A NEW IMPROVED WEIGHTED ASSOCIATION RULE MINING WITH DYNAMIC PROGRAMMING APPROACH FOR PREDICTING A USER'S NEXT ACCESS

S.A.Sahaaya Arul Mary[1] and M.Malarvizhi[2]

[1]Department of Computer Science and Engineering, Jayaram College of Engineering and Technology, Tiruchirappalli, Tamilnadu, India, Pin: 621 014.
samjessi@gmail.com
[2]Department of computer Applications, J.J. College of Engineering and Technology, Tiruchirappalli, Tamilnadu, India, Pin:620 009.
malarbas@yahoo.co.in

ABSTRACT

With the rapid development of Internet, Web search has been taken an important role in our ordinary life. In web search, mining frequent patterns in large database is a major research area. Due to increase of user activities on web, web-searching methods, to predict the next request of user visits in web pages plays a major role. Web searching methods are helpful to provide quality results, timely answer and also offer a customized navigation. In web search, Association rule mining is an important data analysis method to discover associated web pages. Most of the researchers implemented association mining using Apriori algorithm with binary representation. The problem of this approach is not address the issue like the navigation order of web pages. To overcome this problem researchers proposed a weighted Apriori to maintain navigation order but unable to produce optimal results. With the goal of a most favorable result we proposed a novel approach which combines weighted Apriori and dynamic programming. The experimental result shows that this approach maintains the navigation order of web pages and achieves a best solution. The proposed technique enhances the web site effectiveness, increases the user browsing knowledge, improves the prediction accuracy and decreases the computational complexities.

KEYWORDS

Web usage mining, Web page prediction, Dynamic Programming, Weighted Association Rule Mining (WARM), Improved Apriori Algorithm.

1. INTRODUCTION

Internet is becoming one of the most important tools for human being**.** It has grown extensively in last decades and also increased the web data access. Every user attempts are collected and

Sundarapandian et al. (Eds): ICAITA, SAI, SEAS, CDKP, CMCA, CS & IT 08,
pp. 105–124, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2510

stored in the web server and is known as log data. The web server log [11] has the visitors' click stream behaviors (pages template, cookie, transfer log, time stamp, IP address, agent, referrer etc.). The visitors should know that how to use and adopt this historical information personalized the user interface and recommend the interesting web pages. This is an issue of critical importance in Web user. So the discovery and analysis of useful information from web log mining is an essential need. The web log mining [8,34] is the application of data mining technique to discover user behavior patterns from web log data in order to understand, improve and serve the needs of web user. It also helps the web designer to enhance the web sites. The Web Mining [26] can be classified into three methods such as Web content mining, Web structure mining and Web usage mining. Web content mining is the process of extracting knowledge from the online content of documents or their descriptions. The web structure mining is examined the structural data of a web site and generate the summary of the particular site. Web structure data are represented in tree structure format. Web usage mining is the third category in web mining.
Web usage mining [3,28] referred to the discovery of user access patterns from web usage logs, which records every click made by the users. This information is often gathered automatically into access logs via the Web server. Web usage mining process is similar to data mining process. The difference is in data collection phase. The data are collected from databases for data mining whereas it is collected from web log files in web mining. Once the data is collected, a three step process is performed in web usage mining namely data preparation, pattern discovery and pattern analysis. In Web Usage Mining [6,9] are the popular techniques to predict the user behaviors, to name a few, clustering, classification, association and sequential pattern mining. The Association rule mining received a significant research attention.

Association Rule Mining (ARM) [27] is one of the strategies that find out association among the pages visited together frequently. The researchers [1,25] introduced the ARM for Market Basket analysis Problem. Association Rule Mining has a wide range of applicability such as Medical research, Website navigation analysis, Homeland security and so on. A number of efficient Association Rule Mining algorithms [33, 36] proposed by researchers in the last few years. Among these, the Apriori algorithm gains more popularity. The Apriori algorithm [32] is not only influenced the association rule mining community, but its impact is also on other areas of data mining. It is the best-known algorithm to mine association rules. The breadth-first search strategy is used to count the support of item sets and also used as a candidate generation function which exploits the downward closure property. Apriori properties state that all the subsets of frequent item sets must also be frequent. This algorithm also used frequent item sets, join, and prune methods to derive strong association rules.

Apriori algorithm [2] introduced a level-wise iterative search to discover all maximal frequent sequential patterns. This sequential pattern-mining algorithm [19] is common in web navigation pattern mining. Variations of Apriori algorithm are partitioning, sampling, Tree projection, Markov Model [29] and FP-Growth algorithm [13] have discussed. An extended version of Prefix span algorithm is used to extract the multi dimensional sequential patterns [14] in web usage mining. A web access pattern is a sequential pattern that practice frequently by users. The sequences are projected as prefixes in the database.

Yue Xe et al.[35] presented a reliable representations in association rules for eliminating the redundancy [15] in large web data. They introduced the frequent closed item set instead of frequent item sets. Tarek et al. [31] proposed an incremental algorithm for maintaining temporal association rules. The problem in time series data is solved, by converting time expressions into association rules.

Other Association mining techniques [33] are K-Optimal Pattern Discovery, Mining frequent sequences, Generalized Association Rules [30], Quantitative Association Rules, Interval Data Association Rules, Maximal Association Rules, Sequential Association Rules [16,42], Filtered Associations, Predictive Apriori, non redundant association Rules [15,23] and Tertus. The traditional [10] Association Rule Mining (ARM) model gave an equal significance for both visited and unvisited pages. By assigning binary values for all pages and the orders of visited pages are not ascertain.

To overcome the issues in Association Rule Mining, researchers focused on Weighted Association Rule Mining. Wang et al. [33] first proposed a method for mining weighted association rules to mine frequent item sets using path traversal graph. Joong et al. [17] introduced weighted sequential pattern mining with time interval in sequence database. More interesting sequential patterns are considered based on the significance of each data element in a sequence database. Time information of data elements can be helpful for finding more interesting sequential patterns. Sequential Pattern Mining Algorithm and applied it in different domains such as Medical Treatment, Fraud Detection, Web Site Design and Telecommunication etc.

