Volume 3, No. 7, July 2012

Journal of Global Research in Computer Science

ISSN-2229-371X

RESEARCH ARTICAL

Available Online at www.jgrcs.info

NOVEL APPROCH FOR OFT BASED WEB DOMAIN PREDICTION

A. Niky Singhai [1], B. Prof Rajesh Kumar Nigam[2]

[1] M-Tech (Computer science and engineering) TIT Bhopal, Bhopal, (M.P.) India
nikysinghai@gmail.com

[2] Associate Professor, (Computer science and engineering) TIT Bhopal, Bhopal, (M.P.) India
rajesh_rewa@hotmail.com

Abstract— In this paper, we present a complete framework and predict the Web page usage patterns from Web log files of a real Web site that has all the challenging aspects of real-life Web usage predict, including evolving user profiles and external data describing ontology of the Web content. Our Studies have been conducted on pre-fetching models based on

Decision trees, Markov chains, and path analysis. However, the increased uses of dynamic pages, frequent changes in site structure and user access patterns have limited the efficacy of these static techniques. One of the techniques that are used for improving user latency is Caching and another is Web pre-fetching. Approaches that bank solely on caching offer limited performance improvement because it is difficult for caching to handle the large number of increasingly diverse files. For perform successful perfecting, we must be able to predict the next set of pages that will be accessed by users. The OFT Page Rank algorithm used by Google is able to compute the popularity of a set of Web pages based on their link structure. In this paper, a novel OFT Page Rank-like algorithm is proposed for conducting Web page prediction**.** As the tool for the algorithm implementations we chose the “language of choice in industrial world” – MATLAB.

Keywords- Decision trees, Markov chains, path analysis**,** Page Rank, pre-fetching, predict, Web page, Web content**.**

INTRODUCTION

The exponential proliferation of Web usage has dramatically increased the volume of Internet traffic and has caused serious performance degradation in terms of user latency and bandwidth on the Internet. The use of the World Wide Web has become indispensable in everybody’s life which has also made it critical to look for ways to accommodate increasing number of users while preventing excessive delays and congestion.

An improved understanding of the customer’s habits, needs, and interests can allow the business to profit by, for instance, “cross selling” or selling items related to the ones that the customer wants to purchase. Hence, reliable knowledge about the customers’ preferences and needs forms the basis for effective business. As businesses move online, the competition between businesses to keep the loyalty of their old customers and to lure new customers is even more important, since a competitor’s Web site may be only one click away. The fast pace and large amounts of data available in these online settings have recently made it imperative to use automated data mining or knowledge discovery techniques to discover Web user profiles. These different modes of usage or the so-called mass user profiles can be discovered using Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user click streams stored in Web log files. These profiles can later be harnessed toward personalizing the Web site to the user or to support targeted marketing.

Although there have been considerable advances in Web usage mining, there have been no detailed studies presenting a fully integrated approach to mine a real Web site with the challenging characteristics of today’s Web sites, such as evolving profiles, dynamic content, and the availability of taxonomy or databases in addition to Web logs.

Web usage patterns with real-world challenges such as evolving access patterns, dynamic pages, and external data describing an ontology of the Web content and how it relates to the business actors (in the case of the studied Web site, the companies, contractors, consultants, etc., in corrosion). The Web site in this study is a portal that provides access to news, events, resources, company information (such as companies or contractors supplying related products and services), and a library of technical and regulatory documentation related to corrosion and surface treatment.

The portal also offers a virtual meeting place between companies or organizations seeking information about other companies or organizations. Without loss of generality, in the rest of this paper, we will refer to all the Web site participants (organizations, contractors, consultants, agencies, corporations, centers, agencies, etc.) simply as companies. The Web site in our study is managed by a nonprofit organization that does not sell anything but only provides free information that is ideally complete, accurate, and up to date. Hence, it was crucial to understand the different modes of usage and to know what kind of information the visitors seek and read on the Web site and how this information evolves with time. For this reason, we perform clustering of the user sessions extracted from the Web logs to partition the users into several homogeneous groups with similar activities and then extract user profiles from each cluster as a set of relevant URLs.

