Integrated Web Recommendation Model with Improved Weighted Association Rule Mining

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

World Wide Web plays a significant role in human life. It requires a technological improvement to satisfy the user needs. Web log data is essential for improving the performance of the web. It contains large, heterogeneous and diverse data. Analyzing g the web log data is a tedious process for Web developers, Web designers, technologists and end users. In this work, a new weighted association mining algorithm is developed to identify the best association rules that are useful for web site restructuring and recommendation that reduces false visit and improve users' navigation behavior. The algorithm finds the frequent item set from a large uncertain database. Frequent scanning of database in each time is the problem with the existing algorithms which leads to complex output set and time consuming process. The proposed algorithm scans the database only once at the beginning of the process and the generated frequent item sets, which are stored into the database. The evaluation parameters such as support, confidence, lift and number of rules are considered to analyze the performance of proposed algorithm and traditional association mining algorithm. The new algorithm produced best result that helps the developer to restructure their website in a way to meet the requirements of the end user within short time span.

KEYWORDS

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

1. INTRODUCTION

Web data access becomes a regular activity in human life. Every user attempts are collected and stored into the web server and is known as log data. Web log data [11] contains several attributes like client address, client name, date, time, server name, server IP, status code, method name and page name. User click events are organized in this work. Web site visitors and developers should analyze the historical information to discover interesting web pages and to recommend it for the feature directions. This is an important [8] issue for researchers and active research area in the field of data mining. To achieve this problem, web log mining plays an important role in web recommendation system. The user behaviors are discovered and analyzed with the help of web log mining and it is an application of data mining techniques to discover interesting patterns.

DOI : 10.5121/ijdkp.2013.3206

Web Log mining [10,11] is a process to perform several tasks such as data collection, preprocessing, pattern discovery and pattern analysis. Web site development is incredible for every human. This technique is used to transform log data into business intelligence. Every phase perform several important tasks. In the first step, data collection different server log files are collected and convert it into a common file format. Second step, perform preprocessing to discover absolute data from raw log data. Third phase is pattern discovery to apply several techniques such as classification, clustering and association mining. It is useful to find an important pattern for web prediction. In last phase, the patterns are analyzed and the best patterns are presented. The following session described that the related work of web mining carried out by the researchers.

2. RELATED WORKS

Web usage mining [33] is most important method for predicting the user's web pages requirements by web server. The web user's wants to find the right web pages within the short duration of time .So the need of demand, the development is required to forecast the correct web pages from the web. Many techniques are implemented to analyze the web log data but the ARM is attracted by the researchers. Preprocessing performs the base work for web usage mining. This work [35] discussed the importance of preprocessing methods; various techniques are identified and compared. A complete preprocessing technique is proposed to preprocess the web log files for extraction of user patterns. Data cleaning algorithm removes the irrelevant entries from web log files and filtering algorithm discards the uninterested attributes from log file. User and sessions are identified. Sanjay Bapu et al [28] also proposed a complete preprocessing technique to access stream data. Before performing preprocessing phase log data are collected from different data sources and discover a meaningful patterns.

Web usage mining [21] extracted the valuable information from the secondary data from the user access logs. It is important for web site organization, improve business services, personalization web traffic and web recommendation. Web usage mining [29] divided into three different phases and these are discussed. Large web traffic data are identified and applied to web mining techniques for discovering an interesting pattern useful from traffic analysis.

ARM is one of the best strategies used to find out correlated itemsets purchased together frequently. It was first introduced in 1993 by Agarwal et al [2] in market basket analysis problem. Frequent patterns are patterns that appear in data set frequently. For example, a set of items such as milk and bread that are appear frequently together in the transaction data set. Ali Mirza et al [3] proposed a pruning approach using expert knowledge. This technique is used to enhancing and retaining the level of accuracy and dramatically reduces the tree size. They handle large data sets and eliminate wrong decision. Several ARM variations are proposed by the researchers in the last few years such as partitioning, Tree Projections, Markov Models FP-Growth etc. Association mining algorithm [18] is not considered the order of transactions. In many applications ordering performed the significant role. In market analysis problem the business man should know whether the customers buy some items in sequence, example buying bread first then buying milk some time later. In this work proposed a navigation order of web pages. Efficiently and accurately discovered frequent item sets in a large uncertain databases using Poisson distribution algorithm. This procedure eliminated re execution of whole algorithm and supports incremental mining. Symmetric association rule mining [2] method is an extensive method for educational web log data. Symmetric mining not only calculate support and confidence factors of Association rule mining, they also evaluated interesting measures like lift, correlation or conviction. Other relevance measures such as Chi square, cosine and contrasting rules and found that the results were learning towards a positive correlation between the item sets.

