Segmentation of Domestic Tourist in Thailand by Combining Attribute Weight with Clustering Algorithm

Prapassorn Hayamin and Anongnart Srivihok
Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand
Email: {g5614401831, fsciang}@ku.ac.th

***Abstract*—The tourism industry is growing up and competing relatively high. In Thailand, tourism is one of the main industries that can generate a large amount of domestic turnover rate. And tourist information in Thailand is stored in large quantities. It is difficult to understand the needs of tourists. Therefore, this study presents segmentation of domestic tourist in Thailand by combining attribute weight with clustering algorithm. The study used two step algorithms, in the first step, Self-Organizing Maps (SOM) was used to determine the optimum number of clusters which an input parameter to K-Means and Fuzzy C-Means. Then, using SOM, K-Means and Fuzzy C-Means algorithms combine with feature weighting techniques based on Correlation Coefficient (CC), Information Gain Ratio (IGR), Gini Index and Principal Components Analysis (PCA) for clustering the tourists clusters. The quality of cluster was measured by Davies Bouldin Index (DB), Root Mean Square Standard Deviation (RMSSTD) and R Square (RS). The results of this study might be used for tourism management and entrepreneur tour and travel can be used for decision making and business planning.**

***Index Terms*—domestic tourism, clustering, attribute weight, SOM, K-means, fuzzy C-means**

I. Introduction

Tourism is one of the most important industries for economic development and job. It can make revenue is causing a turnover in the country. Key Points for domestic tourism is sustainable tourism development. Therefore, tourism organizations need to understand the tourism patterns and preferences of tourists. To be used in tourism planning and managing to design products or services in accordance with the needs of tourists. However, it is difficult to understand the data collected by the Department Tourism, Ministry of Tourism and Sports of the Kingdom of Thailand. Thus, the purpose of this study was to extract knowledge from tourism data using clustering techniques, which are important tools for large data base that will be set into small number of clusters by considering similarities within clusters to discover useful knowledge. Therefore, this present experiment introduced (1) algorithms for clustering tourists using SOM (Self-Organizing Maps), K-Means and Fuzzy C-Mean and (2) attributes weighting techniques base on Correlation Coefficient (CC), Information Gain Ratio (IGR), Gini Index and Principal Components Analysis (PCA) were used to segment tourists into clusters and to analyze the statistics of each tourists segment.

This paper is organized as follows. In section II, discusses about related work, Section III presents related clustering algorithms. The attribute weighting techniques shows in section IV. Section V describes the measurements of clustering. For section VI, we will explain the experiment and results. Finally, Section VII shows our conclusions.

II. Related Work

In the clustering algorithms, K-Means, Fuzzy C-Means (FCM) and Self-Organizing Maps (SOM) are popular and have more powerful performance in clustering algorithms which are widely used in real world applications. S. Ghosh *et al*. [1] proposed comparative analysis of K-Means and FCM algorithms which shows the time complexity of the K-Means algorithm is better than FCM algorithm. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm. C. Lu *et al*. [2] apply two kinds of cluster method (K-Means and DBSCAN) to the data from Dagang oil well for analyze specific wells to see what can be done to improve the efficiency. So, the clustering method is only used to segment the different types of oil pumping unit or identify the quality of oil. H.L.T. Trang *et al*. [3] introduced clustering techniques on inbound tourism market segmentation of the Andaman cluster of Thailand by using two step clustering technique which combine with HAC and K-Means (HACKM). In the first step, HAC method indicated that the data set divided into three segments and then using K-Means to partition the data set. W. Niyagas *et al*. [4] proposed a clustering technique for e-Banking customers by combining algorithm between SOM and K-Means. K-Means used for clustering data and find the number of clusters or k value is derived from SOM guided by computation of the Standard Deviation, Root Mean Square Standard Deviation (RMSSTD) and R Square (RS). Then, customer data are segmented based on

Manuscript received February 7, 2018; accepted May 15, 2018.

doi: 10.12720/jait.9.2.39-44

customer transactions and their behaviors. W. Yotsawat *et al*. [5] proposed partition the inbound tourists by K-Means clustering and including statistical analysis in each tourist segment. For clustering techniques using SOM to determine the best number of cluster guided by voting the optimum value of Silhouette index, Root Mean Square Standard Deviation (RMSSTD) and R Square (RS). Then, using K-Means to refine the tourist clusters. W. Yotsawat *et al.* [6] propose the development of tourist segment models, and analyzed each segment, based on domestic tourists to Phranakhon Si Ayutthaya province, Thailand. They used two step cluster approach that consisted of the Hierarchical technique and K-Means algorithm. Hierarchical clustering was used to find the optimal number of clusters and the initial seeds. And then, the number and seeds from the previous step are used as input to the K-Means algorithm. Characteristics of these segments were analyzed for tourism stakeholders to plan new products and services of Ayutthaya Tourism.

