PERSONALIZED BOOK RECOMMENDATION SYSTEM USING MACHINE LEARNING TECHNIQUES

BY

SHAKIL AHMED
ID: 171-15-9376

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori
Associate professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised By

Md. Zahid Hasan
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

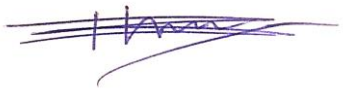
MAY 2021

APPROVAL

This Project/internship titled **"Personalized Book Recommendation System Using Machine Learning Techniques"**, submitted by Shakil Ahmed, ID No: 171-15-9376 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 31-05-2021.

BOARD OF EXAMINERS

Chairman



Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Gazi Zahirul Islam

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Raja Tariqul Hasan Tusher

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

External Examiner

Dr. Dewan Md. Farid

Associate Professor

Department of Computer Science and Engineering

United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Associate professor & Associate Head, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Sheak Rashed Haider Noori
Associate professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Zahid Hasan
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

(Shakil Ahmed)
ID: 171-15-9376
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Dr. Sheak Rashed Haider Noori**, **Associate professor & Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Recommendation System to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Recommendation system plays a tremendous role in this information era. In the world of data, it helps to increase online user experience by filtering information in automated ways. It provides personalized view of information. Recommendation system research has included a wide variety of machine learning techniques including data mining, user modeling, case based reasoning etc. Personalized recommendations are an important part of many online ecommerce platforms where the system is responsible for displaying potential item to user. The principle point regarding this research work is to find interest based book for increasing sales by machine learning techniques.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: Introduction	**10-11**
1.1 Introduction	10
1.2 Motivation	10
1.3 Objective	11
1.4. Expected Outcome	11
CHAPTER 2: Background	**12-15**
2.1 Introduction	12
2.2 Research Summary	12
2.3 Challenges	13
2.4. Related Works	13

CHAPTER 3: Research Methodology **16-24**

3.1 Introduction 16

3.2 Research Subject and Instrumentation 16

3.3 Data Collection 18

3.4 Statistical Analysis 19

3.5 Proposed Methodology 19

3.6 Implementation Requirements 23

CHAPTER 4: Experimental Results and Discussion **25-26**

4.1 Introduction 25

4.2 Experimental Setup 25

4.3 Implementation Requirements 26

CHAPTER 5: Conclusion and Future Scope **27-28**

5.1 Conclusions 27

5.2 Scope of Further Study 28

APPENDIX

REFERENCES

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Different ways of information filtering	13
Figure 3.2: Recommendation goals to achieve	14
Figure 3.3: Dataset analysis	16
Figure 3.4: Predicting recommendation using collaborative filtering	18
Figure 3.5: Computing item-item similarity in column oriented data	19
Figure 3.6: Prediction flow using weighted sum and regression	20

LIST OF TABLES

TABLES	PAGE NO
Table 3.1: Sample dataset	14

CHAPTER 1

Introduction

1.1 Introduction:

Recommendation systems are used to increase sales by identifying potential customers. Commercial companies take the benefit of such system to help users to find product or services that they might not have found by themselves. With this approach companies can efficiently use their resources and make the best use of it. This target oriented approach helps organizations to generate high conversion rate. Our study topic of this work is to find out best way of recommending such product or service. We'll use book as an example recommended item and we're going to build a recommendation engine that can recommend books to the potential customers.

We have used item based collaborative filtering in this work and our goal is to fine tune and make user oriented improvement based on some existing algorithm and propose a new system that generate recommendation. As a result, we will be able to identify how much a book will be liked by a user if he/she does not rate that book yet. With this result we will show users top recommended items. To build this core system we'll take help of various python machine learning tools and libraries that will help us to collect implicit or explicit feedback rating, clean the dataset, visualize to gain insight, analyze the data and finally make prediction. We'll also use popular datasets that has already aggregated by large companies and community as well as our own system data to train and test the proposed model.

