

Prime video insightful recommender : Unraveling patterns with TF-IDF and cosine similarity

Ribhika *

Shilpee Srivastava [†]

Department of Mathematics

Chandigarh University

Mohali 140413

Punjab

India

Geeta [§]

Department of Mathematics

School of Sciences

Geeta University

Panipat 132145

Haryana

India

Abstract

Amazon Prime Video, a prominent on-demand broadcast platform, has achieved a remarkable global reach, extending its services to more than 200 countries and territories. This study delves into the dynamic landscape of Amazon Prime Video, utilizing a dataset of 9,668 records obtained from Kaggle. Our approach involves an exploratory analysis to distill essential insights into the available content on this prominent on-demand broadcast platform. In tandem, we implemented the recommendations using well-designed TF-IDF and cosine similarity algorithms in natural language processing (NLP). Our exploration unveils intriguing data on current content trends. Despite current limitations in the recommendation system, its potential shines through when considering the integration of additional features. This research endeavors to contribute nuanced perspectives on Amazon Prime Video's content landscape, offering insights into emerging trends and the promise held by recommendation algorithms within the platform's dynamic ecosystem.

Subject Classification: 68T09, 68T50, 68W27.

Keywords: Prime video, EDA, Recommendation, NLP, TF-IDF, Cosinesimilarity.

* E-mail: ribhikab19@gmail.com (Corresponding Author)

[†] E-mail: Shilpee2007@rediffmail.com

[§] E-mail: geeta0584@gmail.com

1. Introduction

In the ever-evolving landscape of entertainment, few platforms have had the transformative impact of Amazon Prime Video, commonly known as Prime Video. Built up as an American membership video-on-request over-the-top gushing and rental benefit, Prime Video has become a powerhouse in the industry, available both as a standalone benefit and as portion of Amazon's comprehensive Prime [1]. Propelled on September 7, 2006, as Amazon Unbox within the Joined together States, the benefit has advanced altogether over the long time, adjusting to the improvement of the Prime membership and rebranding as Amazon Moment Video on Request [2].

Prime Video's history is marked by acquisitions and expansion. In 2011, it acquired British broadcast media company LoveFilm [3]. From 2014 DVD by mail service facilitated integration of Prime Video in the UK and Germany for Prime membership [4]. Prime Video steadily grew, culminating in its worldwide launch on December 14, 2016.

The service's history reflects not only its evolution but also its commitment to diverse content. Over the years, Amazon has secured rights to broadcast live sporting events, acquired significant awards for its original productions like "Transparent," and expanded its international reach. Key milestones include the announcement of The Terrific Visit, a motoring appear highlighting Jeremy Clarkson, Richard Hammond, and James May, and the worldwide dispatch of Prime Video into 200 extra nations [2].

Content is at the heart of Prime Video's success, with a vast library of original programming, movies, and TV shows. The platform's foray into original programming began with comedies like Alpha House and Betas in 2013 [5], followed by acclaimed productions such as Transparent. Amazon Studios, the content production arm, holds worldwide TV adjustment rights to The Lord of the Rings [6], showcasing its commitment to high-profile projects.

In the realm of sports programming, Prime Video made strategic acquisitions, securing rights to broadcast NFL's Thursday Night Football, ATP World Tour, and even venturing into live broadcasting of Premier League football matches. The platform continues to invest in sports-related content, with recent announcements regarding UEFA Champions League football and NBA broadcasting rights.

The study outlined in this paper embarks on an exploratory analysis of Prime Video's content landscape, utilizing a dataset from Kaggle [7].

The implementation of a content recommendation system, leveraging advanced algorithms like TF-IDF and Cosine Similarity, underscores the platform's commitment to giving a personalized seeing involvement.

In conclusion, Amazon Prime Video stands as a testament to the dynamism of the entertainment industry. Its rich history, strategic expansions, diverse content offerings, and technological innovations position it as a global player with a profound impact on how audiences consume entertainment. This paper will delve into related works, methodology, results, and discussions, offering insights specific to Amazon Prime Video.

2. Approach

In our approach, we strive to uncover the prevailing patterns in the content offered by Prime Video through an exploratory information investigation (EDA). Hence, we set up a recommendation framework using two Natural Language Processing (NLP) techniques—TF-IDF and Cosine Similarity—to improve content recommendations based on user preferences.

