Movie Recommendation System Using Content Based Filtering

Kusuma.S
Dept. of ISE
M S Ramaiah Institute of Technology,
Bengaluru, India
kusuma.s@msrit.edu

Charunayana V
Dept. of ISE
M S Ramaiah Institute of Technology,
Bengaluru, India
charunayana@msrit.edu

Shruthi J R
Dept. of ISE
M S Ramaiah Institute of Technology,
Bengaluru, India
jrshruti@msrit.edu

Vinutha G K
Dept. of CSE
R N S Institute of Technology,
Bengaluru, India
vinuthaharishk@gmail.com

Shivaraj
Dept. of ISE
M S Ramaiah Institute of Technology,
Bengaluru, India
1ms22is410@msrit.edu

Mohamed Waseem
Dept. of ISE
M S Ramaiah Institute of Technology,
Bengaluru, India
1ms22is406@msrit.edu

Abstract

This research outlines a movie recommendation system that recommends movies based on the user's input. It analyses the movie descriptions to recommend thematically related films based on content-based filtering. The system pre-processes movie data from a CSV file. It selects relevant features like genres and cast descriptions from it. These features are combined and transformed using TF-IDF (Term Frequency-Inverse Document Frequency) to capture the importance of words within descriptions. Cosine similarity then calculates the thematic similarity. This is between the user's chosen movie (potentially corrected for typos) and other movies. Finally, the system recommends the top N (default 30) most similar movies to the user. This recommendation system provides movie descriptions, recommends movies that share similar themes based on the user inputs.

Keywords. Content-based filtering, Thematically related films, TF-IDF (Term Frequency-Inverse Document Frequency), Cosine similarity.

1. Introduction

In today's ever-expanding cinematic landscape, navigating the vast array of movies can be a daunting task. Choosing the perfect film can feel like searching for a hidden gem. This is where movie recommendation systems come in, offering personalized suggestions to guide viewers towards their next cinematic adventure.

This research delves into a content-based movie recommendation system that transcends mere genre suggestions. The proposed work is beyond the surface level and delve into the rich tapestry of movie descriptions. By analysing the thematic weight of words within these descriptions, the system fosters a more content-driven approach to recommend movies of user interest.

This approach provides several advantages. It allows users to discover movies that share similar themes and artistic expressions with their initial selection, even if they fall outside the typical genre boundaries. Furthermore, it empowers viewers to explore hidden gems and embark on unexpected cinematic journeys that resonate with their specific preferences.

The power of TF-IDF and cosine similarity application are used in this research to find the thematically related movies. This system aims to empower viewers to embark on a personalized cinematic odyssey, filled with films that resonate with their unique tastes.

2. METHODOLOGY

The methodology section outlines the dataset, pre-processing steps, model implementation, training, evaluation, and the ensemble method used to improve classification performance. The methodology for this system takes many steps as discussed below.

7.1 Data pre processing

The data pre-processing step involves cleaning the dataset, handling missing values, and scaling features as shown in the Figure 2.1.

