Click Stream Analysis in e-Commerce Websites-a Framework

*1A. Vijaya Bharathi, *2Jyothi M. Rao, #$3Amiya K. Tripathy
*Computer Enginering, K.J. Somaiya College of Engineering, Mumbai, India
#Computer Enginering, Don Bosco Institute of Technology, Mumbai, India
$School of Science, Edith Cowan University, Perth, Australia
1vijaya.a@somaiya.edu, 2jyothirao@somaiya.edu, 3amiya@dbit.in

***Abstract*—The growth and proliferation of Internet has generated a revolution in retail practice. People nowadays prefer virtual shopping over Brick and mortar. So "Customer Retention" is a vital issue in today's e-commerce market. In order to boost customer loyalty, it is crucial for any e-commerce company to have an extensive understanding of online user behavior to strengthen the bond with their "e-customers". Though click stream analysis has been solving e-business problems, still recommendation system and digital marketing are far from perfect. In this article, different pattern discovery methods are addressed to identify various navigation patterns from weblogs to better understand users' behavior in e-commerce websites. The integrated approach of cognitive science and data mining on click stream could provide deeper insights about customer's thinking patterns, perceptions and their decision-making styles which could be utilized for effective customer retention.**

Keywords—weblogs, pattern discovery, click stream analysis, markov model, cognitive model, web usage mining.

I. INTRODUCTION

Today's e-commerce applications supposed to fulfil the demands of thousands of customers failing which can cause huge loss of revenues. Hence, the success of any online company highly depends on potential to captivate visitors. It is feasible for the company to track the data about customer interaction through the so-called click stream data. It is the principal source of information for the companies to adapt their service according to their customers. Click stream analysis may help these organizations determine customer loyalty, improve marketing strategies, effectiveness of promotional campaigns, provide more customized data to visitors, effective website structure, etc. Hence, understanding user's behaviour in Web applications has become necessary for ecommerce.

While the expectation for customer level data analysis is high, there are still problems such as customers receiving significant amount of uninterested mail advertisements and online recommendations are still far from absolute. To build more accurate consumer behaviour models for customers, firms need to recognize their customers better. This includes understanding customers' preferences and customers' behaviour through web history data.

Various pattern discovery algorithms are used by different researchers for identifying web usage patterns. Temporal logic model approach is used in [1] as an option to data mining techniques for the evaluation of structured weblogs. Complex user behavioural patterns were identified by checking temporal logic formulas against the log model developed using SPOT libraries to improve the structure of a website. The K-Nearest-Neighbour (KNN) has been used for successful classification of a real time recommendation system [2]. KNN algorithm is adapted to classify Frequent Access pattern [4]. Several data mining techniques namely association rule mining and decision tree were applied on click stream data to determine user interests and product associations for effective recommendation [5]. The interested users on web were identified using Naïve Bayes classification [7]. The concepts of Hidden Markov Model (HMM) has been used to predict if the user has the intention to buy something or not by the appearance of shopping-cart page in that session [9]. Markov models are also used to create usage profiles so as to optimize the structure and reduce the operational costs in maintenance [10]. With the help of the navigation pattern web users can be grouped based on their cognitive style. It can be used for modelling users to assist in adaptive websites for better organization of information [11].

II. BACKGROUND

Web Usage Mining is the discovery of useful patterns from the weblog data for better understanding of web users. It helps to know about users' behaviours and patterns which can be useful for effective management and construction of the site [13, 14]. The various sources of web usage data include the proxy server logs, web server logs, browser logs, user profiles, mouse clicks, user sessions, user queries, registration data, cookies and any other data as a result of web interactions. The web log files are primary source of data which can be collected from web Servers, proxy servers and Client browsers. A sample raw log file entry is shown below.

2016-02-13 00:12:27 128.230.247.37 GET clothing 80 74.111.18.59

Mozilla/5.0+(iPad;+CPU+OS+9_2_1+like+Mac+OS+X) + AppleWebKit/601.1

http://group0.ist722.ischool.syr.edu/beats-pill-20-wireless-speaker 200 687

The web log contains Date, Time, server IP, HTTP method, URI-query, Server Port, Client IP, User Agent, Referrer Agent, Status and Time taken. User Agent contains the client operating system and browser information whereas Referrer Agent contains the source from where this user arrives. These log attributes provide useful knowledge about navigation behaviour of users [14].

