Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Webpage Accessibility Pattern Algorithms

[1]Mansi Yadav, [2]Pankaj Dalal
[1]M.Tech (S.E.) Scholar, [2]Professor
Shrinathji Institute of Technology & Engineering,
Nathdwara, India

Abstarct- During past few years the World Wide Web (WWW) serves as a platform for exchanging various kinds of information. It is most popular way of communication and information distribution. There are millions of web pages are available on the web but still requirements for the new websites arising every day and to accomplish that requirement, new web page are added day by day. With the large amount of information available on the web, the access time is crucial for better performance of websites.
This issue can be addressed by suitable pattern arrangements of web pages. The log data is used for analysis to resolve the accessibility time and pattern search. In this paper, we suggest algorithms for patterns, Most Frequent pattern, Maximum hits pattern and our suggested pattern with their access time comparison.

Keywords - Web Mining, Web Usage Mining (WUM), Server Log File, Data Pre-Processing, Most Frequent pattern, Maximum hits pattern and Suggested pattern.

I. INTRODUCTION

Web Mining is the process of extracting meaningful data from the Web Warehouse. It is classified into three categories: [1][2][6]
Web Content Mining (WCM**)** is the process of extracting from the contents of web documents such as text, audio, video, and image thus it is also known as Text Mining. [1][3][4]
Web Structure Mining (WSM) is the process of mining from Web hyperlink structure. Hyperlink Induced topic search and Page Rank Algorithms are used in WSM. [1][3][4]
Web Usage Mining (WUM) is the process of extracting usage pattern from Web Log Files. It is also known as web Log Mining. [1][3][4]

II. WEB USAGE MINING

WUM is the process of extracting meaningful user's access information from Web data. It is used to understand and improved Web based applications to users. WUM is a powerful tool to analyzing, designing and modifying a websites according to user's access patterns. WUM main three phases are Data Pre-Processing (involve Data Cleaning, User Identification, Session Identification and Path completion steps), Pattern Discovery and Pattern Analysis. [1]
Data Pre-Processing: Data Cleaning removes irrelevant records which are not required in pattern design process. Log file size is reduced and increased the accuracy of log file after cleaning phase. **User Identification** is second steps, find website users assuming new IP-Address represents new User. **Session Identification** finds session of particular users default timeout for single session is 30 minutes. **Path Completion** is last step of Pre-processing, finding complete user access paths and the missing paths are added. **[1]**
Pattern Discovery is the second phase of WUM process, finds out users' access patterns from cleaned log files using different techniques like Sequential Patterns, Association Rules, Clustering and Classification rules. [1][5][6]
Pattern Analysis is the final phase removes uninteresting patterns form pattern design and mined most frequent pattern using knowledge query mechanism such as Structure Query Language (SQL) and Visualization Techniques. [1][5][6]

III. PROPOSED ALGORITHM

Our previous work, we have mentioned about pre-processing algorithms, we have discussed about pattern design algorithms from log data using Most Frequent Pattern algorithm, Max-hits Pattern algorithm and our Suggested Pattern algorithm. [1]
Pattern Design: After pre-processing phase, design users access pattern.

1) URL/Webpage indexing

We have done webpage indexing after finding session of users. We found web pages URL of website and assign number (ID) for easily identify patterns. Thereafter web page URL is replaced by number. Algorithm-1 is for webpage indexing as present below.

Algorithm 1: URL/Webpage indexing

Input: - LogUser File

Output: - PageIndexing file
Step 1: Open LogUser File //in read mode
Step 2: Open PageIndexing File // in write mode
Step 3: read line
Step 4: path=breakPath() // breakPath function is used for break path and remove data after question mark (?)
Step 5: str=strBreakN(path,6) // split the word after given number from line or 6 is given for web page url exist after 6th space
Step 6: ch=read file
Step 7: if(!strstr(ch,str))
Then
Write str in PageIndexing file
End if
Step 8: repeat above 6 & 7 steps until end of file
Step 9: Close both files.

The above algorithm is implemented in C language. 29 web pages are in website and result is shown through the table 1.

