Design Document For
Techniques in building a Recommender System

Team:

1. Snehit Reddy (2017A7PS0868H)
2. Dhanush Karupakula (2017B3A71011H)
3. Sai Satvik Vuppala (2017B4A71449H)
4. Amit Raj Reddy D (2018B4A70813H)

Topic

This assignment is aimed at implementing and comparing various techniques for building a Recommender System.

Introduction

This application can predict ratings a user might give to a particular movie based on their interest that is by the previous history of ratings given by various users for various movies, and thus it can recommend new movies to the user. Ratings are given as a whole number in the range of 1 to 5. Several techniques like below were used to achieve this functionality -

a. Collaborative filtering
b. SVD (Singular Value Decomposition)
c. CUR

For each method used, RMSE error measures, Precision on Top k, Spearman Rank Correlation have been calculated.

Libraries Used

These are the python libraries used in this project :

1. **Math:** This library of python is used generally for all the mathematical functions.
2. **NumPy:** This library is used to store the data in the form of an array and is used for easier array computations.
3. **Random:** We use this module to generate random values.
4. **Linalg:** We use this library to compute the eigenvalues from an ordinary or generalized eigenvalue problem. The return values of this function are the eigenvalues and the eigenvectors.
5. **Time:** This was used to keep track of the amount of time that each process was taking.

Techniques

- **Collaborative Filtering**

  In our first approach, collaborative filtering was used to predict user ratings; to be more specific item-item collaborative filtering was used. Item-based collaborative filtering is used, as a similarity between items is more accurate and meaningful than the similarity between users. To account for strict and lenient users, each row was subtracted by its mean to make each row mean-centered at 0. To predict the rating of a particular movie given by a particular user, 15 movies that are most similar to the given movie and which are rated by the user are used.

  $$\mathrm{sim}(x, y) = \Sigma(r_{xs}-\mu_{x})\,(r_{ys}-\mu_{y}) \;/\; \sqrt{}[\Sigma(r_{xs}-\mu_{x})_{2} \times \Sigma(r_{ys}-\mu_{y})_{2}]$$

- **Collaborative filtering with baseline estimates**

  This approach was an enhanced version of the approach specified above. Here global and regional baseline estimates were also used as a part of computations. Strict and lenient raters were automatically handled as the baseline estimate accounts for the user’s and the movie’s mean ratings.

  $$b = \mu + bx + bi$$

µ: Overall Mean, bx: Deviation for user x, bi: Deviation for item i

- **SVD Decomposition**

  Singular Value Decomposition is a factorization method to decompose or factorize a real-valued matrix. SVD factorizes a matrix M into U, Sigma, and V* (V-transpose). Columns of U represent the left-singular vectors of the matrix M. Sigma is an m X n rectangular diagonal matrix, with each diagonal element as the singular values of the utility matrix. Columns of V represent the right-singular vectors of the matrix M. Dimensionality reduction is performed on each of the 3 matrices to bring the vectors to a lower-order dimension space. The dot product U.Sigma.V* will generate a matrix which is a close approximation of the matrix M.

  A = U * Sig * V' ,

  Where A = original data matrix (users * items), U = users to concept matrix, V = ms to concept matrix, Sig = concept strength matrix containing eigenvalues in decreasing order.

- **CUR Decomposition**

  CUR matrix decomposition is a low-rank matrix decomposition algorithm that uses a lesser number of columns and rows than the data matrix. This number is represented by the variable k. The rows and columns are selected randomly based on their probability distributions. The probability distribution is based on a statistical leverage score which represents the row/column importance. Matrix C consists of the randomly picked columns and matrix R consists of the randomly selected rows. The intersection of R and C gives us the intermediate matrix W. On W, SVD is applied and U is obtained. Finally, the product CUR gives us approximations.

$$\mathrm{P(x)} = \sum_i (m_{i,x})^2 \ / \ \|M\|_F^2$$

Error Measures

RMSE - Root Mean Square Error

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i-Actual_i)^2}{N}}$$

Precision in Top K

Gives an estimate of how many of the predicted ratings are present in the top K ratings of the user since only the good ones count in the error measure.

Spearman Correlation

Spearman's correlation measures the strength and direction of the monotonic association between two variables

$$\rho = 1-\frac{6\sum d_i^2}{n(n^2-1)}$$

Functionalities provided by the application

- Reads the dataset containing the ratings of all movies given by the users.
- The rating dataset is split into training and test sets
- Generate the User-Movie matrix for each filtering model
- Design Algorithms for SVD, CUR, and Collaborative Filtering models
- Pass the User-Movie matrix to all the above algorithms.
- Generate the predicted User_Movie matrix for each model
- Pass original and predicted matrices to Error measures such as RMSE
- Precision on Top k and Spearman Rank Correlation are also calculated
- Results are tabulated along with the time taken for model development

Advantages and disadvantages of the models

- CUR decomposition

  Advantages:

  - Easy interpretation of the matrices C, U and R
  - Sparse basis

  Disadvantages:

  - Possibility of selecting duplicate columns and rows.
  - CUR decomposition

- Collaborative Filtering

  Advantages:

  - Work even for a new user.

