IDENTIFYING TOURIST ROUTE PATTERNS USING DATA MINING TECHNIQUES

P. Warintarawej, K. Chaikong, P. Kadedaiwang, P. Onsrithong,
P. Laksanajan and S. Siwyew

Faculty of Liberal Arts and Management Sciences
Prince of Songkla University, Surat Thani Campus
31 Moo 6 Makhamtia District, Muang, Surat Thani
84000 Thailand

ABSTRACT

In this study, researchers applied data mining techniques to reveal tourist route patterns to popular destinations in Surat Thani Province in southern Thailand. Data mining refers to the process of discovering patterns in large data. Two data mining techniques were employed: 1) Cluster analysis was used to identify unique clusters of tourists with common behavioral trends. 2) Association rule mining was used to determine tourist route patterns. From these two data mining techniques, the researchers were able to identify unique clusters of tourists who followed common patterns of travel. The main implications of this study are: 1) that data mining may be used to explain the movement of tourists in any region in the world, and 2) that different facets of the tourism industry can use this information to understand and respond to tourists' needs and interests.

.

Keywords: Data mining, Cluster analysis, Association rule mining, Tourist behavior, Tourism route patterns

Introduction

Tourist behavior plays an important role in tourism sectors; it enhances tourism businesses to understand the tourists' needs and interests in order to develop the appropriate traveling packages. Strategic planning and marketing solutions for new products require knowledge of tourist factors; characteristics, preferences, patterns of tourist demand. The study of tourist behavior can forecast tourism trends and improve the product design and development (Liao, Chen & Deng, 2010; Bramwell, 1998; Witt & Witt, 1995). In addition, the behavior of tourists could be translated into meaningful information such as "what are the factors (such as gender, age, job category, income, education) affect tourist's preferences?" Or "How different are the traveling route patterns between European and Asian tourists in Thailand?".

To extract knowledge in large data, data mining is the process of automatically discovering useful information such as patterns, associations, changes and significant structures. (Liao, Chen, & Wu, 2008; Liao, Hsieh & Huang, 2008; Tan, Steinbach, and Kumar, 2006). Data mining techniques have been widely used in tourism data. Bose (2009) concluded three main uses of data mining for the tourism industry, which are (1) forecasting expenditures (2) analyzing tourists'profiles (3) forecasting the number of tourist arrivals. As well as, Samarasinghe (2013) mentioned that data mining and tourism industry have a strong relationship. Due to the fact that, the tourism industry involves people with different needs. Min, Min, & Emam (2002) used data mining techniques to find their customers' preferences to decide customer retention strategy. For instance, high-end tourists are likely to have a unique lifestyle, they will only visit the same destinations which fulfill their needs. Hence, data mining techniques can be used as a tool to mine into tourist data (such as the lifestyle factors and regional details) with regard to dig any patterns available on tourist behavior. Dev, Klein, & Fisher (1996) used association rule techniques to analyze the market analysis of hotels, airlines and other services among visitors for the principle of partner selection and marketing alliances. Lau, Lee, Lam, & Ho. (2001) applied clustering techniques to segment travelers into different clusters based on personal information mined from personal websites. The results show that traveling businesses can take this information into account to offer specially designed packages through email.

As mention above, to understand tourist behavior is the vital information for tourism firms, in this study, we focus on tourist's preferences and their route patterns in order to design the appropriate package tours or activities and also give a benefit to tourism sectors to enhance tourism policy to support local area. herefore, this study applied two data mining techniques for two principal aims: (1) Cluster analysis was used to identify unique clusters of tourists with common behavioral trends and (2) Association rule mining was used to determine tourist route patterns. The study was taken place in Surat Thani province, the popular destinations in southern Thailand. The main implications of this study are: 1) that data mining may be used to explain the movement of tourists in any region in the world, and 2) that different facets of the tourism industry can use data mining to understand and respond to tourists' needs and interests.

The rest of this paper is organized as follows. Section 2 provides a background of the literature on data mining in the part of clustering and association rule technique. Section 3 presents the methodology to identify unique clusters of tourists and determine tourist route patterns and their results. Finally, Section 4 concludes the study.