index	budget	genres	homepage	id	keywords	original_la	original_ti	overview	popularity	productior	productior	release_d	revenue	runtime	spoken_la	status	tagline	title	vote_aver	vote_coun	cast	crew	director
0	2.37E+08	Action Adv	http://ww	19995	culture cla	en	Avatar	In the 22n	150.4376	[{"name":	[{"iso_316	#######	2.79E+09	162	[{"iso_639	Released	Enter the W	Avatar	7.2	11800	Sam Wortl	[{'name': '	James Cameron
1	3E+08	Adventure	http://disn	285	ocean drug	en	Pirates of t	Captain Ba	139.0826	[{"name":	[{"iso_316	#######	9.61E+08	169	[{"iso_639	Released	At the end	Pirates of t	6.9	4500	Johnny Dep	[{'name': '	Gore Verbinski
2	2.45E+08	Action Adv	http://ww	206647	spy based	en	Spectre	A cryptic m	107.3768	[{"name":	[{"iso_316	#######	8.81E+08	148	[{"iso_639	Released	A Plan No	Spectre	6.3	4466	Daniel Cra	[{'name': '	Sam Mendes
3	2.5E+08	Action Crir	http://ww	49026	dc comics	en	The Dark K	Following t	112.313	[{"name":	[{"iso_316	#######	1.08E+09	165	[{"iso_639	Released	The Legen	The Dark K	7.6	9106	Christian B	[{'name': '	Christopher Nolan
4	2.6E+08	Action Adv	http://mov	49529	based on n	en	John Carte	John Carte	43.927	[{"name":	[{"iso_316	3/7/2012	2.84E+08	132	[{"iso_639	Released	Lost in our	John Carte	6.1	2124	Taylor Kits	[{'name': '	Andrew Stanton
5	2.58E+08	Fantasy Ac	http://ww	559	dual identi	en	Spider-Ma	The seemi	115.6998	[{"name":	[{"iso_316	5/1/2007	8.91E+08	139	[{"iso_639	Released	The battle	Spider-Ma	5.9	3576	Tobey Mag	[{'name': '	Sam Raimi
6	2.6E+08	Animation	http://disn	38757	hostage m	en	Tangled	When the	48.68197	[{"name":	[{"iso_316	#######	5.92E+08	100	[{"iso_639	Released	They're tal	Tangled	7.4	3330	Zachary Le	[{'name': '	Byron Howard
7	2.8E+08	Action Adv	http://mar	99861	marvel cor	en	Avengers:	When Ton	134.2792	[{"name":	[{"iso_316	#######	1.41E+09	141	[{"iso_639	Released	A New Age	Avengers:	7.3	6767	Robert Do	[{'name': '	Joss Whedon
8	2.5E+08	Adventure	http://han	767	witch mag	en	Harry Pott	As Harry b	98.88564	[{"name":	[{"iso_316	7/7/2009	9.34E+08	153	[{"iso_639	Released	Dark Secre	Harry Pott	7.4	5293	Daniel Rac	[{'name': '	David Yates
9	2.5E+08	Action Adv	http://ww	209112	dc comics	en	Batman v S	Fearing th	155.7905	[{"name":	[{"iso_316	#######	8.73E+08	151	[{"iso_639	Released	Justice or r	Batman v S	5.7	7004	Ben Affled	[{'name': '	Zack Snyder
10	2.7E+08	Adventure	http://ww	1452	saving the	en	Superman	Superman	57.92562	[{"name":	[{"iso_316	#######	3.91E+08	154	[{"iso_639	Released		Superman	5.4	1400	Brandon R	[{'name': '	Bryan Singer
11	2E+08	Adventure	http://ww	10764	killing und	en	Quantum c	Quantum c	107.9288	[{"name":	[{"iso_316	#######	5.86E+08	106	[{"iso_639	Released	For love, fc	Quantum c	6.1	2965	Daniel Cra	[{'name': '	Marc Forster
12	2E+08	Adventure	http://disn	58	witch fortu	en	Pirates of t	Captain Ja	145.8474	[{"name":	[{"iso_316	#######	1.07E+09	151	[{"iso_639	Released	Jack is bad	Pirates of t	7	5246	Johnny Dep	[{'name': '	Gore Verbinski
13	2.55E+08	Action Adv	http://disn	57201	texas hors	en	The Lone R	The Texas	49.04696	[{"name":	[{"iso_316	7/3/2013	89289910	149	[{"iso_639	Released	Never Tak	The Lone R	5.9	2311	Johnny Dep	[{'name': '	Gore Verbinski
14	2.25E+08	Action Adv	http://ww	49521	saving the	en	Man of Ste	A young bc	99.39801	[{"name":	[{"iso_316	#######	6.63E+08	143	[{"iso_639	Released	You will be	Man of Ste	6.5	6359	Henry Cavi	[{'name': '	Zack Snyder
15	2.25E+08	Adventure Family Fan		2454	based on n	en	The Chroni	One year a	53.9786	[{"name":	[{"iso_316	#######	4.2E+08	150	[{"iso_639	Released	Hope has a	The Chroni	6.3	1630	Ben Barne	[{'name': '	Andrew Adamson
