Information Retrieval

Lecture 3: Evaluation methodology

Computer Science Tripos Part II

UNIVERSITY OF CAMBRIDGE

Simone Teufel

Natural Language and Information Processing (NLIP) Group

sht25@cl.cam.ac.uk

Today

1. General concepts in IR evaluation
2. The TREC competitions
3. IR evaluation metrics

Evaluation: difficulties

- IR system
  - in: a query
  - out: relevant documents
- Evaluation of IR systems
- Goal: predict future from past experience
- Reasons why IR evaluation is hard:
  - Large variation in human information needs and queries
  - The precise contributions of each component are hard to entangle:
    - Collection coverage
    - Document indexing
    - Query formulation
    - Matching algorithm

Evaluation: "the laboratory model"

- Test only "system parameters"
  - Index language devices for description and search
  - Methods of term choice for documents
  - Matching algorithm
  - Type of user interface
- Ignore environment variables
  - Properties of documents → use many documents
  - Properties of users → use many queries

What counts as acceptable test data?

- In 60s and 70s, very small test collections, arbitrarily different, one per project
  - in 60s: 35 queries on 82 documents
  - in 1990: still only 35 queries on 2000 documents
- not always kept test and training apart as so many environment factors were tested
- TREC-3: 742,000 documents; TREC Web-track: small snapshot of the web
- Large test collections are needed
  - to capture user variation
  - to support claims of statistical significance in results
  - to demonstrate that systems scale up → commercial credibility
- Practical difficulties in obtaining data; non-balanced nature of the collection

Today's test collections

A test collection consists of:

- Document set:
  - Large, in order to reflect diversity of subject matter, literary style, noise such as spelling errors
- Queries/Topics
  - short description of information need
  - TREC "topics": longer description detailing relevance criteria
  - "frozen" → reusable
- Relevance judgements
  - binary
  - done by same person who created the query

TREC

- Text REtrieval Conference
- Run by NIST (US National Institute of Standards and Technology)
- Marks a new phase in retrieval evaluation
  - common task and data set
  - many participants
  - continuity
- Large test collection: text, queries, relevance judgements
  - Queries devised and judged by information specialist (same person)
  - Relevance judgements done only for up to 1000 documents/query
- 2003 was 12th year
- 87 commercial and research groups participated in 2002

Sample TREC query

<num> Number: 508
<title> hair loss is a symptom of what diseases

<desc> Description:
Find diseases for which hair loss is a symptom.

<narr> Narrative:
A document is relevant if it positively connects the loss of head hair in humans with a specific disease. In this context, ”thinning hair” and ”hair loss” are synonymous. Loss of body and/or facial hair is irrelevant, as is hair loss caused by drug therapy.

TREC Relevance Judgements

Humans decide which document–query pairs are relevant.

Evaluation metrics

	Relevant	Non-relevant	Total
Retrieved	A	B	A+B
Not retrieved	C	D	C+D
Total	A+C	B+D	A+B+C+D

Recall: proportion of retrieved items amongst the relevant items ($\frac{A}{A+C}$)

Precision: proportion of relevant items amongst retrieved items ($\frac{A}{A+B}$)

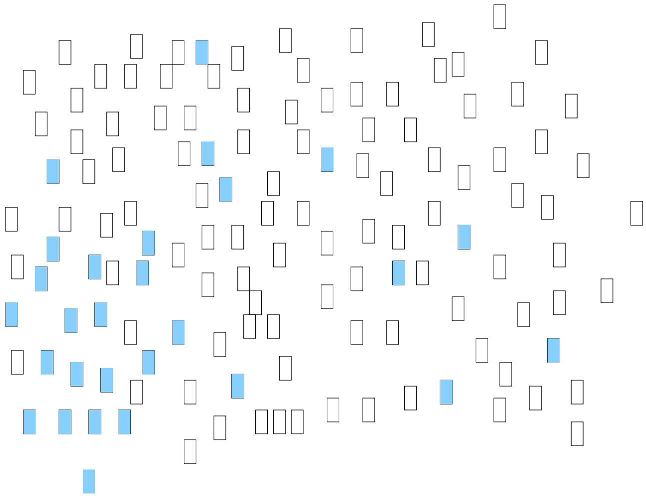
Accuracy: proportion of correctly classified items as relevant/irrelevant ($\frac{A+D}{A+B+C+D}$)

