Two Levels of Prediction Model for User's Browsing Behavior[1]

Chu-Hui Lee, Yu-Hsiang Fu

***Abstract*—Owing to the popularity of World Wide Web, many enterprises have changed the ways of doing business, which enhance the rapid development of E-commerce directly and makes the development of web usage mining skills important. It becomes a crucial issue to predict exactly the ways how users and customers browse websites. The prediction result can be used for personalization, building proper websites, promotion, getting marketing information, and forecasting market trends etc. Markov model is assumed to be a probability model by which users' browsing behaviors can be predicted at category level. Bayesian theorem can also be applied to present and infer users' browsing behaviors at webpage level. In this research, Markov models and Bayesian theorem are combined and a two-level prediction model is designed. By the Markov Model, the system can effectively filter the possible category of the websites and Bayesian theorem will help to predict websites accuracy. The experiments will show that our provided model has noble hit ratio for prediction.**

***Index Terms*—Web Usage Mining, Prediction, Markov models, Bayesian theorem.**

I. Introduction

Owing to the popularity of World Wide Web, many enterprises have changed the ways of doing business, which enhance the rapid development of E-commerce directly and make the development of web usage mining skills important. Web usage mining is a technology of web mining. Web usage mining applies the technology of data mining on the Internet and web services. It is used to find out and extract the pattern and the knowledge in the web usage automatically, for example, association rule, clustering algorithm, and sequential pattern analysis etc [6].

The goal of web usage mining is to find out the useful information from web data or web log files. The other goal is to enhance the usability of the web information and apply the technology to the web application, for instance, pre-fetching and caching, personalization, target advertisement, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise, and marketing analysis etc [2][6][7][12].

CRM (Customer Relationship Management), such as banking, is a quite popular field in which web usage mining is applied. The cost of on-line transaction is getting less than traditional commerce. CRM is further to lead customers to online transactions in the visual bank by applying web usage mining technology. Meanwhile, CRM and web usage mining technology also can be used on the E-commerce web site. The goal, such as customer attraction, customer retention and cross sales, will be achieved [1].

In order to achieve the purpose, it is necessary to understand the customers' browsing behavior and apply web usage mining technology. Web usage mining is used to predict users' and costumer's browsing behavior through mining the web data or web log files. Prediction is a way of analyzing historical information to calculate the most possible probability of next request; the browsing pattern when the future users and customers are surfing the web site is matched. There are many advantages to implement the prediction, for example, personalization, building proper web site, improving marketing strategy, promotion, product supply, getting marketing information, forecasting market trends, and increasing the competitive strength of enterprises etc.

In many portal of World Wide Web, web pages will be classified well for different purposes. In this paper, users' browsing behavior will be observed at two levels to meet the nature of the World Wide Web. One is category level and the other is web page level. Each category contains many web pages on web site. The transaction probabilities of at category level are more stable than web page level. In the level one of prediction model, Markov model is used to predict the category of users' next state with the first order transaction probability [8][13]. It will help to reduce the operation scope in level two just in some specific categories instead of all. After that, the web pages in suitable categories are predicted by Bayesian theorem in the level two of prediction model. It is expected that the two levels of prediction model can reduce the operation scope and increase the accuracy precision.

The rest of this paper is organized as follows: Section 2 is the related work. The two levels of prediction model are proposed in Section 3. Session 4 is the result of experiment. The conclusion and future work will be discussed in Session 5.

II. Related work

Because World Wide Web is rapid developed, Internet is a golden mount with a lot of valuable information. Web usage mining is a powerful and useful tool to find out the information in a systematic, efficient and automatic way. The

[1] This work was supported by National Science Council under Grant NSC 96-2221-E -324-043.

Chu-Hui Lee is with the Chaoyang University of Technology, Wu-Fong Township Tai-Chung County, 41349 Taiwan (R.O.C.). Phone: (04)2332-3000 ext. 4388; fax: (04)2374-2373; e-mail: chlee@ cyut.edu.tw.

