JOIN (Jurnal Online Informatika)
Volume 7 No. 1 | June 2022: 110-115
DOI: 10.15575/join.v7i1.819

p-ISSN: 2527-1682
e-ISSN: 2527-9165

Implementation of Apriori Algorithm for Music Genre Recommendation

Michael Henry[1], Wiryanata Chandra[2], Amalia Zahra[3]
[1,2,3] Computer Science Department, BINUS Graduate Program, Master of Computer Science,
Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received October 3, 2021
Revised March 30, 2022
Accepted April 21, 2022
Published June 30, 2022

Keywords:

Music genre
Association rule
Apriori algorithm
Recommendation system
Data mining

ABSTRACT

Music interest is diverse yet enticing to be a part of knowledge discovery. It influences how people feel, study, work, etc. A lot of things are to be considered in producing brand new music with its correlation to its genre. We have already collected the dataset that we can utilize in this research, which is the history of every song listened to by several users in a total of 20.000 records from a million song dataset. This study implements the Apriori algorithm which can handle a large amount of data while simplifying the data to create a recommendation system where the result is a pattern from the music genre according to the interests of each user with the help of the RapidMiner tool. The purpose of this research is that the pattern which has been found can become a reference for music producers in terms of making or distributing their brand-new music. The result of the best combination of genres states that listeners of the rock genre will also hear the pop genre with a combination frequency of 50, support value of 21.2%, and confidence value of 51%.

Corresponding Author:

Wiryanata Chandra
Computer Science Department, BINUS Graduate Program,
Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia
JL. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
Email: wiryanata.chandra@binus.ac.id

1. INTRODUCTION

Music has become a part of an individual's life. It is a means of communication, a form of language between musicians and people who enjoy every genre of the music they put their heart into [1]. Nowadays, the technology of data mining is implemented in every field of research including music. By extracting a lot of data, new knowledge yet useful information and rules hidden behind massive amounts of data are gained effectively [2]. Data mining is divided into 2, namely supervised learning and unsupervised learning. The difference is in the function of the two where supervised learning is used to make predictions while unsupervised learning is used to find patterns from the data [3]. In this study, the data mining process that will be used is unsupervised learning. So that the input data does not require labels and trains the machine based on the structure of the data. Some of these unsupervised algorithms are k-means clustering [4], association rule [4, 5], fuzzy c-means [7]. It is possible to extract a lot of data at a lower cost due to the contribution of data collection and storage technology [8].

One possible practice of data mining is the implementation of the recommender system. Recommender system is a system which aims to provide recommendations to someone regarding something that is being sought based on existing experience [9]. In this case, the recommender system can help music producers find out about song trends in society and provide recommendations on what songs are suitable for production through extensive information [10].

Previous research on data mining stated its purpose for recommender systems. Recommender system is a process of giving suggestions to users according to their interest in the things that become recommendation material. This recommender system was also applied by the company for requirements engineering which stated that implementing FP Growth can show good performance in terms of extraction time and space

consumption [11]. Other than that, Content-based filtering and collaborative filtering techniques are also used to create a system that can provide recommendations about social networks to someone on a website [12] as well as a system capable of recommending local businesses which are most frequently visited by the community with the help of GPS data [13]. A lot of researchers use this method in various data mining processes. Various kinds of information on a company can also be processed to analyze various transactions that occur in the company. This data is then used to provide package promotions to consumers according to the rules/patterns obtained from the association rule mining process according to the value of confidence and support from each of the rules obtained [14]. For instance, the research conducted by [15] is aimed to improve association rule mining. They use this association rule algorithm to create a recommendation system for film data. The method used to modify the association rule algorithm is Potent-rule and community-aware recommender system (PRUCARS).

In this study, the proposed method is to implement apriori algorithm as a model to find a pattern where later, this pattern will be in the form of a set of song genres. This apriori algorithm is then used as a recommendation system. Other algorithms used for recommendation systems are content-based filtering and collaborative filtering as mentioned in the third paragraph. Those algorithms differ from apriori algorithm and therefore, this is what drives the idea of creating a recommendation system based on another algorithm where the algorithm is a priori. The difference between apriori algorithms and existing recommendation system algorithms is that apriori algorithms are descriptive analytics not predictive analytics. So that the factors cannot be directly compared between these algorithms. In addition, this apriori algorithm is used in this study because it is able to process large amounts of data and is also able to simplify that data. So, the purpose of this study is to create a recommendation system based on the apriori algorithm.