This procedure is repeated in subsequent new periods of Web logging (such as biweekly), then the previously discovered user profiles are tracked, and their evolution pattern is categorized. When clustering the user sessions, we exploit the Web site hierarchy to give partial weights in the session similarity between URLs that are distinct and yet located closer together on this hierarchy. The Web site hierarchy is inferred both from the URL address and from a Web site database that organizes most of the dynamic URLs along an “is-a” ontology of items. We also enrich the cluster profiles with various facets, including search queries submitted just before landing on the Web site, and inquiring and inquired companies, in case users from (inquiring) companies inquire about any of the (inquired) companies listed on the Web site, which provide related services

LITERATURE SURVEY

In this research paper [1] initiatives have addressed the need for improved performance of Web page prediction accuracy that would profit many applications, e-business in particular. Different Web usage mining frameworks have been implemented for this purpose specifically Association rules, and Markov model. Each of these frameworks has its own strengths and weaknesses and it has been proved that using each of these frameworks individually does not provide a suitable solution that answers today's Web page prediction needs. Endeavors to provide an improved Web page prediction accuracy by using a novel approach that involves integrating clustering, association rules and Markov models according to some constraints. In That research paper [1] they try to improves the Web page access prediction accuracy by integrating all three prediction models Markov model, Clustering and association rules according to certain constraints. Integrates the three models using 2-Markov model computed on clusters achieved using k-means clustering algorithm and Cosine distance measures for states that belong to the majority class and performing association rules mining on the rest. The IPM model could be extended to a completely automated system. Currently, some human intervention is required especially during the features selection process.

In this research paper [1] all clustering experiments were developed using MATLAB statistics toolbox. Since k-means computes different centroids each run and this yields different clustering results each time, the best clustering solution with the least sum of distances is considered using MATLAB k-means clustering solutions. Therefore, using Cosine distance measure with the number of clusters leads to good clustering results while keeping the number of clusters to a minimum. Merging Web pages by web services according to functionality reduces the number of unique pages and, accordingly, the number of sessions. The categorized sessions were divided into 7 clusters using the k-means algorithm and according to the Cosine distance measure. Markov model implementation was carried out for the original data in each cluster. The clusters were divided into a training set and a test set each and 2-Markov model accuracy was calculated accordingly.

Then, using the test set, each transaction was considered as a new point and distance measures were calculated in order to define the cluster that the point belongs to. Next, 2-Markov model prediction accuracy was computed considering the transaction as a test set and only the cluster that the transaction belongs to as a training set. Since association rules techniques require the determination of a minimum support factor and a confidence factor, we used the experimental data to help determine such factors. We can only consider rules with certain support factor and above a certain confidence threshold.

In research [7] web prediction is a classification problem in which we attempt to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages. Predicting user’s behavior while serving the Internet can be applied effectively in various critical applications. Such application has traditional tradeoffs between modeling complexity and prediction accuracy. In this paper, we analyze and study Markov model and all-*K*th Markov model in Web prediction. They propose a new modified Markov model to alleviate the issue of scalability in the number of paths. In addition, they present a new two-tier prediction framework that creates an example classifier *EC*, based on the training examples and the generated classifiers. They show that such framework can improve the prediction time without compromising prediction accuracy. They have used standard benchmark data sets to analyze, compare, and demonstrate the effectiveness of our techniques using variations of Markov models and association rule mining. Our experiments show the effectiveness of our modified Markov model in reducing the number of paths without compromising accuracy. Additionally, the results support our analysis conclusions that accuracy improves with higher orders of all-*K*th model.

In this paper [7], they have reviewed the current state-of-the-art solutions for the WPP. They analyzed the all-*K*th Markov model and formulated its general accuracy and PR. Moreover, they proposed and presented the modified Markov model to reduce the complexity of original Markov model. The modified Markov model successfully reduces the size of the Markov model while achieving comparable prediction accuracy. Prediction process in the two-tier model is show on figure 1. Additionally, they proposed and presented a two-tier prediction framework in Web prediction. They showed that our two-tier framework contributed to preserving accuracy (although one classifier was consulted) and reducing prediction time.

