A Hybrid Collaborative and Content-Based Approach for Personalized Movie Recommendations

Chaimaa Abi

[a]*Mohamed bin Zayed University of Artificial Intelligence*

Abstract

Recommendation systems are crucial in the era of big data, significantly enhancing user experience by personalizing content suggestions across diverse digital platforms. This report details the development and simulation of a content-based recommendation system, alongside an exploratory study of collaborative filtering methods, all implemented within the framework of the expansive MovieLens 25M dataset. By focusing on item attributes and user interactions, the content-based system is meticulously crafted to enhance recommendation accuracy and effectively address challenges such as the cold start problem. Additionally, our study involves a comprehensive exploratory data analysis (EDA) and methodological simulations to assess the performance and integration potential of each approach in a hybrid recommendation model. This structured exploration aims to optimize the strengths of both methodologies to improve overall system efficacy.

1. Introduction

Recommendation systems have become an integral part of modern digital platforms, helping users navigate the vast sea of available content and products. With the ever-growing volume of data, these systems face the challenge of efficiently analyzing and processing large datasets, a core issue in big data technologies. As part of our coursework in the big data course, we have implemented two popular approaches to recommendation systems: collaborative filtering and content-based filtering.

Collaborative filtering is a technique that relies on analyzing user behavior and interactions with items, such as ratings or viewing history. The underlying assumption is that users who shared similar preferences in the past are likely to agree on future preferences as well. This method is particularly effective when there is abundant user interaction data, enabling highly personalized recommendations.

On the other hand, content-based filtering recommends items by analyzing their inherent characteristics, such as genres or descriptions, and matching them to the user's previously expressed preferences. This approach is advantageous when user data is sparse or when dealing with new users, as it relies solely on item attributes rather than user interaction patterns.

However, collaborative filtering is not without its limitations. One significant challenge is the cold start problem, which arises when new users or items lack sufficient historical data to generate reliable recommendations. This scenario can significantly hinder the system's ability to provide relevant suggestions, especially in the early stages of user or item adoption.

To address these limitations, our project aims to develop a hybrid recommendation system that combines the strengths of collaborative filtering and content-based filtering. By utilizing the rich interaction data from the extensive MovieLens 25M dataset, as well as leveraging the descriptive power of content-based filtering, we hope to create a system that not only provides personalized recommendations but also effectively handles new or obscure items and accommodates new users.

2. Dataset Description

In our project, we utilize the MovieLens 25M dataset provided by GroupLens, a research lab at the University of Minnesota. This dataset is designed for educational and research purposes and includes a wide array of data crucial for analyzing movie preferences and enhancing recommendation systems. It features interactions from over 160,000 unique users and data on more than 62,000 movies, amounting to nearly 25 million ratings. The dataset encompasses 19 distinct genres, providing a comprehensive framework to explore genre-based preferences and trends. Additionally, it contains 1086 unique tags, enriching the movie metadata and offering refined pathways for content-based filtering method

Rating Distribution Analysis

The distribution of ratings within our dataset as illustrated in Figure 1 shows a clear preference for higher ratings, with the most frequent ratings being 4.0, followed by 3.0 and 5.0. This pattern suggests that users are generally satisfied with the movies they choose to rate, tending to give positive feedback. Notably, ratings below 3.0 are less common, indicating either a tendency among users to watch and rate movies they believe they will like or that most movies generally meet the viewers' expectations.

The skew towards higher ratings is important for our recommendation system. It suggests a strategy where positive ratings could be weighted more heavily, recognizing their greater frequency and potential significance in reflecting user satisfaction.

Conversely, the rarity of lower ratings could highlight the importance of critically poor ratings as significant indicators of content that is less desirable to users.

