# An Intuitive Approach for Web Scale Mining using W-Miner for Web Personalization

R.Lokeshkumar[1], Dr.P.Sengottuvelan[2]
[1 2] Department of Information Technology
Bannari Amman Institute of Technology
Sathyamangalam - 638401
Erode District, Tamil Nadu, India
[1]rlokeshkumar@yahoo.com
[2] sengottuvelan@rediffmail.com

**Abstract— Web usage mining performs mining on web usage data or web logs. It is now possible to perform data mining on web log records collected from the web page history. A web log is a listing of page reference data/click stream data. The behavior of the web page readers is imprinted in the web server log files. By looking at the sequence of pages a user accesses, a user profile could be developed thus aiding in personalization. With personalization, web access or the contents of web page are modified to better fit the desires of the user and also to identify the browsing behavior of the user can improve system performance, enhance the quality and delivery of Internet Information services to the end user and identify the population of potential customers. With clustering, the desires are determined based on similarities. In this study, a Fuzzy clustering algorithm is designed and implemented. For the proposed algorithm, meaningful behavior patterns are extracted by applying efficient Fuzzy clustering algorithm, to log data. It is proved that performance of the proposed system is better than that of the existing best algorithm. The proposed Fuzzy clustering w-miner algorithm can provide popular information to web page visitors.**

**Keyword-** Web mining, click stream data, data mining, Fuzzy clustering algorithm.

## I. INTRODUCTION

The world is becoming flat and the competition is becoming more and more scorching in various domains comparing with similar products that have the same features such as price, function and quality is also quite necessary to the consumers who are about to purchase a certain commodity (Nasraoui *et al*., 2006). Web server register a (web) log entry for every single access they get in which they save the URL requested, the IP address from which the request originated and a timestamp. With the rapid progress of World Wide Web (WWW) technology, a huge number of web log access log records are being collected. It is not easy to perform systematic analysis on such huge amount of data, however many people realized the potential usage of data to make effective use of web access history for server performance, system design improvement, or customer targeting in electronic commerce (Nasraoui and Goswami, 2006). With site mining, the overall quality and effectiveness of the pages at the site can be evaluated. The different modes of usage called user profiles can be discovered using a clustering that extract access patterns from the clickstreams stored in web log files. Using web log files, studies have been conducted on analyzing system performance, improving system design, understanding the nature of web traffic and understanding user reaction and motivation. Web sites that improve themselves by learning from user access patterns (Cross, 2004). Most of the web log analysis tools have limitations with regard to the size of the web log files. Different assumptions are made for each web analysis tools results in different statistics with the same log file. Web server log files contain useful information from which a well-designed can discover beneficial information. Web server log files customarily contain:

The domain name (or IP address) of the request; the date and time of the request; the method of the request (GET or POST); the name of the file requested; the result of the request (success, failure, error); the URL of the referring page (Ziegler *et al*., 2004). A log entry is automatically added each time a request for a resource reaches the web server.

In this study, data mining techniques are proposed to analyze web log records. Mass profiling is based on general trends of usage patterns compiled from all users on a site and can be achieved by mining user profiles from the historical data stored in server access logs (Nasraoui *et al*., 2003). I have presented an evolutionary approach, called Hierarchical

Unsupervised Niche Clustering (H-UNC), for simultaneously mining Web navigation patterns and maximally frequent context-sensitive URL item sets from the historic user access data stored in Web server logs. H-UNC necessitates fixing the number of clusters in advance, is insensitive to initialization, can handle noisy

data, general non-differentiable similarity measures and automatically provides profiles at multiple resolution levels. Unlike content based association methods, this approach also discovers associations between different Web pages based only on the user access patterns and not on the page content. Its hierarchical mode, very small population sizes contributed to making H-UNC very economical from a computational viewpoint, especially when compared to standard evolutionary computation based data mining techniques.

## II. Background References

### *A. Design of data mining system for web log records*

The Profile discovery based on web usage mining which starts with the integration and pre-processing of Web server logs and server content databases includes data cleaning and sessionization and then continues with the data mining/pattern discovery via clustering (Cooley *et al*., 1999). This is followed by a post processing of the clustering results to obtain Web user profiles and finally ends with tracking profile evolution. Web server log files contain useful data from which a well-designed data mining system can discover beneficial information. In this mining evolving user profiles project, the data collected in the web logs goes through three stages. In the first stage, the data is filtered to remove irrelevant information and a database is created containing the meaningful remaining data. This database facilitates information extraction based on individual attributes like client-IP, resource, day etc. In the second stage, it continues with pattern discovery via clustering and summarizes session clusters into user profiles.