The data collected from web server log is often defective and unreliable [12]. Hence it needs pre -processing. It involves tasks such as removing references to embedded objects such as style files, graphics, or sound files, removal of at least some of the data fields (e.g. number of bytes transferred or version of HTTP protocol used, etc.) that may not provide useful information in analysis [13]. Every new IP address is considered as a new user. To accurately identify unique users, combination of IP addresses and other information such as user agents and referrers can be used. A

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

series of pages viewed by a user at a particular visit is known as a Session. [14] Session identification can be done by using timestamp of consecutive log entries. Click stream analysis is performed using statistics, data mining or machine learning algorithms [14]. The meaningful patterns are then analysed using online analytical processing(OLAP) or Visualization techniques.

III. PATTERN DISCOVERY APPROACHES

Some of the widely used classification/prediction techniques are KNN, Decision tree, Markov model, Naïve Bayes and cognitive model.

Using KNN

Identifying the interest of customers becomes necessary for an online company to serve them better. K-Nearest Neighbour algorithm compares a particular test sample with a set of training data that are similar to it [3,4]. Depending on the class of their closest neighbours, the category of the page visited by a user can be determined. The K-NN classifies the tuple based on similarities or distance to the stored training tuples [2, 3].

The Euclidean distance between a training tuple and a test tuple can be derived as follows:
Let X_i be input tuple with p attributes ($x_{i1}, x_{i2}, \ldots, x_{ip}$)
Let n be the number of input tuples (i =1, 2, …., n)
Let p be the number of features (j =1,2, …, p)
The Euclidean distance between Tuple X_i and X_t is

$$d(x_i, x_t) = \sqrt{(x_{i1} - x_{t1})^2 + (x_{i2} - x_{t2})^2 + \ldots + (x_{ip} - x_{tp})^2}$$

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

Let us consider an e-commerce site of an A-mart store with click stream as a vector of four attributes: users, source, page accessed, category with users represented by U1, U2, U3. . ., U7 as shown in Table 1. To determine the category of product purchased by user U3, we have to compute the Euclidean distance between the vector U3 and all other vectors. The Euclidean distance between two tuples U1 and U3 where U1 = (U_{11}, U_{12}, U_{13}) and U3 = (U_{31}, U_{32}, U_{33}). From Table 1, U1= (direct, Amart/home/grocery, Home) and U3 = (search engine, Amart/home/footwear/sport, footwear).

TABLE I. A-MART'S TRAINING TUPLES

UId	source	Page accessed	Category	class
U1	direct	Amart/home/grocery	Home	Home, personal care
U2	Search engine	Amart/kids	Kids	Kids apparel
U4	direct	Amart/footwear	footwear	footwear
U6	Search engine	Amart/ladies garments/kurti	Ladies	Ladies garments
U7	direct	Amart/men's apparel/t-shirt	Men's Apparel	Men's apparel
U8	Search engine	Amart/ladies garments/	Ladies	Ladies garments
U9	direct	Amart/kids	Kids	Kids apparel
U3	search engine	Amart/home/footwear/sport	footwear	?

For categorical attributes, the difference ($U_{11}U_{31,}$) can be computed by simply comparing the corresponding value of the attributes in tuple U1 with U3. If the values are the same then the differences taken to be zero (0), otherwise, the difference is taken to be one (1). So, for ($U_{1,1}$ and $U_{3,1}$) i.e. (direct, search engine), the difference is 1, for (U_{12}andU_{32}) i.e. (Amart/home, Amart/footwear) the difference is 1, likewise for (U_{13} and U_{33}) i.e., (home, footwear) the difference is 1.

The same process is repeated with all other tuples U2, U4, . ., U9, and the result produced a stream of data sorted by their Euclidean distance to the user U3 which is shown in Table 2. Thus, the user U3 has visited footwear related page. Similarly, whether a visitor is seasonal or regular, week end /night visitors can be found out to better understand the users' behaviour. This knowledge about user can be used for customized marketing.

TABLE II. DISTANCE TO USER U3

User	Class	Distance to User U3
U4	Footwear	1.00
U2	Kids Apparel	1.414
U6	Ladies Garments	1.414
U8	Ladies Garments	1.414
U7	Men's Apparel	1.732
U1	Home and personal	1.732
U9	Kids Apparel	1.732

Using Decision Tree

Any online company need to know their potential customers in order to optimize traffic and spend effectively on digital marketing.one of the popular classification algorithm is a decision tree in which each non-leaf node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each leaf node denotes a class prediction [5]. The information gain measure can be used to select the test attribute at each node. The attribute with the highest gain is chosen as the test attribute for the current node [6].

