Information Retrieval (5LN712)

Evaluation in Information Retrieval

Ali Basirat

Department of Linguistics and Philology
Uppsala University

April 15, 2020

Table of contents

Evaluation Needs

- We want to know which IR technique is more effective for an application
- We need a test collection consisting of:
  - A document collection
  - A test suite of information needs, expressible as queries
  - A set of relevance judgements, a binary assessment of either relevant or non-relevant for each document-query pair

Table of contents

Gold Standard

- A document in the test collection is either relevant or non-relevant
- The relevancy class of a document is determined based on a user information need
- The decision about the relevancy of a document is referred to as the gold standard or ground truth judgment of relevance

Gold Standard

Information need versus query

- Information need: A complete sentence describing what an end user is looking for
- Query: a translation of an information need into terms and Boolean operators

Example

- Information need: *Information on whether drinking red wine is more effective at reducing your risk of heart*
- wine AND red AND white AND heart AND attack AND effective

Gold Standard

Information need versus query

- Relevance is assessed with respect to an information need, not a query.
- An information need can be expressed with different queries
- If a document addresses an information need, then it is a relevant document to the information need and its query

Gold Standard

Data Split

- An IR model may have many tuning parameters
- One or more development test collections should be used to tune the parameters of a model
- A test collection is used for final evaluation
- The IR model may overfit to its development set
- Hence an unused test collection is needed to report the actual performance of the model
- In ML terminology, the data is split to train, validation (development), and test data

Table of contents

Standard Test Collections

- Cranfield: A small collection consisting of 1398 abstract aerodynamics journal articles, a set of 255 queries, and exhaustive relevance judgements of all (query, document) pairs.
- Text REtrieval Conference (TREC) consists of many tracks over different test collections. The best known test collections are the ones used for the TREC Ad Hoc track with millions of documents and relevance judgements for 450 information needs, called topics. This collection itself is divided into multiple sub collections in which the collection of Foreign Broadcast Information Service articles is the largest and the topics are more consistent.

Standard Test Collections

Cont.

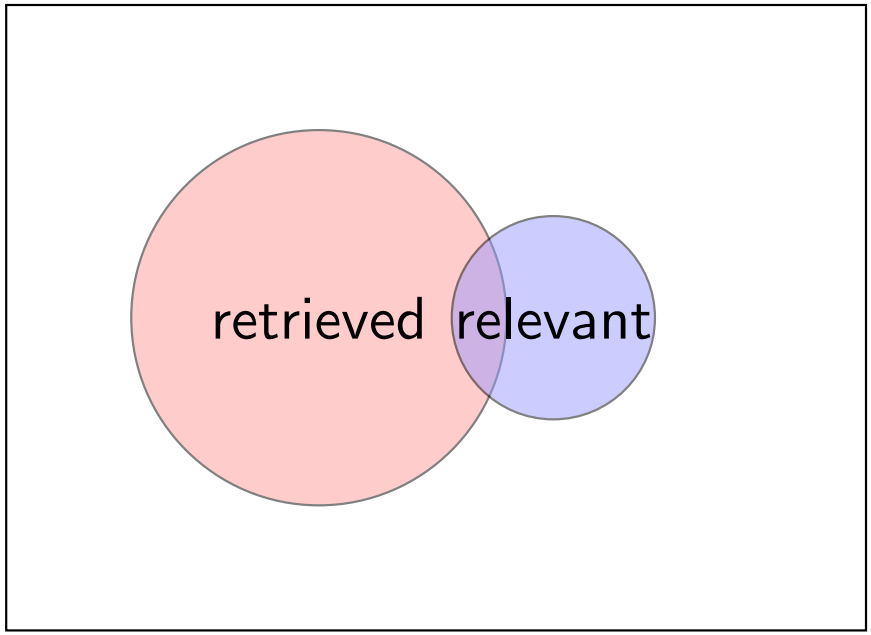
- Gov2: a large collection of web pages. This collection is the largest web collection for the research purpose.
- Reuters: a collection of news wire articles with a rich annotations.
- For more detailed information about test collections read Sec. 8.2 of the book.