2.1 Data Acquisition

For the groundwork of this study, data were sourced from Kaggle [7], a leading data repository platform renowned for its diverse datasets. The dataset comprises a dozen columns, encompassing critical details such as the show's title, categorization as a movie or TV show, particulars about the director and key actors, the countries where the show is accessible on Prime Video, the assigned rating on the Prime Video platform, the show's duration, and a concise summary.

2.2 Data Cleaning and Transformation

Prior to delving into the exploratory analysis of the dataset, preparatory steps were taken to refine and organize the data effectively. This involved handling missing values, standardizing formats, and performing necessary transformations. The refined dataset serves as a reliable foundation for our subsequent analyses and recommendation system implementation

2.3 Exploratory Information Investigation (EDA)

Exploratory information investigation was performed utilizing the Google Colab on Amazon Prime Video's dataset. Python libraries including NumPy, Pandas, Seaborn, and Matplotlib were employed for comprehensive data analysis. The EDA included generating word clouds to visualize program descriptions, and implementing TF-IDF and cosine similarity algorithms for program recommendations. The analysis also examined missing values, imputation of unknown data, and visualization of trends by year and program category.

2.4 Text Frequency-Inverse Document Frequency

The TF-IDF calculation serves to evaluate the importance of words inside a literary corpus is shown in Fig. 1. This noteworthiness is determined by the recurrence of a word inside a record, with a better recurrence contributing to more noteworthy significance. Simultaneously, it considers the inverse proportionality of the word's frequency across the entire corpus. TF (Term Frequency) represents how often words appear in a corpus, signifying their frequency. It is computed by calculating the number of events of a word separated by the entire number of words within the corpus [8].

IDF (Inverse Document Frequency) measures the significance of a term over the complete corpus. It includes computing the logarithm of the inverse of the extent of documents containing the term. This is often done by determining the entire number of records within the corpus partitioned by the number of records where the word is display.

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$

$$\text{TF-IDF} = \text{TF} * \log(N/\text{DF})$$

Where:

- . TF is the term frequency of a word in a document
- . N is the total number of documents in the corpus
- . DF is the document frequency of a word in the corpus (i.e., the number of documents that contain the word)

Figure 1
The TF-IDF calculation

The resulting algorithm constitutes the IDF esteem. The TF-IDF weight is at that point inferred by multiplying these two measures, showing that a higher weight compares to a greater importance of the word inside the corpus.

2.5 Cosine Similarity

The cosine similarity calculation gages likeness by examining the cosine angle between two vectors, indicated as A and B. These vectors can speak to different components such as words, sentences, passages, or total records. A higher cosine similarity value, approaching 1, indicates a smaller angle between the two vectors, say A and B. When applied to textual corpora, this algorithm becomes instrumental in assessing the degree of similarity between documents [9].

$$\cos(\theta) = \frac{A \cdot B}{|A| |B|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

3. Results

3.1 Exploratory Information Investigation (EDA)

In this area, we reveal the results of our in-depth examination of the information post.

3.1.1 Categorization Ranking on Prime Video

The content landscape on Amazon Prime Video exhibits interesting trends when segmented into movies and TV shows. The data showcases the percentage distribution of these categories over the years. In the year 2000, movies dominated with 88.89%, while TV shows constituted 11.11%. The subsequent years witnessed fluctuations, with movies maintaining a significant presence but gradually making way for the growing popularity of TV shows. Throughout the early 2000s, movies consistently held a higher percentage, but as we progressed into the mid-2010s, the balance started to shift. By 2015, TV shows had gained momentum, constituting 25.4% of the content (figure 2). This trend continued, reaching a notable turning point in 2018 when TV shows surpassed movies, marking a shift in user preferences. The data reflects this evolving pattern, with TV shows steadily increasing their share. By 2021, movies accounted for 78.99%, while TV shows constituted 21.01% of the content on Amazon Prime Video (figure