TABLE III. A-MART'S TRAINING TUPLES

User id	Session Id	Session time(mins)	Method used	No of pages	class
U1	1	10 (less)	Get	8 (more)	Casual
U2	2	25 (more)	post	10 (more)	Potential
U3	3	30 (more)	Post	6 (more)	Potential
U4	4	14 (less)	Get	4 (more)	casual
U5	5	12 (less)	Get	9 (more)	casual
U6	6	25 (more)	Get	10 (more)	potential
U7	7	27 (more)	Post	12 (more)	Potential
U8	888	35 (more)	Post	15 (more)	?

To identify potential customers from large volume of big data, consider the set of attributes (session id, session time, no. of pages accessed, method used) from Table 3. The basic idea is to segregate users on their purchase interest and those who simply explore the site. Generally interested users spend long time on web pages and use the HTTP POST mode if they are interested in registering with web sites. The uninterested simply accesses many pages quickly to browse contents [5, 6]. These users do not often use POST method because they are not interested in registering at web sites.

The best splitting attribute Info (D) is calculated as

Gain (A)=Info (D) - Info_A(D)
Info (D) = - $\sum p_i \log_2(p_i)$
Number of tuples belong to potential (yes class) =4
Number of tuples belong to casual (no class) =3
Info (D) =- (3/7log (3/7) +4/7 log (4/7)) =0.984

$$\text{Info}_{session}(D) = \sum \frac{|D_j|}{|D|} \times \text{Info}(D_j)$$

=3/7Info (session <25) +4/7 Info (session >25)
=3/7*(-(0/3) log (0/3) -(3/3) log ((3/3)) + 4/7*(-4/4 log4/4-0)
= (3/7) *0 + (4/7) *0=0
Gain (session) =0.98-0= **0.98**

Info_{method}(D)=4/7*Info(method='G')+3/7*Info (method='P')
= (4/7(1/4 log 4 + ¾ log 4/3)) + 3/7 (-3/3 log 3/3 -0)
=0.46+0
Gain (method) =0.98-0.46=**0.52**

Info_{Number} of pages (D) =7/7 * Info (number='more') + 0/7 * Info (number='less')
=7/7* (-4/7 log 4/7-3/7 log 3/7) +0=0.98
Gain (number of pages) =0.98-0.98=**0**

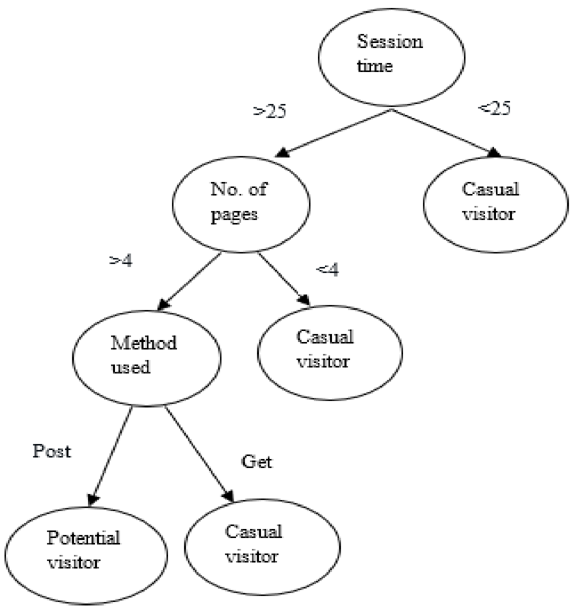

Fig. 1. Decision tree generation

It is observed that session time attribute has the highest information gain (0.98). The Users are classified as "Potential" and" casual" based on the parameters Time Stamp, method used (GET/ POST), number of pages referred. The decision tree generated is shown in Fig 1.

From the decision tree generated, rules can be easily interpreted classifying User U8 as potential user. Similarly, we can classify as new or returning visitor of the site with the help of an attribute 'frequency of visit' (difference between two timestamps).

Using Markov Model

Predicting user's next page request on the World Wide Web is currently an urgent issue. Different methods exist that can look at the user's page views and predict what next page the user is likely to view. On such method is Markov process in which states represents the web pages and edges represents transition probabilities. A trained Markov model can be used to predict the next state, given a set of *p* previous states [9,10]. Markov models can be denoted by three parameters < A S T >, where A represents all actions performed by the user; S represents all possible states; and T is a |A| X |S| Transition Probability Matrix (TPM), where Tij represents the probability of performing action j when the process is in state i.