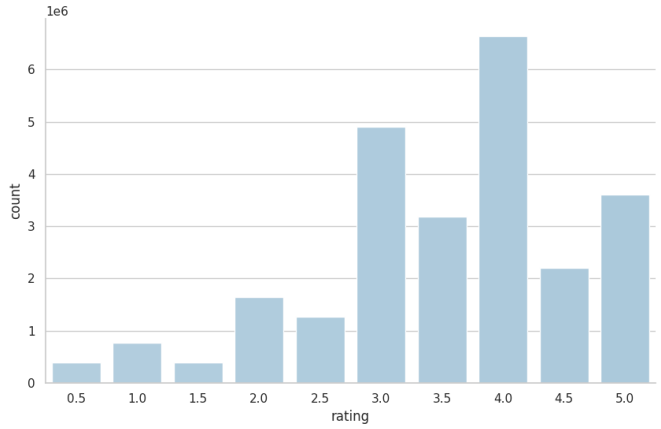

Figure 1: Distribution of Ratings.

Analysis of Average Rating Over Time

The Average Rating Over Time graph as shown in Figure 2 provides insights into the evolution of user ratings from 1995 to 2020. We observe considerable variability in ratings between 1995 and 2005, suggesting a period of adaptation to the rating system or fluctuations in movie quality. The ratings reach their lowest in 2005, potentially indicating shifts in the movie industry or user preferences during this time.

Subsequently, there is a noticeable recovery in average ratings, peaking around 2015, which could correlate with enhancements in movie selections or an increase in user engagement. Following this peak, the ratings show a trend towards stabilization, suggesting that users' expectations and the quality of movies have found a consistent balance in recent years.

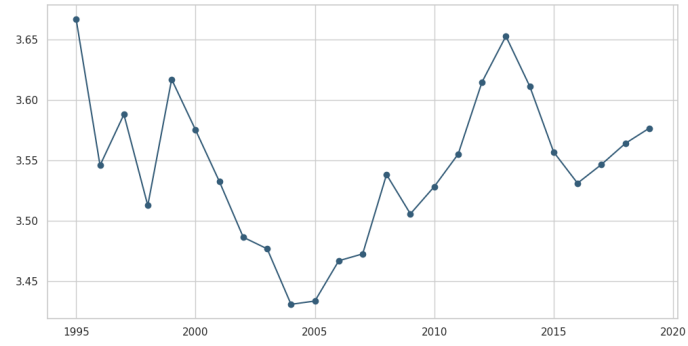

Figure 2: Average Rating Over Time.

Analysis of Highest Rated and Most Popular Movies

Figure 3 shows the highest-rated movies in our dataset, dominated by critically acclaimed titles and documentary series such as "Planet Earth". This reflects a strong preference among viewers for high-quality production and storytelling. Interestingly, the top-rated entries also include influential series and films, indicating a significant appreciation for narrative depth and factual presentation.

Figure 4 illustrates the most popular movies based on the number of ratings. Notable entries like "Forrest Gump" and "The Shawshank Redemption" dominate, suggesting that their cultural impact and accessibility contribute significantly to their popularity. This reveals that while high ratings are an indicator of quality, the frequency of ratings can also reflect a movie's enduring appeal and relevance.

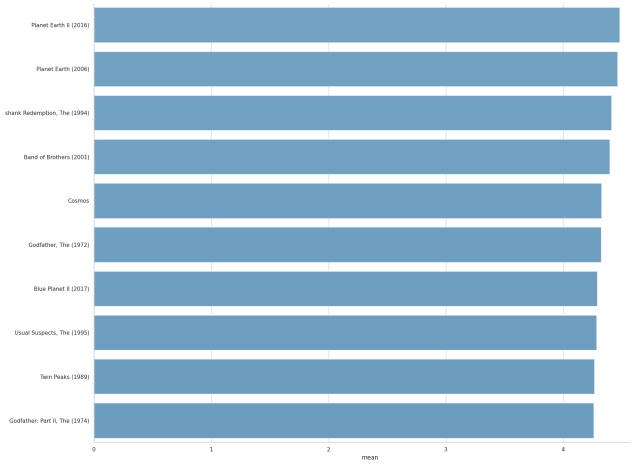

Figure 3: Highest Rated Movies.

