Intelligent Book Recommendation System Based on Collaborative Filtering and Association Rule Mining

Aditya G. Tate
Student, CSE Department, DES’sCOET Dhamangaon Rly, India.

Yogesh R.Gupta
Student, CSE Department, DES’sCOET Dhamangaon Rly, India.

Sharedchandra P. Kawane
Student, CSE Department, DES’sCOET Dhamangaon Rly, India.

Abstract – The increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets grow in size and complexity, direct “hands-on” data analysis has increasingly been augmented with indirect, automated data processing. Over the years, libraries in universities and other educational institutions have gathered a lot of data on books borrowed by students, yet the valuable knowledge embedded in these data has remained untapped. In many cases, students do not find required books in the library or probably the books have been borrowed by some other students. There are even a lot of books that have never been read by students. In many other cases, library management are faced with the challenge of what book to buy that would maximally benefit the students, and also how to place these books in shelves. There is therefore an urgent need for systems that can help the library management make informed decisions so as to address these issues. This paper presented a book recommender system that mines frequently hidden and useful patterns from the book library records and make recommendations based on the pattern generated using associated rule mining technique.

Index Terms – Association rule, Classification, Recommendation system, User based collaborative filtering.

1. INTRODUCTION

With the increasing of universities' library resources, and comparing to the past passive learning, active learning becomes a mainstream trend now. However, facing that so many library resources, how to choose resources becomes a major problem of college students. Effecting of universities' library is not only providing students with books, but also providing realistic books to students, and helping students to rational use the library resources. The concept of recommendation system in the 90s of last century is proposed by the United States firstly. After several years of researching, the recommendation system gradually began to use in the commercial areas. And now recommendation system is almost everywhere. It is estimated that the recommendation system improved 20% of product sales at least. In the book recommendation system, the most representative of the site is Amazon Books. The recommendation system is an intelligent platform built on the massive data mining to help site's users to solve them personalize decision support and information services. This paper presents a book recommendation system, introduced the factors of college, grades, books' borrowing records and so on. And it realized a book recommendation system based on the user's collaborative filtering algorithm and Association rules.

Basically this paper explains the new approach for recommending books to the students and academician. Here the features of classification are considered to classify the books of proper price range and authors. The features of user based collaborative filtering and association rule are considered for selecting the highly rated top N- books. Thus the combine feature of classification, user based collaborative filtering and association rule mining has considered for recommending books to buyers. The rest of the paper is organized as follows. Section 1 describes classification; Section 2 describes user based collaborative filtering; and also describes major point of association rule mining; discusses our novel approach for recommending books to the college students according to price range and publishers.

1. CLASSIFICATION

Classification techniques provide a way to extract rules and patterns from data that can be used for predictions. Classification techniques are used to classify data records into one among a set of predefined classes. In classification, analysis of a particular data set is considered, to generate a set of grouping rules which can be used to classify future data.

Classification is basically, one of the data mining fields that helps in interference learning, knowledge mining, prediction and decision making. A learning set is developed, in classification cases which results a classifier algorithm. Each sample in learning set contains some variable fields and a class number which specifies its class. After completing the learning and test phase, classifier gives the predicted class of incoming
new data. In this paper we used classification technique; We have defined three classes for publishers and declare the variable Pb. Three classes are Reputed with Pb value 0.5, Average with Pb value 0.4 and New with Pb value 0.2.

II. USER BASED COLLABORATIVE FILTERING

Collaborative filtering (CF) provides a way to do recommendation on the web. Collaborative filtering creates a database of preferences for items by users. In the design of the user preference data set, different attributes are considered to describe the various degrees of detail of user’s profile. User based collaborative filtering is a method of personalized recommendations from a dataset, to a user based on similarity between the interest profile of a user and those of other users. In other words, it can be explained as a method to predict the ratings for a product by a target user with the help of other similar user’s rating behavior and buying behavior. Basically CF recommendation system finds the aggregation of nearest neighbor. Suppose there are k users who are nearest similar to target user these users will be considered as k nearest neighbors. There are two ways to calculate the similarity: correlation and cosine. This paper takes correlation method (Pearson’s correlation) to calculate the similarity of users [3]. Pearson’s correlation, as following formula, measures the linear correlation between two vectors of ratings.

