Designing a Retrieval-Augmented Movie Recommendation System

Matteo Capalbo, Davide Di Stefano, Alessandro Pesare

1 5.1 Project Plan

1.1 5.1.1 Users

The proposed system targets a broad user base composed primarily of movie enthusiasts and streaming platform users. Within this group, the system is particularly designed for individuals who actively seek personalized and context-aware movie recommendations that go beyond conventional rating-based suggestions and recommendation system's techniques. The primary goals of these users are to discover **new content** that aligns with their **personal tastes**, **save time** in content selection, and **explore connections** between movies that may not be immediately evident through traditional filtering mechanisms. Typical workflows include consulting recommendation lists, reading plot synopses, exploring related films, and comparing multiple options before making a choice.

From a professional standpoint, such a system could also support **media analysts or recommender system researchers** interested in exploring how retrieval-augmented approaches can improve recommendation quality. Given the ubiquity of digital media consumption, this user group is both relevant and representative of modern content discovery habits, making the system potentially impactful in real-world settings.

1.2 5.1.2 Data

The data required for this project will be drawn from the **MovieLens 25M dataset**, a Kaggle's dataset containing approximately 25 million ratings provided by 162,000 users on 62,000 movies. This structured dataset will be augmented with additional unstructured textual information scraped from **IMDb**, including movie plots, genres, directors, and cast members.

Each movie will thus be represented as a composite textual document containing:

- Title and release year
- Director and main cast members
- Genres and keywords

- Full plot description

The collected information exhibits various degrees of organization along the following dimensions:

- **Granularity:** Each movie is represented as an atomic unit of information, making the dataset highly granular.
- **Connections:** Relationships among movies (e.g., shared actors, genres, directors) are implicit and will need to be inferred via embeddings or similarity measures.
- **Completeness:** The textual data may vary in detail and consistency; IMDb descriptions differ in length and richness, leading to partially fragmented representations.
- **Context:** Metadata such as release date and user ratings provide temporal and evaluative context.
- **Heterogeneity:** The dataset includes both structured (ratings, IDs) and unstructured (textual descriptions) data across multiple sources.

The chosen dataset size strikes a balance between diversity and manageability, enabling both quantitative evaluation and qualitative testing of the system's recommendations without excessive data preparation effort. Approximately one week of work is anticipated for data collection, cleaning, and preprocessing.

1.3 5.1.3 The Problem

The core challenge addressed by this project is the limitation of traditional recommender systems in producing **contextually rich and semantically grounded suggestions**. Classical approaches often rely on surface-level similarities or co-occurrence patterns, which may fail to capture deeper narrative or stylistic relationships between movies.

Users frequently experience problems such as:

- **Retrieval:** "I know I liked a movie with a similar theme, but I cannot recall it."
- **Connection:** "I didn't realize that these movies share similar plots."
- **Synthesis:** "I have rated many movies, but I can't see the bigger picture of my preferences."

Thus, the proposed system aims to **bridge the semantic gap** between user preferences and movie features by leveraging a retrieval-augmented generative model. The goal is not merely to recommend movies with similar metadata, but to enable the system to process rich textual descriptions, providing recommendations that are actually relevant.

1.4 5.1.4 The Solution

Concept

The proposed solution consists of a **Retrieval-Augmented Movie Recommendation System** that integrates traditional content-based filtering with the generative reasoning abilities of a Large Language Model (LLM).

At a conceptual level, the system performs the following tasks:

- **Data Ingestion:** Parse and embed textual descriptions of movies into vector representations using a pre-trained sentence transformer.
- **Data Retrieval:** Perform similarity searches in a vector database to identify movies semantically related to a user's query or previously liked titles.
- **Connect:** Combine retrieved data to expose implicit relationships between movies.
- **Generate:** Use an LLM (e.g., LLaMA-3 8B) to generate personalized recommendations and textual explanations that summarize why each suggested movie is relevant.

This approach aims to enhance the effectiveness of recommendations.

User Interface

The initial goal is to enable command-line interaction with the system, and potentially extend it later, as this approach offers the simplest way to achieve a fully functioning end-to-end prototype. The system might be extended as a web application, which provides a natural and accessible interface for most potential users. Users will be able to:

- Input movie titles or descriptions they enjoyed.
- Receive a ranked list of recommended movies.
- View concise textual explanations generated by the LLM.

The choice of a web interface aligns with the users' typical workflow—browsing, reading, and interacting with media databases online and facilitates rapid testing and deployment.

Technical Approach

From a technical perspective, the project presents three main challenges:

1. **Efficient Information Retrieval:** Determining optimal embedding and chunking strategies for textual movie data, ensuring high-quality vector representations that preserve semantic meaning.

2. **Prompt Engineering and Context Integration:** Designing effective prompts that incorporate retrieved information without overloading the LLM context window.

3. **Evaluation of Generative Recommendations:** Establishing appropriate quantitative metrics to assess the alignment between generated recommendations and user interests.

The system architecture will combine the following components:

- A **vector store** (e.g., Milvus or FAISS) for similarity search;
- A **sentence embedding model** (e.g., all-MiniLM-L12-v2);
- A **Large Language Model** (e.g., LLaMA-3 8B) for text generation and reasoning;
- An orchestration framework (e.g., LangChain) to manage the RAG pipeline.

1.5 5.1.7 Evaluation

The project's evaluation strategy will focus on assessing both the **accuracy** and the **relevance** of the generated movie recommendations. "Success" will be defined as the system's ability to produce recommendations that align with user preferences more effectively than a baseline content-based model.

The following quantitative metrics will be employed:

- **Precision:** Fraction of recommended movies that are relevant to the user's past preferences.
- **Recall:** Fraction of relevant movies successfully retrieved by the system.
- **F-Measure:** Harmonic mean of precision and recall, offering a balanced performance indicator.

A controlled experimental setup will be implemented, simulating user profiles from the MovieLens dataset and comparing recommendations against known user ratings. We will then compare the RAG-based approach to a traditional collaborative filtering baseline. The evaluation will not focus on usability or user satisfaction, but rather on measurable improvements in the recommendation.

Ultimately, the project aims to demonstrate that retrieval-augmented methods can substantially improve both the **semantic depth** and the **trustworthiness** of recommendations in the movie domain, paving the way for more generalizable AI-driven recommendation systems.