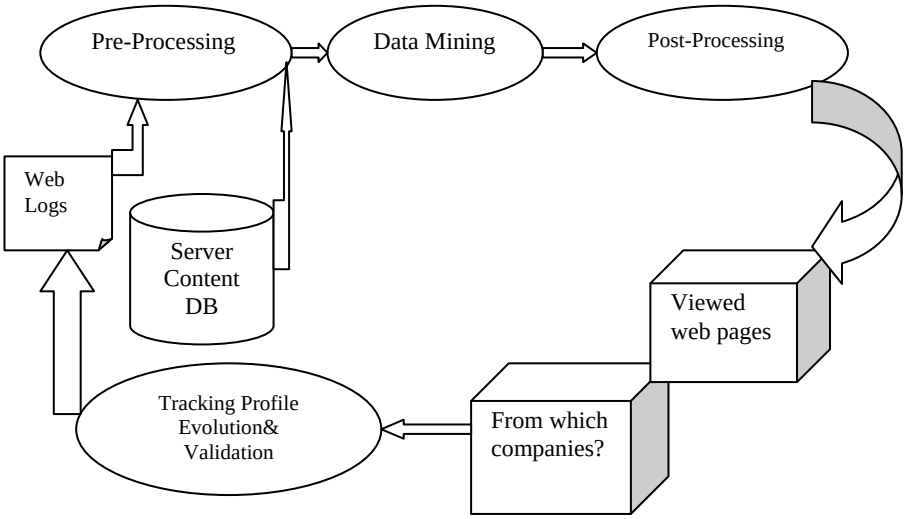


Fig. 1. Web usage mining process

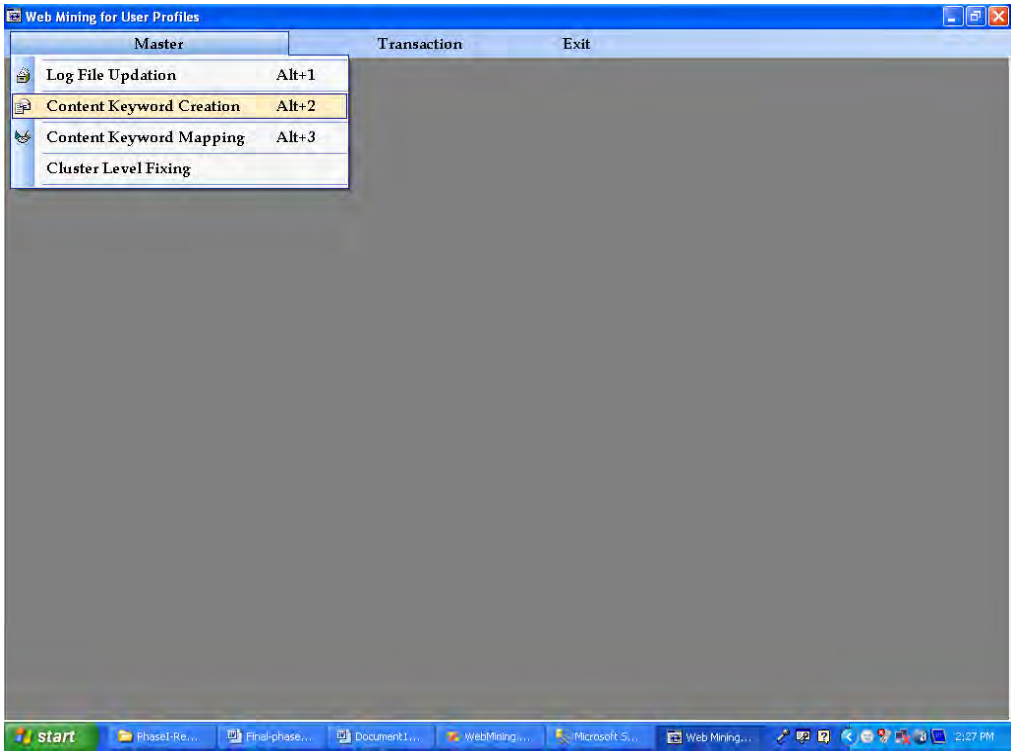


Fig. 2. Master files for web mining

Finally, in the third stage the current profiles are tracked with existing profiles. The log entries with Figure (gif, jpg) are removed. A common technique for a server site to divide the log records into sessions. A session is a set of page references from one source site during one logical period (Agarwal *et al*., 2000). A session would be identified by a user logging into a computer, performing work and then logging off. The login and logoff represent the logical start and end of the session. A transaction with few items may still be a good hub if all component items are top ranked. Conversely, a transaction with many ordinary items may have a low hub weight.