Yu-Hsiang Fu is with the Chaoyang University of Technology, Wu-Fong Township Tai-Chung County, 41349 Taiwan (R.O.C.). e-mail: s9514611@cyut.edu.tw.

terminology of web mining was proposed by Etzioni in 1996. Web mining is applied to web pages and services of Internet in order to find and extract the available knowledge [6].

Web mining can be categorized into three categories (as Fig. 1), web content mining, web structure mining and web usage mining. Web content mining focuses on useful knowledge which is extracted from web pages. Web structure mining is used to analyze the links between web pages through the web structure to infer the knowledge. Web usage mining is extract the information from web log file which is accessed by users.

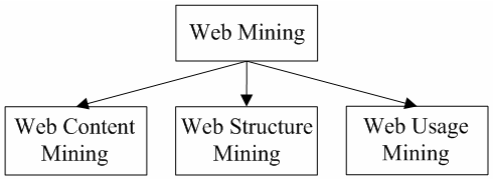

Figure1. Taxonomy of Web Mining

A. Web Log File

Web usage mining is used to find out the interrelated information from web log file which is involved. All of users' browsing behavior is clearly recorded in the web log file with users' name, IP address, date, and request time etc [16].

TABLE 1. DESCRIPTION OF COMMON LOG FILE

Content of log file	Description
163.17.9.84	Address or DNS
-	RFC931
-	Authuser
[08/Nov/2007:20:35: 52 +0800]	TimeStmp
"GET /menu.htm HTTP/1.1"	Http request
200	Status code
1266	Transfer volume
-	*Referrer URL
-	*Usage agent

The format of web log file can be divided into two types, common log files (as Table 1) and extended log [1][6][14]. Common log file is constructed by access log and error log, referrer log and agent log are appended to it, and form extended log. After web log file is acquired, the procedure of web usage mining must be executed.

B. Web Usage Mining Procedure

The first process of web usage mining procedure is preprocessing in the web log files. The second process is mining algorithm which is to find out the pattern. The final process is analyzing the pattern which is mined by mining algorithm. The details of process of web usage mining procedure (as Fig. 2) are as follows [2][12][14]: users' session, as a sequence pattern, is gathered from web log data and recognized through preprocessing process. Mining algorithm is used to find out the rules and patterns from the sequence pattern, for example, association rule, clustering algorithm, and sequential pattern analysis.

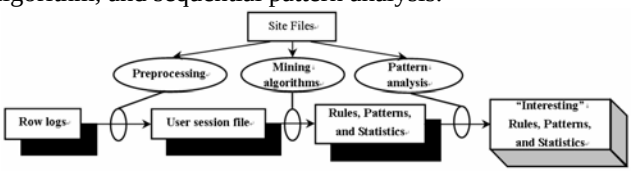

Figure2. Web Usage Mining Procedure

C. Markov model

Markov model is assumed to be a probability model by which users' browsing behaviors can be predicted; their behaviors can be presented by how they visit during browsing [5][8][10]. Markov model has a set of state space $s=\{s_1,s_2,...,s_n\}$. If the sequential pattern is presented as a state sequence $\{s_t;t=1,2,...\}$, then s_t denotes the sequence of state at time t. The transition of any two states in the sequence is based on transition probability p_{ij}. p_{ij} denotes the probability of state j which depends on pervious state i. The state j at a time t depends on state i at a time $t-1$ which is corresponding to first-order transition probability of Markov model. It also can denotes by $p_{ij}=\Pr\{S_t=j\mid S_{t-1}=i\}$. First-order transition matrix is established by calculating all of transition probability of all states. Any two elements in the translation matrix of Markov model are independent.

