A Comparative Study of Centroid and Medoid based Categorical Data Clustering Methods for Solving Cold-start Recommendation Problem

Noor Ifada
Informatics Department
University of Trunojoyo Madura
Bangkalan, Madura, Indonesia
noor.ifada@trunojoyo.ac.id

M. Eko Ariyanto
Informatics Department
University of Trunojoyo Madura
Bangkalan, Madura, Indonesia
m.eko7229@gmail.com

Mochammad Kautsar Sophan
Informatics Department
University of Trunojoyo Madura
Bangkalan, Madura, Indonesia
kaustar@trunojoyo.ac.id

Moh Nikmat
Informatics Department
University of Trunojoyo Madura
Bangkalan, Madura, Indonesia
moh.nikmat12@gmail.com

***Abstract*—One efficient solution to solve the cold-start recommendation problem is by exploiting the user demographic information using a clustering method. As the user demographic information contains categorical data, the choice of the clustering method to be used must naturally suitable to the particular data characteristic. There are two popular heuristic categorical data clustering algorithms, i.e., centroid and medoid based. This paper conducts a comparative study towards the implementation of K-Modes of the centroid-based method and K-Approximate Modal Haplotype (K-AMH) of the medoid-based method for solving the cold-start recommendation problem. The experiment results on the MovieLens dataset show that K-AMH achieves the average performance increase of 0.51% in terms of Precision and 0.40% in terms of Normalized Discounted Cumulative Gain (NDCG) to K-Modes. Yet, K-Modes is more lenient to use due to its scalability.**

Keywords— categorical data, centroid-based clustering, cold-start, k-modes, k-amh, medoid-based clustering, recommendation

I. INTRODUCTION

The implementation of a recommendation system is beneficial to narrow down the list of available options that are offered by the system to its users. The list of recommendations is usually generated based on the user' previous rating history [1, 2]. However, a system will struggle when a new user is registered. In this case, the recommendation system must take extra effort to understand the target user's interest due to no previous rating history recorded from the user. Such a condition is common to occur and is well-known as the cold-start problem [2-5].

One efficient way out to solve the cold-start problem is by exploiting the user demographic information using the clustering methods, i.e., for finding related users that might influence the target user interest [6-8]. As the user demographic information contains categorical data, the choices of the clustering methods to be used must naturally suitable to the particular data characteristic. There are two popular heuristic categorical data clustering algorithms, i.e., centroid and medoid based. The main difference between the centroid and medoid based methods is on how the centroid is represented [9]. The centroid-based clustering method represents the centroid of each cluster by the centre of the inclination calculation of data points. On the other hand, the medoid-based clustering method represents the centroid of a cluster as an object.

This paper conducts a comparative study towards the implementation of K-Modes of the centroid-based method [10-13] and K-Approximate Modal Haplotype (K-AMH) of the medoid-based method [9, 14] for solving cold-start recommendation problem. Our goal is to know which method is best used given the challenge. To our best knowledge, such a comparative study has not been done before. The experiments are conducted using a real-world movie dataset and the recommendation qualities are evaluated in terms of Precision and Normalized Discounted Cumulative Gain (NDCG) metrics. The recommendation task is to generate Top-N list of movies to a new or cold-start target user based on his or her user's demographic information. The categorical data clustering methods are implemented to find in which cluster does the target user belongs to. Hence, the list of recommendations generated for a target user is influenced by the list of movies rated by those users of the same cluster. The comparisons are analyzed in terms of the sensitivities, scalabilities, and performances.

The subsequent sections of this paper are: Section II lists the notations used in this paper. Section III details the concepts and algorithms of the K-Modes and K-AMH methods. Section IV describes the experiment setup. Section V details the results and discussion; while Section VI concludes the paper.

II. NOTATION

We assume that the demographic information of users is represented by a set of categorical attributes. In this case, let $A = \{a_1, a_1, \cdots, a_b\}$ be the set of b user categorical attributes. Assumed that $\dot{\boldsymbol{U}} = \{U_1, U_2, \cdots, U_m\}$ be the set of m users, therefore a user U_i is represented as $[u_{i,1}, u_{i,2}, \cdots, u_{i,b}]$. Table I show a toy example of the user demographic information. In this case, $A = \{gender, age, occupation\}$ and $\dot{\boldsymbol{U}} = \{U_1, U_2, U_3, U_4, U_5\}$ in which $U_1 = [\text{"}M\text{"}, \text{"25"}, \text{"}Writer\text{"}]$.

