ISSN 2347 - 3983

Volume 8. No. 6, June 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter70862020.pdf

https://doi.org/10.30534/ijeter/2020/70862020

Clustering Hostels Data for Customer Preferences using K-Prototype Algorithm

Abba Suganda Girsang[1], FachrulHijriah Usman[2], RintisMardika Sunarto[3]

[1]Computer Science Departement, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, agirsang@binus.edu

[2]Computer Science Departement, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, fachrul.usman@binus.ac.id

[3]Computer Science Departement, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, rintis.sunarto@binus.ac.id

ABSTRACT

Reviews on hostel booking platforms can be used to collect rating data from customers which can be useful for service providers and customers in the future to determine hostel choices based on their individual preferences. The rating data can be used to determine the pattern of hostel selection by customers and can provide suggestions according to customer preferences in choosing hostels in certain areas. In this study, the data will be grouped using the k-prototype algorithm which is a combination of k-means and k-mode algorithms so that it is possible to group mixed attributes. The results of this study are to determine the data segment in accordance with user behavior in selecting the hostel at the time, location of the hostel and certain conditions, so that by knowing the segmentation profile, the hostel service provider can easily provide product promotions to segmented customers.

Key words: K-Prototypes, Clustering, Customer Preferences, KDD

1. INTRODUCTION

The increasing number of tourists and the increasing choice of tourism throughout the world makes the hospitality industry more competitive, one of which is the hostel industry which is spread throughout the world such as the hostel industry in Japan. The hostel industry is competing in providing the best service to each of its customers, one of the methods applied is by providing hostel booking services on various digital platforms to facilitate customers in booking hostel facilities according to the interests and needs of their customers.

There are many digital platforms that can facilitate customers in booking hostel facilities, such as Hostelworld. The digital platform has a review feature that can allow the system to collect rating data provided by customers after using hostel facilities.

The rating data usually contains the hostel cleanliness rating, the facilities provided, the experience felt by the customer, the comfort of the staff in serving customers, and the safety rating felt by the customer of the hostel

The data can be used by hostels to help them understand the behavior of their customers. This can help and be an evaluation material for the hostel in developing, designing and marketing the products they offer to make it even better and can be adapted to the behavior and needs of their customers. In addition, the data can also be utilized by companies in analytic form to help customers find their interests and needs based on the desired preferences. So, it can be an incentive to increase hostel revenue, because they already have a picture regarding the behavior, interests and needs of their customers.

There are many studies related to clustering using the k-prototype algorithm. Regarding the hostel industry and clustering for customer preferences, [1],[2]which shows how the growth of the hostel industry has increased very rapidly. The hostel industry has become an industry that has a very high demand frequency [3], [4] but there are still minimal steps that they take in a structured way to utilize the rating data from customers they have in providing offers to the customers they are targeting. [5], [6] grouping of potential products and knowing sales in certain areas. [7] conduct research that aims to use data mining to provide product category recommendations based on profile assessment of potential customer groups. [8], [9], [10], [11]the trend of providing advice on a targeted service to potential customers is also utilized by several other industries, such as e-commerce that can provide product recommendations to their potential customers.

The k-prototype algorithm will be used to do the grouping, because the data objects consist of mixed and numerical categorical attributes. The use of this algorithm is because the dataset used is a hostel review data set in Japan consisting of mixed attributes, both numerical such as hostel start price, summary scores and categorical locations such as hostels in various cities and the band rating of the hostel.

By using this algorithm, hostel recommendations based on customer preferences are generated into several clusters of scenarios 1, 2 and 3. To produce hostel recommendations, based on preferences, the selected dataset is a hostel dataset containing data about review of hostels in Japan.

The purpose of this study is to determine the data segment of a hostel that can be a recommendation for potential customers based on their user behavior, so that it can be a reference for them when they want to choose a specific time/hostel/condition to be used according to their needs. By knowing the data segment, hostel service providers can more easily provide and determine product services to their customers who have been segmented.

2. RELATED WORK

Data review from users on a service collected by various digital platforms that contain information related to the review of a service that has been felt by users. In this research case study, from the perspective of the hostel industry, review data is collected from customers who have booked hostel facilities, for example hostel cleanliness ratings, facilities provided, customer experience, staff comfort in serving customers, and security ratings felt by customers from the hostel. The review data can be collected and then converted into several data grouping matrices that can be used for analytics.

Clustering is a method that can be used to carry out the process of grouping an object data set into several groups of attributes that you want to use or are interrelated [12]. The process of grouping these object data sets is a fundamentally important task in machine learning, data visualization and computer vision [12]. The k-means algorithm is one of the most popular machine learning methods to use for grouping data objects because this algorithm is simple and easy to implement [13], [14].