*B. Database construction from server log files*

The data filtering step may filter out requests in order to concentrate on data pertaining to actual page hits. The data filtering was adopted mainly transforms the data into a more meaningful representation. Cleaning the data and time field of the log entry, it is simply restructured in a set of fields to specify the day, month, year, hour, minute and seconds. The transformation process replaces the request sequence by the representative URL. After the cleaning and transformation of the web log entries, the web log is loaded into a relational database. To cluster user sessions, a divisive hierarchical version of a robust clustering approach (UNC) that uses Genetic Algorithm to evolve the population of candidates.

Table 1: Profile evolution for June 2008-September 2008

| Profile | June 2008 | July 2008 | August 2008 | September 2008 |
|---|---|---|---|---|
| 1 | Start | Persistence | Persistence | Persistence |
| 2 | Start | Persistence | Persistence | End |
| 3 | Start | Persistence | Persistence | End |
| 5 | Start | Atavism | Atavism | |
| 7 | Start | Persistence | Persistence | |
| 8 | Start | Atavism | | |

GA starts by applying the operators' selection, crossover and mutation. Special Crossover is the specialization performs an independent crossover for each sub chromosome in the low level. First, a measure of the distance between the sub chromosomes of the parents is computed and each sub chromosome from one parent is paired with the most similar unpaired sub chromosome from the other parent. Next, a one point crossover between the paired sub chromosomes is done (the entire sub chromosomes participate in the crossover) (Nasraoui and Krishnapuram, 2002; Masand and Spiliopoulou, 1999). Finally, the activation data of the high

level is crossed, by performing one point crossover between each pair of corresponding high level activation strings (the correspondence is obtained from the matching between the low level paired sub chromosomes).

