Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 460−465
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1236

Movie Recommendation System Using Knowledge-Based Filtering and K-Means Clustering

Kurnia Drajat Wibowo, Z K A Baizal*

School of Computing, Informatics Study Program, Telkom University, Bandung, Indonesia
Email: [1] kurniadrajatw@student.telkomuniversity.ac.id, [2,*]baizal@telkomuniversity.ac.id
Correspondence Author Email: baizal@telkomuniversity.ac.id
Submitted: **26/01/2022**; Accepted: **24/02/2022**; Published**: 31/03/2022**

Abstract−The movie recommender system has an important role in providing movie recommendations for users, but new users have difficulty choosing movies that are given by the recommender system because of the cold start problem. This study aims to overcome the cold start problem using a knowledge-based recommender system, i.e association rule mining using an apriori algorithm. The apriori algorithm aims to extract correlations between product itemsets, but the problem in the apriori algorithm is the large number of association rules that make the complex computation. To overcome this problem, we combine the apriori algorithm and k-means to produce more accurate recommendations, because the items are grouped before the recommendation process using the k-means algorithm. In this study, we use a dataset of movies and ratings from the Kaggle website. This study uses a minimum value of 0.5 confidence, and a minimum value of 4 lifts. To produce the best itemset in the form of antecedents and consequents of the Beauty and the Beast item with The Passion of Joan of Arc which has a value of 0.107981 support, 0.779661 confidence, 4.151695 lift.

Keywords: Apriori; K-Means; Association Rule; Recommender System; Cold Start Problem

1. INTRODUCTION

Currently, data is spreading very fast, causing information overload in various business sectors, including the film industry. The information overload makes users feel difficult for choosing the relevant movies. Therefore, a recommender system is needed to avoid information overload. The role of the recommender system in everyday life has been widely applied, such as Facebook in recommending friends, Twitter in recommending status, and Instagram in recommending videos through the reels feature. Many buying and selling transactions implement a recommender system so that it makes it easy for users to choose products, and sellers are also easy to promote products [1]. Collaborative filtering is one of the paradigms in the recommender system area. However, this paradigm has drawbacks, i.e cold start problem [2]. This problem occurs when new items or users enter the recommender system. To overcome these shortcomings, we develop a knowledge-based recommender system. The knowledge-based recommender system has the advantage that it can overcome the cold start problem [3]. The knowledge-based filtering recommender system is an effective method because it provides recommendations based on user preferences by measuring existing variables [4].

Various types of association rules in data mining algorithms such as apriori, partition, pincer-search [5]. This study, we use and combine the k-means clustering algorithm with apriori. The implementation of the clustering algorithm is widely used in fraud detection applications and email crackers with their behavior based on characteristics [6], so it is suitable to be applied in this study to group items based on their characteristics. The k-means clustering algorithm functions to find and group data that have data similarities between one another [7], by calculating the centroid distance using Euclidean distance serves as a determinant of the location where the centroids will be grouped, the closer the distance between the centroids and other centroids is, the closest cluster will also be selected. Furthermore, principal component analysis (PCA) is used to reduce the complexity better [8]. Apriori algorithm to create association rules, one of its functions is to search for and find the same relationship or correlation between itemsets [9]. However, the weakness of this algorithm is that it has a very large number of rules so that it requires a long execution time [10], but this can be overcome by the k-means algorithm, by grouping items by rule [9], the Apriori algorithm relies on frequent itemset iterations and involves making association rule [11].

Fauzan, et al. [12] stated that the apriori association rule method resulted in a good course recommender system. The result of the research is to find the best parameters, which are found to be 0.01 for minimum support, and 0.6 for minimum confidence, with these two parameters producing a canvas network of 110 rules and an average lift of 19,055, while the Harvard-MITx dataset produces 48 rules and the average lift ratio is 3,662. The conclusion from this study is that the difference in the minimum support and minimum confidence values does not mean that the lift ratio value is smaller if the minimum support and minimum confidence values are greater. Dharsinni, et al. [13] stated that after comparing the apriori algorithm with a combination of apriori + k-means in the dataset, then calculating the computational time and comparing the results, the combination of apriori algorithm with k-means resulted in a faster computational time compared to apriori alone. The combination of the apriori algorithm with k-means resulted in 17.41 minutes while apriori alone was 21.93 minutes.