YongSeog Kim et al. [39] found a new web-mining algorithm by using Streaming Association Rule (SAR). The SAR model is introduced a novel way to incorporate a weighted navigation order for finding interesting associations among the web pages. The divide and conquer technique is implemented to reduce the database scanning time and redundancy. The author has improved the prediction accuracy when compared to traditional techniques. Most of the research work addressed the issue in time complexity.

The proposed approach combines Weighted Association Rule Mining and Dynamic Programming. Weighted association mining addressed the issue like navigation order of web pages. Dynamic programming technique handles large web log data and produces a local optimal solution. Proposed approach combines to address the issues like navigation order of web pages, which has a significant impact in rule accuracy, prediction accuracy and also handles large Web Log data.

In this work, Dynamic Programming (DP) is applied for both Association Rule Mining with Apriori and Weighted Association Rule Mining with improved Apriori technique. These techniques are implemented by using java. The results of these two techniques are compared and best one is proposed. Based on the following implementation Dynamic Programming with Weighted Association Rule Mining will be yielded the best optimal solution. The proposed approach discovered patterns are highly important for web site developer for link restructuring, users for quick searching and site owners for content improvement. The subsequent section of this paper is organized as follows: Section 2 presents some basic considerations of the prediction techniques, Section 3 details the methodology of web Usage mining frameworks, Section 4 focuses the proposed algorithms, Section 5 discusses the results obtained in proposed algorithms and Section 6 Concludes the paper.

2. BACKGROUND OF THE PREDICTION TECHNIQUES

Several algorithms have been applied to solve the issues in web page prediction models. In this section the behavior of the most representative algorithms should be investigated. After analyzing

their drawbacks a novel method can be developed. The Association Rule Mining (ARM), Weighted Association Rule Mining (WARM) and Dynamic Programming approach are used to implement this proposed method.

2.1. Preliminaries of Association Rule Mining

This section presents some preliminaries to facilitate the presentation of proposed method. Association Rule Mining is one of the most important techniques in data mining. Association rule mining technique is used to find the frequently visited web pages from the user access sequences and constructs a set of rules based on those visits. The ARM has two separate phases: the first phase is to find the frequent item sets and the second is to determine the rules form these item sets. D is a database with different transactions. P= (P_1, P_2, …,P_n}be a set of n distinct web pages. An association is an implication in the form of $P_1 \Rightarrow P_2$ where $P_1 \subset P$, $P_2 \subset P$ and $P_1 \cap P_2 = \varnothing$. P_1, (or P_2) is a web page. P_1 is called antecedent of page P_2 where as P_2 is called consequent page. The quality of the rule is measured by its support, confidence, comprehensibility, J-measures and prediction accuracy. The support S is measured by the following.

$$P(P1 \Rightarrow P2) = P(P1 \cup P2) = S$$

A rule $P_1 \Rightarrow P_2$ is satisfied in the set of transactions with confidence factor C if at least C% of the transaction in D that satisfies P_1 also satisfies P_2.

$$\text{Confidence}(P1 \Rightarrow P2) = P\left(\frac{P2}{P1}\right) = \frac{Support(P1 \cup P2)}{Support(P1)} = C$$

$$\text{Comprehensibilty} = \log(1 + y) + \log(1 + x \cup y)$$

$$\text{Prediction Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions Prepared}}$$

Both support and confidence are fractions in the interval [0,1]. The support is a measure of statistical significance, whereas confidence is a measure of the strength of the rule. The rule is said to be “interesting” if its support and confidence are greater than user defined threshold Sup_{min} and Con_{min} respectively. There are two thresholds: P_s and P_a where P_s is a lower bound on the support of the rule and P_a is a lower bound on the accuracy of the rule. A pattern gets a score of 1 if it satisfies both of the threshold conditions and gets a score of 0 otherwise. The goal is to find all rules (pattern) with a score of 1. This technique is applied in the Pattern discovery phase of Web usage mining.

Association rule mining with Boolean order representation method [10], the navigation sequences is a series of binary values “1” and “0”. Visited web pages are represented as “1” and unvisited web pages are represented as “0”. In this research work, navigation patterns are collected from a web site that has three different web pages P_1, P_2 and P_3. Let’s take a pattern of the form $P_1 \rightarrow P_3 \rightarrow P_2$. In Boolean representation, this pattern is represented as P_1=1, P_3=1and P_2=1 which means that the visitors who have visited P_1 and P_3 also visited P_2. The patterns are interesting

based on the characteristics are stability, novelty and action ability. This paper demonstrates and compares the support and confidence values. ARM model handles web pages that are static in nature and not suitable for dynamic web pages.

2.2. Preliminaries of Weighted Association Rule Mining (WARM)

WARM addressed the issue of binary relationships of visited pages in association rule mining model. It allows different weights to be assigned to different pages based on their order of visit, and it is the best approach for improving the ARM model. In weighted association rule mining [17,36] each navigated page is assigned an integer value between "P" to "1". P is assigned to the first visited page, decrement it by 1 for the next visited page and continue up to 1 for the lastly visited page. Early visited page acquired more weight indicating the priority.

This technique eliminates the problems in existing Association Mining method and also handles both static and dynamic web pages. It requires only a single scan of data sets. This in turn eliminates data redundancy and maintains navigation order of web pages. Let the Sample log data S have Transactions $T=\{T_1,T_2,\ldots,T_n\}$ with set of pages $P=\{P_1,P_2,\ldots P_n\}$ and a set of positive real number of weights $W=\{W_1,W_2,\ldots,W_n\}$ attached with each visited page P. The proposed algorithm Weighted Association Rule Mining is scalable and efficient in discovering significant relationships between web pages.

