Surabaya Tourism Destination Recommendation Using Fuzzy C-Means Algorithm

Raymond Sutjiadi[1], Edwin Meinardi Trianto[2] and Adriel Giovani Budihardjo[1]
[1]Informatics Department, Faculty of Information Technology, Institut Informatika Indonesia.
[2]Management of Informatics Department, Faculty of Information Technology, Institut Informatika Indonesia.
raymond@ikado.ac.id

Abstract—**Determining tourism destination requires various criteria that suitable for the travelers' needs. Usually, travelers explore tourism destination through the internet or have recommendation from their relatives. That way is not informative because most people will recommend well-known tourism destination based on their experience only. In this research, C-Means Fuzzy Clustering is used to build a decision support system in selecting tourism destination in Surabaya, Indonesia. By using this application, the travelers, who want to visit Surabaya city, is not only provided the information related to tourism destination, but also nearest hotel and restaurant that suitable to traveler's criteria and budget. This application processes the input into the desired output in the form of recommendation based on the calculation of the degree of membership and center of the cluster**

Index Terms—**Clustering; Fuzzy C-Means; Recommendation System; Tourism Destination.**

I. INTRODUCTION

The fast-moving information in tourism world nowadays is imbalanced with the use of information technology. Only well-known tourism destinations have many visitors and the others are abandoned. The main reason behind this is because the lack of information to promote various tourism objects. On the other hand, the information of accommodation facilities, like hotel and restaurant, near tourism object is important as one criteria to invite many tourists.

Surabaya as the second largest and populous city in Indonesia after Jakarta has many tourism destinations [1]. They are grouped as shopping center, museum, cultural and religious center, monument, theme park and nature tourism. Many tourism destinations in Surabaya are not published massively. Only the famous and city center places can gather many tourists.

To overcome this problem, developing complete digital information related to tourism destination is not enough. It will be more useful if the application is equipped with a feature to give certain recommendations using expert system algorithm. The recommendations are based on certain input criteria, like budget, number of travelers, accommodation expenses and distance.

In this research, a desktop application is developed as the source of information of tourism destinations in Surabaya. Not only that, this application has a feature to give recommendation of accommodations (hotel and restaurant) which is match with allocated budget and daily expenses. This is useful for budget travelers to minimize their budget without forfeiting the ease of traveling.

The recommendation system used in this application is based on Fuzzy C-Means algorithm. Tourism destination data is clustered into 3 clusters. In addition, data of distance and price of hotel and restaurant are clustered into 3 clusters as well. This system provides output in the form of list of hotels, restaurants and tourism objects to be visited by the tourist to minimize traveling budget.

II. FUNDAMENTAL THEORY

A. Fuzzy Clustering

Clustering is a process of partitioning or grouping a given set of unlabeled patterns into a number of clusters such that, similar patterns are assigned to one cluster [2]. Clustering techniques have been widely used in many areas such as data mining, artificial intelligence, pattern recognition, bioinformatics, segmentation, and machine learning [3]. There are 2 common methods of clustering, i.e. crisp clustering (hard clustering) and fuzzy clustering (soft clustering) [4].

Crisp clustering is an absolute partition and unsupervised clustering, like K-Means algorithm [5]. This clustering method is not really accurate and slower performance. In addition, this method is only suitable to be implemented on simple dataset to prevent ambiguity among clusters.

On the other hand, fuzzy clustering is a process determining membership degree and use it to classify element of data into one or more clusters. Fuzzy clustering is more robust compared to crisp clustering in terms of local minima of the objective function [6].

One of the fuzzy clustering algorithms used in this research is Fuzzy C-Means Clustering Algorithm. Fuzzy C-Means is a soft algorithm clustering fuzzy data in which an object is not only a member of a cluster but member of many clusters in varying degree of membership as well [7].

Vector of fuzzy clustering, $V=\{v_1, v_2, v_3, \ldots, v_c\}$, is an objective function which is defined by degree of membership from data X_j and center of cluster V_j. Fuzzy C-Means Clustering Algorithm separates provided data from each finite element and input it into part of clusters collection, based on various inputted criteria.

