PREFETCHING NEXT PAGE ACCESS BY ZIPF ESTIMATOR USING DATA MINING TECHNIQUES

Deepika*

Sachin Kathuria**

ABSTRACT

World Wide Web can be considered as a large distributed information system that provides access to shared data objects. The World Wide Web is of an exponential growth in size, which results in network congestion and server overloading. as more and more information services move onto web. The result of all this is increased access latency for the users. The success of WWW depends on the response time. Due to the fast development of internet services and a huge amount of network traffic, it is becoming an essential issue to reduce World Wide Web user-perceived latency. Web caching has been recognized as one of the effective schemes to alleviate the service bottleneck and reduce the network traffic, thereby minimize the user access latency. Although web performance is improved by caching, the benefit of caches is limited. World Wide Web is an important area for data mining research due to the huge amount of information. The success of WWW depends on the response time. Due to the fast development of internet services and a huge amount of network traffic, it is becoming an essential issue to reduce World Wide Web user-perceived latency. To further reduce the retrieval latency, web prefetching becomes an attractive solution to this problem. Performance measurement of prefetching techniques is primarily in terms of hit ratio and bandwidth usage. A significant factor for a prefetching algorithm in its ability to reduce latency is deciding which objects to prefetch in advance. Zipf's law governs many features of the Internet. Observations of Zipf distributions, while interesting in and of themselves, have strong implications for the design and function of the Internet. This dissertation implements a Zipf law based novel approach for the determination of next page likely to be accessed by specific client.

Keywords: *Data Mining Techniques, Rank Analysis, Web Mining, Zipf's Estimator.*

*Assistant Professor, Department of Computer Engineering, YMCA University of Science and Technology, Faridabad, India

**Research scholar, Department of Computer Engineering, YMCA University of Science and Technology, Faridabad, India

INTRODUCTION

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. One of the main directions of research in the Web is to reduce the time latencies users experience when navigating through Websites. Caching is a technique that is already being used in the Web domain. Caches reduce latencies, because they allow fast retrieval of potentially frequently access documents However, recent studies indicate that the benefits from this technique are rather limited. The vast numbers of documents available in the Web, the quick rate of their change and the diverse needs across users and overtime, are the main factors that cause Web caching to perform poorly. Thus, another method that was first applied in operating systems, prefetching, is now being studied in the Web context. When prefetching is employed, Web pages that the user is likely to access in the near future are transferred to her cache without being requested. If the user does request one of the prefetched pages, it will already be in the local cache, thus reducing the latency to minimum. Prefetching is complementary to the caching mechanism. It may take place while the I/O system is idle, and if the predictions are accurate enough the performance of the cache can be significantly improved.

The organization of the paper is as follows: In section II, challenges in existing approaches are discussed. In section III, we discuss various proposals given by the authors, followed by section IV, it presents the idea for next page access by Zipf estimator. In section V, we draw a conclusion and address the future work.

CHALLENGES IN EXISTING APPROACHES

Because of the expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently The following list of challenges shows the inefficiencies and limitations that have to be overcome in WWW for easily accessing the page.

- The response time perceived by the user is too long.
- The explosive growth of the Web has imposed a heavy demand on networking resources and Web servers.

- Hence, an obvious solution in order to improve the quality of Web services would be the increase of bandwidth, but such a choice involves increasing economic cost.
- Web caching scheme has three significant drawbacks: If the proxy is not properly updated, a user might receive stale data, and, as the number of users grows, origin servers typically become bottlenecks.
- The several factors diminish the ideal effectiveness of Web caching. The obvious factors are the limited system resources of cache servers (i.e., memory space, disk storage, I/O bandwidth, processing power, and networking resources). However, even if the cache space is unlimited, there are significant problems that cannot be avoided by such an approach. Specifically, large caches are not a solution because, the problem of updating such a huge collection of Web objects is unmanageable.
- Main drawback of systems which have enhanced prefetching policies is that some prefetched objects may not be eventually requested by the users. . In such a case, the prefetching scheme increases the network traffic as well as the Web servers' load.

RELATED WORK

The problem of efficiently accessing the page from WWW has received considerable attention by researchers. In this section, some of these contributions are presented.

It was Etzioni who first coined the term web mining .Etzioni starts by making a hypothesis that the information on the web is sufficiently structured and outlines the subtask of web mining. His paper describes the web mining process.

This study provides an overview of research work related to web usage mining. Access histories of users visiting a web server are automatically recorded in client and server side web access log using any of the log formats. Extended common log format is one of the universal server log format. Each entry represents a single request for resource and contains following details : <IP address><Remote log name><Authenticated user ID><Date and time of request><URL request><Status code><Content length of response><Referrer><Agent>.

