Lecture 6: Evaluation

Information Retrieval

Computer Science Tripos Part II

Helen Yannakoudakis[1]

Natural Language and Information Processing (NLIP) Group

UNIVERSITY OF CAMBRIDGE

helen.yannakoudakis@cl.cam.ac.uk

2018

[1] Based on slides from Simone Teufel and Ronan Cummins

Overview

1. Recap/Catchup
2. Introduction
3. Unranked evaluation
4. Ranked evaluation
5. Benchmarks
6. Other types of evaluation

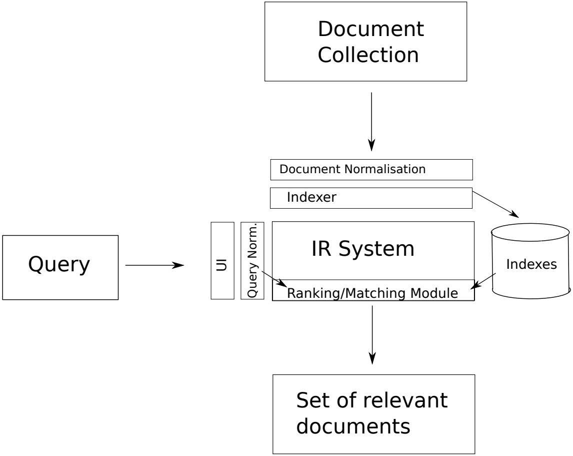
Overview

1. Recap/Catchup
2. Introduction
3. Unranked evaluation
4. Ranked evaluation
5. Benchmarks
6. Other types of evaluation

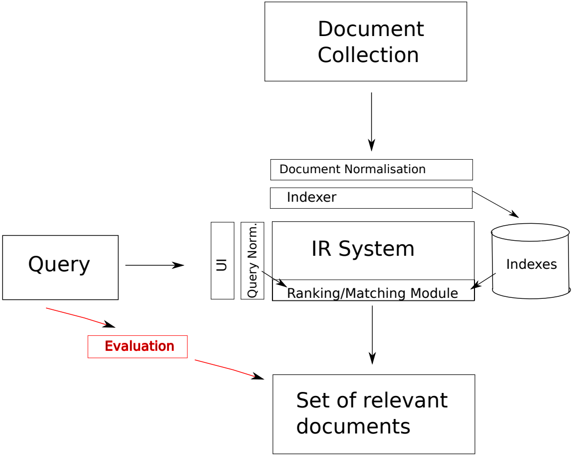
Recap: Ranked retrieval

- In VSM, one represents documents and queries as weighted tf-idf vectors
- Compute the cosine similarity between the vectors to rank
- Language models rank based on the probability of a document model generating the query

Today


Document Collection
Document Normalisation
Indexer
Query
UI
Query Norm.
IR System
Ranking/Matching Module
Indexes
Set of relevant documents

Today



Today: how good are the returned documents?

Overview

Measures for a search engine

- How fast does it index?
  - e.g., number of bytes per hour
- How fast does it search?
  - e.g., latency as a function of queries per second
- What is the cost per query?
  - in dollars
- All of the preceding criteria are measurable: we can quantify speed / size / money

Measures for a search engine

- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
  - Speed of response
  - Size of index
  - Uncluttered UI
  - We can measure:
    - Rate of return to this search engine
    - Whether something was bought
    - Whether ads were clicked
  - Most important: relevance
    (actually, maybe even more important: it's free)
- User happiness is equated with the relevance of search results to the query.
- Note that none of the other measures is sufficient: blindingly fast, but useless answers won't make a user happy.

Most common definition of user happiness: Relevance

- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements:
  1. A benchmark document collection
  2. A benchmark suite of queries
  3. A set of relevance judgments for each query–document pair (gold standard or ground truth judgement of relevance)
     - We need to hire/pay "judges" or assessors to do this.

Relevance: query vs. information need

- Relevance to what? The query?

Information need

"I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine."

- translated into:

Query q

[red wine white wine heart attack]

- So what about the following document:

Document d'

At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.

- d' is an excellent match for query q . . .
- d' is not relevant to the information need.

