HYBRID FRAMEWORK USING ANT COLONY CLUSTERING AND K-MEANS GENETIC ALGORITHM (ANT-KGA) FOR OPTIMIZATIONS OF WEB USAGE PATTERN

Seema C.P.
ASECT University
(MP) India

Dr. Pradeep Chouksey
Department of CSE
TIT College, Bhopal (MP) India

Gajendra K. Vaiker
Department of CS
AISECT University, Bhopal(MP) India

Abstract- *The web usage Mining (WUM) is the process of discovering hidden patterns from the web usage. The proper analysis of web log file is beneficial to manage the websites effectively for administrative and users' prospective. In this research paper, a hybrid framework is proposed Using Ant Colony Clustering and Linear Genetic Programming Approach (ANT-LGP) for optimizations of web usage pattern. The hybrid framework uses an ant colony optimization algorithm to cluster Web usage patterns. In this work, we investigate the performance aspect of proposed hybrid method with well famous K-means genetic algorithm (K-GA) based ant colony clustering method for web usage log pattern analysis. We present a comparative performance evaluation model of ANT-LGP versus ANT-KGA by means of an analytical approach. Quantitative measurements are performed using number clusters, error rate, number of iteration and execution time performance parameters in this research work. The results proposed framework offers a superior performance compared to existing L-GP based system.*

Keywords-*Web Usage Mining, Antcolony clustering ,K-Means, Linear Genetic Programming*

I INTRODUCTION

Web mining is primarily aimed at deriving actionable knowledge from the Web through the application of various data mining techniques [1].Web mining is further divided into three broader classes such as Content Mining; Structure Mining; and Web Usage Mining [2]. Due to the continuous increase in growth and complexity of WWW, web site publishers are facing increasing difficulty in attracting and retaining users. In order to design attractive web sites, designers must understand their users' needs. Therefore analyzing navigational behavior of users is an important part of web page design. Web Usage Mining is the discovery of user access patterns from Web server access logs [2].Web Usage Mining (WUM) is the application of data mining techniques to web usage data in order to discover the patterns that can be used to analyses the user's navigational behavior. A WUM methodology is divided into three steps such as data preprocessing; pattern discovery; and pattern analysis [3,4,5]. These three steps or phases are sequentially connected to each other to form complete WUM methodology. Clustering is a natural way to group the similar objects based on some common properties (similarity measure).The elements of a cluster are more similar to each and have similar properties as well. Clustering is a prominent data mining technique and is used for various applications such as pattern discovery; data analysis; prediction; visualization; and personalization. Web session clustering is an emerging and common technique at preprocessing level of WUM [4,6], which not only extract the hidden behavior from web usage but also groups the sessions based on some common properties (Similarity). Web log analysis can be single level and multilevel. Multilevel log mining searches interesting relationship among the values from different levels in a log database. There are several possible challenges to explore efficient log mining of multiple-level including multiple scans of transaction database [21]. So for simplicity, this work revolves around single level log pattern analysis.

Nature inspired algorithms are problem solving techniques that attempt to simulate the occurrence of natural processes. Some of the natural processes that such algorithms are based on include the evolution of species [7,8]. Ant Colony Optimization (ACO) algorithm [9] belongs to the natural class of problem solving techniques which is initially inspired by the efficiency of real ants as they find their fastest path back to their nest when sourcing for food.

Early approaches in applying ACO to clustering [10,11,12] are to first partition the search area into grids or clusters but this may result in too many clusters as there might be missed or wrong calculated. Therefore, some other algorithms such as K-means, genetic, fuzzy are normally combined with ACO to minimize categorization errors [13].More recently, variants of ant-based clustering have been proposed, such as using k-means genetic algorithm (K-GA), fuzzy-ACO, fuzzy k-means

with genetic algorithm etc. This paper proposes an improved ant colony cluster algorithm based on Linear Genetic Programming Approach (ANT-LGP) for optimizations of web usage pattern. It enables the ants to consult historical information when conveying objects by importing adjusting process and short period memory, and it also does iterative regulating to the cluster formed by the ants. Thus, it advances the convergence speed of the algorithm and the efficiency of the cluster.The rest of paper is organized as follows. In section 2, we present a review on existing web session clustering techniques. Section 3, introduction about ant colony. Section 4, explains the proposed methodology of WUM. Section 5 presents the experimental results of proposed methodology. Section 5 concludes the paper.

