

Recommendation System Based on Content Based Filtering

Ms. Jisna P Antony

PG Scholar

Department of Computer Application
Amal Jyothi College of Engineering
Kanjirappally, India
jisnapantony@gmail.com

Mr. Jinson Devis

Assistant Professor

Department of Computer Application
Amal Jyothi College of Engineering
Kanjirappally, India
jinsondevis@amaljyothi.ac.in

Abstract—A recommendation system is a subclass of information filtering systems that provide or suggests products to its target audience. Recommendation systems are widely used these days. It may be in the form of friend suggestions on Facebook, suggesting similar products on e-commerce sites, etc. Every time we use an e-commerce website, we receive product suggestions based on our prior search activity. There are numerous ways to implement recommendation systems, such as collaborative filtering, content-based filtering, hybrid filtering. This paper developed a book recommendation engine that uses a content-based filtering approach based on their previous actions. Inverse Frequency Function (TF-IDF) (Document Frequency) and cosine similarity were used to implement content-based filtering and It determines how relevant a product is to a user's interests.

Keywords— Content-based filtering, TF-IDF, Cosine similarity

I. INTRODUCTION

Nowadays, the amount of knowledge available on the Internet is expanding quickly. There is a massive amount of information, but the majority of which does not interest the user, either as unwanted information or as content irrelevant to his/her interests. However, each user has certain interests that may only be related to a small portion on the Web. As a result, it is now even harder and takes longer for users to find information in which they are interested. The Web can be customized by employing recommender systems to help people find the information that is relevant to their interests. We can quickly access relevant information without manually searching the web with the help of recommendation systems. As a result, a lot of websites today use recommendation systems to market and sell their goods. Many products, including music, movies, books and etc can be recommended to a client based on that person's online shop or social network profile, browsing habits such as links clicked, browsing activity and other online behaviors. Utilizing such tools, online stores are expanding their sales.

In this paper we propose using recommendation systems for recommending books. In this the books are recommended for the user based on the user preferences title and description of the book.

II. LITERATURE REVIEW

Ramni Harbir Singh, Sargam Maurya, Tanisha Tripathi, Tushar Narula, Gaurav Srivastav [1] proposed a movie recommendation using the content-based filtering. The authors use cosine similarity principle and the KNN algorithm are both used in this model because they provide

greater accuracy than other distance metrics while also having a relatively low level of complexity.

Vikas Sindhwani and Prem Melville [2] introduce a paper on "Recommender Systems". They define several methods and approaches for recommendations. They also attempted to identify the typical issues and constraints with the recommendation system.

Sujoy Datta, Das, Laxman Sahoo and Debashis [3] introduce a paper on "A survey on recommendation system". The paper explains about different types of recommendation systems and their general information. It is a survey paper on recommendation. Both non-personalized and personalized systems were mentioned by the authors. A very good example was used to clarify user-based collaborative filtering versus item-based collaborative filtering. The advantages and disadvantages of various recommendation systems have also been discussed by the authors.

Ayush Jain, Kaza Sai Vineeth, Abhiraj Biswas and Mohana proposed a paper "Development of Product Recommendation Engine By Collaborative Filtering and Association Rule Mining Using Machine Learning Algorithms" [4] using collaborative filtering and association rule mining. The use of association rule mining and collaborative filtering increased cross-selling of products. For the creation of effective similarity scores, algorithms including cosine similarity, jaccard similarity, and pearson correlation are used.

M. Chandrashekhar, V. Subramaniaswamy, R. Logesh, Anirudh Challa and V. Vijayakumar [5] developed a method for individualized movie recommendations that makes use of collaborative filtering. In order to determine which user is the most similar, the Euclidean distance metric has been applied. The user with the smallest Euclidean distance value is identified. The final factor in movie recommendations is the user's overall rating. The authors even stated that the recommendations change over time in order to help the system adapt to the user's evolving preferences.

III. RECOMMENDATION TECHNIQUES

A subcategory of information filtering systems called recommendation systems recommends products to customers based on user ratings and predictions made in the past. There are 3 main types; they are content based recommendation system, collaborative filtering, and hybrid systems.

A. Content based Filtering

The selection and determination of items by a content-based filtering system is based on the relationships and correlations between the content of the items in the dataset. In this case, it describes the book's content and the user's history of book purchases. It selects a variety of characteristics from the book in order to suggest other books with a related content. The user will be given an overview of the book's content. In order for the user to quickly locate the book they wish to use or purchase. The complete dataset of books is filtered by a content-based recommendation engine based on the book's content that the customer is most interested in purchasing. The recommendation algorithm separates books from other books with the same content using content-based filtering.