Relevance: query vs. information need

- User happiness can only be measured by relevance to an information need, not by relevance to queries.
- Sloppy terminology here and elsewhere in the literature: we talk about query–document relevance judgments even though we mean information-need–document relevance judgments.

Overview

Precision and recall

- Precision (P) is the fraction of retrieved documents that are relevant:

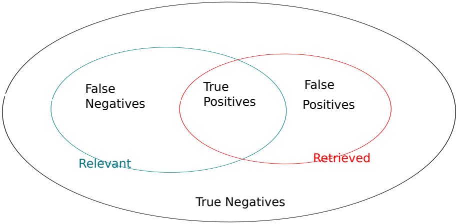
$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- Recall (R) is the fraction of relevant documents that are retrieved:

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

Precision and recall: 2×2 contingency table

WHAT THE SYSTEM THINKS		THE TRUTH: Relevant	THE TRUTH: Non relevant
	Retrieved	true positives (TP)	false positives (FP)
	Not retrieved	false negatives (FN)	true negatives (TN)



$$P = TP/(TP + FP)$$
$$R = TP/(TP + FN)$$

Precision/recall trade-off

- Recall is a non-decreasing function of the number of docs retrieved.
- You can increase recall by returning more docs.
- A system that returns all docs has 100% recall! (but very low precision)
- The converse is also true (usually): It's easy to get high precision for very low recall.

A combined measure: F measure

- F measure: single measure that allows us to trade off precision against recall (weighted harmonic mean):

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

- $\alpha \in [0,1]$ and thus $\beta^2 \in [0,\infty]$
- Most frequently used: balanced F_1 with $\beta = 1$ (or $\alpha = 0.5$):
  - This is the harmonic mean of P and R: $F_1 = \frac{2PR}{P+R}$
- Using β, you can control whether you want to pay more attention to P or R.
- Why don't we use the arithmetic mean?

Example for precision, recall, F_1

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

- $P = \frac{TP}{(TP+FP)} = \frac{20}{(20+40)} = \frac{1}{3}$
- $R = \frac{TP}{(TP+FN)} = \frac{20}{(20+60)} = \frac{1}{4}$
- $F_1 = \frac{2 \times \frac{1}{3} \times \frac{1}{4}}{\frac{1}{3} + \frac{1}{4}} = 2/7$

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
Time	matters	matters less
Tolerance to cases of overlooked information	a lot	none
Information Redundancy	There may be many equally good answers	Information is typically found in only one document
Examples	web search	legal search, patent search

Difficulties in using precision, recall and F

- We need relevance judgments for information-need–document pairs – but they are expensive to produce.
- We should always average over a large set of queries.
  - There is no such thing as a "typical" or "representative" query.
- For alternatives to using precision/recall and having to produce relevance judgments – see end of this lecture.

Why not accuracy?

- Why do we use complex measures like precision, recall, and F?
- Why not something simple like accuracy?
- Accuracy is the fraction of decisions (relevant/non-relevant) that are correct.
- In terms of the contingency table above:
  accuracy $= \frac{(TP+TN)}{(TP+FP+FN+TN)}$
- Limit case:

	relevant	not relevant
retrieved	0	0
not retrieved	10	90

- High accuracy, but the system hasn't returned anything!
- Not suitable when the data is extremely skewed.

Why not accuracy?

- In IR, normally over 99.9% of the documents are in the non-relevant category.
- You then get 99.9% accuracy on most queries by simply saying that all documents are not relevant.
- Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk.
- It's better to return some bad hits as long as you return something.
- $\rightarrow$ We use precision, recall, and F for evaluation, not accuracy.

Overview

Moving from unranked to ranked evaluation

- Precision/recall/F are measures for unranked sets.
- We can easily turn set measures into measures of ranked lists.
- Just compute the set measure for each "prefix": the top 1, top 2, top 3, top 4 etc. results.
- This is called Precision/Recall @ Rank.
- Rank statistics give some indication of how quickly the user will find relevant documents from a ranked list.

