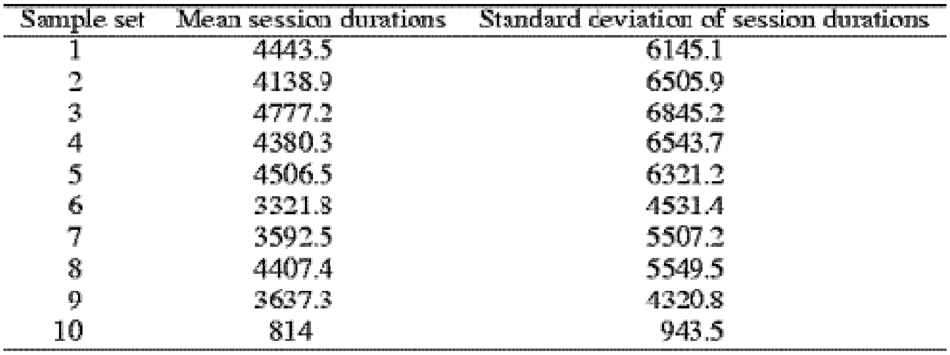
The difference between the mean of queries per user of the first nine samples and the entire data set are statistically insignificant, whereas the difference between the last sample and the entire data set is statistically significant. This result is valid for both at the 5% significance level and 1% significance level (noting that the *z* value for 5% and 1% are 1.96 and 2.575, respectively and the *t* value for 5% and 1% significance levels are 2.101 and 2.878 with the degrees of freedom being 18, respectively). In other words, a sample size less than 30 is not very reliable in determining the mean number of queries, while a sample size equal to or over

30 is. This finding is not unusual as the importance of a sample size larger than 30 is a widely known theoretical fact in statistical applications (Montgomery, 1991).

5.1.2. Session duration

The mean session duration for the entire dataset is 4254.901 and the standard deviation for session durations is 6189.653. Table 5 provides the mean session durations and the standard deviation of session durations for the 10 samples.

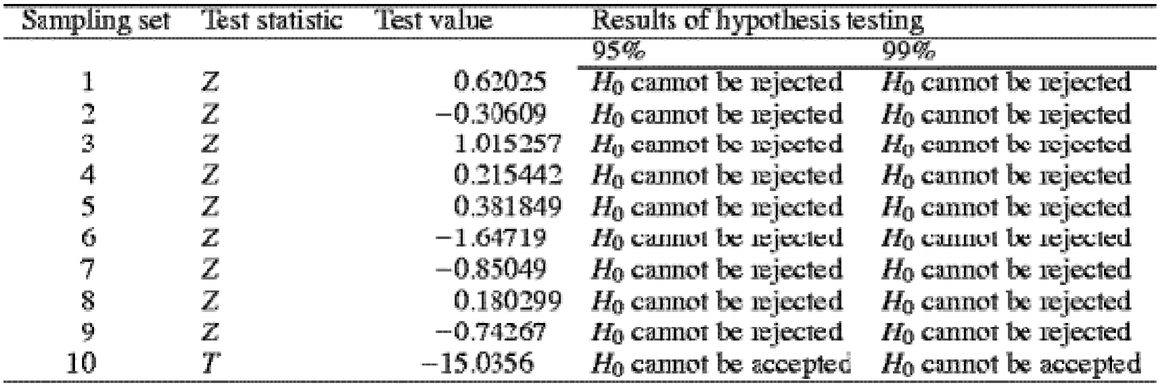
Table 5. Mean and standard deviation of session durations for the sample sets



Sample set	Mean session durations	Standard deviation of session durations
1	4443.5	6145.1
2	4138.9	6505.9
3	4777.2	6845.2
4	4380.3	6543.7
5	4506.5	6321.2
6	3321.8	4531.4
7	3592.5	5507.2
8	4407.4	5549.5
9	3637.3	4320.8
10	814	943.5

The hypothesis testing used for the mean number of queries per session is also applied to the session durations at the 5% and 1% significance levels. The results are given in Table 6.

