COM3110/4115/6115: Text Processing

Information Retrieval: Evaluating IR systems

Mark Stevenson

Department of Computer Science
University of Sheffield

Overview

- Definition of the information retrieval problem
- Approaches to document indexing
  - manual approaches
  - automatic approaches
- Automated retrieval models
  - boolean model
  - ranked retrieval methods (e.g. vector space model)
- Term manipulation:
  - stemming, stopwords, term weighting
- **Evaluation**

Evaluation of IR systems – Why?

- There are various retrieval models/algorithms/IR systems
  - How determine which is the best?
- What is the best component/technique for:
  - Ranking? (cosine, dot-product, . . .)
  - Term selection? (stopword removal, stemming, . . .)
  - Term weighting? (binary, TF, TF.IDF, . . .)
- How far down the ranked list will a user need to look to find some/all relevant items?

Evaluation – Relevance

- Evaluation of effectiveness in relation to the relevance of the documents retrieved
- Relevance is judged in a binary way, even if it is in fact a continuous judgement
  - Impossible when the task is to rank thousands or millions of options: too subjective, too difficult
- Other factors could also be evaluated:
  - User effort/ease of use
  - Response time
  - Form of presentation

Evaluation – Relevance (Benchmarking)

- In IR research/development scenarios, one cannot afford humans looking at results of every system/variant of system
- Instead, performance measured/compared using a pre-created benchmarking corpus, a.k.a. gold-standard dataset, which provides:
  - a standard set of documents, and queries
  - a list of documents judged relevant for each query, by human subjects
  - relevance scores, usually treated as binary
- Example: TREC IR evaluation corpora (`http://trec.nist.gov/`)
  - TREC has run annually since 1991

Evaluation of IR systems – Metrics

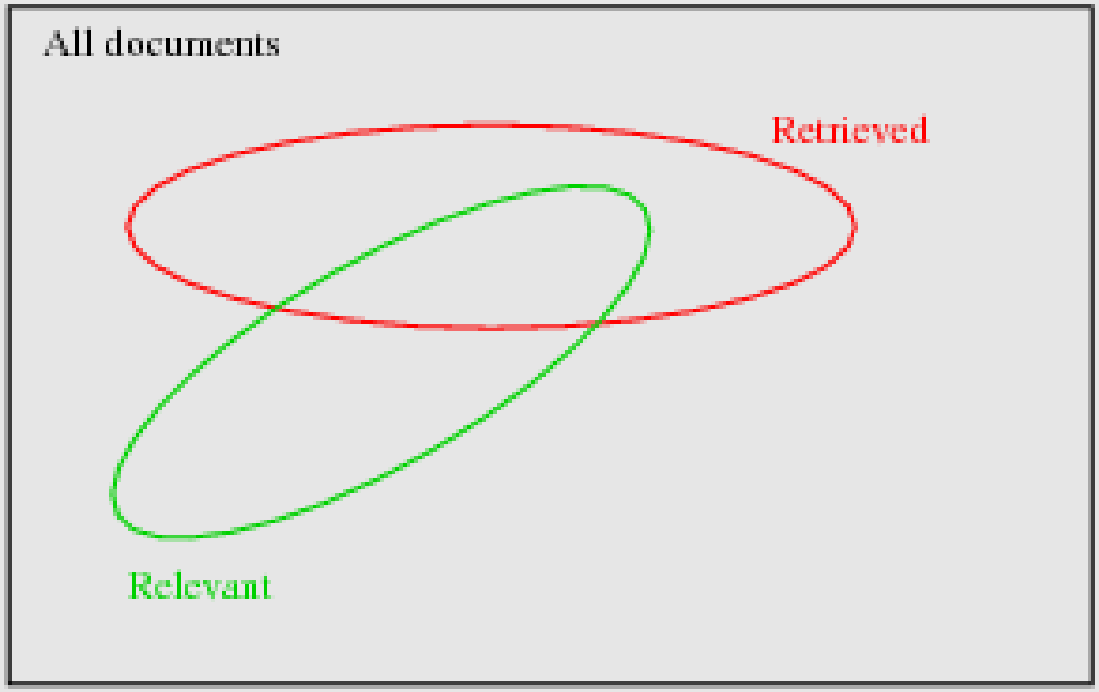
- **AIM**:
  1. get as much good stuff as possible
  2. get as little junk as possible
- The two aspects of this aim are addressed by two separate measures — recall and precision