Table of contents

Evaluation of Unranked Retrieval Sets

The two most frequent evaluation metrics:

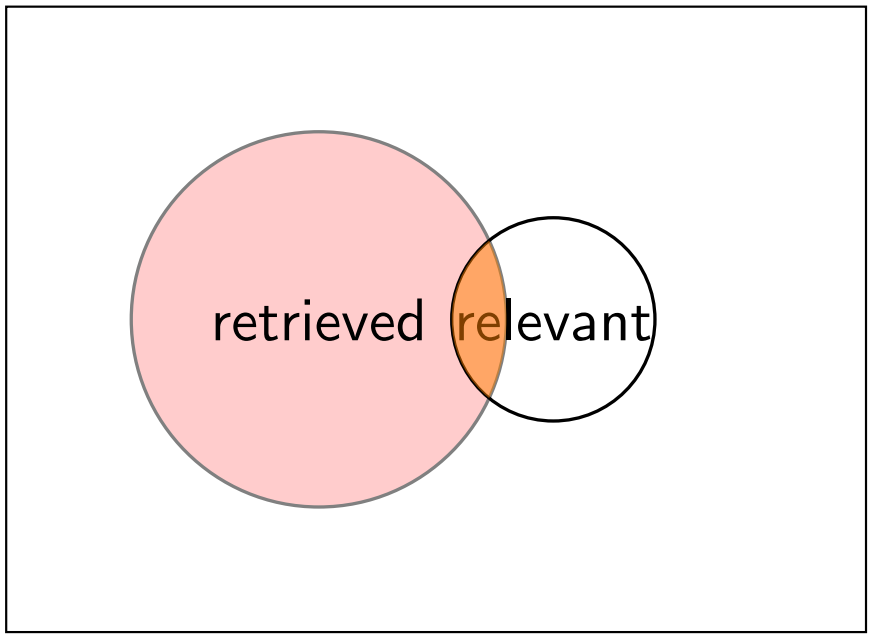
- Precision: what fraction of the retrieved documents are relevant to the information need?
- Recall: what fraction of the relevant documents in the collection are retrieved by the system.



Evaluation of Unranked Retrieval Sets

Precision

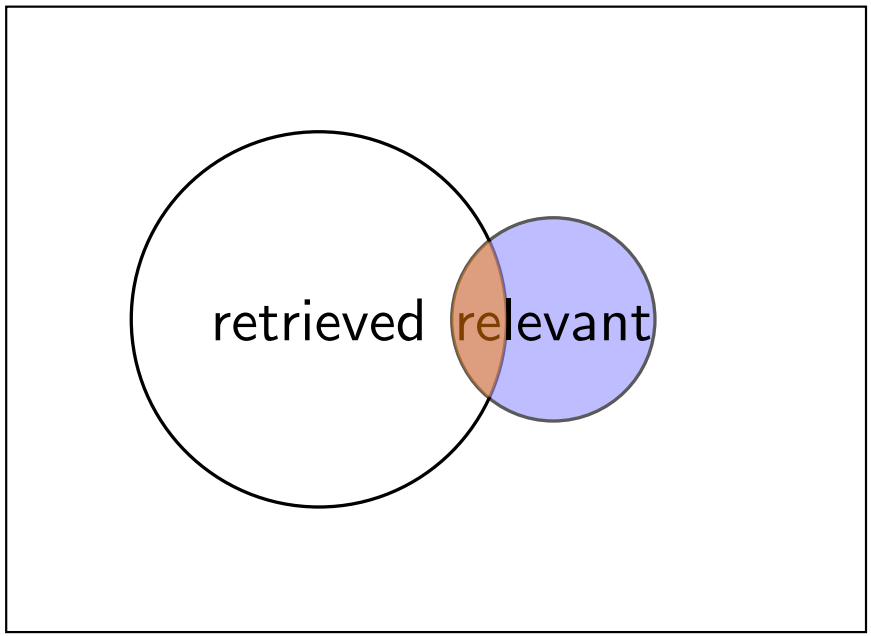
$$\begin{aligned}\text{Precision} &= P(\text{relevant}|\text{retrieved}) \\ &\approx \frac{\#\text{relevant items retrieved}}{\#\text{retrieved items}}\end{aligned}$$



Evaluation of Unranked Retrieval Sets

Recall

$$\begin{aligned}\text{Recall} &= P(\text{retrieved}|\text{relevant}) \\ &\approx \frac{\#\text{relevant items retrieved}}{\#\text{relevant items}}\end{aligned}$$



Evaluation of Unranked Retrieval Sets

Precision & Recall

- A high value of precision means a large part of retrieved documents are relevant
- web surfers would like every result on the first page to be relevant (high precision)
- a high value of recall means a large part of relevant documents are retrieved
- professional searchers are very concerned with trying to get as high recall as possible

