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

$$= \frac{2\text{*recall*precision}}{\text{precision+recall}}$$

[‡] The harmonic mean *h* of two numbers *m* and *n* satisfies $(n-h)/n = (h-m)/m$. Also $(1/m) -(1/h) = (1/h)-(1/n)$

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!

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/ #
6. **Hazardous Waste - US Environ. Protection Agency**
   www.epa.gov/ebtpages/wasthazardouswaste.html
7. **toxic waste — Infoplease.com** toxic waste is waste ...
   www.infoplease.com/ce6/sci/A0849189.html
8. **Toxic Waste Clothing** Toxic Waste Clothing is a trend...
   www.toxicwasteclothing.com/ a

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

X 4. **Toxic Waste Candy Online** Toxic Waste Sour Candy ...
www.candydynamics.com/ #

X 5. **Toxic Waste Candy Online** Toxic Waste … chew bars...
www.toxicwastecandy.com/ #

✓ 6. **Hazardous Waste - US Environ. Protection Agency**
www.epa.gov/ebtpages/wasthazardouswaste.html

✓ 7. **toxic waste — Infoplease.com** toxic waste is waste ...
www.infoplease.com/ce6/sci/A0849189.html

X 8. **Toxic Waste Clothing** Toxic Waste Clothing is a trend...
www.toxicwasteclothing.com/ a

precison at rank

✓ 1. 1 **ic waste - Wikipedia, the free encyclopedia**
ikipedia.org/wiki/Toxic_waste

✓ 2. **ic Waste** Household toxic and hazardous waste ...
1 .urbanedpartnership.org/target/units/recycle/toxic.html

✓ 3. **ic Waste Facts, Toxic Waste Information**
1 ronment.nationalgeographic.com/.../toxic-waste-overview.html

X 4. **ic Waste Candy Online** Toxic Waste Sour Candy ...
3/4 .candydynamics.com/ #

X 5. **ic Waste Candy Online** Toxic Waste … chew bars...
3/5 .toxicwastecandy.com/ #

✓ 6. **ardous Waste - US Environ. Protection Agency**
2/3 .epa.gov/ebtpages/wasthazardouswaste.html

✓ 7. **ic waste — Infoplease.com** toxic waste is waste ...
5/7 .infoplease.com/ce6/sci/A0849189.html

X 8. **ic Waste Clothing** Toxic Waste Clothing is a trend...
5/8 .toxicwasteclothing.com/ a

Plot: precision versus recall

- Choose standard recall levels: r_1, r_2 …
  - r_j increasing
    e.g. 10%, 20% …
- Define “precision at recall level r_j”
  $p(r_j)$ = max over all r with $r_j \le r < r_{j+1}$ of
  precision when recall r achieved
- Smooth: “interpolated precision”
  $p_{interp}(r_i)$ = max over all r_j with $j \ge i$ of $p(r_j)$

See precision vs recall plot in the presentation “Overview of TREC 2004” by Ellen Voorhees.

available from TREC presentations Web site:
trec.nist.gov/presentations/TREC2004/04overview.pdf

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?

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

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

X 4. **Toxic Waste Candy Online** Toxic Waste Sour Candy ...
www.candydynamics.com/ #

X 5. **Toxic Waste Candy Online** Toxic Waste … chew bars...
www.toxicwastecandy.com/ #

✓ 6. **Hazardous Waste - US Environ. Protection Agency**
www.epa.gov/ebtpages/wasthazardouswaste.html

✓ 7. **toxic waste — Infoplease.com** toxic waste is waste ...
www.infoplease.com/ce6/sci/A0849189.html

X 8. **Toxic Waste Clothing** Toxic Waste Clothing is a trend...
www.toxicwasteclothing.com/ a

query: "toxic waste"

✓9. **Jean Factory Toxic Waste Plagues Lesotho**
www.cbsnews.com/stories/2009/08/02/.../main5205416.shtml

X10. **Ecopopulism: toxic waste and the movement for environmental justice** - Google Books Result
books.google.com/books?isbn=0816621756..

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

even more single number characterizations

Reciprocal rank:

Capture how early get relevant result in ranking

reciprocal rank of ranked results of a query

$$= \frac{1}{\text{rank of highest ranking relevant result}}$$

- perfect = 1 → worse → 0
- = average precision if only one relevant document

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

Example

✓ = relevant

rank	rel.	rel.	rel.
1	✓		
2		✓	✓
3			
4	✓	✓	✓
5	✓	✓	✓
6			
7			
8			
9	✓	✓	✓
10	✓	✓	

precision at rank 5 = 3/5 for all

reciprocal rank = 1

reciprocal rank = 1/2

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 doc.s?
  - 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 ($2^{\text{assessor's score}}$ -1)
- Let $d_1, d_2, \ldots d_k$ be returned docs in rank order
- $G(i)$ = gain value of d_i
  - determined by relevance class of d_i
- $DCG(i) = \sum_{j=1}^{i} \left(G(j) / (\log_2 (1+j) \right)$

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

rank	gain	
1	4	DCG(1) = $4/\log_2 2$ = 4
2	0	DCG(2) = 4 + 0 = 4
3	0	DCG(3) = 4 + 0 = 4
4	1	DCG(4) = 4 + $1/\log_2 5$ = 4.43
5	4	DCG(5) = 4.43 + $4/\log_2 6$ = 5.98
6	0	DCG(6) = 5.98 + 0 = 5.98
7	0	DCG(7) = 5.98 + 0 = 5.98
8	0	DCG(8) = 5.98 + 0 = 5.98
9	1	DCG(9) = 5.98 + $1/\log_2 10$ = 6.28
10	1	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) / (¼(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.)

As appeared in "Overview of TREC 2007" E. M. Voorhees, *Sixteenth Text REtrieval Conference Proceedings*.

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.

As appeared in "The TREC 2006 Terabyte Track" Büttcher, Clarke and Soboroff, *Fifteenth Text REtrieval Conference Proceedings*.

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 that 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