First-order transition probability can be denoted by $p_{ij}=p_{ij}^1$. It would not change with time. p_{ij}^n is obtained when the n-order transition probability denotes by p_{ij}^n. p_{ij}^n is a probability which state j at a time t depends on previous state i at a time $t-n$. The n-order transition probability of Markov model also denotes by $p_{ij}^n=\Pr\{S_t=j\mid S_{t-n}=i\}$. The n-order transition probability is calculated by well-know Champman-Kolmogorov equation as follows:

$$p_{ij}^n=\sum_{l=1}^{k}p_{il}^{n-1}\cdot p_{lj} \tag{1}$$

$$p_{ij}^{(m+n)}=\sum_{l=1}^{k}p_{il}^m\cdot p_{lj}^n=P_{ij}^{(m+n)} \tag{2}$$

$$P^{(n)}=P^{(n-1)}\cdot P=P^n \tag{3}$$

If P be a finite transition matrix and $P^{(n)}$ be a n-order transition matrix of Markov model, then $p^{(n)}=p^n$. The second-order and third-order transition matrix can be established by above equations. According to Champman-Kolmogorov equation, $P^{(2)}=P\cdot P=P^2$ or $P^{(3)}=P^2\cdot P=P^3$ will be acquired.

D. Bayesian theorem

Bayesian theorem as well as Markov model also can be used to predict the most possible users' next request. It is assumed that the sample spaces S, A and B are two events of sample spaces S and $P(B)>0$. The condition probability $P(A\mid B)$ means the probability of event A while event B occurs [15]. It is denoted as follows:

$$P(A\mid B)=\frac{P(A\cap B)}{P(B)} \tag{4}$$

In Bayesian theorem, some of the information is used to revise the prior probability and obtain the posterior probability. The above procedure is called Bayesian theorem [15] and the equation is follows:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \quad (5)$$

If A is a partition, $A=\{A_1,A_2,...,A_n\}$, then Bayesian theorem can be inferred as follows:

$$P(A_j|B)=\frac{P(B|A_j)P(A_j)}{P(B)}=\frac{P(B|A_j)P(A_j)}{\sum_{r=1}^{n}P(A_r)P(B|A_r)} \quad (6)$$

III. Two Levels of Prediction Model

The two levels of prediction model are designed by combining Markov model and Bayesian theorem (as Fig. 3). In level one, it is to predict the category at time t, which depends on users' category at time $t-1$ and time $t-2$.

Bayesian theorem is used to predict the most possible web pages at a time t according to users' states at a time $t-1$. Finally, the prediction result of two levels of prediction model is released. The result can be applied in pre-fetching and caching on web site, personalization, target sales, and improving web site design etc.

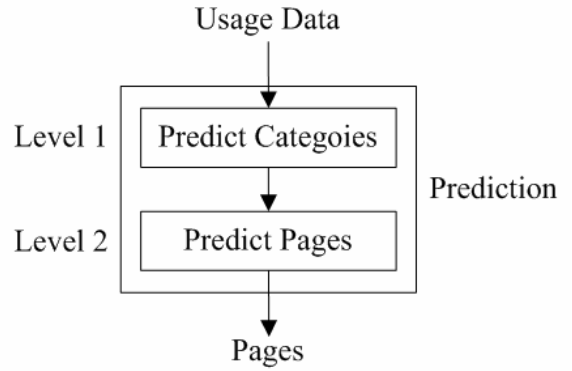

Figure3. Two Levels of Prediction Model

A. Preprocessing

The steps of research framework (as Fig. 4) are described as follows. At first, the similarity matrix S of category is established. The approach of establishing similarity matrix is to gather statistics and to analyze the users' behavior browsing which can be acquired from web log data. In step two, it is to establish the first-order transition matrix P and second-order transition matrix P^2 of Markov model. The transition matrix of Markov is established by the same approach, statistical method, from web log file.

In step tree, the relevance matrix R is computed from first-order and second-order (or n-order) transition matrix of Markov model and similarity matrix. In our proposed method, the relevance is an important factor of prediction. Relevance can be used to infer the users' browsing behavior between web categories.

