Winter 12-8-2014

Web Usage Mining to Extract Knowledge for Modelling Users of Taiwan Travel Recommendation Mobile APP

Guang-Feng Deng

Yu-Shiang Hung

Chi-Ta Yang

Nien-Chu Wu

WEB USAGE MINING TO EXTRACT KNOWLEDGE FOR MODELLING USERS OF TAIWAN TRAVEL RECOMMENDATION MOBILE APP

Guang-Feng Deng, Institute for Information Industry, Taiwan, raymalddeng@iii.org.tw
Yu-Shiang Hung, Institute for Information Industry, Taiwan, garhung@iii.org.tw
Chi-Ta Yang, Institute for Information Industry, Taiwan, gary1122@iii.org.tw
Nien-Chu Wu, Industrial Technology Research Institute, Taiwan, ncwu@itri.org.tw

ABSTRACT

This work presents the design of a web mining system to understand the navigational behavior of passengers in developed Taiwan travel recommendation mobile app that provides four main functions including "recommend by location", "hot topic", "nearby scenic spots information", "my favorite" and 2650 scenic spots. To understand passenger navigational patterns, log data from actual cases of app were collected and analysed by web mining system. This system analysed 58981 sessions of 1326 users for the month of June, 2014. Sequential profiles for passenger navigational patterns were captured by applying sequence-based representation schemes in association with Markov models and enhanced *K*-mean clustering algorithms for sequence behavior mining cluster patterns. The navigational cycle, time, function numbers, and the depth and extent (range) of app were statistically analysed. The analysis results can be used improved the passengers' acceptance of app and help generate potential personalization recommendations for achieving an intelligent travel recommendation service.

Keywords: Taiwan travel recommendation mobile app, web mining system, sequential pattern, cluster analysis

INTRODUCTION

For tourists, the decision on which destination or product to choose requires a considerable time and effort because tourism services are a class of product regarded as high risk and consumers are often led to engage in extensive information search. The tourism industry has experienced a shift from offline to online travelers. Experts underlined many years ago that the Internet is the main source of information in the tourist domain. An increasing number of travelers are no longer dependent on travel agencies to look for information for their next trip; they have replaced using agencies by the use of the Internet and mobile app. Steinbauer and Werthner (2007) affirmed that the development of information communication technologies during the last decade has affected the tourism industry, as a growing number of travelers have begun to look for tourism information online[1]. As the experts pointed out in the 2013 e-Tourism conference held in Innsbruck in January 2013, 'these systems have significantly changed the travel industry'. As a consequence, travel mobile apps are used increasingly worldwide to provide travel information and services to the passengers. Many firms begin to provide travel app as to interact with tourists in order to promote a destination and provide information on it and, furthermore, they should extract knowledge from this interaction. As an app serve as platforms where consumers can be inspired, get all the information they need about the desired destination and eventually book the holiday, the presence of destinations in the app is crucial. Several studies have investigated the usability evaluation of travel mobile app design [1, 2].

"Taiwan travel recommendation mobile app" was developed by Taiwan Institute for Information Industry Innovative DigiTech-Enabled Applications & Services Institute (IDEAS) in 2013. The app provide four main functions including " recommend by location", " hot topic", " nearby scenic spots information", "my favorite" and 2650 scenic spots(Fig. 1). To ensure effective usefulness of Taiwan travel recommendation mobile app, this paper aims to build a web mining system to analyse the navigational behavior of this app. The empirical period has been underway since June 2013.This process requires a data acquisition and pre-processing stage. The machine learning techniques are mainly applied in the pattern discovery and analysis phase to find groups of app users with common characteristics related to the Internet and the corresponding patterns or user profiles. Finally, the patterns detected in the previous steps are used in the operational phase to adapt the system and make navigation more efficient for new users or to extract important information for the service providers.

