Information Retrieval

Lecture 3: Evaluation methodology

Computer Science Tripos Part II
Lent Term 2004

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 $\rightarrow$ use many documents
  - Properties of users $\rightarrow$ 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
- Large test collections are needed
  - to capture user variation
  - to support claims of statistical significance in results
  - to demonstrate that performance levels and differences hold as document file sizes grow $\rightarrow$ 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' $\rightarrow$ reusable
- Relevance judgements
  - binary
  - done by same person who created the query

Relevance Judgements

- Relevance is inherently subjective, so we need humans to do them
- Problem: relevance is situational
  - Information needs are unique to a particular person at a particular time
  - Judgements will differ across judges and for the same judge at different times

    $\rightarrow$ need extensive sampling to counteract natural variation: large populations of users and information needs
- Guidelines given to assessors, in order to define relevance as a reasonably objective property of the document–query pair
  - not fulfillment of information need, not novel information
  - Relevance is defined to be irrespective of information contained in other documents (redundancy)
- These guidelines ensure that each relevance decision can be taken independently

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
- 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 agreement

- Queries devised and judged by information specialist (same person)
- Relevance judgements done only for up to 1000 documents/query
- Annotators don't agree on relevance judgements
- Nevertheless the relative ordering of systems is stable:

> "The comparative effectiveness of different retrieval methods is stable in the face of changes to the relevance judgements" (Vorhees, 2000)

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 ($\frac{A}{A+C}$)

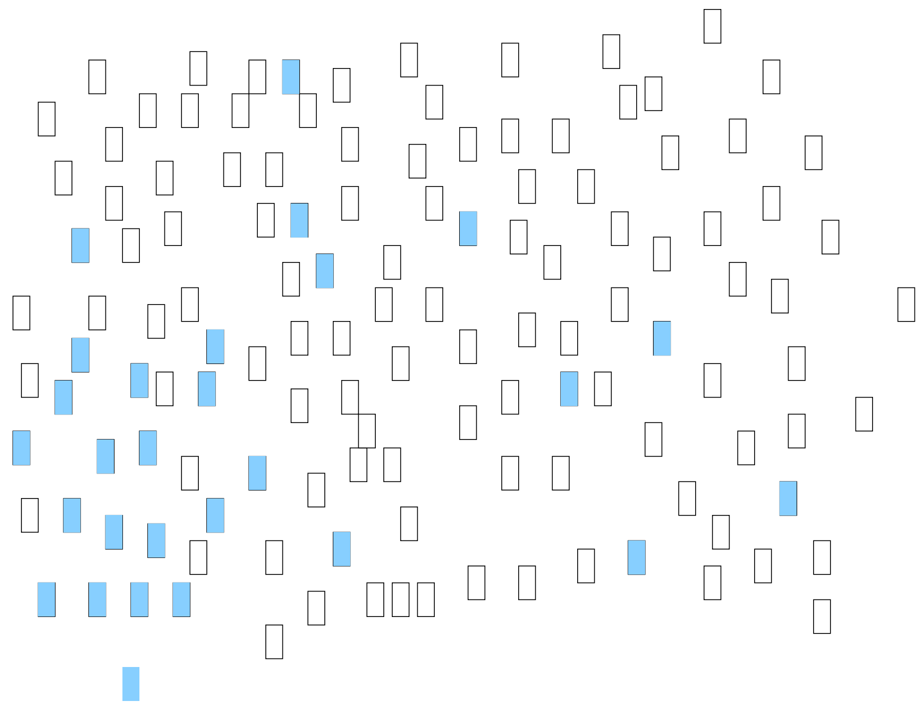
Precision: proportion of relevant items amongst retrieved items ($\frac{A}{A+B}$)

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]

