International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015) pp 37488-37494
© Research India Publications. http://www.ripublication.com

Web Usage Mining: Discovery Of The User’s Navigational Patterns Using ELM And SKPCM

D.Anandhi

Department of Computing, Coimbatore Institute of Technology,
Affiliated to Anna University,Coimbatore-641014,India.
anandhiphd2014@gmail.com

M.S. Irfan Ahmed

Department of Computer Applications, Sri Krishna College of Engineering and Technology,
Affiliated to Anna University,Coimbatore-641008,India.
msirfan@gmail.com

ABSTRACT

Web log mining is the emerging technique which is widely used in E-Commerce business, in particular area such as site-reorganization, link prediction, pre-fetching. Web log mining is the process of retrieving the necessary pattern from the log files which is stored in the web server. This paper proposes a novel technique to extract the required information from log files by data cleaning, classification and clustering process. The Extreme learning machine is used for the classification process. The Similarity and Kernel based PCM (SKPCM) clustering technique is applied for the data clustering. The Experimental result indicates that the SKPCM provides the efficient output.

Keywords: Classification, Data cleaning, User identification, Sessions, Kernel.

INTRODUCTION

Web log mining is the process of analyzing, retrieving the user access pattern from the web log files [1][2][3]. The identification of the service provided by a particular web site can be identified easily with the web log mining process.

Nowadays the E-Commerce businesses growth mainly focused in web log mining [4] to discover the user behavior, buyer activities and visitors profiles [5].It will be used to improve the information of problem occurred to the user and security of the websites, which comes under the site reconstruction application and some of the other application of web log mining are web pre-fetching [6] [7] and link prediction. The site reconstructions application is done by extracting user profile from the web log file which is located in web server. The log files contains information such as client ip address, user name, request type, bytes transferred, visiting path, path traversed and user agent.

The basic phase of the web log mining is data preprocessing, pattern discover and pattern analysis [8][9]. The data preprocessing phases involves the data cleaning [10] and filtering process. The noisy data, inconsistent data will be removed in the cleaning process and the filtering techniques normalizes the cleaned data. The pattern discovery can be performed by various clustering techniques [11] [12]. The major pattern discovery is based on the web pages [13] [14] and frequent pattern [15]. Pattern analysis is performed by comparing the pattern discovery result on the web log mining [10] [16].

The phases of this work is data cleaning which is an initial process of web log mining, the second phase is classifying the navigation pattern and the final phase is online navigation pattern prediction. Classification is a major problem in data mining process, several classification techniques like decision tree, support vector machine, fuzzy logic [17] is introduced by the various researchers. In the proposed work Extreme machine learning [18] [19] technique is applied for classification, the weight vector is calculated and based on the weight the performance establishment is checked. A Similarity and Kernel based PCM (SKPCM) [20] is used for clustering the most relevant user information in order to obtain the navigation pattern. The SKPCM is derived from the SPCM by using the Radial basis function kernel, to cluster the navigational pattern without requesting user to provide the cluster number.

RELATED WORKS

The pattern discovery is the major issue in web log mining it can be performed based on varies factor on the log files. According to the researches most patterns is discovered on traversal pattern, sessions, navigation pattern, web pages and frequent pattern. Most approaches are used to mine these patterns effectively.

Rao, V.V.R.M et al (2010) proposed a novel techniques to extract the navigation pattern. A hybrid predictive model is used by the author to retrieve the web user navigation pattern. Combination of markov model and Bayesian theorem is used with hybrid predictive model which will finds the pattern effectively. Filtering the possible categories, the operation scope is reduced by the markov model and the accuracy of the web page is increased by the Bayesian technique. The author concludes that this work is more efficient with respect to the accuracy.

Sisodia, M.S, et al (2009) proposed a frequent traversal pattern mining algorithm to discover the web user traversal pattern with a weight constraint. In the sequential traversal pattern the weight constrains is added to maintain the downward closure property which is used to assign the minimum and maximum weight. During the scanning process of the session database the proposed method fetch the traversal pattern with the minimum weight by applying the downward closure property.