It is assumed that D denotes a database, which contains m users' usage record. It means that the users' session is recorded and $D=\{session_1,session_2,...,session_m\}$ is obtained. Each user's session can also be recorded as a sequential pattern of n web pages which is browsed by time order, and $session^p=\{page_1,page_2,...,page_n\}$, where page i represents the user's visiting page at time j, is obtained. If a web site has k categories, then the user's session can be reorganized by $session^c=\{c_1,c_2,...,c_k\}$, where $c_i=1$ denotes that user visited the web page of category i otherwise $c_i=0$. After giving the definitions, more details for the prediction model are described in the following sections.

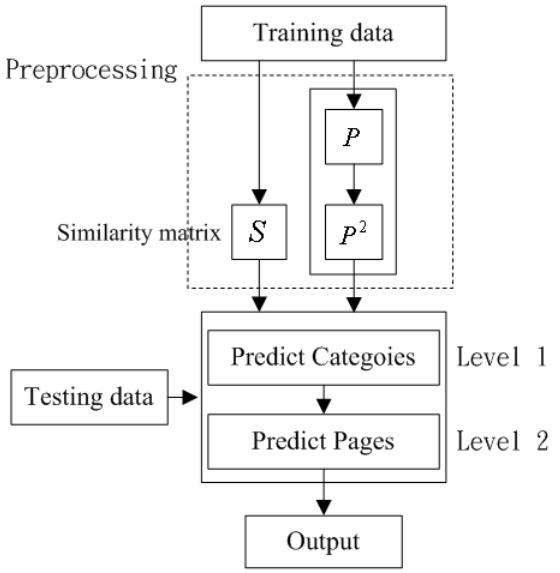

Figure4. Research Framework

1) Similarity matrix of web categories

Research framework of step one is to create the similarity matrix from web log file. At first, the situation of categories in each user's session has to be understood. The $vector_i=<v_{1,i},...,v_{h,i},...,v_{m,i}>$ for each category i is gather the i th element of $session^c$ from all m user sessions, $v_{h,i}=1$ means user h visited web page of category i otherwise $v_{h,i}=0$. Two categories can be calculated the Set similarity [3] and Euclidean distance [8], respectively, by equation (7) and (8), Euclidean distance is further normalized by equation (9). The results are computed by similarity and Euclidean distance. They are combined into a weight total similarity equation as equation (10).

Set similarity:

$$SetSim(A,B)=\frac{|A\cap B|}{|A\cup B|} \quad (7)$$

Euclidean distance:

$$D(A,B)=\sqrt{\sum_{i=1}^{m}(A_i-B_i)^2} \quad (8)$$

Normalization:

$$N(D(A,B))=1-\sqrt{\frac{\sum_{i=1}^{m}(A_i-B_i)^2}{m}} \quad (9)$$

Weight total similarity:

$$S(A,B)=SetSim(A,B)\cdot W_{SS}+N(D(A,B))\cdot W_D \quad (10)$$

where $W_{SS}+W_D=1$, $W_D=1-W_{SS}$.

After the similarity is calculated, the similarity matrix S is

a k x k matrix of categories similarity, where S_{ij} is the similarity between C_i and C_j that is established by above steps.

$$S = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \end{array}$$

2) Transition matrix of Markov model

Research framework of step two is to create the transition matrix of Markov model P, which is based on web log file as well as similarity matrix. The P matrix is first-order transition matrix of Markov model and it is presented as follows:

$$P = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} \end{array}$$

where $P_{ij} = \dfrac{Number(i, j)}{\sum_{j=1}^{k} TotalNumber(i, j)}$. (11)

Each element in the P matrix presents a transition probability between any two categories. P_{ij} presents a transition probability which is calculated by equation (11) between category i and category j. The numerator of equation (11) is the number of transition times between category i and category j, and the denominator is the total number of transition times between category i and every category k. The transition matrix of $P^2,...,P^n$ can be calculated by equation (3).

