International Journal of Scientific Research in Computer Science and Engineering **Research Paper**
Vol.4, Issue.4, pp.1-6, August (2016) **E-ISSN:**2320-7639

Internet Surfing Prediction System using Association Rule Mining based on FP-Growth

A. Singh[1*], N. Jain

[1]Department of Computer Science & Application, Sagar Institute of Research & Technology, Bhopal, India
[2]Department of Computer Science & Application, Sagar Institute of Research & Technology, Bhopal, India

Corresponding Author: arti0024@gmail.com

Available online at www.isroset.org

Received: Apr/26/2016, Revised: May/19/2016, Accepted: Jul/12/2016, Published: Aug/30/2016

Abstract- Web Usage mining is nascent field of research which deals with extraction of interesting knowledge from the web log files. This paper describes the recent works on the field of Web Usage Mining(WUM) it is for the benefit of research on the future request and personalization of web based information services. Web Usage Mining puts an effort to determine valuable information from the secondary data obtained from the communications of users with the web. WUM involves three steps, which includes Preprocessing, Pattern discovery and Pattern analysis.

Keywords: Web Usage Mining Frequent Pattern Tree Association Rule Web Logs.

1. Introduction.

In the information age and electronic age there is a great transition from manual business to the Electronic Business .Now a days we do business electronically which is faster accurate and authentic .The business can be termed as E-commerce or Mobile Commerce. Hence and so forth to do business of this nature, the navigational patterns of user should be understood. However, Web traversal patterns obtained by traditional Web usage mining approaches are ineffective for the content management of websites. They do not provide the big picture of the intentions of the visitors. The Web navigation patterns, termed throughout-surfing patterns as defined in this paper, are a superset of Web traversal patterns that effectively display the trends toward the next visited Web pages in a browsing session. Evolution of designing and developing websites from static to dynamic approach has let them update easily. A lot of research has been done on web usage mining. When the user browses the web pages, the leaves some valuable information in web logs [1]. This web log information is very helpful to find out the web navigation behavior of the user. Through his behavior, we can find out what kind of information user wanted from the web sites. Web usage mining automatic discover the knowledge from the data collected in log file. Many web analysis tools exist but they are limited and the efficiency of these tools is a state of perfection. Clustering, classification and association rule mining are active areas of learning research that is proving to be promising to help with this problem. Web mining can be divided into three different categories according to the kinds of data to be mined [2]. The vast size of the World Wide Web (WWW) nowadays makes it the largest database ever existed. Back to the beginning of this decade it was estimated to contain over 350 million pages while a couple of years ago, it had been estimated that only the indexed part of WWW by a web search engine consists of at least 11.3 billion pages [15]. Every attempt to shape these volumes of data, that follows a very loose schema, is quite difficult and extreme challenging. According to the application of data mining techniques in order to extract useful information that implicitly lay among web data is a very essential task. Web data may be either web data pages or data describing the activity of users. Actual web data consists of the web pages, the web page structure, the linkage structure between the web pages, the surfing navigational behavior of the users and the user profiles including demographic and registration information about the users .Web data mining can be divided in three general categories: web content mining, web structure mining and finally web they are web content mining, web usage mining and web structure mining. These three categories are:

1. Web Content Mining (WCM), involves the mining extraction and integration of useful data in the content of web pages, e.g. Structured text data (plane text content) , semi-structured data (html code), pictures and downloadable files.

2. Web Structure Mining (WSM), focuses on the inner document structure which means discovering the link structure of the hyperlinks at the inter-document level.

3. Web Usage Mining (WUM) (or web log mining) operates on the data from server access logs, information from users, registration application forms, user profile

1. Preprocessing involves three phases: (a) cleaning, which means that useless entries are discarded. (b) session

identification by assign all request from one user to one unique session.(c) data conversion into the format specific for the software tool.

2. Pattern discovery, means applying the presented algorithm with defined constraints by the user to the data [8] the following are the pattern discovery methods-

1. Statistical analysis
2. Association rules
3. Clustering
4. Classification
5. Sequential patterns

Web data are those that can be collected and used in the context of Web Personalization. These data are classified in four categories [2].

Content: The content should be presented to the user of the Web-Sites in an organized way or in a structured way. The organization can in a form of Text, Images or any data retrieved from the databases.

Structure: The Structure should be that the navigational pattern should not be tedious, which in can in a form of hyperlinks, HTML Tags etc.

