# A Hierarchical Cluster with similarity Measure based Pre-processing approach for Web Usage Mining

S.Kartiga[#1], K.Indumathi[#2]

[#1#,2]*M .Tech (Networking) , Pondicherry University*

[#1,#2]*Sri Manakula Vinayagar Engineering College, Madagadipet , Pondicherry-605107*

[#1]*sparkarthi@gmail.com*,[#2]*indumathy2610@gmail.com*

*Abstract*— **Web usage mining has been used effectively as an approach to automatic personalization and as a way to overcome deficiencies of traditional approaches such as collaborative filtering. Clustering is one of the important functions in web usage mining. The likelihood of bad or incomplete web usage data is higher than the conventional applications. Clustering is a process of forming groups (clusters) of similar objects from a given set of inputs. Good clusters have the characteristic that objects belonging to the same cluster are "similar" to each other, while objects from two different clusters are "dissimilar". The idea of clustering originates from statistics where it was applied to numerical data. However, computer science and data mining in particular, extended the notion to other types of data such as text or multimedia. Clearly, web snippets belong to the class of textual data and hence it is the text clustering algorithms that should be regarded as a base for the web search results clustering systems. Recent attempts have adapted the K-means clustering algorithm based on rough sets to find interval sets of clusters. So far, the researchers are not contributed to improve the cluster quality after grouping. In this paper, In hierarchical clustering methods we present enhanced K-Mean and ID3, a method to cascade enhanced k-Means clustering and the ID3 decision tree learning methods for web usage mining. Our proposed method uses the enhanced K-Means Clustering and ID3 algorithm to improve the efficient similarity measures and accuracy results based upon preprocessing task approach for web usage mining. The algorithm is tested with web access logs and shows that pre processing refinement of clusters definitely lead to improved solutions.**

*Keywords*— **Web Usage Mining, Pre-processing, Clustering and classification, k-mean, ID3.**

## I. INTRODUCTION

World Wide Web (WWW) is expanding tremendously everyday in the number of websites and also the population of users. The basic purpose of website is to deliver useful information to its users efficiently and timely; at the same time websites are competing to acquire their own shares of visitors. Websites are striving to improve themselves by offering personalized contents and services that supposedly will match best of the users' tastes or needs. In addition to explicit methods like questionnaires and options of setting users' profiles, one subtle approach is to probe into web log files for revealing patterns of users' interests on the websites. It is well know that users' online interactions with the website are recorded in server web log files that serve as a valuable pool of information. By applying the data mining techniques on web log file, we obtain good insights about the users' behaviors; thereby we can customize the contents and services on the website to better suit the users. We can analyze the web log files for various aspects of website enhancements. Furthermore, proper analysis of web log unleashes useful information for webmaster or administrator for numerous advantages such as web personalization, website schema modification, user surfing behaviors, website structure modification and we can tackle the issue of web server performance as well.

According to Hussain [12] web mining is an important research discipline of data mining and drawing huge interest from academia and software industry. It is an import part of the online Knowledge Discovery Process where data mining techniques are harvesting knowledge over the data collected from World Wide Web. In general, Web mining is further divided into three broad areas, such as Content Mining; Structure Mining; and Web Usage Mining (WUM).

Specifically, Khasawneh [14] defined WUM as applying data mining techniques to discover interactions between users and a website from web logs.A WUM methodology is divided into three essential steps: data preprocessing; pattern discovery; and pattern analysis [3,20,16]. These three steps or phases are sequentially connected to each other to form a complete WUM methodology.Clustering is a natural way to group similar objects based on some common properties, which is called similarity measure. The elements within a cluster are relatively more similar to each because they have similar properties or attributes. Clustering is one of the most prominent data mining techniques that has been 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 [1, 14], which not only extracts the hidden behaviour from web usage but also groups the users' visit sessions based on some common properties (Similarity). Catellano [6] defined the session as a set of user visits to a website in a particular visit and session identification from web log is indeed a complex job [19].

