Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110129

Discovery of Effective Relationships among Web Pages Using Hybrid Weighted PSO

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

[1] Sapthagiri College of Engineering, Periyanahalli, Dharmapuri, Tamilnadu, India.
[2] Jayaram College of Engineering and Technology, Pagalavadi, Tamilnadu, India.

Received: 7 Jul. 2016, Revised: 19 Oct. 2016, Accepted: 22 Oct. 2016
Published online: 1 Jan. 2017

Abstract: With the explosive growth of resources in the World Wide Web, web users are in need to make use of the computerized tools to discover the required resources, to find and analyze their usage patterns. The solutions are generated by traditional techniques but these solutions are not an optimized pattern. Several researchers focused on this issue, but still there is no established solution. Traditional and non-traditional techniques are utilized to generate the Usage patterns, but the solution generated is not in a form suitable to retrieve the required information. From the literature, it is found that Particle Swarm Optimization (PSO) is used in various fields to yield the global optimized solution. In this work, PSO is applied to generate an optimized rule pattern to retrieve the necessary resources from the web. The main objectives of this work is to generate an optimized rule sets and the results of the PSO have to be compared the objective performance measures such as Support, Confidence, Correlation and Lift values with existing traditional and non-traditional techniques to proposed the best technique

Keywords: Association rule Mining, Weighted Association Mining, Dynamic Programming, PSO, Web Knowledge discovery

1 Introduction

With the increase in growth of internet, pattern mining plays a most important role to prevent unexpected outcomes of web users visits. Mining these association relationships is not an easy task due to the complexity with log data and the irregular page visits of users. In order to overcome these problems, an efficient algorithm is needed to mine their causal relationships among pages in efficient manner. In this work, the log data is initially partitioned using dynamic programming technique and the frequent patterns are found by means of weighted association rule mining algorithm and applied the PSO to optimize the associations among web pages in a well-organized way.

2 Literature review

In this section, we focus on the prior research on web usage mining techniques. The web usage mining have been suggested to analyze the large volume of web usage data, it is a very important problem for web users and has attracted much attention by many researchers.

Determine unknown and useful information from Web log data is called Web usage mining. Discover a frequent pattern from web log data is to acquire information about the navigational behavior of the web users.

The time interval in sequence database is proposed [1] using weighted sequential pattern mining technique. Based on the importance of the user needs the weightage is given to web pages and obtain the best patterns from the sequence data base.

The best association rules are found by Streaming association rule (SAR) [2] and it is used to find the correct order by using weighted association rule mining technique. Proposed a new weighted association mining technique based on visited order and integrates divide and conquer technique. Data base scanning time and redundancy were reduced in this technique.

An optimized the fuzzy parameters using swarm intelligence based on Exponential Particle Swarm Optimization (EPSO) [3] has been proposed to form a new optimization technique. It is improved the Mining Infrequent Causal Associations among Drugs and their

* Corresponding author e-mail: malarbas@yahoo.co.in

reactions of selected patient records. It is also improved the identification or detection ADR results the causal and uncausal relationship of the drug and symptoms of the patients are mined exactly.

A rule filtration technique is [4] to discover the best trends of the year and improved the web site ranks, it is guided to increase the web site visitors. The best frequent patterns is found by applying Apriori Algorithm and also applied two more optimization algorithms one is Ant Colony optimization (ACO) and another is Particle swarm Optimization(PSO), to found the month wise optimal search patterns.

Apriori algorithm and modified PSO [5] algorithm are to provide feasible threshold values for minimal support and confidence and also find the best frequent patterns. It is analyzed and present an important rules discovered in databases with the help of interestingness measures. In each generation reserved a pbest and gbest values to generating a new populations for next generation. Chaotic PSO [6] algorithm was to modify the velocity function to introduced chaotic operators. This work main focus is to improve the computational efficiency and accuracy of mining rules. The results of PSO and CPSO algorithm are compared to achieve the better prediction accuracy.

HPSO-TS-ARM [7] has been proposed by using web association rule mining. PSO will fetch the web search data in its optimized form, which is further computed by Tabu Search to prepare balance data arrangement followed by Association rule mining on processed web search data. Better fitness values with less elapsed time that proposed algorithm would perform better results.