Below is the formula to calculate group of data and center of data. Suppose one group finite data $X = \{X_1, \ldots, X_n\}$ and center of cluster:

$$J_m(X, U, V) = \sum_{j=1}^{n} \sum_{i=1}^{c} (\mu_{ij})^m d^2(X_j, V_i) \qquad (1)$$

where: μ_{ij} = Membership degree of X_j and center of cluster is a part of matrix membership $[\mu_{ij}]$.
d^2 = Root of Euclidean distance.
m = Fuzzy parameter, where the average of

obscurity from each degree of membership data is not higher than 1.0.

The output of Fuzzy C-Means is a central of cluster series and membership degrees of each data point. This information can be used to build a Fuzzy inference system.

B. Fuzzy C-Means Clustering Algorithm

In this section will be elaborated step-by-step how to implement Fuzzy C-Means Clustering Algorithm as depicted in Figure 1:

1. Input data, X, in the form of matrix sized n x m (n = number of data sample, m = number of attribute for each data).
   X_{ij} = the i-th data sample (i = 1, 2, …, n), the j-th attribute (i = 1, 2, …, n).
   Determining this value:
   c   = number of cluster.
   w   = weight
   MaxIter   = maximum iteration
   ξ = expected smallest error
   Po = 0   = initial objective function
   t = 1   = initial iteration

2. Generate random value μ_{ik}, i = 1, 2, …, n; k = 1, 2, …, c as initial matrix partition elements. μ_{ik} is degree of membership as a reference to the probability a data can be member of cluster. Matrix position and value are built in random, where value of membership resides between interval 0 and 1. In initial position of matrix, μ partition is not really accurate as well as center of cluster.

$$Q_i = \sum_{k=1}^{c} \mu_{ik} \qquad (2)$$

3. Calculate center of k-th cluster: V_{kj}, where k = 1, 2, …, c and j = 1, 2, …, m. X_{ij} is fuzzy variable and w is the weight.

$$V_{kj} = \frac{\sum_{i=1}^{n}((\mu_{ik}{}^{w} * X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}} \qquad (3)$$

Objective function is used as a requirement of iteration to get accurate center of cluster. Thus, it results in the trend of data to join in a particular cluster at the end of stages.

4. Calculate objective function at t-th iteration, P_t

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right](\mu_{ik})^w\right) \qquad (4)$$

where: X_{ij} = value of fuzzy variable
        V_{kj} = center of cluster
        μ_{ik} = degree of membership
        P_t = objective function

5. Count the partition matrix alteration using i = 1, 2, …, n and k = 1, 2, …, c.

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}} \qquad (5)$$

6. At the end of this process, it will be assessed two conditions:
   a) If ($| P_t - P_{t-1} | < \xi$) or (t > MaxIter), the process will be stopped.
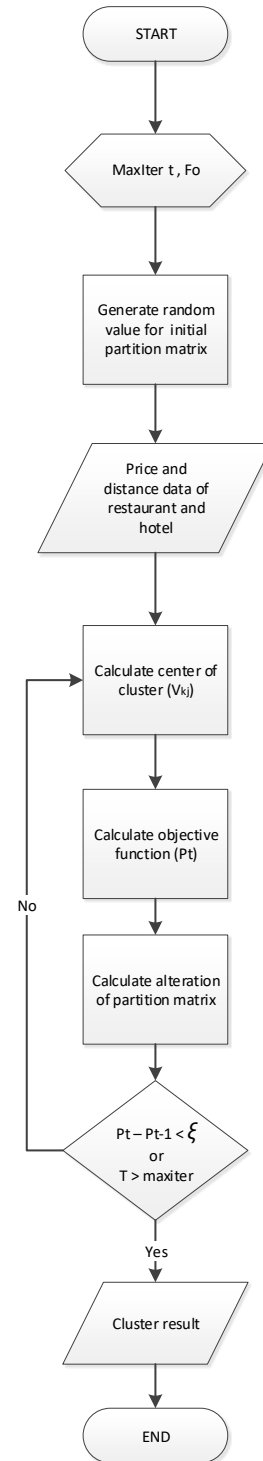   b) If not, continue the process to t = t + 1, then repeat the fourth stage above.