978-1-7281-8283-4/20/$31.00 ©2020 IEEE

TABLE I. TOY EXAMPLE OF USER DEMOGRAPHIC INFORMATION

$\dot{\boldsymbol{U}}$	$\boldsymbol{a_1}$	$\boldsymbol{a_2}$	$\boldsymbol{a_3}$
$\boldsymbol{u_1}$	M	25	Writer
$\boldsymbol{u_2}$	F	35	Artist
$\boldsymbol{u_3}$	M	35	Doctor
$\boldsymbol{u_4}$	F	18	Sales
$\boldsymbol{u_5}$	F	35	Doctor

III. METHODS

This section describes the two categorical data clustering methods implemented to investigate the cold-start recommendation problem, i.e.: K-modes of the centroid-based clustering and K- Approximate Modal Haplotype (K-AMH) of the medoid-based clustering.

Given a set of $\dot{\boldsymbol{U}}$ and an integer number k ($\leq m$), the purpose of the categorical data clustering methods are to group $\dot{\boldsymbol{U}}$ into k clusters.

A. Centroid-based Clustering: K-Modes

The centroid-based clustering method represents the centroid of each cluster by the centre of the inclination calculation of user data points [9]. K-Means is the most popular centroid-based clustering algorithm. However, it cannot be implemented for categorical data. Huang introduces K-Modes as the modification of K-means for clustering categorical data [10, 11]. A centroid in K-modes is represented as a vector of $Q = [q_1, q_2, \cdots, q_b]$ where Q is not necessarily an element of $\dot{\boldsymbol{U}}$.

The distance between a user U and a centroid Q in K-Modes is calculated as:

$$d(U, Q) = \sum_{j=1}^{b} \delta(u_j, q_j) \tag{1}$$

where

$$\delta(u_j, q_j) = \begin{cases} 0, & (u_j = q_j) \\ 1, & (u_j \neq q_j) \end{cases} \tag{2}$$

The objective of K-modes is to minimize the following cost function [10, 11]:

$$F(V, \boldsymbol{Q}) = \sum_{l=1}^{k} \sum_{i=1}^{m} \sum_{j=1}^{b} v_{i,l} \delta(u_{i,j}, q_{l,j}) \tag{3}$$

Subject to

$$\begin{matrix} \sum_{l=1}^{k} v_{i,l} = 1, & 1 \leq i \leq m \\ v_{i,l} \in \{0,1\}, & 1 \leq i \leq m, 1 \leq l \leq k \end{matrix} \tag{4}$$

Solved by

$$\begin{matrix} v_{i,l} = 1 & \text{if } d(U_i, Q_l) \leq d(U_i, Q_t) \text{ for } 1 \leq t \leq k \\ v_{i,t} = 0 & \text{for } t \neq l \end{matrix} \tag{5}$$

where $V = [v_{i,j}] \in \mathbb{R}^{m \times k}$ and $\boldsymbol{Q} = \{Q_1, Q_2, \cdots, Q_k\}$ is a set of centroid vectors. Fig. 1 shows the algorithm of K-Modes clustering.

Algorithm: K-Modes Clustering
Input: User demographic information
Step:
1. Randomly select k initial cluster centroids
2. Measure the distance between each user to each centroid according to Equation (1) and (2)
3. Assign each user to the cluster that has the shortest distance to the user. Repeat until all users are assigned to clusters
4. Update the centroids based on the modes
5. Compare the new centroids to the previous ones. Repeat Step 2 if they are different; otherwise, stop

Output: k users clusters

Fig. 1. Algorithm of K-Modes clustering [10, 11]

B. Medoid-based Clustering: K-Approximate Modal Haplotype (K-AMH)

The medoid-based clustering method represents the centroid of a cluster as a user [9]. One of the recent medoid-based clustering methods is the K-Approximate Modal Haplotype (K-AMH) [9, 14]. A medoid in K-AMH is represented is a vector of $H = [h_1, h_2, \cdots, h_b]$ where $H \in \dot{\boldsymbol{U}}$.

