A DAY IN THE LIFE OF WEB SEARCHING: AN EXPLORATORY STUDY

Seda Ozmutlu
Department of Industrial Engineering
Uludag University
Gorukle Kampusu, Bursa, 16059, Turkey
Tel: (90-224) 442-8176 Fax: (90-224) 442-8021
E-mail: seda@uludag.edu.tr

Amanda Spink*
School of Information
510 IS Building, 135 N. Bellefield Avenue
Pittsburgh PA 15260
Tel: (412) 624-5230 Fax: (412) 624-5231
Email: spink@mail.sis.pitt.edu

Huseyin C. Ozmutlu
Department of Industrial Engineering
Uludag University
Gorukle Kampusu, Bursa, 16059, Turkey
Tel: (90-224) 442-8176 Fax: (90-224) 442-8021
E-mail: hco@uludag.edu.tr

* To whom all correspondence should be addressed.

ABSTRACT

Understanding Web searching behavior is important in developing more successful and cost-efficient Web search engines. We provide results from a comparative time-based Web study of US-based Excite and Norwegian-based Fast Web search logs, exploring variations in user searching related to changes in time of the day. Findings suggest: (1) fluctuations in Web user behavior over the day, (2) user investigations of query results are much longer, and submission of queries and number of users are much higher in the mornings, and (3) some query characteristics, including terms per query and query reformulation, remain steady throughout the day. Implications and further research are discussed.

INTRODUCTION

Search engines are one of the most frequently used tools to retrieve information from the Web. This paper reports findings from a comparative time-based analysis of search queries submitted by US-based Excite and Norwegian-based Fast Web users. A time-based analysis, which investigates patterns of Web user behavior with respect to different hours of the day, can provide valuable data on Web user behavior. This data can be incorporated into techniques geared towards improving Web search engines. A similar study of direct marketing (Ling & Li, 1998) found that marketing campaign profits are optimized by intelligently reducing the targeted population by consumer profiling using time-based techniques. Currently, Web search engines are not designed to differentiate according to the user's profile. Time-based analysis can be applied to Web user logs to clarify the user profiles. This would enhance user-oriented selection of advertisements, promotions, etc. at the search engine web site. In addition, an intelligent allocation of resources for the Web search engine could be crafted for different time frames. Intelligent use of available resources would mean more efficient operations and lower costs for a Web search engine.

A time-based query search analysis can also be incorporated to automatic indexing methods. As far as we know, no indexing mechanisms that consider time-based query arrivals have been developed or publicized. Automatic indexing models assume the uniformity of query characteristics, and therefore, might work less effectively in cases where the query characteristics vary with respect to time. A new indexing method can be developed where the prioritization of the results is altered with respect to time of the day a query is submitted. For example, let's assume that a time-based study reaches the

conclusion that search queries are more business oriented in the morning hours and recreation oriented in the evening hours. In this case, for a given query, an time-based indexing mechanism could provide results more related to business in the morning hours and more related to recreational aspects in the evening hours. So, a search engine which has time-based search preferences of users and user profiles could therefore provide more satisfactory and user-oriented search results.

Another area where time-based query analysis can be used is user-oriented clustering in information retrieval (IR) systems (Bhatia & Deogan, 1998). The clustering models can be modified to more realistic models; capturing the change in the characteristics of Web search engine user queries with respect to time.

This study considers one day's data from both the Excite and Fast search engines, in other words, this study can be seen as an exploratory study of Web user behavior.

RELATED STUDIES

There are a number of studies performing large-scale search query analysis (Silverman, et al., 1999; Spink, et. al., 2002). However, the number of studies analyzing the Web searching within a time-based frame is limited. The current comparative study provides a time-based analysis including data from the Excite and Fast Web search engines.

Software and hardware limitations gain importance in handling databases of millions of queries. Moreover, many studies on Web user query sessions require context-wise interpretation of data that require manual analysis (Spink, et al., 2000). Manual analysis of large datasets might require selecting a representative sample of the entire

data set. The problems with sampling strategies are faced in other research fields such as analysis of Web page dynamics and updates (Brewington & Cybenko, 2000).

The problem is to select a sample data set that will statistically represent the whole data set, and that will also be convenient to handle and analyze. 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, et al., 2002). Ozmutlu, et al., (2002) discuss Poisson sampling as superior to systematic sampling and provide samples of the data set that is both statistically representative of the data set and small enough to be analyzed conveniently. During the sample selection process, every data point should have a chance of selection with the same probability, which requires the use of a random sampling strategy. The Poisson sampling process is the most suitable random sampling process as it includes the following properties:

Unbiased sampling: 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. Changes in the effective parameters create different stages of the stochastic process of the arrival process, such as morning stage, afternoon 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 $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ 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 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 stamps on the search engine transaction logs are 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.

RESEARCH DESIGN

Excite Query Set

The Excite search engine provided a set of 1.7 million queries [http://www.excite.com] for our analysis. These queries are the entire log of queries submitted to the Excite search engine between 9.00 AM until 5:00 PM on December 20, 1999. Using Poisson sampling we generated a representative sample consisting of 3188 queries from 1064 users from the 1.7 million Excite queries. This sample will be used to estimate the time-based characteristics of the entire 1.7 million-query data log. In the Excite data log structure, the entries are given in the order they arrive. It is possible to identify new sessions through a user ID and each query is given time stamps in hours, minutes and seconds based on United States west coast time (Excite is located in

California). Previous studies (Spink, et al., 1999) show that approximately 84% of Excite users were located in the United States.

Fast Data Set

The Fast search engine (http://www.alltheweb.com) provided a query log of 1,257,891 queries for our analysis. These queries are the entire log of queries submitted to the FAST search engine between 12:00 AM on February 6, 2001 and 12:00 AM February 7, 2001. We selected a sample of 10,007 queries from 964 users from the total of 1,257,891 queries. The sample was selected using Poisson sampling to provide a sample dataset that is both statistically representative of the entire data set and small enough to be analyzed conveniently. Poisson sampling provides a basis for sampling from large-scale data logs, such as the Excite search query log, while preserving the characteristics of the main dataset (Ozmutlu, Spink & Ozmutlu, 2002).

In the Fast data log structure, the entries are given in the order they arrive. It is possible to identify new sessions through a user ID and each query is given time stamps in hours, minutes and seconds. After the sample for analysis is selected, it is divided with respect to hours of the day to facilitate time-based analysis of the queries.

The data used in this study was obtained from two Web search engines – FAST, and Excite. Obtaining large-scale data from other major commercial Web search engines such as Google is a difficult process. At the time of this study investigators were not able to secure data from other Web search engines, e.g., Google. There was also no opportunity to sample Web queries over time, such as across days or weeks. The authors also had little control over the quantity or structure of the data provided by the Web search

engines. However, we are grateful to Excite and FAST for providing the data we had to analyze.

Query Analysis

Our analysis emphasizes the following concepts (Spink, et al, 2001):

(1) *Quer*y: a set of one or more search terms; it may include advanced search features, such as logical operators and modifiers.

(2) *Sessio*n: the entire set of queries by the same user over time. A session could be as short as one query or contain many unique and repeat queries.

*Ter*m: any unbroken string of alphanumeric characters entered by a user. Terms include words, abbreviations, numbers, and logical operators (AND, OR, NOT). URLs and e-mail addresses were treated as single terms.

The query analysis performed in this study is divided to three parts.

The *first part of the query analysis* investigated the changes in hourly query and session arrivals, and mean query and session durations in both the Excite and Fast query sets.

The *second part of query analysis* investigated the search terms in the Excite query set. The number of search terms used within queries are provided along with an in depth analysis of the topics of the search terms.

The *third part of the query analysis* provides a Markov analysis of the queries to investigate the specifications of the transition from previous queries to subsequent queries and the long-term ratios of unique queries, modified queries or next page queries in the Excite query set. For the Markov analysis, the identification of the queries as

unique, modified or next page queries has been done manually (further emphasizing the importance of Poisson sampling). Otherwise, it is impossible to differentiate unique (U) and modified (M) queries. To be accepted as a modified query, the difference between the modified query and the previous query of the session can be the addition or subtraction of new terms or advanced search terms (+, -, etc.), or correction of previously used terms. If the change in the subsequent query cannot be qualified as modified query then it is registered as a unique query. Next page queries (P) should be exactly same as the previous query.

RESULTS

This paper extends a preliminary analysis reported in Ozmutlu and Spink (2001). The results reported in this paper are provided in three parts: (1) *basic session and query characteristics*, such as hourly query and session arrivals and mean query and session durations for Fast and Excite data sets, (2) *search terms*, such as the number of search terms used within queries and the topics of the search terms for the Excite and Fast data sets, and (3) Markov analysis of the Excite queries.

User Session and Query Arrivals

Excite Query Set

The mean queries per user session was 2.9. This number is close to the mean queries per user session of 2.4 reported in Spink, et al., (2002). The findings related to the queries per user session, the arrival of sessions and queries to the search engine with respect to hours of the day, are shown in Table 1, Figure 1 and Figure 2.

