Cinematic Data Retrieval: Pursuing a Better Search Mechanism

Fernando da Silva Pereira Borges Afonso
up202108686@up.pt
Faculty of Engineering - University of Porto
Porto, Portugal

João Francisco da Rocha Sequeira Alves
up202006281@up.pt
Faculty of Engineering - University of Porto
Porto, Portugal

João Pedro Sá Torres Neiva Passos
up202108833@up.pt
Faculty of Engineering - University of Porto
Porto, Portugal

Pedro Cancela da Silva
up202400230@up.pt
Faculty of Engineering - University of Porto
Porto, Portugal

Figure 1: Movies And Series

Abstract

The rapid growth of digital media has increased the need for effective information organization and retrieval systems. This project addresses this challenge in the context of the movie and television domain, integrating structured and unstructured data from multiple sources. This report describes the construction of a reproducible data processing pipeline, including data cleaning, integration, and quality assessment, forming the basis for the development of a search engine for movies and series.

G36, Porto, Portugal

ACM ISBN 978-1-4503-XXXX-X/2018/06

CCS Concepts

• **Information systems** → **Information retrieval**.

Keywords

Movies, Series, Cinema, Film, Reviews, Dataset, Search Engine, Pipeline, Data Retrieval, Data Preparation, Data Analysis, Data Processing, Data Refinement

ACM Reference Format:
Fernando da Silva Pereira Borges Afonso, João Francisco da Rocha Sequeira Alves, João Pedro Sá Torres Neiva Passos, and Pedro Cancela da Silva. 2025. Cinematic Data Retrieval: Pursuing a Better Search Mechanism. In *Proceedings of PRI (G36).* ACM, New York, NY, USA, 6 pages.

1 Introduction

This project's primary goal is to apply the principles and techniques of information processing to a real-world domain, exploring the challenges of collecting, organizing, and analyzing heterogeneous data for retrieval purposes.

The choice of movies and television series as the project's central theme was motivated by their cultural significance, the abundance of available data, and the diversity of attributes they encompass. This domain offers a rich combination of structured data, such as titles, genres, release years, and ratings, alongside unstructured textual information, including plot summaries and user-generated reviews. These characteristics make it a valuable case study for exploring the interaction between metadata and natural language content within information retrieval systems.

Despite the magnitude of available data in this area, existing information retrieval systems often remain limited, focusing primarily on keyword-based searches or rigid metadata filters. Such systems fail to explore the deeper semantic and contextual connections within textual content, leaving room for improvement in the way users discover and relate media content. This project aims to address this limitation by constructing a coherent, high-quality dataset that integrates structured and unstructured sources, establishing the foundation for more expressive retrieval approaches in later stages.

This document is structured in the following way: **Data Extraction and Sources** presents the data sources and extraction process. **Data Preparation and Quality Assessment** details data preparation procedures and quality assessment, outlining the reproducible pipeline employed. **Exploratory Data Analysis and Characterization** explores the datasets through statistical and textual characterization. **Conceptual Model** introduces the conceptual model of the domain, followed by **Document Definition**, which defines the notion of documents adopted for information retrieval. **Information Needs and Search Scenarios** discusses potential information needs and search scenarios, and **Conclusions and Future Work** concludes with reflections and directions for future work.

2 Data Extraction and Sources

When selecting the dataset to be used in this project, we chose **IMDb (Internet Movie Database)** [1] as our primary source. IMDb offers a vast collection of movies and series, providing numerous attributes for each title, such as release year, genre, cast, and ratings. In addition, it serves as one of the most widely recognized platforms for user-generated reviews and ratings, making it a reliable and comprehensive data source.

The information provided is available through various structured non-commercial datasets that can be freely accessed for research purposes, with the exception of movie and series descriptions. To address this limitation, web scraping was employed using data from the **StreamWithVPN** [2] website, which provided the missing textual descriptions. This complementary source ensured a complete dataset by filling the descriptive gap present in IMDb's publicly available data.