Recall: [0..1]; Precision: [0..1]; Accuracy: [0..1]

Accuracy is not a good measure for IR, as it conflates performance on relevant items (A) with performance on irrelevant (uninteresting) items (D)

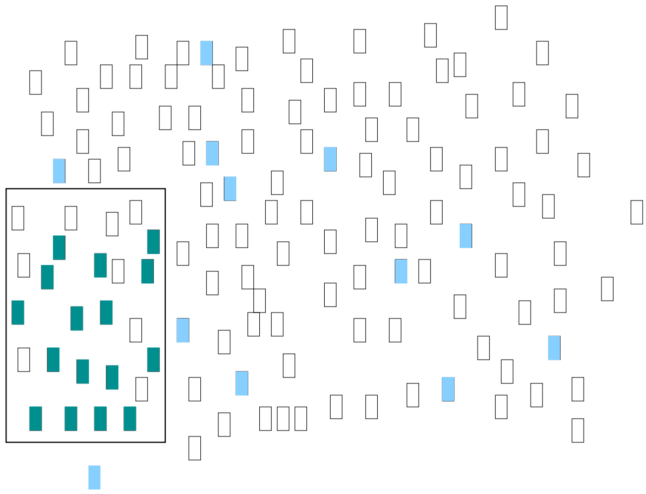
Recall and Precision

- All documents: A+B+C+D = 130
- Relevant documents for a given query: A+C = 28



Recall and Precision: System 1

- System 1 retrieves 25 items: $(A+B)_1 = 25$
- Relevant and retrieved items: $A_1 = 16$

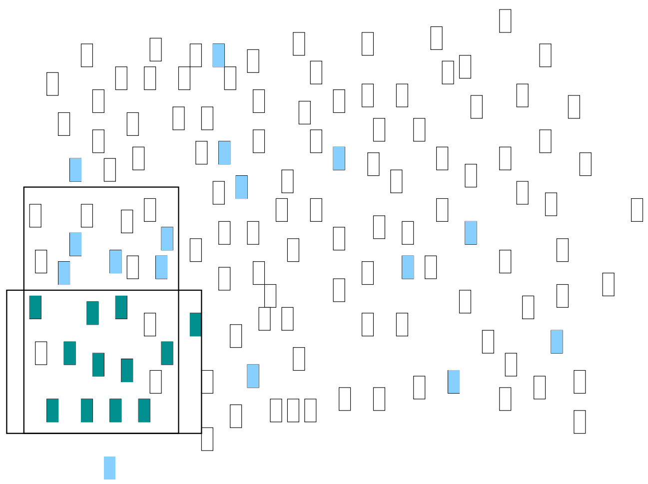


$R_1 = \frac{A_1}{A+C} = \frac{16}{28} = .57$

$P_1 = \frac{A_1}{(A+B)_1} = \frac{16}{25} = .64$

$A_1 = \frac{A_1+D_1}{A+B+C+D} = \frac{16+93}{130} = .84$

Recall and Precision: System 2



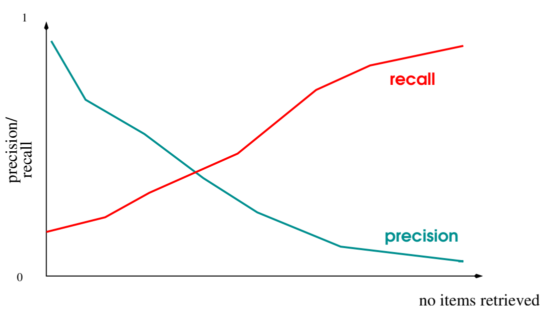
- System B retrieves set $(A+B)_2$ = 15 items
- A_2 = 12

