IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Understanding User's Navigation Behavior using Web Mining Algorithm

Sana M. Deshmukh[1], **Krishnakant P. Adhiya**[2]

Department of Computer Engineering, SSBT's COET, Jalgaon, Maharashtra[1, 2]

Abstract: Web mining is application of data mining which is useful to extract the knowledge. So, wecan use web mining algorithm in understanding users' navigation behavior. In the proposed architecture first step is the preprocessing of web log data. In pre-processing step, modified data cleaning algorithm removes all irrelevant entries from weblog file. After data cleaning step user identification is performed depends on user's domain name or IP addresses. After preprocessing step, web mining algorithms of proposed architecture are applied to study the user navigation behavior. In the proposed architecture three algorithms are proposed. Those algorithms are, Modified Data Cleaning Algorithm, Proposed Modified Clustering Algorithm-I and Proposed Algorithm-II based on preferred path mining. Modified data cleaning algorithm is uses to remove all irrelevant entries and all multimedia files from weblog file. Proposed modified clustering algorithm-I is adapts to cluster users based on their similarity. Proposed algorithm-II based on preferred path mining accomplishes path mining on the clusters of different users found in clustering step. The results of proposed algorithms show improvement in memory usage and execution time.

Keywords: Web Usage Mining, Pre-Processing, Weblog File.

I. INTRODUCTION

World Wide Web there is a huge, connected, semi-structured, widely distributed, highly heterogeneous and hypertext information repository. Web continues to grow at an incredible rate as information gateway. Web mining technologies are the proper solutions for knowledge discovery on the web. Web mining is application of data mining techniques to discover patterns from the web [1]. Web mining classified into three categories.

Web Content Mining: It mines the data Sources on the web and extract the patterns from web data sources.
Web Structure Mining: It mines the structures on the web and use linkage information to improve search engines.

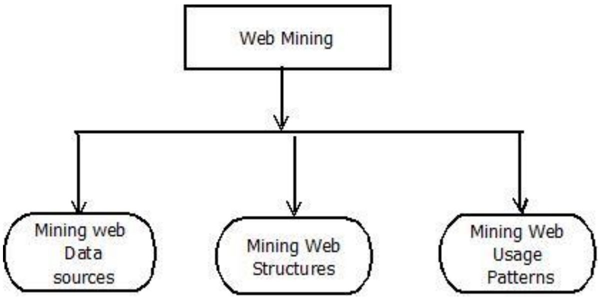

Figure1.Structure of Web Mining

Web Usage Mining: It mines the usage Patterns on the web and improves web usability and user experience.

To understand user's navigation behavior and area of interest recent researches are focusing on web usage mining in the proposed method also web usage mining is applied to extract the usage patterns.

Web Usage Mining
Web usage mining is used to extract the usage patterns from web data. It employs improvementin user's web search experience and web usability. Web usage mining has three different tasks to perform. Those three processes of web usage mining are as follows,

1. Pre-processing
2. Pattern discovery
3. Pattern analysis

DOI10.17148/IJARCCE.2017.6837

Web usage mining is the process which extracts the useful usage patterns from web data. In recent search web usage mining is use to understand user's interest in web search. Thereforeit adduced on weblog file. Weblog file which contain user history and his/her earlier access records. To understand user's behavior web usage mining techniques has employed.

Pre-processing: It is a first step of web usage mining process which performs cleaning of weblog file. In preprocessing step unnecessary data or noisy data from weblog file is cleans and log file size is reduces.
Pattern discovery: Second step of web usage mining process is pattern discoverythat performs on cleaned log file generated in the prepossessing step to discover web patterns.
Pattern analysis:Final step of web usage mining process is pattern analysis. It accomplishes pattern analysis on pattern discovered from second step of web usage mining to generate more useful information related to the user behavior pattern.

II. LITERATURE SURVEY

K. R. Suneetha, Dr. R. Krishnamoorthi in [1], proposed an approach using web usage mining technique. In this proposed technique pre-processing is accomplished on NASA weblog file after, this pattern discovery and pattern analysis has been applied. This approach proposed to find top errors, potential visitors of the web site. This method is very useful for web site designer to improve their system by determining occurred systems errors, corrupted and broken links by using web using mining.