Table 6. The hypothesis tests used for testing the statistical significance of difference in duration of sessions between the entire data set and sample sets and the results of the hypothesis tests



Sampling set	Test statistic	Test value	Results of hypothesis testing	
			95%	99%
1	Z	0.62025	H_0 cannot be rejected	H_0 cannot be rejected
2	Z	−0.30609	H_0 cannot be rejected	H_0 cannot be rejected
3	Z	1.015257	H_0 cannot be rejected	H_0 cannot be rejected
4	Z	0.215442	H_0 cannot be rejected	H_0 cannot be rejected
5	Z	0.381849	H_0 cannot be rejected	H_0 cannot be rejected
6	Z	−1.64719	H_0 cannot be rejected	H_0 cannot be rejected
7	Z	−0.85049	H_0 cannot be rejected	H_0 cannot be rejected
8	Z	0.180299	H_0 cannot be rejected	H_0 cannot be rejected
9	Z	−0.74267	H_0 cannot be rejected	H_0 cannot be rejected
10	T	−15.0356	H_0 cannot be accepted	H_0 cannot be accepted

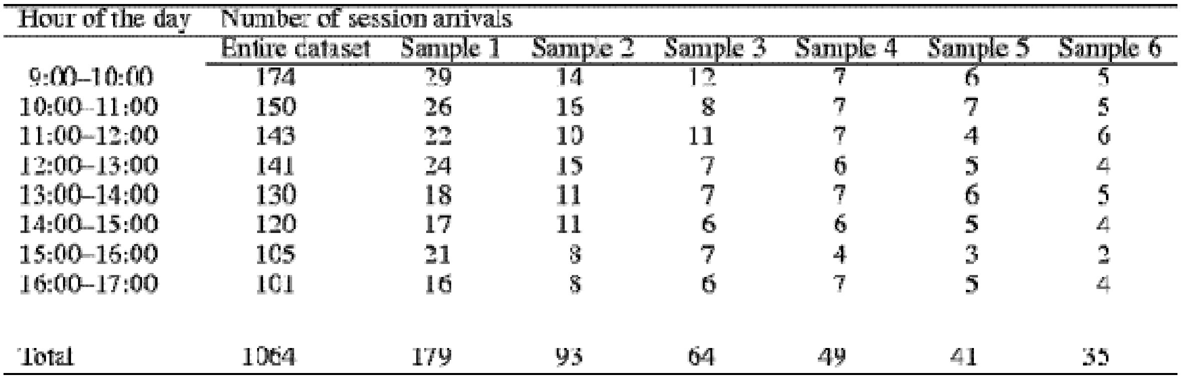
The findings are consistent with the results from the mean number of queries per session analysis. The difference between the mean session durations of the sample sets and the entire data set are statistically insignificant except sample 10. This result is valid for both at the 5% significance level and 1% significance level using the same statistical testing criteria as used in the mean number of queries study.

5.2. Time-based session and query arrivals

5.2.1. Time-based analysis of session arrivals

The session arrivals for the entire dataset of 3188 queries and the six samples are given in Table 7.

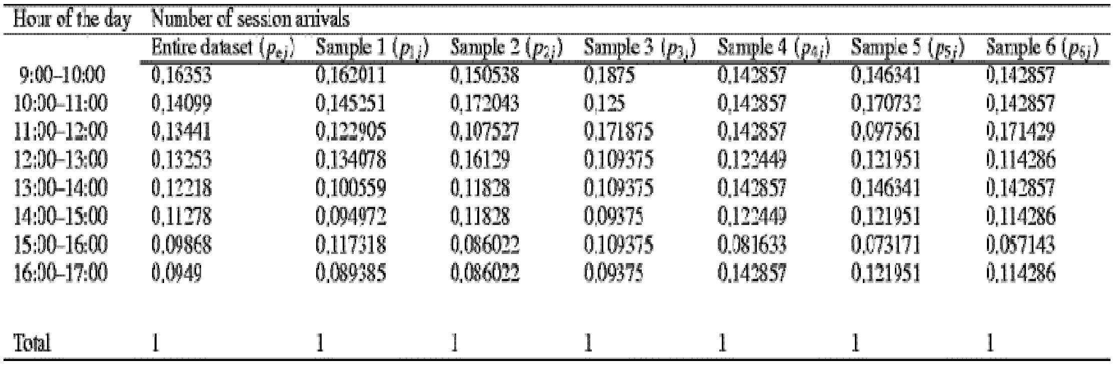
Table 7. Number of session arrivals with respect to hours of the day for the entire dataset and six samples



