Home appliances recommendation system based on weather information using combined modified k-means and elbow algorithms

Basim Amer Jaafar[1], Methaq Talib Gaata[2], Mahdi Nsaif Jasim[3]
[1,2]Computer Science Department, University of Mustansiriya, Iraq
[3]University of Information Technology and Communications, Iraq

Article Info

Article history:

Received Feb 1, 2020
Revised Mar 12, 2020
Accepted Apr 16, 2020

Keywords:

Clustering
Elbow method
K-means
Optimal number of clusters
Recommendation system

ABSTRACT

The recommendation system is an intelligent system gives recommendations to users to discover the best interesting items. The purpose of this proposed recommendation system is to develop a system to find the best electrical devices according to weather conditions and user preferences. The proposed solution relies on the characteristics of electrical appliances and their suitability to weather conditions in any city. The proposed solution is the first recommendation system combines devices properties, weather conditions, and user preferences using a new combination of algorithms. The clustering algorithms are the most applicable in the field of recommendation system. The proposed solution relies on a combination of Elbow method, proposed modified K-means and Silhouette algorithm to find the best number of clusters before starting the clustering process. Then calculate the weights for each cluster and compare them with the weather weights to find the required clusters sorted from the near to far according to a computed threshold. The empirical results showed that the proposed solution demonstrated a 94% accuracy to match the characteristics of the recommended devices with the climatic characteristics of the region and user preferences. The accuracy is measured using Silhouette algorithm.

Corresponding Author:

Basim Amer Jaafar,
Computer Science Department,
University of Mustansiriya, Baghdad, Iraq.
Email: basim200711@gmail.com

1. INTRODUCTION

Because of the great development in the home appliances industry, which causes a rapid increase in the demand for these appliances. Users should choose electrical appliances from a large number of options that can be confusing to any user. As a result, recommendation systems have gained great popularity in every field in digital systems. Where automated recommendation systems with different approaches can be useful to recommend users with appropriate devices. In this paper, a machine learning-based recommending devices using a clustering approach based on combined devices features of weather information. The goal of this paper is to put devices in a group by using the clustering algorithm in order to find the devices that are most suitable to the user geographic area. The recommendation system is known as an intelligent system that tries to recommend the most appropriate items (products or services) to specific users (individuals or businesses) by predicting user's interest in any filed by relying on information related to the elements [1, 2]. Recommendations are intended to provide suggestions to the user. The goal of recommendation techniques is to generate a user result when makes a decision while facing different options [3]. Recommendation systems really help customers to find their requirements, so this meets customer requirements in a short time [4]. It applied in different applications such as music, movies, books, news, search queries, research articles and products. Many of the filtering methods used in the recommendations system exploit the user information to provide the appropriate elements [5]. Although the various methods of recommendation systems have been developed in recent years, this field remains fertile to more research due to the increasing demand for practical applications, which is used to provide customized recommendations and deals with the information overload [6].There is a group of different techniques for recommendation systems, one of the most important of these methods is the clustering [7]. The clustering algorithm is a method of partitioning a physical or theoretical object into a group of similar objects. A cluster is an assortment of information objects; the objects in a cluster are like one another and are not the same objects in other clusters [8]. For the clustering mission, the objects as close as a conceivable inside cluster. However, the random selection of the center point of the sample will make cluster collection not converge [9]. The most commonly utilized is the K-means clustering algorithm because of its effortlessness [10]. In this paper, a modified K-means used with the Elbow strategy to determine the actual number of clusters appropriate to the user database and achieve a high efficiency to fit user requirements to choose the devices appropriate for both weather conditions, device properties, and user preferences.

2. RELATED WORKS

Recently, there are some of researches invistigating the development of recommendation systems based on clustering techniques. Akanksha Jyoti1 et al 2019: applied collaborative filtering to find the user's rate score to make relations with other users and Euclidean distance similarity score distinguish similarity between users. Integrate with a map interface to find the shortest distances among stores whose products were recommended. The result showed a better approach towards the recommendation of products among local stores within a region [11].

MA Syakur et al 2018: used K-means method with Elbow to improve effective and efficient k-means performance of big quantities of data. Elbow and K-means methods that the determination the best value of the clusters [12].