1.2 Motivation:

The internet is full of information. Most of the cases a user doesn't need all of them. Unnecessary information makes a user experience bad which turns into losing profit for businesses. It is very crucial, especially for business because user tend to buy book when they are in front of them. A good information filtering system is necessary for this type of work which will show personalized content to a user. The fundamental inspiration for the proposed procedure is to improve the system output using machine learning techniques.

1.3 Objective:

The model recommended in this undertaking relates to the most significant problem of internet surfing which is finding exact information. This problem can easily be eliminated by a recommender system which is able to recommend content based on user interest. The primary objective of this study is to build such system to make user experience better.

1.4 Expected Outcome:

Comprehensively, the proposed methodology comprises four important steps: Collecting explicit or implicit feedback from user, preprocess the collected data, train & test the recommender model and finally recommend items to users. In most cases we'll use numerous datasets like goodreads, amazon product sales, movie lens, bookkeeping to examine our model output.

CHAPTER 2

Background

2.1 Introduction:

The primary goal of recommender systems is to assist users to find the most relevant items in that context. To do this job done these systems generally used several filtering techniques. Frequently used methods are User-Based collaborative filtering, Item-Based collaborative filtering, content based filtering, demographic filtering etc. In user-based collaborative filtering users are getting matched by other users and create a neighborhood of similar taste people. Top items of that neighborhood is then recommended to other users. On the other hand, item-based collaborative filtering create similarity between items and then recommended to users.

The dataset and recommendation algorithm plays a big role in making such system. How much we can recommend is decided by the data we have. Dataset contains item information's and user ratings which may be implicit or explicit. There can also be some other data such as demographics etc. Later we should apply correct algorithm and find the recommendations.

2.2 Research Summary:

Recommendation system is being popular in the internet world where information filtering is most important. From different types of recommendation system approached we followed collaborative filtering in this work. Collaborative filtering system utilizes the information about other users with similar preference and make prediction. We classified the process of recommendation into four steps.

Firstly, constructing user-item rating matrix by gathering browsing or rating history of a user. After processing the data, we get a user-item rating matrix.

Secondly, we find the similarity by computing the collected data from different users. Two major approaches are user-user filtering and item-item filtering. As user-user

filtering is much more computation heavy method we used item-item filtering in this work.

Thirdly, we select the user neighbor according to the similarity ranking we got from previous steps. Finding the neighbor can be done using various algorithm and we studied many of them in this work. The optimal solution is generated through k-nearest neighbors.

Finally, we need to predict the rating for item recommendation after getting the result from previous steps.

2.3 Challenges:

As recommendation will be made by the data collected from user interaction, it can be very hard to cope with user taste as they vary from time to time. A user who buy a product frequently may or may not also buy in future. As a human, the user taste changes. So monitor every user activity and predict such move from the data can be challenging.

Cold start problem is another challenge to build such system. When the system doesn't know very much about a user, it can't provide good recommendations. It occurs in the case of new user or a user with less interactions.

2.4 Related Works:

Two of the most widely used techniques in recommendation system is based on User and based on item. These systems are widely used and studied in both industry and academia but lot more problem to be solved with high accuracy. Such of problems are cold start problem, data sparsity problem etc. Many novel approach is taken to handle such problems.

2.4.1 Item-Based Collaborative Filtering Recommendation Algorithms:

The most promising works of collaborative filtering is presented on the paper named "Item-Based Collaborative Filtering Recommendation Algorithms" is done by Badrul Sarwar, George Karypis, Joseph Konstan and John Riedl. They are fellow of GroupLens Research Group. In this paper they have shown techniques of making recommendation in live interaction. As producing real-time recommendation on high volume is challenge, this paper showed a way of effective filtering. The techniques used in this work is called item-item collaborative filtering and later they also compare the result with k-nearest neighbor approach. It has successfully eliminated two common problems. They are sparsity and scalability.

