

# Evaluation and Relevance (2)

Lecture 9

CS 410/510

Information Retrieval on the Internet

## Limitations of test collections

- Scale
  - Early ones too small
  - Large ones too expensive
  - Pooled relevance judgments incomplete
- Doesn't reflect the way users work
  - Interactivity
  - Query formulation
- Dependence on relevance judgments

## Assumptions of Cranfield Paradigm

- All relevant documents are known
  - Violated in large test collections
- Single set of judgments for a topic is representative of user population
- Relevance can be approximated by topical similarity; implies:
  - All relevant docs equally desirable
  - Relevance of one doc independent of other docs
  - User information need is static

Based on: Voorhees, EM. *The Philosophy of Information Retrieval Evaluation*. CA Peters et al. (Eds): CLEF 2001, LNCS 2406, pp 355-370, 2002.

CS 510 Winter 2007

3

## Incomplete relevance judgments

- Use sufficiently deep pools
  - 100 documents in TREC
- Use metric that is more robust to incomplete relevance judgments
  - bpref
- Use selection instead of pooling
  - Select document most likely to discriminate between systems being compared (based on its effect on AP)
  - Stop judging when desired confidence level is reached

CS 510 Winter 2007

4

## Inconsistency of relevance judgments

- Sormunen, 2002
  - Reassessed 38 TREC topics to assign graded relevance criteria
  - Of documents judged irrelevant by TREC assessors, 94% judged irrelevant again
  - Of documents judged relevant by TREC assessors, 25% rated irrelevant and 36% rated marginally relevant

CS 510 Winter 2007

5

## Inconsistency of relevance judgments

- TREC assessor agreement
  - Overlap (intersection of relevant document sets from each assessor) < 50%
  - Testing permutations based on the relevance judgments of 3 assessors
    - Values of metrics changed
    - System rankings remained highly correlated
  - Assessor disagreement unlikely to alter results of system performance comparisons

CS 510 Winter 2007

6

# What is relevance?

*“a measure of the effectiveness of a contact between a source and a destination in a communication process”*

- Saracevic, JASIS 1975

- **Aboutness:** document *d* is *about* topic *T*
  - Intellectual assessment of whether *d* is about *T*
- **Pertinence** – as perceived by a user
  - User interpretation of information need, of document
- **Situational relevance**
  - Usefulness in a particular situation or context
  - Related to a particular task,
  - Related to user’s existing knowledge

CS 510 Winter 2007

7

# What is relevance?

- **Multidimensional**
  - Relevance assessment differs among users
    - Perspective, aspects of interest
    - Level of knowledge
- **Dynamic**
  - Relevance describes relationship between information and need at a particular time
  - Relevance changes over time for same user
    - As user accumulates additional information
    - As context changes

CS 510 Winter 2007

8

## Is relevance binary?

- Some documents provide more information than others
  - Cover more aspects of an information need
  - Cover the topic in more depth
  - Cover the topic from a more desirable perspective (for a particular user)
- Graded relevance judgments

CS 510 Winter 2007

9

## Graded relevance judgments: Example

- Irrelevant document
  - Contains no information about topic
- Marginally relevant document
  - Only points to the topic. Does not contain more or other information than the topic statement
- Fairly relevant document
  - Contains more information than topic statement but is not exhaustive. If topic multi-faceted, only covers some sub-themes
- Highly relevant document
  - Discusses topic exhaustively. If topic multi-faceted, covers most sub-themes

Paraphrased from: Järvelin and Kekäläinen. Cumulated gain-based evaluation of IR techniques. ACM TOIS, Vol 20, pp 422-446, 2002

CS 510 Winter 2007

10

- Relevance assessment exercise

## Using graded relevance judgments

- Only consider “highly relevant” documents
  - Apply a threshold to create binary judgments
  - Fewer relevant documents, less stable results
- Evaluate a ranked list by cumulated gain
  - Cumulative gain (CG)
    - Highly relevant documents contribute more value than marginally relevant documents
  - Discounted cumulative gain (DCG)
    - And, relevant documents contribute more value appearing at higher ranks than at lower ranks

## Cumulative Gain (CG)

Let  $G$  be a vector of values representing the graded relevance judgments for each document in a ranked list. Let  $G[i]$  be the graded relevance of the document in the  $i$ -th position of a results list.

$$CG[i] = \begin{cases} G[1] & \text{if } i = 1 \\ CG[i - 1] + (G[i]) & \text{otherwise} \end{cases}$$

CS 510 Winter 2007

13

## Discounted Cumulative Gain (DCG)

Let  $G$  be a vector of values representing the graded relevance judgments for each document in a ranked list. Let  $G[i]$  be the graded relevance of the document in the  $i$ -th position of a results list.

$$DCG[i] = \begin{cases} CG[i] & \text{if } i < b \\ DCG[i - 1] + (G[i] / \log_b i) & \text{if } i \geq b \end{cases}$$

Choice of  $b$  allows modeling user impatience vs. persistence

Smaller values of  $b$  cause greater discounting of documents retrieved at lower ranks.  $b = 2$  models a more impatient user while  $b = 10$  models a more persistent user, willing to examine more documents.

