ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Examining Frequent URLs for Speeding up the Web Access

Jaswinder Kaur
Dept. of Computer Science & Applications
Kurukshetra University, Kurukshetra
Haryana, India

Dr. Kanwal Garg
Dept. of Computer Science & Applications
Kurukshetra University, Kurukshetra,
Haryana, India

Abstract: These days web is a vast collection of information and is growing rapidly. In the real world, enterprises and organizations have challenges to keep a clear and well organized Websites. Web master may need to know about visitor's access information such as mostly visited path for frequently accessed pages. Web usage mining process helps in identifying usage pattern from web data in pattern discovery phase. Usage pattern plays a significant role to predict user access pattern while the user is interacting with web sites. In this paper, researcher proposed the procedure to examine those web pages which are browsed by the user most frequently and to term those URL's as frequent pattern, that may be handed over to the web master.

***Keywords*:** Frequent Pattern; Web Page; Minimized Frequent Pattern; Web Usage Mining; Web Log

I. INTRODUCTION

World Wide Web also known as the Web which has grown at the rapid rate from where a user can access valuable information since the technology is available in every field such as business, e-commerce, medical, education. Web mining is classified into three area of interest which are content mining, structure mining and web usage mining. Web usage mining objective is to understand the user navigational pattern and their use of web resources, which requires user's access information to understand and better serve the web based applications[9]. Web usage mining consists of several stages such as data collection as a web log, data preprocessing, pattern discovery and analysis or interpretation of frequent pattern as shown in Figure 1. Web server is computer that runs the software application (HTTP Server Software) such as Apache and IIS to transfer the Web pages on the internet or intranet. When a user visits the web pages, some valuable information pertaining to the user is stored in Web log files on web server such as IP address, user name, time stamp and referer[7]. These Log files are collected on various locations like web server side, proxy server side, and client browser side. For getting optimum results, need to extract user's data from these log file in data collection phase. Access logs can exist in various formats such as Common log file format, Combined log format and IIS log format[10]. In second phase, preprocessing remove incomplete, inconsistent, noisy data from raw web log file to obtain processed data, which will be used for pattern mining. Data preprocessing includes various steps such as data cleaning, User identification and session identification[5]. In Web usage mining, pattern discovery is the third phase that applies various techniques on processed data such as statistical analysis, association rules, classification and clustering to discover frequent patterns. Researcher had used association techniques to determine the frequent pattern. Analysis or Interpretation of frequent pattern is the final phase of web usage mining. This phase consists of filtering of interesting patterns from patterns that were found in the earlier phase, that in turn helps the web designer or developer to understand the need of a user, it also assists in predicting the user interested pages and customizes web pages to the user navigational knowledge.

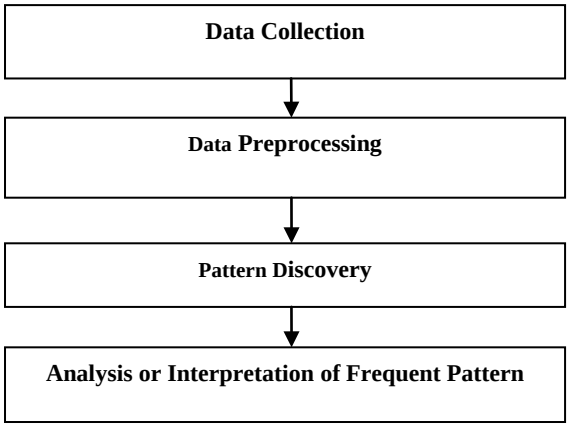

Figure 1. Steps For Web Usage Mining

II. RELATED WORK

Pattern analysis is the last step, that is applied on the frequent pattern to filter out the interesting pattern in Web usage mining. Yongjian Fu et.al (2001)[2] classified web pages into index page and content page based on user access information. After this, site reorganization approach is used to find better ways to reorganize the Web pages on the site. Mrs. Geeta R.B et.al (2009)[6] explored that web log file can be used to collect statistics that helps in knowledge discovery. This knowledge is used to take decision in the Web application on various factors such as popular pages, navigation pattern and time. The website is reorganized based on user navigation behavior or hit counts of the web page that provide quick response to the web user. M.