They conducted extensive set of experiments using different prediction models, namely, Markov, ARM, all-*K*th Markov, all-*K*th ARM, and a combination of them. They performed our experiments using three different data sets, namely, NASA, UOFS, and UAEU, with various parameters such as rank, partition percentage, and the maximum number of *N*-grams.

Our comparative results show that large number of *N*-grams in the all-*K*th model does not always produce better prediction accuracy. The results also show that smaller *N*-gram models perform better than higher *N*-gram models in terms of accuracy. This is because of the small number of experiences / sessions obtained during data processing of large *N*-grams. They have also applied ranking to improve the prediction accuracy and to enhance its applicability. Our results show that increasing the rank improves prediction accuracy.

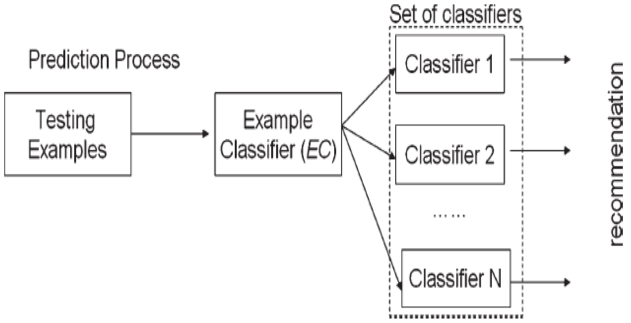
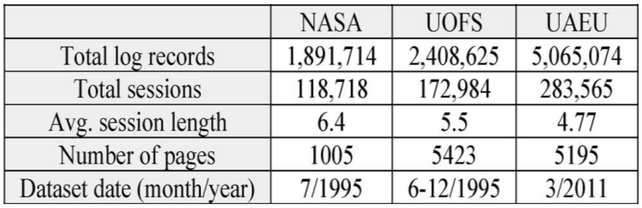


Figure 1. Prediction process in the two-tier model [7].

However, they have found that individual higher ranks have contributed less to the prediction accuracy. In addition, the results clearly show that better prediction is always achieved when combining all-*K*th ARM and all-*K*th Markov models. Finally, for the two-tier framework, our results show the efficacy of the *EC* to reduce the prediction time without compromising the prediction accuracy.



	NASA	UOFS	UAEU
Total log records	1,891,714	2,408,625	5,065,074
Total sessions	118,718	172,984	283,565
Avg. session length	6.4	5.5	4.77
Number of pages	1005	5423	5195
Dataset date (month/year)	7/1995	6-12/1995	3/2011

Figure 2. Summary of NASA and UOFS DATA Sets [7].

They considered three data sets, namely, the NASA data set, the University of Saskatchewan's (UOFS) data set, and the United Arab Emirates University (UAEU) data set [9]. figure 2 shows a brief statistics of each data set. In addition to many other items, the preprocessing of a data set includes the following: grouping of sessions, identifying the beginning and the end of each session, assigning a unique session ID for each session, and filtering irrelevant records. In these experiments, they follow the cleaning steps and the session identification techniques introduced in [8].

PROPOSED TECHNIQUE

Links are made by Web designers based on relevance of content and certain interests of their own. In our method, we classify Web pages based on hyperlink relations and the site structure. We use this concept to build a category based dynamic prediction model. For example in a general portal www.njiffy.com all pages under the movies section fall under a single unique class.

We assume that a user will preferably visit the next page, which belongs to the same class as that of the current page. To apply this concept we consider a set of dominant links that point to pages that define a particular category. All the pages followed by that particular link remain in the same class. The pages are categorized further into levels according to the page rank in the initial period and later, the users' access frequency [2] [3] [4].