Meanwhile, AlZoubi et al. [14] show that Association Rule Mining is a tool for creating a recommender system for a set of students in a subject. This study looks at several different factors in the recommender system, especially the confidence value and the matching rule may not be the best decision, but by choosing a relatively high value of the confidence or match rule we can get better performance. Pradana, et al. [15] results from research looking for

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 460−465
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1236

support and confidence values obtained more accurately using data mining. Based on system testing using the apriori algorithm, it produces a derivative rule from a combination of two items, with the highest association rule, that is New Balance with Adidas getting a value of 87.5% and the lowest being Adidas with Puma, which is 18.2%.

Based on the studies previously mentioned, we combined the apriori (knowledge-based) algorithm with k-means clustering to deal with the cold start problem, therefore this research hopes to reduce the complexity and increase accuracy. Our contribution in this research is that we first designed a recommender system at the beginning with the latest research literature review, made documentation, until the results of later research, then we searched for datasets that will be used in this study, furthermore, we began to implement a recommender system. has been designed, then testing and evaluating the results of the implementation, finally making research documentation. This research is composed of 4 steps, we make an introduction in step 1 which contains research problems, related research, advantages, and disadvantages of research. Step 2 the research methodology contains the stages of research and related research stages. Step 3 is the results of the study, and the last step is conclusions from the research.

2. RESEARCH METHODOLOGY

2.1 System Design

In this study, we use the apriori association rule algorithm which is one of the algorithms in knowledge-based recommender system. The first step in this research we use the K-means clustering algorithm to group similar items, furthermore we use the apriori algorithm to build association rules for each itemset that has been grouped by the k-means algorithm. An overview of the system diagram is shown in Figure 1.

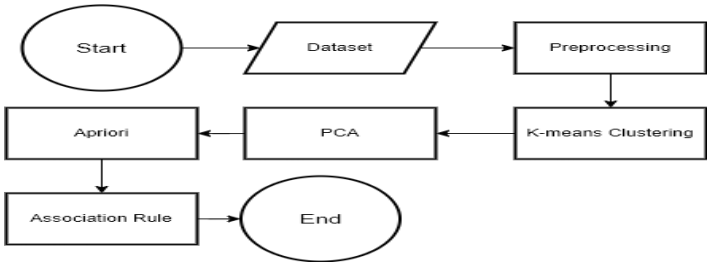

Figure 1. System Design

Based on the system design algorithm in Figure 1, in the first step we imported the dataset and library. The second step is we preprocessed data such as missing value handler by changing the null value to zero, dropping columns or features that were not needed. The third step is we use the k-means clustering algorithm to group the data, furthermore PCA is used to separate the main components by reducing the dimensions to 2 dimensions, x and y.The fifth step is the apriori algorithm used to generate the itemset of antecedents and consequents, and the next step is using apriori algorithm. The last step is the association rule to calculate the support, confidence, and lift values from the itemset.

2.2 Dataset

In this study, we use 2 datasets from the kaggle website which contains 45467 movie metadata and 100005 rating data. Movie dataset has attributes id, title, genre, while rating dataset has attributes userid, movieid, rating, timestamp as shown in table 1 and table 2.

Table 1. Movie Meta Dataset

Id	Title	Genre
862	Toy Story	[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name': 'Family'}]
884	Jumanji	[{'id': 12, 'name': 'Adventure'}, {'id': 14, 'name': 'Fantasy'}, {'id': 10751, 'name': 'Family'}]
15602	Grumpier	[{'id': 10749, 'name': 'Romance'}, {'id': 35, 'name': 'Comedy'}]

Table 2. Rating Dataset

UserId	MovieId	Rating	Timestamp
1	31	2.5	1260759144
1	1029	3	1260759179
1	1061	3	1260759182

Table 2 shows an example of rating data, in the table, there are columns for userid, movieid, rating, and timestamp. userid is the id of each user, while movieid is the id of each movie that is rated by the user, furthermore rating is the value given by the user for each movieid, and timestamp is the time when giving a rating.