Gnanavel and Dr. E. R. Naganathan (2012)[3] explained how Web visualization techniques such as standard 2D/3D display, geometrically-transformed display, iconic display and stacked display can be used in customization. Joy Shalom Sona & Asha Ambhaikar (2012)[8]) proposed a reconciling website system which includes mining the web architecture, determining user log and obtaining website browsing efficiency to improve web navigation efficiency. Kamika Chaudhary and Santosh Kumar Gupta (2013)[1] focuses on different types of tools such as Webalizer, Naviz and WebViz. It also present pattern analysis techniques such as knowledge Query mechanism and OLAP is used in various applications such as personalization, system improvement, site modification and e-commerce. Mr. P.G.Vedaprakash et.al (2016)[11] adapted tire algorithm to construct a tree structure that also captures user visit frequencies, which is called trial tree algorithm. In which, a complete path from the root to the leaf node is called a trial. Clustering is classified into structure-base, attribute-base and structure/attribute. This is used to identified the frequent and semifrequent customers for on-line shopping.

III. Frequent Pattern using Modified Apriori Algorithm

The researcher have used the web server log data of National Philosophical Counseling Association of size 94,646 KB. After preprocessing phase, obtained processed file size is of 33,491 KB. Researcher found frequent patterns from processed data of different pattern length by using modified Apriori algorithm in which pattern length 5 has 26250 frequent pattern out of 97423 pattern and pattern length 6 has 82504 frequent pattern out of 315362 pattern as shown in Figure 2[4].

Pattern Length	No. of Pattern	No. of Frequent Pattern
1	17	5
2	191	37
3	1985	507
4	17712	4364
5	97423	26250
6	315362	82504

Figure 2. Number of Frequent Pattern in Different Pattern Length

This approach proves to be more efficient than already existing Apriori algorithm as proposed algorithm generates frequent patterns with large pattern length, especially if large number of users exist in web log file. Figure 3 shows frequent patterns with support of pattern length 4. One row indicates one frequent pattern that has four page and last page is the target page for example, in normal-root-includes-content pattern, content is the target page. After examining the frequent pattern, the researcher found that a large number of different patterns are visited by the user to reach a particular target page and sometime same web pages can be visited more than once in a pattern or navigational path which is not useful to web master for particular target page such as includes-content-content-journal, and journal-includes-content-journal.

Support	Page1	Page2	Page3	Page4
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....
566	normal	root	includes	content
561	normal	includes	content	journal
520	includes	content	docs	docs
519	includes	content	content	docs
491	normal	journal	includes	includes
486	normal	includes	content	includes
477	normal	journal	content	content
474	normal	root	content	includes
468	normal	includes	docs	docs
458	journal	includes	docs	docs
457	includes	content	content	journal
446	root	content	includes	training
441	includes	content	journal	journal
440	root	includes	training	training
435	normal	journal	includes	journal
428	journal	includes	content	journal
424	normal	journal	content	includes
421	root	includes	content	content
418	normal	content	includes	training
418	journal	journal	journal	journal
415	root	includes	content	training
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....

Figure 3. Frequent Patterns of Pattern Length 4

IV. Proposed Methodology

Researcher proposed a methodology to find out mostly visited navigational path for a target webpage without repetition of the same web page in the frequent pattern.

1. Proposed Methodology to find out mostly visited navigational path for a target page.

Input: Pattern of X Path
X Path= $p_i, p_{i+1,}$……,p_x Pages, p denotes page number and k denotes the level number.

a) Step 1: Select one page_type for target page example p_x. Then, select all frequent patterns that has target page. From which select the frequent pattern that has support value greater than average support value.

