RULES MINING BASED ON CLUSTERING OF INBOUND TOURISTS IN THAILAND

Wirot Yotsawat and Anongnart Srivihok
Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand
E-Mail: g5314401258@ku.ac.th

ABSTRACT

Tourism industries are growing up rapidly with more competition. So, travel agencies or tourism organizations must have a good planning and provide campaign for tourist's needs. This study proposes the usage of data mining for tourism industries in Thailand. Data clustering and association rule mining were chosen as the data mining methods in order to discover useful knowledge. Two-level clustering with decision tree bagging was applied to construct the segments of tourist. Apriori algorithm was then used to find the rules on each cluster. The experimental results indicated that the tourists data was separated into eleven differently segments and decision tree bagging for attributes weighting can enhance the quality of clusters. The eleven segments were analyzed in order to identify tourists' behavior patterns and their preferences. Association rule mining was applied to each segment in order to find the relationship among the features of tourist data. The rules were filtered again by experts. The clustering and association rule results can be served to tourism organization in order to support their strategic and market planning.

Keywords: tourism, rule mining, cluster.

INTRODUCTION

Tourism industries are growing up rapidly in many countries. There are many supporting policy and developing plans from their governments. Activities and destinations are promoted for the attraction of tourists and tourism investors from foreign countries. Tourists receive more choices for selecting the best interesting places. So, travel agencies or tourism organizations must have a good planning and provide campaigns for tourist's needs. They have to know tourists' behavior patterns and their preferences. This paper proposes data mining techniques on inbound tourists in Thailand by using segmentation and association rule techniques. Two-level clustering with factors weighting by Decision Tree bagging, was a methodology for applying cluster analysis to explore the tourists patterns for market planning, promotion and package design for each group. Association rule mining was then applied to each segment. The results of association rule can be applied for tourists' recommendation system. Thus, this study can serve as useful knowledge for travel agencies and other tourism organizations.

LITERATURE REVIEW

Data mining is the process of automatically discovering useful information in large data base [1]. Since data mining was introduced, it has developed by many researchers. Data mining technique has been applied in many fields such as accounting, medicine, law, and so forth. Some researchers focused on the implementation of data mining for tourism such as Wong, Chen, Chung, and Kao [2], they proposed the usage of three data mining techniques to analyze the travel patterns of Northern Taiwan tourists. RFM (Recency, Frequency and Monetary Value) was applied to identify valuable travelers, C4.5 decision tree and association rule were then applied to discover the traveling pattern and rule. Gul Gokay Emel, Cagatan Taskin and Omer Akat [3] applied Apriori rule mining to profile the domestic tourists of Bursa, Turkey and provided suggestions for relevant tourism enterprises. Moreover, Brida *et al.* [4] implemented two-level approach to conduct cluster analysis based on Italian Christmas Market visitors.

Some researchers used the Decision Tree weighting for improving the accuracy of classifier such as Kaewchinporn C., Vongsuchoto N. and Srisawat A. [5] used Decision Tree bagging to weight features for improve the predictive performance of K-Means. Hall M. [6] applied Decision Tree bagging to weight attributes for improve the performance of Naive Bayes classifier. In this study, the researchers enhanced the quality of cluster by applying the Decision Tree bagging weighted to the features of data. After clusters were constructed, tree bagging and attributes weighting were applied. K-Means and Self Organizing Map (SOM) algorithm were then applied to refine the clusters. Useful knowledge can be served to tourism enterprises.

RELATED ALGORITHMS

Two-level clustering

There were some researchers who used two-level clustering for market segmentation. Conventional approach of two-level clustering is the combination of Hierarchical and non-Hierarchical techniques. The advantages of first level are solved the limitation of algorithm on the second level. For example, Punj and Steward [7] presented two-step clustering by binding of Hierarchical Clustering (HAC) and K-Means approach. The HAC was used to find the number of cluster and initial seeds for the input of K-Means algorithm. However, HAC cannot handle the large data set. Moreover, once a decision is made to combine two clusters, it cannot be undone. R.J. Kuo [8] proposed two-stage clustering by combining of SOM and K-Means algorithms. In the first stage, the researcher replaced HAC with SOM for solve the limitation of HAC. His proposed method is slightly better than the conventional two-stage method.

