Implementation of Apriori and FP-Growth Algorithms on Movie Dataset

Muhammad Shafeen 22P-9278

September 6, 2024

Abstract

This document outlines the application of the Apriori and FP-Growth algorithms to analyze a dataset of movies. These algorithms are used to find frequent itemsets and association rules among movie genres or user ratings.

1 Dataset Description

The dataset consists of movie titles, genres, ratings, and user information. It is sourced from a popular movie database (e.g., MovieLens, IMDb). Each record in the dataset represents a user's rating for a particular movie and includes metadata about the movie.

2 Methodology

2.1 Preprocessing

Data preprocessing steps include:

- Cleaning: Handling missing values and removing duplicates.
- Transformation: Encoding genres and user ratings for algorithm compatibility.

2.2 Apriori Algorithm

The Apriori algorithm was implemented to identify the most common combinations of genres that co-occur in the dataset. The steps include:

1. Setting up the minimum support and confidence thresholds.
2. Finding all frequent itemsets.
3. Deriving strong association rules from these itemsets.

2.3 FP-Growth Algorithm

The FP-Growth algorithm was used to perform a similar analysis but is optimized to handle larger datasets more efficiently. The implementation steps are:

1. Building the FP-tree from transactions.
2. Extracting frequent itemsets directly from the FP-tree.

3 Results

The results section would detail the frequent itemsets and association rules discovered by both algorithms, presented in tables and charts. Comparative analysis of performance metrics like runtime and memory usage between Apriori and FP-Growth would also be included.

4 Conclusion

Conclusions drawn from the analysis, implications for movie distributors, and suggestions for further research would be presented here.