Hybrid GA/PSO [8] was to obtained better results for web structure mining. This work is eliminated unrelated search results to obtain quality web links and also handled high dimensional order clustering of web search.

A hybrid Genetic-PSO [9] algorithm is used to find membership functions which are suitable for mining problems by a strong cooperation of GA and PSO. It also used in fuzzy data quantitative transactions. These algorithms is integrated with two techniques for entire run of a simulation in each of the iteration, a part of population are substituted by new ones generated by means of GA, while in other part is the same of previous generation but moved on the solution space by PSO. At the end, best final sets of membership functions in all the populations are gathered to be used for mining fuzzy association rules.

Based on the literature survey, it is identified that the PSO is not used to generate the optimized rule set with weighted order association mining. In this work, PSO is used to generate optimized rule set in weighted order and the outcome is analyzed with other traditional and non-traditional techniques.

3 Problem description

Web Log data are massive amount of uncertain, redundant and incomplete data [15]. Web log data are pre-processed before apply the Web data mining. The sample log data are taken from the literature proposed by Kim [2]. The msnbc sample log data describes the page visits of users who visited msnbc.com. The page visits are recorded by visited order. The sample web log data are taken from msnbc.com anonymous web log data and the data types are in discrete sequence order. The msnbc.com visited web page clusters are associated with an integer number starting from 1 to 17.

Table 1: First 10 lines of the ASCII data file.

1	1	4	5	7	17	1	1	2	2	2	3	3
2	1	1	2	2	4	4	4	17	17	17		
3	2	2	4	2	2	2	3	3	3	3	3	3
5	8	8	8	16	16	16	16					
1	17	17	17	12	12	12	12					
6	13	1	14	14	14	12						
1	1	2	2	2	2	2	2					
6	7	7	7	6	6	8	8	8	8			
6	9	4	4	4	10	3	10	5	10	4	4	4
12	12	17	17	17	16	16						

Web log sample are sequence of visits to find the associations between web pages on the basis of sequential patterns over a period of time. Each sequence of page visit is represented as an ordered list of numbers and each number represents one of the possible clusters of web page requested by the user. Duplicate reference to a page in a web access sequence is removed.

First line of the table shows the access sequence 1 1 4 5 7 17 1 1 2 2 2 3 3, in which 1, 2 and 3 are three duplicate web page visits are eliminated. Each row in Table 1 shows the visited order of web pages by the web users.

4 Rule generation techniques

The Weighted Association Rule Mining (WARM) technique generalizes the association rule mining by assigning weights to the visited pages based on visited order. It is one of the rule generation techniques which are used in this work to generate frequent patterns for predicting the user needs such as to reduce execution time and eliminates redundancy. The frequent patterns are evaluated with the help of the factors such as support, confidence, correlation and lift values. Support is a primary measuring factor to decide whether a set of web pages occurs repeatedly in a web log data or not. The support factor is used to determine the associated web page frequency. Confidence is another significant factor in

rule generation and it is used to decide whether a derived rule is interesting or not.

These two measures are insufficient to sort out uninteresting rules, it does not assess the real strength of associated web pages. To overcome this weakness, a correlation is used to enhance the support-confidence framework for association rules. Correlation is a symmetric measure. A correlation around 0 indicates that X and Y are not correlated, a negative figure indicates that P_1 and P_2 are negatively correlated and a positive figure that they are positively correlated. Note that the denominator of the division is positive and smaller than 1.

Lift is correlation measure between the two visited pages. If the resulting value is less than 1, then occurrence of one web page is negatively correlated with occurrence of another web page. If the resulting value is greater than 1, then two pages are positively correlated. If the resulting value is equal to 1, then two web pages are independent. These four are the important measuring parameters used to decide whether a set of web pages are interesting or not.

5 Optimization techniques

The optimization Research is a scientific approach to solve the mathematical models and the researchers are used different techniques for solving their mathematical models. Based on the objective function and the solution accuracy, the suitable optimization techniques are utilized for web optimization prediction. The proposed work utilizes two major optimization techniques Dynamic Programming and PSO framework.