Cooley et al; Srivastava et al. [13] define Web usage mining as a three-phase process, consisting of preprocessing, pattern discovery, and pattern analysis. Their prototype system, Web SIFT, first performs intelligent cleansing and preprocessing for identifying users, server sessions, and inferring cached page references through the use of the referrer field, and also performs content and structure preprocessing [14]. Pattern discovery is accomplished through the use of general statistic algorithms and data mining techniques such as association rules, sequential pattern analysis, clustering, and classification. The results are then analysed

through a simple knowledge query mechanism, a visualization tool, or the information filter, that makes use of the preprocessed content and structure information to automatically filter the results of the Knowledge discovery algorithm.

Padmanabhan [2] use dependency graph for prediction and prefetching. Their prediction algorithm construct a dependency graph that depicts the pattern of accesses to different file stored at the server. The graph has a node for every file that has ever been accessed. There is an arc from node A to B if and only if B was accessed with in w(look ahead window size) access after A.

Prefetching and caching are effective techniques for improving the performance of file systems. Pei Cao [6] show that the two integrated prefetching and caching strategies are indeed close to optimal and that these strategies can reduce the running time of applications by up to 50%.

Prefetching and caching are techniques commonly used in I/O systems to reduce latency. Many researchers have advocated the use of caching and prefetching to reduce latency in the Web. T. M. Kroeger [8] found that for these traces, local proxy caching could reduce latency by at best 26%, prefetching could reduce latency by at best 57%, and a combined caching and prefetching proxy could provide at best a 60% latency reduction.

M. Eirinaki [9] use Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between groups of users with specific interest. This information can be used for example for restructuring Web sites by adding links between those pages which are visited together. Using this knowledge the trends of the activity of the users can be determined and predictions to the next visited pages can be calculated.

S. Jespersen [4] said that Markov assumptions are used as the basis to mine the structure of browsing patterns. Markov-based structures for web usage mining are best suited for tasks demanding less accuracy such as pre-fetching, personalization, and targeted ads. Many of the papers proposed using association rules or Markov models for next page prediction. Faten Khalil [5] proposes an improved approach, based on a combination of Markov models and association rules that result in better prediction accuracy and more coverage. They used low order Markov models to predict multiple pages to be visited by a user and then applied association rules to predict the next page to be accessed by the user based on long history data.

Zipf's law, an empirical law formulated using mathematical statistics, refers to the fact that many types of data studied in the physical and social sciences can be approximated with a

Zipfian distribution, one of a family of related discrete power law probability distributions. The law is named after the linguist George Kingsley Zipf (pronounced /zɪ f/) who first proposed it (Zipf 1935, 1949).

Several researchers have observed that the relative frequency with which web pages are requested follows Zipf's law [12]. Zipf's law states that the relative probability of a request for the i'th most popular page is proportional to 1/i. Glassman [7] was perhaps the first to use Zipf's law to model the distribution of webpage requests, and several other authors have also applied Zipf's law to the distribution of web requests [3], [1]. However, several recent studies have investigated whether the requests do indeed follow Zipf's law and concluded otherwise.

Lee Breslau [13] showed the evidence that web request follow a Zipf-like distribution. He first investigates the page request distribution seen by web proxy cache using traces from a variety of sources. Furthermore he finds that the there is only a weak correlation between the access frequency of a web page and its size and a weak correlation between the access frequency of the web page and its rate of change. He then produces a model where the web access is independent and the reference probability of the document follows a zipf like distribution.

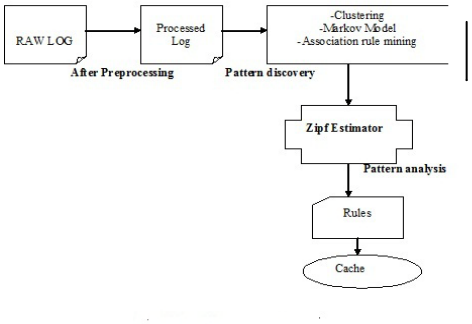
Lei Shi [8] presents the Web object popularity based model on zipf –like law, introduces the stability concept of the web system, and calculate the upper bound for the minimum length of the request stream in order to get stability. Zipf's law holds the promise of more effective design and use of web cache resources.

Huberman [14] used a spreading activation model to address another universal finding in studies of WWW activity that of Zipf's like distribution in the number of hit per page. He ran spreading activation simulations on random graphs of 100 nodes each, with an average of five links per node, using various initial conditions. The resulting probability distribution of the number of hits received over the collection of page followed a Zipf' law.