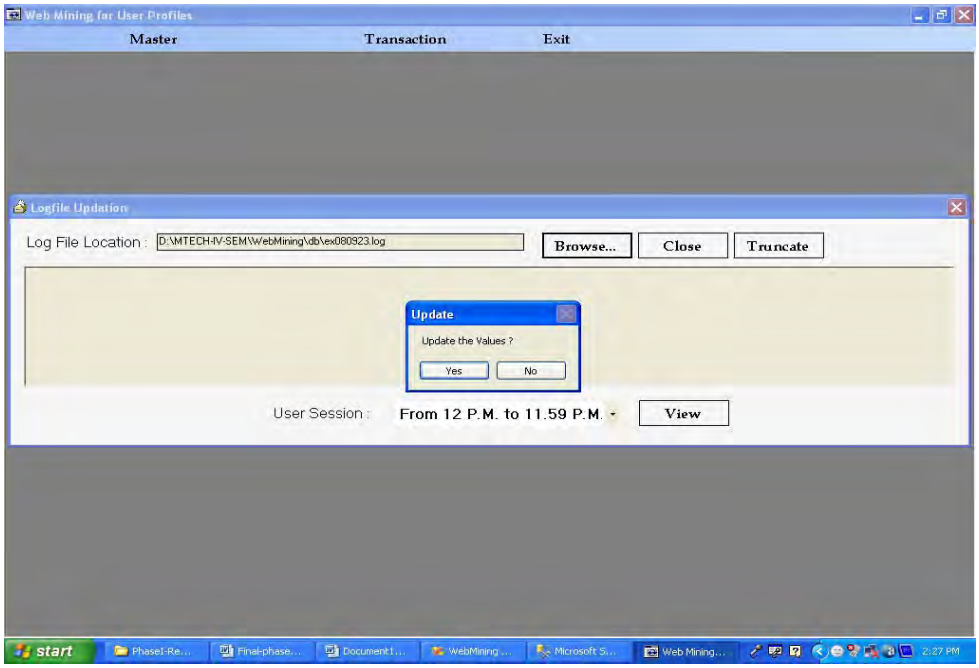


Fig. 3. Locating log file

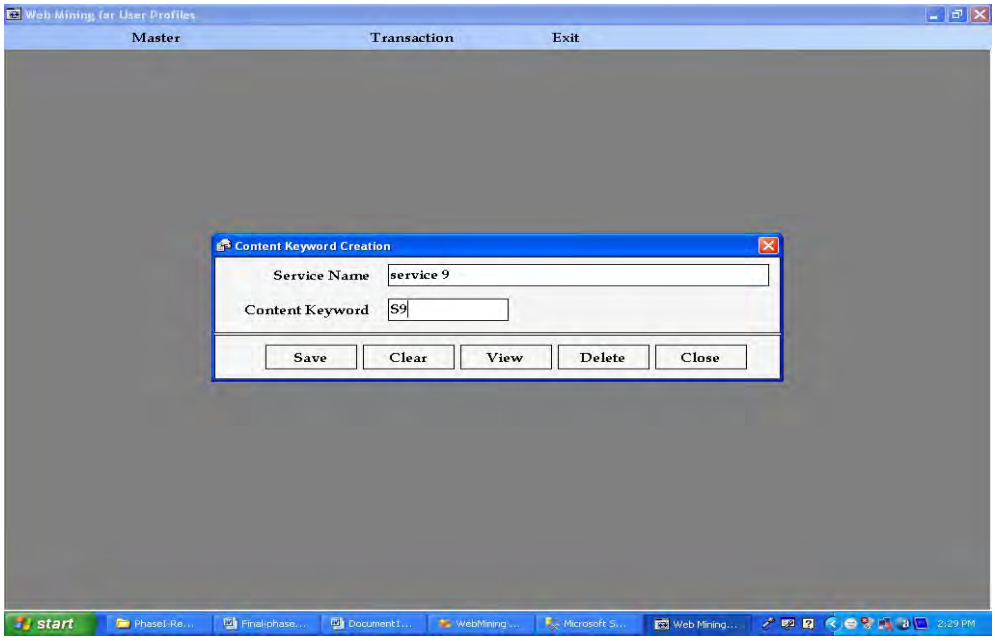


Fig. 4. Content keyword mapping

Hierarchical Unsupervised Niche Clustering (H-UNC), for mining both user profile clusters and URL associations. This approach proved to be successful in mining clusters from large web session data (Cooley *et al*., 1999). H-UNC can handle noise in the data and automatically determines the number of clusters. It is mentioned elaborately in section 3.

## III. IMPLEMENTATION

### *A. Tracking Evolving User Profiles:*

Tracking various user profile events across different time periods generate better understanding of the evolution of user access patterns. Both user profiles and clickstreams are typically evolving; each profile $p_i$ is discovered with a measure of scale $\sigma_i$ that represents the amount of variance or dispersion of the user sessions in a given cluster around the cluster representative (Zaiane et al., 1998). This measure is used to determine the boundary around each cluster and thus allows us to automatically determine whether two profiles are compatible.

The notion of compatibility between profiles is essential for tracking evolving profiles. After mining the Web log of a given period, perform an automated comparison between all the profiles discovered in the current batch and the profiles discovered in the previous batch by a sequence of SQL queries on the profiles that have been stored in a database.

