Evaluation in Information Retrieval (Chapter 8)

Definition 1 (Recall)
Recall describes how many of the relevant documents are retrieved.

$$\text{recall}(Q) = R = \frac{\#relevant\ retrieved}{\#relevant}$$

Definition 2 (Precision)
Precision describes how many of the retrieved documents are relevant.

$$\text{precision}(Q) = P = \frac{\#relevant\ retrieved}{\#retrieved}$$

Definition 3 (*F*-measure)
An F-measure (F_1-measure) defines a recall-precision relationship represented by their weighted harmonic mean:

$$\text{F} = \frac{2 \cdot R \cdot P}{R + P}$$

Definition 4 (Ranking)
Let $C = \{d_1, d_2, ..., d_n\}$ (set of documents d_i) be the corpus for document retrieval. Ranking is defined a total ordering of this set $d_a, d_b, d_c, ..., d_z$ according to index set $I = \{a, b, c, ..., z\}$, usually provided by the retrieval system.

Definition 5 (Cutoff@K)
Every metric defined in this section can be computed in non-exhaustive K-cutoff fashion, considering only top $-$ K items of the ranking. For instance

$$\text{recall@K} = \frac{\#relevant\ retrieved\ in\ top\text{-}K}{min(\#relevant, K)}$$

Example 1 (Precision@25 and Recall@25)

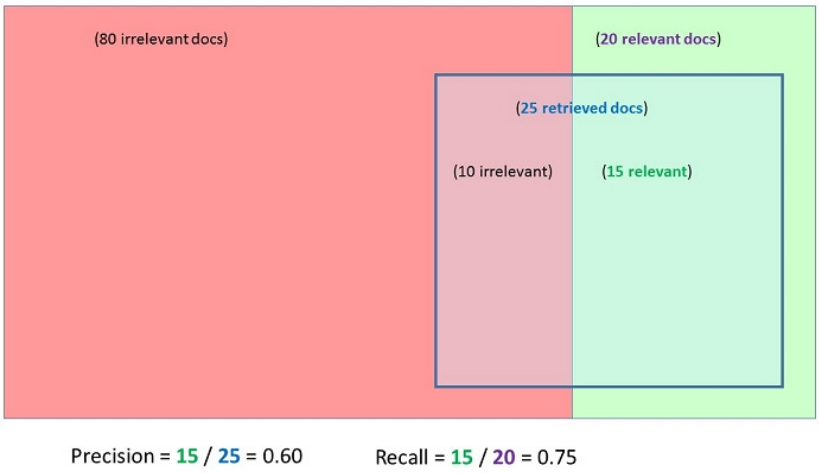

Figure 1: Example of precision@25 and recall@25.

Definition 6 (Mean Reciprocal Rank)
MRR expresses the average reciprocal rank of the first relevant item. Consider query $q \in Q$ from dataset of queries Q and $rank_q$ being the index of the first relevant item in ranking.

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{rank_q}$$

Example 2 (MRR)

Example of MRR computation is shown below.[1]

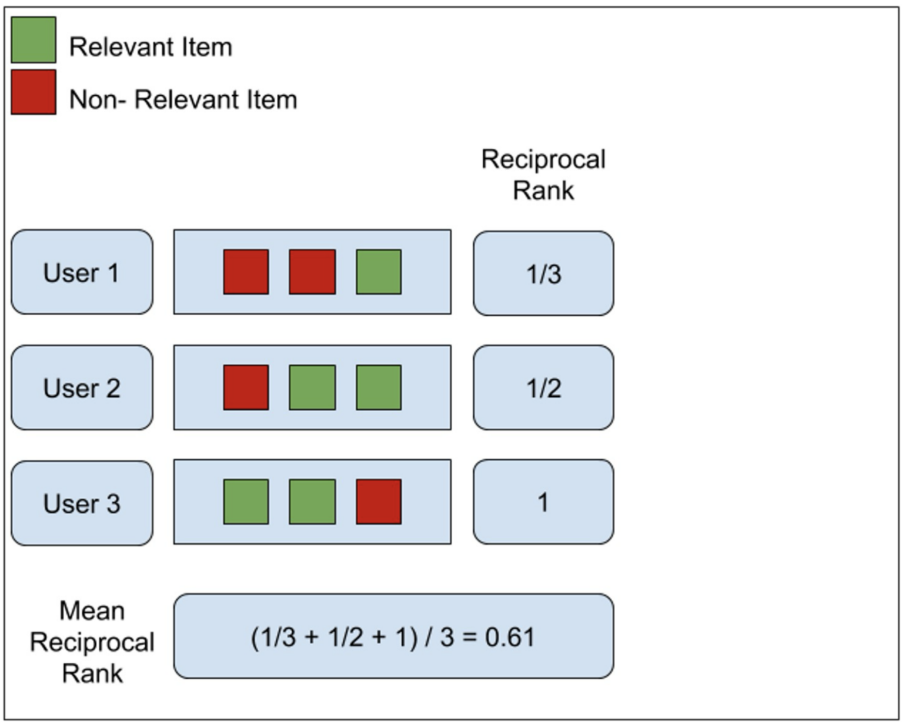

Figure 2: Computation of mean reciprocal rank.