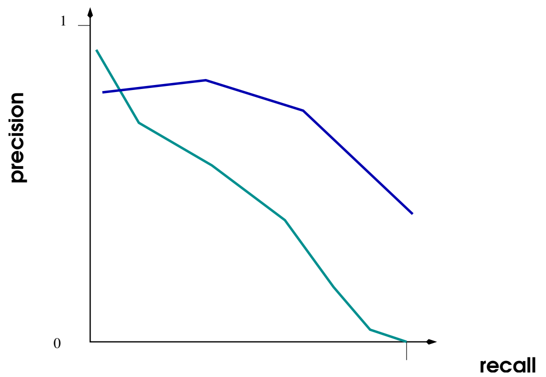
$R_2 = \frac{12}{28} = .43$

$P_2 = \frac{12}{15} = .8$

$A_2 = \frac{12+99}{130} = .85$

Recall-precision curve





- Plotting precision and recall (versus no. of documents retrieved) shows inverse relationship between precison and recall
- Precision/recall cross-over can be used as combined evaluation measure
- Plotting precision versus recall gives recall-precision curve
- Area under normalised recall-precision curve can be used as evaluation measure

Recall-criticality and precision-criticality

- Inverse relationship between precision and recall forces general systems to go for compromise between them
- But some tasks particularly need good precision whereas others need good recall:

Precision-critical task	Recall-critical task
Little time available	Time matters less
A small set of relevant documents answers the information need	One cannot afford to miss a single document
Potentially many documents might fill the information need (redundantly)	Need to see *each* relevant document
Example: web search for factual information	Example: patent search

The problem of determining recall

- Recall problem: for a collection of non-trivial size, it becomes impossible to inspect each document
- It would take 6500 hours to judge 800,000 documents for one query (30 sec/document)
- Pooling addresses this problem

Pooling

Pooling (Sparck Jones and van Rijsbergen, 1975)

- Pool is constructed by putting together top N retrieval results from a set of n systems (TREC: N = 100)
- Humans judge every document in this pool
- Documents outside the pool are automatically considered to be irrelevant
- There is overlap in returned documents: pool is smaller than theoretical maximum of $N \cdot n$ systems (around $\frac{1}{3}$ the maximum size)
- Pooling works best if the approaches used are very different
- Large increase in pool quality by manual runs which are recall-oriented, in order to supplement pools

F-measure

- Weighted harmonic mean of P and R (Rijsbergen 1979)

$$F_\alpha = \frac{PR}{(1-\alpha)P + \alpha R}$$

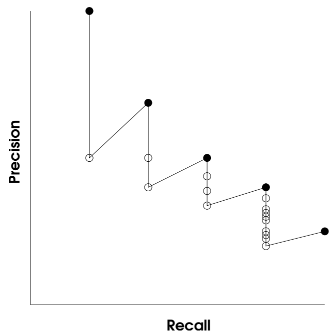
  - High α: Precision is more important
  - Low α: Recall is more important
- Most commonly used with α=0.5 $\rightarrow$

$$F_{0.5} = \frac{2PR}{P+R}$$

- Maximum value of $F_{0.5}$-measure (or F-measure for short) is a good indication of best P/R compromise
- F-measure is an approximation of cross-over point of precision and recall

Precision and recall in ranked IR engines

- With ranked list of return documents there are many P/R data points
- Sensible P/R data points are those after each new relevant document has been seen (black points)



Query 1			
Rank	Relev.	R	P
1	X	0.20	1.00
2		“	0.50
3	X	0.40	0.67
4		”	0.50
5		”	0.40
6	X	0.60	0.50
7		”	0.43
8		”	0.38
9		”	0.33
10	X	0.80	0.40
11		”	0.36
12		”	0.33
13		”	0.31
14		”	0.29
15		”	0.27
16		”	0.25
17		”	0.24
18		”	0.22
19		”	0.21
20	X	1.00	0.25

Summary IR measures

- Precision at a certain rank: P(100)
- Precision at a certain recall value: P(R=.2)
- Precision at last relevant document: P(last_relev)
- Recall at a fixed rank: R(100)
- Recall at a certain precison value: R(P=.1)

Summary IR measures over several queries

- Want to average over queries
- Problem: queries have differing number of relevant documents
- Cannot use one single cut-off level for all queries
  - This would not allow systems to achieve the theoretically possible maximal values in all conditions
  - Example: if a query has 10 relevant documents
    * If cutoff $>$ 10, $P < 1$ for all systems
    * If cutoff $<$ 10, $R < 1$ for all systems
- Therefore, more complicated joint measures are required

11 point average precision

$$P_{11\_pt} = \frac{1}{11} \sum_{j=0}^{10} \frac{1}{N} \sum_{i=1}^{N} \tilde{P}_i(r_j)$$