2.4.2 Collaborative Filtering for Implicit Feedback Datasets:

This work has been done by Yifan Hu, Chris Volinsky from AT&T Labs and Yehuda Koren from Yahoo Research. In this work authors explained some mechanism of collaborative filtering when there is not explicit data from user. Generally, most of the production application nowadays depends on implicit data. Though explicit data is more accurate but most user don't provide data by themselves. For example, a study of an ecommerce site shows that only 15% - 20% user rate an item after they get a product. So in this scenario we need to depends on implicit data to recommend next content for user.

There are different types of challenges when measuring recommendation with implicit feedback dataset. No negative feedback, Noise, Numeric values indicate confidence instead of preferences are the primary challenges. Author of this work has shown the way this challenges can be eliminated.

Authors used latent factor algorithm to address the preferences-confidence paradigm.

2.4.3 Collaborative Filtering via Euclidean Embedding:

This work is done by Mohammad Khoshneshin and W. Nick Street from University of Iowa. In this work authors have shown usage of Euclidean embedding method as an alternative of popular latent factor model.

In this method, users and items are embedded in a unified Eculidean space where the distance between a user and an item is inversely proportional to the rating.

2.4.4 Restricted Boltzmann Machines for Collaborative Filtering:

Most approaches of collaborative filtering can't handle large dataset while real-time computation. In this paper Ruslan Salakhutdinov, Andriy Mnih and Geoffrey Hinton from University of Toronto has shown how a class of two layer undirected graphical models can be used to predict user preferences.

CHAPTER 3

Research Methodology

3.1 Introduction:

An overview of my research and the techniques I experienced will be in this part of the report. It gives us point by point information about the strategies I utilized and how they were actualized. It will likewise examine the sources we have drawn my information from and how I intend to utilize those to recommend books to a user. Instruments utilized in this research are additionally discussed in this section.

3.2 Research Subject:

The goal of this research is to see an efficient method that would recommend books with better precision and less mistake rate. To accomplish it, we applied various content filtering methods and tried to find out which method performs well in a given scenario.

Information filtering can be done in various ways. They can be primarily divided into six major parts.

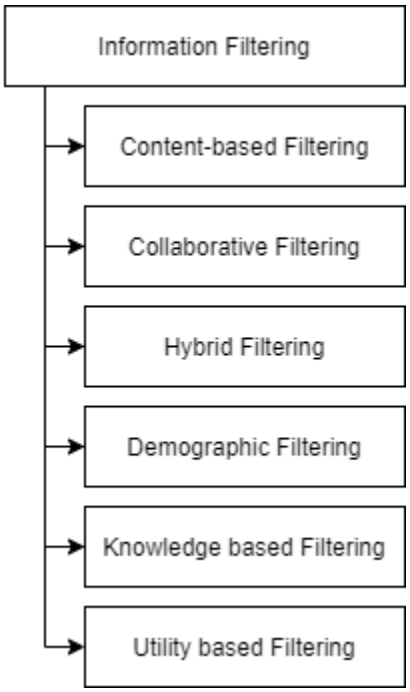


Figure 3.1: Different ways of information filtering

Collaborative filtering is the most interesting one to study about because it works with buying pattern of all users in a system. It's "collaborative" because it predicts a given customer's preference on the basis of other customers. Here two types of collaborative is mostly popular. User-User collaborative filtering and Item-Item collaborative filtering. In user-user collaborative filtering method the system identifies similar users by buying pattern and cluster them in a group. Then recommend items which isn't reviewed or bought yet by another user. It works well but it has massive drawbacks in maintenance as user preference changes over the time. In a research on amazon.com finds that user preferences can be changed in a matter of a day. Clustering users based on preference required offline analysis and should update regularly which is computationally expensive. On the other hand, item-item collaborative filtering gives superior result with less overhead as item information doesn't change frequently and it is possible to recommend books on the fly.

