Visualizing the Navigational Patterns from Web Log Files using Web Mining Applications

R. Gobinath, M. Hemalatha*

Abstract—Web Usage Mining is the application of Data Mining process used for understanding the behavioural patterns of the web user. The web usage patterns are stored in web server and inthe form of web access log files. These web access log files carries full blue print of the usage data of the web pages. The web usage mining researchers uses this web log files to analysis the behaviour of the web user. The extracted patterns from the web log files are clustered for analyzing using proposed PSO-K-Harmonic technique. The clustered navigational patterns are analyzed with proposed negative association rules and pruned with negative selection algorithm.

Index Terms—Data Mining, Negative Association Rules, Negative Selection Algorithm, Web Mining, Web Usage Mining,.

————————————◆————————————

1 INTRODUCTION

THE term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest. Web data mining is not only focused to gain business information, but also used by various organizational departments to make the right predictions and decisions for things like business development, work flow, production processes and more by going through the business models derived from the data mining. In general Web mining is the application of data mining techniques to discover patterns from the web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. The path used by the website visitor carries navigational history along with the patterns which is much useful for website designers to analyze and redirect the website [1]. Web usage mining is the process of extracting useful information from server logs, e.g. use web usage mining is the process of finding out what users are looking for on the internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be further classified depending on the kind of usage data considered:

- Web Server Data: The user logs are collected by the web server. Typical data includes IP address, page reference and access time.
- Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

- Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

2 RELATED WORK

Detailed approaches followed for the extraction of necessary information from the web servers by R. Cooley, B. Mobasher and J. Srivastava (1997) [2] clearly shows the extraction of necessary web log files from the web server. B. Mobasher, N. Jain, E. Han, and J. Srivastava (1996) [3] , shows the discovery of necessary patterns from the log files by collecting sample web log files. They introduce a framework which includes the applications of data mining and techniques of knowledge discovery process. Many researchers presented the preprocessing phase for the web personalization in several steps. The steps followed by various researchers are not common in many ways. For example steps for data preprocessing followed by the researchers Yan LI, BoqinFeng and Qinjiao Mao (2008) [4] is different from the step followed by the researchers DoruTanasa and Brigitte Trousse (2004) [5]. The strategy followed by each and ever researchers may differ from each others.

The strategies focus in this paper is on the method of getting the session details which contains the details about the (1) Fields divided from the web log files; (2) Cleaning necessary data from the irrelevant data; (3) The User from the web log files is identified; (4) The Session of the user is identified; (5) The Feature extraction process is carried out from the collected data. The pattern analysis process has to be carried out after the feature extraction process. K. R. Suneetha, and R. Krishnamoorthi, (2009) [6] analysis the NASA websites log files from the world wide web server. The errors and information about websites are stored and it can predict the potential visitors that enters the websites. The proposed Personalized Electronic Catalogue (PEC) by P.C. Benjamin Yena, C.W. Robin Kong (2002) [7] shows how effectively we can build a website by customizing the information content. The information

helps the web designers in developing the most efficient website for the consumer, organization, and display for electronic catalogs. B. Mobasher, N. Jain, E. Han, and J. Srivastava (1996) [8] has introduced a different technique for pattern discovery from the web log files using some of data mining applications. The Yu-Hui Tao, Tzung-Pei Hong, Wen-Yang Lin and Wen-Yuan Chiu, (2009) [9] processed entire data set with the international browsing data by delivering unified intention based web transaction mining algorithm in e-commerce websites and increases the customers. The algorithm utilities used is different from traditional web usage mining algorithm. The preprocessing of log files stages including cleaning of data, identifying the user, session identification and path competition are represented by ArshiShamsi, Rahul Nayak, PankjPratap Singh and Mahesh Kumar Tiwar (2012) [10]. The researchers K Sudheer Reddy, G. ParthaSaradhiVarma and I. Ramesh Babu, (2012) [11] worked out an algorithm for joining log files after extracting from the web server. T.T. Aye, (2011) [12] removes irrelevant data and noisy data in the preprocessing stage by following some techniques like field extraction, where the single lined log files are separated into fields for better pattern discovery. The algorithms used in the cleaning process in useful in removing the unwanted information in the analyzing of data. The access stream process is started by the Sanjay Bapu Dakar and Z. Gawali (2010) [13] even before the actual mining process begins and improves the complete processing in web usage mining in an effective way. This process undergoes different preprocessing stages by collecting the log files from the different sources and finding useful patterns and access stream users' relationships plays a role in this data preprocessing. The categorization process improvement given by Aburrous, Maher, Hossain, Alamgir, Dahal, Keshav and Thabtah, Fadi, (2010) [14], where the gathered rules are sent into the fuzzy rule engine for increasing the classification technique in data mining.Sandeep Singh Rawat and Lakshmi Rajamani, (2010) [15], introduced Custom-Apriori algorithm instead of using the traditional Apriori algorithm for analyzing patterns. S. Taherizadeh and N. Moghadam, (2009) [16], combines web content mining with web usage mining for finding patterns and user behavior prediction for the e-commerce websites. The Apriori algorithm is applied in the dataset for predicting the association rule and the user gathers information according to the behavior patterns of previous user's.

