
Evaluation in Information Retrieval

Mandar Mitra

Indian Statistical Institute

Kolkata

- Background
- Standard measures
 - set-based metrics
 - metrics for ranked retrieval
- Issues in evaluation

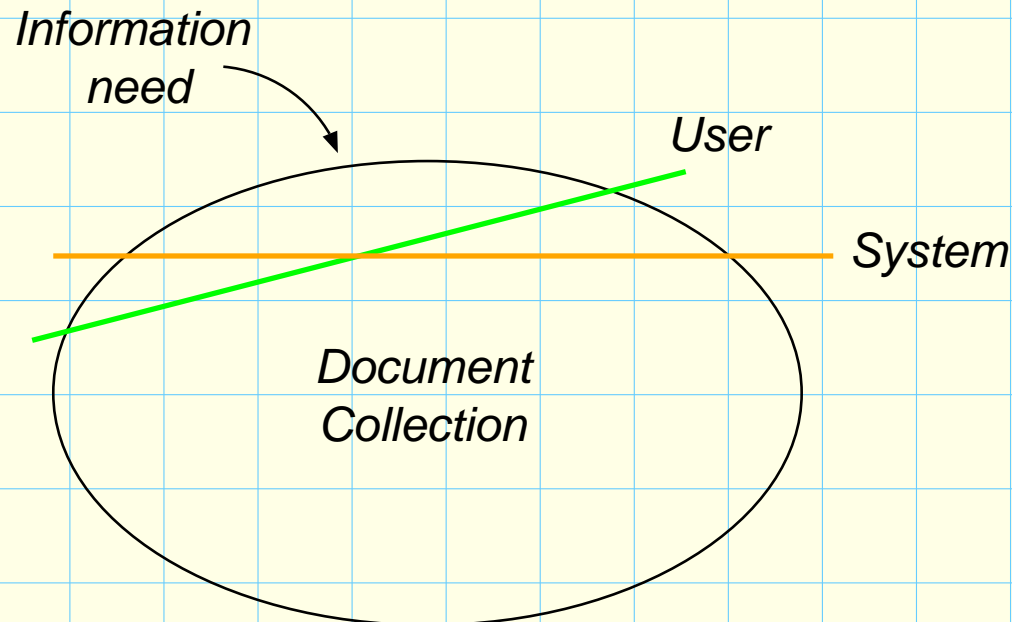
- IR is an *empirical* discipline.
- Proposed techniques need to be validated and compared to existing techniques.
- Intuition can be wrong!

Background

- User has an information need.
- Information need is converted into a **query**.
- Documents are **relevant** or **non-relevant**.
- Ideal system retrieves all and only the relevant documents.

Background

- User has an information need.
- Information need is converted into a **query**.
- Documents are **relevant** or **non-relevant**.
- Ideal system retrieves all and only the relevant documents.



Set-based metrics

$$\begin{aligned}\text{Recall} &= \frac{\#(\text{relevant retrieved})}{\#(\text{relevant})} \\ &= \frac{\#(\text{true positives})}{\#(\text{true positives} + \text{false negatives})}\end{aligned}$$

$$\begin{aligned}\text{Precision} &= \frac{\#(\text{relevant retrieved})}{\#(\text{retrieved})} \\ &= \frac{\#(\text{true positives})}{\#(\text{true positives} + \text{false positives})}\end{aligned}$$

$$\begin{aligned}\mathbf{F} &= \frac{1}{\alpha/P + (1 - \alpha)/R} \\ &= \frac{(\beta^2 + 1)PR}{\beta^2 P + R}\end{aligned}$$

Metrics for ranked results

Non-interpolated average precision

$$AvgP = \frac{1}{N_{Rel}} \sum_{d_i \in Rel} \frac{i}{Rank(d_i)}$$

	<i>R</i>	<i>P</i>
1. Relevant		
2. Non-relevant	0.2	1.00
3. Relevant	0.4	0.67
4. Non-relevant	0.6	0.50
5. Non-relevant	0.8	0.00
6. Relevant	1.0	0.00