Figure 1 Taiwan travel recommendation app

Web usage mining is an area of web mining that deals with extracting interesting and useful knowledge from logging information produced by Web servers [3-5]. Many researchers have applied web usage mining for characterizing usage based on navigation patterns [6, 7], for behavior prediction[8], for personalized recommendation[9-11] and for web service improvement[12]. This study applied web usage mining techniques for mining usage information from app server logs to understand user behavior patterns. To capture sequence information for click behavior patterns using app, sequence-based representation schemes in association with Markov models are combined with an ART2-enhance *K*-mean algorithm for mining cluster patterns. A statistical analysis of each cluster such as user use cycle, time, function numbers is evaluated the depth and extent range of app use. The analysis results can be used improved the passenger's acceptance of app and help generate potential recommendations for achieving an intelligent travel recommendation app. The system will automatically record various types of user operating behavior, including number of clicks, use time, pause time, order of operation, and sequence of clicks. This data will be transferred to a data analysis platform for analysis of user behavior, which will guide subsequent optimization of service design.

METHODOLOGY

This section first describes the data log preprocessing procedure and then presents sequence-based representation suitable for capturing user navigational behavior. Finally, the ART2 neural network and K-mean clustering algorithm used in this study are introduced.

Web-log preprocessing

Data log preprocessing transforms the original logs so that all web access sessions can be identified. The Web server usually registers the access activities of app users in Web server logs. Different server parameters settings result in many different web log types, but log files typically share the same basic information, including client IP address, request time, requested URL, HTTP status code, referrer, etc. Generally, several preprocessing tasks are required before performing web usage mining algorithms on the Web server logs. The tasks in this work include data cleaning, user differentiation and session identification.

These preprocessing tasks resemble those in any other web usage mining problem and are discussed in detail in Hussain et al.

(2010) [13] . The original server logs are cleansed, formatted, and then grouped into meaningful sessions before use in web usage mining. A session can be described as the navigational activities performed by an user between the start and the end of the app usage session. Therefore, session identification is the process of segmenting the access log of each user into individual access sessions. The activity stay time-based method developed by Liu and Kešelj (2007) was applied for session identification in this study [14].This method limits the time spent on a function of app to a specified threshold. If the time between the request most recently assigned to a session and the next request from the user exceeds the threshold, a new access session is assumed. A 10 min activity-stay time is considered a conservative threshold for capturing the time for loading and studying page content. However, this study uses the average use time of the function of all sessions as a threshold for function use time.

Sequence-based representation

After the preprocessing step, behavior representation is performed for the derived access sessions of user navigation in app. Frequency-based representations use a frequency vector in which each element corresponding to a given navigational function is computed by counting the uses for that function and dividing it by the total number of uses in a session. Clustering studies of web usage mining often apply frequency-based representations because of their easiness for representing user sessions and for calculating the distances between sessions. Here, sequence information for user navigational behavior is captured by a sequence matrix, i.e., a matrix representation constructed based on a transition matrix of a Markov chain similar to that developed by Park et al. [15].

A state of a transition matrix of a Markov chain can be defined as either a sub-function (ex. scenic spots) or as a main-function (a category of sub-function) that can be used by the user for navigation. In [15]., user sessions are represented using categories of general topics, not pages themselves, and each Markov chain represents the behavior of a particular subgroup. Fu et al. showed that clustering can also be performed in the sub-function category by using a distance-based hierarchical clustering algorithm to cluster user sessions. Although each app sub-function can be considered a state, considering a main-function as one state is preferable for two reasons. First, the navigational behavior of the app participant within the same category of sub-functions may not provide sufficient information to identify the user clusters. Secondly, since the performance of the clustering algorithm decreases as the number of sub-functions grows and app contains 2560 scenic spots (sub-functions), the clustering problem may be unmanageable if all scenic spots are considered states. In this discussion, therefore, the term ‘‘sub-function” actually refers to a main function.

