ENHANCED PREFIX TREE MINING: SEQUENTIAL AND NON SEQUENTIAL PATTERNS

Joshila Grace.L.K. [1], Maheswari.V [2]
[1]Research Scholar, Department of Computer Science and Engineering
[2]Professor and Head, Department of Computer Applications
[1,2]Sathyabama University, Chennai, India
Email: [1]joshilagracejebin@gmail.com

Abstract

The entire work focuses on the development of E-commerce which in turn increases the business of the individual or a group. In this fast moving environment most of the necessities are satisfied by online. Thus the web site that reaches the viewer in an efficient way would be the first to come out in this competitive world. How this can be analyzed for the web site is by finding the efficiency of the web site by using the parameters retrieved from the log file. The parameters that are considered for the proposed work are utility, frequency, downloads and book marks. Using the web log details the patterns are extracted and enhanced prefix tree is is generated. For a given pattern of the web developer the mining is done and the patterns are extracted for efficiency. A mathematical model is generated for finding the weight values of each a pattern and for finding the efficiency. It considers both sequential and non sequential patterns.

Key words: Enhanced Prefix tree, Sequential patterns, non sequential patterns

I. INTRODUCTION

In early days the knowledge about products and services are attained only by analysis of the product directly or by media. But in the emerging fast moving world the knowledge is gained by online. People don't waste time in knowing about the product or service by moving to its place rather it is product to the peoples place. This is done by various web services provided by the web site developers. These web site developers are also in need of periodical analysis of the web site so that they would be updated according to the web site visitors need. Many analytical engines exist to work on this factor. These analysis are basically started its process from the log file of the web site.

Log files contain data of the web site visitor who visit the web site and traverse through the web site. Additional information like the users IP address, time spent, traversing path, down loads made, book marks made, copies made etc., are also present. These log files are the basic starting point of any analytical engine.

[10]Web mining is a vast area of research done. There are three types of web mining techniques involved web content mining, web structure mining and web utility mining. Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the Web pages. Web structure mining tries to discover the model underlying the link structures of the web. Model is based on the topology of hyperlinks with or without description of links. Web usage mining refers discovery of user access patterns from Web servers. Web usages data include data from web server access logs, proxy server logs, browser logs, user profiles, registration data, user session or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls or any other data as result of interaction.

The path in which the user traverses through the web site is called as the pattern. They can be both sequential and non sequential patterns. If the web site has a path (a,b,c,d,e) 'a' to 'e' via "bcd' where "abcde" are web pages then these type of path traversed are called as the sequential path. The pattern generate by using this path is called as sequential pattern. If the web site visitor moves through (a,b,e) "a" to "e" via "b" where "abe" are web pages then this is called as non sequential paths. Since none of the intermediate paths are visited and the user directly moves to the desired destination the patterns generated by this is called as non sequential patterns. The non sequential patterns occur if the user types the address directly or back tracking through the web site. Since both the sequential and the non sequential patterns are considered it can be called as hybrid system. In the proposed work both sequential and non sequential patterns are mined and calculation of efficiency of a web pattern is generated

II. RELATED WORK

Many research carried out in the field of web mining. The purpose for which path traversing is done are for predicting the visitors future click sequence, for learning the web user intention of visiting the web site and mining the sequential patterns. The various systems in existence are discussed below

Prefix Span integrated with pseudo projection is the fastest among all the tested algorithms. Furthermore, this mining methodology can be extended to mining sequential patterns with user-specified constraints [1]. The high promise of the pattern-growth approach may lead to its further extension toward efficient mining of other kinds of frequent patterns, such as frequent substructures.. Prefix Span recursively projects a sequence database into a set of smaller projected sequence databases and grows sequential patterns in each projected database by exploring only locally frequent fragments. This deals only with the frequency parameter. This is applicable only for sequential patterns and not for non sequential patterns.

The EXT-Prefixspan mines the complete set of patterns but greatly reduces the efforts of candidate subsequence generation. [2] Moreover, prefix – projection substantially reduces the size of projected database and leads to efficient processing. We show that the EXT-Prefixspan algorithm is more flexible at capturing desired knowledge than previous Algorithm. This also deals with the frequent sequential patterns only. No other parameters or parameters can be considered.