	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: $\frac{A}{A+C}$ = proportion of relevant documents returned
- Precision: $\frac{A}{A+B}$ = proportion of retrieved documents that are relevant
  - ◇ Both measures have range: $[0 \ldots 1]$

Retrieved vs. Relevant Documents

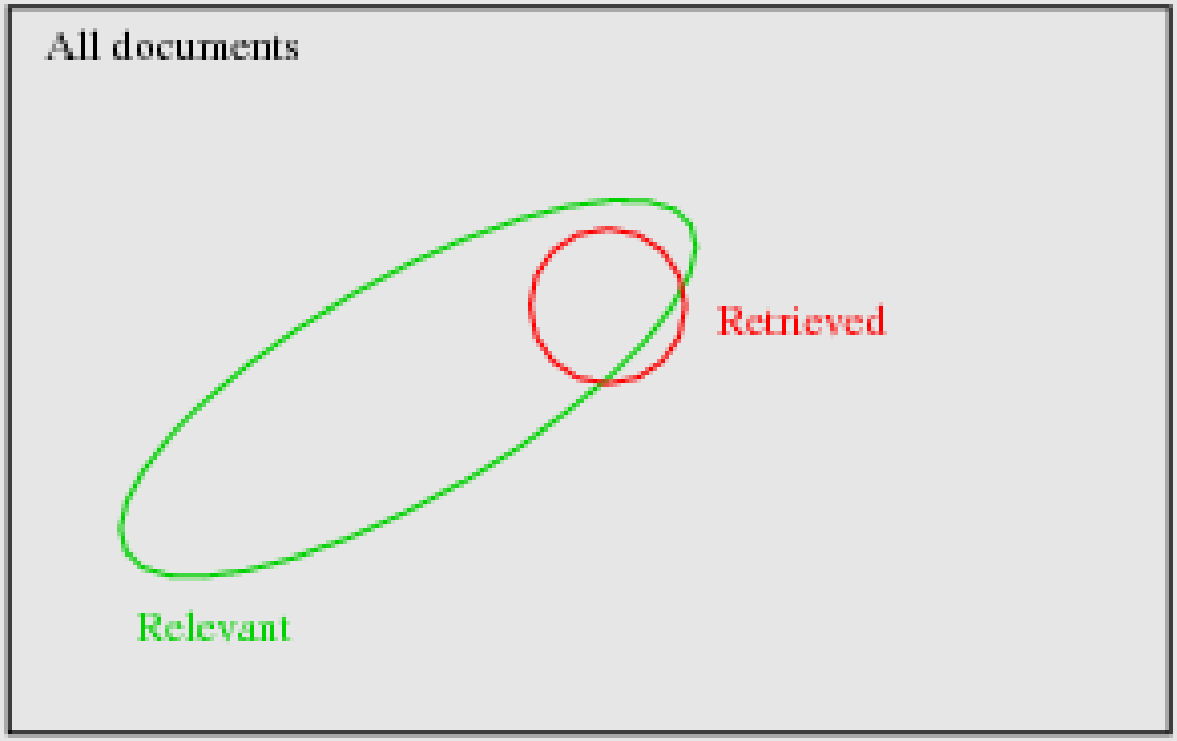
- Precision and Recall address the relation between the *retrieved* and *relevant* sets of documents