2.3. Preliminaries of Dynamic programming Technique

Web log data are massive in nature and unable to process the entire data base data in single scan. So a new technique is required to divide the log data into several sub datasets and to solve the independent sub sets. The divide and conquer method is applied in existing Association mining techniques. The divide and conquer algorithm is inefficient, because it repeatedly solved the same sub problems and are not independent. These introduce redundancy, increase processing time and memory space. This technique requires exponential time for trivial computation.

To overcome these problems, Dynamic programming approach [4,5] is used to solve each sub problems only once and the results are stored in a table. So recursion is not required and time efficient for single scan of data sets. It requires only linear time for computation. Dynamic programming techniques found suitable for solving complex sub problems. It is divided into multistage decision sub problems and solved the same. Dynamic programming algorithm has three basic components.

- The recurrence relation (for defining the value of an optimal solution)
- The tabular computation (for computing the value of an optimal solution)
- The trace back (for delivering an optimal solution).

So the Dynamic programming technique is used in the proposed methodology. This can be implemented in pattern analysis phase of web usage mining. It will be improved the time efficiency and redundancy of Weighted Association Mining Technique. The new approach combines the weighed association mining technique and dynamic programming to obtain the

most excellent navigational sequences. The proposed web recommendation model is illustrated in the following sections.

3. WEB USAGE MINING FRAMEWORK

In this section, the Web usage mining [3, 26] can be divided into several phases as shown in the figure 1.

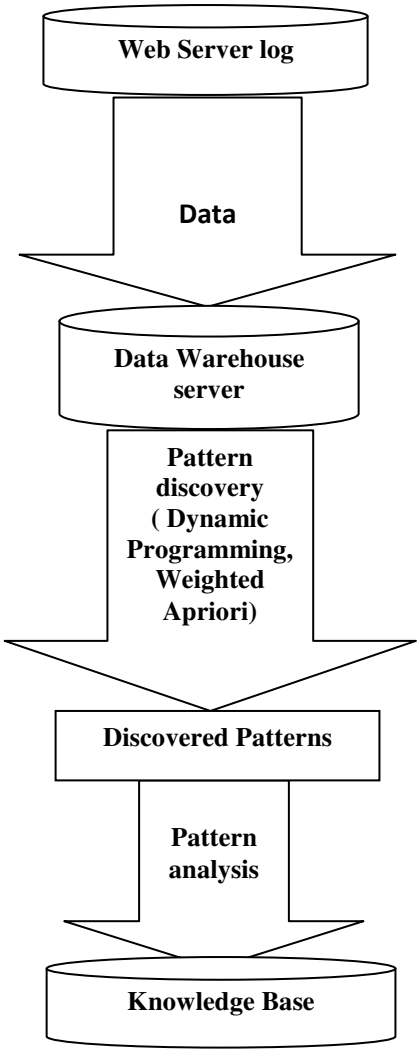

Figure 1 Web Usage Mining Process

Web usage mining is a process to collect the Web Log data from the Web Server or Application Server. Once a data is collected from various resources, they are integrated and stored in data warehouse server. The collected data are unclean data and these data should not give quality result. So we need a pre-processing phase to improve the quality of web log data.

The Raw Web Log data are converted into abstract data using data cleaning and filtering techniques. In Pattern discovery phase, Dynamic programming approach is applied to divide the

web log data and Weighted Apriori algorithm is applied to extract the hidden patterns from partitioned web log data. The final stage of the process extracts the frequently visited pages and stored it in a knowledge base.

3.1. Web Server Log

Web log files may contain a large amount of erroneous, misleading, and incomplete information. Sample server log data is shown in figure 2. The log data described the page visits of users. Visits are recorded at the level of URL category in time order. The information available in the log files are: IP address, Date and Time, HTTP request, HTTP response, Response size, Referring document and User agent string. This is referred as raw log data and need to be processed. The test data sets are taken from the literature and it is proposed by YongSeog Kim [39]. These data sets are used to generate the prediction rules by applying Weighted Association Rule Mining technique. The data set used in this study is msnbc data set. The summary of the description is available at Kdd.ics.uci.edu/databases/msnbc/msnbc.html.

```
123.456.78.9-[25 /Apr /98/03:04:41-05000] GET A.html HTTP/1.0/200-3290-Mazilla /3.04(Win95.1)
123.456.78.9-[25/Apr/ 98: 03:05:34:05000] GET B.html HTTP/1.0 200 123.456.78.9-[25/Apr
/98:03:06:39: 5000] L.html HTTP/1.0 200 3290 L.html Mozilla/3.04 123.456.78.9-[25/ Apr/ 98:
03:06:58-05000] GET R.html HTTP/1.0 200 2050 R.html Mozilla /3.04(Win95, 1) 209.456.78.3-
[25/Apr/98: 5:06:03-0500]GET D.html HTTP/1.0.200 A.html Mozilla/3(Win95,1)
```

Figure 2 Sample Web Server Log

3.2. Data Selection and Preprocessing

In data selection phase, data relevant to the analysis task are retrieved from the server logs. The selected data must be preprocessed. In this preprocessing step web log data are prepared for mining process. The output of the preprocessing phase is abstract data. The data mining algorithms are applied to this abstract data. The preprocessing step contains three separate phases. First, the collected data must be cleaned. For example data such as the graphics and multimedia data are removed. Secondly, the different sessions belonging to different users should be identified. In the third step, convert the raw data into a format needed by the mining algorithm. Duplicate, incomplete, noisy and inconsistent data items are eliminated in this phase.

3.3. Pattern Discovery and Analysis

The abstract data produced by the above phase is used for pattern discovery. Pattern discovery phase contains several techniques such as statistical analysis, association rules, clustering, classification, sequence patterns and dependency modeling to extract the hidden patterns.This work mainly aimed at proposing an improved Association Rule Mining to extract the hidden pattern from the abstract data that is generated by the preprocessing phase. With the discovered patterns, pattern analysis is performed to find the frequently visited page. This analytical report helps user and designer of web site.