Krishnakant P. Adhiya et al., in [2], presented an approach to group web sessions. Web access sequences of web usage mining can used to find behavior or intent of set of users. Proposed technique carries, grouping of similar web user sessions. Grouping of web sessions is based on similarity and consists of maximizing the intra-group similarity while minimizing the inter-group similarity has achieved by using sequence alignment method. In this technique web session global and local alignment techniques of similarity measurement has considered. But sequence alignment method can further improved by using affine gaps in computing global alignment value.

Naga Lakshmi et al., in [3], discussed about data preprocessing technique because web is growing as a dominant platform for retrieving information and discovering knowledge from weblog file. Web usage analysis or click stream analysis is the process of extracting useful knowledge from database logs, user queries, cookies and user profiles in order to analyze web users' behavior. But web usage analysis depends on data abstraction of a weblog file.To resolve these issues proposed technique uses different formats of web server log files. Proposed preprocessing approach has focused on different formats. But preprocessing can be improved by removing all irrelevant multimedia files.

Krishnakant P. Adhiya et al., in [4], proposed an efficient sequential access patternmining algorithm(SAP), based on CSB-mine. Sequential access pattern mining algorithmshas used to discover interesting and frequent patterns from web data. Algorithm focuses onconstructing WAS list, US list, and generation of SAP table without using WAP trees atany phase. Algorithm does not require any extra data structure to find first appearance ofeach symbol, thus saving the space and time.

Priyanka Patel et al., in [5], presented a method to identify a web user session on a weblog file of NCSA format to find average time spent by user on pages. Proposed method considers time spent by user on web site. To cluster a web user session rough set clustering algorithm has used. Degree of page access and the initial time out is calculated for understanding users' navigation behavior. But average time spent by user on website has considered always.

Krishnakant P. Adhiya, Satish R. Kolhe in [6], presented a research work to consider users individual need for personalized web search. In this approach proposed the architecturefor web search personalization using web usage mining without user's explicit feedback. Anefficient sequential access pattern mining algorithm has based on CSB-mine algorithm usingthis new approach for sessionization and modified user profile has introduced. Result ofproposed technique has carried out on synthetic data. In approach users individualfeedback has considered.

Monika Dhandi et al., in [7], discussed about web personalization technique using FPgrowth algorithm. To produce better relevancy to user for the website designer should knowabout user behavior. Finding of user behavior, interest, statics of user is known as webpersonalization. In this technique it works on FP growth algorithms for locating frequentuses pages efficiently without candidate set generation, relative weighted rule for determiningrelative weight of every page respect to other pages in order to improve server reply fast andmarkov model used to store relative weight of page to their relative position.

DOI10.17148/IJARCCE.2017.6837

S. Uma Maheswari et al., in [8], proposeda technique enhanced user behavior(EUB) algorithm to recognize user's behavior on a website. The quality and accuracy of pattern mining has improved using better proposed preprocessing method. In the proposed method preprocessinghas performed on NCSA weblog file format. To Identify user navigation behavior andclusters similar kind of access user pattern using k-means clustering.

Xiaojing Li, Yanzhen Cheng et al., in [9],discussed that previous mining algorithm onlyconsider the users access frequency, without focusing on interest of users in their visitingpath. This will be motivation for the web site designer. In this approach clustering algorithm combines the Jacques ratio coefficient and the longest public path coefficient. Inthis proposed method focused is on to estimate user similarity of page interest and website accessstructures matrix is correct for the element value based on the three tuple model multiplication. Proposed method works on improved mining algorithm for preference and interestcalculation, the bad impact of mining has removed due to pages idle and links.

