Evaluation in Information Retrieval

Reading:
Textbook, Chapter 8

April 15, 2010

1

Written HW 2 to be assigned on Tuesday.

2

Example: Google vs. Yahoo vs. Bing

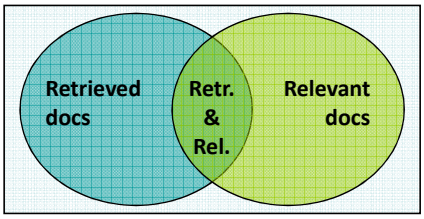
3

How to evaluate the results of a search engine or to compare different search engines?

4

Evaluation: Metrics

- Two basics:



Recall = $\frac{\text{\# documents retrieved and relevant}}{\text{\# documents relevant}}$

Precision = $\frac{\text{\# documents retrieved and relevant}}{\text{\# documents retrieved}}$

5

Definitions

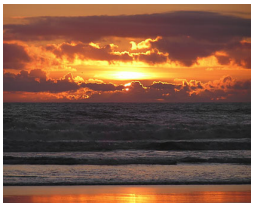
- "True positive" (retrieved and relevant)
- "False positive" (retrieved and not relevant)
- "True negative" (not retrieved and not relevant)
- "False negative" (not retrieved and relevant)

6

Examples in Classification Applications

Classified as "ocean"
True positive

Classified as "ocean"
False positive



Classified as "not ocean"
False negative

Classified as "not ocean"
True negative

7

Boolean search ("relevant"/ "nonrelevant") evaluation measures

- Precision

$$P = \frac{TP}{TP + FP} = \frac{\text{relevant and retrieved documents}}{\text{all retrieved documents}}$$

- Recall

$$R = \frac{TP}{TP + FN} = \frac{\text{relevant and retrieved documents}}{\text{all relevant documents}}$$

8

Examples

9

- In what cases would we care more about precision?
- In what cases would we care more about recall?

10

Precision-Recall Curve

- For search engines, we have more than just "relevant / nonrelevant" information for each document

  We also have "rank".

- Can define the *k* top ranked documents as "relevant".

- Can plot precision versus recall as we vary *k*, to see how they trade off against each other.

11

Precision-Recall Curve

Recall	Precision
0%	
10%	
20%	
30%	
40%	
50%	
60%	
70%	
80%	
90%	
100%	

12

Interpolated precision

For any recall level r, plot maximum precision at any equal or higher recall level $r' \geq r$

- This smoothes out "sawtooth" shape of graph
- Justification: "almost anyone would be prepared to look at a few more documents if it would increase the percentage of the viewed set that were relevant".

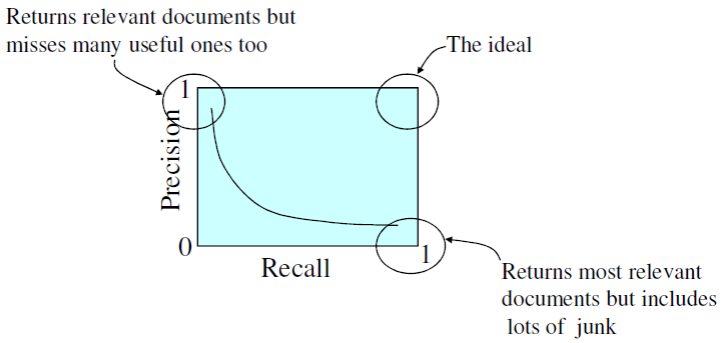
13

Plotting recall and precision

- Single query performance not necessarily representative of system
  - Compute recall and precision for multiple queries
  - Average the interpolated values at each recall level