[Place Table 1 Here]

[Place Figure 1 Here]

[Place Figure 2 Here]

The mean queries per Excite user session decreased from 3.9 around 9:00 AM to 2.22 around 4:00PM (United State west coast time). The decrease in the mean queries per user session is about 43.09% between these hours. The mean queries per morning hours are higher than the mean of 2.9, whereas in the afternoon all the values demonstrate a decline. The query arrivals per hour also decreases sharply from 679 around 9:00 AM to 224 around 5:00 PM – about 67.01% - as the day progresses, also accompanied by a decrease in the number of session arrivals from 174 around 9:00 AM to 101 around 5:00 PM– about 42.3%. The decreasing trend is noticeable in all the session and query arrival criteria. These results demonstrate that as the hours progress, the use of the Excite Web search engine is diminished. Moreover, the decrease in the queries per session indicates that Web search engine users might spend less effort on retrieving their information needs later in the day.

Since the search queries are recorded in U.S. West coast time, the total three hour difference in time zones in the U.S. might affect the analysis slightly. Some stronger effects of Web user patterns in hourly increments might be normalized by the contribution of queries made at different time zones. In this study the statistical values in each hourly increment can be seen as a moving average of their relevant values coming from three time zones in the US. In addition, there could be some truncation effects in the analysis of the last hour of the log data (16:00-17:00 for Excite and 23:00-24:00 for FAST). Since the analysis is truncated at 17:00 for Excite and 24:00 for FAST, some trends can be reflected weaker and stronger than they are. For example, for a session that

begins between 16:00-17:00 hours for Excite and continues beyond 17:00 hours, this study only considers the queries made between 16:00 and 17:00, which might cause a lower query per session value. Considering these difficulties, it should be noted that the objective of the paper is to provide insight in the general trends of Web use throughout the day, i.e. comparing morning to evening or late night, instead of comparing hourly or more detailed time segments. Despite the time difference between queries made by users in different time zones, the general trend of Web use throughout the day will still be evident, though some sharp effects might be leveled and there might be some noise in last hour data.

Fast Query Set

For the entire Fast sample query log of largely European users, the mean queries per session throughout the day was 10.3. The query and session arrival statistics with respect to hours of the day are provided in Table 2.

[Place Table 2 Here]

The number of queries and sessions are at quite a low number with 295 queries and 33 sessions around midnight. The number of arrivals for queries and sessions decreases further into the early morning hours, gets as low as 118 for queries and 12 for sessions around 5:00 AM. As the morning progresses, the number of session and query arrivals display a quite steady increase. The number of queries increases as high as 789 and the number of sessions as high as 57 during the late morning hours. After late morning and into the afternoon hours, the query and session arrivals decrease significantly, about 30-40% down to around 500 queries per hour and 50 sessions per hour around 2:00 PM. The number of queries and sessions further decrease into the

evening and night hours to 200 queries per hour and 26 sessions per hour around midnight February 7, 2001. The change in the number of hourly query and session arrivals can be seen in Figures 3 and 4, respectively.

[Place Figure 3 Here]

[Place Figure 4 Here]

These findings suggest that on average the level of interest by European users in Web searching during the day is higher than during the night. The trends in both Figures 3 and 4 are very similar. Both the number of queries and number of sessions submitted per hour seem to be higher for hours of the day than the hours of the evening. The mid-morning to noon hours seem to be the prime-time for Web searching. As the hours progress into afternoon, a lower level of Web searching is observed. This shows that on average a European user's need for Web searching diminishes during the day. The decrease in Web searching continues and is actually more emphasized later into the day and during the night, reaching very low levels later into the night. These results might suggest that Web users might be doing a considerable amount of Web searching throughout their daily routines, but prefer other forms of entertainment in the evening, like TV. It should be noted that the effect of queries submitted to the Excite search engine by Web users in different time zones is also valid for the Fast search engine; therefore general inferences are drawn from the analysis such as comparing morning hours to evening hours instead of making specific hourly comments.

, [Place Figure 5 Here]

The trend in the mean queries per session can be seen in Figure 5. The mean queries does not display a certain trend, but rather is around 9-11 queries/session with some increases and decreases during the day.

Session and Query Duration Analysis

Excite Query Set

An analysis of session and query durations is useful in order to support the finding that information retrieval needs decrease as the day progresses and depict whether the information demand is also decreasing in terms of durations. The mean Excite session duration was 2265.3 for 1064 sessions submitted throughout the day. The session duration with respect to day hours is shown in Table 3 and Figure 6.

[Place Table 3 Here]

[Place Figure 6 Here]

Excite session durations decline sharply from 3988 around 9:00 am to 328.4 around 5:00 pm – about 91.7% - as the day progresses. This takes place as the sessions per hour also decreases. It can be concluded that utilization of the Web search engine decreases sharply both in terms of time and number of session arrivals as the day progresses. Hence, an allocation of resources for the Web search engine is appropriate with respect to hours of the day. Intelligent use of the available resources will contribute to more efficient operations and lower costs.

The duration analysis can be completed with the investigation of the duration of queries. The mean query duration throughout the day was 600.3 for 3188 queries. The query duration with respect to day hours is given in Table 4 and Figure 7.

[Place Table 4 Here]

[Place Figure 7 Here]

Excite query durations also display the strong tendency to decline from 979.2 to 138.8-– about 85.8%- as the day progresses. This tendency is also consistent with the decreasing number of query arrivals. The query and session duration analyses and query and session arrival analyses strongly indicate that the use of the Excite Web search engine decreases significantly as the day progresses. The use of the Web search engine diminishes both in terms of quantities (of search queries) and also in terms of durations. It is clearly obvious that the characteristics of Web search engine user sessions demonstrate variations depending on the time of the day.

Fast Query Set

In the previous section, we have found that Web searching is more prevalent during the morning hours and decrease later into the day for the Fast search engine users. A duration analysis would be helpful in further analyzing Web user behavior. The mean Fast session duration throughout the day was 8238.9 for 903 sessions. The number of sessions included in the duration analysis is less than the total number of sessions in the dataset, since the single query sessions cannot be included in the duration analysis. The mean query duration throughout the day was 827.7 for 9037 queries. The session and query duration with respect to workday hours is given in Table 5.

[Place Table 5 Here]

The mean duration of both Fast queries and sessions are higher during the day and lower later in the day, i.e. during the afternoon and evening hours. Both mean query and sessions durations are quite high during the early morning and mid-morning hours;

between 1000-1500 seconds for queries and 10,000-15,000 seconds for sessions. A lower level of mean session and query durations is observed during the afternoon, evening and night hours, session durations only reaching 3000-4000 seconds and query durations 300-500 seconds. This is a decrease around 70% for query and session durations. .The mean query and session durations are shown in Figures 8 and 9.

[Place Figure 8 Here]

[Place Figure 9 Here]

The results are similar to those observed for the Excite search engine. The use of the Fast Web search engine decreases in terms of search queries submitted and also in terms of durations of queries and sessions.

Analysis of Search Terms

The arrival and the duration analysis of search sessions and queries provide information on the technical characteristics of the search procedure. Besides the technical characteristics, it is important to capture the insight to a search session. This insight can only be obtained by a term analysis. We present a detailed analysis of terms used during the Web search sessions of various users. We begin the investigation with the analysis of the number of search terms used within queries, and continue with an in depth analysis of the most frequently used search terms.

Terms per Query

The analysis of terms per query sheds light to the length of the queries made by the Web search engine users.

Excite Terms Per Query

The mean terms per Excite query within day hours and the mean change in the number of terms in consecutive queries are presented in Table 6. The results can also be seen Figure 10 and Figure 11.

[Place Table 6 Here]

[Place Figure 10 Here]

[Place Figure 11 Here]

The first result derived from this analysis shows no significant change in the number of terms per Excite query with respect to hours of the day.. The change in the number of terms in the consecutive queries is also consistent from the beginning of the day to the end of the day. There is a sharp decrease in the number and durations of sessions and queries, but the quality of the queries in terms of the number of search terms stay the same as the day progresses.

Fast Terms Per Query

The mean terms per query for the Fast search engine is presented in Table 7. The results can also be seen Figure 12.

[Place Table 7 Here]

[Place Figure 12 Here]

In the Fast dataset, there is no significant change in the number of terms per query with respect to hours of the day. The mean number of terms per query is around 2.5 with slight variations throughout the day. Similar findings as in the Excite dataset are also observed for the Fast dataset.

Top Terms Analysis

A further look into the term analysis involves looking into the top terms used in the Web searches. The top terms analysis opens the window to the content analysis of queries. The most frequently used terms with respect to hours of the day are identified within the entire sample data log. The objective of this portion of this study is to determine whether there is a pattern or trend in the contents of the top terms with respect to hours of a day. We will only report the first ten top terms in the searches. The list contains only the terms, which depict some search meaning. Such a list has been generated using a two-step procedure and through eliminating some characters and terms. The first step is about eliminating characters such as "?", "&", "%", "$", "@", "#", ".", ",", ";", ":", """, "'", "!", "*", "(", ")", "[", "]", "{", "}", "-", "+", "=", "_", "~", "\", "/", "<", ">".

