Analysis of Melaka tourists' destinations using k-means clustering

U.S. Muslim[1], N. Ahmad[1,*], Z.I.M. Yusoh[1], N. Mohamad[2]
[1]Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
[2]Centre for Telecommunication Research and Communication, Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

[*]Corresponding author's email: norashikin@utem.edu.my

ABSTRACT: The term Tourism 4.0 has been used to relate to the tourism industry in the industrial revolution 4.0 era that aims to improve the data discovery and enhance tourist experience utilizing automation and innovative technology including data mining. Diverse data mining techniques have extensively been adopted in tourism research to discover patterns that are useful for example in segmenting tourists and promoting related attractions. This paper presents the analysis of tourists' destinations in Melaka, Malaysia using the k-means algorithm. The algorithm utilizes the tourist information as well as the meteorological data to examine its effect on the choice of tourist destinations. The results show that k-means could give some insights in discovering knowledge about the tourists' destinations given the time, number of tourists and meteorological data.

Keywords: *tourism data mining; k-means clustering; meteorological data*

1. INTRODUCTION

Tourism 4.0 may be referred to as a new tourism value eco-system built upon a highly technology-based service production paradigm and supported by the common principles of Industry 4.0 [1]. It involves interconnectivity, automation and using innovative and smart technology including data mining.

Tourism is one of the vital industries in Malaysia and Melaka in particular. It is important to attract tourists to visit and stay in Melaka as it contributes to the socio-economic development of the country and the state. The analysis of tourist data is beneficial for the state's tourism planning to ensure a steady flow and an increased number of tourists visiting Melaka.

Data mining techniques have been widely used in tourism for example in analyzing tourist profiles for market segmentation and recommend tourist destinations. In previous studies, Rodriguez et al. [2] has developed and applied a hierarchical clustering technique for geo-localized data from smartphones to discover significant market groups relating to tourism. Abbasi-Moud et al. [3] developed a system of tourist recommendations that analyses the preferences of users to give customized suggestions using semantic clustering and sentiment analysis. In another study, Roshan et al. [4] used biometeorology clusters to analyse climate change in Iran and climate-conditioned place for tourism at a certain time.

2. CLUSTERING TOURISM DATA USING K-MEANS

2.1 K-means algorithm

K-means clustering algorithm groups the objects into k number of clusters based on attributes or features. The grouping of objects is done by minimizing the sum of squares of distances, i.e., a Euclidean distance between data and the corresponding cluster centroid [5]. Each sample will be assigned to the cluster with the closest centroid in each iteration and the centroid values will be recalculated. The iteration stops when there is no change in the assignment of the clusters. In k-means clustering, the number of clusters, k must be decided in advance.

2.2 Experimental Setup

The dataset used in the experiment consists of 2365 entries of Melaka tourism data combined with the meteorological data from 2016 to 2020. The attributes or features are month, year, number of domestic tourists, number of foreign tourists, average humidity, average temperature, and total rainfall. Each sample has a destination as a label for example Zoo Melaka, Water World, Melaka International Bowling Centre, Melaka Planetarium and Memorial Kemerdekaan and others. For cluster analysis, the tourist destinations are classified into five main categories based on Webster et al. [6]. The categories are Heritage (1080 samples), Amusement (468 samples), Recreational (732 samples), Commercial (24 samples) and Industrial (60 samples).

The k-means clustering was carried out using Rapidminer. Several parameters need to be initialized before processing the data, such as the number of clusters, k, number of maximum runs, and the measure type. We used the Mixed Euclidean Distance measure as there are non-numeric attribute values in the dataset. In the experiment, k=5 and k=10 have been used. We choose k=5 as there are five main categories in the dataset. The selection of k=10 will enable us to observe if there are new patterns in the formation of the clusters when we have a greater number of clusters.

3. RESULTS AND DISCUSSION

The results obtained from the experiments are summarized in Table 1 (for k=5) and Table 2 (for k=10). The values shown in the tables are the percentage of data

that belongs to the five categories in each cluster. In this discussion, the centroid values for the rainfall attribute (Rainfall Ct.) are examined and discussed.