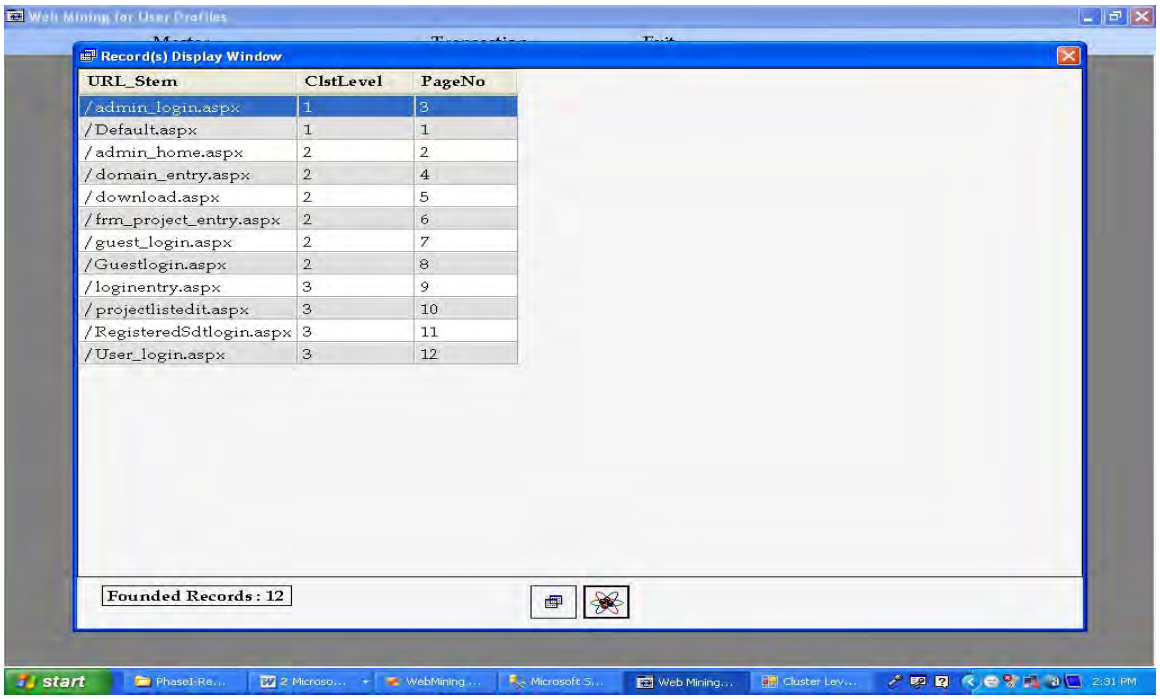


Fig. 5. Assigning cluster level and Page Number

*B. Extracting user sessions from the web log file*

The access log of a Web server is a record of all files (URLs) accessed by users on a Web site. Each log entry consists of the access time, IP address, URL viewed. The first step in pre-processing (Borges and Levene, 1999; Kleinberg, 1999) consists of mapping the $N_U$ URLs on a Web site to distinct indices.

A user session consists of requests from the same IP address within a predefined time period. Each URL in the site is assigned a unique number, where $N_U$ is the total number of valid URLs. The ith user session is then encoded as a binary attribute vector with the following property:

$$S^{(i)}j = \begin{cases} 1 & \text{if the user accessed the jth URL} \\ 0 & \text{Otherwise} \end{cases}$$

If the user is accessing specified webpage (URL) at a particular session, the value of user session is encoded as a binary attribute vector with value 1 otherwise 0.

**Algorithm:** W-miner algorithm:

**Input:** User Sessions, Maximum number of hierarchy levels Lmax minimum allowed cluster cardinality Nsplit and minimum allowed scale

Output:-User profiles

{

- Partition of the user sessions into clusters (each session is assigned to closet profile)
- Set current resolution to L //cluster representative pi is taken for consideration
- Encode binary session vectors
- Increment resolution level L = 1
- For each parent cluster representative Pi
- Apply NC on the only data records