Singh, A.K et al (2014) compares the Apriori and FP-growth algorithm to discover the frequent item set on web log file. From the various kinds of data such as web server data, application level data and application server data the author choose the web server data which includes the log files for this work. Ip address, access time of the user and page references are taken as the main input for this comparison process from the log files. The author compares the properties such as memory size, prefetching and scalability of these inputs to check the efficiency of this work.

PROPOSED METHODOLOGY

The data mining is applied in web log mining to discover the pattern of browsing nature of the web site users. The navigation pattern is extracted to trace the browsing pattern, and the structure of the web site is improved accordingly. While dealing with the browsing nature of the web sites the important factor that should be concentrated are, duration of using the web page or frequent access of the web page. These data can be retrieved from log file which contains session information about the web pages.

The steps in web log mining to extract the navigation pattern is, preprocessing, classification and clustering. Preprocessing is also known as data preparation that cleans the data by removing the unnecessary information (log entries) which is not needed for mining process and provides the required input i.e. user session files to the web log mining process and also it improves the quality of the mining result.

Classification is an important factor in data mining that creates a set of class label for an unclassified data. The input data set is considered as the training data set that consists of multiple numbers of records with several attributes. Each and every record is identified by class label. Then the classifier creates a model based on attributes or feature for each class. The proposed work used the Extreme Machine Learning classifier. The Clustering group the set of meaningful classes by partitioning the set of data or object. The proposed work groups the web pages that are visited by most number of users. The Similarity and Kernel based PCM is used in this work to cluster the web pages. The proposed work flow is showed in figure 1.

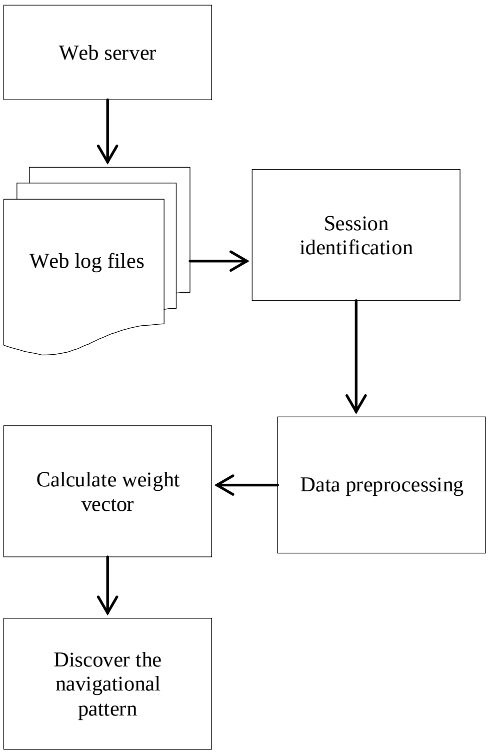

Figure 1 Flow diagram of proposed work

A. *Preprocessing*

The preprocessing is the initial step of the proposed work which contains data cleaning and filtering process. Preprocessing removes the unnecessary information and provides the required input for web log mining process and also it improves the quality of the mining result.

Data Cleaning

In data cleaning the inconsistent data, noisy data, missing values and the outliers is identified and a cleaning work is proceeded to remove all the unwanted information in the data. The input to the web log mining is a user log files, this files contains user name, web pages history, data entries etc. Among these, the data entry and null values in column field is not needed for the proposed work, because this work focus on mining the Web pages, Age, Gender, Product and Region of the web users from the dataset so these kind of information is removed during the preprocessing step. Figure 2 shows the preprocessor steps in detail. After the data cleaning, the next process is page identification, session identification and the user identification