5.1 Dynamic Programming

Web growth is very rapid and it is a major source of information provider. Also, massive amount of log data is generated due to user interactions with web sites. To utilize and solve the entire log data at once is impossible in web mining techniques. So a new technique is required to divide the log data into several components. Dynamic programming is an optimization technique to solve a complicated problem by split into a collection of smaller sub problems, solve each sub problems just once, and storing their solutions using a memory-based data structure. The same sub problem occurs again, instead of resolve its problem, acquired the previously computed solution, and as a result computation time is reduced.

Dynamic programming is a stage-wise development technique suitable for optimization problems whose solutions may be viewed as the result for a series of conclusions. The most attractive property of this strategy is that during the search for a solution it avoids full enumeration by pruning early partial decision solutions that cannot possibly lead to optimal solution. So, Dynamic programming technique is used in the proposed methodology and it will be improved the time and space efficiency.

5.2 Particle Swarm Optimization (PSO)

Particle swarm optimization is one of the non-traditional techniques, which is motivated by the social behaviour of fish schooling or birds flocking for food. The PSO is established to solve various continuous functions by Kennedy and Eberhart. It is the fast convergence technique as compared to other optimization techniques and is easy to implement all the applications when compared with other evolutionary algorithms while only a few parameters need to be adjusted.

Initially, the particles are randomly generator based on the parameters bounds and they are updated by velocity function. The particles are updated by two best values i.e. "pbest"(p_{ij}) and "gbest"(p_{gi}). The new particles are identified its velocity and the corresponding equations are given below:

$$\begin{aligned} v_{ij}(t+1) &= v_{ij}(t) + c_1 r_1 (p_{ij}(t) - x_{ij}(t)) \\ &\quad + c_2 r_2 (p_{gj}(t) - x_{ij}(t)) \\ x_{ij}(t+1) &= x_{ij}(t) + v_{ij}(t+1) \end{aligned}$$

where $i, j = 1, 2, \ldots, d$ and c_1, c_2 are constants (1 to 4) and r_1, r_2 are random numbers in the range [0, 1]. The general structure of PSO algorithm is shown in Fig. 1

5.3 Algorithm

Step 1. Particles are randomly generated for initial population.

Step 2. Find the objective function for each particle that is called as fitness value. The best fitness value is selected in each particles i.e. p_{best} in history. Set current value as the new p_{best}.

$$\begin{aligned} \text{Fitness}(x) = \big(&x_1 \times \text{support}(x) + x_2 \times \text{confidence}(x) \\ &+ x_3 \times \text{correlation}(x) + x_4 \times \text{lift}(x)\big) \\ &/(x_1 + x_2 + x_3 + x_4) \end{aligned}$$

where x_1, x_2, x_3, x_4 are support, confidence, correlation and lift weights.

Step 3. Select the best fitness value of all the particles and it is called as the g_{best}.

Step 4. For each particle, velocity function is calculated by using the following equation

$$\begin{aligned} v[] = v[] &+ c_1 * rand() * (p_{\text{best}}[] - present[]) \\ &+ c_2 * rand() * (g_{\text{best}}[] - present[]) \end{aligned}$$

and $present[] = present[] + v[]$

where $v[]$ = particle velocity,

$present[]$ = current particle,

$rand()$ = random number between 0 to 1,

c_1, c_2 = learning factors.

