International Journal of Engineering & Technology, 7 (2.12) (2018) 171-173

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

A study on website log data analysis methodology by transition probability

Jae-Kyeong Lee [1], Mi-Hwan Hyun [2], Dong-Gu Shin [3 *]

[1] Author's Position, Korea Institute of Science and Technology Information, 02456, Republic of Korea
[2] Korea Institute of Science and Technology Information, 02456, Republic of Korea
[3] Corresponding author, Korea Institute of Science and Technology Information, 02456, Republic of Korea
**Corresponding author E-mailgngee@kisti.re.kr:*

Abstract

Background/Objectives: To measure occupancy using transition probability matrix as a data analysis method to predict future requirements for web use. From this study, Executives facing business challenges can enhance the decision-making process for management and can be provided quantified evidence.
Methods/Statistical analysis: Transition matrix and transition probability matrix are estimated if web users' webpage use patterns are tied with frequency, using web log data. Occupancy is forecasted based on a Markov chain model.
Findings: Data analysis from the perspective of web log-based marketing mostly focuses on increasing traffic and improving transition rates. However, general-purpose tools such as Google Analytics provide diverse web log data. In assumption of independence on users' page reload, occupancy can be easily estimated through matrix on page reload (transition). As a result, we obtained slightly different results from the usual method that reported only frequency. In particular, rather than making business decisions with the frequency of absolute concepts, we were able to identify the top priority services through the percentage value of relative concepts.
Improvements/Applications: The occupancy prediction using transition matrix is about future prediction based on previous information. However, it differs from marketing techniques in that it is estimated based on probability. In addition, it is able to predict more accurately through a probability model.

Keywords: *Weblog; Markov Chain; Transition Matrix; Probability Transition Matrix; Steady State*

1. Introduction

In the analysis of web use information data, users' behavior patterns are analyzed by applying general data-mining technique and methodology from marketing perspective to the website [5] [6]. Then, individual- or group-customized services are provided [3] [4] [8]. After that, the growth and development of user environments are pursued by monitoring traffic and diagnosing problems just like interactions between the web operator and users. The conventional web log analysis techniques are mostly just a diagnosis on past issues as a performance-centered typology using numerical data. Therefore, it is hard to find any clue on future operational decisions. This study proposes more accurate prediction or forecast, using a Markov chain model after expressing singular values on page reload among web log data in matrix and implementing transition probability matrix [2]. These predicted outcomes could play a key role in decision-making and provide optimized services to users [7].

2. Proposed work

In particular, predicting the requirements using the results when a user visited a website is a very important factor in terms of information provision and communication with the service provider. Regarding analysis of user behavior patterns, clustering by service session, creation of association rules and network analysis are usually used. In Markov chain model-based analysis, the page change by a user's click is defined as an independent variable incident before and after a certain point in time. In this study, among the dimensions provided by Google Analytics, PagePath values were used. For the measurement of page variables, pageview was referred to for metric. Depending on analysis purposes, association and migration-deemed services were sorted and grouped. If n-by-n matrix is expressed in pageview according to the number of the classified services, it is defined as transition matrix. In this matrix, then, 'row' and 'column' refer to point in time before and after change respectively. Since row sums are a sum of the position before the change, therefore, transition matrix probability can be completed through division with this value. In other words, the assumption that the status before the change independently determines current conditions regardless of the past situation according to Markov attribute with the application of conditional probability distribution is accepted [9] [10].
A discrete-time Markov chain is a sequence of random variablesX_nwith markov property as known as probability of moving to the next state depends only on the present state and not on the previous states. Markov chains are often described by a sequence of directed graphs, where the edges of graph n are labeled by the probabilities of going from one state at time 'n' to the other states at time 'n+1'. The same information is represented by the transition matrix from time n to time 'n+1'. However, Markov chains are frequently assumed to be time-homogeneous, in which case the graph and matrix are independent of 'n' and are thus not presented as sequences.
Markov chains are processes where

Copyright © 2018 Jae-Kyeong Lee et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

$\Pr(X_n = x_n | X_{n-1} = x_{n-1})$

For positive integer 'n'.
The probability of going from state i to state j in 'n' time steps is

