Web Usage Patterns Using Association Rules and Markov Chains

Amnat Sawatnatee
Chandrakasem Rajabhat University, Thailand
amnats.cru@gmail.com

***Abstract* - The objective of this research is to illustrate the probability of web page using at a period of time using two statistical techniques. SME Nonthaburi province handicraft goods e-commerce web site was selected as a case study in this research. Web pages were categorized into three portions: SME firm section's News, Goods details and Customer activities. Markov chain technique was applied in order to present probability of each event. Association rule technique was also used to derive the paths of web page visiting. This complementary results from two techniques should support web administrator to spot what web pages are the interested web pages of his web site architecture. Association rules show the path of consequence visited web page while Markov chains give and information of possibility of events that should be visited in specific assigned period of time.**

***Keywords* - Web Mining, Markov Chains, Association Rule**

I. INTRODUCTION

A website administrator (also known as admin) has responsibilities to create, maintain and fulfill the important data and information of organization in order to support the enterprise mission. Some contents are interested but another one are bored. The contents that are mostly simultaneous visiting may cause the downgrade of system performance. Administrator must deeply analyze the software design, database design, server or even network bandwidth adjustment in order to address these problems. In case of less visited web page, administrator has to modify, increase some contents or deleted them. This research gathered web usage from sys-log database. The e-commerce website was used as an experiment.

The content in website was categorized into three types such as SME firm section's News (event A), Goods details (event B) and Customer activities (event C). The 120 web usage data observations were gathered during 10-31January 2018. These observations were used to find out their patterns of web page visiting by using Association rule technique.

Meanwhile, Markov chains were also calculated in order to obtain the probability of events in defined period of time. From the results, both techniques provide information about web usage patterns and possibility of occurring so that web site administrator can use these in content management.

II. RELATED THEORY AND RESEARCH

A. Related Research

1) Association rule technique was applied in web usage mining. Observations were gathered from web usage log file of web page VTSNS, the "Advanced School of Technology Novi Sad Serbia, web site". The experiment was repetitive pruning huge received rules by setting value of "support" and "confidence". The derived rules were used in web site map (architecture) modification.

2) Two web browsers, under exposed, that were used and experimented in the server of "Department of Mathematics and Statistics, Sagar University" that which one was most popularity and appreciate. Markov chains model was used to illustrate probability of their two browsers usage after passing "*t*" time unit

after start time "$t=0$". The less preferred browser, small value probability of existence, was inspected for its problems, performance, etc. After this less popular web browser was modified about undesired features, such as adjust-add on-re configuration, Markov chains was then re processing. The modified web browser was become more increase in probability of state existence than before.

B. Association Rule [3]

Association rule is a statistical technique that used to find out the dependency of attributes. For example, if attribute "A" is occurred while attribute "B" is also occurred then it can define the rule as "$A \rightarrow B$". The criteria of decision making from the rule can be accepted are two metrics: "support" and "confidence" that can be computed by (1) and (2).

$$\sup port(A \rightarrow B) = (A \cap B) \tag{1}$$

$$confidence\,(A \rightarrow B) = \frac{n(A \cap B)}{n(A)} \tag{2}$$

where

$n(A \cap B)$ is the number of observations that "A" (source) and "B" (destination) are both occurred.

$n(A)$ is total observations that state "A" is presence.

Normally, a support value is calculated from total data set therefore this value is used to find out the paths have the frequently happening of user's web site traveling. If the support value is too high then there may a few rules are discovered. On the other hand, there may get many rules if the support value is too small. For more detail, the confidence value covers the total observations that attribute "A", or source "B" attributes, is occurred.

Therefore, the confidence value presents the probability of specific rules that the happening of destination attributes when the source attributes of observations are totally occurred. If this value is high, it means that these rules have more happening in case of the specific occurring source attribute observations dataset.

C. Markov Chains [4]

"Markov chains" is a statistical technique that is used to calculate for the probability of transition of two events between two periods in a time. There should have more states in the studying problems. Thus, the specific state could travel to another state, or even itself, under prior probability of transition matrix (P). At any time period passed from start point, the probability of all states could be calculated from the system of first order difference equation.

Let "Π_0" is a "$n \times 1$" size of vector that describes the possibility of all states being at time $t=0$. "Π_t" describes the possibility of all states being at the time t. The "Π_t" should be calculated by (3).

$$\Pi_t = P\Pi_{t-1} \tag{3}$$

"Π_0" is an initial possibility vector of all states. It is used to derive the value of all parameters in a system of the first order difference equation that is used to find out all possible events at a period of time without equation (3) in consequence from t = 0, 1, 2, 3, t-1. Thus, the first order difference equation provides more comfortable in computing than typical processing.

