Retrieval Evaluation

Lecture 8, September 17, 2019

Exercise #1

Compare the effectiveness of System A and System B on a test collection consisting of three queries. Table 1 contains the rankings generated by the two systems as well as the ground truth. We assume that relevance is binary, i.e., the ground truth column contains a set of the relevant documents.

Query	System A ranking	System B ranking	Ground truth
Q1	1, 2, 4, 5, 3, 6, 9, 8, 10, 7	2, 4, 3, 10, 5, 6, 7, 8, 9, 1	1, 3
Q2	1, 2, 4, 5, 3, 9, 8, 6, 10, 7	5, 6, 4, 1, 7, 8, 9, 10, 2	2, 4, 5, 6
Q3	1, 7, 4, 5, 3, 6, 9, 8, 10, 2	2, 4, 3, 7, 5, 6, 1, 8, 9, 10	7

Table 1: Document rankings produced by two systems and binary relevance judgments.

We highlighted the relevant documents in Table 1 for a better overview.

Solution

First we compute effectiveness metrics for individual queries (rows 1–3 in Table 2). Then, we average these number over the set of queries (row 4)

	System A				System B			
Query	P@5	P@10	(M)AP	(M)RR	P@5	P@10	(M)AP	(M)RR
Q1	$\frac{2}{5}$	$\frac{2}{10}$	$(\frac{1}{1}+\frac{2}{5})/2$	$\frac{1}{1}$	$\frac{1}{5}$	$\frac{2}{10}$	$(\frac{1}{3}+\frac{2}{10})/2$	$\frac{1}{3}$
Q2	$\frac{3}{5}$	$\frac{4}{10}$	$(\frac{1}{2}+\frac{2}{3}+\frac{3}{4}+\frac{4}{8})/4$	$\frac{1}{2}$	$\frac{3}{5}$	$\frac{4}{10}$	$(\frac{1}{1}+\frac{2}{2}+\frac{3}{3}+\frac{4}{10})/4$	$\frac{1}{1}$
Q3	$\frac{1}{5}$	$\frac{1}{10}$	$(\frac{1}{2})/1$	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{1}{10}$	$(\frac{1}{4})/1$	$\frac{1}{4}$
Average	0.4	0.233	0.601	0.666	0.333	0.233	0.455	0.527

Table 2: Effectiveness measures.

Exercise #3

Evaluate a given system in terms of NDCG@5 and NDCG@10 on a test collection consisting of three queries. Table 3 contains the rankings generated by the system as well as the ground truth. Documents are judged on a 4-point scale: non-relevant (0), poor (1), good (2), excellent (3).

Query	System ranking	Ground truth		
		Excellent (3)	Good (2)	Poor (1)
Q1	$2^{(1)}$, $1^{(2)}$, $3^{(0)}$, $4^{(3)}$, $5^{(0)}$, $6^{(0)}$, $10^{(0)}$, $7^{(0)}$, $9^{(0)}$, $8^{(0)}$	4	1	2
Q2	$1^{(2)}$, $2^{(1)}$, $9^{(0)}$, $4^{(3)}$, $5^{(0)}$, $6^{(0)}$, $7^{(0)}$, $8^{(1)}$, $3^{(3)}$, $10^{(0)}$	3, 4	1	2, 8
Q3	$1^{(3)}$, $7^{(2)}$, $4^{(3)}$, $5^{(2)}$, $3^{(0)}$, $6^{(1)}$, $9^{(0)}$, $8^{(1)}$, $10^{(0)}$, $2^{(0)}$	1, 4	7, 5	6, 8

Table 3: Document rankings produced by a systems and graded relevance judgments.

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i} \tag{1}$$

We added the gain value for each document in parentheses in the superscript.

Solution

Qry	gain values	DCG values	gains for perfect ranking	ideal DCG values	NDCG@5	NDCG@10
Q1	1,2,0,3,0, 0,0,0,0,0	1,3,3,4.5,4.5, 4.5,4.5,4.5,4.5,4.5	3,2,1,0,0, 0,0,0,0,0	3,5,5.6,5.6,5.6, 5.6,5.6,5.6,5.6,5.6	4.5/5.6 =0.799	4.5/5.6 =0.799
Q2	2,1,0,3,0, 0,0,1,3,0	2,3,3,4.5,4.5, 4.5,4.5,4.8,5.8,5.8	3,3,2,1,1, 0,0,0,0,0	3,6,7.3,7.8,8.2, 8.2,8.2,8.2,8.2,8.2	4.5/8.2 =0.549	5.7/8.2 =0.705
Q3	3,2,3,2,0, 1,0,1,0,0	3,5,6.9,7.9,7.9, 8.3,8.3,8.6,8.6,8.6	3,3,2,2,1, 1,0,0,0,0	3,6,7.3,8.3,8.7, 9.1,9.1,9.1,9.1,9.1	7.9/8.7 =0.907	8.6/9.1 =0.948
Avg.					0.751	0.817

Table 4: NDCG computation.

The values are in two lines, corresponding to ranks 1–5 and 6–10, for better visibility.