Literature review

Data mining

Data mining refers to extracting or "mine" knowledge from large amounts of data. Many literatures treat data mining as synonym of Knowledge Discovery from Data (KDD). Alternatively, the boundaries of the data mining step in the KDD process are not clearly defined. Fayyad, Piatetsky-Shapiro & Smyth (1996) shows that data mining can be organized such an essential step in KDD process which shows in Figure 1:

1) Learning the application domain: includes relevant prior knowledge and the goals of the application parameters.
2) Select target dataset: focusing on a subset of variables or data samples on which discovery is to be performed.
3) Preprocessed data: includes basic operations, such as removing noise or outliers if appropriate, handling missing data fields.
4) Transformed data: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5) Data mining process: choosing data mining techniques for searching patterns in data, such as classification, association rules, clustering etc.
6) Interpretation and Evaluation: interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users.

Data mining techniques can be grouped into two types of models as:

- Predictive model: the model tries to assign an unknown data based on known data such as classification, regression etc.
- Descriptive model: the model tries to describe the characteristics of data by finding the patterns in data, such as clustering, association rules, sequential pattern discovery, etc.

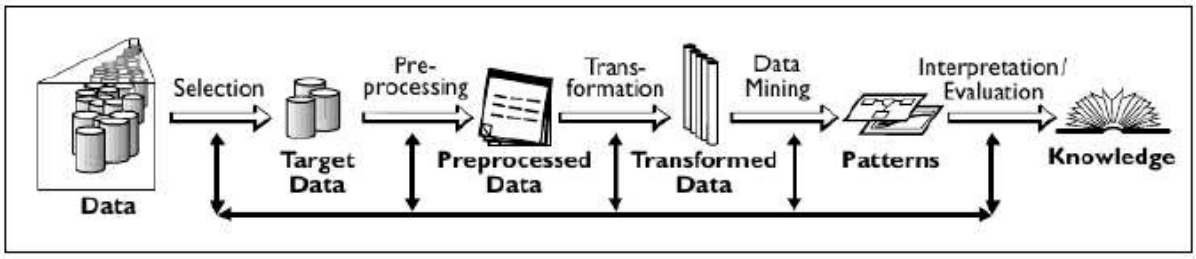

Figure 1. Overview of KDD Process

Cluster Analysis

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters (Jiawei & Kamber, 2006). Cluster analysis is a form of learning by observation with little or no prior knowledge. Based on statistics, k-means

algorithm is the most well-known partitioning method by a distance-based between the object and the cluster mean. Given an input number of clusters, *k*, and partition a set of *n* object into *k* clusters. The k-means algorithm has two steps as follows.

1) Randomly select *k* objects as a center or mean of each cluster. For each remaining object, an object will be assigned to the cluster to which it is the most similar based on the distance between the object and the cluster mean.
2) After assigning all objects into their clusters, the mean of each cluster will be recomputed. The process iterates until the mean of each cluster doesn't change.

Association Rules

Association rule mining is one of data mining techniques which aims to find patterns that occur in a dataset frequently enough to be interesting. Hence the association or correlation of data attributes is considered within instances, rather than between instances. These correlations are then expressed as rules: if X appears in an instance, then Y also appears.

The association rules technique was introduced by Agrawal (1993) which help to uncover hidden patterns in large datasets; the idea comes from the market basket analysis. The association rule algorithm is to determine the relationship between items or features that occur synchronously in the database. For example, if customers buy item X also buy item Y as well. The rules can be written such as "40% of the customers who buy a dozen of eggs also buy milk; 60% of all transactions that contain both of these item". As the example, association rules are explained by two statistical measure scores called *support* and *confidence*, here 40% is confidence and 60% is support. The rules that satisfy user-specified minimum support and minimum confidence constraints are extracted from the database. To understand well, let's see the overview of the basic concept of association rules mining as follow.