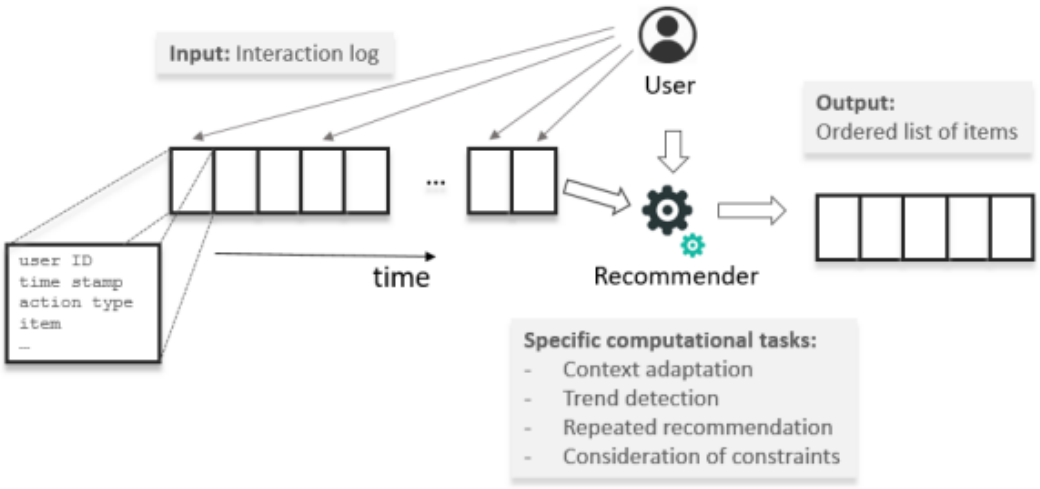


Figure 3.2: Recommendation goals to achieve

3.3 Data Collection:

In this work we've used Goodread book dataset which is open source and the reason of selecting this dataset is, it is production dataset which is collected from real scenario. All the ratings are given by real user of goodread. Using such type of datasets gives us big advantage because it reflects real behavior of users of the system.

The dataset contains 7.8 million reviews which is large enough to split into train-test set and evaluate our model perfectly. The dataset contains anonymous user id, product id, rating provided by the user and timestamp.

Table 3.1: Table of sample dataset

userId	**productId**	**rating**	**timestamp**
AKM1MP6P0OYPR	0132793040	5.0	1365811200
A2CX7LUOHB2NDG	0321732944	5.0	1341100800
A2NWSAGRHCP8N5	0439886341	1.0	1367193600
A2WNBOD3WNDNKT	0439886341	3.0	1374451200
A1GI0U4ZRJA8WN	0439886341	1.0	1334707200

