AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

19CSE337 Social Networking and Security

Lecture 22

Topics to Discuss

- Collaborative filtering

- The idea behind collaborative filtering is to recommend new items based on the similarity of users.

- Collaborative filtering use a user-item matrix where each entry either missing or a rating given by each user to item.

- The rating scale ranges from 0 to 5, such that 0 being the worst and 5 the best.

	ARGO	Seven	Righteous Kill	Heat
John	5	1	3	5
Tom	?	?	?	2
Alice	4	?	3	?

- In collaborative filtering, one aims to predict the missing rating and recommend the item with the highest predicted rating to the user.

- This prediction can be performed directly by using previous ratings in the matrix.

- This approach is called memory-based collaborative filtering because it employs historical data available in the matrix.

- Alternatively, one can assume an underlying model governs the way user rate items.

- This will be a learning based model.

- The second approach is called model based collaborative filtering.

Memory Based Collaborative Filtering

- Memory based collaborative filtering has basically two approaches:
  - user-based and item-based.
- User-based collaborative filtering is based on the user similarity or neighborhood. (Users with similar previous rating for items are likely to rate future items similarly).
- Item-based collaborative filtering is based on similarity among items.(Items that have received similar ratings previously from users are likely to receive similar ratings from future users).

Memory Based Collaborative Filtering

- In both cases, users or items collaboratively help filter out irrelevant content (dissimilar users or items).

- To determine similarity between users or items, two commonly used measures are cosine similarity and Pearson correlation.

- Let $r_{u,i}$ denote the rating assigned by user u to item i, $\bar{r}_u$ denote the average rating for user u, $\bar{r}_i$ denote the average rating for item i.

- Cosine similarity between the users u and v is

$$\cos(u_u, u_v) = \frac{\sum_i r_{u\,i} * r_{v\,i}}{\sqrt{\sum_i r_{u\,i}^2}\sqrt{\sum_i r_{v\,i}^2}}$$

- And Pearson correlation is

$$\text{sim}(u_u, u_v) = \frac{\sum_i (r_{u\,i} - \bar{r}_u) * (r_{v\,i} - \bar{r}_v)}{\sqrt{\sum_i (r_{u\,i} - \bar{r}_u)^2}\sqrt{\sum_i (r_{v\,i} - \bar{r}_v)^2}}$$

User Based Collaborative Filtering

- In this method, we predict the rating of user u for item i by
  - finding users most similar to u and
  - using a combination of the ratings of these users for item i as the predicted rating of user u for item i.
- To remove noise and reduce computation, it is advised to limit the number of similar users to some fixed number (say for example top 10 or top 5 etc).
- This list include the neighbours of user u, and its size is N(u).

- In user based collaborative filtering, the rating of user u for the item i is calculated as

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in N(u)} sim(u,v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N(u)} sim(u,v)}$$

- Calculate the missing rate using user CF?

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

- Step 1- Calculate the average ratings of each user.

$$\bar{r}_{John} = \frac{3+3+0+3}{4} = 2.25$$

$$\bar{r}_{Joe} = \frac{5+4+0+2}{4} = 2.75$$

$$\bar{r}_{Jill} = \frac{1+2+4+2}{4} = 2.25$$

$$\bar{r}_{Jane} = \frac{3+1+0}{3} = 1.33$$

$$\bar{r}_{Jorge} = \frac{2+2+0+1}{4} = 1.25.$$

- Step 2- Calculate similarity between Jane and other users using cosine similarity (or Pearson correlation).

$$sim(Jane, John) = \frac{3 \times 3 + 1 \times 3 + 0 \times 3}{\sqrt{10} \sqrt{27}} = 0.73$$

$$sim(Jane, Joe) = \frac{3 \times 5 + 1 \times 0 + 0 \times 2}{\sqrt{10} \sqrt{29}} = 0.88$$

$$sim(Jane, Jill) = \frac{3 \times 1 + 1 \times 4 + 0 \times 2}{\sqrt{10} \sqrt{21}} = 0.48$$

$$sim(Jane, Jorge) = \frac{3 \times 2 + 1 \times 0 + 0 \times 1}{\sqrt{10} \sqrt{5}} = 0.84.$$