Evaluation of Unranked Retrieval Sets

Precision & Recall

From a binary classification point of view:

- Positive: the relevant documents
- Negative: the non-relevant documents
- True: a document is classified correctly
- False: a document is classified incorrectly

	Relevant	Non relevant
Retrieved	TP	FP
Not retrieved	FN	TN

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Evaluation of Unranked Retrieval Sets

F-score

- The weighted harmonic mean of precision and recall
- $F = \frac{1}{\frac{\alpha}{P} \times \frac{1-\alpha}{R}}$ where $\alpha \in [0, 1]$
- $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$ where $\beta^2 = \frac{1-\alpha}{\alpha}$ and $\beta^2 \in [0, \infty)$
- $F_{\beta=1} = \frac{2PR}{P+R}$ gives equal weights to precision ans recall $(\alpha = 0.5)$
- A $\beta < 1$ emphasizes precision
- A $\beta > 1$ emphasizes recall

Evaluation of Unranked Retrieval Sets

Accuracy

- The fraction of correctly classified documents
- accuracy $= \frac{TP+TN}{N}$ where N is the total number of documents in the collection
- In reality, most of the documents (99.9%) are non-relevant, so it would be relatively easy to find non-relevant documents
- If an IR system is tuned to maximize its accuracy, it is likely to classify all documents as non-relevant
- It means the system will have a high TN
- Users want to see some documents (TP+FP) with a relatively large TP and small FP

Evaluation of Unranked Retrieval Sets

Accuracy

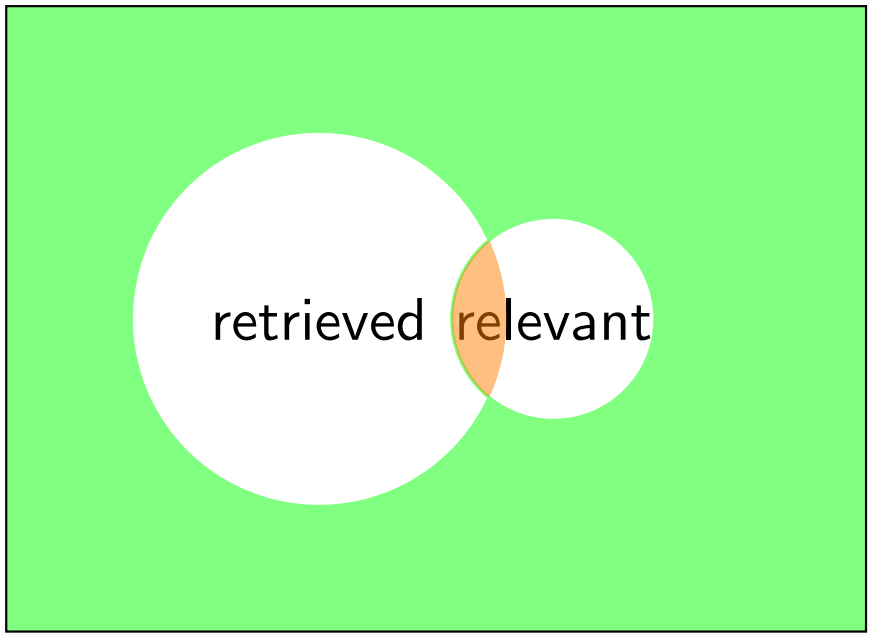


Figure: The accuracy is the fraction of TP (orange) plus TN (green) to the whole area.

