Information Retrieval

Evaluation

Intelligent Information Retrieval

Why Evaluation?

- **Reason 1: So that we can assess how useful an IR system/technology would be (for an application)**
  - **Measures should reflect the utility to users in a real application**
  - **Usually done through user studies (interactive IR evaluation)**
- **Reason 2: So that we can compare different systems and methods (to advance the state of the art)**
  - **Measures only need to be correlated with the utility to actual users, thus don't have to accurately reflect the exact utility to users**
  - **Usually done through test collections (test set IR evaluation)**

What to measure?

- **Incomplete list (Cleverdon 66):**
  1. ***"The ability of the system to present all relevant documents***
  2. ***The ability of the system to withhold non-relevant documents***
  3. ***The interval between the demand being made and the answer being given (i.e. time)***
  4. ***The physical form of the output (i.e. presentation)***
  5. ***The effort, intellectual or physical, demanded of the user (i.e. effort)."***
- **IR evaluation has so far focused more on 1. & 2. (unique challenges for IR)**

History of Test Set Evaluation: Pre-TREC

- **Early 1950s – 1960s (Cranfield): Establishment of the test collection evaluation methodology**
- **1960s- early1990s (Pre-TREC): Initial development of test collections and measures**
  - **Documents: mostly catalogue information about academic papers; later, full text news articles**
  - **Measures were strongly focused on *high recall* search: finding as many relevant items as possible.**

History of Test Evaluation: The TREC Era

- **Early 1990s – early 2000s ("TREC ad hoc" period): Larger-scale and standardization of evaluation**
  - **Documents: mostly news articles**
  - **Measures: still focused on high-recall search**
  - **A variety of retrieval tasks that go beyond standard ad hoc task**
- **Early 2000s – present (Post TREC ad hoc period): Diverse search applications + Larger scale**
  - **Documents: beyond news articles (web pages,…)**
  - **Measures: more reflecting a user's common preference for finding a small number of relevant item**
  - **A new form of evaluation research: studying test collection methodologies**

Cranfield Test Methodology

- Specify a retrieval task
- Create a collection of sample documents
- Create a set of topics/queries appropriate for the retrieval task
- Create a set of relevance judgments (i.e., judgments about which document is relevant to which query)
- Define a set of measures
- Apply a method to (or run a system on) the collection to obtain performance figures

Sample Test Collections (1960s-1970s)

Name	Docs.	Qrys.	Year	Size, Mb	Source document
Cranfield 2	1,400	225	1962	1.6	Title, authors, source, abstract of scientific papers from the aeronautic research field, largely ranging from 1945-1962.
ADI	82	35	1968	0.04	A set of short papers from the 1963 Annual Meeting of the American Documentation Institute.
IRE-3	780	34	1968	-	A set of abstracts of computer science documents, published in 1959-1961.
NPL	11,571	93	1970	3.1	Title, abstract of journal papers
MEDLARS	450	29	1973	-	The first page of a set of MEDLARS documents copied at the National Library of Medicine.
Time	425	83	1973	1.5	Full text articles from the 1963 edition of Time magazine.

Evaluating a Boolean Retrieval System

Cleverdon and Keen (1966, p. 34)

	Relevant	Not-relevant	
Retrieved	a (w)	b	a+b (m)
Not retrieved	c	d	c+d
	a+c (x)	b+d	a+b+c+d (n)