Usage: data represent a web site's usage such as the IP address, time and date of access, files or directories that have been accessed .The above factors are included in the Web access logs.

User Profile: The User Profile discusses the demographic information, age, Qualification interest, and other factors related to the taste of navigation of the Web Sites. Often these information is collected by Questionnaire or can be analyzed Usage of Web Logs.

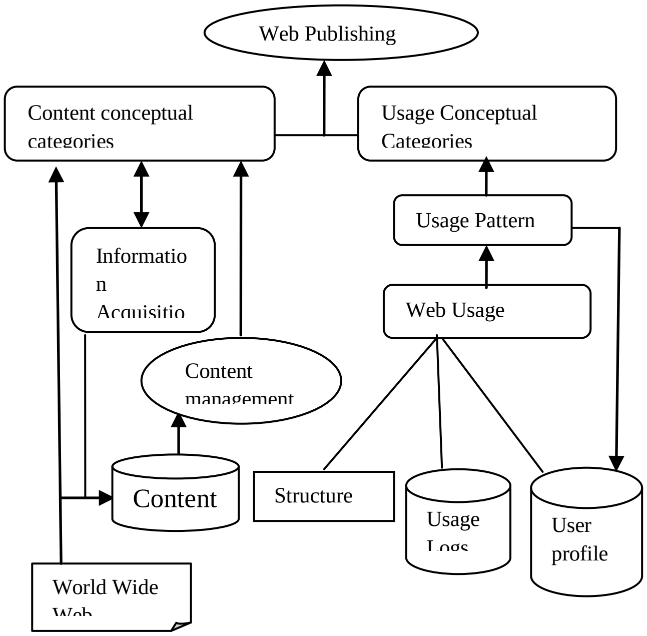

Fig: 1.1 Overview of Web-Personalization [1]

2. Historical Background of Web Usage Mining:

This Part of the paper discusses about how web Usage Mining Developed and led to Web Personalization which in turn gave birth to navigational pattern of User which again in turn gave boost for the E-business because the user can access the Information which is useful and valuable.

This Section describes in chronological order the development story of Web Usage Mining. The Development Web Usage Mining was from 90's.

The paper by jaideep Shrivastav et.al (2000) [3] discusses three steps that are being used in the Web Usage Mining are-:

2.1 Preprocessing

Preprocessing involves the usage, content and structure information contained in the various forms into abstraction necessary for Pattern discovery.

2.1.1 Usage Preprocessing:

This is the most complex part in the Web Usage Mining due the incompleteness of the available data. Unless a client side tracking mechanism is used, only the IP address, agent and server side clicking stream are available to identify users and server sessions.

The problems that are being identified from preprocessing that are being discussed by Jaideep et.al are

(a) Single IP/Multiple Server Session

(b) Multiple IP/Single Server Session.

(c) Multiple IP/Single Users

(d) Multiple Agent/Single Users

The problem is that to infer cached page reference. These are problems that are being discussed by the author/(s).

Second part according to the author/(s) is

2.2 Content Preprocessing-:

The content like Images, text, scripts and other files such as multimedia files are converted into useful data for Web Mining Processes. The process involves classification and clustering. Result of a classification is such that what type of pages has been visited or what class of products has been searched.

This paper goes further and discusses how to run the content mining. first of all the information is to be converted into quantifiable form like text files can be broken up into vector of words and keywords or text descriptions can be substituted for image or multimedia files. The author discusses that the classification of dynamic Web Pages gives the challenges of multiple sessions.

2.3 Structure Preprocessing-:

A well organized page should have well organized hyperlinks. Therefore the content of the pages and referenced pages are structured preprocessed as content of site.

2.3.1 Pattern Discovery-:

Pattern discovery has wide range of applications like on statistical data, data mining and machine learning .The author has limited the coverage of pattern Discovery in the field of Web Domain.

The Pattern Discovery in the Web Usage Mining to analyze and Discover the Pattern that has been generated by Server sessions which is the sequence of pages requested by the user.

The Pattern Discovery involves the following steps:

2.3.2 Statistical Analysis-:

It is the most common technique to get the information about how the user is using the Web Sites like the page views, viewing time and the length of navigational path. Many times the user does not visit all the pages in the Web sites. The Statistical Analysis involves simple Arithmetic mean, median and mode.

2.3.3 Association Rule-:

Association rule generations can be used to relate pages that are most often referenced together in single server session. According to the author association rule refer to set of pages that are accessed together with a support value exceeding some specified threshold.

