Association for Information Systems
AIS Electronic Library (AISeL)

Proceedings of the 2016 Pre-ICIS SIGDSA/IFIP WG8.3 Symposium: Innovations in Data Analytics

Special Interest Group on Decision Support and Analytics (SIGDSA)

12-11-2016

Predictive Analytics with Sequence-based Clustering and Markov Chain

Sungjune Park
University of North Carolina at Charlotte, supark@uncc.edu

Vinay Vasudev
University of North Carolina at Charlotte, vkvasude@uncc.edu

Follow this and additional works at: http://aisel.aisnet.org/sigdsa2016

Recommended Citation
Park, Sungjune and Vasudev, Vinay, "Predictive Analytics with Sequence-based Clustering and Markov Chain" (2016). *Proceedings of the 2016 Pre-ICIS SIGDSA/IFIP WG8.3 Symposium: Innovations in Data Analytics*. 12.
http://aisel.aisnet.org/sigdsa2016/12

This material is brought to you by the Special Interest Group on Decision Support and Analytics (SIGDSA) at AIS Electronic Library (AISeL). It has been accepted for inclusion in Proceedings of the 2016 Pre-ICIS SIGDSA/IFIP WG8.3 Symposium: Innovations in Data Analytics by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Predictive Analytics with Sequence-based Clustering and Markov Chain

Research-in-Progress

Sungjune Park
Department of Business Information Systems and Operations Management
The Belk College of Business
UNC Charlotte
supark@uncc.edu

Vinay Vasudev
Department of Business Information Systems and Operation Management
The Belk College of Business
UNC Charlotte
vkvasude@uncc.edu

Abstract

This research proposes a predictive modeling framework for Web user behavior with Web usage mining (WUM). The proposed predictive model utilizes sequence-based clustering, in order to group Web users into clusters with similar Web browsing behavior and Markov chains, in order to model Web users' Web navigation behavior. This research will also provide a performance evaluation framework and suggest WUM systems that can improve advertisement placement and target marketing in a Web site.

Keywords: Predictive analytics, sequence-based clustering, Markov chain

Introduction

Within the past two decades, many organizations have begun implementing value-added services on the Web to gain competitive advantages by attracting loyal customers and providing targeted messages. In order to make the Web more user-friendly for individuals and create long-term relationships with them, companies now realize that providing personalized web services is crucial. In addition, online advertising has become major source of revenue for many business organizations with large websites with heavy user traffic. Web usage mining (WUM) allows extraction of knowledge about such navigation patterns, identifies targeted smart online advertising, and thus potentially leads to better Web experiences for the users (Ho et al. 2011). WUM analyzes the data generated by the Web users' interactions with the Web including Web server access logs, user queries, and mouse-clicks, in order to extract patterns and trends in Web users' behaviors. A growing interest in the business use of 'intelligent' Web, also known as, Web 3.0, and social networking sites accentuate the importance of utilizing such patterns and trends for the purpose of creating effective marketing tools as well as enhancing user experiences on the Web.

This research proposes a predictive modeling framework for Web user behavior with Web usage mining. The proposed predictive model utilizes a sequence-based clustering approach, in order to group Web users into clusters with similar Web browsing behavior, and Markov chains (MC), in order to model Web users' navigation behavior. Clustering Web users aims at facilitating the prediction of Web users' navigation behavior and its resulting transition state while the users are browsing within a website (Park et al. 2008). Sequence-based clustering enables full consideration of sequential activities on the Web such as page visits or content views, which is a significant improvement over the usual practice of considering the frequency of visits to web pages (Kim 2007; Park et al. 2008; Hung et al. 2013). The Markov model has also been shown to be effective in predicting Web user's sequential navigation patterns. The use of MC allows calculation of transition probabilities at any given time of active user sessions, and along with sequence-based clustering we expect that this will lead to a higher accuracy of prediction of Web users'

next steps. The ability to predict more accurately will lead to a better Web personalization and a more effective online advertising outcome.

Therefore, the main objectives of this research are 1) to develop a Web user behavior prediction model that integrates sequence-based clustering and the use of MCs, 2) to provide a performance evaluation framework, and 3) to suggest a WUM system that can be used to improve online advertising and target marketing, which are important subsets of Web personalization applications and revenue management for business.