Use of the Markov chain (*MC*) assumes that the states likely to be used in the next navigation depend on only the sub-function currently in use by users. Therefore, values are assigned to the sequence matrix so that each element in the sequence matrix indicates the proportion of use in state *j* at the next transition given the present state *i*. A user not currently on app is described as being in state 0. Suppose the sequence of use by the first user, s_1 = Start–bus Transfer query service–nearby dining information –interactive picture –Taipei cultural creativity information–bus real-time information service–nearby dining information – End, is represented by the sequence vector x_1 = (0, 5, 1, 2, 3, 4, 1, 0) where 0 is Start and End, 1 is nearby dining information, 2 is interactive picture, 3 is Taipei cultural creativity information, 4 is bus real-time information service, 5 is bus Transfer query service and 6 is bus route query service. The sequence matrix s_1 can be constructed such that each transition, 0–5, 5–1, 1–2, 2–3, 3–4,4-1 and 1–0, is counted and divided (normalized) by the frequency of all the transitions made from the same state. The resulting sequence is

	0	1	2	3	4	5	6
0						1	
1	1/2		1/2				
2				1			
3					1		
4		1					
5		1					
6							

For sequences such as s_1 where not all states are visited, states not included in a sequence (row 6 and column 6) can be omitted from the sequence matrix. For distance calculation purposes, however, all states have a value of 0. This violates the Markov model transition matrix rule that the sum of each row must equal 1. To satisfy this rule, the only alternative is assigning a value of 1/*M* to all elements in such rows. For app service containing an *M* main-function, an *M* * *M* sequence matrix (S_n) can be constructed from the corresponding sequence vector x_n = (0, x_1, . . . , x_h, 0) as follows:

Step 1. Initialize $S_n(i, j)$ with zeroes, for all *i* and *j*.
Step 2. Set $t = 0$.
Step 3. $S_n(x_t,x_t+1) \leftarrow S_n(x_t,x_t+1) + 1$.
Step 4. $t \leftarrow t + 1$.
Step 5. Repeat steps 3 and 4 if $t <= h$.
Step 6. For each *i* and *j*, $S_n(i, j) \leftarrow S_n(i, j) / \sum_{j=0}^{M} S_n(i, j)$

Clustering analysis

In Web usage mining, clustering finds groups that share common properties and behavior by analyzing the data collected in web servers. Given the transformation of user navigational access sessions into a multi-dimensional space as sequence-based representation matrices of functions, a clustering algorithm was applied to the derived user navigational access sessions. Since access sessions are the images of activities by users, representative user navigational patterns can be obtained by clustering. These patterns also facilitate profiling of users of the app service. This section describes how session clustering is performed and how cluster number is determined.

Optimizing the number of clusters

Since the used clustering algorithm is a supervised clustering method, an *ART*2 neural network [16] is needed to determine the number of clusters. The ART2 neural network architecture is designed for processing both analog and binary input patterns. An *ART*2 neural network consists of F_1 and F_2 layers. The F_1 layer has seven nodes (W, X, U, V, P, Q). The input signal is processed by the F_1 layer and then passed from the bottom to the top value (b_{ij}). The result of the bottom-to-top value is an input signal of the F_2 layer. The nodes of the F_2 layer compete with each other to produce a winning unit, which returns the signal to the F_1 layer. The match value is then calculated with the top-to-bottom value (t_{ji}) in the F_1 layer and compared with the vigilance value. If the match value exceeds the vigilance value, then the weights of b_{ij} and t_{ji} are updated. Otherwise, the reset signal is sent to the F_2 layer, and the winning unit is inhibited. After inhibition, the other winning unit is found in the F_2 layer. If all F_2 layer nodes are inhibited, the F_2 layer produces a new node and generates the initial weights corresponding to the new node.

