Book Recommendation System Using Collaborative Filtering

Aniqa Zaida Khanom
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Sheikh Mastura Farzana
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Sushmita Roy Tithi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

***Abstract*- The online recommendation system is widely used in today's world for various purposes. Many e-commerce websites give relevant suggestions to the end users based on this recommendation system. This is a powerful way of helping people to find the things they required. The proposed book recommendation system uses collaborative filtering to recommend books to the end user based on book ratings of the user and similar books liked by other users who have rated those books already. Additionally, the system will analyze the user behavior by using the features of various recommendation techniques such as content based, collaborative and demographic. This paper explores a few recommendation approaches and demonstrates a comparison between them in order to provide users with the satisfaction of the best and efficient books recommendations.**

***Keywords*- Book recommendation, collaborative filtering, Pearson's Correlation Coefficient, k-NN algorithm, matrix factorization.**

I. INTRODUCTION

Recommendation systems generate results to the end users in the form of recommendations. It can be implemented in any field of e-commerce sector to make the best choice among plenty of options. Additionally, with the increment of the use of internet in the present world, recommendation systems have become important and is widely used in various aspects in our everyday life. Recommendation systems provide benefit to both the consumer and the manufacturer, by suggesting items to consumers, which can't be demanded until the recommendations [1]. In order to help an individual to find a book of their interest, this recommendation system will suggest books based on user ratings and choices of books that they read and rated before. Every time a user rates a new book, it is added to the dataset, making the system a dynamic one.

Book recommender systems today require specialized expertise in analytics, machine learning and software engineering [2]. In this paper, some basic fundamental techniques and implementations in Python was covered. Later, more sophisticated methods like content-based filtering and collaborative based filtering were applied and compared to the previous approaches. The techniques used in this paper include Rating Count based, Content based and Demographic based systems using k-NN algorithm, Pearson's Correlation Coefficient, and Collaborative Filtering.

In this paper, section two deals with the existing systems, section three talks about the methodology; about the dataset, the feature extraction and the algorithms used. Section four talks about the results of the implementations of different techniques. Lastly, the conclusion talks about the future works and improvements that can be made on this project.

II. EXISTING SYSTEM

Existing recommender systems almost exclusively utilize a form of computerized match-making called collaborative filtering [3, 4]. The system maintains a database of the ratings of individual books by individual users, finds other users whose known preferences correlate significantly with the given patron, and recommends to a person other items enjoyed by his or her matched patron [5]. This approach assumes that a given user's tastes are generally the same as another user of the system and that a sufficient number off user ratings are available. Apart from that, in content based recommendation engines, system generates recommendations from source based on the features associated with products and the user's information [7]

III. METHODOLOGY

A. Dataset

The dataset being used in this paper is named Book Crossing Dataset. It was originally collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. The dataset contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings about 271,379 books. The dataset is divided into 3 parts, BX-Users, BX-Books, BX-Book-Rating. The first part contains user information such as location, age and UserID. The next part has 8 fields, ISBN, Book-Title, Book-Author, Year-Of Publication, Publisher, Image-URL-S, Image-URL-M, Image-URL-L which are all individual book related information. The last part of the dataset contains book ratings of individual

books by individual users. All ratings are done in a scale of 0-10. Since every user has not rated all books, NaN values appear frequently in the dataset. As input data, columns such as ISBN, UserID and rating is used. The output column demonstrates ISBN number of the recommended book which can be converted into Book name and author name using BX-Books part of the dataset.

B. Feature Extraction

The dataset contains more than 0.27 million users and 1.15 million books. Since all individual users have not rated all individual books, calculating recommendation based on all books or all users makes the system less efficient. Due to this reason books that have more than 200 ratings and users who have rated more than 100 books are considered to optimize the system. Additionally, variables such as Year-Of-Publication, Publisher, Image-URL-S, Image-URL-M and Image-URL-L has no relation with the books being recommended hence, these variables are excluded during system implementation. New features are extracted from the existing features. These features are age based user count. Another hybrid variable is total rating count based on all books that have more than 50 ratings and the users who rated them. In the demographic recommendation approach, only the users of Canada and US have been considered.

C. Algorithms Used

The fundamental algorithms that have been used in this project are discussed in this portion of the paper.

i. Pearson's Correlation Coefficient or Pearson's R: Pearson's Correlation Coefficient is the covariance of two variables divided by the product of their standard deviations [11]. The formula of Pearson's R :

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

In this case, the two variables X and Y represent the ratings of two books.

ii. K-Nearest Neighbors (k-NN): kNN is a machine learning algorithm to find clusters of similar users based on common book ratings, and make predictions using the average rating of top-k nearest neighbors [2]. In pattern recognition, the *k*-nearest neighbors algorithm (*k*-NN) is a non-parametric method used for classification and regression [12].

iii. Single Value Decomposition (SVD) is a model for identifying latent factors through Matrix Factorization. SVD is a factorization of a real or complex matrix. It is the generalization of the eigen decomposition of a positive semidefinite normal matrix to any matrix via an extension of the polar decomposition.

