IR Evaluation

• **Efficiency**
  – storage
  – response time

• **Effectiveness**
  – recall
  – precision

• **Usability**
  – assistance in formulating queries
  – document presentation
  – functionalities
Test Collection

- a collection of test documents
- a set of test queries
- a set of relevant documents for each query
Relevance

• it is hard to precisely define relevance
• a relevant document is judged in context of a query
• judgement of a relevant document depends on who judges it and even when it is judged
• relevance of a document is fuzzy
• with real collections, one never knows the true set of relevant documents
Recall And Precision

• a collection of test documents $D$
• a query $q$
• a set of relevant documents $Relevant$ for $q$ from $D$
• a set of retrieved documents $Retrieved$ using a retrieval method
• a set of retrieved relevant documents $RR$

$$RR = Relevant \cap Retrieved$$
Recall And Precision

- Recall: proportion of the relevant documents that are retrieved
  \[
  \text{Recall} = \frac{|RR|}{|Relevant|}
  \]

- Precision: proportion of the retrieved documents that are relevant
  \[
  \text{Precision} = \frac{|RR|}{|Retrieved|}
  \]
Ranked Retrieval

- Recall-Precision Graph
- Recall-Precision Table
- Compute recall-precision at each relevant document
- Compute precision at fixed recall points (20%)
- Compute precision at fixed recall cutoffs (rank 20)
<table>
<thead>
<tr>
<th>Recall</th>
<th>Method 1</th>
<th>Method 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>88.3</td>
<td>90.2</td>
</tr>
<tr>
<td>10</td>
<td>84.7</td>
<td>88.9</td>
</tr>
<tr>
<td>20</td>
<td>78.6</td>
<td>82.5</td>
</tr>
<tr>
<td>30</td>
<td>70.3</td>
<td>73.2</td>
</tr>
<tr>
<td>40</td>
<td>65.0</td>
<td>66.6</td>
</tr>
<tr>
<td>50</td>
<td>58.5</td>
<td>60.3</td>
</tr>
<tr>
<td>60</td>
<td>55.5</td>
<td>57.2</td>
</tr>
<tr>
<td>70</td>
<td>46.6</td>
<td>49.2</td>
</tr>
<tr>
<td>80</td>
<td>35.7</td>
<td>38.2</td>
</tr>
<tr>
<td>90</td>
<td>23.2</td>
<td>25.7</td>
</tr>
<tr>
<td>100</td>
<td>13.4</td>
<td>14.2</td>
</tr>
</tbody>
</table>
Average And Interpolation

• Evaluation is often conducted over many different queries, it is hard to compare individual tables and graphs.

• Average over all queries to get an average precision at each recall level

• Interpolate precision at standard recall levels
Other Measures

• Fallout: proportion of non-relevant documents that are retrieved

\[ Fallout = \frac{|\text{Retrieved}| - |RR|}{|D| - |Relevant|} \]

• Miss

\[ \text{Miss} = 1 - \text{Recall} \]
## Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Not Retrieved</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

Total Document = A + B + C + D

Relevant = A + C  \quad \text{Not Relevant} = B + D

Retrieved = A + B  \quad \text{Not Retrieved} = C + D

Recall = A/(A+C)  \quad \text{Precision} = A/(A+B)

Fallout = B/(B+D)  \quad \text{Miss} = C/(A+C)
E Measure

- used to emphasize precision or recall
- a weighted average precision and recall
- a larger $\alpha$ increases the importance of precision
- a smaller E value means better performance
- $\alpha = 0.5$, precision and recall are equally important

$$E = 1 - \frac{1}{\alpha \frac{1}{p} + (1 - \alpha) \frac{1}{r}}$$
F Score (Harmonic Mean)

- F = 1 - E, when $\alpha = 0.5$
- a larger F means better performance

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$
Queries

• Ad Hoc vs. Long Term (User Profile)

• Different views of queries
  – a small relevant document
  – a rule for defining relevant documents
  – an expression of a user information need
Query vs Information Need

- Information need is internal to a searcher
- Query is operational form of information need
- Information need is often not precisely defined
- Information need changes with search process
- Many different kinds of information needs
  - Topic (e.g., sports)
  - Item (e.g., about a person or a city)
  - Event (e.g., 911)
Query Formulation

- Query Languages
  - key words
    - boolean
    - free text
    - context
  - pattern
  - structural

- User Interfaces
Query Processing

- Stopwords
- Stemming
- Query expansion
- User interaction
Boolean Query

• a query is a boolean expression and is formed using boolean operators \( \text{\textit{and}} \), \( \text{\textit{or}} \), \( \text{\textit{not}} \).

• the following IR models use boolean queries
  – boolean model
  – fuzzy set model
  – extended boolean model
Free Text

• a query is viewed as a document

• a query is represented as a vector of term weights

• used by vector and probabilistic models

• facilitates relevance feedback
Context

- phases
- word distance
Pattern Matching

- retrieval of pieces of text that satisfy the pattern
- partial matching
- specifying patterns
  - words and substrings
  - prefixes and suffixes
  - ranges
  - regular expressions
Structural Queries

• a document consists of a list or a hierarchy of subcomponents

• queries specify not only content also where (which subcomponent) to look for the content