Phongsavanh Phorasim and Lasheng Yu 2017: used K-means and collaborative filtering to movies recommendation proposed, a user-based recommendation method using Euclidian distance to calculate two users of the cluster dataset [13].

Garg and Tiwari 2016: Proposed an efficient Massive Online Open Courses (MOOCs) recommendation system based on K-means and collaborative. The rating created from the activity of users. The system produces the neighborhood clusters from the user database. The system has being trained for predicting the user [14].

Oyelade et al 2010: used a k-means clustering algorithm was implemented to analyze student results based on cluster analysis and used statistical algorithms to rank their grade data according to their level of performance [15].

From the above related work, a combination of more than one algorithm and a combination of weather condtions and device features are not tackeled, So this reseach is proposed to achieve the objectives of the current research.

3. PROBLEM STATEMENT

Wide varaiety of home appliances developed by several companies, offering different features of the same devices but in different working conditions, that making it difficult to find the best device fit. The proposed recommendation system helps users to find the best fit devices that match their needs, interests and weather conditions. This proposed system is the only one system to give recommendations that are more accurate using new combination of algorithems and new combination of parameters to achieve these objectives.

4. RESEARCH OBJECTIVES

Design and implement a recommendation system has the ability to deal with large number of devices and recommend the best choice regarding to the user preferences, weather conditions and devices properties.

5. RESEARCH METHODOLOGY

To build the devices recommendation system, we need to create a database that contains the devices properties and weather conditions (temperature and humidity). The data collected and stored in the database designed for this purpose.The block diagram of the proposed recommendation system is shown in Figure 1.

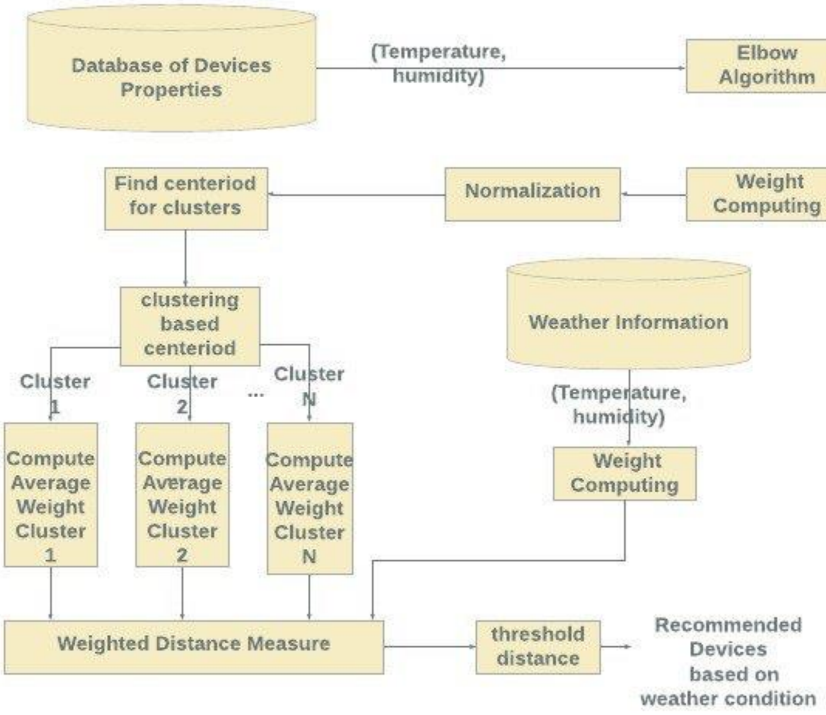

Database of Devices Properties
(Temperature, humidity)
Elbow Algorithm
optimal Number of Cluster K
Weight Computing
Normalization
Find centeriod for clusters
clustering based centeriod
Cluster 1
Cluster 2
...
Cluster N
Compute Average Weight Cluster 1
Compute Average Weight Cluster 2
Compute Average Weight Cluster N
Weather Information
(Temperature, humidity)
Weight Computing
Weighted Distance Measure
threshold distance
Recommended Devices based on weather condition