III. RESEARCH METHODOLOGY

A. Data Preparation

- **Data Source:** A number of 120 records of the observation were collected from SME Nonthaburi Province handicraft goods e-commerce Website. Website map is composed of three menus. The first portion explains about the mission of private enterprise. The second portion is an important menu since it gives information about enterprise's goods. And the third portion information and activities of order, payment and goods receive. Each menu is composed of many sub menus thus this research considered only in three groups in

order to reduce computational complexity.

- **Web Usage Log:** Our research case studied web site application has designed a database that was used to keep all users' actions at the choosing menu from start event, or state (invocate), and other traveling states until they logoff the website.

TABLE I
DATA DETAIL OF WEB USAGE LOG DATABASE

Attributes	Description	Data Type
IP address	IP address of external user	999.999.999.999
Date	Day: month: year	99:99:99
Time	Hour: min	99:99
Menu-chosen	Menu type S=start, L=logoff, A=Firm's news, B=Goods detail, C=Customer activity.	S, A, B, C, L

Note: Data collection period: Web usage logs were gathered during 1-30 January 2018.

B. Association Rules

- From web usage log data base, each user's data, observation, were coding and cleaning before further data processing step.

Objective of data preparation was to present the absence and presence of event, or state, during period of time, for an example, five observations were shown in Table II. If the user selects any menu (S, A, B, C, L) then chosen menu wascoded as "1" (presence), The "0" was done if the absence.

TABLE II
PARTIAL DATA ABOUT USER'S MENU SELECTION

Observation#	Start	A	B	C	Logoff
1	1	1	0	0	1
2	1	1	1	1	1
3	1	1	0	1	1
4	1	0	1	0	1
5	1	0	0	1	1

Some combination of paths do not happen such as $L \to A, A \to S$ Thus this kind of path must be deleted.

The amount of all possible paths, or rule, in an "Association rule" is shown in (4).

$$\#ofPossibleRule = 3^d - 2^{d+1} + 1 \quad (4)$$

While "d" is an amount of event, or attribute, an interest experiment. For example, if there are 3 events as: A - B - C, then there are 12 possible rules.

$$A \to B,\ A \to C,\ A \to B,C$$
$$B \to C,\ B \to A,C$$
$$C \to B,\ C \to A,B$$
$$A,B \to C,\ A,C \to B,\ B,C \to A$$

In this research, there were five events (d=5) then there were one hundred possible association rules. Some of these rules were measured by "support" and "confidence" metrics while some paths did not pass. In order to address the problem of numerous rules about important and less important contents. Thus, the support and confidence values should be the high score. In this research, the acceptable support value was set to "0.3" and the confidence value was set to "0.5".

- There were four discovered rules that passed both criteria (descending order) as shown in Table III.

The mostly happen path is $S \to A \to L$, support = 0.33. Many customers start visit this website, read News then log off. The enterprise goods might not be interested so that they suddenly leave this web site.

TABLE III
USER'S MOSTLY OCCURRED PATHS

#	Support	Confidence	Rule or Path
1	0.33	0.56	$S \to A \to L$
2	0.31	0.52	$S \to B \to L$
3	0.30	0.51	$S \to B \to C \to L$

C. Markov Chain

- All data of user's menu selection, during they spent their time in the website, were summarized in working table, as shown in Table IV). For example - according to Table

IV, after login to the website, the 30% of users (p=0.30) chose to visit menu "A". And the logoff was chosen at 10%. In case of "A" state, after "A" was chosen, 10% of user still chose to stay in "A" and the 30% chose to travel to B, C, and L (logoff).

The experimental transition probabilities matrix (P) detail is shown in Table IV.

TABLE IV
SUMMARY PROBABILITY
OF USERS' BEHAVIORS

Start	A	B	C	Logoff	
0.0	0.0	0.0	0.0	0.2	Start
0.3	0.1	0.2	0.4	0.0	A
0.4	0.3	0.5	0.3	0.0	B
0.2	0.3	0.1	0.2	0.0	C
0.1	0.3	0.2	0.1	0.8	Logoff

According to transition probability matrix, there were connecting path between some node, or state, with other nodes or even its self while some connecting paths were absence since there were not transition probability between them. The whole relation paths were illustrated, as shown in Fig. 1.