- Various situations that arise can be pictorially represented in these terms

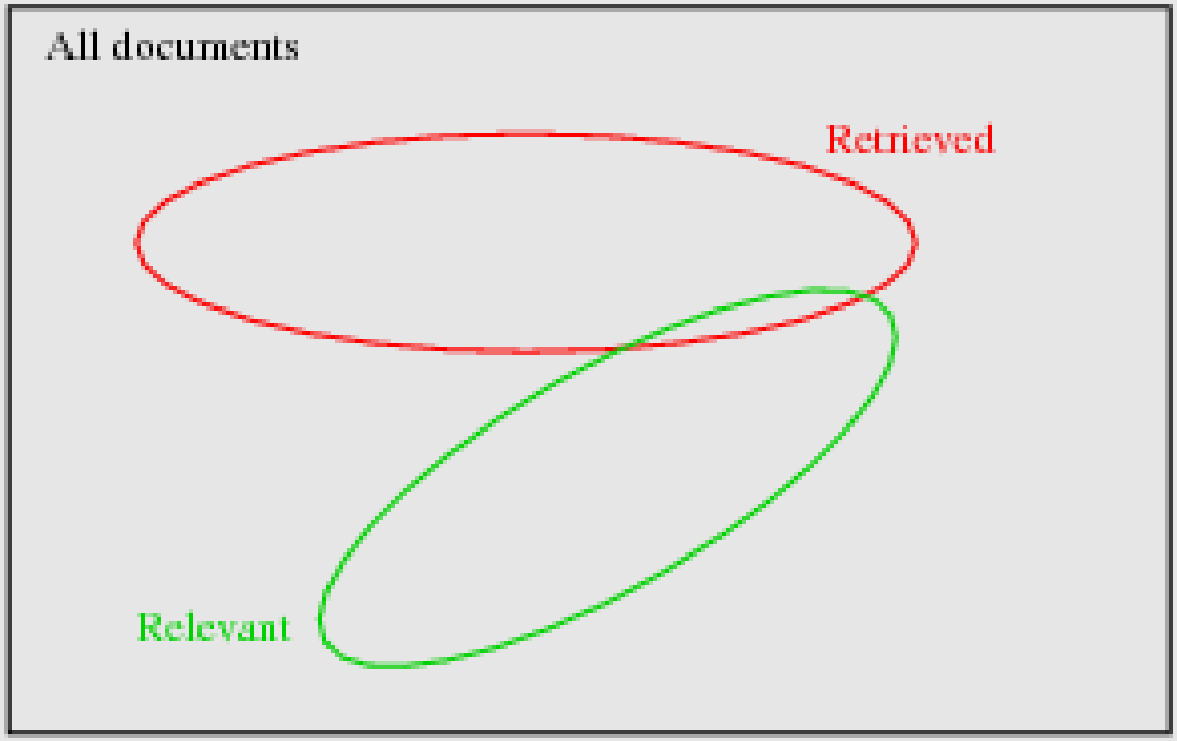
Retrieved vs. Relevant Documents (contd)

- High precision, low recall:



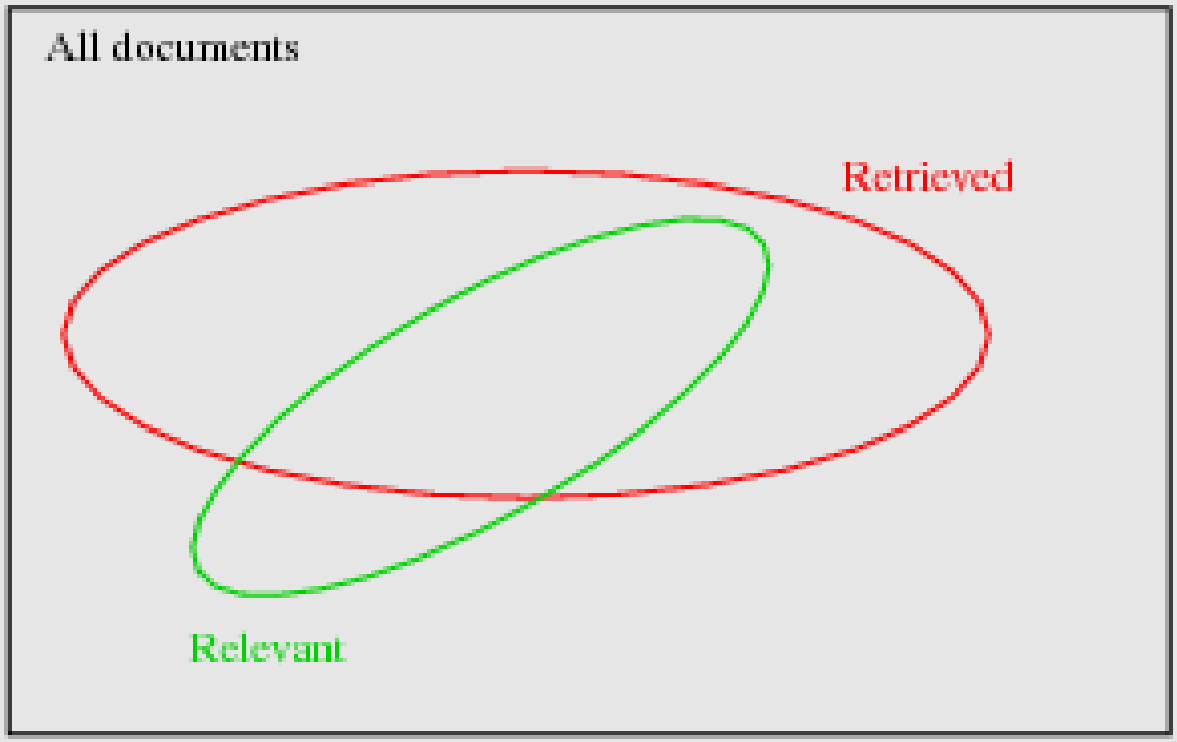
Retrieved vs. Relevant Documents (contd)

