A Systematic Review On Mining Web Navigation Pattern Using Graph Based Techniques

Hemraj K. Varachhia, Ankur N. Shah

Abstract: Today internet is widely used in every aspect of day to day life. People get more and more dependent on internet for getting any kind of information. Record of different web user's web using pattern are get stored in web log repository, which are great source of knowledge about user's navigation. With increasing the use of internet, number of web sites and web pages are increasing rapidly. So analyze and discovering user's interesting patterns are necessary for web administrator and recommendation system. Web usage mining is a part of data mining. In this, mining techniques are applied to web data for finding user's interesting patterns. That means in which patterns user want to access web pages and web-sites. In this review paper I reviewed several pattern reorganization techniques such as graph traversal algorithm, efficient sequential pattern mining algorithm, mining a web usage data mining graph. From this entire techniques graph traversal algorithm is best.

Index Terms: web log repository, filtered data, recommendation, web usage graph, Apriori & Eclat, TSPs, via-link

1 Introduction

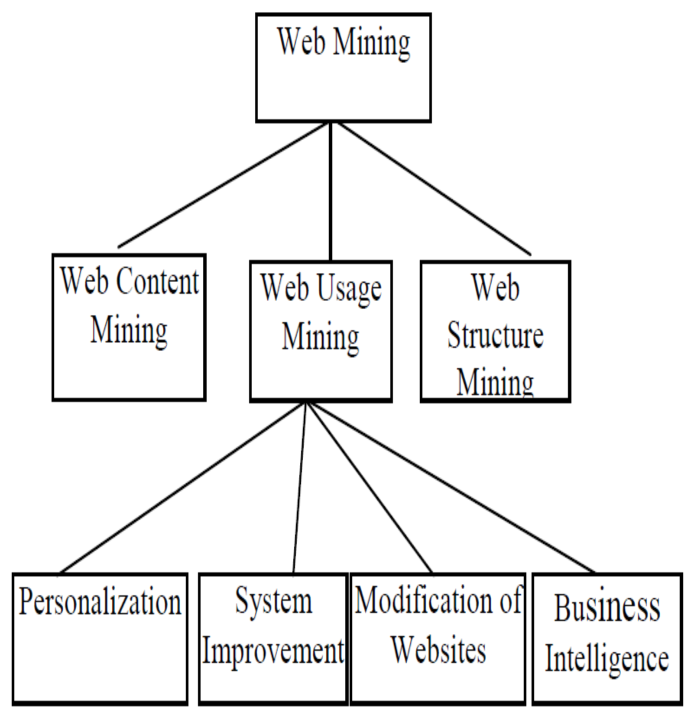


Figure 1: Web mining categories [2]

Discover the hidden data from large set of database is called data mining. We get accurate data if we extract it from last five years data from database. In database huge amount of data are available, from all of available data user interested data extracted is the process of data mining. Data mining techniques are applied on web data is called web mining. Web mining is useful for websites developer or administrator to modify its contents and structure, etc. Web mining contains web content mining, web usage mining and web structure mining. Web structure and content mining related with primary data of web, while usage mining is related with secondary data, it contain interaction of user with the web. Web log repository or web logs contain huge amount of data about interaction of user with the web. It is main source of knowledge for discovering interested pattern. Web usage mining (WUM) is a process to discover hidden interesting and usage pattern from the web data. It is useful to get understand user's interest and better to serve them that type of data. Web usage mining is useful to understand customer's profile and strength and weakness of web. By using this developer can forecast user motivation and after that provide better recommendation for customers. There some methods are available based on tree structure for mining useful patterns like FP-growth, FTP. But they all are suffering from repeatedly scan of database and extra space requirement to store those copies. In this paper I reviewed some graph based techniques which explain below.

- *Hemraj Varachhia is currently pursuing masters degree program in Computer science and engineering in Gujarat Technical University, India, PH-09426333311.E-mail: hamraj_793@gmail.com*
- *Ankur Shah is currently working as Assistant Professor in Computer Science & Engineering Department in Parul Institute of Technology, Gujarat, India.*
- *PH: 8511152907. E-mail: ankur11586@gmail.com*