Although k-means is popular because it is easy and simple to implement, but this algorithm also has disadvantages because it can only be used on data objects with numeric attribute types so this algorithm cannot be used for grouping data objects with categorical attribute types [15].

Thus, to make groupings of data with numeric data types, [16] proposes to use the k-mode algorithm which is the result of a k-means modification that is faster and can be used on categorical attributes.

Although the k-mode algorithm can do data grouping on data objects with categorical attribute types, but this algorithm cannot do data grouping on data objects with mixed attribute types (numeric and categorical).To group data on data with mixed attribute types, [17] proposes using the k-prototypes algorithm which can be a solution for classifying data on data objects with mixed attribute types.

this study, the k-prototype algorithm is used to group review data of hostels in Japan. The k-prototype algorithm is a combination of k-means and k-mode algorithms that can be used to group data objects with mixed attribute types consisting of numerical and categorical. This algorithm is more useful for data on objects that are often found in the real world because usually the data consists of various types of attributes in it.

3. PROPOSED METHOD

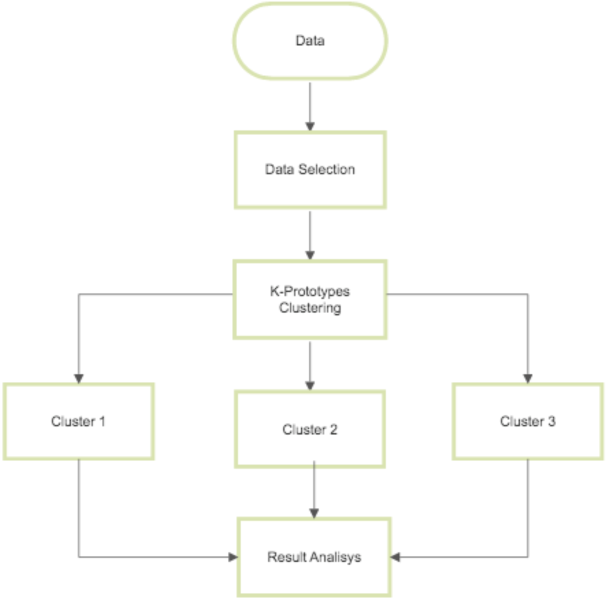

Figure 1: Research Steps

Figure 1 shows the steps that are taken in this study. The first step is fetching the hostel dataset in Japan from Kaggle. After that, we will select the data before the clustering process. The results of the data selection will be stored in separate parts of the operational database.Then we will preprocess the data to be clustered using the k-prototype algorithm. The results of the clustering will then be analyzed.

In this study, grouping analysis on data sets of 298 rows of hostel data in Japan. The data contains a lot of information about hostels which are then filtered so that only the necessary information is taken, such as score categories, prices, and cities.

From all the data, we selected the data we will use with a total of 298 rows of data, as follows (table 1):

Table 1: Sample data after data selection

City	Price From	Summary Score	Rating Band
Osaka	3300	9.2	Superb
Tokyo	3600	8.7	Fabulous
Tokyo	2600	7.4	Very Good
Tokyo	1500	9.4	Superb
Tokyo	2100	7.0	Very Good
Tokyo	3300	9.3	Superb
Tokyo	2200	7.7	Very Good
Osaka	1600	9.2	Superb
Tokyo	2000	8.5	Fabulous
Tokyo	2200	10.0	Superb

After all data is selected, the data is ready to be grouped.

Based on the literature on clustering [16], [17],[18], there are 4 steps in implementing the k-prototype algorithm:

Step 1: read parameters, in this step, all parameters in the data will be read, such as n = amount of data, k = maximum number of each grouping, all attributes with numeric data types and number of attributes with categorical data types, and attribute names and attribute type.

Step 2: Initial prototype selection, in this step, all k objects will be randomly selected as the initial prototype for cluster k.

Step 3: make an initial allocation, in this step, all objects in the dataset will be assigned to clusters that have a minimum difference from the prototype that was done in the previous stage. After that, the prototype will be updated after being assigned.

Step 4: Reallocate, in this step, all prototypes that have been updated in the previous stages will be renewed, after executing using the algorithm in the Python program, the console will provide a data movement that shows that some objects have changed the cluster in the process. If the data movement is equal to zero, then it indicates that the algorithm that has been run has obtained the best results.

4. RESULTS

In order to find out the data segment, in this case the user's behavior in selecting hotels at certain times, hotels and conditions, cluster analysis is performed using the k-prototypes algorithm. By grouping this data, the results can provide a representation and summary of how to provide hostel recommendations that are grouped by score and price categories as well as prices and cities.