Table 1 URL/Webpage Indexing

Index ID	URL/Path	Index ID	URL/Path
1	/aaa.php	16	/Adolescent-Counseling%20.php
2	/contact-us.php	17	/sitemap.php
3	/profile.php	18	/feedback.php
4	/ask-questn.php	19	/faq2.php
5	/faq1.php	20	/refer-friend.php
6	/registration.php	21	/faq3.php
7	/Menopause.php	22	/faq.php
8	/msg.php	23	/Infertility.php
9	/calendar_form.php	24	/Family-Planning.php
10	/due-date-calu.php	25	/index.php
11	/home.php	26	/profile.php/RK=0
12	/pregnancy-delivery.php	27	/profile.php/Menopause.php
13	/appointment.php	28	/googleaab9bbb70d0e1e27.html
14	/Testimonials.php	29	/blog
15	/Laparoscopic-Surgeries.php		

2) Delete Unused Data

This step cleans out single visited URL, extra data and continuous repeated web page URL from LogSession File. Below algorithm-2 removes Unused Data from log file.

Algorithm-2 : Delete Unused Data

Input: - LogSession File (In text format)
Output: - LogAbstract File (In text format)
Step 1: Open LogSession File //in read mode
Step 2: Open LogAbstract File // in write mode
Step 3: if IP have single URL
Then
Remove
End if
Step 4: if same URL repeated in single session
Then
remove
End if
Step 5: else
Write in LogAbstract file upto URL

//only IP address,time and URL write in LogAbstract file
Step 6: repeat above 3,4 & 5 steps until end of file
Step 7: Close both files.

Fig. 1 Users' access patterns

The implementation of the above algorithm removes sessions which have single URL, replaces URL with index ID and removes extra data apart from IP address, time and path (URL). The 86 patterns are found, shown through the snapshot in figure 1.

3) Most Frequent Pattern:
Most Frequent pattern is found from cleaned log file by using algorithm-3. This algorithm finds number of hits frequency of particular pattern and pattern having highest frequency is our most frequent pattern.

Algorithm-3: Most Frequent Pattern

Input: - LogPattern File (In text format)
Output: - Most Frequent pattern, LogDisplay File
Step 1: Open LogPattern File //in read mode
Step 2: Open LogDisplay File // in write mode
Step 3: if patterns have less than 3 pages (URL)
Then
Remove patterns
End if
Step 4: read pattern from file and calculates frequency (number of hits) of pattern
Step 5: repeat above 4th step until end of file
Step 6: which pattern have highest frequency, display that pattern as a most frequent pattern.
Step 7: Close both files.

The implementation of the above algorithm have found Pattern "Start→ 22→11→3→8→12→4→13→4→End" is MF pattern in drgoyal website its means users first access faq.php webpage then goes to home.php webpage and the rest. "Faq.php→ home.php→ profile.php→ msg.php→ pregnancy-delivery.php→ ask-questn.php→ appointment.php→ ask-questn.php" pattern is used by mostly clients. The snapshot of most frequent pattern is shown in figure 2.

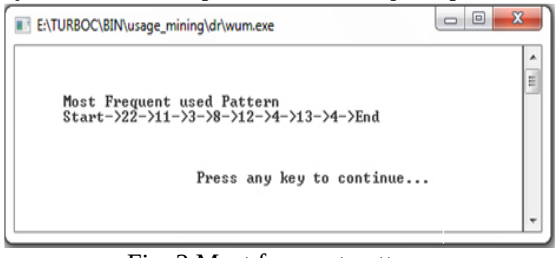

Fig. 2 Most frequent pattern

4) Maximum Hits Pattern:
Max-hits pattern is pattern that represents the maximum hits webpages in descending order. It means webpage that is hits maximum time has highest priority and webpage that has less hits have less priority. Max-hits pattern is used to rearrange the webpage's on the basis of hit priority.

The algorithm-4 is as shown for finding Max-hits pattern. This algorithm first find average length of pattern by using minimum length of pattern and maximum length of pattern.

Algorithm-4: Max-hits pattern

Input: - LogPattern File (In text format)
Output: - Max Hits pattern
Step 1: Open LogPattern File //in read mode
Step 2: if patterns have less than 3 pages (URL)
Then
Remove patterns
End if
Step 3: find maximum and minimum length of patterns and calculate average length of pattern

$$\text{Length}_{Avg} = \frac{min+max}{2}$$

Step 4: read pattern from file and calculates frequency (number of hits) of pages
Step 5: repeat above 4th step until end of file
Step 6: which page have highest frequency, display that page first(arrange in decreasing order)
Step 7: pattern length is equal to Length_{Avg}
Step 8: Close file.