$$\text{Precision} = \frac{a}{a+b} \qquad \text{Recall} = \frac{a}{a+c}$$

$$\text{Fallout} = \frac{b}{b+d}$$

Summarizing precision and recall to a single value

- **Why summarizing?**
- **How to summarize?**

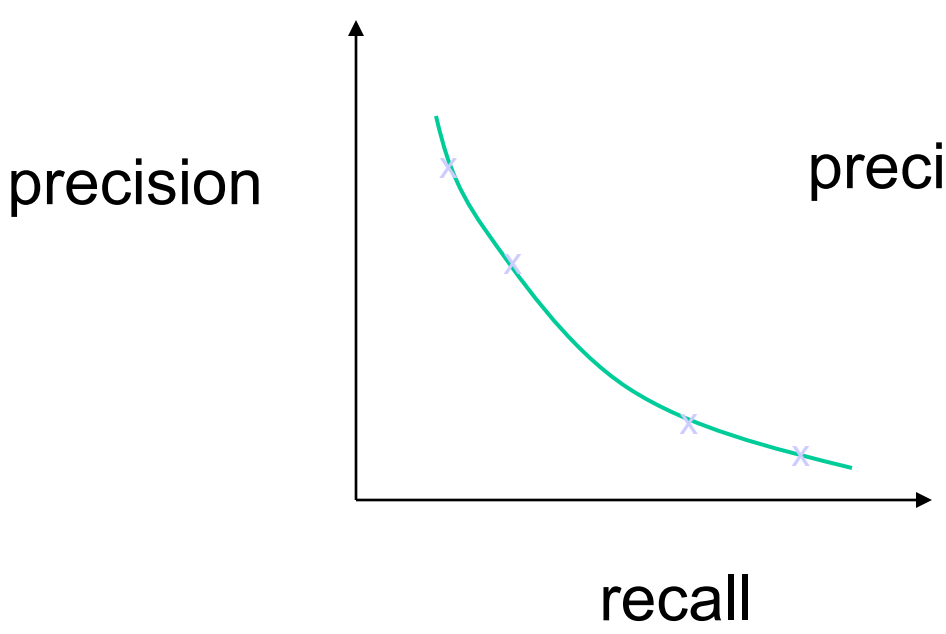
$$f = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)} \qquad F1 = \frac{1}{\frac{1}{2}\left(\frac{1}{R} + \frac{1}{P}\right)} = \frac{2PR}{P+R}$$

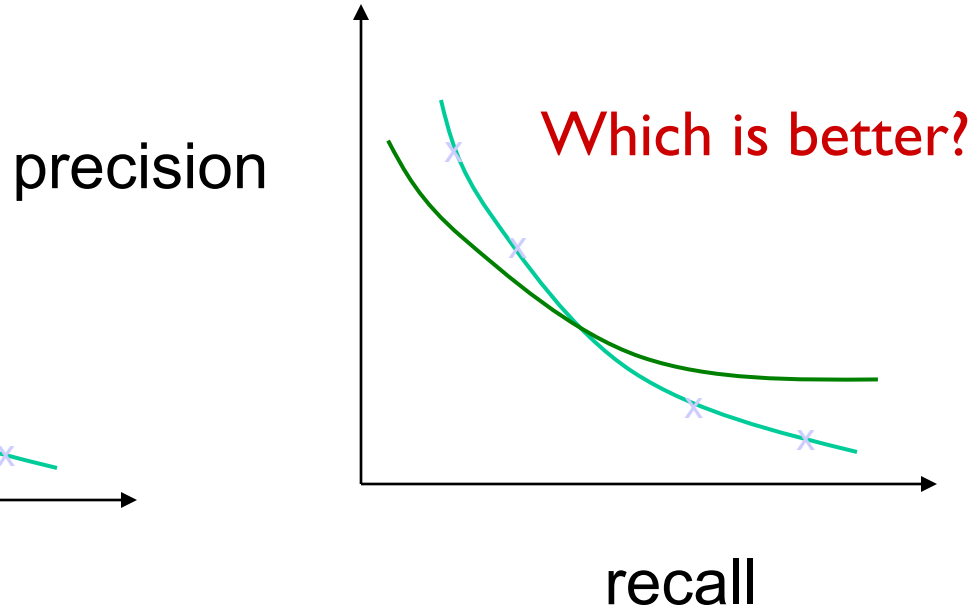
F-measure [Rijsbergen 79]

- **What's the implied tradeoff between precision and recall in the F measure?**

How to measure a ranking?

- **Compute the precision at every recall point**
- **Plot a precision-recall (PR) curve**