Hour of the day	Number of session arrivals						
	Entire dataset	Sample 1	Sampe 2	Sample 3	Sample 4	Sample 5	Sample 6
9:00–10:00	174	29	14	12	7	6	5
10:00–11:00	150	26	15	8	7	7	5
11:00–12:00	143	22	10	11	7	4	6
12:00–13:00	141	24	15	7	6	5	4
13:00–14:00	130	18	11	7	7	6	5
14:00–15:00	120	17	11	6	6	5	4
15:00–16:00	105	21	8	7	4	3	2
16:00–17:00	101	16	8	6	7	5	4
Total	1064	179	93	64	49	41	35

The proportion of session arrivals per hour to the total number of session arrivals (p_{ij} and p_{ej}) is given in Table 8. As described in the data analysis for time-based session arrivals, the proportion of session arrivals per hour to the total number of session arrivals was used to develop a common comparison ground between the sample sets and the entire dataset.

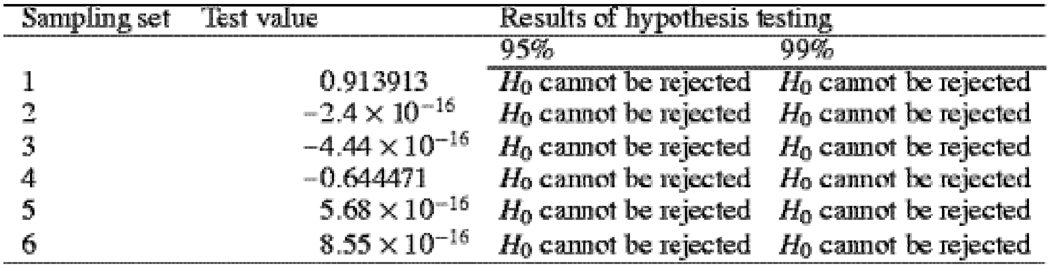
Table 8. Proportion of session arrivals each hour to the total number of session arrivals



Hour of the day	Number of session arrivals						
	Entire dataset (p_{ej})	Sample 1 (p_{1j})	Sample 2 (p_{2j})	Sample 3 (p_{3j})	Sample 4 (p_{4j})	Sample 5 (p_{5j})	Sample 6 (p_{6j})
9:00–10:00	0,16353	0,162011	0,150538	0,1875	0,142857	0,146341	0,142857
10:00–11:00	0,14099	0,145251	0,172043	0,125	0,142857	0,170732	0,142857
11:00–12:00	0,13441	0,122905	0,107527	0,171875	0,142857	0,097561	0,171429
12:00–13:00	0,13253	0,134078	0,16129	0,109375	0,122449	0,121951	0,114286
13:00–14:00	0,12218	0,100559	0,11828	0,109375	0,142857	0,146341	0,142857
14:00–15:00	0,11278	0,094972	0,11828	0,09375	0,122449	0,121951	0,114286
15:00–16:00	0,09868	0,117318	0,086022	0,109375	0,081633	0,073171	0,057143
16:00–17:00	0,0949	0,089385	0,086022	0,09375	0,142857	0,121951	0,114286
Total	1	1	1	1	1	1	1

Table 9 displays the values for the hypothesis tests, and the results of the hypothesis tests at the 5% and 1% significance levels.

Table 9. The values and the results for the paired *t*-tests of time-based session arrivals



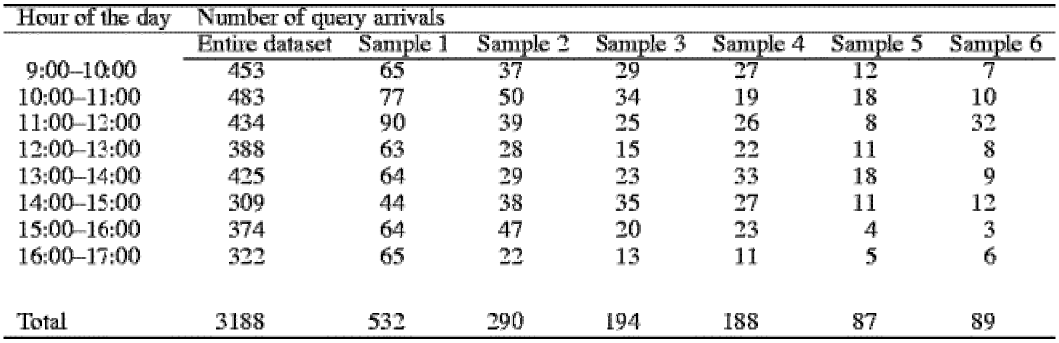
Sampling set	Test value	Results of hypothesis testing	
		95%	99%
1	0.913913	H_0 cannot be rejected	H_0 cannot be rejected
2	-2.4×10^{-16}	H_0 cannot be rejected	H_0 cannot be rejected
3	-4.44×10^{-16}	H_0 cannot be rejected	H_0 cannot be rejected
4	−0.644471	H_0 cannot be rejected	H_0 cannot be rejected
5	5.68×10^{-16}	H_0 cannot be rejected	H_0 cannot be rejected
6	8.55×10^{-16}	H_0 cannot be rejected	H_0 cannot be rejected