Definition 7 (Mean Average Precision)
MAP expresses the precision in each point a new relevant document is included in the result. For each query q from the set of all queries Q:

- *let r_q be the ranking for query q*
- *let $\mathbb{I}[p]$ be an indicator function, producing 1 iff predicate p holds, else 0*
- *let R_q be number of documents relevant to q*
- *let $Precision_q@K$ denote the precision for single query q cutoff at K*

[1]Source: `https://medium.com/swlh/rank-aware-recsys-evaluation-metrics-5191bba16832`

then define ***average precision***

$$\text{AveP}(q) = \frac{1}{R_q} \sum_{i=1}^{n} Precision_q@K \cdot \mathbb{I}[r_q(i) \in R_q]$$

and define ***mean average precision*** *as*

$$\text{MAP} = \frac{1}{Q} \sum_{q \in Q} \text{AveP}(q)$$

where rel_q is the number of relevant documents for query q and $prec_i$ is the precision at the i-th relevant document.

Example 3 (MAP)

Example of MAP computation is shown below.[2]

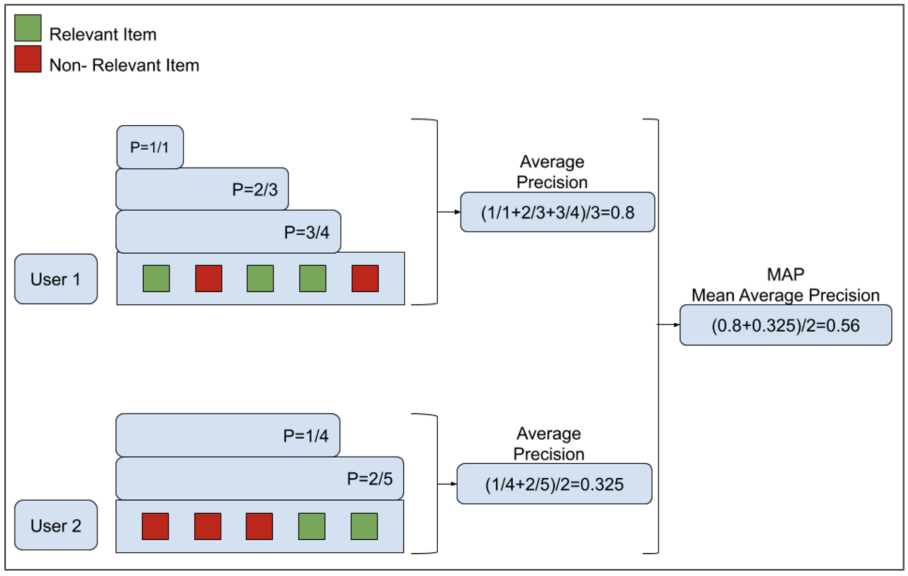

Figure 3: Computation of mean average precision.

Definition 8 (Normalized Discounted Cumulative Gain)
Let r_q be the ranking for query q and assume graded relevance annotation—the relevance of k-th item in the ranking $r_q(k)$ (so-called gain for graded relevance) isn't binary, but graded e.g., from 1 to 5. nDCG@K allows to asses the performance of system using graded annotation. We can derive the metric while expanding definitions as follows: Define ***Cumulative Gain@K*** *as*

$$\text{CG@K}(q) = \sum_{i=1}^{K} r_q(i).$$

Next define ***Discounted Cumulative Gain*** *as*

$$\text{DCG@K}(q) = \sum_{i=1}^{K} \frac{2^{r_q(i)} - 1}{\log_2(i+1)}.$$

[2]Source: `https://medium.com/swlh/rank-aware-recsys-evaluation-metrics-5191bba16832`

Notice the DCG is unbounded. Queries with more annotations would thus influence the mean aggregate through their higher expected magnitude. To fix this, next let $REL@K$ be the ideal set of items contained in ideal relevance ranking cutoff at K.
Define ***Ideal Discounted Cumulative Gain***

$$\text{IDCG@K}(q) = \sum_{i=i}^{|REL@K|} \frac{2^{r_q(i)} - 1}{\log_2(i+1)}.$$

Now, the ***Normalized Discounted Cumulative Gain*** *can be defined as*

$$nDCG@K = \frac{1}{|Q|} \sum_{q \in Q} \frac{DCG@K(q)}{IDCG@K(q)}$$

Example 4 (nDCG)

Example of nDCG computation is shown below.[3] Notice this is slightly different formulation of nDCG then what we defined. What is the difference?