Literature Review

Many data analytics techniques such as clustering, classification, association rules, sequence pattern analysis, and dependency modeling have been applied to Web server logs (Facca et al. 2005; Pierrakos et al. 2003; Dimopoulos et al. 2010; Sen et al. 2003). Past research on the use of cluster analysis to identify Web user groups has primarily focused on clustering web users based on the frequency of their page visits. Cluster analysis based on sequences of Web navigation remains a relatively undeveloped area (Kim 2007; Kumar et al. 2007; Park et al. 2008; Shahabi et al. 2003). This is probably due to the dimensional complexity resulting from sequential data representation. A recent study by Hung et al. (2013) showed that sequence-based clustering effectively finds meaningful groups that share common interests and behaviors of Web users. Another interesting finding is that many studies discuss the need for a dynamic and adaptive clustering system, where clustering adapts to the continuous flow of new inputs in real-time. But, only a few studies (Borges et al. 2005; Da Silva et al. 2006; Dimopoulos et al. 2010) presented implementations of dynamic clustering systems. The emergence of Web 3.0 and continued enrichment in Web 2.0 are expected to empower existing Web personalization applications. The significance of this research is thus widely acknowledged because it provides improvements for those applications through the knowledge discovered from sequence-based and dynamic clustering methods.

Also, while the efficacy of using Markov chains to model Web user navigation behavior has been stated repeatedly in many past WUM studies (Cadez et al. 2003; Deshpande et al. 2004; Facca et al. 2005; Sarukkai 2000; Sen at al. 2003), the WUM literature rarely addresses the integration of a Markov chain-based prediction model and cluster analysis. Some research has claimed to achieve a kind of clustering effect of Web pages using a mixture of first-order Markov models without explicitly employing any clustering techniques while performing well (Sen et al. 2003). Additionally, some researchers have looked at the usefulness of modeling Web page navigation behaviors using higher-order Markov chains and some with variable-order Markov chains (Borges et al. 2005; Borges et al. 2007; Singer et al. 2014). Developing such an integrated model, along with an evaluation framework that enables systematic comparison of the various techniques, can help to close obvious gaps in the WUM literature. Web user cluster formation research should address development of clear clustering methodology that can handle sequential information effectively and efficiently. The methodology should be tested with real data in such a way that practical implications are highlighted, for example, the effectiveness of integration of user clusters in a Web user behavior prediction model.

Unlike the existing Markov models focusing primarily on link prediction, i.e., the next page visit by a Web user, the proposed research tries to improve accuracy of prediction by first clustering the sequences in similar clusters and then improve the accuracy by applying the first-order Markov chains. The proposed research also goes beyond this step by developing an experimental framework for applying variable-order Markov chain to improve the prediction accuracy. Improved accuracy will provide business organizations managing large website with heavy traffic an optimal way to place online advertising. To the best of our knowledge, there is not much research discussing integrated clustering and Markov chain approach for next page prediction.

Model

Although a user's navigation pattern recorded in a web server may include the time they spent on each web page in addition to the sequence of pages visited, we restrict ourselves to the problem of identifying and clustering web users' navigation patterns based only on the sequence of page visits. Hence, the primary objective of the model is to predict the next page visit given a Web user navigation history. The

Web user sessions are simplified as a collection of page visits (x_t). The following notation is used throughout this paper.

M : the number of web pages (or page categories)

N : the number of observations

K : the number of clusters

L, l : the length of a sequence vector representing page (page category) visits; L is the random variable and l is a realization.

X_t, x_t : the state visited at time t. X_t is the random variable and x_t is a realization; $x_{l+1} = x_0 = 0$, where 0 indicates off-site (pre-entry or after-exit) state.

$\mathbf{x_n}$: the n[th] sequence vector (Web user session); $\mathbf{x_n} = (x_0, \dots, x_{l+1})$

$\mathbf{S_n}$: the n[th] sequence matrix corresponding to $\mathbf{x_n}$

$p_{ij}(k)$: the transition probability for cluster k defined as $\Pr\{X_{t+1} = j | X_t = i\}$ for i,j = 0,1,..., M

$\mathbf{P}(k)$: the transition matrix for cluster k. $\mathbf{P}(0)$ represents the general transition matrix, for the generation of which no clustering is performed.

Web user clustering methods usually assume that that web server access logs can be preprocessed so that a sequence of pages navigated by a user is available as a list of pages visited in order. Therefore, the first step is to devise a representation scheme that is flexible, meaningful, and easy-to-use when applying clustering algorithms.

Sequence-based Web User Clustering

Considering that most of the clustering algorithms require a distance measure, it is critical to find a way to best represent the sequential web visits. In order to accomplish the research objectives mentioned above, we first adopt the transition-based sequence representation and use fuzzy similarity as the distance measure for clustering. Using the replicated clustering approach, a widely-accepted method for comparison of clustering algorithms, we then investigate whether the clusters, i.e., groups of Web users who follow the same Markov process, are correctly identified and improve the prediction of the next page visits. We develop a performance evaluation framework with the performance metric briefly explained below. Finally, we will conduct a series of experiments in order to determine whether prediction performance is affected by factors such as sequence representation scheme or clustering method as well as by other factors such as the number of actual Web user clusters, the number of pages, similarity between clusters, minimum session length, the number of user sessions, and the number of clusters to form.