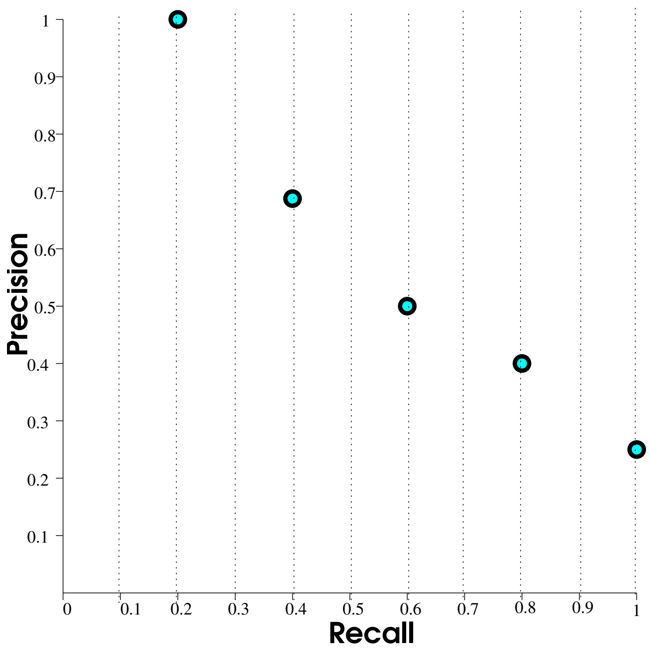
with $\tilde{P}_i(r_j)$ the precision at the jth recall point in the ith query (out of N queries)

- Define 11 standard recall points $r_j = \frac{j}{10}$: $r_0 = 0$, $r_1 = 0.1$ ... $r_{10} = 1$
- We need $\tilde{P}_i(r_j)$; i.e. the precision at our recall points
- $P_i(R = r)$ can be measured: the precision at each point when recall changes (because a new relevant document is retrieved)
- Problem: unless the number of relevant documents per query is divisible by 10, $\tilde{P}_i(r_j)$ does not coincide with a measurable data point r
- Solution: interpolation

$$\tilde{P}_i(r_j) = \begin{cases} max(r_j \leq r < r_{j+1}) P_i(R = r) & \text{if } P_i(R = r) \text{ exists} \\ \tilde{P}_i(r_{j+1}) & \text{otherwise} \end{cases}$$

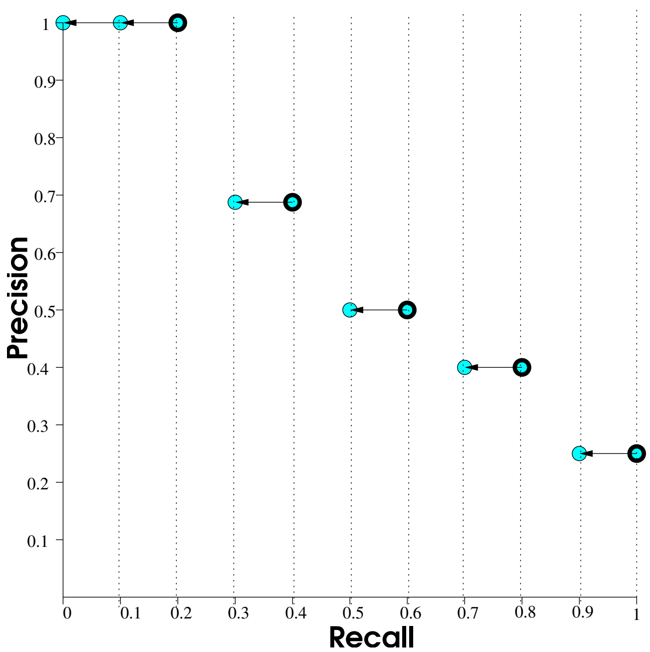
- Note that $P_i(R = 1)$ can always be measured.