The rest of this paper is organized as follows. Section 2 explains the proposed methods, and Section 3 provides an explanation and analysis of the results obtained after using the proposed method. Finally, Section 4 outlines our conclusions.

2. METHOD

This study implements Apriori algorithm for the benefit it provides which is the ability of handling large amounts of data and simplifying the data itself. The dataset in this study is massive and therefore, the implementation of Apriori algorithm and association rule mining use an open-source application that can perform data mining processes with this Apriori algorithm, namely RapidMiner 9.9. This study uses RapidMiner 9.9 because of its simplicity in usage and its fast-computing performance, especially in implementing the Apriori algorithm. Figure 1 explains what steps are taken to implement the Apriori algorithm.

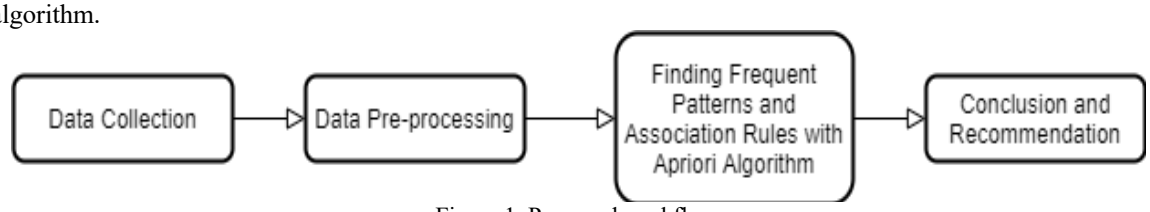

Figure 1. Proposed workflow

2.1. Data Collection

The purpose of this study is to provide a rule in the form of a song genre, which later will be given to music producers to produce music according to the interests of listeners so that the music they produce can be accepted by listeners. Therefore, to support the creation of a rule in the form of genre, the required dataset in this study is a playlist history dataset of several people with their preferred genre. With the need for such a dataset, the dataset is obtained from UCI machine learning repository, namely YearPredictionMSD Data Set in the form of user playlist history with user id, song id, and the number of plays for the song. The reason that dataset is used is because it is the appropriate dataset for this study. Lastly, This dataset contains 20.000 records history, 76.353 users and 10.000 song id.

2.2. Data pre-processing

After obtaining the dataset, the next step is data pre-processing where the dataset will be converted to fit the needs of this study. In this case, the process that will be carried out includes data cleaning, data reduction, set role, numerical to binomial and finally finding frequent genre and association rule.

The first step, we remove some noisy data. This process is called data cleaning. Noisy data contained in this dataset is in the form of missing song title data and song id data that is not in the playlist history so that such data is removed.

e-ISSN 2527-1682
p-ISSN 2527-9165

Table 1. Genre distribution

Genre	Total records
Blues	3
Country	33
Electronic / Dance	58
Hip-hop	39
Jazz	12
K-pop	26
Metal	46
Rock	98
Pop	146
Reggae	11
R&B	44
Gospel	16

As stated in table 1, the amount of data from 20,000 history records is reduced to 532 history records only due to the fact that the number of records of 20,000 will make the association rule search process take longer so reducing the data will speed up the rule mining process.

After cleaning the data, the user id is changed from the regular data type to id data type. This is done with the aim that when applying the Apriori, it will not affect the user id attribute. In other words, user id will not be used as one of the attributes that will be searched for.

Then after the set role process, the numerical to binomial operator will be carried out. This Numerical to Binomial operator, as from the meaning of the name itself, changes the type of the dataset attribute from a numerical data type to a binomial data type. For example, the pop genre attribute in user 001 is number 1 (numerical data type), so after entering the numerical to binomial operator, the number 1 will change to true/false depending on the specified min/max. In the dataset obtained, we did not determine the min/max because the contents of the dataset are already binary numbers, namely 0 and 1, where 0 means that the genre has never been heard and 1 means that the genre has been heard.