Weighted Association rule mining [36] method first proposed by Wang et al. in 2000. They have mined frequent item sets using path traversal graph using depth first search method and also maintained the order of page visit and improved the prediction accuracy. Joong et al [14] also proposed a weighted sequential pattern mining algorithm. This procedure is identified the time interval and maintains the navigation order that leads to produce interesting sequential patterns. Weighted Sequential mining application areas are identified in Web Design, Medical, Fraud Detection, Tele communication etc. Steaming Association Rule (SAR) mining algorithm [40] combined the weighted association mining and divide-and-conquer technique. Compared to traditional mining algorithm the author has to improve prediction accuracy, rule accuracy and reduce the database scanning time. It is handled large log data and eliminated redundancy that leads to navigation order of web pages.

Sequential pattern mining techniques are essential for pattern discovery phase. Sequence of web page accesses are made by different web users over a period of time. The Mohbubul et al. [19] proposed a weighted access pattern tree (WAP-tree) algorithm to access patterns effectively from log data. WAP-tree recursively calculated the trees to reconstruct an intermediate trees ie starting with suffix sequences and ending with prefix sequences. Avrillia et al. proposed [5] a new sequential mining algorithm called FLAME (Flexible and Accurate Motif Detection). It is a tree based algorithm for fast, scalable and best performance technique. This method is addressed several frequent constraints. Shrivastva Neeraj et al [30] proposed a new integrated method Closed Sequence generator mining (CSGM) to combine sequential generators and closed sequential patterns. This algorithm scanned the database only once and discovering non-redundant sequential association rules from sequential datasets with higher accuracy, less memory and time.

The subsequent section of this paper is organized as follows: Section 3 presents background study of the prediction techniques, Section 4 details the proposed methodology of integrated web Usage mining frameworks, Section 5 discusses the results obtained in proposed algorithms and Section 6 Concludes the research work.

3. BACKGROUND STUDY

Before described the elements of this work, let us first discuss the basic concepts used about the proposed methodology. After analyzing the drawbacks of a novel method can be developed. This work is integrated the important features of Association Rule Mining, Weighted Association Rule Mining, Sequential Pattern Mining and Dynamic Programming techniques.

3.1. Association Rule Mining

Association Rule Mining is one of the most important techniques in web 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:

(i). To find the frequent item sets

(ii). Determine the rules form these item sets.

Let D is a database with different transactions. D=(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 interestingness of the rule is measured by its support, confidence and lift. The evaluation factor support S is measured by the following formula.

$$\mathrm{P}(P1 \Rightarrow \mathrm{P2}) = P(P1 \cup P2) = S$$

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

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

Both support and confidence are fractions between [0,1]. The support is a measure of statistical implication, whereas confidence is used to measure the strength of the rule. The rule is said to be “interesting” if its support and confidence are greater than user defined thresholds, Support minimum and Confidence minimum respectively. 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 with a score of 1.

Confidence alone may not be enough to assess the descriptive interest of a rule. A rule with high confidence occurs with chance. Such rules can be detected by determining whether the antecedent and consequent are statistically independent.

$$Lift(P1 \Rightarrow \mathrm{P2}) = \frac{Confidence(P1 \Rightarrow \mathrm{P2})}{Support(p2)}$$

It ranges within [0,&] Values close to 1 imply that P1 and P2 are independent and the rule is not interesting. The Value far from 1 indicates that the evidence of P1 provides information about p2. These three factors are used to determine the interestingness of the rules. These measures are generally application dependent that are making use of the Pattern discovery phase.

3.2. Weighted Association Rule Mining (WARM)

Weighted association rule mining [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 is 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,…,T_n} with set of pages P={P_1,P_2,…P_n} and a set of positive real number of weights W={W_1,W_2,…,W_n} attached with each visited page P. Let's take a pattern of the form $P_1P_3P_2$. In weighted representation, this pattern is represented as P_1=3, P_3=2 and P_2=1. First visited page has the highest priority (P_1=3), intermediate pages hold highest priority less than one and last visited page contains the one (P_2=1). The visitor's web page sequences are retained based on the weights assigned to the visited web page order. The proposed algorithm Weighted Association Rule Mining is scalable and efficient in discovering significant relationships between web pages.