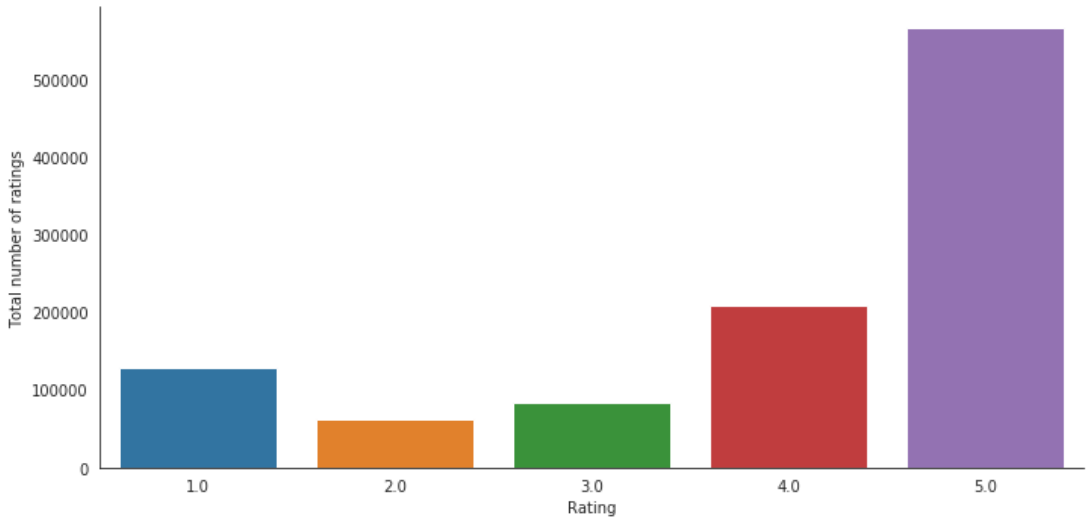


Figure 3.3: Dataset analysis

3.4 Statistical Analysis:

In the collaborative recommendation system, the main thing is to calculate the closeness between users. Here are the techniques we tried to get the work done.

The first approach we followed to measure the similarity between users is using Cosine Similarity method where we need to find the cosine angle between two users. The user rating is pointed as a vector in two dimensional vector spaces. If the cosine similarity between two users is closer, then we can cluster them in a similar neighbor.

Another popular approach to measure the similarity is using SVM which performs well when data are dense and continuous.

Finding neighbors with K Nearest Neighbor algorithm also works well. We applied this algorithm with Surprise and Scikit-learn machine learning toolkit. The result is accurate with high performance rate.

According to these analysis, we propose a new approached to find similarity.

3.5 Proposed Methodology:

In traditional recommendation system finding the similarity between users is done by rating information. There is some gap in this traditional approaches. The factor of user confidence and time context is not taken into account.

Time context is also important in recommending item to user. Recently visited item plays an important role of finding user preferences as time is dynamic attribute and preferences can be changed based on time. In this work we've introduced time aware cosine similarity algorithm to calculate the similarity between users and finding neighbors.

In a given timeframe user prefer specific item. For example, students prefer budget specific item but after a while when they got a job their preference should change. Again junior engineer tends to buy books or course which helps them to learn fundamental ideas of their domain but when they got senior their preference should be changed. These preference is relatively short-lived because these need is no longer required in future but we can identify the lifecycle loop as other users may be gone through these path. For these reason our neighbor should be updated time to time and our algorithm should aware about time.

Our algorithm should also find the domain expert in user. Every user in their personal life is some sort of expert in their field. If our algorithm can find out the work domain of a user, it is more convenient to find neighbors.

In various season and festival time user prefer some item for that time only. For example, in summer user prefer refreshing wearing like shorts and winter user prefer to have warm cloth, coats. Festival like Eid users buy festival centric items. These are short-lived preferences and our algorithm should also aware about it.

After experimenting with various techniques we have finalized item item collaborative filtering for real-time capability and scalability issue. As it performs well in business scenario with high accuracy, it is the expected one.

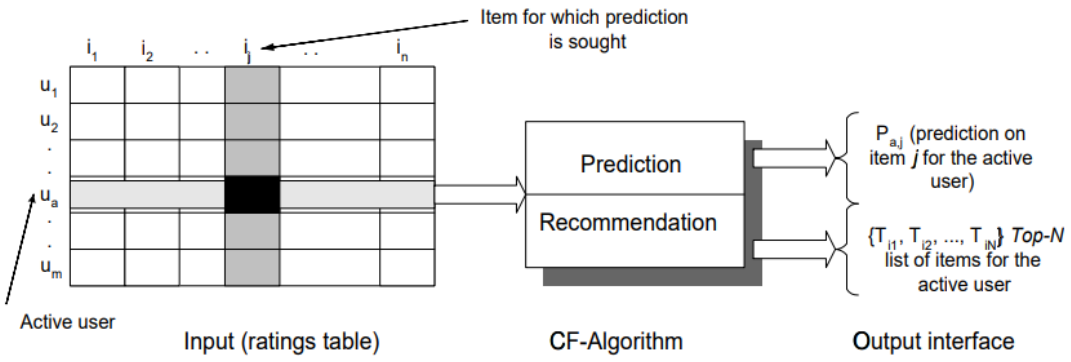


Figure 3.4: Predicting recommendation using collaborative filtering

Unlike user based collaborative filtering, item based collaborative filtering seek similarity between items. As a production system may have many users but not all of them rate large portion of item, finding similarity between items is less computation intensive and time efficient. Also we can eliminate sparsity problem with this technique.