3 METHODOLOGY AND IMPLEMENTATION

3.1 Collect Web Server Log

Web access log files are single lined data present in web server, holding detailed information about the website user. The log files are very important in web usage mining for identifying the user navigational patterns for final web personalization process. Basically a web log file has nine attributes, which stores information about the usage data. The web log files which record the user who accesses the websites. The identifying information about particular user and particular session are very difficult job for researchers to analysis the web log files.

- IP address of the visitor
- User ID
- The session (date and time)
- The requested file
- Used protocol
- Status of the process (success/failure)
- Bytes send
- The site from the visitor came
- The used browser

The format of the web log files usually follows W3C extended, NCSA and IIS extentions. The W3C format allows the user to log for each request where as NCSA and IIS has fixed format of logged each request.
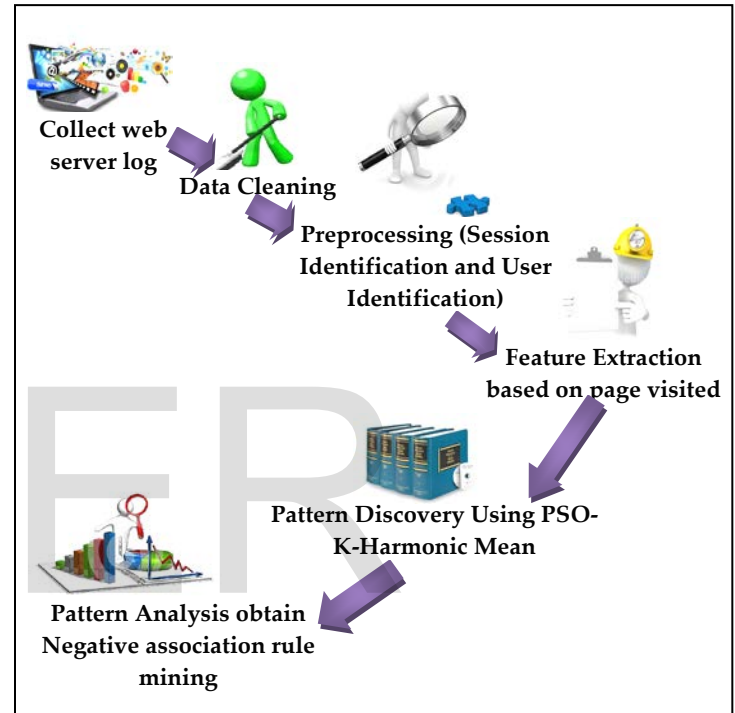

Fig 1. Discover Web log Files Using data Mining Techniques

3.2 Data Cleaning

The preprocessing of web log file are most important process followed in every web usage mining process. Unwanted and irrelevant data for the personalization process are removed or cleaned preprocessing stage. The irrelevant data such as multimedia files, page style files, successful status code, duplicate system generated entries, etc are removed in cleaning stage. The single lined web log files are separated into attributes for convenient data mining process. The attributes are maintained in Server Log Table (SLT). The used algorithmic steps for cleaning process are given below

    Read Entries from SLT
    For each Entry in SLT

    Read fields (Status code, method)
If Status code='200'and method= 'GET'
    Then
Get IP_address and URL_link
If suffix.URL_Link= {*.gif,*.jpg,*.css} Then

Remove suffix.URL_link
Save IP_sddress and URL_Link
End if
Else
Next Entry
        End if

The cleaned data are taken for identifyingnecessary user and session for web usage mining process. The necessary user is identifyied by taking IP addresss into consideration.The session are the time duration spent by the user in the website. The usage data fall into every 30 minutes are divided into different sessions.