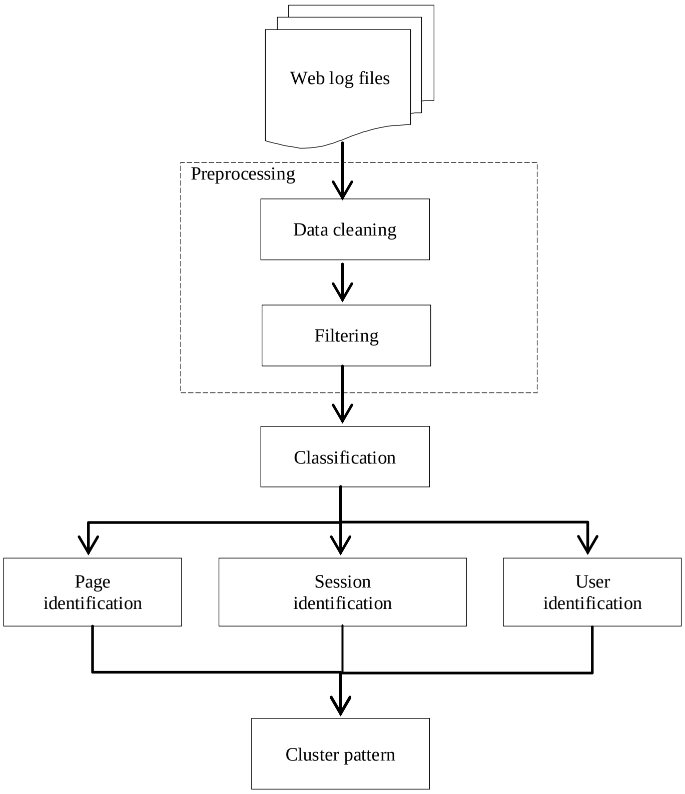

Figure 2.Preprocessing of web log files.

Filtering

Filtering is the important factor which improves the quality of the mined pattern, it is the final step of preprocessing; the result of the data cleaning is filtered by reducing the attributes. This work focus on limited number of attributes so the unnecessary attributes such as user agent, visiting path, success rate and Request type is filtered.

B. Classification

The input which is the result of the preprocessing is considered as the training data set that consists of multiple numbers of records with several attributes. The proposed work used the Extreme Machine Learning classifier to create a set of class label for an unclassified data. Figure 2 explains the working procedure of the web log mining.

The learning algorithm, Extreme learning machine (ELM) for the single hidden layer feed forward neural networks is faster than the other traditional algorithm. During the preprocessing some of the unwanted data is removed already, now the ELM classify the exact requirement such as page, session and User identification from the unclassified data. For Q training data set (p_i, qi$_{)}$ where p∈ R^{d1} and q∈ R^{d2}, the SLFN with K hidden unit, the output weight is obtained by the following algorithm.

Algorithm for ELM

```
s(x): activation function
$W_i$ : input weight
$\tau_i$ : Biases
$\rho_i$ : Output weight
Q= $(p_i, q_i)$- input set
Begin
Find s(x) for Q hidden unit
$\sum_{i=1}^{Q} \rho_i s_i(x_j) = \sum_{i=1}^{Q} \rho_i s(W_i . x_j + b_i) = o_j$
Assign input weight randomly $W_i$ and bias $\tau_i$
Assign K= $s(W_i . x_j + b_i)$
Compute K$\rho$=T
Calculate $\rho = K * T$
End
```

Page identification is the process of finding the number of mostly visited pages by the web users, with this page details navigational pattern is improved. The user id, time stamp, user name can be retrieved in the user identification step. Based on this information only the user session will be identified. The web user whenever enter into the web site and the time spend on that particular web site is consider as a session. Identifying this user session is such a difficult process, so all the page request is divided into sub sequence.

Proposed work focus on improving the navigational prediction pattern by extracting the required feature samples. The sample of this feature is given for training. The ELM classifier is trained based on it and test the results for remaining pages. This makes the system to classify the pages and order them easily which reduces the time.

C. *Similarity and Kernel Based PCM*

The SKPCM is the clustering algorithm applied in the proposed work for clustering the navigational pattern. Proposed work uses Radial Basis Function (RBF) kernel instead of mountain method. The merit of this algorithm is, it will automatically generate the required cluster according to the similarity matrix without specify the cluster number. This special feature of SKPCM is obtained by RBF kernel method which searches the center of each cluster. For the given data points$\{x_1, x_2 \dots \dots x_n\}$,$x_{tj}$ is consider as j^{th} coordinate of t^{th} point.The maximum value of the Radial basis kernel function is consider as the initial center for these data and the next center is found by modifying RBF kernel , this will repeated until it reaches the termination condition. The RBF kernel on two point P and $p'$is given below

