TRACING EFFICIENT PATH USING WEB PATH TRACING

L.K. Joshila Grace

Sathyabama University, Department of Computer Science and Engineering, Jeppiaar Nagar,
Chennai, Tamil Nadu 600 119, India
joshilagracejebin@gmail.com

V. Maheswari

Sathyabama University, Department of Computer Applications, Jeppiaar Nagar,
Chennai, Tamil Nadu 600 119, India

Abstract

In the fast improving society, people depend on online purchase of goods than spending time physically. So there are lots of resources emerged for this online buying and selling of materials. Efficient and attractive web sites would be the best to sell the goods to people. To know whether a web site is reaching the mind of the customers or not, a high speed analysis is done periodically by the web developers. This works helps for the web site developers in knowing the weaker and stronger section of their web site. Parameters like frequency and utility are used for quantitative and qualitative analysis respectively. Addition to this down loads, book marks and the like/dislike of the particular web site is also considered. A new web path trace tree structure is implemented. A mathematical implementation is done to predict the efficient pattern used by the web site visitors.

***Keywords*: Qualitative analysis; Quantitative analysis; web path trace.**

1. Introduction

Web mining is a part of data mining .Web mining is broadly classified as web structure mining, web content mining and web usage mining [1]. Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web Structure Mining is the process of discovering structure information from the Web. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data [7]. These mining procedures are used mainly for improvising the web site. Similar to this new technique of mining procedures is being used to identify the quality of a web site. If the user of the web site has used the particular web page for a long duration then it is concluded that the web page has been used qualitatively. On the other way the web page has been visited by different users a number of times then it can be concluded as the web page has been used quantitatively. So there is a necessity of qualitative and quantitative analysis of the web site to find the efficiency of web site.

There are various constraints that have to be noticed each time a qualitative analysis and the quantitative analysis is being done. The main constraint in the qualitative analysis is done. The user would have opened a web page and may not have worked with the page. In that case the time duration that is being calculated will be including the idle time of the user. So to monitor this idle time a constraint must be maintained.

In the case of the quantitative analysis the visit made by the user would be for few minutes or seconds. This visit may or may not have a great impact on the quantitative parameter. The user would have found this as the intermediate web page of the target web page. This particular information must be noticed in deep for a detailed analysis of the web site.

In this work the constraints that are discussed above are analyzed for a single web site. This also can be applied for any web site for which the analysis has to be done. The basic requirement of the analysis is the log details of the web site. The log file is the location where all the visitor information of the web site is being recorded. The log file contains raw data about the web site visitors these data has to be converted in a readable format which would be the input of the system.

Patterns are generated to know path traced by the user during the visit to the web site. These patterns are of two type continuous patterns and discontinuous patterns. The consecutive web pages being visited by the user is called as continuous patterns. Example the visitor moves A → B → C →D then they are called as sequential patterns. Where A, B, C, D are web pages. The non consecutive web pages being visited by the user are called as discontinuous patterns. A → D → F and A → C → A are examples for discontinuous patterns. The discontinuous patterns will also include the back tracking of web pages.

If the entries in the log file are more, then the web site can be assumed to reach more number of visitors. This type of web site can be even more scrutinized to find the efficiency of the web path of the web site. There are several traversal techniques and prediction techniques used widely only for finding the efficiency. This efficiency factor helps the web developers to improvise the efficiency of their weaker section of the web site. This type of analysis helps in improving the e-commerce.

Typically, an admired website may register hundreds of megabytes of web log records every day that offers rich information about web dynamics. From the web log databases, the repeated sequential web accessing patterns were determined by path traversal pattern mining. However, it fails to reflect the different impacts of different Web pages to different users. In internet information service applications, the variation made on web pages makes a strong impact on decision making. These log files are extracted by the web developers for tracing the type of viewers visiting the web site. There various factors that are being recorded including the IP address of the user using the system. Therefore any type of analysis to be made in the web site the Log files has to be extracted by converting the raw data to a readable form and then the necessary processing is done. This conversion process is carried out are called as pre processing.

2. Related work

1.1. Preferred Navigation Tree

Several techniques are used based on either frequency or utility for traversal patterns mining. The HITS (hypertext induced topic selection) and PNT(preferred navigation tree) concepts are together used by Jieh-Shan Yeh et al. [10] .HITS are used for ranking the web page. User preferred patterns are being calculated by using the PNT. Combining these two concepts a new concept is being introduced as PNTH (preferred navigation tree with HITS) algorithm which is an extension of PNT. This is one of the algorithm used the concept of PNT and considered the relationship among web pages using HITS algorithm. Their algorithm was suitable for e-commerce applications like improving web site design and web server performance.