This step was performed by replacing all occurrences of non-letter characters with a space. Elimination of the character "." provided an interesting finding. The terms "com", "www" and "http" are in the top six terms among most common 10 terms. The high occurrence of the terms "com", "www" and "http" can not be detected unless “.” are replaced with a space. Web page searches are usually submitted in the form "http://www.######.com". Since the mid part of the web addresses vary significantly, each search query for web pages will be counted as a unique term unless the search is for the exact same web page. The second step of the term analysis involved determining the pure search terms. After elimination of the characters, the list of most frequently used terms included words such as “http”, “com”, “and”, “or”, etc. Such terms do not really reflect any information about the content of the search query; hence they have been

eliminated from the study. Considering these guidelines, the results of the top terms analysis is as below.

Excite Top Terms Analysis

This portion of the analysis presents the top term analysis for the Excite dataset. The most frequently used ten pure terms in the Excite dataset with respect to hours of the day is reported in Table 8.

[Place Table 8 Here]

The analysis of the most frequently used Excite terms did not yield any specific results. We have not detected any trend in the content of the Web searches with respect to hours of the day.

Fast Top Terms Analysis

This portion of the analysis presents the top term analysis for the Fast dataset. Since the Fast dataset expands over 24 hours, it enables an analysis of terms used during the day and night hours. The most frequently used ten pure terms in the Fast dataset with respect to hours of the day is reported in Tables 9 and 10.

[Place Table 9 Here]

[Place Table 10 Here]

There seems to be a slight increase in use of sexual terms during the night hours compared to the day hours. There are a few sexual terms in top ten pure terms during night and early morning hours, whereas there are almost no sexual terms among the top ten terms between 9:00 AM and 3:00 PM. Sexual terms are replaced by other terms such as business or computer related terms, etc. during work hours. Consequently, it could be concluded that the topics of Web searches could change based on time of the day.

Type of Excite Queries

Another term analysis relates to the type of Excite queries, i.e. whether a query is unique, has been modified from the previous query or is the next page for the preceding query. The number of unique, next page and modified queries and their percentages with respect to total queries are given in Table 11. The chart for the percentages of the queries can be seen in Figure 13.

[Place Table 11 Here]

[Place Figure 13 Here]

The unique queries form about 50% of the total number of queries submitted in an hour, while the next page queries and modified queries form about 35 % and 15 %, respectively, of the total number of queries in an hour. The percentage of both unique and next page queries and the modified queries do not show a significant change as the day progresses. The range for the percentage to the total number of queries in a specific hour is 45.7% - 49.1% for unique queries, 30.7% - 40.5% for next page queries and 12.2 % - 20.1% for modified queries. Hence, as the day progresses the quality of the queries stay the same not only in terms of the number of terms per query; but also in terms of the type of queries.

Markov Analysis of Excite Queries

This portion of the study analyzes hourly changes in the Markov matrix for the queries of the user sessions of the Excite data log. Initially, the frequencies of transitions from one type of query to another type of query within a session are investigated,

followed by the limiting probabilities for different types (states) of queries, i.e. unique query, next page and modified query states.

Tables 12 and 13 show the hourly frequency matrices for transitions between unique, next page and modified query states (and from these states to the end state).

[Place Table 12 Here]

[Place Table 13 Here]

The number of transitions from a certain type of query to another type of query is investigated, such as from unique queries to unique queries, from unique queries to modified queries, etc. The rows of each portion of the table show the previous query state of consecutive queries, whereas the columns show the subsequent query state of consecutive queries. Within the same row and column format, Tables 12 and 13 also show the initial ratio of the transition of queries from one state to another. The initial ratios for transitions from state *i* to state *j* is calculated by dividing the number of transitions from state *i* to state *j* to the total number of the queries originating from state *i*.

There is not much difference in the transition of Excite queries originating from states P and M to other states with respect to hours of the day. Hence, no matter what the hour is, if a user is already in a next page query or has submitted a modified query, the probability of going to other states stays more or less the same. Regardless of the hours of the day, about 10-15% of next page queries are followed by unique queries, about 50%-60% followed by more next page queries and about 5%-10% followed by modified queries. Similarly for unique queries, regardless of the hours of the day, 10%-15% are followed by more unique queries, 25%-30% followed by next page queries and 10%-15% followed by modified queries. The change in hours of the day causes a slight change in

the transition of queries originating from the modified queries state. However, the change does not have a specific pattern, so it might be deduced that the slight change is random.

Once a user views the next page, he/she will continue looking at the following pages of results with over 50% chance, and with approximately 25% chance they will end the session. In other words, once they begin looking at the next pages, with over 75% chance they will not modify the query or submit a new query. In addition, generally, users submit unique queries before ending the session. The second portion of the Markov analysis emphasizes limiting probabilities for unique, next page and modified query states (for this portion of the study the end state is ignored to be able to calculate the limiting probabilities of the other three states U, M and P).

The limiting probabilities of queries show the proportion of time that the Markov chain visits the respective query states over an extended period of time, in other words the limiting probabilities provide the average number of times users select each query type. The long-term ratios also provide information on the popularity of the query types over a long term. The limiting probabilities (π_j's, $j \geq 0$) for unique, next page and modified queries are calculated using the initial ratios (P_{ij}'s, the probabilities of going state j given the system is in state i (i, j = 1,2,3), which are given in Table 8) as follows (whereas j = 1 for unique query state, j = 2 for next page state, and j = 3 for modified query state):

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, \quad j \geq 0$$

$$\sum_{j=0}^{\infty} \pi_j = 1$$

The hourly limiting probabilities for unique queries, next page queries and modified queries are given in Table 14 and Figure 14.

[Place Table 14 Here]

[Place Figure 14 Here]

The long-term ratio of Excite unique queries has a declining trend as the day progresses. This finding actually verifies the findings of the previous sections of the results. Web users tend to spend less effort on Web searching as the day progresses. Instead of showing the effort of making new queries, Web users tend to continue the already submitted queries through next page queries or submit slightly modified queries.

Paired t-tests have been performed in order to compare the set of long-term ratios for different types of queries within different hours. For example, the statistical significance of the difference between the long-term ratios for unique queries, next page queries and modified queries for 9:00 AM-10:00 AM and the same set of long-term ratios for 10:00 AM-11:00 AM is tested. The testing is performed through conventional paired t-testing and has been repeated for all possible combinations of the hours of the day. Separate paired t-tests have been applied for pairs of 9:00-10:00 and 10:00-11:00, 9:00-10:00 and 11:00-12:00,……, and 15:00-16:00 and16:00-17:00.The t values for the paired t-tests are given in Table 15.

[Place Table 15 Here]

Table 15 shows the t values for paired t-testing of long-term ratios of query types among hours are very low. At the 95% confidence level, none of the t values demonstrate statistical significance. Hence, it can be concluded that the difference of the set of limiting probabilities for unique, next page and modified queries from any hour of the

day to any other hour is statistically insignificant. However, as mentioned previously, differently from next page and modified queries, the limiting probability of unique queries seems to be on the decline as the day progresses.

The top 20 web user session types might provide a deeper insight to the Markov analysis for query types and are provided in Table 16

[Place Table 16 Here]

Clearly, Excite sessions with single query (type "U") are the most common type in all hours, and the ratios of session types UU, UP, UM (sessions with two queries) are the following most common session types. The distribution of other session types differs with respect to the hours of the day. However, the number of sessions with more than three queries is insignificant compared to number of sessions with one or two queries. Therefore, hourly changes in the types of sessions with three or more queries are not a significant finding.

The ratio of UU and UM are declining compared to the use of session type UP within the top four Excite session types. In other words, users are more reluctant to submit new queries or modify their original queries as the day progresses.

DISCUSSION

Our findings suggest that Web user behavior might fluctuate from the beginning of the day to the end of a day. There are sharp changes in some characteristics of Web querying based on time of the day. More search queries are submitted to Excite and Fast during the morning hours, than during the afternoon or evening hours. The session and query durations have also decreased as the day progressed from the morning hours.

While the technical characteristics of the Web search queries are affected as the day progresses, the quality of Web search queries may remain almost the same from the beginning of the day to the end of the day. For both the Excite and Fast datasets, the terms per query analysis did not reveal any interesting changes in the quality of Web searches within different time frames. There is also no significant difference in the most frequently used terms during the day for the Excite dataset, whereas sexual terms seems to be more prevalent in the evening hours for the Fast dataset. In addition, in the Excite dataset, query reformulation, ratio of unique, modified and next-page queries remained steady through different hours of the day. A summary of the changes in Excite and Fast datasets from morning hours to afternoon and evening hours can be seen in Table 17.

[Place Table 17 here]

In addition, the Markov analysis of queries with respect to hours did not yield any specific trend between the hours of the day. The behavior of Web user searches in terms of transition from one type of query to another in consecutive queries within a session seems to stay the same as well as the long-term ratios of unique, modified and next page queries. This finding further supports the fact that while the technical characteristics of the Web search queries, such as query and session arrivals and durations are affected as the day progresses, the quality of the search sessions stays the same both in terms of search terms and in terms of the structure of the queries.