$p_{n,ij} = \Pr(X_n = j \mid X_0 = i)$

And the single-step transition is

$p_{ij} = \Pr(X_1 = j \mid X_0 = i)$

Fora time-homogeneous Markov chain is

$p_{n,ij} = \Pr(X_{k+n} = j \mid X_k = i)$

$p_{ij} = \Pr(X_{k+1} = j \mid X_k = i)$

The n-step transition probabilities satisfy the Chapman-Kolmogorov equation, that for any k such that 0< k <n,

$p_{n,ij} \sum_{t \in S} p_{k,it}\, p_{n-k,tj}$

Where S is the state space of the Markov chain.
The Marginal distirution $\Pr(X_n = x)$ is the distribution over states at time 'n'. The initial distribution is $\Pr(X_0 = x)$. the evolution of the process through one time stepis described by

$\Pr(X_n = j) = \sum_{t \in S} p_{tj} \Pr(X_{n-1} = t) = \sum_{t \in S} p_{n,tj} \Pr(X_0 = t)$

In this study, the web log data of an online idea platform [1] was adopted. About 4 pageviews were observed per person (session). As stated in the equation above, 'n' is confirmed in most data from '1' to '4'. However, it should include page landing for the first time. Since there is no limitation in 'n', 'n' would actually be infinite from '0'. Such data are divided by 'n' and 'n-1', and the frequency in which the point in time is converted in one on one is expressed in matrix. Then, the following transition matrix is developed.

Table 1: Pageview Transition Matrix

	Exit	Main	Idea community	Mentoring ~	Search	Etc.	Sum
Entrance	-	180,235	45,893	13,873	345	56,501	636,712
Main	97,452	47,280	37,904	26,032	3,923	11,763	470,915
Idea community	57,675	23,431	376,712	5,001	295	1,482	494,821
Mentoring ~	34,711	21,388	5,845	259,337	652	6,933	380,723
Search	1,439	1,335	188	813	8,380	1,093	18,652
Etc.	46,014	8,802	2,199	4,380	735	120,505	212,773
Sum	629,002	470,236	495,986	382,784	18,341	213,778	3,955,419

If idea plant services are divided into 14 categories including entrance and exit, all transitions can be expressed. According to Table 1 above, the transition from entrance to exit has no frequency because no session is created. A diagonal matrix component refers to page transition or browser refresh in the service. In other words, in case of ideacommunity, the service is very crowded, which means low transition to other services. With a 1-dimension review on frequency only, it is able to get multiple insights on operation. If the transition matrix is converted into probability at 'n-1', the values would appear as follows:

Table 2: Transition Probability Matrix of Pageview

	Exit	Main	Idea community	Mentoring ~	Search	Etc.	Sum
Entrance	0.000	0.283	0.072	0.022	0.001	0.089	1
Main	0.207	0.100	0.080	0.055	0.008	0.025	1
Idea community	0.117	0.047	0.761	0.010	0.001	0.003	1
Mentoring ~	0.091	0.056	0.015	0.681	0.002	0.018	1
Search	0.077	0.072	0.010	0.044	0.449	0.059	1
Etc.	0.216	0.041	0.010	0.021	0.003	0.566	1

Table 2 reveals the division of the sum of rows in a transition matrix. A concept of standardization was applied. This probability matrix has the meaning equivalent to the frequency in Table 1. However, it can provide additional insight as a relative concept. The probability matrix was calculated for a more advanced concept. The repetitive self-multiplication of probability matrix reaches a steady state with the application of limit theory to transition actions. Then, steady state probability is calculated.
If the Markov chain is a time-homogeneous Markov chain, so that the process is described by a single, time-independent matrix pij, then the vector π is called a stationary distribution(or invariant measure) if$\forall j \in S$.