In this present study, we implement SOM and K-Means approach to segment tourist's dataset. SOM algorithm was discovered by Kohonen. The SOM algorithm is used to find the optimum number of cluster by calculating two criterions consist of RMSSTD and RS [9] which defined as:

$$RMSSTD = \sqrt{\frac{\sum_{\substack{i=1..n_c\\ j=1..d}} \sum_{k=1}^{n_{ij}} (x_k - \bar{x}_j)^2}{\sum_{\substack{i=1..n_c\\ j=1..d}} (n_{ij} - 1)}} \qquad (1)$$

$$RS = \frac{\sum_{j=1}^{d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2 - \sum_{\substack{i=1..c\\ j=1..d}} \sum_{k=1}^{n_{ij}} (x_k - \bar{x}_j)^2}{\sum_{j=1}^{d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2} \qquad (2)$$

C	c is the number of cluster
d	d is the number of dimension
$\bar{x}_j$	$\bar{x}_j$ is the mean value of j^{th} jth dimension
n_{ij}	n_{jj} is the number of sample in i^{th} cluster, j^{th} dimension

After the optimum number of cluster was found, the initial seeds were determined. They were used to input to K-Means method in the second step. For further calculation of SOM and basic K-Means algorithm are available in Tan [1].

Decision Tree Bagging and features weighting

Decision Tree bagging is used for features weighting because the factors may have different importance. This method may improve the quality of cluster. Tree bagging uses a Decision Tree algorithm to construct *n* models based on a different diversity of training data. The bagging algorithm was shown as following [5].

Algorithm: Bagging
Input: D: data set;
n: the number of models;
- a learning scheme (e.g., decision tree);
Output: A composite model, M*.

Method

```
a) For i=1 to n do
b) create bootstrap sample, $D_i$, by sampling D with replacement
c) use $D_i$ to derive a model, $M_i$;
d) End for
```

The attributes weighting technique uses the features which appear in the trees. After decision trees were constructed, attributes which appeared in each trees were selected and computed the weight. The weight of attribute was varying on a size of tree and the position of that attribute appearing in that tree. The computation of weight for each attribute was defined as [5].

$$w_{(k.i)} = \left(height\_M_i - j + 1\right) \Big/ \left(height\_M_i + 1\right) \qquad (3)$$

where $w_{k,i}$ is the weight of attribute *k* in tree i, *height_M_i* is the height of tree *i* and j is the level of attribute node *k* in tree *i*.

All attribute weights from each tree models were calculated for an average weight by the following equation.

$$w_k = \left(w_{k,1} + w_{k,2} + ... + w_{k,n}\right) \Big/ n \qquad (4)$$

where w_k is the total weight of attribute k and n is the number of tree models.

Apriori algorithm

Association rule was discovered by Agrawal *et al*. [10]. It is used in the recommendation systems such as www.amazon.com. Apriori was well-known algorithm and popular usage in market basket data analysis. It was used to find the relationship between two or more attributes in the large database. There were two standard measurements such as minimum support (Minsup) and minimum confidence (Minconf). Support was used to evaluate the statistical importance of a set of transactions in database such as Sup (X, D) represented the rate of transactions in D containing the item set X. Confidence represented the rate of transactions in D that contain item set X and also item set Y. It was defined as Conf(X->Y) = Sup (X∩Y)/Sup (X, D). The first step of Apriori algorithm was to detect a large item set with greater than minimum support and the second step was to generate association rules with greater than minimum confidence. Moreover, lift value was used to reduce the possible biases when used the support and confidence values. Lift was defined as Lift = Conf(X->Y)/Sup(Y) [11]. The association rules were useful for many applications such as tourism marketing [3], tourism recommendation systems [12, 13], new product development and customer relationship management [14].

METHODOLOGY

Data collection

This study, the researchers used secondary data accumulated from the Department of Tourism, Ministry of Tourism and Sports, Thailand. The total number of inbound tourist in the data files is 83,402 who arrived to Thailand in a period from 2008 through 2010. Data pre-processing were applied by removing unreliable, missing values and outliers. So, the total number of observation in the sample was 79,473. The attributes of data are the length of stay (days), age (years), gender, occupations, annual incomes (US Dollars), average expenditures (Baht per day), purpose of visit, types of accommodation, tourist origins, places of residence and types of transportation.

Study framework

The specification of this study required the applying of data mining technique to partition the data into segments and to discover the relationship among the features of tourist in each segment, respectively. Using cluster analysis and association rule, this study analyzed tourist behaviors and then extracted knowledge to explore useful information for tourism businesses. The useful knowledge can be helpful for tourism organizations in order to understand the tourist needs and their preferences.