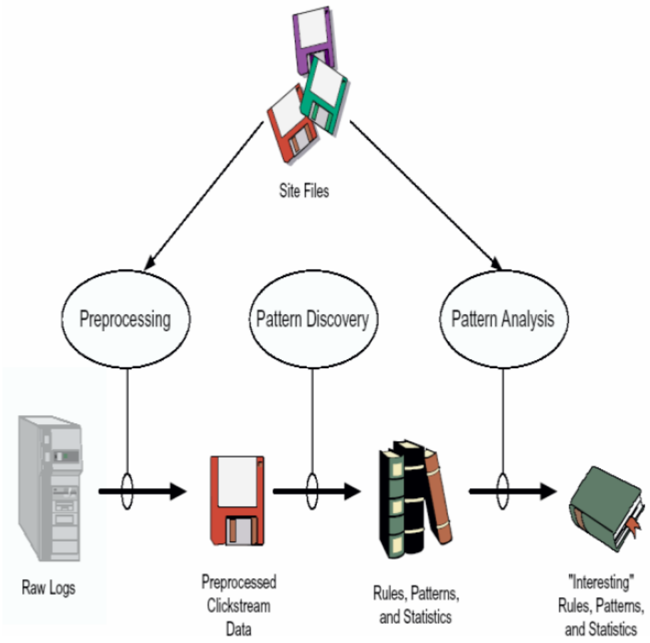
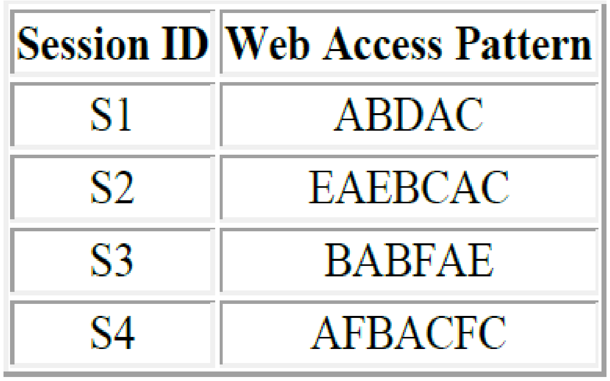

Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure 2: Process of Web Usage Mining [6]

2 Related work

In [Yao-Te Wang, Anthony J.T.Lee][1], this paper author describes the concept of TSP and via-link. Author also explains cyclic TSP, acyclic TSP or partial cyclic TSP. From the available data initial path traversal graph is created. In this graph which vertex connected directly with edge and via-link are displayed. Concept of Via-link is that "from-to-via" link. After creating initial path traversal graph irrelevant data are deleted from that graph. Vertex and edges whose count is less then minimum support count called irrelevant data. After removing it frequent path traversal graph is generated. Form this as per DFS each vertex is selected with its number of via-link and push it into Untracedstack one-by-one. After this step last vertex of via-link is pop from Untracedstack and push into Iteration. Each vertex pop from Untracedstack is replaced with available vertex with whom it connected. Iteration is used for creating TSP. Each Iteration creates different TSPs. It is frequent patterns which get created from dataset. In [Dlilip Singh Sisodia, Shrish Verma][2], this paper author describes the concept web usage mining and review the process of discovering useful patterns from web server log files of academic institute. Author explains the process of web log mining. Author also explain the common web logs contents which need to be analyzed like Access log, Error log, Browser log, Referred log etc. Author also displays different type of log file format. And they show the result of log which contains number of users access a system with different unique IP, visitors, hits, countries, etc. In [Dheeraj Kumar Singh, Varsha Sharma, Sanjeev Sharma][3], this paper author describes the new approach for mining the web usage data by creating graph and generate useful sequential access pattern. In this author takes log data as input and get filtered data by removing multimedia and irrelevant data. By using filtered data web usage graph is created. After it pruned graph is created as per requirement. And from pruned graph frequent access patterns is created according to their length. Thus they introduce a new way of web usage mining. And they showed how to to discover useful access pattern with using graph data structure to accomplished complex user's browsing behavior. Following is the example discussed in the paper



