University of Augsburg, Institute of Computer Science
Prof. Dr. P. Fischer
J. Kastner, L. Rudenko

SS 2019
08. Juli 2019
Solution 9

Analyzing Massive Data Sets

Exercise 1: Fuzzy IR-model (homework)

a) Determine the **Jaccard indices**:

	Augsburg	**Europe**	**soccer**	**Bundesliga**
Augsburg	1	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{4}$
Europe	$\frac{2}{3}$	1	$\frac{1}{3}$	0
soccer	$\frac{1}{4}$	$\frac{1}{3}$	1	$\frac{1}{3}$
Bundesliga	$\frac{1}{4}$	0	$\frac{1}{3}$	1

b) Compute **fuzzy degree of membership** $W(D_j, t_i)$:

fuzzy degrees of membership for D_1:

- $W(D_1, Augsburg) = 1 - (1-1) * (1-2/3) * (1-1/4) = 1$
- $W(D_1, Europe) = 1 - (1-2/3) * (1-1) * (1-1/3) = 1$
- $W(D_1, soccer) = 1 - (1-1/4) * (1-1/3) * (1-1) = 1$
- $W(D_1, Bundesliga) = 1 - (1-1/4) * (1-0) * (1-1/3) = 1/2$

fuzzy degrees of membership for D_2:

- $W(D_2, Augsburg) = 1 - (1-1/4) * (1-1/4) = 7/16$
- $W(D_2, Europe) = 1 - (1-1/3) * (1-0) = 1/3$
- $W(D_2, soccer) = 1 - (1-1) * (1-1/3) = 1$
- $W(D_2, Bundesliga) = 1 - (1-1/3) * (1-1) = 1$

fuzzy degrees of membership for D_3:

- $W(D_3, Augsburg) = 1 - (1-1/4) * (1-1) = 1$
- $W(D_3, Europe) = 1 - (1-0) * (1-2/3) = 2/3$
- $W(D_3, soccer) = 1 - (1-1/3) * (1-1/4) = 1/2$
- $W(D_3, Bundesliga) = 1 - (1-1) * (1-1/4) = 1$

fuzzy degrees of membership for D_4:

- $W(D_4, Augsburg) = 1 - (1-1) * (1-2/3) = 1$
- $W(D_4, Europe) = 1 - (1-2/3) * (1-1) = 1$
- $W(D_4, soccer) = 1 - (1-1/4) * (1-1/3) = 1/2$
- $W(D_4, Bundesliga) = 1 - (1-1/4) * (1-0) = 1/4$

c) Evaluate this query Q and determine the **Top-2** results:

- $\mu_Q(D_1) = \min(\max(1;1);(1-1)) = 0$
- $\mu_Q(D_2) = \min(\max(\frac{7}{16};\frac{1}{3});(1-1)) = 0$
- $\mu_Q(D_3) = \min(\max(1;\frac{7}{9});(1-\frac{1}{2})) = \frac{1}{2}$
- $\mu_Q(D_4) = \min(\max(1;1);(1-\frac{1}{2})) = \frac{1}{2}$

$\rightarrow$ **Top-2**: D_3, D_4

Exercise 2: Vector Space Model (live)

The solution was discussed in the exercise.

Exercise 3: Effectiveness Metrics (homework)

a) Precision, Recall and Fallout:

First of all the classification results have to be determined:

- true negatives $\bar{A} \cap \bar{B}$: 160
- false negatives $A \cap \bar{B}$: 180-160=20
- false positives $\bar{A} \cap B$: 15
- true positives $A \cap B$: 300-180-15=105

After that the evaluation metrics can be calculated easily.

- Precision: $P_Q = \frac{|A \cap B|}{|B|} = \frac{105}{105+15} = 0.875 = 87.5\%$
- Recall: $R_Q = \frac{|A \cap B|}{|A|} = \frac{105}{105+20} = 0.84 = 84\%$
- Fallout: $F_Q = \frac{|\bar{A} \cap B|}{\bar{A}} = \frac{15}{15+160} = 0.086 = 8.6\%$

b) F_1-measure: $F_\beta = \frac{(\beta^2+1)\cdot R \cdot P}{R+\beta^2 \cdot P}$, $\beta = 1$

$F_1 = \frac{(1^2+1)\cdot R \cdot P}{R+1^2 \cdot P} = \frac{2 \cdot 0.84 \cdot 0.875}{0.84+0.875} = \frac{1.47}{1.715} = 0.857 = 85.7\%$

c) In order to **maximize the F-measure**, **Precison and Recall** have to be **maximized**. In order to improve **Precision**, the **number of retrieved and relevant documents** (true negatives, documents which are retrieved but not relevant for Q) proportional to **all retrieved documents** has to increase. In reverse the **false positive** documents(Type I errors) should be **minimized**.
In order to **maximize Recall**, the **number of retrieved relevant** documents proportional to **all relevant** documents has to **increase**. So the number of **false negative documents** (Type II errors, documents which are relevant, but not retrieved for Q) have to be **minimized**.

