DATA MINING: Jester - Movie Ratings Data Sets (Collaborative Filtering Dataset)

Omphemetse Senna - u23884224

October 10, 2024

1 Research Questions

1.1 Problem Statement:

Our current recommendation systems now depend on collaborative filtering and content-based filtering, which can sometimes overlook the deeper connections between various movies. Consequently, viewers might miss out on recommendations that resonate with their movie genres. This research will bridge that gap by using Association Rule Learning (ARL) by identifying the viewers hidden patterns and using the hidden patterns in user behaviour, aiming to provide more relevant and enjoyable viewing suggestions

1.2 Objective:

- Implement the Association Rule Learning algorithms like Apriori and FP-Growth to analyse users's viewing data.
- Discover sets of patterns and association rules that highlight connections between different movies.
- Evaluates how effective these patterns are in improving recommendation systems to provide a more tailored viewing experience for viewers.

1.3 Why is it interesting:

This topic is very interesting to me because the large number of movies available can be a bit overwhelming for viewers and this often leads to decision fatigue or even disappointment in our own movie choices. A solid

movie recommendation system can really make a difference by using data-driven insights to help with suggestions based on what other viewers like and how they watch movies. By digging deep into user behaviour with methods like association rule learning, we can find patterns that we might help us to create more tailored and satisfying recommendations that resonate with viewers.

2 Data:

2.1 Data to be used:

This study makes use of the Jester Dataset from the Jester Online Joke Recommender System, which includes anonymous ratings obtained from viewers of the platform.

2.2 How Big Is the Data?

The dataset is approximately 3.9MB in size and is divided over three data files compressed in a .zip format. When unzipped, the files are available in Excel (.xls) format.

2.3 Attributes:

The dataset includes ratings from 73,421 users which are grouped in a matrix format with each row representing a single user. The first column represents the number of jokes rated by that user, whilst the next 100 columns represent the ratings for jokes numbered 01-100. Ratings are actual numbers ranging from -10.00 to +10.00, with ”99” denoting a ”null” rating for unrated jokes. A special focus will be on a thick sub-matrix including ratings for jokes 5, 7, 8, 13, 15, 16, 17, 18, 19, and 20, as virtually all users have rated these jokes, allowing for full examination of universal jokes preferences.

3 Approach:

3.1 What methods, algorithms, techniques will you be using?

Data Preprocessing:
To maintain accuracy and consistency, data from user viewing history and

ratings will be cleaned and normalized. This will include handling missing values and filtering out very rarely viewed movies.

2. Association Rule Learning (ARL):
Algorithms: The most important algorithms will be the Apriori Algorithm and the FP-Growth Algorithm. The Apriori Algorithm is very useful when creating frequent item groups and association rules by discovering the relationships between movies based on user behaviors.

The FP-Growth Algorithm will be used as a comparison tool, because it adeptly handles big datasets without the requirement to build possible item sets, allowing for a faster detection of common patterns.

Evaluation Metrics for Association Rules:
To evaluate the effectiveness and significance of the relationships we have established with our algorithms, we will use measures like support, confidence, and lift as our evaluative benchmarks.

3.2 Expectation:

From the application of these methods, algorithms, and techniques, we expect to achieve the following outcomes:

- Enhanced the quality of recommendations by making movie choices more relevant to people's watching patterns, which should lead to more engagement.
- Get the significant insights by uncovering previously unknown relationships between films, which might help influence targeted marketing campaigns or content creation projects.

4 Evaluations

4.1 How will you measure success?

Success will be measured from the following: We will calculate the User Engagement Rate and then compare the average number of movies viewed per user before and after the implementation of association rule learning

(ARL)-based recommendations. We will examine Retention Rates to evaluate user retention over the period of time following the implementation of the new recommendation system.

4.2 Baseline:

To create baselines for our study, we will collect the initial data from the existing recommendation system, such as average engagement rates and user satisfaction scores, before applying the ARL methodology.

5 Expected Output:

The expected outputs of this research include:

- A detailed Analysis Report outlining the results of the ARL application.
- Accurated selection of highly rated movies selected as the best suggestions for viewers.