Session ID	Web Access Pattern
S1	ABDAC
S2	EAEBCAC
S3	BABFAE
S4	AFBACFC

Figure 3: Web Access Sequence Table

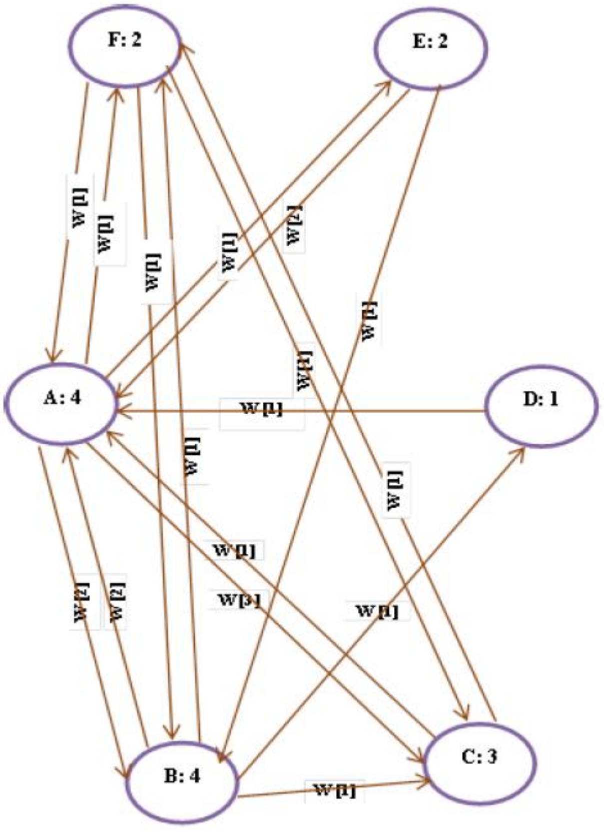


Figure 4: Web Usage Graph

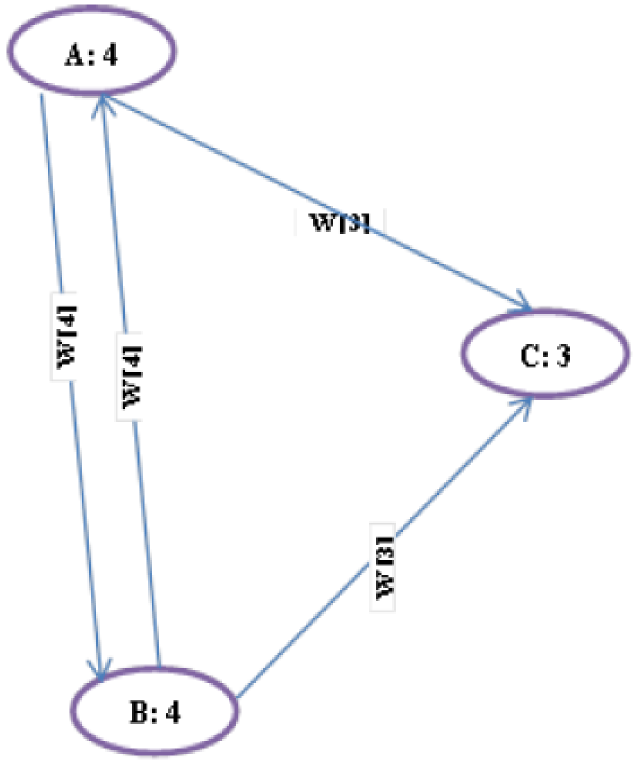
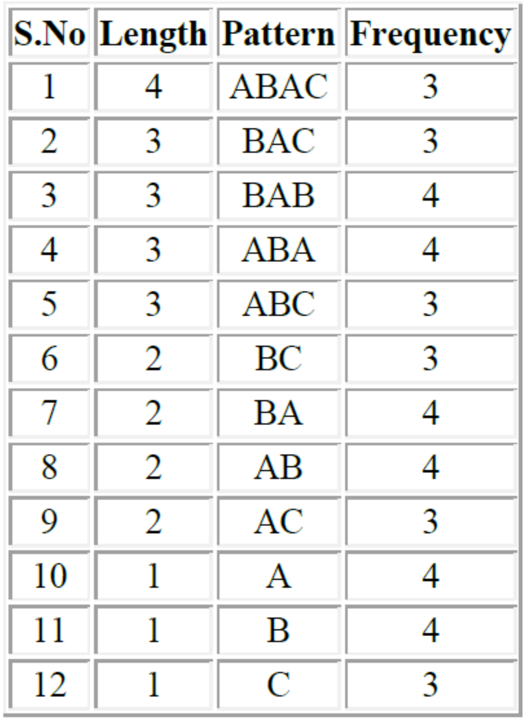


Figure 5: Pruned Graph



S.No	Length	Pattern	Frequency
1	4	ABAC	3
2	3	BAC	3
3	3	BAB	4
4	3	ABA	4
5	3	ABC	3
6	2	BC	3
7	2	BA	4
8	2	AB	4
9	2	AC	3
10	1	A	4
11	1	B	4
12	1	C	3