Let D be a set of $\boldsymbol{n}$ transactions, $D = \{T_1, T_2, \dots, T_n\}$ and let I be a set of items, $I = \{I_1, I_2, \dots, I_n\}$ Each transaction is a set of itemset, i.e., $T_i \in I$. An association rule is an implementation of the form $X \rightarrow Y$ (support, confidence), where $X, Y \in I$, and $X \cap Y \neq \emptyset$; X is called the antecedent and Y is called the consequent of the rule.The support of an association rule $X \rightarrow Y\ (denoted\ as\ Supp\ (X \rightarrow Y))$ is the fraction of all transactions which contain both $X\ and\ Y$. The confidence of the rule $X \rightarrow Y\ (denoted\ as\ Conf\ (X \rightarrow Y))$ is the proportion of transactions containing X which also contain Y. Support and confidence can be written in probability terms as

$$Supp\ (X \rightarrow Y) = P\ (X \cap Y)$$

$$Conf(X \rightarrow Y) = P\ (X \mid Y) = \frac{P\ (X \cap Y)}{P\ (x)}$$

The procedures of mining association rules are breakdown into two steps:

1. Finding frequent itemsets: to find all combinations of items which whose support is greater than or equal to the minimum support threshold (called Minsup).
2. Generate association rules: to combine all frequent itemsets and calculate its confidence, the rules that whose confidences are greater than or equal than minimum confidence are retrieved (called MinCof).

Methodology and the results

Data Collection

The study used questionnaires as a tool to collect data from tourists in Surat Thani province covered a period from 14th December 2013 to 13rd January 2014. The sample size is 245 which calculated by using the formula developed by Cochran (1953) at 95% confidence interval. The tourist sampling group was selected by convenience sampling.

Demographic characteristics

The demographic characteristics of the sample are shown in Table 1. The total number of tourists in the sample was 245. The samples comprised of Thai (62%), European (27.35%), Asian (4.08%), Australian (3.67%) and American

(3.27%) visitors. The majority of respondents were aged 20-30 (45.3%) and 31-40 (25.71%). Moreover, 43.3% were business owners, 24.08% were officers in private companies. Over 24 Thais earned 15,000-25,000 (Thai baht) per month. On the other hand, 21.63 % of foreigners earned monthly income more than 50,000.

Table 1. Demographic characteristics

Items	Thai		Foreigner		Total	
	n	%	n	%	n	%
1. Gender						
Male	65	26.53	44	17.96	109	44.49
Female	87	35.51	49	20.00	136	**55.51**
Total	152	62.04	93	37.96	245	100
2. Age						
< 20	10	4.08	3	1.22	13	5.31
20-30	59	24.08	52	21.22	111	**45.30**
31-40	42	17.14	21	8.57	63	25.71
41-50	27	11.02	11	4.49	38	15.51
> 50	14	5.71	6	2.45	20	8.16
3. Occupation						
Students	24	9.80	13	5.31	37	15.1
Government officers	28	11.43	5	2.04	33	13.46
Business Owners	48	19.59	11	4.49	59	**43.40**
Officers	49	20.00	47	19.18	96	24.08
Husbandry/Housewife	1	0.41	1	0.41	2	0.81
Retired	2	0.82	16	6.53	18	7.34
4. Income (baht/month)						
< 15,000	36	14.69	14	5.71	50	20.4
15,000-25,000	60	24.49	4	1.63	64	26.12
25,000-35,0000	25	10.20	9	3.67	34	13.87
35,000- 50,000	14	5.71	13	5.31	27	11.02
> 50,000	17	6.94	53	21.63	70	**28.57**

Clustering Analysis- Analysis Tourist Clusters

In order to use k-means algorithm, data has to present as a matrix. In our case, we represent tourism data for each row as a tourist containing *p* variables consist of demographic characteristics and tourist's preferences (shown in Table 2.).

Table 2. Tourist's variables

No.	Demographic characteristics
1	Gender
2	Age
3	Occupation
4	Income

No.	Tourist's preferences
1	Number of times used to visit
2	Period
3	Preferred tourism attractions
4	The purpose of traveling
5	Number of nights stay in Surat Thani
6	Information sources
7	Average daily cost per person in Surat Thani

The structure of the matrix is shown as in Table 3., when x_{ij} is the data of i^{th} tourist at j^{th} variable.

Transaction/Variables	V_1	V_2	…	V_p
1	x_{11}	x_{12}		x_{1p}
2	x_{21}	x_{22}		X_{2p}
…				
n	x_{n1}	X_{n2}		x_{np}

The result from k-means algorithm by Weka data mining tool by using Euclidean distance formula (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009) is shown in Table 3.