3.3. Sequential pattern mining

Sequential patterns [14, 25] indicated the correlation between transactions while association rule represented intra transaction relationships. Sequential pattern is a sequence of item sets that are frequently occurred in a specific order, all items in the same item sets are supposed to have the same transaction time value or within a time interval.

Let D be a database [26] of transactions, I= $\{i_1, i_2, \ldots, i_k\}$be a set of k distinct attributes called itemsets. A sequence S=<t1,t2,…,tm> is an ordered list. The length l(s) is the total number of items in the sequence. S be a set of sequences that consists of an ordered list of itemsets S1,S2….Sn. In Sequential pattern mining extracting sequential patterns whose support exceed a predefined minimum support threshold. The numbers of sequences are very large, users have different interests and requirements to find the interesting sequential patterns. Minimum support is defined by the uses. In sequential pattern mining algorithm is applied in this work.

3.4. 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 introduced the redundancy, increase processing time and memory space. This technique required exponential time for trivial computation.

To overcome these problems, Dynamic Programming is an algorithm design technique for optimization problems. Dynamic programming approach [4,6] 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).

Let the sequence $K = k_1 < k_2 < \cdots < k_n$ of n sorted keys, with a search probability p_i for each key k_i. We want to build a binary search tree (BST) with minimum expected search cost. Apply the condition actual cost is equal to number of items examined. If it is satisfied key k_i, cost = $\text{depth}_T(k_i)+1$, where $\text{depth}_T(k_i)$ = depth of k_i in BST T . When we reach the goal node, to find the cost of an optimal path and can trace back through the grid to discover the associated path. DP has both time and space complexity, storage costs for all nodes and computation cost at each node is constant.

Dynamic programming is generalized to k dimensions, where the time and space complexity is $0(l^{k})$ for a hypercube with length l. 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.

4. INTEGRATED MODEL

This work proposed a hybrid weighed association mining model that combines the strengths of Weighted association mining, dynamic programming and Improved Apriori. In this section, the hybrid weighed Association mining can be divided into several phases as shown in the figure 1.