2.3.4 Clustering.

The similar data item when grouped the grouping is done on the basis of similar characteristics, it can be called Clustering. The clustering in the field of WUM is that to group similar pages having same type of content.

2.3.5 Classification

Classification is to classify the given data. In WUM Classification is to classify the user according to the web site visited. Like the online shopping of computer peripherals belong to age group of 16 to 25.The classification can be done by different algorithm like decision tree classifier ,Bayesian classifier etc.

2.3.6. Sequential Pattern.

It is the pattern used by the advertiser in which the prediction is made by advertiser that what next page is to be visited by the user.

2.3.7 Dependency Modeling

This model uses the Hidden-Markov Model and Bayesian Belief Network. The dependency Model predicts the consumption of the Web Resources [9].

In the above paper author has discussed different steps involved in the Web Mining but has not discussed the merits and demerits of the steps involved.

The discussion on Web Usage Mining is incomplete without the name of Cooley et.al [4] .The trio Cooley, Mobasher and Srivastav did in depth study of procedure of WUM. The work of Cooley focused on the pre-processing Techniques which separates the User's Navigational data and or content purposes. The mining of Web Server Logs for better Web Site Design was the work of Drott(1998)[5] .This led to boosting of advertising world to give the users of Web coorectly and in time. The paper by Das et.al [6] gives the detailed study of Web Usage Mining and also the application of Web Usage Mining

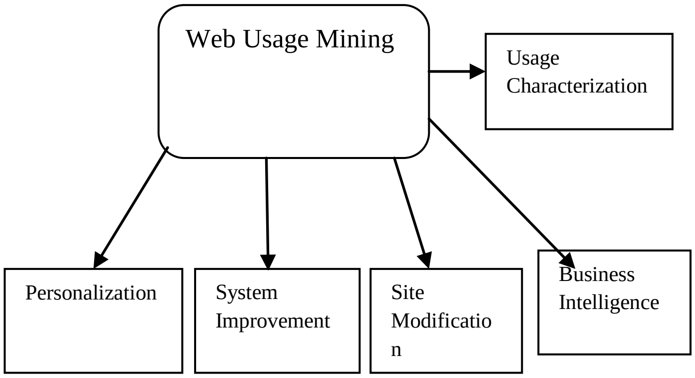

Fig-2.1 Main Applications of Web Usage Mining [6]

Web Usage Mining uses on the secondary Web Data such as Web server logs, proxy server logs browser logs user profiles registration data user sessions, transactions, cookies user queries, bookmarks mouse clicks or any other data generated by interaction of user with the web.

The paper discusses about the Path analysis, the methodologies that have been by the author is the methods that have been already discussed in the above section.

A path analysis is a hierarchical multiple regression analysis used to test the fit of the co-relation matrix against two or more causal model. The methodology adapted by Das et.al can be better be understood by the figure below which can be self-explanatory.

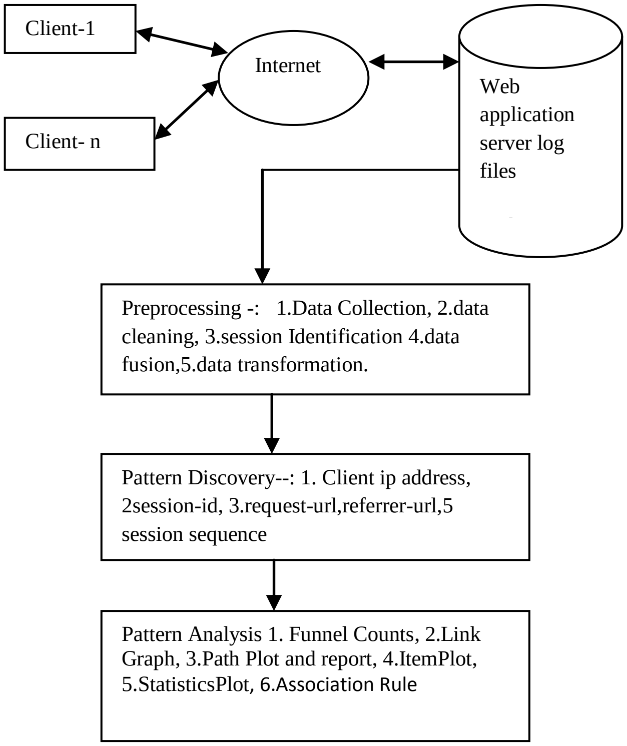
Client-1
Client- n
Internet
Web application server log files
Preprocessing -: 1.Data Collection, 2.data cleaning, 3.session Identification 4.data fusion,5.data transformation.
Pattern Discovery--: 1. Client ip address, 2session-id, 3.request-url,referrer-url,5 session sequence
Pattern Analysis 1. Funnel Counts, 2.Link Graph, 3.Path Plot and report, 4.ItemPlot, 5.StatisticsPlot, 6.Association Rule