For attribute weighting can be used in conjunction with a data mining according to these studies. J. Wu *et al*. [7] studies about complex structure models for BNCs and then carry out experimental studies to investigate the effectiveness of the attribute weighting strategies for complex BNCs. In order to learn proper weight values for BNCs, they have proposed many methods to evaluate the importance of attributes, including Gain Ratio, CFS (Correlation-based Feature Selection) and each other. Experiments and comparisons benchmark data sets demonstrate that attribute weighting technologies just slightly outperforms unweight complex BNCs. J. H. Choi *et al*. [8] modified correlation coefficient (CC) by giving the geodesic distance weights to the reconstructed sources to reflect the geometric information of cortical surface. The new evaluation metric named Weighted Correlation Coefficient (WCC) was proposed to combine the advantages of both types of evaluation metrics and showed enhanced performances compared to the conventional metrics.

III. Clustering Techniques

A. Self-Organizing Maps (SOM)

SOM was developed by Kohonen in 1982. SOM is an algorithm for unsupervised competitive learning [9]. It is a neural network algorithm that networks have only input layers and output layers and we can use SOM for clustering data without knowing the class memberships. Therefore, SOM is used to solve the problem of K-Means by determining the number of clusters for input into K-Means and can provides good clustering results and large data sets can be clusters. SOM is a popular algorithm and has been extensively used in different fields of research. The basic algorithm of SOM as follows [10].

Algorithm Basic SOM Clustering.

1. Randomly initialize all weights.
2. Select input vector.
3. Calculate distance between input vectors and all weight to find the closest output node.
4. Define a neighborhood function that allows to identifying the output node, to be updating in the next step.
5. Update the winner's weight.

B. K-Means

K-Means clustering was introduced by James MacQueen in 1967. K-Means is a technique of clustering that aims to partition n data points into k clusters in which each point belongs to the cluster with the nearest mean [11]. It is popular for cluster analysis in data mining but required assigning the number of clusters. Therefore, two step clustering was proposed by some researchers which can find the input requirement of K-Means in the first step [5]. The basic algorithm of K-Means as follows [10].

Algorithm Basic K-Means Clustering.

1. Randomly select k points as initial centroids.
2. Repeat.
3. From k clusters by assigning each point to its closest centroid.
4. Recomputed the centroid of each cluster.
5. Until centroids do not change.

C. Fuzzy C-Means (FCM)

FCM algorithm is one of the most popular fuzzy clustering techniques which able to determine, and in turn, iteratively update the membership values of a data point with the pre-defined number of clusters. Thus, a data point can be the member of all clusters with the corresponding membership values [12].

FCM is useful when the required numbers of clusters are predefined. Therefore, the algorithm tries to put each of the data points to one of the clusters. It does not decide the absolute membership of a data point to a given cluster but it calculates the likelihood which a data point will belong to that cluster. Performance depends on initial centroids which calculated as being the mean of all points and weighted by their degree of belonging to the cluster. The advantages of the FCM are simple and fast algorithms that can be applied to a wide variety of data types. Providing the best result for overlapped data set and comparatively better than K-Means algorithm.

IV. Attribute Weighting

Attribute weighting is often used for data mining and the way to learn the attribute weights is the most important part for attribute weighted. Therefore, assigning different weight values to attributes can potentially help improve the clustering performance. In this section, we refer to the four attribute weighting techniques as follows.

