iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45736

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Personalized Book Recommendation System using TF-IDF and KNN Hybrid

Rashika S[1], Namit S Gouranna[2], Nishanth Nayak T[3], Prajwal C R[4], Mr. Prashanth J[5]
[1, 2, 3, 4]*BNM Institute of Technology, Bangalore, Karnataka*
[5]*Assistant Professor, Department of Computer Science and Engineering, BNM Institute of Technology, Bangalore, Karnataka*

Abstract: ***A recommendation system helps an organization to create loyal customers and build trust by offering their desired products and services. These systems today are so powerful that they can handle the new customer too who has visited the site for the first time. With the increasing number of books, people prefer to use e-books. Today, online businesses have emerged that are dedicated only for e-books. They allow their users to purchase any books of their interest or even read them online. This improves their business targets. To make their users engaged, they use machine learning models that recommend users the books based on their preferences. Such a system is called Book Recommendation System. Over the past, a large number of book recommendation systems have been built, most of them are found to be useful for both the organization and the users, and are being put into use in the real world. In this proposed system, we build a Book Recommendation System which recommends a set of books to users based on their previous ratings and readings using content-based filtering, collaborative filtering, and hybrid filtering model. This will save users time in searching the books of their interest.***
Keywords: Hybrid recommendation, content-based filtering, collaborative filtering, personalized recommendation.

I. INTRODUCTION

Every recommender system comprises of two entities, one is user and other is item. A user can be any customer or consumer of any product or items, who get the suggestions. Input for recommendation algorithms can be a data in database of user and items and output obliviously will be the recommendations. As in our case, inputs consist of database of customer and database of books and output denotes the book recommendations.

There are many approaches for recommender systems used in the development of the machines through Content-based approach and Collaborative approach or Hybrid approach.

1) Collaborative Filtering Approach: Collaborative filtering approach uses collecting the data and analysing data based on the user' behaviours, preferences or activities and predicting what users will like based on their similarities with other users.
2) Content-based Approach: This approach is based on the item description and user profile. In a content-based recommender system, keywords are used to describe the items and a user profile is built to indicate the type of item this user likes.
3) Hybrid Recommender System: This system is combination of content based and collaborative filtering approach on the data.

Personalized recommendation system are seek to predict the preference based on the user's interest, behaviour and other information. Personalized recommendation is not only can provide the user needs, but also to help users explore and discover new hobbies. Now-a-days many book selling websites are available on the internet. Many of them are having their own recommendation system to recommend books to the buyers.

In this project, we build a Personalized Book Recommendation System which recommends a set of books to users based on their previous ratings and readings using content-based filtering, collaborative filtering, and hybrid filtering model. This will save users time in searching the books of their interest.

II. DATASET DESCRIPTION

This dataset contains ratings for ten thousand popular books. As to the source, let's say that these ratings were found on the internet. Generally, there are hundreds of reviews for each book, although some have less - fewer - ratings. Ratings go from one to five. Both book IDs of books and user IDs of user are contiguous. For books, they are one to ten thousand for user, one to fifty three thousand. All users have made at least two ratings. Median number of ratings per user is eight. There are also books marked to read by users, book metadata and tags.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

III. RELATED WORKS

As a ground work, we examined and studied about different kinds of existing recommendation engines and their uses and limitations. We have presented a concise version of our study here. [1] YougChangWang et al. gives the methodology used, that is PCA (Principle Component Analysis) and SVD (Singular Value Decomposition) for Dimensionality reduction. With the evolution of large-scale, complex and high dimension data, it is very much required to reduce this dimensionality. PCA is used to make the variance of the data distribution maximum, for this it uses Eigen values and Eigen vectors concept, We can also find the concept of SVD in this paper. [2] Bin Li et al. gives the methodology used, that is K-Nearest Neighbours (K-NN), where the Root Mean Squared Error (RMSE) value of the same is calculated. In experiments, the methods of Top-10 recommended mainly refer to the score on the basis of prediction. We recommend the items whose score is highest. The model extracts keyword information of items at first and then calculates their weights by using different methods like Term Frequency (TF), Document Frequency (DF) and Term Frequency - Inverse Document Frequency (TF-IDF). [3] Mohammed Fadhel Aljunid et al. uses the Alternating Least Square algorithm. The ALS algorithm uses least squares computation to minimize the estimation errors, and alternates between solving for product factors and solving for user factors. This paper also mentioned an improved ALS approach based on apache spark. This improved ALS model is proposed to be more efficient when compared with the existing ALS algorithm. Both the ALS and the improved ALS algorithm work in a similar fashion but the major difference between these two methods is that while ALS algorithm splits the entire dataset into test and train data only once, the Improved ALS model splits the dataset k times to reduce the Root Mean Squared Error (RMSE). [4] Muhammad Zuhdi Fikri Johari et al. uses the Indonesian Online Marketplace dataset to provide recommendations of relevant items to the users by using IMDB's weighted rating formula to generate the top-n products. This also proves that by using the user ratings and IMDB weighted rating formula and the demographic filtering method, top-n items of any category could be found. [5] N. Muthurasu et al. uses the Term-Frequency Inverse Document Frequency to provide recommendations according to the user's preferences. Each data record is converted into a vector by using the TF-IDF vectorization algorithm. For each vector, a similarity measure is computed using the cosine similarity method.