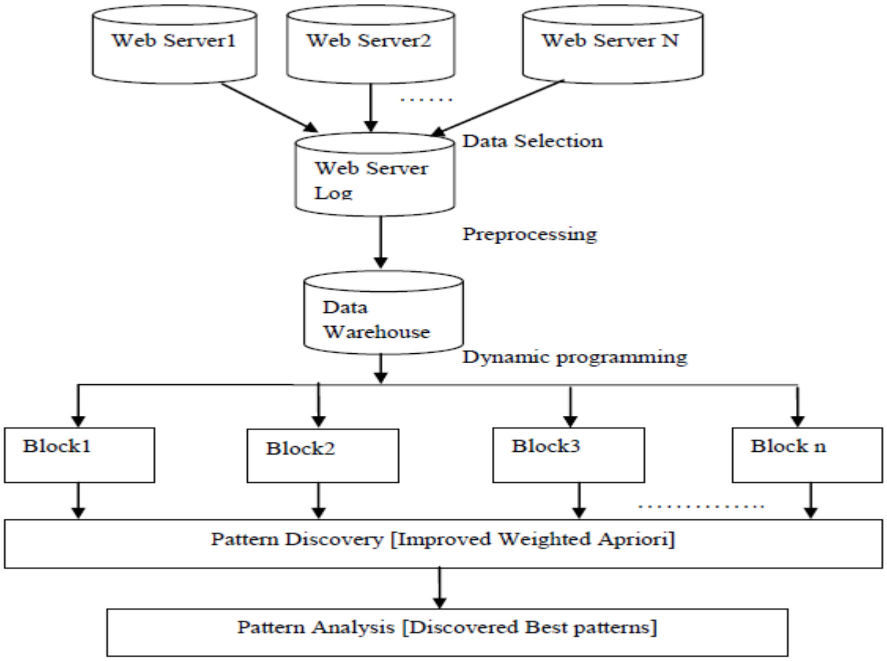


Figure 1 Web Usage Mining Process

The proposed work consists of several phases and is given below

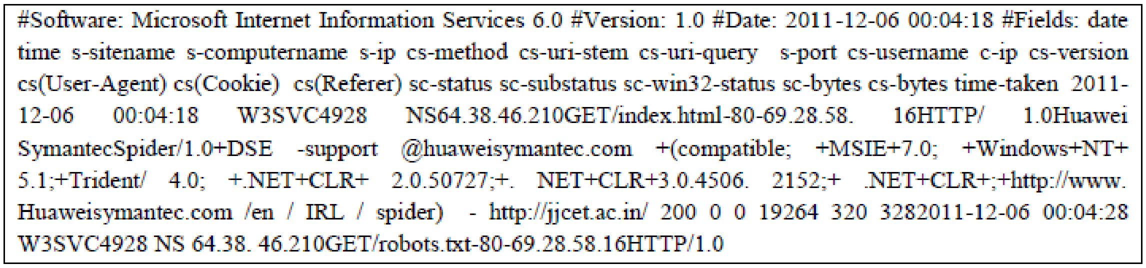
- Retrieving Web log data
- Merging
- Preprocessing
- Applying dynamic programming Technique
- Pattern discovery Phase
- Analyzing discovered Patterns

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.

4.1. Retrieving Web Log data

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. Web log data contains several attributes such as 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. Sample Server Log data are collected from Microsoft Internet Information Services. www.jjcet.ac.in website log data are collected and analyzed in this work. This work is implemented in java.

The data of the years 2009 and 2010 log data are combined to find source log data. This web site focuses on engineering education and also provided useful information about education such as teaching, research and sports etc. This Log File is the input for the pre-processing phase. Nearly 4566 visitors visited this college web site every year. In 2009, the web log data size is 48 MB and in 2010 the data size is 59 MB. So the total web log size is 107 MB. The sample data ranges from 10.09.2010 to 30.09.2010 from the file are taken which contains 21 days web log data. The total number of records found is 1,47,273 from this sample. The results of the preliminary analysis of these log files are reported in table 1, which shows that the information provided by the data summarization module that contains total number of records and size of individual log files.



```
#Software: Microsoft Internet Information Services 6.0 #Version: 1.0 #Date: 2011-12-06 00:04:18 #Fields: date
time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query  s-port cs-username c-ip cs-version
cs(User-Agent) cs(Cookie)  cs(Referer) sc-status sc-substatus sc-win32-status sc-bytes cs-bytes time-taken  2011-
12-06    00:04:18    W3SVC4928    NS64.38.46.210GET/index.html-80-69.28.58.    16HTTP/    1.0Huawei
SymantecSpider/1.0+DSE   -support   @huaweisymantec.com   +(compatible;   +MSIE+7.0;   +Windows+NT+
5.1;+Trident/   4.0;   +.NET+CLR+   2.0.50727;+.   NET+CLR+3.0.4506.   2152;+   .NET+CLR+;+http://www.
Huaweisymantec.com  /en  /  IRL  /  spider)   -  http://jjcet.ac.in/  200  0  0  19264  320  3282011-12-06  00:04:28
W3SVC4928 NS 64.38. 46.210GET/robots.txt-80-69.28.58.16HTTP/1.0
```

Figure 2 Sample Web Server Log

4.2. Merging

All the log file entries from 10.09.2010 to 30.09.2010 are put together into a single log file that contains the entire preprocessed requests. The combined or merged log file of size 1,82,534 records are arranged in ascending order based on time stamp of each record (or date). All the Log Files (LF) is collected from different sources and is put together in joint log file L. The set of joint

Log files LF={l_1, l_2, l_3,…….,l_m} are merged into a single log file L. In this work, Log files for 21 days are joined together to construct a new Log file that is the input of data cleaning process.

4.3. 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 user belongings to different sessions 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. In data cleaning step, from the 1,47,273 records, there are 38,676 image records, 7,684 failed request and 917 incomplete requests which are eliminated. Nearly 80% of records are found to be unnecessary and the remaining 20% of suitable records are used in pattern discovery process. Finally 33,004 useful sample datasets are identified in this phase.