Figure 1: Flowchart of Fuzzy C-Means Clustering Algorithm

III. SYSTEM ANALYSIS AND DESIGN

As the implementation of Fuzzy C-Means Clustering Algorithm, a desktop application is developed. The aim of this application is to process the input data from a user using Fuzzy C-Means Clustering Algorithm to result in some tourism destination recommendations that suitable to the

needs.

There are 4 data to be inputted by the user to this application:

a) Total Budget: sum of money that a tourist may allocate for a trip.
b) Total Day Trip: total day to be spent on a trip.
c) Number of People: total person who join in a trip.
d) Tourism Destination: location of a trip.

To process the input into the output data, an administrator should enter the data of hotels, restaurants and tourism destinations located in Surabaya. For this research, 30 data of tourism destinations, 50 data of hotels, and 50 data of restaurants were collected.

Detailed activities between administrator and tourist as a user are depicted at Use Case Diagram and Activity Diagram on Figure 2 and 3 respectively.

The needed data for hotels in Surabaya are name, address, contact number, and average price for standard room. Then the needed data for restaurants in Surabaya are name, address, contact number, and average price for standard meal. Whilst for tourism destinations, the needed data are name, address,

and contact number. All these data are gathered from Google Search Engine, Surabaya Tourism Board, or by on-site survey, and saved in database to be processed further by clustering algorithm.

After all data are complete, there are 3 steps of clustering. The first is clustering process for latitude and longitude of tourism destinations. This will group 30 tourism destinations into 3 clusters based on the distance. After the first clustering, each cluster will be grouped again based on the distance of hotel and restaurant to tourism destination. And the last, there is one more clustering process for average price of hotel and meals in restaurant. The flowchart of this process is depicted in Figure 4.

The input data above will be processed and clustered into several output recommendations below:

a) Tourism Destination Location, where tourist should visit the tourism object.
b) Hotel Location, where tourist should live in during a trip.
c) Restaurant Location, where tourist should go for a meal (breakfast, lunch and dinner) during a trip.
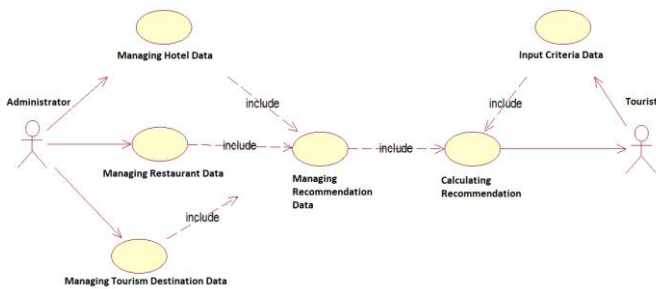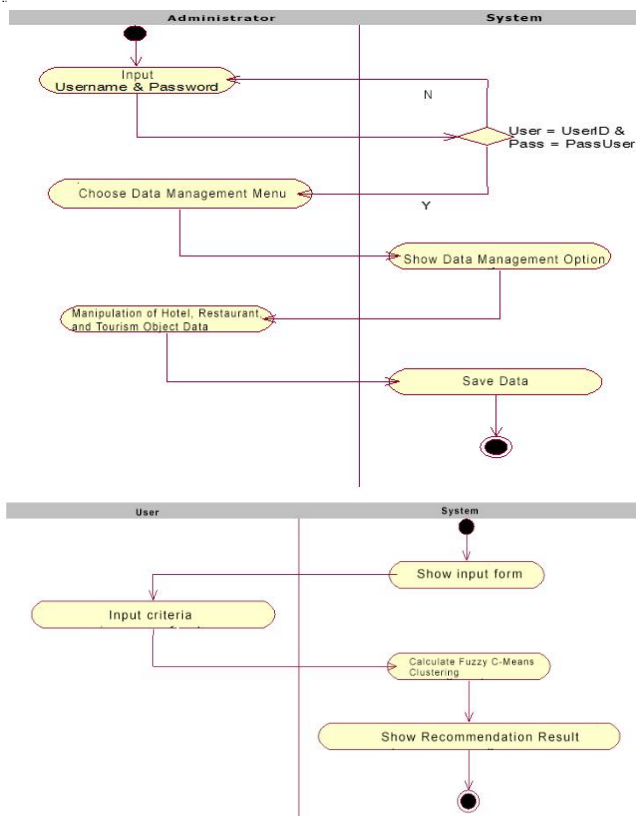