For this work, the following datasets were used: the IMDb core files (`title.basics`, `title.ratings`, `title.principals`, and `name.basics`), supplemented with the extracted textual descriptions from StreamWithVPN.

Table 1: Summary of Movie/Series datasets.

Dataset	Features	Entries	Size (MBs)
Title Basics IMDb	9	11962455	987.0
Title Ratings IMDb	7	1619879	26.9
Title Principals IMDb	6	95180101	3950.0
Name Basics IMDb	6	14742901	864.0

Was also considered it relevant to include movie and series reviews to further enrich our data collection. For this purpose, we selected two different datasets: one from **Kaggle**[3] for movie reviews and another from **OpenDataBay**[4] for series reviews. Reviews directly from IMDb were not used because they are not provided in the available IMDb datasets. The chosen datasets are presented below:

Table 2: Summary of the Reviews datasets.

Dataset	Features	Titles	Reviews	Size (MBs)
Movie Reviews	8	15951	664436	308.28
Series Reviews	13	3136	52895	38.5

3 Data Preparation and Quality Assessment

The dataset construction followed a rigorous and reproducible pipeline (Figure 3) inspired by standard information retrieval methodologies. The consolidated dataset was obtained by integrating IMDb core files—`title.basics`, `title.ratings`, `title.principals`, and `name.basics`—with enriched textual data gathered through targeted web scraping. Intermediate results were exported in TSV format to ensure traceability and reproducibility throughout the process.

The **initial cleaning phase** consisted of the systematic removal of incomplete, null, or uninformative records to ensure dataset integrity. Subsequently, categorical restrictions were applied, limiting the corpus to `feature films` and `television series` while excluding other title categories. Entries lacking essential rating information were discarded, and a threshold of 1,000 minimum votes was imposed to mitigate sparsity and popularity bias. To further refine representativeness, a Bayesian weighted-rating formula was implemented, using the global mean rating and the 75th percentile of vote counts as priors, resulting in a ranked subset of 10,000 titles for enrichment.

The **normalization phase** addressed structural and textual inconsistencies across datasets. Titles were standardized by removing extraneous metadata such as release years or additional descriptors. For instance,`Dekalog (1988)` was normalized to `Dekalog`, and `The Office (UK): Season 3` to `The Office`, ensuring uniformity across entries. This process reduced variability and improved cross-source matching accuracy. Similarly, non-standard delimiters and encoding inconsistencies were resolved to achieve a unified schema.

Given that the reviews dataset originated from a distinct source, a **matching procedure** was required to associate reviews with the correct movie or series entities. This alignment employed the

`fuzzywuzzy`[5] library, applying an 80% similarity threshold to reconcile discrepancies caused by non-printable characters and hidden formatting. Upon successful matching, the "title" column in the reviews dataset was replaced by the unique IMDb identifier (**tconst**), adopted as a foreign key to enable consistent cross-referencing between reviews and titles.

Further **enrichment procedures** involved joining principal-role records with corresponding person metadata to extract up to three unique actors or actresses per title along with their respective character names. All string attributes were cleaned to remove escape sequences and malformed arrays. Web scraping was conducted in batch mode with retry logic and polite delays to acquire textual descriptions of movies and series. Each operation was logged to maintain an auditable record of success and failure rates.

The **final normalization and validation** steps included deduplication based on unique identifiers, profiling of missing values, and standardization of data formats, such as enforcing canonical genre delimiters. The resulting dataset exhibited high structural coherence and consistency across all attributes.

Despite extensive preprocessing, **known issues** persisted. A minor subset of titles lacked descriptions due to failed retrieval attempts, and occasional parsing errors were observed in third-party metadata fields. Additionally, certain user-generated texts contained non-ASCII characters that were retained as NaN values for transparency. These anomalies were documented and preserved in compliance with best practices for data integrity and reproducibility.

4 Exploratory Data Analysis and Characterization

This section presents a comprehensive analysis of the processed dataset, examining multiple dimensions to understand the characteristics and patterns within movie and series reviews. The graphs and tables were obtained using the Matplotlib [6], Numpy [7] and NLTK [8] libraries.