3) Relevance matrix

Research framework of step three is to create the relevance matrix. The element R_{ij} of relevance matrix is equal to product of S_{ij} and P_{ij}, which are acquired from similarity matrix and transition matrix of Markov model respectively. In this paper, the relevance is an important factor of prediction between any two categories. The relevance can be used to infer the users' browsing behavior between categories. The relevance matrix is presented as follows:

$$R^n = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} R^n_{11} & R^n_{12} & \cdots & R^n_{1k} \\ R^n_{12} & R^n_{22} & \cdots & R^n_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ R^n_{k1} & R^n_{k2} & \cdots & R^n_{kk} \end{bmatrix} \end{array}$$

where $R^n_{ij} = S_{ij} \cdot P^n_{ij}$. (12)

R^n_{ij} presents a relevance, which is calculated by equation (12), between category i at time $t-n$ and category j at time t; More high the value of $R^n{}_{ij}$ means more relevance between category i and category j.

B. Two levels of prediction model

In the steps of prediction, the first-order and the second-order ($n = 2$) transition matrix of Markov model and relevance matrix are used to predict users' browsing behavior. In this paper, the establishment of prediction (as Fig. 5) is to predict the position C which depends on the position B and A, now and past respectively. The two levels of predictions model are proposed in the paper. The category is predicted by level one and web page is predicted by level two.

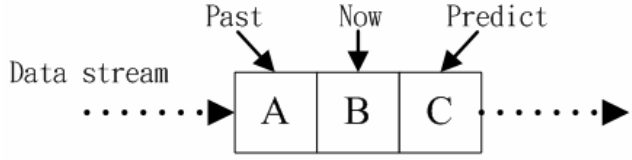

Figure5. Schematic Diagram of Prediction

1) Level one: predicting category

The level one prediction is try to find out the most possible category C_t, that the previous one state is category C_{t-1} or the previous two state is category C_{t-2}. The θ is a set of categories which are obtained by level one. After θ is obtained, only the web pages belong to category θ will be considered in level two instead of all. Thus, it will reduce the scope of the level two predictions. $R^n_{C_{t-n}}$ is the row vector for category C_{t-n}, that is $[R^n_{C_{t-n,1}}, R^n_{C_{t-n,2}}, \cdots, R^n_{C_{t-n},k}]$, from relevant matrix R^n, where we choose $n=1$ or $n=2$. The category $C_t \in \theta$ that are the corresponding categories which has top r value in the vector $R^1_{C_{t-1}}$ or $R^2_{C_{t-2}}$. θ is defined as follows:

$\theta = \{C_t \mid C_t$ is the corresponding category that has top r value in the vector $R^1_{C_{t-1}}$ or $R^2_{C_{t-2}}\}$

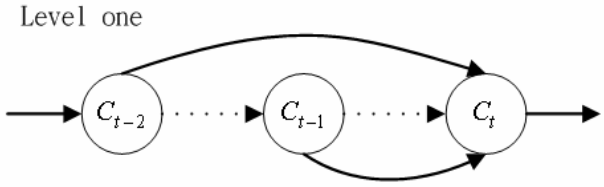

Figure6. Level One of Categories Prediction

2) Level two: predicting web pages

In the level two, Bayesian theorem is used to calculate probability that candidate $page_b$ will be the successor (as Fig. 7). τ are the results of prediction in the level two, which are most r possible web pages that the users will visit.

