↳ The heart of neighbourhood based Collaborative filtering is the ability to find people similar to you.

Ways to measure Similarity:

1. Cosine similarity ⇒ used to measure how similar two items are

items are user behaviour data.

$$\text{CosSim}(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

Here each attribute is a dimension. So if have n attributes it forms n-dimensional space.

↳, angle blw two attributes is the similarity cos of that two attributes.

↳ These attributes are for user behaviour only not for content

(2.) Sparsity:

Data is sparse (small amount of data is spreaded over a large area)

↳ Because, they are many movies in the world and no one sees all the movies.

↳ So

↳ So, one person who has seen a particular movie; and that particular movie is may/may not be seen by other persons.

↳ To find similar person who watched the movie is difficult: So, its tough for Collaborate filtering to find similar person.

↳ This Collaborate filter is working well for Amazon, netflix because they are having more no. of users; so they will generate meaningful relations.

Similarity Metrics:-

① Adjusted Cosine:-

$$\text{CosSin}(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

ratings are different for opinion to opinion

↳ Some Country are more brutal with their ratings than others

↳ so, adjusted cosines attempts to normalize these differences.

↳ If we have less data (people who rated the movie are very less) then we don't get a meaningful average.

↳ so, the difference blw user & users mean is "0"

↳ If we have large data & many people have rated means then Adjusted Cosine is useful.

(2) (Item based) pearson Similarity:

$$\text{CosSim}(x,y) = \frac{\sum (x_i - \bar{i})(y_i - \bar{i})}{\sqrt{\sum_i (x_i - \bar{i})^2} \sqrt{\sum_i (y_i - \bar{i})^2}}$$

Here we look at difference blw ratings & the average from all users for that given item.

$\bar{i}$ → average ~~item~~ rating of the item.

↳ If we have sparse data, it is a good metric.

③ Spearman Rank Co-relation :-

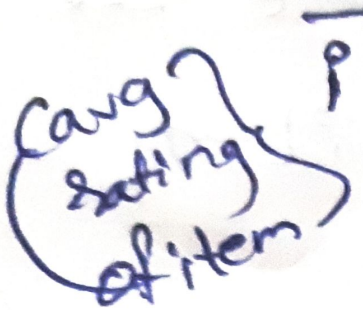
Pearson Similarity ~~ranking~~ based on ranks

not ratings

↳ same idea as pearson's Similarity

but instead of ratings scores directly

we use ranks instead.



$\bar{r}$ ⟶ rank amongst all movies

individual ratings for a movie ⟶ rank that movie among all that individual ratings

(4) Mean squared Difference

$$MSD(x,y) = \frac{\sum_{i \in I_{xy}} (x_i - y_i)^2}{|I_{xy}|}$$

$$MSD\,sim(x,y) = \frac{1}{MSD(x,y) + 1}$$

↳ Here you just directly comparing how two people rated the same set of things.

↳ ~~Its Very much same idea~~

Its same idea as mean absolute error

$|I_{xy}|$ → no. of items each user had in common that we summed accross to get the average/mean.

↳ The problem is, we computed how different users $x \& y$ are ($MSD(x,y)$)

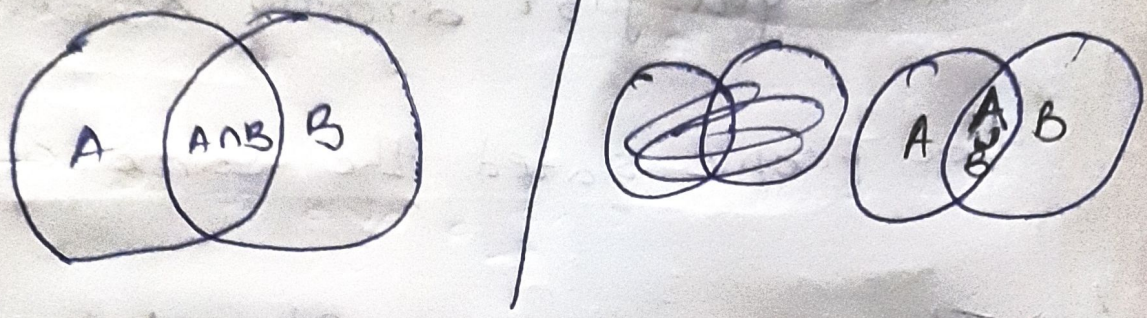
↳ To get Similarity we will take inverse +1

ie So, $MSD\,Sim\,(x,y)$

in order to avoid 0

$x \& y$ → two diff things; not diff people

(5) Jaccard Similarity :-



A & B → two users.

$$= \frac{A \cap B \text{ ratings}}{A \cup B \text{ ratings}}$$

⇒ no. of users A ratings & intersection with no. of users B ratings intersection

(as we are counting things & we are not looking the actual ratings values. So, we are throwing some information

↳ It is used for implicit ratings

- because we know their interests not by by taking their ratings (so, ratings won't come