CS47300: Web Information Search and Management

Prof. Chris Clifton
30 August 2020

Material adapted from course created by Dr. Luo Si, now leading Alibaba research group

AD-hoc IR: Basic Process

? Information Need

Representation

Query

Retrieval Model

Indexed Objects

Retrieved Objects

Evaluation
Evaluation: What do we Evaluate?

• Effectiveness
  – How do we define effective?
  – Where can we find the correct answers?

• Efficiency
  – Retrieval speed?
  – Storage space?
  *Particularly important for large-scale real-world system*

• Usability
  – What do real users really want?
  – Is user interface important to IR evaluation?

Evaluation Criteria

• Effectiveness
  – Favor returned document ranked lists with more relevant documents at the top
  – Objective measures
    • Recall and Precision
    • Mean-average precision
    • Rank based precision

For documents in a subset of a ranked lists, if we know the truth

<table>
<thead>
<tr>
<th></th>
<th>Retrieved</th>
<th>Not retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>Relevant docs retrieved</td>
<td>Relevant docs not retrieved</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>Irrelevant docs retrieved</td>
<td>Irrelevant docs not retrieved</td>
</tr>
</tbody>
</table>

Precison = \( \frac{\text{Relevant docs retrieved}}{\text{Retrieved docs}} \)

Recall = \( \frac{\text{Relevant docs retrieved}}{\text{Relevant docs}} \)
Evaluation: “Ground Truth”

<table>
<thead>
<tr>
<th>Retrieved</th>
<th>Not retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>Relevant</td>
</tr>
<tr>
<td>Relevant does not retrieved</td>
<td>Relevant does not retrieved</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>Irrelevant</td>
</tr>
<tr>
<td>Irrelevant does not retrieved</td>
<td>Irrelevant does not retrieved</td>
</tr>
</tbody>
</table>

Question: How to find all relevant documents?

Difficult for Web, but possible on controllable corpus

- How to find all relevant documents? (difficult to check one by one)
- Judgers may have inconsistent decisions (subjective judgment)

The Pooling process

Evaluation: Inconsistent Judgement

- People may not agree on the “right” answer
  - Some think document is relevant to query, others don’t
- Discussion among multiple judgers to reduce bias
- Combine judgments from multiple judgers
  - Majority vote
- *If it is hard to decide for human judges, it is likely to be hard for an automatic system*
Evaluation: Pooling Strategy

- Retrieve documents using multiple methods
- Judge top \( n \) documents from each method
- Whole retrieved set is the union of top retrieved documents from all methods
- Problems: the judged relevant documents may not be complete

- *It is possible to estimate the total number of relevant documents by random sampling*
Unranked Measures:

- Precision: $\frac{\# Relevant Retrieved}{\# Retrieved}$
- Recall: $\frac{\# Relevant Retrieved}{\# Relevant}$
- F1 score: $\frac{2PR}{P+R}$

Evaluation

- Evaluate a ranked list
  - Precision at Recall
- Evaluate at every relevant document

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.667</td>
<td>0.2</td>
</tr>
<tr>
<td>0.75</td>
<td>0.3</td>
</tr>
<tr>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>0.667</td>
<td>0.4</td>
</tr>
<tr>
<td>0.714</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Not Retrieved: ++++++
Ranked Metrics

Single number

- Mean average precision
  - Calculate precision at each relevant document; average over all precision values
  - Mean average precision – average over many queries
- 11-point interpolated average precision
  - Calculate precision at standard recall points (e.g., 10%, 20%...); smooth the values; estimate 0% by interpolation
  - Average the results

Evaluation:

Single Value Metrics

- Rank based precision
  - Calculate precision at top ranked documents (e.g., 5, 10, 15…)
  - Desirable when users care more for top ranked documents
- Mean Reciprocal Rank
  - Reciprocal Rank: 1/rank (position in list) of first relevant document
  - MRR: Average Reciprocal Rank over many queries
TREC collections with queries and relevance judgment

- **TREC CDs 1-5**: 1.5 millions docs, 5GB, news and government reports (e.g., AP, WSJ, Dept of Energy abstracts)
- **TREC WT10g**: crawled from Web (open domain), 1.7 million docs, 10GB
- **TREC Terabyte**: crawled from U.S. government Web pages, 25 million docs, 426 GB
- *All have more than 100 queries with relevance judgment*
Evaluation: TREC

• TREC query example

  <title> airport security
  <desc> Description:
  What security measures are in effect or are proposed
to go into effect in airports?
  <narr> Narrative:
  A relevant document could identify a specific airport
  and describe the security measures already in effect
  or proposed for use at that airport. Relevant items
could also describe a failure of security that was
cited as a contributing cause of a tragedy which came
to pass or which was later averted. Comparisons between
  and among airports based on the effectiveness of the
  security of each are also relevant.

Evaluation: TREC

• TREC relevance judgment example

  451 WTX058-B50-85 0
  451 WTX059-B06-411 0
  451 WTX059-B07-154 0
  451 WTX059-B09-203 0
  451 WTX059-B11-245 0
  451 WTX059-B30-262 1
  451 WTX059-B37-11 0
  451 WTX059-B37-149 1
  451 WTX059-B37-217 0
  451 WTX059-B37-268 0
  451 WTX059-B37-27 0
Review to date:

- Basic Concepts of Information Retrieval:
- Task Definition of Ad-hoc IR
  - Terminologies and Concepts
  - Overview of Retrieval Models
- Text representation
  - Indexing
  - Text preprocessing
- Evaluation
  - Evaluation methodology
  - Evaluation metrics