The implementation of the above algorithm have found Pattern "Start→11→12→13→22→2→7→3→4→8→21→23→14→6→24→25→10→17→19→5→9→20→18→15→16→26 →End" is Max-hits pattern in drgoyal website. The snapshot of Max-hits pattern is shown in figure-3. Average length of pattern is 8 hence our max-hits pattern is "Start→11→12→13→22→2→7→3→4" its means "home.php → pregnancy-delivery.php → appointment.php → faq.php → contact-us.php → Menopause.php → profile.php → ask-questn.php" is pages in our max-hits pattern.

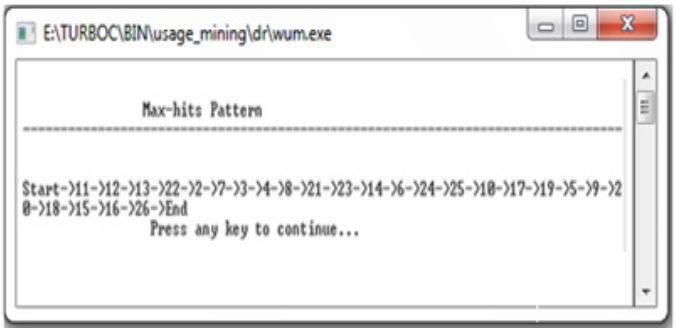

Fig. 3 Max-hits pattern

5) Our Suggested Pattern:

Suggested pattern reduces the user's access time of websites. This algorithm-5 finds page hits, max hit page to min hits page, forms our suggested pattern and calculates average length of pattern.

Algorithm-5: Suggested pattern

Input: - LogPattern File (In text format)
Output: - Suggested pattern
Step 1: Open LogPattern File //in read mode
Step 2: if patterns have less than 3 pages (URL)
Then
Remove patterns
End if
Step 3: find maximum and minimum length of patterns and calculate average length of pattern

$$\text{Length}_{Avg} = \frac{min+max}{2}$$

Step 4: read pattern from file and calculates highest frequency (number of hits) of page.
Step 5: repeat above 4th step until end of file
Step 6: which page have highest frequency in column, display that page
Step 7: pattern length is equal to Length_{Avg}
Step 8: Close file.

The implementation of the above algorithm have found Pattern "Start→11→7→3→2→12→4→13→18→19→15" which is our suggested pattern for drgoyal website. The snapshot of suggested pattern is shown in figure-4. Average length of pattern is 8 hence our suggested pattern is "Start→11→7→3→2→12→4→13→18" its means "home.php → Menopause.php →profile.php →contact-us.php→ pregnancy-delivery.php→ ask-questn.php→ appointment.php →feedback.php" is our suggested pattern.

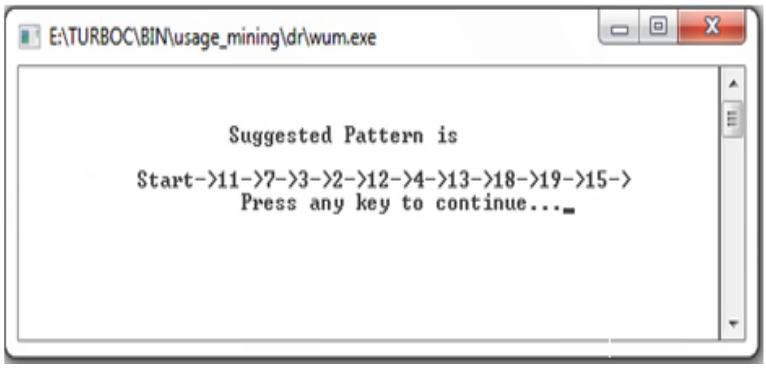

Fig. 4 Suggested pattern

IV. EXPERIMENTAL RESULTS

Table 2 shows pattern access time for original, max-hits, most frequent and suggested pattern. U1, U2….Un represents number of users, S1,S2….Sn represents number of session of particular users. The access time comparison of the entire patterns, Original (website is in the beginning), Max-hits user's access (website is arranged according to the found max-hits pattern), Most frequent user's access pattern (website is arranged in the most frequent pattern), and Suggested access pattern (website is arranged according to the Suggested pattern).

