Evaluation of Retrieval Systems

Performance Criteria

1. Expressiveness of query language
   - Can query language capture information needs?
2. Quality of search results
   - Relevance to users’ information needs
3. Usability
   - Search Interface
   - Results page format
   - Other?
4. Efficiency
   - Speed affects usability
   - Overall efficiency affects cost of operation
5. Other?

Quantitative evaluation

- Concentrate on quality of search results
- Goals for measure
  - Capture relevance to user information need
  - Allow comparison between results of different systems
- Measures define for sets of documents returned
- More generally “document” could be any information object

Core measures: Precision and Recall

- Need binary evaluation by human judge of each retrieved document as relevant/irrelevant
- Need know complete set of relevant documents within collection being searched
- Recall = \( \frac{\text{# relevant documents retrieved}}{\text{# relevant documents}} \)
- Precision = \( \frac{\text{# relevant documents retrieved}}{\text{# retrieved documents}} \)

Combine recall and precision

- F-score (aka F-measure) defined to be: harmonic mean of precision and recall
- \[ F = \frac{2 \times \text{recall} \times \text{precision}}{\text{precision} + \text{recall}} \]

Use in “modern times”

- Defined in 1950s
- For small collections, these make sense
- For large collections,
  - Rarely know complete set relevant documents
  - Rarely could return complete set relevant documents
- For large collections
  - Rank returned documents
  - Use ranking!

\[ (n-h)/n = (h-m)/m. \text{ Also } (1/m) - (1/h) = (1/h) - (1/n) \]
Ranked result list

- At any point along ranked list
  - Can look at precision so far
  - Can look at recall so far
    - if know total # relevant docs
- Can focus on points at which relevant docs appear
  - If $m^{th}$ doc in ranking is $k^{th}$ relevant doc so far, precision is $k/m$
  - No a priori ranking on relevant docs

query: "toxic waste"

1. Toxic waste - Wikipedia, the free encyclopedia
   en.wikipedia.org/wiki/Toxic_waste
2. Toxic Waste Household toxic and hazardous waste ...
   www.urbanedpartnership.org/target/units/recycle/toxic.html
3. Toxic Waste Facts, Toxic Waste Information
   environment.nationalgeographic.com/toxic-waste-overview.html
4. Toxic Waste Candy Online Toxic Waste Sour Candy ...
   www.candydynamics.com/
5. Toxic Waste Candy Online Toxic Waste … chew bars...
   www.toxicwastecandy.com/
   www.epa.gov/ebtpages/wasthazardouswaste.html
7. toxic waste — Infoplease.com toxic waste is waste ... 
   www.infoplease.com/cel/sci/A0849189.html
8. Toxic Waste Clothing Toxic Waste Clothing is a trend...
   www.toxicwasteclothing.com/
9. Toxic waste - Wikipedia, the free encyclopedia
   en.wikipedia.org/wiki/Toxic_waste
10. Toxic Waste Household toxic and hazardous waste ...
    www.urbanedpartnership.org/target/units/recycle/toxic.html
11. Toxic Waste Facts, Toxic Waste Information
    environment.nationalgeographic.com/toxic-waste-overview.html
12. Toxic Waste Candy Online Toxic Waste Sour Candy ...
    www.candydynamics.com/
13. Toxic Waste Candy Online Toxic Waste … chew bars...
    www.toxicwastecandy.com/
    www.epa.gov/ebtpages/wasthazardouswaste.html
15. toxic waste — Infoplease.com toxic waste is waste ... 
    www.infoplease.com/cel/sci/A0849189.html
16. Toxic Waste Clothing Toxic Waste Clothing is a trend...
    www.toxicwasteclothing.com/

Plot: precision versus recall

- Choose standard recall levels: $r_1$, $r_2$ ...
  - $r_j$ increasing
  - Define "precision at recall level $r_j"$
    \[ p(r_j) = \max \text{ over all } r \text{ with } r_{j-1} < r < r_j \text{ of precision when recall $r$ achieved} \]
- Smooth: "interpolated precision"
  \[ p_{\text{inter}}(r_j) = \max \text{ over all } r_j \text{ with } j \text{ ai of } p(r_j) \]
Single number characterizations

- Can look at precision at one fixed critical position of ranking: "Precision at k"
  - If know are T relevant documents can choose k=T
  - May not want to look that far even if know T
  - Can choose set of R relevant docs, and calc.
    precision at k=R only with respect to these docs
    - "R-precision" of Intro IR
    - can only do with some prior analysis of collection
  - For Web search
    - Choose k to be number pages people look at
    - k=? What expecting?