According to Chaofeng [7], web session clustering embraces a sequence of the following three steps: data *preprocessing; measurement of similarity among sessions;* and *clustering algorithm.*To cater the issues of web session clustering, we proposed a complete preprocessing methodology, which covers all the above stated phases. In preprocessing phase, we selected a sample web server log file and applied proposed data cleaning and data filtering algorithms. We also proposed algorithms to identify the users on the basis of IP Address (an attribute of web log). By applying proposed session identification algorithm, we obtained the user sessions from web log and transform them into session vectors. We applied the *"Angular Separation", "Canberra Distance"* and *"Spearman Distance"* similarity measures to compute the similarity among the session vectors. In the last phase, we applied the proposed algorithm based on Swarm and Agglomerative algorithms to obtain the hierarchical sessionization of user sessions.

The rest of paper is organized as follows. In section 2, we present a review on existing K-mean algorithms in clustering. In Section 3, explains the proposed method of preprocessing steps required similarity measures for WUM. In Section 4, we represents the contribution of enhanced k-mean and ID3 classification. Section 5, we concludes clustering and classification in our proposed methods.

## II. LITERATURE REVIEW

In this section, we review the existing literature on preprocessing methodology and particularly, the web session clustering for WUM. Web log file is primary the source for WUM and contains a large amount of "irrelevant information". Due to this inherent deficiency, original web log file cannot be directly used in WUM process. Hence, preprocessing of web log file becomes imperative. Preprocessing step helps to improve the quality of data [12], and consequently, improves the efficiency and effectiveness of subsequent steps of WUM process.

By reviewing the preprocessing techniques of web log file we learn the strengths and shortcoming of the existing techniques that enables us to innovate a suitable new technique. The objective of preprocessing step is to prepare the most relevant information for the next steps of WUM and, on the other hand, to improve the quality, and structure the information through grouping it, based on common properties for pattern mining step of WUM.

Khasawneh [14] applied data cleaning method on log files to remove irrelevant entries. The authors identified the users based on IP Address, date, and time of visit and set of records visited by user in that period of time. For session identification, an ontology based algorithm was designed especially for structures of website.

Castellano [6] developed a tool LODAP (Log Data Preprocessor) for the preprocessing of web log file. Tool performs data cleaning, data structuration (session identification) and data filtering along with summary at each step of processor. Weighing the pros and cons, we suppose that it is a good effort by the authors to support their work by using proper MS Access based tool. In summing up it can be said that if authors were able to perform some sort of classification through LODAP, it would be an effective tool for preprocessing of WUM.

For web session clustering, Alam [2] used the *"Euclidean Distance"* to calculate the similarity and applied the Particle Swarm Optimization (PSO) for web session clustering. For PSO, sessions were taken as particles and ParticleID; Distance FromEachSession;WonSessionVector;SessionAttributeValues; and P-Best attributes were used.

Alam [2] performed the session clustering by applying the Euclidean Distance (ED) measure. The authors conducted the experiment and compare the results with Kmean. While PSO and K-Mean are different in nature and produce different results, the authors did not compare the results with any other PSO based session clustering. In our proposed methodology of preprocessing, we applied the two different similarity measures and generated the hierarchical clusters.

Lu and Nguyen [16] proposed the PSO based on sequence clustering technique which is also adopted in our model. Similarity measure for sequences clustering was defined as ratio of common items and unique items in two sequences, and then set the similarity in the order of occurrence of items in two sequences . In this paper, only similarity measure was changed from Euclidean Distance to S3M and the authors calculated the similarity based on the longest common sequence (LCS). For large amount of data, the time and space complexity can be big issue. The authors compared the results with KMean only.

Mayil and Duraiswamy [18] performed the session clustering by applying agglomerative hierarchical clustering algorithm. *Alignment Score (Sa)* and *Local Similarity (Sb)* are two major components to calculate the similarity between sessions. Mayil and Duraiswamy [18] applied dynamic programming on sessions and hierarchical clustering technique to pick the results. No comparative study and measure calculation justification was given. It is important to mention that no other preprocessing techniques were adapted for the complete preprocessing phase of WUM.

