# 1 Evaluation Measures for ranked results

**Reminder: Precision and Recall**

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

**P/R for ranked lists** Precision and recall are single-value metrics based on the whole list of documents returned by the system. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented by providing the P/R curve.

**Average Precision** Average precision emphasizes ranking relevant documents higher. It is the average of precisions computed at the point of each of the relevant documents in the ranked sequence:

$$AP = \frac{\text{Sum of all precision values at relevant documents}}{\text{Number of relevant document in the list}}$$

**Interpolated Precision** Interpolation is used to remove the P/R curve's jiggles:

$$P_{interp}(r) = max_{r' \geqslant r}\ P(r')$$

the interpolated precision at a certain recall level $r$ is defined as the highest precision found for any recall level $r' \geqslant r$.

usually the recall levels are fixed (11 point interpolated precision – 0, 10, 20 ... 100 percent)

**Mean Average Precision** for a set of queries is the mean of the average precision scores for each query.

$$MAP = (\sum_{i=1}^{Q} AP(q))/Q$$

where Q is the number of queries and AP is the average precision.

## 2 Inter-annotator Agreement

**Kappa agreement** how much judges agree or disagree.

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$[0.8 - 1]$ (good agreement) – $[0.67 - 0.8]$ (fair agreement) – $[\cdot - 0.67]$ (dubious basis for an evaluation).

**Example**

| | | Judge 2 Relevance | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Judge 1 Relevance | Yes | 300 | 20 | 320 |
| | No | 10 | 70 | 80 |
| | Total | 310 | 90 | 400 |

Observed proportion of the times the judges agreed $P(A) = (300 + 70)/400 = 370/400 = 0.925$

Pooled marginals $P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

Probability that the two judges agreed by chance
$P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2 = 0.665$

Kappa statistic $\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$ (still in acceptable range)

# 3 Exercises

**Exercise 1** Consider an information need for which in a collection there are 4 relevant documents. Contrast two systems, A and B; their results have been judged for relevance as below.

| Rank | A | B |
|---|---|---|
| 1 | R | N |
| 2 | N | R |
| 3 | R | N |
| 4 | N | N |
| 5 | N | R |
| 6 | N | R |
| 7 | N | R |
| 8 | N | N |
| 9 | R | N |
| 10 | R | N |

Compute of each system:

- The P and R without considering the ranking.
- The P and R curve (i.e. considering the ranking.)
- The AP
- The interpolated precision
- The 11 point interpolation precision

Compare the different results.

**Exercise 2** Consider the ranked list below retrieved out of a collection of 10,000 documents. The system has retrieved 6 relevant documents but there were 8 relevant documents in the whole collection.

| Rank | Judgment |
|---|---|
| 1 | R |
| 2 | R |
| 3 | N |
| 4 | N |
| 5 | N |
| 6 | N |
| 7 | N |
| 8 | N |
| 9 | R |
| 10 | N |
| 11 | R |
| 12 | N |
| 13 | N |
| 14 | N |
| 15 | R |
| 16 | N |
| 17 | N |
| 18 | N |
| 19 | N |
| 20 | R |

- What is the P@20?
- What is the uniterpolated precision of the system at 25% recall?
- What is the interpolated precision at 33% recall?

**Exercise 3** Assume that you have run a system over 4 queries and obtained the following results

| Queries | Rank | | | | | | Tot doc rel |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1 | R | R | R | R | R | R | 6 |
| 2 | R | N | N | N | N | N | 5 |
| 3 | R | R | N | N | N | N | 5 |
| 4 | N | N | N | R | N | R | 7 |

Compute the MAP.

**Exercise 4** Below is a table showing how two human judges rate the relevant of a set of 12 documents to a particular information need. Let us assume that you've written a system that for this query returns the set of documents $\{4, 5, 6, 7, 8\}$.

| docID | Judge 1 | Judge 2 |
|---|---|---|
| 1 | N | N |
| 2 | N | N |
| 3 | R | R |
| 4 | R | R |
| 5 | R | N |
| 6 | R | N |
| 7 | R | N |
| 8 | R | N |
| 9 | N | R |
| 10 | N | N |
| 11 | N | R |
| 12 | N | R |

Calculate:

- the kappa agreement between the two judges.
- calculate the P and R of your system if a document is considered relevant only if the two judges agree.
- calculate the P and R of your system if a document is considered relevant only if either judge thinks it is relevant.