The distance between a user U and a medoid H in K-AMH is calculated as:

$$d(U, H) = \sum_{j=1}^{b} \delta(u_j, h_j) \tag{6}$$

where

$$\delta(u_j, h_j) = \begin{cases} 0, & (u_j = h_j) \\ 1, & (u_j \neq h_j) \end{cases} \tag{7}$$

The objective of K-AMH is to maximize the following cost function:

$$F(W, D) = \sum_{l=1}^{k} \sum_{i=1}^{m} w_{l,i}^{\alpha} d_{l,i} \tag{8}$$

To satisfy

$$F(W, D)^s > F(W, D)^t, s \neq t; \forall t, 1 \leq t \leq (m - k) \tag{9}$$

where

$$w_{l,i}^{\alpha} = \left(\begin{matrix} 1, & \text{if } U_i = H_l \\ 0, & \text{if } U_i = H_z, z \neq l \\ \dfrac{1}{\sum_{z=1}^{k} \left[\frac{d(U_i, H_l)}{d(U_i, H_z)} \right]^{\frac{1}{\alpha - 1}}} & \text{otherwise} \end{matrix} \right)^{\alpha} \tag{10}$$

Subject to

$$w_{l,i}^{\alpha} \in [0,1], \quad 1 \leq i \leq m, 1 \leq l \leq k \tag{10a}$$

and

$$0 < \sum_{l=1}^{m} w_{l,i}^{\alpha} < m, \quad 1 \leq l \leq k \tag{10b}$$

where

$$d_{l,i} = \begin{cases} 1.0, & \text{if } w_{l,i}^{\alpha} = \max^{w_{l,i}^{\alpha}, 1 \le l \le k} \\ 0.5, & \text{otherwise} \end{cases} \quad (11)$$

Subject to

$$d_{l,i} \in \{1,0.5\}, \quad 1 \le i \le m, 1 \le l \le k \quad (11a)$$

and

$$1.0 < \sum_{l=1}^{k} d_{l,i} < k, \quad 1 \le l \le m \quad (11b)$$

$$0.5 < \sum_{l=1}^{m} d_{l,i} < m, \quad 1 \le l \le k \quad (11c)$$

where:

- $W = \left[w_{l,i}^{\alpha}\right] \in \mathbb{R}^{k \times m}$ is the degree of membership matrix
- $\alpha \in [1, \infty)$ is a weighting exponent that is used to escalate the accuracy of the degree of membership
- $D = \left[d_{l,i}\right] \in \mathbb{R}^{k \times m}$ is the dominant weighting matrix

The complete algorithm of K-AMH clustering is presented in Fig. 2.

Algorithm: K-AMH Clustering
Input: User demographic information
Step:
1. Randomly select k initial cluster medoids
2. Measure the distance between each user to each medoid according to Equation (6) and (7)
3. Generate W matrix according to equation (10) that subjects to Equation (10a) and (10b).
4. Generate D matrix according to equation (11), (11a), (11b), and (11c)
5. Calculate cost function $F(W, D)$ according to Equation (8)
6. Test for each initial medoid by the other users. If the condition satisfies Equation (9), then replace the medoid
7. Repeat Step 2 up to Step 6 for each U and H
8. When the final medoids are acquired for all clusters, assign the users to their corresponding clusters

Output: k users clusters

Fig. 2. Algorithm of K-AMH clustering [9, 14]

IV. Experiment Setup

We use the MovieLens 100K dataset [15] in which the users have rated at least 20 movies. The dataset is filtered such that we only use the top-300 users that have rated at least 100 movies. It now contains 300 users, 1640 movies, and 62161 ratings. The recommendation task is to generate Top-N list of movies to a new or cold-start target user based on the user's age, gender, and occupation. The K-Modes and K-AMH methods are implemented to find in which cluster does a target user belong to. Hence, the list of recommendations generated for a target user is influenced by the list of movies rated by those users of the same cluster.

The experimentations implement the 5-fold cross-validation method in which the dataset is randomly split five times into 80% training and 20% test data. We make sure that the users of test data have no ratings in the training data to simulate the cold-start problem scenario. The recommendation performances are evaluated based on the Precision and Normalized Discounted Cumulative Gain (NDCG) metrics. The reported results are the average of the recommendation performances of all users in the test data.

V. Results and Discussion

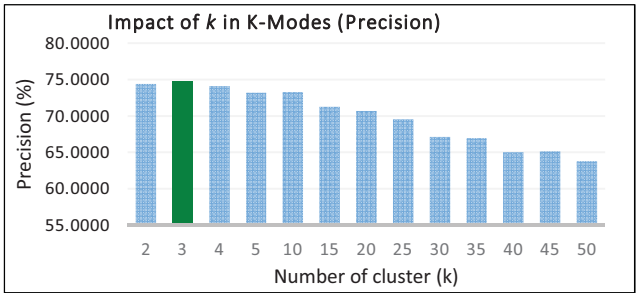
Through a series of experiments, we empirically analyze and compare the sensitivities,scalabilities, and performances of K-Modes and K-AMH for solving the cold-start recommendation problem.