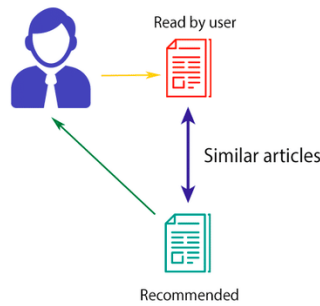


Fig. 1. Content based filtering

B. Collaborative Filtering

The quality of the item cannot be determined by content-based screening. Because they are based on the opinions of other users, collaborative filtering techniques are employed to solve this issue. The core concept is that users rate items first, then the system compares these ratings with those of other users to suggest products based on shared interests to the user. This is mainly used in e-commerce. The websites that make product recommendations based on other products user reviews. As a result, high-quality products are advised.

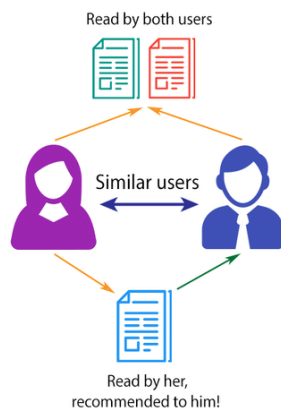


Fig. 2. Collaborative filtering

C. Hybrid Filtering

Hybrid recommendation the products are recommended using both content-based and collaborative filtering. This hybrid strategy was developed to address an issue with traditional recommendation systems. Numerous techniques have been calculated utilizing hybrid techniques include

switching, weighted, mixed, and feature cascade, feature augmentation, meta level and combination

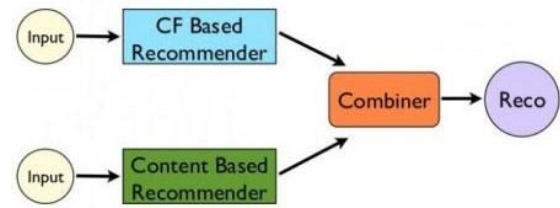


Fig. 3. Hybrid filtering

IV. METHODOLOGY

The goal of this recommendation system is to suggest books to buyers based on their interests. This approach uses the content describing the items and the consumers' preferences or wants instead of relying on other users' ratings. It presents the idea of TF-IDF and determines the item's similarity from its description. There are three ways to calculate how similar two item vectors are: Cosine similarity, Euclidian distance, and Pearson's correlation. In this case, the cosine similarity is used to determine the similarity between the items.

- Data Collection: User have a dataset for book, there 3802 rows and 3 columns. The columns are id, book_title and genres. The id denotes the uniquely identified id of a book, title is the title of book and genres is the genres of each book.

	A	B	C	D	E	F	G	H	I	J	K
1	id	book_title	genres								
2		1 The Hungry	Young Adult Fiction Science Fiction Dystopia Fantasy Science Fiction								
3		2 Harry Pott	Fantasy Young Adult Fiction								
4		3 To Kill a M	Classics Fiction Historical Historical Fiction Academic School								
5		4 Pride and	Classics Fiction Romance								
6		5 Twilight	Young Adult Fantasy Romance Paranormal Vampires Fiction Fantasy Paranormal								
7		6 The Book	Historical Historical Fiction Fiction Young Adult								
8		7 The Chron	Fantasy Classics Fiction Young Adult Childrens								

Fig. 4. Dataset

- Keyword-Based Vector-Space Model: This model is based on the TF-IDF weighing technique to represent each book as a vector of weights, where each weight denotes the degree to which the relationship between the research paper and a term or keyword.
- Item Representation: The items are represented by a group of features. The attributes are: title of the book, id and the genres.
- Keyword extraction and cosine similarity: There are 3 methods for keyword extraction such as CountVectorizer, Tf-Idf Vectorizer, and Rake. Tf-Idf Vectorizer representation is used in this.