Avg support=sum (support)/no. of frequent pattern

b) Step 2: Set i=1and k=0, calculate the support for each page_type in p_i and arrange all page_type in descending order. Set p_i as $level_k$.

Support = No. of occurrences of page_type.

c) Step 3: If target page has largest support value than other pages in $level_k$ then select target page with support value under the heading "Minimized Frequent Pattern." Then goto step 11. Otherwise, set all page_type as valid except target page in $level_k$

d) Step4: Calculate the support for each page_type in P_{i+1} and arrange all page_type in descending order. Set all page_type as valid in p_{i+1}. Set C1= $level_k$ and C2=p_{i+1}

e) Step 5: Make a combination of each valid page_type in C1 with all valid page_type in C2 and each combination has unique page_type. After each combination, count the support for a combination. The page_type in C1 and C2 that has support less than combination's support, that will come under invalid page_type. After this, Arrange all combination with support in descending order.

f) Step 6: Increase k by one for next level. Set combinations with the largest support as next level $level_k$.

g) Step 7: If all combination has the support value not greater than one in $level_k$. Print all combination of C1 under the heading "Minimized Frequent Pattern" Then goto step 11 .

h) Step 8: Otherwise, one condition is true out of two conditions in $level_k$. The first condition is, one combination exist in $level_k$ and combination contain target page in $level_k$. Then print the combination with support under "Minimized Frequent Patten." Then goto step 11. Second, any combination contains target page in $level_k$. Then select all combination except target page from $level_k$. If these two conditions do not exist then select all combination with support from $level_k$

i) Step 9: Set all combination as valid page_type in $level_k$ and increase i by one for next page.

j)Step 10: Repeat step from step 4 to step 9 and until k< x-1.

k) Step 11: Exit

V. EXPERIMENTAL RESULTS

Proposed procedure performed for frequent patterns those having target page journal in pattern length 4 in C # .net.

Iteration 1:

Support	Page1	Page2	Page3	Page4
561	normal	includes	content	journal
457	includes	content	content	journal
441	includes	content	journal	journal
435	normal	journal	includes	journal
428	journal	includes	content	journal
418	journal	journal	journal	journal
397	normal	includes	journal	journal
357	normal	root	includes	journal
349	normal	journal	journal	journal
330	root	includes	journal	journal
329	normal	root	journal	journal
323	includes	includes	content	journal
311	root	journal	journal	journal
309	journal	includes	journal	journal
307	root	includes	content	journal
302	normal	content	includes	journal
277	root	content	includes	journal
273	content	includes	journal	journal
266	includes	journal	journal	journal
257	normal	content	journal	journal
253	root	content	journal	journal
.....	.....	.....	.....	.....
.....	.....	.....	.....	.....

Figure 4. Frequent Patterns For A Target Page Whose Support Value > Average Support Value

Select frequent patterns of target page journal with support. 625 frequent patterns of journal exist out of 4364 pattern in pattern length 4. After this, calculate average support for journal target page. In Figure 4, frequent patterns is selected which has support value greater than average support value.

Average Support= 40977/625= 65.5

$Level_0$	
Page1	**Support**
normal	47
includes	43
root	41
content	37
journal	13
about	6
application	3
docs	1

Figure 5. Page_Type With Corresponding Support Value In Page1

Proposed procedure

In Figure 5, page 1 has eight different page_type such as normal, include, root and content. Count support for each page_type and arrange all page_type with support in descending order under $level_0$ for example, normal has largest support 47 then includes , so on. Journal page_type support is not highest. So that, select all page type as valid except journal page_type.

C2	
Page2	**Support**
includes	46
content	42
journal	20
application	18
about	17
root	16
training	13
conferences	7
docs	7
contacts	2
people	2
documents	1

Figure 6. Different Page_Type With Corresponding Support Value In Page2

Page 2 in pattern length 4, has 12 different page_type such as includes, content and journal. Then count support for each page_type and arrange all page_type with support in descending order. Then, select all page type as valid under C2 as shown in Figure 6.