3.2.1.   User identification

The individual user from the web log files has been extracted by identifying the IP address. The new entry is detected when the reading IP address from server log table is in process. The already present IP address but different operating system or browser is also considered to be a new entry. The user identified algorithmic representation is given below.

Distinct user identification algorithm

Read each entry in SLT
If an IP address not exist then
Consider the user as new user
End if
If IP address exists andthe ((browser version or Operating System) is not exist)
then
Consider the user entry as new user
Else if
Next entry

3.2.2 Session identification

The time spend by the user on the website is known to be a session. The session time is categorized by every 30 minutes working with the user. The referrer based method is used for identifying the new session in the web log files. Basically information is taken when IP address, browser version and operating system are same. The next session is calculated if the time duration exceeds more than 30 minutes interval.

Time Oriented Heuristic Algorithm for Session Identification
For each entry in SLT
Sort the log data by IP address , agent and time
Next entry
Read each entry in SLT
If the IP address and agent not exist then
If (requestedtime$_i$ -requestedtime$_{i-1}$)> Session Time Out or Session time outdoes not belong toSession history
then
Increment the value of session by 1
Consider the user session as new
Endif
    Endif
    Next entry

The user and session identified in such a manner are to be checked for quality with clustering models than taken for feature extraction process.

3.3 Feature Extraction

Feature extraction is the final stage of every web usage mining process, which is optional in much web usage mining process. The necessary features are identified based on applying certain rules by considering time factor, most frequently visited pages and most frequently navigated paths.

3.4 Pattern Discovery

Pattern discovery plays most important part in the web personalization process. The website designers and owners are very much interested in knowing the behaviroul patterns of the website users. The correct identification of the navigational patterns will show the behavior of the web page user. Many web usage mining researchers pay more attension in pattern discovery process, because clustering of similar navigational paths are succeded in this stage. Data mining algorithms used in this process helps in clustering the similar navigational paths.Proposed PSO-K-Harmonic mean based clustering techniqueis used for clustering similar navigational patterns. The Particle Swam Optimization (PSO) algorithm is used for clustering the navigational paths. The K- Harmonic mean based clustering concept is used for finding the shortest distance from one centroid to another centroid. The K-Harmonic mean (KHM) objective function uses the harmonic mean of the distance from each point to all data centers. The distance between the navigational paths are calculated by considering number of pages visited by the visitor.

eg:
Path 1   :Frontpage->Frontpage
Path 2   :Frontpage->misc->misc->Frontpage

The distance of the given navigational path 1 is 1 and distance of the given navigational path of the path 2 is 2 (consecutive pages are not taken into account).

$$KHM(X,C) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}} \qquad (1)$$

Here p is an input parameter, and typically $p \geq 2$. The harmonic mean gives a good (low) score for each data point when that data point is close to any one center. This is a property of the harmonic mean; it is similar to the minimum function used by KM, but it is a smooth differentiable function. The velocity and position of the particles are modified using the following equations

$$V(t+1) = w.v(t) + c_1 r_1 \big(pbest(t) - X(t)\big) + c_2 r_2 \big(gbest(t) - X(t)\big)$$
$$X(t+1) = X(t) + V(t+1) \qquad (2)$$

Where, V and X are velocity and positions of particles. W is for inertial weightc_1,c_2 are positive constant called acceleration coefficient which controls influence of pbest and gbest on the search processP represent no of particle in the swarmr1 and r2 are random value ranges from [0 ,1].The sample pseudo code for proposed PSO-K-Harmonic mean is given below along with the k-harmonic mean based clustering pseudo code.

Pseudo code for K-harmonic mean

Data Structures:
N: n by d+1 array - contains static information about data set

Nmin: n element array which holds the minimum distance to any center

K: k by d array that holds information about centers

M: n by k array that holds distance from all point in N to all points in K

Temporary Arrays (Could be reduced but shown for simplicity)
U: n element array
Q: n by k temporary array
QQ:k element array
R: n by k temporary array
RR:k element array
T: n by k
p: power associated with the distance calculation