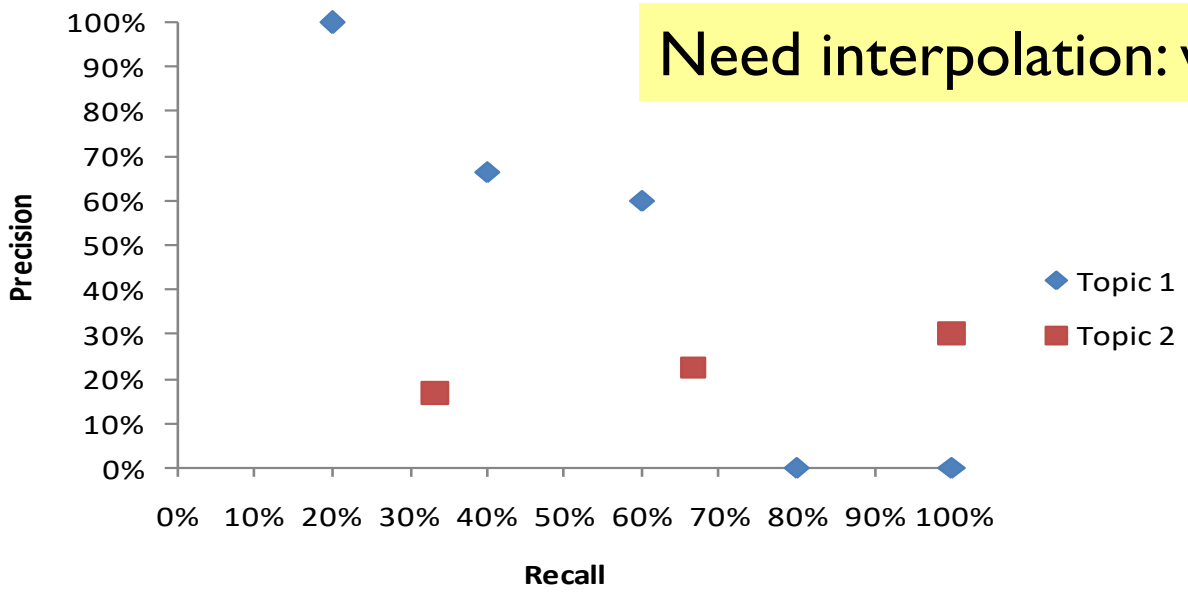




Evaluating ranked retrieval results: Precision-Recall curve

Rank	Rel	Pr	Rcl
1	1	100%	20%
2	0		
3	1	67%	40%
4	0		
5	1	60%	60%
6	0		
7	0		
8	0		
9	0		
10	0		
∞	1	0%	80%
∞	1	0%	100%

Rank	Rel	Pr	Rcl
1	0		
2	0		
3	0		
4	0		
5	0		
6	1	17%	33%
7	0		
8	0		
9	1	22%	67%
10	1	30%	100%



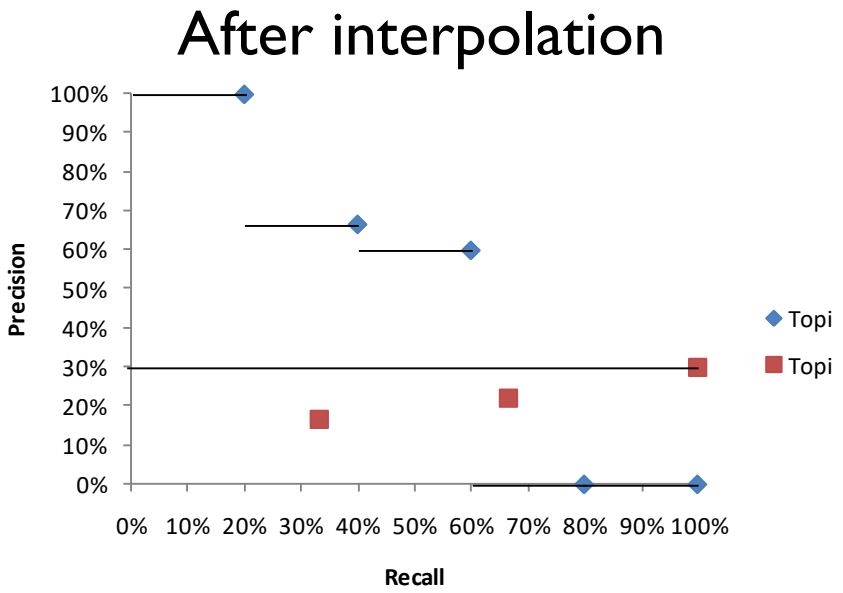
Need interpolation: why & how?