4.1 Descriptive Analysis

From the IMDb dataset, we selected the 10,000 best-rated titles, comprising 3,216 series and 6,784 movies. Among these, we successfully retrieved plot descriptions for 7,976 titles through web scraping, achieving a 79.76% success rate. This process ensured that the majority of our selected entries contained comprehensive textual metadata suitable for further analysis.

Regarding the review datasets, the s Reviews dataset originally included 52,895 reviews across 3,136 unique titles. From these, we successfully matched 1,038 series titles to our selected IMDb subset, resulting in a total of 28,920 corresponding reviews, resulting in a correspondence rate of 54, 90%. Similarly, in the Movie Reviews dataset, which contained 664,436 reviews, we were able to link 5,532 movie titles to our selected IMDb movies, which represents 358,152 reviews with a correspondence rate of 53, 9%.

In general, the data scraping process to obtain plot descriptions was satisfactory, achieving nearly 80% success. The correspondence between the review datasets and the selected IMDb titles was also acceptable, with more than half of all reviews successfully matched to the chosen titles. These results indicate that the integrated dataset maintains a solid level of completeness, combining both structured information and review-based content suitable for subsequent analysis.

4.2 Review Distribution Analysis

Analysis of the most reviewed titles (Figure 4) reveals superhero and franchise films dominating user engagement. Zack Snyder's Justice League leads with approximately 1,800 reviews, followed by Wonder Woman and The Dark Knight Rises with over 1,600 reviews each. This pattern suggests heightened engagement with blockbuster releases and established franchises, which typically generate more polarized opinions and discussion.

The correlation analysis between ratings and review counts (Figure 5) yields a coefficient of 0.03, indicating virtually no linear relationship between a title's rating and its review volume. This finding suggests that review quantity is driven more by factors such as marketing reach, franchise popularity, and cultural impact rather than perceived quality.

4.2.1 Genre Distribution Analysis. The genre distribution analysis (Figure 6) demonstrates clear patterns in review engagement across different content categories. Science Fiction leads with an average of 173 reviews per title, followed by Adventure (139) and Action (133). In contrast, niche genres such as Talk-Show, Reality-TV, and Short Films average fewer than 35 reviews per title.

Interestingly, when examining average ratings by genre, an inverse pattern emerges. Short Films achieve the highest average rating (8.27), followed by Talk-Shows (8.26) and Game-Shows (8.15), while mainstream genres typically receive lower average scores. This suggests specialized content attracts more selective, appreciative audiences, while mass-market genres face broader, more critical viewership.

These analytical results provide valuable guidance for developing our system as understanding how engagement and ratings vary across genres enables more informed data structuring and feature prioritization.

4.3 Analysis of Word Clouds

To acquire a better understanding of the key terms and themes within our datasets we generated the following word clouds:

4.3.1 Description Word Cloud (Figure 7). The first word cloud, generated from the collection of movie /series descriptions in our dataset reflects the thematic and narrative structure of the movies and series themselves. Words like life, world, family, love, find, and live dominate, pointing to recurring motifs around human relationships, personal struggles, and existential experiences. The prominence of terms such as man, woman, son, father, and friend underscores the centrality of interpersonal dynamics in most story lines.

4.3.2 Reviews Word Cloud (Figure 8). The second word cloud highlights the most frequent terms used by users when expressing opinions about movies and series. Words such as movie, film, good, great, story, and characters dominate, suggesting that user discussions are primarily centered on the overall quality of storytelling, character development, and general enjoyment. Positive sentiment

terms like best, love, and amazing also appear frequently, indicating a generally engaged and expressive audience that emphasizes emotional and qualitative evaluations rather than technical aspects.

5 Conceptual Model

The conceptual model represents the logical organization of the data used in this project, serving as the foundation for subsequent information retrieval and analysis tasks. It establishes how different entities interact within the domain of movies and television series.

As illustrated in Figure 2, the dataset is composed of two main entities: **Movie/Series** and **Review**. The *Movie/Series* entity stores structured metadata collected from authoritative sources, including identifiers, titles, release years, genres, ratings, cast information, and textual descriptions. The *Review* entity encapsulates unstructured textual data written by users, associated with the corresponding movie through the attribute `review_tconst`.