The critical *t* values for the 5% and 1% significance levels are 2.365 and 3.499, respectively. The mean difference between the entire dataset and each sample in terms of the proportion of number of sessions per hour to the total number of sessions is statistically insignificant at both the 5% and 1% significance levels. Although 1% significance level is more than adequate for statistical purposes, the test values are so low that the samples would even provide satisfactory results at stricter significance levels. The results are also consistent with the first portion of the data analysis. The size of each sample was greater than 30, providing reliable results.

5.2.2. Time-based analysis of query arrivals

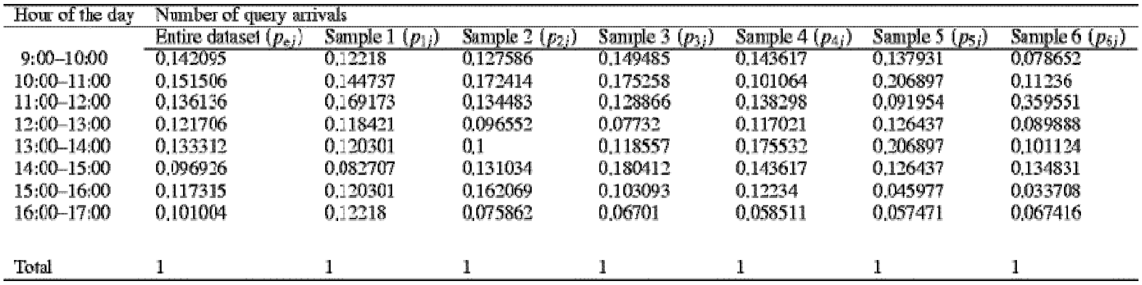
The hourly query arrivals for the entire dataset of 3188 queries and the six samples are given in Table 10 followed by the proportion of number of query arrivals per hour to the total number of queries in Table 11.

Table 10. Number of query arrivals with respect to hours of the day for the entire dataset and six samples



Hour of the day	Number of query arrivals						
	Entire dataset	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
9:00–10:00	453	65	37	29	27	12	7
10:00–11:00	483	77	50	34	19	18	10
11:00–12:00	434	90	39	25	26	8	32
12:00–13:00	388	63	28	15	22	11	8
13:00–14:00	425	64	29	23	33	18	9
14:00–15:00	309	44	38	35	27	11	12
15:00–16:00	374	64	47	20	23	4	3
16:00–17:00	322	65	22	13	11	5	6
Total	3188	532	290	194	188	87	89

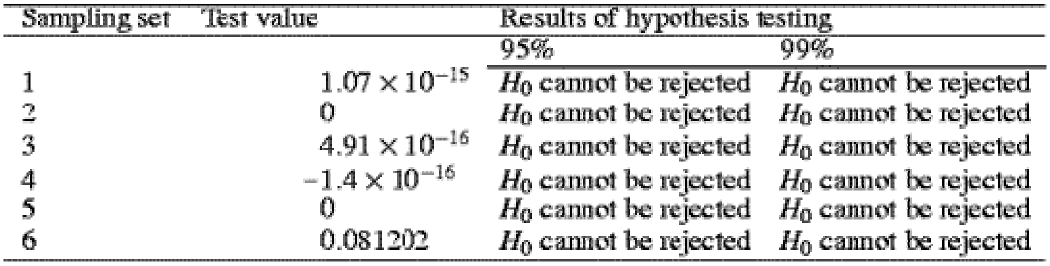
Table 11. Proportion of query arrivals each hour to the total number of query arrivals



