OPENING MINDS:
RESEARCH FOR SUSTAINABLE
DEVELOPMENT

Enhancing the Hyperspace of the OUSL with a Web Page Prediction Model

U. I. Katupitiya, D. D. M. Ranasinghe and G. S. N. Meedin

Department of Electrical and Computer Engineering, Open University of Sri Lanka, Nugegoda, Sri Lanka

**Corresponding author: Email: ddran@ou.ac.lk*

1 INTRODUCTION

The problem of predicting a user's behaviour on a web-site has gained importance due to the rapid growth of the world-wide-web and the need to personalize and influence a user's browsing experience. In addition, the busy lifestyles of users makes them unlikely to tolerate latency in their browsing. Since web sites are constantly competing to attract and retain new visitors, it is of much importance to provide the best possible service to the user who logs onto the site. Therefore, feedback on navigation patterns that occur will significantly aid site owners to efficiently organize the hyperspace they present to their visitors.

The Open University of Sri Lanka is an education institute with a huge number of students. One of its fundamental issues is to provide the information requested by the users, efficiently and accurately. Therefore, the focus of this paper is to develop a system to predict the next web page a user is likely to visit which will aid the university to organize the hyper space in a more efficient manner.

1.1 Literature survey

The focus of this survey was to study and compare the different techniques available to predict the next web page a user is likely to visit. Jalali M, *et al.* (2010), proposed an online and offline phase recommendation system called WebPUM. In Chimphlee *et al*., (2010) an integrated prediction method was proposed by combining the Markov model, Association rules and Fuzzy Adaptive Resonance Theory (Fuzzy ART). Sonavane (2012), address the problem of matching pattern sequences by considering the maximal length common to both sequences and uses the Longest Common Subsequence (LCS) algorithm to overcome this short coming. Kaur et al., 2013 proposed a system architecture for predictions based on Fuzzy Clustering. Langhnoja *et al.* (2013) have used the association rule mining technique combined with the DBSCAN clustering algorithm to predict user behaviour from large data sets. Jarkad, and Bhonsle (2015) have used a graph partition algorithm for clustering the data. Charpate *et al.* (2015) have used a higher order Markov model for predictions along with page ranking. Bohra and Sharma (2016) have performed a comparative analysis of web-mining approaches for efficient mining of server log formats using Apriori and FP Growth techniques. Based on the literature review it is evident that the use of clustering over classification is preferred mainly due to

less the fact that it is less complex and the apriori algorithm is used mostly for association rule mining while predictions are mostly done using higher order Markov models.

2 METHODOLOGY

Considering the facts obtained from the above survey on predicting the next web page users are likely to visit on the OUSL website, association rule mining was used along with kth order Markov model. Figure 1 contains the overall design of the proposed methodology for the prediction system. E-sources for web usage mining are server log files, cookies, data tags, login information., client or server-side scripts etc. files used in this project are in NCSA (Common or Access) combined log format. A portion of the log file format is given in Figure 2.

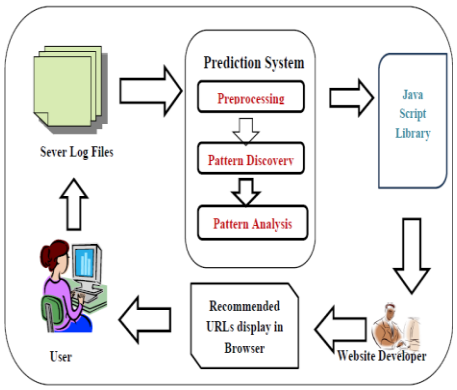

Figure 1: Proposed system architecture

```
175.157.62.8 - - [04/Sep/2016:12:40:13 +0530] "GET /home/index.php/exam-results HTTP/1.1" 200
8782 "-" "Mozilla/5.0 (Windows NT 6.1; rv:47.0) Gecko/20100101 Firefox/47.0"

103.21.166.51 - - [04/Sep/2016:12:40:14 +0530] "GET /home/index.php/exam-results HTTP/1.1" 200
8782 "http://www.ou.ac.lk/home/index.php/exam-results" "Mozilla/5.0 (Linux; Android 4.4.4; E2105
Build/24.0.A.5.14) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.81 Mobile
Safari/537.36"
```

Figure 2: Log file in combined log format

The fields in the NCSA Combined Log Format are:

<host><rfc931><username><date:time>
<request><statuscode><bytes>
<referrer> <user_agent> <cookie>

Data fields obtained from log files as shown in Figure 2 have to be separated before applying the cleaning procedure. This process of separating out different data fields from single server log entry is identified as data field extraction. OUSL log files use space as the separator.