[3]In order to discover the frequently occurred sequential patterns from databases basically, the existing studies on finding sequential patterns can be roughly classified into two main categories. In the First category, the discovered patterns are continuous patterns, where all the elements in the pattern appear in consecutive positions in transactions. The second category is to mine discontinuous patterns, where the adjacent elements in the pattern need not appear consecutively in transactions. Here both sequential and non sequential are mined but considering only one parameter frequency.

Path traversal graphs are generated and are used for understanding the web site visitors aim in surfing the web site [4]. There are throughout surfing patterns re generated (TSP) for mining [5].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.

To effectively provide online prediction, we have developed a recommendation system called WebPUM, an online prediction using Web usage mining system and propose a novel approach for classifying user navigation patterns to predict users' future intentions [6]. The approach is based on the new graph partitioning algorithm to model user navigation patterns for the navigation patterns mining phase. Furthermore, longest common subsequence algorithm is used for classifying current user activities to predict user next movement. This would consider the entire set of items even though only some are needed.

Utility-based WAS tree (UWAS-tree) and incremental UWAS-tree (IUWAS-tree) for mining WASs in static and incremental databases, respectively. [7] IT can handle both forward and backward references, static and incremental data, avoids the level-wise candidate generation-and-test methodology, does not scan databases several times, and considers both internal and external utilities of a web page. The IUWAS-tree is also applicable for interactive mining. Extensive performance analyses show that our approach is very efficient for both static and incremental mining of high utility WASs.

[8] An algorithm for utility-based web path traversal mining. This technique prunes a huge number of candidates by using a pattern growth sequential mining approach. It efficiently divides the search space by small projected databases recursively using the divide and conquer technique. This needs a number of scans of the database.

Lexicographic quantitative sequence tree is generated to extract the complete set of high utility sequences and design concatenation mechanisms for calculating the utility of a node and its children with two effective pruning strategies. [9] Substantial experiments on both synthetic and real datasets show that USpan efficiently identifies high utility sequences from large scale data with very low minimum utility. This considers only utility and not any other additional parameters.

III. PROPOSED WORK

A. Data Extraction

The main input for the analysis is the log file of the web site for which the analysis has to be made. These log file has details of all the activities carried out by the web site visitors. This work is limited to the parameters like frequency, utility, down loads, book marks. The raw data present in the log file will be converted to the machine readable format. The input is retrieved from the log file whenever the developer is in need of it. These data is put up in the existing data base. There for the data are modified periodically. This type of database we use is called as incremental database.

B. Enhanced prefix tree construction

Prefix tree is an ordered data structure used for dynamically changing data. Since the information retrieved from the log file makes the data base to modify periodically. This type of Enhanced prefix tree can be used so that each time the tree is also modified. The parameters present in the node are web page name, time, frequency, down loads, Book marks.

A	750s	5	3	4

Fig. 1. Node structure

The Fig. 1. Shows the node structure of the prefix tree generated. Where A represents the web page of the web site considered, 750s is the time utilized by the visitors visited the web page, 5 is the number of times the web page has been visited, 3 is the number of downloads made in the web site and 4 is the number book mark made by the visitors of the web page. These parameters are not confined to a particular user it is considered for the various users using the web page. This act as the nodes of the prefix tree generated and the parameter values changes dynamically as the data base change occurs.

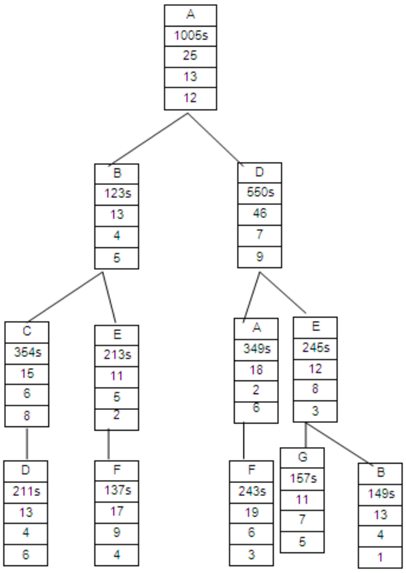

Fig. 2. Sample tree structure

The Fig 2. Shows the sample tree structure of the prefix tree that is generated. The values of the various parameters present in the tree are dynamically changed according to the log data that are extracted dynamically. This tree is generated for both sequential patterns and non sequential patterns. Considering the path A → B → C → D corresponds for the sequential patterns. The path a → B → E → F corresponds for the non sequential patterns.