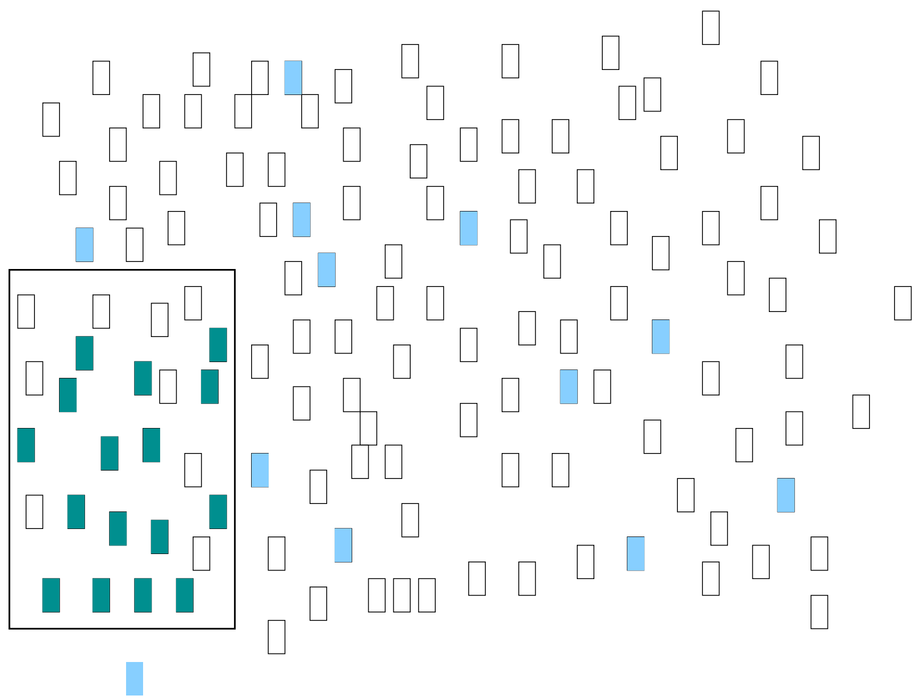
Accuracy is not a good measure for IR, as it conflates performance on relevant items (A) with performance on irrelevant items (D) (which we are not interested in)

- 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: $A_1 = 16$



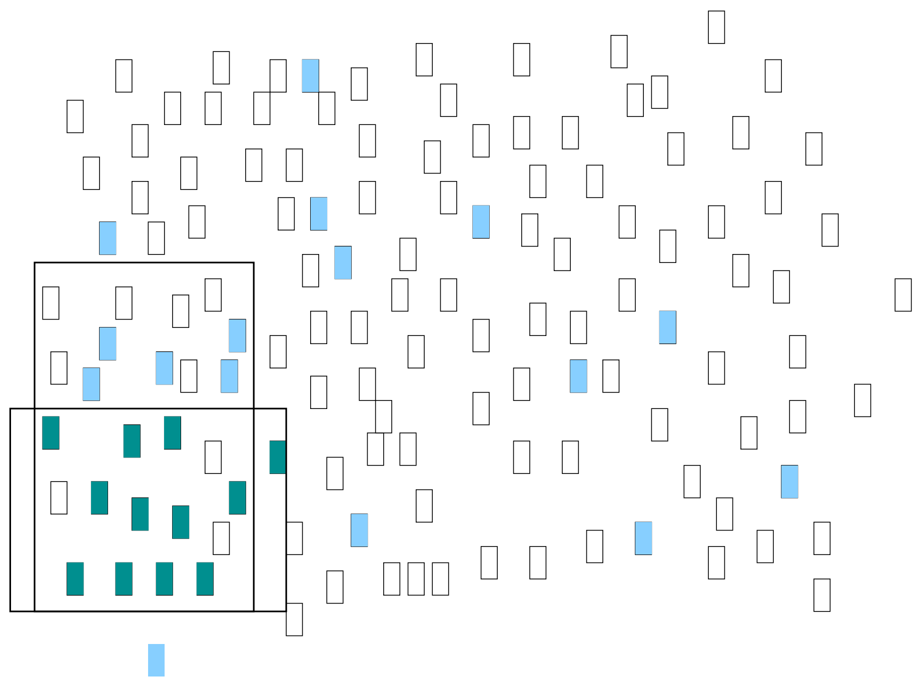
$$R_1 = \frac{A_1}{A+C} = \frac{16}{28} = .57$$