Hour of the day	Number of query arrivals						
	Entire dataset (p_{ej})	Sample 1 (p_{1j})	Sample 2 (p_{2j})	Sample 3 (p_{3j})	Sample 4 (p_{4j})	Sample 5 (p_{5j})	Sample 6 (p_{6j})
9:00–10:00	0.142095	0.12218	0.127586	0.149485	0.143617	0.137931	0.078652
10:00–11:00	0.151506	0.144737	0.172414	0.175258	0.101064	0.206897	0.11236
11:00–12:00	0.136136	0.169173	0.134483	0.128866	0.138298	0.091954	0.359551
12:00–13:00	0.121706	0.118421	0.096552	0.07732	0.117021	0.126437	0.089888
13:00–14:00	0.133312	0.120301	0.1	0.118557	0.175532	0.206897	0.101124
14:00–15:00	0.096926	0.082707	0.131034	0.180412	0.143617	0.126437	0.134831
15:00–16:00	0.117315	0.120301	0.162069	0.103093	0.12234	0.045977	0.033708
16:00–17:00	0.101004	0.12218	0.075862	0.06701	0.058511	0.057471	0.067416
Total	1	1	1	1	1	1	1

Table 12 displays the test values and the results of the hypothesis testing on the 5% and 1% significance levels.

Table 12. The values and the results for the paired *t*-tests of time-based query arrivals



Sampling set	Test value	Results of hypothesis testing	
		95%	99%
1	1.07×10^{-15}	H_0 cannot be rejected	H_0 cannot be rejected
2	0	H_0 cannot be rejected	H_0 cannot be rejected
3	4.91×10^{-16}	H_0 cannot be rejected	H_0 cannot be rejected
4	-1.4×10^{-16}	H_0 cannot be rejected	H_0 cannot be rejected
5	0	H_0 cannot be rejected	H_0 cannot be rejected
6	0.081202	H_0 cannot be rejected	H_0 cannot be rejected

Using the same critical *t* values for hypothesis testing (2.365 and 3.499 for 5% and 1% significance levels, respectively), it can be observed that the mean difference between the entire dataset and each sample in terms of the proportion of number of queries per hour to the total number of queries is statistically insignificant at both the 5% significance level and 1% significance level. Again the level of the *t* values are worth notice; depicting that the samples strongly represent the entire dataset.

6. Comparison of systematic sampling and Poisson sampling

The benefits of Poisson sampling increases as the size of the dataset increases and the size of the sample set remain constant (in other words, the number of samples per day decreases). The ultimate goal in applying Poisson sampling is to be able to analyze very large data sets with a time span in terms of weeks, months, seasons, years, etc. To be able to analyze a data log of years of Web user query sessions, only few samples per day can be taken in order to keep the sample set in manageable size. In other words, there won't be enough samples for each hour of every day.

The data used in the following (data analysis) section of this study (to demonstrate the effectiveness of Poisson sampling) contains Web user search query logs between 8 a.m. and 5 p.m. of a single day, and it is not suitable to demonstrate the difference between Poisson and systematic sampling with low sampling rate over a long period of time. Therefore, the effectiveness of Poisson sampling is demonstrated on a randomly generated dataset with 200 days. The first day is taken directly from the actual data and the number of session arrivals per hour of the actual data is used as the mean number of arrivals of that hour for the remaining 199 days. We allowed 5% uniform variation on both directions (plus or minus) in the number of session arrivals.

After generating the dataset for 200 days, we applied Poisson and systematic sampling so that there will be one sample per day on average. We randomly created a dataset of 338,869,542 query sessions. The interval for systematic sampling (*n*) and the mean for Poisson sampling were equal to 1,693,888 that were the mean number of arrivals per day for the dataset for 200 days. We then observed the hourly distribution of the number of sample points within the workdays. Ideally the distribution of the sample points among hours of the day should reflect the ratio between the mean number of session arrivals of hours, in other words a successful sampling should reflect the general daily trends and proportions.

6.1. Distribution of samples

We analyzed the samples taken from the first 50 days to observe the effects of the warm-up period. Fig. 1 demonstrates the distribution of sample points (chosen with Poisson and systematic sampling) with respect to hours of the workday.