II LITERATURE SURVEY

A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining" A framework for web session clustering is given by applying preprocessing level of web usage mining. The framework here cover the data preprocessing steps to prepare the web log data and convert the categorical web log data into numerical data. A session vector is obtained, so that appropriate optimization could be applied to cluster the web log data. The hierarchical cluster based approach here enhances the existing web session techniques for more structured information about the user sessions. The three different measures "Angular Separation", "Canberra Distance" and "Spearman Distance" instead of "Euclidean Distance" are used. The PSO algorithm based on "Angular Separation" and "Canberra Distance" and then agglomerative to obtain hierarchical sessionization of sessions is applied. The results of AS and CD are providing more structured information as compare to Alam [5] ED and SD.

By Zahid Ansari, A. Vinaya Babu, Waseem Ahmed and Mohammad Fazle Azeem[15]**" A Fuzzy Set Theoretic Approach to Discover User Sessions from Web Navigational Data"**describe web navigational data using fuzzy logic through web usage mining. The session files are filtered to remove very small sessions in order to eliminate the noise from the data. But direct removal of these small sized sessions may result in loss of a significant amount of information especially when the number of small sessions is large. A "Fuzzy Set Theoretic" approach is applied to deal with this problem. Instead of directly removing all the small sessions below a specified threshold, weights are assigned to all the sessions using a "Fuzzy Membership Function" based on the number of URLs accessed by the sessions. After assigning the weights a "Fuzzy c-Mean Clustering" algorithm is applied to discover the clusters of user profiles.

By Weihui Dai, Shouji Liu and Shuyi Liang[16], **"An Improved Ant Colony Optimization Cluster Algorithm Based on Swarm Intelligence"** Proposed an improved ant colony optimization cluster algorithm based on a classics algorithm- LF algorithm. By the introduction of a new formula and the probability of similarity metric conversion function, as well as the new formula of distance, this algorithm can deal with the category data easily. It also introduces a new adjustment process, which adjusts the cluster generated by the carry process iteratively. Experiments show that the improved ant colony algorithm can form more accurate and stability clusters than the K-Modes algorithm, Information Entropy-Based Cluster Algorithm. They also describe the process and idea of the algorithm usage by a mobile customer classification case and analyze the cluster results. This algorithm can handle large category dataset more rapidly, accurately and effectively, and keep the good scalability at the same time.

By A. Azadeh, A. Keramati and H. Panahi[17], **"A hybrid GA-ant colony approach for exploring the relationship between IT and firm performance"** A hybrid Genetic Algorithm (GA) Ant Colony Optimization (ACO) approach is proposed for data clustering. This is because of the need for the application of meta heuristic algorithms parallel to deterministic approaches. This study discusses and analyses data from 90 companies in a unique supply chain. The data includes 26 indices about IT and 11 indices about performance. The companies are classified with respect to the IT and performance indices (indicators). Then, IT clusters and performance clusters are mapped to one another and, consequently, the relationship between them is explored. This is the first study which integrates ant colony approach and GA for exploring the relationship between IT and firm performance[18] .They improve the slow speed of the AntClass algorithm by proposed new algorithm named DBAntCluster. Firstly, the high density clusters are got in the dataset by using DBSCAN algorithm, and then these high density clusters are scattered in the grid board as a special kind of data object with other single data objects in the dataset. In DBAntCluster algorithm, the ants can avoid many unnecessary movements by using the data attribute of density and distribution well, and the speed is greatly accelerated.

