International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 5, No. 5, (2022), pp. 570-577

Implementation of Algorithms For Frequent Itemset In Forming Association Rules In Movie Recommendation System

Ilham Prayudha[1], Muhammad Habib Algifari[2]

[1,2]Informatics/Computer Science, Institut Teknologi Sumatera, Lampung, Indonesia

[1]ilham.14115001@student.itera.ac.id, [2]muhammad.algifari@if.itera.ac.id
**Corresponding author: [2]muhammad.algifari@if.itera.ac.id*

Abstract

A large number of movies around the world, causing a person to take a long time to find the movie they want to watch, not only that the audience will be confused to determine which movie suits their interests. A recommendation system is defined as a decision-making strategy for a user under complex information environments. From the perspective of e-commerce, the recommendation system was described as a tool that can help users decide related to user interest and preference [5]. The recommendation system is intended primarily for individuals who have no experience evaluating the number of alternative items offered, such as movie selection. This study will implement a recommendation system to form association rules from the two algorithms for frequent itemset, namely Apriori and FP-Growth.

Keywords: *Algorithms for Frequent Itemset, Association rules, Movie Recommendation System*

1. Introduction

A movie is one of the literary forms containing a story, play, history, culture, incidents, science, etc., that is recorded as a video and shown in cinema, television, theaters, or other broadcast media that is as entertaining as the primary purpose [1]. The development of movie production increases from year to year based on a report from the MovieLens website. Based on the MovieLens website, there were 62,423 movies in the dataset released in 2019, and on the Wikipedia website, in 2021, there were more than 200 new movies in the related dataset. A large number of movies around the world, causing a person to take a long time to find the movie they want to watch, not only that the audience will be confused to determine which movie suits their interests.

Association rule is a data mining technique to find associative rules between a combination of items [2]. The association rule technique will produce rules that can determine the tendency of subsequent transactions by different users from the count of the number of occurrences of item set combinations with the same pattern in a dataset. The primary step in the association rule is to find out how often a variety of things appears in the database [3]. The combination of these items can be found using a Frequent Itemset search technique, which is a technique to find out which items often appear as one or a collection of objects with a minimum display of support in a dataset [4].

A recommendation system is defined as a decision-making strategy for a user under complex information environments. From the perspective of e-commerce, the recommendation system was described as a tool that can help users decide related to user interest and preference [5]. The recommendation system is intended primarily for individuals who have no experience evaluating the number of alternative items offered, such as movie selection. This study will implement a recommendation system to form association rules from the two algorithms for Frequent Itemset, namely Apriori and FP-Growth.

ISSN : 2580-7250

2. Research Methodology

2.1. Recommendation System

A recommendation system is a system that provides a sequence of choices for the user to help them choose their desired product. Another definition, the recommendation system is defined as a decision making strategy for user under complex information environments. From the prespective of e-commerce, the recommendation system was defined as a tool that can help user to make a decision related to user interest and preference. The options provided by the recommendation system can help users in determining the decisions taken [6]. The simplest form of a recommendation system is to create a list of things based on user preferences. Information systems can overcome users who have difficulty choosing services or items by implementing an algorithmic approach such as collaborative filtering. Therefore, the recommendation system will act as an information filter to predict user preferences, provide choices or items related to users, and predict items liked by users.

2.1.1. Recommendation Techniques

Various methods have been proposed to develop recommendation systems, two of which form the basis for the development of other approaches, namely content-based filtering and collaborative filtering [7]. These techniques are classified as content-based filtering, collaborative filtering, knowledge-based filtering, and hybrid techniques as shown in Figure 1.