Overall, Web searching needs were at their highest levels during morning hours and decreased into the afternoon hours and evening hours. There is a strong indication that the other characteristics of Web search might also vary based on the time of the day. The analysis on the remaining characteristics of the Web user inquiry sessions is proposed as

the extension of this research. We believe that time sensitive Web search engine development will provide more intelligent and cost efficient systems by redesigning the allocation of their sources and user interfaces.

CONCLUSION

Our research suggests changes in user Web searching based on the time of the day. Overall, some characteristics of Web search queries, such as arrival and durations of sessions and queries, are at their highest during the morning hours and decrease later into the day. Other characteristics, such as the quality of queries in terms of number of terms per query and reformulation of queries, stay the same throughout the day. Our findings and the analysis of further data sets can be useful to Web search engines in reconstructing their search structure and reallocating their resources with respect to different time frames.

ACKNOWLEDGMENT

The authors thank Doug Cutting, Jack Xu and Soo Young Rieh from Excite@Home.Com and Per G. Auran from Fast.com for providing the Web query data sets.

REFERENCES

Bhatia, S. K., & Deogun, J. S. (1998). Conceptual clustering in information retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 28*(3), 427-436.

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

Brewington, B. E., & Cybenko, G. (2000). How dynamic is the Web? *Proceedings of the 9th World Wide Web Conference, May 2000, Amsterdam, Netherlands.*

Ling, C. X., & Li, C. (1998). *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 73-79.

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

Ozmutlu, H. C., & Spink, A. (2001). Time-based analysis of search data logs. *Proceedings of Internet Computing'01 Conference on Internet Computing, June 25-28, 2001,* Vol.1, pp. 41-46.

Ozmutlu, H. C., Spink, A., & Ozmutlu, S. (2002). Analysis of large data logs: An application of Poisson sampling to Excite Web queries. *Information Processing and Management,* 38(3), 473-490.

Silverstein, C., Henzinger, M., Marais, H., &Morris, M. (1999). Analysis of a very large Web search engine query log, *ACM SIGIR Forum,* 33(3).

Spink, A., Bateman, J., & Jansen, B. J. (1999). Searching heterogeneous collections on the Web: A survey of Excite users. *Internet Research: Electronic Networking Applications and Policy, 9*(2), 117-128.

Spink, A., Jansen, B. J., & Ozmutlu, H. C. (2000). Use of query reformulation and relevance feedback by Web users. *Internet Research: Electronic Networking Applications and Policy, 10*(4), 317-328.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer, **35***(3), 133-135.

Spink, A., Wolfram, D., Jansen, M.B.J. and Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3): 226–234

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

Table 1. Mean queries per user session, number of session arrivals and number of query arrivals with respect to hours of the day - Excite Query Set

Hour of the Day	Mean Queries Per Session	Number of Hourly Session Arrivals	Number of Hourly Query Arrivals
9:00-10:00	3.9	174	679
10:00-11:00	3.2	150	486
11:00-12:00	3.06	143	437
12:00-13:00	2.6	141	367
13:00-14:00	2.7	130	358
14:00-15:00	2.7	120	333
15:00-16:00	2.9	105	304
16:00-17:00	2.2	101	224

Figure 1. Mean queries per user session with respect to hours of the day - Excite Query Set.

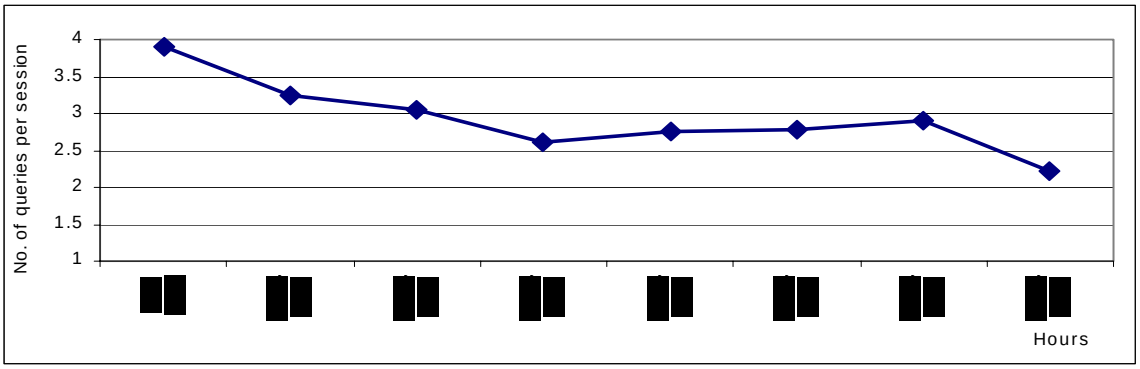

Figure 2. Session and query arrivals with respect to hours of the day - Excite Query Set.

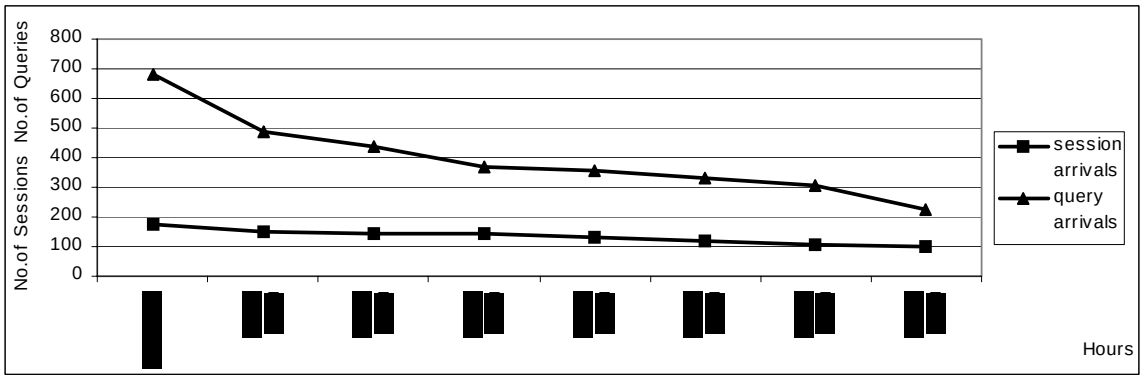

Table 2: The mean queries per user session, total number of session arrivals and total number of query arrivals with respect to hours of a day - Fast Query Set

Hour of the day	Mean Queries per Session	Number of Hourly Session Arrivals	Number of Hourly Query Arrivals
12:00-1:00	8.9	33	295
1:00-2:00	11.1	32	358
2:00-3:00	9.7	25	244
3:00-4:00	6.6	18	119
4:00-5:00	11.4	14	160
5:00-6:00	9.8	12	118
6:00-7:00	12.2	13	159
7:00-8:00	12.2	35	430
8:00-9:00	11.08	65	720
9:00-10:00	10.5	56	589
10:00-11:00	13.8	57	789
11:00-12:00	11.7	46	539
12:00-13:00	9.3	53	495
13:00-14:00	10.6	48	513
14:00-15:00	10.1	56	570
15:00-16:00	9.3	55	516
16:00-17:00	10.6	47	500
17:00-18:00	10.7	39	392
18:00-19:00	10.6	56	596
19:00-20:00	9.9	46	459
20:00-21:00	11.5	44	507
21:00-22:00	8.6	48	416
22:00-23:00	8.07	40	323
23:00-24:00	7.6	26	200

Figure 3: Query arrivals with respect to hours of a day - Fast Query Set

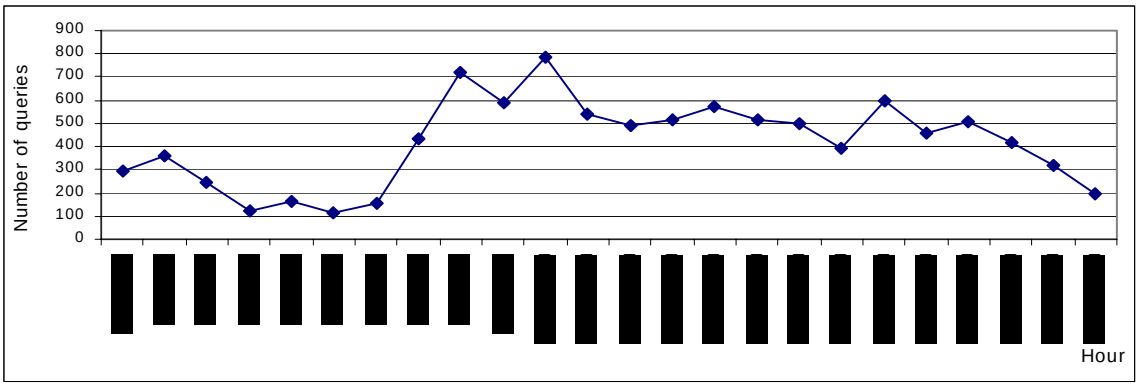

Figure 4: Session arrivals with respect to hours of a day - Fast Query Set.