Fig-2.2 Methodology adapted by Das et.al

The above work does not explain about the accuracy level, it also does not say about the limitations of the work.

The next part of the development story and the approaches made by the authors are elaborated in this part.

The paper by Pawel Weichbroth et.al [7] has shown a different framework for the Web Usage Mining. It comprises of Six Steps. Some steps that has been given are common like Database, Data Access Service Controller is the steps that has been used differently it says that implementation of framework main logic it controls all data flow and main functionalities.

The next step which is described is the User Interface which is as usual human Computer Interaction. The next step in this paper is the algorithm, the algorithm applied is Apriori – algorithm. Then comes the File controller, which consists of two parts (a) parse log files and next (b) is save the output of algorithm result.
The conclusion of this paper says that the author has not included a Holistic approach.

2.4 The last paper which is been included in the historical development of WUM is by V.Sujatha et.al [8]. In her paper she proposed a novel algorithm and named it as predicting User Navigation pattern using Clustering and Classification from Web Log Data (PUCC) to predict user's navigation.

In this paper the algorithm and the process is given by the following figure.

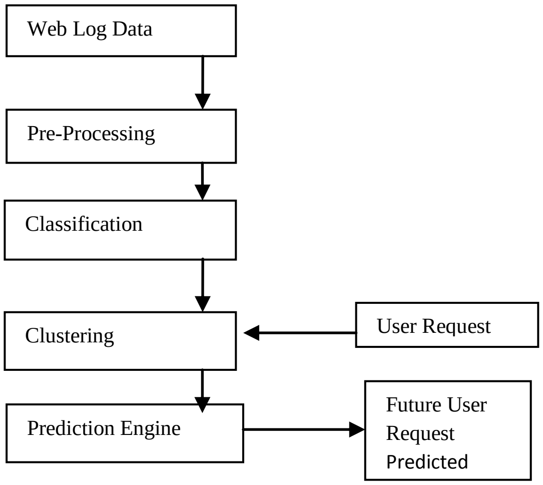

Fig-:2.3 Methodology of WUM [8]

The Processes are the same but up to Clustering in this paper adapted the process of Longest Common Sub Sequence (LCS) in the prediction Engine part. The main aim of LCS is to find longest subsequence common to all sequence in a set of sequences.

3. Proposed Approach:

The proposed approach consists of the following steps:

Step1: In the first steps data is being collected from the Web log file and then Preprocessing is applied. In the Preprocessing the Data is being loaded and it is being converted in to the Data set having fields Client-IP,Session_ID,Country,AccessDateTime,Method,URL,URL- ID,Protocol,Status,Bytes. In the Preprocessing Data cleaning by removing the Image which many consists of jpeg formats, in this step Session Identification is also performed .This step also performs Data Transformation .This is done because Data is coming from different Sources. The session is calculated in 30 minutes interval of time, after 30 minutes the system will recognize the same user as next user.

Step 2: In this step there is Pattern Discovery which is performed by the Frequent Pattern (FP) which involves FP Tree which in turn FP growth FP tree method is used in Data Mining .It consists of two passes over the Data Set .In the first Pass it scans data and find the minimum support for the each item. The item set whose support is less than minimum is discarded .The Data item that is included is the Web Site or the URL that is being visited by the User.

Next steps in the First Pass in the FP tree are to generate a decreasing order on the basis of frequency of occurrence of the Item Set Which is the URL visited by the User. In the Second Pass of the FP Tree Transaction is being read .In this work the Transaction is the number of user visited the particular Web Site.

The Read Transaction is iterated until all the Transaction is being completed. After Reading all the Transaction discards all the transaction which has lees support or support than the minimum threshold value which is 20 percent in this case.

Identification is done for the frequent Item-set from the Frequent Item-Set, minimal frequent Item-set is classified.