E. Session clustering

After the *ART2* neural network determines the number of clusters, standard clustering algorithms can partition this space into groups of sessions that are close to each other based on a distance measure. The well-known *K*-means algorithm is used as the base method for clustering interest-based representation sessions and sequence-based representation sessions. The *K*-means clustering algorithm groups sessions by attributes/features into a *k* (positive integer) number of groups by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Additionally, the most popular Euclidean distance is used as the distance measure. The *K*-means clustering algorithm is performed in the following steps: Step 1: Generate initial random cluster centroids for *k* clusters and *k* obtained by ART2 neural network. Step 2: Assign each session to its closest cluster centroid in terms of Euclidean distance. Step 3: Compute new cluster centroids. Step 4: If cluster memberships differ from the last iteration, repeat steps 2–3. Step 5: Stop and store clustering result.

Session clustering obtains a set of clusters, $C = \{c_1, c_2, \ldots, c_k\}$ in which each c_i ($1 <= i <= k$) is a subset of the set of user access sessions *S* where *k* is the number of clusters. A mean vector mc of a sequence-based representation (a mean matrix mc of an interest-based representation) is computed as a representation for each session cluster $c \in C$. Each mean vector represents the representative user navigational pattern for a cluster in which a particular set of functions are accessed. The mean value for each function in the mean vector is computed as the average weight of the functions across total access sessions in the cluster. Therefore, the mean value is also between 0 and 1. Meanwhile, a weight threshold for the mean vector of each session cluster, w_{min}, is set as a constraint to filter out functions with mean values below the threshold for the cluster. The remaining navigational functions in each cluster are considered of greatest interest to users and are used as representative navigational patterns for the cluster. Since the least mean value is always far smaller than the second least and the third least mean values, the second least mean value of each mean vector is used as the w_{min} for each session cluster.

In our research user navigational patterns are described in terms of the common usage characteristics for a group of users. Since many users may have common interests up to a point during their navigational navigation, navigation patterns should capture the overlapping interests or the information needs of these users. In addition, navigational patterns should also be capable to distinguish among functions based on their different significance to each pattern. This work defines a user navigational pattern *np* as a pattern that captures an aggregate view of the behavior of a group of users based on their common interests or sequence information. After session clustering, $NP = \{np_1, np_2, \ldots, np_k\}$ represents the set of user navigational patterns, in which each np_i is a subset of *F*, the set of functions.

The system will automatically record various types of user operating behavior, including number of clicks, use time, pause time, order of operation, and sequence of clicks. This data will be transferred to a data analysis platform for analysis of user behavior, which will guide subsequent optimization of service design.

BEHAVIORAL ANALYSIS

This section describes the user navigational pattern of four clusters found by applying sequence-based representation schemes in association with Markov models combined with *ART2*-enhanced *K*-mean algorithm. The app data server was used to obtain a record of each function used by a passenger from the start of the app until the present. For all of June, 2014, 58981 sessions of 1326 users were identified for analysis. Matlab Language is used to perform the web usage mining algorithm. Figure 2 shows four clusters that described the main navigational sequential path, and Table 1 presents the characteristics of each

cluster.