Table of contents

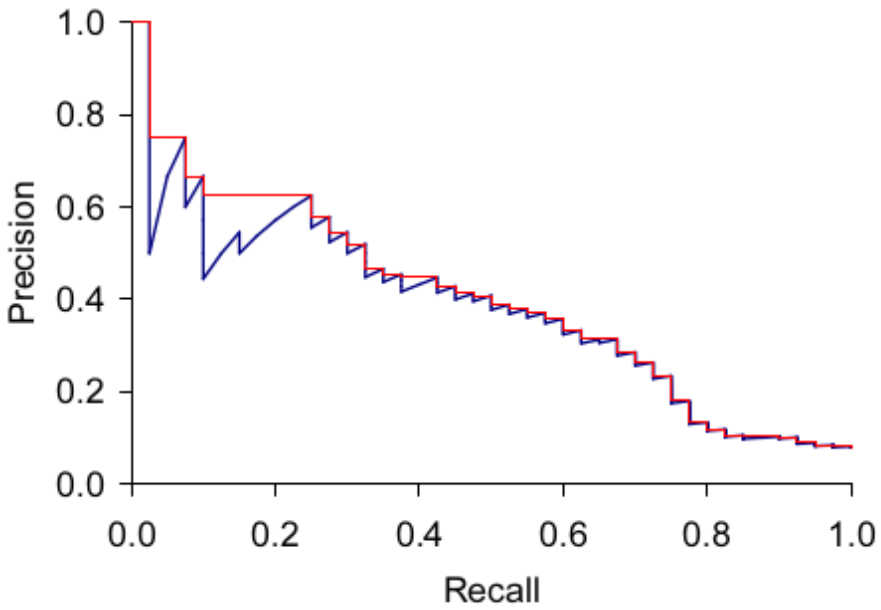
Evaluation of Ranked Retrieval Results

- Precision and recall are set-based measures with no regard to the order of the documents
- In ranked retrieval, the order of retrieved documents is important
- We need to extend them to evaluate the ranked retrieval

Evaluation of Ranked Retrieval results

Precision-recall Curve

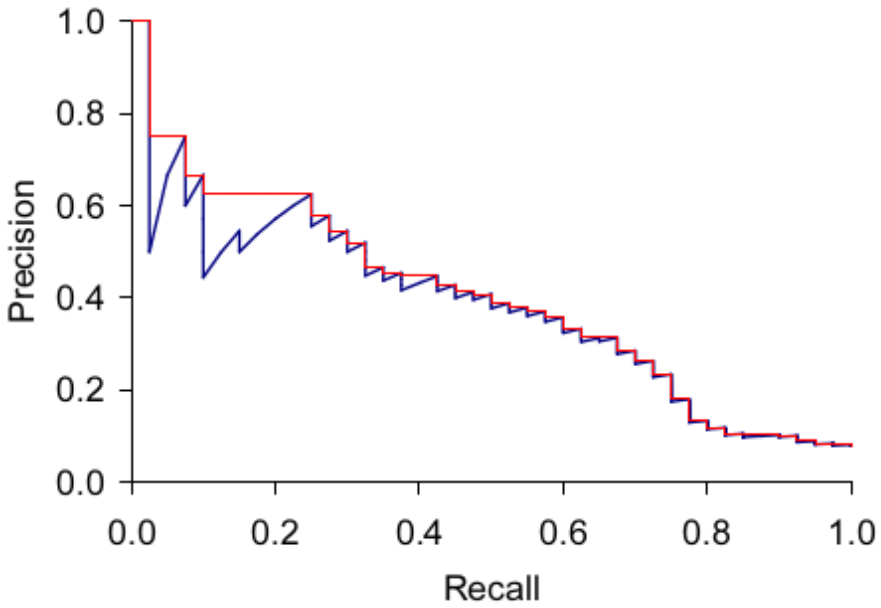
- In ranked retrieval, we evaluate a set of top k retrieved documents
- The plot of the precision and recall values of such sets



Evaluation of Ranked Retrieval results

Precision-recall Curve

- Once a document is retrieved, it is either relevant (TP) or non-relevant (FP)
- If the $(k+1)^{\text{th}}$ document retrieved is non-relevant (FP), the precision decreases, but the recall remains with no change
- If the $(k+1)^{\text{th}}$ document retrieved is relevant (TP), then both precision and recall increase