\[
Sim(m,n) = \frac{\sum_{a \in I_{nm}} (R_{m,a} - A_m)(R_{n,a} - A_n)}{\sqrt{\sum_{a \in I_{nm}} (R_{m,a} - A_m)^2 \sum_{a \in I_{nm}} (R_{n,a} - A_n)^2}}
\]

(1) Where \(R_{m,a}\) is the rating of the item a by user m, \(A_m\) is the average rating of user m for all co-rated items, and \(I_{nm}\) is the items set both rating by user m and user n. Select all the neighbors who will be considered as recommenders. For the selection of neighbors, Threshold based selection technique is used. According to this technique, users whose similarity exceeds a certain threshold value are considered as neighbors of the target user. The rating \(P(x,k)\) of the target user x to the target item k is as following:

\[
P_{x,k} = A_x + \frac{\sum_{m=1}^{c} (R_{m,k} - A_m)^2 Sim(x,m)}{\sum_{m=1}^{c} Sim(x,m)}
\]

(2) Where \(A_x\) is the average rating of the target user x to the items, \(R_{m,k}\) is the rating of the neighbor user m to the target item k, \(A_m\) is the average rating of the neighbor user m to the items, Sim(x,m) is the similarity of the target user x and the neighbor user m, and c is the number of the neighbors.

2. RELATED WORK

Association rule mining

Association is the discovery of association relationships or correlations among a set of items. This problem was introduced in. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Over the last several years, the problem of efficiently generating large numbers of association rules has been an active research topic in the data mining community. Mining for association rules can help in business decision making, and the development of customized marketing programs and strategies. Generally according to, an association rule mining algorithm contains the following steps. The set of candidate k itemsets is generated by 1-extensions of the large (k 1) item sets generated in the previous iteration.

- Supports for the candidate k item sets are generated by a pass over the database.
- Item sets that do not have the minimum support are discarded and the remaining item sets are called large k item sets.

This process is repeated until no larger item sets are found.

Data mining is an integral part of Knowledge Discovery in Databases (KDD), which is the overall process of converting a series of transformation steps, from data preprocessing to post-processing of data mining results. Figure 1 below shows the process of knowledge discovery in databases.

Figure 1. Process of Knowledge discovery

2.0.1 Overview of the Collaborative Filtering Process

The goal of a collaborative filtering algorithm is to suggest new items or to predict the utility of a certain item for a particular user based on the user’s previous likings and the opinions of other like-minded users. The user has expressed his/her opinions about. Opinions can be explicitly given by the user as a rating score, generally within a certain numerical scale, or can be implicitly derived from purchase records, by analyzing. Memory-based Collaborative Filtering Algorithms.
Memory-based algorithms utilize the entire user-item database to generate a prediction. These systems employ statistical techniques to find a set of users, known as neighbors, that have a history of agreeing with the target user (i.e., they either rate different items similarly or they tend to buy similar set of items). Once a neighborhood of users is formed, these systems use different algorithms to combine the preferences of neighbors to produce a prediction or top-N recommendation for the active user. The techniques, also known as nearest-neighbor or user-based collaborative filtering, are more popular and widely used in practice.

Model-based Collaborative Filtering Algorithms. Model-based collaborative filtering algorithms provide item recommendation by first developing a model of user ratings. Algorithms in this category take a probabilistic approach and envision the collaborative filtering process as computing the expected value of a user prediction, given his/her ratings on other items. The model building process is performed by different machine learning algorithms such as Bayesian network, clustering, and rule-based approaches. The Bayesian network model formulates a probabilistic model for collaborative filtering problem. Clustering model treats collaborative filtering as a classification problem and works by clustering similar users in same class and estimating the probability that a particular user is in a particular class C, and from there computes the conditional probability of ratings. The rule-based approach applies association rule discovery algorithms to find association between co-purchased items and then generates item recommendation based on the strength of the association between items.