14

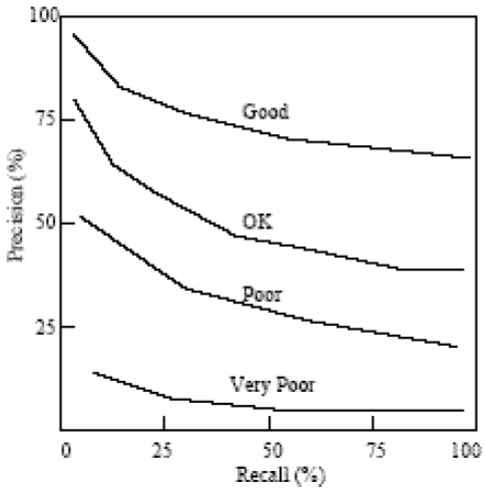
Trade-off between Recall and Precision



From: www.cs.utexas.edu/~mooney/ir-course/slides/Evaluation.ppt

15

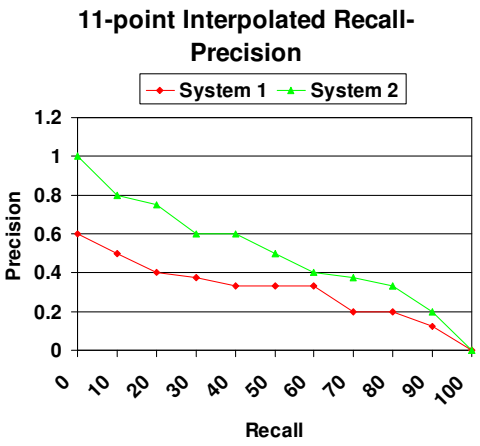
Precision / Recall Curves



16

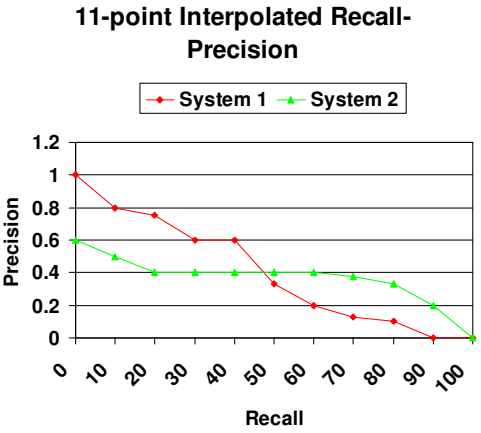
Which system is better?

Recall level	Average interpolated precision	
	System 1	System 2
0	0.6	1.0
10	0.5	0.8
20	0.4	0.75
30	0.375	0.6
40	0.33	0.6
50	0.33	0.5
60	0.33	0.4
70	0.2	0.375
80	0.2	0.33
90	0.125	0.2
100	0.0	0.0



Which system is better?

Recall level	Average interpolated precision	
	System 1	System 2
0	1.0	0.6
10	0.8	0.5
20	0.75	0.4
30	0.6	0.4
40	0.6	0.4
50	0.33	0.4
60	0.2	0.4
70	0.125	0.375
80	0.1	0.33
90	0.0	0.2
100	0.0	0.0



Mean average precision (MAP)

- Calculate average (uninterpolated) precision of top k documents
- Average this over a set of queries Q

19

R-precision

- Suppose we have a set of known relevant documents *Rel*.
- Calculate precision for top |*Rel*| documents retrieved.
- *and*
- Also called "break-even point": Value at which P and R are equal.
  - Suppose r of the |*Rel*| documents retrieved are actually relevant.
    - Then: $P = r / |Rel|$
    - But: $R = r / |Rel|$

20

F-measure: Weighted harmonic mean of precision and recall

F-measure

$$F_{\beta=1} = \frac{2PR}{P+R}$$

More generally:

$$F_{\beta} = \frac{(\beta^2+1)PR}{\beta^2 P + R}, \beta \in [0,\infty]$$

Idea: Avoid getting high score on mean by simply retrieving all documents (i.e., all are "relevant").

E.g., what would happen if all documents are retrieved under arithmetic versus harmonic mean of P and R?

Examples of F-measure.

21

Receiver Operating Characteristic (ROC) Curves

- Alternative to precision/recall curves
- Shows tradeoff between true positive rate and false positive rate.

True positive rate = TP/(TP + FN) ("sensitivity")
False positive rate = FP/(TN + FP)

22

Example ROC curve

23

How to obtain ground truth for judging relevance?

24

System quality and user utility

25