In Figure 7, Make a combination of first valid page_type in C1 with all valid page_type in C2 without repetition of same page until all valid page_type of C1 has made a

combination with all valid page_type of C2, for example firstly valid page_type normal of C1 make a combination with includes valid page_type of C2 then count support for a combination normal, includes. Those valid page_type of C1 and C2 has support less than support of normal, includes that came under invalid page_type such as about, application, docs and conferences. After this take next valid page_type of C1.

C1	Support	C2	Support	Combination	Support
normal	47	includes	46	normal, includes	12
includes	43	content	42	normal, content	11
root	41	journal	20	normal, journal	4
content	37	application	18	normal, application	2
about	6	about	17	normal, about	4
application	3	root	16	normal, root	11
docs	1	training	13	normal, training	0
		conferences	7		
		docs	7	includes, content	12
		contacts	2	includes, journal	4
		people	2	includes, application	6
		documents	1	includes, about	3
				includes, root	4
				includes, training	4
				.....	.....
				.....	.....

Invalid Page_Type

Figure 7. Combination of Valid Page Type of C1 And C2

$Level_1$	Support
normal, includes	12
includes, content	12
content, includes	12

Figure 8. Combination of Page_Type With Corresponding Support Value In $Level_1$

Figure 8, shows those combinations, which has largest support value than others under $level_1$. These are all valid page_type.

Iteration 2:

Calculate support for different page_type of page 3 under C2 such as journal, application and conferences. Set all page_type as valid and arrange in descending order in Figure 9. Combinations of $level_1$ assigned to C1. Make a combination of C1 and C2 as said above in iteration 1 as shown in Figure 10, such as normal, includes, journal and normal, includes, application.

C2	
Page 3	**Support**
journal	45
application	19
conferences	19
training	19
docs	16
includes	16
contacts	13
content	13
about	11
people	9
root	7
documents	4

Figure 9. Different Page_Type With Corresponding Support Value In Page3

C1	Support	C2	Support	Combination	Support
normal, includes	12	journal	45	normal, includes, journal	1
includes,content	12	application	19	normal, includes, application	1
content,includes	12	conferences	19	normal, includes, conferences	1
		training	19	normal, includes, training	1
		docs	16	normal, includes, docs	1
		includes	16	normal, includes, contacts	1
		contacts	13	normal, includes, content	1
		content	13	normal, includes, about	1
		about	11	normal, includes, people	1
		people	9	normal, includes, root	1
		root	7	normal, includes, document	1
		documents	4		
				includes, content, journal	1
				includes, content, application	1
				includes, content, conferences	1
				includes, content, training	1
				.....	.....
				.....	.....

Figure 10. Combination of Valid Page Type of C1 And C2

In Figure 11, $level_2$ has combinations, which has largest support value. Support value is one. So, print all combination of C1 under the heading “Minimized Frequent Pattern” such as normal, includes, journal and includes, content, journal as shown in Figure 12.

Level$_2$	Support
normal, includes, journal	1
normal, includes, application	1
normal, includes, conferences	1
normal, includes, training	1
normal, includes, docs	1
normal, includes, contacts	1
normal, includes, content	1
normal, includes, about	1
normal, includes, people	1
normal, includes, root	1
normal, includes, document	1
includes, content, journal	1
includes, content, application	1
includes, content, conferences	1
includes, content, training	1
includes, content, docs	1
includes, content, contacts	1
. . . . .	. . . . .
. . . . .	. . . . .

Figure 11. Combination of Page_Type With Corresponding Support Value In Level$_2$

Minimized Frequent Pattern
normal, includes, journal
includes, content, journal
content, includes, journal

Figure 12. Minimized Frequent Pattern

In Table 1, Target page_type journal and training has 625 and 758 frequent pattern. Only two frequent patterns are taken into consideration for each target page such as journal, training, docs and content that are having pattern length 4 and 5.