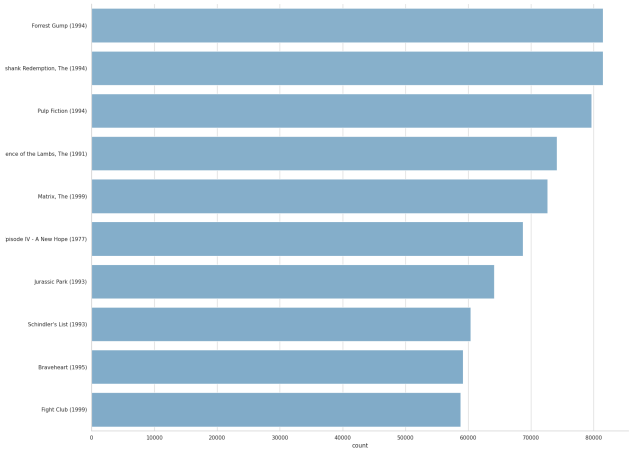

Figure 4: Most Popular Movies.

Relationship Between Average Rating and Number of Ratings

The scatter plot in Figure 5 presents the relationship between the average rating and the number of ratings for movies. There is a noticeable trend where movies with higher average ratings also tend to have more ratings, suggesting a positive correlation between a movie's perceived quality and its popularity. This pattern highlights that movies which are well-received generally attract more viewers and ratings, reinforcing their visibility in recommendation systems.

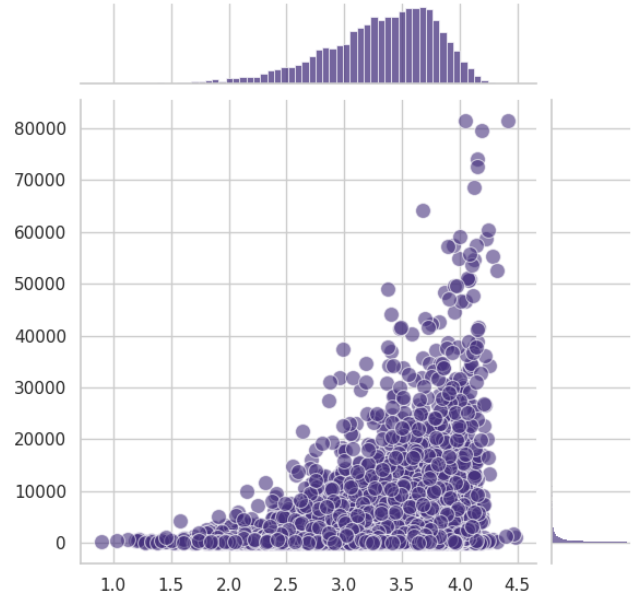

Figure 5: Relationship Between Average Rating and Number of Ratings.

Tag Relevance Distribution

The histogram of tag relevance scores in the MovieLens 25M dataset reveals a pronounced leftward skew, with the majority of tags assigned lower relevance scores, predominantly clustering between 0 and 0.2. This distribution suggests that while many tags are applied to movies, they are often only marginally relevant. The presence of a long tail extending towards higher relevance scores highlights the rarity of highly relevant tags. This pattern indicates that the tagging system, though generous in the application of tags, finds only a fraction of these to be strongly related to the movies.