To implement this method, at first we need to find similarity between items. To do this we can follow various approach. Most popular ways are:

1. Cosine-based similarity.
2. Correlation-based similarity.
3. Adjusted Cosine Similarity.

3.5.1 Cosine-based Similarity:

Here two items are considered separate vector in different dimension. Then measuring the cosine of angle between them tells us how similar both items are.

$$sim(i,j) = \cos(\bar{\imath},\bar{\jmath}) = \frac{\bar{\imath}.\bar{\jmath}}{||\imath| * |\jmath||}$$

3.5.2 Correlation-based Similarity:

Here pearson-r correlation take place to calculate similarity between items.

3.5.3 Adjusted Cosine Similarity:

In dataset users are aligned on row and items are aligned in columns so user-based collaborative filtering is done by calculating along rows. But in item-based collaborative filtering measuring similarity between items is done by columns.

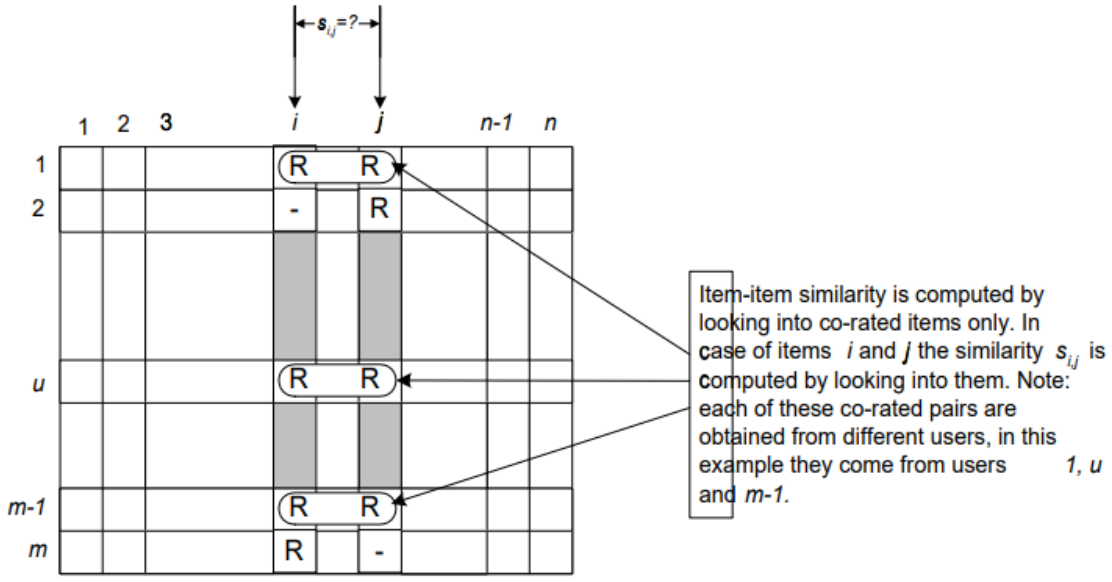


Figure 3.5: Computing item-item similarity in column oriented data

After the filtering techniques, the most important part is predicting the output. Here we can follow two approach. They are:

1. Weighted Sum
2. Regression

3.5.4 Weighted Sum:

In this method, all the rating given by a user U is summed up. Then they are weighted by the corresponding similarity between items.

3.5.5 Regression:

It is also similar with weighted sum method but uses approximation of rating based on regression model.