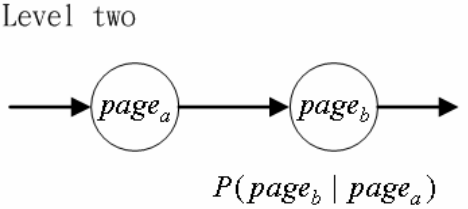

Figure7. Level Two of Pages Prediction

$page_a$ is the current visiting page and $page_b$ is the candidate for the next user's visiting web pages and limited by θ, that is $page_b \in$ Category θ, where θ is the results of level one prediction. It is assume that $P_\theta = \{page_{b_1}, Page_{b_2}, \cdots Page_{b_r}\}$ is a set of all Pages in category θ.

$$P(page_{b_i} \mid page_a) = \frac{P(page_a \mid page_{b_i})P(page_{b_i})}{\sum_{j=1}^{r} P(page_a \mid page_{b_j})P(page_{b_j})} \quad (13)$$

$\tau = \{Page_{b_i} \mid$The value of $P(page_{b_i} \mid page_a)$ is top r in set $P_\theta\}$

IV. Experiment

The source of database in the experiment is from UCI Dataset Repository [17]. This repository contains a lot of different research fields. Therefore, this repository is very popular in the research field. The dated of the database is September, 28, 1999 on msnbc.com website and a part of news data is from msn.com website. Each sequence data is corresponding to users' page views. The time interval is 24 hours. The categories are presented in sequential data. The 17 types of categories are frontpage, news, tech, local, opinion, on-air, music, weather, health, living, business, sports, summary, bbs, travel, msn-news, and msn-sports. The number of category is represented from 1 to 17 in the sequential data.

The UCI Dataset Repository is a quite popular database, but the users' browsing behavior is just recorded by category in the database, the user's web pages visiting are simulated by Gaussian distribution. The numbers of web pages in each category are assumed equal.

The experiment environment in this paper is HP dx5150. The hardware is performed on AMD Athlon™ 64 X2 Dual Core Processor 3800+ CPU, 1GB RAM and Microsoft Windows XP operating system. The software is performed on Java 1.6.0 Update 2.

Hit Ratio is calculated by equation (14) [11]. Hit_{ratio} is a ratio of cache access to $request_{all}$. $cache_{access}$ is a number of the web page is found in the prediction. $request_{all}$ is total number of the web pages which are browsed by users.

$$Hit_{ratio} = \frac{cache_{access}}{request_{all}} \quad (14)$$

The users' browsing record is obtained and divided into training data and testing data. Training data is 80% which are 791,855 and testing data is 20% which are 197,963 of the web log file. The training data is used to establish the two levels of prediction model and the testing data is used to calculate in Hit Ratio.

There are five cases in level one prediction, which are Top-1, Top-2, Top-3, Top-4 and Top-5. The Hit Ratio is from 66.74% on top-1 relevance to 89.80% on Top-5 relevance (as Fig. 8).

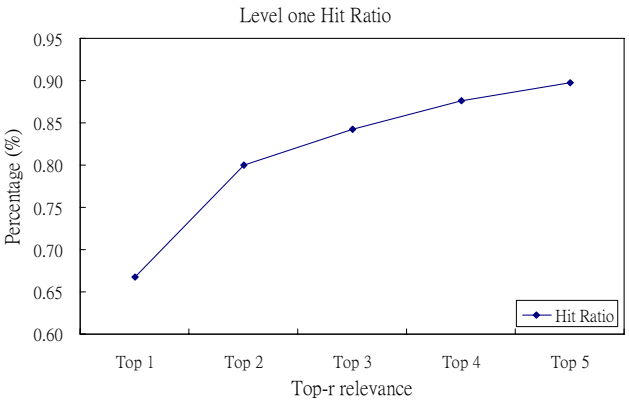

Figure8. Hit Ratio of Level One

In the experiments for level two, the number of web pages number is 40, 50 and 60 respectively. There are three cases for the prediction result, which are Top-10, Top-15, Top-20. The results are shown in the Fig. 9. For example, the Hit Ratio is from 74.71% (Top-10) to 94.30% (Top-20) when the number of web pages is 40.

