Modelling and Predicting Web Page Accesses Using Burrell’s Model

Devanshu Dhyani, Sourav S. Bhowmick, and Wee Keong Ng

School of Computer Engineering, Nanyang Technological University,
Singapore 639798
{assourav, awkng}@ntu.edu.sg

Abstract. The significance of modeling and measuring various attributes of the Web in part or as a whole is undeniable. In this paper, we consider the application of patterns in browsing behavior of users for predicting access to Web documents. We proposed two models for addressing our specification of the access prediction problem. The first lays out a preliminary statistical approach using observed distributions of interaccess times of individual documents in the collection. To overcome its deficiencies, we adapted a stochastic model for library circulations, i.e., Burrell’s model, that accounts for differences in mean access rates of Web documents. We verified the assumptions of this model with experiments performed on a server log of accesses recorded over a six month period. Our results show that the model is reasonably accurate in predicting Web page access probabilities based on the history of accesses.

1 Introduction

The importance of measuring attributes of known objects in precise quantitative terms has for long been recognized as crucial for enhancing our understanding of our environment. In this paper, we focus on characterizing the usage of Web resources from the perspective of modeling and predicting Web page accesses. We begin by relating Web page access modeling to prediction for efficient information retrieval on the WWW and consider a preliminary statistical approach that relies on the distribution of interaccess times. The deficiencies are addressed by adapting a more refined stochastic model due to Burrell [1,2,3] to the context of Web page accesses. We discuss in detail results of experiments for verifying and applying Burrell’s model for access prediction.

Let us first elucidate the general access prediction problem. Our basic premise is that page accesses should be predicted based on universally available information on past accesses such as server access logs. Given a document repository and history of past accesses, we would like to know which documents are more likely to be accessed within a certain interval and how frequently they are expected to be accessed. The information used for prediction, typically found in server logs comprises the time and URL of an HTTP request. The identity of the client is necessary only if access prediction is personalized for the client. From this information about past accesses, several predictor variables can be determined,
for example, the frequency of accesses within a time interval and inter-access times.

2 Statistical Prediction

An obvious prediction is the time until the next expected access to a document, say \( A \). The duration can be derived from a distribution of time intervals between successive accesses. This kind of statistical prediction relates a predictor variable or a set of predictor variables to access probability for a large sample assumed to be representative of the entire population. Future accesses to a document can then be predicted from the probability distribution using current measurements of its predictor variable(s). A variant of this approach is to use separate distributions for individual documents measured from past accesses.

Let us illustrate the above approach for temporal prediction using interaccess time. Suppose \( f(t) \) is the access density function denoting the probability that a document is accessed at time \( t \) after its last access or its interaccess time probability density. Intuitively, the probability that a document is accessed depends on the time since its last access and duration into the future we are predicting. At any arbitrary point in time, the probability that a document \( A \) is accessed at a time \( T \) from now is given by \( \Pr\{A \text{ is accessed at } T \} = f(\delta + T) \) where \( \delta \) is the age or the time since the last access to the document. The function \( f(t) \) has a cumulative distribution \( F(t) = \sum_{t' = 0}^{\infty} f(t') \) which denotes the probability that a document will be accessed within time \( t \) from now. Since \( f(t) \) is a probability density, \( F(\infty) = 1 \), meaning that the document will certainly be accessed sometime in the future. If \( f(t) \) is represented as a continuous distribution, the instantaneous probability when \( \delta, T \to 0 \) is zero, which makes short term or immediate prediction difficult. To find the discrete density \( f(t) \) from the access logs, we calculate the proportion of document accesses that occur \( t \) time units after the preceding access for \( t \) ranging from zero to infinity. This approach assumes that all documents have identical interaccess time distributions, that is, all accesses are treated the same, irrespective of the documents they involve and that the distributions are free from periodic changes in access patterns (such as weekends when interaccess times are longer.) The implication of the first assumption is that the prediction is not conditioned on frequency of past accesses since all documents in the observed repository are assumed equally likely to be accessed giving rise to identical frequency distributions.