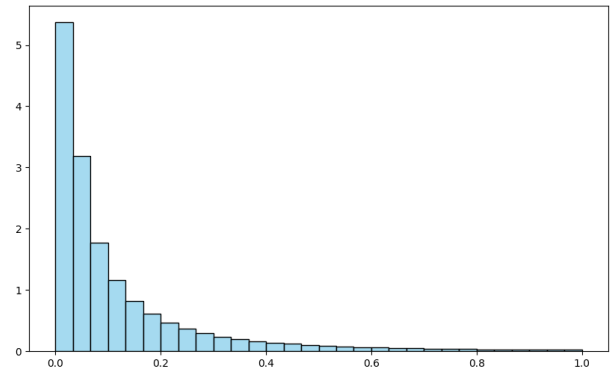

Figure 6: Histogram of Tag Relevance Scores

Tag Rank and Relevance Analysis

To further refine our understanding of tag relevance, we rank the relevance of tags within each movie (see appendix A.9). The most relevant tag receives a rank of 1, with subsequent tags receiving higher ranks. We visualize this data to analyze the distribution and median trend of relevance scores across the top 60 relevance ranks. Our visualizations include a box plot for distribution analysis and a line plot showing the median relevance score trend, which helps in identifying the optimal cutoff for tag relevance.

Optimization of Tag Relevance for Recommendations

Our analysis indicates a descending trend in the median relevance as tag rank increases, with a noticeable leveling off around the 15th rank. This finding suggests that the 15th rank may serve as an optimal cutoff point for considering tag relevance in recommendations. Tags ranked above 70 show a significant decrease in relevance, and their variability suggests they may introduce more noise than valuable information. By setting this cutoff, we enhance the quality of our recommendations, focusing on tags that are most informative and relevant, thereby maximizing computational efficiency and recommendation quality.

3. Methodology

3.1. Content-Based Filtering

Content-based filtering serves as a cornerstone methodology that relies extensively on the descriptive attributes of items—specifically, movie metadata such as genres and tags. This approach is particularly advantageous for addressing challenges such as the cold start problem with new movies, where traditional collaborative filtering might falter due to a lack of user interaction data.

The foundation of our content-based filtering process begins with a meticulous preparation of data, where movie tags are extracted and evaluated based on their relevance scores. To ensure both relevance and computational efficiency, we focus on the top 70 tags per movie. Using this refined data, we apply various similarity measures to assess the relationships between movies. These measures include:

- **Cosine Similarity:** This measure calculates the cosine of the angle between two vectors of tag relevance, providing insight into the orientation and magnitude of these vectors in the tag space. Mathematically, it is defined as:

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{||A||||B||} \tag{1}$$

  Here, A and B are vectors representing the tag relevance scores for two movies. A higher cosine similarity value (closer to 1) indicates that the two movies have similar tag patterns, suggesting they may be relevant to similar user interests.

- **Jaccard Similarity:** Useful for binary data, this measure evaluates the similarity between sets by comparing the size of the intersection to the union of the sets. The formula is:

$$\text{jaccard\_similarity}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

In this context, A and B represent the sets of tags associated with the two movies. A higher Jaccard similarity value (closer to 1) indicates that the two movies share a larger proportion of common tags, making them more similar in terms of content.

- **Overlap Coefficient:** This coefficient measures the overlap between two sets, normalized by the size of the smaller set, thus focusing on the proportion of overlap relative to the smaller dataset. It is calculated as:

$$\text{overlap\_coefficient}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{3}$$

Similar to Jaccard similarity, A and B represent the sets of tags for the two movies. The overlap coefficient focuses on the proportion of overlap relative to the smaller dataset, which can be useful when dealing with movies with significantly different numbers of associated tags.

Through the implementation of the above measures, we construct a tag matrix for all movies in the dataset. Each movie is represented as a row in the matrix with columns corresponding to tags, filled based on the relevance of the tag to that particular movie. This structured approach allows us to compute detailed similarity scores between all pairs of movies.
The user-item matrix was constructed with users as rows and items as columns, where each entry represents the rating a user gives to a specific movie. Unrated items are represented by NaN, which are addressed in our preprocessing for similarity calculations.

3.2. Collaborative Filtering

Collaborative Filtering (CF) is a technique utilized in recommendation systems to predict user preferences based on the aggregation of preference information from multiple users. The underlying hypothesis is that if a user A has a similar rating pattern to user B, user A is likely to have similar preferences to user B on unknown items.