16	2.2E+08	Science Fic	http://mar	24428	new york s	en	The Aveng	When an u	144.4486	[{"name":	[{"iso_316	#######	1.52E+09	143	[{"iso_639	Released	Some asse	The Aveng	7.4	11776	Robert Do	[{'name': '	Joss Whedon
17	3.8E+08	Adventure	http://disn	1865	sea captai	en	Pirates of t	Captain Ja	135.4139	[{"name":	[{"iso_316	#######	1.05E+09	136	[{"iso_639	Released	Live Forev	Pirates of t	6.4	4948	Johnny Dep	[{'name': '	Rob Marshall
18	2.25E+08	Action Con	http://ww	41154	time trave	en	Men in Bla	Agents J (V	52.03518	[{"name":	[{"iso_316	#######	6.24E+08	106	[{"iso_639	Released	They are b	Men in Bla	6.2	4160	Will Smith	[{'name': '	Barry Sonnenfeld
19	2.5E+08	Action Adv	http://ww	122917	corruption	en	The Hobbi	Immediate	120.9657	[{"name":	[{"iso_316	#######	9.56E+08	144	[{"iso_639	Released	Witness th	The Hobbi	7.1	4760	Martin Fre	[{'name': '	Peter Jackson
20	2.15E+08	Action Adv	http://ww	1930	loss of fath	en	The Amazi	Peter Park	89.86628	[{"name":	[{"iso_316	#######	7.52E+08	136	[{"iso_639	Released	The untold	The Amazi	6.5	6586	Andrew Ga	[{'name': '	Marc Webb
21	2E+08	Action Adv	http://ww	20662	robin hood	en	Robin Hoo	When sold	37.6683	[{"name":	[{"iso_316	#######	3.11E+08	140	[{"iso_639	Released	Rise and ri	Robin Hoo	6.2	1398	Russell Cro	[{'name': '	Ridley Scott
22	2.5E+08	Adventure	http://ww	57158	elves dwar	en	The Hobbi	The Dwarv	94.37056	[{"name":	[{"iso_316	#######	9.58E+08	161	[{"iso_639	Released	Beyond da	The Hobbi	7.6	4524	Martin Fre	[{'name': '	Peter Jackson
23	1.8E+08	Adventure	http://ww	2268	england cc	en	The Golde	After overl	42.99091	[{"name":	[{"iso_316	#######	3.72E+08	113	[{"iso_639	Released	There are	The Golde	5.8	1303	Dakota Bl	[{'name': '	Chris Weitz
24	2.07E+08	Adventure Drama Act		254	film busine	en	King Kong	In 1933 Ne	61.22601	[{"name":	[{"iso_316	#######	5.5E+08	187	[{"iso_639	Released	The eighth	King Kong	6.6	2337	Naomi Wa	[{'name': '	Peter Jackson
25	2E+08	Drama Ro	http://ww	597	shipwreck	en	Titanic	84 years la	100.0259	[{"name":	[{"iso_316	#######	1.85E+09	194	[{"iso_639	Released	Nothing or	Titanic	7.5	7562	Kate Winsl	[{'name': '	James Cameron
26	2.5E+08	Adventure	http://mar	271110	civil war w	en	Captain Ar	Following t	198.3724	[{"name":	[{"iso_316	#######	1.15E+09	147	[{"iso_639	Released	Divided W	Captain Ar	7.1	7241	Chris Evan	[{'name': '	Anthony Russo
27	2.09E+08	Thriller Action Advent		44833	fight u.s. n	en	Battleship	When mar	64.92838	[{"name":	[{"iso_316	#######	3.03E+08	131	[{"iso_639	Released	The Battle	Battleship	5.5	2114	Taylor Kits	[{'name': '	Peter Berg
28	1.5E+08	Action Adv	http://ww	135397	monster di	en	Jurassic W	Twenty-tw	418.7086	[{"name":	[{"iso_316	6/9/2015	1.51E+09	124	[{"iso_639	Released	The park is	Jurassic W	6.5	8662	Chris Pratt	[{'name': 'Steven Spielberg', 'g	
p'	'job': 'Set I	'credit_id'	'id': 1463	{'name': '(	'gender': (	'departme	'job': 'Set I	'credit_id'	'id': 1463	{'name': '	'gender': (	'departme	'job': 'Proj	'credit_id'	'id': 1463	{'name': '	'gender': (	'departme	'job': 'Fole	'credit_id'	'id': 1463	{'name': '	'gender': ('departı
29	2E+08	Action Adv	http://ww	37724	spy secret	en	Skyfall	When Bon	93.00499	[{"name":	[{"iso_316	#######	1.11E+09	143	[{"iso_639	Released	Think on y	Skyfall	6.9	7604	Daniel Cra	[{'name': '	Sam Mendes
30	2E+08	Action Adv	http://ww	558	dual identi	en	Spider-Ma	Peter Park	35.14959	[{"name":	[{"iso_316	#######	7.84E+08	127	[{"iso_639	Released	There's a f	Spider-Ma	6.7	4321	Tobey Mag	[{'name': '	Sam Raimi