A. Correlation Coefficient (CC)

CC [8] is a statistical technique that can show how two variables in a data set are related. It can calculate the relevance of the attributes by computing the value of correlation for each attribute of the input data set. The weight by CC is based upon correlation and absolute or squared value of correlation as attribute weight. The higher the weight of an attribute, the more relevant it is considered. Result of a CC it ranges from -1 to 1. If it

closes to 1, it would indicate that the variables are positively linearly related. For -1, it indicates that the variables are negatively linearly related. And for 0, it would indicate a weak linear relationship between the variables.

B. Gini Index

The Gini Index is a measure of inequality of a distribution which often used to measure income distribution or wealth distribution among a population. The Gini Index expressed as a percentage ranges from 0 to 1. Here, 0 representing perfect equality and 1 representing perfect inequality. We propose the weight by Gini Index using Rapid Miner, which calculates the weight of attributes with respect to the label attribute by computing the Gini index of the class distribution, if the given data set would have been split according to the attribute.

C. Information Gain Ratio (IGR)

IGR used by Quinlan in 1993 [13]. It was proposed to solve the drawback of information gain by dividing each attribute's IG score by the information encoded in each attribute itself. In 2004, Zhang and Sheng [14] argued that an attribute with a higher gain ratio value deserves a higher weight. In their studies, they proposed a gain ratio weighted method that calculates the weight of an attribute from a data set, as shown in the following Eq. (2).

The information gain ratio *Gain Ratio(S,A)* can be calculated as shown in the following.

$$\text{GainRatio(S, A)} = \frac{\text{Gain(S, A)}}{\text{Split\_info(S, A)}} \tag{1}$$

The information gain *Gain (S,A)* is the amount of decrease of the entropy about the probabilistic distribution of the data set *S*. The split information *split_info(S,A)* is the entropy about partitioning the attribute *A*. The IGR *Gain Ratio(S,A)* is defined as the ratio of the information gain to the split information.

The information gain ratio weighted method can be calculated as shown in the following.

$$\text{w}_\text{j} = \frac{\text{GainRatio}(\text{A}_\text{j}) \times \text{n}}{\sum_{\text{j=1}}^{\text{n}} \text{GainRatio}(\text{A}_\text{j})} \tag{2}$$

Therefore, from equation (2) showing the calculated the weight of attributes with respect to the label attribute by using the information gain ratio. The higher the weight of an attribute, the more relevant it is considered.

D. Principal Components Analysis (PCA)

PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It's often used to visualize genetic distance and relatedness between populations. We using Rapid Miner, for create attribute weights of the dataset by using a component created by the PCA.

V. Performance Measurement

In order to have some measure to help deciding on or help confirming our impression about the number of clusters, the Root Mean Square Standard Deviation (RMSSTD), R-Squared (RS) and Davies Bouldin Index (DB) statistics can be employed. The details of those measurements described follows:

A. Root Mean Square Standard Deviation (RMSSTD)

RMSSTD is an index of the variance of clusters that measure of the homogeneity of the clusters that have been formed. The less value of RMSSTD means better clustering [15]. The formula of RMSSTD defined on (3).

$$RMSSTD = \sqrt{\frac{\sum_{j=1..d}^{i=1..c} \sum_{k=1}^{n_{ij}} (x_k - \bar{x}_j)^2}{\sum_{j=1..d}^{i=1..c} (n_{ij} - 1)}} \tag{3}$$

B. R Square (RS)

RS is an index for measurement dissimilarity of clusters. The values have ranging from 0 to 1 where 0 means high similarity among the clusters, 1 means in opposite [15]. The formula of RS defined on (4).

$$RS = \frac{\sum_{j=1}^{d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2 - \sum_{j=1..d}^{i=1..c} \sum_{k=1}^{n_{ij}} (x_k - \bar{x}_j)^2}{\sum_{j=1}^{d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2} \tag{4}$$

Notation of RMSSTD and RS follow:
c is the number of cluster, d is the number of dimension,
n_j is the number of sample in j[th] dimension,
n_{ij} is the number of sample in i[th] cluster, j[th] dimension,
x_k is the sample k[th] and $\bar{x}_j$ is the mean value of j[th] dimension.