Exercise 4: Mean Average Precision (MAP) (homework)

Q1	**r**	n	n	n	**r**	**r**	n	**r**	**r**	**r**
P	$\frac{1}{1} = \mathbf{1}$	$\frac{1}{2} = 0,5$	$\frac{1}{3} = 0,33$	$\frac{1}{4} = 0,25$	$\frac{2}{5} = \mathbf{0,4}$	$\frac{3}{6} = \mathbf{0,5}$	$\frac{3}{7} = 0,43$	$\frac{4}{8} = \mathbf{0,5}$	$\frac{5}{9} = \mathbf{0,56}$	$\frac{6}{10} = \mathbf{0,6}$
R	$\frac{1}{6} = \mathbf{0,17}$	$\frac{1}{6} = 0,17$	$\frac{1}{6} = 0,17$	$\frac{1}{6} = 0,17$	$\frac{2}{6} = \mathbf{0,33}$	$\frac{3}{6} = \mathbf{0,5}$	$\frac{3}{6} = 0,5$	$\frac{4}{6} = \mathbf{0,67}$	$\frac{5}{6} = \mathbf{0,83}$	$\frac{6}{6} = \mathbf{1}$

Q2	n	**r**	**r**	n	n	n	n	**r**	n	n
P	$\frac{0}{1} = 0$	$\frac{1}{2} = \mathbf{0,5}$	$\frac{2}{3} = \mathbf{0,67}$	$\frac{2}{4} = 0,5$	$\frac{2}{5} = 0,4$	$\frac{2}{6} = 0,33$	$\frac{2}{7} = 0,29$	$\frac{3}{8} = \mathbf{0,38}$	$\frac{3}{9} = 0,33$	$\frac{3}{10} = 0,3$
R	$\frac{0}{3} = 0$	$\frac{1}{3} = \mathbf{0,33}$	$\frac{2}{3} = \mathbf{0,67}$	$\frac{2}{3} = 0,67$	$\frac{2}{3} = 0,67$	$\frac{2}{3} = 0,67$	$\frac{2}{3} = 0,67$	$\frac{3}{3} = \mathbf{1}$	$\frac{3}{3} = 1$	$\frac{3}{3} = 1$

Q3	**r**	**r**	**r**	**r**	n	**r**	n	n	n	n
P	$\frac{1}{1} = \mathbf{1}$	$\frac{2}{2} = \mathbf{1}$	$\frac{3}{3} = \mathbf{1}$	$\frac{4}{4} = \mathbf{1}$	$\frac{4}{5} = 0,8$	$\frac{5}{6} = \mathbf{0,83}$	$\frac{5}{7} = 0,71$	$\frac{5}{8} = 0,63$	$\frac{5}{9} = 0,56$	$\frac{5}{10} = 0,5$
R	$\frac{1}{5} = \mathbf{0,2}$	$\frac{2}{5} = \mathbf{0,4}$	$\frac{3}{5} = \mathbf{0,6}$	$\frac{4}{5} = \mathbf{0,8}$	$\frac{4}{5} = 0,8$	$\frac{5}{5} = \mathbf{1}$	$\frac{5}{5} = 1$	$\frac{5}{5} = 1$	$\frac{5}{5} = 1$	$\frac{5}{5} = 1$

Average Precision Query $1 = (1 + 0,4 + 0,5 + 0,5 + 0,56 + 0,6)/6 = 0,59$

Average Precision Query $2 = (0,5 + 0,67 + 0,38)/3 = 0,52$

Average Precision Query $3 = (1 + 1 + 1 + 1 + 0,83)/5 = 0,97$

Mean Average Precision $= (0,59 + 0,52 + 0,97)/3 =$ **0,69**

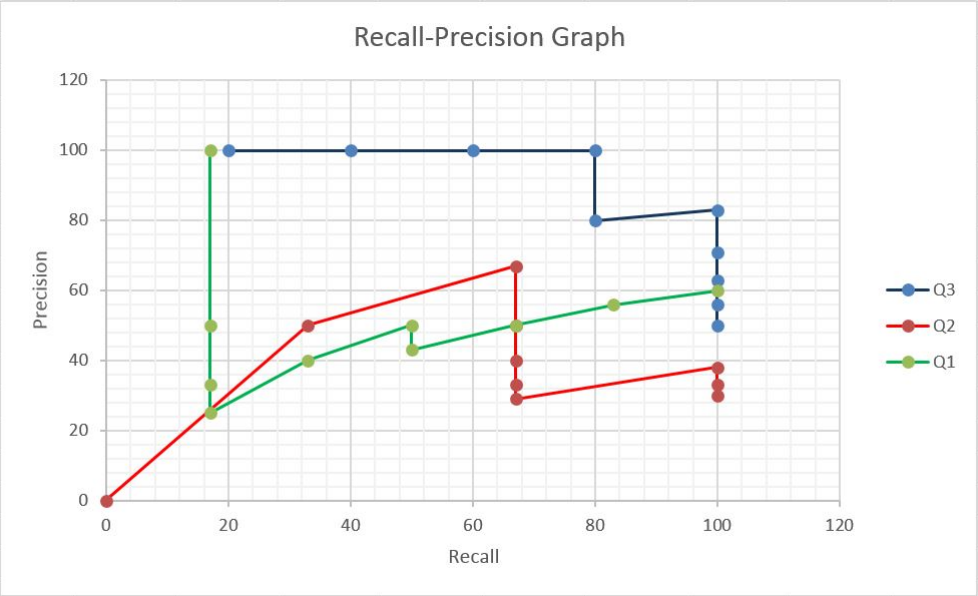

Abbildung 1: Recall-Precision Graph

Exercise 5: PageRank (live)

The solution was discussed in the exercise.