Table 3. The result of k-means: Tourist Clusters

Factors	Cluster 1 (148) Mean/Majority	Cluster 2 (97) Mean/ Majority
Nationality	Thai	Foreigner
Age	21-30	21-30
Occupation	Business Owners	Officers in private companies
Income (Thai Baht)	15,000-25,000	> 50,000
Number of times used to visit	1-2	Never (First time)
Period	Depend on opportunities	Depend on opportunities
Preferred tourist attraction	Sea	Sea
Purpose of traveling	Relaxation	Relaxation
Number of nights stay in Surat Thani	2.17	3.43
Information source	Friends	Websites
Average daily cost per person in Surat Thani (Thai Baht)	2,720.10	4,134.02

The analysis result presents tourists' behavior in Surat Thani into two clusters. Cluster 1 shows the majority of visitors are Thais who earn monthly income of 15,000-25,000 Baht. Most of them are business owners and spent about 2 days in Surat Thani per visit. Thai visitors were used to visit Surat Thani 1-2 times. Morever, they spent around 2,700 baht for their trip. Thai tourists prefer to get the information about Surat Thani from friends. On the other hand, Cluster 2 describes the foreign tourist's behavior. Mainly foreign tourists who visited Surat Thani were officers in private companies who earned twice the monthly income of the Thai tourists. The average daily expenditure per person for cluster 2 is about 4,000 baht because they stayed a bit longer than Thai tourist. The main information sources for foreigner are websites.

Nevertheless, the common behaviors for both clusters are presented; age, preferred traveling period, preferred tourist attraction and the objective of traveling. Generally, Surat Thani is popular among Thai and foreign visitors. Surat Thani is well-known for its hospitality, tropical climate, relaxing and enjoyable vacation. Therefore, most of visitors come for the beaches and relaxation.

Association Rules - Identify Tourist Route Patterns

Given a set of tourism attractions in Surat Thani, we provided 2 main groups of attractions in the questionnaire which are natural attractions and cultural attractions. The list of tourist attractions is shown in Table 4.

Table 4. Tourism attractions

Code	Natural attractions
N1	Koh Samui
N2	Koh Phagan
N3	Angthong National Marine Park
N4	Khaosok National Park
N5	Ratchaprapha Dam Surattani

Code	Cultural attractions
C1	Wat Phra Boromathat Chaiya
C2	Wat Suan Mokkh
C3	Thai silk Phumriang Village
C4	Phrathat Sri Surat
C5	Singkron Cave

Let the samples give a list of places that they plan to visit. The example of tourist route patterns is presented in Table 5.

Table 5. The example of tourist route patterns

Transaction	Route pattern
	N1,N2
2	N1,N3,N4
…	…
n	N1,C2,C3

The results of association rules by an Apriori algorithm with *minSupp* = 10% and *minConf* = 30% are shown in Table 6 for cluster 2 (Foreigners) and Table 7 for cluster 1 (Thai)

Table 6. Tourist Route Patterns of foreigner (cluster 2)

Rules	Route pattern	Interpret the rule	(Supp,Conf)
1	N1 -> N2	Koh Samui -> Koh Phangan	(95.70%, 55.91%)
2	N3 -> N1	Angthong National Marine Park -> Koh Samui	(10.75%, 10.75%)
3	N3 -> N2	Angthong National Marine Park -> Koh Phagan	(10.75%, 7.53%)

The first rule says 55.91% of the tourists who go to Koh Samui will go to Koh Phangan too; 95.70% of all tourist routes that contain Koh Samui and Koh Phangan (Koh means island in Thai language). In fact**,** they are not far from each other (20 km., 10 minutes by boat from koh Samui to Koh Phagan). The most popular natural attraction which is near Koh Samui is Angthong National Marine Park. Tour operators often provide the trip from Koh Samui to Angthong National Marine Park, where tourists can have many activities such as snorkeling, trekking, canoe. The rule 2 and 3 show that Angthong National Marine Park and Koh Samui (also Koh Phangan) frequently happen together with the traveling route of foreigners.