4. PROBLEM SOLVING METHODOLOGIES

The proposed work comprises of three phases and is given below

4.1. Preprocess and divide the web log data

4.2. Discover patterns using weighted association rule mining algorithm.

4.3. Find an optimal rule using dynamic programming Approach.

The first phase is concerned about the preprocessing and web log data partition. The second phase dealt with pattern discovery using weighted association rule mining with the enhancement on Apriori algorithm. The third phase of the pattern discovery is to find an optimal pattern.

4.1. Preprocess and divide the web log data

In this work, the data sets are taken from web server logs of msnbc.com web site. It consists of nearly one million visitors and the particular session contains seventeen categories. Each category has 5000 records. Preprocessing techniques are applied to sample web log data of size 85,000 records and have eliminated the graphics files, duplicated data, incomplete data, noisy data and inconsistent data. As a result of the preprocessing phase each session is having only 5000 records From the large data set, a particular session data has been taken. For these 5000 sample log data, Association Rule Mining with Apriori and Weighted Association Rule Mining with improved Apriori techniques have been applied. It seems to be tedious task to present the output for the 5000 records. So the work is illustrated by considering 20 sample data from the entire data set. Table 1 is formulated based on the web log data from figure 2.

Table1: Weighted Order Representation of Navigation Patterns

Blocks	Name of Transaction	Web Pages Visited order	Weighted order Representation		
			P_1	P_2	P_3
B1	T1	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1
	T2	$P_3 \rightarrow P_2 \rightarrow P_1$	1	2	3
	T3	$P_1 \rightarrow P_2$	3	2	0
	T4	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1
	T5	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1
B2	T6	$P_2 \rightarrow P_3$	0	3	2
	T7	$P_1 \rightarrow P_3$	3	0	2
	T8	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1
	T9	$P_1 \rightarrow P_2$	3	2	0
	T10	$P_2 \rightarrow P_1$	2	3	0
B3	T11	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1
	T12	$P_3 \rightarrow P_2 \rightarrow P_1$	1	2	3

	T13	$P_2 \rightarrow P_3$	0	3	2
	T14	$P_3 \rightarrow P_2 \rightarrow P_1$	1	2	3
	T15	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1
B4	T16	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1
	T17	$P_2 \rightarrow P_3$	0	3	2
	T18	$P_1 \rightarrow P_2$	3	2	0
	T19	$P_3 \rightarrow P_2 \rightarrow P_1$	3	2	1
	T20	$P_1 \rightarrow P_2$	3	2	0

The log data is divided into four blocks B_1, B_2, B_3 and B_4. Every Block has five different transactions T_1, T_2, T_3, T_4 and T_5. All the transactions are represented in weighted order and each block consist of three navigation records with web pages P_1, P_2 and P_3. The page that visited first must given a highest priority, then the next visited page given a value which is one less than the highest priority. Example P_1=3, P_3=2 and P_2=1, meant that visitors who have visited "P_1" first and "P_3" second also visited "P_2". However a web page in each rule is showed different weights of visited order. Based on the Weighted Order Representation the researchers come across the navigation order of web pages appropriately. After applying the Weighted Association Rule mining technique with dynamic programming in web log data, we get the subsequent table with blocks of records, number of transactions and navigation order of web pages. The pruning rule is then applied in each block of data. The improved pruning rules are developed and that will eliminate redundancy, insignificant association rules. The following are the rules applied for data cleaning. This proposed technique enhances the association mining by applying the following.

Rule1: Pages with zero weight does not generate association rules.

Rule2: Eliminate pages that are not immediately visited.

Rule3: The antecedent weight is always consequent weight plus one.

Rule4: Eliminate pages with less search time span.

Rule5: Eliminate special antecedent and consequent pages.

The algorithm in section.4.1.1 designed using dynamic programming approach. It is also assigned weights according to the visited order of each page in the transaction. Each block generated by this algorithm is passed as an input parameter to the WApriori algorithm in section 4.2.1. The output of the WApriori is stored in Optimal Binary Search (OBST) table and from that an optimized solution is generated.

4.1.1. DWAssociation algorithm: Find frequently visited pages using a dynamic programming with Weighted Apriori algorithm.

Input: D, a Database of Transactions

N_t , Total number of transactions in D

N_s, Number of subsets

P, Maximum number of pages visited

Output: Find the Frequently visited pages

Method:

1. **Scan** the Database D, calculate the subset size $D_s=N_t/N_s$

2.**Partition** the dataset D into N_s number of subsets $D_{s1},D_{s2},\ldots,D_{sNs}$

3.**Initializ**e the weight of each subset in D transactions into zero

4.**for** i = 1 to N_s **do**

5.**for** j = 1 to D_s **do**

6. **If** first visited page **then**

7. W←P

8. **Else if** next visited page **then**

9. P=P-1

10. W←P

11 **end if**

12. **end if**

13.D_s (j)=W

14.**end for**

15.**end for**

16.**for** i =1 to N_s **do**

17. R (i)=**WApriori**(D_s(i))

18.**end for**

19.Apply Dynamic programming technique to store R(i) into the OBST table.

20.**Repeat**

21. Scan the OBST data sets and **call WApriori** algorithm

22.**Until** reach the goal state

23.Obtain the best rule R_b

24.**Return** R_b

4.2. Discover Patterns using Weighted Association Rule Mining Algorithm

Apriori algorithm is a classical algorithm of association rule mining. A lot of algorithms for mining association rules are available and they are proposed on basis of Apriori algorithm with binary representation. Instead of binary representation in Apriori, we use weights with numerical values to maintain the order of visit. This is an important modification step done in Apriori and denoted as Weighted Apriori (WApriori). In weighted association, rules are denoted by $(P_1 => P_2)^W$, which is obtained by two main concepts namely weighed support and weighted confidence. W is an access sequence table of the web log data denoted by W= {w_1, w_2, w_3…wn}. Table 2 shows the computational result of support and confidence values for the two association rules such as $P_1 \rightarrow P_2$ and $P_1 \rightarrow P_2 \rightarrow P_3$ in B_1, B_2, B_3 and B_4. For instance, to the block B_1 the Weighted order representation for the two pages is P_1=3 and P_2=2 and its corresponding support and confidence values are 80% (=4/5) and 80%(=4/5). For the three pages with weights such as P_1=3, P_2=2 and P_3=1, its support and confidence values can be represented as 60%(=3/5) and 60%(=3/5).