So, the research design was divided into two phases included data clustering phase and association rule mining phase. Preprocessing methods were used to clean the data. Clustering phase was then conducted by two-level clustering and weighting features by decision tree bagging. SOM algorithm was performed to find the optimum number of cluster. The appropriate number of cluster was defined by the computation of RMSSTD and RS. The cluster labels were assigned to build decision tree bagging for attributes weighting. Attributes weighting by decision tree bagging included four cases. Case one, case two and case three consisted of ten trees which each tree was constructed by twenty five, fifty and seventy five percentage of random with replacement tourist data, respectively. Case four was built by overall data for one tree. The features set were weighted and the data set was refined by the clustering algorithms. Each cluster was analyzed when the clustering phase was finished. After some attribute was preprocessed in order to qualify the requirement of Apriori algorithm, association rule mining was performed for each segment. The study framework was illustrated in Figure-1.

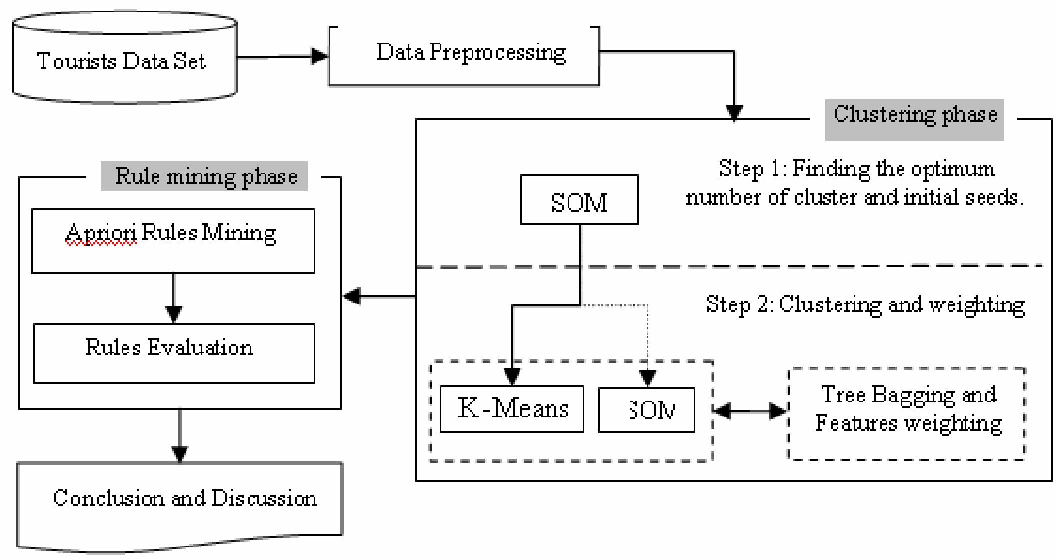
Tourists Data Set
Data Preprocessing
Clustering phase
Step 1: Finding the optimum number of cluster and initial seeds.
SOM
Rule mining phase
Apriori Rules Mining
Rules Evaluation
Step 2: Clustering and weighting
K-Means
SOM
Tree Bagging and Features weighting
Conclusion and Discussion

Figure-1. Study framework.

Table-1. Finding the optimum number of cluster by SOM clustering.

#Cluster	RMSSTD	RS	#Cluster	RMSSTD	RS
2	2.7653	0.1092	9	2.4865	0.2798
3	2.6827	0.1616	10	2.4487	0.3015
4	2.6279	0.1955	11	2.4279	0.3134
5	2.5873	0.2202	12	2.4320	0.3111
6	2.5678	0.2319	13	2.4523	0.2995
7	2.5173	0.2618	14	2.4359	0.3089
8	2.4880	0.2790			

EXPERIMENTAL RESULTS

Clustering results

Finding the optimum number of cluster, Table-1, eleven clusters are obtained from SOM method by the computation of RMSSTD and RS. After cluster labels were assigned to each record, decision tree bagging was constructed for calculate the weight of each attribute. After that, tourists' data was recomputed and compared the RMSSTD and RS between data with weighted attributes and data without weighted attributes. The experimental result indicated that data with weighted attributes give a better quality. It was illustrated on Table-2.

Table-2. The comparison of RMSSTD and RS between SOM and K-Means for data with weighted and not weighted attributes (number of cluster = 11).

