Effective Book Recommendation System using Machine Learning Algorithm

Sarangthem Nirupama Devi
School of Computer and Information Sciences
University of Hyderabad)
Hyderabad,India
20mcmt29@uohyd.ac.in

Sumit Thombare
School of Computer and Information Sciences
University of Hyderabad
Hyderabad,India
sumit.thombare07@gmail.com

***Abstract*—We saw an exponential rise in online activities like online movie watching, online book reading, online shopping etc. due, to the covid-19 pandemic. Recommending users with the correct item from the vast data is the biggest challenge for e-commerce websites. Some systems are biased in their rating technique because they only count the user's ratings from their databases, they are not using data of other users. Hence, the recommendation made by these systems is not precise, so people tend to use many sites to get personalised recommendations.**

We are proposing an effective personalized book recommendation system where we gather data from many sources such as Goodreads, Amazon, etc. based on the user rating and interest we will recommend the most related books using machine learning. We used the K-nearest neighbors clustering algorithm, cosine distance function to measure the distance and cosine similarity function to find similarity between clusters.

The model could also remove the boring books when the average Specificity is higher than sensitivity. We evaluated the accuracy of the model by comparing it with the actual data found that the model performs remarkably well. We are using web scraping to give the details where users can buy these books, so they don't need to go to different sites eventually will save the users time.

***Index Terms*—Machine learning,K-nearest neighbors clustering algorithm, cosine distance function,web scraping**

I. Introduction

With the exponential rise in online shopping due to Covid-19 pandemic, all the ecommerce websites are trying to make personalized recommendation in hopes of increasing their sales and to guide user to find good books. Huge increase in online book readers during Covid-19 pandemic. Recommendation system helps to finding relevant books from a vast e-book space which becomes a tremendous challenge. With the development in the field of data mining we can mine data from a vast amount of data available. This mined data sets can be used for machine learning.

Personal recommendation systems are used to mine related books based on user rating and interest. Most of these existing systems are user-based ratings where content-based and collaborative based learning methods are used.

To developed recommendation systems there are three main techniques commonly used. They are collaborative, content-based and hybrid algorithms. Firstly, in collaborative filtering technique users' interest on items information and opinions about the item are considered to make recommendation on the product. It can make predictions by collaborating information from many users which is based on users opinion. For example, collaborative filtering could make predictions about book based on the ratings of the user. Accuracy is only 88% and requires vast amount of data and doesn't consider domain dependencies .

Secondly, in content-based filtering method recommendation of items is based on the Similarity among articles. this method don't consider users' ratings which is also an important factors when suggesting new items like for recommending new books or journals. So, the content-based filtering has low accuracy prediction in term of books or journals recommendation. Thirdly, in hybrid-based filtering which combines two or more filtering techniques to produce the output. The accuiracy of prediction of hybrid filtering is better comparing to collaborative and content-based filtering.

This paper proposed an effective system for recommending books for online users using collaborative filtering techniques. First collected data frame which consist of three data set namely users info dataset , rating info dataset and book dataset. To ensure statistical significant output prediction, the unpopular books are removed. Only the top 1% of highest rating and popular book is considered which increases accuracy. We then compute Compressed sparse row matrix to calculate sparse matrix to increase efficiency of calculation. K- nearest neighbors and K-means Cosine distance function is used for recommending the books. K- nearest neighnors an unsupervised machine learning algorithm which finds similarity of the user input book to suggest or recommend new books. The proposed system used the K-means Cosine Distance function to measure distance and Cosine Similarity function to find Similarity between the book clusters. We calculate Pearson's R correlation coefficient and compared the result with our system which shows high accuracy between 0.9 to 1. The result concludes that recommendations, based on a particular book, are more accurately effective than a user-based recommendation system

Identify applicable funding agency here. If none, delete this.

II. METHODOLOGY

In this section we present our method of clustering by using maching learning to recommend books to online user. As depicted in fig. 1 our process consists of three main modules. The first one data acquisition where we collect book data set, user dataset and rating of books dataset from vast data set and merging the dataset. Second module consist of calculating rating of all the books and considering only the top 1% high rated books. Then compressed sparse matrix is calculated for efficient calculation. Third module K-nearest neighbors clustering is applied along with cosine function to calculated accurate prediction to recommend. finally, book is recommented.