2.3. Finding association rules

After preprocessing the data, the dataset will be incorporated into this a priori algorithm. The architecture of this a priori algorithm can be seen in Figure 2. The first stage is that we have to determine the minimum support value to be used, which is 0.2. Then it will calculate the support value of each item in the dataset. This support value will later be used as a basis for whether this item will be eliminated or not. The item will be removed if it does not meet the minimum support value that has been determined. Furthermore, after selecting items that have a support value higher than the minimum support, then the next step is to determine the minimum confidence, which is 0.2. After that, the confidence value of each existing item will be calculated after passing the previous minimum support selection. If the item does not have a confidence value that matches the criteria, then the item will be eliminated. And the last stage is the process of forming the rules themselves which will be the main result of this research.

This study also provides the formula for support value and confidence value calculation. The formula for the support itself is defined in formula (1) below.

$$Support\ (X) = \frac{freq(X)}{N} \tag{1}$$

In (1), freq(X) indicates the frequency of the item whose value is to be calculated, while N indicates the total transactions in the dataset. The frequency of transactions where the item exists divided with the total number of transactions is the way to obtain the support value. In addition to support, there is also a parameter called confidence. Confidence is a comparison of the support value of a collection of items in a rule with the support value of an item that precedes it. So the formula for confidence is defined in formula (2) below.

$$Confidence\ (X \rightarrow Y) = \frac{support\ X\ \cup\ Y}{support\ X} \tag{2}$$

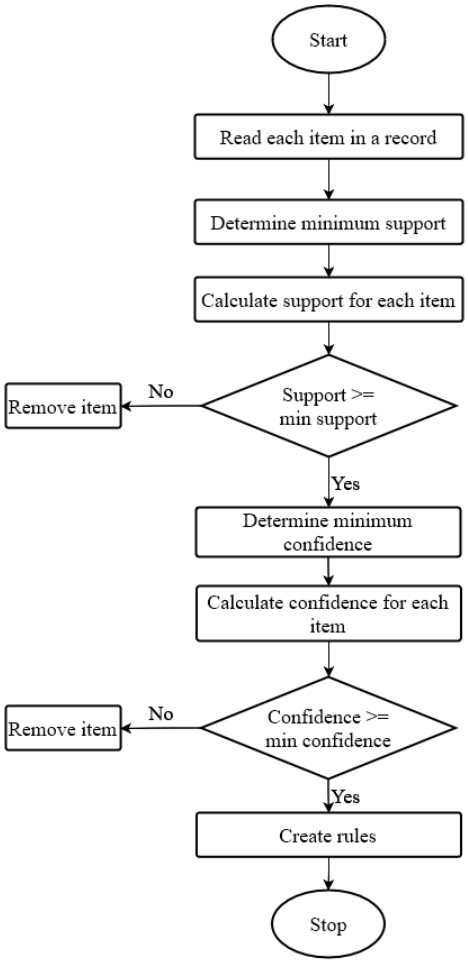

Figure 2. Association rules flowchart

In (2), support X indicates the support value of item X, and support Y indicates the support value of item Y. For the calculation, every rule that meets the requirements is collected at the rule table, i.e. if the resulting confidence of the rule has exceeded the minimum confidence that was defined at the beginning. With confidence value, we can measure how precise the relationship between itemset is in the Apriori/association rule. By determining the minimum support, we limit the percentage of frequent items set to suit our interest. Greater support of the item specified increases the possibility of the item combination to be selected [16]. We also determine the minimum confidence to 0.2 which means every relationship that has confidence value below 0.2 will be eliminated in this research. However, the important parameters that play a role in running Apriori/association rule are minimum support and minimum confidence. Confidence is a percentage of the relationship between several data/items.

