LIBRARY BOOK RECOMMENDATION SYSTEM ANALYSIS

Introduction:

Recommendation system is a software program that helps students to find products according to their needs and interests by using the student’s rating of each item and the student’s preferences. A recommender system works as a helper in finding relevant and related items based either on their explicitly mentioned preferences or objective behaviours. This way, it is a big source of reducing information overload in finding relevant items in several domains including books.

In order to recommend items including books to the student. The book recommendation system would help the student to borrow a book, by recommending books based on collaborative filtering and k-nearest neighbor classification algorithm.

Motivation:

· What is the Problem?

Library Book Recommendation System

· Why is the Problem?

Due to wide application of management system and information density, information data grows rapidly day by day. On one hand, students have a large number of information resources. On the other hand, the time cost and difficulty of student in finding the proper information increases. To tackle the problems, book recommendation is one of the solutions for college libraries which possess huge volumes of books and reading-intensive users.

**Preliminaries:**

This section summarizes related work and briefly defines the fundamental concepts needed to facilitate the presentation of the proposed algorithm.

A. Related work:

Different models have been developed in order to generate book recommendation. Many approaches rely on collaborative filtering (CF) methods based on the main idea that people have similar preferences and interests, so similarities of users or books are calculated. Basically, each user gives rating scores for a list of items and these scores are used to predict the rating active user.

B. User-based collaborative filtering:

Collaborative Filtering algorithm is based on the main idea that people have similar preferences and interests. One user’s behavior is compared with other user’s behavior and his/her nearest neighbors, and according to his/her neighbor’s preferences or interest to predict his/her preferences or interest.

Suppose that U = {u1, u2, ..., um} is a list of m users and I = {i1, i2, ..., in} is a list of n items. Each user Ui gives rating scores for a list of items Iui. The prediction problem is to predict the rating active user Ua will give to an item Iua from the set of all items that Ua has not yet been rated. The CF technique composes of 3 steps as follows: 1) users similarity calculation 2) top N nearest neighbors selection(knn) and 3) prediction.

1. Similarity

2. k-nearest neighbour classification

3. Recommendation

Area of Application:

Education -

* As we have said this is mainly for Libraries in college or universities
* This can be used in other places like public library or book stores.
* Book stores it will suggest the books for the new reader by getting to know the interest of that buyer.

Objective:

To recommend user’s books of their interest. The recommendation or suggestion being accurate to the user’s interest. Recommendation systems are software programs that help a student to find products according to their needs and interests.

Literature Survey:

Recommender systems have become a vital research field since the emergence of the first paper on collaborative filtering in the mid- 1990s. In general, these systems are stated as the support systems which help users to find content, products, or services (such as books, movies, music, TV programs, and websites) by gathering and examining suggestions from other users, which means reviews from various establishments, and users. These systems are broadly classified into collaborative filtering (CF) and content-based filtering (CB). CF is an information filtering practice that is based on the user’s evaluation of items or previous purchases records. However, this method has been known to expose two major issues that are sparsity problem and scalability problem. CB examines a set of items rated by an individual user and then uses the content of these items, as well as the provided ratings, to deduce a profile that can be used to recommend additional items of interest. However, the syntactic nature of Content based filtering to detect the similarity between items that share the same attributes or features causes overspecialized recommendations that only comprise very similar items to those the user already knows.

Analytical Research: The researcher has to use facts/data that are already available and analyse these to make a critical decision or evaluation of the material

Reference

|  |  |  |  |
| --- | --- | --- | --- |
| Sl. No | Title of the  Paper Authors | Details of  Publication | Summary of the Paper |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Methods of Data Collection:

The method used for data collection is Secondary Data Collection

When the data is collected by someone else for a purpose other than the researcher’s current project and has already undergone the statistical analysis is called as Secondary Data. Features of Secondary Data:

• Secondary data are published data.

• They give the latest information.

• They can be easily collected from various internal and external sources.

• They are relatively cheaper; they need less effort, time, and money.

• They have been collected by other people for their own problems and situations in the past.

• They are used as a supplementary to primary data. Mostly, they are used for defining and understanding problems.

• The use of secondary data is optional. Research can be conducted even without the use of this type of data.

The secondary data is readily available from other sources and there are no specific collection methods. The researcher can obtain data from the sources both internal and external to the organization which may include censuse, information collected by government departments, organizational records and data that was originally collected for other research purposes.

Data Set-

The dataset of books is collected from IIF (Institute for Informatik Freiburg).

The dataset comprises of 3 tables.

Books:

The original dataset contains 2,71,361 number of books. After data preprocessing, the number of books that are available are 86,581 books. The table is having features ISBN, bookTitle, bookAuthor, yearofpublication, publisher.

Users:

The original dataset contains 2,78,858 number of users. Some users have rated the books, some have not.