- Step 3- Fix the neighbour size. Let N(u)=2. The top 2 similarity scores are for joe and Jorge. Then Jane's rating for Aladdin computed from user based collaborative filtering is

$$
\begin{aligned}
r_{Jane,Aladdin} \quad &= \quad \bar{r}_{Jane} + \frac{sim(Jane, Joe)(r_{Joe,Aladdin} - \bar{r}_{Joe})}{sim(Jane, Joe) + sim(Jane, Jorge)} \\
&\quad + \frac{sim(Jane, Jorge)(r_{Jorge,Aladdin} - \bar{r}_{Jorge})}{sim(Jane, Joe) + sim(Jane, Jorge)} \\
&= \quad 1.33 + \frac{0.88(4 - 2.75) + 0.84(2 - 1.25)}{0.88 + 0.84} = 2.33
\end{aligned}
$$

Item-Based Collaborative Filtering

- In user based CF, we compute average rating of different users and find the most similar users to the user for whom we are seeking recommendations.

- Unfortunately, in most online systems, users do not have many ratings. Therefore, the averages and similarities may be unreliable.

- The value may change when more users are added to the system.

- But when we consider items, the similarity between them and their averages are more stable.

- Therefore, we move to item based CF.

- The rating of user u for item i using Item-Based CF is calculated as

$$r_{u,i} = \bar{r}_i + \frac{\sum_{j \in N(i)} sim(i, j)(r_{u,j} - \bar{r}_j)}{\sum_{j \in N(i)} sim(i, j)}$$

where $\bar{r}_i$, $\bar{r}_j$ are the average ratings for the items i and j.

- Calculate the missing rate using item CF?

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

- Step 1- Find the average ratings of items.

$$\bar{r}_{Lion\ King} = \frac{3+5+1+3+2}{5} = 2.8.$$

$$\bar{r}_{Aladdin} = \frac{0+4+2+2}{4} = 2.$$

$$\bar{r}_{Mulan} = \frac{3+0+4+1+0}{5} = 1.6.$$

$$\bar{r}_{Anastasia} = \frac{3+2+2+0+1}{5} = 1.6.$$

- Step 2- Calculate item similarity using cosine similarity or (Pearson Correlation).

$$sim(Aladdin, Lion\ King) = \frac{0 \times 3 + 4 \times 5 + 2 \times 1 + 2 \times 2}{\sqrt{24}\ \sqrt{39}} = 0.84.$$

$$sim(Aladdin, Mulan) = \frac{0 \times 3 + 4 \times 0 + 2 \times 4 + 2 \times 0}{\sqrt{24}\ \sqrt{25}} = 0.32.$$

$$sim(Aladdin, Anastasia) = \frac{0 \times 3 + 4 \times 2 + 2 \times 2 + 2 \times 1}{\sqrt{24}\ \sqrt{18}} = 0.67.$$

- Step 3- Fix the neighbour size, N(u)=2. Top 2 similar items are Lion King and Anastasia. Then Jane's rating for Aladdin using item based CF is

$$
\begin{aligned}
r_{Jane,Aladdin} &= \bar{r}_{Aladdin} + \frac{sim(Aladdin, Lion\ King)(r_{Jane,Lion\ King} - \bar{r}_{Lion\ King})}{sim(Aladdin, Lion\ King) + sim(Aladdin, Anastasia)} \\
&+ \frac{sim(Aladdin, Anastasia)(r_{Jane,Anastasia} - \bar{r}_{Anastasia})}{sim(Aladdin, Lion\ King) + sim(Aladdin, Anastasia)} \\
&= 2 + \frac{0.84(3 - 2.8) + 0.67(0 - 1.6)}{0.84 + 0.67} = 1.40.
\end{aligned}
$$

Pros:

- Works for any items.
- No need for profile information.

Cons:

- Cold start problem.
- Sparsity.
- First rater (unrated items can't be recommended).
- Can't recommend items for someone with unique taste.

Thanks………..