C. Davies Bouldin Index (DB)

DB [16] is an index for measuring the quality of clustering that is very popular. To define the DB index, we need to define the dispersion measure, the cluster similarity measure and the dissimilarity measure. In the dispersion S_i of cluster C_i (5) and d_{ij} is the dissimilarity measure between clusters C_i and C_j (6) are defined as:

$$S_i = \frac{1}{\|C_i\|} \sum_{x \in C_i} d(x, V_i) \tag{5}$$

where V_i denote the center of the cluster C_i and $\|C_i\|$ is the size of the cluster C_j, and:

$$d_{ij} = d(V_i, V_j) \tag{6}$$

Then, the formula of DB is defined as:

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \tag{7}$$

where n_c is the number of clusters at a given step in hierarchical clustering and R_i is defined as:

$$R_i = \max_{i=1...n_c, i \neq j} (R_{ij}), i = 1...n_c \tag{8}$$

where R_{ij} is the similarity measure between clusters C_i and C_j, and defined as:

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \quad (9)$$

DB index quantifies the average most similarity between a cluster by small values of DB correspond to clusters that are compact, and whose centers are far away from each other. Thus, the number of clusters that minimizes DB is taken as the optimal number of clusters.

VI. Experiments and Results

A. Data Set

The data used in this study are domestic tourist data from Department Tourism, Ministry of Tourism and Sports of the Kingdom of Thailand. A total of dataset are 36,399 rows and 20 attributes of data are Region, Sex, Age, Status, Education, Occupation, Stay (stay overnight or not), Number of travel, Travel companions, Decision maker in travel, Objective, Activity, Travel range, Travel type, Vehicle, Accommodation, Number of days, Expenditure, Expenditure type and Income were selected.

B. Framework

This study involves the clustering and attributes weighting techniques. The clustering techniques are using SOM, K-Means and FCM. Attribute weighting techniques using CC, IGR, Gini Index and PCA. The framework explains follow (Fig. 1):

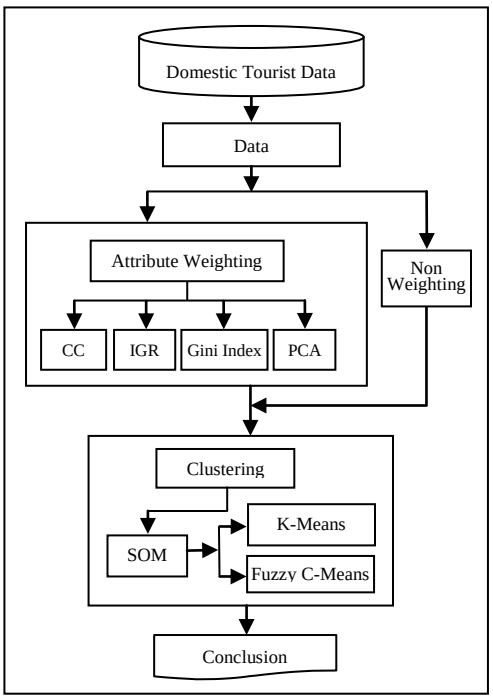

Figure 1. Study Framework.

1) Data preprocessing by removing unreliable, missing values and outliers. Duplicate attributes are excluded: resident data, such as district and sub-district, which are features that are too specific to describe. We transformed some attribute values in order to qualify for the requirements of the algorithms and appropriate formats.

2) Find the best number of clusters. we used SOM algorithm to determine the optimum number of clusters guided by voting the best value of DB, RMSSTD and RS. We processed the number of clusters from 2 to 12. The best number of cluster in the term of DB was 4 clusters (0.793) showed in Fig. 2 but only DB is not enough to identify the natural number of domestic tourist cluster. So, we consider the value of RMSSTD and RS as majority to vote the best number of clusters. The best results of RMSSTD and RS are in 8 clusters which the value of dissimilarity within cluster only 0.902 when the value of dissimilarity between cluster was 0.830. The plots of RMSSTD and RS were showed in Fig. 3 and Fig. 4. Therefore, we summarized that the best number of cluster generated by SOM was 8 clusters. It would be the input number of cluster to the next step.

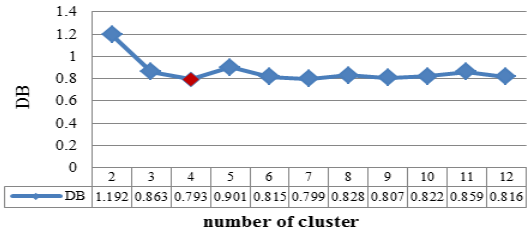

Figure 2. DB for the number of cluster by SOM.