III. PROPOSED ARCHITECTURE

In structure of proposed architecture overall flow of algorithms has described. In this method NASA weblog has used in preprocessing. When user visits on website, his/her entry has created in the log file. From log file history, it is very easy to understand the user's navigation behavior. But in weblog file it also has irrelevant information. Therefore data cleaning is needed. After preprocessing remain two algorithms has applied. Figure 2 shows the architecture of proposed architecture. Proposed architecture uses three algorithms. These algorithms are as follows,
1) Modified Data Cleaning Algorithm
2) Proposed Modified Clustering Algorithm-I
3) Proposed Algorithm-II Based on preferred path mining

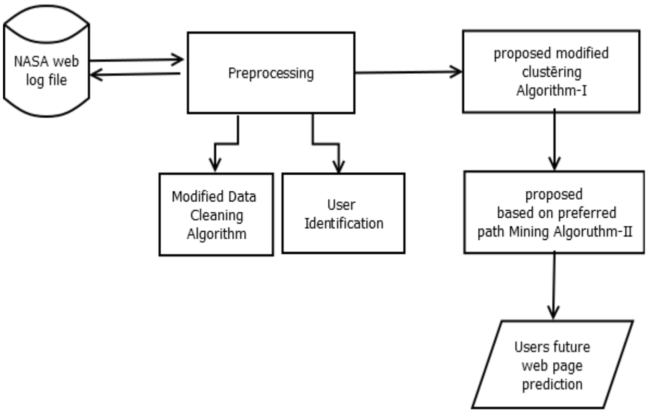
NASA web log file
Preprocessing
proposed modified clustēring Algorithm-I
Modified Data Cleaning Algorithm
User Identification
proposed based on preferred path Mining Algoruthm-II
Users future web page prediction

Figure2. Proposed Architecture for Understanding Users Navigation Behavior

Algorithm1:Modified Data Cleaning

```
Start
While(NOT end of file (LogFile)) do
    LogRecord = Read (LogFile)
    If LogRecord.Cs-url-stem <>gif, jpeg, jpg, xbm,wav,mp3,mp4,mpeg,mpg 314, 302,
    404, 403, 500, 204,502,501,400,315 then
        Add entry count into failures
        If LogRecord.Cs-method= GET then
            Add entry into text file
        End if
        If LogRecord.Sc-status = (200) then
            Add entry into text file
        End if
        If LogRecord.User-agent <>Crawler, Spider,Robot then
            Remove entry from weblog file
        End if
    End if
End while
End
```

DOI10.17148/IJARCCE.2017.6837

This algorithm removes all irrelevant entries from weblog file. By using this algorithm weblog file size is also reduces.

Proposed Modified Clustering Algorithm-I
Using weblog file it is possible to cluster same users. But same users have joined using two different sub-formulae. Those derived formulae are s' and s". Where s' is used to find the cluster of similar user access path but not common path of users.

S '= twice of intersection of users similar path/ Union of users access path

Therefore s" is required which calculates similar user access with the common path.
S"=Users Common access path/ users total access path.

By multiplying s' and s" outcome is in the form of clusters having common users withsimilar interest and similar user access path.

S= Multiplication of s' and s"

Algorithm2:Proposed Modified Clustering Algorithm-I

```
1.Similar path clustering c=ci and c= {φ}
Where C is similar path clustering constant
2. While NOT ends of file do {
3. Read content
4. While NOT end of file do {
5. For each entry of weblog file does
6. Calculate s
7. S= Multiplication of s' and s"
8. Endfor}}
9. If s >Ѳthen
10. {
11. Keep path as it is
12.}
13. Initialize Ci
14. If ci is not subset of set c then
15.{
16. Add ci into set c
17.}
18. End if
19. End if
20. End while
21.End while
22. Result is in C.
```

This algorithm produces outcome in the form clusters of users having same area of interest.