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

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

Recall and Precision: System 2

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

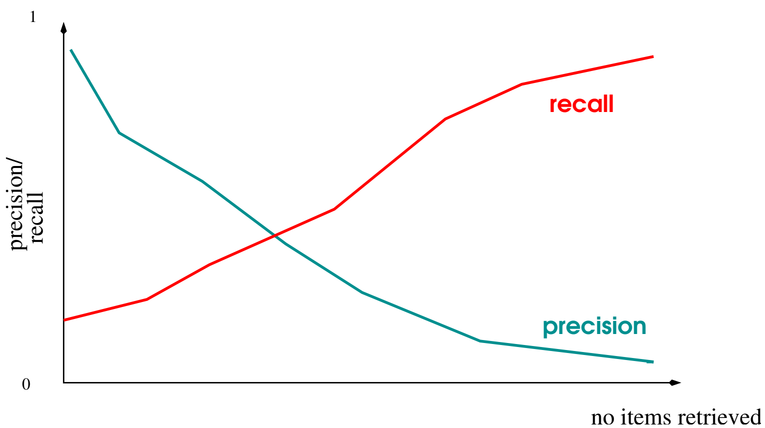


$$R_2 = \frac{12}{28} = .43$$

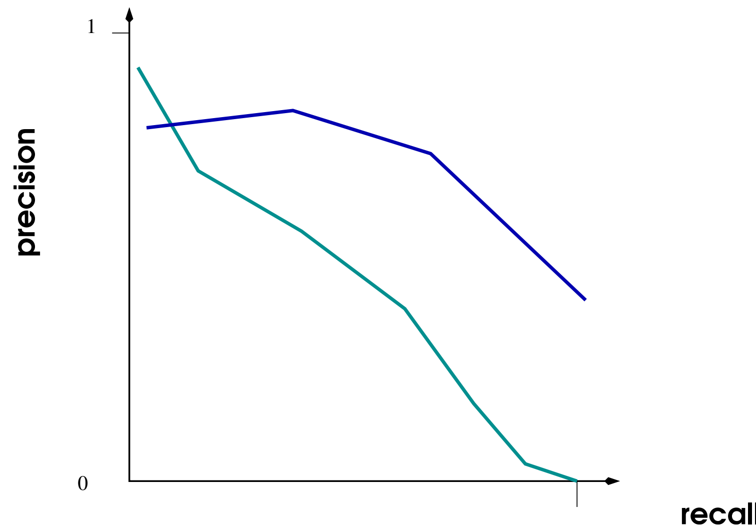
$$P_2 = \frac{12}{15} = .8$$

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

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 conflated 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 (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

- Rijsbergen (1979)

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

  - High α: Recall is more important
  - Low α: Precision is more important

- Most commonly used with α=0.5 $\rightarrow$ Weighted harmonic mean of P and R

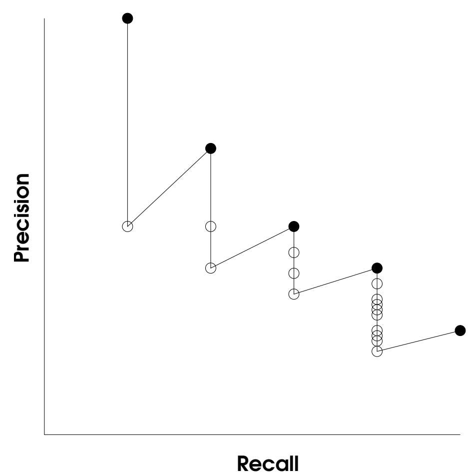
$$F_{0.5} = \frac{2PR}{P+R}$$

- 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

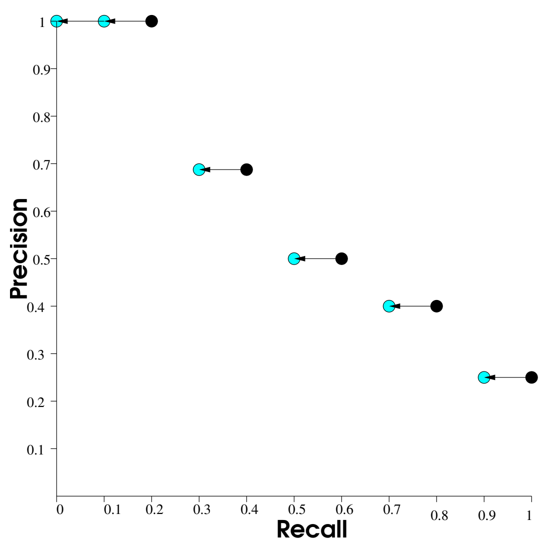
Summary IR measures

- 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)

