Book Recommendation System

Shantanu Houzwala
Data science trainee,
AlmaBetter, Bangalore

Abstract:

A recommendation system is one of the top applications of data science. Every consumer Internet company requires a recommendation system like Netflix, Youtube, a news feed, etc. What you want to show out of a huge range of items is a recommendation system. A recommendation engine is a class of machine learning which offers relevant suggestions to the customer. Before the recommendation system, the major tendency to buy was to take a suggestion from friends. But Now Google knows what news you will read, Youtube knows what type of videos you will watch based on your search history, watch history, or purchase history.

1.Problem Statement

Recommender systems are algorithms aimed at suggesting relevant items to users (movies, books, products). Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or can also be a way to stand out significantly from competitors.
In this project, we create a book recommendation system for users through Unsupervised Machine Learning using three datasets - Books.csv, Users.csv, Ratings.csv

2. Introduction

A recommendation system helps an organization to create loyal customers and build trust by them desired products and services for which they came on your site. The recommendation system today are so powerful that they can handle the new customer too who has visited the site for the first time. They recommend the products which are currently trending or highly rated and they can also recommend the products which bring maximum profit to the company.

3. Types of Recommendation System

A. Popularity Based Recommender

It is a type of recommendation system which works on the principle of popularity and or anything which is in trend. These systems check about the books which are in trend or are most popular among the users and directly recommend them.

B. Content-Based Filtering

The algorithm recommends a product that is similar to those which used as watched. In simple words, In this algorithm, we try to find item look alike. For example, a person likes to watch Sachin Tendulkar shots, so he may like watching Ricky Ponting shots too because the two videos have similar tags and similar categories.

C. Collaborative-based Filtering

Collaborative based filtering recommender systems are based on past interactions of users and target items. In simple words here, we try to search for the look-alike customers and offer products based on what his or her lookalike has chosen. Let us understand with an example. X and Y are two similar users and X user has watched A, B, and C movie. And Y user has watched B, C, and D movie then we will recommend A movie to Y user and D movie to X user.

D. Hybrid Filtering Method

It is basically a combination of both the above methods. It is a too complex model which recommends product based on your history as well based on similar users like you.

There are some organizations that use this method like Facebook which shows news which is important for you and for others also in your network and the same is used by Linkedin too.

4. How Book Recommender System works?

A book recommendation system is a type of recommendation system where we have to recommend similar books to the reader based on his interest. The books recommendation system is used by online websites which provide ebooks like google play books, open library, good Read's, etc.

In this project I have used Popularity Based approach and Collaborative Based approach.

5. Understanding the Data:

The Book-Crossing dataset comprises 3 files:

- **Users**

Contains the users. User IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.

- **Books**

Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. In the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-S,

Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon website.

- **Ratings**

  Contains the book rating information. Ratings (Book-Rating) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

6. Steps involved:

- **Exploratory Data Analysis**

  The primary goal of EDA is to support the analysis of data prior to making any conclusions. It may aid in the detection of apparent errors, as well as a deeper understanding of data patterns, the detection of outliers or anomalous events, and the discovery of interesting relationships between variables**.**

- **Data Cleaning**

  In the data cleaning section, we drop off the features which do not contribute towards making a good recommendation system. The Image-URL-S and Image-URL-L do not contribute towards our final goal. So, we will drop off these columns and also dropping these columns we will not be losing any information.

- **Feature Engineering**

  From the analysis part we get that the Year-Of-publication was wrongly mentioned for some of the rows. There are some incorrect entries in the Year-Of-Publication field. Publisher names 'DK Publishing Inc' and 'Gallimard' have been incorrectly loaded as Year-Of-Publication in the dataset due to some errors in the csv file.

  Also, since these entries were few, we could manually correct the column values for these particular rows.

- **Recommender System**

  For building recommendation system I used two algorithms:

1. **Popularity Based Recommendation**
2. **Collaborative Based Filtering**

7. Algorithms:

1. **Popularity Based Recommendation:**
   It is a type of recommendation system which works on the principle of popularity and or anything which is in trend. These systems check about the books which are in trend or are most popular among the users and directly recommend them.
   For example, if a product is often purchased by most people then the system will get to know that that product is most popular so for every new user who just signed it, the system will recommend that product to that user also and chances become high that the new user will also purchase that. Using this popularity metric we can calculate the top books that could be recommended to a user.
   The **Harry Potter series** was the most popular book series. Other than that, books like **To Kill A Mockingbird**, **The Lord of the Rings** and **Dune** were among the top books.

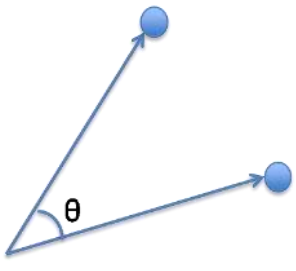
2. Collaborative Based Filtering:

It is considered to be one of the very smart recommender systems that work on the similarity between different users and also items that are widely used as an e-commerce website and also online movie websites. It checks about the taste of similar users and makes recommendations.

The similarity is not restricted to the taste of the user, moreover there can be consideration of similarity between different items also. The system will give more efficient recommendations if we have a large volume of information about users and items

In Data Mining, similarity measure refers to distance with dimensions representing features of the data object, in a dataset. If this distance is less, there will be a high degree of similarity, but when the distance is large, there will be a low degree of similarity.
In this project I have used **'Cosine Similarity'** which is a popular similarity measure. **Cosine similarity** is a metric, helpful in determining, how similar the data objects are irrespective of their size. In cosine similarity, data objects in a dataset are treated as a vector.

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



8. Conclusion:

That's it! We reached the end of our exercise.
Starting with loading the data so far we have done EDA, data cleaning, feature engineering and then building recommendation system.
The recommendation system works well and recommends accurately.

Given more information regarding the books dataset, namely features like Genre, Description etc., we could implement a content-filtering based recommendation system and compare the results with the existing collaborative-filtering based system.

We would like to explore various clustering approaches for clustering the users based on Age, Location etc., and then implement voting algorithms to recommend items to the user depending on the cluster into which it belongs.

References-

1. Towards Data Science
2. GeeksforGeeks
3. Analytics Vidhya