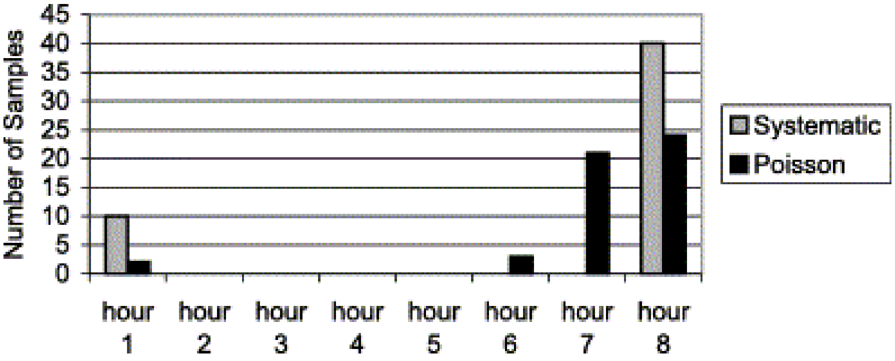


Fig. 1. Distribution of sample points from the first 50 days with respect to hours of the workday.

Although the distributions of samples (from the first 50 days) of both sampling strategies over hours of the day are not proportional to the mean arrival rate for the corresponding hours (due to effects of the warp-up period), Poisson sampling manages a better distribution of the samples over the hours of the day. In systematic sampling, the sample points have gathered in two hourly intervals out of the eight hourly intervals of the workday, the concentration of sample points (40 sample points) being in hourly interval 8 and 10 sample points in hourly interval 1. In Poisson sampling, the sample points have gathered in four different hourly intervals, with the concentration again being in hourly intervals 7 and 8 (2 days out of 50 were in interval 1, 3 out of 50 days were in interval 6, 21 out of 50 days were in interval 7 and 24 out of 50 days in interval 8).

The distributions of the samples (obtained by applying Poisson and systematic sampling) over the hours of the day change significantly as the sampling process continued until samples are taken from 200 days. From the dataset of 200 days, application of systematic sampling resulted in 200 samples whereas Poisson sampling selected 206 samples. The distributions of the samples (from Poisson and systematic sampling) over the hours of the day are demonstrated in Fig. 2.

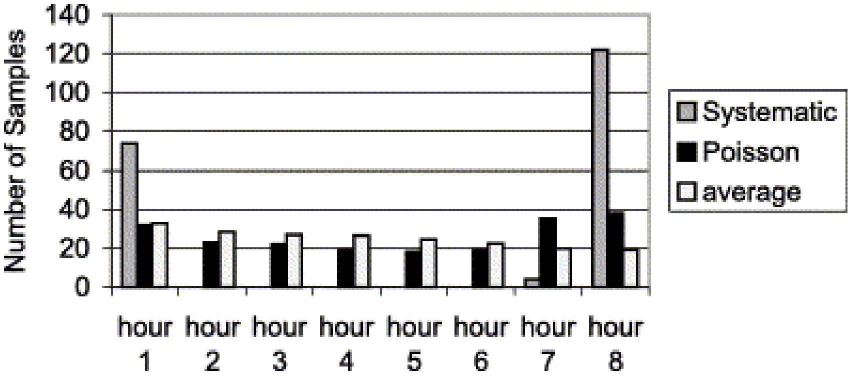


Fig. 2. Distribution of samples taken from 200 days with respect to hours of the workday.

In Fig. 2, the first two columns in each hourly interval represent systematic and Poisson sampling, respectively. The third column in Fig. 2 represents the desired number of samples in each hour given that the total number of samples is 200. As it can clearly be observed from Fig. 2, the performance of Poisson sampling in capturing the desired proportion of the samples throughout hours is significantly superior to that of systematic sampling, since systematic sampling mostly sampled from two out of the eight hourly intervals. Although the effects of the samples from the first 50 days affects the number of samples in hourly intervals 7 and 8, Poisson sampling captured the desired decrease in the number of samples from hourly interval 1 to hourly interval 6. As the number of samples increased (by continuing sampling after 200 days), Poisson sampling will converge to the desired distribution of number of samples over the hours of the workday.