$$KF^1(p,p') = exp\left(-\frac{\|p-p'\|^2}{2\sigma^2}\right) \tag{1}$$

Where $\|p-p'\|^2$ is Square Euclidean distance between data point p and grid nodep', σ is a free parameter. The termination condition for the RBF kernel is

$$\frac{KF(p,p')}{Max(K(p,p'))} < \varphi \tag{2}$$

Where $Max(K(p,p'))$denotes the maximum value of RBF kernel function, this will be taken as the initial center andφ is given parameter. When this condition is reached the RBF gets terminated. In order to find the similarity between the each cluster the RBF kernel is modified. Let KF be the total similarity of xt data points then the correlation comparison procedure is written as

$$KF(xt)gm = \sum_{j=1}^{n} f(xt)^{gm} \tag{3}$$

In the above equation g is taken as 5 and m=1, 2, 3. The idea and algorithm for the proposed work is presented in this section.

Based on the given data set the similarity matrix SMhas to generate before executing the SPCM algorithm. In data set the similarity between each pair of objects will be stored in the similarity matrix SM. According to the similarity matrix the SKPCM performs the cluster process without requesting the number of cluster from the user. The aim of the proposed work is to collect the web user details of the particular organization web site and analyze the mostly visited pages, time spend on that pages, products, which age group people visits the web site frequently and also gender wise and region wise user details is gathered. By having these details navigational pattern is provided for the new users of those organizational web sites to improve quality and creates a best impression for the new users.

Algorithm: SKPCM

```
$m$:Fuzzifier
Q: cut point
C: cluster
$\rho$: Minimal enhancement
$\sigma$: Minimal decreasing ratio
C_COUNT: cluster counter
l: Iteration count
$b^{(0)}$: Possibilistic of c partition
$D_i$: Distance of membership degree
Begin
Set C-COUNT =1
$KF^{C-COUNT\ 1^*} = MAX(KF^1(p,p'))$
Set l=1
Initialize $b_c^{(0)}$
Estimate $D_i$ value
$K(p,p') = exp\left(-\frac{\|p-p'\|^2}{2\sigma^2}\right)$
Choose initial center as $KF^*$
Select next center by SQD
SQD=$exp(\gamma\|p-p'\|^2)$
Else
Terminate RBF function
Set l=i+1.
If $\|b_c^{(i-1)} - b_c^{(t)}\| > \rho$
```

```
Update Squared Euclidean distance
Set C-COUNT = C-COUNT +1
do
Update RBF kernel value
Set
$KF^{C-COUNT*} = MAX(value\ of\ KF^{C-COUNT})$
While($\frac{K(p,p')}{Max(K(p,p'))} < \varphi$ )
```

EXPERIMENTAL RESULT

The web log files are collected from the E-commerce, then it is preprocessed to remove the noisy and inconsistent data, then it is classified and ranked, This makes the system to classify the pages and order them easily which reduces the time. This section describes the input data, taken for the proposed work and the statistical analysis of the result compared with some existing work.

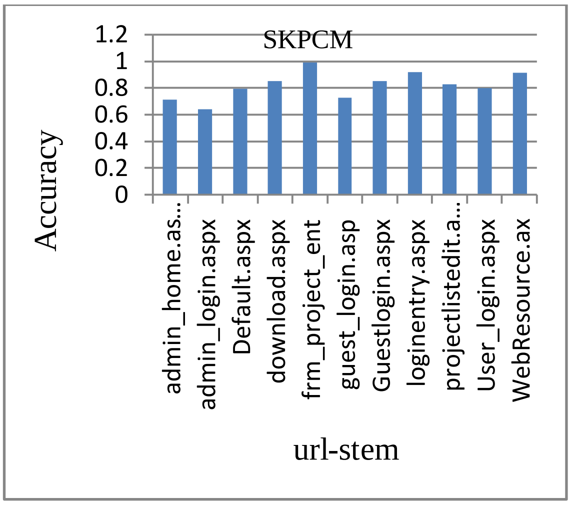

Figure 3 Accuracy report for extracted pages

Accuracy of web pages extracted from log files is shown in the figure 3. The web page frm-project-ent obtainthe higher accuracy compare to all other web pages. The accuracy of frm-project-ent is 99%.