A. Sensitivity

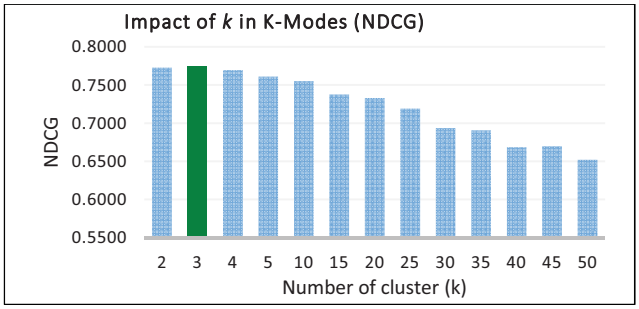
K-Modes. The sensitivity of K-Modes is analyzed in terms of the impact of the number of clusters (k) to the recommendation performances. We evaluate the performances based on the variations of $k = \{2,3,4,5,10,15,20,25,30,35,40,45,50\}$.

Fig. 3 displays the impact of k to K-Modes in terms of Precision and NDCG. The results show that K-Modes performs the best when $k = 3$ and that the larger k gradually deteriorates the K-Modes performances. These findings suggest that the least number of clusters is sufficient to leverage the performance of K-Modes. Moreover, a similar pattern results in both evaluation metrics suggest that the performance of K-Modes is stable on any evaluation metrics.

K-AMH. The sensitivity of K-AMH is analyzed in terms of the impacts of the weighting exponent (α) and the number of clusters (k) to the recommendation performances. We evaluate the performances based on the variations of $\alpha = \{1.1,1.2,1.3,1.4,1.7,1.8,1.9,2.0\}$, following the approaches in [14, 16], and $k = \{2,3,4,5,10,15,20,25,30,35,40,45,50\}$.



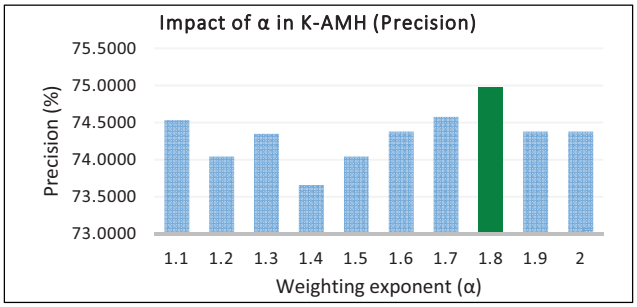
(a) Precision



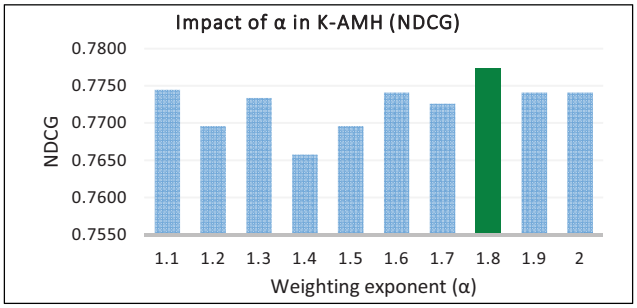
(b) NDCG

Fig. 3. Impact of k in K-Modes: (a) Precision and (b) NDCG

Fig. 4 shows that K-AMH oscillates until it performs the best when $\alpha = 1.8$ and remains stagnant afterwards, in terms of both Precision and NDCG. These behaviours confirm that the best α is typically between 1.1 and 2.0 [14, 16]. Meanwhile, Fig. 5 shows K-AMH performs the best when $k = 10$ and progressively decreases after that. These findings point out that K-AMH requires more number of clusters compared to K-Modes to leverage its performance. Additionally, the comparable shapes exposed based on the results in both Precision and NDCG evaluation metrics advise that the performance of K-AMH is stable on any evaluation metrics.