- Low precision, low recall:



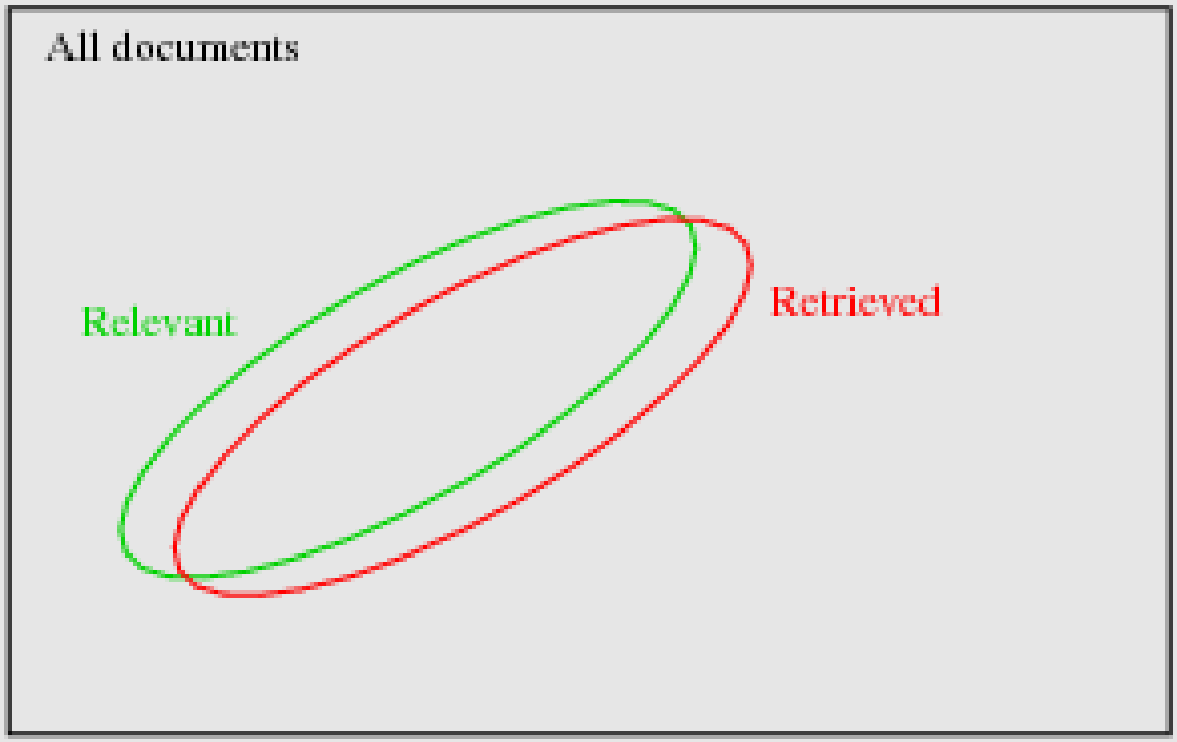
Retrieved vs. Relevant Documents (contd)