}

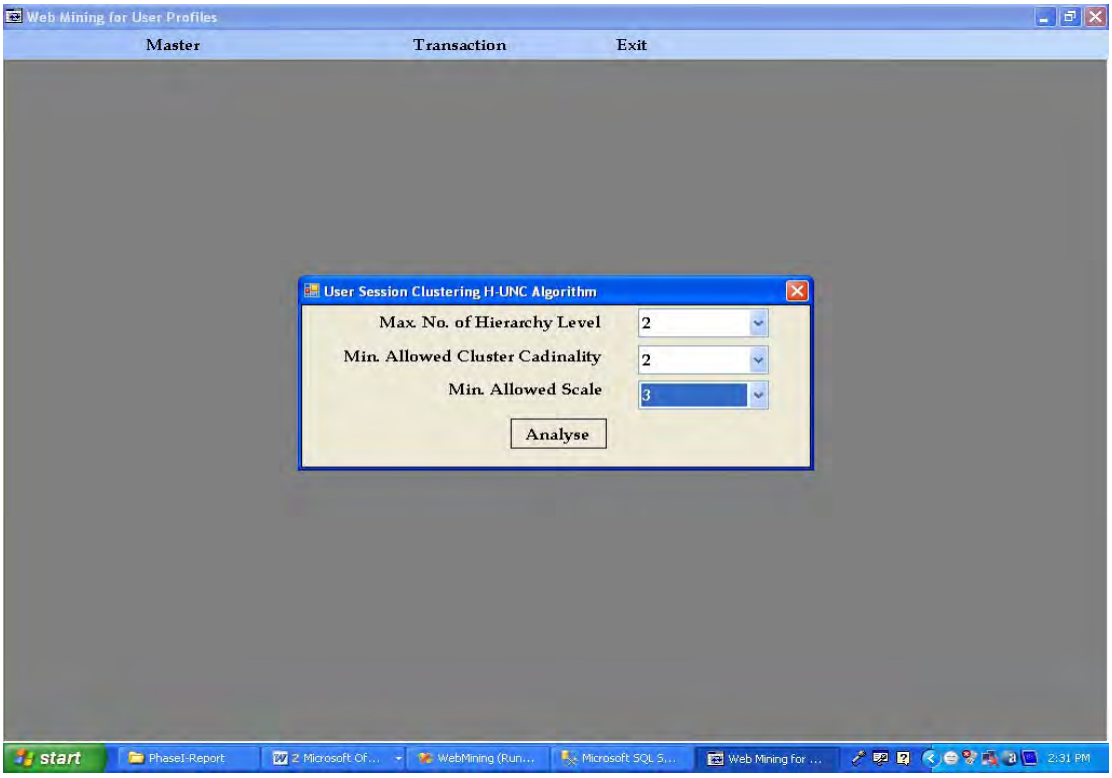


Fig. 6. User session clustering algorithm

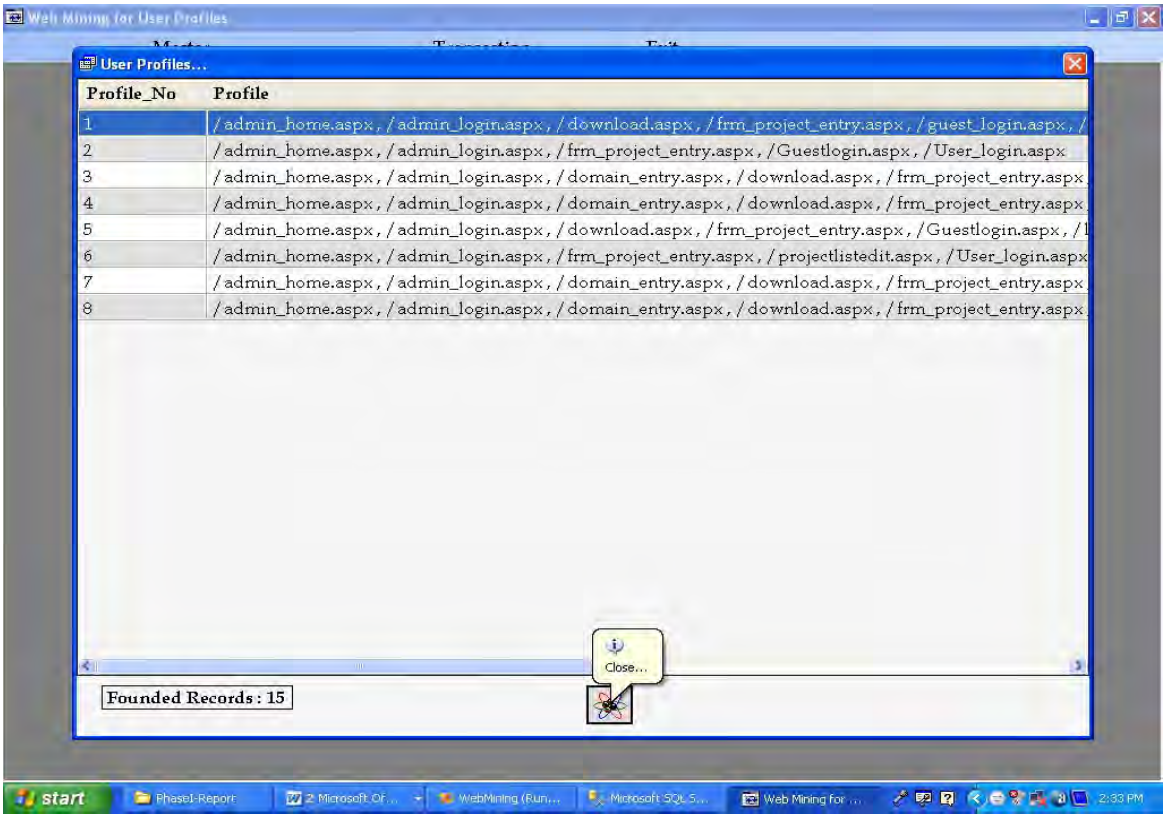


Fig. 7. Set of User Profiles

### *C. Steps to discover user profiles*

In Hierarchical Unsupervised Niche Clustering Algorithm, user profiles are generated by getting user sessions, maximum number of hierarchy levels Lmax, minimum allowed cluster cardinality Nsplit and minimum allowed scale σsplit. A user profile is a set of URLs with corresponding scale value σi. User visited web pages are encoded as binary session vectors (1). Initiate by setting current resolution level to provide

profiles at multiple resolution levels. By applying Unsupervised Niche Clustering (UNC) to the data set, this results in cluster representatives and corresponding scales. The steps are repeated recursively until it reaches the maximum number of hierarchy levels by incrementing the resolution level. Each cluster representative found at each level. In Unsupervised Niche Clustering Algorithm, user profiles are generated by getting only user sessions. Initially set scale σi. Update the distance dij of each data record to each cluster representative pi. The robust weight wij of each record to each cluster representative pi is updated using the formula.