Since frequency distributions are more likely to vary between documents than not, it is clear that the above assumptions make this analysis suitable only on a per-document basis. However, the approach still holds, notwithstanding that the distribution \( F(t) \) is now specific to a particular document. To predict the probability of access within time \( T \) from now, for a particular document \( A \), we may use \( A \)’s distribution function \( F_A(t) \) to obtain \( F_A(\delta + T) \) where \( \delta \) is the age at the current time. If the interaccess time distributions are similar but not identical, we could condition these distributions on the parameters and find distributions of these parameters across the documents.
Our use of a single predictor, the interaccess time, obtained from the age $\delta$ and prediction interval $T$ does not imply that the technique is univariate. The use of multiple predictors, such as the frequency of accesses in a given previous interval can easily be accommodated into a multidimensional plot of access probability. The method becomes complicated when several dimensions are involved. To alleviate this, we may derive a combined metric from the predictor variables, transforming the problem back to univariate prediction. However, this requires empirical determination of correlation between predictors and subsequently a combination function.

Given the statistical principle, one might naturally be led to ask how the distribution $F(t)$ (or its variant $F_A(t)$) can be used for actionable prediction. Recall that $F(t)$ is a cumulative probability distribution. For a given document age, it tells us the probability that a document is accessed within a certain interval of time. If a single probability distribution is used, this probability is an indicator of overall document usage with respect to time interval. If we use individual distributions $F_A(t)$, it can be used to compare the relative usage of documents. The expected time to next access, $\bar{T}$ is given by the mean of the distribution: $E[\bar{T}] = \sum_{t=0}^{\infty} t \cdot f(t)$. The expected time $\bar{T}$ before the next access to a document, if it is known for all documents, can be used as a criteria for populating server side caches.

The temporal approach discussed above bases prediction on interaccess times. Equally, we may use a frequency based alternative for predicting access. A frequency distribution denotes the probability of a certain number of accesses to a document or a sample of documents over a fixed time interval. Using an analogous method to that discussed earlier, we can answer the following for prediction over the next time interval: (1) What is the probability that exactly $N$ documents will be accessed? (2) What is the probability that $N$ or more documents will be accessed? (3) How many documents are expected to be accessed?

This approach has the same drawbacks as discussed previously—it does not account for periodic changes in access rates, rather it aggregates them into a single distribution and accesses to all documents are treated the same. Finally, both temporal and frequency prediction may be combined to ascertain probabilities of a certain number of accesses during a given time period in the future.

3 Burrell’s Stochastic Model

The temporal and frequency-based approaches for statistical access prediction aggregate access statistics—interaccess times and frequency of access—into a single distribution. By treating all document accesses identically, they do not distinguish documents that are more frequently accessed than others. Our experiments on server access logs collected over a six month period show that Web documents have highly varying average frequency of access ranging from several hundred to just a few. This observation confirms the intuition that aggregated

---

1 Center for Advanced Information Systems Web server, April to October 2000.
statistical distributions are not ideal for predicting document access in the Web environment. Adopting interaccess or frequency distributions for individual documents alleviates this problem, but makes the solution impractical due to the overhead in computing and storing distributions for all documents on the server.

To address the problem of differential access rates, we investigate a stochastic model for library loans originally introduced by Burrell [1,2,3]. Burrell’s model accounts for the fact that some items are borrowed more often than others and explains the observed geometric distribution of circulation frequencies. We present the model in a generic context below.