Table 7. Tourist Route Patterns of Thai (cluster 1)

Rules	Route pattern	Interpretation of the rule	(Supp,Conf)
1	C2 -> C1	Wat Suan Mokkh -> Wat Phra Boromathat Chaiya	(34.21%, 25%)
2	N5 -> N1	Ratchaprapha Dam Surattani -> Koh Samui	(21.05%, 17.11%)
3	N2 -> N1	Koh Phagan -> Koh Samui	(13.82%, 11.84%)
4	C1,C3-> C2	Pra Boromathat Chaiya Temple, Thai silk Phumriang Village -> Wat Suan Mokkh	(11.84%, 8.55%)
5	C2,N1-> C1	Suan Mokkh Temple, Koh Samui -> Wat Phra Boromathat Chaiya	(11.84%, 6.58%)

On the other side, the route patterns of Thai visitors in Surat Thani are different from foreigners. As a result shows in Table 7, Thai tourists came to Surat Thani not only for relaxing on the beach but also visiting the famous cultural places in Chaiya which was the center of *Srivijaya* empire from the 5th to 13th century. The evidences of *Srivijaya* empire can be found in Wat Phra Boromathat Chaiya Rat (wat means temple in Thai language). The central Chedi in Srivijaya style with its many golden ornaments is even the iconic symbol of the province of Surat Thani (My unseen Thailand, 2010). Moreover, there are a few cultural attractions in Chaiya District, such as Wat Suan Mokkh and Thai Silk Phumriang Village. The Wat Suan Mokkh is the famous place for meditation practice. As well as, Thai silk Phumriang Village, the place has the reputation of a notable local product made of handwoven silk cloth from the coastal village Phumriang Village.

Conclusion

Tourists' needs and interests are the treasure information for tourism, business, if the firm can understand and develop a new product that fulfil their preferences, then the tourists will be more appreciated and become a royalty customer. This study applied two data mining techniques to extract the hidden pattern from tourist's behavior. The cluster analysis and association rules are presented on observatory tourism data by using questionnaire. In order to identify unique clusters of tourists, k-means algorithm for cluster analysis was performed. Association rule mining by the Apriori algorithm was used to determine tourist route patterns. The results explain the characteristics of tourists and also their movement in Surat Thani. The results can be used by different aspects the tourism industry to create a tour package and also the Tourism Authority of Thailand can launch a campaign to promote the unknown places to tourists or provide information to augment the tourism marketing plan.

5. References

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In ACM SIGMOD Record (Vol. 22, No. 2, pp. 207-216). ACM.

Bose, I. (2009). Data Mining in Tourism.

Bramwell, B. (1998). User satisfaction and product development in urban tourism. *Tourism Management*, *19*(1), 35-47.

Dev, C. S., Klein, S., & Fisher, R. A. (1996). A market-based approach for partner selection in marketing alliances. Journal of Travel Research, 35(1), 11-17.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth P., "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Commun. ACM, vol. 39, no. 11, pp. 27–34, 1996.

Jiawei, H., & Kamber, M. (2006). Data mining: concepts and techniques. San Francisco, CA, itd: Morgan Kaufmann, pp.383.

Lau, K. N., Lee, K. H., Lam, P. Y., & Ho, Y. (2001). Web-site marketing: for the Travel-and-Tourism Industry. The Cornell hotel and restaurant administration quarterly, 42(6), 55-62.

Liao, S. H., Chen, Y. J., & Deng, M. Y. (2010). Mining customer knowledge for tourism new product development and customer relationship management. *Expert Systems with Applications*, *37*(6), 4212-4223.

Liao, S. H., Chen, C. M., & Wu, C. H. (2008). Mining customer knowledge for product line and brand extension in retailing. Expert systems with Applications,34(3), 1763-1776.

Liao, S. H., Hsieh, C. L., & Huang, S. P. (2008). Mining product maps for new product development. Expert Systems with Applications, 34(1), 50-62.

My unseen Thailand. (2010). Retrieved October 9, 2014, from http://myunseenthailand.blogspot.com/2010/07/wat-phra-boromathat-chaiya.html

Tan, P.-N., Steinbach, M., Kumar, V. (2005). Introduction to data mining, Addison Wesley.

Samarasinghe, I. A. K. C., Kodituwakku, S., & Yapa, R. D. (2013). Data Mining and Service Customization in Leisure and Hospitality. International Journal of Soft Computing and Engineering. 3(5).

Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. International journal of Forecasting, 11(3), 447-475.