11 point average precision: measured data points, Q1



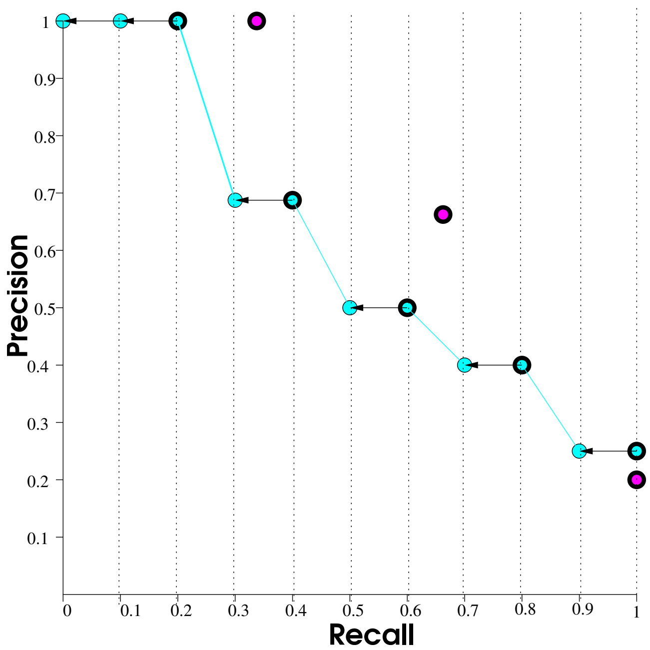
- Blue for Query 1
- Bold Circles measured
- Five r_js coincide with datapoint

11 point average precision: interpolation, Q1



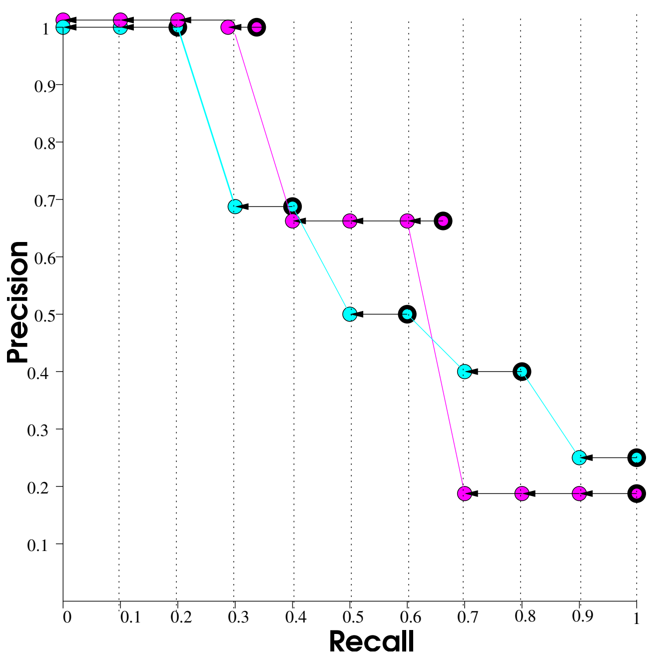
- Blue for Query 1
- Bold Circles measured
- Thin circles interpolated

11 point average precision: measured data points, Q2



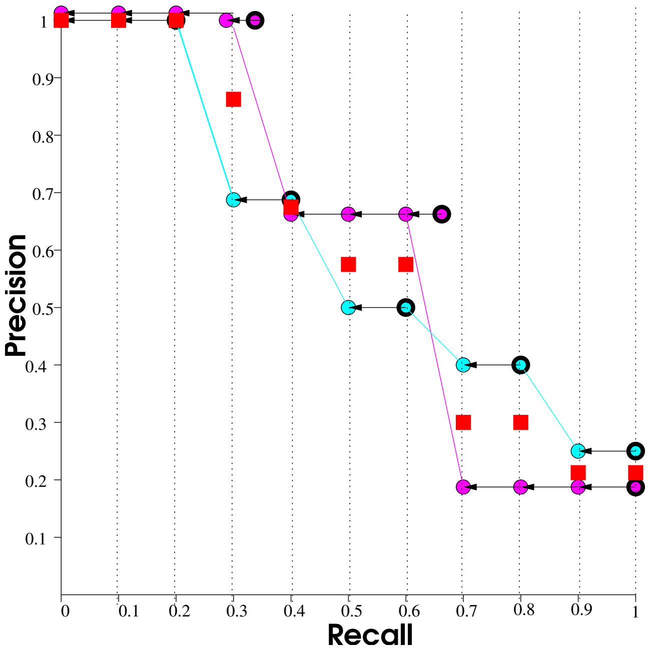
- Red for Query 2
- Bold Circles are measured
- Only $r_1 0$ coincides with a data point

