Raeesa Youssef
Jakob Erben
Johannes Sindlinger

Problem Set 2
Mining Massive Datasets, 22/23

2022-11-06

1 Exercise

Feature	A	B	C
Processor Speed	3.06	2.68	2.92
Disk Size	500	320	640
Main-Memory-Size	6	4	6

(a) If $\alpha = \beta = 1$, all features keep the same values when scaling. So the cosinus similarity is given by

$$\begin{aligned}\cos(A, B) &= \frac{A \cdot B}{||A|| \cdot ||B||} \\ &= \frac{3.06 \cdot 2.68 + 500 \cdot 320 + 6 \cdot 4}{\sqrt{3.06^2 + 500^2 + 6^2} \cdot \sqrt{2.68^2 + 320^2 + 4^2}} \\ &= 0.9999973\end{aligned}$$

For A, C and B, C follows using the same calculation:

$$\begin{aligned}\cos(A, C) &= 0.9999953 \\ \cos(B, C) &= 0.9999878\end{aligned}$$

(b) For $\alpha = 0.01$ and $\beta = 0.5$ the adapted features yield:

Feature	A	B	C
Processor Speed	3.06	2.68	2.92
Disk Size	5	3.2	6.4
Main-Memory-Size	3	2	3

Using the same formula as above in part a), we obtain the following results:

$$\begin{aligned}\cos(A, B) &= 0.9908815 \\ \cos(A, C) &= 0.9915547 \\ \cos(B, C) &= 0.9691779\end{aligned}$$

(c) The averages of A, B, C are given as

$$\begin{aligned}avg(A) &= 2.887 \\ avg(B) &= 486.667 \\ avg(C) &= 5.333\end{aligned}$$

So $\alpha = \frac{1}{486.667} = 0,0021$ and $\beta = \frac{1}{5.333} = 0.1875$. The value for scaling feature A would be $\frac{1}{avg(A)} = \frac{1}{2.887} = 0,3464$. So the feature table changes to

Feature	A	B	C
Processor Speed	1.060	0.9284	1.0115
Disk Size	1.027	0.6575	1.3151
Main-Memory-Size	1.125	0.75	1.125

And the cosine values then result as

$$\begin{aligned}\cos(A, B) &= 0.9898552 \\ \cos(A, C) &= 0.9915270 \\ \cos(B, C) &= 0.9692788\end{aligned}$$

2 Exercise

How can a competitor - in principle - try to steal the valuable data for recommendation from this website?
He could try to extract the rating values by liking items and determining the recommended items. With the main goal of recreating the data from on which those recommendations are calculated in the first place.

Does this work better when the web shop implemented a content based or a collaborative filtering system?
This would work better with a collaborative system as the content based one is focused on the preferences of the user, which the competitor would then define himself by the first likes that he chooses. A collaborative system on the other hand draws conclusions from the whole user/item base.

What data would the competitor be able to infer?

- Clusters of simular items
- List of popular items (recommended quite heavely)
- List of unpopular item (rarely recommended)

Would this technique has an impact on the recommendation system, i.e., would this attack create a bias on the data?
Only if it is a small system with a very limited amount of users. Because only than, the impact of single preference is high. This however does not apply when comparing between items. Or when working against a collaborative system. In which case, the attacker would only confuse his own recommendations.

Why is this attack probably not viable in any case?
The target system would need to be in a quite extreme dependency for a singular user profile. This would speak against the recommender system heavily, in which case it is probably not even giving good recommendations in the very first place.

3 Exercise 3

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

(a) Taking the utility matrix as a boolean matrix, the Jaccard similarity lead to

$$\begin{aligned} JacSim(A,B) &= \frac{|A \cap B|}{|A \cup B|} \\ &= \frac{|\{b,d,e,g\}|}{|\{a,b,c,d,e,f,g,h\}|} \\ &= 0.5 \end{aligned}$$

$$\begin{aligned} JacSim(A,C) &= \frac{|\{a,d,g,h\}|}{|\{a,b,c,d,e,f,g,h\}|} \\ &= 0.5 \end{aligned}$$

$$\begin{aligned} JacSim(B,C) &= \frac{|\{c,d,f,g\}|}{|\{a,b,c,d,e,f,g,h\}|} \\ &= 0.5 \end{aligned}$$

