

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

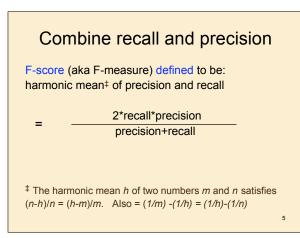
3

More generally "document" could be any information object

Core measures: Precision and Recall

2

- 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 =
 # relevant documents retrieved
- # relevant documents
 Precision =
- # relevant documents retrieved
- # retrieved documents



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
 - Google's "about N results" inadequate estimate
- Can focus on points that 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

Plot: precision versus recall

- Choose standard recall levels: r₁, r₂...
 Eg 10%, 20% ...
 - Define "precision at recall level ri"
 - $p(r_j) = max \text{ over all } r \text{ with } r_j \leq r < r_j + 1 \text{ of}$
 - precision when recall r achieved
 - Similar to Intro IR "interpolated precision"

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 I

- 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?

10

8

Single number characterizations II 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 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

11

7

9

Single number characterizations III Reciprocal rank:

Capture how early get relevant result in ranking

1

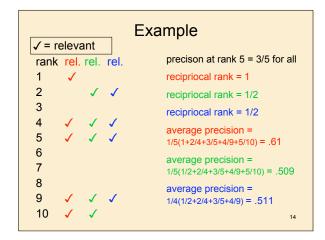
reciprocal rank of ranked results of a query

- 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 queries

Summary so far

- 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
 - recipriocal rank of first relevant doc



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?

15

13

Discounted cumulative gain (DCG)

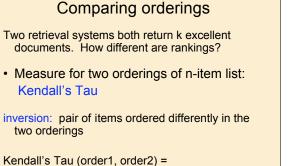
- Assign a gain value to each relevance class
 e.g. 0 (irrel.), 1, 2, 3, 4 (best) assessor's score
 how much difference between values?
 - text uses (2^{assessor's score}-1)
- Let $d_1, d_2, \dots d_k$ be returned docs in rank order
- G(i) = gain value of d_i
 determined by relevance class of d_i

•
$$DCG(i) = \sum_{j=1}^{i} (G(j) / (\log_{b} (1+j)))$$

 parameter b: how much doc retrieved lower down in ranking is penalized – text uses b=2

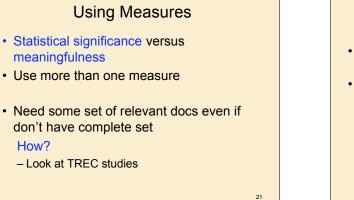
Using Discounted Cumulative Gain can compare retrieval systems on query by plotting values of DCG(i) versus i for each plot gives sense of progress along rank list choosing fixed k and comparing DCG(k) if one system returns < k docs, fill in at bottom with "irrel" can average over multiple queries text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

Example			
rank gain			
1	4	$DCG(1) = \frac{4}{\log_2 2} = 4$	
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 + <mark>4/log₂6</mark> = 5.98	
6	0	DCG(6) = 5.98 + <mark>0</mark> = 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	



1 - ((# inversions) / (¼(n)(n-1)))

Example doc rank1 rank2 А 1 3 В 2 4 С 3 1 D 4 2 # inversions: A-C, A-D, B-C, B-D = 4 Kendall tau = 1 - 4/3 = -1/3 20



19

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

22

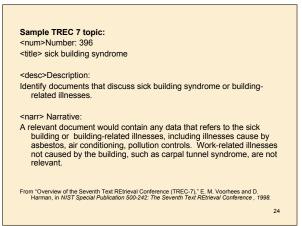
Sample TREC 3 topic:

<num> Number: 168 <title> Topic: Financing AMTRAK

<desc> Description:

How?

- A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).
- <narr> Narrative: A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with the substance AMTRAK substance to a provide the provide the substance. with those provided to AMTRAK would also be relevant. </top>
- As appeared in "Overview of the Sixth Text REtrieval Conference (TREC-6)," E. M. Voorhees and D. Harman, in NIST Special Publication 500-240: The Sixth Text REtrieval Conference, 1997.



Pooling

- Human judges can't look at all docs in collection: thousands to millions
- 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

26

 (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 evaluationKinds of searched 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 gave similar categories
 - Information
 - Broad research or single fact?
 - Transaction
 - Navigation

29

25

27

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