This relationship is an optional one-to-many association, where each movie or series can be associated with zero or more reviews, and each review is related to exactly one movie or series. This structure allows flexible querying, supporting retrieval at both the document level (e.g., specific reviews) and the item level (e.g., aggregated movie information).

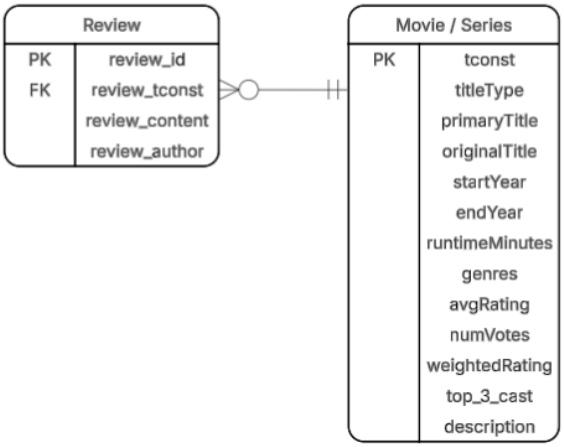

Figure 2: Conceptual Model

6 Document Definition

In our search result document, we include the following elements:

- Movie/Series ID: Unique identification string that represents the movie/series
- Title Type: Distinguishes if the entry is a movie or a series
- Primary Title: Title of the movie/series in English
- Original Title: Title of the movie/series in the original language (if it applies)
- Start Year: movie/series release year
- End Year: Series release year (movies don't apply)
- Runtime: Time the movie lasts (or average time of a series' episode)
- Genres: List of genres that can describe the movie/series
- Average Rating: Rating given by the people to the movie/series in IMDB
- Number of Votes: The amount of votes/ratings a movie/series has
- Top 3 Cast: The top 3 cast members in the movie/series and the respective characters they play
- Description: Text that describes the movie/series (most around 200/300 words)

We also have separate result documents for the reviews, since each movie/series has multiple reviews. These are the elements in them:

- Review ID: Unique identification string that represents the review
- Reviews tconst: Unique Identification from related Movie/Series
- Review Content: Text expressing author's opinion
- Review Author: Author of the review

7 Information Needs and Search Scenarios

To design an effective search engine for movies and series, it's important to understand users' information needs. Below are several key types of queries that reflect different aspects of the system's functionality.

- Similarity Search: "Find movies similar to Inception"
- Theme Based Discovery: "Retrieve movies about artificial intelligence"
- Era and Genre Filtering: "List comedies from the 90s"
- Runtime Finder: "Get movies with a runtimes below 90 minutes"
- Actor Based Search: "Series where Bryan Cranston is featured"
- Rating-Weighted Discovery: "Best reviewed content during pandemic era (2020-2022)"
- Reviewer Trust Metrics: "Films recommended by Roger Ebert"
- Length-Aware Review Filtering: "Movies with detailed analytical reviews (average review above 500 words)"
- Cross-Media Recommendations: "Films similar to Game of Thrones but standalone"

8 Conclusions and Future Work

This milestone concludes the data preparation phase, in which heterogeneous data from multiple sources were collected, cleaned, integrated, and analyzed to produce a coherent and reproducible dataset. The resulting collection combines structured metadata from IMDb with unstructured textual content from external reviews, forming a solid foundation for subsequent information retrieval experiments.

Future enhancements may include the integration of additional data sources, normalization of entity references, and the application of sentiment analysis to user reviews.

With a consistent and comprehensive dataset in place, the next stage will focus on developing a functional search system capable of connecting users with relevant titles according to their queries. In the long term, the aim of the system is to provide an intuitive and effective search experience that leverages both factual metadata and descriptive content to better reflect users' interests and intentions.