Banerjee and Ghosh [3] calculated the similarity by using longest common subsequence (LCS) applied the clustering algorithm to cluster the sessions. It is concluded from the literature review that for session clustering, we so far lack of a complete preprocessing methodology. Hence our proposed methodology will not only improve the quality and efficiency of data for later steps but also to enhance the log file visibility and to structure the information in hierarchical clustering. It is concluded from the literature review that for session clustering, we so far lack of a complete preprocessing methodology. Hence our proposed methodology will not only improve the quality and efficiency of data for later steps but also to enhance the log file visibility and to structure the information in hierarchical clustering. The above survey covering the concepts to related of web usage mining access pattern behavior and complete pre-processing stages applied on particular application.

## III. PROPOSED PREPROCESSING APPROACH

It may be known that applying, one or several preprocessing techniques individually, cannot guarantee the reliability of overall results of WUM process. Preprocessing phase is a set of inter-connected, coherent, and integrated techniques, applied in a sequence to produce clear and well-defined results. From literature review, we observed the need of a complete preprocessing methodology at preprocessing level of WUM [12].

To cater this issue, we have proposed a complete methodology at preprocessing level of WUM. The objectives of proposed methodology is to drag the most relevant and structured information about the usage sessions for the next steps of WUM; and at a same time, improve the quality and structure of information by applying clustering algorithm on web log data based on well-established similarity metrics.

In this section, we elaborate the main components of proposed preprocessing methodology of WUM process, which are Data Cleaning; Data Filtering; User Identification; Session Identification; Session Clustering and Similarity measures.

### *A) Data Cleaning*

Web log contains large amount of irrelevant entries, which are required to be removed from the web log for preparation prior to data mining. Data Cleaning Algorithm removes such irrelevant entries from the web log file.

In addition, different web logs have different formats, so cleaning means shall be chosen according to the actual demand. Data cleaning concerned with removing all the data tracked in web logs that are useless for mining purposes e.g. requests for graphical page content (e.g. JPG,.GIF, and css); Request for any other file which might be included in to web page ; or even navigation sessions performed by robots and web spiders.

Robots and web spider navigation patterns must be explicitly identified. This is usually done for instance by referring to the remote host name, by referring to the user agent or by checking the access to the robots.txt file. However some robots actually send a false user agent in HTTP request. The navigational behavior can be used to separates robot sessions from actual user's sessions. An algorithm for cleaning the entries of server logs is presented below:

```
Read record in database.
For each record in database
Read fields (URI – stem) //URI- stem indicates
The target URL//
If fields = {*.gif,*.jpg,*.css} then
Remove records
Else
Save records
End if
Next record
```

The task of data cleaning is to remove the irrelevant and redundant log entries for the mining process. There are three kinds of irrelevant or redundant data to be removed.
They are Additional Requests:
a) A user's request to view a particular page often results in several log entries. Graphics and scripts are downloaded in addition to the HTML file, because of the connectionless nature of the HTTP protocol. Since the main intention of Web Usage Mining is to get a picture of the user's behavior, it does not make sense to include file requests that the user did not explicitly request. Suffix part of an URL is checked and eliminates suffixes like gif, jpg, GIF, JPEG, css, map etc.
b) Robots requests: Web robots are software tools that scan a Web site to extract its content. Spiders automatically follow all the hyperlinks from a Web page. To remove robots" request, we can look for all hosts that have requested the page "robots.txt", which is checked by robot while browsing.
c) Entries with error: Status code shows the success or failure of a request. Entries with status code less than 200 and greater than 299 are failure entries which are to be removed. Only necessary fields like date, time, IPaddress, User Agent, URL requested, URL referred, time taken are considered for further experiments to reduce the processing time So attribute subset selection is done.

### *B) Log File Filtering*

There are different sources of web log file and each source contains different number of log attributes. All the attributes of web log did not take part in web session clustering and there may be attributes which contain no data at all. To further purify the web log for more structured information, we proposed a log file filtering algorithm.