Construction of transition-based sequence matrix

There are many ways to represent navigation patterns, and each method may show different performance depending on the problem context. We adopt a matrix-based representation and call it *sequence matrix*. As we are interested in prediction of a user's future navigation behaviors (page visits), we consider using a sequence matrix that resembles the transition matrix of the Markov chain model. The representation then has the benefit of being compatible with many existing clustering algorithms.

The use of the Markov chain works under the assumption that what states are likely to be visited in the next navigation depends only on what page a web user is viewing now. Therefore, each element $\mathbf{S}(i, j)$ of a sequence matrix **S** indicates the probability of visiting j at the next transition, given the present state i and provided that there is only one observation.

Suppose the sequence of visits by the first web user is <1-3-2-3-2-1>. The sequence matrix of the sequence vector $\mathbf{x_1}$ = (0,1,3,2,3,2,1,0), denoted by $\mathbf{S}(i, j)$, can be constructed such that each transition, 0-1, 1-3, 3-2, 2-3, 3-2, 2-1, and 1-0, is counted and divided (normalized) by the frequency of all the transitions

from the same state. Likewise sequence matrix of the second sequence <1,3,1> can be constructed the same way after setting $\mathbf{x_2} = (0,1,3,1,0)$.

$$\mathbf{S}_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ .5 & 0 & 0 & .5 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}, \quad \mathbf{S}_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Applying Clustering Algorithms

In order to give as much flexibility as possible in clustering, distance measures are first defined between a pair of sequence matrices each matrix representing two web users' navigation patterns.

Among the choice of distance measures the most popular one is the Euclidean distance measure. Any sequence of visits among M possible pages, regardless of the length of that sequence, can be converted into a $(M+1)\times(M+1)$ vector as shown above, and the distance between sequence 1 and sequence 2 can be defined as:

$$d_{12} = \frac{\sum_{i,j}[\mathbf{S}_1(i,j) - \mathbf{S}_2(i,j)]^2}{(M+1)^2}$$

As an alternative to distance measure, a similarity measure may be used for clustering. Because we are interested in the patterns of web user's navigation, the following fuzzy similarity ρ can be defined and used as an alternative to a distance measure because 1- ρ is a dissimilarity measure.

$$\rho_{12} = \frac{\|\mathbf{S}_1 \wedge \mathbf{S}_2\|}{\|\mathbf{S}_1 \vee \mathbf{S}_2\|},$$

where $\wedge$ is the fuzzy AND operator: $\mathbf{S}_1 \wedge \mathbf{S}_2 = \boldsymbol{min}\{\mathbf{S}_1(i,j) - \mathbf{S}_2(i,j)\}$; $\vee$ is the fuzzy OR operator: $\mathbf{S}_1 \vee \mathbf{S}_2 = \boldsymbol{max}\{\mathbf{S}_1(i,j) - \mathbf{S}_2(i,j)\}$; and $\|\cdot\|$ is the sum of elements in the matrix.

The distance between sequence 1 and sequence 2 is then computed so that$d_{12} = 1 - \rho_{12}$, which takes values between 0 and 1 . Zero distance indicates an exact match and a distance of 1 indicates a perfect mismatch.

Once a distance measure is defined, commonly used are traditional hierarchical clustering algorithms such as single linkage, average linkage, and Ward methods. However, they are practically prohibited when the number of observations is large due to the storage and computation requirement. Hence, k-means algorithm, which is widely used for fast and efficient clustering, is recommended for this problem. The k-means algorithm however require additional classification model because clustering is designed to cluster existing user sessions, not the upcoming ones (Pierrakos et al. 2003). An appropriately designed classification module makes it possible to use the result of clustering in predicting a web user's future web visits.

An Integrated Prediction Model with Sequence-based Clustering

Prediction of future visits by one user or a group of users can be easily modeled using a Markov chain. Assuming Web users share some browsing patterns, each cluster formed from sequence-based clustering is represented as one Markov chain. In other words, each cluster representing a class of web users with similar navigation patterns has its own transition matrix. In order to calculate the transition matrix for a cluster k, denoted by $\mathbf{P}(k)$, each transition probability $p_{ij}(k)$ of the cluster is estimated from the frequency $\mathbf{S}_n(i,j)$ for all $n \in C_k$, which is normalized so that the row sums equal to 1. For the sake of prediction, we use the general transition probability $p_{ij}(0)$ as a surrogate probability if no transitions occurred in from a given state.