- 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($R = n$) is precision at that point where recall has first reached n
- Define 11 standard recall points $P(r_0)$, $P(r_1)$, ... $P(r_{10})$
- $P(r_n) = P(R = \frac{n}{10})$
- $P(r_2)$ measures precision at the point where R=0.2
- This might not coincide with a data point, in which case interpolation is necessary:



$$P_{ip}(r_i) = max(r_i < r \leq r_{i+1})P(r)$$

11 standard recall points for our example

			$P_1(r_i)$	**$P_2(r_i)$**			
Query 1			$P_{ip,1}(r_0)$ = 1.00	$P_{ip,2}(r_0)$ = 1.00			
#		**R**	$P_{ip,1}(r_1)$ = 1.00	$P_{ip,2}(r_1)$ = 1.00			
1	X	0.20	$P_1(r_2)$ = 1.00	$P_{ip,2}(r_2)$ = 1.00	**Query 2**		
2			$P_{ip,1}(r_3)$ = 0.67	$P_{ip,2}(r_3)$ = 1.00	**R**		**#**
3	X	0.40	$P_1(r_4)$ = 0.67		0.33	X	1
4				$P_{ip,2}(r_4)$ = 0.67			2
5			$P_{ip,1}(r_5)$ = 0.50	$P_{ip,2}(r_5)$ = 0.67	0.67	X	3
6	X	0.60	$P_1(r_6)$ = 0.50	$P_{ip,2}(r_6)$ = 0.67			4
7							5
8							6
9			$P_{ip,1}(r_7)$ = 0.40	$P_{ip,2}(r_7)$ = 0.20			7
10	X	0.80	$P_1(r_8)$ = 0.40	$P_{ip,2}(r_8)$ = 0.20			8
11							9
12							10
13							11
14			$P_{ip,1}(r_9)$ = 0.25	$P_{ip,2}(r_9)$ = 0.20			12
15							13
16							14
17				$P_2(r_{10})$ = 0.20	1.00	X	15
18							
19							
20	X	1.00	$P_1(r_{10})$ = 0.25				

$P_{ipol}(r_i)$ values (blue) have been interpolated, $P(r_i)$ values(black) have been exactly measured