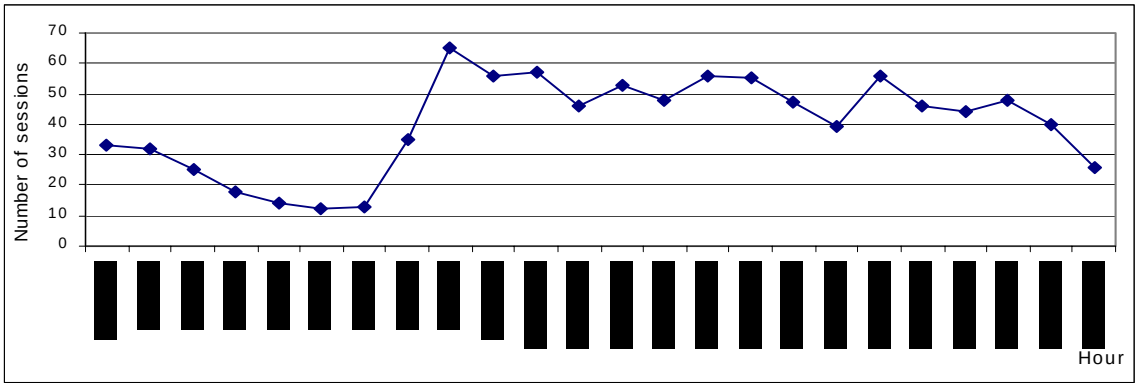

Figure 5: Mean queries per user session with respect to hours of a day - Fast Query Set.

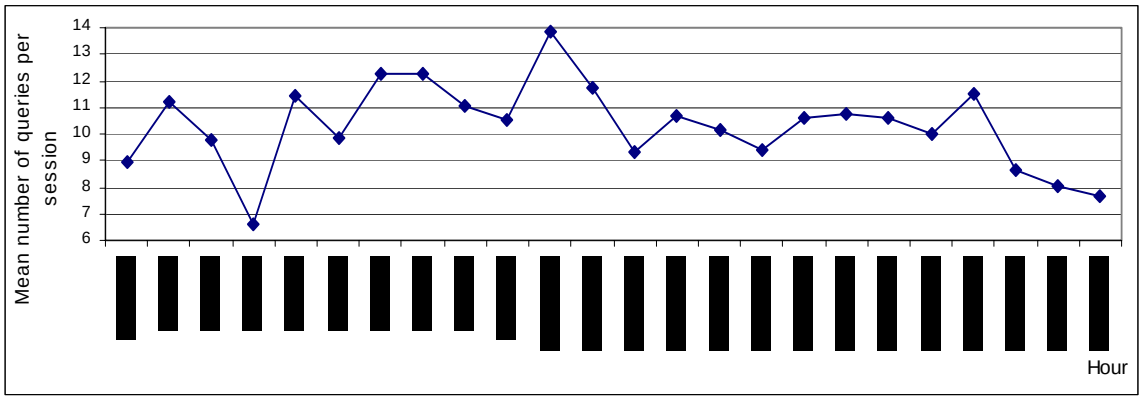

Table 3. Mean session durations with respect to hours of the day (in seconds) - Excite Query Set.

Hour of the Day	Mean Session Duration
9:00-10:00	3988.6
10:00-11:00	3487.5
11:00-12:00	2167.7
12:00-13:00	1914.7
13:00-14:00	1317.2
14:00-15:00	1183.6
15:00-16:00	756.2
16:00-17:00	328.4

Figure 6. Mean session durations with respect to hours of the day (in seconds) - Excite Query Set.

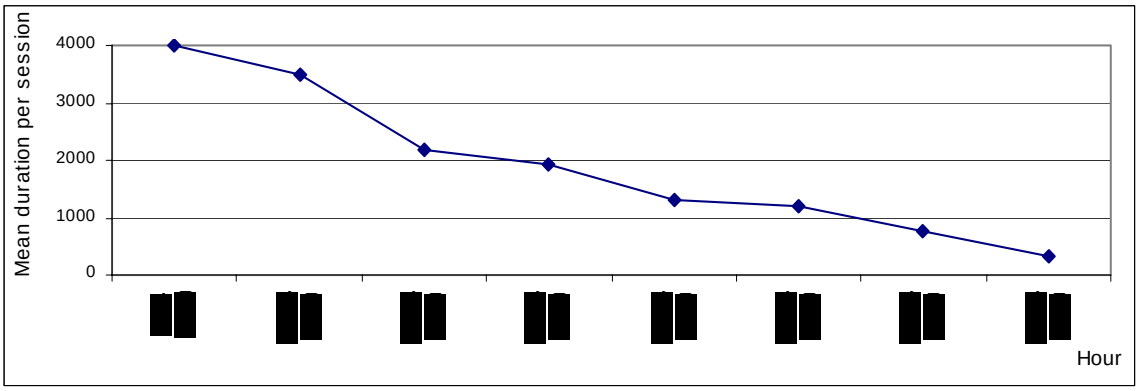

Table 4. Mean query durations with respect to hours of the day (in seconds) - Excite Query Set.

Hour of the Day	Mean Duration Per Query
9:00-10:00	979.2
10:00-11:00	975.6
11:00-12:00	676.7
12:00-13:00	635.4
13:00-14:00	433.3
14:00-15:00	344.5
15:00-16:00	228
16:00-17:00	138.8

Figure 7. Mean query durations with respect to hours of the day (in seconds) - Excite Query Set.

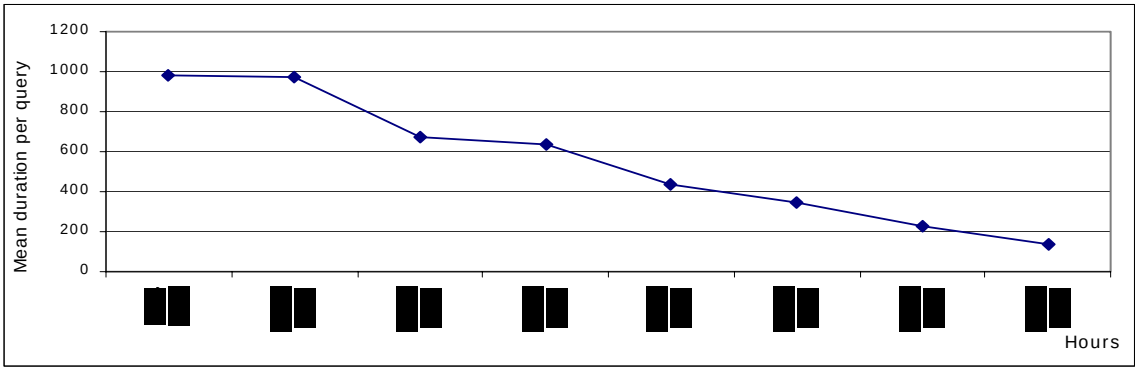

Table 5: The mean session and query durations with respect to hours of the day- Fast Query Set.

Hour of the Day	Mean Duration Per Query (seconds)	Mean Duration Per Session (seconds)
12:00-1:00	709.3	5994.8
1:00-2:00	1588.4	17261.4
2:00-3:00	1628.8	15509.09
3:00-4:00	1214.4	7215.4
4:00-5:00	1181.1	12317.4
5:00-6:00	1475.8	14222.1
6:00-7:00	1343.7	15091.8
7:00-8:00	1347.5	16633.4
8:00-9:00	1300.3	13636.09
9:00-10:00	1304.8	12879.3
10:00-11:00	944.8	12574.6
11:00-12:00	577.7	6473.7
12:00-13:00	839.9	7424.8
13:00-14:00	565.2	5973.1
14:00-15:00	714.5	6929.6
15:00-16:00	595.8	5722.8
16:00-17:00	427.3	4501.6
17:00-18:00	694.02	6999.6
18:00-19:00	554.1	5400.8
19:00-20:00	365.1	3427.3
20:00-21:00	277.08	2651.9
21:00-22:00	456.9	3577.6
22:00-23:00	476.3	3456.7
23:00-24:00	691.7	5335.8

Figure 8: The mean query durations with respect to hours of the day - Fast Query Set.

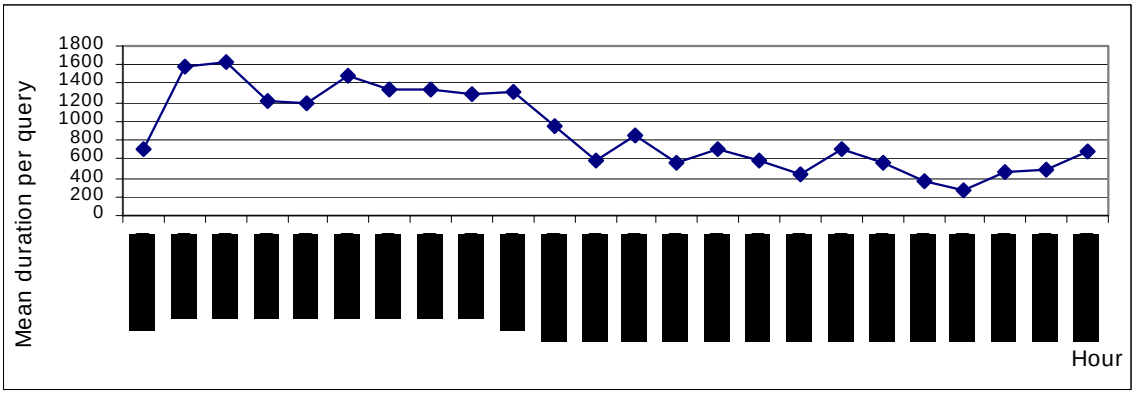

Figure 9: Mean session durations with respect to hours of the day- Fast Query Set.