Initialization
1. Create an initial K:
      Choose any k points from N
Main Loop
2. Fill Matrix M:
Calculate distances from all points in N to all centers in K
3. Compute Nmin:
Find minimum distance for to any center for each point in N
4. Recompute Harmonic Averages and Update K:

```
For each point (j = 0 to n)
     For each center (i = 0 to k)
             U[j] = U[j] + (Nmin[j]/N[j,I])
       U[j] = U[j] – 1;

For each center (i = 0 to k)
  For each point (j = 0 to n)
      Q[j,I] = [(Nmin[i]^(p-2) * (Nmin[i]/N[j,i]))^(p+2)] /
                     [(1 + U[j]^p)^2]
For each center (i = 0 to k)
  For each point (j = 0 to n)
      QQ[i] += Q[j,I]
For each center (i = 0 to k)
  For each point (j = 0 to n)
      R[j,I] = Q[j,i] / QQ[i]

For each center (i = 0 to k)
  For each point (j = 0 to n)
K[i] = K[i] + R[j,i]*N[j]
```
5. If no center is updated in step 4 then stop
      The algorithm has converged
Pseudo code for PSO-K-Harmonic
For each particle
  Initialize particle using K-harmonic mean
END
Do
  For each particle

Calculate fitness value
If the fitness value is better than the best fitness value (pBest) in history
      set current value as the new pBest
End
Choose the particle with the best fitness value of all the particles as the gBest
For each particle
   Calculate particle velocity according equation (2)
   Update particle position according equation (1)
End

The clustered navigational patterns are taken for analyzing and visualization process. The error rate is reduced after each and every iterations.

3.5 Pattern analysis

The association between each and every discovered patterns explains the behavior of website visitor. The extraction of interesting rules by removing irrelevant rules are performed in pattern analyzing process and a clear statergy for improveing the websites are given [17].The extracted patterns are pruned with negative selection algorithm for improving the quality. The negative association rule mining technique for rule mining is differentfrom traditional positive association rule mining technique [20]. The correlation coefficient measurement for extraction negative association rules and pruning strategies are designed. The extraction of negative association rules can be possible with the rules extracted using associated technologies. The negative itemset mining is explained in following algorithmic steps

Negative Association Rules
Input: TDB-Transactional Database
MS-Minimum Support
MC-Minimum Confidence
Output: Negative Association Rules
Method:
N←Φ
Find F1← Set of frequent 1- itemsets
for (k=2;Fk-1!=Φ; k++)
{
Ck= Fk-1 join Fk-1
// Prune using Apriori Property
for each i ε Ck, any subset of i is not in
Fk-1 then Ck = CK - { i }
        for each i ε Ck find Support(i)
        {
        for each A,B (A U B= i)
   {
   QA,B= Association(A,B);
   if Q<0
   {
if(supp(A←¬ B)≥MS&&conf(A←¬ B)≥MC)
then
N←N U { A←¬ B}
if(supp(¬ A   B)≥MS&&conf(¬ A   B)≥MC)
then
N←N U { ¬ A← B}
}
}

AR← P U N

The collected rules are again used for pruning process with the help of adapted technique known as negative selection algorithm. The acquired immune system of mammalian is a major inspiration for developing the negative selection algorithm. The self discrimination character plays a key role in the negative selection algorithm.The immunity acquired by clonal selection theory, which means adaptive behaviour of the immune system including ongoing selection and proliferation of cells that select-for potentially harmful (and typically foreign) material in the body [18]. The behaviour of user can be clearly understand with the final pattern analysing stage. The results given in the section 4 clearly explain the behavioural patterns of the user based on the historic web log data.

4 EXPERIMENTAL RESULTS

The web access log files collected from the online repositories were used for the pruning and testing purpose of the proposed algorithms. The web log files namely Kdlog (data set 1), online shopping (data set 2) and msnbc.com (data set 3) are used in various stages of iterations to test the algorithm.TheKdlog data set was analyzed, which is a Web server of Apache Combined Log entry format and pre-processing and cleaning process was performed and clusterd. The result shows that the performance of Make density based algorithm works fine than the remaining algorithms in the proposed Distinct User Identification algorithm(DUI)using server log table and Time Oriented Heuristic Algorithm(TOH) for Session Identification model [19].

The kdlog file sizeis 39974 Instances after applying the proposed model for cleaning process the percentage found was 50.2.The numberof usersidentifiedwas2307,the number of sessions identified was1787 with the accuracy 92.7.When the same data set is used in traditional based web topology models, the record obtained after cleaning process are 14678, the percentage of reduction is 36.7, number of users identified are 1508, number of sessions identified are 678 and the accuracy is 86.8 which is much lesser compared to the proposed model. The feature extraction process proceeded with testing web log files namely online shopping web access log files and kdlog files. The categories followed in features are explained in previous section and results are shown. The top 10 frequently used entries, top 10 frequently used visitor, top 9 visitors entry pages, top 10 most frequently time spend pages, top 10 frequently used path for online shopping website and top 10 frequently used navigational paths for Kdlog files are found using feature extraction process. The PSO-K-Harmonic mean based clustering process is used for pattern discovery process and useful patterns are collected.