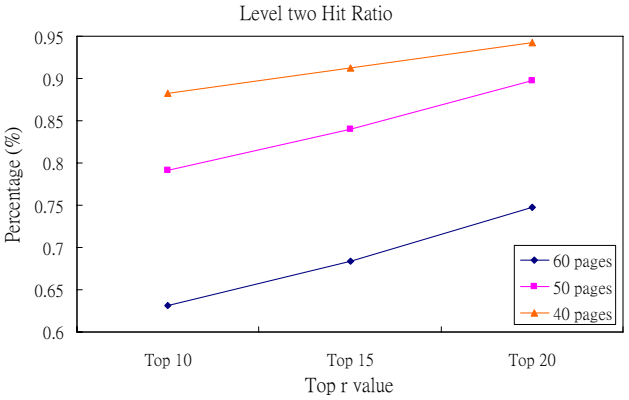

Figure9. Hit Ratio of Level Two

V. Conclusion and Future Work

Because of the huge quantity of data of web pages on many portal sites, for convenience, are to assemble the web page based on category. In this paper, users' browsing behavior will be observed at two levels to meet the nature of the portal. One is category level and the other is web page level. In level one is to predict category. The unnecessary categories can be excluded. The scope of calculation is massively reduced. Next, using Bayesian theorem in the level two to predict the users' browsing page is more effective and accurate. The results of experiment prove the Hit Ratio is well in both levels.

References

[1] S. Araya, M. Silva and R. Weber, "A Methodology for Web Usage Mining and Its Application to Target Group Identification," *Fuzzy Sets and Systems 148*, pp. 139-152, 2004.

[2] W. Bin and L. Zhijing, "Web Mining Research," *ICCIMA'03 IEEE*, pp. 84-89, 2003.

[3] S.K. De and P.R. Krishna, "Clustering web transactions using rough approximation," *Fuzzy Sets and Systems 148*, pp. 131-138, 2004.

[4] D. Dhyani, W. K. Ng and , S. S. Bhowmick, "A Survey of Web Metrics," *ACM Computing Surveys*, Vol. 34 2002.

[5] D. Dhyani, S. S. Bhowmick and W. K. Ng, "Modeling and Predicting Web Page Accesses Using Markov Processes," *DEXA'03 IEEE*, 2003.

[6] F. M. Facca and P. Luca Lanzi, "Mining Interesting Knowledge from Weblogs: A Survey," *Data and Knowledge Engineering 53*, pp. 225-241, 2005.

[7] R. Kosala, H. Blockeel, "Web Mining Research: A Survey," *SIGKDD Explorations*, Volume 2, Issue 2, pp. 115.
[8] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, Vol. 31, No. 3, 1999.
[9] S. Jespersen, T. B. Pedersen and J. Thorhauge, "Evaluating the Markov Assumption for Web Usage Mining," *WIDM'03*, pp.82-89, 2003.
[10] G. Pallis, L. Angelis, A. Vakali, "Vaildation and interpretation of Web users' sessions clusters," *Information Processing and Management 43*, pp. 1348-1367, 2007.
[11] L. Shi, Z. M. Gu, L. Wei and Y. Shi "Popularity-based Selective Markov Model," *WI'04*, pp.504-507, 2004.
[12] J. Srivastava, R. Cooley, Mukund Deshpande and Pang-Ning Tan, "Web Usage Mining:Discovery and Applications of Usage Patterns form Web Data," *SIGKDD Explorations*, Vol. 1, Issue 2, pp. 12-23, Jan 2000.
[13] R. R. Sarukkai, "Link prediction and path analysis using Markov chains," *Computer Networks*, pp. 377-386, 2000.
[14] R. M. Suresh and R. Padmajavalli," An Overview of Data Preprocessing in Data and Web Usage Mining," *Digital Information Management IEEE*, pp. 193-198, 2006.
[15] R. Walpole, R. Myers, S. Myers and K. Ye, "Probability and Statistics for Engineers and Scientists ," in *Paperback*, 7 ed., Pearson Education, 2002, pp.82-87.
[16] W3C Extended Log File , http://www.w3.org/TR/WD-logfile.html.
[17] UCI KDD archive, http://kdd.ics.uci.edu/.