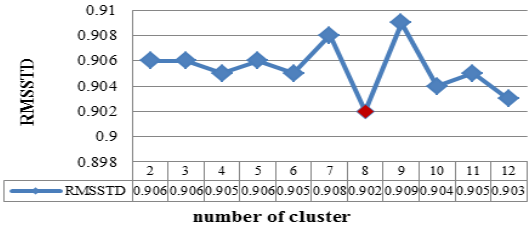

Figure 3. RMSSTD for the number of cluster by SOM.

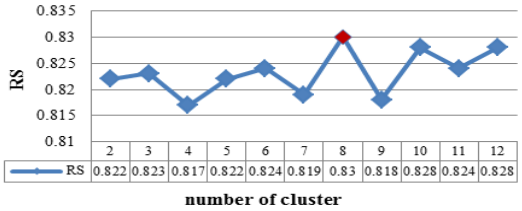

Figure 4. RS for the number of cluster by SOM.

3) We using SOM, K-Means and FCM algorithms to clustering of domestic tourists data. Using k value obtained from the previous step as input parameter for K-Means and FCM. Cluster quality measured by three performance indices are DB, RMSSTD and RS. We comparing the result clustering of SOM, K-Means and FCM algorithms with feature weighting techniques such as CC, Gini Index, IGR and PCA. From Table I, K-Means (IGR) outperforms SOM (IGR) in the terms of DB and RMSSTD while RS was the best with SOM (IGR). So, K-Means (IGR) would be able to partition the domestic tourists into 8 clusters. The percentage of domestic tourists of each clusters were illustrate in Fig. 5.

TABLE I. COMPARE THE QUALITY OF CLUSTERING WITH SOM, K-MEANS AND FCM ALGORITHMS COMBINED WITH CC, GINI INDEX, IGR AND PCA WEIGHTING TECHNIQUES

Algorithm	Evaluators		
	DB	RMSSTD	RS
SOM	0.828	0.902	0.830
SOM (CC)	0.403	0.296	0.976
SOM (Gini Index)	0.398	0.305	0.942
SOM (IGR)	0.410	0.294	**0.993**
SOM (PCA)	0.407	0.382	0.985
K-Means	0.423	0.726	0.527
K-Means (CC)	0.074	0.242	0.791
K-Means (Gini Index)	0.058	0.243	0.792
K-Means (IGR)	**0.053**	**0.241**	0.794
K-Means (PCA)	0.067	0.267	0.790
Fuzzy C-Means	0.456	0.728	0.531
Fuzzy C-Means (CC)	0.082	0.243	0.789
Fuzzy C-Means (Gini Index)	0.063	0.243	0.790
Fuzzy C-Means (IGR)	0.057	0.243	0.791
Fuzzy C-Means (PCA)	0.071	0.267	0.789

C. *Segmentation of Domestic Tourist in Thailand*

Study on domestic tourist behavior in Thailand. Using the result of the previous step.

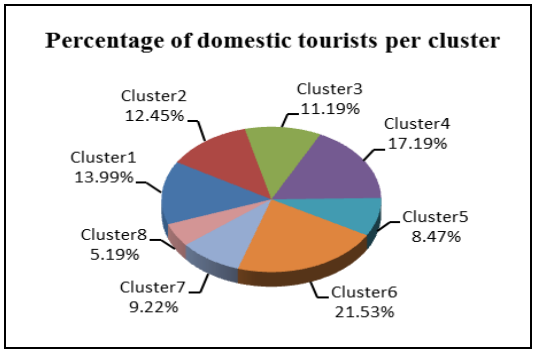

Figure 5. Pie chart shows the percentage of domestic tourists in each cluster.

The pie chart indicates that segment 6 is the biggest cluster. It comprises over 21.53% (its member is 7,835) of overall 36,399 tourists. Cluster 8 is the smallest cluster and consists of 5.19% (its member is 1,888). The character of domestic tourist cluster can be summarized as follows.