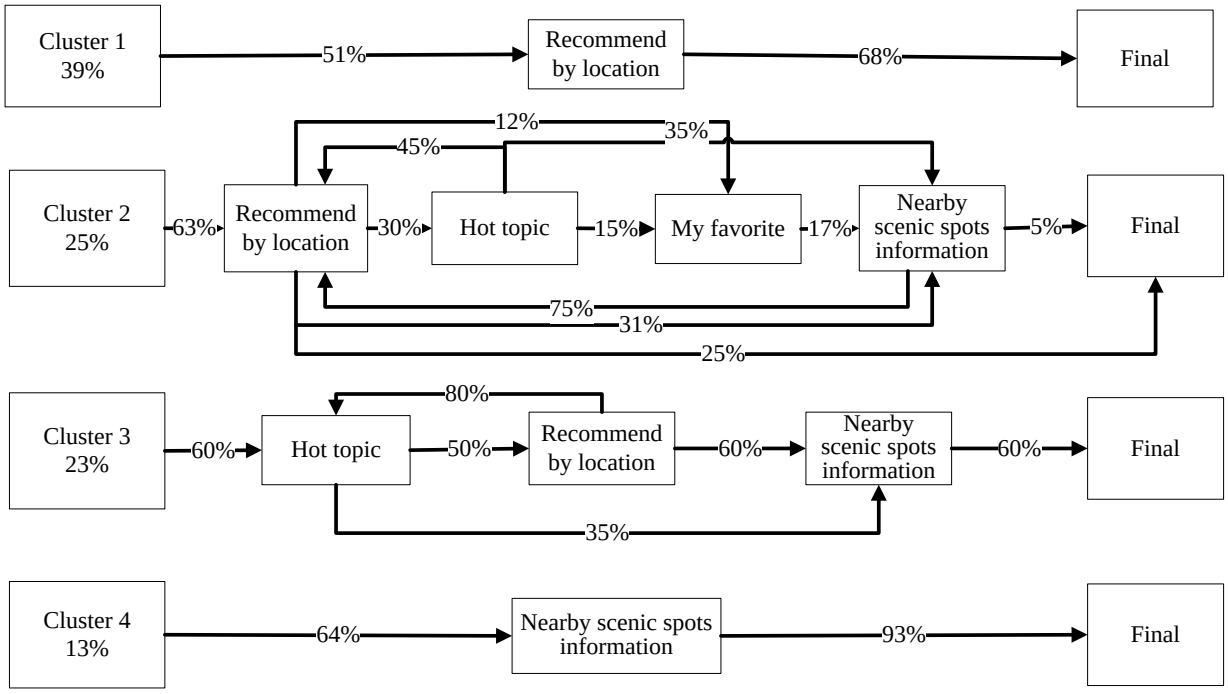

Figure 2 User navigational sequential clustering patterns

Cluster 1, the largest cluster, contains 39% session records; cluster 2 contains 25% session records; cluster 3 contains 23% session records. Based on the observed means and proportions of characteristic, C_1 can be labelled as "Task orientation type user", which enables provisional identification of cluster 1 as the more "typical" cluster. The "recommend by location" function trigger C_1 users at a rate thousands of times greater than other clustering users. The cluster C_2 can be labelled "Classic tourism type user". The "recommend by location" function triggers C_2 users, and the contiguous navigational sequence is "hot topic", "my favourite", and "nearby scenic spots information" functions. The cluster C_4 users can be labelled as "Convenient orientation type". The "nearby scenic spots information" function triggers C_2 users, and then leave app. The cluster C_3 users can be labelled as "Entertainment type". The "hot topic" function is requested at a much higher rate by C_3 users compared to other clustering users, and the contiguous navigational sequence is "recommend by location", and "nearby scenic spots information" functions.

Clusters 1 to 4 show that the "recommend by location" and "hot topic" functions are the most common triggers of app utilization by users. Table 1 show that the Classic tourism type C_2 use the most functions and spend longest time. The Convenient orientation type C_4 uses the fewest functions. Task orientation type users C_1 use app most frequently. The Convenient orientation type users C_4 spend the least time every time on app service.

Table 1 Characterization of user navigational sequential clustering patterns

Cluster ID	Type	People Account	percentage	Time (sec)	Function number
C_1	Task orientation type	520	**39%**	282	7.5
C_2	Classic tourism type	328	25%	**3142**	**124**
C_3	Entertainment type	299	23%	1307	44
C_4	Convenient orientation type	179	13%	266	7

CONCLUSION AND FUTURE WORK

The tourism industry has experienced a shift from offline to online travelers and this has made the use of intelligent systems in the tourism sector crucial. These information systems should provide tourism consumers and service providers with the most relevant information, more decision support, greater mobility and the most enjoyable travel experiences.