IV. RESULT

To start with, a basic fundamental recommendation system was built using Rating Counts. After implementing, it was noticed that since the books recommended are based on their previous ratings, the user might or might not like the book. It does not have any relation with the type of books liked by separate readers. That is, it recommends books based on which books have higher average ratings and does not put into consideration what type of reader the user is. So, it is not a very efficient program.

Afterwards, a recommendation system was applied using Pearson's Correlation Coefficient, or Pearson's R. In this method, there was a correlation established between two books for all cases. As an example, book recommendations were asked for readers who have read "The Lovely Bones: A Novel". The book is about a teenage girl who was killed and could see her parents and close ones dealing with her death. It was found that the books recommended were similar to the above referred book. They had similar plots surrounding crime and thriller. Hence, it can be said that this recommendation system was better than the previous one based on Rating Counts. However, it does not implement machine learning techniques and thus takes a longer time to process and is not very dynamic.

Thirdly, a system was built using k-Nearest Neighbors (k-NN) to apply collaborative filtering. Instead of using the whole dataset, this method was implemented on a set of popular books only. After data analysis, 2713 books were taken to build another data frame. Using this data frame, k-NN was implemented twice; the first time it was executed without any filters, and the second time using demographical locations. Both the systems give good recommendations. However, when the users were filtered and only users of USA and Canada were selected, it was observed that the system can give a better recommendation. This is mainly because the machine can train better with a specific set of data with more common variables, and ultimately give a better recommendation.

Lastly, collaborative filtering using Matrix Factorization was implemented. Matrix Factorization has a number of models for identifying latent factors. Among them. Singular Value Decomposition (SVD) was used for this project. After fitting the model into SVD through compression and dimensionality reduction, the Pearson's R was used to find the correlation between the books. Of all the implementations, Matrix Factorization gives the best results. While it takes a little more time compared to collaborative filtering using k-NN, it gives a more specific set of recommendations and is more accurate. Also, it is possible to build a very dynamic system using matrix factorization meaning every time a new book is rated it can be added to the dataset. The new additions will be brought into account based on correlation coefficient every time a book is recommended, making the system dynamic and flexible.

V. CONCLUSION

In conclusion, people are increasingly becoming internet dependent and prefer to find all their necessities across web. Amongst all the contents and products offered by various merchandises it is not enough efficient to recommend people

just the good things available. Rather, it has become important to recommend people things they prefer. Book rating sites such as Goodreads solely depend on users input, they use user preferences to recommend new books to the users. There are many approaches to build a user based recommendation systems. This paper has explored a few of them. According to the results obtained, it can be said that a machine learning approach is better for making a preference based recommendation system. Collaborative filtering is a very effective and efficient way to deduce predictions, making a system more accurate and dynamic at the same time. With more variables, one being genres, it is possible to make this approach more feasible.

REFERENCES

1. Li, S. (2017, September 17). *How Did We Build Book Recommender Systems in an Hour Part 1—The Fundamentals*. Retrieved from https://towardsdatascience.com
2. Li, S. (2017, September 20). *How Did We Build Book Recommender Systems in an Hour Part 2—k Nearest Neighbors and Matrix Factorization*. Retrieved from https://towardsdatascience.com
3. D. Goldberg, D. Nichols, B. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. Communications of the Association for Computing Machinery, 35(12):61–70, 1992.
4. P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Reidl. GroupLens: An open architecture for collaborative filtering of netnews. InProceedings of the 1994 Computer Supported Cooperative Work Conference, New York. ACM.
5. Mooney, R. J., & Roy, L. (2000, June). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries* (pp. 195-204). ACM.
6. Sohail, S. S., Siddiqui, J,. Ali, R. (2013). Book recommendation system using opinion mining technique, presented at 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, India, 2013 August 22-23. IEEE
7. Sase, A., Varun, K., Rathod, S., Patil, D. (2015). A Proposed Book Recommender System. *International Journal of Advanced Research in Computer and Communication Engineering,4*(2). DOI 10.17148/IJARCCE.2015.42108
8. Rajpurkar, S., Bhatt, D., Malhotra, P. (2015). Book Recommendation System. *International Journal for Innovative Research in Science & Technology*, *1*(11).
9. Sarwar, B., Karypis, G., Konstan, J., Riedl, J. (2001). Item-Based Collaborative Filtering Recommendation Algorithms. Retrieved from http://www.grouplens.org/papers/pdf/www10_sarwar.pdf
10. Amatriain, X., Jaimes, A., Oliver, N., Pujol, J. (2011). Data Mining Methods for Recommender Systems. Retrieved from http://www.newbooks-services.de/MediaFiles/Texts/7/9780387858197_Excerpt_001.pdf
11. Wikipedia
12. Altman, N. S. (1992). An Introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician.* 46(3): 175-185.