ISSN:0975-9646

Neelam Sain et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4) , 2012,4858-4862

A Survey of Web Usage Mining based on Fuzzy Clustering and HMM

Neelam Sain[*1], Prof. Sitendra Tamrakar[#2]

[#2] *Asst. Prof. Dept. of Computer Science*
NRI Institute of information
Science and technology, Bhopal(India)

[*1] *Dept. of Computer Science, RGPV*
NRI Institute of information
Science and technology, Bhopal(India)

***Abstract*-- Web usage mining is a kind of data mining method that can be useful in recommending the web usage patterns with the help of users' session and behavior. Web usage mining includes three process, namely, preprocessing, pattern discovery and pattern analysis. There are different techniques already exists for web usage mining. Those existing techniques have their own advantages and disadvantages. This paper presents a survey of over 34 research papers dealing with Web usage Mining technique based on Fuzzy clustering and HMM (Hidden Markov Model) the advantage of the technique is that it can measure the similarity efficiently among the users on the basis of their browsing characteristics and it also accurately predict the user patterns.**

***Keywords*- Web Usage Mining, Web Log mining, pre-fetching, fuzzy clustering, HMM**.

I. INTRODUCTION

Web mining, can be categorized into three types, Content Mining, Structure Mining and Usage Mining. Web content mining can be described as the process of extracting knowledge from the content or descriptions of web documents. In web content mining there are two dominant groups of strategies: Web page content mining and Search result mining. Web content mining has to do with the retrieval of information (content) available on the web into more structured forms as well as its indexing for easy tracking information locations.

In[1], Supervised Fuzzy Clustering for Rule Extraction, the application of orthogonal transforms
and fuzzy clustering to extract fuzzy rules from data. The orthogonal least squares method to supervise the progress of the fuzzy clustering algorithm and remove clusters of less importance with respect to describing the data. Clustering takes place in the product space of systems inputs and outputs and each cluster corresponds to a fuzzy IF-THEN rule (2000).
In [2], Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, overview of the Web SIFT system as an example of a prototypical Web usage mining system is given.

An important constituent category of Web Mining is Web Log mining also known as Web Usage mining, is the process of extracting interesting patterns from web access logs [Zaiane (2001)]. Web Usage mining imitates the actions of humans as they interact with the Internet [Vellingiri and Pandian (2011)]. It can also be defined as the process of identifying browsing patterns by analyzing the user's navigational behavior. This information takes as input the usage data, i.e. the data residing in the Web server logs, recording the visits of the users to a Web site. Extensive research in the area of Web usage mining led to the appearance of a related research area, that of Web personalization. Web personalization utilizes the results produced after performing Web usage mining, in order to dynamically provide recommendations to each user [Kosala and Blockeel (2000)].
In [3]A Data Clustering Algorithm Based On Single Hidden Markov Model, which identifies a suitable number of clusters in a given dataset without using prior knowledge about the number of clusters. Initially, the dataset is partitioned into windows of fixed size based on the HMM log likelihood values. This provides a framework for identifying the most appropriate number of clusters (windows of varying sizes). After determining the number of clusters, the data values are then labeled and allocated to clusters. The algorithm is tested using a number of benchmark datasets. The proposed algorithm for both small and large datasets (KDD 1999 Intrusion Detection dataset) performed significantly better compared to other commonly used clustering algorithms. [Md. RafiulHassan (2006)].
B. de la Ossa, J. A. Gil, In [4], Improving Web Prefetching by Making Predictions at Prefetch, reduce the user's perceived latency with no additional cost over the basic prefetch mechanism.
The process of web usage mining model falls into four sections Ya-Xiu Yu [5]. Those are source data collection phase, data pretreatment phase, pattern mining phase and pattern analysis phase. The process is shown in figure.