TABLE IV. SAMPLE PAGE VIEWS

User	Page View
U1	p2→p3→p2→p1→p5
U2	p2→p1→p3→p2→p1→p5
U3	p1→p2→p5
U4	p1→p2→p5→p2→p4
U5	p1→p2→p1→p4
U6	?

The simplest Markov model predicts the next action by only looking at the previous action performed by the user [9]. A markov process is represented as a directed acyclic graph in which every node denotes a state corresponding to a page view, and edges labelled with probabilities represents transitions between the connected states. All transition probabilities are stored in a transition probability matrix $\mathbf{P}n \times n$, where n is the number of states in the model [10]. Consider the set of transactions presented in Table 5.

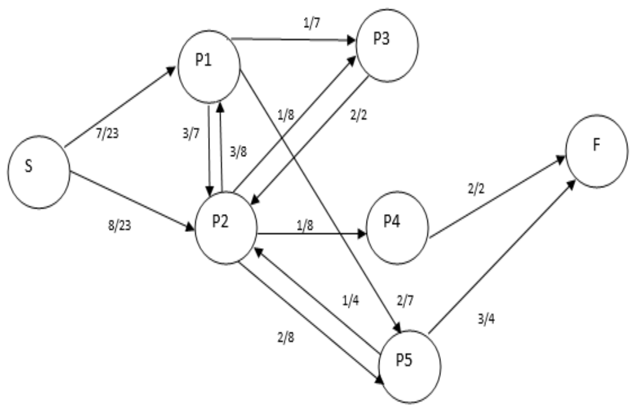

Fig. 2. Markov chain for web transactions

To build a Markov chain start with an initial state (S) into the chain and a final state (F) at the end. The probabilities associated with the edges are obtained by counting the number of times the transaction occurs in the trails. The probability to move from the initial state S to a state p1 represents the page p1 is about 7/23 (0.31), where 7 is the number of times that page p1 occurs, and 23 is the total number of requests.

TABLE V. TRANSITION PROBABILITY MATRIX

	P1	P2	P3	P4	P5
P1	0	0.43	0.14	0.14	0.28
P2	0.5	0	0.125	0.125	0.25
P3	0	1	0	0	0
P4	0	0	0	0	0
P5	0	0.25	0	0	0

Using the same process, the probability to move from pagep1 to page p2 is 3/7(0.42), where 3 is the number of times that p2 occurs after p1, and 7 is the number of times p1 occurs. Finally, the probability to move from page p4 to the final state F is 2/2(1), where 2 is the number of trails where p4 is the final state, and 2 is the number of times that p4 occurs.

The Markov chain generated from such transactions is depicted in fig 2. Assume that a user U6 browsed through the sequence of page views <p2→p5→p1→p3>. Looking at **P** in Table 4, there is 100% probability that the user will view page p2 next. A problem that could arise here is contradicting prediction, for example, there is an equal probability a user will view page p3 or p4 after viewing page p1. Thus, the prediction capability of the system will not be accurate and will be ambiguous in such cases [10].

Using Naïve Bayes

Since Naïve Bayes algorithm works best on large volume of data, it is addressed here to identify the same pattern discussed earlier using decision tree. P (H | Q) represents the probability that hypothesis H holds given the "evidence". Let us consider our training data set (Table 3.) attributes as: session id, time taken, number of pages visited then P (H | Q) is the probability that the session id may be a potential user or not given the time taken and number of pages viewed [7,8]. P (H) is called as priori probability of H. To classify User U8 as potential or casual user, compute the conditional probabilities s follows.

P(class='casual') =3/7=0.428
P(class='Potential') =4/7=0.571
P (session time='more'| class='casual') =1/3=0.333
P (session time='more '|class='potential') =3/4=0.75
P (method used='post'| class='casual') =1/3=0.333
P (method used='post'| class='potential') =3/4=0.75
P (pages accessed='more' |class='casual') =3/3=1
P (pages accessed='more'| class='potential') =4/4=1
P(U8|Class='casual') =0.333*0.333*1=0.111
P (U8|class='Potential') =0.75*0.75*1=0.563
P (Ci/X) =P(X/Ci) P (Ci)
P (U8|class='casual')
P (class='casual') =0.428*0.111= 0.0475
P (U8|class='Potential')
P(class='Potential') =0.571*0.563=**0.321** **(Maximum probability)**

The maximum probability obtained for the User U8 is with the 'Potential' class. Hence U8 will be predicted as a Potential customer of A-mart's store.