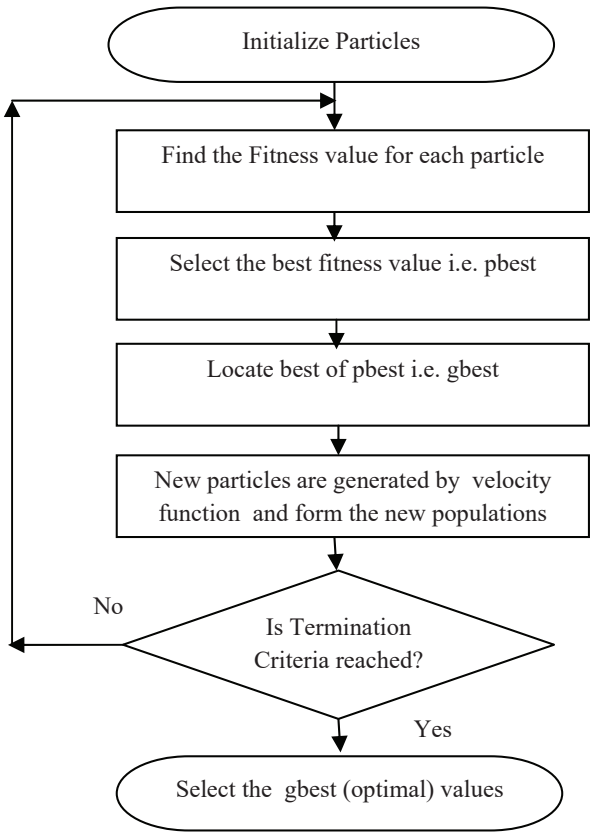

Fig. 1: General structure of PSO.

Step 5. Maximum velocity V_{max} calculated by sum of each particle velocity.
If the sum of the velocity on that dimension to cross V_{max} (specified by the user) and it is limited to Vmax.
Step 6. Terminate the procedure if the condition is satisfied.
Step 7. Else, goto Step 2.

6 Proposed methodology

The general architecture of this proposed system is given in Fig. 2

The important part of the proposed system is web log data collection, which contain the entire successful visit made on the web while browsing. The browsing data are automatically accumulated and can be obtained from web server. Collected web log file are preprocessed and implemented with the help of preprocessing technique. In the preprocessing stage to eliminate unnecessary, incomplete page visits of log files. Cleaned log data are in 0 and 1 format that can be converted into weighted log data.

Table 2: Initial parameter settings about GA and PSO.

GA	PSO
Population size : 100	Population size : 100
No. of generations: 1000	No. of iterations: 1000
Length of the Chromosome: 50	Acceleration coefficient
Selection operator: Roulette wheel	C_1 : 0.5
Cross over operator: Single point	C_2 : 0.5
Cross over probability: 0.7	Fitness function:
Mutation probability: 0.05	Maximize $f(x)$
Fitness function : Maximize $f(x)$	

In 8,00,000 web log data are taken for further processing. Weighted association mining technique is applied and to find frequent item sets from that web log data. Frequent item sets and the corresponding support, confidence, correlation and lift values are given as the input of the PSO algorithm. Finally PSO algorithm produces the optimal solution.

7 Experimental results and discussion

This work combines different techniques to create a new hybrid approach for getting better results. It avoids limitations of the existing individual technique. In this work, the existing web mining method of weighted association rule mining is to describe, analyze, implement and upgrade. The Weighted Association Rule Mining and PSO algorithm are combined to discover the optimal solution.

The datasets considered for the experiments is taken from literature 8,00,000 web log data and initial parameter values for PSO and GA set is listed in Table 3 The initial velocity set was 0 for all the datasets and the learning factors c1 and c2 between 0 to 1. Maximum number of iterations carried out is 1000 and the population size is 100.

The fitness of the particle is calculated by using following five formulas.

1. Support $(X \to Y) = P(X \cap Y)$
2. Confidence $(X \to Y) = P(X \cap Y)/P(X)$
3. Correlation
$$(X \to Y) = \frac{P(X,Y) - P(X) - P(Y)}{\sqrt{(P(X)P(Y)(1-P(X))(1-P(Y)))}}$$
4. Lift$(X \to Y)$ = Confidence$(X \to Y)$/Support(Y)
5. $F(X) =$
$$\frac{X_1 \times \text{Support}(X) + X_2 \times \text{Confidence}(X) + X_3 \times \text{Correlation}(X) + X_4 \times \text{lift}(X)}{X_1 + X_2 + X_3 + X_4}$$

Fig. 3 shows the screen shot of the PSO program. This work is implemented by visual Basic the results are stored in a database.