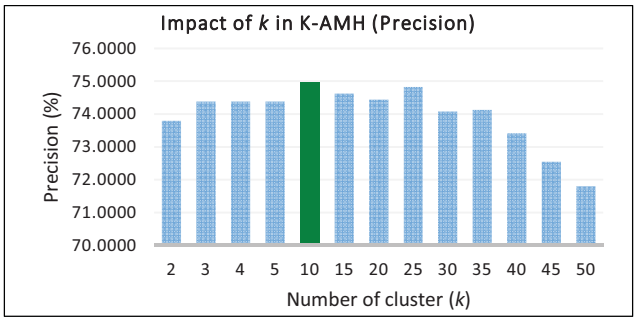


(a) Precision

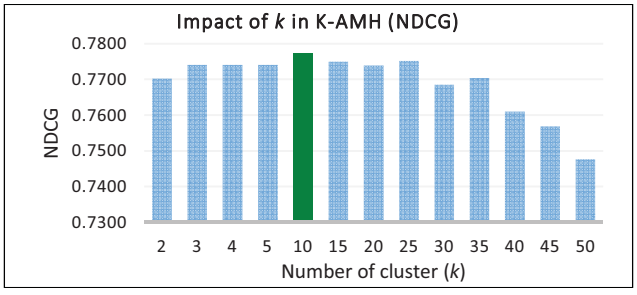


(b) NDCG

Fig. 4. Impact of α in K-AMH: (a) Precision and (b) NDCG



(a) Precision



(b) NDCG

Fig. 5. Impact of k in K-AMH: (a) Precision and (b) NDCG

B. *Scalability*

The scalabilities of K-Modes and K-AMH are compared based on the running time required to cluster 300 users into a various number of clusters (k). From the comparison shown in Fig. 6, we can visibly see that the clustering time of K-AMH is a lot higher than that of K-Modes. In fact, these results confirm those of the previous studies on K-Modes [12] and K-AMH [14] that the side-by-side comparison of their running time graphs indicates that the slope of the latter is significantly greater than that of the former. In this case, we can state that K-Modes is more scalable than K-AMH. Moreover, it is worthwhile to note that K-Modes is essentially the same as K-Means and consequently it is suitable for clustering large categorical data [11].

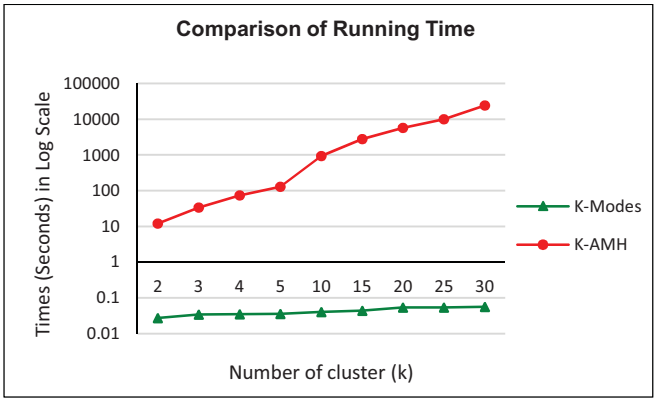


Fig. 6. The comparison of running time to cluster 300 users into various k

TABLE II. PERFORMANCE COMPARISON IN TERMS OF PRECISION

Top-N	Precision		Increase Percentage (K-AMH to K-Modes)
	K-Modes	*K-AMH*	
1	88.4790	88.4790	0.00%
2	85.7030	85.6981	-0.01%
3	83.3568	83.5496	0.23%
4	80.7234	81.1728	0.56%
5	79.5058	80.2730	0.97%
6	78.5529	79.0564	0.64%
7	76.4136	78.0800	2.18%
8	75.3810	76.0369	0.87%
9	75.5742	75.1123	-0.61%
10	74.7473	74.9702	0.30%
AVERAGE	**79.8437**	**80.2428**	**0.51%**

TABLE III. PERFORMANCE COMPARISON IN TERMS OF NDCG

Top-N	NDCG		Increase Percentage (K-AMH to K-Modes)
	K-Modes	*K-AMH*	
1	0.8848	0.8848	0.00%
2	0.8633	0.8633	0.00%
3	0.8453	0.8467	0.16%
4	0.8256	0.8288	0.38%
5	0.8152	0.8207	0.67%
6	0.8069	0.8109	0.49%
7	0.7912	0.8027	1.46%
8	0.7825	0.7879	0.70%
9	0.7817	0.7801	-0.20%
10	0.7748	0.7774	0.34%
AVERAGE	**0.8171**	**0.8203**	**0.40%**

C. Performance

Table II and Table III respectively list the performance comparisons between K-Modes and K-AMH in terms of Precision and NDCG at various $N = \{1,2,3,4,5,6,7,8,9,10\}$. We can observe that K-Modes and K-AMH have a comparable performances when $N \leq 2$, though the latter is very slightly left behind in terms of Precision when $N = 2$. K-AMH constantly over performs K-Modes to some extent when $N \geq 3$, except when $N = 9$. The Precision and NDCG percentage increases displayed in Table II and Table III show that the average of increases from K-AMH to K-Modes are respectively 0.51% and 0.40%. In this case, the outperformance of K-AMH towards K-Modes is not significant. Moreover, recall that the scalability analysis presented in the previous section concludes that K-Modes is more scalable than K-AMH according to the comparison of their clustering running times. In other words, indeed K-AMH performs slightly better than K-Modes for solving the cold-start recommendation problem; however, K-Modes is still more worth it to be implemented due to scalability reason.