11 point average precision: interpolation, Q2



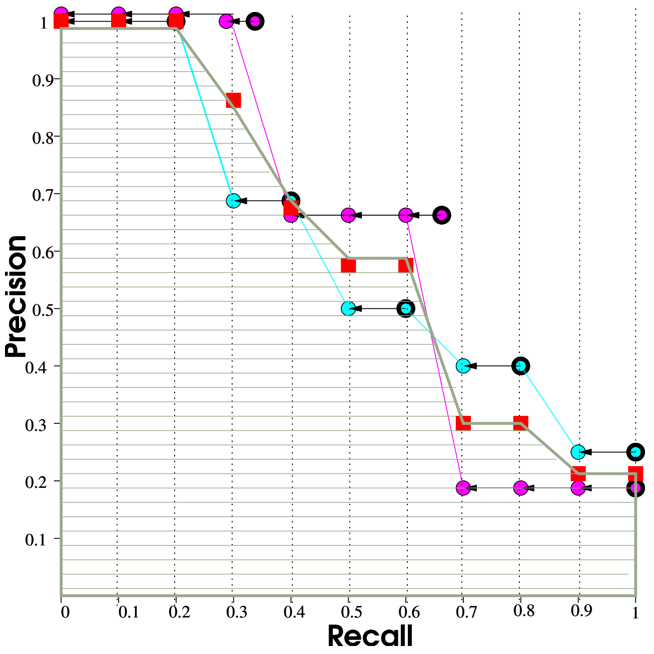
- Red for Query 2
- Bold Circles measured
- Thin circles interpolated

11 point average precision: averaging



- Now average at each p_r
- over N (number of queries)
- $\rightarrow$ 11 data points

11 point average precision: area/result



- End result:
- 11 point average precision
- (representation of area)

11 point average precision: same example as a table

			$P_1(r_i)$		$\Sigma_{j=1}^{N} P_j(r_i)$		$P_2(r_i)$			
Query 1			$\tilde{P}_1(r_0)$ = 1.00	→	1.00	←	$\tilde{P}_2(r_0)$ = 1.00			
#		**R**	$\tilde{P}_1(r_1)$ = 1.00	→	1.00	←	$\tilde{P}_2(r_1)$ = 1.00			
1	X	0.20	$\tilde{P}_1(r_2) = P_1(R=.2)$ = 1.00	→	1.00	←	$\tilde{P}_2(r_2)$ = 1.00	**Query 2**		
2			$\tilde{P}_1(r_3)$ = 0.67	→	0.84	←	$\tilde{P}_2(r_3)$ = 1.00	**R**		**#**
3	X	0.40	$\tilde{P}_1(r_4) = P_1(R=.4)$ = 0.67	→	0.67			0.33	X	1
4						↖	$\tilde{P}_2(r_4)$ = 0.67			2
5			$\tilde{P}_1(r_5)$ = 0.50	→	0.59	←	$\tilde{P}_2(r_5)$ = 0.67	0.67	X	3
6	X	0.60	$\tilde{P}_1(r_6) = P_1(R=.6)$ = 0.50	→	0.59	←	$\tilde{P}_2(r_6)$ = 0.67			4
7										5
8										6
9			$\tilde{P}_1(r_7)$ = 0.40	→	0.30	←	$\tilde{P}_2(r_7)$ = 0.20			7
10	X	0.80	$\tilde{P}_1(r_8) = P_1(R=.8)$ = 0.40	→	0.30	←	$\tilde{P}_2(r_8)$ = 0.20			8
11										9
12										10
13										11
14			$\tilde{P}_1(r_9)$ = 0.25	→	0.23	←	$\tilde{P}_2(r_9)$ = 0.20			12
15										13
16										14
17					0.23	←	$\tilde{P}_2(r_{10}) = P_2(R=1.0)$ = 0.20		X	15
18										
19				↗	↓					
20	X	1.00	$\tilde{P}_1(r_{10}) = P_1(R=1.0)$ = 0.25		P_{11_pt} = 0.61					

$\tilde{P}_i(r_j)$ is (interpolated) precision of ith query, at jth recall point. $P_i(R = r_j)$ (black) are exactly measured precision values.