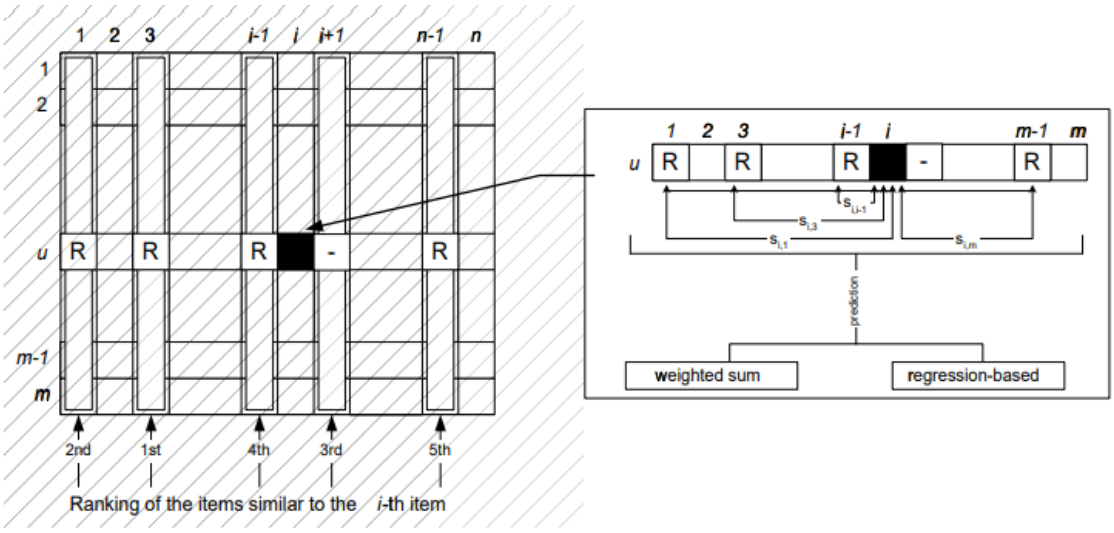


Figure 3.6: Prediction flow using weighted sum and regression

3.6 Implementation Requirements:

The details system information is given below:

Hardware Requirement (minimum):

- Processor:
  2.4 GHz (minimum) with 4 cores CPU and also multithreading enabled
- GPU
  Better to have NVIDIA GTX1050
- Storage
  Minimum 25 Gigabyte of free space secondary memory (SSD/HDD)

Software Requirements (Operating System):

- Linux – Ubuntu 16.04 LTS or higher
- Windows – Windows 10 Pro Edition V1609 or higher

Required Environments:

- Python 3.6 or higher
- Anaconda
- Jupyter Notebook

- Django

Packages:

- Numpy
- Pandas
- Matplotlib
- Scikit-learn
- Surprise

CHAPTER 4

Experimental Results and Discussions

4.1 Introduction:

Each method of hypothesis needs to be thoroughly tested. We've therefore used our proposed approach as an active framework to test our vision. This section of the report governs the internal and external definition of our framework used. This section further shows the installation parameters and the production parameters of the frame.

4.2. Evaluation:

Until now, numerous scientists in the report have introduced numerous measurements to assess the exhibition of suggestion frameworks. As a rule, the measurements for assessing the proposal framework quality essentially incorporate two classes. They are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)

MAE is perhaps the most generally utilized measurements to assess the proposal exactness and is characterized as the normal of outright contrast between expectation esteems and real evaluations. Accepting that the arrangement of anticipated client appraisals is $\{P_1, P_2, P_3 \ldots .P_n\}$ and the relating set of genuine client evaluations is $\{R_1, R_2, R_3 \ldots .R_n\}$, the MAE is determined as follows:

$$\text{MAE} = \frac{\sum |Pi - Ri|}{n}$$

4.3 Experimental Results:

In this part, tests are executed on the Amazon Product Dataset. We contrast our proposed calculation and different calculations by utilizing the measurements.

As is notable to us every one of us, number of selected nearest neighbors critically affects the quality of recommendation output. In this way, the quantity of the nearest neighbor k

varies from 10 to 60 in experiment. As of now, the regularly used methods to figure the similarity between users incorporate COS.