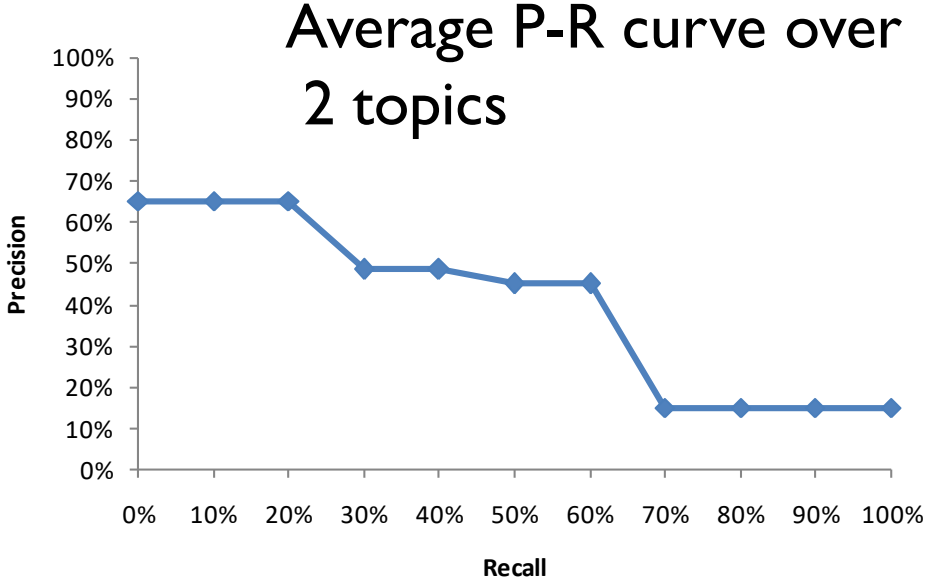
Keen's semi-Cranfield (neo-Cleverdon) interpolation

- **The interpolated value of precision measured at any recall level (r_i) was defined to be the highest precision measured at any level (r') greater than or equal to r_i.**

$$P_{interp}(r_i) = \max_{r' \geq r_i} P(r') \quad \text{[Manning et al. 08]}$$

- **Advantages?**





Cooper's Expected Search Length (ESL)

[Cooper 68]

- **Motivation:**
  - ***"most measures do not take into account a crucial variable: the amount of material relevant to [the user's] query which the user actually needs"***
  - ***"the importance of including user needs as a variable in a performance measure seems to have been largely overlooked"***
- **ESL = the amount of a ranking that had to be observed by a searcher in order to locate a pre-specified quantity of relevant documents**
  - **ESL for finding one relevant document = rank-of-the-1st-rel-doc**
- **ESL motivated cumulative gain measures, which are now quite influential**

Challenging assumptions in early collections

- **Assumption 1: queries sent to a computer system would be the same as those sent to a librarian (i.e., sentence-length request), and users want to have high recall**
  - **Fairthorne (63): query can be an ambiguous phrase**
- **Assumption 2: relevance = independent topical relevance**
  - **Verhoeff et al (1961): different users may judge different documents as relevant for the same query**
  - **Goffman (64): redundant information is useless**
  - **Cooper (71): utility vs. topical relevance (include other factors: credibility of sources, recency of docs, writing styles,…)**
  - **Saracevic's well known survey about relevance (1975)**
- **These suggestions mostly overlooked until recently**

Assessor consistency

- Relevance judgments are subjective: is inconsistency of assessors a concern?
- Studies mostly concluded that the inconsistency didn't affect relative ranking of systems
  - Lesk & Salton (1968): assessors mostly disagree on documents at lower ranks, but measures are more affected by top-ranked documents
  - Cleverdon (1970), Burgin (1992): similar conclusions
  - Harman (1994): 80% agreement between TREC assessors
- Judgments on relative ranking of documents are more consistent (Schultz 67)

Challenges in creating early test collections

- Challenges in obtaining documents:
  - Salton had students to manually transcribe Time magazine articles
  - Not a problem now!

- Challenges in distributing a collection
  - TREC started when CD-ROMs were available
  - Not a problem now!
- Challenge of scale – limited by qrels (relevance judgments)
  - The idea of "pooling" (Sparck Jones & Rijsbergen 75)

Larger collections created in 1980s

Name	Docs.	Qrys.	Year	Size, Mb	Source document
INSPEC	12,684	77	1981	-	Title, authors, source, abstract and indexing information from Sep-Dec 1979 issues of Computer and Control Abstracts.
CACM	3,204	64	1983	2.2	Title, abstract, author, keywords and bibliographic information from articles of Communications of the ACM, 1958-1979.
CISI	1,460	112	1983	2.2	Author, title/abstract, and co-citation data for the 1460 most highly cited articles and manuscripts in information science, 1969-1977.
LISA	6,004	35	1983	3.4	Taken from the Library and Information Science Abstracts database.