To improve understanding of the navigational behavior of Taiwan travel recommendation app service, this work presents a web mining system to analyse real-world data by applying sequence-based representation schemes in association with Markov models combined with *ART2*-enhance *K*-mean algorithm. The analysis results show that user navigational behavior can be

classified by a sequence-based representation scheme into four distinct cluster types from different use sequence. Each type displays different navigational time, function numbers and needs. This research shows that the use of sequence-based clustering in web usage mining effectively finds meaning groups that share common interests and behaviors and effectively extracts knowledge needed to understand the motivation for travel recommendation app service. The analysis results can be used by experts in tourism and advertisers to further research and to assist in policy-making in the public traffic domain. Future research will apply the sequence-based clustering results for the travel recommendation app to providing personalized advertisement to users.

ACKNOWLEDGMENT

This study is conducted under the" Smart LOHAS Service Development/Technology Applications and Multi-field Validation Project (1/4)" of the Institute for Information Industry which is financially supported by the Ministry of Economy Affairs of the Republic of China.

REFERENCES

[1] Harding, M., Finney, J., Davies, N., Rouncefield, M., & Hannon, J. (2013) 'Experiences with a social travel information system', *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 173-182.

[2] Norgate, S. H., & Smith, L. (2012) 'Re-thinking app design processes: applying established psychological principles to promote behaviour change-a case study from the domain of dynamic personalized travel planning', *Proceedings of the 6th ACM workshop on Next generation mobile computing for dynamic personalised travel planning*, pp. 5-6.

[3] Sajid, N. A., Zafar, S. & Asghar, S. (2010) 'Sequential pattern finding: A survey', *Proceedings of International Conference on Information and Emerging Technologies (ICIET)*, pp. 1-6.

[4] Facca, F. M. & Lanzi, P. L. (2005) 'Mining interesting knowledge from weblogs: a survey', *Data & Knowledge Engineering*, Vol. 53, pp. 225-241.

[5] Wang, Y.-T. & Lee, A. J. T. (2011) 'Mining Web navigation patterns with a path traversal graph', *Expert Systems with Applications*, Vol. 38, pp. 7112-7122.

[6] Bayir, M. A., Toroslu, I. H., Demirbas, M. & Cosar, A. (2012) 'Discovering better navigation sequences for the session construction problem', *Data & Knowledge Engineering*, Vol. 73, pp. 58-72.

[7] Chen, L., Bhowmick, S. S. & Nejdl, W. (2009) 'COWES: Web user clustering based on evolutionary web sessions', *Data & Knowledge Engineering*, Vol. 68, pp. 867-885.

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

[9] Pierrakos, D., Paliouras, G., Papatheodorou, C. & Spyropoulos, C. D. (2003) 'Web usage mining as a tool for personalization: a survey', *User Modeling and User - Adapted Interaction*, Vol. 13, pp. 311-372.

[10] Mobasher, B., Cooley, R. & Srivastava, J. (2000) 'Automatic personalization based on Web usage mining', *Communications of the ACM*, Vol. 43, pp. 142-151.

[11] Park, D. H., Kim, H. K., Choi, I. Y. & Kim, J. K. (2012) 'A literature review and classification of recommender systems research', *Expert Systems with Applications*, Vol. 39, pp. 10059-10072.

[12] Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J. &García, S. (2012) 'Web usage mining to improve the design of an e-commerce website: OrOliveSur.com', *Expert Systems with Applications*, Vol. 39, pp. 11243-11249.

[13] Hussain, T., Asghar, S. & Masood, N. (2010) 'Web usage mining: A survey on preprocessing of web log file', *Proceedings of International Conference on Information and Emerging Technologies (ICIET 2010)*, pp. 1-6.

[14] Liu, H. & Kešelj, V. 'Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests', *Data & Knowledge Engineering*, Vol. 61, pp. 304-330.

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

[16] Kuo, R. J. & Lin, L. M. (2010)"Application of a hybrid of genetic algorithm and particle swarm optimization algorithm for order clustering," *Decision Support Systems*, Vol. 49, pp. 451-462.