Figure 1. Block diagram of the proposed recommendation system

5.1. Database of devices

The database contains a set of devices (200 devices). It is collected from different companies' websites; each device has a set of properties, including temperature and humidity of operation. These properties represent the standard conditions in which the device works. In some cities, the temperature is the effective factor in the recommended system, and humidity is less effective and vice versa. Therefore, we need to calculate the special weights for each factor. The weight of the properties of each device is calculated according to the as shown in 1.

$$\text{The weight of device} = \big((\text{device working temperature} \times \text{temperature degree of importance}\ ) + (\text{device working humidity}\ \times \text{working humidity degree of importance}\ )\big) \quad (1)$$

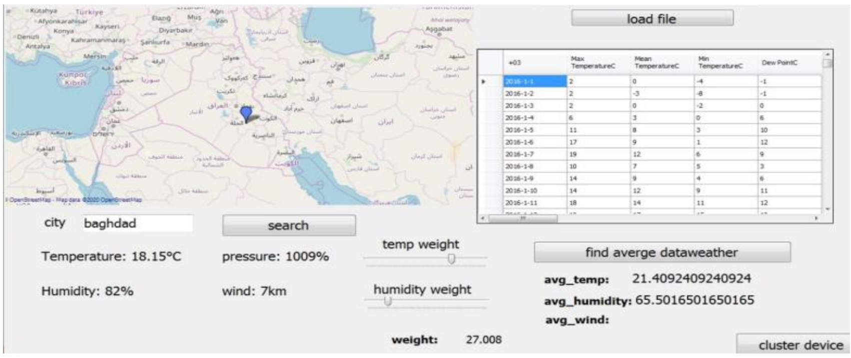
device working temperature, device working humidity : are read from DB.
temperature degree of importance, humidity degree of importance: are user input.

5.2. Geolocation information weather

The goals of data mining are to provide accurate knowledge in the form of rules, techniques, visual charts and useful models for weather parameters throught data sets [16]. The proposed recommendation system relies on weather information (temperature and humidity) for different cities. In some cities, the temperature is the biggest factor in the recommended system, and humidity is constant and vice versa. Therefore, we need to calculate the special weights for weather information for the city and the calculation of the special weights for each device (i.e. resulting from the temperature and humidity of the device) as in the as shown in 2 and Figure 2.

$$\text{The weight of weather} = \big((\ \text{temperature of the city} \times \text{temperature degree of importance}\ ) + (\ \text{humidity of the city}\ \times \text{working humidity degree of importance}\ )\big) \quad (2)$$

temperature of the city, humidity of the city: read from weather DB for several years or requst online.
temperature degree of importance, humidity degree of importance: are user input as shown in Figure 2.



load file

	+03	Max TemperatureC	Mean TemperatureC	Min TemperatureC	Dew PointC
▸	2016-1-1	2	0	-4	-1
	2016-1-2	2	-3	-8	-1
	2016-1-3	2	0	-2	0
	2016-1-4	6	3	0	6
	2016-1-5	11	8	3	10
	2016-1-6	17	9	1	12
	2016-1-7	19	12	6	9
	2016-1-8	10	7	5	3
	2016-1-9	14	9	4	6
	2016-1-10	14	12	9	11
	2016-1-11	18	14	11	12

city baghdad

search

Temperature: 18.15°C

pressure: 1009%

Humidity: 82%

wind: 7km

temp weight

humidity weight

weight: 27.008

find averge dataweather

avg_temp: 21.4092409240924

avg_humidity: 65.5016501650165

avg_wind:

cluster device

Figure 2. Geolocation information weather extraction and weight computing

5.3. Elbow method

The basic concept of the Elbow method is to utilize the square of the distance between the sample points focuses on each cluster and the centroid of the cluster to give a progression of K (i.e. number of cluster) value. The sum of squared errors (SSE) is utilize as a performance show that each cluster is closer. At the point when clusters number is set close to the number of the real cluster, SSE shows quick downhill. However, will turn out to be slower rapidly [17]. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters as shown in Figure 3 [18].