Algorithm3:Proposed Algorithm-II Based on Preferred Path Mining Algorithm

```
1. Set U, Ri.e. U=set of Users, R=set of requests
2.While(end of file) {
3. U={all users}
4. R={All Request}
5.While(end of file set U) {
6.While(end of file set R) {
7. Uci=count all request of each user}}
```

First find all unique users and requests from set U and Set R and count interest of eachuser. For each user sum of interest expressed is calculated i.e. coun. Then check sij>0 ornot if sij is positive then increment sij.

```
8. i=0
9.While(i<end of records in table) {
10. m=non-zero number of data records
11.Coun=0 where coun represents the sum of interest expressed in the each record.
12. j=0
```

```
13.While(j<end of records in table) {
14.If(Sij>0)
15. coun+=Supij
16. j++}
17. i++} (merge the same preferred path)
18. To Find time Session Duration 1 hr.
19.Date1 [] =array of date i.e. from 1 July to 30 July
20. While(i<8) {
21. If(date==date1 [j]) (select sum of rebyte,count of request)
22. j++}
23. To Find time Session Duration 3 hr.
24.While (i<8) {
25. If(date==date1 [j]) (select sum of Rebyte, Count of request)
26. j++}
27. While(end of table entries) {
28. While(end of table entries) {
29.Count=Count+ No request
30. If(count>3) {
Get all users, date,time,request,rebyte} }
To find date wise request hit
31. While(end of table Entries) {
(Select date and there count group by date)}
```

This algorithm performs path mining on the clusters found using algorithm2. Outcome of algorithm3 is in the form of prediction of web pages in future each interested user may visit.

IV. RESULT AND DISCUSSION

A. Results of Preprocessing Algorithm

In this paper we have analyzed NASA web server log file of size 200,432 MB, various analysis has been carried out to identify the user interest.Table1 shows Status Codes of Hypertext Transfer Protocol according to Table1 errors and irrelevant entries which arise in web surfing were removed from web server log file.

Table1. Status Codes of Hypertext Transfer Protocol [1]

200	**OK (Relevant Data)**	**401**	**Unauthorized Access**
205	Reset Content of Data	402	Required Pay to access
206	Partial Content of Data	403	Forbidden
300	Multiple Choices to Select	404	Data Not Found
301	Moved Data Permanently	405	Method Not Allowed
302	Moved Data Temporarily	500	Server Error
303	See Other Options	501	Not Implemented
304	Not Modified	502	Bad Gateway
305	Use Proxy	503	Out of Resources
400	Bad Request	504	Gateway Time-Out

Table2. Results of Preprocessed data

Server Log File	NASA Jul-95
Duration	1 - 15days
Original Length	119.00 MB
Reduced length After Preprocessing	33.00 MB
Percentage in Reduction	72.00
Total No. of Unique Users	19156

In this paper NASA weblog of year 1995 has used. Preprocessing is performed for two week of July 1995. After preprocessing reduction of original weblog file size has shown in Table 2.

By observing the Table3 the system administrators and website designer can able to understand the how many web log entries was there and how much was failure with number of request hit on particular day. Also minimum and maximum traffic time with bytes transfers on that time has achieved. Most busy day, least busy day, number of request hits on most busy day as well as least busy day. Count of unique user on particular day also has achieved.