Step3: In this step Pattern analysis is done and in this

Candidate rule is generated and on the basis of candidate rule confidence is generated.
The above steps are being described diagrammatically.

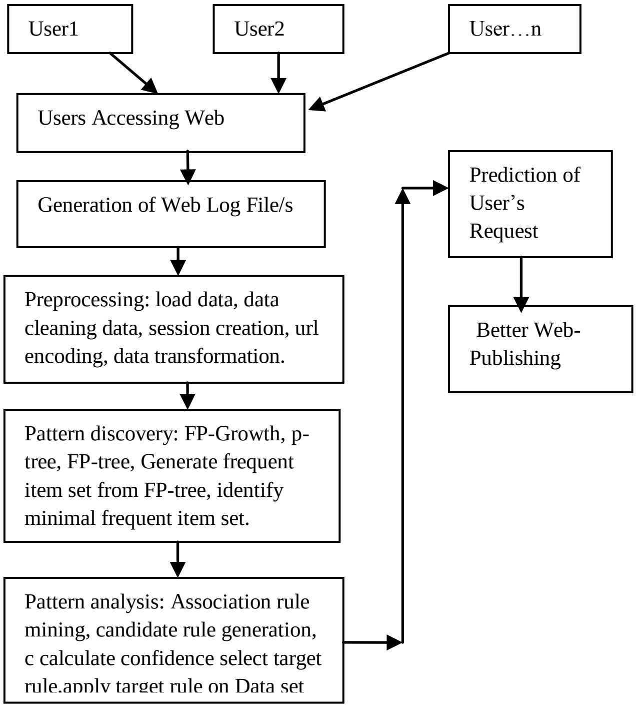
User1
User2
User…n
Users Accessing Web
Generation of Web Log File/s
Preprocessing: load data, data cleaning data, session creation, url encoding, data transformation.
Pattern discovery: FP-Growth, p-tree, FP-tree, Generate frequent item set from FP-tree, identify minimal frequent item set.
Pattern analysis: Association rule mining, candidate rule generation, c calculate confidence select target rule,apply target rule on Data set
Prediction of User's Request
Better Web-Publishing

Fig: 3.1 Methodology of Proposed Approach

The description of the proposed approach is given by the above figure. The result of the above approach is discussed in the next section.
In the result of Web Usage Mining the paper discusses about the three sets of Parameter. The first set of Parameter is the (a) Time V/s number of Transactions (b) Confidence V/s Number of Rule generations.

The third set of parameter is the (c) Support V/s Number of Rule Generations.

The graph of the above discussed Parameter is given subsequently.

(a)

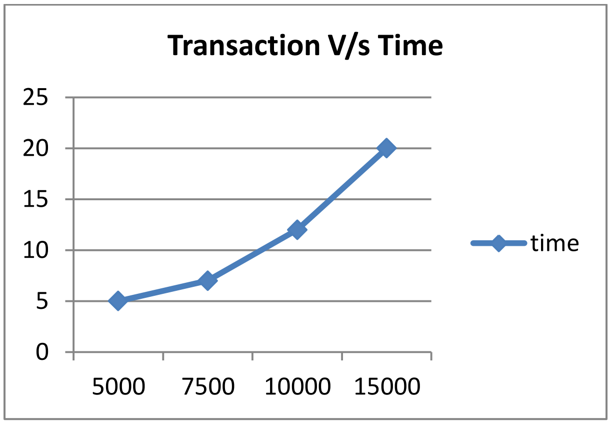

Fig-3.2 Graph of Transaction V/s Time

In the above graph the Time is taken in the Y-Axis and it is seconds and in the X-Axis there is Transactions.
(b)

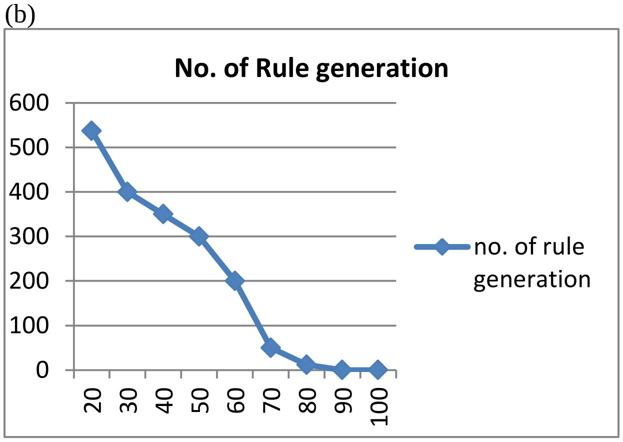

Fig-3.2 Graph of Confidence V/s Rule Generation

In this graph X-Axis is used for the No. of Confidence and in the Y-Axis represents the No. of Rule Generations.
(c)