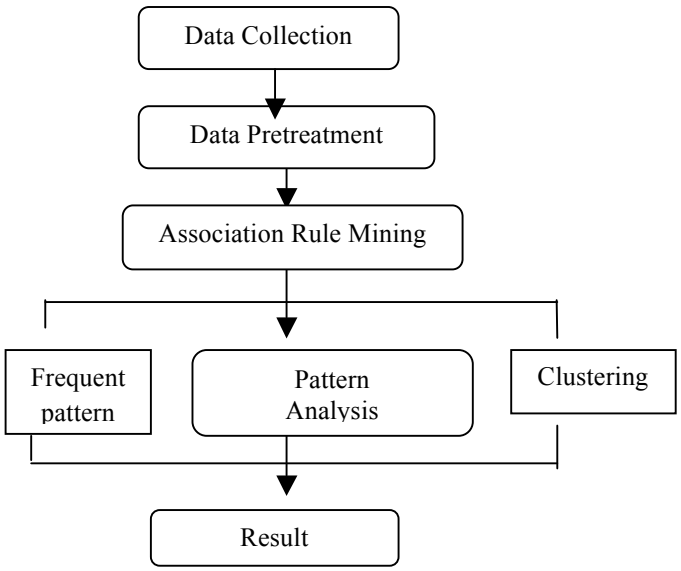

Fig. The process of web usage mining model (Data mining)

II. LITERATURE REVIEW

It is mention here that, although the Fuzzy clustering, HMM and web usage mining is shown above, yet different authors have given their contributions in various techniques /components as discussed below**.**

Web usage information mining could help to engage new customers, maintain current customers, track customers who are leaving web site, and so on [6]. Usage information can be extracted to increase web server efficiency by prefetching and caching strategies [7].Based on several researches done in the area of web mining, we can broadly classify it into three domains: web content mining, web structure mining, and web usage mining. Web content mining is the process of extracting knowledge from web documents such as text and multimedia.

Knowledge extraction from the structure of web and hyperlink references is called web structure mining. Web usage mining is the process of knowledge exploitation from the secondary data [2]. By secondary data, we mean browser logs, user profiles, web server access logs, registration data, user sessions or transactions, cookies, user queries, mouse clicks and any other data that is the result of interaction with the Web.

In [8] A Fuzzy Rule-Based Clustering Algorithm the FRBC employs a supervised classification approach to do the unsupervised cluster analysis. It tries to automatically explore the potential clusters in the data patterns and identify them with some interpretable fuzzy rules. Simultaneous classification of data patterns with these fuzzy rules can reveal the actual boundaries of the clusters. To illustrate the capability of FRBC to explore the clusters in data, the experimental results on some benchmark datasets are obtained and compared with other fuzzy clustering algorithms. The clusters specified by fuzzy rules are human understand able with acceptable accuracy.

In [7] They proposed GFCR becomes an alternative model of the generalized FCM (GFCM) that was recently proposed by Yu and Yang. To advance theoretical study, they have the following three considerations. 1) We give an optimality test to monitor if GFCR converges to a local minimum. 2) We relate the GFCR optimality tests to Occam's razor principle, and then analyze the model complexity for fuzzy clustering algorithms. 3) We offer a general theoretical method to evaluate the performance of fuzzy clustering algorithms. Finally, some numerical experiments are used to demonstrate the validity of our theoretical results and complexity analysis.

Qiang Yang, [10] they present an application of web log mining to obtain web-document access patterns and use these patterns to extend the well-known GDSF caching policies and pre-fetching policies. Using real web logs, they show that this application of data mining can achieve dramatic improvement to web-access performance.

In [11] Qiang Yang and Haining Henry Zhang have implement Web-log mining method for caching Web objects and use this algorithm to enhance the performance of Web caching systems. They develop an n-gram-based prediction algorithm that can predict future Web requests. The prediction model is then used to extend the well-known GDSF caching policy. Show that the system performance is improved using the predictive-caching approach. such as caching, for dynamically changing database-driven Web sites. For such Web sites, query-level prediction algorithms need to be studied.