After field extraction a portion of the separated log is given in Figure 3. Next, we will look at the different stages of data pre-processing in detail.

IP Address	Date	Request	Code	Size	Country	Referer	UserAgent
141.0.15.40	[04/Sep/2016:0	GET /home/index	200	1097	Norway	-	Mozilla/4.0 (Windows 98; US) Opera 12.16 [en]

Figure 3: Log data after field extraction

2.1 Data Cleaning

Accessorial resources embedded in the HTML file, robot requests and error requests are to be cleaned. In doing so the following steps were followed.

1. Identify web log records with filename extensions that includes *.gif, *.js, *.jpg, *.jpeg, *.png and *.css, in the requested field and remove those records from the web logs database.
2. Identify records with failed HTTP status code by examining the status code field of every record in the web access logs that has status code greater than 299 and status code less than 200 and remove those records from the web logs database.
3. Identify records with value “GET” in the method field and only retain those in the web logs database. That is, delete the records with the value “POST” or “HEAD” in the method

ISSN 2012-9916 © The Open University of Sri Lanka

field as they do not represent the request from common users.

4. Identify records with text "robot" or "spider" or crawler" in the cs (User-Agent) field and remove those records from the web logs database.

2.2 User and Session Identification

If there are two requests with the same IP address and with different browser name or operating systems then it is considered as requests from two different users. It was assumed that each user has a unique IP address while browsing the website. The same IP address can be assigned to other users after the user completes browsing. A set of user clicks is usually referred to as a click stream. A click stream across a web server is defined as a user session. The timeout mechanism is used as the session identification method. In doing so the followings rules are being employed to identify a user session.

1. If there is a new user, and hence, there is a new session;

2. If the refer page is null in one user session, there is a new session;

3. If the time between two page requests exceeds 30 minutes period, it is assumed that the user is starting a new session. In line with these rules, the web logs of OUSL were analysed and entries were sorted by session ID.

2.3 Path Completion and pattern identification

In order to identify patterns of usage, the complete path of one's browsing session has to be recorded. Sometimes several reasons such as local and agent cache 'post" techniques and the hitting of the browser's back button can result in incomplete paths. In order to find patterns, data with similar features has to be clustered. Based on the literature survey, the k-means algorithm was employed for clustering. To find the co-existence relations between clusters, an apriori algorithm in association rule mining is applied. The final stage of the implementation is identification of the patterns that can be used for next page access predictions.

3 RESULTS AND DISCUSSION

Log files pertaining to the period from the 4th of September 2016 to the 8th of September 2016 of the Open University website was used for the analysis. Figure 4 shows the spider hits and normal hits along with the total hits before pre-processing. Figure 5 shows the total number of hits after pre-processing and the same comparison is given in Table 1 as well.

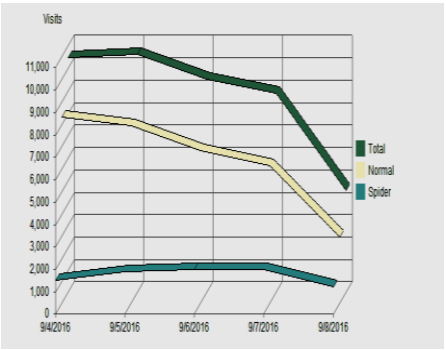

Figure 4: Visits before pre-processing

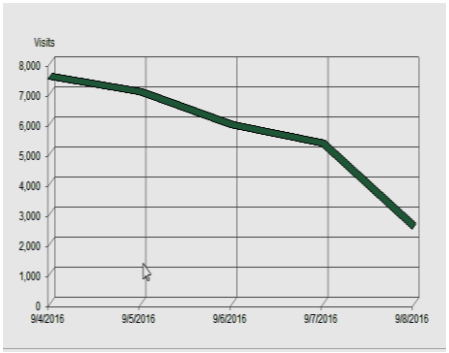

Figure 5: Visits after pre-processing

As Table 1 indicates there are no spider hits and failed hits after pre-processing. Therefore, the total number of hits are reduced. During this period there were 91,418 log records after cleaning log files and there were 5,165 different users and 5,704 different sessions.

