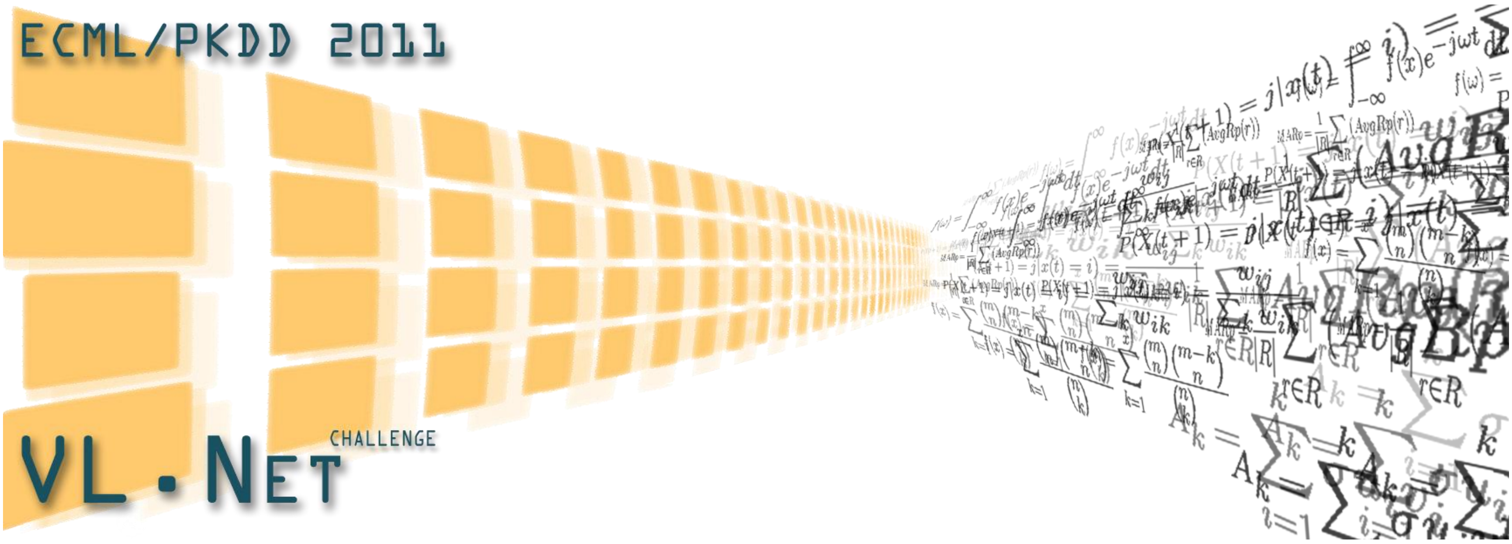

Evaluation metrics example

Mean Average R-precision

http://tunedit.org/challenge/VLNetChallenge

ECML PKDD 2011 videolectures.net exchange ideas & share knowledge rapid-i REPORT THE FUTURE

Example 1:

Let's assume that the solution S is a set of two solution lists $S = \{s_1, s_2\}$ of size ten.

s_1:

Lecture	1	2	3	4	5	6	7	8	9	10
CVS score	100	99	98	97	96	95	94	93	92	91

s_2:

Lecture	101	102	103	104	105	106	107	108	109	110
CVS score	50	49	48	47	46	45	44	43	42	41

Let's assume that $R = \{r_1, r_2\}$ is a set of two recommendation (submission) lists size ten.

r_1:

Lecture	1	2	3	20	10	6	7	8	21	22

r_2:

Lecture	101	102	103	50	30	106	107	108	109	52

To calculate average R-precision for recommendation list r_1 for cut-off lengths $Z = \{5,10\}$ we proceed as follows:

relevant$_{z=5}$=$\{1,2,3,4,5\}$; retrieved$_{z=5}$=$\{1,2,3,20,10\}$,

relevant$_{z=10}$=$\{1,2,3,4,5,6,7,8,9,10\}$; retrieved$_{z=10}$=$\{1,2,3,20,10,6,7,8,21,22\}$

$$Rp@z(\boldsymbol{r}) = \frac{|relevant \cap retrieved|_z}{|relevant\,|_z} \quad => \quad Rp@5(r_1) = \frac{3}{5}\ ; \quad Rp@10(r_1) = \frac{7}{10}$$

$$AvgRp(r_1) = \sum_{z \in Z} \frac{Rp@z(r_1)}{|Z|} = \left(\frac{3}{5} + \frac{7}{10}\right)\frac{1}{2} = 0.65$$

The same is done for the recommendation list r_2:

relevant$_{z=5}$=$\{101,102,103,104,105\}$, retrieved$_{z=5}$=$\{101,102,103,50,30\}$,