Commercial systems then routinely support searching over millions of documents

➔ Pressure for researchers to use larger collections for evaluation

The Ideal Test Collection Report [Sparck Jones & Rijsbergen 75]

- **Introduced the idea of pooling**
  - **Have assessors to judge only a pool of top-ranked documents returned by various retrieval systems**
- **Other recommendations (the vision was later implemented in TREC)**

1.that an ideal test collection be set up to facilitate and promote research;
2.that the collection be of sufficient size to constitute an adequate test bed for experiments relevant to modern IR systems…
3.that the collection(s) be set up by a special purpose project carried out by an experienced worker, called the Builder;
4.that the collection(s) be maintained in a well-designed and documented machine form and distributed to users, by a Curator;
5.that the curating (sic) project be encouraged to, promote research via the ideal collection(s), and also via the common use of other collection(s) acquired from independent projects."

TREC (Text REtrieval Conference)

- 1990: DARPA funded NIST to build a large test collection
- 1991: NIST proposed to distribute the data set through TREC (leader: Donna Harman)
- Nov. 1992: First TREC meeting
- Goals of TREC:
  - create test collections for a set of retrieval tasks;
  - promote as widely as possible research in those tasks;
  - organize a conference for participating researchers to meet and disseminate their research work using TREC collections.

The "TREC Vision" (mass collaboration for creating a pool)

"Harman and her colleagues appear to be the first to realize that if the documents and topics of a collection were distributed for little or no cost, a large number of groups would be willing to load that data into their search systems and submit runs back to TREC to form a pool, all for no costs to TREC. TREC would use assessors to judge the pool. The effectiveness of each run would then be measured and reported back to the groups. Finally, TREC could hold a conference where an overall ranking of runs would be published and participating groups would meet to present work and interact. It was hoped that a slight competitive element would emerge between groups to produce the best possible runs for the pool." (Sanderson 10)

The TREC Ad Hoc Retrieval Task

- **Simulate an information analyst (high recall)**
- **Multi-field topic description**
- **News documents + Government documents**
- **Relevance criteria: *"a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document)"***
- **Each run submitted returns 1000 documents for evaluation with various measures**
- **Top 100 documents were taken to form a pool**
- **All the documents in the pool were judged**

An example TREC topic

```
<top>

<num> Number:  200

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description:  Document must report on how the importation of foreign
textiles or textile products has influenced or impacted on the U.S. textile
industry.

<narr>  Narrative:  The impact can be positive or negative or qualitative.
It may include the expansion or shrinkage of markets or manufacturing volume
or an influence on the methods or strategies of the U.S. textile industry.
"Textile industry" includes the production or purchase of raw materials;
basic processing techniques such as dyeing, spinning, knitting, or weaving;
the manufacture and marketing of finished goods; and also research in the
textile field.

</top>
```

Main TREC ad hoc retrieval measure: Mean Average Precision

- Given that n docs are retrieved
  - Compute the precision (at rank) where each (new) relevant document is retrieved => p(1),…,p(k), if we have k rel. docs
  - E.g., if the first rel. doc is at the 2nd rank, then p(1)=1/2.
  - If a relevant document never gets retrieved, we assume the precision corresponding to that rel. doc to be zero
- Compute the average over all the relevant documents
  - Average precision = (p(1)+…p(k))/k
- This gives us an average precision, which captures both precision and recall and is sensitive to the rank of each relevant document
- Mean Average Precisions (MAP)
  - MAP = arithmetic mean average precision over
  - gMAP = geometric mean average precision over by difficult topics)

$$gMAP = \sqrt[n]{\prod_{i=1}^{n} AP_i}$$

	S1	S2
Q1 AP	0.02	0.04
Q2 AP	0.4	0.38
MAP	0.21	0.21
gMAP	0.089	0.123

Other TREC Measures

- Precision at k documents (e.g., prec@10doc):
  - more meaningful to users than MAP
  - tends to be higher for a larger collection
  - also called breakeven precision (R-precision) when k is the same as the number of relevant documents
- Mean Reciprocal Rank (MRR):
  - Same as MAP when there's only 1 relevant document
  - Reciprocal Rank = 1/Rank-of-the-relevant-doc
- R-precision: Precision at the rank that is equal to the total number of relevant documents