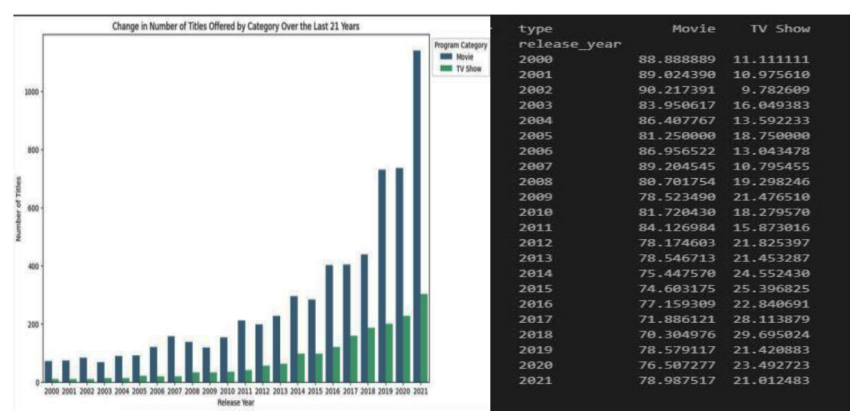


Figure 2
Evolution in Title Count and Category Distribution across Two Decades

2). This transformation in the platform’s content offering suggests a broader shift in viewer preferences over the years, with TV shows progressively becoming more prominent. The year 2021, marked by the global COVID-19 pandemic, might have influenced user behaviors and contributed to the dynamic nature of the streaming landscape.

3.1.2 Global Landscape of Content on Prime Video

A word cloud actsshown in Figure 3 as a sort of condensed summary of a body of text. It visually displays the words that occur most frequently, with the size of each word indicating its frequency – the larger the word, the more often it appears. Typically, a word cloud serves as an initial visualization step, providing a broad understanding of the main themes within a text. It precedes the application of more sophisticated Natural Language Processing (NLP) algorithms, which can offer more detailed insights into the semantic nuances of the document.



Figure 3
Word Cloud Analysis of Program Descriptions on Prime Video

In the context of Prime Video content, it's crucial to recognize that the significance of words extends beyond mere frequency. While a word cloud reveals prevalent terms like "life," "one," "fine," "family," "love," "world," and "story," understanding their contextual relevance requires more advanced Natural Language Processing (NLP) technique. However, relying solely on word frequency has limitations in capturing contextual significance. The Bag-of-Words model, based on frequency vectors, simplifies document representation but lacks depth in understanding context. Advanced Natural Language Processing (NLP) techniques like Word Embedding [10], Word2vec [11] or TF-IDF are preferred in the NLP domain. These methods delve into word relationships and context, offering a more nuanced interpretation of a document's semantics use [12].

3.2 Content Recommendation Framework Utilizing TF-IDF and Cosine Similarity

Prime Video's recommendation framework operates with a nuanced approach, leveraging algorithms like TF-IDF and Cosine Similarity to suggest content based on contextual similarity. The output for 'The Boys' reveals calculated scores for various titles, indicating their resemblance to the selected series. Notably, the recommendations span diverse genres, offering viewers a well-rounded selection beyond the immediate context.

Examining the growth trend over the last 21 years, the data showcases a consistent increase in Prime Video's content offerings. The evolution from 88.89% movies and 11.11% TV shows in 2000 to a more balanced distribution in recent years signifies the platform's adaptability and commitment to diverse content creation. The Boys, a standout series on Prime Video, has garnered a noteworthy calculated score in the recommendation system. It has achieved a notable calculated score, reflecting the intricacies of Prime Video's recommendation algorithm. This score is a quantitative representation of the similarity between The Boys and other titles within the platform's extensive library. The calculated score is a crucial metric as it determines the relevance and affinity of recommended titles to The Boys. A higher score indicates a closer thematic or narrative alignment, enhancing the precision of the recommendation system. In the case of The Boys shown in Figure 4, the calculated score serves as a guide for users, offering transparency into why specific titles are suggested. Now, transitioning to the recommendations themselves, Prime Video provides a curated list of titles that share thematic elements,

	Title	Calculated Score	Recommendations for 'The Boys':
739	The Big Sleep	0.166238	8979 Gamma smoke test
4313	Razorteeth	0.161962	8597 Rammstein: in Amerika - Live from Madison Squa...
8202	I Am Dragon	0.158210	8415 X-Ray 1C Test Title
5555	The Ultimate Gift	0.134934	5043 Jabillikosam akasamalle
1777	Madame	0.133835	3430 A Company of Heroes
...	...	...	1260 Prime Rewind: Inside The Boys
7316	Dhogs	0.000000	4622 Wired Outdoors
3980	What Haunts Us	0.000000	3251 Anger Management
7314	Domestique	0.000000	5000 Isha Sakhi
3982	Uratu: The Forgotten Kingdom	0.000000	
9667	Harry Brown	0.000000	
[9667 rows x 2 columns]			Name: title, dtype: object

Figure 4

listed of calculate score and recommendation for “The Boys”

genre nuances, or narrative intricacies with The Boys. These recommendations extend beyond conventional genre categorizations, enriching the user experience by introducing a diverse range of content.

