# Batch Evaluation Measures

Evaluation, session 3

# Relevance Judgements

The goal of retrieval effectiveness evaluation is to rank IR systems.

In order to compare them, we use standard document collections with queries for which relevance judgements have already been collected.

In recall-oriented retrieval, the judgements are typically binary. Precision-oriented retrieval often uses graded relevance judgements.

- **Grade 0:** Non-relevant documents. These documents do not answer the information need.
- **Grade 1:** Somewhat relevant documents. These documents are on the right topic, but have incomplete information.
- **Grade 2:** Relevant documents. These documents do a reasonably good job of answering the query, but the information might be slightly incomplete.
- **Grade 3:** Highly relevant documents. These documents are an excellent reference on the information need and completely answer it.
- **Grade 4:** Nav documents. These documents are the “single relevant document” for navigational queries.

**Possible Relevance Grade Scheme**

# Relevance Judgement Ambiguity

Expert human judges often disagree on the correct relevance grade for a document.

- They may have different interpretations of the information need.
- They may have different understandings of the document.
- They may disagree on whether a document is “relevant” or “highly relevant.”

However, studies so far suggest that this has a negligible affect on the system ranking.

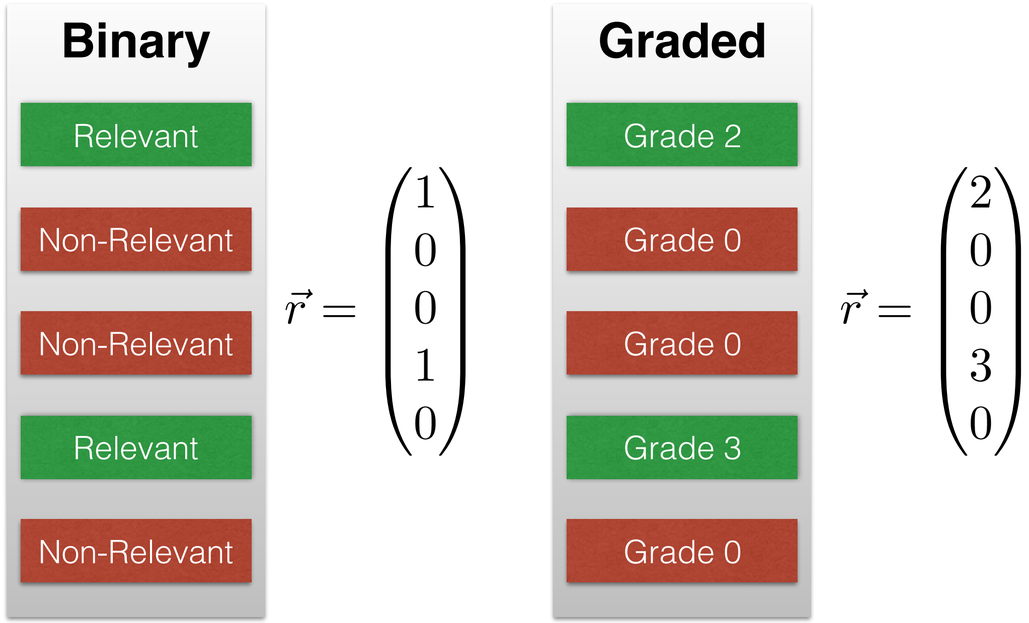
# Evaluating Rankings

Given a ranking and a relevance grade for each ranked document, we build a vector of relevance grades to use for evaluation.

Given a binary vector (and, for recall, the total number of relevant documents $R$):

$$recall(\vec{r}, R) = \frac{1}{R} \sum_i \vec{r}_i$$

$$precision(\vec{r}) = \frac{1}{|\vec{r}|} \sum_i \vec{r}_i$$




**Converting binary and graded rankings into vectors of grades**

# F-Measure

*F-Measure* combines both recall and precision, so systems that favor are penalized for whichever is lower.

$$F(\vec{r}, R, \beta) = \frac{(\beta^2 + 1) \cdot \mathit{precision}(\vec{r}) \cdot \mathit{recall}(\vec{r}, R)}{(\beta^2 \cdot \mathit{precision}(\vec{r})) + \mathit{recall}(\vec{r}, R)}$$