Figure 2: Use Case Diagram.
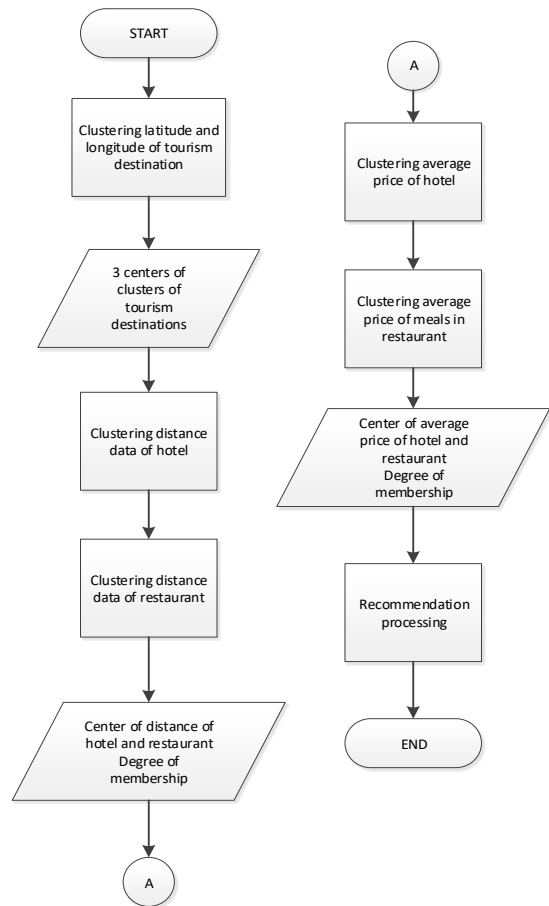

Figure 3: Activity Diagram.

Figure 4: Flowchart of Clustering Stages.

IV. IMPLEMENTATION AND TESTING

A desktop application is developed using Visual Basic .NET and Microsoft SQL server as the database. This application can be installed in Microsoft Windows Operating System with preinstalled .NET Framework 3.5 or above.

Before using this application, the administrator or user need

to authenticate themselves by typing username and password on the first page. This will enable authorization whether the client is administrator privilege-enabled client or user privilege-enabled client.

After login, user can input criteria as needed in a form as shown in Figure 5 and click process button to run clustering process.

Figure 5: Input Criteria Form.

As a result, recommendation will be suggested to user in 3 separate tables, each for tourism object, hotel, and restaurant as shown in Figure 6. In this form, user may see forecast budget allocation and distance of each place as well. They can also view detailed information about each suggested place, like detailed address, contact number, short description, photo, by clicking context menu at each row.

The system also can save all the recommendation result as an archive. User can see the last result without need to re-input all criteria from beginning and print the result for documentation purpose.

Figure 6: Recommendation Result Page.

To test the accuracy of algorithm, a scenario should be run, where user is simulated to input criteria to the application and it will output recommendation as the result of clustering process. Table 1 below is the scenario as an input to the system:

Table 1
Testing Scenario

Budget	IDR 10,000,000
Trip Duration	3 days
Trip Destination	Mangrove Gunung Anyar
Number of Traveler(s)	4 people
Suggestion Priority	By Price

After clustering process, the recommendation results are shown as below (Figure 7):
- The nearest tourism destination from Mangrove Gunung Anyar are:
  a. Mangrove Wonorejo Rungkut, distance: 5.5 Km.
  b. THP Kenjeran (Kenpark), distance: 16.9 Km.
  c. Kenjeran Beach, distance: 19.5 Km.