4.4. Applying dynamic programming technique

Association Rule Mining with Apriori and Weighted Association Rule Mining with improved Apriori techniques have been applied for 33,004 sample log data. These precise resources play a major role for the web site analysis. It seems to be tedious task to present the output for the 33,004 records. So the work is illustrated by considering 20 sample data from the entire data set. 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 Binary and 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. Table 1 is formulated based on the web log data from fig.2

Table1: Binary and Weighted Order Representation of Navigation Patterns

Blocks	Name of Transaction	Web Pages Visited order	Weighted order Representation			Binary Representation		
			P_1	P_2	P_3	P_1	P_2	P_3
B1	T1	$P_3 \rightarrow P_2 \rightarrow P_1$	1	2	3	1	1	1
	T2	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1	1	1	1
	T3	$P_1 \rightarrow P_2$	3	2	0	1	1	0
	T4	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1	1	1	1
	T5	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1	1	1	1
B2	T6	$P_2 \rightarrow P_3$	0	3	2	0	1	1
	T7	$P_1 \rightarrow P_3$	3	0	2	1	0	1
	T8	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1	1	1	1
	T9	$P_1 \rightarrow P_2$	3	2	0	1	1	0
	T10	$P_2 \rightarrow P_1$	2	3	0	1	1	0
B3	T11	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1	1	1	1
	T12	$P_3 \rightarrow P_2 \rightarrow P_1$	1	2	3	1	1	1
	T13	$P_2 \rightarrow P_3$	0	3	2	0	1	1
	T14	$P_3 \rightarrow P_2 \rightarrow P_1$	1	2	3	1	1	1
	T15	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1	1	1	1
B4	T16	$P_1 \rightarrow P_2 \rightarrow P_3$	3	2	1	1	1	1
	T17	$P_2 \rightarrow P_3$	0	3	2	0	1	1
	T18	$P_1 \rightarrow P_2$	3	2	0	1	1	0
	T19	$P_3 \rightarrow P_2 \rightarrow P_1$	3	2	1	1	1	1
	T20	$P_1 \rightarrow P_2$	3	2	0	1	1	0

4.5. Pattern Discovery

Apriori algorithm is the best known algorithm for 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= $\{w_1, w_2, w_3 \ldots wn\}$. The algorithm in section.4.5.1 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 4.5.2. The output of the WApriori is stored in Optimal Table and from that an optimized solution is generated.

4.5.1. 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.Initialize 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 $R_b$
```

Table 2 shows the computational result of support, confidence and lift 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, Confidence and Lift values

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

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.

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

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

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
```

```
if $N_{o(1)}$ < min-sup then
        Prune $N_{o(1)}$
else
        Add to list $L_2$
End if
Discover 3-page-visit, 4-page-visit etc. until found all frequently visited pages ($F_p$)
Return $F_p$.
```

4.6. Analyzing discovered Patterns

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, Confidence and Lift values

Blocks	$P_1 \rightarrow P_2$			$P_1 \rightarrow P_2 \rightarrow P_3$		
	Support	Confidence	Lift	Support	Confidence	Lift
B_1+B_2	3/5	2/3	9/10	2/3	2/3	1
B_2+B_3	2/5	1/2	4/5	3/10	2/3	9/20
B_3+B_4	1/2	5/8	4/5	3/10	1/2	3/5

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, confidence and Lift 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, Confidence and Lift values

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

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, Confidence and Lift values

Block No.	No. of Rules	Name of the Rules	Rule Count	Support	Confidence	Lift
1	1	$P_1 \rightarrow P_2$	4	4/5	4/5	4/5
	2	$P_2 \rightarrow P_3$	3	3/5	3/5	3/4
	3	$P_2 \rightarrow P_1$	1	1/5	1/5	1/5
	4	$P_3 \rightarrow P_2$	1	1/5	1/4	1/4
2	1	$P_1 \rightarrow P_2$	2	2/5	2/4	5/8
	2	$P_1 \rightarrow P_3$	1	1/5	1/4	5/12
	3	$P_2 \rightarrow P_3$	2	2/5	2/4	5/6
	4	$P_2 \rightarrow P_1$	1	1/5	1/4	5/16
3	1	$P_1 \rightarrow P_2$	2	2/5	2/4	1/2
	2	$P_2 \rightarrow P_3$	3	3/5	3/5	3/5
	3	$P_2 \rightarrow P_1$	2	2/5	2/5	1/2
	4	$P_3 \rightarrow P_2$	2	2/5	2/5	2/5
4	1	$P_1 \rightarrow P_2$	3	3/5	1	1
	2	$P_2 \rightarrow P_1$	1	1/5	1/5	1/4
	3	$P_2 \rightarrow P_3$	2	2/5	2/5	2/3
	4	$P_3 \rightarrow P_2$	1	1/5	1/3	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, Confidence and Lift values