Reciprocal rank:
Capture how early get relevant result in ranking

more single number characterizations

1) Record precision at each point a relevant document encountered through ranked list
   - Don’t need know all relevant docs
   - Can cut off ranked list at predetermined rank
2) Average the recorded precisions in (1)
   = average precision for a query result

Mean Average Precision (MAP):
For a set of test queries, take the mean (i.e. average)
Of the average precision for each query
- Compare retrieval systems with MAP

even more single number characterizations

Reciprocal rank:
Capture how early get relevant result in ranking

reiprocal rank of ranked results of a query

get mean reciprocal rank of set of test queries

Summary

- Collection of measures of how well ranked search results provide relevant documents
- based on precision
- based to some degree on recall
- single numbers:
  - precision at fixed rank
  - average precision over all positions of relevant docs
  - reciprocal rank of first relevant doc

query: "toxic waste"

1. Toxic waste - Wikipedia, the free encyclopedia
   en.wikipedia.org/wiki/Toxic_waste
2. Toxic Waste - Household toxic and hazardous waste ...
   www.urbanedpartnership.org/target/units/recycle/toxic.html
3. Toxic Waste Facts, Toxic Waste Information
   environment.nationalgeographic.com/toxic-waste-overview.html
4. Toxic Waste Candy Online Toxic Waste Sour Candy ...
   www.candydynamics.com/
5. Toxic Waste Candy Online Toxic Waste … chew bars...
   www.toxicwastecandy.com/
   www.epa.gov/osp/pages/wastehazardouswaste.html
7. toxic waste — Infoplease.com toxic waste is waste ...
   www.infoplease.com/crd/sci/A0849180.html
8. Toxic Waste Clothing Toxic Waste Clothing is a trend...
   www.toxicwasteclothing.com/

query: "toxic waste"

10. Ecopolitism: toxic waste and the movement for environmental justice - Google Books Result
    books.google.com/books?id=0816821756...

THEN precision at rank 10 is 0.6 and average precision at rank 10 is 0.84

= 1/1+2/2+3/3+4/6+5/7+6/9
Example

<table>
<thead>
<tr>
<th>rank</th>
<th>rel.</th>
<th>rel.</th>
<th>rel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

precision at rank 5 = 3/5 for all
reciprocal rank = 1
reciprocal rank = 1/2
average precision = 1/5(1+2/4+3/5+4/9+5/10) = .61
average precision = 1/5(1/2+2/4+3/5+4/9+5/10) = .509
average precision = 1/4(1/2+2/4+3/5+4/9) = .511

Beyond binary relevance

• Sense of degree to which document satisfies query
  – classes, e.g.: excellent, good, fair, poor, irrelevant
• Can look at measures class by class
  – limit analysis to just excellent docs?
  – combine after evaluate results for each class
• Need new measure to capture all together
  – does document ranking match "excellent, good, fair, poor, irrelevant" rating?

Discounted cumulative gain (DCG)

• Assign a gain value to each relevance class
  – e.g. 0 (irrel.), 1, 2, 3, 4 (best) assessor’s score
  – how much difference between values?
  – text uses (2^assessor’s score -1)
• Let d_1, d_2, … d_k be returned docs in rank order
• G(i) = gain value of d_i
  – determined by relevance class of d_i
• DCG(i) = ∑ ( G(j) / (log_2 (1+j) ) )

Using Discounted Cumulative Gain

can compare retrieval systems on query by
• plotting values of DCG(i) versus i for each
  – plot gives sense of progress along rank list
• choosing fixed k and comparing DCG(k)
  – if one system returns < k docs, fill in at bottom with “irrel”
• can average over multiple queries
  – text “Normalized Discounted Cumulative Gain”
  • normalized so best score for a query is 1