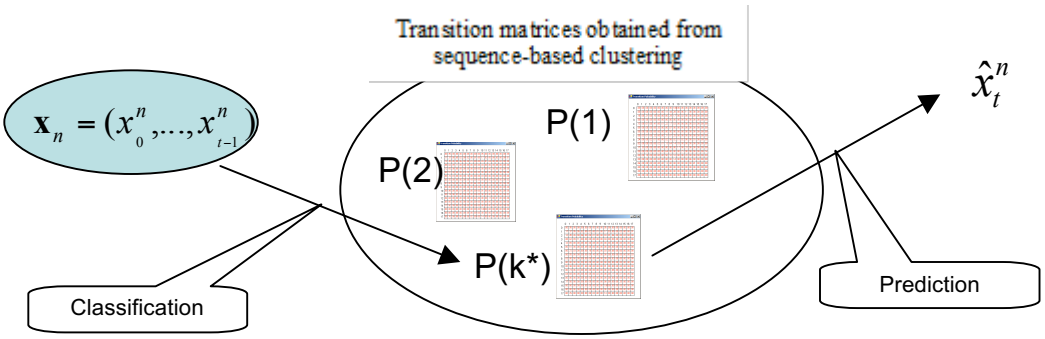

Figure 1. Overview of the Integrated Prediction Model

Given all transition matrices, an integrated prediction model is illustrated in Figure 1. A single prediction of the next visit at time *t* of n^{th} user denoted by $\hat{x}^n_t$ is calculated as:

$$\hat{x}^n_t = argmax_j\{p_{\hat{x}^n_{t-1}j}(k^*)|j = 0, \ldots, M\}$$

where k^* is the cluster that n^{th} user is assigned to, based on $\mathbf{x_n} = (x^n_0, \ldots, x^n_{t-1})$, which represents visit sequence till time *t-1*. A classification module for this assignment will be developed based on the static clustering algorithm such as the *k*-means algorithm.

Clustering performance depends on the problem context and the characteristics of data source, therefore a series of experiments will be conducted by varying factors such as sequence representation scheme, clustering methods, and the number of clusters to form.

Performance Evaluation

In order to test the effectiveness of the proposed prediction model, we first generate a synthetic data set and calculate the prediction accuracy while varying the number of clusters. We generate 10,000 synthetic Web user sessions from three dissimilar Markov chains, where navigation behaviors three Web user clusters are described with transitions between 15 states (page or page categories). The user sessions are then divided into a training set (80%) and a test set (20%) for performance evaluation. After partitioning the data set, overall performance of the prediction model is measured through prediction score denoted by Ω, which is simply the proportion of correct predictions. The prediction score at transition *t*, can also be defined as:

$$\Omega(t) = \frac{\sum_{n \in D} \gamma(\hat{x}^n_t)}{\sum_{n \in D} \delta(\hat{x}^n_t)}$$

where *D* is a partitioned test set. The functions γ and δ are defined as follows:

$$\gamma(\hat{x}^n_t) = \begin{cases} 1, & if\ \hat{x}^n_t = x^n_t \\ 0, & otherwise \end{cases} \text{ and } \delta(\hat{x}^n_t) = \begin{cases} 1, & if\ \hat{x}^n_{t-1} > 0 \\ 0, & otherwise \end{cases}$$

The prediction score with respect to *t,* is expected to provide insights on required time to observe the Web user behavior for accurate prediction of future behaviors.

Preliminary Results

The preliminary results summarized in Table 1 indicate that restoring the true clusters and applying MC models can improve the predictive power. Even when the number of clusters does not match between true and generated clusters, a significant improvement in prediction score was observed compared to the case where only one Markov chain is constructed.

	One MC	3 Clusters	6 Clusters	9 Clusters	12 Clusters
Synthetic Data (6 true clusters)	0.1609	0.2709	0.3762	0.3473	0.3497
Real-world Data (MSNBC Web log)	0.5104	0.5101	0.5186	0.5179	0.5214

Table 1. Prediction Score

We also applied the proposed clustering and prediction methodology to a publicly available large Web log at http://archive.ics.uci.edu/ml/ . The data set was again divided into a training set and a test set with the same ratio (80-20) for performance evaluation. The data set is generated from Internet Information Server (IIS) logs of msnbc.com for the entire day of September, 28, 1999. User sessions are recorded at the level of page category. We used the first 10,000 user sessions and predicted next visits on the test set. The result from this real-world data set also suggested that using our approach can improve the predictive power if we find the correct number of clusters. Whether forming clusters further improve the performance was, however, inconclusive as shown in Table 1. This may be in part due to the fact that we do not know the right number of Web user clusters. Another interpretation is that the navigation behavior may not follow the Markov process or follows higher-order Markov process.