2.0.2 Challenges of User-based Collaborative Filtering Algorithms

User-based collaborative filtering systems have been very successful in past, but their widespread use has revealed some potential challenges such as:

In practice, many commercial recommender systems are used to evaluate large item sets (e.g., Amazon.com recommends books and CDnow.com recommends music albums). In these systems, even active users may have purchased well under 1% of the items (1% of 2 million books is 20,000 books). Accordingly, a recommender system based on nearest neighbor algorithms may be unable to make any item recommendations for a particular user. As a result the accuracy of recommendations may be poor.

Scalability. Nearest neighbor algorithms require computation that grows with both the number of users and the number of items. With millions of users and items, a typical web-based recommender system running existing algorithms will suffer serious scalability problems.

Purpose of this book recommendation system is to recommend books to the student according to their comfort price range and publishers. This recommendation system stores recommendation in student’s web profile and works when user remains offline. This system has following steps:

1. From buyer profile records find out the category of the book that the buyer has bought earlier like CSE, ELECTRICAL, ECE, CIVIL etc.
2. Find out the subcategory of the book if there is any in the step 1 found category like for CSE subcategory will be C, C++, Data structure, Operating System etc.
3. Extract the profile of all buyers whose year of joining is less than target user’s year of joining but not lesser than 5 years (senior students of target user). From the transaction database find all those transactions done by the above buyers and whose category and subcategory (if there is any) is same as found in step 1 and step 2. Apply association rule on those transactions and find out the books that the target buyer can buy afterward. Adjust the support and confidence parameters to get the stronger rules. This step gives the name of the subject which user is expected to buy and also gives the list of all books of given subject.
4. With the help of classification techniques find out the class of publisher’s name and price range for each book, present in the List of books (output of step 3). The output of this step publisher value(Pb) and price(Pr) for each book.
5. Perform user based collaborative filtering to find the rating(R) of the target user to the books found in step 3. For selecting the nearest neighbors to the target user, consider the profile of all buyers whose year of joining is less than the target user’s year of joining but not lesser than 5 years (senior students of target user). This step will give the predicted rating(R) by target user to the books (list of books from output of step 3).
6. Now, apply mathematical formula to get list of books for recommendation.

\[ D = \frac{10}{p_r} + P_b + R \]

Where, \( p_r \) implies price of the book, \( P_b \) implies the publisher value, \( R \) implies the predicted rating for this book. \( D \) is a variable. We, recommend top 5 books according to the descending values of \( D \). Book having highest value with \( D \) will be recommended first. From table 1, we can see values of \( D \) for different books and Book_1 will be placed in first rank for recommendation. Similarly this calculation can be done for many numbers of books. This book recommendation system is represented by block diagram.

**Fig. Block diagram of Book recommendation system**

<table>
<thead>
<tr>
<th>Name of book</th>
<th>Predicted rating(( R ))</th>
<th>Price (( P_r ))</th>
<th>Publisher value (( P_b ))</th>
<th>( D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book_1</td>
<td>5</td>
<td>500</td>
<td>0.5</td>
<td>5.52</td>
</tr>
<tr>
<td>Book_2</td>
<td>5</td>
<td>500</td>
<td>0.4</td>
<td>5.43</td>
</tr>
<tr>
<td>Book_3</td>
<td>5</td>
<td>600</td>
<td>0.2</td>
<td>5.22</td>
</tr>
</tbody>
</table>

**REFERENCES**


[8] Qiang Ding , Qin Ding, William Perrizo,”Decision Tree Classification of Spatial Data Streams Using Peano Count Trees”, SAC 02:413-4


4. CONCLUSION

This study focused on the design and implementation of a Book Recommender System for the Library using data mining techniques. This paper proposes a book intelligent commendation system based on user and association rules, uses MATLAB to data preprocessing and calculation part of the collaborative filtering algorithm the users. In future, we proposed that software engineers should be able to develop a system that not only recommends but also prompts a user as soon as there is a new recommendation based on the user former preference, and also a system that not only recommends books, but all materials in the library, and finally, a hybrid system of association and classification algorithms in building a recommender system. The future research work will set a more reasonable association rule and its threshold, enhance the accuracy of recommendation system meet the more personalized service of digital library.