III ANT CLUSTERING

The ant colony optimization algorithm (ACO) is a probabilistic technique for solving computational problems. In

the natural world, ants (initially) wander randomly, and upon finding food return to their colony while laying down pheromone trails. If other ants find such a path, they are likely not to keep travelling at random, but to instead follow the trail,

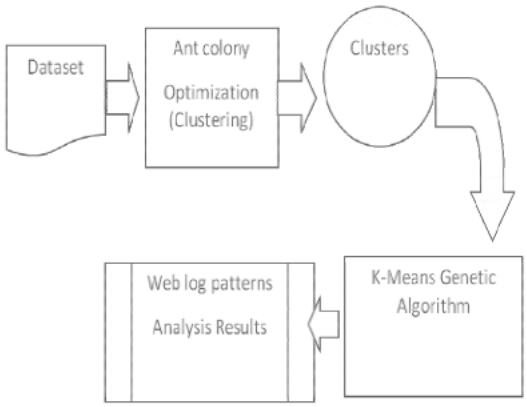

Figure 1: Proposed Hybrid Approach

returning and reinforcing it if they eventually find food. Over time, however, the pheromone trail starts to evaporate, thus reducing its attractive strength. The more time it takes for an ant to travel down the path and back again, the more time the pheromones have to evaporate. A short path, by comparison, gets marched over more frequently, and thus the pheromone density becomes higher on shorter paths than longer ones. Pheromone evaporation also has the advantage of avoiding the convergence to a locally optimal solution. If there were no evaporation at all, the paths chosen by the first ants would tend to be excessively attractive to the following ones. In that case, the exploration of the solution space would be constrained. Figure 1 shown the basic steps of ant colony algorithm.

```
Procedure ACO
  While (not_termination)
    GenerateSolutions()
    AttractivenessCalculation()
    TrailLevelUpdate()
  End while
End procedure
```

Figure 1: Ant Colony Algorithm

At each iteration of the algorithm, each ant moves from a state (x) to state (y), corresponding to a more complete intermediate solution. Thus, each ant (k) computes a set $A_k(x)$ of feasible expansions to its current state in each iteration. For ant (k), the probability $p_{xy}^{\ k}$ of moving from state (x) to state (y) depends on the combination of two values (attractiveness η_{xy} of the move, and the trail level T_{xy} of the move). The trail level represents a posteriori indication of the desirability of that move In general, the k th ant moves from state (x) to state (y) with probability:

$$p_{xy}^{\ k} = \frac{(T_{xy}^{\ \alpha})(\eta_{xy}^{\ \beta})}{\sum (T_{xy}^{\ \alpha})(\eta_{xy}^{\ \beta})}$$

Where α Is a parameter to control the influence of T_{xy} ($\alpha \geq 0$)

β Is a parameter to control the influence of η_{xy} ($\beta \geq 1$)

IV CLUSTERING ALGORITHMS WORK

Ant-based clustering algorithms are based upon the brood sorting behavior of ants. Its dissimilarity-based evaluation of the local density make it suitable for data clustering and it has subsequently been used web log analysis. Genetic programming (GP) is burning issue now days. It is able to produce accurate results without having much of analytical knowledge related to problem domain. GP does impose restrictions on how the structure of solutions should be formulated. So it is best suitable for enhancing formation of cluster and analysis of patterns in web log analysis. There are several variants of GP, some of them are: Linear Genetic Programming (LGP), Gene Expression Programming (GEP), Multi Expression Programming (MEP), Cartesian Genetic Programming (CGP), Traceless Genetic Programming (TGP). Proposed work try to introduced new variation with K-Means. It is represented by (KGA)

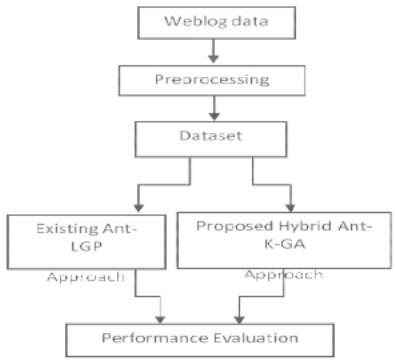

Figure 2: Proposed framework for web log analysis and performances evaluation

Throughout in this work. The structure of proposed hybrid approach is shown in figure 2. The hybrid framework uses an ant colony optimization algorithm to cluster Web usage patterns. In this research work KGA is associated with ant colony clustering algorithm for web log pattern analysis. This hybrid approach is represented by (Ant-K-GA). One of the benefits of this type of hybridization is the implicit re-use of

posteriori knowledge in the cluster formation. The Basic idea is that isolated items should be picked up and dropped efficiently at cluster where more items of that type are present. Proposed (Ant-K-GA) algorithm follows real antlike behaviors as much as possible.