On the other hand, systematic sampling failed to converge to the desired distribution of number of samples over the hours of the workday. An analysis on the sample set obtained by systematic sampling is misleading since the dataset contains the characteristics of two (out of eight) hourly intervals. Since the characteristics of the Web user search sessions do change according to the hour of the day, systematic sampling is not a suitable sampling strategy for examining large datasets consisting of days, weeks, years, etc. Any analysis performed on the sample data (performed in the manner explained above using systematic sampling) would not be statistically representative of the entire dataset.

As we mentioned earlier, systematic sampling may provide satisfactory samples depending on the stochastic arrival process and the sampling rate. However, systematic sampling is not as robust sampling process as Poisson sampling that has proven robustness for applicability on any kind of stochastic arrival process and any kind of sampling rate (Wolff, 1982). Therefore, systematical sampling is not suitable for analysis of large data logs of Web user query sessions that include sessions for days, weeks, years, etc.

7. Discussion

7.1. The problem

Successful modeling of Web search engine interaction can be partly achieved by more effective analysis of queries submitted to Web search engines. Millions of queries submitted to the search engines create enormous data logs, the sizes of which are increasing exponentially as the number of Web users increases over time. With current computational tools, even the simplest of analytical tasks are difficult to perform. To avoid such difficulties, developing a sampling methodology for analyzing large data sets was considered in this study. The problem involves choosing an effective sample from the entire data set. Such a sample will reduce the size of the data to be handled by researchers and carry the statistical characteristics of the entire data set. This paper proposed a random sampling strategy based on Poisson sampling that can tackle the task of choosing a sample of manageable size representative of the whole data set. The sampling strategy considered is Poisson sampling, since it does not require the query arrivals to conform to a particular stochastic arrival process and provides unbiased sampling.

7.2. Data analysis and comparison of Poisson sampling and systematic sampling

To demonstrate the properties of Poisson sampling we used a data log of Web search queries from Excite. The data analysis was performed in two sections. The first section of the data analysis concentrated on the number of queries per session and session durations. We have chosen 10 different samples from 1000 user sessions, using different Poisson sampling means, and have tested the statistical significance of the difference between the mean number of queries per session and session durations. The mean number of queries obtained from the sample sets greater than 30 observations statistically represent the mean number of queries of the entire dataset. The same result is also valid for session durations.

The second section of data analysis emphasized the time-based arrival of queries and sessions. The number of sessions and queries arriving during each hour of a workday are investigated. Six different samples were generated from 1064 user sessions and the statistical significance of the difference in the hourly number of session and query arrivals was tested. The hourly number of session and query arrivals obtained from the sample sets greater than 30 observations statistically represent the corresponding value from the entire dataset.

We also compared Poisson sampling to systematic (fixed interval) sampling. We generated a random dataset for 200 days (using the number of arrivals per hour of the first day (existing data) as mean values for number of arrivals, and allowing 10% (uniform) variation). We applied Poisson and systematic sampling with a sampling rate of one sample per day. The distributions of the sample points (in terms of hour of the day) from Poisson sampling and systematic sampling were similar for the samples of the first 50 days (which are biased due to the initial warm up period of sampling techniques). For Poisson sampling, the bias of the samples of the first 50 days diminished as the sampling continued for the remaining 150 days, and the ratio of the number of samples per hour converged to the ratio of the mean number of session arrivals per hour. However, we did not observe similar convergence in systematic sampling. The samples taken using systematic sampling neither captured the ratio between the mean number of arrivals per hour, nor achieved selection of at least one sample from each hour (there were no samples for 75% of the hours). Given that the properties of Web user query sessions differentiate according to the hour of the day, using systematic sampling will provide misleading information of the characteristics of the Web user query sessions.

8. Application areas

A successful sampling strategy can reduce the large data sets to manageable size, making detailed data analysis possible. Many companies, and particularly e-commerce businesses, keep large databases as records of their processes that they would like to analyze in detail to learn customers' behaviors and patterns of consumption for their and others' products. The proposed sampling strategy can be valuable in establishing the basis for efficient analysis for many companies conducting large-scale Web data analysis.

In this study we used Excite Web query data. Excite, a major Web media public company that offers free Web searching and a variety of other services to millions of customers every day and keeps logs of query data. These data logs are of enormous size and difficult to analyze statistically. The amount of data grows year to year. For example, the number of queries submitted to Excite was 1 million per day in 1997 (Spink, Wolfram, Jansen, & Saracevic, 2001) and 1.7 million queries were submitted in 8 h in 1999 (Wolfram, Spink, Jansen, & Saracevic, accepted for publication). It is logical to assume that the number of queries will increase exponentially in the upcoming years as the Web is growing at an exponential rate (Brewington & Cybenko, 2000).