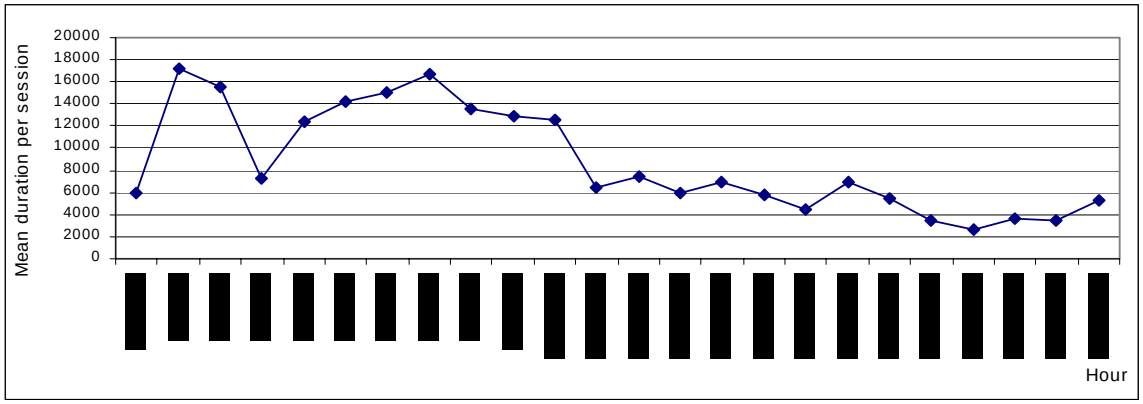

Table 6. Mean terms per query and mean changes in terms used in consecutive queries with respect to hours of the day - Excite Query Set.

Hour of the Day	Mean Terms Per Query	Mean Changes in Terms Per Query in Consecutive Queries
9:00-10:00	2.3	0.4
10:00-11:00	2.4	0.8
11:00-12:00	2.5	0.5
12:00-13:00	2.6	0.5
13:00-14:00	2.5	0.5
14:00-15:00	2.4	0.4
15:00-16:00	2.5	0.4
16:00-17:00	2.2	0.3

Figure 10. Mean terms per query with respect to hours of the day - Excite Query Set.

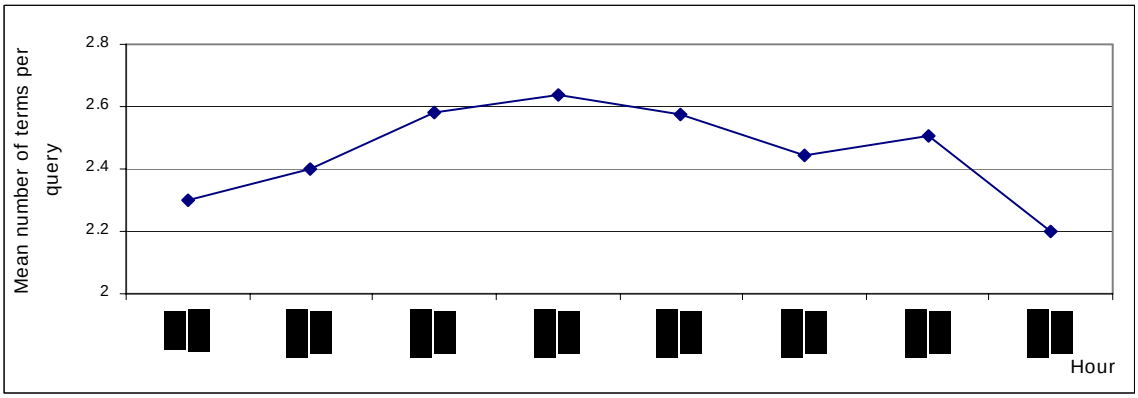

Figure 11. Mean changes in the number of terms in consecutive Excite queries with respect to hours of the day - Excite Query Set.

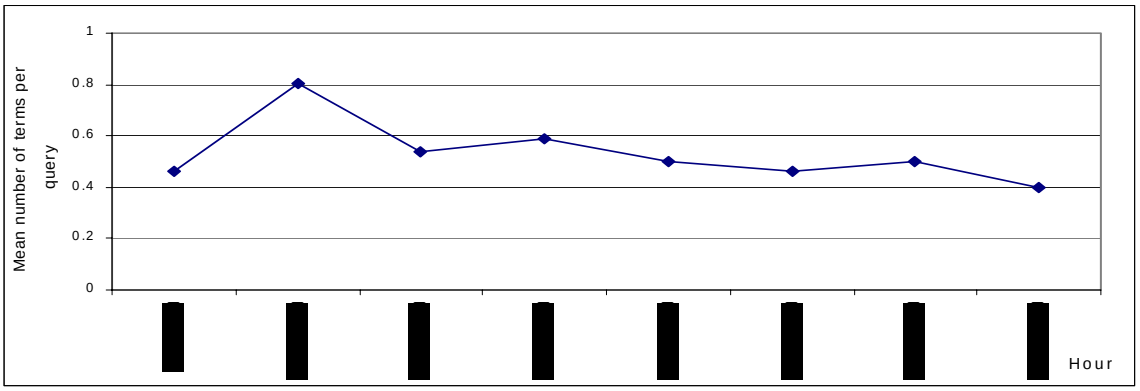

Table 7: Mean terms per query with respect to hours of the day - Fast Query Set.

Hour of the Day	Mean Terms Per Query
12:00-1:00	2.3
1:00-2:00	2.5
2:00-3:00	2.6
3:00-4:00	2.8
4:00-5:00	2.5
5:00-6:00	3.1
6:00-7:00	1.8
7:00-8:00	2.4
8:00-9:00	2.4
9:00-10:00	2.4
10:00-11:00	2.6
11:00-12:00	2.5
12:00-13:00	2.6
13:00-14:00	2.3
14:00-15:00	2.7
15:00-16:00	2.4
16:00-17:00	2.3
17:00-18:00	2.4
18:00-19:00	2.6
19:00-20:00	2.4
20:00-21:00	2.4
21:00-22:00	2.1
22:00-23:00	2.5
23:00-24:00	2.6

Figure 12: Mean terms per query with respect to hours of the day - Fast Query Set.

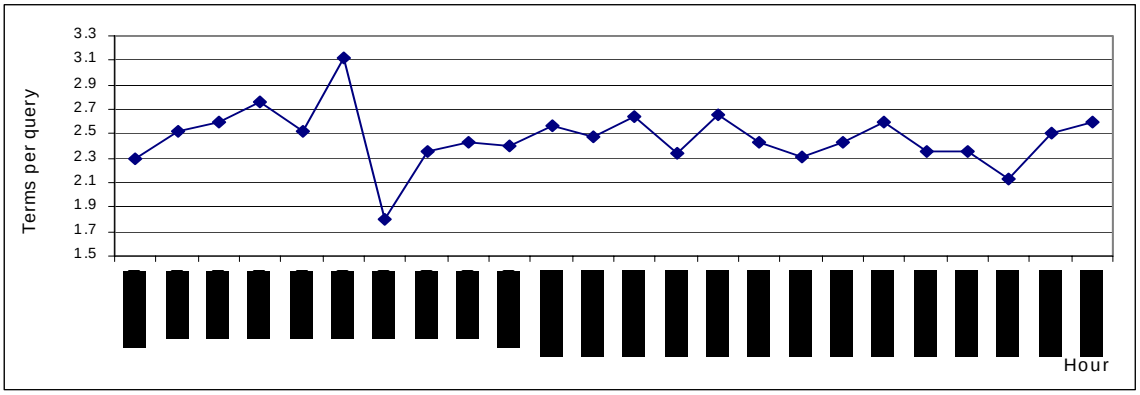

Table 8. List of the most frequently used ten terms with respect to hours of the day - Excite Query Set.

