Subject: Big Data Analytics

Collaborative Filtering

i. A significantly different approach to recommendation, using features of items to determine their similarity, focusing on the similarity of the user ratings for two items.
ii. Place of the item-profile vector for an item, we use its column in the utility matrix.
iii. Contriving a profile vector for users, we represent them by their rows in the utility matrix.
iv. Users are similar if their vectors are close according to some distance measure such as Jaccard or cosine distance.
v. Recommendation for a user U is then made by looking at the users that are most similar to U in this sense, and recommending items that these users like.
vi. The process of identifying similar users and recommending what similar users like is called collaborative filtering.

A. Measuring Similarity

- ✓ The first question we must deal with is how to measure similarity of users or items from their rows or columns in the utility matrix.
- ✓ The utility matrix is as shown below

Subject: Big Data Analytics

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Fig: The utility Matrix

- ✓ The above data is too small to draw any reliable conclusions, but its small size will make clear some of the pitfalls in picking a distance measure.
- ✓ Specifically the users A and C. They rated two movies in common, but they appear to have almost diametrically opposite opinions of these movies.
- ✓ Expect that a good distance measure would make them rather far apart.
- ✓ The alternative measures to consider are:

a. Jaccard Distance

- ✓ Ignore values in the matrix and focus only on the sets of items rated.
- ✓ If the utility matrix only reflected purchases, this measure would be a good one to choose.
- ✓ When utilities are more detailed ratings, the Jaccard distance loses important information.

b. Cosine Distance

Subject: Big Data Analytics

- ✓ Treat blanks as a 0 value.
- ✓ This choice is questionable, since it has the effect of treating the lack of a rating as more similar to disliking the movie than liking it.

c. Rounding the Data

- ✓ Try to eliminate the apparent similarity between movies a user rates highly and those with low scores by rounding the ratings.
- ✓ For instance, we could consider ratings of 3, 4, and 5 as a "1" and consider ratings 1 and 2 as unrated. The utility matrix would then look as shown below.

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	1			1			
B	1	1	1				
C					1	1	
D		1					1

Fig: Utilities of 3,4 and 5 have been replaced by 1, while ratings of 1 and 2 are omitted.

- ✓ Now, the Jaccard distance between A and B is 3/4, while between A and C it is 1; i.e., C appears further from A than B does, which is intuitively correct.
- ✓ Applying cosine distance to Figure above allows us to draw the same conclusion.

d. Normalizing Ratings

- ✓ If we normalize ratings, by subtracting from each rating the average rating of that user, we turn low ratings into negative numbers and high ratings into positive numbers.
- ✓ If we then take the cosine distance, we find that users with opposite views of the movies they viewed in common will have vectors in almost opposite directions, and can be considered as far apart as possible.
- ✓ Users with similar opinions about the movies rated in common will have a relatively small angle between them.

B. The Duality of Similarity

- ➢ The utility matrix can be viewed as telling us about users or about items, or both.
- ➢ There are two ways in which the symmetry is broken in practice.
  - ✓ We can use information about users to recommend items. We can base our recommendation on the decisions made by these similar users, e.g., recommend the items that the greatest number of them have purchased or rated highly. There is no symmetry. Even if we find pairs of similar items, we need to take an additional step in order to recommend items to users. This point is explored further at the end of this subsection.

Subject: Big Data Analytics

✓ There is a difference in the typical behavior of users and items, as it pertains to similarity. Intuitively, items tend to be classifiable in simple terms. For example, music tends to belong to a single genre.

C. Clustering Users and Items

- It is hard to detect similarity among either items or users, because we have little information about user-item pairs in the sparse utility matrix.
- Even if two users both like a genre or genres, they may not have bought any items in common. · One way of dealing with this pitfall is to cluster items and/ or users. · There may be little reason to try to cluster into a small number of clusters immediately.

	HP	TW	SW
A	4	5	1
B	4.67		
C		2	4.5
D	3		3

Fig: Utility matrix for users and clusters of items.

- A hierarchical approach, where we leave many clusters unmerged may suffice as a first step.
- For example, we might leave half as many clusters as there are items.