# Mining massive Datasets WS 2017/18

**Problem Set 4**

Rudolf Chrispens, Marvin, Daniela Schacherer

November 20, 2017

## Exercise 01

We use the formula $cos\phi = \frac{a*b}{||a||*||b||}$ where $\phi$ is the angle between the vectors $a$ and $b$. The weighting vector $w$ which is multiplied to $a$ and $b$ before calculating the cosine angle is $\begin{pmatrix} 1 \\ \alpha \\ \beta \end{pmatrix}$.

a) Here we have $\alpha = 1$ and $\beta = 1$. We receive the following cosine angles, which indicate that all three vectors point in almost the same direction:

- $\phi_{AB} = 0.13°$
- $\phi_{AC} = 0.17°$
- $\phi_{BC} = 0.28°$

b) Here we have $\alpha = 0.01$ and $\beta = 0.5$. The weighted vectors are thus

$$\begin{array}{cc} & \begin{array}{ccc} A & B & C \end{array} \\ \begin{array}{c} PS \\ DS \\ MMS \end{array} & \begin{pmatrix} 3.06 & 2.68 & 2.92 \\ 5 & 3.2 & 6.4 \\ 3 & 2 & 3 \end{pmatrix} \end{array}$$

We receive the following cosine angles:

- $\phi_{AB} = 7.74°$
- $\phi_{AC} = 7.45°$
- $\phi_{BC} = 14.26°$

c) If we want to select $\alpha$ and $\beta$ as the invers proportional of the average in the respective component we receive $\alpha = \frac{1}{\frac{500+320+640}{3}} = \frac{1}{487}$ and $\beta = \frac{1}{\frac{6+4+6}{3}} = \frac{1}{5.34}$.

$$\begin{array}{cc} & \begin{array}{ccc} A & B & C \end{array} \\ \begin{array}{c} PS \\ DS \\ MMS \end{array} & \begin{pmatrix} 3.06 & 2.68 & 2.92 \\ 1.03 & 0.66 & 1.31 \\ 1.12 & 0.75 & 1.12 \end{pmatrix} \end{array}$$

With possible rounding errors during the calculation we receive for the angles:

- $\phi_{AB} = 6.01°$
- $\phi_{AC} = 5.25°$
- $\phi_{BC} = 10.67°$

## Exercise 02

Consider a web shop that sells furniture and uses a recommendation system. When a new user creates an account and likes one product, he will be presented with similar products on his next visit.

How can a competitor - in principle - try to steal the valuable data for recommendation from this website?

-Write a Bot that likes an Item (or a random set of items) and looks into the recommended list.

-Repeat this process numerous times for every item.

-Using this data gathered, the competitor can recreate his own recommendation data out of the suggestions of the web shop.

Does this work better when the web shop implemented a content- based or a collaborative filtering system?

-It works better in a content- based filtering system, because it uses only the user data of one user.

What data would the competitor be able to infer?

-The competitor could infer what kind of items would be recommended for specific users, because of the gathered data.

-He can create sets of recommendations for all combinations of the current set of data from the web shop.

Would this technique have an impact on the recommendation system, i.e., would this attack create a bias on the data?

-If the bot data of the web shop is treated as a user (not filtered out) then it will bias the data with random not useful data. For example: Like on a Deathpunk band and also Like on Justin Bieber, which would not be ”normal” user data.

Why is this attack probably not viable in any case?

-Because of the huge real user Dataset and the modern Algorithms for filtering those kind of attacks would not be relevant, they get filtered out.

## Exercise 03

a) The Jaccard distance between two sets $C_1$ and $C_2$ is defined as $d(C_1, C_2) = 1 - \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$.
For instance regarding user A and B we need to sum up the elements which have been rated by at least one of the two users (which are all given elements and thus 8) and in a second step count the elements which have been rated by both, user A and B (which is the case for element b, d, e, g and thus 4 elements). We receive: $d(C_A, C_B) = 1 - \frac{4}{8} = 0.5$. We perform this calculation for all pairs of users and get:

- $d(C_A, C_B) = 1 - \frac{4}{8} = 0.5$
- $d(C_A, C_C) = 1 - \frac{4}{8} = 0.5$
- $d(C_B, C_C) = 1 - \frac{4}{8} = 0.5$

b) The cosine distance is defined as: $sim(A,B) = cos(A,B) = \frac{A*B}{||A||*||B||}$. For every missing rating of a user a zero is inserted, thus the vector for user A would be $A = (4{,}5{,}0{,}5{,}1{,}0{,}3{,}2)$. We get the following results:

- $sim(A,B) = 0.601$
- $sim(A,C) = 0.615$
- $sim(B,C) = 0.514$

c) If we treat 3,4 and 5 as 1 (interpretation True) and 1,2, and blank as 0 (interpretation False) we receive the following values for the Jaccard distance:

- $d(C_A, C_B) = 1 - \frac{2}{5} = 0.4$
- $d(C_A, C_C) = 1 - \frac{2}{6} = 0.3$
- $d(C_B, C_C) = 1 - \frac{1}{6} = 0.17$

d) If we do the same as in c) we receive for the cosine distances:

- $sim(A,B) = \frac{2}{2*\sqrt{2}} = 0.707$
- $sim(A,C) = \frac{2}{2*2} = 0.5$
- $sim(B,C) = \frac{1}{\sqrt{2}*2} = 0.354$

e) The average user vales are: $avg_A = 2.5, avg_B = 1.75, avg_C = 2.25$. Normalizing the matrix with those averages one receives:

$$\begin{array}{c|cccccccc} & a & b & c & d & e & f & g & h \\ \hline A & 1.5 & 2.5 & - & 2.5 & -1.5 & - & 0.5 & -0.5 \\ B & - & 1.25 & 2.25 & 1.25 & -0.75 & 0.25 & -0.75 & - \\ C & -0.25 & - & -1.25 & 0.75 & - & 1.75 & 2.75 & 0.75 \end{array}$$

f) With the matrix calculated in part e) we finally end up with the following cosine distances:

- $sim(A,B) = 0.689$
- $sim(A,C) = 0.223$

- $sim(B,C) = -0.41$