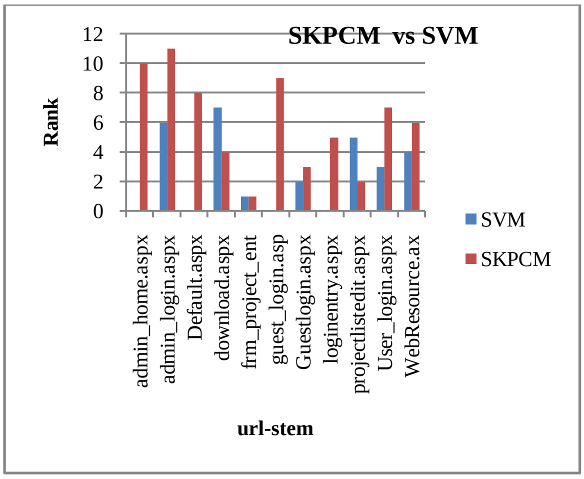

Figure 4. Comparative Page Rank for the Pages

The efficiency of the SKPCM is given in the Figure 4. Among all the page Frm-project-ent obtain the highest rank compared to other pages and the admin_login page has the lowest rank. From this it is clear that Frm-project-ent is visited mostly.

Table 1. Comparison of Web Page Rank

Pages	Existing technique page rank	proposed technique page rank
admin_home.aspx	0	10
admin_login.aspx	6	11
Default.aspx	0	8
download.aspx	7	4
frm_project_ent	1	1
guest_login.asp	0	9
Guestlogin.aspx	2	3
loginentry.aspx	0	5
WebResource.ax	5	2
projectlistedit.aspx	3	7
User_login.aspx	4	6

Table 1 describes the comparison of page ranks with the Existing SVM algorithm. Several variation is there while comparing both algorithm. one similarity is frm_project_ent obtain the highest rank. The existing work shows admin, default, guest and login entry page got 0 ranks. But the proposed SKPCM perform the efficient cluster process and shows those are viewed rarely and they got 10th, 8th, 9th and 5th rank respectively.

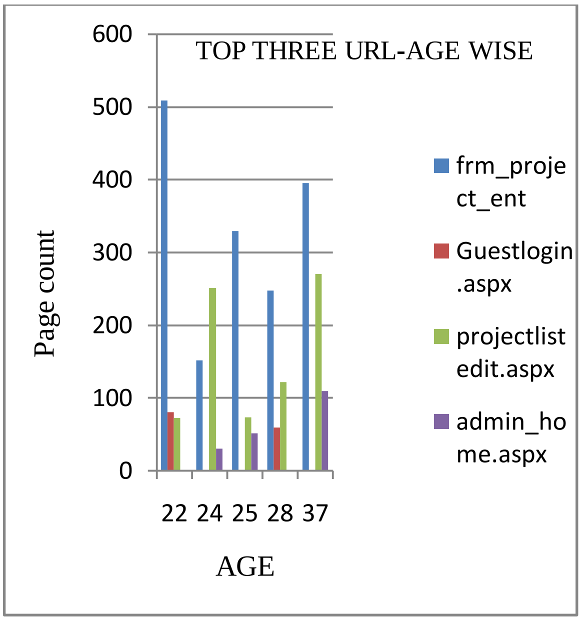

Figure 5. Age wise Page count measurement

The proposed work suggests the navigation pattern to the web user according to the result shown in figure 5. The frm_project_ent has been viewed by most of the user. The Admin_home.aspx page has been viewed by least count of the user.From this it is clear thatfrm_project_ent page will be suggested to the user with any age.