11 point average precision

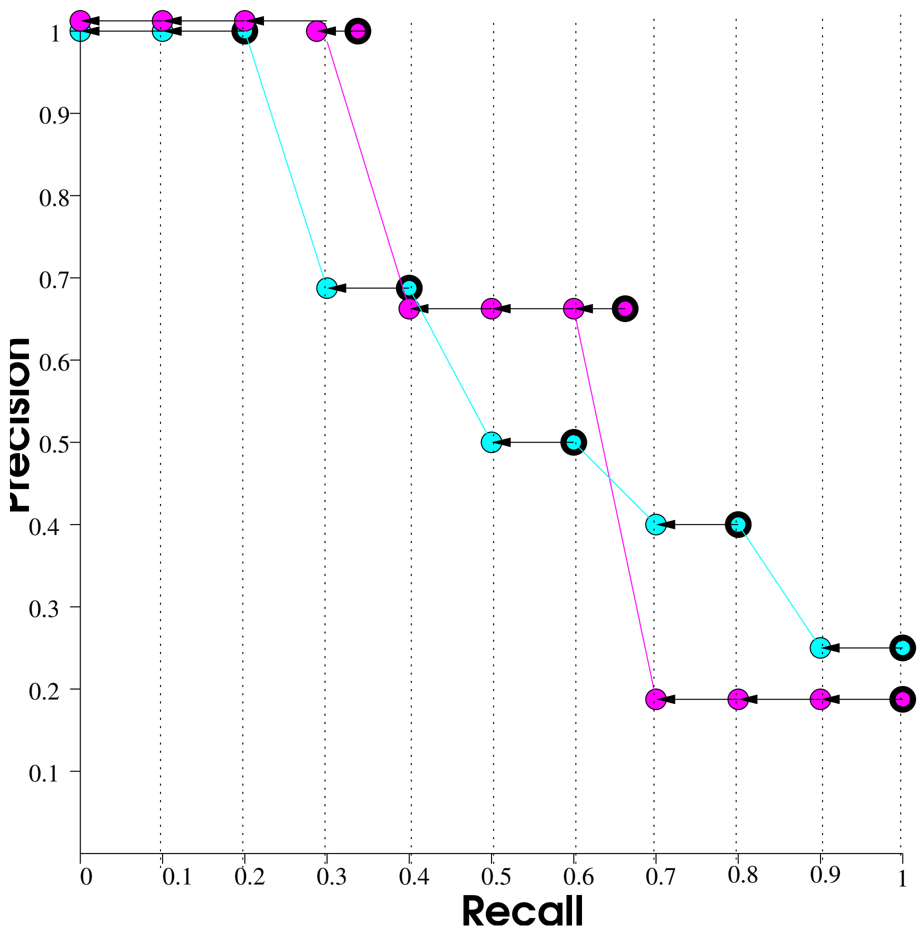
$$P_{11\_pt} = \frac{1}{11} \sum_{j=0}^{10} \frac{1}{N} \sum_{i=1}^{N} P_{ip,i}(r_j)$$

with $P_{ip,i}(r_j)$ the jth interpolated recall point in the ith query (out of N queries)

In our example:

	Query 1	Query 2	Avg. (Queries)
$P_i(r_0)$	1.00	1.00	1.00
$P_i(r_1)$	1.00	1.00	1.00
$P_i(r_2)$	1.00	1.00	1.00
$P_i(r_3)$	0.67	1.00	0.84
$P_i(r_4)$	0.67	0.67	0.67
$P_i(r_5)$	0.50	0.67	0.59
$P_i(r_6)$	0.50	0.67	0.59
$P_i(r_7)$	0.40	0.20	0.30
$P_i(r_8)$	0.40	0.20	0.30
$P_i(r_9)$	0.25	0.20	0.23
$P_i(r_{10})$	0.25	0.20	0.23
			P_{11_pt}:0.61

Graphic representation of example



- Blue for Query 1
- Red for Query 2
- Bold Circles measured
- Thin circles interpolated

Mean precision at seen relevant documents

- Also called “mean average precision”
- 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

$$P_{srd} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(rel = i)$$

with:

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

Mean precision at seen relevant documents: example

Query 1		
Rank	Relev.	P
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
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

- Mean precision at seen relevant documents favours systems which return relevant documents fast
- Precision-biased

$$P_{srd} = \frac{0.564+0.623}{2} = 0.594$$

TREC: IR system performance

- Fully automatic searches in TREC-7 and 8: P(30) between .40 and .45, using long queries and narratives (one team even for short queries) $\rightarrow$ Systems optimised for long queries
- Manual searches: best results between .55 and .60.
- Several systems achieved almost 50% P(10) even with very short queries; several exceed 50% with medium length queries. (Manual searching can lead to 70%)
- TREC-3: best results in .55 to .60 range (but only for long queries)
- TREC-4, 5, and 6: less favourable data conditions (less relevant documents available, less information on topics given) $\rightarrow$ results declined
- Better performance in TREC-7 and 8 must be due to better systems, as the manual performance remained on a plateau
- The best systems are statistically not significantly different $\rightarrow$ plateau reached

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 methodology is very stable and mature
  - Relevance problem solvable (in principle) by extensive sampling
  - Recall problem solvable (in practice) by pooling methods
  - Provable that these methods produce stable evaluation results
  - Host of elaborate performance metrics available
    * 11 point average precision
    * Mean precision at seen relevant documents

Literature

- Teufel (2005, To Appear): Chapter *IR and QA evaluation*. In: Evaluation Methods in Speech and NLP. Kluwer.