In that sense, genetic behavior is incorporated into the web log cluster analysis system, avoiding randomly moving agents without interest. The proposed framework for web log analysis and performances evaluation is shown in figure 3.

Since trail movement is controlled by attractiveness (The probability to put item in correct cluster) so this is called transition probabilities and it depends on the spatial distribution of pheromone across the environment (behavior of users). The variation of genetic programming (KGA) is incorporated in this research work to effectively control the spatial distribution. There are two major factors that should influence any local action taken by the ant: the number of objects in his neighborhood, and their similarity. KGA strategy not only allows guiding ants to find clusters of objects in an adaptive way also develop short-term memories to overcome local minima. Proposed framework for web log analysis and performances evaluation is shown in Figure 3 and pseudo code of proposed (Ant-K-GA) approach is shown in figure 4.

Initialization

a. Set initial parameters: variable, states, function, input

b. Set initial pheromone trails value

c. Each ant is individually placed on initial state with empty memory.

While (not_termination)

a. Construct Ant Solution:

b. Calculate attractiveness of next move

c. Apply Local Search through KGA (Avoid local minima)

d .If there is an improvement-. Update Trails

e. calculate evaporation through genetic operators

f. select the population with a probability based on fitness.

End While

pseudo code of proposed (Ant-K-GA) approach

V EXPERIMENTATION AND RESULT ANALYSIS

To show this procedure we take live web log records of Maulana Azad National Institute of Technology, Bhopal, India, for processing our possible approach, first we parse these log records by the tool WebLogExpert [19], in Figure 5 show daily search phrases for the Maulana Azad National Institute of Technology, Bhopal, and server. For measuring performance, accuracy of the proposed method, we perform our operation on the 500 records of the MANIT, Bhopal, India log records. To support our methodology, we designed and implemented in MATLAB 7.8. This experiment is handled with null values efficiently in data preprocessing steps.

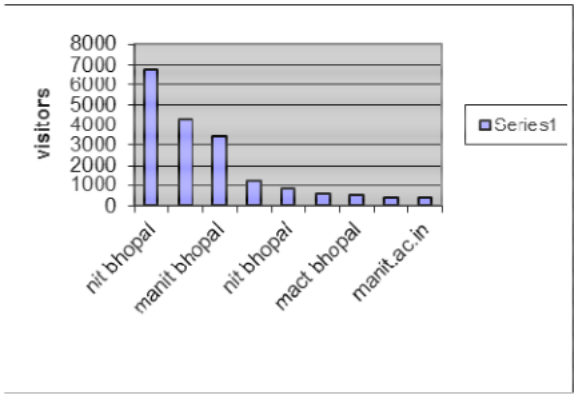

Figure 3: Top Search Phrases In MANIT Log Records.

We present a comparative performance evaluation model of ANT-LGP versus ANT-KGA by means of an analytical approach. Quantitative measurements are performed using number clusters, error rate, number of iteration and execution time performance parameters in this research work.

(a) Number of cluster

Method for choosing the number of clusters based on distortion, a quantity those measures the average distance, per dimension, between each observation and its closest cluster center. Our technique is computationally efficient and straightforward to implement. This approach is guaranteed to produce the correct answer provided the clusters do not overlap too severely. The more clusters the better quality (the smaller variance within cluster). The correct choice of k is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set. In addition, increasing k without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster (i.e., when k equals the number of data points, n). Intuitively then, the optimal choice of k will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. There are several categories of methods for making this decision. One simple rule of thumb sets the number:

$$k \approx \sqrt{n/2}$$

Where

n : The number of objects (data points).

Proposed ant-KGA method gives more concise results. The less number of cluster generated by ant-KGA as shown in table 1 and table 2.

(b) Error rate

The true error rate (Err) is a measure of how accurately the classification, built with the learning sample, would be if they were applied to the whole universe. In this paper, we test the false positive rates in which a clustering is presented with a large number of samples that do not belong to any of cluster. Error rate is major factor for calculating performance. Ant-KGA shows better performance compared to Ant-LGP. Results are shown in table 1 and 2.