Cluster 1 is consisted of only 5,094 tourists or about 13.99% of overall tourists. The range of ages in this cluster is 55-65 years and most of the tourists are women. They have a marital status as a housewife, and are retire from work. The average monthly income of this group is less than 10,000 Baht. The total expenditures of this cluster are in highly, about 10,714.07 Baht. They spend on food and drinks, souvenirs and vehicle costs throughout the travel. The main purposes of traveled are to visit relatives or friends. They travel with family or relatives on weekend and do not stay overnight. Mainly tourists in this cluster travel in the Central of Thailand by car or private car.

Cluster 2 is consisted of only 4,530 tourists or about 12.45% of overall tourists. The range of ages in this cluster is 35-44 years and mostly are women which are farmers or have small business. The average monthly income of this cluster is less than 10,000 Baht. The tourists in this cluster are the least important in term of economics because they spend a small expenditure are about 454.79 Baht. Mostly spend on vehicle costs throughout the travel. The main purpose of trip is to visit relatives or friends by car or private car. They select to stay with relatives or friends about 2 days on weekend. Mainly tourists in this cluster travel in the Lower North of Thailand.

Cluster 3 is consisted of only 4,354 tourists or about 11.96% of overall tourists. The range of ages in this cluster is 35-44 years and mostly are women which are employees or have small business. Earned 10,000–15,000 Baht or more than 15,000 Baht. They spend the medium expenditures are about 1,496.46 Baht. Mostly spend on food and drinks, souvenirs and vehicle costs throughout the travel. Mainly tourists in this cluster travel in the Central of Thailand by car or private car to visit relatives or friends and shopping on weekend.

Cluster 4 is consisted of only 6,257 tourists or about 17.19% of overall tourists. The range of ages in this cluster is 25-34 years and most of the tourists are women. The tourists in this cluster have a variety of occupations and earned more than 15,000 Baht. The total expenditures are about 991.04 Baht. They spend on food and drinks, souvenirs and vehicle costs throughout the travel. The main purposes of traveled to visit relatives or friends and travel to the general on weekday. Mainly tourists in this cluster travel in Southern of Thailand.

Cluster 5 is consisted of only 3,083 tourists or about 8.47% of overall tourists. The range of ages in this cluster is 25-34 years and mostly are employees. Earned 10,000-15,000 Baht or more than 15,000 Baht. The total expenditures of this cluster are in medium rate, about 2,948 Baht. They spend on food and drinks, souvenirs and vehicle costs throughout the travel. Mainly tourists in this cluster travel in Bangkok, Nonthaburi, Pathum Thani and Samut Prakan by car or private car to visit relatives or friends and travel about 3 days on long weekend.

Cluster 6 is the biggest cluster. It consisted of 7,835 tourists or about 21.53% of overall tourists. The range of ages in this cluster is 35-44 years and mostly are women which are farmers. Although the average monthly income of this group is less than 10,000 Baht but the total expenditures of this cluster are in higher rate, about 5,022 Baht. They spend on food and drinks, souvenirs and vehicle costs throughout the travel. Mainly tourists in this cluster travel in the Northeastern of Thailand by car or private car with family or relatives to visit relatives or friends and travel to the general about 3 days on weekend. They stay overnight with relatives or friends.

Cluster 7 is consisted of only 3,358 tourists or about 9.22% of overall tourists. The range of ages in this cluster is 35-44 years and mostly are women which are farmers or have personal business. The average monthly income of this group is less than 10,000 Baht. They spend the medium expenditures are about 2,000 Baht. Mostly they

spend on food and drinks, souvenirs and vehicle costs throughout the travel. Mainly tourists in this cluster travel in the Northern of Thailand by car or private car to travel about 3 days on weekday. They stay overnight with relatives or friends and may be go back.

Cluster 8 is consisted of only 1,888 tourists or about 5.19% of overall tourists. The range of ages in this cluster is 35-44 years and mostly are women which are employees or have small business. Earned 10,000-15,000 Baht or more than 15,000 Baht. It is the highly important in term of economics because the expenditures of tourist are in highest rate, about 26,989 Baht. They spend on food and drinks, souvenirs and vehicle costs throughout the travel. The main purposes of traveled to travel and visit relatives or friends by car or private car. They stay overnight with relatives or friends about 5 days on long weekend. Mainly tourists in this cluster travel in the Eastern of Thailand.