- Low precision, high recall:



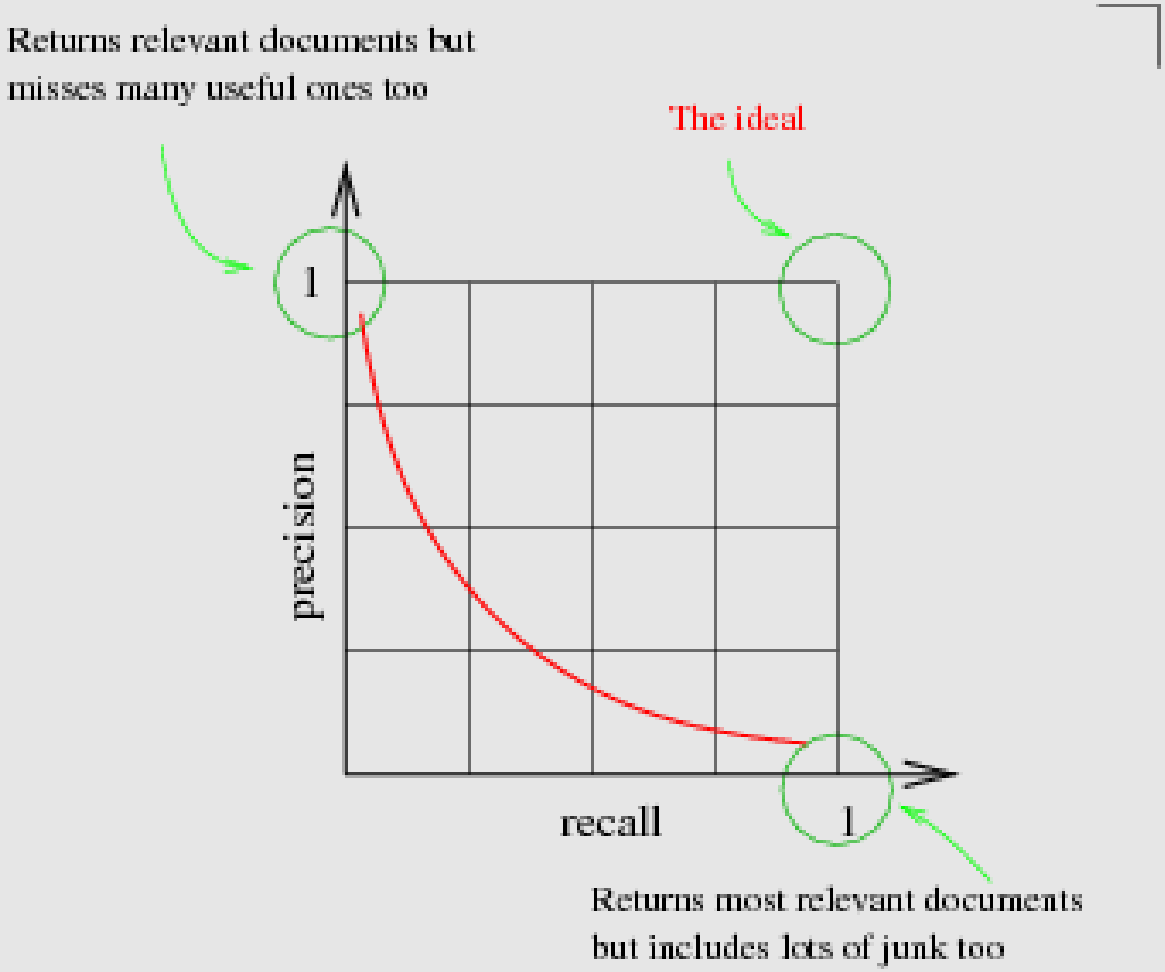
Retrieved vs. Relevant Documents (contd)

- High precision, high recall:

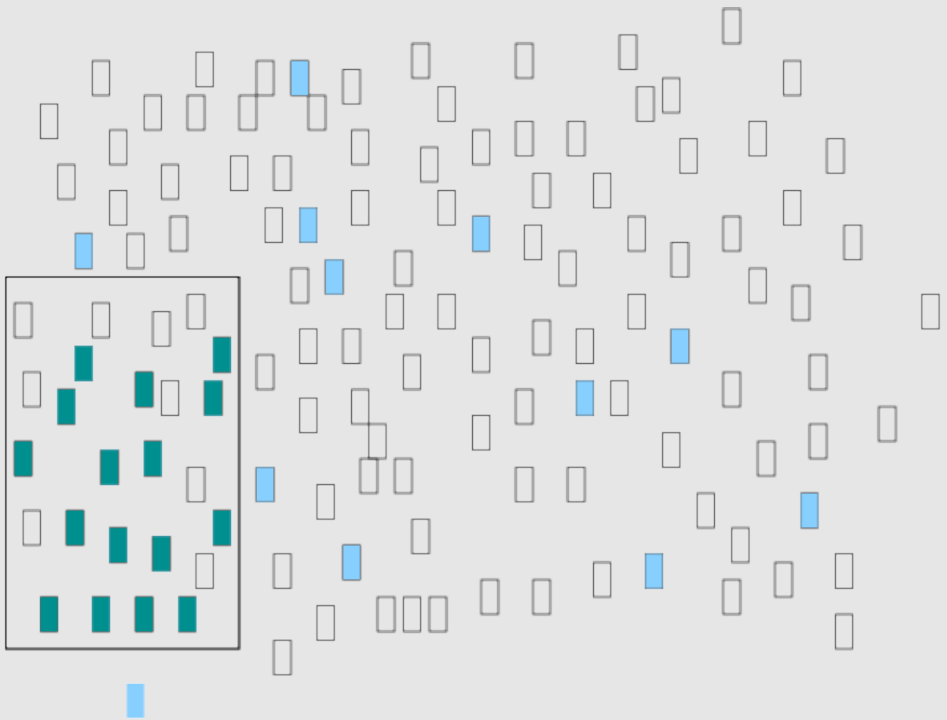


Trade-off Between Recall and Precision

- There is always a trade-off between precision and recall
  - For IR: as more results are considered down the list, precision generally drops, while recall generally increases