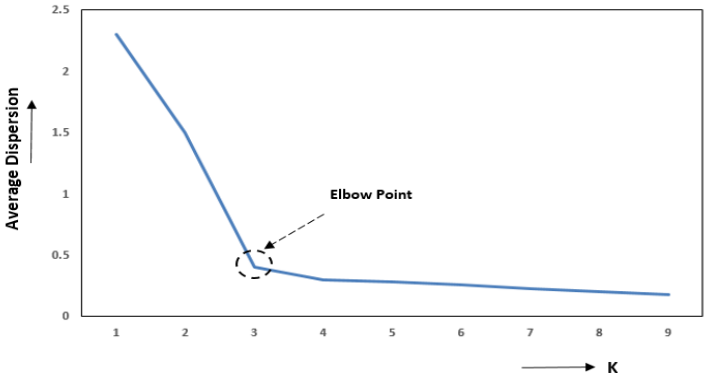


Figure 3. Elbow method for optimal value of K

```
Algorithm 1: Elbow method for determining K
1. Initialize k = 1
2. Start
3. Increase the value of k
4. Measure the cost of optimal quality solution
5. If the cost of the solution at some point decreases dramatically
6. This is real k
7. End
```

5.4. K-means algorithm

K-means is one of the most popular and oldest clustering techniques and can be applied even to large data sets [19]. The K-means algorithm gives a simple method to execute an approximate solution. The purposes behind the publicity of K-means are the simplicity and easiness of execution, adaptability to spare data, intermingling speed and scalability [20]. The distance will use as the scale given for K classes in the dataset, calculate the distance mean, the initial centroid given, with each category described by the centroid. For a given data set X that contains multidimensional data points and a class K to be divided, Euclidean distance is defined as an indicator of similarity and group targets reduce the sum of squares of different objects; this means that it reduces [21]. K-Means algorithm is a widely used algorithm for identifying clusters because it has accurate calculations, easy to use and meets the needs of use because it is flexible to modify [22].

A clustering algorithm, it gathers various information dependent on the features and properties of that information and the clustering procedure by diminishing the separations between data center. The block diagram of the K-means shown in Figure 2.

```
Algorithm 2: k-means [23]
Input: D = {d1, d2, d3,…, dn} : set of n numbers of data.
K: The number of desire groups.
Output: A set of k clusters.
Step 1: Select k points as primary centroids.
Step 2: Repeat.
Step 3: From K groups by assigning every data point to the nearest centroid.
Step 4: Calculate the centroid of each cluster so that the centroid does not change.
```

The main weaknesses of K-means the number of clusters to be determine by the user and the random number selection affecting the accuracy of the cluster results.

5.4.1. Suggested solving of k-means weakness points

The greatest challenging problem in the pattern recognition field has distinguished is the finding of the ideal number of clusters for the discretionary dataset collection [24]. To choose the K clusters number and the centroid of the perfect cluster that can be provid for the K-means. Elbow technique can determine the appropriate number of clusters for the dataset [25].

5.4.2. The proposed algorithm (modified k-means clustering algorithm)

Model building require the using of a modified clustering algorithm used to aggregate household appliances for each customer. The proposed method works based on the weights average, that arrange the clusters according to which the first cluster would contain the most suitable devices and the last cluster would contain the least appropriate devices for the user. During the first stage, a data pre-processing technology that added data weights and normalization process adopted. During the second stage, the average weight for each cluster is calculated. During the third stage, the Euclidean distance with geographic area weather weights applied to arrange these clusters. Fourth stage the threshold value applied to the distances from the third stage.

```
Algorithm 3: Name: Modified K-Means Clustering Algorithm
Inputs:
      Weights of device calculated from Equation (1),
      Weights of weather calculated from Equation (2),
       Number of clusters (k) computed from algorithm (1),
      Threshold value.
Output: clusters that only contain the required devices.
Strat
      Step1: Weights reading.
      Step2: Normalization.
      Step3: A set of weights as Centroids of clusters (k) randomly assigned.
      Step4: Calculate the distance between each weight and all Centroids, this process
      done by using the
       Euclidean distance.
      Step5: Collect weights to the nearest Centroids.
      Step6: Calculating new Centroids for each cluster.
      Step7: Repeat steps 3 through 5 until stability occurs.
      Step8: Calculate the average weights for each cluster by calculating the total weight
      of the cluster
       divided by the number of elements in the cluster.
      Step9: use the Euclidean distance between the weights generated by from Equation (2)
      with the
       average weight for each cluster.
      Step10: Arrange the clusters from the lowest distance to the largest distance.
      Step11: Cluster suggestion to the user where the distance value from step 9 is less
      or equal to the
       threshold value.
END
```