Wei-Guan and Teng, In [3] they propose an innovative cache replacement algorithm, which not only considers the caching effect in the Web environment, but also evaluates the prefetching rules provided by various prefetching schemes. In 2005 specifically, they formulate a normalized profit function to evaluate the profit from caching an object (i.e., either a non implied object or an implied object according to some prefetching rule). Based on the normalized profit function devised, they devise an innovative Web cache replacement algorithm, referred to as Algorithm IWCP (standing for the Integration of Web Caching and Prefetching). Using an event-driven simulation, we evaluate the performance of Algorithm IWCP under several circumstances. The experimental results show that Algorithm IWCP consistently outperforms the companion schemes in various performance metrics.

WENYING FENG and HUA CHEN in [13] have introduced a method for Matrix Algorithm for Web Cache Pre-fetching they present a new Web Cache Pre-fetching scheme, the matrix algorithm. There method is simple to implement and adopts the idea of machine learning into caching systems. To develop the simulation program, they propose and implement a topic request model on the client side that is also by matrix application. Results from the simulation show that our new algorithm significantly improves cache performance measured by hit rates

Some necessary modifications were presented to Dynamic Web Pre-fetching Technique for Latency Reduction by Achuthsankar S. Nair, Jayasudha J.S. in [14]. A dynamic pre-fetching technique in which web caching and pre-fetching techniques are integrated. In this technique, number of subsequent links to be pre-fetched depends on the user's interest in accessing the documents, cache contents, current bandwidth usage and maximum capacity of the existing network. Preference lists are used for maintaining user's preferences. A hash table is used for storing the list of accessed URLs and its weight information. Intelligent agent monitors the bandwidth usage and helps the prediction engine to decide the number of web pages to be pre-fetched. The simulation result shows that dynamic pre-fetching technique provides better utilization of bandwidth and reduces latency. Using dynamic pre-fetching technique, cache hit ratio is increased to 40% –75% and latency is reduced to 20% – 63%. Web pre-fetching techniques are used for reducing latency, but it increases web traffic. In dynamic web pre-fetching technique, subsequent links are pre-fetched only if bandwidth usage of existing network is less than a predefined threshold. For each web page request, the retrieved page is parsed to identify the subsequent links and URL's corresponding to these links is searched in the hash table to get its weight information. Intelligent agents monitor the bandwidth usage, user's preferences and hash table weights to identify the number of URLs to be pre-fetched. That dynamic pre-fetching browser maintains almost constant web traffic even if pre-fetching is done. Since dynamic pre-fetching technique increases cache hit ratio, reduces latency and maintains almost constant traffic, it is preferred to all the existing pre-fetching techniques.

B. de la Ossa, J. A. Gil, J. Sahuquillo and A. Pont in [15] They explains how a pre-fetching technique can be extended to include our P@P proposal on real world conditions without changes in the web architecture or HTTP protocol. To show how that proposal can improve pre-fetching performance an extensive performance evaluation study has been done and the results show that P@P can considerably reduce the user's perceived latency with no additional cost over the bas pre-fetch mechanism.

They discussed on detail what characteristics are required in the web browser and the prediction engine in order to perform Prediction at Pre-fetch in a safely way. Mozilla is a well known web browser that satisfies all the requirements; thus, it can be used without any modification. Regarding the web server and the prediction engine, we proposed and implemented Prediction at Pre-fetch on Delfos, and tested it with Mozilla on real world usage. The additional aggressiveness allowed by the prediction engine when using the proposed technique reduces the latency per object perceived by the user at expenses of increasing the traffic. The effectiveness of Prediction at Pre-fetch with different prediction thresholds has been checked. The results of the experiments show that a properly configured prediction engine provides a good ratio cost-benefit. A latency reduction up to 14% was achieved while increasing by about 8% the byte traffic.

Payal Gulati, Dr. A. K. Sharma, Dr. Amit Goel and Jyoti Pandey in [16] reduction of World Wide Web user perceived latency has assumed importance in the wake of the fast development of Internet services and a huge amount of network traffic and hence adaptation of web pages to the needs of a specific user is today's trend of web technologies. Although web performance can be improved by caching, the benefit of using it is rather limited owing to filling the cache with documents without any prior knowledge. Web prefetching becomes an attractive solution wherein forthcoming page accesses of a client are predicted, based on access log information. They propose a Zipf's Law based novel approach for the determination of next page likely to be accessed by specific client.

found that by using the Zipf estimator, the probability of accessing the next page has been computed efficiently. Depending upon the probability of the next page, the web pages are pre-fetched locally on the proxy server so as to reduce the retrieval latency.