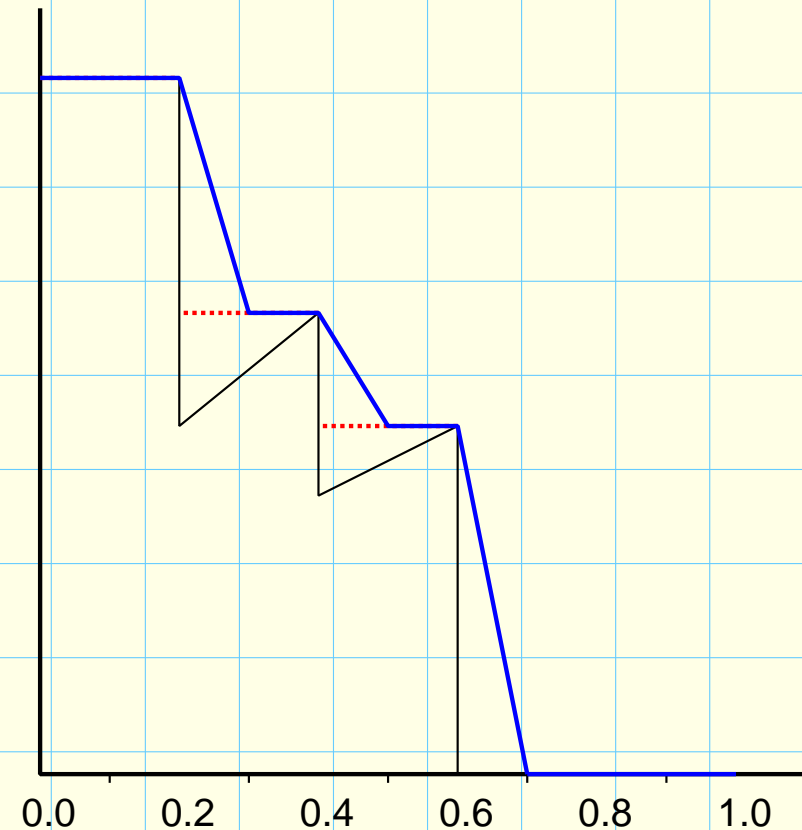
$$AvgP = \frac{1}{5} \left(1 + \frac{2}{3} + \frac{3}{6} \right)$$