Evaluation of Ranked Retrieval results

Interpolated Precision

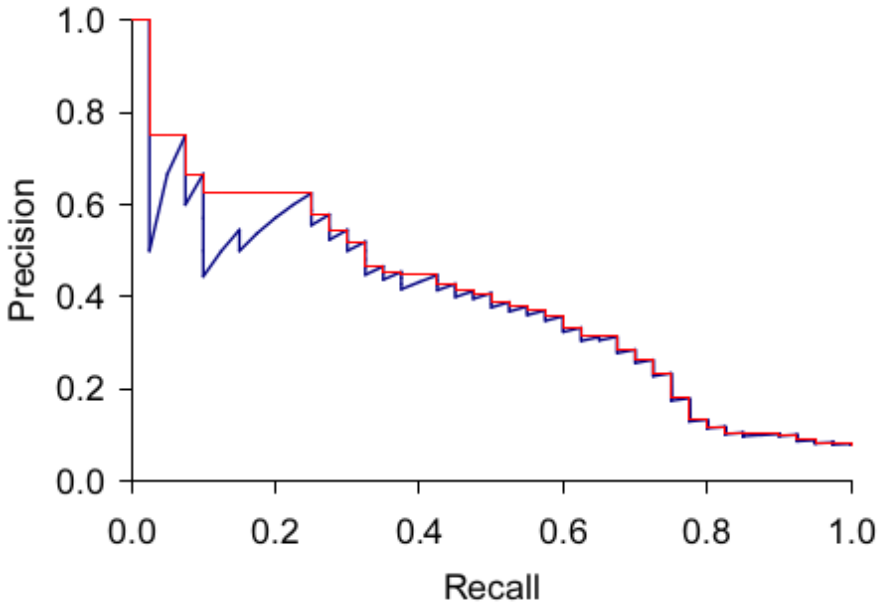
- We want to remove the jiggles of the precision-recall curve
- The user are prepared to look at a few more documents if the precision become higher
- The interpolated precision at a recall value of r is the highest precision for any recall value $r' \geqslant r$

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

Evaluation of Ranked Retrieval results

Interpolated Precision

The red lines show the interpolated precision



Evaluation of Ranked Retrieval results

11-point interpolated average precision

For each information need, the interpolated precision is measured at the 11 recall levels $0.0, 0.1, \ldots, 1.0$.

R	P_{interp}
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

Table: Interpolated precision at 11 recall levels

Evaluation of Ranked Retrieval results

11-point interpolated average precision

For each recall level, we calculate the arithmetic mean of the interpolated precision at that recall level for each information need in the test collection

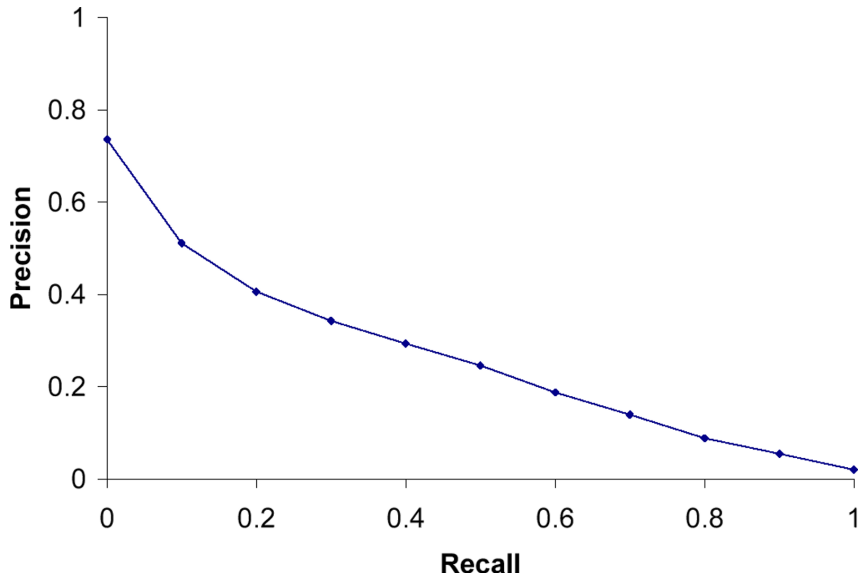


Figure: Averaged 11-point precision/recall graph across 50 queries for a representative TREC system.

Evaluation of Ranked Retrieval results

Average Precision (AP)

- The average precision of a single information need is the average of the precision value obtained for the set of top k documents existing after each document is retrieved
- Let $d_1, \ldots, d_{mj}$ be the set of relevant documents for the information need $q_j \in Q$
- Let R_{jk} be the set of ranked retrieval results from the top results until the document d_k is included
- The average precision of q_j is:

$$\text{AP}(q_j) = \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

Evaluation of Ranked Retrieval results

Mean Average Precision (MAP)

- The mean average precision (MAP) is a single-figure quality metric across recall levels
- The mean average precision is the average of the average precision (AP) over all information needs:

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \text{AP}(q_j)$$

Evaluation of Ranked Retrieval results

Mean Average Precision (MAP)