Precision/Recall @ Rank

Rank n	Doc
1	d_{12}
2	d_{123}
3	d_{4}
4	d_{57}
5	d_{157}
6	d_{222}
7	d_{24}
8	d_{26}
9	d_{77}
10	d_{90}

- Blue documents are relevant.
- P@n: P@3=0.33, P@5=0.2, P@8=0.25
- R@n: R@3=0.33, R@5=0.33, R@8=0.66

Another idea: Precision @ Recall r

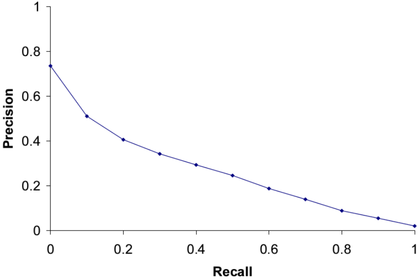
Rank	S1	S2
1	X	
2		X
3	X	
4		
5		X
6	X	X
7		X
8		X
9	X	
10	X	

$\rightarrow$

	S1	S2
P@r 0.2	1.0	0.5
P@r 0.4	0.67	0.4
P@r 0.6	0.5	0.5
P@r 0.8	0.44	0.57
P@r 1.0	0.5	0.63

X denotes the relevant documents.

11-point Interpolated Average Precision



- Compute (interpolated) precision at recall levels / recall points 0.0, 0.1, 0.2, 0.3, . . . 1.0
- Do this for each of the queries in the evaluation benchmark.
- For each recall level, average over queries.
- Figure: example graph of such results from a representative good system at TREC (more later).

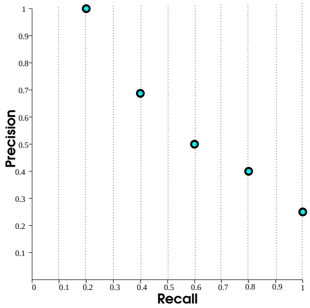
11-point Interpolated Average Precision more formally

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

where $\tilde{P}_i(r_j)$ is the precision at the jth recall level for the ith query (out of N)

- Define 11 standard recall points $r_j = \frac{j}{10}$: $r_0 = 0$, $r_1 = 0.1$ ... $r_{10} = 1$
- To get $\tilde{P}_i(r_j)$, we can use $P_i(R = r_j)$ – but what if there is no point with r_j recall (i.e., there is no relevant document at exacty r_j)?

Worked Example avg-11-pt prec: Query 1, measured data points



- Blue for Query 1
- Bold Circles measured

Query 1				
Rank		R	P	
1	X	0.2	1.00	$\tilde{P}_1(r_2) = 1.00$
2				
3	X	0.4	0.67	$\tilde{P}_1(r_4) = 0.67$
4				
5				
6	X	0.6	0.50	$\tilde{P}_1(r_6) = 0.50$
7				
8				
9				
10	X	0.8	0.40	$\tilde{P}_1(r_8)= 0.40$
11				
12				
13				
14				
15				
16				
17				
18				
19				
20	X	1.0	0.25	$\tilde{P}_1(r_{10}) = 0.25$

- Five r_js $(r_2, r_4, r_6, r_8, r_{10})$ coincide directly with datapoint

11-point Interpolated Average Precision more formally

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

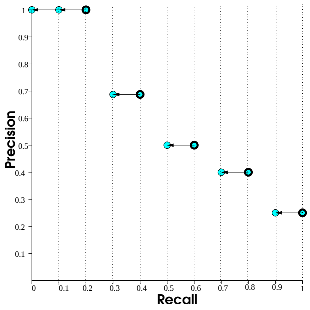
where $\tilde{P}_i(r_j)$ is the precision at the jth recall level for the ith query (out of N)

- Define 11 standard recall points $r_j = \frac{j}{10}$: $r_0 = 0$, $r_1 = 0.1$ ... $r_{10} = 1$
- To get $\tilde{P}_i(r_j)$, we can use $P_i(R = r_j)$ – but what if there is no datapoint with r_j recall (i.e., there is no relevant document at exacty r_j)?
- Interpolated precision: the highest precision found for any recall level $r' \geq r_j$:

$$\tilde{P}_i(r_j) = \max_{r' \geq r_j} P_i(r')$$

  Now we have a value for every recall level.
- Note that $P_i(R = 1)$ can always be measured.