Using Cognitive Model

Cognitive science is an interdisciplinary advance towards the understanding of human behaviour. Cognitive sciences can have direct application to web usage mining. More recently, economists have applied such concepts to explain consumers' behaviour [11]. Cognitive styles describe the way users process and organize information. Previous relevant works had identified a number of dimensions in which the users' cognitive styles may differ. (Chen and Macredie, 2002; Liuand Ginther, 1999). Serialists and wholists have different characteristics among them (Pask, 1976). The wholists opt for global way of processing while serialists prefer to understand step by step. Researchers had proposed specific user interaction metrics to inspect how users navigate (i.e., linearly/non-linearly) based on the sequence of hyperlinks visited. Clustering techniques can be performed to determine users' navigation behaviour and their relation to cognitive styles. In order to measure the linearity of user interactions with the website the following interaction metrics are used.

Absolute Distance of Links (ADL): It represents the sum of total absolute distance between the hyperlinks visited by a user.

Average Sequential Links (ASL), denotes the number of sequential links visited by the user .

Average non-sequential Groups of Links (AGL), indicates the number of non-sequential links visited by the user. The above metrics can be used to find whether the user had followed a linear/nonlinear navigation path. For example, consider the sequence of pages visited by users in Table 6.

Session id 1 contains the page sequence A, B, C, D, E (where A is first link from homepage referred as 1, 2 ,3, 4, 5 for easy understanding) for which the interaction metrics ADL, ASL, AGL can be calculated as ADL= (|1-1|+|2-1|+|3-2|+|4-3|+|5-4|)/N=4/5=0.8 where N =Number of total links clicked. ASL=M/N=5/5=1 where M=number of sequential links visited. AGL=B/N=0/5=0 where B=number of non-sequential links visited. Cognitive style ratio based on ADL ranging between 0 and 1.667 indicates a linear approach of navigation, range between 3 and 4 indicates a non-linear approach. There is a link between cognitive style dimension (i.e., Wholist–Intermediate–Analyst and Verbal–Intermediate–Imager) and navigation style (i.e., linear and non- linear) [11].

TABLE VI. SAMPLE PAGE VIEWS

Session Id	Transactions
1	A->B-> C-> D ->E
2	A ->B-> C
3	A-> B-> C-> E
4	C-> D-> E
5	C-> D-> E-> B
6	C ->D-> A-> E
7	D-> A-> B-> E

It is found that wholist type of users follow linear navigation behaviour. The identification of users with specific cognitive and navigation style will ultimately help e-retail companies to derive new strategies and methods to provide better service for the customers. [11].

Discussion

From the detailed survey done on various pattern discovery approaches, some useful patterns are identified and listed in Table 7.

TABLE VII. PATTERNS IDENTIFIED AND MODELS

Parameters	**Useful Patterns**	**Model Recommended**
User ID, Session ID, Pageview, Timestamp, Visit Duration	New User / Returning User, Potential / Casual Visitor, Evening Visitor / Week End Visitor Seasonal Visitors	Classification
User ID, Session ID, Page View, Item Bought, Product ID, Product Category, Price, Quantity	Interested Product Category for Each User	Classification
ser ID, Session ID, Age, Gender Product Category, Item Bought, Number Of Items Bought	Most Interested Product by Certain Age Group / Gender	Classification / Clustering
User ID, Products Purchased, Category	Product Association to Users	Classification / Association Rules
User ID, Session ID, Item Bought, Price, Frequent Visitor, Category	Predict Buy / Not	Markov Model / Classification
User ID, Session ID, Page View, Mostly Visited Page	Link Prediction, Predict Next Click, Frequent Sequence Pages	Markov Model / Sequence Pattern Mining
User ID, Session ID, Page View	Linear / Nonlinear Path, User Cognitive Styles	Cognitive Model

It is also clear that cognitive science in web usage mining has a scope to get deeper insights about consumer shopping psychology. The customer's thinking patterns, perceptions and their decision-making styles can be interpreted with the help of cognitive science and data mining as an integrated approach. Depending on the type of personality better customized marketing can be done to captivate customers. Similarly based on customer's cognition, relevant recommendations of product can be given more precisely to customers.