For the real-world data, the second-order Markov model with one MC showed that the prediction score improves significantly to 0.5948. We plan to test whether clustering can further increase the predictive power. Because of dimensional complexity of calculating distances with high-order Markov model, an efficient clustering algorithm for higher-order Markov chain is currently under development.

Conclusions

As the Web grows exponentially, there are increasingly greater opportunities to discover and utilize useful information and knowledge through web usage mining. The proposed research makes it possible the conversion of these opportunities into successes. In addition, the research still in progress is expected to contribute to the web usage mining literature by

- providing effective sequence-based clustering method for identifying Web user clusters,
- offering novel predictive modeling and performance evaluation frameworks through systematic experiments and sensitivity analysis,
- suggesting a WUM system that can improve ad placement and target marketing, and
- providing possible analysis and subsequent model with the transition matrix in Markov chain for a better Web personalization and revenue management.

References

Borges, J., and Levene, M. 2005. "Generating dynamic higher-order markov models in web usage mining," in *Knowledge Discovery in Databases: PKDD 2005*, Springer, pp. 34-45.

Borges, J., and Levene, M. 2007. "Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions," *IEEE Transactions on Knowledge and Data Engineering* (19:4), pp. 441-452.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. 2003. "Model-based clustering and visualization of navigation patterns on a web site," *Data Mining and Knowledge Discovery* (7), pp 399-424.

Da Silva, A., Lechevallier, Y., de Carvalho, F., and Trousse, B. 2006. "Mining web usage data for discovering navigation clusters," Computers and Communications, 2006. ISCC'06. Proceedings. 11th IEEE Symposium on, IEEE2006, pp. 910-915.

Deshpande, M., and Karypis, G. 2004. "Selective Markov models for predicting web-page accesses," *ACM Transactions on Internet Technology* (4:2), pp 163-184.

Dimopoulos, C., Makris, C., Panagis, Y., Theodoridis, E., and Tsakalidis, A. 2010, "A web page usage prediction scheme using sequence indexing and clustering techniques," *Data & Knowledge Engineering* (69), pp 371-382

Facca, F. M., and Lanzi, P. L. 2005. "Mining interesting knowledge from weblogs: A survey," *Data and Knowledge Engineering* (53:3), pp 225-241.

Ho, S. Y., Bodoff, D., and Tam, K. Y. 2011. "Timing of adaptive web personalization and its effects on online consumer behavior," *Information Systems Research* (22:3), pp 660-679.

Hung, Y.-S., Chen, K.-L. B., Yang, C.-T., and Deng, G.-F. 2013. "Web usage mining for analysing elder self-care behavior patterns," *Expert Systems with Applications* (40:2), pp 775-783.

Kim, Y. 2007. "Weighted order-dependent clustering and visualization of web navigation patterns," *Decision Support Systems* (43:4), pp 1630-1645.

Kumar, P., Krishna, P. R., Bapi, R. S., and De, S. K. 2007. "Rough clustering of sequential data," *Data & Knowledge Engineering* (63:2), pp 183-199.

Park, S., Suresh, N. C., and Jeong, B.-K. 2008. "Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm," *Data & Knowledge Engineering* (65:3), pp 512-543.

Pierrakos, D., Paliouras, G., Papatheodorou, C., and Spyropoulos, C. D. 2003. "Web usage mining as a tool for personalization: A survey," *User Modeling and User - Adapted Interaction* (13:4) Nov 2003, pp 311-372.

Sen, R., and Hansen, M. H. 2003. "Predicting Web Users' Next Access Based on Log Data," *Journal of Computational and Graphical Statistics* (12:1), pp 143-155.

Sarukkai, R. R. 2000. "Link prediction and path analysis using Markov chains," *Computer Networks* (33:1), pp 377-386.

Singer, P., Helic D., Taraghi, B., and Strohmaier, M. 2014. "Detecting Memory and Structure in Human Navigation Patterns Using Markov Chain Models of Varying Order," *PLoS ONE* 9(7): e102070. doi:10.1371/journal.pone.0102070

Shahabi, C., and Banaei-Kashani, F. 2003. "Efficient and anonymous web-usage mining for web personalization," *INFORMS Journal on Computing* (15:2) Spring 2003, pp 123-147.