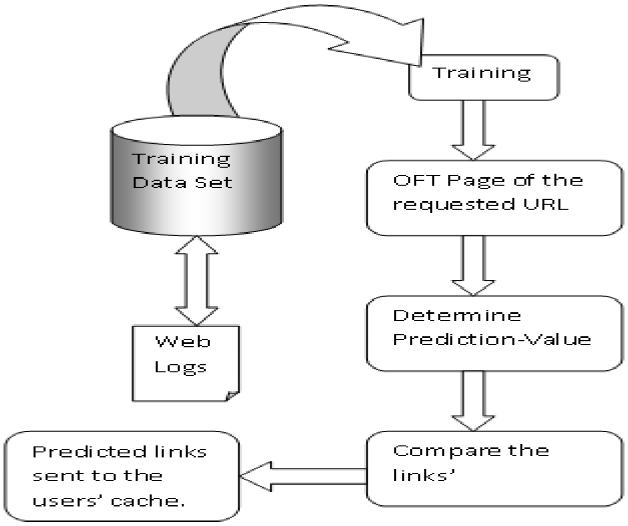

Training
Training Data Set
OFT Page of the requested URL
Web Logs
Determine Prediction-Value
Predicted links sent to the users' cache.
Compare the links'

Figure 3. OFT based prediction model

The major problem in this field is that, the prediction models have been dependent on history data or logs [5]. They were unable to make predictions in the initial stages [6]. We present the structure of our OFT based prediction model in Figure 3. To begin with, HTTP requests arrive at the Predictor Algorithm. The Predictor Algorithm uses the data from the data-structure for prediction. In our prediction model shown in Figure 3, we categorize the users on the basis of the related pages they access. Our model is divided into levels based on the popularity of the pages. Each level is a collection of disjoint classes and each class contains related pages. Each page placed in higher levels has higher probability of being predicted.

In our approach, the OFT Page ranking system resides on the server side and all the information that is required for the computation of our personalized OFT Page Rank algorithm can be derived from the website's Web logs file. We assume that these Web logs are preprocessed and all user sessions are identified. The access oft of a page *m*, and the number of times page *n* was visited right after page m can be obtained simply by counting the number of times page *m* appears and the number of times pages *m* and *n* appear consecutively in all user sessions respectively.

Moreover, if page *m* and *n* were visited consecutively in a user session, and m_t and n_t are the times the user requested them respectively, then m_t-n_t is the approximate time-length the user spent on visiting page *m* in this session. In the case that page *m* is the last page of a user session, for which the access duration cannot be calculated, we can compute the average time-length spent on page *m* from all user sessions as its access time-length. Therefore, when the access time-length spent on a Web page by a user exceeds the average time-length spent on this page by a large percentage, we use this average access time-length as the user's access time-length on this Web page.

Let us consider a web domain as a directed graph *G*, where the nodes represent the web pages and the edges represent the links. Both nodes and edges carry weights, the weight w_m on node *m* is the total time-length all previous users spent on browsing page *m,* while the weight $w_{(n,m)}$ on edge *n*→*m* represents the sum of the time–lengths spent on visiting page *m* when page *n* and *m* were visited consecutively. If we consider all user sessions as 1st-order Markov Chains (in this case, the next page to be visited by a user only depends on the page the user is visiting currently), then $w_{`}$ is the sum of the weights of edges that point to node *m.* Let B_m be the set of pages that point to page *u*, we have the equation:

$$w_m = \sum_{n \in Bm} w(n, m)$$

From the definition of $w_{(n,m)}$ we can see that if more previous users follow the path *n*→*m* and stay on page *m* for a longer time, the value of $w_{(n,m)}$ will be larger, thus $w_{(n,m)}$ covers both information of access time-length and access frequency of a page *m*.

In order to include access oft and accessing time-length of a page to conduct the computation of our personalized OFT Page Rank algorithm O*PR*, we adopt $w_{(n,m)}$ as the biasing factor. When distributing its ranking value to its out links, page *n* will now propagate:

$$\frac{w(n, m)}{\sum_{n \in Fn} w(n, m)}$$

Units of its importance to page *m* in a non-uniform way, where F_n is the set of pages that page *n* points to. We also guarantee the web domains directed graph *G* is strongly connected so that the calculation of *OPR* can converge to a certain value by including the damping factor *(1*-α *)*. Then we eliminate all dangling pages from *G* by adding a link to all other pages in *G* for pages with no out links.