$0 \leq \pi_j \leq 1.$

$\sum_{j \in S} \pi_j = 1$

$\pi_j = \sum_{j \in S} \pi_i p_{ij} = 1$

An irreducible chain has a positive stationary distribution if and only if all of its states are positive recurrent.In that case, π is unique and is related to the expected return time:

$\pi_j = \frac{C}{M_j}$

Where C is the normalizing constant. Further, if the positive recurrent chain is both irreducible and aperiodic, it is said to have a limiting distribution
For any i and j,

$\lim_{n \to \infty} p_{n,ij} = \frac{C}{M_j}$

If a state i is periodic with period k > 1 then the limit

$\lim_{n \to \infty} p_{n,ii}$

Does not exist, although the limit

$\lim_{n \to \infty} p_{kn+r,ii}$

Does exist for every integer r.
A Markov chain need not necessarily be time-homogeneous to have an equilibrium distribution. If there is a probability distribution over states πsuch that

$\pi_j = \sum_{i \in S} \pi_i \Pr(X_{n+1} = j \mid X_n = i)$

for every state j and every time n then πis an equilibrium distribution of the Markov chain. Such can occur in Markov chain Monte Carlo (MCMC) methods in situations where a number of different transition matrices are used, because each is efficient for a particular kind of mixing, but each matrix respects a shared equilibrium distribution.

$$\pi = \{\pi_1, \pi_2, \pi_3, \dots, \pi_n\}$$

$$\pi = \pi \times \mathrm{P}$$

$$\pi = \pi \times 1 = 1$$

The probability vector 'π' refers to steady state probability. The multiplication of transition action means the repetition of acts. In other words, future acts are also independently included. According to Markov chain theory, therefore, steady state probability means future occupancy. Steady state probability can be exponentiated until the transition probability matrix 'P' is converged, or it could be estimated, using singular value decomposition (SVD) under linear algebra. In this study, it was calculated by multiplication with unlimited at limited. The calculation would appear as follows:

Table 3: Steady State Probability of Pagevie

	Exit	Main	Idea-community	Mentoring	~	Search	Etc.	Sum
steady state	0.159	0.118	0.126	0.098	~	0.005	0.054	1.000

3. Conclusion

The prediction of future occupancy using a transition matrix is useful in diverse fields. Since the matrix is divided into two parts (before, after), the overall aspect can be understood with a section. With the understanding of numbers only, it is able to check if services are properly operated. If economic theories are applied, in addition, elasticity and sensitivity can be measured through a correlation model. Furthermore, using the probability-based Markov chain model, a future state (occupancy) can be predicted. Therefore, it would be helpful in planning operation. However, the results can slightly differ depending on the data editing standards of the web log used in developing a transition matrix. In this study, dimension and metric provided by Google Analytics were used. However, the results can vary depending on page path setting conditions, use of pageview and metric count on entrance and exit. Therefore, it needs to be cautious in setting data depending on service characteristics and analysis purposes.

References

[1] Online Creative Economy Town, South Korea online idea commercialization platform, www.creativekorea.or.kr

[2] Y.H. Kim, U.M. Kim, M.S. Jung, W.J. Kang, A Web Usage Predicton model by Transition Probability Matrix, Communications of the Korean Institute of Information Scientists and Engineers, Vol. 31, No.2, pp. 31.

[3] D.S. Kang, A Study on the Determinants of User Satisfaction of e-Government Services, skku, 2009

[4] American Customer Satisfaction Index, www.theacsi.org

[5] Robert Cooley, Pang-Ning Tan, and JaideepSrivastave. Discovery of interesting usage patterns from web data. Technical Report TR 99-022, University of Minnesota, 1999.

[6] V.N. Padmanabham and J.C.Mogul. Using predictive prefetching to improve world wide web latency. Computer Communication Review, 199

[7] Ramesh R. Sarukkai. Link prediction and path analysis using markov chains. In Ninth International World Wide Web Conference, 200I

[8] gor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. Vidualization of navigational patterns on web site using model based clustering. Technical Report MSR-TR-0018, Microsoft Research, Microsoft Corporation, 2000.

[9] Mukund Deshpande and George Karypis. Selective Markov Models for Predicting Web-Page Accesses, ACM Transactions on Internet Technology, Volume 4, Isuue 2, May 2004.

[10] Jianhan Zhu, Jun Hong, and John G. Hughes, Using Markov Chains for Link Prediction in Adaptive Web Sites, Springer-Verlag Berlin Heidelberg, 2002.