Retrieval Evaluation

Lecture 3, September 4, 2017

Task 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.

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)

Query	System A				System B			
	P@5	P@10	(M)AP	(M)RR	P@5	P@10	(M)AP	(M)RR
Q1								
Q2								
Q3								
Average								

Table 2: Effectiveness measures.

Task 2

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, 3, 4, 5, 6, 10, 7, 9, 8	4	1	2
Q2	1, 2, 9, 4, 5, 6, 7, 8, 3, 10	3, 4	1	2, 8
Q3	1, 7, 4, 5, 3, 6, 9, 8, 10, 2	1, 4	7, 5	6, 8

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

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

Solution

Qry	gain values	DCG values	gains perf. ranking	ideal DCG values	NDCG@5	NDCG@10
Q1						
Q2						
Q3						
Avg.						

Table 4: NDCG computation.