Relevant$_{z=10}$=$\{101,102,103,104,105,106,107,108,109,110\}$,
retrieved$_{z=10}$=$\{101,102,103,50,30,106,107,108,109,52\}$

$$Rp@5(r_2) = \frac{3}{5}\ ; \quad Rp@10(r_2) = \frac{7}{10}$$

$$AvgRp(r_2) = \sum_{z \in Z} \frac{Rp@z(r_2)}{|Z|} = \left(\frac{3}{5} + \frac{7}{10}\right)\frac{1}{2} = 0.65$$

Finally, we calculate mean average R-precision over all recommendations in R:

$$MARp = \frac{1}{|R|} \cdot \sum_{r \in R} AvgRp(r) = (0.65 + 0.65)\frac{1}{2} = 0.65$$

Example 2.

Assume that the solution S is a set of two solution lists $S = \{s_1, s_2\}$ of size ten or less.

s_1:

Lecture	1	2	3	4	5	6	7	8	9	10
CVS score	100	99	98	97	96	96	96	93	92	91

s_2:

Lecture	1	2	3	4	5	6	7
CVS score	50	49	48	47	46	45	44

Assume we have the following set of recommendation lists $R = \{r_1, r_2\}$ of size ten.

r_1:

Lecture	1	2	3	7	20	6	21	8	22	23

r_2:

Lecture	23	2	3	20	10	6	7	8	21	22

Average R-precision for recommendation list r_1, for cut-off lengths $Z = \{5,10\}$ is:

Relevant$_{z=5}$=$\{1,2,3,4,5,6,7\}$, retrieved$_{z=5}$=$\{1,2,3,7,20\}$,

Relevant$_{z=10}$=$\{1,2,3,4,5,6,7,8,9,10\}$, retrieved$_{z=10}$=$\{1,2,3,7,20,6,21,8,22,23\}$

As the lectures 6 and 7 have the same CVS score as the lecture 5, they are considered as relevant items for the cut-off lenght z=5. Number of relevant items at cut-off length z $|relevant\,|_z$ is defined as min(m,z), where m is the total number of relevant items for recommendation list (in this case 10).

$$Rp@z(\boldsymbol{r}) = \frac{|relevant\ \cap retrieved|_z}{|relevant\,|_z} = \frac{|relevant\ \cap retrieved|_z}{\min(m,z)},$$

$Rp@5(r_1) = \frac{4}{5}$; $Rp@10(r_1) = \frac{6}{10}$

$$AvgRp(r_1) = \sum_{z\in Z} \frac{Rp@z(r_1)}{|Z|} = \left(\frac{4}{5} + \frac{6}{10}\right)\frac{1}{2} = 0.7$$

For the recommendation list r_2 , the size of the solution list s_2 is $m = 7$.

s_2:

Lecture	1	2	3	4	5	6	7
CVS score	50	49	48	47	46	45	44

r_2:

Lecture	23	2	3	20	10	6	7	8	21	22

$\text{Relevant}_{z=5}=\{1,2,3,4,5\}$, $\text{retrieved}_{z=5}=\{23,2,3,20,10\}$,

$\text{Relevant}_{z=10}=\{1,2,3,4,5,6,7\}$, $\text{retrieved}_{z=10}=\{23,2,3,20,10,6,7,8,21,22\}$

In this case $|relevant\ |_{z=10}$, defined as min(m,z) equals 7 !

$$Rp@z(\boldsymbol{r}) \ = \frac{|relevant\ \cap retrieved|_z}{|relevant\ |_z} = \frac{|relevant\ \cap retrieved|_z}{\min(m, z)},$$

$$Rp@5(r_2) = \frac{2}{5}; \qquad Rp@10(r_2) = \frac{4}{7}$$

$$AvgRp(r_2) = \sum_{z \in Z} \frac{Rp@z(r_2)}{|Z|} = \left(\frac{2}{5} + \frac{4}{7}\right)\frac{1}{2} = 0.4857$$

Again, average R-precision over all recommendations in R is:

$$MARp = \frac{1}{|R|} \cdot \sum_{r \in R} AvgRp(r) = (0.7 + 0.4857)\frac{1}{2} = 0.5928$$

Why this measure ?

We introduced R-precision because it is more apt to our situation: it adjusts for the size of the set of relevant documents. As an example, if there were only 4 items (lectures) in the whole collection relevant to the particular query, a perfect recommender system would score 1, measured by Rp@10, whereas its p@10 would be only 0.4. Using this measure for our application makes more sense, as the number of relevant items can vary from 1 to above 30, and in such situations $Rp@z$ expresses the quality of retrieval more fairly at some predefined retrieval (cut-off) length, than $p@z$.

The reason why we use $Avg\_Rp(\boldsymbol{r})$ over set of different $Rp@z$, is that through the averaging we can take into account ranking and at the same time improve the ability to differentiate between similar solutions (recommenders).