### *C) User Identification*

User identification plays a significant role to identify the distinct and unique users of website. Although users alone play no role in web session clustering, they provide significant information about who the distinctive website users are. We proposed a user identification algorithm based on IP Address (an attribute of web log).
The strategy of User Identification based on the log entries without considering the topology structure of site. The description of concrete strategy algorithm is as follow:

**Input:** N records of web log file
**Output:** User sets identified
*Step 1:* Compare IP address of first log entry with IP address of second log entry.
*Step 2:* If both are same compare the user agent of both entries Else assume as different users.
*Step 3:* If both user agents are same identify both entries are from same user.
Until last entry

User's IP addresses of two consecutive entries are compared. If the IP address is the same, user's browser and operating system is verified and if both are same, both the records are considered from the same user. These experiments prove that the algorithm significantly improves the efficiency and the accuracy of user identification without usage of site topology.

In web usages mining does not require knowledge about a user history because the users visit or request given more than one time to the server. If we visit more than one time, then it generate multiple sessions for each user. It is also known as User activity Records. User Identify by using IP address and User Agent in log files. Client request to server then it generate log files at that time client also send user agent to server.

### *C) Session Identification*

A group of the activities of a single user within a certain period of time from the web log files is called a session. As long as a user is connected to the website, he is within the same 'session' regardless of his surfing patterns. In most of the research, 30 minutes timeout is taken as a default session timeout being assumed [20].For proposed session identification algorithm; we selected the uses of the following log file attributes: IP Address; Data; Time; URL accessed; Data downloaded; and User Agent. We calculate the Start Time (a beginning time when new user first interacted with website) and End Time (The time when user left the website) from user accesses and convert them into 30-minutes sessions. We also obtain the multiple sessions (episodes) for users who stay more than 30-minutes with the website. After obtaining users sessions, we convert the user sessions into session vectors. The proposed session identification algorithm in pseudo codes as follows:

**Input:** User sets with N records, BTmin, BTmax, 2D matrix
**Output:** Constructed Sessions
**Algorithm:**

```
Repeat steps
    Step 1: Calculate the browsing time of a web page by a user by finding the difference between two consecutive entries and subtract the time taken value
    Step 2: Compare the browsing time with minimum and maximum time of each web page
    Step 3: If the browsing time is less than minimum time fix the weight as "0" else if it is between minimum and maximum, then weight is fixed as "1" , if the weight exceeds maximum fix as 10 and if referrer URL is null weight is fixed as 100.
    Step 4: If the same page is visited by the user again in users set increment the corresponding entry.
    Step 5: Weights are stored in the matrix in the corresponding cells. The value aij represents a weight based on users browsing time in page_j. Until last row in users set.
```

Session captures in two ways: 1) Time oriented
2) Structure oriented

**Time Oriented:** Time oriented is depends on the Time stamps or date and time of request in the server log file. In the time oriented session there are two types: i) The difference between First request and last request is < =30 minutes. ii) The difference between First request and next request is <= 10. Using these two points we judge time oriented sessions.

**Structure Oriented:** Structure oriented capture in the referrer fields of the server logs. Structure oriented depends on Referrer fields is currently open or that user currently login referrer. Means it's belonging to more than one "open" constructed session.

### *D) Path supplement*

Due to buffer memory of client-side, user may use back function of the browser when browsing, therefore, ratiocination shall be conducted according to the backwards and forwards pages visited by the user to make supplement to user's access path. In this way, website structure can be adjusted and optimized, so that users can visit pages more simply and faster. It also can be used in intelligent recommendation and customized electronic commerce activities according to user's typical browse mode.

### *E) Transaction Identification*

User session is the only element with natural transaction characteristic of web log mining, but for some mining algorithms, the data granularity of the user session may be too large and needs to be converted into smaller transaction by using segmentation algorithm to identify.

### *F) Similarity measure*

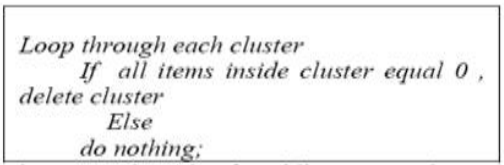
All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this paper, we introduce a hierarchical cluster with similarity measure based pre-processing approach for web usage mining. Using clustering algorithm enhanced k-mean and ID3, more informative assessment of similarity could be achieved in web usage mining. The below section we will be describe entirely the steps under clustering algorithm for enhanced k-mean and ID3.