Recall and Precision: System 1



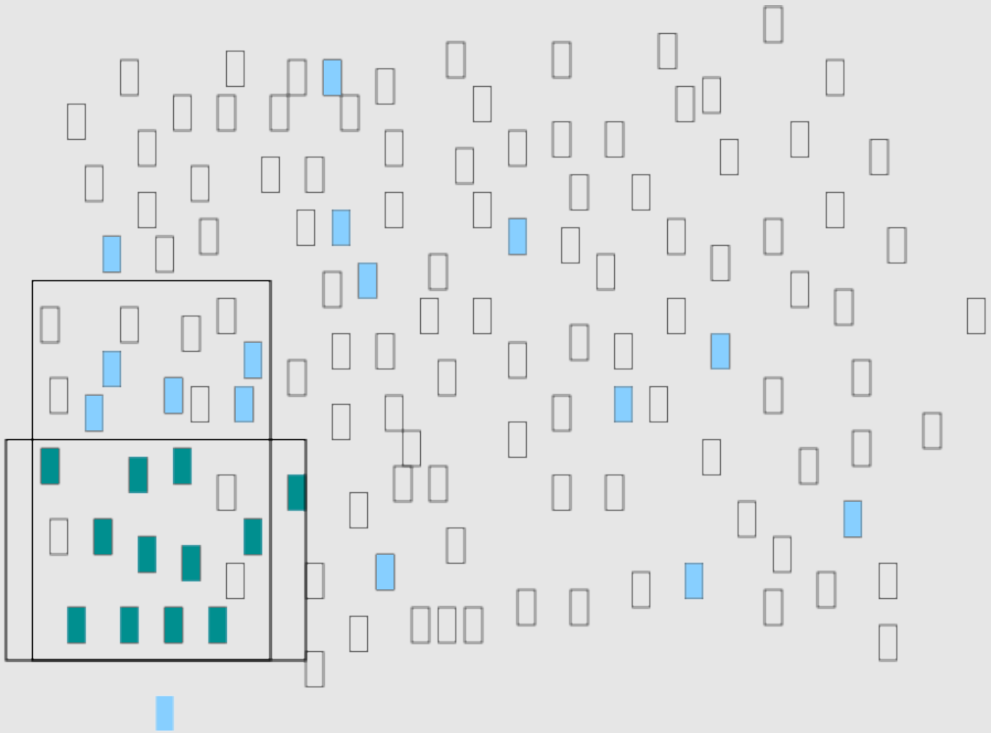
	Rel	Non-rel
Ret	A	B
Not ret	C	D

$$R_1 = \frac{A_1}{A_1+C_1} = \frac{16}{28} = .57$$

$$P_1 = \frac{A_1}{A_1+B_1} = \frac{16}{25} = .64$$

- System 1 retrieves 25 items: $A_1+B_1 = 25$
- Relevant and retrieved items: $A_1 = 16$
- Relevant documents for query: $A_1+C_1 = 28$

Recall and Precision: System 2



	Rel	Non-rel
Ret	A	B
Not ret	C	D

$$R_2 = \frac{A_2}{A_2+C_2} = \frac{12}{28} = .43$$

$$P_2 = \frac{A_2}{A_2+B_2} = \frac{12}{15} = .8$$

- System 2 retrieves 15 items: $A_2+B_2 = 15$
- Relevant and retrieved items: $A_2 = 12$
- Relevant documents for query: $A_2+C_2 = 28$

Recall and Precision: Which is the better system?

$$R_1 = \frac{A_1}{A_1+C_1} = \frac{16}{28} = .57$$
$$P_1 = \frac{A_1}{A_1+B_1} = \frac{16}{25} = .64$$

$$R_2 = \frac{A_2}{A_2+C_2} = \frac{12}{28} = .43$$
$$P_2 = \frac{A_2}{A_2+B_2} = \frac{12}{15} = .8$$

- Which did better: System 1 or System 2?

F-measure

- A way to combine precision and recall into a single figure, giving both equal weight:

$$\frac{2PR}{P+R}$$

- F is a harmonic mean – penalises low performance in one value more than averaging does (behaves differently to *arithmethic* mean):

	values	mean	$F - measure$
e.g.	P=0.5, R=0.5	0.5	0.5
	P=0.1, R=0.9	0.5	0.18

- Previous example
  - System 1 F-measure: 0.603
  - System 2 F-measure: 0.559

Precision at a cutoff

- Measures how well a method ranks relevant documents before non-relevant documents
- E.g. there are **5 relevant documents = d1,d2,d3,d4,d5** – compute precision at top 5

	System 1		System 2		System 3	
	d1:	√	d10:	×	d6:	×
	d2:	√	d9:	×	d1:	√
	d3:	√	d8:	×	d2:	√
	d4:	√	d7:	×	d10:	×
rank 5:	d5:	√	d6:	×	d9:	×
	d6:	×	d1:	√	d3:	√
	d7:	×	d2:	√	d5:	√
	d8:	×	d3:	√	d4:	√
	d9:	×	d4:	√	d7:	×
rank 10:	d10:	×	d5:	√	d8:	×
precision at rank 5:	1.0		0.0		0.4	
precision at rank 10:	0.5		0.5		0.5	

Precision at a cutoff (ctd)

- Note precision at top 5 for System 1: inner order of relevant documents doesn't matter as long as they are all relevant

	System 1		System 2		System 3	
	d5:	√	d10:	×	d6:	×
	d4:	√	d9:	×	d1:	√
	d3:	√	d8:	×	d2:	√
	d1:	√	d7:	×	d10:	×
rank 5:	d2:	√	d6:	×	d9:	×
	d6:	×	d1:	√	d3:	√
	d7:	×	d2:	√	d5:	√
	d8:	×	d3:	√	d4:	√
	d9:	×	d4:	√	d7:	×
rank 10:	d10:	×	d5:	√	d8:	×
precision at rank 5:	1.0		0.0		0.4	
precision at rank 10:	0.5		0.5		0.5	