Let us consider a static (not growing) collection of items over a fixed period of time, long enough to even out cyclic variations in accesses. The distribution of accesses, known as the frequency of circulation distribution (or foc distribution) is observed to be geometric. That is, if the random variable $X$ denotes the number of accesses to a randomly chosen item, then the probability of $x$ accesses is given by

$$\Pr\{X = x\} = (1 - p)^{x-1}p$$  \hspace{1cm} (1)

for $x = 1, 2, \ldots$ and where $0 < p < 1$. Acknowledging the differences in access patterns of individual items, we term the average number of times an item is accessed in a unit time as its desirability $\lambda$. Burrell makes the following assumptions regarding foc and desirability distributions:

- Given an item with desirability $\lambda$, the number of times it is accessed is a Poisson process of rate $\lambda$. Thus if we denote with $\Lambda$ the desirability of a randomly chosen item and by $X$ the number of accesses to it in a unit period of time, the probability conditioned on $\Lambda$ that this item is accessed $x$ times in a unit time period is given by

$$\Pr\{X = x|\Lambda = \lambda\} = \frac{e^{-\lambda} \lambda^x}{x!}$$  \hspace{1cm} (2)

for $x = 0, 1, 2, \ldots$

- Accordingly, in order for the overall distribution of $X$ to be geometric, the distribution of desirabilities must be negative exponential. That is,

$$\Pr\{\Lambda \leq \lambda\} = 1 - e^{-\alpha \lambda}$$  \hspace{1cm} (3)

for $\lambda \geq 0$.

The desirability distribution of Equation 3 is a cumulative distribution. The continuous exponential density function obtained by differentiation is $\alpha e^{-\alpha \lambda}$. The overall foc is then defined in terms of the two densities using the law of total probability as follows:

$$\Pr\{X = x\} = \int_0^\infty \frac{e^{-\lambda} \lambda^x}{x!} \lambda e^{-\alpha \lambda} d\lambda = \left( \frac{1}{1 + \alpha} \right)^{x-1} \left( 1 - \frac{1}{1 + \alpha} \right)$$

2 The overall distribution is actually derived by convolving the generating functions of the desirability and conditional foc distributions.
Comparing with Equation 1, we see that the above is equivalent to a geometric distribution as observed, with \( p = \alpha/(1+\alpha) \). Burrell’s model has been generalized in [3] to account for the negative binomial distribution of loan frequencies by modeling the desirability as a gamma distribution. This model, known as the gamma mixture of Poisson processes has been widely applied in other contexts such as analyzing accident data, industrial sampling, and absenteeism data.

Burrell’s model also accounts for items that have been implicitly out of circulation, (equivalent to “dead” Web pages.) Estimation of usage such as the expected number of items loaned out \( x \) times during a period and the probability that an item will not circulate during the period are discussed further in [4]. Finally, Burrell incorporates aging whereby the desirability of an item decays exponentially with time according to the relationship: \( \lambda(t) = \lambda e^{-\beta t}; \ t > 0 \) and \( \beta > 0 \). The resulting distribution of loan frequencies is found to be negative binomial.

A limitation of Burrell’s model is that it does not accommodate cyclic variations in access rate. Desirabilities computed over too short an interval may not be reflective of the longer term access patterns. Consequently, it can only be used to explain access patterns for durations long enough to iron out such variations.

3.1 Verification of Burrell’s Model

We noted earlier that Web page access rates show high variance. Could this suggest a geometric distribution for frequency of access or FOA as addressed by Burrell’s model? Figure 1 shows a histogram of access frequencies for a collection of over 900 pages observed over one week. The distribution is approximately geometric, confirming an earlier observation by Pitkow and Pirolli [6].
We briefly discuss some of the issues related to analysis of server log data. Server logs record each HTTP request explicitly received by the server. For the purpose of studying document access frequencies, only two fields are of relevance, namely the URL of the page requested and the time of the request. Server logs do not record repeat accesses to pages because these are served instead by browser caches residing on the client machine. If the time dimension in this analysis is made less granular, the effect of this missing data is mitigated due to the periodic cleaning of client-side caches. Personalized access prediction which requires client identification must in addition devise ways of uniquely identifying users behind proxies and firewalls.