2.3 Preprocessing

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 460−465
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1236

The preprocessing stage, the system process movie, and rating data before being processed into a recommender system. First, the system combines 2 datasets into 1 table, then perform missing values handlers such as filling in null columns with zero values, furthermore we drop columns or features that are not needed, remove negative values, and data encoding changes the value to 1 or 0, data will be worth 1 if the movie has a rating, while 0 if the movie does not have a rating.

2.4 K-means Clustering

The system uses k-means clustering algorithm to group items based on attributes into several partitions, where the number of k <number of items [16]. The first step k-means clustering algorithm is to determine the number of clusters, set the initialization of the cluster, calculate the distance between the centroids with the euclidean distance, create a new cluster with the closest distance, repeat the step of calculating distance centroids.

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_i - x_j)^2} \quad (1)$$

Formula 1 has several arithmetic operations and variables, variable d as the level of object difference, then n as the number of vector objects, k as cluster centroid starting from 1, then x_i is the input vector minus the comparison vector x_j.

2.5 Principal Component Analysis (PCA)

PCA is used to separate the main components after grouping using the k-means clustering algorithm. It aims to reduce the complexity and divided into 2, the x and y axes, where the x axis is the coordinate on the horizontal line, while the y axis is the coordinate on the vertical line.

2.6 Apriori

This study, we use the apriori algorithm to create association rules on the grouped itemset. The apriori algorithm is useful for reducing the search for the number of candidate items in the itemset creation process [12]. With the association rule mining technique, we can analyze several itemset that may be recommended, for example, users will buy bread with milk drinks, by taking advantage of this, placing goods in supermarkets will be more profitable [15]. The principle of this algorithm is bottom-up and breadth-first, if an itemset appears frequently then a subset of the itemset must also appear frequently [17].

2.7 Association Rule

The association rule process is carried out to obtain the strongest association rule based on the minimum value of support and confidence itemset, the system performs a selection process for each itemset [12].

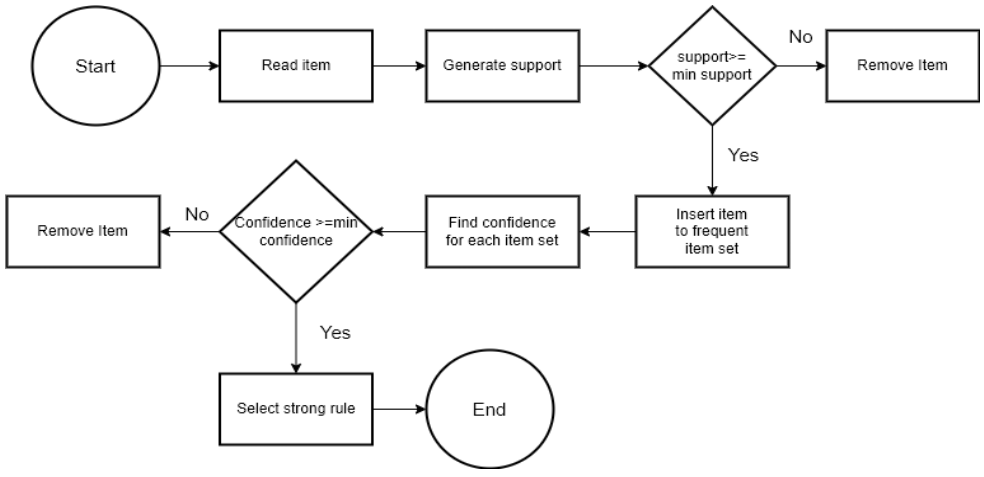

Figure 2. The Flow of Association Rule Mining

Figure 2 shows the flow of association rule mining, first, the apriori algorithm will read items and generate support for each item. We provide a minimum support value to produce the best support value, if the support value is more than the minimum support then insert into frequent itemset and otherwise. The system gives a confidence value from each itemset. We provide a minimum confidence value to produce the best confidence value, if the confidence value is below the minimum confidence it will be eliminated, and otherwise.

There are support and confidence values, both metrics are used to measure the level of association rule. Support is the level of how often the combination transaction of antecedent and consequent items appear together, while confidence is the credibility value of the strength of the association rule with the probability of a combination of two items [18] with the following formula (2-3),

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 460−465
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1236

$$support(A,B) = P(A \cup B) = \frac{A \cup B}{D} \quad (2)$$

$$confidence(A,B) = \frac{support(A,B)}{support(A)} = \frac{P(A \cup B)}{P(A)} \quad (3)$$

$P(A \cup B)$ is probability of total items A and B, where A and B are the total of transactions from the different items, meanwhile, D is the total transactions of all items.