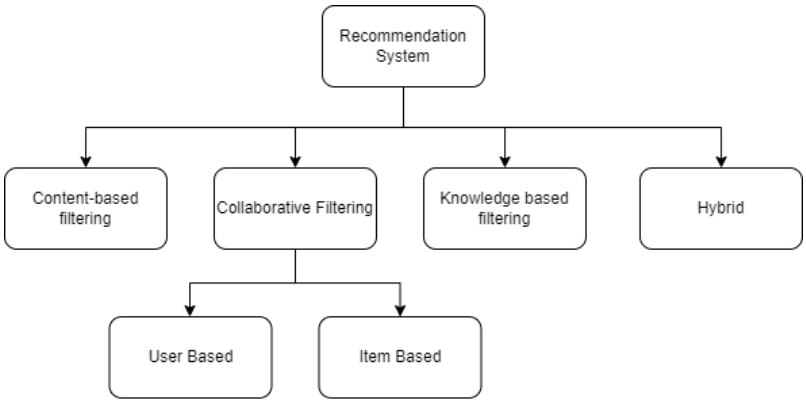


Figure 1. Recommendation System Techniques

This research uses a collaborative filtering recommendation system technique, the most popular and frequently used recommendation technique. In the combined filtering technique, users with similar interests prefer to choose the new items. This technique works on two points. First, this technique serves as a criterion for selecting a group of similar people whose opinions will be collected as a basis for recommendations (nearest neighbors). Second, this technique also calculates the occurrence pattern to form a larger itemset and significantly impacts the recommendations [7].

2.2. Association Rule

Association rules have the main computations in the dataset, which include**:**

A. Support

Support is the percentage of items, either one or a combination of several items, which means the number of times in the dataset. The support value measures how often the item set occurs in an association [8].

$$P(A \Rightarrow B) = \frac{\sum A \text{ and } B \text{ Transaction}}{\sum \text{Transaction}} \quad (1)$$

B. Confidence

Confidence is the accuracy of the probability of one item to another item occurring. The confidence value is in the range of zero to one, which means the higher the confidence value, the more likely the two items occur together [8].

$$P(A \Rightarrow B) = \frac{\sum A \text{ and } B \text{ Transaction}}{\sum A \text{ Transaction}} \quad (2)$$

C. Lift

Lift is the ratio of the observed support to that which would be expected if X and Y were independent. Lift shows the value of the rule based on random events X and Y [9].

$$P(A \Rightarrow B) = \frac{A \text{ and } B \text{ Support}}{(\text{Support } A) * (\text{Support } B)} \quad (3)$$

2.3. Algorithms for Frequent Itemset

In the association rule, it is known to use the frequent itemset technique, which is a technique to find out which items often appear as one or a collection of items with a minimum display of support in a data set [4].

A. Apriori

The apriori algorithm is the most commonly used algorithm to find the frequency pattern of items. The apriori principle is that if an item appears frequently, all subsets formed with that item must also appear frequently in a data set [10]. The processes that occur in the algorithm are:

1) First, the algorithm will create a path over the data set to determine the support for each item then find each frequent itemset set.
2) The algorithm generates a new item set and discards unnecessary candidates from the previous iteration.
3) Then, the algorithm will create an additional path to calculate the candidate's support.
4) After that, all candidate item sets that do not meet the minimum support will be eliminated. The algorithm will stop when there is no new frequent itemset set.

B. FP-Growth

FP-Growth or Frequent Pattern Growth is an alternative algorithm that can be used to determine frequent itemset in the dataset, FP-Growth uses the concept of building a tree in finding regular item sets [4]. Using a tree for finding items causes the FP-Growth algorithm to be faster than the apriori algorithm. The FP-Growth algorithm, unlike apriori which scans and creates a candidate set to all data, FP-Growth requires two scans on the dataset to retrieve the frequent itemset in the dataset. The FP-Growth process is as follows:

1) The algorithm scans the entire dataset once to calculate each support item, ignore sparse items, and sort the resulting set items in descending order.
2) The algorithm builds a tree by creating a second path over the data set. The first transaction is created as a node between those items where a path is created to encode this relation, and each node along the path has a frequency count of 1.
3) This process continues until all generated transactions are mapped to their relative paths that have been built in the tree.

2.4. Research Flow

In this research there are 4 work stages, such as:

a) Problem identification.
b) Data acquisition and pre-processing.
c) Generate association rule.
d) Association rule implementation results.