Frequent patterns for journal and training are normal-includes-content-journal, content-includes-content-journal, root-includes-root-training and includes-includes-application-training. After applying above proposed procedure on all frequent patterns that have one type of target page of same pattern length, researcher discovered only three minimized frequent pattern of length 3 for the journal such as normal-includes-journal, includes-content-journal and content-includes-journal.

Five minimized frequent pattern of length 3 for the training such as root-includes-training, root-content-training, normal-includes-training, includes-content-training and content-includes-training.

Table I. Minimize The Frequent Pattern Length For A Target Page

Pattern Length	***Target Page***	***No. of Frequent Pattern***	***Frequent Pattern***	***No. of Minimized Pattern***	***Minimized Frequent Pattern***
4	journal	625	normal, includes, content, journal content, includes, content, journal	03	normal, includes, journal includes, content, journal content, includes, journal
	training	758	root, includes, root, training includes, includes, application, training	05	root, includes, training root, content, training normal, includes, training includes, content, training content, includes, training
5	docs	2646	normal, root, includes, journal, docs root, includes, content, includes, docs	02	normal, root, includes, docs normal, root, content, docs
	content	1072	normal, root, includes, application, content root, content, includes, application, content	01	normal, root, includes, content

Similarly, proposed procedure is applied to frequent patterns for docs and content that are having pattern length 5. normal-root-includes-docs, normal-root-content-docs, are the minimized frequent pattern for docs. Only one minimized frequent pattern is discovered for content that is normal, root, includes, content.

Figure 13. showing number of patterns of frequent and minimized frequent pattern in pattern length 4.

A number of patterns are less and pattern length is reduced in minimized frequent pattern as compared to frequent patterns for one target page. These minimized frequent patterns are mostly visited navigational path for one target page without repetition of the web page. These minimized patterns are more useful for the webmaster to restructure their website in term to reduce web access time while user accessing the website.

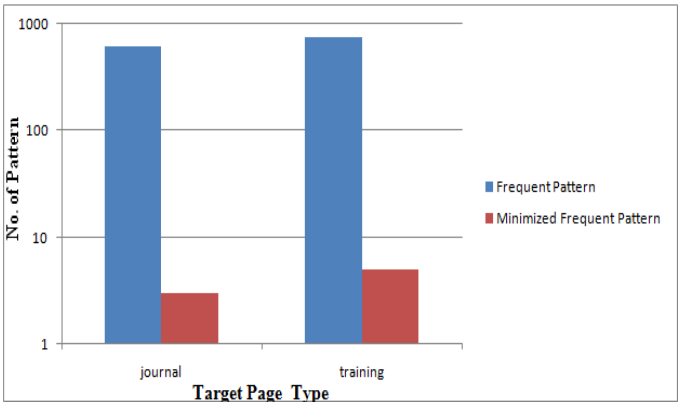

Figure 13. Comparison Between Frequent And Minimized Frequent Pattern in Pattern Length 4

VI. SPEED UP THE WEB ACCESS

In order to increase the speed of web access, essential thing is that a user spends minimum time on the web. Here researcher have developed a web crawler to calculate the time for a web page means how much time a pattern will take to open on the web. The Researcher found minimized frequent pattern take less time than the frequent pattern. In Table 2, normal-includes-content-journal pattern of length 4 taken 0.443 ms but normal-includes-journal pattern of length 3 taken 0.366 ms for a target page journal. On the other hand, normal-root-includes-application-content pattern taken 0.514 ms but normal-root-includes-content pattern of length 4 taken 0.453 ms for a target page content at a bandwidth of 100.0 Mbps and download speed of approx. 1.90 Mbps.

Web master can use these minimized frequent pattern for restructuring their website to speed up the web access. Thus, the web user can access the targeted page easily and spend less time while interacting with the website.