Mean Average Precision (MAP)

- Also called "average precision at seen relevant documents"
- Determine precision at each point when a new relevant document gets retrieved
- Use P=0 for each relevant document that was not retrieved
- Determine average for each query, then average over queries

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j}P(doc_i)$$

with:

Q_j	number of relevant documents for query j
N	number of queries
$P(doc_i)$	precision at ith relevant document

Query 1		
Rank	Relev.	$P(doc_i)$
1	X	1.00
2		
3	X	0.67
4		
5		
6	X	0.50
7		
8		
9		
10	X	0.40
11		
12		
13		
14		
15		
16		
17		
18		
19		
20	X	0.25
AVG:		0.564

Query 2		
Rank	Relev.	$P(doc_i)$
1	X	1.00
2		
3	X	0.67
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15	X	0.2
AVG:		0.623

- MAP favours systems which return relevant documents fast
- Precision-biased

$$MAP = \frac{0.564+0.623}{2} = 0.594$$

- Relevance is subjective $\rightarrow$ Judgements differ across judges
- Relevance is situational $\rightarrow$ Judgements also differ across time (same judge!)
- Problem: Systems are not comparable if metrics compiled from different judges or at different times will differ
- Countermeasure, Part A: Use guidelines
  - Relevance defined independently of novelty
  - Then, relevance decisions are independent of each other
- Countermeasure, Part B: counteract natural variation by extensive sampling; large populations of users and information needs
- Then: Relative success measurements on systems stable across judges (but not necessarily absolute ones) (Voorhees, 2000)
- Okay if all you want to do is compare systems

TREC: IR system performance, future

- TREC-7 and 8: P(30) between .40 and .45, using long queries and narratives (one team even for short queries); P(10) = .5 even with short queries, $>$ .5 with medium length queries
- Systems must have improved since TREC-4, 5, and 6 $\rightarrow$ manual performance (sanity check) remained on a plateau of around .6
- The best TREC-8 ad-hoc systems not stat. significantly different $\rightarrow$ plateau reached? Ad hoc track discontinued after TREC-8.
- New tasks: filtering, web, QA, genomics, interactive, novelty, robust, video, cross-lingual,. . .
- 2006 is TREC-15. Latest tasks: spam, terabyte, blog, web, legal

TREC tracks 1992-2002

TRACK	TREC													
	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Ad Hoc	18	24	26	23	28	31	42	41						
Routing	16	25	25	15	16	21								
Interactive			3	11	2	9	8	7	6	6	6			
Spanish			4	10	7									
Confusion				4	5									
Database Merging				3	3									
Filtering				4	7	10	12	14	15	19	21			
Chinese					9	12								
NLP					4	2								
Speech						13	10	10	3					
Cross-Language						13	9	13	16	10	9			
High Precision						5	4							
Very Large Corpus							7	6						
Query							2	5	6					
Question Answering								20	28	36	34	33	28	33
Web								17	23	30	23	27	28	
Video										12	19			
Novelty Detection											13	14	14	
Genomic												29	33	41
HARD												14	16	16
Robust												16	14	17
Terabyte													17	23
Enterprise														19
Spam														13
	22	31	33	36	38	51	56	66	68	87	93	93	103	117

Summary

- IR evaluation as currently performed (TREC) only covers one small part of the spectrum:
  - System performance in batch mode
  - Laboratory conditions; not directly involving real users
  - Precision and recall measured from large, fixed test collections
- However, this evaluation methodology is very stable and mature
  - Host of elaborate performance metrics available, e.g. MAP
  - Relevance problem solvable (in principle) by query sampling, guidelines, relative system comparisons
  - Recall problem solvable (in practice) by pooling methods
  - Provable that these methods produce stable evaluation results

Literature

- Teufel (2006, To Appear): Chapter *An Overview of evaluation methods in TREC Ad-hoc Information Retrieval and TREC Question Answering*. In: L. Dybkjaer, H. Hemsen, W. Minker (Eds.) Evaluation of Text and Speech Systems. Springer, Dordrecht, The Netherlands.

  Copyrighted – please go to Student Administration for a Copy