Metrics for ranked results

11-point interpolated average precision

$$P_{int}(r) = \max_{r' \geq r} P(r')$$

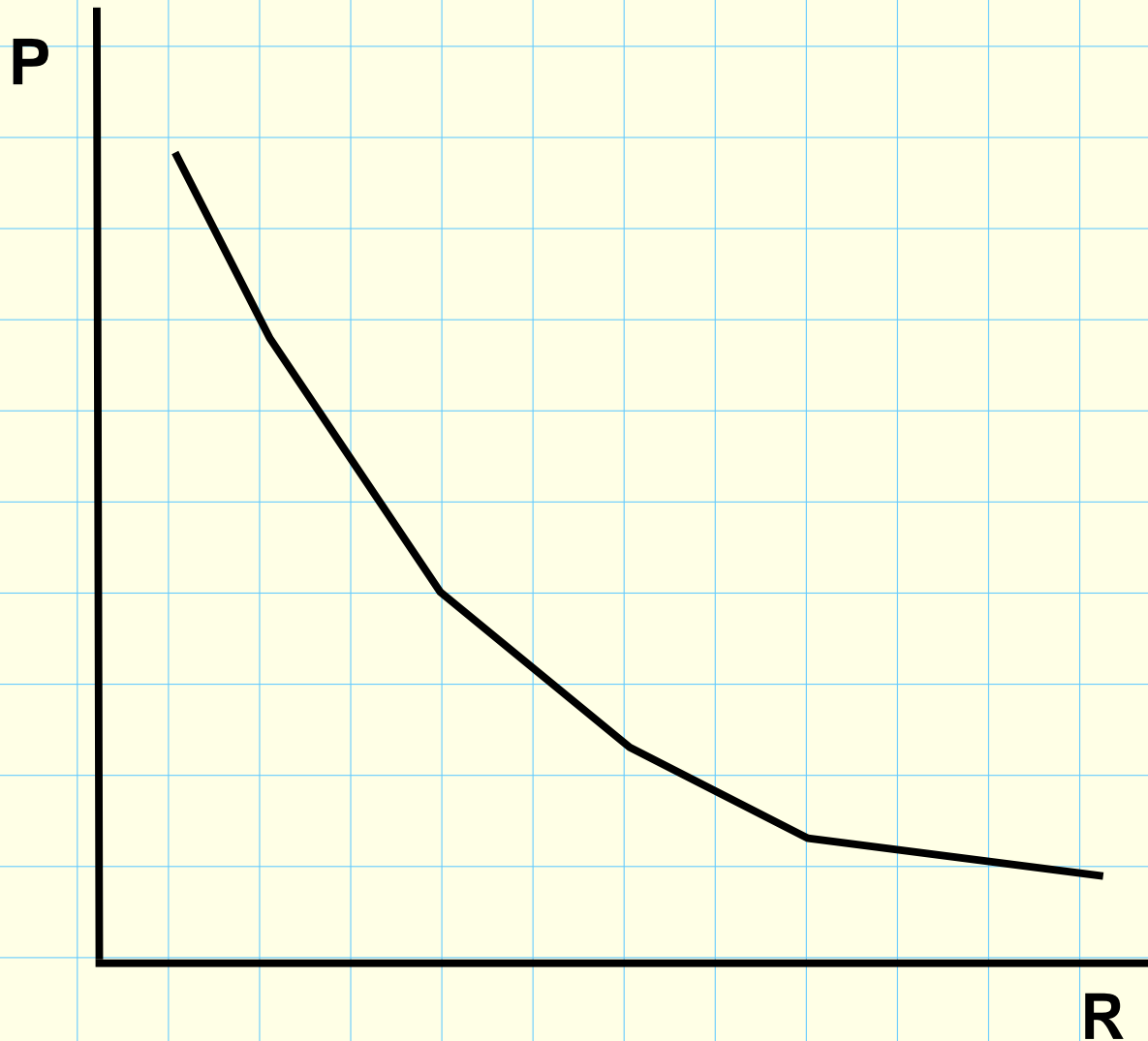
	<i>R</i>	<i>P</i>
	0.0	1.00
1. Relevant	0.1	1.00
2. Non-relevant	0.2	1.00
3. Relevant	0.3	0.67
4. Non-relevant	0.4	0.67
5. Non-relevant	0.5	0.50
6. Relevant	0.6	0.50
	0.7	0.00
	0.8	0.00
	0.9	0.00
	1.0	0.00



Metrics for ranked results

11-point interpolated average precision

- Averages across queries can be meaningfully computed



Metrics for sub-document retrieval

Let p_r - document part retrieved at rank r

$rsize(p_r)$ - amount of relevant text contained by p_r

$size(p_r)$ - total number of characters contained by p_r

T_{rel} - total amount of relevant text for a given topic

$$P[r] = \frac{\sum_{i=1}^r rsize(p_i)}{\sum_{i=1}^r size(p_i)}$$

$$R[r] = \frac{1}{T_{rel}} \sum_{i=1}^r rsize(p_i)$$

Metrics for ranked results

- **Precision at k (P@k)** - precision after **k** documents have been retrieved
 - easy to interpret
 - not very stable / discriminatory
 - does not average well
- **R precision** - precision after N_{Rel} documents have been retrieved

Idea:

- Highly relevant documents are more valuable than marginally relevant documents
- Documents ranked low are less valuable

Idea:

- Highly relevant documents are more valuable than marginally relevant documents
- Documents ranked low are less valuable

$$Gain \in \{0, 1, 2, 3\}$$

$$G = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots \rangle$$

$$CG[i] = \sum_{j=1}^i G[j]$$

$$DCG[i] = \begin{cases} CG[i] & \text{if } i < b \\ DCG[i - 1] + G[i] / \log_b i & \text{if } i \geq b \end{cases}$$

Mean Reciprocal Rank

- Useful for *known-item* searches with a single target
- Let r_i — rank at which the “answer” for query i is retrieved.