The Table 3 shows that the sample of 50 best iterations results obtained by PSO. The particles are updated by Present best (P-best) and Global best (G-best) values obtained in each and every iteration. The objective

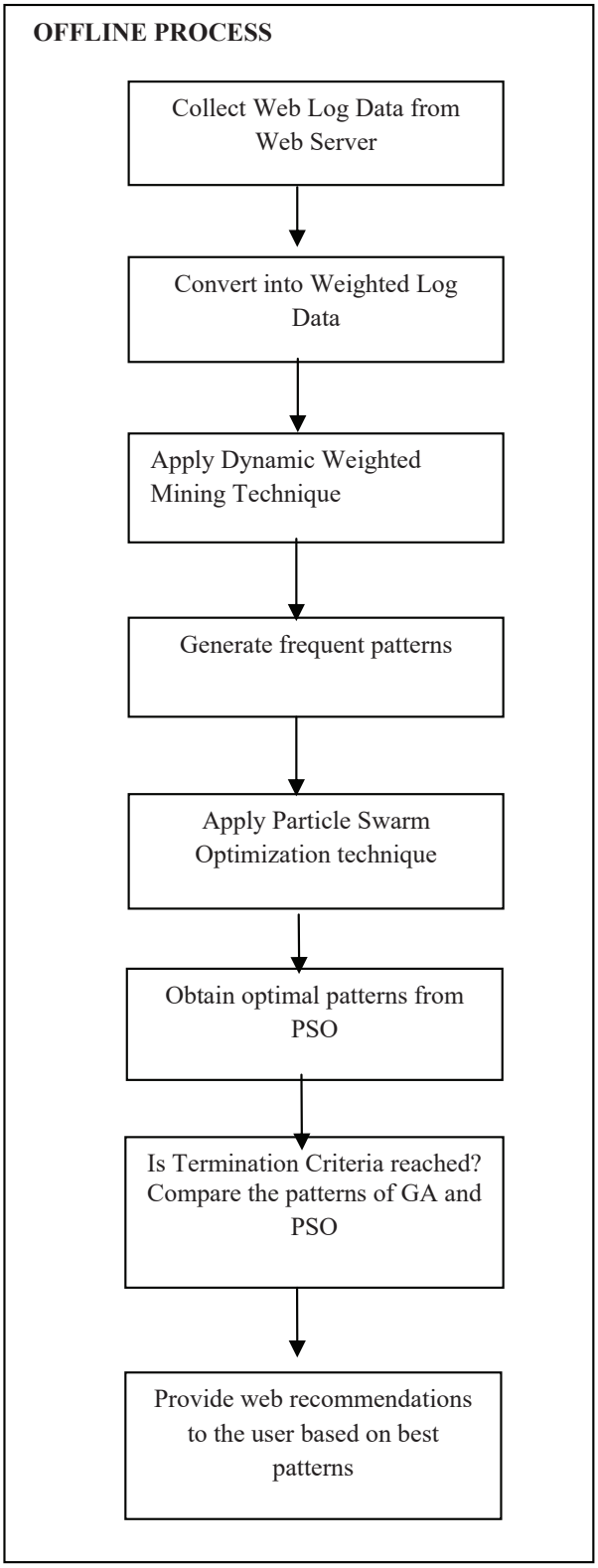

Fig. 2: Schema diagram for proposed methodology.

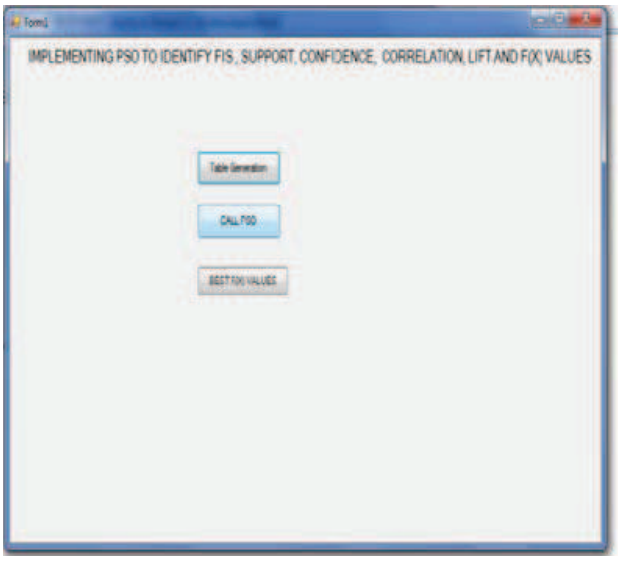

Fig. 3: Screen shot.

