Evaluation of Retrieval Systems

1

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?

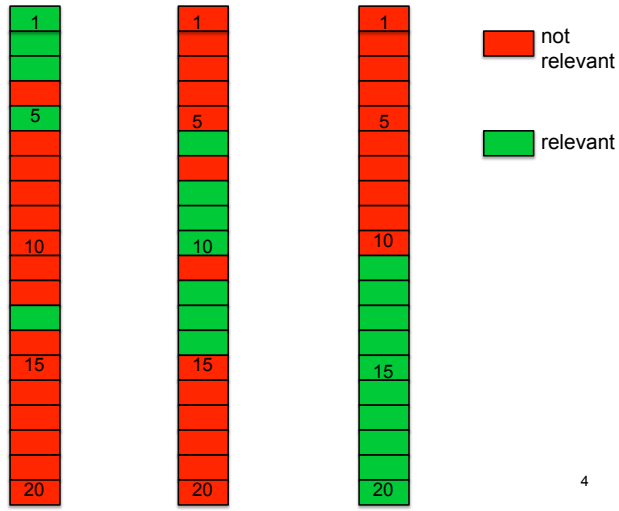
2

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

3

Example: 3 different search results – 1st 20



4

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 =

$$\frac{\text{\# relevant documents retrieved}}{\text{\# retrieved documents}}$$

5

Combine recall and precision

F-score (aka F-measure) defined to be:
harmonic mean‡ of precision and recall

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

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

6

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!

7

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

8

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

9

precison at rank

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

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

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

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

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

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

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

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

10

Single number characterizations

- "Precision at k": look at precision at one fixed critical position k of ranking
- Examples:
  - If know are T relevant docs can choose k=T
    - May not want to look that far even if know T
  - For Web search
    - Choose k to be number pages people look at
    - k=? What expecting?

11

more single number characterizations

average precision for a query result

1) Record precision at each point a relevant document encountered through ranked list
   - Can cut off ranked list at predetermined rank
2) Divide the sum of the recorded precisions in (1) by the total number of relevant documents
   = average precision for a query result
   - need know how many relevant docs in collection

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

12

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

13

query: "toxic waste"

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

X **10. Ecopopulism: toxic waste and the movement for environmental justice** - Google Books Result
books.google.com/books?isbn=0816621756..

Suppose there are 15 relevant documents in the collection

THEN precision at rank 10 is 0.6 and

average precision at rank 10 is 0.337

= (1/1+2/2+3/3+4/6+5/7+6/9)/15

14

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

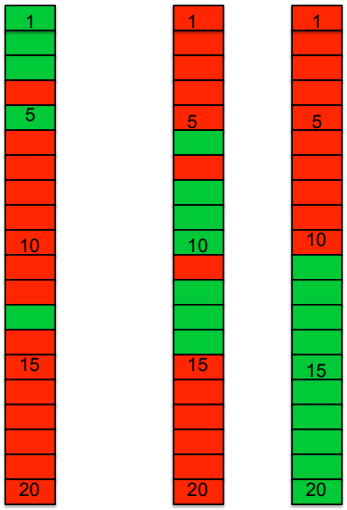
15

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

16

Example: 3 different search results – 1st 20



reciprocal rank:
1 1/6 1/11

precision at rank 5:
4/5 0 0

average precision at 20 if only 10 relevant docs in collection:
1/10*(1/1+2/2+3/3+4/5+5/13)= .418

1/10*(1/6+2/8+3/9+4/10+5/12+6/13+7/14)=.253

1/10*(1/11+2/12+3/13+4/14+5/15+6/16+7/17+8/18+9/19+10/20 =.331

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
- $DCG(i) = \sum_{j=1}^{i} (G(j) / (\log_2 (1+j)))$

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

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

21

Expected reciprocal rank (ERR)

Chapelle et al of Yahoo Labs, 2009 CIKM

- Models "expected reciprocal length of time that user will take to find a relevant document" [authors]
- DCG assumes user's response to document at rank i is independent of documents at rank < i
- research using click behavior shows likelihood user examines doc at rank i depends on quality of docs higher up list (lower rank)

22

Definition ERR

$$ERR = \sum_{j=1}^{n} \left((1/j) * \prod_{k=1}^{j-1} (1-R(score_k)) * R(score_j) \right)$$

where

$R(score) = (1/2^{max}) * (2^{score} - 1)$

for scores 0, 1, …, max

23

using ERR

- calculate one ERR score for desired number returned results
- ERR correlates better with click data
- Primary effectiveness measure for recent TREC

24

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) =

$$1 - ((\#\text{ inversions}) \; / \; (¼(n)(n-1)))$$

25

Example

ranking 1	rank	ranking 2
A	1	C
B	2	D
C	3	A
D	4	B

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

26

Comparing orderings

- Second measure:
  Spearman's rank order correlation (rho)

d_i = difference between ranks of item i in two orderings (distance)

Spearman's rho (order1, order2) =

$$1 - ((6 \sum_i d_i^2) \; / \; (n(n^2-1)))$$

Approx. based on Pearson correlation coefficient:

$$\sum_i (x_i-\bar{x})(y_i-\bar{y}) \; / \; \text{sqrt}(\sum_i (x_i-\bar{x})^2 \sum_i (y_i-\bar{y})^2)$$

x_i is position of item i in order 1; y_i is position of item i in order 2;

27

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

28

Relevance by TREC method

Text Retrieval Conference 1992 to present

- Fixed collection per "track"
  - E.g. "*.gov", CACM articles, Web
- 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*

29

Sample TREC topic from 2010 Blog Track

- Query: chinese economy
- Description: I am interested in blogs on the Chinese economy.
- Narrative: I am looking for blogs that discuss the Chinese economy. Major economic developments in China are relevant, but minor events such as factory openings are not relevant. Information about world events, or events in other countries is relevant as long as the focus is on the impact on the Chinese economy.

As appeared in "Overview of TREC-2010" by Iadh Ounis, Graig Macdonald, and Ian Soboroff, *NineteenthText REtrieval Conference Proceedings*.

30

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

31

How construct pool for a topic? Let competing search engines decide:

- Choose a parameter k
  k=30 for 2012 TREC Web track (48 entries)
- 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

32

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
  - r = 10,000 for 2012 TREC Web track
- Entries compared by quantitative measures

33

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?

34

Web search evaluation

- Different kinds of tasks identified in TREC Web Track – some are:
  - Ad hoc
  - Diversity: "return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list"
  - Home page: # relevant pages = 1 (except mirrors)
- Andrei Broder gave similar categories (2002)
  - Information
    - Broad research or single fact?
  - Transaction
  - Navigation

35

More web/online issues

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

36

Other issues in evaluation

- Are there dependences not accounted for?
  - ad placement?
- Many searches are interactive

37

Google v.s. DuckDuckGo

- Class experiment – Problem set 2

From duck.co/help/results/sources:

- over one hundred sources, some are
  - DuckDuckBot (own crawler),
  - crowd-sourced sites (like Wikipedia)
  - Bing
- intelligence layer attempts improve results, e.g
  - pick the best source
  - remove spam
  - re-rank

38