VI. Conclusion and Future Work

We have presented a comparative study of K-Modes and K-AMH as respectively the centroid and medoid based categorical data clustering methods for solving cold-start recommendation problem. Our series of experiments on the MovieLens dataset show that K-AMH achieves the average performance increase of 0.51% in terms of Precision and 0.40% in terms of NDCG to K-Modes. K-AMH performs slightly better than K-Modes. Yet, K-Modes is more lenient to use due to its scalability.

Future works are to use other centroid and medoid based clustering methods in the comparative study as well as to use other datasets with different sizes.

Acknowledgment

This study is supported by the Directorate of Research, Technology and Higher Education (Indonesia), grant scheme: “Penelitian Dasar”, financial year: 2020.

References

[1] J. A. Konstan and J. Riedl, "Recommender systems: from algorithms to user experience," *User Modeling and User-Adapted Interaction,* vol. 22, pp. 101-123, 2012.

[2] C. C. Aggarwal, *Recommender Systems: The Textbook*. Switzerland: Springer International Publishing, 2016.

[3] G. Guo, J. Zhang, and D. Thalmann, "Merging trust in collaborative filtering to alleviate data sparsity and cold start," *Knowledge-Based Systems,* vol. 57, pp. 57-68, 2014.

[4] S. Loh, F. Lorenzi, R. Granada, D. Lichtnow, L. K. Wives, and J. P. de Oliveira, "Identifying Similar Users by their Scientific Publications to Reduce Cold Start in Recommender Systems," in Proceeding of *5th International Conference on Web Information Systems and Technologies (WEBIST 2009)*, pp. 593-600, 2009.

[5] N. Ifada and R. Nayak, "An Efficient Tagging Data Interpretation and Representation Scheme for Item Recommendation," in Proceeding of *The 12th Australasian Data Mining Conference*, Brisbane, Australia, pp. 205-215, 2014.

[6] L. Yanxiang, G. Deke, C. Fei, and C. Honghui, "User-based clustering with top-n recommendation on cold-start problem," in Proceeding of *2013 Third international conference on intelligent system design and engineering applications* Hong Kong, pp. 1585-1589, 2013.

[7] A. L. V. Pereira and E. R. Hruschka, "Simultaneous co-clustering and learning to address the cold start problem in recommender systems," *Knowledge-Based Systems,* vol. 82, pp. 11-19, 2015.

[8] D. Zhang, C. H. Hsu, M. Chen, Q. Chen, N. Xiong, and J. Lloret, "Cold-start recommendation using bi-clustering and fusion for large-scale social recommender systems," *IEEE Transactions on Emerging Topics in Computing,* vol. 2, pp. 239-250, 2013.

[9] A. Seman, Z. A. Bakar, A. M. Sapawi, and I. R. Othman, "A medoid-based method for clustering categorical data. ," *Journal of Artificial Intelligence,* vol. 6, pp. 257-265, 2013.

[10] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery,* vol. 2, pp. 283-304, 1998.

[11] J. Z. Huang, "Clustering categorical data with k-Modes," in *Encyclopedia of Data Warehousing and Mining (Second Edition)*, ed Hershey, New York: IGI Global, 2009, pp. pp. 246-250.

[12] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the k-Modes clustering algorithm," *Knowledge-Based Systems,* vol. 26, pp. 120-127, 2012.

[13] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-modes clustering," *Expert Systems with Applications,* vol. 40, pp. 7444-7456, 2013.

[14] A. Seman, Z. A. Bakar, and M. N. Isa, "An efficient clustering algorithm for partitioning Y-short tandem repeats data," *BMC Research Notes,* vol. 5, p. 557, 2012.

[15] F. Harper and J. Konstan, "The movielens datasets: History and context," *ACM Transactions on Interactive Intelligent Systems,* vol. 5, pp. 1-19, 2015.

[16] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE transactions on Fuzzy Systems,* vol. 7, pp. 446-452, 1999.