The Root node is started with the web page represented by A since this may be the first page visited by the first user. The user information is not recorded rather the path details are recorded. For the next modification in the prefix tree the web page is compared starting from the root if the root matches then the details like utility, frequency, book mark, down loads are updated from the root node till the match path is there in the tree. If the root node is not matching with the initial path of the next user then the child is compared, the comparison is carried out until any of the child node matches. If a match of one node is found then all the parameters are updated of done starting from that position to the entire path following it. If the initial node matches and the following nodes are not matched then a new branch is formed and the parameters are added to the nodes. Once all the patterns are formed as a tree the next step is the pattern mining process.

B. Mining the pattern considering their values

Sequential patterns are known to the web developers where as the non sequential patterns are unknown. The mining is done by comparing a particular path data with the Enhanced prefix tree generated. The mining of the sequential patterns are done separately and the non sequential patterns are done separately. Since the test data for the sequential pattern is known to the web developer. The pattern length is considered as one of the constant. Example A→ B→ C → D are said to be having four as the pattern length similarly A → B → C is said to have a pattern length of three. The web developer would have a number of pattern coming under the four length pattern similarly it is applicable in each of length that is considered. For each and every constant length pattern the mining is done in the prefix tree. The first five high utilized pattern and five least utilized patterns is extracted as the necessary pattern.

Definition 1: For NW number of web pages in the web site there are NP number of web paths traversed by the visitor.

A web site is a composition of many web pages. According to the proposed work the web site is considered to have web pages named starting from

'A' to 'T'. Therefore the website contains twenty web pages. These twenty web pages are traversed by one thousand four hundred and sixty two visitors during the extraction done for experimentation. Since each visitor constitute for the path traversed the NP will take up the value.

$NW=20$ [1]

$NP=1562$ [2]

Definition 2: For NP number of web paths there are same NP number of patterns generated to form the prefix tree.

The path traversed by the user can also be considered as the patterns generated during the traversal. Therefore both take up the same value.

Definition 3: For NP number patterns generated there are N number of nodes present in the prefix tree.

Being considering the prefix tree the advantage is that the path that already exists in the tree does not need a new branch of nodes. They can just update the parameters of the existing nodes present in the prefix tree. The mining would be easier only if there is a small tree formed by using the patterns. In the proposed work for the considered NP value there are two hundred and thirty eight nodes formed in the prefix tree.

$N=238$ [3]

Definition 4: There is Ns number of patterns available for varying s length patterns considered for efficiency calculation.

s is said to be the length of pattern considered for the mining from the prefix tree. This s takes up varying values like 3, 4, 5, 6, 7, 8. For this varying length considered the number of sequential patterns Ns available for comparison.

$N2=190$ [4]

$N3=1140$ [5]

$N4=4845$ [6]

$N5=15504$ [7]

The equations [4,5,6,7] shows the number of total combinations that can be present in the web site. Considering the web site with twenty web page in it.

The combination values include sequential patterns and non sequential patterns.

For a web developer would have these many types of sequences for which efficiency to be calculated. The entire path for which the efficiency to be calculated is given by the web developer. The efficiency of that particular is found by the formulas generated in the next efficiency calculation. The traversing of these given paths are done in the Enhanced prefix tree that is generated.

B. Efficiency calculation

Once the desired path for which the efficiency has to be found is given by the web developer the next activity is to apply all the values of each parameters and apply it in the formula

$PW= (TD+TB)*TF*\sqrt{TU}$ [8]

$\text{Path efficiency}= (PW/100)*(NL/NW)$ [9]

The equation [1, 2] shows the equation for finding the path weight and efficiency of path respectively.

Where PW ➔Path weight, TD ➔Total number of downloads, TF➔total number of Bookmarks, TU ➔ Total utility in seconds. After calculating the path weight the result value is used in finding the efficiency of the path.

Where NL➔ Total length of the path considered, NW ➔ Total number of web pages found in the web site considered. This efficiency value enables the developer to know about how much is the path reached the customers who traverse the web site.