Hours of the Day							
9:00-10:00		10:00-11:00		11:00-12:00		12:00-13:00	
Search Term	Number	Search Term	Number	Search Term	Number	Search Term	Number
botanical	45	Free	14	braindumps	21	international	16
shampoo	45	Web	14	windows	21	site	14
hotel	16	Maids	13	cna	18	clubs	13
mpeg	15	Music	13	banking	16	free	13
pokemon	13	Christmas	12	risks	13	Stoughton	13
City	12	galleries	11	sex	12	travel	13
Free	12	themes	11	pc	11	Middle	12
illegal	10	black	10	anywhere	10	school	12
photos	10	bugs	10	demo	10	web	11
schoolgirls	10	fetish	10	free	9	Canon	10
13:00-14:00		14:00-15:00		15:00-16:00		16:00-17:00	
Search Term	Number	Search Term	Number	Search Term	Number	Search Term	Number
Free	16	Pictures	34	sex	23	free	15
New	13	Rape	33	millenium	14	nude	10
Pics	13	Banners	19	pc	14	Mortal kombat	9
bands	10	CRAFTS	12	poems	12	home	7
Steel	10	Angels	11	FREE	10	dams	6
Beach	9	Free	10	playstation	10	downey	6
jones	9	Quartet	10	Rat	10	usa	6
Lee	9	String	10	Terrier	10	built	5
Myrtle	9	Angel	9	toro	10	DCCC	5
SC	9	Clipart	9	snowthrowers	9	last	5

Table 9. List of the most frequently used ten pure terms with respect to hours of the day - Fast Query Set.

Hours of the Day							
24:00-1:00		1:00-2:00		2:00-3:00		3:00-4:00	
Search Term	Number	Search Term	Number	Search Term	Number	Search Term	Number
black	26	animal	20	berlin	28	channel	17
pantyhose	25	in	20	managerial	12	country	17
ardilla	19	testing	20	nose	11	music	17
clipper	18	bear	16	of	9	free	10
cars	17	gay	16	plastic	8	interacial	9
used	17	the	16	surgery	8	job	8
girls	14	hedland	15	lln	7	nude	8
fat	12	port	15	airport	6	of	8
crazy	11	condoms	14	pengantar	6	according	7
mixed-up	11	on	14	the	6	gospel	7
4:00-5:00		5:00-6:00		6:00-7:00		7:00-8:00	
Search Term	Number	Search Term	Number	Search Term	Number	Search Term	Number
preteen	29	visual	14	banners	27	lolitas	21
girl	22	i	9	powerdvd	15	download	18
is	17	sex	9	Maw	11	h	15
what	17	directory	6	taurus	11	2	12
cattleya	12	gallery	6	testnews	9	capi	12
celebrity	12	nude	6	wasserstrahlpumpen	9	mp3	12
wanted	12	parent	6	bilder	7	linux	12
photos	10	good	4	nikki	7	rotor	12
gis	9	programz	4	nova	7	gas	11
nude	7	so	4	tattoo	7	oil	11
8:00-9:00		9:00-10:00		10:00-11:00		11:00-12:00	
Search Term	Number	Search Term	Number	Search Term	Number	Search Term	Number
freud	25	mp3	24	download	45	uk	31
lucian	25	halford	24	uk	35	free	28
pictures	23	and	21	full	34	business	18
webvirgins	21	of	19	version	32	leads	18
isles	19	pl	17	appz	28	medway	18
photographs	19	n	17	shemale	24	acdsee	16
scilly	19	immobilien	17	nantwich	23	donny	16
takraf	19	gql	15	r	20	osmond	16
asphalt	17	treatment	14	corporate	19	800	15
in	17	sally	14	governance	19	201xp	15

Table 10. List of the most frequently used ten pure terms with respect to hours of the day - Fast Query Set.

Hours of the Day							
12:00-13:00		13:00-14:00		14:00-15:00		15:00-16:00	
Search Term	Number	Search Term	Number	Search Term	Number	Search Term	Number
2000	29	film	34	university	35	de	34
windows	28	free	24	editor	29	compiler	28
server	27	umbrella	23	registry	29	ayuntamiento	27
mp3	26	conveyor	22	Tips	29	sexshare	23
appz	21	omen	21	Carr	20	and	20
printing	19	the	21	Cory	20	yahoo	20
a	18	folding	19	Of	18	glockenspiel	19
lba	17	in	15	barron	17	norma	19
mode	17	modular	15	S	17	stitz	19
of	17	dnv	13	3d	14	ibo	17
16:00-17:00		17:00-18:00		18:00-19:00		19:00-20:00	
Search Term	Number	Search Term	Number	Search Term	Number	Search Term	Number
index	24	mcmahon	17	and	51	sex	27
manual	22	shane	17	free	31	ficken	23
fence	22	car	16	de	20	free	23
electric	22	mp3	16	garters	18	schule	23
kiedis	20	lolita	15	pictures	17	the	23
anthony	20	print	14	transsexual	17	fax	20
designer	18	pro	14	resume	15	internet	20
archive	18	thumb	14	uniform	15	der	19
noble	16	agua	13	c	14	loudspeaker	18
barnes	16	care	13	lavoratori	14	via	18
20:00-21:00		21:00-22:00		22:00-23:00		23:00-12:00	
Search Term	Number	Search Term	Number	Search Term	Number	Search Term	Number
ciego	40	girls	51	and	24	blade	13
del	40	webcam	21	of	23	runner	13
la	40	banks	19	eltron	21	battery	10
noche	40	outer	19	sex	20	intelligent	10
terror	40	pizza	19	adult	18	powerbook	10
madrid	18	recipe	19	cards	16	public	10
a	17	bicicletas	17	greeting	16	3	9
for	17	trigger	17	seventeen	16	decq	9
half	17	bitters	16	club	15	driver	9
lords	17	copperplate	16	luis	15	gratification	9

Table 11. Number of unique, modified and next page queries and their percentages to total number queries with respect to hours of the day - Excite Query Set.

Hour of the Day	Unique Queries		Next Page Queries		Modified Queries	
	Number	%	Number	%	Number	%
9:00-10:00	321	47%	275	40%	83	12%
10:00-11:00	236	48%	186	38%	64	13%
11:00-12:00	213	48%	157	35%	67	15%
12:00-13:00	180	49%	116	31%	71	19%
13:00-14:00	176	49%	110	30%	72	2%
14:00-15:00	156	46%	128	38%	49	14%
15:00-16:00	139	45%	123	40%	42	13%
16:00-17:00	124	55%	72	32%	28	12%

Figure 13. Percentage of unique, next page and modified queries to total number queries with respect to hours of the day - Excite Query Set.

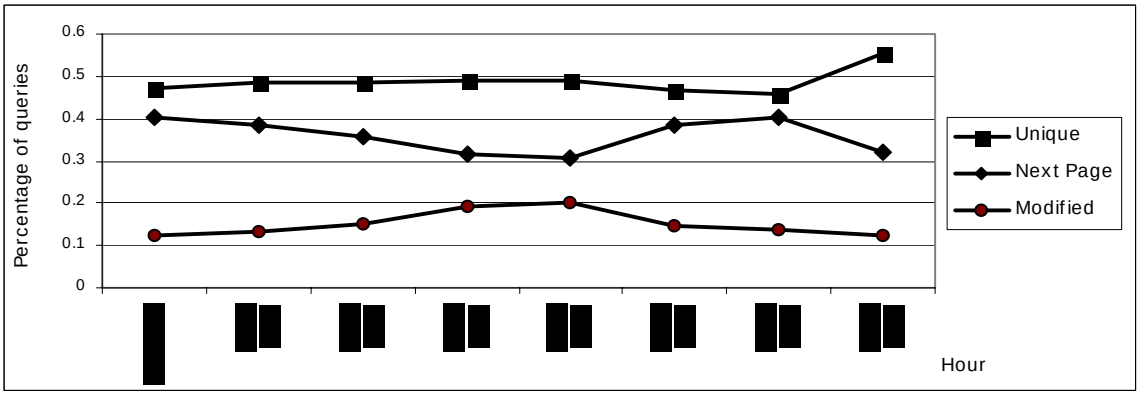

Table 12. The frequency matrix for transition from one type of query to another and the initial ratios for the transitions (U: Unique queries, P: Next Page queries, M: Modified Queries) - Excite Query Set.

9:00 – 10:00									
Number of Query Transitions					Initial Ratios of Queries				
to / from	U	P	M	END	to / from	U	P	M	END
U	93	96	41	91	U	0.28972	0.299065	0.127726	0.283489
P	37	156	17	65	P	0.134545	0.567273	0.061818	0.236364
M	17	23	25	18	M	0.204819	0.277108	0.301205	0.216867
END	0	0	0	1	END	0	0	0	1
10:00 – 11:00									
Number of Query Transitions					Initial Ratios of Queries				
to / from	U	P	M	END	to / from	U	P	M	END
U	46	82	27	81	U	0.194915	0.347458	0.114407	0.34322
P	27	91	20	48	P	0.145161	0.489247	0.107527	0.258065
M	13	13	17	21	M	0.203125	0.203125	0.265625	0.328125
END	0	0	0	1	END	0	0	0	1
11:00 – 12:00									
Number of Query Transitions					Initial Ratios of Queries				
to / from	U	P	M	END	to / from	U	P	M	END
U	31	63	34	85	U	0.14554	0.295775	0.159624	0.399061
P	30	75	17	35	P	0.191083	0.477707	0.10828	0.22293
M	9	19	16	23	M	0.134328	0.283582	0.238806	0.343284
END	0	0	0	1	END	0	0	0	1
12:00 – 13:00									
Number of Query Transitions					Initial Ratios of Queries				
to / from	U	P	M	END	to / from	U	P	M	END
U	23	45	32	80	U	0.127778	0.25	0.177778	0.444444
P	11	51	13	41	P	0.094828	0.439655	0.112069	0.353448
M	5	20	26	20	M	0.070423	0.28169	0.366197	0.28169
END	0	0	0	1	END	0	0	0	1
13:00 – 14:00									
Number of Query Transitions					Initial Ratios of Queries				
to / from	U	P	M	END	to / from	U	P	M	END
U	29	47	29	71	U	0.164773	0.267045	0.164773	0.403409
P	13	48	15	34	P	0.118182	0.436364	0.136364	0.309091
M	4	15	28	25	M	0.055556	0.208333	0.388889	0.347222
END	0	0	0	1	END	0	0	0	1

Table 13. The frequency matrix for transition from one type of query to another and the initial ratios for the transitions (U: Unique queries, P: Next Page queries, M: Modified Queries) - Excite Query Set.

