CADT

Cambodia Academy of Digital Technology

Information Retrieval

Evaluation

Lecturer

Mr. Natt KORAT

Lecturer-Researcher,

Institute of Digital Research & Innovation (IDRI),

Cambodia CADT Academy of Digital Technology (CADT)

Recap: an unresolved disagreement

- **Lab 3:** TF-IDF ranked d04 first for “python”; BM25 ranked d18 first
- Which one is better?
- We don’t know yet, we only compared 2 lists by eyes

Why we need evaluation

- Every design choice (stemming rules, k1, b, …) needs a way to compare “**before**” and “**after**”
- Without measurement, tuning is just a guessing.

Test collections

- Three parts:
    1. Document (our 30 docs)
    2. Queries
    3. Relevant judgments (qrels)
- Query relevance judgments (qrels) tell us which documents satisfy each information neeed.
- Judments are made by a HUMAN, not computed

Example: relevance judgments (qrels)

- **Query: “python”**
- D04 -> relevance (e.g., a useful Python introduction)
- D18 -> relevant (e.g., a brief Python code snippet)
- D05 -> not relevant for this query
- A judgment tells us what counts as a correct result.

Binary vs. graded judgments

- Binary: relevant or not (yes/no)
- Graded: HOW relevant (e.g., 0-3)
- Graded judgments can reward ranking the BEST document first, not just relevant one

Where judgments come from

- Small projects: the team itself, or a few volunteers
- Large evaluation compaigs (e.g., TREC): trained annotators, following detailed guidelines
- Judgments are the form of DATA , and data has noise

TREC: https://trec.nist.gov/

Precision, recall, F1 (recap)

- Precision: of what you returned, how much is relevant?
- Recall: of what is relevant, how much did you return?
- F1: harmonic mean, punishes a system that is very good at only one of the two

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

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

$$F1 = 2 \times \frac{Precison \times Recall}{Precision \ + \ Recall}$$

Precision@k and Recall@k

- Only evaluate the first k results returned by the system.
- $P@k = \frac{relevant\ doc\ in\ top\ k}{k}$
- $R@k = \frac{relevant\ docs\ in\ top\ k}{all\ relevant\ docs}$
- K represents how far a user is likely to inspect.

Worked example: top 3 results

Rank	Document	Relevant?
1	d04	yes
2	d18	yes
3	D15	no

- **Ground truth**: relevant documents = {d04, d18} -> 2 in total.

Calculate P@3, R@3, and F1@3

- Relevant in top 3 = 2; return in top 3 = 3
- $P@3 = \frac{2}{3} = 0.67$
- $R@3 = \frac{2}{2} = 1.00$
- $F1@3 = 2 \times \frac{0.67 \times 1.00}{0.67+1} = 0.80$
- High recall, but one result in the top 3 is irrelevant.

Changing k changes the score

Cutoff	Relevant found	P@k	R@K
K = 1	1	1.00	0.50
K = 2	2	1.00	1.00
K = 3	2	0.67	1.00

Which *k* matches the user's real search behavior?

Why not rely on one cutoff?

- P@1 can reward a system that finds only one useful result.
- P@10 evaluates more results, but many users never look that far.
- A single cutoff may hide where relevant documents appear.
- Average Precision summarizes relevant hits across the randking.

Average precision (AP)

- $AP = \frac{1}{R}\sum_{k=1}^{N} P(k)rel(k)$ or

- $AP = \frac{1}{R}\sum_{k=1}^{N} \frac{rel_n@k \cdot rel(k)}{k}$

- **Relevant documents at earlier ranks increase AP.**

AP: step-by-step example

Rank	Document	Relevant?	P@rank if relevant
1	d04	yes	1/1 = 1.0
2	d18	no	0
3	105	yes	2/3 = 0.67

- $AP = \frac{1.00+0.67}{2} = 0.83$

Mean Average precision (MAP)

- $MAP = \frac{1}{|Q|}\sum_{q \in Q} AP_q$

- **A single score across mutiple information needs and its queries.**

Limits of AP and MAP

Ranking A	Ranking B
1. d04 (very relevant)	1. d18 (somewhat relevant)
2. d18 (somewhat relevant)	2. d04 (very relevant)
AP = 1.00	AP = 1.00

Both are binary-relevant. AP cannot reward the better document at rank 1.

Graded relevance: example 0-3

Grade	Meaning	Illustrative document
0	Not relevant	d05
1	Somewhat relevant	d18
2	Useful	-
3	Highly relevant	d04

The labels and scale must be explained consistently to annotators.

NDCG: key ideas

- **Graded gain**: a highly relevant document earns more credit.
- **Rank discount**: the same result earns less credit if it appears lower.
- Normalize against the ideal ranking for that query.
- NDCG combines all three ideas into a score from 0 to 1.

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

***NDCG**: Normalized Discounted Cumulative Gain*
https://en.wikipedia.org/wiki/Discounted_cumulative_gain

DCG: formula and intuition

- Gain from grade r: $gain(r) = 2^r - 1$
- Discount at rank i: $\log_2(i+1)$
- $DCG@k = \sum \frac{2^r - 1}{\log_2(i+1)}$, $i = 1 \ \ldots k$
- High grades and early ranks contribute the most.

DCG step 1: relevance -> gain

Relevance grade r	Gain = $2^r - 1$
0	0
1	1
2	3
3	7

Moving from grade 2 to grade 3: gain increases from 3 to 7.

DCG step 2: apply rank discount

Rank	Doc	Grade	Gain	$\log_2(i+1)$	Contribution
1	d18	1	1	1.00	1.00
2	d04	3	7	1.58	4.43
3	d05	0	0	2.00	0.00

$$\boldsymbol{DCG@3 = 1.00 + 4.43 + 0 = 5.43}$$

IDCG and NDCG: normalize the score

- Ideal order: d04 (grade 3), d18 (grade 1), d05 (grade 0)
- $IDCG@3 = \frac{7}{1} + \frac{1}{1.58} + \frac{0}{2} = 7.63$
- $NDCG@3 = \frac{5.43}{7.63} = 0.71$

NDCG = 1.0 when the ranking achieves ideal DCG.

Why NDCG can see what AP misses

Metric	A: d04 -> d18 -> d05	B: d18 -> d04 -> d05
Relevant at ranks	1, 2	1, 2
AP(binary)	1.00	1.00
DCG@3	7.63	5.43
NDCG@3	1.00	0.71

A puts the best document first. Only the graded metric reflects that.

Which metric should we choose?

Retrieval goal	Useful metric
Show a few good results immediately	P@k, NDCG@k
Find as many relevant items as possible	Recall / R@k (appropriate k)
Compare binary-ranked systems over queries	MAP
Reward highly relevant results near the top	NDCG

Annotator disagreement

- Two people can disagree about a borderline document.
- Those choices change the qrels, and therefore **MAP** or **NDCG**.
- Use clear guidelines, example cases, and repeated judgments.
- Discuss disagreements before claiming a tiny score difference is meaningful.

Recap

- Test collection = documents + queries + relevance judgments.
- P@k, R@k and F1@k measure success at a cutoff.
- AP rewards early relevant hits; MAP averages across queries.
- NDCG uses graded relevance and discounts lower ranks.
- Choose metrics for the task; check how trustworthy the qrels are.