The commonly-used *F1-Measure* is the harmonic mean of recall and precision.

$$\begin{aligned} F1(\vec{r}, R) &= F(\vec{r}, R, 1) \\ &= \frac{2 \cdot \mathit{precision}(\vec{r}) \cdot \mathit{recall}(\vec{r}, R)}{\mathit{precision}(\vec{r}) + \mathit{recall}(\vec{r}, R)} \end{aligned}$$

**Example**

$$\vec{r} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{aligned} \mathit{precision}(\vec{r}) &= 0.4 \\ \mathit{recall}(\vec{r}, 20) &= 0.1 \end{aligned}$$

$$\begin{aligned} F1(\vec{r}, 20) &= \frac{2 \cdot 0.4 \cdot 0.1}{0.4 + 0.1} \\ &= 0.16 \end{aligned}$$

# R-Precision

As you move down the ranked list, recall increases monotonically. Precision goes up and down, with a general downward trend.

*R-Precision* is the value of recall and precision at the rank where they are equal.

$$rprecision(\vec{r}, R) := precision@k(\vec{r}, R)$$

Note:

$$precision@k(\vec{r}, R) = recall@k(\vec{r}, R, k)$$

**Example**

$$\vec{r} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

$$precision@k(\vec{r}, k = 5) = 0.4$$
$$recall@k(\vec{r}, R = 5, k = 5) = 0.4$$
$$rprecision(\vec{r}, R = 5) = 0.4$$

# Average Precision

*Average Precision* combines the precision at relevant documents, so it combines recall and precision in a different way.

It is the mean of the $precision@k$ scores for every rank containing a relevant document.

$$ap(\vec{r}, R) = \frac{1}{R} \sum_{k:\vec{r}_k=1} precision@k(\vec{r}, k)$$

**Example**

$$\vec{r} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \qquad prec@k = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/2 \\ 2/5 \end{pmatrix}$$

$$\begin{aligned} ap((1,0,0,1,0)^T, 2) &= \frac{1}{R} \sum_{k:\vec{r}_k=1} precision@k(\vec{r}, k) \\ &= 0.5 \cdot (1 + 0.5) \\ &= 0.75 \end{aligned}$$

# Precision-Recall Curves

A precision-recall curve, or PR-curve, plots precision versus recall at increasing ranks.

The red line is an interpolated version of the plot. It plots recall versus the maximum precision for any higher rank.

AP is approximately the area under the interpolated PR curve. R-precision (rp) is the area under the piecewise linear approximation connecting (0,1) to (rp, rp) and (rp, rp) to (1, 0).