*D. W-miner algorithm*

In hierarchical Unsupervised Niche Clustering Algorithm, user profiles are generated by getting user sessions, maximum number of hierarchy levels $L_{max}$, minimum allowed cluster cardinality $N_{split}$ and minimum allowed scale $\sigma_{split}$. A user profile is a set of URLs with corresponding scale value $\sigma_{i.}$ User visited web pages are encoded as binary session vectors (1). Initiate by setting current resolution level to provide profiles at multiple resolution levels (Ramkumar *et al*., 1998; Liu *et al*., 1998). By applying Unsupervised Niche Clustering (UNC) to the data set, this results in cluster representatives and corresponding scales. The steps are repeated recursively until it reaches the maximum number of hierarchy levels by incrementing the resolution level. Each cluster representative found at each level.

In Unsupervised Niche Clustering Algorithm, user profiles are generated by getting only user sessions. Initially set scale $\sigma_i$ (Cai *et al*., 1998). Update the distance $d_{ij}$ of each data record to each cluster representative $p_{i.}$ The robust weight $w_{ij}$ of each record to each cluster representative $p_i$ is updated using the Formula:

**Algorithm:**

- Encode binary session vectors;
- Set current resolution Level L = 1;
- Start by applying UNC to entire data set.
- Repeat recursively until $L=L_{max}$ OR all cluster cardinalities $N_i<N_{split}$ OR all scales

  $\sigma_i< \sigma_{split.}$
- Increment resolution level : L=L+1.
- For each parent cluster found at L-1;
- If cluster cardinality $N_i>N_{split}$ OR all scales $\sigma_i> \sigma_{split}$. THEN

Reapply UNC on only data records $x_j$ assigned to cluster representatives' $p_i$.

$$W_{ij} = e^{-d_o} f(2\sigma_i)$$

The scale $\sigma_{i..}$ of each record to each cluster representative is updated using the formula:

$$\sigma_i = \begin{cases} \frac{\sum w_o d_w}{\sum w_s} \end{cases}$$

The fitness $f$i of each cluster representative is updated using the formula:

$$f_i = \frac{\sum w_o}{\sigma_i}$$

After pairing of 2 parents, each parent is replaced by the most similar child only if the latter higher fitness. If child's fitness is greater than closest parent fitness child replaces parent in the new population, other parent remains in the new population.

*E. Tracking User Profiles*

It describes how profiles discovered during as certain period relate to profiles discovered in another period. It determines which new profiles are compatible with old profiles and which new profiles are incompatible with previous profile. Removing irrelevant information from the web log file to avoid confusion. Pre-processing of web server logs and server content databases includes data cleaning process. The system should be able to extract set of user sessions. An event start indicates that a new profile has never before been observed, an end states a cluster of user activities that vanishes (Cooley *et al*., 1997).

Atavism is labelled that a profile disappears temporarily then re-emerges again at a later period; persistence is that the same profile is observed to continue. Collection of Web Server Logs from the web server. It is necessary for selecting user sessions. Log entry consists of the access time, IP address. The system should retrieve the user sessions from web log record and it should be stored in the database. Web server automatically generates web log records both on client side and server side. The users' activities/click streams which are recorded in web server logs. First the web log file should be located for a particular session. If it is considered to be an essential one, it may be included and database should be updated. Otherwise, the collected web log details can be truncated. It displays the details of client IP, URL stem, status, method, date-time.

## IV. CONCLUSION

In this paper a novel approach for Web scale mining is found out. User profile summarizes a group of users with similar access activities and consists of visited web pages. Mapping new session to persistent profiles and updating these profiles confine the search towards interestingness. Tracking different profile events across various time periods generate a better understanding of the evolution of user access patterns. By identifying the potential customers, reliable knowledge about the customer preferences aids in improving the target pages to retain the customers. This traditional method has computation cost problems so the link-based model is useful in adjusting the mining results given by the traditional techniques. Some interesting patterns may be discovered when the transactions are taken into account. The future development will continue with a main focus to implement task process enhancement in Web Mining (Data clustering is the task of partitioning a multivariate data set into groups maximizing intra-group similarity and inter-group dissimilarity). Efforts to improve efficiency by reducing the effects of interference by the sub-agents will also be aimed at execution time estimation model for Web Mining jobs.

## REFERENCES


[1] Agarwal, R., C. Aggarwal and V.V.V. Prasad, 2000. A tree projection algorithm for generation of frequent itemsets. J. Parallel and Distributed Comput.

[2] Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94). pp: 487-499.

[3] Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large datasets. Proc. ACM SIGMOD' 93. pp: 207-216.