Worked Example avg-11-pt prec: Query 1, interpolation



- Bold circles measured
- **thin circles interpolated**

Query 1			
Rank		R	P
1	X	.20	1.00
2			
3	X	.40	.67
4			
5			
6	X	.60	.50
7			
8			
9			
10	X	.80	.40
11			
12			
13			
14			
15			
16			
17			
18			
19			
20	X	1.00	.25

$\tilde{P}_1(r_0) = 1.00$

$\tilde{P}_1(r_1) = 1.00$

$\tilde{P}_1(r_2) = 1.00$

$\tilde{P}_1(r_3) = .67$

$\tilde{P}_1(r_4) = .67$

$\tilde{P}_1(r_5) = .50$

$\tilde{P}_1(r_6) = .50$

$\tilde{P}_1(r_7) = .40$

$\tilde{P}_1(r_8) = .40$

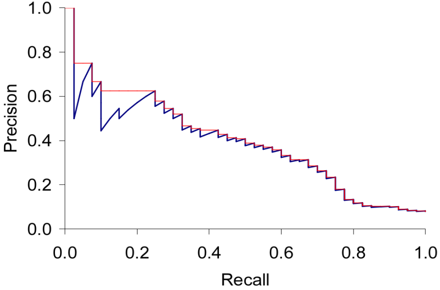
$\tilde{P}_1(r_9) = .25$

$\tilde{P}_1(r_{10}) = .25$

- The six other r_js (r_0, r_1, r_3, r_5, r_7, r_9) are interpolated.

(Worked avg-11-pt prec example for supervisions at the end of slides.)

Another example



- Each point corresponds to a result for the top k ranked hits ($k = 1, 2, 3, 4, \ldots$)
- Interpolation (in red): Take maximum of all future points
- Rationale for interpolation: The user is willing to look at a few more documents if that would increase both precision and recall.

Mean Average Precision (MAP)

- Also called "average precision at seen relevant documents"
- Determine precision at each point when a new relevant document gets retrieved
- Calculate average precision for each query, then average over queries:

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

where:

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

- Use P=0 for each relevant document that was not retrieved

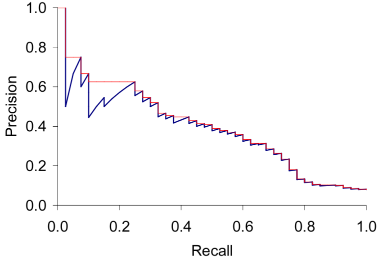
Mean Average Precision: example ($MAP = \frac{0.564+0.623}{2} = 0.594$)

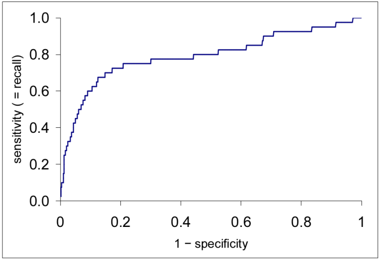
Query 1		
Rank		$P(doc_i)$
1	X	1.00
2		
3	X	0.67
4		
5		
6	X	0.50
7		
8		
9		
10	X	0.40
11		
12		
13		
14		
15		
16		
17		
18		
19		
20	X	0.25
AVG:		0.564

Query 2		
Rank		$P(doc_i)$
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

No need for fixed recall levels, and no interpolation.

ROC curve (Receiver Operating Characteristic)





- y-axis: TPR (true positive rate): TP/total actual positives (also called sensitivity ≡ recall)
- x-axis: FPR (false positive rate): FP/total actual negatives;
    - FPR = fall-out = 1 - specificity (TNR; true negative rate)
- But we are only interested in the small area in the lower left corner (blown up by prec–recall graph)
- For a good system, the graph climbs steeply on the left side

Variance of measures like precision/recall

- For a test collection, it is usual that a system does badly on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and really well on others (e.g., $P = 0.95$ at $R = 0.1$).
- Indeed, it is usually the case that the variance of the same system across queries is much greater than the variance of different systems on the same query.
- That is, there are easy information needs and hard ones.

Overview

What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs, expressible as queries
  - ... which we will often incorrectly refer to as queries
  - Information needs must be representative of the information needs we expect to see in reality.
- Human relevance assessments (relevance assessed relative to the information need)
  - We need to hire/pay "judges" or assessors to do this.
  - Expensive, time-consuming
  - Judges must be representative of the users we expect to see in reality.

First standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK
- 1,398 abstracts of aerodynamics journal articles, a set of 225 queries, exhaustive relevance judgments of all query–document-pairs
- Too small, too untypical for serious IR evaluation today

Second-generation relevance benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST)
- TREC is actually a set of several different relevance benchmarks.
- Best known: TREC Ad Hoc, used for first 8 TREC evaluations between 1992 and 1999
- 1.89 million documents, mainly newswire articles, 450 information needs
- No exhaustive relevance judgments – too expensive
- Rather, NIST assessors' relevance judgments are available only for the documents that were among the top k returned for some system which was entered in the TREC evaluation for which the information need was developed.

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.

Example of more recent benchmark: ClueWeb09

- 1 billion web pages
- 25 terabytes (compressed: 5 terabyte)
- Collected January/February 2009
- 10 languages
- Unique URLs: 4,780,950,903 (325 GB uncompressed, 105 GB compressed)
- Total Outlinks: 7,944,351,835 (71 GB uncompressed, 24 GB compressed)

Inter-judge agreement at TREC

information need	number of docs judged	disagreements
51	211	6
62	400	157
67	400	68
95	400	110
127	400	106

Impact of inter-judge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless? No.
- Large impact on absolute performance numbers
- Virtually no impact on ranking of systems
- Suppose we want to know if algorithm A is better than algorithm B
- An information retrieval experiment will give us a reliable answer to this question . . .
- . . . even if there is a lot of disagreement between judges.

Overview

Evaluation at large search engines

- Recall is difficult to measure on the web
- Search engines often use precision at top k, e.g., $k = 10$ ...
- ... or use measures that reward a system more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures:
  - Clickthrough on first result (frequency with which people click on the top result)
  - Not very reliable if you look at a single clickthrough (you may realize after clicking that the summary was misleading and the document is non-relevant) ...
  - ... but pretty reliable in the aggregate.
  - A/B testing

A/B testing

- Purpose: Test a single innovation
- Pre-requisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an "automatic" measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most

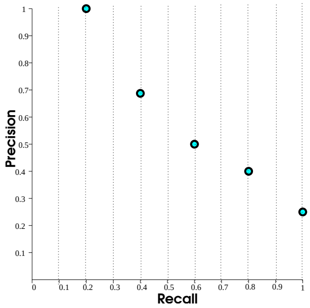
Take-away

- Focused on evaluation for ad-hoc retrieval
  - Precision, Recall, F-measure
  - More complex measures for ranked retrieval
  - Other issues arise when evaluating different tracks, e.g. Question Answering (QA), although typically still use P/R-based measures
- Evaluation for interactive tasks is more involved
- Significance testing is an issue
  - Could a good result have occurred by chance?
  - is the result robust across different document sets?
  - slowly becoming more common
  - Underlying population distributions unknown, so apply non-parametric tests such as the sign test

Reading

- MRS, Chapter 8

Worked Example avg-11-pt prec: Query 1, measured data points

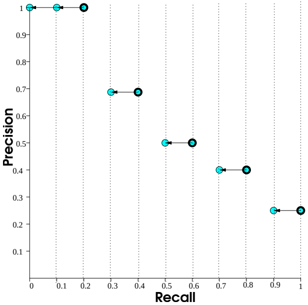


- Blue for Query 1
- Bold Circles measured

Query 1				
Rank		R	P	
1	X	0.2	1.00	$\tilde{P}_1(r_2) = 1.00$
2				
3	X	0.4	0.67	$\tilde{P}_1(r_4) = 0.67$
4				
5				
6	X	0.6	0.50	$\tilde{P}_1(r_6) = 0.50$
7				
8				
9				
10	X	0.8	0.40	$\tilde{P}_1(r_8) = 0.40$
11				
12				
13				
14				
15				
16				
17				
18				
19				
20	X	1.0	0.25	$\tilde{P}_1(r_{10}) = 0.25$