Sarina Sulaiman, Siti Mariyam Shamsuddin, Ajith Abraham and Shahida Sulaiman in [17]. They express the discussion on what is the Web caching and pre-fetching, why we have to opt its and how to pertain of these two technologies. Caching and pre-fetching can work individually or combined. The blending of caching and pre-fetching enables doubling the performance compared to single caching. These two techniques are very useful tools to reduce congestion, delays and latency problems. Consequently, basic knowledge on how these both techniques work; architectures/deployment scheme, placement and replacement algorithms and lastly how to measure its performance are essential to realize an accomplishment of the Web caching and pre-fetching

Heung Ki Lee, Baik Song An, and Eun Jung Kim in [18] They attempt to design an adaptive web pre-fetch scheme by predicting memory status more accurately and dynamically. First, they design Double Prediction-by-Partial-Match Scheme (DPS) that can be adapted to the modern web framework. Second, they propose Adaptive Rate Controller (ARC) to determine the pre-fetch rate depending on the memory status dynamically. Finally, they suggest Memory Aware Request Distribution (MARD) that distributes requests based on the available web processes and memory. For evaluating the pre-fetch gain in a server node, we implement an Apache module in Linux. In addition, they build a simulator for verifying our scheme with cluster environments. Simulation results show 10% performance improvement on average in various workloads.

They implement the prototype of web prefetch engine using PAPI and Apache web server in the Linux. Also, they perform the simulation for verifying the benefit of our scheme in cluster environments. there experimental results show that our prefetch scheme improves the performance of web cluster system up to 40% in various web workloads.

Introduced the new method by Johann M´arquez, Josep Dom`enech, Jos´e A. Gil and Ana Pont in [19] they propose a

novel global framework for performance evaluation in scenarios where different parts of the web architecture interact. Unlike existing proposals our approach is a fast and flexible tool that allows to represent faithfully the behavior of each element of the architecture in order to study, reproduce, evaluate and design web strategies to decrease the user´s perceived latency when surfing the web.

Ya-Xiu Yu and Xin-Wei Wang in [5] attempt To cluster similar web user, by considering two factors that the page-click number and web browsing time, which is stored in the web log, and the different degree of influence of the two factors. They proposed a technique that can help web site organizations to recommend web pages, improve web structure, so that can attract more customers, and increase customers' loyalty.

They introduced a novel web usage fuzzy clustering method, our initial analysis of the clustering result suggests that this clustering strategy can effectively group similar queries together. There proposed strategy can help web site owners to provide personalized service. A possible future enhancement in there work can be that the application environment of this method is a single E-commerce web site, how to apply it to whole internet needs to do more improvement. So a hidden Marko model can be very useful as an addition to there work to be applied on whole internet.

By A. Anitha in [20] They proposed to predict next page access identify similar access patterns from web log using pair-wise nearest neighbor based clustering and then sequential pattern mining is done on these patterns to determine next page accesses. The tightness of clusters is improved by setting similarity threshold while forming clusters. In traditional recommendation models, clustering by non-sequential data decreases recommendation accuracy. they proposed to integrate Markov model based sequential pattern mining with clustering. A variant of Markov model called dynamic support pruned all kth order Markov model is proposed in order to reduce state space complexity. Mining the web access log of users of similar interest provides good recommendation accuracy. the proposed model provides accurate recommendations with reduced state space complexity. The resulted in good prediction accuracy with less state space complexity. The drawback of this work is, loosely connected access sequences are not considered for mining process. it is suggested to extend this work by considering noncontiguous access sequences also.

R.Khanchana and M. Punithavalli in [21] they enhances the two levels of Prediction Model to achieve higher hit ratio. The uses Fuzzy Possibilistic algorithm for clustering. The experimental result shows that the proposed techniques results in better hit ratio.