DETERMINING NEXT PAGE ACCESS BY ZIPF ESTIMATOR

A Model consisting of various modules such as Data Mining techniques, Zipf Estimator used for determining next page is shown in Figure 1.

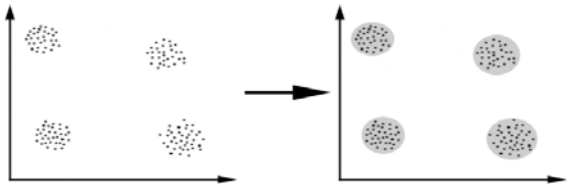
Figure 1: Model For Determining Next Page



For finding the page to be prefetched first the proxy server raw log is processed using the data preprocessing technique of web usage mining. On the preprocessed log various web mining techniques such as rough set theory for finding clusters, markov and association rule mining is applied on the clustered set of the pages. Applying all these techniques the frequently visited pattern are determined. On these patterns the zipf estimator is applied to determine the probability of next page access and that pages are prefetched in the cache.

Clustering

The clustering could be the process of organizing objects into groups whose members are similar in some way. A Cluster is therefore a collection of objects "similar "between them and are "dissimilar "to the object belonging to the other clusters.



A rough set first described by Zdislaw I.Palwak, is a formal approximation of a crisp set in term of a pair of sets which give lower and upper approximation of the original set. On the cleaner version of the web log the rough set theory can be applied to form the clusters.

Let W= (U, A), be an information system, where U is a non empty set of finite objects (the universe), and A is non-empty, finite set of attributes such that a: $U \rightarrow V_a$ for every $a \in A$. V_a is the set of values that attribute a may take [5]. The next section gives the Rough Set Nomenclature.

Equivalence Set: Equivalence set is a set of attributes whose elements are related to each other by an equivalence relation. Two objects are in the same class if they have the same value for attribute in S. x, y $\in$ same equivalence class if they have same values for every

attributes in S.

Lower Approximation and Positive Region: The P-lower approximation or positive region is the union of all equivalence classes in $[x]_p$ which are contained by the target set.

$$\underline{P}X = \{x \mid [x]_p \subseteq X\}$$

Upper Approximation and Negative Region: The P-upper approximation is the union of all equivalence class in $[x]_p$ which has nonempty intersection with the target set .

$$\overline{P}X = \{x \mid [x]_p \cap X \neq \phi\}$$

Threshold: Threshold is defined here as L/S where L is the total number of Links and S is the total number of sessions. A domain expert can choose the threshold based on his experience.

Target Set: Target set is a set that consists of number of sessions whose value is equal to the threshold value.

K-ORDER MARKOV MODEL AND ASSOCIATION RULE

After dividing user sessions into a number of clusters, Markov model analyses are carried out on each of the clusters. Markov models are used in the identification of the next page to be accessed by the Web site user based on the sequence of previously accessed pages. Let $P = \{p1, p2,\ldots\ldots\ldots, pm\}$ be a set of pages in a Web site. Let W be a user session including a sequence of pages visited by the user in a visit. Assuming that the user has visited l pages, then prob (p_i |W) is the probability that the user visits pages p_i next. Page P_{l+1} the user will visit next is estimated by:

$$P_{l+1} = \text{argmax}_{p \in IP}\{P(P_{l+1}=p|W)\}$$

$$P_{l+1} = \text{argmax}_{p \in IP}\{P(P_{l+1}=p|p_l,p_{l-1},\ldots p_1)\} \quad (1)$$

This probability, prob (p_i |W) is estimated by using all sequences of all users in history (or training data), denoted by W. Naturally, the longer l and the larger W, the more accurate prob (p_i |W). However, it is infeasible to have very long l and large W and it leads to unnecessary complexity. Therefore, a more feasible probability is estimated by assuming that the sequence of the Web pages visited by users follows a Markov process that imposes a limit on the number of previously accessed pages k. In other words, the probability of visiting a page pi does not depend on all the pages in the Web session, but only on a small set of k preceding pages, where $k << l$.

The equation become

$$= \text{argmax}_{p \in IP}\{P(P_{l+1}=p|p_l,p_{l-1},\ldots p_{l-(k-1)})\} \quad (2)$$

where *k* denotes the number of the preceding pages and it identifies the order of the Markov model. The resulting model of this equation is called the all *kth* order Markov model. Of course, the Markov model starts calculating the highest probability of the last
page visited because during a Web session, the user can only link the page he is currently visiting to the next one.