Tf-Idf counts the number of times a word appears in a document and then compares this number to the number of other documents in the collection where the term also appears. The system then assigns each word a rank based on how frequently it appears in that particular document while not appearing frequently in any of the other texts. TF-IDF is given by:

$$tf-idf(t, d, D) = tf(t, d) \times idf(t, D)$$

Where,

t = term in the user's query

d = a document in the collection,

D = a collection of documents

TF=Term Frequency given by:

$$tf(t, d) = Nt, d Nd (2)$$

IDF= Inverse Document Frequency given by:

$$idf(t, D) = \log N |d \in D: t \in d| (3)$$

Where,

N = number of documents in the collection

Nd = Number of terms in the document d

Nt, d = Number of times term t appears in document d

The cosine similarity evaluates the cosine of the angle between two vectors in an inner product space to determine how similar they are. This is based on how closely a book matches a user's search or a book the user has previously enjoyed. Each weight in the vectors of weights used to represent the books indicates the strength of the relationship between the book and the term.

Similarity is measured by

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

Fig. 5. Cosine similarity formula

Where,

A_i, B_j is two different books

V. RESULT

In this paper, we develop a recommendation system that uses Cosine Similarity and content-based filtering to recommend books to users based on their interests. The code is executed in Google Colab using the NumPy and panda libraries. The dataset contains different books in different genres and it is classified and correctly recommended to users based on their interests.

```
genre_recommendations('Harry Potter and the Order of the Phoenix').head(20)
```

50	Harry Potter and the Sorcerer's Stone
87	Harry Potter and the Deathly Hallows
89	The Golden Compass
103	Harry Potter and the Prisoner of Azkaban
112	Harry Potter and the Goblet of Fire
127	Harry Potter and the Half-Blood Prince
155	Harry Potter and the Chamber of Secrets
211	His Dark Materials
288	Inkheart
392	The Amber Spyglass
409	The Subtle Knife
446	Harry Potter Series Box Set
914	Lirael
944	Abhorsen
957	The Harry Potter Collection 1-4
1683	The Amber Spyglass
1746	The Singing
1960	Gregor the Overlander Box Set
2310	Harry Potter and the Deathly Hallows
2433	La Belle Sauvage

Name: book_title, dtype: object

Fig. 6. Recommendation of books similar to 'Harry Potter and the Order of the Phoenix'

```
genre_recommendations('The Da Vinci Code').head(20)
```

461	The Lost Symbol
606	Angels & Demons
2765	Jack Reacher's Rules
1779	Double Impact: Never Say Die / No Way Back
2753	A Pound of Flesh
3660	The Next Right Thing: A Novel
71	The Girl with the Dragon Tattoo
187	The Girl Who Played with Fire
266	The Girl Who Kicked the Hornet's Nest
328	Gone Girl
3013	Gone Girl
3181	Touch & Go
3170	To Play the Fool
2593	Gideon
1673	I Am Watching You
2265	Immeiden kabinetti
856	The Bone Collector
3564	Unseen
333	A Time to Kill
3211	The Chalk Man

Name: book_title, dtype: object

Fig. 7. Recommendation of books similar to 'The Da Vinci Code'

VI. CONCLUSION

In the internet era, recommendation systems have emerged as the single most important component of an accurate and trustworthy information source.

Most recommendation systems try to anticipate the buyer's interests and then suggest books in that manner. To give recommendations to the intended consumers, this paper used a content-based filtering technique. The availability of the contents characterising the objects and users' interest profiles resulted to a solution to this issue. Although they rely on these contents, content-based approaches are not influenced by user ratings. In addition, an algorithm for offering or suggesting recommendations based on user queries is presented in this work. The approach uses the cosine similarity metric in addition to TF-IDF weighing.

The outcomes of the research demonstrate that the suggested strategy offers relevant suggestions. The presented work can be used to suggest other domains like movies, music, and other media in various sectors.

REFERENCES

- [1] Gaurav Srivastav, Sargam Maurya, Ramni Harbir Singh, Tanisha Tripathi, Tushar Narula: "Movie Recommendation System using Cosine Similarity and KNN"(2020)
- [2] Vikas Sindhwani and Prem Melville and: "Recommender System " (2019)
- [3] Sujoy Datta, Das, Laxman Sahoo and Debashis: "A survey on recommendation system." International Journal of Computer Applications 160.7 (2017)
- [4] Ayush Jain, Kaza Sai Vineeth, Abhiraj Biswas and Mohana: "Development of Product Recommendation Engine By Collaborative Filtering and Association Rule Mining Using Machine Learning Algorithms"(2020)
- [5] M. Chandrashekhhar, V. Subramaniaswamy, R. Logesh, Anirudh Challa and V. Vijayakumar: "A personalised movie recommendation system based on collaborative filtering."(2017)