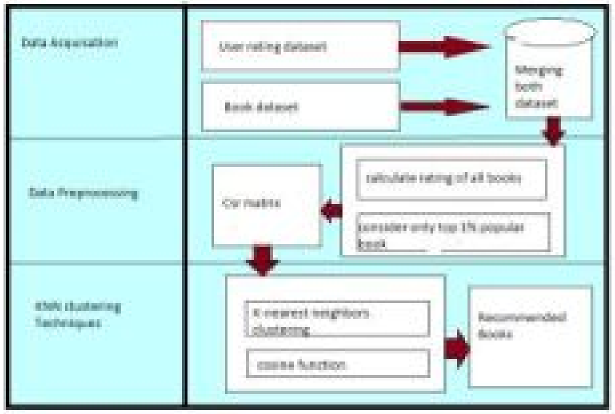

Fig. 1. diagram for purposed method

A. DATA Acquisition

We used Book-Crossings dataset which is a book ratings dataset. The dataset content of three tables that is ratings, books info, and user's info. The ratings are in range from 0 to 10. The rating data set contents a list of ratings given by users to the books available. It includes three fields namely userID, ISBN, and book Rating. The book data set gives book details like ISBN, book title, book author, publisher and so on up to 8 fields. The user's info data set includes three fields namely user id, location, and age.

B. Data pre-processing

To find real and reliable outcome of recommendation we consider the popular books rather than unnecessary books. We can find out the popular books by combining book data with the rating dating and choosing book data with good ratings. we consider only top 1

Compressed Sparse Row (CSR) Matrix for Machine learning Many users don't read many of the books so the previously consider book matrix contain many sparse or zero values. Large sparse matrices when applied in machine learning increases computational cost and degrade performance. So sparse matrices use encoding schemes which is used for data preparation. For our study we use Machine learning libraries that takes NumPy data structure which can operate transparently on SciPy sparse arrays i.e., scikit-learn. Csr matrix() function is used to convert dense matrix stored in NumPy array which contains missing values to sparse matrix by storing missing values with zeros. We transform the values of the matrix dataframe into a scipy sparse matrix for more efficient calculations.

C. K-Nearest Neighbours (KNN) Clustering

K- nearest neighbors algorithm is a supervised machine learning algorithm which can be used for both classification and regression problems. In our study we used KNN to find clusters of similar users based on common book ratings, and make predictions using the average rating of top k nearest neighbors.

To find the nearest neighbors, we use KNN algorithms with sklearn.neighbors. The algorithm we use to compute the nearest neighbors is "brute", and we specify "metric=cosine".the metric cosine will calculate the cosine similarity between rating vectors. After that we fit in the model.

we will described how recommendation is done. The kNN algorithm measures distance to determine the "closeness" among the instances. After the instance's closeness are determine we group them depending on their closeness distance. When a user search for a book, we can set book recommendation by selecting the most popular class among the neighbors.

III. Results

We select a book and calculate the Pearson's R correlation coefficient for every book pair in our final matrix. The result obtained is compared with the result obtained from applying KNN algorithm by selecting same book. We found that there is high correlation between these two i.e., 0.9 to 1.0.

IV. Conclusion

This study used K- nearest neighbors clustering algorithms to increase the prediction accuracy of the recommendation system. The datasets were collected from the Book-crossings books repository of Cai-Nicolas Ziegler. About 2173 books were processed by using machine learning algorithms (k-NN clustering and cosine function). The results show that our proposed system can remove boring books from the recommendation list more efficiently. In future we can improved our accuracy by using other filtering like hybrid based filtering .

References

[1] A. Gazdar and L. Hidri, "A new similarity measure for collaborative filtering based recommender systems," Knowledge-Based Systems, vol. 188, p. 105058, 2020.

[2] S. B. Shirude and S. R. Kolhe, "Improved Hybrid Approach of Filtering Using Classified Library Resources in Recommender System," in Intelligent Computing Paradigm: Recent Trends: Springer, 2020, pp. 1-10

[3] A. Gholami, Y. Forghani, and M. Branch, "Improving Multi-class CoClustering-Based Collaborative Recommendation Using Item Tags Improving Multi-class Co-Clustering-Based Collaborative Recommendation Using Item Tags."