## IV CONTRIBUTION OF ENHANCED K-MEAN AND ID3 ALGORITHM CLASSIFICATION

### *A) K-Mean Algorithm*

After investigating the most common algorithms for data clustering, there is enhancement for the most familiar algorithm which called k-means trying to solving some problems of k-mean algorithm. More features will be added to enhanced algorithm such as; change centriod point from random points to center points for data and adding step to avoid empty clusters when visualize data it shows also in steps of algorithm. Figure 5.31 shows the flowchart of enhanced K-men algorithm.

Step 1: Assume number of cluster “K”.
Step 2: Calculate „K‟ at center points of data set
Step 3: Calculate the distance between a data sample and clusters.
Step 4: Assign a data sample to closest cluster center.
Step 5: Calculate new cluster center.
Step 6: Repeat step 3.4 and 5 until no objects move group.
Step 7: Avoid empty clusters



```
Loop through each cluster
    If  all items inside cluster equal 0 ,
delete cluster
        Else
    do nothing;
```

Fig 7: Pseudo code of avoiding empty clusters

8. Perform visualization

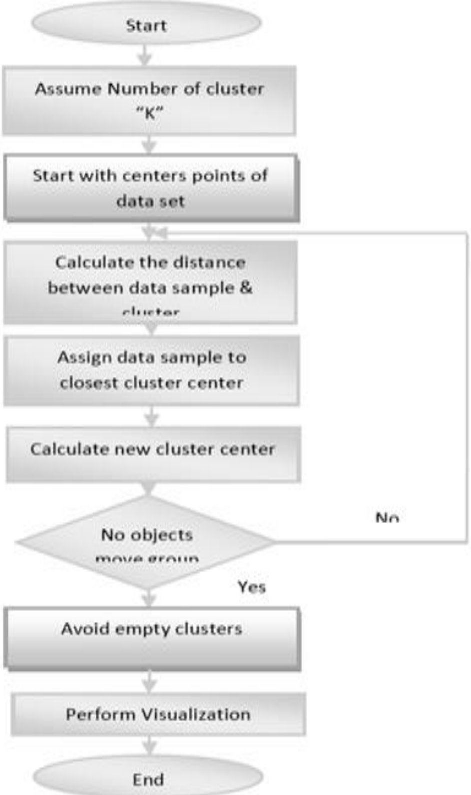
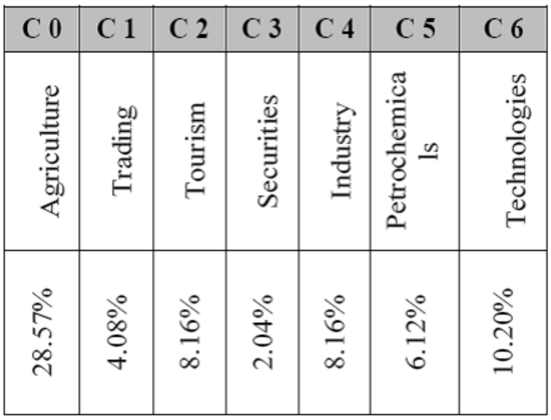

Start
Assume Number of cluster "K"
Start with centers points of data set
Calculate the distance between data sample & cluster
Assign data sample to closest cluster center
Calculate new cluster center
No objects move group
No
Yes
Avoid empty clusters
Perform Visualization
End

Fig 8: E_K-mean algorithm flowcharts

After applying enhanced k-mean algorithm to avoid empty clusters and re clustering data more accurate using change center point for K, it gives the research more accurate results appear in next section. Table 7 presents the set of data used in the implementation experiments for training data to set a best result and choosing an effective number of clusters and percentage of data set to apply the technique for the system. These data are divided among 7 clusters representing the different percentage of the set of data used using enhanced K-Mean algorithm E_K-M.