For instance, titles like ‘The Big Sleep’ and ‘Razor teeth’ not only echo The Boys’ themes but also introduce viewers to different genres and storytelling approaches. This holistic approach to recommendations ensures that users encounter a tailored selection that aligns with their unique preferences, fostering a more engaging and satisfying streaming experience.

4. Discussion and Conclusion

In conclusion, this study provides a thorough exploration of the dynamic landscape of Amazon Prime Video, utilizing a dataset of 9,668 records sourced from Kaggle. Through an extensive exploratory analysis and the implementation of a recommendation framework utilizing TF-IDF and Cosine similarity calculations, we have revealed crucial insights into the content distribution and trends on the platform. Segmenting the content into movies and TV shows uncovered an evolving pattern, with TV shows progressively gaining prominence over the years. The significant shift in 2018, where TV shows surpassed movies, reflects changing viewer preferences. The global context of works available on Prime Video is visually summarized through a word cloud, underscoring the importance of advanced Natural Language Processing (NLP) techniques for contextual understanding. The recommendation framework, based on TF-IDF and Cosine Similarity, demonstrated calculated scores for titles related to ‘The Boys,’ showcasing the precision of Prime Video’s recommendation algorithm. ‘The Boys,’ with a notable calculated score, serves as a

benchmark for understanding the thematic and narrative alignment of recommended titles.

In summary, Amazon Prime Video stands as a dynamic player in the entertainment industry, marked by strategic expansions, diverse content offerings, and technological innovations. This study contributes nuanced perspectives on the platform's content landscape, highlighting emerging trends and the promise held by recommendation algorithms in delivering personalized viewing experiences. The integration of advanced NLP techniques underscores Prime Video's commitment to providing content tailored to diverse global audiences, solidifying its position as a leading on-demand broadcast platform.

References

- [1] Team91, "Amazon Prime plans 2023: monthly, quarterly, and yearly membership price in India, benefits, Jio and Airtel offers, and more," Apr. 27 (2023).
- [2] Amazon Prime Video, Wikipedia, "History." Available: https://en.wikipedia.org/wiki/Amazon_Prime_Video#cite_note-PR1-13.
- [3] Amazon takes full control of Lovefilm," The Guardian, Jan. 20 (2011). Available: <https://www.theguardian.com/business/2011/jan/20/amazon-lovefilm-streaming>.
- [4] P. Sawers, "Amazon Launches Prime Instant Video in UK & Germany," The Next Web, Feb. 21 (2014). Available: <https://thenextweb.com/media/2014/02/21/amazon-launches-prime-instant-video-in-uk-germany/>.
- [5] "Amazon's Original Series 'Alpha House' Debuts Friday," The Motley Fool, Associated Press, Nov. 12 (2013). Available: <https://www.fool.com/investing/general/2013/11/12/amazons-original-series-alpha-house-debuts-friday.aspx>.
- [6] M. Sweney, "Amazon's \$1bn bet on Lord of the Rings shows scale of its TV ambition," The Guardian, Nov. 21 (2017). Available: <https://www.theguardian.com/media/2017/nov/21/amazon-1bn-bet-lord-of-the-rings-shows-scale-tv-ambition>.
- [7] Kaggle, "Amazon Prime Movies and TV Shows," (2021). Available: <https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows>.

- [8] A. Pradeep, "Understanding TF-IDF: A comprehensive guide to term frequency-inverse document frequency," *Journal of Natural Language Processing*, vol. 10, no. 2, pp. 123-145 (2023).
- [9] Y. Liu, X. Qi, and T. Zhang, "Research on Text Classification Method based on PTF-IDF and Cosine Similarity," *Journal of Information and Communication Engineering*, vol. 6, no. 1, pp. 335-338 (2020).
- [10] E. H. Mohamed, M. E. Moussa, and M. H. Haggag, "An Enhanced Sentiment Analysis Framework Based on Pre-Trained Word Embedding," *International Journal of Computational Intelligence and Applications*, vol. 19, no. 5, p. 2050031 (2020).
- [11] T. Mikolov, W. T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in Proc. 2013 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (2013).
- [12] A. Jain, S. Tiwari, and N. Mani, "Qualifying Lemon Scab Severity: Autoencoder and XGBoost in a Hybrid Approach," in *Doctoral Symposium on Computational Intelligence*, pp. 531-542 (2024).

Received September, 2024