1.2 Throughout-Surfing Patterns

A new technique called throughout-surfing patterns (TSPs) was introduced by Yao-Te Wang and Anthony J.T. Leeb [2]. An efficient method to mine the web path traversal patterns. Path traversal graphs are generated and are used for understanding the web site visitors aim in surfing the web site [8]. The path is been split to a length of three as maximum length. There is a necessity create more number of intermediate trees in this technique. More time and resources are used in this method.

1.3 High Utility Path Traversal Mining

Many utility based algorithms are being used. These are very much helpful in finding the importance of the web page. One such algorithm which is being used to find the qualified web site was introduced by Lin Zhou et al. [4] They implemented their 'high utility path traversal mining' algorithm on weblog database and compared it with the high utility path traversal patterns with the frequent traversal patterns by traditional path traversal technique. There are two types of tree techniques proposed [6][9] namely utility based web access sequence (UWAS) tree and incremental UWAS (IUWAS) tree to mine the web path patterns. It efficiently handles both forward and backward web access sequences. It can work on both static and incremental web access sequences. It needs three database scans for the mining process. It also avoids level wise candidate generation

1.4 Markov Chain

The prediction of the next web site that the user would click on is given by the method proposed by V.ValliMayil [5]. The prediction process is done with the help of the log data of the web site and the concept of markov chain. They concentrate on high probability trials of the user who have already used the web site. Since working on probability they need a high amount of data to be considered. The markov chain is even more reformatted by using a depth first search. This result in the web prediction for the user is done the previous users probability result.

1.5 DSM-PLW

A projection-based, single-pass algorithm, called DSM-PLW (Data Stream Mining for Path traversal patterns in a Landmark Window) [3], for online incremental mining of path traversal patterns over a continuous stream of maximal forward references generated at a rapid rate. Each maximal forward reference of the stream is projected into a set of reference-suffix maximal forward references, and these reference-suffix maximal forward references are inserted into a new in-memory summary data structure, called SP-forest (Summary Path traversal pattern forest), which is an extended prefix tree-based data structure for storing essential information about frequent reference sequences of the stream.

3. Proposed Work

The log file data of a web site is being considered and they are converted to a machine readable format. Only the essential details are being considered out of the entire set of details. A set of pre processing is done on the log data. Then a web path tracing tree (WPT) is generated with the extracted details as shown in the table 2. The web path tracing tree is generated for both the sequential patterns and the non sequential patterns. This tree is similar to a prefix tree format but with additional parameters.

For a particular web site to be analyzed a pattern has to be generated from the log file details. This pattern is analyzed for effectively predicting the quality of each and every part of the web site. The qualitative analysis and the quantitative analysis is done taking up these factors time, frequency, down loads, book marks and like/dislike. Here both continuous and discontinuous patterns are used for analysis. The continuous patterns are known to the developer itself where as the discontinuous patterns are the one which is new for the developer to analyze. The projected database concepts are used here for maintaining the patterns.

Table 1. Generated for individual user

User	Patterns	Time (sec.)	Frequency	Downloads	Book Marks	Like/dislike
U1	a,b,c,d	320	2	0	1	1
U2	b,c,f,d,e	657	1	1	1	0
U3	a,d,e	558	1	1	0	1

The table1. shows the values for individual user. The user is being named as U1,U2,U3,.....Un. Each user is not being saved as user by itself instead the address of the system is noted and assuming that a single user would have used the web page for not more 30 minutes if it is idle without any records of any operation. The next move made by the user on the same system will be considered as the next user. Even though the user is the same user. This is helpful in the other way that many user would use the same system. So the user are just considered as the visitors who come at that time. There is no records of the user is being maintained which is not useful for the work.

The time is calculated in seconds denotes the total time taken by the individual user traversing through the path. Frequency is calculated from the number of times the same path is being viewed by each individual web site visitors in the single visit. Since no history of the users are maintained. The rest of the factors like the Downloads, Book marks, Like/dislike are given in binary values. If the action is performed it is given is given the value 1else given as 0. Downloads shows the downloads done by the particular visitor whenever the particular web path is being traversed. Book mark option shows the book marks done whenever the particular web path is being traversed by this user. The last option Like/Dislike is given as a Boolean value. If the likes option is clicked by the user then it is represented as 1 otherwise it is given as zero. The draw back behind this is many users may not have clicked any option. In that case they are not considered as like or dislike.