Table2: Example Association Rules with Support and Confidence values

Blocks	$P_1 \rightarrow P_2$		$P_1 \rightarrow P_2 \rightarrow P_3$	
	Support	Confidence	Support	Confidence
B_1	4/5	4/5	3/5	3/5
B_2	2/5	2/4	1/5	1/5
B_3	2/5	2/4	2/5	2/5
B_4	3/5	3/4	1/5	1/5

The following algorithm in section.4.2.1 is used to find the frequently visited pages by applying join and prune techniques. It initially starts with 1-page-visit. Then it is generated 2-page-visit, 3-page-visit and so on up to n-page-visits. Every page visit is generated by compare it with minimum support count value. If the constraint is not satisfied it is pruned else generate a new candidate page visit. This algorithm is invoked by DWAssociation algorithm.

4.2.1. Algorithm WApriori: Find frequently visited pages using weighted order representation.

Input **:** T_s, a Transaction database Min-sup, the minimum support count threshold.

Output **:** F_p, the frequently visited pages in D

1.**Scan** T_s and count the number of occurrences (N_o) of 1-page-visit from visited pages (V_p) using page weights.

2.**Compare** N_o with minimum support count.

3.**if** N_o < min-sup **then**

4. Prune V_p

5.**else**

6. Add to list L_1

7. **End if**

8. Discover 2-page-visit L_2 by joining L_1 with L_1

9. Scan T_s and count number of occurrences of 2-page-visit $N_{o(1)}$ using weighted order method

10. Compare $N_{o(1)}$ with minimum support count

11.**if** $N_{o(1)}$ < min-sup **then**

12. Prune $N_{o(1)}$

13.**else**

14. Add to list L_2

15. **End if**

16. Discover 3-page-visit, 4-page-visit etc. until found all frequently visited pages (F_p)

17. Return F_p**.**

4.3. Find an optimal rule using dynamic programming Approach.

In Traditional Association mining techniques several models were used for rule generation. They are Simple Model, Model without Rule merge, Model with Rule Merge and Popularity voting. These techniques have a drawback that they scanned the same database multiple times. But in this work we used dynamic programming approach and derived rules are stored in table. So we eliminated the multiple data scans and repetitive calculation. An integrated model is used to combine all data sets and filtered to maintain a new compact set of rules. Table 3 shows the support and confidence values for combined blocks and the final results are stored an Optimal Binary Search table. From this table, the optimal solution is attained.

Table 3: Combined Blocks Association Rules with Support and Confidence values

Blocks	$P_1 \rightarrow P_2$		$P_1 \rightarrow P_2 \rightarrow P_3$	
	Support	Confidence	Support	Confidence
B_1+B_2	3/5	2/3	2/3	2/3
B_2+B_3	2/5	1/2	3/10	2/3
B_3+B_4	1/2	5/8	3/10	1/2

5. RESULTS AND DISCUSSIONS

The DWAssociation and WApriori algorithm is successfully implemented using Java. The inputs for this algorithm are 20 sample transactional data items $T=\{T_1, T_2, \ldots, T_{20}\}$ as given in table 1. Table 4 contains the rule count or frequently visited page count, their support and confidence values for 1- page-visit.

The values in table 4 are pruned based on minimum support count, which is represented as min-sup and assigned a value 2. In pruning we compare the support count value with the minimum support count. The minimum confidence value also used to compare the confidence values for the best rule selection. If support count and confidence values are smaller eliminates the pattern else continue the candidate generation.

Table 4: 1-Page-Visit count with Support and Confidence values

Block No.	No. of Rules	Name of the Rules	Rule Count	Support	Confidence
1	1	P_1	5	1	1
	2	P_2	5	1	1
	3	P_3	4	4/5	1
2	1	P_1	4	4/5	1
	2	P_2	4	4/5	1
	3	P_3	3	3/5	1
3	1	P_1	4	4/5	1
	2	P_2	5	1	1
	3	P_3	5	1	1
4	1	P_1	4	4/5	1
	2	P_2	5	1	1
	3	P_3	3	3/5	1

From the table 4, we observed that no rules are pruned because the table support and confidence values are higher than the minimum support, confidence values and it is called as the first level. This level also eliminated the pages that are not immediately visited and lesser time span pages. After applying pruning rule in the data set to get pruned one page resultant patterns. Using one page patterns apply the same procedure to obtain 2-page-visit patterns. The resultant page visits are shown in table 5.

Table 5: 2-Page-Visit navigation patterns with Support and Confidence values

Block No.	No. of Rules	Name of the Rules	Rule Count	Support	Confidence
1	1	$P_1 \rightarrow P_2$	4	4/5	4/5
	2	$P_2 \rightarrow P_3$	3	3/5	3/5
	3	$P_2 \rightarrow P_1$	1	1/5	1/5
	4	$P_3 \rightarrow P_2$	1	1/5	1/4
2	1	$P_1 \rightarrow P_2$	2	2/5	2/4
	2	$P_1 \rightarrow P_3$	1	1/5	1/4
	3	$P_2 \rightarrow P_3$	2	2/5	2/4
	4	$P_2 \rightarrow P_1$	1	1/5	1/4
3	1	$P_1 \rightarrow P_2$	2	2/5	2/4
	2	$P_2 \rightarrow P_3$	3	3/5	3/5
	3	$P_2 \rightarrow P_1$	2	2/5	2/5
	4	$P_3 \rightarrow P_2$	2	2/5	2/5
4	1	$P_1 \rightarrow P_2$	3	3/5	1
	2	$P_2 \rightarrow P_1$	1	1/5	1/5
	3	$P_2 \rightarrow P_3$	2	2/5	2/5
	4	$P_3 \rightarrow P_2$	1	1/5	1/3

Apply the pruning process with 2-page-visit navigation record that is in table 5. As a result, the record set 3 and 4 in Block 1, 2 and 4 in block 2, 2 and 4 in block 4 are eliminated. The candidate item sets are generated using pruned data set and the resultant 3-Page-visit navigation records are stored in table 6.