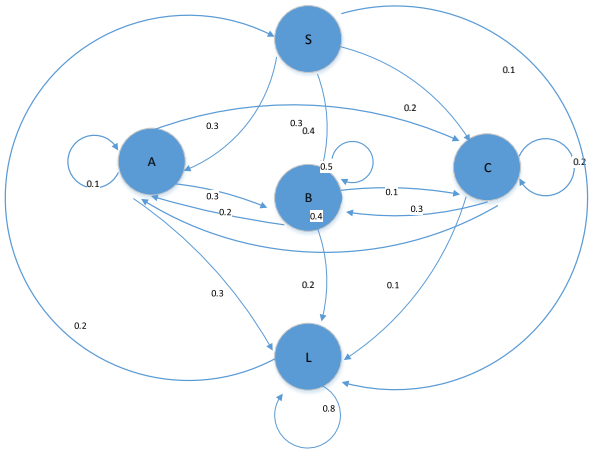

Fig. 1 Markov Model of Transition Probability Matrix (P)

- From Table IV, data was called "transition probability Matrix (P)". This matrix presents an explanation about users' behaviors about the web site's menu choosing. Markov chains method was then applied to illustrate the model of all probably events (or state) in a specific time period. Markov chains are cumbersome in calculation cause the next probability of interested event is depend on prior event probability thus these can made simplified by another technique, the system of first order difference equation method. After the data preparation was finished, it was summarized that initial state (Π_0) has probability column vector as shown in (5), the probability event vector at time "t" as shown in (6) and the experimental transition probability matrix (P) as shown in (7).

$$\Pi_0 = \begin{bmatrix} 0 \\ 0.2 \\ 0.2 \\ 0.5 \\ 0.1 \end{bmatrix} \tag{5}$$

$$\Pi_t = \begin{bmatrix} S_t \\ A_t \\ B_t \\ C_t \\ L_t \end{bmatrix} \tag{6}$$

$$P = \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.2 \\ 0.3 & 0.1 & 0.2 & 0.4 & 0.0 \\ 0.4 & 0.3 & 0.5 & 0.3 & 0.0 \\ 0.2 & 0.3 & 0.1 & 0.2 & 0.0 \\ 0.1 & 0.3 & 0.2 & 0.1 & 0.8 \end{bmatrix} \tag{7}$$

After a long trial of mathematical calculation, the system of first order difference equation for all events at time period "t" was computed then all event's probability or Π_t vector was presented in equation (8), (9), (10), (11), and (12).

$$S_t = 0.25 - 0.16(0.56)^t - 0.03(-0.20)^t + 0.07(0.02)^t + 0.02(0.21)^t \tag{8}$$

$$A_t = 0.12 + 0.15(0.56)^t - 0.15(-0.20)^t - 0.01(0.02)^t + 0.11(0.21)^t \tag{9}$$

$$B_t = 0.21 + 0.33(0.56)^t + 0.03(-0.20)^t - 0.03(0.02)^t - 0.34(0.21)^t \tag{10}$$

$$C_t = -0.09 + 0.12(0.56)^t + 0.12(-0.20)^t - 0.04(0.02)^t + 0.18(0.21)^t \tag{11}$$

$$L_t = 0.48 - 0.44(0.56)^t + 0.03(-0.20)^t - 0.01(0.02)^t + 0.02(0.21)^t \tag{12}$$

IV. RESEARCH SUMMARY AND SUGGESTION

A. Summary

The proposed research techniques in this research could present the co-occurrence among events under the arbitrary defined dependency level such as "support" and "confidence". Website administrator can choose the value of rules as well as that whether result rules should be sufficient to explain the customers' behaviors or not.

According to the calculation results, Markov chains can present the probability of particular events. While the association rules give all possible consequence paths that are related to the occurred events but no any occurrence of probability about each event. Significance association rule path informs web site owner about user or customers' behaviors. Path #2 and #3 present how often users visit website from menu "A", "B" then go to "L". Some customers visit the events "B" and "C" then go to "L". In practice, if there is an amount of total customers "N" that count the number of website visiting at a period of time, such as the first day of any month, then the roughly possible amount of customers "M" that should follow the path #3 that could predicted by (13).

$$M = N * B_t * C_t \tag{13}$$

The main objective of e-commerce web site is to provide the interesting goods to huge number of customers so that many potential visitors decide to buy web site's goods. Therefore, the percentage of web site success could be calculated from (14).

$$SuccessPercentage = \frac{M}{N} * 100\% \tag{14}$$

B. Further Research

Based on techniques that were used here, all significant event dependences might (or not) meet the statistical significance criteria ($\alpha 0.05$) since they give no any information about type I error. In alternative, Bayesian theorem under joint probability density function, Information theory, Causal model analysis and etc should be considered as other possible techniques that can be used to solve this problem.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Dimitrijević, M. (2011). "Web Usage Association Rule Mining System". Interdisciplinary Journal of Information, Knowledge, and Management, Vol. 6.

[2] Shukla, D. (2011). "Analysis of Users Web Browsing Behavior Using Markov chain Model". Department of Mathematics and Statistics, Sagar University, Sagar M.P., 470003, India.

[3] Kumar, V. (2005). "Introduction to Data Mining". Pang-Ning Tan, Michael Steinbach, Vipin Kumar Addison-Wesley, ISBN: 0321321367.

[4] Fewster. "Markov chains". Auckland University, New Zealand.