Algorithms	Weighting	RMSSTD	RS
SOM	No	2.4279	0.3134
K-Means	No	2.4277	0.3135
SOM	Yes	1.6933	0.4395
K-Means	Yes	1.7001	0.4350

The result of clustering phase was shown on Table-3. The differences of attribute among segments were illustrated on Table-3. The result showed no significant difference on gender, age and occupation. Segment 1 was relatively dominant among the eleven clusters. It comprised over 31 percentage (n=24, 777) of overall tourists (N=79, 473) and can be considers as a homogeneous cluster. The majority of visitor in Cluster 9 traveled to Thailand for business purpose. The expenditure per day of Cluster 9 was highest around 4, 747 Baht by average. The tourists of cluster 11 stayed in Thailand for a longest time when compare with the tourists in other clusters. Thus, they used a variety of transportation' types. Cluster 3 chose domestic airplane for transportation. Moreover, clustering results suggested that the expenditure per day was inverse variation with the length of stay. In other word, the expenditure of cluster 11 was the least but cluster 11 stayed in Thailand for the longest term, the expenditure of cluster 9 was the highest but cluster 9 stayed in Thailand for a short time. Cluster analysis results shown that the tourism organizations can apply the useful knowledge for market planning, promotion design and other related tourism developments.

Association rule mining results

Clustering methods provided the distinct characteristics of segment but cannot show the association among the features of data. Thus, Apriori rule mining was applied to each tourist cluster. The rules with lift value ranging from 1.00 to 1.34 are obtained with minimum rule support (Minsup) of 25% - 35% and minimum rule confidence (Minconf) of 80% - 85%. Table-4 described a part of association rules which discovered from each tourist segment. Four significant rules were demonstrated on Table-4.

The experimental results indicated that the tourist segments provided association rules which were related to the characteristics of each segment. The rules could be transformed to if-then clause. Finally, the rules were evaluated by tourism professionals. Some rule was accepted by minimum criterions but it was rejected by experts filtering. However, the most difficult of this study was the translation of the segmentation and association rules to the suggestion of tourism management. Table-5 shows the activity which tourism organizations should be focused.

Table-3. The significant characteristics of cluster.

Factors	**Clusters**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Sample size (%)	31.18	2.50	5.6	15.47	5.65	3.59	4.41	7.50	10.33	10.14	3.61
Stay' length (days)	6	12	13	8	9	9	11	11	6	8	17
Income $US (%)											
< 20k	43	40	25	44	50	34	36	32	32	32	42
20-40k	36	29	34	33	30	31	34	34	31	36	33
40-60k	12	14	24	14	12	18	17	18	17	22	14
Expenditure (THB)	4,439	3,242	4,551	4,174	3,322	4,268	3,568	4,147	4,747	4,215	2,815
Purpose (%)											
Holiday	100	85	91	92	94	85	93	90		92	92
Business		3	2	3	1	6	2	2	55	2	1
Other		12	7	5	5	9	5	8	45	6	7
Accommodation (%)											
Hotel	100		94			100			91	99	
Resort		35		100	60		32		1		89
Guesthouse		38			18		57	100	1		
Apartment		23	5		12		7		4	1	9
Zone (%)											
America	5	14	11	5	10	12	10	9	5	7	16
East Asia	28	21	11	30	30	22	29	20	21	19	8
Europe	13	27	46	20	24	25	34	38	10	30	55
Oceania	4	7	11	3	6	9	4	12	4	6	8
South Asia	1	5	4	12	6	3	4	3	17	3	1
SEA	33	17	6	21	19	20	12	11	34	26	5
Destination (%)											
Central											
Southern	100		96		100				97	91	94
Northern		88				100		95			
Eastern				100			100				
Transportation (%)											
Plane	6	36	100	4	15	29	10	41	5	0	53
Bus	11	37	23	15	45	14	38	25	10	12	68
Train	6	16	10	3	21	6	8	10	6	4	30
Ferry	7		36	9	26		28	30	5	9	62
Other	65	60	72	45	76	48	63	54	80	42	71

Table-4. Four significant rules for each tourist cluster with Minsup = 25% - 35%, Minconf = 80% - 85% and lift = 1.00 - 1.34.