Table 6: 3-page-visit navigation patterns with Support and Confidence values

No. of Rules	Name of the Rules	Rule Count	Support	Confidence
1	$P_1 \rightarrow P_2 \rightarrow P_3$	7	7/20	7/11
2	$P_3 \rightarrow P_2 \rightarrow P_1$	2	2/20	2/4

Finally the data sets in table 6 are pruned and best rule is generated and stored in table 7. It contains the best rule with corresponding support and confidence values.

Table 7: Final navigation table with support and confidence values

Block No.	No. of Rules	Name of the Rules	Rule Count	Support	Confidence
1	1	$P_1 \rightarrow P_2 \rightarrow P_3$	3	3/5	3/ 4
2	1	$P_1 \rightarrow P_2 \rightarrow P_3$	1	1/5	1/3
3	1	$P_1 \rightarrow P_2 \rightarrow P_3$	2	2/5	2/2
	2	$P_3 \rightarrow P_2 \rightarrow P_1$	2	2/5	2/2
4	1	$P_1 \rightarrow P_2 \rightarrow P_3$	1	1/5	1/3

To show the effect of a weighted order representation, the first two frequent rules are taken into account out of 20 transactions. The generated rules must be stored in table 8. Similarly we should find the rules for all the blocks and must be stored in the same table. This table contains support and confidence values for all the generated rules. Finally the optimal rule is generated from the table 8.

Table 8: Optimal Binary Search table using weighted order representation

Blocks	B1		B2		B3		B4	
	Sup	Conf	Sup	Conf	Sup	Conf	Sup	Conf
B1	3/5	3/4	4/10	4/6	6/15	6/8	**7/20**	**7/11**
B2			1/5	1/2	3/10	3/4	4/15	4/7
B3					2/5	2/2	3/10	3/5
B4							1/5	1/3

The resultant rule is $P_1 \rightarrow P_2 \rightarrow P_3$. The support count 7/20 (35%) and confidence value is 7/11 (64%). The same procedure is applied in binary representation method and its optimal results are stored in table 9.

Table 9: Optimal Binary Search table using Binary representation

Blocks		B1		B2		B3		B4	
	Rules	Sup	Conf	Sup	Conf	Sup	Conf	Sup	Conf
B1	R1	4/5	4/5	1/2	5/7	3/5	3 /4	11/20	11/15
	R2	4/5	1	1/2	5/6	3/5	9/11	11/20	11/14
	R3	4/5	1	1/2	5/6	3/5	9/10	11/20	11/12
	R4	4/5	1	1/2	5/6	3/5	3/2	11/15	11/12
	R5	4/5	1	1/2	5/7	9/5	9/11	11/20	11/15
	R6	4/5	1	1/2	5/7	3/5	9/11	11/20	11/15

B2	R1			1/5	1/2	1/2	5/7	7/15	7/10
	R2			1/5	1/2	1/2	5/7	7/15	7/9
	R3			1/5	1/2	1/2	5/6	7/15	7/8
	R4			1/5	1/3	1/2	5/7	7/15	7/8
	R5			1/5	1/3	1/2	5/7	7/15	7/10
	R6			1/5	1/3	1/2	5/7	2/5	6/10
B3	R1					4/5	4/5	3/5	3/ 4
	R2					4/5	4/5	3/5	3 / 4
	R3					4/5	1	3/5	1
	R4					4/5	1	3/5	1
	R5					4/5	1	3/5	3 /4
	R6					4/5	1	3/5	3 /4
B4	R1							2/5	2/3
	R2							2/5	2/3
	R3							2/5	1
	R4							2/5	1
	R5							2/5	1/2
	R6							2/5	1/2

The results generated by Weighted Order representation and Binary representations are compared.

Table 10: Rules comparison Table

Block Names	Binary representation (No. of Rules)	Weighted order representation (No. of Rules)
B1	6	1
B2	6	2
B3	6	1
B4	5	1

From the table 10 we observed that weighted order representation is more accurate than Binary Representation. The reason is that more number of rules generated for every block in Binary representation. This leads to increase in memory and time utilization. For Every block, six rules are generated by the binary method and these support and confidence values are shown in table 9. The generated rules are $P_1 \rightarrow P_2 \rightarrow P_3$, $P_1 \rightarrow P_3 \rightarrow P_2$, $P_2 \rightarrow P_1 \rightarrow P_3$, $P_2 \rightarrow P_3 \rightarrow P_2$, $P_3 \rightarrow P_1 \rightarrow P_2$, and $P_3 \rightarrow P_2 \rightarrow P_1$. The drawback of the binary representation is that the same record set is considered for processing repeatedly and no order is followed for page navigation too. These results are

produced inappropriate prediction of page visits. In weighted order method, the page visits are represented using weights. So, only one rule is generated in weighted order representation for every block whose support and confidence values are stored in table 8. The generated rule is $P_1 \rightarrow P_2 \rightarrow P_3$.