There are numerous difficulties in handling such enormous data logs. The simplest analysis tasks cannot be tackled with available software tools even for daily data logs. The cross-examination of weekly or monthly data is an even bigger challenge. Reducing the size of the data set could create significant benefits in convenience of analysis, and allow the analysis of even bigger data sets such as a week's or even year's worth of data to allow the observation of users' search behavior characteristics within different hours of the day, day of the week or months of the year. In addition, analyzing data sets of large time periods may yield trends that may be unrevealed by analysis of data of small time periods. However, the sample data set should keep the characteristics of the whole data set for correct analysis.

Another possible application area for Poisson sampling is monitoring changes in Web pages. Web monitoring is a difficult task due to the Web's enormous size: 800 million pages and growing (Lawrence & Giles, 1999). Brewington and Cybenko (2000) investigated over two million Web pages (over seven months at 100,000 pages per day) to gather statistical data. Applying an effective sampling strategy, such as Poisson sampling, will reduce the volume of Web pages to be monitored and decrease the amount of effort spent on monitoring or more Web pages could be monitored in less time. Another advantage of Poisson sampling is in determining the sampling rate to investigate changes in and duration of Web site inertness.

9. Conclusion

This study demonstrates that Poisson sampling is a sampling strategy that allows the opportunity to analyze large data sets using significantly reduced sample sizes without losing the statistical characteristics of the entire data set. We also demonstrated that Poisson sampling is superior to the most common sampling strategy, systematic sampling. Using Poisson sampling, in depth analysis of data can be performed with less computational effort (compared to analyzing the full dataset), that can provide the opportunity to apply types of analysis that are impossible due to today's software and hardware capacity limitations. Better and more detailed analysis of data can lead to better understanding of Web users' information searching behavior, and result in more effectively designed search engines.

References

Bilinkis and Mikelsons, 1992. I. Bilinkis and A. Mikelsons *Randomized signal processing*, Prentice-Hall, New York (1992).

Brewington and Cybenko, 2000. Brewington, B. E., & Cybenko, G. (2000). How dynamic is the Web? In *Proceedings of the Ninth World Wide Web Conference*. May.

Lawrence and Giles, 1999. S. Lawrence and C.L. Giles , Accessibility of information on the web. *Nature* **400** (1999), pp. 107–109.

Mann et al., 1974. N.R. Mann, R.E. Schafer and N.D. Singpurwalla *Methods for statistical analysis of reliability and life data*, Wiley, New York (1974).

Montgomery, 1991. Montgomery, D. C. (1991). *Design and analysis of experiments*, 3rd ed. New York: Wiley.

Ozmutlu et al., 2001. Ozmutlu, H. C., Spink, A., & Hurson, A. (2001). Time-based web mining of search logs: implications for efficient operations. In *Proceedings of IC2001: International Conference on Internet Computing*. June 25–28, Las Vegas, NV.

Spink et al., 1999. A. Spink, J. Bateman and B.J. Jansen , Searching the web: survey of excite users. *Internet Research: Electronic Networking Applications and Policy* **9** 2 (1999), pp. 117–128.

Spink et al., 2000. A. Spink, B.J. Jansen and H.C. Ozmultu , Use of query reformulation and relevance feedback by web users. *Internet Research: Electronic Networking Applications and Policy* **10** 4 (2000), pp. 317–328.

Spink et al., 2001. A. Spink, D. Wolfram, B.J. Jansen and T. Saracevic , Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology* **53** 2 (2001), pp. 226–234.

Tomaiuolo and Packer, 1996. N.G. Tomaiuolo and J.G. Packer , An analysis of Internet search engines: assessment of over 200 search queries. *Computers in Libraries* **16** 6 (1996), pp. 58–62.

Wolff, 1982. R.W. Wolff , Poisson arrivals see time averages. *Operations Research* **30** 2 (1982), pp. 223–231.

Wolfram et al., accepted for publication. Wolfram, D., Spink, A., Jansen, B. J., & Saracevic, T. (accepted for publication). The public searching of the web. *Journal of the American Society for Information Science and Technology*.