Forecasting the user's browsing pattern is a significant technique for many applications. The Forecasting results can be utilized for personalization, building proper web site, enhancing marketing strategy, promotion, product supply, getting marketing data, forecasting market trends, and enhancing the competitive strength of enterprises etc. The uses web usage mining technique for predicting the user's browsing behavior. One of the effective existing techniques for web usage mining is the usage of hierarchical agglomerative clustering to cluster users' browsing behaviors. The usage of Two Levels of Prediction Model framework is explained in this paper which works better for general cases. However, Two Levels of Prediction Model suffer from the heterogeneity user's behavior. To overcome this difficulty, this paper uses Fuzzy Possibilistic algorithm for clustering. The experimental result shows that the proposed technique results in higher hit rate.

III. CONCLUSIONS

In this paper we have presented a survey of over 34 research papers analyzing method/ techniques / phases for Web mining, the application of data mining and Web Personalization techniques.

We have made a systematic statement on web usage mining, and the introduced a Hidden Markov model, web usage fuzzy clustering method, our initial analysis of the clustering result suggests that this clustering strategy can effectively group similar queries together. This proposed strategy can help web site owners to provide personalized service.

The proposed method resulted in good prediction accuracy. The user's browsing pattern is a significant technique for many applications. The results can be utilized for personalization, building proper web site, enhancing marketing strategy, promotion, product supply, getting marketing data, market trends, and enhancing the competitive strength of enterprises etc.

IV. REFERENCES

[1] Magne Setnes, "Supervised Fuzzy Clustering for Rule Extraction", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 8, NO. 4, AUGUST 2000.

[2] Jaideep Srivastava_ y , Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Volume 1, Issue 2, ACM SIGKDD, Jan 2000.

[3] Md. Rafiul Hassan, Baikunth Nath and Michael Kirley, "A Data Clustering Algorithm Based On Single Hidden Markov Model", Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 57 – 66, ISSN 1896-7094, © 2006.

[4] B. de la Ossa, J. A. Gil, J. Sahuquillo and A. Pont, "Improving Web Prefetching by Making Predictions at Prefetch", 1-4244-0857-1/07/$25.00 ©2007 IEEE.

[5] Ya-Xiu Yu, Xin-Wei Wang, "Web 978-0-7695-3600-2/09 $25.00 © 2009 IEEE DOI 10.1109/IFITA.2009.

[6] A. Abraham., "Natural Computation for Business Intelligence from Web Usage Mining", Proceeding of Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNAC2005), pp. 3-11, 2005.

[7] F. Masseglia, P. Poncelet, and R. Cicchetti, "An Efficient Algorithm for Web Usage Mining", Networking and Information Systems Journal (NIS), 2(5-6), pp. 571-603, 1999.

[8] Eghbal G. Mansoori, "FRBC: A Fuzzy Rule-Based Clustering Algorithm", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 19, NO. 5, OCTOBER 2011

[9] Jian Yu and Miin-Shen Yang, "A Generalized Fuzzy Clustering Regularization Model With Optimality Tests and Model Complexity Analysis", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 15, NO. 5, OCTOBER 2007.

[10] Qiang Yang, Haining Henry Zhang and Tianyi Li, "Mining Web Logs for Prediction Models in WWW Caching and Prefetching", In The Seventh ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining KDD'01, August 26 - 29, 2001 San Francisco, California, USA.

[11] Qiang Yang and Haining Henry Zhang, "Web-Log Mining for Predictive Web Caching", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 15, NO. 4, JULY/AUGUST 2003

[12] Wei-Guang Teng, Cheng-Yue Chang, and Ming-Syan Chen, "Integrating Web Caching and Web Prefetching in Client-Side Proxies", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 16, NO. 5, MAY 2005.

[13] WENYING FENG and HUA CHEN, "A Matrix Algorithm for Web Cache Pre-fetching", 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007) 0-7695-2841-4/07 $25.00 2007

[14] Achuthsankar S. Nair and Jayasudha J.S., "Dynamic Web Pre-fetching Technique for Latency Reduction", International Conference on Computational Intelligence and Multimedia Applications 2007.

