CS6530 Analytic Databases

Assignment: Recommendation System

Marks: 60

Start Date: 25-01-2018

Due Date: 18-02-2018

Overview

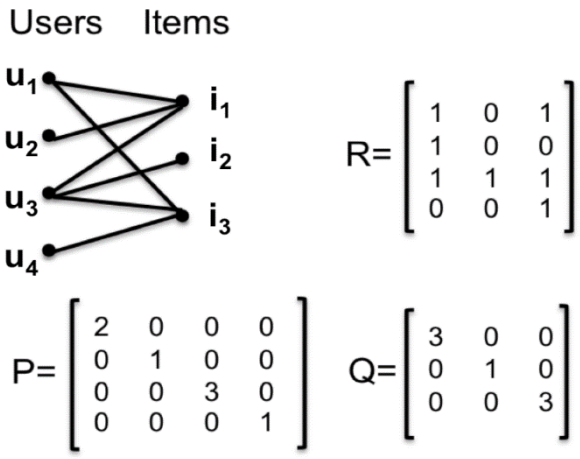
This project is aimed at developing your own recommendation system and comparing the results of the system built by you with a standard recommender system library. The data source to be used is: User-TVShow, an overview of the dataset is given later in the below sections. MyMediaLite library is used for comparing the results. You can do the assignment using Java or Python.

Problem Description

Consider a user-item rating matrix R defined as

$$R_{ij} = \begin{cases} 1 & if\ user\ i\ likes\ item\ j \\ 0 & otherwise \end{cases}$$

Let there be 'm' users and 'n' items, hence R is (m x n) matrix. Now let P be an (m x m) diagonal matrix whose i^{th} diagonal element represents the number of items that are liked by the i^{th} user. Let there be another diagonal matrix Q of order (n x n) whose j^{th} diagonal element represents the number of users who have liked the item 'j'. Below is an illustrated example.



Exercises

Exercise 1 [5 points] Consider the matrix $T = R * R^T$. What is the difference between matrix entries T_{ii} and T_{ij} where $i \neq j$?

Exercise 2 [5 points]

The cosine similarity of two vectors u and v can be defined as

$$cosSim(u, v) = \frac{u.v}{||u||.||v||}$$

Item similarity matrix S_I can be defined as: $S_I = Q^{-1/2} R^T R Q^{-1/2}$

User similarity matrix S_U can be defined as: $S_U = P^{-1/2} R R^T P^{-1/2}$

Give justification for the above definitions of item and user similarity matrix? Would the same expression hold true if we had user ratings in the scale 1-5 instead of binary 0-1 rating?

Exercise 3 [10 points]

For both item-item and user-user collaborative filtering the recommendation matrix *Γ*, which is an (m x n) matrix, can be defined as:

$$\Gamma(i, j) = r_{ij}$$

For user-user collaborative recommendation r_{ij} is defined as

$$r_{ij} = \sum_{x \in users} cosSim(x, i) * R_{xj}$$

For item-item collaborative recommendation r_{ij} is defined as

$$r_{ij} = \sum_{x \in items} R_{ix} * cosSim(x, j)$$

The matrix Γ can thus be defined as:

$$\Gamma = \begin{cases} S_U * R & \text{for user-user filtering} \\ R * S_I & \text{for item-item filtering} \end{cases}$$

Give justification for why the above equation for matrix *Γ* holds true. To recommend the top-*k* items for a user *i*, you need to pick the k items with highest r_{ij}.

Exercise 4 [20 points]

In this question, you will evaluate the two recommendation algorithms (item-item and user-user filtering) described in Exercise 3. For testing, we give you a dataset about TV shows that contains information about 9985 users and 563 popular TV shows. The dataset contains two files:

- **User-shows.txt**: This is the ratings matrix R, where each row corresponds to a user and each column corresponds to a TV show. $R_{ij} = 1$ if user i watched the show j over a period of three months. The columns are separated by a space.
- **Shows.txt**: This is a file containing the titles of the TV shows, in the same order as the columns of R.

Consider the user corresponding to the 20th row of User-shows.txt. For this user erase the first hundred entries by replacing them with 0s. This means we do not know which of the first 100 shows this user has watched. Based on the other shows that this user has watched recommend the shows that this user would watch in the first 100 shows. You will compare your recommendations with the shows that the user had actually watched.

Compute the recommendation matrix Γ for both item-item and user-user filtering. Let S be the set of first 100 shows for the user corresponding to row 20. What are the five shows that have highest similarity score using the two filtering methods? What are their similarity scores? In case of ties between two shows, choose the one with smaller index. Do not write the index of the TV shows. Write their names using Shows.txt.

For a given number k, the true positive rate at top-k is defined as follows: using the matrix Γ computed previously, compute the top-k TV shows in S that are most similar to the user (break ties as before). The true positive rate is the number of top-k TV shows that were watched by the user in reality, divided by the total number of shows he watched in the held-out 100 shows.

- Plot the true positive rate at top-k (defined above) as a function of k, for $k \in [1,19]$, with predictions obtained by the user-user collaborative filtering.
- On the same figure, plot the true positive rate at top-k as a function of k, for $k \in [1,19]$, with predictions obtained by the item-item collaborative filtering.

Exercise 5 [20 points]

For the 20th user in the dataset list out the Show IDs of the top-10 recommended TV shows using the following methods (here you don't need to erase the 100 entries as you did in exercise 4):

1. Item-item collaborative filtering (as discussed in the assignment)
2. User-user collaborative filtering (as discussed in the assignment)
3. ItemKNN (use MyMediaLite's Item Recommendation program)
4. WRMF (use MyMediaLite's Item Recommendation program)

Use the following format to list the result:

Item-item	User-user	ItemKNN	WRMF
ShowID1	ShowID1	ShowID1	ShowID1
ShowID2	ShowID2	ShowID2	ShowID2
.	.	.	.
.	.	.	.
.	.	.	.
ShowID10	ShowID10	ShowID10	ShowID10

Also provide the Show names for the above table as follows:

ShowID	ShowName
ShowID1	Show Name 1
.	.
.	.

Now compare the four ranked lists obtained using Kendall Tau distance and report your findings in the form of a matrix given below. List down the rankings that are closest based on this measure:

	Item-item	User-user	ItemKNN	WRMF
Item-item				
User-user				
ItemKNN				
WRMF				