In Figure 2, the flow of this research can be seen as follows:

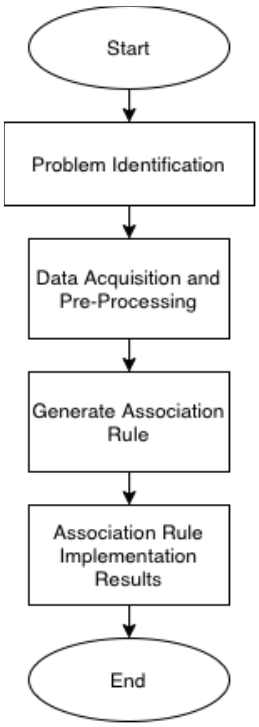


Figure 2. Research Flow

3. Results and Discussion

3.1. Problem Analysis

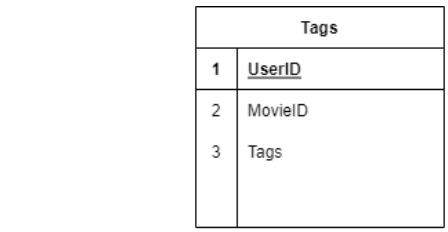
At this stage, identifying problems in the movie case study is carried out. The problem identified is that many movie productions continue to increase every year, making it difficult for users to determine which movies are suitable. To solve this problem, the researcher designs a recommendation system that can assist users in choosing movies based on the movie they have watched.

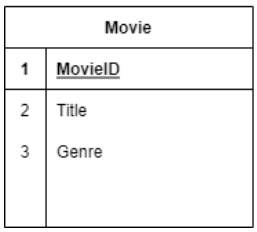
3.2. Data Acquisition and Pre-Processing

The data acquisition stage is the data preparation process, from data retrieval to data transformation and storing data into a database.
The steps taken are as follows:

A. Searching or determining dataset

The data to be used is in the form of a movielens dataset which can be retrieved at *https://grouplens.org/datasets/movielens*. Movielens data has three types of datasets, namely ratings.csv, tags.csv, and movies.csv. The movielens data used is in the form of small data released in 2018. This data contains 100,000 ratings and 3600 tag applications for 9,000 movies by 600 users. The structure of the Movilens dataset can be seen in Figure 3.





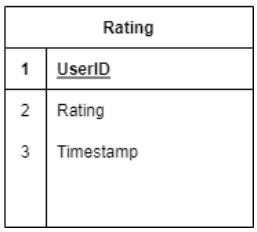


Figure 3. Dataset Structure

B. Perform data cleaning

After the data is obtained, the analysis process is carried out in advance, where information that is not needed will not be stored in the database. Several steps are taken at this stage, including deleting the tag.csv file and the timestamp column in the rating data because they are irrelevant to the data used. Then in this recommendation system, considering the rating value given by the user where the rating is equal to or less than 2, it is assumed that the user does not like the movie, so that the rating data similar to or smaller than 2 is deleted.

C. Save data into a database

After the data cleaning process, the user.csv and ratings.csv data are created and stored through scripting by reading each line in the CSV file and inserting it into a table in the database.

3.3. Generate Association Rule

The formation of association rules is when the rating data is processed in the database. The association rules are formed between each movie the user watches using the algorithms for frequent item (apriori and FP-growth). The data in the rating table is joined with the data in the movies table and then combined based on the user id. In the movie title column, strings are connected using the delimiter '|' and sorted by the movies frequency of appearance. The data will be entered into the rule table with the same support and confidence configuration. The Generate association rule process is as follows:

a) Prepare configuration related on parameters and implemented formulas specifying the values for MIN_SUPPORT = 2, MIN_CONFIDENCE = 0.75 and MAX_COL = 5.
b) Determine the relationship of the movie data using the collaborative filtering technique with the user id as a reference for grouping the itemset in the combined movies and rating table data.
c) Create association rules using the Apriori algorithm using the apriori library version 1.1.2 python.
d) Create association rules using the FP-Growth algorithm using the python version 1.0 pyfpgrowth library with customization of the generate_association_rules() function that has modified the data form on the return of the function values.

3.4. Association Rule Implementation Results

The association rules formed using the apriori algorithm and the FP-growth algorithm can recommend a movie. The association rule recommendation system will use collaborative filtering techniques. Collaborative filtering has the concept that the user will be advised of an item that is the same or related to his item. Each algorithm has the same configuration in minimum support and minimum confidence. The analysis will be carried out on the association results using a support value of 20% and a confidence value of 75%. The results of the algorithm produce different association rules. In the FP-growth algorithm, each transaction data will be processed to create a frequency tree of an item. From this tree, the confidence of the formed association rules will be calculated to find the best and follow the specified confidence value.