After running the clustering process using K-prototype, we get the results into 3 clusters, the results can be seen in figures 2 and 3 which illustrate the results of clustering based on the data used.

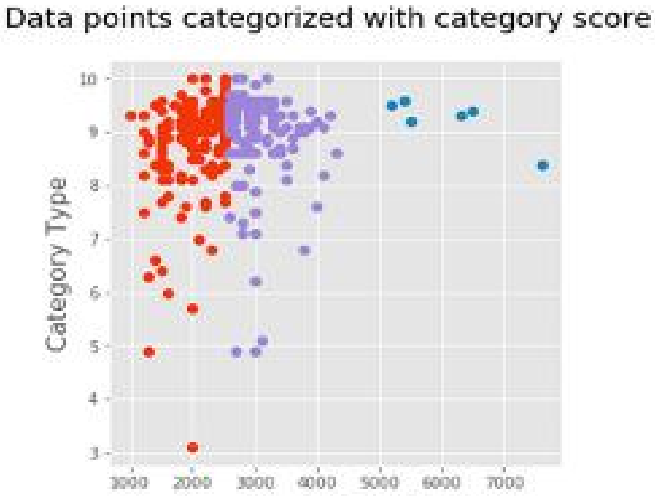

Figure 2: Clustering based on data score category

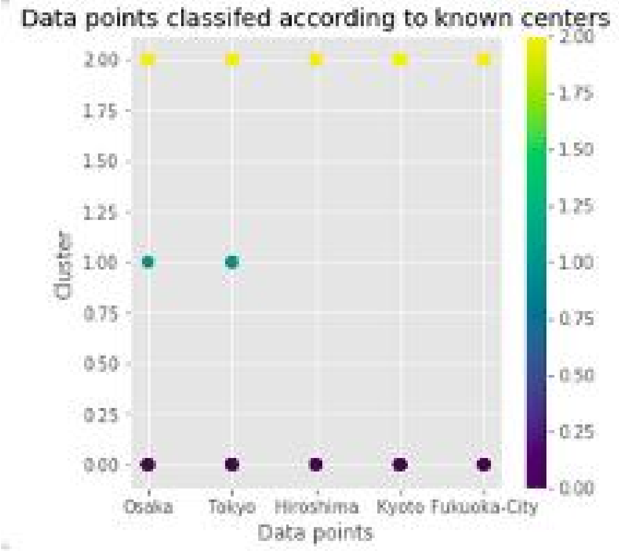

Figure 3: Clustering based on data city

After running the k-prototypes clustering processshows that each cluster contains data on score, price, and city categories. and each cluster produced will be named according to their content that represents them the best. The description for each cluster is as follows:

From these 3 clusters, we conducted an analysis with the following results:

1. The expensive capital city hostel cluster (dark blue) is a cluster of 2 cities, Osaka and Tokyo. This cluster is a collection of hostels that have a high enough price, with an average of 6084. However, this hostel is a hostel that has the highest average score of 9.3. Of course, hostel prices are comparable to prices of hostels.

2. Midrange city hostel cluster (purple in figure 2 & yellow in picture 3) is a cluster consisting of medium hostel groups in all cities with an average price of 3090. This cluster gets an average score of 8.8, the same as the cluster budget city hostel.

3. Budget city hostel clusters (red in figure 2 & purple in figure 3) are clusters consisting of cheap hostel groups in all cities with an average price of 2029. These hostels are hostels that we can visit if we want to stay with the price is quite cheap. In terms of scores, this cluster gets the same value as the midrange city hostel cluster, which is 8.8. This makes the hostels in this cluster may be more desirable than the midrange cluster.

Based on all existing clusters, all of them have the highest band rating result, "Superb". Figure 2 shows that there are quite a number of hostels that are included in the cluster budget city hostel and have a fairly high summary score. This indicates that hostels in this cluster have many enthusiasts and can be the best choice for potential customers with a limited budget but still get the best service quality.

5. CONCLUSION

All the analysis presented in this study is focused on knowing the data segment, in this case to recommend hostels based on user behavior in choosing hotels at certain times, hotels and conditions. Segmentation is done based on score categories, prices, and cities that produce 3 clusters. Of course by knowing the results of the segmentation, hostel service providers are easier to provide and determine product services to their customers who have been segmented. For example, when a customer is looking for a hostel with a low budget but still wants the best service the data shown is on a budget cluster city hostel, and vice versa if you want to find a hostel with a high budget and good service, you can choose hostels in the cities of Osaka and Tokyo that are the data is in the expensive capital city hostel cluster.The results of this study can be applied to Web or Android applications in recommended features to customers according to customer behavior. Future work should focus on larger scale analysis, using much larger datasets for more detailed analysis of user behavior.