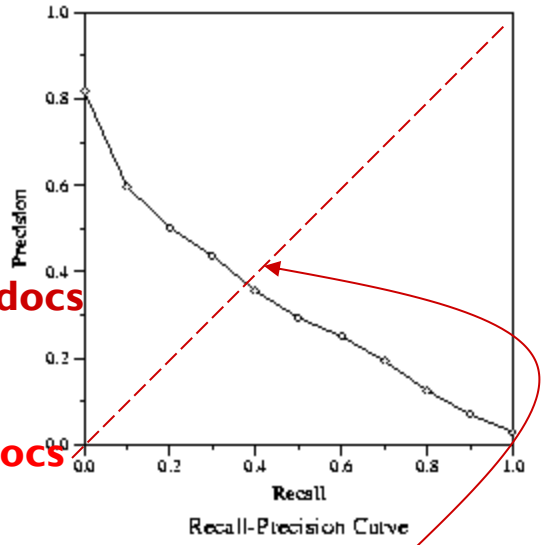
Typical TREC Evaluation Result (from "trec_eval")

Ad hoc results — Microsoft Research Ltd

Summary Statistics	
Run Number	ok8amxc
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel-ret:	3212

Out of 4728 rel docs, we've got 3212

Precion-Recall Curve



Recall-Precision Curve

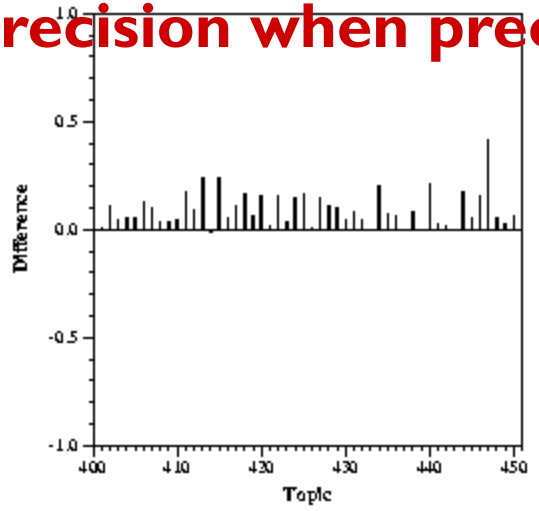
Recall Level Precision Averages	
Recall	Precision
0.00	0.8190
0.10	0.5975
0.20	0.5032
0.30	0.4372
0.40	0.3561
0.50	0.2936
0.60	0.2511
0.70	0.1941
0.80	0.1257
0.90	0.0696
1.00	0.0296
Average precision over all relevant docs	
non-interpolated	0.3169

Document Level Averages	
	Precision
At 5 docs	0.5800
At 10 docs	0.5500
At 15 docs	0.4987
At 20 docs	0.4650
At 30 docs	0.4253
At 100 docs	0.2680
At 200 docs	0.1921
At 500 docs	0.1085
At 1000 docs	0.0642
R-Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.3470

Precision@10docs

about 5.5 docs in the top 10 docs are relevant

Breakeven Precision (precision when prec=recall)