CS 510 Winter 2007

14

## Discounted Cumulative Gain (DCG)

$$\text{DCG}[i] = \begin{cases} \text{CG}[i] & \text{if } i < b \\ \text{DCG}[i - 1] + (G[i] / \log_b i) & \text{if } i \geq b \end{cases}$$

Rank	DocID	Relev level
1	0234	0
2	0132	2
3	0115	3
4	0193	0
5	0123	1
6	0345	3
7	0387	0
8	0256	2
9	0078	1
10	0311	2

### Example:

CG vector: <0, 2, 5, 5, 6, 9, 9, 11, 12, 14>

DCG vector: <0, 2, 3.9, 3.9, 4.3, 5.5, 5.5, 6.2, 6.5, 7.1>  
( $b=2$ ; divide by na, na, 1.58, 2, 2.32, 2.58 ...)

Average the vectors over a set of queries to get average performance

CS 510 Winter 2007

15

## Cumulative Gain metrics

- Relevance assessments may be made from an ordinal scale
- Assign weights to levels of relevance
  - Turns an ordinal scale into a ratio scale
- Previous example shows a 4-point relevance scale with weights 0, 1, 2, 3
- Could assign 0, 1, 10, 100
  - Or any other weighting that fits use scenario

CS 510 Winter 2007

16

## normalized DCG (nDCG)

- Normalize cumulative gain (or discounted cumulative gain) by comparing results to theoretical best results for each query.
  - Create ideal vector
    - Fill first  $i$  positions with value for highest relevance level
    - Fill next  $j$  positions with value for next relevance level ...
    - Where  $i$  = num docs at highest relevance
    - And  $j$  = num docs at next highest level ...
  - Does **not** assume a theoretical best result of all docs being relevant
- Divide CG or DCG vector by the ideal vector to get normalized vector

CS 510 Winter 2007

17

## Normalized DCG (nDCG)

$$DCG[i] = \begin{cases} CG[i] & \text{if } i < b \\ DCG[i-1] + (G[i] / \log_b i) & \text{if } i \geq b \end{cases}$$

Rank	Relev level
1	0
2	2
3	3
4	0
5	1
6	3
7	0
8	2
9	1
10	2

**Example ( $b = 2$ ):**

CG vector: <0, 2, 5, 5, 6, 9, 9, 11, 12, 14>

DCG vector: <0, 2, 3.90, 3.90, 4.33, 5.49, 5.49, 6.16, 6.47, 7.08>

Ideal vector: <3, 3, 2, 2, 2, 1, 1, 0, 0, 0>

Ideal CG vector: <3, 6, 8, 10, 12, 13, 14, 14, 14, 14>

Ideal DCG: <3, 6, 7.26, 8.26, 9.12, 9.51, 9.87, 9.87, 9.87, 9.87>

nDCG: <0, 0.33, 0.54, 0.47, 0.47, 0.58, 0.56, 0.63, 0.66, 0.72>

CS 510 Winter 2007

18

## Rpref

- Proposed generalization of bpref for use with graded relevance judgments
  - Weighted counts of documents ranked higher than documents judged more relevant
  - A mis-ordered pair of document incurs a penalty proportional to the difference in their relevance values
  - New metric, not well-studied yet

CS 510 Winter 2007

19

## Do batch results predict user results?

- Hersh, et al., 2001 TREC Interactive track
  - Instance recall task
    - e.g. *Find all the discoveries made by the Hubble telescope*
  - Two systems with identical user interfaces
    - Similarity algorithm in experimental system shown to be better (17.6% better instance precision) than in baseline system (for *this* document collection, using description of task as a query in batch mode)
  - All 24 subjects searched 6 topics, 3 with each system
  - No difference in user performance between systems
  - Queries? Reading speed? Reading comprehension?

CS 510 Winter 2007

20

## Do batch results predict user results?

- Turpin and Hersh, 2001
  - Design similar to instance recall study: same UI, different underlying system performance
    - Improved system 67% higher MAP on users' queries
  - New task: Find answers to 2 types of questions
    - Find all small number of answers for a topic, e.g. *Name four films in which Orson Welles actually appeared*
    - Select correct answer (of two), e.g. *Is Denmark larger or smaller in population than Norway?*
  - Measured user's rate of answering questions correctly
  - Found user performance 6% worse with "better" IR system (statistically insignificant)

CS 510 Winter 2007

21

## Do batch results predict user results?

- Turpin and Scholer, 2006
  - Two tasks
    - Find one relevant document (precision-oriented)
    - Find as many relevant docs as possible in 5 minutes (recall-oriented)
  - 30 students searched on 50 TREC topics
    - 10 systems (users searched 5 topics each system)
    - Each system returned doc list with a **known AP**
      - 0.55 – 0.95
      - Regardless of the user's query

CS 510 Winter 2007

22

## Do batch results predict user results?

- Turpin and Scholer, cont.
  - Precision-based task: *no* relationship between user performance and system performance
  - Recall-based task: *weak* relationship between user performance and system performance
    - Significant difference comparing MAP 55% to 75% and 65% to 75%; magnitude of effect very small
  - 47.2% of searches found no relevant documents in 5 minutes despite high AP!

CS 510 Winter 2007

23

- Next: Query expansion and relevance feedback

CS 510 Winter 2007

24