Let S^k_j be a state containing *k* pages, $S^k_j = (p_{l-(k-1)}, p_{l-(k-2}, \ldots p_l)$ the probability of $P(p_i \mid S^k_j)$ is estimated as follows from a history (training) data set.

$$P(p_i \mid S^k_j) = \frac{\text{Frequency}(<S^k_j, p_i>)}{\text{Frequency}(S^k_j)} \qquad (3)$$

The fundamental assumption of predictions based on Markov models is that the next state is dependent on the previous *k* states. The longer the *k* is, the more accurate the predictions are.

ASSOCIATION RULE

Let I= {I1, I2, … , Im} be a set of m distinct attributes, T be transaction that contains a set of items such that T ⊆ I, D be a database with different transaction records Ts. An association rule is an implication in the form of X=>Y, where X, Y ☐ I are sets of items called itemsets, and X ∩ Y =∅. X is called antecedent while Y is called consequent, the rule means X implies Y.

There are two important basic measures for association rules, support(s) and confidence (c). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively. Support(s) of an association rule is defined as the percentage/fraction of records that contain X ☐Y to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item.

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X ☐Y to the total number of records that contain X.

Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X=>Y is 80%, it means that 80% of the transactions that contain X also contains Y together.

RANK ANALYSIS AND PROBABILITY CALCULATOR

Rank is analysed on the basis of frequency count. Rank will be less, if frequency is more. Probability Calculator calculates the probability for accessing the next page. Assuming that there exists i number of moves (m_i) in the transaction T where m_i is the subset of T, then the probability of $(i+1)^{th}$ move i.e. the next page to be accessed can be estimated by using

$$P(m_1 \cup m_2 \cup \ldots \cup m_{i-1} \cup m_i \cup m_{i+1}) \quad (1)$$

where P is the probability of union of all the moves, $m_1, m_2, \ldots, m_{i-1}, m_i, m_{i+1}$. According to information theory

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *and*

$P(A \cap B) = P(A).\ P(B/A)$

$\therefore$ $P(A \cup B) = P(A) + P(B) - P(A).\ P(B/A)$

$\therefore$ Using above, Eq. (1) can be rewritten as:

$\therefore$ $P((m1 \cup m_2 \cup \ldots \cup m_i) \cup m_{i+1})$ (2)

$\therefore$ $= P(m_1 \cup m_2 \cup \ldots \cup m_{i-1} \cup m_i) + P(m_{i+1}) - P((m_1 \cup m_2 \cup \ldots \cup m_i) \cap m_{i+1})$

$\therefore$ $= P(m_1 \cup m_2 \cup \ldots \cup m_i) + P(m_{i+1}) - P(m_1 \cup m_2 \cup \ldots \cup m_i)\ P(m_{i+1}/m_1 \cup m2 \cup \ldots \cup m_i)$ **(3)**

In (3), $P(m_1 \cup m_2 \cup \ldots m_i)$ can be measured precisely by analyzing previously downloaded pages in the earlier moves while $P(m_{i+1})$ which is the probability of the next move is to be estimated which is done by *Zipf estimator*.

Zipf ESTIMATOR

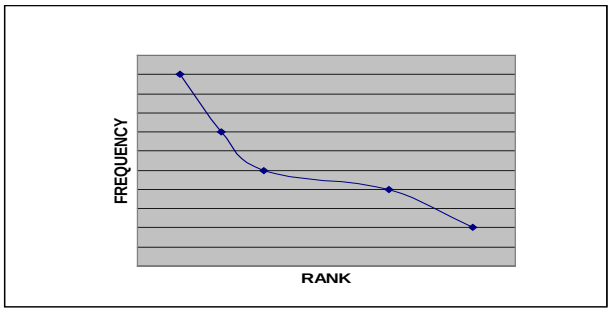
Zipf estimator is based on Zipf law. Zipf's Law is one of the famous linguistic laws, which describes the statistical distribution of words with different ranks by means of frequency. Zipf's Law uncovers the relationship between word frequency and its position in the word list.

The law discovered empirically by Zipf (1949) for word tokens in a corpus states that if *f* is the frequency of a word in the corpus and *r* is the rank, then:

$$f=k/r \qquad (1)$$

where *k* is a constant for the corpus.

Figure 2: Zipf curve



Zipf Law states that frequency of terms in a set of text collection follows a power law distribution. This fact was derived from an experiment, which was originally done on a corpus of natural language. In this, frequency count of the frequently occurring words like *the, and, of,* etc was calculated. On the basis of calculated frequency, the rank of these words was established. From the results it was drawn that the frequency is inversely proportional to the rank i.e. the higher the frequency, the lower is the rank as shown in Figure.