In our approach, we implemented three different similarity measures to calculate the similarity between users based on their historical ratings:

1. **Pearson Correlation Coefficient (PCC):** The Pearson correlation between two users u and v is defined by the formula:

$$PCC(u, v) = \frac{\sum(r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum(r_{ui} - \bar{r}_u)^2}\sqrt{\sum(r_{vi} - \bar{r}_v)^2}}$$

where r_{ui} and r_{vi} are the ratings given by user u and v to item i, and $\bar{r}_u$ and $\bar{r}_v$ are the mean ratings of user u and v respectively.
2. **Cosine Similarity:** We used the cosine similarity measure as defined earlier in the content-based filtering section.
3. **Jaccard Similarity:** We used the Jaccard similarity measure as defined earlier in the content-based filtering section. In the context of collaborative filtering, A and B represent the sets of items rated by users u and v respectively.

The user-item matrix was constructed with users as rows and items as columns, where each entry represents the rating a user gives to a specific movie. Unrated items are represented by NaN, which are addressed in our preprocessing for similarity calculations.

Collaborative filtering predicts a user's rating for an item based on the ratings of other similar users, utilizing the wisdom of the crowd to recommend items. The predicted rating $S(u, i)$ for a user u on an item i can be expressed as:

$$S(u, i) = \bar{r}_u + \frac{\sum_{v \in V}(r_{vi} - \bar{r}_v) \cdot w_{uv}}{\sum_{v \in V} |w_{uv}|}$$

Where:

- $S(u, i)$: The predicted rating for user u on item i.
- $\bar{r}_u$: The average rating given by user u.
- v: Other users who have rated item i and whose tastes are similar to user u.
- r_{vi}: The rating user v has given to item i.
- $\bar{r}_v$: The average rating of user v.
- w_{uv}: The similarity weight between user u and user v (e.g., calculated using PCC, cosine, or Jaccard similarity).

The prediction adjusts the user's average rating based on how similar users rated the item. It considers the difference between each similar user's rating for that item and their average rating, weights this difference by their similarity to the user in question, and normalizes this sum by the total of the absolute similarity weights. This weighted sum adjusts the user's average rating to predict what they might rate the item.

To assess the accuracy of our recommendation system, we computed the Root Mean Squared Error (RMSE) between the actual and predicted ratings. The RMSE is formulated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(predicted_i - actual_i)^2}{N}}$$

where $predicted_i$ and $actual_i$ are the predicted and actual ratings respectively, and N is the total number of predictions made.

4. Discussion

4.1. Content-Based Filtering

In order to evaluate the effectiveness of our content-based recommendation system, we simulated a user with specific movie preferences. By providing a hypothetical list of movies that the

user likes, we aimed to test the system's ability to recommend similar movies. The user's preferences were given as follows:
User Preferences:

- *Toy Story* (Animation/Family)
- *Finding Nemo* (Animation/Adventure)
- *Shrek* (Animation/Adventure)
- *The Lion King* (Animation/Drama)
- *Tarzan* (Animation/Drama)

The outcome of these computations is manifested in the recommendations provided for each similarity measure, as outlined below:

- **Top 10 Jaccard Recommendations:** A Bug's Life, Ice Age, Toy Story 2, The Jungle Book, Monsters, Inc., The Emperor's New Groove, Hercules, Bolt, Frozen, and Aladdin.
- **Top 10 Cosine Recommendations:** Ice Age, A Bug's Life, Toy Story 2, Hercules, The Jungle Book, The Tigger Movie, The Emperor's New Groove, Monsters, Inc., Antz, and Bolt.
- **Top 10 Overlap Recommendations:** Ice Age, A Bug's Life, Toy Story 2, The Jungle Book, The Emperor's New Groove, Hercules, Monsters, Inc., Bolt, Frozen, and Aladdin.

The implementation of Jaccard, Cosine, and Overlap similarity measures demonstrated their effectiveness in identifying movies with similar thematic and genre attributes, specifically within the animated and family-friendly categories. The top recommendations from each measure included films such as "Ice Age," "A Bug's Life," "Toy Story 2," "The Jungle Book," and "Monsters, Inc.," underscoring a consistent alignment with these genres across different computational approaches.