IV. PROPOSED IDEA

The proposed framework shown in Fig 3. adopts the conceptual framework of cognitive architecture namely ACT-R (adaptive control of thought-Rational) along with data mining which provide cognitive behaviours of online customers.

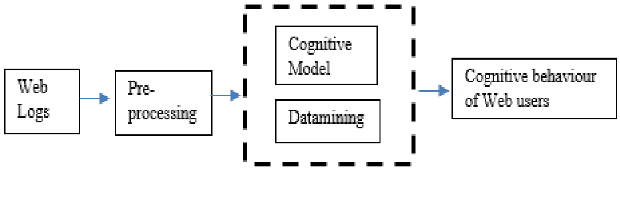

Fig. 3. Proposed methodology

The motivation behind this idea is the lack of study of cognitive processes on consumer model using click stream data. The click stream data can be effectively used to identify various styles of users' decision making and their perceptions. This customer analytics could help e -retailers for effective digital marketing and online recommendations.

V. CONCLUSION

Customer retention is a critical current issue faced by online retail companies. Online recommendation system and digital marketing are few business problems which are still far from perfect. From the comprehensive study, it is observed that the application of cognitive science in web usage mining is in its infancy, further investigation could reveal more relevant relationships between cognitive styles and navigation behaviour of user. Getting deeper insights about consumer psychology through navigation behaviour from weblogs can help e-companies improve their customer retention rate by providing more personalized marketing and relevant recommendations to user.

REFERENCES

[1] Sergio Hernandez, Pedro Alvarez, Javier Fabra, Jaoquin Ezpeleta "Analysis of Users' behaviour in structured e-Commerce Websites", IEEE Access, vol. 5, pp. 11941–11958, May 2017.

[2] D.A. Adeniyi, Z. Wei, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics, Vol. 12, pp. 90–108, 2016.

[3] Manisha Kumari, "A Review of Classification in Web Usage Mining using K-Nearest Neighbor", Advances in Computational Sciences and Technology, Vol. 10, No. 5, 2017.

[4] Suharjito, Diana and Herianto, "Implementation of Classification Technique in web usage mining of banking company", Proc. IEEE seminar on Intelligent technology and its applications, pp. 211-218, Jul. 2016.

[5] Yoon Ho Cho, Jae Kyeong Kim, Soung Hie Kim, "A personalized recommender system based on web usage mining and decision tree induction", Expert Systems with Applications, Vol.23, No.3, pp. 329-342, Oct 2002.

[6] Rianto, Lukito Edi Nugroho, P. Insap Santosa, "Pattern Discovery of Indonesian Customers in an Online Shop: A Case of Fashion Online Shop", Proc. IEEE Conference on Information Tech, Computer, and Electrical Engineering (ICITACEE), pp. 313-316, Oct. 2016.

[7] K. Santra, S. Jayasudha, "Classification of Web Log Data to Identify Interested Users using Naïve Bayesian Classification", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, Jan. 2012.

[8] Mahdi Khosravi, Mohammad J. Tarokh, "Dynamic Mining of Users Interest Navigation Patterns Using Naive Bayesian Method", Proc. IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 119-122, Aug. 2010.

[9] Chun-Jung Lin 1, Fan Wu, I-Han Chiu, "Using Hidden Markov Model to Predict the Surfing User's Intention of Cyber Purchase on the Web", Journal of Global Business Management, 2009.

[10] Alice Marques and Orlando Belo, "Discovering Student Web Usage Profiles Using Markov Chains", Portugal Electronic Journal of e-Learning, Vol. 9 Issue 1, 2011.

[11] Marios Belk, Efi Papatheocharous, Panagiotis Germanakos, George Samaras, "Modeling users on the World Wide Web based on cognitive factors, navigation behaviour and clustering techniques", The Journal of Systems and Software, Vol. 86, Issue 12, pp. 2995-3012, 2013.

[12] P. Dhana Lakshmi, Dr. K. Ramani, Dr. B. Eswara Reddy, "The Research of Preprocessing and Pattern Discovery Techniques on Web Log files", Proc. IEEE 6th International Conference on Advanced Computing, pp. 138-145, Feb. 2016.

[13] J. Srivastava, Robert Cooley and Mukund Deshpande, "Web Usage Mining: Discovery and Applications of Usage Patterns from web data", SIGKDD Explorations, Vol. 1, Issue 2, Jan. 2000.

[14] Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Data-Centric Systems and Applications, Springer, 2007.