We have calculated access time for all patterns and found users access time are less when website is arranged according to the our suggested pattern. Some time may be possible the access times are less in other patterns but mostly access time is less in suggested pattern.

Table 2 Pattern Access time

Users	Sessions	Users Access Patterns	Access Time			
			Original Pattern	Max Hit Pattern	Most Frequent Pattern	Suggested Pattern
U1	S1	Start-> 3->2->3->2->End	00:18	00:20	00:10	00:08
U2	S1	Start-> 2->13->14->15->16->17->8->18->19->End	00:58	00:56	00:48	00:41
U3	S1	Start-> 6->20->3->21->22->10->End	00:42	00:50	00:27	00:23
U4	S1	Start-> 4->7->23->24->5->End	00:21	00:30	00:27	00:18
U4	S2	Start-> 24->7->5->End	00:09	00:12	00:10	00:10
U4	S3	Start-> 8->24->7->5->End	00:18	00:19	00:22	00:15
U5	S1	Start-> 11->22->12->2->12->23->13->24->22->4->3->14->12->16->12->4->13->End	01:36	01:59	01:29	00:59
U6	S1	Start-> 12->8->11->4->End	00:12	00:14	00:18	00:09
U6	S2	Start-> 17->25->11->12->14->13->End	00:45	00:33	00:22	00:22
U7	S1	Start-> 11->8->4->End	00:09	00:05	00:09	00:05
U8	S1	Start-> 22->11->3->8->12->4->13->4->End	00:34	00:31	00:37	00:26
U9	S1	Start-> 18->8->17->2->End	00:14	00:10	00:13	00:10
U10	S1	Start-> 3->10->22->21->End	00:17	00:12	00:17	00:11
U11	S1	Start-> 8->11->12->End	00:08	00:08	00:06	00:07
U12	S1	Start-> 11->3->8->12->4->End	00:16	00:19	00:19	00:12
U13	S1	Start-> 4->22->12->End	00:12	00:09	00:09	00:08
U14	S1	Start-> 2->3->8->4->End	00:13	00:07	00:25	00:09
U15	S1	Start-> 11->14->13->2->3->8->4->End	00:27	00:19	00:27	00:17
U16	S1	Start-> 11->13->3->12->22->5->6->9->14->10->17->18->End	01:01	00:46	00:44	00:41
U17	S1	Start-> 11->2->3->13->18->8->4->End	00:36	00:20	00:23	00:18
U18	S1	Start-> 2->3->13->18->8->4->14->11->End	00:40	00:19	00:27	00:21
U19	S1	Start-> 15->8->3->10->18->20->17->22->5->19->21->End	01:05	00:45	00:55	00:39

V. CONCLUSION

Our work, presents the Most frequent pattern algorithm, Max-hits pattern algorithm and Suggested pattern algorithm which have been designed to access patterns in log file with respect to the users and frequency of webpage hits. The live website is changed as per the pattern received by our algorithms and found the suggested algorithm pattern is giving lowest access hit time.

REFERENCES

[1] Mansi Yadav and Pankaj Dalal, "Algorithms for Web Log Data: WUM Pre-Processing phase" International Journal of Engineering Research & Technology (IJERT),Vol. 3 Issue-12, December-2014.

[2] Vijay Kumar Padala, Sayeed Yasin, Durga Bhavani Alanka, " A Novel Method for Data Cleaning and User- Session Identification for Web Mining", International Journal of Modern Engineering Research, Vol. 3, Issue. 5, Sep - Oct. 2013.

[3] Shaily G.Langhnoja,Mehul P. Barot,Darshak B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery", International Journal of Data Mining Techniques and Applications, Vol 02, Issue 01,June 2013.

[4] Devinder Kaur, Ravneet Kaur, "Minimizing the Repeated Database Scan Using an Efficient Frequent Pattern Mining Algorithm in Web Usage Mining", International Journal of Research in Advent Technology, Vol.2, No.6, June 2014.

[5] K.S.R. Pavan Kumar, L. Manoj Chowdary, V.V. Sreedhar, "A Critique on Web Usage Mining", International Journal of Computer Science and Information Technologies, Vol. 3.

[6] L.K. Joshila Grace, Dhinaharan Nagamalai, V.Maheswari, "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