function is maximize the fitness value and it is displayed in the last column of Table 3 The global best value in current iteration is compared with previous best values obtained in preceding iterations and the best fitness value is updated. The PSO is executed for 1000 iterations and the overall best fitness value is selected as optimal fitness value.

Table 3: Best 10 results of PSO approach.

No	Frequent Item sets	Support	Confi-dence	Correl-ation	Lift	$F(x)$
1	73	0.73	0.89	2.48	2.62	1.68
2	80	0.8	0.98	2.21	2.12	1.53
3	46	0.46	0.56	2.02	2.95	1.50
4	55	0.55	0.9	1.34	2.37	1.29
5	42	0.42	0.51	1.51	2.33	1.19
6	26	0.26	0.9	0.75	2.64	1.14
7	39	0.39	0.48	1.28	2.16	1.08
8	16	0.16	0.55	0.59	2.9	1.05
9	29	0.29	1	0.69	2.17	1.04
10	38	0.38	0.79	0.81	2.08	1.02

Although there are many parameters in PSO and GA, in this section Number of transactions, support, confidence lift, correlation and fitness function are investigated to observe the changes on optimal solutions.

Both algorithms are compared on optimization of the membership function. There are 1000 iterations were executed for each algorithm and the results are compared. The GA has taken more computational time in all iteration as compared to PSO but it has yielded best result and is given in Table 4

Table 4: Comparison of Weighted GA and Weighted PSO approach.

Descriptions	Weighted GA	Weighted PSO
Number of transactions per data set	8,00,000	8,00,000
Support percentage	80	99
Confidence percentage	97	45
Lift Percentage	57	53
Correlation Percentage	47	38
$F(x)$	8.8	5.6

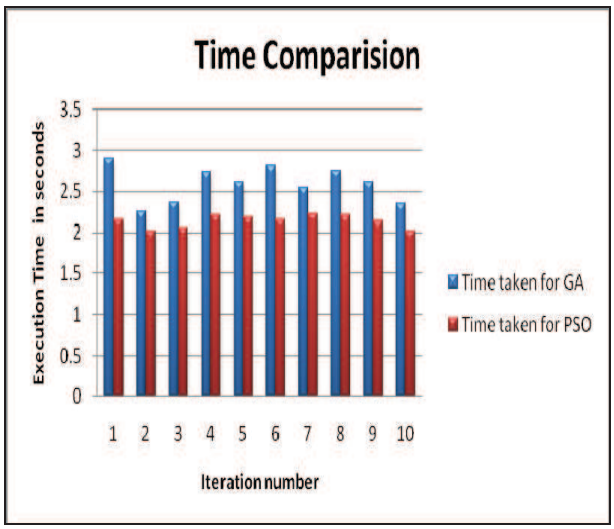

Fig. 4: Execution time comparision of Weighted GA and Weighed PSO.

In this work, the useful association rules are retrieved for improving the prediction accuracy and also the rule set size are reduced. The fitness value of the problem is improved in succeeding iterations and it is stopped when a stagnant value is obtained continuously. User navigation pattern prediction system was presented in this work which is predicting the user future request.

8 Conclusion

This work proposed a hybrid PSO technique to enhance the prediction accuracy of weighed association mining. The hybrid PSO algorithm is more efficient which takes less computational time and maintain a high confidence, also providing the user with high quality rules when compared to traditional methods. The selected rules are increased to 15% support and confidence, lift and correlation values. The reduced frequent patterns are arranged in ascending order based on the fitness values and the best 15% patterns are selected for web recommendation. Based on the results obtained in experiments, the Weighted PSO's performance is significantly improved and provided the best performance when compared with Weighted GA.