IV. PROPSED MODEL

Our proposed system is a hybrid book recommendation that uses IMDB weighted rating formula, cosine similarity algorithm and for recommending books. The main advantage of this system is that, the algorithm is designed to work efficiently even for a small set of data. The recommendations are based on the genre, authors, likes and dislikes of the user. It allows the users to save time in searching books. Better and more efficient recommendation systems also increase market reach and create a flux of recurring customers for the site.

A. IMDB weighted rating formula

The IMDb Weighted rating, which was formerly used for the calculation of the top-n ranking of films by IMDb, is used to calculate the top-n products from the book dataset. Mathematically IMDb's weighted rating formula is represented as follows:

$$\text{WR} = \frac{v}{v+m} \times R + \frac{m}{v+m} \times C$$

Where,

- *WR* is Weighted Rating.
- *v* is the number of ratings for the book.
- *m* is the minimum ratings required to be listed in the chart.
- *R* is the average rating of the book.
- *C* is the mean rating across the whole report.

The next step is to determine an appropriate value for *m*, the minimum ratings required to be listed in the chart. We will use 95th percentile as our cutoff. In other words, for a book to feature in the charts, it must have more ratings than at least 95% of the books in the list.

B. Cosine Similarity

This algorithm converts a text document as a vector of terms. By this model, the similarity between two dataset can be found by determining cosine value between two vectors. Application of this algorithm can be performed on any two texts such as documents, sentence or paragraph. In case of search engines, the similarity value between user query and documents are determined and then it is categorized from highest to lowest one. Higher the similarity score between the user query vector and document vector means more relevancy between query and document.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Similarity measurement between the user query and document should analyze the meaning of the term. Cosine similarity on the other hand still can't deal with the semantic meaning of the query very well. Semantic meaning problem does not meet the difference of syntax matching.

Mathematically cosine similarity formula is represented as follows:

$$(t_a, t_b) = \sum(t_a . t_b) / (\sum t_a . \sum t_b)$$

Where,

- t_a is a term as a key.
- t_b is the term comparison.

C. K-Nearest Neighbour

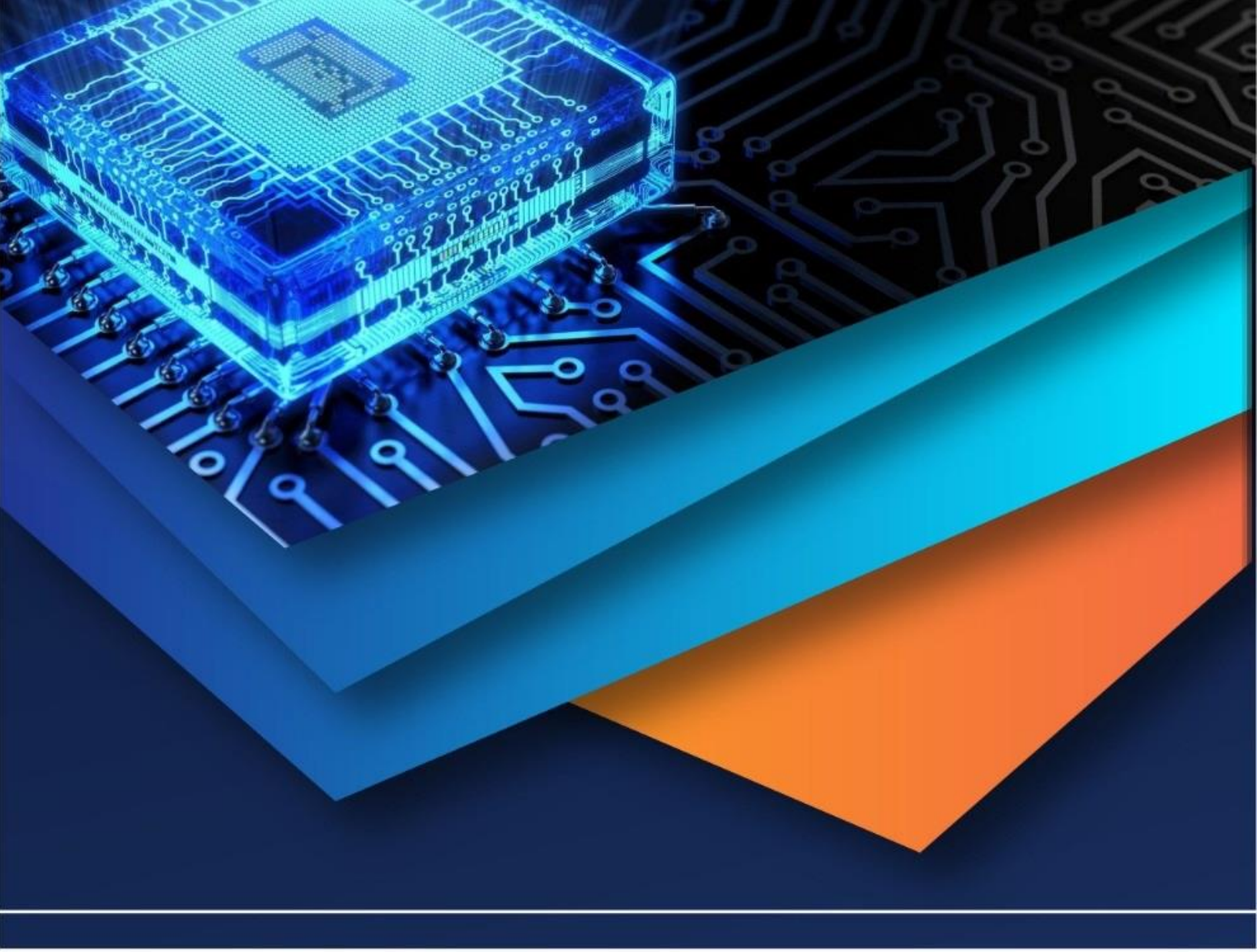
KNN implements an item based collaborative filtering, KNN is a perfect go-to model and also a very good baseline for recommender system development. KNN is a nonparametric, lazy learning method. It uses a dataset in which the data points are separated into several clusters to make inference for new samples. KNN does not come up with any assumptions on the underlying data distribution but it relies on item feature similarity. When KNN makes inference about a book, KNN will calculate the "distance" between the target book and every other book in its dataset, then it ranks its distances and returns the top K nearest neighbour books as the most similar book recommendations.