(Uninterpolated) Average Precision

- Aggregates many precision numbers into one evaluation figure
- Precision computed for each point a relevant document is found, and figures averaged

System 1			System 2			System 3		
d1:	√	(1/1)	d10:	×		d6:	×	
d2:	√	(2/2)	d9:	×		d1:	√	(1/2)
d3:	√	(3/3)	d8:	×		d2:	√	(2/3)
d4:	√	(4/4)	d7:	×		d10:	×	
d5:	√	(5/5)	d6:	×		d9:	×	
d6:	×		d1:	√	(1/6)	d3:	√	(3/6)
d7:	×		d2:	√	(2/7)	d5:	√	(4/7)
d8:	×		d3:	√	(3/8)	d4:	√	(5/8)
d9:	×		d4:	√	(4/9)	d7:	×	
d10:	×		d5:	√	(5/10)	d8:	×	

precision at rank 5:	1.0	0.0	0.4
precision at rank 10:	0.5	0.5	0.5
avg. prec:	1.0	0.354	0.573

Interpolated Average Precision

- Compute an *interpolated precision* score for each of a range of different recall levels
  - usually, for the 11 recall levels: 0%, 10%, 20% ... 90%, 100%
  - if a given recall level is not achieved, its precision is 0
  - these 11 scores are then averaged
  - this score more directly based on recall achieved
- **Steps**:
  - for the given recall level, compute the *highest* precision score observed *after* that recall level is reached
  - based on assumption that typical user willing to look at more docs if would increase the %age of relevant docs amongst those viewed
  - note: typically, precision of docs viewed will *tend* to go down as move down ranking, but sometimes it can go up

Interpolated Average Precision – example

- Compute interpolated precision scores for different recall levels, as given in right-hand table, using data in left-hand table – then average
  - e.g. for recall level 30%: discount data points for recalls below 30% (first two rows left-hand table), highest remaining precision is 0.667
  - e.g. for recall level 50%: discount data points for recalls below 50% (first five rows left-hand table), highest remaining precision is 0.625

Ranking 3		Recall		Precision	
d6:	×	0	(0%)	0	(0)
d1:	√	1/5	(20%)	.5	(1/2)
d2:	√	2/5	(40%)	.667	(2/3)
d10:	×	2/5	(40%)	.5	(2/4)
d9:	×	2/5	(40%)	.4	(2/5)
d3:	√	3/5	(60%)	.5	(3/6)
d5:	√	4/5	(80%)	.57	(4/7)
d4:	√	5/5	(100%)	.625	(5/8)
d7:	×	5/5	(100%)	.55	(5/9)
d8:	×	5/5	(100%)	.5	(5/10)

Rec.Level	Int.Prec
0%	.667
10%	.667
20%	.667
30%	.667
40%	.667
50%	.625
60%	.625
70%	.625
80%	.625
90%	.625
100%	.625
interp. avg. prec:	0.644

Interpolated Average Precision – example #2

- As noted, score for this metric directly affected by recall
- Illustrate this with following example:
  - ◇ similar to last example, except one document (d5) is *not* found

Ranking 4		Recall		Precision	
d6:	×	0	(0%)	0	(0)
d1:	√	1/5	(20%)	.5	(1/2)
d2:	√	2/5	(40%)	.667	(2/3)
d10:	×	2/5	(40%)	.5	(2/4)
d9:	×	2/5	(40%)	.4	(2/5)
d3:	√	3/5	(60%)	.5	(3/6)
d11:	×	3/5	(60%)	.429	(3/7)
d4:	√	4/5	(80%)	.5	(4/8)
d7:	×	4/5	(80%)	.44	(4/9)
d8:	×	4/5	(80%)	.4	(4/10)

Rec.Level	Int.Prec
0%	.667
10%	.667
20%	.667
30%	.667
40%	.667
50%	.5
60%	.5
70%	.5
80%	.5
90%	0
100%	0
interp. avg. prec:	0.439

Summary

- Evaluation of Information Retrieval Systems
  - Comparison against gold standard
- Evaluation measures
  - Precision
  - Recall
  - F-measure
  - Precision at N
  - Average Precision
  - Interpolated Average Precision