Block No.	No. of Rules	Name of the Rules	Rule Count	Support	Confidence	Lift
1	1	$P_1 \rightarrow P_2 \rightarrow P_3$	3	3/5	3/ 4	5/4
2	1	$P_1 \rightarrow P_2 \rightarrow P_3$	1	1/5	1/3	5/6
3	1 2	$P_1 \rightarrow P_2 \rightarrow P_3$ $P_3 \rightarrow P_2 \rightarrow P_1$	2 2	2/5 2/5	1 1	5/3 5/2
4	1	$P_1 \rightarrow P_2 \rightarrow P_3$	1	1/5	1/3	5/6

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 confidence and lift values.

Table 7: Final navigation table with support, confidence and Lift values

No. of Rules	Name of the Rules	Rule Count	Support	Confidence	Lift
1	$P_1 \rightarrow P_2 \rightarrow P_3$	7	7/20	7/11	7/11
2	$P_3 \rightarrow P_2 \rightarrow P_1$	2	1/10	1/2	1/2

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%), confidence value is 7/11 (64%) and lift (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	Rules	B1		B2		B3		B4	
		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 $P_1 \rightarrow P_2 \rightarrow P_3$.

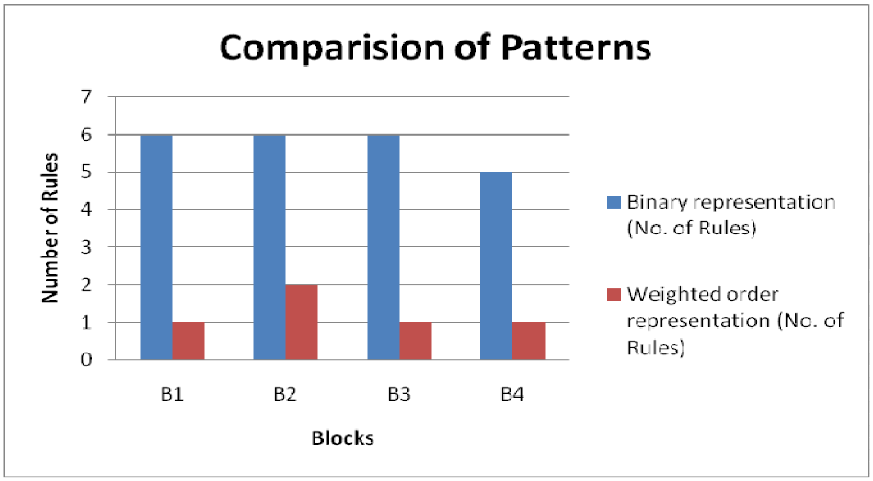


Figure 3: Rule Comparison of Binary and weighted order representation

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

The proposed technique effectively captures the navigation patterns and efficiently reduces the search time and space. It is experimentally proved that the proposed system is a new web recommendation system. In this research work, the parameters such as support, confidence, lift and number of rules that are essential in web page design have been significantly improved. The experimental result proves that this method guarantees 35% in support, 63% in confidence, 64% in lift, and number of rules which is better than the conventional association rule mining models. The proposed technique provides more significance to the order of user's visit, which is more helpful to the users and developers.

REFERENCES

[1] Agathe Merceron & Kalina Yacef, (2007) "Revisiting interestingness of strong symmetric association rules in educational data", *Proceedings of the International Workshop on Applying Data Mining in e-Learning*, pp. 3-12.

[2] 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.

[3] Ali Mirza Mahmood & Mrithyumjaya Rao Kuppa,(2012) " A novel pruning approach using expert knowledge for data-specific pruning", *Engineering with Computers* , Vol.28, pp.21–30, 2012.

[4] 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.