A. Compute average weights for each cluster

After determining the appropriate number of k in the Elbow algorithm method and performing proposed method K-means, it will produce a set of cluster with the same predefined k number. The goal of the clustering process is treat the devices as a group of similar weight devices instead of treating them individually. The sum of the weights of the devices for each cluster are calculated and then divided by the number of devices of each cluster to produce a set of weights for each cluster.

B. Weighted distance measure

Euclidean distance between the weighted temperature and humidity for a given region and the average weight of the respective devices is calculated for each cluster. After the Euclidean distance equation is performed, the clusters will be rearrange according to the result of this process in ascending order, the first cluster is the lowest difference in Euclidean distance between the weight of the region and the weights for clusters and next clusters with increasing difference. The first cluster contains the best devices suitable for a specific region until we reach the last cluster, which contains the largest difference, which reflects least fit devices and it is not preferred to work in these weather conditions.

C. Determine the best-fit devices cluster

To determine the best fit clusters, a threshold is used to determine groups that contain devices closest to a specific region. The threshold value is computed and then the result from applying the Euclidean distance between the average weights of each cluster and the weight of the specified region. If the resulting value is less than or equal to the threshold value, then devices in this cluster have approached user requirements.

6. RESULTS AND DISCUSSION

The proposed system developed using C# language, and SQL databases used to store the input dataset and the clustering results. The proposed system is a blend of modified K-means and Elbow method to improve the clustering process to promote clustering quality. The devices properties dataset is submit to the clustering process, the intial number of clusters (K) is determind by Elbow method. K passed to the modified K-means clustering algorithm. The modified K-means algorithm calculates and finds the most fitted clusters of devices. After the completion of the work clusters for similar devices in temperature and humidity then calculated, the average weight of each cluster compared with the average weight of the region weather and find the best cluster that contains the most suitable devices.

After clustering process, efficiency test done to determine the results accuracy. Total sum of squared errors (SSE) for each cluster show the biggest decrease in K = 5 (as shownin Figure 4 and Table 1). In this test, it is clear to discover the effect of intial clusters number which computed by Elbow method on the accuracy of clustering process. Table 1 shows the SSE value in the test number of clusters in the range of 1 to 10 clusters.

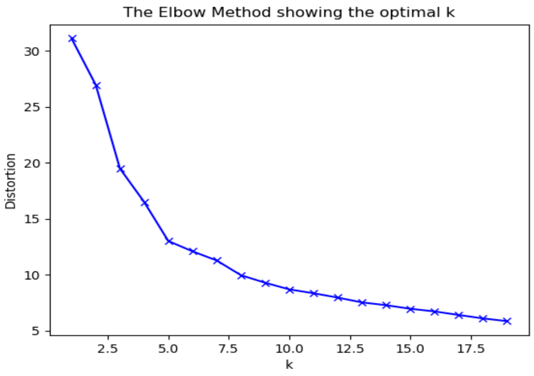


Figure 4. The effect of K on clustering results (k=5)

Table 1. Sum of square error results from each number of clusters

Number of Clusters	Result of sum square error for 200 devices
K=1	269981.28
K=2	181363.59595959596
K=3	106348.37306211118
K=4	73679.78903948834
K=5	44448.45544793371
K=6	37265.86520484347
K=7	30259.65720728547
K=8	25095.703209997548
K=9	21830.041978049434
K=10	20736.679938924124

After performing the clustering procedure, the threshold cuts off to five clusters of devices. The average weight of each cluster calculated by finding the total weights of devices then divided it by the number of devices in each cluster as shown in Table 2. Suppose we want to know what the appropriate devices for the city of Baghdad. We can make a query about the temperature and humidity for a known period and got the results shown in Table 3.