Table3. User profile

Day	Total no. of web log entries	Total no. of preprocessed web log entries	Removed Entries in preprocessing	No. of unique users	No. of hits	Maximum traffic		Minimum traffic	
						Duration	No of bytes transferred	Duration	No of bytes transferred
1	64714	43693	21021	4620	19939	17:00:00 To 18:00:00	14184699	07:00:00 To 08:00:00	4951045
2	54959	36596	18363	3980	17579	17:00:00 To 18:00:00	11100635	04:00:00 To 05:00:00	4149058
3	89584	60183	29401	6460	28217	13:00:00 To 14:00:00	19064116	04:00:00 To 05:00:00	8214328
4	70452	46794	23658	4875	22717	15:00:00 To 16:00:00	14568900	06:00:00 To 07:00:00	8484254
5	94575	64086	30489	6630	29249	14:00:00 To 15:00:00	23342798	04:00:00 To 05:00:00	5459217
6	100960	68680	32280	6991	30811	14:00:00 To 15:00:00	23147335	04:00:00 To 05:00:00	5535723
7	87162	59033	28129	6998	26820	14:00:00 To 15:00:00	23447335	04:00:00 To 05:00:00	5535723
8	38867	25798	13069	2553	12367	14:00:00 To 15:00:00	8490867	04:00:00 To 05:00:00	3039572
9	35272	23319	11953	2243	11328	20:00:00 To 21:00:00	10710079	05:00:00 To 06:00:00	1609813
10	72860	47885	24975	4016	23754	16:00:00 To 17:00:00	26177066	04:00:00 To 05:00:00	2786143
11	80407	53266	27141	4443	25842	14:00:00 To 15:00:00	19500726	04:00:00 To 05:00:00	4978977
12	92536	61808	30728	4858	29539	14:00:00 To 15:00:00	20813591	04:00:00 To 05:00:00	4165028
13	134203	87535	46668	6319	45122	09:00:00 To 10:00:00	40429965	19:00:00 To 20:00:00	9474993
14	9673	6360	3313	600	3165	00:00:00 To 01:00:00	10668120	04:00:00 To 05:00:00	1735455
15	65392	37456	23975	3570	2275	14:00:00 To 15:00:00	15813591	04:00:00 To 05:00:00	3065028

B. Results Analysis of Proposed Algorithm-I and Proposed Algorithm-II

In the process of experiment analysis Proposed Algorithm-I and Proposed Algorithm-II has combined to understand users' navigation behavior and to predict users' future web page request. After preprocessing NASA weblog file has used in the four weblog set of July 1995 from 1 to 31 July. Each weblog set is of 7 days. Set 1 is from 1 to 7 July, set 2 is from 8 to 15 July, set 3 is from 16 to 23 July and set 4 is from 24 to 31 July. Both proposed algorithm-I and algorithm-II has applied on four weblog sets and their results are compared with Improved Browsing Pattern Mining algorithm (IBPM)[9].

In this analysis proposed algorithm and Improved Browsing Pattern Mining algorithm (IBPM) are applied on four weblog sets and their results of both techniques are compared with each other as shown in figure3. Comparison is based on numbers of unique users found in each weblog set using both techniques. Figure3 shows relevant number of unique users using proposed algorithm.

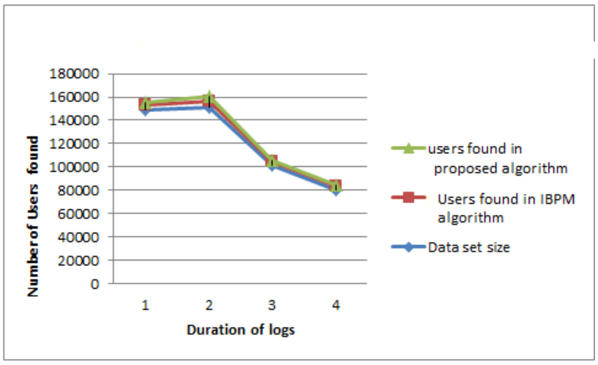

Figure3. Number of Unique Users Found

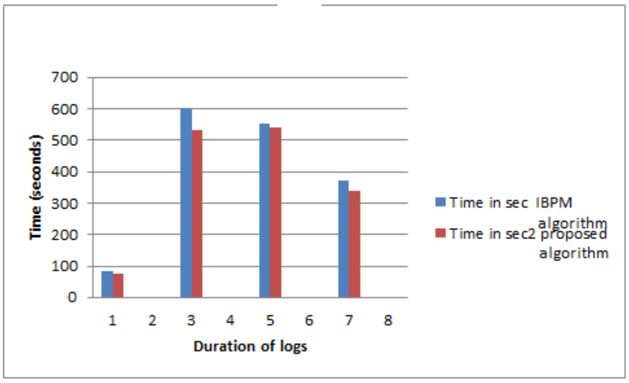

Figure4. Execution Time Comparisons

Another parameter for comparison of both techniques is execution time required for both techniques to run algorithms as shown in Figure4 As we know each web log sets is of 7 days so results of each weblog set for each 7 day set has achieved. Time required in IBPM algorithm has more than proposed algorithm.