- MAP is not based on fixed recall levels
- MAP is not based on the precision interpolation
- MAP values vary widely across different information needs for a system
- MAP values agree more with a single information need across different systems rather than different information needs for the same system
- This means: MAP is an effective metric if it is applied to a test set that is large and diverse enough
- If the information needs are similar then MAP results in relatively similar results for all systems

Evaluation of Ranked Retrieval results

Precision@k

- MAP is an average over all recall levels
- In web search, users want many good results in the first page
- In web search, the recall is not an effective measure since the size of the relevant documents can be very large
- The web users want high precision at a small number of retrieved documents, say 10 or 30 documents
- Precision at k (precision@k) corresponds to the number of relevant documents among the top k documents
- Advantage: no need to know the size of the relevant documents
- Disadvantage: it is highly influenced by the total number of relevant documents; if the number of relevant documents is smaller than k, the precision@k is always less than 1

Evaluation of Ranked Retrieval results

R precision

- R-precision is based on a set of known relevant documents, *Rel*
- We calculate the precision of the top *Rel* documents returned
- For a set *Rel* of relevant documents, we examine top $|\text{Rel}|$ documents retrieved by the system.
- If r documents out of the examined documents are relevant, then R-precision is $\frac{r}{|\text{Rel}|}$
- The R-precision and the recall of the top $|\text{Rel}|$ retrieved are equal

Evaluation of Ranked Retrieval results

Precision@k and R-precision

- Both precision@k and R-precision describe a single point on the precision-recall curve
- Neither of the metrics summarize the entire precision-recall curve
- It is empirically found that the R-precision is highly correlated with MAP

Table of contents

Relevance Assessment

- To assess the quality of the relevance between information needs and documents
- The random combination of query terms as an information need is not good
- The judgment process by human might be expensive
- The human judgments are idiosyncratic and variable
- A good IR system should be able to satisfy the needs of idiosyncratic human

The Relevance Assessment

Kappa Statistic

- How much agreement is there between the relevance judgements
- The kappa statistic is based on the probability that judges agree, P(A), and the probability that they would expected to agree by chance, P(E)

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

The Relevance Assessment

P(E)

- In a näive binary classification, where a document is either relevant or not, P(E) can be considered as 0.5.
- $P(E)$ can be estimated based on the marginal distribution of data

$$P(E) = P(\text{relevant})^2 + P(\text{non-relevant})^2$$

The Relevance Assessment

kappa Example

Example

		Judge 2 relevance		
		yes	no	total
Judge 1 relevance	yes	300	20	320
	no	10	70	80
	total	310	90	400

$$
\begin{aligned}
P(A) &= \frac{300 + 70}{400} = 0.925 \\
P(\text{non-relevant}) &= P(\text{non-relevant}|\text{judge 1})P(\text{judge 1}) \\
&\quad + P(\text{non-relevant}|\text{judge 2})P(\text{judge 2}) \\
&= 0.5 \times \frac{80}{400} + 0.5 \times \frac{90}{400} = 0.2125 \\
P(\text{relevant}) &= P(\text{relevant}|\text{judge 1})P(\text{judge 1}) \\
&\quad + P(\text{relevant}|\text{judge 2})P(\text{judge 2}) \\
&= 0.5 \times \frac{320}{400} + 0.5 \times \frac{310}{400} = 0.7878 \\
P(E) &= P(\text{relevant})^2 + P(\text{non-relevant})^2 \\
&= 0.665 \\
\text{kappa} &= \frac{P(A) - P(E)}{1 - P(E)} = 0.776
\end{aligned}
$$

The Relevance Assessment

The Interpretation of Kappa Statistic

- kappa is 1 if the judges always agree
- kappa is 0 if they agree only at the rate given by chance
- kappa is negative if the judgement agreement is worse than random

As a rule of thumb:

- $0.8 <$ kappa indicates a good agreement
- $0.67 <$ kappa < 0.8 indicates a fair agreement
- kappa < 0.67 indicates that the data is not probably good for an evaluation

Table of contents

Summary

- Which IR technique is more effective
- We need a gold standard to evaluate the effectiveness of an IR system
- There are some standard test collections
- The evaluation of unranked retrieval is based on the precision and recall
- The evaluation of ranked retrieval is based on the precision-recall curve
- Some standard evaluation metrics: MAP, Precision@k, and R-precision
- The goodness of the relevance judgements is measured by the kappa statistics

References I

Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich.
Introduction to Information Retrieval.
Cambridge University Press, 2008.