Figure 6: List of Frequent Sequential Access Pattern

In [Bina Kotiyal, et. al.][4], this paper author describes the concept of two sequential pattern mining algorithm namely Apriori and Eclat. Author explain that algorithms for predicting user's behavior. They also compare performance of both algorithm in terms of time and space. They applied these algorithms on filtered data which comes from storage and gives analyzed patterns as an output. They apply theses algorithms on same data and display the result. Which shows that Eclat algorithm is good in compare of Apriori. Because in Eclat algorithm no need to level wise search of database and no need to multiple scan of database. In Eclat no need to extra computation overhead and this algorithm is easily work with huge amount of data. So Eclat is better to use in compare of Apriori. In [Mirghani A. Eltahir, Anour F.A Dafa-Alla][5], this paper author describes the process of getting knowledge from web server log files where all users' history of navigation is registered. In this paper they explain different type of logs from where data get be collected. They describe whole life cycle of web usage mining which contain data collection, preprocessing, pattern discovery and analysis. They consider different log file format on the basis of number of hits, number of visitors, browser type, cookies, platform, etc. They analyze the log of "www.interactivegt.com" and show the result according to user information registered in log files, such as top visited pages, popular paths through sites, top search engine, errors, etc. In [Manisha Valera, Uttam Chauhan][6], this paper author describes the concept of efficient sequential pattern mining algorithm to find frequent sequential web access pattern. In that algorithm from web access behavior of user web usage graph is created and then this graph is used to find out frequent sequential web access patterns. After that on the basis of that web recommendation system is created. In this algorithm preprocessing is done on input web server log and get list of different web access patterns visited in different patterns and after that web usage graph is created after that applying mining technique to mine useful sequential access pattern. And using this pattern web recommendation system is getting created.

3 PROPOSED METHOD

In previous section we discussed many graph based techniques to mine the frequent sequential access pattern. But those techniques are time consuming to generate frequent pattern. I am going to overcome this problem by applying my new technique, which I named advanced graph based technique. Flowchart of this technique is as below.

4 Flowchart

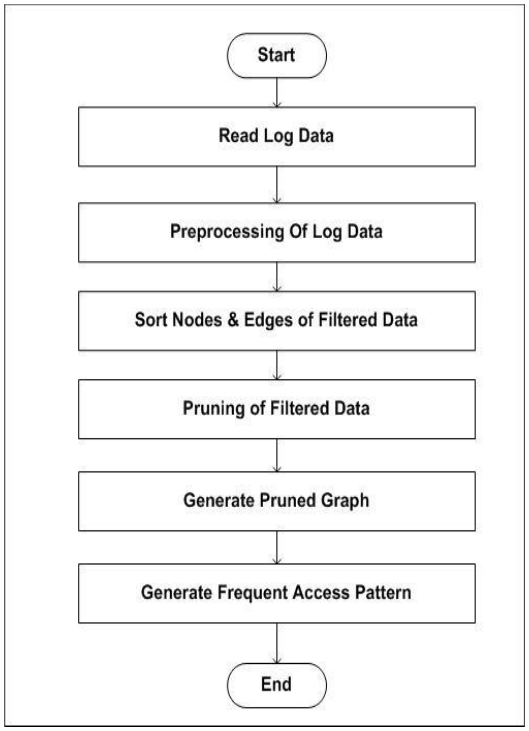

Start
Read Log Data
Preprocessing Of Log Data
Sort Nodes & Edges of Filtered Data
Pruning of Filtered Data
Generate Pruned Graph
Generate Frequent Access Pattern
End

Figure 7: Flow chart of Proposed Work

5 Conclusion

By applying advance graph based technique problem of time complexity that appears in graph based techniques can be easily solved. Furthermore by using this new technique administrator or developer can easily predict the navigation behavior of user which will help modifications of web sites or web pages.

References

[1] Yao-Te Wang, Anthony J.T.Lee, “Mining Web navigation patterns with a path traversal graph”, ELSEVIER(2011), pages (7112-7122)

[2] Dilip Sing Sisodia, Shrish Verma. “Web Usage Pattern Analysis Through Web Logs: A Review”, IEEE(2012), pages(49-53)

[3] Dheeraj Kumar Singh, Varsha Sharma, Sanjeev Sharam, “Graph based Approach for Mining Frequent Sequential Access Patterns of Web pages”, IJCA(2012) pages (33-37)

[4] Bina Kotiyal, et. al., “User Behavior Analysis in Web Log through Comparative Study of Eclat and Apriori” IEEE(2012) pages (421-426)

[5] Mirghani A. Eltahir, Anour F. A. Dafa-Alla, “Extracting Knowledge from Web Server Logs Using Web Usage Mining” IEEE(2013), pages(413-417)

[6] Mnaish Valera, Uttam Chuhan, “An Efficient Web Recommender System based on Approach of Mining Frequent Sequential Pattern from Customized Web Log Preprocessing”, IEEE(2013)