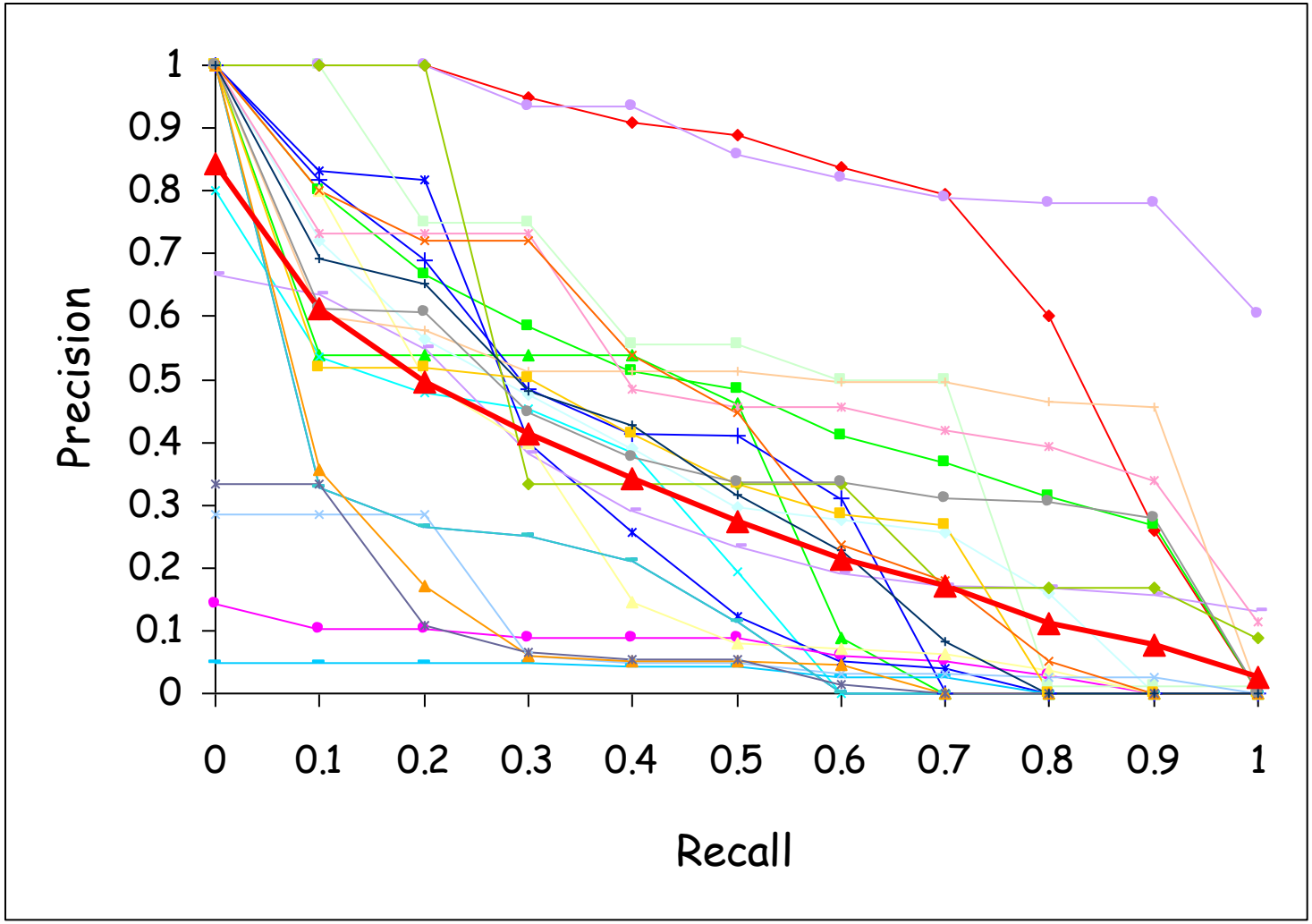
Difference from Median in Average Precision per Topic

Mean Avg. Precision (MAP)

D1 +
D2 +
D3 –
D4 –
D5 +
D6 -

Total # rel docs = 4
System returns 6 docs
Average Prec = (1/1+2/2+3/5+0)/4

What Query Averaging Hides



Slide from Doug Oard's presentation, originally from Ellen Voorhees' presentation

Statistical Significance Tests

- **How sure can you be that an observed difference doesn't simply result from the particular queries you chose?**

Experiment 1

Query	System A	System B
1	0.20	0.40
2	0.21	0.41
3	0.22	0.42
4	0.19	0.39
5	0.17	0.37
6	0.20	0.40
7	0.21	0.41
Average	0.20	0.40