For the purpose of our experiments, we used server logs containing over 50,300 records of accesses over six months from April to October 2000. We considered only requests to HTML documents as valid accesses, discarding image and script accesses since these are either not consciously requested by the user or present a different URL to the server each time. Requests originating from the same domain as the server were also removed since these are more likely to represent internal development activity than authentic user visits. After the cleaning phase, the combined log contained nearly 38,200 entries. An important parameter in log analysis is the size of time window over which statistics are aggregated. Each window acts as one time unit in modeling. To even out the influence of cyclical variations such as those over weekends in access rates, we chose weekly time windows. The final preprocessing step is to organize the log as a matrix, \( \text{Log}_{ij} = \text{number of accesses to page } j \text{ in week } i \).

From Figure 1, it is clear that the FOA distribution is similar to the FOC distribution of Burrell’s model for library loans. However, the application of this model to FOA is not justified until its assumptions outlined by Equations 2 and 3 are deemed reasonable for Web page accesses. We conducted further experiments on the processed Web logs to plot the distribution of desirabilities or mean number of accesses to documents per unit time period, in this case one week. For a page \( j \), the desirability is given by \( \sum_{i}^{[i]} \text{Log}_{ij}/[i] \). Figure 2 shows the distribution of desirabilities computed over the six month period. It resembles the highly concentrated negative exponential distribution of Equation 3.

Given the similarity of the FOA and desirability distributions to their equivalents in Burrell’s model, it is reasonable to expect the access frequency conditioned on desirability to be a Poisson process. Before proceeding to measure this distribution we examine two important properties of Poisson processes in modeling Web page accesses. Let us model the sequence of accesses to a Web page repository as a counting process, i.e., a set of random variables \( \{N(t), t > 0\} \) where \( N(t) \) counts the number of accesses that have occurred at or up to \( t \).

1. **Stationarity:** In a stationary counting process, the probability of a certain number of events is the same for any two intervals of equal length. Our choice of one week as the unit time period satisfies this condition, since it is safe to assume that there is little variation in access trends in between weeks (though accesses certainly vary within a week.) Web page accesses can therefore be con-

---

3 This problem is detailed in [5] along with possible solutions.
sidered a stationary process. (2) **Independent increments**: A counting process is said to have independent increments if the occurrence of a certain number of events in an interval is independent of the number of events in previous intervals. Because previous accesses to Web pages do not in any way affect current or future accesses, it is reasonable to classify the process as having independent increments.

The Poisson process allows us to calculate the probability of \( x \) accesses in any period of length \( t \) (i.e., \( t \) weeks)

\[
\Pr\{X(t) = x\} = \frac{(\lambda t)^x e^{-\lambda t}}{x!}
\]

(4)

where \( \lambda t \) is the expected number of accesses in \( t \) weeks. This implies that \( E[X(t)] \) need not be determined separately. The only unknown parameter is average accesses in one time unit or \( \lambda \). This result is used for access prediction in the next section.

Measuring the conditional distribution of access frequency poses a challenge. Although the access frequency is a discrete random variable (denoted by \( X \) in the previous section), the desirability is necessarily continuous. To plot the distribution \( \Pr\{X = x|\Lambda = \lambda\} \), desirability measurements must be discretized. Figure 3 shows the corresponding histogram for the conditional distribution \( \Pr\{X = x|\Lambda = 2\} \) with desirability discretized in the range \([1.5, 2.5]\) (having a mean of \( \lambda = 2 \)). The conditional distribution compares well with the Poisson function having an average arrival rate of 2. The empirical verification of Figure 3 becomes the basis for applying Burrell’s model to Web document accesses and predicting access.
The observation that the accesses to a set of documents having the same desirability are Poisson distributed has an interesting consequence. The Poisson process has the property that the time interval between successive events has an exponential density. We noted in the previous section that statistical prediction using individual interaccess time distributions for documents is impractical. However, having shown that Web page access for a group of documents having similar desirability is a Poisson process, we could categorize documents according to their desirabilities and maintain a single distribution per category. Knowing that the density is exponential, we may fit an analytical form to the observed distributions to further reduce the amount of data required for prediction.