Type of hits	Before Pre-process	After Pre-process
Total Hits	2,159,986	233,037
Normal Hits	641,100	233,037
Spider Hits	27,731	0
Average hits / Day	431,997	46,607
Failed Request	68,793	0

As Table 1 indicates there are no spider hits and failed hits after pre-processing. Therefore, the total number of hits are reduced. During this period there were 91,418 log records after cleaning log files and there were 5,165 different users and 5,704 different sessions.

Table 1: Comparison of no. of hits before and after processing

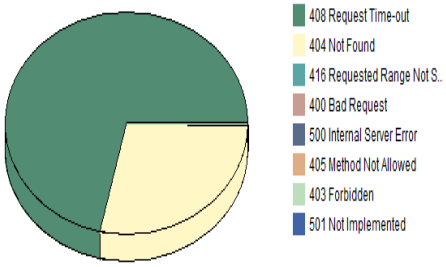

Figure 6: Types of http error codes

In addition, during the pre-processing stage various http error codes were also removed from log files. It was noticed that the highest error code is 408 (Request Timeout) and the next is 404 (Request Not Found). This is diagrammatically shown in Figure 6. The outcome of the pattern analysis was that the next page a user is most likely to visit is the exam results page with a 72% prediction rate. When compared with the actual page views as given in Table 2, it is evident that the prediction is correct.

Table 2: Actual page access results

Pages	Page Views
/home/	28507
/home/index.php/exam-results	26726
/apps/examresults/examresultfiles/getresult.php	20465
/home/index.php/find-a-programme/ undergraduate-programmes	5496
/home/index.php/find-a-programme/ up-comming-programmes	4599
/home/index.php/find-a-programme/ diploma-programmes	4599

ISSN 2012-9916 © The Open University of Sri Lanka

4 CONCLUSONS

The results obtained through this analysis will be particularly important for the Open University of Sri Lanka in organizing the hyperspace of the university, which is the main window of communication. The period where the log data was accessed was an exam result announcing period. Therefore, it is clearly evident that the students will access the exam result page after logging on to the home site of the university. The accuracy of the prediction could have been improved further by applying the techniques in different time lags for predictions. Considering the importance of getting to know the next web page, future work will incorporate this system as a java Script library to allow the web developer to integrate it into the website.

Acknowledgement

We thank the IT division of the Open University of Sri Lanka who provided the web logs from 4th of September 2016 to 8th of September 2016 of the Open University website.

REFERENCES

Charpate, A**.,** Bramhankar, C., Gaikawad, P. and Londhe, A.D**. (**2015**)**. Prediction of Link and Path for User's Web Browsing Using Markov Model.

Chimphlee, S., Salim, N., Ngadiman, M.S.B. and Chimphlee, W. (2010). Hybrid Web Page Prediction Model for Predicting a User's Next Access. Inf. Technol. J. 9, 774–781.

Jalali, M., Mustapha, N., Sulaiman, M.N. and Mamat, A. (2010). WebPUM: A Web-based recommendation system to predict future user movements. Expert Syst. Appl. 37, 6201–6212.

Jarkad, M.P. and Bhonsle, P.M. (2015). Improved Web Prediction Algorithm Using Web Log Data. Int. J. Innov. Res. Comput. Commun. Eng. 3.

Kaur, D., Kaur, A.S. and Punjab, F.S. (2013). User Future Request Prediction Using KFCM in Web Usage Mining. Int. J. Adv. Res. Comput. Commun. Eng. IJARCCE 2

Langhnoja, S.G., Barot, M.P. and Mehta, D.B. (2013). Web usage mining using association rule mining on clustered data for pattern discovery. Int. J. Data Min. Tech. Appl. 2.

Sonavane, V.P. (2012). Study and Implementation of LCS Algorithm for Web Mining