14:00 – 15:00									
Number of Query Transitions					Initial Ratios of Queries				
to / from	U	P	M	END	to / from	U	P	M	END
U	18	46	20	72	U	0.115385	0.294872	0.128205	0.461538
P	11	73	14	30	P	0.085938	0.570313	0.109375	0.234375
M	7	9	15	18	M	0.142857	0.183673	0.306122	0.367347
END	0	0	0	1	END	0	0	0	1
15:00 – 16:00									
Number of Query Transitions					Initial Ratios of Queries				
to / from	U	P	M	END	to / from	U	P	M	END
U	19	46	19	55	U	0.136691	0.330935	0.136691	0.395683
P	13	70	6	34	P	0.105691	0.569106	0.04878	0.276423
M	2	7	17	16	M	0.047619	0.166667	0.404762	0.380952
END	0	0	0	1	END	0	0	0	1
16:00 – 17:00									
Number of Query Transitions					Initial Ratios of Queries				
to / from	U	P	M	END	to / from	U	P	M	END
U	14	34	16	60	U	0.112903	0.274194	0.129032	0.483871
P	3	30	7	32	P	0.041667	0.416667	0.097222	0.444444
M	6	8	5	9	M	0.214286	0.285714	0.178571	0.321429
END	0	0	0	1	END	0	0	0	1

Table 14. Hourly long-term ratios for unique queries, next page queries and modified queries – Excite Query Set.

	Types of Queries		
Hour of the Day	Unique (π_U)	Next Page (π_P)	Modified (π_M)
9:00-10:00	0.253447	0.592505	0.154048
10:00-11:00	0.241728	0.555489	0.202783
11:00-12:00	0.236007	0.545937	0.218055
12:00-13:00	0.144447	0.562393	0.293159
13:00-14:00	0.157657	0.493626	0.348717
14:00-15:00	0.155253	0.605526	0.239221
15:00-16:00	0.142427	0.639886	0.217687
16:00-17:00	0.144913	0.651248	0.203838

Figure 14. Hourly long-term ratios for unique queries, next page queries and modified queries - Excite Query Set

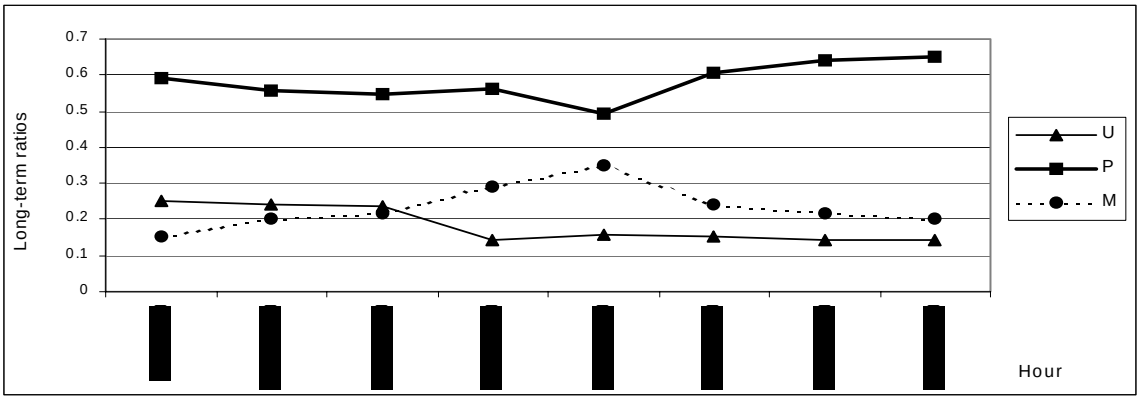

Table 15. Paired t-testing results for comparing the long-term ratios of different query types with respect to hours - Excite Query Set.

Hour / Hour	10:00-11:00	11:00-12:00	12:00-13:00	13:00-14:00	14:00-15:00	15:00-16:00	16:00-17:00
9:00-10:00	2.791 -06	5.712 -06	1.949 -06	-1.97 -06	-2.58 -06	2.393 -08	3.42 4-06
10:00-11:00		1.529 -05	1.32 -06	-3.59 -06	-4.8 -06	-1.301 -06	2.069 -06
11:00-12:00			-9.493 -07	-5.79 -06	-7.79 -06	-3.467 -06	-5.25 -08
12:00-13:00				-9.16 -06	-9.81 -06	-3.2 -06	8.425 -07
13:00-14:00					8.5 -07	2.4 -06	4.313 -06
14:00-15:00						7.986 -06	1.349 -05
15:00-16:00							2.501 -05

Table 16. Top twenty session types with respect to hours of the day (U: Unique queries, P: Next Page queries, M: Modified Queries) - Excite Query Set.

Type of Session	Hours of the Day							
	9:00-10:00	10:00-11:00	11:00-12:00	12:00-13:00	13:00-14:00	14:00-15:00	15:00-16:00	16:00-17:00
U	50	56	58	66	55	58	45	49
UP	28	17	14	15	12	15	13	18
UU	12	12	7	10	8	4	5	2
UM	6	7	10	10	10	8	5	4
UPP	6	3	2	5	5	1	3	3
UUU	5	1	2	1	1	1	2	3
UUP	4	2	1	0	1	0	3	0
UMM	3	2	2	1	3	0	5	1
UPPP	2	4	1	2	1	5	1	1
UMU	2	1	2	0	0	1	0	0
UUUU	2	1	0	0	1	0	0	0
UPPPP	1	3	0	0	1	1	0	0
UPU	1	1	5	2	1	1	3	1
UPM	1	1	1	1	2	2	1	2
UUM	1	1	1	1	0	0	1	0
UPPUM	1	1	1	0	0	0	0	0
UPPMP	1	1	0	0	0	0	0	0
UUPPP	1	1	0	0	0	0	0	0
UPUU	1	0	2	1	0	1	0	0
UPPPPPP	1	0	0	0	1	1	0	0

Table 17. Summary of changes in percentage in various query and session characteristics from morning hours to later in the day, i.e. afternoon, evening and night hours - Excite and Fast query sets.

Query or session characteristic	Excite Query Set	Fast Query Set
Arrival of queries	67% decrease from morning to afternoon	75% decrease from morning to evening and night
Arrival of sessions	42% decrease from morning to afternoon	50% decrease from morning to evening and night
Duration of queries	86% decrease from morning to afternoon	70% decrease from morning to evening and night
Duration of sessions	92% decrease from morning to afternoon	70% decrease from morning to evening and night
Mean queries per session	43% decrease from morning to afternoon	No significant changes
Mean terms per query	No significant changes	No significant changes
Most frequently used terms	No significant changes	Sexual terms among top terms in evening hours; versus no sexual terms among top terms in morning hours

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 ~ Exp (λ)). 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 Equation 1.

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \begin{cases} 1-e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (1)$$

If x is generated according to an exponential distribution, then the outcome of cdf, $F(x)$ for $x \geq 0$, has a Uniform (0,1) distribution. Since random variables u ~ Uniform (0,1) are fairly easy to obtain, it is logical to use a formula where the interarrival time x ~ Exp (λ) can be obtained by a variable u ~ Uniform (0,1). By calculating the analytical inverse of the exponential cdf in Equation 1, we can develop the desired formula, which is stated in Equation 2. After each sample point, a new uniform number u has to be generated to calculate the next exponentially distributed interarrival time using Equation 2.

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

The other case of the 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 ~ 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 ≥ 0}. Note that the inter-arrival times of samples are distributed according to Poisson process, not the inter-arrival times of the process from where the samples are taken. The probability mass function of the Poisson process is given in Equation 3.

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

However, the analytical inverse of the Equation 3 is not available. Therefore the following algorithm is used to generate the Poisson variate *n* (Mann, et al, 1974)

Step 1: **Set $j = 0$ and $y_j = u_0$, where $u_j \sim$ Uniform (0,1), $j = 0,1,\ldots$,**

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

Step 3: **$j = j + 1$, and $y_j = u_j y_j - 1$**

Goto Step 2

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

The Excite and FAST 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.