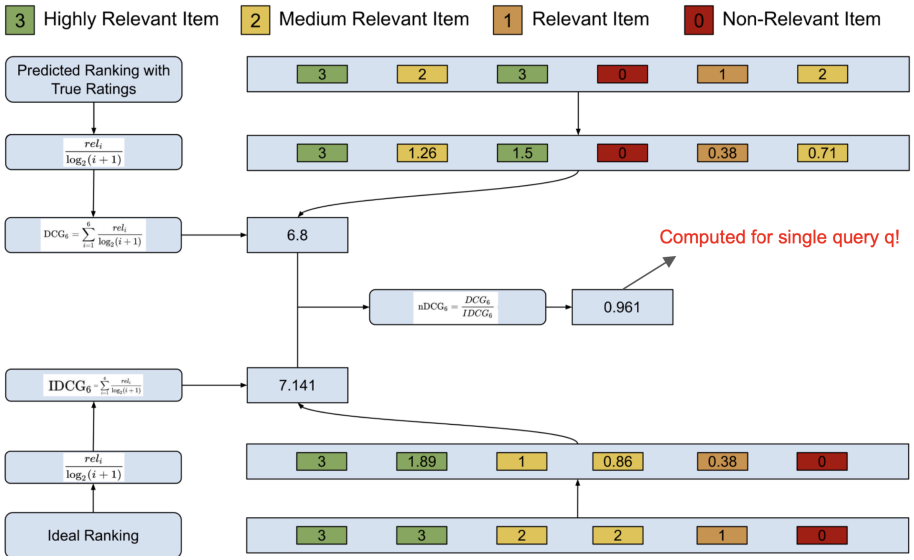

Figure 4: Computation of nDCG.

Definition 9 (κ statistic/ Cohen's kappa coefficient)
Let N be the total number of documents, J is a set of judges, let $|J| = 2$, and $P(A) = \frac{\#agree}{N}$ the proportion of documents on which the judges agree. Let also define R_j and NR_j be the proportion of relevant and non-relevant documents, respectively, according to the judge $j \in J$ and

$$P(R) = \frac{\sum_{j \in J} R_j}{|J| \cdot N} \quad \text{and} \quad P(NR) = \frac{\sum_{j \in J} NR_j}{|J| \cdot N}$$

as the number of relevant and non-relevant documents, respectively. Let finally define

$$P(E) = P(R)^2 + P(NR)^2$$

[3]Source: `https://medium.com/swlh/rank-aware-recsys-evaluation-metrics-5191bba16832`

is the proportion of the documents the judges would be expected to agree by chance. Then the κ statistic is defined as the measure of agreement between the judges

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

Exercise 8/1

The following ordered list of 20 letters R and N represents relevant (R) and non-relevant (N) retrieved documents as an answer for a query on a collection of 10 000 documents. The leftmost document is expected to be the most relevant. The list contains 6 relevant documents. Assume that the collection contains 8 documents relevant to the query.

$$RRNNNNNNRNRNNNRNNNNR$$

a) What is the precision on the first 20 results?

b) What is the F-measure on the first 20 results?

c) What is the non-interpolated precision of the system at 25% recall? (R=25%)

d) What is the interpolated precision of the system at 33% recall? (R>33%)

e) Assume that these 20 documents are the complete list of retrieved documents. What is the MAP@20 of the system?

Now assume that the system returned all 10,000 documents in an ordered list and above is the top 20.

f) What is the highest possible MAP the system can achieve?

g) What is the lowest possible MAP the system can achieve?

h) What is the MRR@20 of the system?

Exercise 8/2

The following ordered list of 5 letters R and N represent relevant (R) and non-relevant (N) retrieved documents as an answer for a query on a collection of 100 documents. The leftmost document is expected to be the most relevant. The list contains 3 relevant documents. Assume that the collection contains 5 documents relevant to the query.

$$RNNRR$$

a) What is the F-measure on the first 5 results?

Assume that these 5 documents is the complete list of retrieved documents.

b) What is the MAP of the system?

Now assume that the system returned all 100 documents in an ordered list and above is the top 5.

c) What are the highest and lowest possible MAPs the system can achieve?

Exercise 8/3

The following two sequences of letters R and N represent the complete lists of relevant (R) and non-relevant (N) retrieved documents as answers for two queries on a collection of 100 documents. The leftmost document is expected to be the most relevant. Assume that the collection contains 10 documents relevant to the first query and 20 documents relevant to the second query. Find the F-measure and the MAP of this system.

$$NRNNNRN \quad \text{and} \quad NNRRR$$

Exercise 8/4

Below is a table showing how two judges judged the relevance (0 = non-relevant, 1 = relevant) of the set of 12 documents with respect to a query. Assume that you developed an IR system, that for this query returns the documents $\{4, 5, 6, 7, 8\}$.

Doc ID	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

Table 1: Judges judging the relevance of documents.

a) Calculate the κ statistic.

b) Calculate the recall, precision and F-measure of your system in which a document is considered relevant if the judges agree.

c) Calculate the recall, precision and F-measure of your system in which a document is considered relevant if at least one of the judges thinks so.