The Jaccard similarity highlighted general thematic similarities by focusing solely on the presence or absence of tags, which is beneficial for broadly grouping movies with similar content. In contrast, Cosine similarity, by considering both the presence and magnitude of tags, captured more nuanced relationships between movies, allowing for a more refined recommendation list that includes films like "The Tigger Movie" and "Antz," which might share more specific thematic elements with other movies in the list. Meanwhile, the Overlap Coefficient, emphasizing the proportion of tag overlap relative to the smaller tag set, proved effective in identifying movies with a high degree of specific thematic overlap, ensuring that the recommendations are not only similar but also intensely relevant to the tags that define user preferences.

The alignment of recommendations across different similarity measures confirms the validity of the content-based approach. The movies recommended using Jaccard, Cosine, and Overlap similarities are closely related to the user's preferences, with significant overlap in the top 10 lists.

A Bug's Life (Animation/Adventure) shares the same animation and adventure genres as "Finding Nemo" and "Shrek," aligning perfectly with the user's interest in animated adventure films. **Ice Age** (Animation/Adventure) combines adventure and comedy in a way similar to "Finding Nemo" and "Shrek," appealing to fans of animated adventures with comedic elements. **Toy Story 2** (Animation/Adventure) continues the beloved "Toy Story" franchise, which the user has already shown an interest in by liking "Toy Story."

The Jungle Book (Animation/Adventure) offers a mix of adventure and musical elements similar to "The Lion King" and "Tarzan," providing an adventurous yet family-friendly viewing experience. **Monsters, Inc.** (Animation/Adventure) is a highly acclaimed animated film that provides comedy and adventure similar to "Shrek" and "Finding Nemo." **The Emperor's New Groove** (Animation/Adventure) offers a humorous animated adventure in a style similar to "Shrek," appealing to fans of light-hearted animated comedy.

Hercules (Animation/Adventure) shares a comedic and adventurous style with "Shrek" and "Finding Nemo," with additional elements of mythology and fantasy. **Bolt** (Animation/Adventure) is an action-packed animated film that aligns with the user's interest in "Shrek" and "Finding Nemo." **Frozen** (Animation/Adventure) appeals to the user's interest in animated adventures, particularly through its strong family-friendly themes, musical elements, and adventurous storyline. **Aladdin** (Animation/Adventure) is similar to "The Lion King" and "Tarzan," with a strong musical component and adventurous storytelling.

(a) Toy Story (b) Finding Nemo (c) Shrek (d) The Lion King

(e) A Bug's Life (f) Ice Age (g) Toy Story 2 (h) The Jungle Book

Figure 7: User Preferences (First Row) and Top Recommendations Across All Similarities (Second Row)

The consistent presence of these recommendations across the different similarity measures illustrates the system's ability to identify popular family-friendly animated films that align with the user's interests. The slight variations in the lists highlight each similarity measure's unique strengths, allowing for tailored recommendations that cater to specific user preferences.

To visually represent the thematic elements of the recommended movies, word clouds were generated for each set of recommendations (see Figure 8). These clouds highlight the most prominent tags associated with the movies recommended by each algorithm, providing an intuitive understanding of the thematic focus of the recommendations.

The word clouds generated from the Jaccard, Cosine, and Overlap recommendations, compared to the user likes, succinctly illustrate the thematic focus of each method. The Jaccard word cloud displays a diverse array of themes like "oscar," and "story," indicating a broad clustering of movies based on shared tags, though possibly lacking depth. In contrast, the Cosine word cloud highlights more specific genres such as "civil war," "comic book," and "sci-fi," reflecting its capability to recognize the intensity and relevance of tags, thus aligning closely with particular narrative themes. The Overlap recommendations share a thematic range similar to Jaccard, focusing on significant tag overlaps but resulting in a broad thematic spectrum. On the other hand, the user likes cloud features terms like "oscar," and "true story," pointing to a preference for narrative-rich and potentially award-winning films. This analysis suggests that the Cosine measure, with its nuanced tag sensitivity, aligns most closely with the user's specific thematic interests, providing tailored recommendations that resonate more deeply with personal tastes.