Example

<table>
<thead>
<tr>
<th>rank</th>
<th>gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

DCG(1) = 4/log_2 2 = 4
DCG(2) = 4 + 0 = 4
DCG(3) = 4 + 0 = 4
DCG(4) = 4 + 1/log_2 5 = 4.43
DCG(5) = 4.43 + 4/log_2 6 = 5.98
DCG(6) = 5.98 + 0 = 5.98
DCG(7) = 5.98 + 0 = 5.98
DCG(8) = 5.98 + 0 = 5.98
DCG(9) = 5.98 + 1/log_2 10 = 6.28
DCG(10) = 6.28 + 1/log_2 11 = 6.57

Comparing orderings

Two retrieval systems both return k excellent documents. How different are rankings?

• Measure for two orderings of n-item list: Kendall’s Tau

inversion: pair of items ordered differently in the two orderings

Kendall’s Tau (order1, order2) = 1 – ( ( # inversions) / (½(n)(n-1) ) )
Example

doc rank1 rank2
A  1  3
B  2  4
C  3  1
D  4  2

# inversions:  A-C, A-D, B-C, B-D  = 4
Kendall tau = 1 - 4/3 = -1/3

Using Measures

• Statistical significance versus meaningfulness
• Use more than one measure
• Need some set of relevant docs even if don’t have complete set
  How?
  – Look at TREC studies

Relevance by TREC method

Text Retrieval Conference 1992 to present

• Fixed collection per "track"
  • E.g. "*.gov", CACM articles
• Each competing search engine for a track
  asked to retrieve documents on several "topics"
  – Search engine turns topic into query
  – Topic description has clear statement of what
    is to be considered relevant by human judge

Sample TREC topic from 2007 Blog Track

• Title: Mutual Funds
• Description: Blogs about mutual funds performance and trends.
• Narrative: Ratings from other known sources (Morningstar) or relative to key performance indicators (KPI) such as inflation, currency markets and domestic and international vertical market outlooks. News about mutual funds, mutual fund managers and investment companies. Specific recommendations should have supporting evidence or facts linked from known news or corporate sources. (Not investment spam or pure, uninformed conjecture.)

Sample from 2006 Terabyte Track, Adhoc Task

• Title: Big Dig pork
• Description: Why is Boston’s Central Artery project, also known as “The Big Dig”, characterized as “pork”?
• Narrative: Relevant documents discuss the Big Dig project, Boston’s Central Artery Highway project, as being a big rip-off to American taxpayers or refer to the project as “pork”. Not relevant are documents which report fraudulent acts by individual contractors. Also not relevant are reports of cost-overruns on their own.

Pooling

• Human judges can’t look at all docs in collection: thousands to billions and growing
• Pooling chooses subset of docs of collection for human judges to rate relevance of
• Assume docs not in pool not relevant
How construct pool for a topic?
Let competing search engines decide:

- Choose a parameter $k$ (typically 100)
- Choose the top $k$ docs as ranked by each search engine
- Pool = union of these sets of docs
  Between $k$ and (# search engines) * $k$ docs in pool
- Give pool to judges for relevance scoring

Pooling cont.

- $(k+1)^{st}$ doc returned by one search engine
  either irrelevant or ranked higher by another search engine in competition

- In competition, each search engine is judged on results for top $r > k$ docs returned

Web search evaluation

Kinds of searches do on collection of journal articles or newspaper articles less varied than what do on Web.

What are different purposes of Web search?

Web search evaluation

- Different kinds of queries identified in TREC Web Track – some are:
  - Ad hoc
  - Topic distillation: set of key resources small, 100% recall?
  - Home page: # relevant pages = 1 (except mirrors)
  - Distinguish for competitors or just judges?
- Andrei Broder (Yahoo! Research) gave similar categories
  - Information
    - Broad research or single fact?
  - Transaction
  - Navigation

More web/online issues

- Are browser-dependent and presentation dependent issues:
  - On first page of results?
  - See result without scrolling?

Other issues in evaluation

- Does retrieving highly relevant documents really satisfy users?
  - Subjectivity?
- Are there dependences not accounted for?
- Many searches are interactive