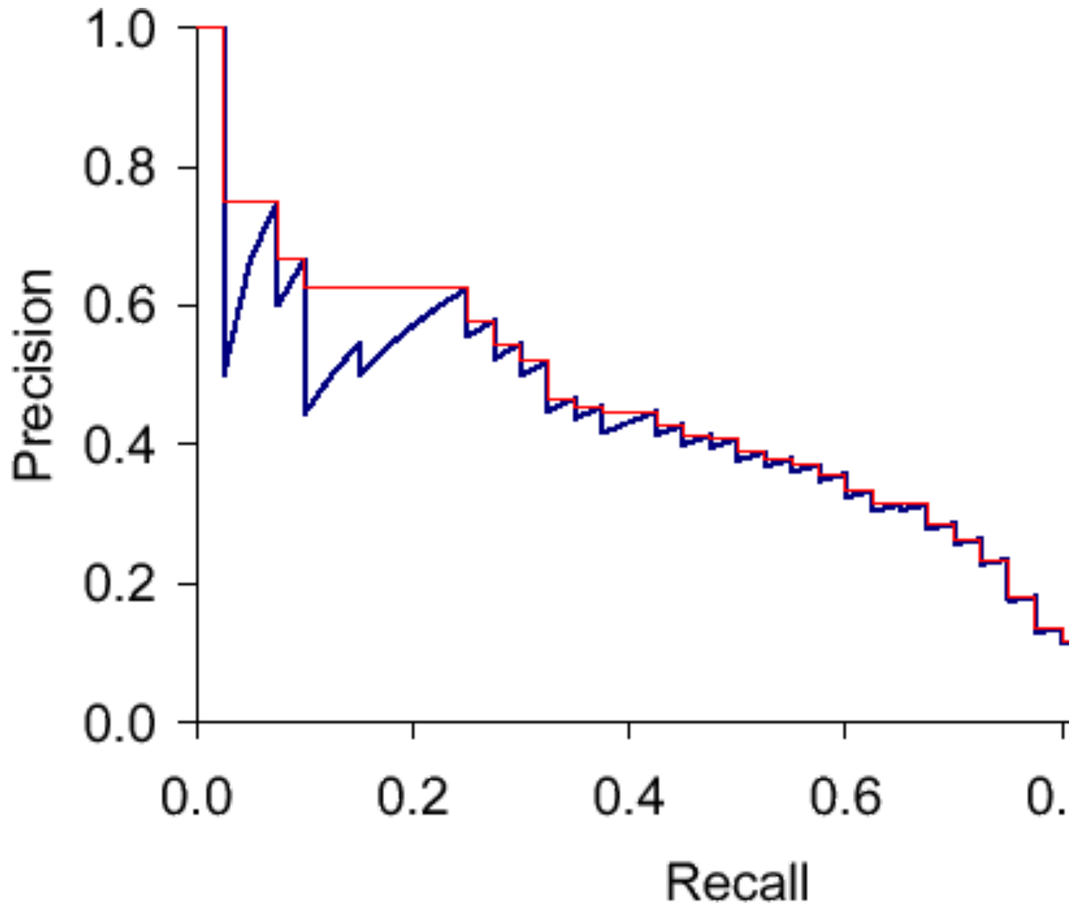



Image courtesy http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html

# Reciprocal Rank

*Reciprocal Rank* is the reciprocal of the rank of the first relevant document. It's equivalent to average precision when there is one relevant document.

It's commonly used for evaluating NAV queries, or high-precision queries.

$$rr(\vec{r}) = \frac{1}{\arg\min_k \{\vec{r}_k \neq 0\}}$$

**Example**

$$\vec{r} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \qquad \begin{aligned} rr(\vec{r}) &= \frac{1}{1} \\ &= 1 \end{aligned}$$

# Discounted Cumulative Gain

*DCG* is used for graded relevance judgments, but can't be compared across different queries.

$$dcg(\vec{r}, k) := r_1 + \sum_{i=2}^{k} \frac{r_k}{\lg i}$$

The normalized version, *nDCG*, fixes that by normalizing with the DCG of the sorted ("ideal") list.

$$ndcg(\vec{r}, k) := \frac{dcg(\vec{r}, k)}{dcg(\textbf{sort-desc}(\vec{r}), k)}$$

**Example**

$$\vec{r} = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 3 \\ 0 \end{pmatrix}$$

$$\begin{aligned} dcg(\vec{r}, 5) &= 2 + \frac{0}{\lg 2} + \frac{0}{\lg 3} + \frac{3}{\lg 4} + \frac{0}{\lg 5} \\ &= 2 + 3/2 \\ &= 3.5 \end{aligned}$$

$$\begin{aligned} ndcg(\vec{r}, 5) &= \frac{dcg(\vec{r}, 5)}{dcg((3, 2, 0, 0, 0)^T, 5)} \\ &= 3.5 / \left( 3 + \frac{2}{\lg 2} + \frac{0}{\lg 3} + \frac{0}{\lg 4} + \frac{0}{\lg 5} \right) \\ &= 0.7 \end{aligned}$$

# Wrapping Up

The measures seen here are the most common, but there are many more to choose from. How do you pick?

- F-measure forces you to optimize for both precision and recall, and lets you choose their relative importance.
- RP and AP are recall-oriented, and approximate the area under the PR curve.
- RR and NDCG are precision-oriented. RR is stricter, but NDCG considers more documents in the list.

Next, we'll try to shed some light on what these measures imply about how users interact with a ranked list.