4.2. Collaborative Filtering

We evaluated the performance of each similarity measure through the RMSE of the generated recommendations:

- **Pearson Correlation Coefficient (PCC):**
  RMSE = 0.7076
- **Cosine Similarity:** RMSE = 0.7076
- **Jaccard Similarity:** RMSE = 1.8958

The evaluation reveals that Pearson and Cosine similarity measures, which account for the magnitude of user ratings, significantly outperform the Jaccard similarity in our model. The higher RMSE for Jaccard indicates that mere knowledge of whether users rated the same items is less informative compared to how similarly they rated them.

Conclusion

In this project, we delved into the vast MovieLens dataset to explore the dynamics of recommendation systems in the realm of big data. Our primary focus was on implementing a content-based recommendation approach while also exploring the principles of collaborative filtering. We navigated through the intricacies of data analytics in the context of movie recommendations, and highlighted the potential of combining content-based and collaborative filtering approaches into a hybrid system, which could offer more powerful recommendations by leveraging the strengths of both methods. Moving forward, further investigations into advanced machine learning techniques could enrich our understanding and refine the efficacy of recommendation systems in accommodating diverse user preferences within the ever-expanding realm of big data analytics.

Figure 8: Word Clouds for Recommended Movies by Each Similarity Measure

Appendix A. Appendix

Appendix A.1. Additional Figures

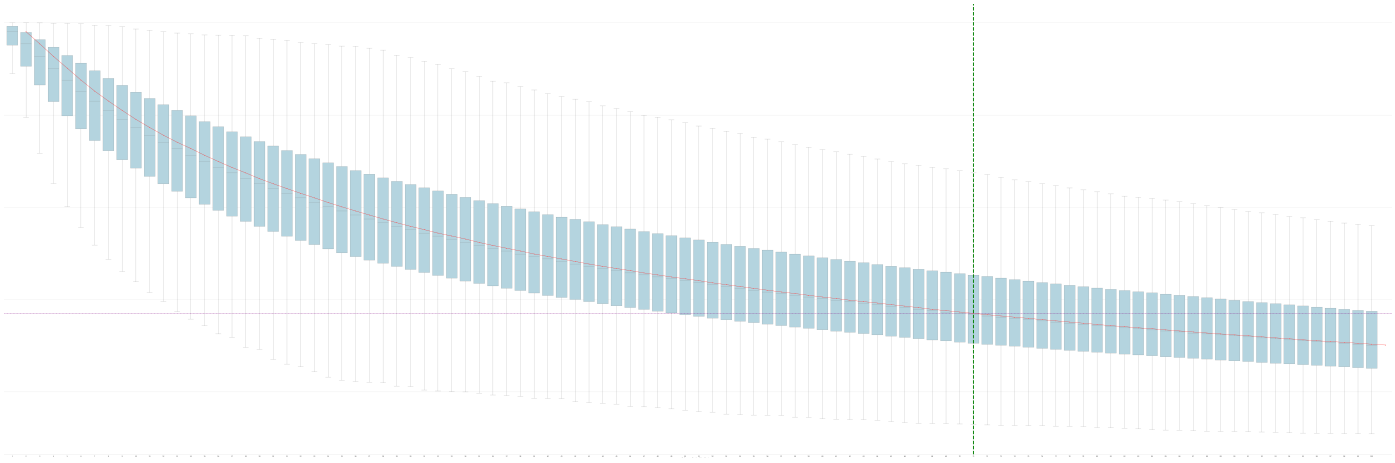

Figure A.9: Distribution of Tag Relevance Scores by Rank Within Movies.