CHAPTER 5

Conclusion and Future Scope

5.1 Conclusions:

In this work, we've presented a collaborating filtering algorithm dependent on item similarity and rating of a user. We've also added time aware techniques where recent and seasonal items get higher priority for recommendation. Our methodology depends on the viewpoint that user taste is not constant and may switch to different domain. For recognizing the best average similarity measure, the essential work of this paper is to acquire a superior measure to adjust. For that reason, we've studied and applied most recent approach to build this system. Subsequently, we perform experiment in our paper which has demonstrated that the time aware algorithm is more suitable and reasonable for calculating the similarity between items or users. In experiments, we assessed the effectiveness of our proposed calculation on MAE, RMSE, accuracy, review, and measure. Furthermore, trial results on the dataset showed our proposed calculation has better execution contrasted and current calculations and some different calculations. All in all, in our proposed calculation, item-based collaborative filtering is powerful to improve the performance of recommendation systems.

5.2 Scope of Further Study:

Later on, we will keep on experimenting the effect of the user behavior and the information of items in proposed methodologies and study the system of the user rating behavior and the effect of the item information.

Reference:

Conference/Journal Papers:

[1] K. Choi and Y. Suh, "A new similarity function for selecting neighbors for each target item in collaborative filtering" IEEE Transl. J. Magn. Japan, vol. 37, no. 1, pp. 146 - 153 2013.

[2] C. Martinez-Cruz, C. Procel, J. Bernabe-Moreno, and Herrera-Viedma, "A model to represent users trust in recommender systems using ontologies and fuzzy linguistic modeling," Information Sciences, vol. 311, pp. 102-118, 2015.

[3] G. Adomaviciaus and A. Tuzhilin, "Toward the next generation of recommender systems; a survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749 2015.

[4] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutierrez, "Recommender systems survey," Knowledge-Based Systems, vol. 46, no. 1, pp. 109-132 2013.

[5] Badrul Sarwar, George Karypis, Joseph Konstan and John Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", vol. 59, no. 2, pp. 155-1198 2014.

[6] Aggarwal, C.C. Wolf, J.L, "Horting Hatches an Egg: A New Graph-theoretic approach to Collaborative Filtering", Proceedings of the ACM KDD'99 Conference, pp. 201 - 212.

[7] Herlocaker, J., Konstan, J. Borchers, "An Algorithmic Framework for Performing Collaborative Filtering." Proceedings of ACM SIGIR'99, ACM press.

[8] Herlocker J., "Understanding and Improving Automated Collaborative Filtering Systems" Ph.D. Thesis, Computer Science Dept., University of Minnesota.

Books:

[1] Kim Falk, Practical Recommender Systems, Manning, 2019, 253

[2] Toby Segaran, Programming Collective Intelligence, O'REILLY, 2007, 325

Websites:

[1] Amazon Science, available at *https://www.amazon.science/the-history-of-amazons-recommendation-algorithm* last accessed on 18 February 2021 at 9.00PM.

Recommender_System_171-15-9376.pdf

ORIGINALITY REPORT

25%	23%	8%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	11%
2	www.hindawi.com Internet Source	5%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
4	Mohammad Khoshneshin, W. Nick Street. "Collaborative filtering via euclidean embedding", Proceedings of the fourth ACM conference on Recommender systems - RecSys '10, 2010 Publication	1%
5	scholarworks.sjsu.edu Internet Source	1%
6	Komal Kapoor, Karthik Subbian, Jaideep Srivastava, Paul Schrater. "Just in Time Recommendations", Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15, 2015 Publication	<1%

7	ict.ewi.tudelft.nl Internet Source	<1%
8	mafiadoc.com Internet Source	<1%
9	www.amazon.science Internet Source	<1%
10	Submitted to King's College Student Paper	<1%
11	Submitted to Middle East Technical University Student Paper	<1%
12	Submitted to Birla Institute of Technology and Science Pilani Student Paper	<1%
13	www.slideshare.net Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		