Table 2. First 5 rows of experimental result

Premises	Premises Frequency	Conclusion	Conclusion Frequency	Premise and Conclusion Frequency	Support	Confidence
Rock	98	Pop	146	50	0.212	0.510
Pop	146	Rock	98	50	0.212	0.342
Electronic/Dance	58	Pop	146	27	0.114	0.466
Metal	46	Pop	146	24	0.102	0.522
RnB	44	Pop	146	20	0.085	0.455

3. RESULTS AND DISCUSSION

In Table 2, the result which delivers patterns of genre interest to music producers who want to distribute their music are presented. This result is given in excel format to make it more perceptible yet viable to read. The columns reviewed are the “Premise Frequency”, “Conclusion Frequency”, "Premise and

Conclusion Frequency", "Support" and "Confidence" columns where these columns have an important role in providing recommendations of what genres interest users the most. To provide a recommendation, the aspect to be looked up for is the “Premise and Conclusion Frequency” with the highest value first. After that, take the highest value based on "Support" and "Confidence" columns.

The explanation for each column is as follows, The “Premise Frequency” column indicates the amount of total frequency of the premise itself. For instance, in the first row of Table 2, the “Premise Frequency” column shows 98 Premise Frequency of the rock genre in the dataset, and so on. The “Conclusion Frequency” column indicates the amount of total frequency of the occurring conclusion. In the first row of Table 2, “Conclusion Frequency” column shows 146 Conclusion Frequencies of the pop genre in the dataset, and so on. The “Premise and Conclusion Frequency” column indicates the total frequency if there are premises and conclusions in a situation. As in Table 2, the first column with the value 50 represents the 50 occurring combination frequency of users who listen to rock and also listen to pop genres. The “Support” column indicates the support value. For instance, in the first row of Table 2, the premise of “rock” with the conclusion of “pop” has the support value of 0.212, which means that 21.2% of the total records of the dataset consist of users who listen to “Rock” and “Pop” genres. The “Confidence” column is the confidence value. For instance, in the first row of Table 2, the premise of “Rock” with the conclusion of “Pop” has a confidence value of 0.51, which means that for every user who listens to rock genre, 51% of them also listens to pop genre. In other words, this value represents the percentage of relationships of people who listen to rock also tend to listen to pop.

Also in table 2, the premise of “Pop” and the conclusion of “Rock” has the support value of 0.212, which means that 21.2% of the total records of the dataset consists of user who listens to “Pop” and “Rock” genre and for every user who listens to “Pop”, 34.2% of them also listens to “Rock” based on the confidence value of 0.342. This combination occurs 50 times based on the “Premise and Conclusion Frequency” column. Other pattern also shown that the premise of “Electronic/Dance” and the conclusion of “Pop” has the support value of 0.114, which means that 11.4% of the total records of the dataset consists of user who listens to “Electronic/Dance” and “Pop” and for every user who listens to “Electronic/Dance”, 46.6% of them also listens to “Pop” based on the confidence value of 0.446. This combination occurs 27 times based on the “Premise and Conclusion Frequency” column and so on.

Based on the results above, this study found that the combination of “Rock” and “Pop” genres is the best combination considering the Premise and conclusion frequency, support value, and confidence value.

4. CONCLUSION

The implementation of the Apriori algorithm for this study started from preprocessing the dataset that has been collected by handling noisy values to converting each value to binomial. Then, the next and important step of this study is finding association rules which started from support value calculation to confidence value calculation. Through the steps above, the results that have been obtained from this study conclude that the best genre combination to produce according to our dataset is the rock and pop genre with 50 combination frequency. Also, the combination of rock and pop has a 21.2% support value and a 51% confidence value. The result also shows other rules after rock and pop combinations which are electronic/dance and pop, metal and pop, and also RnB and Pop.

For future research, it can be accomplished by increasing the size of the dataset used, realizing that data size greatly affects the performance of Apriori algorithm/association rule mining. In addition, it is possible to improve such a performance by determining minimum support and minimum confidence. Finally, other algorithms can also be implemented in this recommender system.

5. REFERENCES

[1] K. E. Barkwell *et al.*, “Big data visualisation and visual analytics for music data mining,” *Inf. Vis. - Biomed. Vis. Vis. Built Rural Environ. Geom. Model. Imaging, IV 2018*, pp. 235–240, 2018, doi: 10.1109/iV.2018.00048.