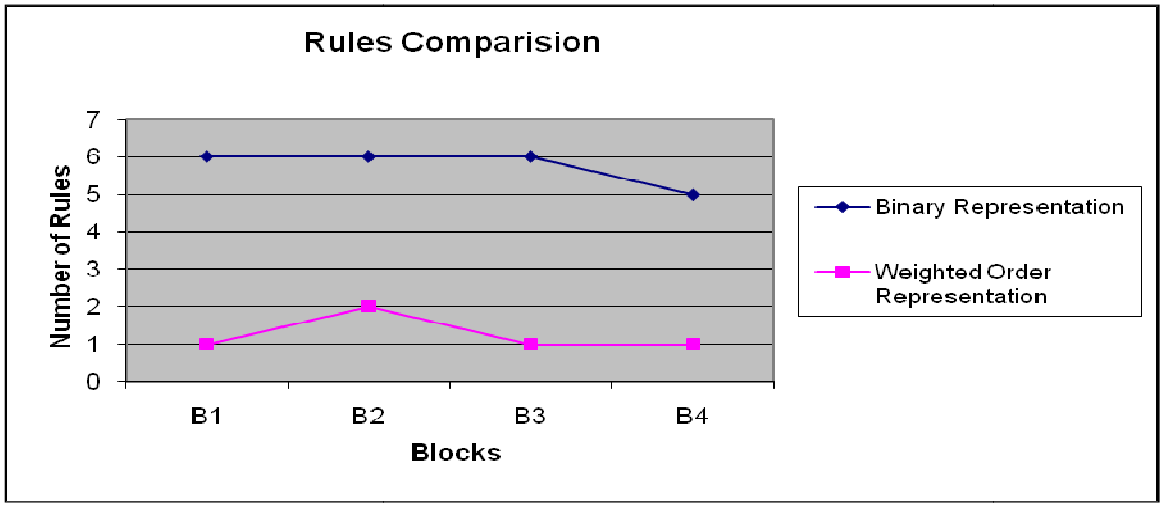

Figure 3: Rule Comparison of Binary and weighted order representation

From the figure 3, we compared four block rules in binary and weighted order representations. Number of rules in Binary representation is greater than weighted order representation. Binary representation seems harder to predict the frequent page visit because larger number of rules generated from data sets and also consume more memory and time. But in weighted order representation only few rules are generated which in turn helps to predict the visiting order easily and also saves time and reduce the memory utilization.

6. CONCLUSIONS

This work is evident and proposes a new web page recommendation system based on the Weighted Apriori algorithm with dynamic programming approach. It extends the association rule mining by assigning significant weights to the pages based on visiting order of each page. The proposed technique provides more weight ages to the order of user's visit, which is more helpful to the users and developers. The System performance is evaluated based on time and space. The proposed technique effectively captures the navigation patterns and efficiently reduces the search time. The experimental result proves that this method guarantees 35% in support and 64% in confidence, which is better than the conventional association rule mining models.

REFERENCES

[1] Agrawal R, Imielinski T & Swami A N, (1993) "Mining association rules between sets of items in large databases", ACM SIGMOD International Conference on Mgt. of Data, Vol.22, Issue 2, pp.207-216.

[2] Agarawal R & Srikant R, (1995) "Mining Sequential Patterns", In proceedings of the 11th International conference on Data Engineering, pp.3-14.

[3] Anitha A & Krishnan N,(2011) "Dynamic Web Mining Frameworks for E-learning Recommendations using Rough Sets and Association Rule Mining", International Journal of Computer Applications,12(11) pp 36-41.

[4] Bertsekas D P, Borkar V, & Nedic A, (2004) "Improved Temporal Difference Methods with Linear Function Approximation", Automatic Control, IEEE Press, pp. 231-255.

[5] Bertsekas.D.P, (2007) "Separable Dynamic Programming and Approximate Decomposition Methods" ,IEEE Transactions on Automatic Ctrl, Vol. 52, Issue 5, pp. 911-916.

[6] Cristina Faba-Peterez & Vicente P et al., (2003) "Data mining in a closed web environment", Scientometrics, Vol 58, No 3, pp.623-640.

[7] Chimphlee S , Salim N & Ngadiman M S B et al.,(2010) "Hybrid Web Page Prediction Model for Predicting User's Next Access", Journal of Information Technology, Vol.9, No.4,pp. 774-781.

[8] Cooley R, Mobasher B, & Srivastava J, (1997) "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, Vol.1, pp.558-567.

[9] Cooley R, Mobasher B, & Srivastava J, (1999) "Data Preparation for Mining World Wide Web Browsing Patterns", Knowledge and Information Systems, vol.1, No.1, pp.5-32.

[10] Hanbing Liu & Baisheng Wang , (2007) " An Association Rule Mining Algorithm Based On A Boolean Matrix", Data Science Journal, Vol. 6, pp.S559-S565.

[11] Hans-Peter Kriegel, & Karsten M et al.,(2007) " Future Trends in Data Mining", Data Mining and Knowledge discovery, pp.15:87–97.

[12] Hardwick J & Stout Q F, (1992) "Optimal Adaptive Equal Allocation Rules" , Computing Science and Statistics, Vol.24, pp. 597-601.

[13] Hengshan Wang, & Cheng Yang et al., (2006) "Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan", Communications of the IIMA, Vol.6, Issue.2, pp.71-86.

[14] Hye-Chung Kum & Joong Hyuk Chang et al.,(2006) " Sequential Pattern Mining in Multi-Databases via Multiple Alignment", Data Mining and Knowledge Discovery, Vol.12,pp.151–180.

[15] James Cheng & Yiping Ke et al., (2008) "Effective elimination of redundant association rules" ,Data Mining and Knowledge Discovery,16:221–249.

[16] Jiawei Han, & Jian Pei et al., (2004) "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, Vol.8, No.1, pp. 53–87.

[17] Joong Hyuk Chang, (2011) "Mining Weighted Sequential Patterns in a Sequence Database with a Time-Interval Weight", Knowledge-Based Systems, Vol.24, Issue 1, pp.1-9.

[18] Ke Wang Senqiang Zhou et al., (2005) "Mining Customer Value: From Association Rules to Direct Marketing" , IEEE International Conference on Data Engineering,Vol.11, pp.57–79.

[19] Kum H C, Paulsen S & Wang W, (2005) "Comparative Study of Sequential Pattern Mining Models", Studies in Computational Intelligence: Foundations of Data Mining and Knowledge Discovery, Springer,Vol. 6, pp. 43–70.

