Evaluating IR System

Information retrieval (IR)

By: Dr. LOUNNAS Bilal

Measures for a search engine

- How fast does it index
  - Number of documents/hour
  - (Average document size)
- How fast does it search
  - Latency as a function of index size
- Expressiveness of query language
  - Ability to express complex information needs
  - Speed on complex queries
- Uncluttered UI
- Is it free?

Measures for a search engine

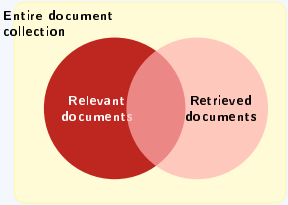
- All of the preceding criteria are *measurable*: we can quantify speed/size
  - we can make expressiveness precise
- The key measure: **user happiness**
  - What is this?
  - Speed of response/size of index are factors
  - But blindingly fast, useless answers won't make a user happy
- Need a way of quantifying user happiness with the results returned
  - Relevance of results to user's information need

Evaluating ranked results

- **Evaluation of a result set:**
  - If we have
    - a benchmark document collection
    - a benchmark set of queries
    - assessor judgments of whether documents are relevant to queries

  $\Rightarrow$ Then we can use **Precision/Recall**
- **Evaluation of ranked results:**
  - The system can return any number of results
  - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*

Precision and Recall



		retrieved	not retrieved
irrelevant		retrieved & irrelevant	Not retrieved & irrelevant
relevant		retrieved & relevant	not retrieved but relevant

$$recall = \frac{\text{Relevant docs retrieved}}{\text{Total relevant docs}}$$

$$precision = \frac{\text{Relevant docs retrieved}}{\text{Total docs retrieved}}$$

Recall/Precision

	R	P
1	R	
2	N	
3	N	
4	R	
5	R	
6	N	
7	R	
8	N	
9	N	
10	N	

Assume 10 rel docs in collection

Precision-Recall Curve

What is it?

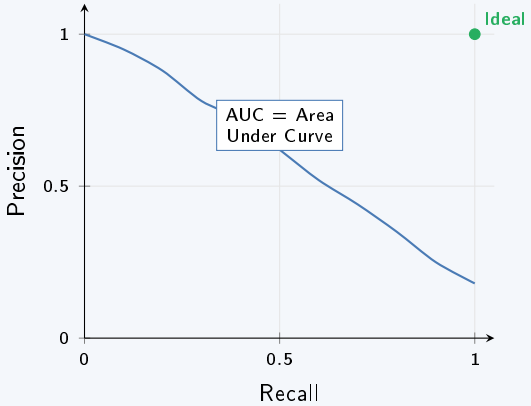
Plots *Precision* (y-axis) vs. *Recall* (x-axis) at varying thresholds. Recall rises as precision typically decreases.

Ideal system:

Achieves Precision = Recall = 1 (top-right corner of the plot).

AUC (Area Under Curve):

Higher AUC ⇒ better system.



F-Measure / F1 Score

Definition

Harmonic mean of Precision and Recall — combines both into one metric:

$$F_1 = \frac{2 \times P \times R}{P + R} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

General F_β measure:

$$F_\beta = \frac{(1 + \beta^2)\, P\, R}{\beta^2 P + R}$$

- $\beta < 1 \Rightarrow$ favours Precision
- $\beta > 1 \Rightarrow$ favours Recall
- $\beta = 1 \Rightarrow$ balanced (standard F_1)

Recall / Precision — Worked Example

Assume **10 relevant documents** in the collection.

Rank	Doc	Recall	Precision
1	R	1/10=0.10	1/1=1.00
2	N	—	—
3	N	—	—
4	R	2/10=0.20	2/4=0.50
5	R	3/10=0.30	3/5=0.60
6	N	—	—
7	R	4/10=0.40	4/7=0.57
8	N	—	—
9	N	—	—
10	N	—	—

Key observations

- Precision fluctuates at each rank
- Recall only increases when a relevant doc is found
- Non-relevant docs reduce precision, not recall
- Only 4/10 relevant docs retrieved $\Rightarrow$ Recall $= 0.40$

MAP — Mean Average Precision

Step 1 — Average Precision per query

$$\mathrm{AP}(q) = \frac{1}{R} \sum_{k=1}^{N} P(k) \cdot \mathrm{rel}(k)$$

Compute precision at each rank k where a relevant document appears, then average over R relevant docs.

Step 2 — Mean over all queries

$$\mathrm{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{AP}(q)$$

Average AP values across all Q queries in the benchmark.

Worked example (10 relevant docs in collection):

- **Query 1** — relevant at ranks 1, 3, 5: $\mathrm{AP} = \frac{1}{3}(1.0 + 0.67 + 0.6) \approx 0.756$
- **Query 2** — relevant at ranks 2, 4: $\mathrm{AP} = \frac{1}{2}(0.5 + 0.5) = 0.500$

$\Rightarrow$ $\mathrm{MAP} = (0.756 + 0.500)/2 \approx 0.628$

Benchmark Datasets for IR Evaluation

Standard benchmarks require three components:

1. **Document collection** — a fixed corpus
2. **Query set** — a set of information-need topics
3. **Relevance judgments** — human-assessed doc–query relevance

TREC (1992–present)

Gold standard by NIST. Annual tracks: newswire, web, genomics, legal.

Cranfield (1966)

1,400 aeronautics abstracts, 225 queries. Introduced Precision-Recall.

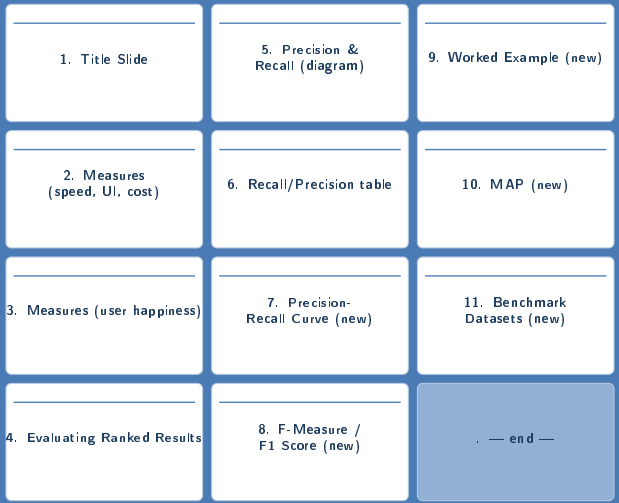
MS MARCO (2016–present)

1M+ Bing queries with human-annotated answers for passage retrieval.

BEIR (2021–present)

18 datasets for zero-shot evaluation of dense retrieval models.

Slide Preview — All Slides at a Glance



Original slides: 1–6 | New extended slides: 7–11