### 3.2 Poisson Prediction of Web Page Accesses

If the conditional distribution $f_{X|\Lambda}(x|\lambda)$ is known to be a Poisson process as in Equation 4, we can use it to predict accesses to a document from the observed mean accesses. That is, given that a document has been accessed on an average $\lambda$ times over a unit time interval the probability of $x$ accesses to it over $t$ units of time can be found as $f_X(x,t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$.

This method only requires knowledge of a document’s desirability to make access predictions anytime into the future, say $t$ units later. The simplest predicted quantity—expected number of accesses over time $t$ is given by $E[f_X(x,t)] = \sum_{x=0}^{\infty} x \cdot f_X(x,t) = \lambda t$ by the stationarity property. Recall that the desirability $\lambda$ is simply the mean number of accesses in a unit time period or $E[f_X(x,1)]$. This quantity can be estimated from a sample of past accesses to a document. In our experiments we found that computing $\lambda$ over eight weeks gives a good
estimate of a document’s desirability. Another predictive measure, the probability of one or more access within time $t$ or the cumulative distribution can be determined as follows: $\Pr\{X > 0\} = \sum_{x=1}^{\infty} f_X(x, t) = 1 - \Pr\{X = 0\} = 1 - e^{-\lambda t}$ using Equation 4.

The results of the above prediction for $t = 1$ for desirability computed over eight weeks are compared with the actual access probabilities determined a-posteriori from the processed access log, Log in Figure 4.

Modeling Web page accesses as a Poisson process helps us predict future accesses to particular items given their desirabilities calculated from past usage. Burrell’s model can also aid in making access predictions about the entire repository. If we denote by $P_r(T)$ the proportion of pages accessed $r$ times during a period $T$ and by $1 - \beta$ the proportion of dead pages, then $P_0(T) = (1 - \beta) + \beta p$ and $P_r(T) = \beta p (1 - p)^r$.

The values $P_r$ are indicators of usage in the future and can be determined using estimates of $\beta$ and $p$. Suppose $f_r(T)$ is the number of pages that have been accessed $r$ times during the period $T$. Then the maximum likelihood estimators of $\beta$ and $p$ are given by $\hat{p} = \frac{N - f_0(T)}{\sum f_r(T)}$ and $\hat{\beta} = \frac{N - f_0(T)}{N(1 - \hat{p})}$.

### 4 Related Work

In this section, we discuss some of the recent research in modeling and predicting Web page accesses. Sarukkai [8] has applied a variation of Markov chains to predict user accesses based on the sequence of previously followed links. Consider a stochastic matrix $P$ whose elements represent page transition probabilities and
a sequence of vectors, one for each step in the link history of a user, denoted \( I^1, I^2, \ldots, I^{t-1} \). The \( \ell^{th} \) element in vector \( I^k \) is set to 1 if the user visits page \( \ell \) at time \( k \), otherwise it is set to 0. For appropriate values of constants \( a_1, a_2, \ldots, a_k \), the state probability vector \( S^t \) for predicting the next link is determined as follows: \( S^t_j = \sum_{k=1}^{n} a_k I^{t-k} P^k \). The next page to be accessed is predicted as the one with the highest state probability in the vector \( S^t \). The same approach can be used to generate tours by successively predicting links of a path. Recker and Pitkow \[7\] have used the human memory model to predict document accesses in a multimedia repository based on the frequency and recency of past document accesses.

5 Conclusions

In this paper we considered the application of patterns in browsing behavior of users for predicting access to Web documents. We proposed two models for addressing our specification of the access prediction problem. The first lays out a preliminary statistical approach using observed distributions of interaccess times of individual documents in the collection. To overcome its deficiencies, we adapted a stochastic model for library circulations that accounts for differences in mean access rates of Web documents. We verified the assumptions of this model with experiments performed on a server log of accesses recorded over a six month period. Our results show that the model is reasonably accurate in predicting Web page access probabilities based on the history of accesses.

References