EXPERIMENT AND RESULT ANALYSIS

In this research paper we use the Web logs of the www.bhumisoft.com website. We obtained the Web logs of a 1 week period in Feb 2012 to March 2012 and used the Web logs from 22/Feb/2012 to 02/Mar/2012 as the training data set. We filtered the records and only reserved the hits requesting Web pages (such as *.htm, *.html, and *.aspx). When identifying user sessions, we set the session timeout to 10 minutes, with a maximum of 20 page views per session.

All required information about the pages of the Website is indexed using their URLs in a table where a URL acts as the key. When a request is received, a search on the table is conducted and the information thus obtained is analyzed in the following manner:

a. Check the OFT Page of the requested URL
b. Get Class number of the requested URL.
c. Get the links associated with the page and also fetch their respective OFT Page and class numbers.
d. Determine Prediction-Value (P-value) pairs for the entire candidate URLs, where a P-value pair is defined as [OFT Page, Class,].
e. Compare the links' OFT Page number with the URLs' OFT Page number.
f. Compare the class numbers of the links with that of the requested URL. The link having the same class number will get preference.
g. The links in the higher OFT Pages are the predicted links to be sent to the users' cache.

We chose the most popular paths because using these most accessed paths allowed us to provide a better representation of the typical navigational behaviors of Web users than those paths that are with low access oft. Here we selected the most popular paths the previous users followed from the training data set, and each path was expanded to construct the corresponding sub-graph according to the sessions in the training data set.

CONCLUSION

The time-length spent on visiting a Web page and the oft the Web page is accessed were used to bias OFT Page Rank so that it favors the pages that were visited for a longer time and more frequently than others. However, in our experimental setup, we only propose the 1st-order Markov Chain model, which is "memory less", to calculate the weights of edges in the directed graph of a website and to expand the sub-graph for a user's current navigational path and also will take into account using higher order Markov Chain models to improve the prediction accuracy and also applying the proposed approach to other data sets to evaluate its reliability and performance.

REFERENCE

[1]. Faten Khalil Jiuyong Li Hua Wang, "Integrating Recommendation Models for Improved Web Page Prediction Accuracy", Thirty-First Australasian Computer Science Conference (ACSC2008), Wollongong, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 74, 2008.

[2]. Brin, S., Page, L.: The Anatomy of a Large-scale Hypertextual Web Search Engine. 7th WWW Int. Conf., Brisbane, Australia (1998) 107-117.

[3]. Kleinberg, J.: Authoritative sources in a hyperlinked environment. 9th ACM-SIAM Symposium on Discrete Algorithms, ACM Press (1998) 668-677.

[4]. Mukhopadhyay, D., Biswas, P.: FlexiRank: An Algorithm Offering Flexibility and Accuracy for Ranking the Web Pages. Lecture Notes in Computer Science, Vol. 3816. Springer-Verlag, Berlin Heidelberg New York (2005) 308 – 313.

[5]. Su, Z., Yang, Q., Lu, Y., Zhang, H.: WhatNext: A Prediction System for Web Requests using N-gram Sequence Models. 1st Int. Conf. on Web Information System and Engineering (2000) 200-207.

[6]. Davison, B.D.: Learning Web Request Patterns. Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Springer-Verlag, Berlin Heidelberg New York (2004) 435-460.

[7]. Mamoun A. Awad and Issa Khalil "Prediction of ser's Web Browsing Behavior: Application of Markov Model", IEEE, 2012.

[8]. R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns," *J. Knowl. Inf. Syst.*, vol. 1, no. 1, pp. 5–32, 1999.

[9]. Internet Traffic Archive. [Online]. Available: http://ita.ee.lbl.gov/html/ traces.html