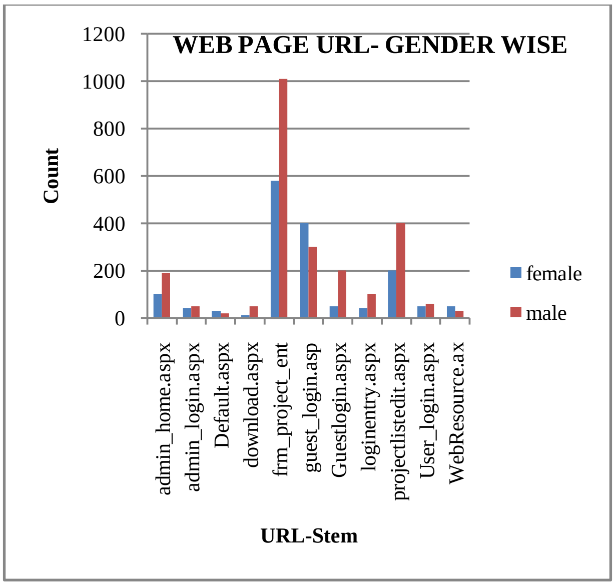

Figure 6. Gender wise Page Count Measurement

In order to provide a better navigational pattern, gender wise data is clustered in the proposed work. Figure 6. Explains the web pages count measurement. Compared to male, female web users viewed the frm_project_ent more.

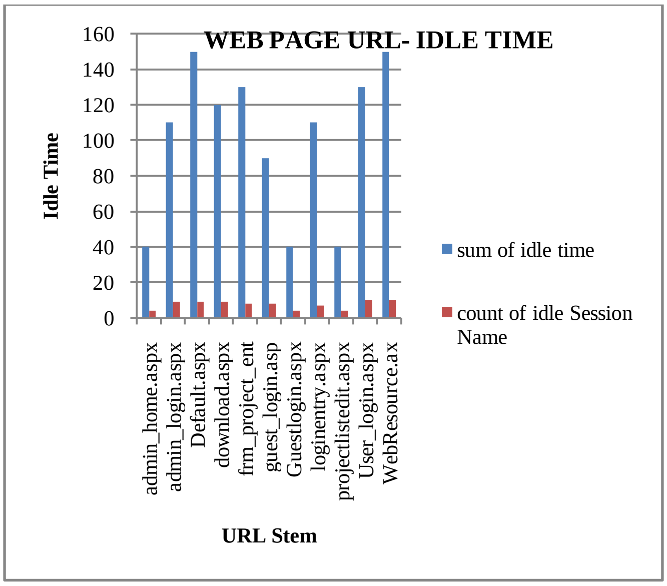

Figure 7. Idle Time Report

Idle time report is generated from the Web Pages during their idleness. This idle time is calculated to identify which Web Pages are not accessed properly by the Web users. By using this information these pages are eliminated from the navigational pattern result to improve thenavigation pattern efficiency. The idle time report is shown in Figure 7.

CONCLUSION

Web log mining is the process of analyzing, retrieving the user access pattern from the web log files. The identification of the service provided by a particular web site can be identified easily with the web log mining process. Nowadays the E-Commerce businesses growth mainly focused in web log mining to discover the user behavior, buyer activities and visitors profiles, which will be used to improve the information of problem occurred to the user and security of their websites. This work focus on improving the navigational prediction pattern by extracting the required feature (web page, session name, age, idle time and gender) samples. The sample of this feature is given for training. The ELM classifier is trained based on it and test the results for remaining pages. This makes the system to classify the pages and order them easily which reduces the time.Then SPCM is applied to rank the pages which are used to improve the navigational pattern in an efficient manner. In future this work will be improved by incorporating our work with an adaptive web site platform to provide the link suggestion for the users