By calculating the Euclidean distance between the weight extracted from the weather information and weight of each cluster, the cluster containing the minimum distance between other clusters will be the best cluster containing the appropriate device for this city as shown in Table 4. The results shown in Table 4 show that the fourth group contains less than the threshold value (threshold value = 5) compared to other groups resulting from the Euclidean distance since the other groups have a distance greater than the threshold value. According to the previous results, the data Cluster 4 containing the recommended devices shown according to the weather data for the city of Baghdad according to the input dataset as shown in Table 5.

Table 2. Number of devices and weight of each cluster

Cluster number	Number of devices	Average Weight of cluster
1	37	72.02162
2	74	53.566233
3	39	85.78974
4	26	26.807692
5	24	38.3583333

Table 3. Weight resulting from weather information in Baghdad

City	temperature degree	Humidity	Weight
Baghdad	20.15	68.9%	29.86066

Table 4. Calculate the distance between each cluster and weight weather

Cluster Number	Average Weight of cluster	Weather Weights	Result of Euclidean distance
1	72.02162	29.86066	42.16096
2	53.566233	29.86066	23.70557
3	85.78974	29.86066	55.92908
4	26.807692	29.86066	3.05296
5	38.3583333	29.86066	8.4976

Table 5. The devices in cluster 4

Device number	Temperature	Humidity	Device number	Temperature	Humidity
1	15 C°	39%	14	20 C°	15%
2	16 C°	66%	15	20 C°	13%
3	17 C°	40%	16	21 C°	35%
4	18 C°	61%	17	23 C°	29%
5	33 C°	44%	18	24 C°	35%
6	34 C°	67%	19	25 C°	40%
7	37 C°	66%	20	28 C°	70%
8	38 C°	61%	21	28 C°	73%
9	39 C°	66%	22	29 C°	71%
10	39 C°	55%	23	30 C°	53%
11	39 C°	48%	24	33 C°	44%
12	19 C°	33%	25	24 C°	70%
13	19 C°	40%	26	24 C°	76%

7. LIMITATIONS OF PROPOSED SYSTEM

Among the difficulties encountered in the paper is the absence of a standard database and often the characteristics of the devices (temperature and humidity) are not available in standardized units of measurement, so the research requires design, implementation, and data collection of targeted devices.

One of the weakness points of this system is that it neglected the consumption of electrical energy and the price of each device. This improvement must be make to the system to be more inclusive of the variables that play a role in determining the user requirements for the devices.

8. CONCLUSION AND FUTURE WORKS

This paper present a recommendation system for the best electrical appliances suitable for a specific city based on weather information, device properties, and user preferences. The system tested on a set of devices (200 devices), where the results obtained from the system showed that they are more accurate than traditional algorithms. The results of the system tested with one of the methods for evaluating the cluster, which is the Silhouette through the calculation of inter and intra cluster and gave a value (0.609) with a small run time (8.45 seconds). The intial number of clusters effects clearly the clustering results and then the accuracy of system recommandations. It is necessary to test the system on a huge database of devices to prove the efficiency of the system.

REFERENCES

[1] Lu, Jie, et al. "Recommender system application developments: a survey," *Decision Support Systems,* vol 74, pp. 12-32, 2015.

[2] Mohd Suffian Sulaiman, Amylia Ahamad Tamizi, Mohd Razif Shamsudin, Azri Azmi. "Course recommendation system using fuzzy logic approach," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* Vol. 17, No. 1, pp. 365-371, 2020.

[3] K.A.F.A. Samah, I.M. Badarudin, et al. "Optimization of house purchase recommendation system (HPRS) using genetic algorithm," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* Vol. 16, No. 3, pp. 1530-1538, 2019.

[4] Naw Naw and Ei Ei Hlaing. "Relevant Words Extraction Method for Recommendation System," *Buletin Teknik Elektro dan Informatika* Vol. 2, No. 3, pp. 169-176, 2013.

[5] Shah, Jaimeel M., and Lokesh Sahu. "A hybrid based recommendation system based on clustering and association," *Binary Journal of Data Mining & Networking,* vol 5, no. 1, pp. 36-40, 2015.