Table1. provides details of individual user and the paths traverse by them through the web site, from which Table 2. is extracted. This corresponds to individual patterns as shown below.

Table 2. Individual patterns

Patterns	Time (sec.)	Frequency[F]	Downloads[D]	Book Marks[B]	Like/dislike[DL]
a,b,c,d	1900	5	20	50	1
b,c,f,d,e	1300	4	16	43	1
a,d,e	2400	8	19	32	0

The Table 2. shows the values for various factors respective of the path traversed by the web site users. The time is calculated in seconds that denote the total time taken by various users. Frequency is calculated from the number of times the same path is being viewed by all the web site visitors. Downloads shows the total number of downloads done whenever the particular web path is being traversed by various user. Book mark option shows the total number of book marks done whenever the particular web path is being traversed by various users. The last option Like/Dislike is given as a Boolean value. If the number likes is greater than the number of dislikes then the value is given as 1 otherwise it is given as zero. It is not considering about the visitor who did not click any option.

When a web site is being created by a developer the sequential web path available for the web site is known. Since they are fixed by the developers them self. But non sequential path may not exactly fall under any of the pattern that is already known to the developer. The back tracking of the web page will also come under the non sequential patterns. With the available known web path the mining process can be done by finding the match. Out of the entire set of data random data is selected of mining. Then Mining the efficiency of each web path by using the equations given below

$$EF = TV + TT + TD + TB / 100 \qquad (1)$$

$$EPATH = (TP * LN)/TW \qquad (2)$$

The efficiency factor of the particular web path is given by the term EF Eq. (2). TV represents the total views for the web path, TT represents the total time utilized by all the web viewers who accessed the path, TD gives the total down loads done by the web site visitors of the particular path and TB gives Total Book marks done while surfing the path.

The efficiency of the particular length of pattern selected are been given by EPATH Eq. (1). EF gives the efficiency factor calculated for the path, LN represents the length constraint applied and TW gives total number of web pages available. The lengths of the patterns are being varied and the matches are found for each of the variation. The other constraint being applied is with the amount of data being considered. Instead of considering the entire set of data a random pick of certain percentage of data is made. The equation is being applied for each case and compared with the other results. This entire processing is applied only for sequential patterns although the tree is generated for both sequential and non sequential patterns. Since extra effort needed for an efficient result of non sequential patterns only sequential is considered for discussion.

4. Results and Discussions

There are many algorithms being developed for the web traversal mining, finding efficiency, predicting the users intention etc., They all indirectly contribute for the web developers to improvise the web site. When the web site is being developed it attracts more number of customers to view the web site. The concept is the same as it is seen in many previous works. But implementation is done entirely in a new way starting from the tree generation till the finding of efficient path. Therefore the comparison of the result is done with the prefix span approach and found significant amount of variation in the result.

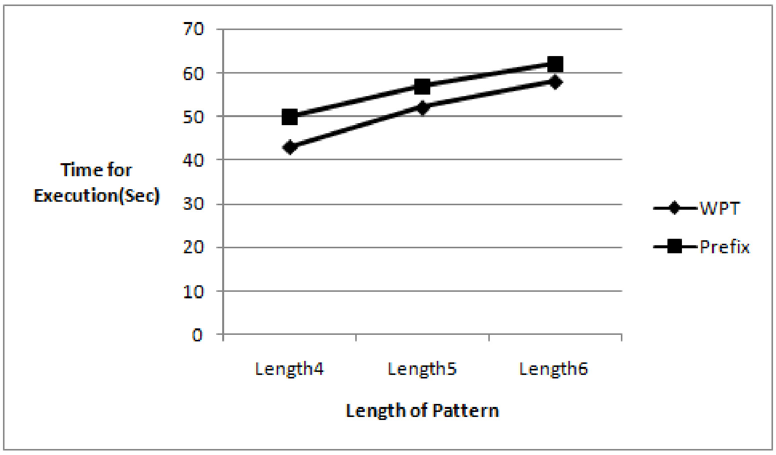

Fig. 1. Execution time corresponding to the length of pattern

The Fig. 1 shows a comparison done with the existing Prefix span tree corresponding to the time of execution. The constraints followed here is the length of the pattern. The Length4 means considering the pattern of size 4. Example A → B → C → D. Similarly it is applicable for the rest of the lengths shown in the fig1. The graph shows a considerable amount of efficiency in the time taken for execution.