[20] Maja Dimitrijevic, & Zita Bosnjak, (2010) "Discovering Integrating Association Rules in the Web Log Usage Data", Interdisciplinary Journal of Information, Knowledge and Management, Vol.5, pp. 191-207.

[21] Michal Munk, Jozef Kapusta, & Peter Svec , (2010)"Data Preprocessing Evaluation for Web Log Mining Reconstruction of Activities of a Web Visitor", International Conference on Computational Science, Vol.1, No.1,pp.2273–2280.

[22] Minhyuk Oh, Jongmoon Baik, Sungwon Kang & Ho-jin Choi,(2008) " An Efficient Approach for QoS-Aware Service Selection Based on A Tree-Based Algorithm" ,Seventh IEEE/ACIS International Conference on Computer and Information Science, pp.605-610.

[23] Mohammed J Zaki,(2004) " Mining Non-Redundant Association Rules", Data Mining and Knowledge Discovery,Vol. 9,Issue3,pp.223248.
[24] Mohd Helmy Abd Wahab, et al., (2008) "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science Engineering and Technology, Vol. 48, pp.190-197.
[25] Renata Ivancsy & Istvan Vajk ,(2006) " Frequent Pattern Mining in web Log Data", Acta Polytechnica Hungarica, Vol. 3, No. 1,pp.77-90.
[26] Ratnesh Kumar Jain & Suresh Jain et al., (2009) "Web Usage Mining Review", Advances in Computational Sciences and Technology, Vol. 2, No. 2, pp. 187-197.
[27] Sotiris Kotsiantis & Dimitris Kanellopoulos , (2006)"Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol.32, No.1, pp.71-82.
[28] Srivastava J, Cooley R, Deshpand M & P N Tan, (2000) "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" ACM SIGKDD Explorations, Vol.2, pp.12-23.
[29] Siriporn Chimplee & Naomie Salim et al., (2006) "Using Association Rules and Markov Model for Predict Next Access on Web Usage Mining", Advances in Systems, Computing Sciences and Software Engineering, pp.371-376.
[30] Srikant.R & Agrawal.R (1996) "Mining sequential patterns: Generalizations and performance improvements", In Proc. 6th International Conference Extending Database Technology, pp 3–17.
[31] Tarek F, Hamed Nassar, Mohamed Taha & Ajith Abraham, (2010) "An Efficient algorithm for incremental mining of temporal association rules",Data and Knowledge Engineering, Vol.69, pp800-815.
[32] Veeramalai S, Jaisankar N & Kannan A, (2010) "Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy", International journal of computer science & information Technology (IJCSIT), Vol.2, No.4, pp. 60-74.
[33] Wang W, Yang J & Yu P, (2000) "Efficient mining of weighted association rules", Proceeding of the sixth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp.270-274.
[34] Wen-Hai Gao, (2010) "Research on Client Behavior Pattern Recognition System Based On Web Log Mining", Proceedings of the Ninth International Conference on Machine Learning and Cybermetics, Vol. 1, pp.11-14.
[35] Yue Xu, Yuefeng Li & Gavin Shaw, (2011) "Reliable representations for association rules", Data Mining and knowledge Engineering, Vol.70, Issue6, pp.555-575.
[36] Yang Bin, Dong Xiangjun & Shi Fufu, (2009) "Research of WEB Usage Mining Based on Negative Association Rules", Computer Science Technology and Applications, IFCSTA, Vol.1, pp.196-199.
[37] YaJun Dua & HaiMing Li, (2010)" Strategy for mining association rules for web pages based on formal concept analysis.", Applied Soft Computing, Vol.10, Issue 3, pp.772–783.
[38] Ya-ling Tang & Feng Qin, (2010) "Research on Web Association Rules Mining Structure with Genetic Algorithm", Proceedings of the 8th World Congress on Intelligent Control and Automation, IEEE, pp.3311-3314.
[39] YongSeog Kim, (2009) "Streaming Association Rule (SAR) Mining with a Weighted Order-Dependent Representation of Web Navigation Patterns", Journal of Export Systems with Applications, Vol.36, Issue 4, pp.7933-7946.
[40] Yanxin Li, (2010) "Study on Application of Apriori Algorithm in Data Mining", International Conference on Computer Modeling and Simulation, Vol.3, pp.111-114.
[41] Yao-Tu Wang & Anthony J T Lee,(2011) " Mining Web Navigation Patterns with a Path Traversal Graph", Expert Systems With Applications, Vol.38, pp.7112-7122.
[42] Zhuo Zhang, Lu Zhang & Shaochung Zhong Jiwen Guan, (2008)"A New Algorithm for Mining Sequential Patterns. Fuzzy Systems and Knowledge Discovery", Fifth International Conference, Vol.5, pp.625.

AUTHORS

Dr. S. A. Sahaaya Arul Mary is the Professor and Head of the Department of Computer Science and Engineering at Jayaram College of Engineering and Technology, Affiliated to Anna University, Chennai. She obtained her Ph.D. in Software Testing from the Bharathidasan Institute of Technology, Trichy in the year 2009 and M.E., in Computer Science and Engineering from the Anna University, Chennai, in the year 2004. She has authored several books in Computer Science and has published many research papers in reputed journals, international and national conferences. She has to her credit several projects in Software Testing and Data mining. Her areas of interest include Software Engineering, Networks, Software Testing, Data Warehousing and Data Mining. She is guiding seven research scholars.

M.Malarvizhi has received her Master of Philosophy (M.Phil) in Computer Science from Manonmani Sundarnor University, India in the year 2003 and also her Post Graduate Degree (MCA) from Bharathidasan University, India in the year 1998. She is working as Professor in the Department of Computer Applications, J.J. College of Engineering and Technology, Trichy Tamilnadu, India. Presently she is a research scholar of Anna University Chennai. She has two international publications in her accounts. She is a keen researcher in web data mining techniques.