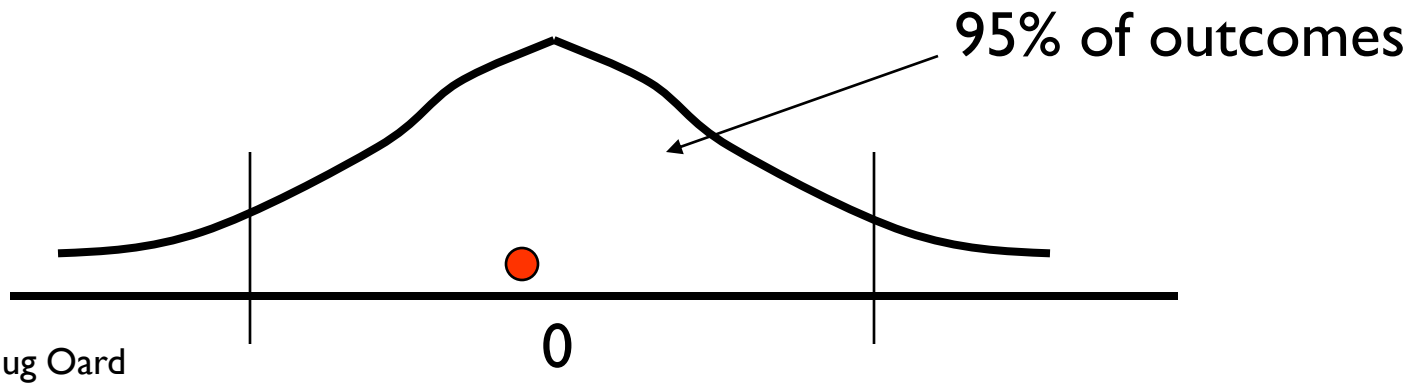
Experiment 2

Query	System A	System B
1	0.02	0.76
2	0.39	0.07
3	0.16	0.37
4	0.58	0.21
5	0.04	0.02
6	0.09	0.91
7	0.12	0.46
Average	0.20	0.40

Slide from Doug Oard

Statistical Significance Testing

Query	System A	System B	Sign Test	Wilcoxon
1	0.02	0.76	+	+0.74
2	0.39	0.07	-	- 0.32
3	0.16	0.37	+	+0.21
4	0.58	0.21	-	- 0.37
5	0.04	0.02	-	- 0.02
6	0.09	0.91	+	+0.82
7	0.12	0.46	-	- 0.38
Average	0.20	0.40	p=1.0	p=0.9375



Slide from Doug Oard

Try some out at: http://www.fon.hum.uva.nl/Service/Statistics.html

Other TREC-like Initiatives

- CLEF – The annual Cross Language Evaluation Forum
  - Initially European languages; later Persian and others
  - imageCLEF: image retrieval, medical retrieval
- NTCIR – The NII Test Collection for IR Systems
  - Asian languages such as Japanese, Chinese and Korean
  - Patent search
- TDT – Topic Detection and Tracking
  - the automatic detection and tracking of important emerging stories in streaming text
- INEX – The INitiative for the Evaluation of XML Retrieval
  - retrieval of document fragments
- TRECVID – an evaluation exercise focused on video retrieval
- A number of other smaller and/or newer evaluation exercises

Post Ad Hoc Collections and Measures

- Many new tasks leading to new collections, new topic sets, new measures
- Major issues addressed
  - Beyond binary judgments
  - Diversity
  - Managing unjudged documents
  - Relevance applied to parts of a document
  - Are all topics equal?
- All still current research topics in IR Evaluation

Multi-level relevance judgments: nDCG

- What if relevance judgments are in a scale of [1,r]? r>2
- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be r_1, r_2, …r_n (in ranked order)
  - CG = $r_1+r_2+\ldots+r_n$
- Discounted Cumulative Gain (DCG) at rank n
  - DCG = $r_1+r_2/\log_2 2+ r_3/\log_2 3+ \ldots r_n/\log_2 n$
  - We may use any base for the logarithm, e.g., base=b
  - For rank positions above b, do not discount
- Normalized Cumulative Gain (nDCG) at rank n
  - Normalize DCG at rank n by the DCG value at rank n of the ideal ranking so that all nDCG values would be within [0,1].
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc

nDCG Example

- **Relevance Scores:**

D1	D2	D3	D4	D5	D6
3	2	3	0	1	2

- **Ranked list:** $D1, D2, D3, D4, D5, D6$
- **CG =** (3, 5, 8, 8, 9, 11)
- **DCG =** (3, 3 + 2/log2, 3 + 2/log2 + 3/log3, 3 + 2/log2 + 3/log3, 3 + 2/log2 + 3/log3 + 1/log5, 3 + 2/log2 + 3/log3 + 1/log5 + 2/log6)
- **Ideal ranking based on relevance scores:** 3, 3, 2, 2, 1
- **IDCG =** (3, 3 + 3/log2, 3 + 3/log2 + 2/log3, 3 + 3/log2 + 2/log3 + 2/log4, 3 + 3/log2 + 2/log3 + 2/log4, 1/log5)
- **NDCG = DCG/IDCG**

More on new measures later in the course…

Questions?