References

[1] Joong Hyuk Chang, "Mining weighted sequential patterns in a sequence database with a time-interval weight", Knowledge-Based Systems, **24(1)**, 1–9, (2011).
[2] Yong Seog Kim, "Streaming Association Rule (SAR) mining with a weighted order-Dependent Representation of Web Navigation Patterns", Expert Systems with Applications, **36(4)**, 7933–7946, (2009).
[3] M. Poorani, P. Nithya and B. Umamaheshwari, "A Method For Mining Infrequent Causal Associations With Swarm Intelligence Optimization For Finding Adverse Drug Reaction", International Journal of Computing, Communications and Networking, **3(1)**, 25–32, (2014).
[4] Prashant Sharm, "An efficient Optimization technique for Web Log Data based on ACO and PSO", International Journal of Modern Computer Science (IJMCS), **2(6)**, 56–62, (2014).
[5] T. Bharathi and P. Krishnakumari, "An Enhanced Application of Modified PSO for Association Rule Mining", IJCSI International Journal of Computer Science, **10(4)**, 49–55, (2013).
[6] K. Indira and S. Kanmani, "Enhancing particle swarm optimization using chaotic operators for association rule mining", Computer Science and Engineering, **46**, 8563–8566, (2012).
[7] A. Parmjeet Kaur, Usvir Kaur and Dheerendra Singh, "Fast and Robust Hybrid Particle Swarm Optimization and Tabu Search Algorithm for Web Data Association Rule Mining", International Journal of Current Engineering and Technology, **4(5)**, 3225–3228, (2014).
[8] B. Rajdeepa And P. Sumathi, "Web Structure Mining For Users Based On A Hybrid GA/PSO Approach", Journal of Theoretical and Applied Information Technology, **70(3)**, 573–578, (2014).
[9] Amin Jourabloo, "Genetic-PSO Fuzzy Data Mining With Divide and Conquer Strategy", International Conference on Artificial Intelligence, ICAI 2011, Las Vegas Nevada, USA.
[10] Veenu Mangat and Renu Vig, "HybridMiner: Effective Rule Mining based on a Hybrid Dynamic Swarm Method", International Conference on Artificial Intelligence, **2**, 232–237, (2013).
[11] M. SelviMohana, B. Rosiline Jeetha, "Methodologies on User Behavior Analysis and Future Request Prediction in Web Usage Mining using Data mining Techniques", International Journal of Data Mining Techniques and Applications, **3**, 369–373, (2014).
[12] R. Gobinath and M. Hemalatha, "Visualizing the Navigational Patterns from Web Log Files using Web Mining Applications", International Journal of Scientific & Engineering Research, **5(5)**, 538–544, (2014).
[13] Ivan S. Mitzev and Nickolas H. Younan, "Time Series Shapelets: Training Time Improvement Based on Particle Swarm Optimization", International Journal of Machine Learning and Computing, **5(4)**, 283– 287, (2015).

[14] Randhir, Hemant N, Ravindra Gupta, and G.R. Selokar. ”Extract Knowledge and Association Rule from Free Log Data using an Apriori Algorithm”, International Journal of Advanced Computer Research (IJACR), **3(12)**, 191–196, (2013).

[15] Lu LIU and Tao PENG, “Post-processing of Deep Web Information Extraction Based on Domain Ontology”, Advances in Electrical and Computer Engineering, **13(4)**, 25–32, (2013).

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 five international publications in her accounts. She is a keen researcher in web data mining techniques.

S. A. Sahaaya Arul Mary is a Doctorate in Computer Science and Engineering and having twenty years of experience in teaching with good knowledge in the area of Computer science and Information Technology, currently working as Dean Academic and HOD/CSE, Jayaram College of Engineering and Technology, Trichy, India. She has authored several books in Computer Science and has published 40 research papers in reputed journals, international and national conferences. She acts as recognized research supervisor in Anna University, Chennai and Manonmaniam Sundaranar University. More than 10 research scholars work under her guidance and two of the have submitted the thesis. Her areas of interest include Software Engineering, Networks, Software Testing, Data Warehousing and Data Mining. She is guiding seven research scholars.