[4] Borges, J. and M. Levene, 1999. Data Mining of user Navigation Patterns, Web usage Analysis and User Profiling, LNCS. Abbas, H.A., R.A Sarker and C.S. Newton, (Eds.) pp: 29-111.

[5] Cai, C.H., A.W.C. Fu, C.H. Cheng and W.W. K .Wong, 1998. Mining association rules with weighted items. Proc. IEEE Int'l Database Eng. Appli. Symp. (IDEAS '98), pp: 68-77.

[6] Cooley, R., B. Mobasher and J. Srivastava, 1997. Web mining: information and pattern discovery on the world wide web. Proceedings of the 9th IEEE Int'l Conf. Tools with AI (ICTAI '97). pp: 558-567.

[7] Cooley, R., B. Mobasher and J. Srivastava, 1999. Data preparation for mining world wide web browsing patterns. Knowledge and Infor. Syst., vol: 1.

[8] Cooley, R., P.-N. Tan and J. Srivastava, 1999. Discovery of interesting usage patterns from web data. Technical Report TR 99-022, University of Minnesota.

[9] Cross, V., 2004. Fuzzy semantic distance measures between ontological concepts. Proceeding of the Annual Meeting North Am. Fuzzy Information Processing Soc. (NAFIPS '04), pp: 392-397, June 2004.

[10] Desikan, P. and J. Srivastava, 2004. Mining temporally evolving graphs. Proceeding of the Workshop Web Mining and Web Usage Analysis (WebKDD' 04).

[11] Han, J., J. Pei and Y. Yin, 2000. Mining frequent patterns without candidate generation. Proceeding of the ACM SIGMOD.

[12] Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. J. ACM, 46: 604-632.

[13] Liu, B., W. Hsu and Y. Ma, 1998. Integrating classification and association rule mining. Proc. ACM SIGKDD '98. pp: 80-86.

[14] Masand, B. and M. Spiliopoulou, 1999. Workshop on Web Usage Analysis and User Profiling, (webKDD).

[15] Nasraoui, O. and C. Rojas, 1993. From Static to Dynamic Web Usage Mining: Towards Scalable Profiling and Personalization with Evolutionary Computation.

[16] Nasraoui, O. and R. Krishnapuram, 2002. A new evolutionary approach to web usage and context sensitive associations mining. Int'l J. Comput. Intelligence and Appli. Special Issue on Internet intelligent Syst., 2: 339-348. Dai, H. and B. Mobasher, 2002. Using ontologies to discover domain- level web usage profiles. Proceeding of the Second ECML/PKDD Semantic Web Mining Workshop.

[17] Nasraoui, O., C. Cardona, C. Rojas and F. Gonzalez, 2003. mining evolving user profiles in noisy web clickstream data with a scalable immune system clustering algorithm. Proceeding of the Workshop Web Mining as a Premise to Effective and Intelligent Web Applications (WebKDD '03). pp: 71-81, Aug. 2003.

[18] Nasraoui, O., C. Rojas and C. Cardona, 2006. A framework for mining evolving trends in web data streams using dynamic learning and retrospective validation. Comput. Networks, special Issue on Web Dynamics, vol: 50. (3) Nasraoui, O. and S. Goswami, 2006. Mining and validating localized frequent itemsets with dynamic tolerance. Proceeding of the Sixth SIAM Int'l Conference Data Mining (SDM '06). pp: 578-582, Apr. 2006.

[19] Nasraoui, O., R. Krishnapuram, H. Frigui and A. Joshi, 2000. Extracting web user profiles using relational competitive fuzzy clustering. Int'l J. Artificial Intelligence Tools, 9: 509-526.

[20] Nasrauoi, O., M. Solman, E. Saka, A. Badia and R. Germain, 2008. A web usage mining framework for mining evolving user profiles in dynamic web sites.IEEE Trans. Knowledge and Data Eng., vol: 20.

[21] Oberle, D., B. Berendt, A. Hotho and J. Gonzalez, 2003. Conceptual user tracking. Proceeding of the First Int'l Atlantic Web Intelligence Conf. (AWIC '03).Ramkumar, G.D., S. Ranka and S. Tsur, 1998. Weighted association rules. Model and algorithm. Proc. ACM SIGKDD.

[22] Zaiane, O.R., M. Xin and J. Han, 1998. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. Proceeding of the Advances in Digital Libraries (ADL'98). pp: 19-29.

[23] Ziegler, C., G. Lausen and L. Schmidt-Thieme, 2004. Taxonomy-driven computation of product recommendations. Proceeding of the 13th ACM Conf. Information and Knowledge Management (CIKM '04). pp: 406-415.