#	Antecedents	Consequent
R_{11}	P*=Holiday, A=Hotel, E*=4,514-6,832 Baht/day	D*=Central
R_{12}	P=Holiday, Occupation=Professional	A*=Hotel, D=Central
R_{13}	A=Hotel, Age=25-34 years	P=Holiday, D=Central
R_{14}	P=Holiday, Zone=East Asia	A=Hotel, D=Central
R_{21}	A=Guesthouse	D=Northern, P=Holiday
R_{22}	D=Northern, Income=less than 20,000 $US	P=Holiday
R_{23}	Income=less than 20,000 $US	D=Northern
R_{24}	T*=Plane	D=Northern
R_{31}	D=Southern, Zone=Europe	T=Plane, P=Holiday
R_{32}	T=Plane, Age=25-34 years	D=Southern
R_{33}	T=Plane, P=Holiday, Gender=Male	D=Southern, A=Hotel
R_{34}	D=Southern, Age=25-34 years	A=Hotel
R_{41}	D=Eastern, Gender=Male, T=Other	A=Hotel
R_{42}	D=Eastern, Income=20,000-39,999 $US	A=Hotel, P=Holiday
R_{43}	D=Eastern, A=Hotel, Age=25-34 years	P=Holiday
R_{44}	P=Holiday, Income=less than 20,000 $US	D=Eastern, A=Hotel
R_{51}	P=Holiday, Age=15-24 years	D=Central
R_{52}	Income=less than 20,000 $US	D=Central, P=Holiday
R_{53}	D=Central, P=Holiday	A=Resort
R_{54}	P=Holiday, T=Other, Income=less than 20,000 $US	D=Central
R_{61}	D=Northern, Gender=Female	A=Hotel, P=Holiday
R_{62}	A=Hotel, P=Holiday, Gender=Female	D=Northern
R_{63}	D=Northern, Income=20,000-39,999 $US	A=Hotel
R_{64}	Occupation=Professional	D=Northern, A=Hotel
R_{71}	Zone=Europe	D=Eastern, P=Holiday
R_{72}	D=Eastern, T=Bus	P=Holiday
R_{73}	D=Eastern, A=Guesthouse	P=Holiday
R_{74}	Age=25-34 years	D=Eastern
R_{81}	T=Ferry	A=Guesthouse, D=Southern
R_{82}	Occupation=Professional	A=Guesthouse
R_{83}	A=Guesthouse, Income=20,000-39,999 $US	P=Holiday
R_{84}	A=Guesthouse, Age=25-34 years	D=Southern
R_{91}	A=Hotel, P=Business	D=Central
R_{92}	T=Other, Gender=Male, P=Business	D=Central, A=Hotel
R_{93}	D=Central, Occupation=Professional	A=Hotel
R_{94}	D=Central, Zone=South East Asia	A=Hotel
R_{101}	P=Holiday, D=Southern, Gender=Female	A=Hotel
R_{102}	Income=20,000-39,999 $US	A=Hotel
R_{103}	P=Holiday, Gender=Female	D=Southern
R_{104}	D=Southern, Gender=Male	A=Hotel, P=Holiday
R_{111}	D=Southern, T=Plane	P=Holiday, A=Resort
R_{112}	P=Holiday, Zone=Europe	D=Southern, A=Resort
R_{113}	D=Southern, P=Holiday, T=Ferry	A=Resort
R_{114}	D=Southern, Income=less than 20,000 $US	A=Resort

*A = Accommodation, E = Expenditure, D = Destination, P = Purpose, T = Transportation

Table-5. Example of implementation of clustering and association rules mining results.

Knowledge founding	Actionable activities which should be focused on
Most of the extracted rules on cluster 1 belong to "Holiday" purpose, "Hotel" accommodation and "Central" destination.	Tour agencies design package tours as a base for segmentation, Tourism recommendation systems.
Cluster 11 is a "price-conscious" segment.	Marketing managers should be focused on pricing strategies and planning.
Cluster 9 is a "hi-end" segment.	Marketing managers should be focused on the quality of services and products.
For cluster 3, if "Destination=Southern" and "Zone=Europe" then "Transportation=Plane" and "Purpose=Holiday".	Information center should be focused on tourist attraction and transportation for holiday on Southern part of Thailand.
For cluster 9, if "Destination=Central" and "Purpose=Business" then "Accommodation=Hotel".	Hotel at Central part of Thailand should be prepared their location and relationship between business activities.

