Evaluation of Retrieval Systems

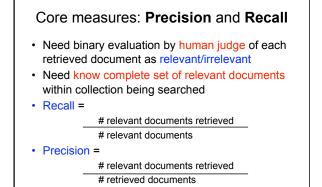
Performance Criteria 1. Expressiveness of query language Can query language capture information needs? 2. Quality of search results Relevance to users' information needs 3. Usability Search Interface Results page format Other? 4. Efficiency Speed affects usability Overall efficiency affects cost of operation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

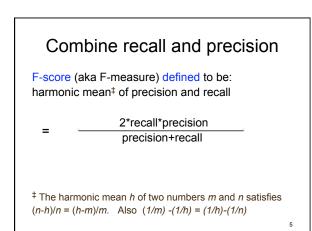
Quantitative evaluation

- · 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

3

 More generally "document" could be any information object





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 mth doc in ranking is kth 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
- Toxic Waste Facts, Toxic Waste Information environment.nationalgeographic.com/.../toxic-waste-overview.html
- Toxic Waste Candy Online Toxic Waste Sour Candy ... www.candydynamics.com/ #
- 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

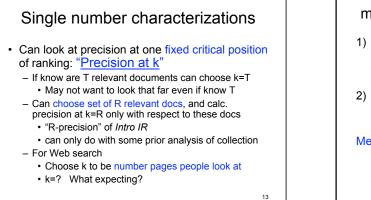
query: "toxic waste"

- ✓ 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



Plot: precision versus recall Choose standard recall levels: r₁, r₂ ... - r_j increasing e.g. 10%, 20% ... Define "precision at recall level r_j" p(r_j) = max over all r with r_j≤r<r_{j+1} of precision when recall r achieved Smooth: "interpolated precision" p_{interp}(r_j) = max over all r_j with j≥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



more single number characterizations

- 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
- Average the recorded precisions in (1)
 - = average precision for a query result
 - 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

14

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



even more single number characterizations

Reciprocal rank:

Capture how early get relevant result in ranking

reciprocal rank of ranked results of a query

rank of highest ranking relevant result

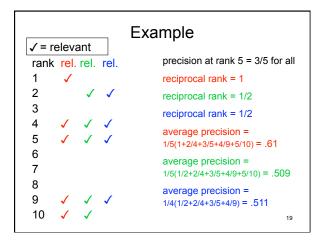
• perfect =
$$1 \rightarrow \text{worse} \rightarrow 0$$

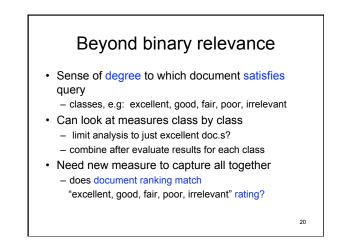
= average precision if only one relevant document

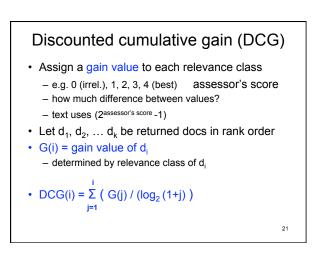
get mean reciprocal rank of set of test queries₁₇

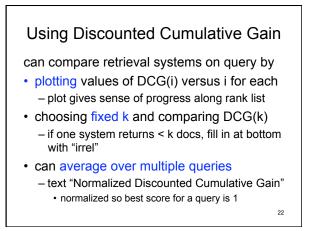
Summary Collection of measures of how well ranked search results provide relevant documents based on precision based to some degree on recall single numbers:

- precision at fixed rank
- average precision over all positions of relevant docs
- reciprocal rank of first relevant doc

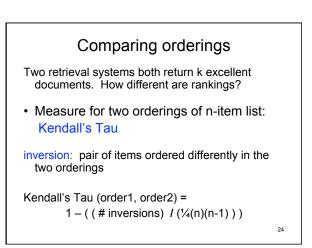




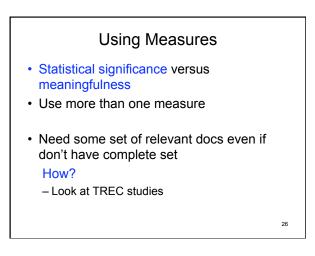


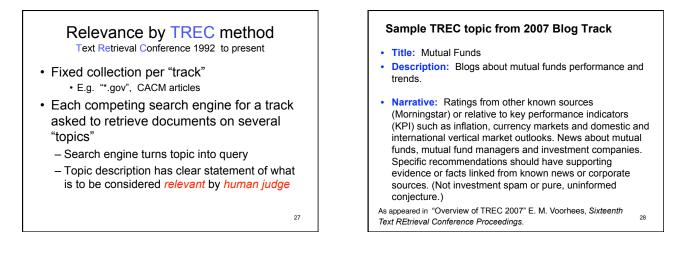


		Example
rank	gain	
1	4	$DCG(1) = 4/log_2 = 4$
2	0	DCG(2) = 4 + 0 = 4
3	0	DCG(3) = 4 + 0 = 4
4	1	$DCG(4) = 4 + \frac{1}{\log_2 5} = 4.43$
5	4	$DCG(5) = 4.43 + \frac{4}{\log_2 6} = 5.98$
6	0	DCG(6) = 5.98 + 0 = 5.98
7	0	DCG(7) = 5.98 + 0 = 5.98
8	0	DCG(8) = 5.98 + <mark>0</mark> = 5.98
9	1	DCG(9) = 5.98 + <mark>1/log₂10</mark> = 6.28
10	1	DCG(10) = 6.28 + 1/log ₂ 11 = 6.57
		23



			Example
doc	rank1	1 rank2	
Α	1	3	
В	2	4	
С	3	1	
D	4	2	
			, A-D, B-C, B-D = 4 4/3 = -1/3





25

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.* 29

Pooling

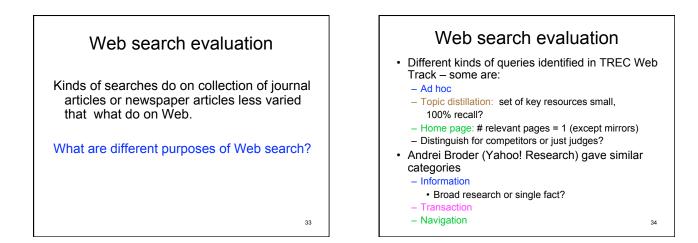
- Human judges can't look at all docs in collection: thousands to billions and growing
- Pooling chooses subset of docs of collection for human judges to rate relevance of
- · Assume docs not in pool not relevant

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



31

35

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?
 - Subjectivity?
- Are there dependences not accounted for?
- · Many searches are interactive