And so the Jaccard distance is given as

$$JacDist(A,B) = JacDist(A,C) = JacDist(B,C) = 1 - 0.5 = 0.5$$

(b) For the cosine distance the missing values of ratings will be set to 0 and so the the cosine distance is given as

$$\begin{aligned}\cos(A,B) &= \frac{A \cdot B}{||A|| \cdot ||B||} \\ &= \frac{4\cdot 0 + 5\cdot 3 + 0\cdot 4 + 5\cdot 3 + 1\cdot 1 + 0\cdot 2 + 3\cdot 1 + 2\cdot 0}{\sqrt{4^2+5^2+5^2+1^2+3^2+2^2}\cdot\sqrt{3^2+4^2+3^2+1^2+2^2+1^2}} \\ &= 0.60104\end{aligned}$$

For A, C and B, C follows using the same calculation:

$$\begin{aligned}\cos(A,C) &= 0.61492 \\ \cos(B,C) &= 0.51387\end{aligned}$$

(c) The modified utility matrix using ratings of 3,4,5 as 1 and 1,2 and blank as 0 is given as

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
A	1	1	0	1	0	0	1	0
B	0	1	1	1	0	0	0	0
C	0	0	0	1	0	1	1	1

And so the Jaccard similarity yields to:

$$\begin{aligned}JacSim(A,B) &= \frac{|A \cap B|}{|A \cup B|} \\ &= \frac{|\{b,d\}|}{|\{a,b,c,d,g\}|} \\ &= \frac{2}{5}\end{aligned}$$

$$\begin{aligned}JacSim(A,C) &= \frac{|\{d,g\}|}{|\{a,b,d,f,g,h\}|} \\ &= \frac{1}{3}\end{aligned}$$

$$\begin{aligned}JacSim(B,C) &= \frac{|\{d\}|}{|\{b,c,d,f,g,h\}|} \\ &= \frac{1}{6}\end{aligned}$$

The Jaccard distance is given as follows:

$$\begin{aligned}JacDist(A,B) &= 1 - \frac{2}{5} = 0.6 \\ JacDist(A,C) &= 1 - \frac{1}{3} = 0.667 \\ JacDist(B,C) &= 1 - \frac{1}{6} = 0.833\end{aligned}$$

(d) Recalculating cosine distance using the modified matrix from part c) gives us

$$\begin{aligned}\cos(A,B) &= \frac{A \cdot B}{||A|| \cdot ||B||} \\ &= \frac{2\cdot 1}{\sqrt{4}\cdot\sqrt{3}} \\ &= 0.57735\end{aligned}$$

For A, C and B, C follows using the same calculation:

$$\begin{aligned}\cos(A,C) &= 0.50000 \\ \cos(B,C) &= 0.28867\end{aligned}$$

(e) The average values for each user considering all non-blank entries are given as:

$$avgModified(A) = 3.333$$
$$avgModified(B) = 2.333$$
$$avgModified(C) = 3.000$$

So the modified matrix by normalizing the data of the non-zero-values is given as

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
A	0.667	1.667	0	1.667	-2.333	0	-0.333	-1.333
B	0	0.667	1.667	0.667	-1.333	-0.333	-1.333	0
C	-1.0	0	-2.0	0.0	0	1.0	2.0	0.0

(f) And the corresponding cosine distance values as

$$\cos(A, B) = 0.58431$$
$$\cos(A, C) = -0.11547$$
$$\cos(B, C) = -0.73957$$

4 Exercise

(a) **Student rating website**

- Users: The students will be the different users
- Items: The system knows different kinds of items: lectures, exams, seminars and professors other likes/dislikes: passing a exam, not finishing a course (dislike), practicals with certain professors, thesis with a certain professors
- Other likes/dislikes: not liking a image (dislike), commenting on a picture(possibly with sentiment analysis), liking a user profile (small appreciation for all said artist pictures) Dating profile

(b) **Art sharing network**

- Users: the artists/art enjoyers
- Items: the art pieces
- Other likes/dislikes: not liking a image (dislike), commenting on a picture(possibly with sentiment analysis), liking a user profile (small apreciation for all said artist pictures)

(c) **Dating profile**

- Users: the users signed up to the plattform
- Items: the attribute-values of the other users // the other users
- Other likes/dislikes: block (dislike), number of messages send (like)

This case is somewhat special since users are also the items. Without the description of there dream partner and thier own answering of some questions this system would highly suffer from the new users problem.