- Five r_js $(r_2, r_4, r_6, r_8, r_{10})$ coincide directly with datapoint

Worked Example avg-11-pt prec: Query 1, interpolation

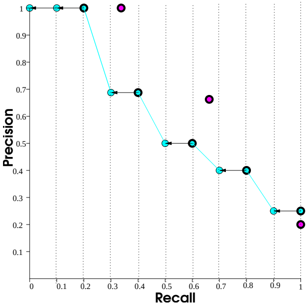


- Bold circles measured
- **thin circles interpolated**

Query 1					$\tilde{P}_1(r_0) = 1.00$
Rank		R	P		$\tilde{P}_1(r_1) = 1.00$
1	X	.20	1.00	$\tilde{P}_1(r_2) = 1.00$	
2					$\tilde{P}_1(r_3) = .67$
3	X	.40	.67	$\tilde{P}_1(r_4) = .67$	
4					
5					$\tilde{P}_1(r_5) = .50$
6	X	.60	.50	$\tilde{P}_1(r_6) = .50$	
7					
8					
9					$\tilde{P}_1(r_7) = .40$
10	X	.80	.40	$\tilde{P}_1(r_8) = .40$	
11					
12					
13					
14					$\tilde{P}_1(r_9) = .25$
15					
16					
17					
18					
19					
20	X	1.00	.25	$\tilde{P}_1(r_{10}) = .25$	

- The six other r_js (r_0, r_1, r_3, r_5, r_7, r_9) are interpolated.

Worked Example avg-11-pt prec: Query 2, measured data points



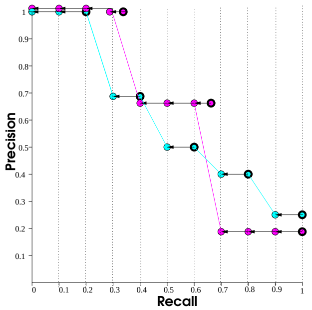
- Blue: Query 1; **Red: Query 2**
- **Bold circles measured**; thin circles interpol.

Query 2			
Rank	Relev.	R	P
1	X	.33	1.00
2			
3	X	.67	.67
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15	X	1.0	.2

$\tilde{P}_2(r_{10}) = .20$

- Only r_{10} coincides with a measured data point

Worked Example avg-11-pt prec: Query 2, interpolation



- Blue: Query 1; **Red: Query 2**
- Bold circles measured; **thin circles interpol.**

Query 2			
Rank	Relev.	R	P
1	X	.33	1.00
2			
3	X	.67	.67
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15	X	1.0	.2

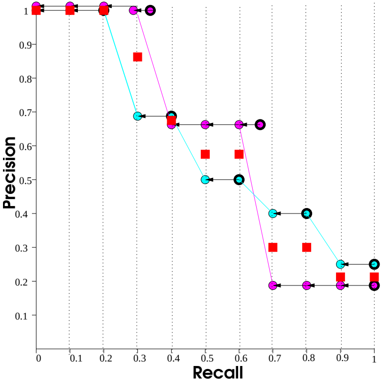
$\tilde{P}_2(r_{10}) = .20$

$\tilde{P}_2(r_0) = 1.00$
$\tilde{P}_2(r_1) = 1.00$
$\tilde{P}_2(r_2) = 1.00$
$\tilde{P}_2(r_3) = 1.00$

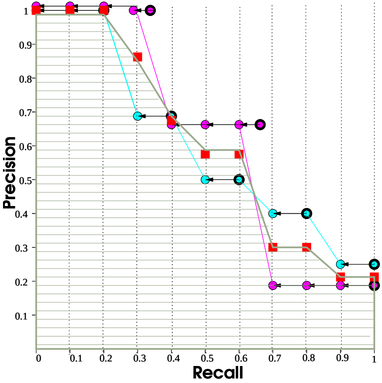
$\tilde{P}_2(r_4) = .67$
$\tilde{P}_2(r_5) = .67$
$\tilde{P}_2(r_6) = .67$

$\tilde{P}_2(r_7) = .20$
$\tilde{P}_2(r_8) = .20$
$\tilde{P}_2(r_9) = .20$

- 10 of the r_js are interpolated



- Now average at each p_j
- over N (number of queries)
- $\rightarrow$ 11 averages



- End result:
- 11 point average precision
- Approximation of area under prec. recall curve