CONCLUSIONS

Data mining can be extracted hidden information and patterns on the inbound tourist data. This study focused on the role of data mining for tourism industries in Thailand. Data clustering and association rule mining were chose as the data mining methods in order to discover hidden knowledge. The clustering results indicated that inbound tourists in Thailand consisted of various segments with different profiles. Tourist data was segmented into eleven clusters by two-level clustering. Decision tree bagging method was then applied in order to enhance the quality of cluster. The eleven segments were analyzed to identify tourists' behavior patterns and their preferences. The experimental results indicated that the distinct characteristics among clusters can be helpful for tourism organizations in order to define the market planning or strategic making such as tourism management or package tour designing. After eleven segments were analyzed, association rule mining was applied in order to discover the relationship among the features of tourist in each segment. Minimum support and minimum confidence were set to filter the rules which generated by Apriori algorithm. Finally, the rules were evaluated by experts. Association rules mining results indicated that the tourist segments provide association rules which related to the characteristics of each segment. The rules can be implemented on recommendation systems or marketing action which should be focused on the rules found.

Future studies of data mining on tourism can focus on other related features such as cultural, socio-economic variables and values added from incompletely data such as RFM analysis. Finding a good way to weight features is very important for optimizing clustering results of many real world dataset. Moreover, the researchers should be suggested the data collector to collect more information for data analysis and researching.

ACKNOWLEDGEMENT

This work was funded by Faculty of Science and the Graduate School of Kasetsart University.

REFERENCES

[1] Punj G. and Steward D.W. 1983. Cluster analysis in marketing research: review and suggestions for applications, Journal of Marketing Research. 20(2): 134-148 (May).

[2] Kuo, R.J., Ho, L.M., Hu and C.M.: Integration of Self-Organizing Feature Map and K-means Algorithm for Market Segmentation, Computers and Operations Research. 29(11), 1475-1493 (September, 2002).

[3] Tan, P.N., Steinbach, M. and Kumar, V.: Introduction to Data Mining, Pearson Education, Inc., USA. (2006).

[4] Kovacs, F., Legany, C., Babos, A.: Cluster Validity Measurement Techniques, World Scientific and Engineering Academy and Society (WSEAS), pp. 388-393, (2006).

[5] Kaewchinporn, C., Nattakan, V., Vongsuchoto, S.: A combination of Decision Tree Learning and Clustering for Data Classification, In: Proceedings of 2011 Eight International Join Conference on Computer Science and Software Engineering (JCSSE), pp. 11-13, (2011).

[6] Agrawal, R., Imilienski,T., Swami, A.: Mining association rules between sets of items in large databases, In: Proceedings of the 1993 ACM SIGMOD international conference on management of data, pp. 207-216. ACM New York, NY, USA (1993).

[7] Emel, G.G., Taskin, C., Akat, O.: Profiling a Domestic Tourism Market by Means of Association Rule Mining, Anatolia: An International Journal of Tourism and Hospitality Research 18(2), 334-342 (2007).

[8] Aghdam A. R., Mostafa, K., Dong, C. et al.: Finding Interesting Places at Malaysia: A data Mining

Perspective, Mathematics and Comuters in Contemporary Science, pp. 89-93, (2013).

[9] Juwattanasamrn P., Supattranuwong, S., Sinthupinyo, S.: Applying Data Mining to Analyze Travel Pattern in Searching Travel Destination Choices, The International Journal of Engineering and Science (IJES), 2(4), 38-43 (2013)

[10] Liao S., Chen, Y., Deng, M.: Mining customer knowledge for tourism new product development and customer relationship management, Expert Systems with Applications 37(6), 4212-4223 (June, 2010).

[11] Wong, J., Chen, H., Chung, P., Kao, N.: Identifying Valuable Travelers and their Next Foreign Destination by the Application of Data Mining Techniques, Asia Pacific Journal of Tourism Research 11(4), 355-373 (2006).

[12] Brida, J. G., Disegna, M., Osti, L.: Segmenting Visitors of Cultural Events by Motivation: A sequential Non-linear Clustering Analysis of Italian Christmas Market Visitors, Expert Systems with Applications 39(13), 11349-11356 (October, 2012).

[13] Hall, M.: A Decision Tree-Based Attribute Weighting Filter for Naïve Bayes, Knowledge-Based Systems 20(2), 120-126 (March, 2007).

[14] Wang, Y. F., Chuang, Y. L., Hsu, M. H., Keh, H. C.: A personalized recommender system for the cosmetic business, Expert Systems with Applications 26(3), 427-434 (April, 2004).