3. RESULT AND DISCUSSION

3.1 Preprocessing Result

Preprocessing step produces two kinds of rating (true and false). The rating is true, if the user gives a rating to the movie title, and otherwise. Then the data encoding results from the rating will be worth 1 if the rating is true, however will be zero if the rating is false. Table 3 shows an example userid 2 is true or one when userid 2 gives a rating to the title of the movie Gator Bait, if it is zero or false then userid 2 does not give a rating to the title of the movie.

Table 3. Preprocessing Result

UserId	Title				
	Women Art Revolution	Gator Bait	Twas the Night Before Christmas	Zozo	eXistenZ
1	1	0	0	1	0
2	0	1	0	0	0
3	0	0	0	0	0

3.2 K-means Clustering

We use the elbow method to determine the number of clusters because this method is the best method for determining or calculating the number of clusters in the k-means clustering algorithm. We use sum square error (SSE) to reduce distortion or square of the number of errors by forming a square graph[19].

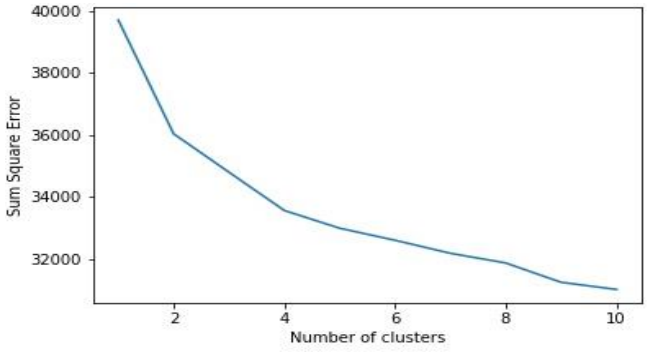

Figure 3. Comparison of SSE Values for Each Number of Clusters

Figure 3 shows the comparison value between SSE and the number of clusters. To determine the best number of clusters, we look at the elbow or the one with the deepest descent, Figure 3 shows that k=4 is the best number of clusters because it has the sharpest angled decline. Based on the results of figure 3, we can group the items with k-means clustering algorithm. Table 4 shows an example of the results from data clustering. For example, userid 15 and 23 are in the same cluster 2.

Table 4. User Cluster Result

UserId	Title					Cluster
	Gator Bait	À nos amours	Items or Less	Things I Hate About You	Zombie Holocaust	
15	1	1	0	0	0	2
23	0	0	1	0	0	2
40	0	0	0	0	0	3

3.3 Principal Component Analysis (PCA)

PCA can separate components into two, the x and y axes aims to reduce the data dimension. Based on the type of cluster, each cluster has different coordinate. Table 5 shows there are 3 examples of PCA data, in cluster 1 it has similar coordinates for example in userid 1 it has the x axis coordinate value -2.304881, meanwhile userid 2 has x axis coordinates at -1.22755.

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 460−465
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1236

Table 5. PCA

Title	UserId	Cluster	X	Y
0	1	1	-2.304881	-0.275198
1	2	3	0.116551	3.607165
2	3	1	-1.227557	-0.170187

Figure 4 shows the separation of the main components based on clusters that have been grouped with the k-means clustering algorithm. Cluster 1 shows in gray and is between 2 to 6 on the y axes.