Figure 7: Recommendation Result for Nearest Tourism Destination.

- The nearest hotel from Mangrove Gunung Anyar based on price priority are (Figure 8):
  a. 88 Hotel, distance: 3 Km, average price: IDR 315.000/night for 1 person.
  b. Inna Simpang Hotel, distance: 3.9 Km, average price: IDR 465.000/night for 1 person.
  c. Veni Vidi Vici, distance: 4.2 Km, average price: 380.431/night for 1 person.

Figure 8: Recommendation Result for Nearest Hotel.

- The nearest restaurant from Mangrove Gunung Anyar based on price priority are (Figure 9):
  a. Ahimsa Vegan Lounge, distance: 2.2 Km, average price: IDR 29.000 per meal for 1 person.
  b. Calibre Coffee Roasters, distance: 3.1 Km, average price: IDR 67.000 per meal for 1 person.

c. Sate Klopo Ondomohen, distance: 3.2 Km, average price: IDR 30.000 per meal for 1 person.

Figure 9: Recommendation Result for Nearest Restaurant.

V. CONCLUSION

This application has been tested by 100 respondents to sum up the performance and accuracy of the clustering system to determine tourism destination plan. The respondents have to fill in a questionnaire related their satisfaction after trying the application.

The application can give recommendation about tourism destination quite accurately, suitable for users' needs. Around 75 percent of respondents feel that the recommendation results are appropriate with their inputted criteria. The other 25 percent gives an average rate and the application still needs more improvements in accuracy and variety of data.

In addition, this application can minimize the traveling budget and time as well by clustering nearest hotel and restaurant. Around 50 percent of respondents tell that inputted criteria, such as budget, trip duration, tourism destination, number of people, and priority, already reflect proper recommendation in order to minimize travelling budget and time. On the other hand, 40 percent of respondents give average rating and 10 percent of respondents are still dissatisfied with the recommendation result related to budget and time efficiency.

Sample data of tourism destinations, hotels and restaurants should be improved. Many tourism destinations data reside in Eastern Surabaya. The effect is clustering result tends to suggest tourism destination in Eastern Surabaya only. Sample data for tourism destinations, hotels and restaurants have to spread evenly across the area of Surabaya with adequate number of data. It will improve the accuracy of recommendation result.

A suggestion for future development, it will be better if this application has mobile version, instead of desktop application because travelers need flexibility and practicality to support their adventure mobility.

REFERENCES

[1] H. Santosa, "Environmental Management in Surabaya with Reference to National Agenda 21 and the Social Safety Net Programme," *Environment & Urbanization*, Vol. 12, No. 2, pp. 175-184, October 2000

[2] H. Ming-Chuan and Y. Don-Lin, "An Efficient Fuzzy C-Means Clustering Algorithm," in *Proceedings 2001 IEEE International Conference on Data Mining*, pp. 225 – 232.

[3] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, Vol. 31, No. 3, pp. 264-323, September 1999.

[4] R.Suganya and R.Shanthi, "Fuzzy C-Means Algorithm-A Review," *International Journal of Scientific and Research Publications*, Vol. 2, Issue 11, November 2012.

[5] J. Chen, D. Pi, and Z. Liu, "An Insensitivity Fuzzy C-means Clustering Algorithm Based on Penalty Factor," *Academy Publisher*, pp. 2379-2384, 2013.

[6] F. Klawonn, "Fuzzy Clustering: Insights and a New Approach," Mathware & Soft Computing, Vol. 11, pp. 125-142, 2004

[7] Z. Cebeci and F. Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures," *Journal of Agricultural Informatics*, Vol. 6, No. 3, pp. 13-23, October 2015.