[6] Sharma, Lalita, and Anju Gera. "A survey of recommendation system: Research challenges," *International Journal of Engineering Trends and Technology (IJETT),* vol. 4, no. 5, pp. 1989-1992, 2013.

[7] Ricci, Francesco, Lior Rokach, and Bracha Shapira. "Introduction to recommender systems handbook," *Springer,* Boston, MA, pp. 1-35, 2011.

[8] Deepana, R. "On Sample Weighted Clustering Algorithm using Euclidean and Mahalanobis Distances," *International Journal of Statistics and Systems*, vol. 12, no. 3, pp. 421-430, 2017.

[9] Yuan, Chunhui, and Haitao Yang. "Research on K-Value Selection Method of K-Means Clustering Algorithm," *Multidisciplinary Scientific Journal*, vol. 2, no. 2, pp. 226-235, 2019.

[10] Ordonez, Carlos. "Programming the K-means clustering algorithm in SQL," *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 823-828, 2004.

[11] Jyoti, Akanksha, et al. "Nearby Product Recommendation System Based on Users Rating," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, vol. 5, pp. 963-968, 2019.

[12] Syakur, M. A., et al. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conference Series: Materials Science and Engineering,* Vol. 336, No. 1, 2018.

[13] Phorasim, Phongsavanh, and Lasheng Yu. "Movies recommendation system using collaborative filtering and k-means," *International Journal of Advanced Computer Research*, vol. 7, no. 29, p. 52, 2017.

[14] Garg, Vishal, and Ritu Tiwari. "Hybrid massive open online course (MOOC) recommendation system using machine learning," *IET Conference Proceedings,* p. 11 (5.)-11 (5.), 2016.

[15] Oyelade, O. J., O. O. Oladipupo, and I. C. Obagbuwa. "Application of k Means Clustering algorithm for prediction of Students Academic Performance," *arXiv preprint arXiv*:1002.2425, 2010.

[16] Talib, M. Ramzan, et al. "Application of Data Mining Techniques in Weather Data Analysis," *International Journal of Computer Science and Network Security,* vol. 17, no. 6, pp. 22-28, 2017.

[17] Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering," *International Journal,* vol. 1, no. 6, pp. 90-95, 2013.

[18] Langtangen, Hans Petter. "Numerical computing in python," *Python scripting for computational science*, pp. 131-181, 2006.

[19] Danganan, Alvincent E., Ariel M. Sison, and Ruji P. Medina. "OCA: overlapping clustering application unsupervised approach for data analysis," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* vol. 14, no. 3, pp. 1471-1478, 2019.

[20] Mohammed Ibrahim and Mahdi Nsaif Jasim. "New Modified Dynamic Clustering Algorithm," *Journal of Engineering and Applied Sciences*, vol. 14, no. 18, pp. 6742-6746, 2019.

[21] Mahdi, Muhammed U. "Determining Number & Initial Seeds of K-Means Clustring Using GA," *Journal of Babylon University and Applied Sciences,* vol. 18, no. 3, pp 1-6, 2010.

[22] Lailiyah, Siti, Ekawati Yulsilviana, and Reza Andrea. "Clustering analysis of learning style on anggana high school student," *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 17, no. 3, pp. 1409-1416, 2019.

[23] Karrar, Abdelrahman Elsharif, Marwa Abdelhameed Abdalrahman, and Moez Mutasim Ali. "Applying K-Means Clustering Algorithm to Discover Knowledge from Insurance Dataset Using WEKA Tool," *The International Journal of Engineering and Science*, vol. 5, no. 10, pp. 35-39, 2016.

[24] Arellano-Verdejo, Javier, et al. "Efficiently finding the optimum number of clusters in a dataset with a new hybrid cellular evolutionary algorithm," *Computación y Sistemas,* vol. 18, no. 2, pp. 313-327, 2014.

[25] Bholowalia, Purnima, and Arvind Kumar. "EBK-means: A clustering technique based on elbow method and k-means in WSN," *International Journal of Computer Applications*, vol. 105, no. 9, 2014.