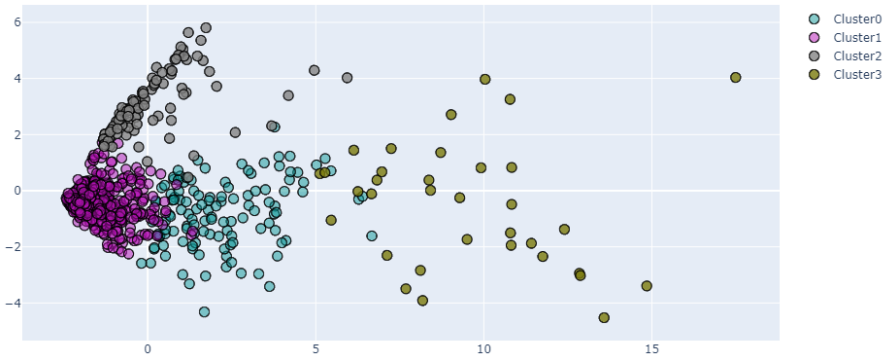

Figure 4 Cluster PCA

3.4 Result

Based on the result of table 4 and table 5, this process is carried out to merge table 4 and table 5 before the apriori algorithm process to create a frequent itemset table. Table 6 shows the merge results of table 4 and table 5, which x and y is the coordinate value of PCA. Furthermore, table 6 shows that userid 1 has the same x and y values. It shows a correlation between cluster and PCA.

Table 6. Merge Table Cluster and PCA

	UserId	Cluster	X	Y	Title	Rating
0	1	1	-2.304881	-0.275212	Confidentially Yours	1
1	1	1	-2.304881	-0.275212	Greed	1
2	1	1	-2.304881	-0.275212	Jay and Silent Bob Strike Back	1

The process to implement the apriori algorithm, we use an itemset table. We apply the minimum support of 0.07 which result 151 association rules. Table 7 shows an example of the support value of itemset table. For example, itemset Monson Wedding with Terminator 3: Rise of the Machines have a support value of 0.071599. The first row and second row shows the correlation because it has the same value of support and has the same item i.e Terminator 3: Rise of the Machines.

Table 7. Support Value Of Itemset

	Support	Itemsets
109	0.071599	(Monsoon Wedding, Terminator 3: Rise of the Machines)
137	0.071599	(Terminator 3: Rise of the Machines, Young and Innocent)
84	0.071599	(Arlington Road, The 39 Steps)

The result of the apriori algorithm is an association rule, which is a rule that connects antecedents and consequents. Antecedents are items from the rule, meanwhile, consequents are related items from the rule. We use the association rule parameters, lift >=4 and confidence >=5, table 8 shows the result.

Table 8. Result of Association Rule

	Antecedents	Consequents	Support	Confidence	Lift
0	(Beauty and the Beast)	(The Passion of Joan of Arc)	0.107981	0.779661	4.151695
1	(The Passion of Joan of Arc)	(Beauty and the Beast)	0.107981	0.575000	4.151695

Based on the results of the association rule, there are Beauty and the Beast items with The Passion of Joan of Arc which have a support value of 0.107981, confidence 0.779661, and lift 4.151695. This value becomes the best itemset because it has the highest confidence and lift value, furthermore has the same itemset in antecedents and consequents.

Building of Informatics, Technology and Science (BITS)
Volume 3, No 4, Maret 2022 Page: 460−465
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i4.1236

4. CONCLUSION

The conclusion of the research that has been implemented is, the movie recommender system using apriori (knowledge-based) recommender system and k-means clustering got the best association rule values, the value is 0.107981 support, 0.779661 confidence, and 4.151695 lift. The best association itemset result is Beauty and the Beast with The Passion of Joan of Arc, this is obtained based on the minimum confidence metric value >=0.5 and the minimum lift value >= 4. This research might be better if using another recommender system algorithm or using a hybrid recommender system. The results of this study are expected to help further researchers in developing a better recommender system and can help the wider community in the future.

REFERENCES

[1] H. El Bouhissi, M. Adel, A. Ketam, and A. B. M. Salem, "Towards an efficient knowledge-based recommendation system," CEUR Workshop Proc., vol. 2853, pp. 38–49, 2021.

[2] S. B. Ud Duja et al., "A proposed method to solve cold start problem using fuzzy user-based clustering," Int. J. Adv. Comput. Sci. Appl., no. 2, pp. 529–536, 2020, doi: 10.14569/ijacsa.2020.0110267.

[3] Z. K. A. Baizal, D. H. Widyantoro, and N. U. Maulidevi, "Design of knowledge for conversational recommender system based on product functional requirements," *Proc. 2016 Int. Conf. Data Softw. Eng. ICoDSE 2016*, 2017, doi: 10.1109/ICODSE.2016.7936151.

[4] Z. K. A. Baizal, D. H. Widyantoro, and N. Ulfa, "Data & Knowledge Engineering Computational model for generating interactions in conversational recommender system based on product functional requirements," *Data Knowl. Eng.*, no. March, p. 101813, 2020, doi: 10.1016/j.datak.2020.101813.