Table 1 Result of k-means clustering with $k = 5$

	Cluster				
Category	0	1	2	3	4
Heritage	**50.4**	**47.4**	31.9	39.1	42.9
Amusement	26.0	18.8	25.4	1.6	0.00
Recreational	23.1	29.1	**42.8**	**59.4**	**57.1**
Industrial	0.4	3.3	0.00	0.0	0.0
Commercial	0.0	1.4	0.00	0.0	0.0
Rainfall Ct.	171	164	160	167	212

Table 2 Result of k-means clustering with $k = 10$

	Cluster				
Category	0	1	2	3	4
Heritage	**49.1**	31.1	32.5	37.0	**57.7**
Amusement	16.9	11.1	23.8	0.0	9.6
Recreational	28.7	**57.8**	**43.7**	**63.0**	32.7
Industrial	3.8	0.0	0.0	0.0	0.0
Commercial	1.6	0.0	0.0	0.0	0.0
Rainfall Ct.	165	160	161	168	162
	5	6	7	8	9
Heritage	33.0	0.0	42.9	**69.5**	33.3
Amusement	30.7	0.0	0.0	5.7	**56.7**
Recreational	**36.3**	**100.0**	**57.1**	23.8	10.0
Industrial	0.0	0.0	0.0	1.0	0.0
Commercial	0.0	0.0	0.0	0.0	0.0
Rainfall Ct.	161	138	212	167	180

As shown in Table 1, cluster 0 has the largest percentage of Heritage category with 50.4%, followed by the Amusement category with the rainfall centroid of 171. Cluster 1 also has the highest percentage of Heritage with 47.4%, followed by Recreational at 29.1%. Cluster 2 has the highest percentage of Recreational with 42.8%, and the rainfall centroid is 160 which is the lowest. This suggests that when the rainfall amount is low, most tourists have chosen to do some recreational activities. Cluster 3 and 4 also show the highest percentage of the Recreational category. The rainfall centroid for cluster 4 which is 212 is the highest, and a closer look at the total number of tourists in cluster 4 also shows that it is the highest among other clusters. This could be due to the school holiday season between September to December.

In Table 2, the samples have been divided into 10 clusters where cluster 0, 4 and 8 have the highest number of percentages for the Heritage category, while Recreational category has the highest percentage in six clusters (cluster 1, 2, 3, 5, 6 and 7). From Table 2, it is found that cluster 6 has the lowest rainfall centroid value. However, it has only one category that is Recreational with one record. This is due to the formation of small clusters when we set k=10 in the experiment. The rainfall centroid value in cluster 7 is the highest and the cluster is populated by only Heritage and Recreational categories. This indicates the tourists' interest in visiting those places despite the wet season. It can be seen from Table 1 and 2 that Heritage and Recreational are the most popular destinations regardless of the rainfall amount in almost all clusters.

4. CONCLUSION

This paper has presented the use of the k-means algorithm in discovering knowledge about tourists' destinations from the tourism and meteorological data. The finding, while preliminary, suggests that more data analysis of the tourists and destinations can be carried out using the k-means clustering. Further research might include an in-depth analysis of the clusters including other attributes centroid values. It would also be interesting to see the hierarchical relationship that might exist between the clusters at different numbers of k. We are also looking forward to using other features and clustering algorithms and compare the results.

ACKNOWLEDGEMENT

We would like to thank Melaka Tourism and the Department of Meteorology, Malaysia for the datasets. This work is supported by a short-term grant from Universiti Teknikal Malaysia Melaka (UTeM) (PJP/2020/FTMK/PP/ S01798).

REFERENCES

[1] T. Pencarelli, "The digital revolution in the travel and tourism industry", *Inf Technol Tourism 22*, 455–476 (2020). https://doi.org/10.1007/s40558-019-00160-3

[2] J. Rodríguez, I. Semanjski, S. Gautama, N. Van de Weghe, and D. Ochoa, "Unsupervised hierarchical clustering approach for tourism market segmentation based on crowdsourced mobile phone data," *Sensors (Switzerland)*, vol. 18, no. 9, 2018, doi: 10.3390/s18092972.

[3] Z. Abbasi-Moud, H. Vahdat-Nejad, and J. Sadri, "Tourism recommendation system based on semantic clustering and sentiment analysis," *Expert Syst. Appl.*, vol. 167, p. 114324, Apr. 2021, doi: 10.1016/j.eswa.2020.114324.

[4] G. Roshan, R. Yousefi, and K. Błażejczyk, "Assessment of the climatic potential for tourism in Iran through biometeorology clustering," *Int. J. Biometeorol.*, vol. 62, no. 4, pp. 525–542, 2018, doi: 10.1007/s00484-017-1462-6.

[5] P. Bathia, "Data Mining and Data Warehousing: Principles and Practical Techniques", Cambridge University Press, 2019.

[6] D. Webster, D. Owens, E. Thomlinson, G. Bird, and G. Tripp, "Introduction to Tourism and Hospitality in BC" - *2nd Edition*, 2015.