(c) Iteration

Clustering is a extensive iterative process. Algorithm must be design in such a way that it can fit in main memory. Normally hybrid approaches are resources consuming but proposed hybrid approach (Ant-KGA) give better performance in term of memory requirement. Ant-KGA gives the same performance in term of iteration as shown in table 1 and table 2. In other words, Ant-KGA does not show any performance degradation.

(d) Execution Time

Usually the efficiency or running time of an algorithm is stated as a function relating the input length to the number of steps. Run-time analysis is a theoretical classification that estimates and anticipates the increase in running time (or run-time) of an algorithm as its input size (usually denoted as n) increases. In term of execution time taken by algorithm, Ant-KGA gives better performance.

Implemented (Ant-KGA) algorithm executed on different threshold values to identify its scalability. We have executed it with two threshold values (0.23 and 0.50). Threshold value must be between 0 to 1. Results are shown in Table 1 and 2.

Parameters	Ant-KGA	Ant-LGP
Threshold :0.23		
Error rate	2.8	4.2
Number of iteration	7	7
Execution time	5.4	6.4
Number of clusters	5	6

Table 1: Performance comparison with threshold value 0.23

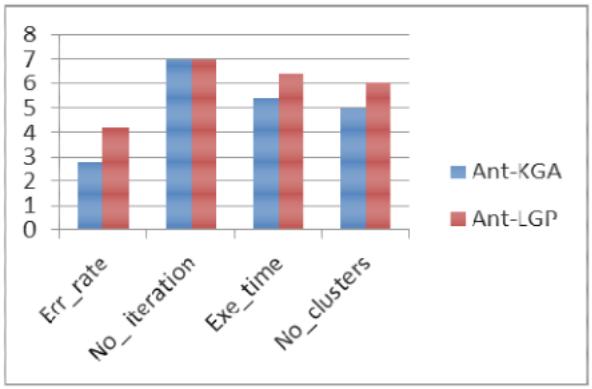

Figure 4: Performance comparison with threshold value 0.23

Parameters	Ant-KGA	Ant-LGP
Threshold :0.50		
Error rate	3.59	4.8
Number of iteration	6	6
Execution time	2.34	3.57
Number of clusters	3	4

Table 2: Performance comparison with threshold value 0.50

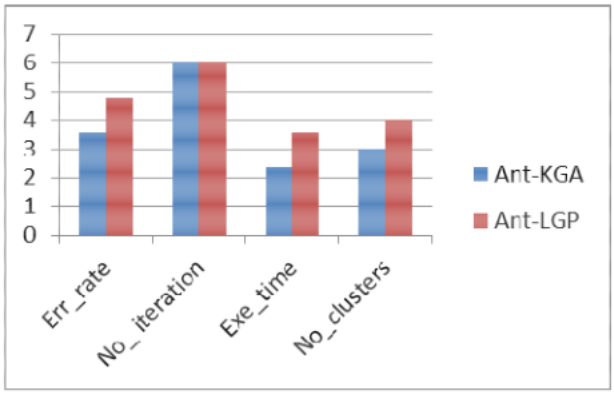

Figure 5: Performance comparison with threshold value 0.50

VI CONCLUSION

The proposed ANT-LGP model seems to work very well for the web log pattern analysis. In this paper, our focus was to develop hybrid approach to analyze the hourly and daily web traffic volume. The results also reveal the importance of using optimization techniques for mining useful information. The genetic nature of the algorithm makes it fairly robust to the effects of outliers within the data. A quantitative comparison has been given with various parameters to show performance improvement. The work show that proposed (Ant-KGA) approach is well applicable in real-world applications including web log pattern analysis. The objective of research was to enhance the web log visualization and structured

information for the next phases of WUM process. Proposed work gives following advantages:

1. More powerful pattern analysis than using conventional GP.
2. Efficient evaluation of cluster boundaries from the intrinsic feature exhibited by (Ant-K-GA).
3. Less complicated formulation of trail and attractiveness via the crossover and mutation genetic operators.

REFERENCES

[1] P. Kolari and A. Joshi, "Web mining: research and practice," Computing in Science and Engineering, vol. 6, no. 4, pp. 49–53, 2004.