[5] S. Vijayarani and S. Sharmila, "Comparative analysis of association rule mining algorithms," *Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016*, vol. 2016, 2016, doi: 10.1109/INVENTIVE.2016.7830203.

[6] H. N. Ravuvar, H. Goda, R. Sumathi, and P. Chinnasamy, "Smart health predicting systemusing K-means algorithm," *2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020*, pp. 22–25, 2020, doi: 10.1109/ICCCI48352.2020.9104206.

[7] M. R. Yudhanegara, S. W. Indratno, and R. K. N. Sari, "Clustering for Item Delivery Using Rule-K-Means," J. Indones. Math. Soc., vol. 26, no. 2, pp. 185–191, 2020, doi: 10.22342/jims.26.2.871.185-191.

[8] C. Langensiepen, A. Cripps, and R. Cant, "Using PCA and K-Means to predict likeable songs from playlist information," *Proc. - 2018 UKSim-AMSS 20th Int. Conf. Model. Simulation, UKSim 2018*, pp. 26–31, 2018, doi: 10.1109/UKSim.2018.00017.

[9] A. Dahbi, M. Mouhir, Y. Balouki, and T. Gadi, "Classification of association rules based on K-means algorithm," Colloq. Inf. Sci. Technol. Cist, vol. 0, pp. 300–305, 2016, doi: 10.1109/CIST.2016.7805061.

[10] S. D. Patil, R. R. Deshmukh, and D. K. Kirange, "Adaptive Apriori Algorithm for frequent itemset mining," *Proc. 5th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2016*, pp. 7–13, 2017,doi:10.1109/SYSMART.2016.7894480.

[11] W. Xueyuan and Y. Bo, "Design and implementation of an apriori-based recommendation system for college libraries," *Proc. - 2018 Int. Conf. Eng. Simul. Intell. Control. ESAIC 2018*, vol. 9, pp. 372–375, 2018, doi: 10.1109/ESAIC.2018.00094.

[12] F. Fauzan, D. Nurjanah, and R. Rismala, "Apriori Association Rule for Course Recommender system," Indones. J. Comput., vol. 5, no. 2, pp. 1–6, 2020, doi: 10.21108/indojc.2020.5.2.434.

[13] N. P. Dharshinni, F. Azmi, I. Fawwaz, A. M. Husein, and S. D. Siregar, "Analysis of Accuracy K-Means and Apriori Algorithms for Patient Data Clusters," *J. Phys. Conf. Ser.*, vol. 1230, no. 1, pp. 0–8, 2019, doi: 10.1088/1742-6596/1230/1/012020.

[14] W. Ahmad AlZoubi, "Cluster Based Association Rule Mining for Courses Recommendation System," *Int. J. Comput. Sci. Inf. Technol.*, vol. 11, no. 6, pp. 13–19, 2019, doi: 10.5121/ijcsit.2019.11602.

[15] H. A. Pradana, . Laurentinus, F. P. Juniawan, and D. Y. Sylfania, "Product Recommendation Systems using Apriori in the Selection of Shoe based on Android," no. Conrist 2019, pp. 311–318, 2020, doi:10.5220/0009909603110318.

[16] S. Gupta and S. Arora, "Handling Cold Start Problem in Recommender Systems By Clustering Demographic Attribute," *Int. J. Eng. Appl. Sci. Technol.*, vol. 1, no. August, pp. 59–63, 2016.

[17] R. B. R, "Recommendation System for Movie Cast and Crew using Datamining Algorithm," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 4, pp. 1495–1497, 2020, doi: 10.35940/ijeat.d7522.049420.

[18] J. Hong, R. Tamakloe, and D. Park, "Discovering Insightful Rules among Truck Crash Characteristics using Apriori Algorithm," *J. Adv. Transp.*, vol. 2020, 2020, doi: 10.1155/2020/4323816.

[19] A. B. A. Alwahhab, "Proposed Recommender System for Solving Cold Start Issue Using k-means Clustering and Reinforcement Learning Agent," *Proc. - 2020 2nd Annu. Int. Conf. Inf. Sci. AiCIS 2020*, pp. 13–21, 2020, doi: 10.1109/AiCIS51645.2020.00013.