REFERENCES

[1] M. Mowforth and I. Munt, **Tourism and Sustainability: Development, Globalisation and New Tourism in the Third World**, United Kingdom: Routledge, 2015. https://doi.org/10.4324/9781315795348

[2] UNWTO, **"Tourism and Poverty Alleviation,"** May 2018. [Online]. Available: http://step.unwto.org/content/tourism-and-poverty-alleviation-1. [Accessed 5 June 2020].

[3] Z. Xiang, Z. Schwartz, J. H. Gerdes Jr. and M. Uysal, "**What can big data and text analytics tell us about hotel guest experience and satisfaction?**," *Elsevier,* vol. 44, pp. 120-130, 2015. https://doi.org/10.1016/j.ijhm.2014.10.013

[4] L. Zhou, S. Ye, P. L. Pearce and M. Y. Wu, "**Refreshing hotel satisfaction studies by reconfiguring customer review data,**" *Elseiver,* vol. 38, pp. 1-10, 2014. https://doi.org/10.1016/j.ijhm.2013.12.004

[5] **E. M. Sipayung, F. Cut and T. R, "Decision Support System for Potential Sales Area of Product Marketing using Classification and Clustering Methods,"** *Proceeding International Seminar on Industrial Engineering and Management,* pp. 33-39, 2015.

[6] E. M. Sipayung, F. Cut and T. R, **"Modeling Data Mining Dynamic Code Attributes with Scheme Definition Technique,"***Proceeding Electrical Engineering, Computer Science and Informatics,* pp. 25-28, 2014.

[7] E. M. Sipayung, H. Maharani and B. A. Paskhadira, **"Designing Customer Target Recommendation System Using K-Means Clustering Method,"***International Journal of Information Technology and Electrical Engineering,* vol. 1, no. 1, pp. 2550-0554, 2017. https://doi.org/10.22146/ijitee.25155

[8] F. Isinkaye, Y. Flajimi and B. Ojokoh, "**Recommendation systems: Principles, methods and evaluation,**" *Egyptian Informatics Journal,* vol. 16, pp. 261-273, 2015.

[9] Y. H. Cho, J. K. Kim and S. H. Kim, "**A personalized recommender system based on web usage mining and decision tree induction**," *Elseiver Expoert Systems with Application,* vol. 23, pp. 329-342, 2002.

[10] P. H. Chour, P. H. Li, K. K. Chen and M. J. Wu, "**Intergrating web mining and neural network for personalized e-commerce automatic service**," *Elseiver Expert Systems with Applications,* vol. 37, pp. 2898-2910, 2010. https://doi.org/10.1016/j.eswa.2009.09.047

[11] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "**Analysis of Recommendation Algorithms for E-Commerce,**" *Proc. of the 2nd ACM conference on Electronic commerce (EC),* pp. 158-167, October 2000. https://doi.org/10.1145/352871.352887

[12] P. Awasthi, M. Chaikar, R. Krishnaswamy and A. K. Sinop, "**The Hardness of Approximation of Euclidean K-means**," *ArXiv Cornell University,* pp. 1-14, 2015.

[13] C. Slamet, A. Rahman, M. A. Ramdhani and W. Darmalaksana, "**Clustering the Verses of the Holy Qur'an using K-Means Algorithm,**" *Asian Journal of Information Technology,* vol. 15(24), pp. 5159-5162, 2016.

[14] G. Gan and M. K.-P. Ng, "**K-means Clustering with Outlier Removal,**" *Elseiver Pattern Recognition Letter,* vol. 90, pp. 8-14, 2017.

[15] F. Jiang, G. Liu, J. Du and Y. Sui, "I**nitialization of K-modes clustering using outlier detection techniques**," *Elseiver Information Sciences,* vol. 332, pp. 167-183, 2016. https://doi.org/10.1016/j.ins.2015.11.005

[16] R. S. Sangam and H. Om, "**An equi-biased k-prototypes algorithm for clustering mixed-type data**," *Indian Academy of Sciences,* vol. 43:37, 2018. https://doi.org/10.1007/s12046-018-0823-0

[17] Z. Huang, "**Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values**," *Data Mining and Knowledge Discovery 2,* pp. 283-304, 1998. https://doi.org/10.1023/A:1009769707641

[18] D. Guo, Y. Chen and J. Chen, "**A K-Prototypes Algorithm Based on Adaptive Determination of the Initial Centroids**," in *ICMLC 2018: 2018 10th International Conference on Machine Learning and Computing,* Macau, 2018.