[2] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: Information and pattern discovery on the world wide web," in Ninth IEEE International Conference on Tools with Artificial Intelligence, 1997. Proceedings. 1997, pp. 558–567.

[3] Castellano, G., A. M. Fanelli, et al. (2007 a). LODAP: A LogData Preprocessor for mining Web browsing patterns. Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Corfu Island, Greece, February 16-19, 2007.

[4] Nichele, C. M. and K. Becker (2006). "Clustering Web Sessions by Levels of Page Similarity." W.K. Ng, M. Kitsuregawa, and J. Li (Eds.): PAKDD 2006, LNAI 3918, pp. 346 – 350, 2006. © Springer-Verlag Berlin Heidelberg 2006.

[5] Srivastava, J., R. Cooley, et al. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations. Copyright 2000 ACM SIGKDD, Jan 2000.

[6] Suneetha, K. R. and D. R. Krishnamoorthi (2009). "Identifying User Behavior by Analyzing Web Server Access Log File." IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[7] E. Yu and K.S. Sung. A genetic algorithm for a university weekly courses timetabling problem. International Transactions in Operational Research, 9(6):703–717, 2002.

[8] E.K. Burke, D.G. Elliman, and R.F. Weare. A genetic algorithm based university timetabling system. In Proceedings of the 2nd East-West International Conference on Computer Technologies in Education, pages 35–40, Crimea, Ukraine, September 1994.

[9] M. Dorigo, V. Maniezzo, and A. Colorni. The ant system: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics – Part B, 26(1):29–42, 1996.

[10] J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chretien. The dynamics of collective sorting robot-like ants and ant-like robots. In Proceedings of the first international conference on simulation of adaptive behavior on from animals to animats, pages 356–363, Cambridge, MA, USA, 1990. MIT Press.

[11] Lumer E. D. and Faieta B. Diversity and adaptation in populations of clustering ants. In Cli D., Husbands P., Meyer J., and Wilson S., editors, Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3, pages 501–508, Cambridge, MA, 1994.MIT Press.

[12] Kuntz P., Layzell P., and Snyers D. A colony of ant-like agents for partitioning in VLSI technology. In P. Husbands and I. Harvey, editors, Proceedings of the Fourth European Conference on Artificial Life, pages 417–424. MIT Press, 1997.

[13] Y. Peng, X. Hou, and S. Liu. The k-means clustering algorithm based on density and ant colony. In IEEE International Conference in Neural Networks and Signal Processing, Nanjing, China, December 2003.

[14] Tasawar Hussain, Sohail Asghar and Nayyer Masood" Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence" in 2010 6th International Conference on Emerging Technologies (ICET).

[15] Zahid Ansari, A. Vinaya Babu, Waseem Ahmed and Mohammad Fazle Azeem" A Fuzzy Set Theoretic Approach to Discover User Sessions from Web Navigational Data" in ieee 2009.

[16] Weihui Dai, Shouji Liu and Shuyi Liang, "An Improved Ant Colony Optimization Cluster Algorithm Based on Swarm Intelligence" Journal of Software vol. 4, No. 4, June 2009.

[17] A. Azadeh, A. Keramati and H. Panahi, "A hybrid GA-ant colony approach for exploring the relationship between IT and firm performance" International Journal of Business Information Systems archive,Pages 542-563 ,Volume 4 Issue 5, May 2009.

[18] Shang Liu,Zhi-Tong Dou and Fei Li, "A new ant colony clustering algorithm based on DBSCAN" International Conference on Machine Learning and Cybernetics, 2004.

[19] WebLogExpert,http://www.weblogexpert.com.

[20] Vivek Tiwari & Vipin Tiwari (2010). Association Rule Mining-A Graph based approach for mining Frequent Itemsets. IEEE International Conference on Networking and Information Technology (ICNIT 2010) Manila, Philippines,IEEE, ISBN:978-4244-7577-3.

[21] Vivek Tiwari & Dr. R.S.Thakur (2011). A level wise Tree Based Approach for Ontology-Driven Association Rules Mining. CiiT International Journal of Data Mining and Knowledge Engineering.