In this work, Zipf's distribution is being used to determine rank of a new page to be accessed by the user. For estimation, each next move m_{i+1} is ranked on the basis of its $P(m_{i+1} | m_1 U m_2 U m_3 U \ldots U m_i)$ value. It is observed that rank and frequency are inversely proportional to each other.

. . [1]

$$f \propto \frac{1}{r} \qquad (4)$$

Mandelbrot [55] generalized Zipf's Law as:

$$P(m_{i+1}) = \alpha[R(m_{i+1}) + \beta]^{-\gamma}$$

. ..(5)

Where P is the probability, $R(m_{i+1})$ is the rank of page m_{i+1} and rank of page are the constants. Then using the known $P(m_i)$ values from previous moves, we can compute parameters $\alpha\beta\gamma$. With the given frequency count and their ranks, the curve can be best fitted for the equation (5). The following values for the parameters yield the best fitting curve: α=0.08, β=0.25, γ=1.15. Once these constants are obtained, the probabilities of any move $P(m_{i+1})$ can be estimated using (5).

For every next move $m_{i+1,}$ transaction set T_i is checked. If m_{i+1} is found in T_i then frequency count and rank is evaluated. Afterwards move m_{i+1}, is placed at the appropriate position in the statistic table S, as per the evaluated rank. On the basis of the known rank, unknown P (m_{i+1}) can be evaluated using (5).

CONCLUSION

This paper views the web as a system for the prediction and prefetching the useful information which is helpful for the user. It emphasis on the zipf estimator (a probability calculator), for estimating the probability of accessing the next page. Depending upon the probability of the next page, the page can be prefetched locally on the proxy server. When the user request for that page the page is given directly to the user rather than going to the web server.

REFERENCES

[1] Virgilio Almeida, Azer Bestavros, Mark Crovella, and Adriana de Oliveira, " Characterizing reference locality in the WWW " , In IEEE International Conference in Parallel and Distributed Information Systems, Miami Beach, Florida, USA, December 1996. http://www.cs.bu.edu/groups/oceans/papers/ Home.html

[2] Padmanbhan, "Using Predictive prefetching to Improve World wide web Latency", V.N, 1996 Comput. Comm. Rev, 26(3):22-36

[3] Carlos Cunha, Azer Bestavros, and Mark Crovella, "Characteristics of WWW client-based traces.", Technical Report TR-95-010, Boston University, Computer Science Dept., Boston, MA 02215, USA, April 1995.

[4] S. Jespersen, T. B. Pedersen, and J. Thorhauge, "Evaluating the markov assumption for web usage mining," in WIDM '03: Proceedings of the 5th ACM international workshop on Web information and data management. New York, NY, USA: ACM Press, 2003, pp. 82-89

[5] Faten Khalil, Jiuyong Li and Hua Wang "A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses" ,Proc. Fifth Australasian Data Mining Conference (AusDM2006), CRPIT Volume 61,177-184

[6] Pei Cao,Edward W. Felten,Anna R. Karlin, Kai Li, **"**A study of integrated prefetching and caching strategies" ,Proceedings of the 1995 ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems. **Pages: 188 - 197, 1995**

[7] Steven Glassman, " A caching relay for the World Wide Web", In First International Conference on the World Wide Web, CERN, Geneva, Switzerland, May 1994.

[8] Lei Shi,Zhimin Gu,Lin Wei,and Yun Shi, "An applicative study of zipf's law on web cache" ,In International Jounal of information Technology,Vol. 12 No.4 2006

[9] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," *ACM* Trans. *Inter. Tech.*, Vol. 3, No. 1, pp. 1-27, 2003

[10] T. M. Kroeger, D. D. Long, and J. C. Mogul, "Exploring the bounds of web latency reduction from caching and prefetching," in Proc. of the 1st USENIX Symposium on Internet Technologies and Systems, Monterey, USA, 1997.

[11] Jyoti Pandey ,Amit Goel, Dr. A K Sharma, A Framework for Predictive Web Prefetching at the Proxy Level using Data Mining, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.6, June 2008

[12]George Kingsley Zipf.,"Relative frequency as a determinant of phonetic change", Reprinted from the Harvard Studies in Classical Philiology, Volume XL, 1929.

[13] Lee Breslau, Pei Cao, Li Fan, Graham Phillips, Scott Shenker. *Web caching and Zipf-like Distributions*: *Evidence and Implications. In* IEEE INFOCOM, VOL. XX, NO. Y, MONTH 1999