One more parameter of comparison is in the form of memory required for each technique to run algorithm. Comparison of memory requirement has shown in Figure5.According to Figure5 memory required in proposed techniques is less than the IBPM technique.

According to results analysis of proposed algorithm it produces more efficient unique users and prediction of future web pages is more accurate also requiresless time andless memory.

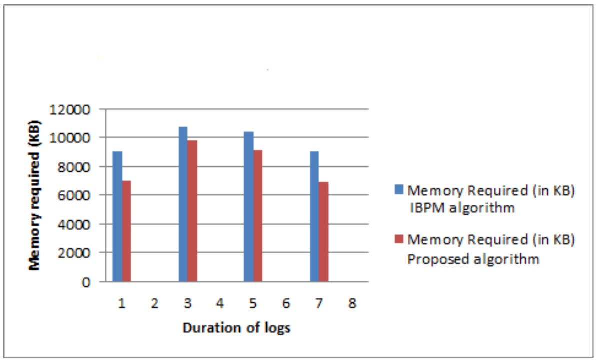

Figure5. Memory Usage Comparisons

V. CONCLUSION AND FUTURE SCOPE

The Proposed architecture focuses on group of the frequently accessed patterns of interested users. It provides better suggestions to website designers to improve performance of the web by giving preference to the patterns accessed by

DOI10.17148/IJARCCE.2017.6837

the interested users. Modified data cleaning algorithm removes all multimedia files and other irrelevant information from weblog file. Proposed modified clustering algorithm-I produces better clustering user groups based on their similarity. Using proposed algorithm-II based on preferred path mining algorithm future prediction of next web pages for user is better than previous existing algorithm. Result of proposed architecture has improved as compared to IBPM algorithm on basis of more accuracy and efficiency. Proposed architecture preserves 6-8% memory as compare to IBPM algorithm. Also in time concern 4-6% accessing time has saved in proposed method.

In future understanding users navigation behavior will be perceives in terms of active user session. Proposed solution analyses user profiles only from his/her weblog history. In future online user navigation behavior will be better exposure in remuneration.

REFERENCES

1. k. Suneetha and R. Krishnamoorthi, “Identifying user behavior by analyzing web server access log file," International Journal of Computer Science and Network Security, vol. 9, pp. 327-332, April 2009.
2. B. S. Chordia and K. P. Adhiya, “Grouping web access sequences using sequence alignment method," Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, Jun-Jul 2011.
3. N. Lakshmi and R. S. Rao, “An overview of preprocessing on web log data for web usage analysis," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 2, p. 329-337.
4. K. P. Adhiya and S. R. Kolhe, “An efficient and novel approach for sequential access pattern mining," Journal of Emerging Technologies in Web Intelligence, vol. 7, pp. 26-30, November 2015.
5. P. Patel and M. Parmar, “Improve heuristics for user session identification through web server log in web usage mining," (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 5, pp. 3562-3565, 2014.
6. K. P. Adhiya and S. R. Kolhe, “An efficient and novel approach for web search personalization using web usage mining," Journal of Theoretical and Applied Information Technology, vol. 73, 20th March 2015.
7. M. Dhandi and R. K. Chakrawarti, “A web personalization technique for recommendation," International Journal of Technology Research and Management, vol. 3, pp. 2348-9006, October 2016.
8. S. U. Maheswari and S. K. Srivatsa, “An application of preprocessing and clustering in web log mining," Int. journal of philosophies in computer science, vol. 1, pp. 21-30, January 2015.
9. X. Li and Y. Cheng, “The improved clustering algorithm for mining users preferred browsing paths," Springer International Publishing Switzerland, pp. 274-279, March, 2013.
10. Sana M. Deshmukh, Krishnakant P. Adhiya,”A Review on Finding Users Navigation Behavior Using Web Mining Algorithm”, 2016 IJSRSET | Volume 2 | Issue 6 | Print ISSN. 2395-1990 | Online ISSN . 2394-4099.

DOI10.17148/IJARCCE.2017.6837