REFERENCES

1. O. Nasraoui, R. Krishnapuram, and A. Joshi, “Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator,” Proc. Eighth Int’l World Wide Web Conf. (WWW ’99), pp. 40-41, 1999.
2. O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi, “Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering,” Int’l J. Artificial Intelligence Tools, vol. 9, no. 4, pp. 509-526, 2000.
3. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” SIGKDD Explorations, vol. 1, no. 2, pp. 1-12, Jan. 2000.
4. OlfaNasraoui,MahaSolimanEsinSaka, Antonio Badia,Memberand 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.
5. F.M. Facca, P.L. Lanzi “Mining interesting knowledge from Weblogs: a survey”, Data and Knowledge Engineering Vol. 53, No. 3, June 2005,pp: 225-241.
6. DiamantoOikonomopoulou, Maria Rigou,SpirosSirmakessis,AthanasiosTsakalidis,, “Full-Coverage Web Prediction based on Web Usage Mining and Site Topology”. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI’04) 0-7695-2100-2/04, April 20, 2009.
7. Junjie Chen and Wei Liu, “Research for Web Usage Mining Model”, International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06) 0-7695-2731-0/06 © 2006 IEEE.
8. JikeGeYuhuiQiuZuqin Chen Shiqun Yin Faculty of Computer and Information Science. Southwest University, Chongqing, China gjkid@swu.edu.cn “Technology of Information Push Based on Weighted Association Rules Mining”.
9. F. Masseglia, P. Poncelet, and M. Teisseire, “Using data mining techniques on web access logs to dynamically improve hypertext structure”. In ACM SigWeb Letters, Pp: 13-19, 1999.
10. Sisodia, M.S, Pathak, M, Verma, B, Nigam, R.K, “Design and Implementation of an Algorithm for Finding Frequent Sequential Traversal Patterns from Web Logs Based on Weight Constraint”,2nd International Conference on Emerging Trends in Engineering and Technology (ICETET),Pp:317 – 322, 2009.
11. Ai-Bo Song ,Zuo-Peng Liang, Mao-Xian Zhao, Yi-Sheng Dong, “Mining Web Log data based on Key path”, Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002.
12. Mike Perkowitz, Oren Etzioni, “Adaptive Web Sites: Automatically Synthesizing Web Pages”, Department of Computer Science and Engineering, 98195, 1998

13. Singh, A.K, Kumar, A,Maurya, A.K, "An empirical analysis and comparison of apriori and FP- growth algorithm for frequent pattern mining" International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), Pp:1599 – 1602, 20 14

14. Noy, N. F, Sintek, M, Decker, S, Crubezy, M, Fergerson, R.W, &Musen, M.A. (2001),"Creating Semantic Web Contents with Protege-2000". IEEE Intelligent Systems 16(2), 60-71.

15. Kudelka, Milos, Lehecka,Ondrej, Snasel, Vaclav,El-QawasmehE, "Web pages clustering based on web patterns," 2nd IEEE International Conference on Digital Information Management, Vol:2, Pp:657 - 664 , 2007.

16. H. Jantan, A.R. Hamdan, Z.A. Othman, Classification and Prediction of Academic Talent Using Data Mining Techniques, International Journal of Technology Diffusion,vol 1, pp. 29-41, 2010.

17. Guang-Bin Huang, Qin-Yu Zhu,Chee-KheongSiew, "Extreme learning machine: Theory and applications",Neurocomputing, Vol. 70, pp. 489-501, 2006.

18. Shifei Ding, Hang Zhao,Yanan Zhang, "Extreme Learning Machine: algorithm, theory and application", Artificial intelligence review, published online, 2013.

19. Vincent S. Tseng, and Ching-Pin Kao," A Novel Similarity-Based Fuzzy Clustering

20. Baek Hwan Cho; Hwanjo Yu; Jongshill Lee; Young JoonChee; In Young Kim; Kim, S.I."Nonlinear Support Vector Machine Visualization for Risk Factor Analysis UsingNomograms and Localized Radial Basis Function Kernels" Information Technology in Biomedicine, IEEE Transactions on Vol: 12, No: 2,Pp: 247 – 256, 2008.

21. Rao, V.V.R.M,Kumari, V.V, "An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User Using Web Usage Mining", IEEE International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom),Pp: 225-230, 2010.

22. Sisodia, M.S,Pathak, M,Verma, B, Nigam, R.K, "Design and Implementation of an Algorithm for Finding Frequent Sequential Traversal Patterns from Web Logs Based on Weight Constraint", IEEE 2nd International Conference on Emerging Trends in Engineering and Technology (ICETET),Pp: 317-322, 2009.

23. Singh, A.K, Kumar, A,Maurya, A.K, "An empirical analysis and comparison of apriori and FP- growth algorithm for frequent pattern mining" IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), Pp: 1599-1602. 2014.