[2] F. Liu, S. Zhang, J. Ge, F. Lu, and J. Zou, “Agricultural Major Courses Recommendation Using Apriori Algorithm Applied in China Open University System,” *Proc. - 2016 9th Int. Symp. Comput. Intell. Des. Isc. 2016*, vol. 1, pp. 442–446, 2016, doi: 10.1109/ISCID.2016.1109.

[3] P. D. Waggoner, “Unsupervised Machine Learning for Clustering in Political and Social Research,” *Unsupervised Mach. Learn. Clust. Polit. Soc. Res.*, 2020, doi: 10.1017/9781108883955.

[4] M. Sandeep Kumar and J. Prabhu, “A hybrid model collaborative movie recommendation system using K-means clustering with ant colony optimisation,” *Int. J. Internet Technol. Secur. Trans.*, vol. 10, no. 3, pp. 337–354, 2020, doi: 10.1504/IJITST.2020.107079.

[5] L. Yao, Z. Xu, X. Zhou, and B. Lev, “Synergies Between Association Rulesand Collaborative Filteringin

Recommender System: An Applicationto Auto Industry," *Data Sci. Digit. Bus.*, pp. 23–40, 2019, doi: 10.1007/978-3-319-95651-0_2.
[6] C. Wang and X. Zheng, "Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint," *Evol. Intell.*, vol. 13, no. 1, pp. 39–49, 2020, doi: 10.1007/s12065-019-00234-5.
[7] A. Gonzalez and F. Forsberg, "Unsupervised Machine Learning : An Investigation of Clustering Algorithms on a Small Dataset," pp. 1–39, 2017.
[8] S. V Hovale and P. G, "Survey Paper on Recommendation System using Data Mining Techniques," *Int. J. Eng. Comput. Sci.*, vol. 0869, no. 4, pp. 18–19, 2016, doi: 10.18535/ijecs/v5i5.60.
[9] M. K. Najafabadi, M. N. Mahrin, S. Chuprat, and H. Sarkan, "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining," *Comput. Human Behav.*, pp. 113–128, 2017, doi: http://dx.doi.org/10.1016/j.chb.2016.11.010.
[10] S. Alzu'Bi, B. Hawashin, M. Eibes, and M. Al-Ayyoub, "A Novel Recommender System Based on Apriori Algorithm for Requirements Engineering," *2018 5th Int. Conf. Soc. Networks Anal. Manag. Secur. SNAMS 2018*, pp. 323–327, 2018, doi: 10.1109/SNAMS.2018.8554909.
[11] M. Muhairat, S. Alzu'bi, B. Hawashin, M. Elbes, and M. Al-Ayyoub, "An intelligent recommender system based on association rule analysis for requirement engineering," *J. Univers. Comput. Sci.*, vol. 26, no. 1, pp. 33–49, 2020.
[12] F. Ali, T. Ahmad, A. M. Martinez-Enriquez, and A. Muhammad, "Data mining based recommendation system using social websites," *Proc. - 2015 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. WI-IAT 2015*, vol. 1, pp. 365–368, 2016, doi: 10.1109/WI-IAT.2015.78.
[13] J. Jooa, S. Bangb, and G. Parka, "Implementation of a Recommendation System Using Association Rules and Collaborative Filtering," *Procedia Comput. Sci.*, vol. 91, no. Itqm 2016, pp. 944–952, 2016, doi: 10.1016/j.procs.2016.07.115.
[14] M. Brilliant, D. Handoko, and Sriyanto, "Implementation of Data Mining Using Association Rules for Transactional Data Analysis," *3rd Int. Conf. Inf. Technol. Bus.*, pp. 177–180, 2017.
[15] I. B. E. Kouni, W. Karoui, and L. B. Romdhane, "Prucars: Improved association rule-based social recommender systems using overlapping community detection," *Procedia Comput. Sci.*, vol. 176, pp. 956–965, 2020, doi: 10.1016/j.procs.2020.09.091.
[16] M. Fauzy, K. R. Saleh W, and I. Asror, "Penerapan Metode Association Rule Menggunakan Algoritma Apriori pada Simulasi Prediksi Hujan Wilayah Kota Bandung," *J. Ilm. Teknol. Inf. Terap.*, vol. II, no. 2, pp. 221–227, 2016.