Fig 2.1: Data Pre Processing

Reading the Script: The system starts by reading movie data from a CSV file. This file is assumed to contain various attributes of each movie, such as title, genre(s), keywords, tagline, cast, and director.

Feature Focus: Not all movie information is equally relevant for recommendations. The system selects specific features that provide rich insights into a movie's content. These features typically include genres (e.g., comedy, thriller), keywords (e.g., dark humor, time travel), tagline (a concise summary of the movie's plot or theme), cast (actors and actresses involved), and director (their filmmaking style and thematic preferences). To understand a movie's essence, the system goes beyond just genres. It delves into richer details. Genres, like comedy or thriller, provide a broad category. But keywords offer a deeper look, with tags like "dark humor" or "time travel" revealing specific themes. The tagline, a concise summary, captures the movie's core message. Even the cast and director provide clues. Actors like Jackie Chan are known for action comedies, while different directors often have signature styles and thematic preferences. By combining all these elements – genres, keywords, tagline, cast, and director – the system creates a comprehensive description for each movie, allowing for more nuanced recommendations.

Missing Pieces, handled with Care: Real-world data often contains missing values. The system addresses missing data points in the chosen features. A common approach is to fill them with empty strings. While this doesn't introduce additional information, it avoids skewing the analysis during later stages.

Weaving the Narrative Tapestry: Once the data is cleaned, the system combines the selected features for each movie. This creates a comprehensive description for every movie, encompassing its genre, thematic keywords, tagline, and information about the cast and director. By combining these elements, the system creates a richer representation of each movie's content.

2.2 Algorithms Used

TF-IDF (Term Frequency-Inverse Document Frequency)

The Vectorizer's Magic: Now, we enter the world of Natural Language Processing (NLP). Here, a TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is created. This tool analyzes the combined movie descriptions and transforms them into numerical feature vectors. Imagine a vector as a special code that captures the importance of different words within the description.

Term Frequency: Shining a Light on Frequent Flyers: TF-IDF considers how often each word appears within a particular movie description. Words that show up more frequently are likely more important for understanding the movie's content. The TF-IDF vectorizer assigns a weight to each word based on its frequency, reflecting its significance within that specific movie.

Inverse Document Frequency: Not All Stars Shine Equally Bright: TF-IDF goes beyond just counting word occurrences. It also considers the inverse document frequency (IDF) of each word. Words that appear very frequently across all movie descriptions (e.g., "the," "a") are less important for distinguishing a particular movie's theme. IDF downplays the weight of such common words, ensuring that words unique and informative for a specific movie have a greater influence in the recommendation process.

The Power of the Vector: Through TF-IDF, each movie description is transformed into a numerical vector. This vector captures the importance of various words within the description, providing a more sophisticated representation of the movie's content compared to simply using raw text.

2.3 Model Training and Evaluation

The models were trained on the TF-IDF vectors of the training set and evaluated using the testing set. The following metrics were used for evaluation:

- Accuracy: [6]

$$\text{Accuracy} = \frac{TP + TN}{\text{Total Count}}$$

- Precision: [6]

$$\text{Accuracy} = \frac{TP}{\text{Predicted Yes}}$$

- Recall: [6]

$$\text{Recall} = \frac{TP}{\text{Actual Yes}}$$

8. RESULTS

```
 Enter your favourite movie name :  spider man
Movies suggested for you :

1 . Spider-Man
2 . Spider-Man 3
3 . Spider-Man 2
4 . The Notebook
5 . Seabiscuit
6 . Clerks II
7 . The Ice Storm
8 . Oz: The Great and Powerful
9 . Horrible Bosses
10 . The Count of Monte Cristo
11 . In Good Company
12 . Finding Nemo
13 . Clear and Present Danger
14 . Brothers
15 . The Good German
16 . Drag Me to Hell
17 . Bambi
18 . The Queen
19 . Charly
20 . Escape from L.A.
21 . Daybreakers
22 . The Life Aquatic with Steve Zissou
23 . Labor Day
24 . Wimbledon
25 . Cold Mountain
26 . Hearts in Atlantis
27 . Out of the Furnace
28 . Bullets Over Broadway
29 . The Purge: Election Year
```

Figure 3.12. Sample Output of Movie Recommendation's

This research work of movie recommendation system, personalizes suggestions based on movie content. The sample result based on the user input is as shown in the Figure 3.1. By utilizing TF-IDF technique and by leveraging cosine similarity it analyses the importance of words within movie descriptions. By considering features like genres, keywords, tagline, cast, and director, the system creates a richer representation of each movie's theme. Comparing these representations, movies with similar thematic content are recommend to a user. It offers a personalized content-based approach that helps users discover movies that share themes and styles.

9. CONCLUSION

In conclusion, this content-based movie recommendation system offers a personalized approach to exploring the vast cinematic landscape. By leveraging TF-IDF and focusing on movie content, the system empowers users to discover hidden gems that resonate with their thematic preferences, taking them on unexpected cinematic journeys.

10. References

[1] M. Kumar and S. Patel, "Movie Recommendation System Using Machine Learning," in *Proc. 2024 IEEE Int. Conf. AI Data Sci. (ICAIDS)*, 2024, pp. 1–6.

[2] A. Sharma and R. Gupta, "Movie Recommender System Based on Generative AI," in *Proc. 2024 IEEE Int. Conf. Comput. Intell. Mach. Learn. (CIML)*, 2024, pp. 1–8.

[3] L. Chen and Y. Wang, "Designing a Movie Recommendation System Through a Transformer-Based Embeddings Space," in *Proc. 2024 IEEE Symp. AI Multimedia Process. (AI-MP)*, 2024, pp. 45–52.

[4] P. Singh and K. Verma, "Enhancing Movie Recommendation Systems with BERT: A Deep Learning Approach," in *Proc. 2024 IEEE Conf. Natural Lang. Process. AI (NLP-AI)*, 2024, pp. 78–85.

[5] S. Das and M. Roy, "Movie Recommendation System Based On Emotions," in *Proc. 2024 IEEE Int. Conf. Human-Comput. Interact. AI (HCI-AI)*, 2024, pp. 112–119.

[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson Education Limited, 2021.

[7] R. Baeza-Yates, M. Ferro, and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and the Algorithms*. ACM Press, 2011.

[8] G. Guo, H. Wang, J. Yin, and S. He, "A novel movie recommendation system based on social network analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 23, pp. 6309–6319, 2013.

[9] J. Srivastava and P. K. Sharma, "A hybrid recommendation system for movies," *Procedia Comput. Sci.*, vol. 46, pp. 1243–1249, 2014.

[10] A. Said and T. K. Mohamed, "A survey on content-based movie recommendation systems," *ACM Comput. Surv. (CSUR)*, vol. 48, no. 3, pp. 1–36, 2015.

[11] S. Ventura and J. P. Silva, "A survey on content-based recommender systems," *J. Artif. Intell. Res.*, vol. 52, pp. 621–670, 2015.

[12] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge Univ. Press, 2009.

[13] G. Piao and J. Kim, "A hybrid collaborative filtering method for movie recommendation," *Expert Syst. Appl.*, vol. 31, no. 4, pp. 786–798, 2006.

[14] T. Grossman and M. Newman, "Spam detection using n-gram co-occurrence statistics," in *Proc. 7th Int. Workshop Inf. Retrieval Asian Lang.*, 2004, pp. 178–186.

[15] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*, 1998, pp. 285–295.

[16] R. H. Sloan, "A method for measuring similarity between documents," *J. Am. Soc. Inf. Sci.*, vol. 48, no. 6, pp. 433–443, 1996.

[17] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–404, 1990.