Table 7: Testing data on a 7 clusters model using E_K-M Algorithm



| C 0 | C 1 | C 2 | C 3 | C 4 | C 5 | C 6 |
|---|---|---|---|---|---|---|
| Agriculture | Trading | Tourism | Securities | Industry | Petrochemica ls | Technologies |
| 28.57% | 4.08% | 8.16% | 2.04% | 8.16% | 6.12% | 10.20% |

*B) ID3 Algorithm*

The ID3 algorithm (Inducing Decision Trees) was originally introduced by Quinlan in [11] and is described below in Algorithm 1. Here we briefly recall the steps involved in the algorithm below:

□ Define $P_1$, $P_2$, …., $P_n$ Parties.(Horizontally partitioned).

□ Each Party contains R set of attributes $A_1$, $A_2$, ….,$A_R$.

□ C the class attributes contains c class values $C_1$, $C_2$,…., $C_c$.

□ For party $P_i$ where i = 1 to n do

□ If R is Empty Then

□ Return a leaf node with class value

□ Else If all transaction in T($P_i$) have the same class Then

□ Return a leaf node with the class value

□ Else

□ Calculate Expected Information classify the given sample for each party $P_i$ individually.

□ Calculate Entropy for each attribute ($A_1$, $A_2$, …., $A_R$) of each party $P_i$.

□ Calculate Information Gain for each attribute ($A_1$, $A_2$,…., $A_R$) of each party $P_i$

□ Calculate Total Information Gain for each attribute of all parties (TotalInformationGain( )).

□ $A_{BestAttribute}$ □ MaxInformationGain( )

□ Let $V_1$, $V_2$, …., $V_m$ be the value of attributes. $A_{BestAttribute}$ partitioned $P_1$, $P_2$,…., $P_n$ parties into m parties

□ $P_1(V_1)$, $P_1(V_2)$, …., $P_1(V_m)$

□ $P_2(V_1)$, $P_2(V_2)$, …., $P_2(V_m)$

□ . .

```
□ . .
□$P_n(V_1)$, $P_n(V_2)$, …., $P_n(V_m)$
□ Return the Tree whose Root is labelled $A_{BestAttribute}$ and has m edges labelled $V_1$, $V_2$, …., $V_m$. Such that for every i the edge Vi goes to the Tree
□ NPPID3(R – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
 End.
```

The goal of the ID3 algorithm has been to investigate the efficiency of different classical clustering algorithms in clustering network traffic data for unsupervised anomaly detection. The clusters obtained by clustering the network traffic data set are intended to be used by a security expert for manual labeling. A second goal has been to study some possible ways of combining these algorithms in order to improve their performance

Here the datasets when partitioned horizontally, vertically and after that the clustering algorithm is applied performs better performance than on the whole dataset. The proposed method is compared with the individual k-Means and ID3 methods and the other proposed approaches based on markovian chains and stochastic learning automata in terms of the overall classification performance defined over number of different performance measures. Results on real evaluation network data sets show that: the proposed method outperforms the individual k- Means and the ID3 compared to the other approaches

Many real world complex systems can be represented as graphs. The entities in these system represent the nodes or vertices and links or edges connect a pair or more of the nodes. We encounter such networks in almost any application domain i.e. computer science, sociology, chemistry, biology, anthropology, psychology, geography, history, engineering.

## V CONCLUSION

In this paper, we developed the enhanced K-Means and ID3 algorithm are hierarchical clustering manner develop with similarity measures based preprocessing approach for web usage mining. The enhanced K-Means and ID3 method is based on cascading two well-known machine learning methods: 1) the enhanced k-Means and 2) the ID3 decision trees. The enhanced k-Means method is first applied to partition the training instances into k disjoint clusters. The ID3 decision tree built on each cluster learns the subgroups within the cluster and partitions the decision space into finer classification regions; thereby improving the overall classification performance. We compare our cascading method with the individual k-Means and ID3 methods in terms of the overall classification performance defined different performance measures.

## VI REFERENCES

[1] Abraham, A. and V. Ramos (2003). Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming. Proc. Of the Congress on Evolutionary Computation (CEC 2003), Canberra, pp. 1384-1391. IEEE.