Table II. Time Comparison Between Frequent Pattern And Minimized Frequent Pattern

Pattern Length	*Target Page*	*Frequent Pattern*	*Time (ms)*	*Minimized Frequent Pattern*	*Time (ms)*
4	journal	normal, includes, content, journal	0.443 ms	normal, includes, journal	0.366 ms
5	content	normal, root, includes, application, content	0.514 ms	normal, root, includes, content	0.453 ms

VII. CONCLUSION

Today World Wide Web is growing rapidly, because online system is popular in the market. Pattern discovery is a significant phase of web usage mining that extracts usage pattern from processed web log data. After this, frequent patterns are examined. The frequent patterns contain several patterns for a target page. In addition to this, sometime same web page is repeated in the pattern for the particular web page. Researcher proposed the procedure to extract mostly visited paths with unique web pages to achieve target web page from the frequent patterns in different pattern length. These patterns can be handed over to the web master for the restructuring of the website. That will enable the user to access goal pages in less time.

VIII. REFERENCES

[1] Kamika Chaudhary and Santosh Kumar Gupta, "Web Usage Mining Tools and Techniques: A Survey", International Journal of Scientific & Engineering Research, ISSN: 2229-5518, Vol. 4, Issue 6, pp. 1762-1768, June 2013.

[2]Yongjan Fu, Mario Creado and Chunhua Ju, "Reorganizing Web Sites Based on User Access Patterns", in CIKM '01 Proceedings of The Tenth International Conference On Information And Knowledge Management, pp. 583-585, 2001.

[3] GM. Gnanavel and Dr. E.R. nagathan, "A study of patternAnalysis Techniques of Web Usage", International Journal of Web Technology, Vol. 01, No. 01, pp. 5-10, June 2012

[4] Jaswinder Kaur and Kanwal Garg, "Jaswinder Kaur and Kanwal Garg, "Discovery of Frequent Usage Pattern For Web Data To Optimized Web Based Applications", in the International Journal of Computer Application, ISSN: 0975-8887, Vol. 145, No. 13, pp.14-17, July 2016.

[5] Naga Lakshmi, Raja Sekhara and Satyanarayana Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis, in the International Journal of

Innovative Technology and Exploring Engineering, ISSN:2278-3075, Vol 2, Issue 4, pp. 274-279, 2013.

[6] Mrs Geeta R.B, Prof. Shashikumar G.Totad and Dr. Prasad Reddy PVGD, "Amalgamation of Web Usage Mining and Web Structure Mining", in the International Journal of Recent Trends in Engineering, ISSN 1797-9617, Vol. 1, No. 2, pp.279-281, May 2009.

[7] Divya Racha, "Web Usage Mining For Extracting User's Navigational Behaviour", International Journal of Engineering And Computer Science, ISSN: 2319-7242, Vol. 3, Issue 5, pp. 5989-5995, May 2014.

[8] Joy Shalom Sona and Prof. Asha Ambhaikar, "Enhancing the Website Structure by Reconciling Website", in the IOSR Journal of Engineering (IOSRJEN), ISSN: 2250-3021, Vol. 2, Issue 9, PP 122-125, 2012.

[9] Anitha Talakok, "A Survey on Web Usage Mining, Application and Tools", in the Computer Engineering and Intelligent Systems, ISSN 2222-1719, Vol. 6, No. 2, 2015.

[10] Priyanka Verma and Dr. Nishtha Kesswani, "Web Usage Mining Framework For Data Cleaning and IP Address Identification", International Journal of Advanced Studies In Computer Science and Engineering (IJASCSE), Vol. 3, No.8, 2014.

[11] Mr. P.G.Vedaprakash, Mr. P.G.Om Prakash, and Mr. M.Navaneethakrishnan, "Analyzing The User Navigation Pattern From Weblogs Using Data Pre-Processing Technique" International Journal of Computer Science and Mobile Computing, ISSN 2320–088X**,** Vol. 5, Issue 5, pp. 90 – 99, May 2016.