The next comparison is made with the same prefix tree methodology varying the amount data set being considered for mining. The data considered are randomly picked from the entire set of data. These are shown in

Fig. 2. as D20%, D30% and D40%. This denotes that twenty percent of data is considered out of the whole set of data available. This is represented as D20%. Similar assumption is applicable for the other data set extracted.

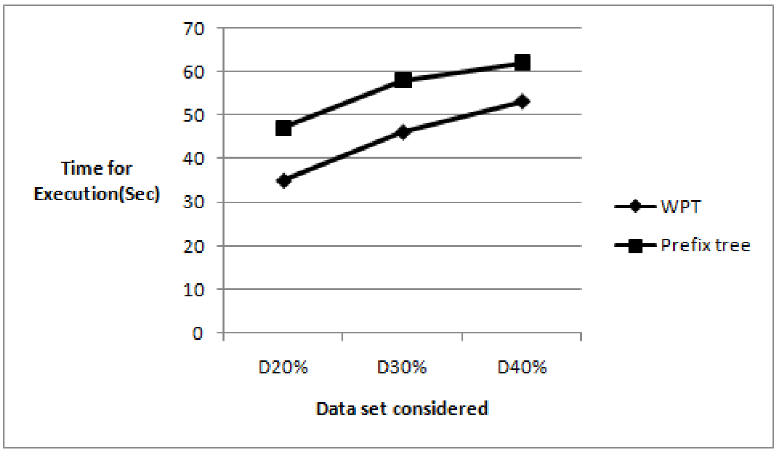

Fig. 2. Execution time corresponding to the data set

These results are discussed only for sequential patterns and not for the non sequential patterns. Even though web path trace (WPT) tree is created for both sequential and non sequential patterns only the sequential patterns are considered for further calculation of the efficiency.

5. Conclusion

The web path traverse tree is developed for both sequential and non sequential patterns**.** The efficiency of the web site is obtained only for sequential patterns. Sequential patterns are known to the developer the pattern is mined in considerable less time Additional to the frequency factor , utility, book marks, downloads etc., are considered. Considering all this factors the efficiency is calculated. Thus helps the web site developer to analyze his own web site, improve it in the necessary areas and increase the satisfaction of the customers.

References

[1] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, (2005), *Web Mining - Concepts, Applicaions & Research Directions*, Foundations and Advances in Data Mining Studies in Fuzziness and Soft Computing, Volume 180, pp 275-307.

[2] Yao-Te Wanga, Anthony J.T. Lee,(2011), *Mining Web navigation patterns with a path traversal graph*, Expert Systems with Applications, vol. 38, pp. 7112–7122.

[3] Hua-Fu Li, Suh-Yin Lee, Man-Kwan Shan, (2006), *DSM-PLW: Single-pass mining of path traversal patterns over streaming Web click-sequences*, Computer Networks, vol. 50, pp.1474–1487.

[4] Zhou L., Liu Y., Wang J., Shi Y.(2007), *Utility-Based Web Path Traversal Pattern Mining*. In Proceedings 7th International Conference on Data Mining Workshops, pp. 373-378.

[5] V.ValliMayil,(2012), *Web Navigation Path Pattern Prediction using First Order Markov Model and Depth first Evaluation*, International Journal of Computer Applications, Vol.45, no.16.

[6] Ahmed, C.F.,Tanbeer, S.K, Byeong-Soo Jeong, (2010), *Mining High Utility Web Access Sequences in Dynamic Web Log Data*, 11th ACIS International Conference on Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD).

[7] Kosala and Blockeel, (2000), *Web mining research: A survey*, ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1– 15.

[8] Istvan K. Nagy and Csaba Gaspar-Papanek, (2009), *User Behaviour Analysis Based on Time Spent on Web Pages*, Web Mining Applications in E-commerce and E-services, pp. 117-136.

[9] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong Soo Jeong,(2011), *A frame work for mining high utility web access sequences*, IEEE Technical Review, vol 28, issue I, pp 3-16.

[10] Jieh-Shan Yeh ,Ying-Lin Lin,Yu-Cheng Chen, (2009), *Mining Preferred Traversal Paths with HITS*, Proceedings of the International Conference on Web Information Systems and Mining,pp.98- 107.