[5] Avrilia Floratou, Sandeep Tata & Jignesh M. Patel,(2011) " Efficient and Accurate Discovery of Patterns in Sequence Data Sets", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 23, No. 8, pp. 1154-1168.

[6] Kerf, Richard, (1985) " Depth-First iterative-deepening : An optimal admissible tree search", *Artificial Intelligence*,Vol. 27, Issue 1, pp.97-109.

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

[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] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Lei Shi, Shen Guo, Deyu Qi & Fufang Li, (2006)“Appling Association Rule to Web Prediction, *Computer and Computational Sciences*, Vol. 2, pp. 522-527.

[18] Liang Wang et al., (2012) “Efficient Mining of Frequent Item Sets on Large Uncertain Databases”, *IEEE Transactions on Knowledge and Data Engineering*, pp 2170-2183, Vol. 24, No. 12.

[19] Mahbubul Arefin Khan et al., (2012) “Pattern Finder–Efficient Framework for Sequential Pattern Mining”, *4th International Conference on Computer Modeling and Simulation*, pp.130-133, vol.22.

[20] Nanopoulos and Y. Manolopoulos,(2000) "Finding generalized path patterns for web log data mining," *Proceedings of the East-European Conference on Advances in Databases and Information Systems Held Jointly with International Conference on Database Systems for Advanced Applications*, pp. 215-228,Springer- Verlag.

[21] Nithya.P et al ,(2012) “A Survey on Web Usage Mining: Theory and Applications”, *International Journal Computer Technology & Applications*,Vol.3 ,issue 4,pp.1625-1629.

[22] Pinar Senkul & Suleyman Salin, (2012) “Improving pattern quality in web usage mining by using semantic information”, *Knowledge Information System*, vol. 30, pp.527–541.

[23] Pinar Senkul & Suleyman Salin, (2012) “Improving pattern quality in web usage mining by using semantic information”, Knowledge and Information Systems, Vol. 30,Issue 3, pp 527-541.

[24] Priyanka Tiwari & Nitin Shukla ,(2012) “ Multi dimensional sequential pattern mining”, *International Journal of Scientific and Research Publications*, Vol.2, Issue 4, pp.1-3.

[25] Qiankun Zhao & Sourav S. Bhowmick , (2003) “Sequential Pattern Mining: A Survey”, *Technical Report, CAIS*, Nanyang Technological University, Singapore, pp 1-27, 2003.

[26] Rajashree Shettar , (2012) “Sequential Pattern Mining From Web Log Data” , *International Journal Of Engineering Science & Advanced Technology*, Vol. 2, Issue-2, pp.204 – 208.

[27] Renata Ivancsy & Istvan Vajk ,(2006) “ Frequent Pattern Mining in web Log Data”, *Acta Polytechnica Hungarica*, Vol. 3, No. 1,pp.77-90.

[28] Sanjay Bapu Thakare et al., (2010) “A Effective and Complete Preprocessing for Web Usage Mining”,*International Journal on Computer Science and Engineering*, Vol. 02, No. 03, pp. 848-851.

[29] Shakti kundu, (2012) “An Intelligent Approach of Web Data Mining*”, International Journal on computer science and engineering*, Vol. 4,No. 05, pp.919-928.

[30] Shrivastva Neeraj & Lodhi Singh Swati, (2012) “Approach to Recover CSGM Method with Higher Accuracy and Less Memory Consumption using Web Log Mining”, *Journal of Engineering Sciences,* Vol.1,issue 1, pp.83-87.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Vijayashri Losarwar & Madhuri Joshi, (2012) “Data Preprocessing in Web Usage Mining” *, International Conference on Artificial Intelligence and Embedded Systems* , pp.1-5, Singapore.

[36] 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.

[37] 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.

[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] Yanxin Li, (2010) "Study on Application of Apriori Algorithm in Data Mining*", International Conference on Computer Modeling and Simulation*, Vol.3, pp.111-114.

[40] 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.

[41] Yue Xu, Yuefeng Li & Gavin Shaw, (2011) "Reliable representations for association rules", *Data Mining and knowledge Engineering,* Vol.70, Issue6, pp.555-575.

[42] Zhiguo Zhu & Liping Kong,(2010) " A Design For Architecture Model Of Web Access Patterns Mining System ",*IEEE International Conference on Computer and Communication Technologies In Agriculture Engineering*, pp.288-292.

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. 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.