[2] Alam, S., G. Dobbie, et al. (2008). Particle Swarm Optimization Based Clustering Of Web Usage Data. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 978-0-7695-3496-1/08 © 2008 IEEE DOI 10.1109/WIIAT.2008.292 2008 IEEE/WIC/ACM International Conference on Web.

[3] Banerjee, A. and J. Ghosh (2001). Clickstream Clustering using Weighted Longest Common Subsequences. In Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, Chicago (2001).

[4] Bayir, M. A., I. H. Toroslu, et al. (2009). Smart Miner: A New Framework for Mining Large Scale Web Usage Data. (IW3C2). WWW 2009, April 20–24, 2009, Madrid, Spain. ACM 978-1- 60558-487-4/09/04.

[5] Cao, L. (2009). Data Mining and Multi-agent Integration Sydney, Springer Science + Business Media, LLC 2009.

[6] Castellano, G., A. M. Fanelli, et al. (2007). LODAP: A Log DAta 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.

[7] Chaofeng, L. (2009). "Research on Web Session Clustering." JOURNAL OF SOFTWARE VOL. 4(NO. 5).

[8] Eberhart, R. and J. Kennedy (1995). A New Optimizer Using Particle Swarm Theory. Sixth International Symposium on
Micro Machine and Human Science 0-7803-2676-8/95 1995 IEEE.

[9] Fu, Y., K. Sandhu, et al. (2000). A Generalization Based Approach to Clustering of Web Usage Sessions. M. Spiliopoulou B. Masand, editor, Web Usage Analysis and User Pro_ling, pages 21.38. Springer, 2000.

[10] Hasan, T., S. Mudur, et al. (2009). A Session Generalization Technique for Improved Web Usage Mining. WIDM'09, November 2, 2009, Hong Kong, China. Copyright 2009 ACM 978-1-60558-808-7/09/11.

[11] Huang, X., F. Peng, et al. (2004). "Dynamic Web Log Session Identification With Statistical Language Models." JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 55(14):1290–1303, 2004.

[12] Hussain, T., S. Asghar, et al. (2010). Web Usage Mining: A Survey on Preprocessing of Web Log File. IEEE, International Conference on (ICIET) 2010.

[13] Izakian, H., A. Abraham, et al. (2009). Clustering Categorical Data Using a Swarm-based Method. 978-1-4244-5612- 3/09/2009 IEEE.

[14] Khasawneh, N. and C.-C. Chan (2006). Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06) 0-7695-2747-7/06 © 2006.

[15] Kumar, R. (2009). Mining Web Logs: Applications and Challenges. KDD'09, June 28–July 1, 2009, Paris, France. ACM 978-1-60558-495-9/09/06.

[16] Lu, H. and T. T. S. Nguyen (2009). Experimental Investigation of PSO Based Web User Session Clustering. 2009 International Conference of Soft Computing and Pattern Recognition 978-0- 7695-3879-2/09 © 2009 IEEE DOI 10.1109/SoCPaR.2009.127.

[17] Makanju, A., A. N. Zincir-Heywood, et al. (2009). Clustering Event Logs Using Iterative Partitioning. KDD'09, June 28–July 1, 2009, Paris, France. Copyright 2009 ACM 978-1-60558-495- 9/09/06.

[18] Mayil, V. V. and Dr.K.Duraiswamy (2008). "Similarity Matrix Based Session Clustering by Sequence Alignment Using Dynamic Programming." Computer and Information Science, Vol. 1, No. 3, August 2008, www.ccenet.org/journal.html

[19] Mr.V.K.Panchal, M. H. Kundra, et al. (2009). "Comparative Study of Particle Swarm Optimization based Unsupervised Clustering Techniques." IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.10, October 2009.

[20] 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.

[21] Pulido, G. T. and C. A. C. Coello (2004). Using Clustering Techniques to Improve the Performance of a Multi-Objective Particle Swarm Optimizer. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'2004), pages 225–237, USA, June 2004.

[22] Rao, V. V. R. M., D. V. V. Kumari, et al. (2010). "Understanding User Behavior using Web Usage Mining." ©2010 International Journal of Computer Applications (0975 – 8887) Volume 1(No. 7).