IV. RESULTS AND DISCUSSION

The traversing technique is compared with the existing prefix tree technique. The prefix tree present in the existing system deals with one parameter frequency. The proposed work deals with multiple parameters like utility, frequency, down loads, Book marks. The difference in traversal is shown in the graph below. This constitutes for both sequential and non sequential path. Other path traversing algorithms deals only with sequential path. In the proposed system multiple parameters with consideration of sequential and non sequential patterns are done.

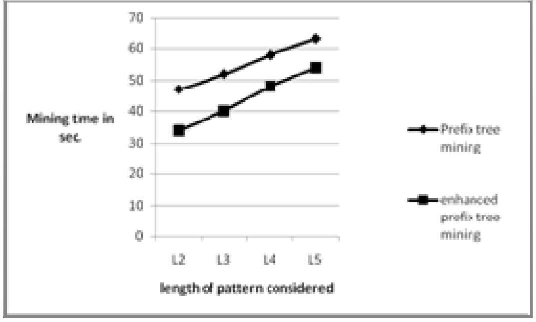

Fig. 3. Comparison with respect length of pattern

The Fig 3. Gives the comparison done between the existing Prefix tree mining with the proposed Enhanced prefix tree mining. The time taken for mining of Patterns is represented. There is considerable difference in the time taken for mining. It also considers both sequential and non sequential patterns for mining. The various length considered for comparison are L2, L3, L4 and L5 which is nothing but pattern length of two, three, four and five, respectively.

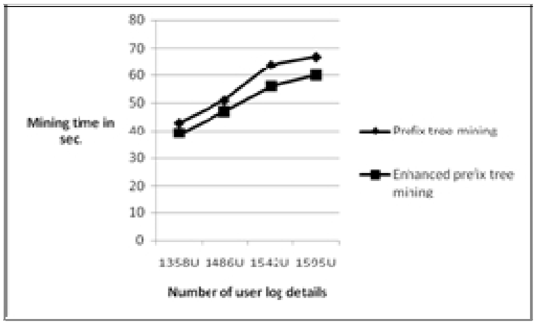

Fig. 4. Comparison with respect to number of user

Fig. 4 gives the comparison of prefix tree mining and Enhanced prefix tree mining with respect to the number of users considered from the log file. There is considerable variation in the time taken for mining of the sequential and the non sequential patterns. The proposed work shows n enhanced result than the existing.

Therefore it is concluded that there is a considerable efficiency in the traversing of the patterns. Additional parameters considered and consideration of both sequential and the non sequential patterns adds weight for the proposed work.

V.CONCLUSION

The proposed work gives an efficient mining result with both sequential and non sequential patterns. It is helpful for establishment of efficient web site. This by default would help in the improvement of the web service provided by the web developers in order to render efficient service.

REFERENCES

[1] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, Mei-Chun Hsu., 2004, Mining Sequential Patterns by Pattern-Growth:The PrefixSpan Approach, IEEE transactions on knowledge and data engineering, vol. 16, no. 10,pp. 1 -17.

[2] S.Vijayalakshmi,V.Mohan, S.Suresh Raja., 2009, Mining Constraint-based Multidimensional Frequent Sequential Pattern in Web Logs, European Journal of Scientific Research, Vol.36 No.3, pp 480-490.

[3] Yen-Liang Chena, Shih-Sheng Chena, Ping-Yu Hsu, 2002, Mining hybrid sequential patterns and sequential rules, Information System, Elsevier, pp. 345–362.

[4] Istvan K. Nagy, 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.

[5] 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.

[6] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat, 2010, WebPUM: A Web-based recommendation system to predict user future movements, Expert Systems with Applications, vol. 38, pp. 6201–6212.

[7] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, 2011, A Framework for Mining High Utility Web Access Sequences, IETE technical review, vol. 28, pp 3-16.

[8] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, and Young-Koo Lee, 2009, Efficient Mining of Utility-Based Web Path Traversal Patterns, ICACT, vol. 28, pp 2215-2218.

[9] Junfu Yin, Zhigang Zheng, Longbing Cao, 2013, USpan: An Efficient Algorithm for Mining High Utility Sequential Patterns , International conference on Knowledge discovery and data mining, pp 660-668.

[10] Chintandeep Kaur, Rinkle Rani Aggarwal, 2012, Web mining tasks and types: A Survey, International Journal of Research in IT & Management, vol. 2, pp. 547-558.