Generic Recommender System

Mateusz Bernart
156072

Patryk Janiak
156053

Marek Seget
156042

November 3, 2025

1 Introduction

Recommender systems are critical components for personalizing user experiences. This report details the foundational phase for building a generic recommender system, which presents a valuable but significant challenge.

The primary goal is to understand the source data and transform it into a flexible, graph-based representation. We will model the data as a heterogeneous graph connecting user, item, and feature nodes. This graph structure effectively captures the complex relationships between these entities and serves as a robust foundation for the development of generic recommendation algorithms.

https://github.com/markopolo139/Generic-Recommender-Semantics-Put/

2 Datasets

To ensure the resulting recommendation framework is truly generic, two distinct and popular public datasets have been selected for this initial analysis:

- MovieLens: Sourced from the GroupLens research lab, this is a benchmark dataset in the recommendation domain. It consists of user-provided ratings and tags for a wide collection of movies, providing clear user-item interactions (ratings) and item features (genres, tags).

  https://grouplens.org/datasets/movielens/

- Steam Reviews: This dataset, part of the UCSD Amazon & Bytedance/TikTok dataset collection, provides a different domain: video games. It contains user reviews, game metadata, and user-owned game information, offering a rich source of implicit and explicit user feedback.

  https://cseweb.ucsd.edu/~jmcauley/datasets.html

By processing and modeling these two different datasets, we can validate that our graph-based representation is flexible enough to handle varied domains.

3 Exploratory analysis

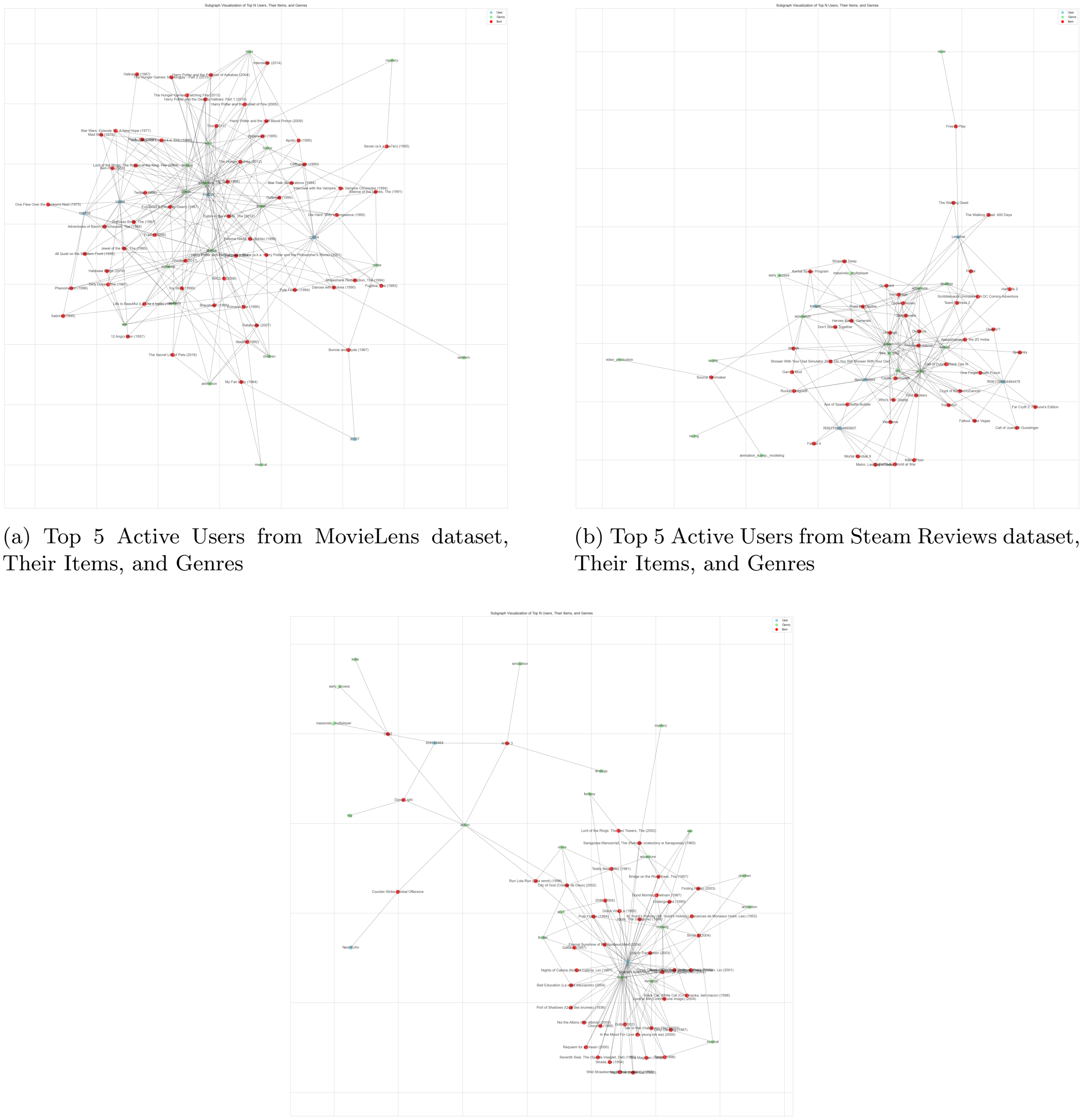
3.1 MovieLens

The MovieLens dataset contains 62,423 unique movies and 162,342 users. These users have generated over 25 million reviews, creating a rich source of interaction data. Movie-level features are robust, with 1,639 distinct genres identified; 'Drama' is the most frequently occurring. User ratings are provided on a scale from 0.5 to 5.0. It should be noted that the rating distribution is skewed towards positive sentiment, as the 25th percentile for scores is 3.0.

3.2 Steam Reviews

The Steam dataset provides a view into the gaming domain, encompassing 30,054 unique games and 22,077 users. The dataset features 883 genres, with 'Action' being the most prevalent. A total of 45,261 game ownership records are present, indicating that users in this dataset own approximately two games on average. This dataset offers a strong basis for analyzing both explicit review-based interactions and implicit ownership-based interactions.

3.3 Visualizations


Subgraph Visualization of Top N Users, Their Items, and Genres
User
Genre
Item
Subgraph Visualization of Top N Users, Their Items, and Genres
User
Genre
Item
Subgraph Visualization of Top N Users, Their Items, and Genres
User
Genre
Item

(a) Top 5 Active Users from MovieLens dataset, Their Items, and Genres

(b) Top 5 Active Users from Steam Reviews dataset, Their Items, and Genres

(c) Naive Merge of the Datasets, displaying Top 5 Active Users from each Dataset, Their Items, and Genres