VII. Conclusion

This study proposed the segmentation of domestic tourist in Thailand by combining attribute weight with clustering algorithm. In the clustering phases, SOM was used to estimate the optimum number of clusters which was an input parameter to K-Means and FCM. The experimental results of SOM provided 8 clusters. Then, data set were clustered by SOM, K-Means and FCM algorithm with feature weighting techniques including Correlation Coefficient (CC), Information Gain Ratio (IGR), Gini Index and Principal Components Analysis (PCA). The measures were DB, RMSSTD and RS. The algorithms performing the best result was K-Means with Information Gain Ratio weighting technique. Based on the clustering results, the cluster 6 was the largest cluster, it is consisted of 21.53% of overall tourists. Tourists in this cluster traveled in the Northeastern of Thailand on weekend and spent the medium expenditures. While, the tourists in cluster 8 was the smallest only 5.19% of overall tourists. They traveled in the Eastern of Thailand on long weekend and spent the highest expenditures. So, the results of clustering can help entrepreneur tour and travel agencies for making the marketing plan of domestic tourism in Thailand.

Acknowledgment

This work was supported by the Department of Computer Science, Faculty of Science, Kasetsart University.

References

[1] S. Ghosh and S. K. Dubey, "Comparative analysis of K-means and fuzzy CMeans algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 4, 2013.

[2] C. Lu, Y. Shi, Y. Chen, S. Bao, and L. Tang, "data mining applied to oil well using K-means and DBSCAN," in *Proc. International Conference on Cloud Computing and Big Data*, July 2016.

[3] H. L. T. Trang and P. Kullada, "Inbound tourism market segmentation of the Andaman cluster, Thailand," M.S. thesis, Prince of Songkla Unive., Songkla, Thailand, 2009.

[4] W. Niyagas, A. Srivihok, and S. Kitisin, "Clustering e-Banking customer using data mining and marketing segmentation," in *Proc. ECTI International Conference*, May 2006, pp. 63-69.

[5] W. Yotsawat and A. Srivihok, "Inbound tourists segmentation with combined algorithms using K-means and decision tree," in *Proc. International Join Conference on Computer Science and Software Engineering*, May 2013, pp. 189-194.

[6] W. Yotsawat and A. Srivihok, "Thai domestic tourists clustering model using machine learning techniques: Case study of phranakhon si Ayutthaya Province, Thailand," *International Information Institute*, vol. 19, no. 2, pp. 413-422, 2016.

[7] J. Wu, Z. Cai, S. Pan, X. Zhu, and C. Zhang, "Attribute weighting: How and when does it work for Bayesian network classification," in *Proc. International Joint Conference on Neural Networks*, July 2014.

[8] J. H. Choi, M. Kim, L. Feng, C. Lee, and H. Jung, "A new weighted correlation coefficient method to evaluate reconstructed brain electrical sources," *Journal of Applied Mathematics*, 2012.

[9] K. J. Cios, W. Pedry, R. W. Swiniarski, and A. Kurgan, "Data mining: A knowledge discovery approach," in *Springer Science + Business Media*, LLC, 2007.

[10] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, USA: Peason Education, Inc., 2006.

[11] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Journal of Applied Statistics*, vol. 28, pp. 100-108, 2013.

[12] S. Chattopadhyay and D. K. Pratihar, "A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms," *Computing and Informatics*, vol. 30, pp. 701–720, 2011.

[13] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[14] H. Zhang and S. Sheng, "Learning weighted naive bayes with accurate ranking," in *Proc. International Conference on Data Mining Series*, 2004, pp. 567-570.

[15] F. Kovács, C. Legány, and A. Babos, "Cluster validity measurement techniques," *World Scientific and Engineering Academy and Society*, pp. 388-393, 2006.

[16] J. C. R. Thomas, M. M. Cofre, and M. Santos, "New version of davies-bouldin index for clustering validation based on hyperrectangles," in *Proc. Chilean Conference on Pattern Recognition*, 2014, p. 13.

Prapassorn Hayamin received her B.S. in information technology with Second Class Honours from Kasetsart University in 2013. She is currently studying for a master's degree of computer science at Kasetsart University, Thailand. Her research interests are data mining.

Anongnart Srivihok is an associate professor at the Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand. She has a doctorate degree in information systems from Central Queensland University, Australia. Her research areas include data mining, knowledge management, decision support.