The generating rules for analyzing stage is the most difficult part of every web usage mining process. Table 1 shows the positive association rules obtained from the clusters of naviga

Table 1. POSITIVE ASSOCIATION RULE

tional patterns collected with applying PSO-K-Harmonic mean based clustering algorithm. The support and confidence for each navigational patterns are also indicated in this table.

The process of generating rules are implemented in clusters

formed due to PSO-K-Harmonic mean based algorithm. A

Positive Association Rules	Support	Confidence
Opinion and misc and travel -> on air	90.26	92.37
Living and sports and business and bbs ->front page	90.00	91.72
News and tech and living and business and sports ->front page	89.00	90.81
Front page and tech and opinion and living and news	87.60	88.59
Front page and tech and living and business ->sports	87.87	88.01
News and living and business and sports ->front page	87.18	88.14
Misc and living and travel ->on air	86.55	87.12
News and tech and misc and bbs ->front page	85.99	84.52
Misc and business and travel ->on air	85.65	85.39
Misc and business and bbs ->front page	85.57	85.41
Tech and living and sports and bbs ->news	85.49	86.09
Local and misc and business and sports ->frontpage	85.32	87.35
News and on air and business ->sports	85.01	85.69
Front page and opinion and living and sports ->news	87.81	87.94
News and misc and business and sports ->front page	86.22	86.84

new method of collecting negative association rules instead of positive association rules are obtained with each navigational paths support and confidence level, which is shown in Table 2.Table 2. NEGATIVE ASSOCIATION RULE

Negative Association rule	Support	Confidence
not Opinion and misc and Not business -> on air	89.21	91.53
Living and not sports and not travel ->front page	91.27	91.45
not News and not tech and living and not business and sports -> front page	93.56	93.94
Not Front page and not tech and not opinion and living -> news	94.28	94.63
Front page and not tech and not living and business ->sports	91.20	94.51
Not News and not living and not business and sports ->front page	89.49	89.53
Not Misc and not living and not travel ->on air	88.53	88.09
News and tech and misc and not bbs ->front page	87.28	87.05
Misc and not business and travel ->on air	86.14	86.13

Not Misc and business and not bbs ->front page	85.94	85.37
Not Tech and living and not sports and bbs ->news	86.31	86.25
Not Local and misc and not business and not sports ->frontpage	85.58	85.63
News and not onair and business ->sports	85.32	85.29
Front page and not opinion and not living and not sports ->news	85.36	85.41
News and not misc and business and not sports ->front page	85.39	85.04

Fig2.Peroformance analysis of proposed work in ROC curve

Fig 2.Peroformance analysis of proposed work in ROC curve
The performance analysis of proposed work is shown in Figure 5. The result of ROC curve with performance of proposed work is displayed with the AUC 0.8684. The maximum specificity and efficiency is shown. The Max Sensitivity Cut-off point= 194.00, Max Specificity Cut-off point= 122.00, Cost effective Cut-off point= 152.00 and Max Efficiency Cut-off point= 161.00.

5 CONCLUSION

Websites are most popularly used medium for business and social medium. The best outcome of every websites depend on the number of visitors uses the website in convenient and usefulness. The study is very much helpful in providing the solution to various questions in website designers and owners. The historical data collected collected from web servers are very much used to understand the behavior of visitor. The PSO-K-Harminic mean based algorithm and Negative association rule based algorithm are very useful in finding meaningful patterns or navigational paths from huge rule set. The analyzed pattern are taken for pruning stage with negative selection algorithm. The study also gives a chance to the web usage mining researchers to use negative selection algorithm for pruning the obtained analyzed patterns.

REFERENCE

[1] J. Nielsen, " Designing Web usability: the practice of simplicity. Indianapolis " IN: New Riders Press, 2000

[2] R. Cooley, B. Mobasher and J. Srivastava, 1997 "Web Mining: Information and Pattern Discovery on the World Wide Web", 1n Proceedings of the 9th 1EEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997

[3] B. Mobasher, N. Jain, E. Han, and J. Srivastava, Web Mining: Pattern Discovery from World Wide Web Transactions. Technical Report TR-96-050, Department of Computer Science, University of Minnesota, 1996.