Annexes

Figure 3: Pipeline

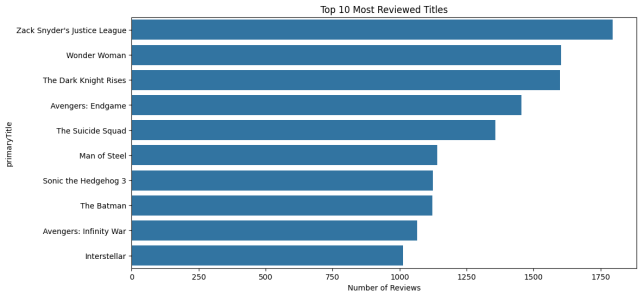

Figure 4: Most Reviewed Titles

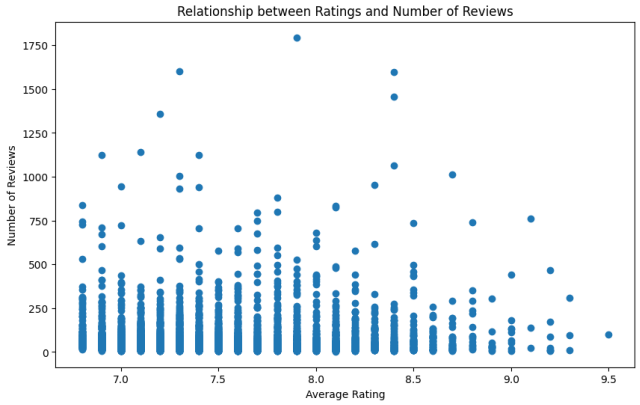

Figure 5: Distribution of Reviews per Title

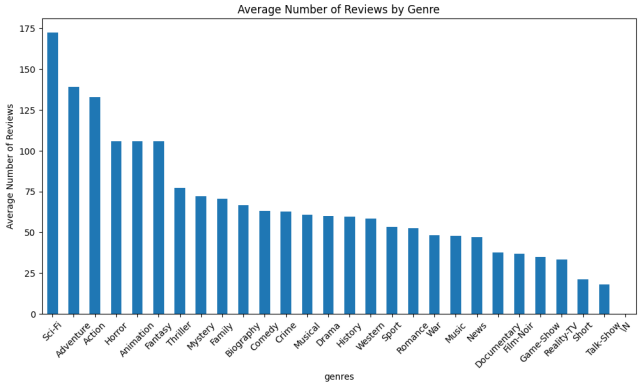

Figure 6: Genre Distribution of Reviews and Ratings

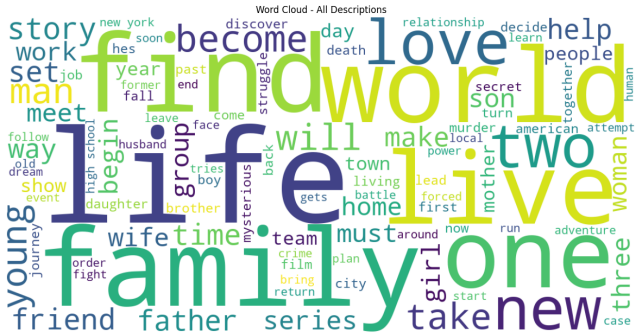

Figure 7: Description Word Cloud

Figure 8: Reviews Word Cloud

References

[1] Imdb non-commercial datasets. https://developer.imdb.com/non-commercial-datasets, 2024.
[2] Streamwithvpn website. https://www.streamwithvpn.com/, 2024.
[3] Kaggle. Movies reviews dataset. https://www.kaggle.com/datasets/davutb/metacritic-movies, 2022.
[4] Open Data Bay. Series reviews dataset. https://www.opendatabay.com/data/consumer/a772e407-653d-4bea-a137-bdf4400cef6f, 2023.
[5] Fuzzywuzzy library. https://github.com/seatgeek/fuzzywuzzy, 2023.
[6] Matplotlib library. https://matplotlib.org/, 2023.
[7] Numpy library. https://numpy.org/, 2023.
[8] Nltk library. https://www.nltk.org/, 2023.
[9] Pandas library. https://pandas.pydata.org/, 2023.