V. CONCLUSION

Recommender systems are an extremely potent tool utilized to assist the selection process easier for users. This paper has covered the personalized book recommendation using hybrid recommendation model. On the bases of this study, Hybrid approach of IMDB weighted rating formula, Cosine similarity algorithm and Alternating Least Square algorithm has been proposed in order to improve the efficiency of basic algorithm.

REFERENCES

[1] YongChangWang, Ligu Zhu, "Research and Implementation of SVD In Machine Learning", 16th International Conference on Computer and Information Science (ICIS), Wuhan, China, 2017.
[2] Bin Li, Hua Xia, Sailuo Wan, Fengshou Qian, "The Research for Recommendation System Based on Improved KNN Algorithm", 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 2020.
[3] Y. N. Bhagirathi, P. Kiran, "Book Recommendation System using KNN Algorithm", International Journal of Research in Engineering, Science and Management, 2019.
[4] Mohammed Fadhel Aljunid, D. H. Manjaiah, "An Improved ALS Recommendation Model Based on Apache Spark", ResearchGate, Kollam, India, 2018.
[5] Muhammed Johari, Arif Laksito, "The Hybrid Recommender System of the Indonesian Online Market Products using IMDb weight rating and TF-IDF", JURNAL RESTI, 2021.
[6] N Muthurasu, Nandhini Rengaraj, Kavitha Conjeevaram Mohan, "Movie Recommendation System Using Term Frequency-Inverse Document Frequency and Cosine Similarity Method", International Journal of Recent Technology and Engineering (IJRTE), 2019.
[7] Suad A. Alasadi and Wesam S. Bhaya, "Review of Data Preprocessing Techniques in Data Mining", Journal of Engineering and Applied Sciences, Babil, Iraq, 2017.
[8] Sandeep Matharia and C.N.S Murthy, "NOVA: Hybrid Book Recommendation Engine", Institute of Electrical and Electronics Engineers (IEEE), Indore, India 2012.
[9] Sunny Sharma, Vijay Rana, Manisha Malhotra, "Automatic recommendation system based on hybrid filtering algorithm", Springer, 2021.
[10] Salil Kanetkar, Akshay Nayak, Sridhar Swamy, Gresha Bhatia, "Web-based Personalized Hybrid Book Recommendation System", IEEE, 2014.
[11] Yassine Afoudi, Mohamed Lazaar, Mohammed Al Achhab, "Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network", ScienceDirect, 2021.
[12] Yonghong Tian, Bing Zheng, Yanfang Wang, Yue Zhang, Qi Wu, "College Library Personalized Recommendation System Based on Hybrid Recommendation Algorithm", ScienceDirect, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)