[4] Yan LI, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique In Web Usage Mining", IEEE International Symposium On Computer Science and Computational Technology, pp. 554-559, 2008.

[5] DoruTanasa and Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining ", Published by the IEEE Computer Society, pp. 59-65, March/April 2004.

[6] K. R. Suneetha and Dr. R. Krishnamoorthi, 2009. "Identifying User Behavior by Analyzing Web Server Access Log File", International Journal of Computer Science and Network Security, Vol (9) 4, pp. 372-332.

[7] P.C. Benjamin Yena, C.W. Robin Kong (2002), " Personalization of information access for electronic catalogs on the web ", Electronic Commerce Research and Applications, vol (1), pp. 20- 40.

[8] B. Mobasher, N. Jain, E. Han, and J. Srivastava, Web Mining: Pattern Discovery from World Wide Web Transactions. Technical Report TR-96-050, Department of Computer Science, University 0f Minnesota, 1996.

[9] Yu-Hui Tao, Tzung-Pei Hong, Wen-Yang Lin and Wen-Yuan Chiu, 2009 "A Practical extension of web usage mining with intentional browsing data toward usage ", Elsevier Ltd, Expert Systems with Applications, 36(3937-3945).

[10] ArshiShamsi, R.N., PankjPratap Singh and Mahesh Kumar Tiwar,"Web Usage Mining by Data Preprocessing," International Journal of Computer Science And Technology, Vol.3, No. 1, 2012.

[11] K Sudheer Reddy, G. ParthaSaradhiVarma and I. Ramesh Babu, 2012. "Preprocessing the Web Server Logs – An illustrative approach for effective usage mining", ACM SIGSOFT Software Engineering Notes, 37(3), PP: 1-5.

[12] Aye.T.T, "Web Log Cleaning for Mining of Web Usage Patterns," International Conference on Computer Research and Development, IEEE, 2011. Vol(2), pp. 490 - 494

[13] Sanjay BapuThakare and Z. Gawali., "An Effective and Complete Preprocessing for Web Usage Mining," International Journal on Computer Science and Engineering, Vol. 2, No. 3, p. 848 - 851, 2010.

[14] Aburrous, Maher, Hossain, Alamgir, Dahal, Keshav and Thabtah, Fadi, 2010," intelligent phishing detection system for e-banking using fuzzy data mining".Journal of Expert Systems with Applications, 37 (12). pp. 7913-7921.

[15] Sandeep Singh Rawat and Lakshmi Rajamani, 2010,"Discovering Potential User Browsing Behaviors Using Custom-Built Apriori Algorithm", International journal of computer science & information Technology, 2 (4), pp : 28-37.

[16] S. Taherizadeh and N. Moghadam, 2009. " Integrating Web Content Mining into Web Usage Mining for Finding Patterns and Predicting
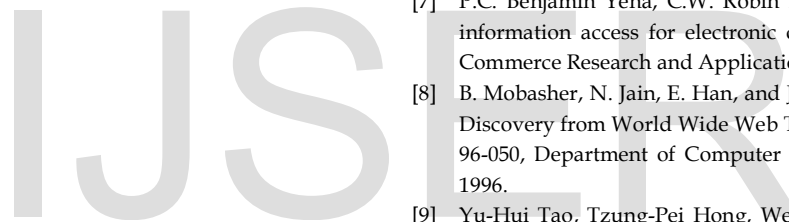
Users' Behaviors", International Journal of Information Science and Management,7(1).PP: 51-66.

[17] Antonie, M., L and Zaıane., O, R, 2004. Mining Positive and Negative Association Rules: an Approach for Confined Rules, Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, pp 27–38.

[18] Jason Brownlee, 2012, Clever Algorithms: Nature-Inspired Programming Recipes, PP. 277-283.

[19] Gobinath, R and Hemalatha, M, (2012), Improved Preprecessing Techniques for Analyzing Patterns in Web Personalization Process, International Journal of Computer Applications, 58(3), pp. 13-20.

[20] Honglei., Z, Zhigang., X, 2009. A Novel Incremental Updating Algorithm for Maintaining Discovered Negative Association Rules, icrccs, International Conference on Research Challenges in Computer Science, pp.164-167.