Data Output | Explain | Messages | Notifications

	rule text	support numeric	confidence numeric	lift numeric
1	Rule: Party Girl (1995) -> Clerks (1994)	0.003	1.0	38.125
2	Rule: Jaws (1975) -> Alien (1979)	0.003	1.0	203.333
3	Rule: Beauty and the Beast (1991) -> Aladdin (1992)	0.005	1.0	122.0
4	Rule: Addams Family Values (1993) -> Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	0.003	1.0	6.489
5	Rule: Before Sunrise (1995) -> Pulp Fiction (1994)	0.005	1.0	9.242
6	Rule: Clear and Present Danger (1994) -> Braveheart (1995)	0.003	1.0	6.63
7	Rule: Reservoir Dogs (1992) -> Braveheart (1995)	0.003	1.0	6.63
8	Rule: My Fair Lady (1964) -> Breakfast at Tiffanys (1961)	0.003	1.0	203.333
9	Rule: Bridges of Madison County, The (1995) -> GoldenEye (1995)	0.003	1.0	5.398
10	Rule: Congo (1995) -> Seven (a.k.a. Se7en) (1995)	0.005	1.0	6.1

Figure 4. Best Association Rule from Apriori Algorithm

As seen in Figure 4 above, the processing time on the apriori algorithm is longer than FP-growth, and the rules formed are more than the rules formed on FP-growth. The difference in results is due to the apriori algorithm process creating and testing all candidates, such as the breadth-first method, while in FP-growth using a tree formation method such as the depth-first method. The FP-growth algorithm avoids making candidates explicitly. In FP-growth it sorts the items in descending order where the more often these items occur in the database, the more likely it is to be generated.

Data Output | Explain | Messages | Notifications

	rule text	support numeric	confidence numeric	lift numeric
1	Rule: Fight Club (1999) -> Star Wars: Episode V - The Empire Strikes Back (1980)	0.003	1.0	67.778
2	Rule: Fight Club (1999) -> Matrix, The (1999)	0.003	1.0	67.778
3	Rule: Fight Club (1999) -> Matrix, The (1999)|Star Wars: Episode V - The Empire Strikes Ba...	0.003	1.0	67.778
4	Rule: Fight Club (1999)|Matrix, The (1999) -> Star Wars: Episode V - The Empire Strikes Ba...	0.003	1.0	67.778
5	Rule: Fight Club (1999)|Star Wars: Episode V - The Empire Strikes Back (1980) -> Matrix, T...	0.003	1.0	67.778
6	Rule: Reservoir Dogs (1992) -> Braveheart (1995)	0.003	1.0	6.63
7	Rule: Bridges of Madison County, The (1995) -> GoldenEye (1995)	0.003	1.0	5.398
8	Rule: Three Colors: Red (Trois couleurs: Rouge) (1994) -> Three Colors: Blue (Trois couleu...	0.003	1.0	101.667
9	Rule: Party Girl (1995) -> Clerks (1994)	0.003	1.0	38.125
10	Rule: Flirting With Disaster (1996) -> Get Shorty (1995)	0.003	1.0	11.296

Figure 5. Best Association Rule from FP-Growth Algorithm

The principle on apriori is to calculate with a candidate test approach and candidate generation. The bottom line is that if an item set occurs infrequently, it's impossible to generate a superset. The first algorithm will scan transaction data to get 1-item patterns that often appear, then generate candidate itemset with the next length or combination of items and test against existing transaction data. The algorithm will stop when no frequent candidate or item set is generated.

This creates a large number of candidates in contrast to FP-growth using a frequent growth pattern or mining without candidate generation. FP-growth will avoid candidate generation explicitly. First, the algorithm will get a 1-item pattern, sort the frequent items in descending order, and sort the transaction data according to the frequent items. Then scan the new transaction data again and create an FP-tree. From the FP-tree, the conditional pattern base will be searched for each item followed by the FP-tree conditional generation and pattern generated. Figure 4 and Figure 5, which contain the 10 best association rules formed by the apriori algorithm and the FP-growth algorithm, have the same association rules. With figures 4 and 5, the rules formed can be concluded by:

1) If a user watches Lord of the Rings: The Two Towers, the (2002), then there is an 89% chance of that user watching Lord of the Rings: The Fellowship of the Ring, The (2001).
2) If a user watches Lord of the Rings: The Two Towers, the (2002), then there is an 89% chance of that user watching Lord of the Rings: The Fellowship of the Ring, The (2001).
3) If a user watches Star Wars: Episode V - The Empire Strikes Back (1980), then there is an 88% chance of that user watching Star Wars: Episode IV - A New Hope (1977).
4) If a user watches Star Wars: Episode VI - Return of the Jedi (1983), then there is an 87% chance of that user watching Star Wars: Episode IV - A New Hope (1977).
5) If a user watches Lord of the Rings: The Fellowship of the Ring, the (2001), then there is an 85% chance that the user watched Lord of the Rings: The Two Towers, The (2002)
6) If a user watches Lord of the Rings: The Fellowship of the Ring, the (2001), then there is an 85% chance of that user watching Lord of the Rings: The Return of the King, The (2003).
7) If a user watch Seven (a.k.a. Se7en) (1995), then there is an 83% chance that the user watched Pulp Fiction (1994).
8) If a user watches Star Wars: Episode VI - Return of the Jedi (1983), then there is an 83% chance of that user watching Star Wars: Episode V - The Empire Strikes Back (1980).
9) If a user watches Jurassic Park (1993), then there is an 82% chance that the user watched Forrest Gump (1994).

4. Conclusion

Based on the research that has been carried out to compare the efficiency of the apriori algorithm and the FP-growth algorithm in the case study movie, it can be concluded that the association rules of each movie were successfully formed by the apriori algorithm and the FP-Growth algorithm, the association rules generated by the apriori and FP-growth algorithms are the same, and the comparison of association rule formation time is faster using the FP-growth algorithm compared to the a priori algorithm. Suggestions for the development of related research in the future is they can use the data processed in this study with the same or different algorithms to improve the quality of the recommendations.

References

[1] Anggraeni, Putri, Januarius Mujiyanto, and Ahmad Sofwan. "The implementation of transposition translation procedures in english-indonesian translation of epic movie subtitle." *ELT Forum: Journal of English Language Teaching*. Vol. 7. No. 2. 2018.
[2] Amrin, “Data Mining Dengan Algoritma Apriori untuk Penentuan Aturan Asosiasi Pola Pembelian Pupuk,” Paradigma, vol. XIX, no. 1, pp. 74–79, 2017, doi: https://doi.org/10.31294/p.v19i1.1836.

[3] W. Aprianti, K. A. Hafizd, and M. R. Rizani, "Implementasi Association Rules dengan Algoritma Apriori pada Dataset Kemiskinan," Limits J. Math. Its Appl., vol. 14, no. 2, p. 57, 2017, doi: 10.12962/limits.v14i2.2933.
[4] I. M. D. P. Asana, I. K. A. G. Wiguna, K. J. Atmaja, and I. P. A. Sanjaya, "FP-Growth Implementation in Frequent Itemset Mining for Consumer Shopping Pattern Analysis Application," Mobile-Based Natl. Univ. Online Libr. Appl. Des., vol. 4, no. 3, pp. 1–7, 2021, [Online]. Available: http://iocscience.org/ejournal/index.php/mantik/article/view/882/595.
[5] Isinkaye, Folasade Olubusola, Yetunde O. Folajimi, and Bolande Adefowoke Ojokoh. "Recommendation systems: Principles, methods and evaluation." *Egyptian informatics journal* 16.3 (2015): 261-273
[6] F. Ricci, L. Rokach, and B. Shapira, Recommender Systems Handbook. 2011.
[7] S. K. Raghuwanshi and R. K. Pateriyai, Recommendation Systems: Techniques, Challenges, Application, and Evaluation, vol. 2, no. January. Springer Singapore, 2019.
[8] S. Panjaitan et al., "Implementation of Apriori Algorithm for Analysis of Consumer Purchase Patterns," J. Phys. Conf. Ser., vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012057.
[9] A. Masnur, "Analisa Data Mining Menggunakan Market Basket Analysis untuk Mengetahui Pola Beli Konsumen," SATIN-Sains dan Teknol. Inf., vol. 1, no. 2, pp. 32–40, 2015.
[10] A. W. Oktavia Gama, I. K. Gede Darma Putra, and I. P. Agung Bayupati, "Implementasi Algoritma Apriori Untuk Menemukan Frequent Itemset Dalam Keranjang Belanja," Maj. Ilm. Teknol. Elektro, vol. 15, no. 2, pp. 21– 26, 2016, doi: 10.24843/mite.1502.04.