  Disadvantages:

  - Need for sufficient number of users (cold start problem)
  - Too sparse matrix
  - Popularity bias
  - Cannot recommend the item that has not been created (first starter problem)
  - Collaborative Filtering

- SVD

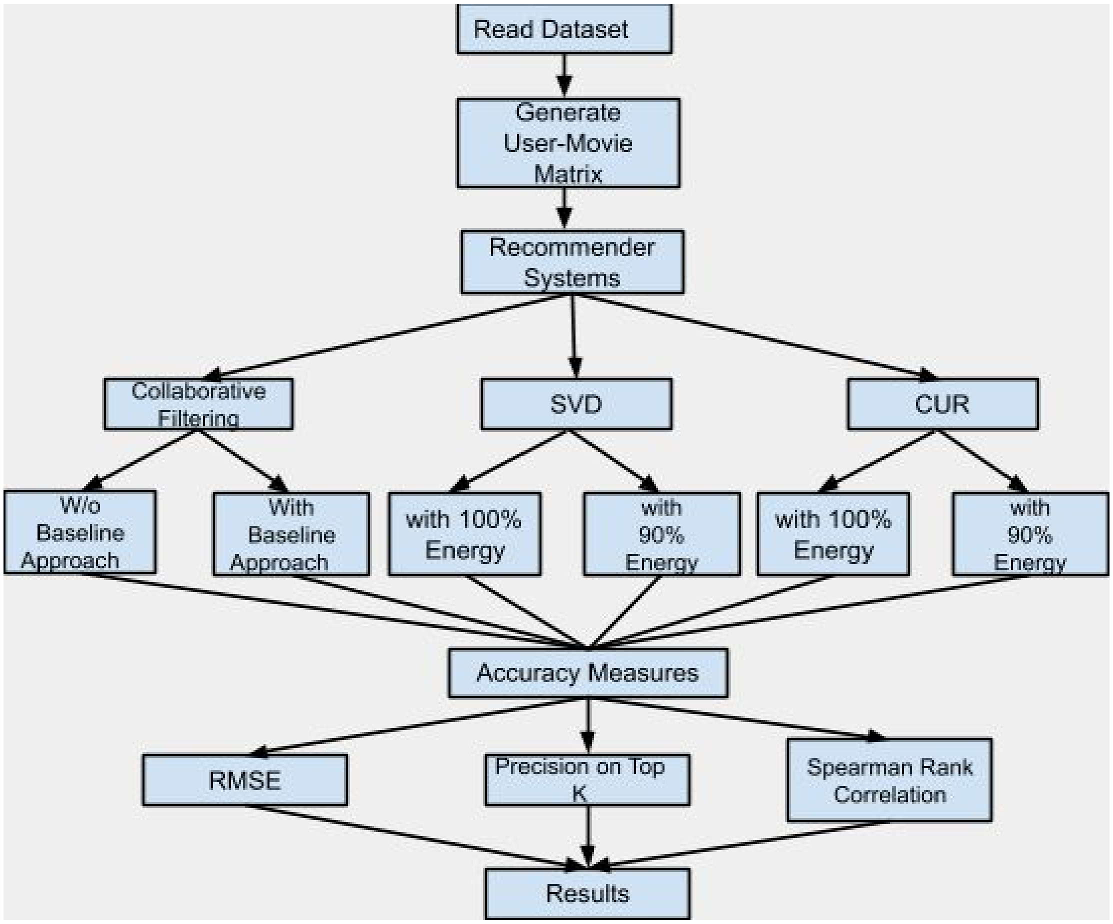
  Advantages:

  - Optimal low rank approximation

  Disadvantages:

  - Interperetability problem
  - Singular vectors are sparse.

Flowchart


Read Dataset
Generate User-Movie Matrix
Recommender Systems
Collaborative Filtering
SVD
CUR
W/o Baseline Approach
With Baseline Approach
with 100% Energy
with 90% Energy
with 100% Energy
with 90% Energy
Accuracy Measures
RMSE
Precision on Top K
Spearman Rank Correlation
Results

Results

Recommender System Technique	Root Mean Square Error (RMSE)	Precision on top K	Spearman Rank Correlation	Time taken for prediction
Collaborative	1.372	0.179	0.999	19.28s
Collaborative along with Baseline approach	1.187	0.194	0.999	18.5s
SVD	1.199	0.310	0.999	30s
SVD with 90% retained energy	1.315	0.310	0.999	28s
CUR	0.798	0.310	0.999	2.5s
CUR with 90% retained energy	0.902	0.31	0.999	5s

References

1. C. D. Manning, P. Raghavan, and H. Schutze. Introduction to Information Retrieval, Cambridge University Press, 2008.
2.