Then reciprocal rank = $1/r_i$

$$\text{Mean reciprocal rank (MRR)} = \sum_{i=1}^n \frac{1}{r_i}$$

The Cranfield method (CLEVERDON ET AL., 60S)

- Document collection
- Query / topic collection
- Relevance judgments - information about which document is relevant to which query

The Cranfield method (CLEVERDON ET AL., 60S)

- Document collection
- Query / topic collection
- Relevance judgments - information about which document is relevant to which query

Assumptions

- relevance of a document to a query is objectively discernible
- all relevant documents contribute equally to the performance measures
- relevance of a document is independent of the relevance of other documents

Assessor agreement

- Judges / assessors may not agree about relevance.

Example (MANNING ET AL.)

	Yes ₁	No ₁	Total ₂
Yes ₂	300	20	320
No ₂	10	70	80
Total ₁	310	90	400

$$P(A) = (300 + 70)/400 = 370/400 = 0.925$$

$$P(\text{nrel}) = (80 + 90)/(400 + 400) = 0.2125$$

$$P(\text{rel}) = (320 + 310)/(400 + 400) = 0.7878$$

$$P(E) = P(\text{non-rel})^2 + P(\text{rel})^2 = 0.665$$

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.925 - 0.665}{1 - 0.665} = 0.776$$

- Rules of thumb:

$\kappa > 0.8$ — good agreement

$0.67 \leq \kappa \leq 0.8$ — fair agreement

$\kappa < 0.67$ — poor agreement

- Exhaustive relevance judgments may be infeasible.
- Pool top results obtained by various systems and assess the pool.
- A wide variety of models, retrieval algorithms is important.
- **Manual interactive retrieval** is a must.
- Unjudged documents are assumed to be non-relevant.

- Exhaustive relevance judgments may be infeasible.
- Pool top results obtained by various systems and assess the pool.
- A wide variety of models, retrieval algorithms is important.
- **Manual interactive retrieval** is a must.
- Unjudged documents are assumed to be non-relevant.

Can unbiased, incomplete relevance judgments be used to reliably compare the relative effectiveness of different retrieval strategies?

- Based on number of times judged nonrelevant documents are retrieved before relevant documents

Let R - set of relevant documents for a topic

N - set of first $|R|$ judged non-rel docs retrieved

$$bpref = \frac{1}{|R|} \sum_{r \in R} \left(1 - \frac{|n \text{ ranked higher than } r|}{|R|} \right)$$

$$bpref10 = \frac{1}{|R|} \sum_{r \in R} \left(1 - \frac{|n \text{ ranked higher than } r|}{10 + |R|} \right)$$

- With complete judgments:
system rankings generated based on MAP and bpref10 are highly correlated
- When judgments are incomplete:
system rankings generated based on bpref10 are more stable

- Use a standard dataset if possible
- Metrics: MAP, P@20
- Use a good baseline
- Use statistical tests of significance

References

- *Introduction to Modern Information Retrieval*. Salton, McGill. McGraw Hill, 1983.
- *An Introduction to Information Retrieval*. Manning, Raghavan, Schutze.
<http://www-csli.stanford.edu/~schuetze/information-retrieval-book.html>
- *Retrieval Evaluation with Incomplete Information*. Buckley, Voorhees. SIGIR 2004.
- <http://trec.nist.gov>
- *Cross-Language Evaluation Forum: Objectives, Results, Achievements*. Braschler, Peters. Information Retrieval, 7:12, 2004.
- <http://research.nii.ac.jp/ntcir>