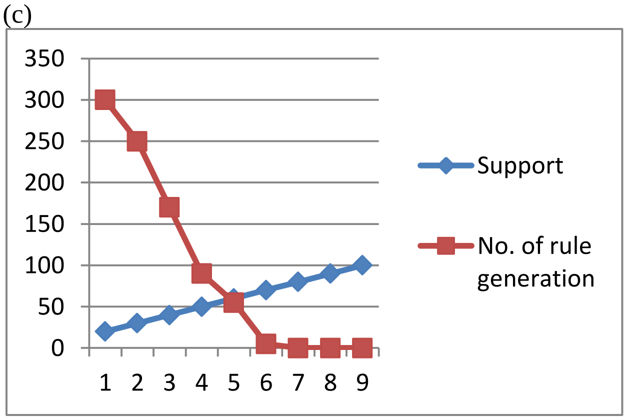

Fig-3.2 Graph of Support V/s Rule Generation

The graph that is being described in this section is Support V/s No of Rule Generation.

Result Analysis of Various Graph that have been given above.

(a) The graph is plotted between Transaction V/s Time .In the graph it is being seen that as the time increases the number of transaction also increases. Hence it can be seen that number of transaction done by the user increases relatively to the time.

(b) In the second graph the plotting is done between Number of Confidences and Number of Rule Generation .In the second graph it can be seen that it is inversely proportional to each other.

(c) The last graph is between support and number of rule generation. The graph states that Number of Rule Generation decreases as number of support increases. The next section is Conclusion.

4. Conclusion:

This Paper is done the Web Usage Mining by the help of three simple steps .The Complexity of the Steps is simple as compared to the other Algorithms or methodology that has done by the various authors.

The Future work can be on Parallel systems or can be implemented on Cloud Technology.

References

[1]. R. Chourasia, P. Choudhary, "*An Approach For Web Log Pre-Processing And Evidence Preservation For Web Mining*", International Journal of Computer Sciences and Engineering, Vol.2, Issue.4, pp.210-216, 2014.

[2]. M. Spilopoulou, L.C. Faultstich, Wum, "*AWeb utilization Miner*", The World Wide Web and Databases, Vol. 1950, pp,184-20, 1998.

[3]. J. Srivastav, R. Colley, M. Despanade, PN. Tan, "*Web Usage Mining: Discovery and Application of Usage Pattern from Web Data",* ACM SIGKDD Explorations Newsletter, Volume.1, Issue.2, pp.12-23, 2000.

[4]. R. Cooley, B. Mobasher, J. Srivasta, "*Web Mining: Information and pattern Discovery on theWorld WideWeb"*, IEEE International Conference on tools with Artificial Intelligence,USA, pp.558-567, 1997.

[5]. O. Etzioni, "*The world wide web Quagmire or gold mine"*, Magazine Communications of the ACM, Vol.39 Issue.11, pp.65-68, 1996.

[6]. R. Das, L. Turkoglu, " *Creatingmeaningful data from web logs for improving the impressiveness of a website by using path analysis method*", Expert Systems with Applications, Vol.36, Issue.3, pp.6635–6644, 2009.

[7]. P. Weichbrth, M. Owoc, M. Pleszkum, "*Web User Navigation Pattern Discovery from WWW Server Log Files"*, Proceeding of the Federal Conference on Computer Science and Information Systems, Wroclaw, pp.1171-1176, 2012.

[8]. V. Sujatha, Punithavalli, "*Improved User Navigation Pattern Preditions From Web Log Data* ", Procedia Engineering, Vol.30, pp.92-99, 2012.

[9]. G. Mansingh, L. Rao, KMO. Bryson, "*Profilling Internet Banking users:A Knowledge discovery in data mining process model based approach*" , Information Systems Frontiers, Vol.17, Issue.1, pp.193–215, 2015.

[10]. U. Patil, S. Pardeshi, "*A Servey on User Future Request Prediction: Web Usage Mining"*, International journal of Emerging Technology and advanced Engineering, Vol.2, Issue. 3, pp. 121-124, 2012.