Ratings:

The original dataset contains 10,31,136 ratings. The ratings are connected to userID and ISBN. A specific book having specific ISBN rated by single or multiple users. The ratings can be implicit or explicit.

Classification Technique is used to give or generate the ratings for the books. Classification according to class intervals to generate the ratings ranging from 0 to 10.

Data Cleaning: The extreme values in the dataset are removed before the processing of the data. To consider the books for recommendation, we have taken the specification that, the person rating the book should have rated more than one book. And the book must be rated by more than one user.

By doing this, the list of books will be reduced based on the conditions. Only the books which are satisfying the conditions are taken into consideration.

Tools used:

· Anaconda

It is a free and open distribution of the Python and R languages for scientific computing, that aims to simplify package management and deployment. It comes with more than 1,500 packages as well as the conda package and virtual environment manager. It also includes a GUI Anaconda Navigator, as a graphical alternative to the command line interface.

· PyCharm

It is an IDE used for computer programming, specifically for Python language. It is developed by Czech company JetBrains. It supports wed development with Django as well as Data Science with Anaconda.

· VSCode

It is an IDE developed by Microsoft. It is based on Electron, a framework which is used to deploy Node.js applications for the desktop running on the Bink layout engine.

· MS-Excel

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications.

**Technology used**

· Python

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python’s elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

· Scikit-learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. Scikit-learn was initially developed by David Cournapeau as a Google summer of code project in 2007.

KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. However, it is more widely used in classification problems in the industry.

Cosine similarity

Cosine similarity is the cosine of the angle between two *n*-dimensional vectors in an *n*-dimensional space. It is the dot product of the two vectors divided by the product of the two vectors' lengths (or magnitudes). Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

· SciPy

SciPy is an Open Source Python-based library, which is used in mathematics, scientific computing, Engineering, and technical computing. SciPy is the most used Scientific library only second to GNU Scientific Library for C/C++ or MATLAB’s.

· Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

Histogram

A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

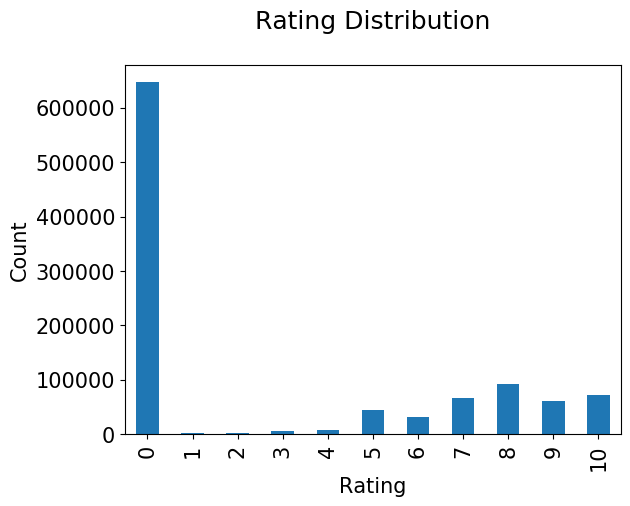
· Pandas

It is software library written for Python language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series.

Dataframes

Pandas allow importing data of various file formats such as csv, excel etc. Pandas allows various data manipulation operations such as group by, join, merge, melt, concatenation as well as data cleaning features such as filling, replacing or imputing null values.

Implementation:



This Histogram is used to represent the distribution of a single attribute. In this image the Rating distribution in the Rating table is represented. The x axis represents the Rating from 0 to 10. The Y axis represents the count of ratings.

In this graph the books which are not rated i.e, having rating 0 are seen having count more than 6,00,000. This helps usse for Data Adjustment.

Result and Conclusion:

This system aims to provide personalized recommendation of books to the students. This system considers big data of books. The system makes use of collaborative filtering algorithm and knn classification algorithm to provides the user with recommendation list. The system tries to predict the ranking by considering the item's similarity as well as user's similarity so that a user can get recommendations for new books.

References in IEEE format:

[1] X. Su and T. M. Khoshgoftaar, “A survey of collaborative filtering techniques,” Advances in Artificial Intelligence archive, 2009

[2] L. Xin, E. Haihong, S. Junde, S. Meina, and T. Junjie, “Collaborative book recommendation based on readers’ borrowing records,” In Proceeding of the 2013 International Conference on Advanced Cloud ang Big Data, Dec. 2013, pp. 159-163.

[3] Y. Koren, R. Bell, and C. Volinsky, “Matrix Factorization techniques for recommender systems,” Computer, vol. 42, Aug. 2009, pp. 42-49

[4] H. Polat and W. Du, “SVD-based collaborative filtering with privacy,” ACM Symposium on Allied Computing, Mar. 2005, pp. 142-146

[5] 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, A Collaborative Filtering Based Library Book Recommendation System.