[15] B. de la Ossa, J. A. Gil, J. Sahuquillo and A. Pont, "Improving Web Prefetching by Making Predictions at Prefetch", IEEE 1-4244-0857-1/07/$25.00 2007.

[16] Dr. A. K. Sharma, Dr. Amit Goel and Jyoti Pandey, "A Novel Approach for Determining Next Page Access", First International Conference on Emerging Trends in Engineering and Technology, 978-0-7695-3267-7/08 $25.00 © 2008 IEEE DOI 10.1109/ICETET.2008.

[17] Sarina Sulaiman, Siti Mariyam Shamsuddin, Ajith Abraham and Shahida Sulaiman, "Web Caching and Prefetching: What, Why, and How?", IEEE 978-1-4244-2328-6/08/$25.00 , 2008.

[18] Heung Ki Lee, Baik Song An, and Eun Jung Kim, "Adaptive Prefetching Scheme Using Web Log Mining in Cluster-basedWeb Systems", IEEE International Conference on Web Services, 978-0-7695-3709-2/09 $25.00,DOI 10.1109/ICWS. 2009.

[19] Johann M´arquez, Josep Dom`enech, Jos´e A. Gil and Ana Pont, "A Web Caching and Prefetching Simulator" IEEE conference, E-ISBN : 978-953-290-009-5, 2008.

[20] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications (0975 – 8887) Volume 8– No.11, October 2010.

[21] R.Khanchana and M. Punithavalli., "Web Usage Mining for Predicting Users' Browsing Behaviors by using FPCM Clustering", IACSIT International Journal of Engineering and Technology, Vol. 3, No. 5, October 2011.

[22] Sotirios P. Chatzis and Theodora A. Varvarigou," A Fuzzy Clustering Approach Toward Hidden Markov Random Field Models for Enhanced Spatially Constrained Image Segmentation", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 16, NO. 5, OCTOBER 2008.

[23] Eghbal G. Mansoori, "FRBC: A Fuzzy Rule-Based Clustering Algorithm", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 19, NO. 5, OCTOBER 2011.

[24] Samia Nefti, Mourad Oussalah and Uzay Kaymak, "A New Fuzzy Set Merging Technique Using Inclusion-Based Fuzzy Clustering", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 16, NO. 1, FEBRUARY 2008.

[25] Olfa Nasraoui, Maha Soliman, Esin Saka, Antonio Badia and Richard Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 2, FEBRUARY 2008.

[26] Andrew Skabar and Khaled Abdalgader, "Clustering Sentence-Level Text using a Novel Fuzzy Relational Clustering Algorithm", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, TKDE-2010-09-0519.R2, IEEE 1041-4347/11/$26.00 © 2011.

[27] Lin Zhu, Fu-Lai Chung, and Shitong Wang, "Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 39, NO. 3, JUNE 2009.

[28] Magne Setnes, "Supervised Fuzzy Clustering for Rule Extraction", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 8, NO. 4, AUGUST 2000.

[29] Sankar K. Pal, Varun Talwar and Pabitra Mitra, "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 5, SEPTEMBER 2002.

[30] Md. Rafiul Hassan, Baikunth Nath and Michael Kirley, "A Data Clustering Algorithm Based On Single Hidden Markov Model", Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 57 – 66, ISSN 1896-7094 © 2006 PIPS.

[31] Kobra Etminani, Mohammad-R. Akbarzadeh-T and Noorali Raeeji Yanehsari, "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method", ISBN: 978-989-95079-6-8, IFSA EUSFLAT 2009.

[32] Sandro Araya, Mariano Silva and Richard Weber, "A methodology for web usage mining and its application to target group identi'cation", Fuzzy Sets and Systems 148 (2004) 139–15 , doi:10.1016/j.fss.2004.03.011, 2004.

[33] Yongjian Fu, Kanwalpreet Sandhu and Ming-Yi Shin, "Clustering of Web Users Based on Access Patterns", In Proceedings of the 1999 KDD Workshop on Web Mining,1999.

[34] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", University of Minnesota Minneapolis, MN 55455, USA, 1997.