Information Retrieval
Tutorial 6: Evaluation

Professor: Michel Schellekens
TA: Ang Gao

University College Cork

2012-11-30

Overview

Precision and recall

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

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

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

- 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

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

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

Precision and recall

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

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

accuracy $= (TP + TN)/(TP + FP + FN + TN)$.

Accuracy

- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.

Accuracy

- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table above, accuracy $= (TP + TN)/(TP + FP + FN + TN)$.

Accuracy

- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table above, accuracy $= (TP + TN)/(TP + FP + FN + TN)$.
- Why is accuracy not a useful measure for web information retrieval?

Accuracy

- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table above, accuracy $= (TP + TN)/(TP + FP + FN + TN)$.
- Why is accuracy not a useful measure for web information retrieval?
- Ans: In IR system normaly only a small fraction of documents in the collection are relevance, as a result $TN >> TP$, even we have a good IR system which only retrieve relevant documents, the accuracy between this good IR system with a poor system(such as always return nothing) is small, thus this measurement can't help us evaluate IR system.

Exercise

- The snoogle search engine below always returns 0 results ("0 matching results found"), regardless of the query. Why does snoogle demonstrate that accuracy is not a useful measure in IR?

Why accuracy is a useless measure in IR

Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing

Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
- You then get 99.99% accuracy on most queries.

Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
- You then get 99.99% accuracy on most queries.
- Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk.

Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
- You then get 99.99% accuracy on most queries.
- 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.

Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
- You then get 99.99% accuracy on most queries.
- 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.

Precision/recall tradeoff

- You can increase recall by returning more docs.

Precision/recall tradeoff

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.

Precision/recall tradeoff

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs has 100% recall!

Precision/recall tradeoff

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

Precision/recall tradeoff

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs has 100% recall!
- The converse is also true (usually): It's easy to get high precision for very low recall.
- Which is better: IR sytem1 P: 63% R: 57%, IR system2 P: 69% R:60%

A combined measure: F

- F allows us to trade off precision against recall.

A combined measure: F

- F allows us to trade off precision against recall.
- $$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

A combined measure: F

- F allows us to trade off precision against recall.
- $$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$
- $\alpha \in [0,1]$ and thus $\beta^2 \in [0,\infty]$

A combined measure: F

- F allows us to trade off precision against recall.
- $$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$
- $\alpha \in [0,1]$ and thus $\beta^2 \in [0,\infty]$
- Most frequently used: balanced F with $\beta = 1$ or $\alpha = 0.5$

A combined measure: F

- F allows us to trade off precision against recall.
- $$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$
- $\alpha \in [0,1]$ and thus $\beta^2 \in [0,\infty]$
- Most frequently used: balanced F with $\beta = 1$ or $\alpha = 0.5$
  - This is the harmonic mean of P and R: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$

A combined measure: F

- F allows us to trade off precision against recall.
- $$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$
- $\alpha \in [0,1]$ and thus $\beta^2 \in [0,\infty]$
- Most frequently used: balanced F with $\beta = 1$ or $\alpha = 0.5$
  - This is the harmonic mean of P and R: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$
  - $F = \frac{2PR}{P+R}$

F: Exercise

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

F: Exercise

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

- $P = 20/(20 + 40) = 1/3$

F: Exercise

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

- $P = 20/(20+40) = 1/3$
- $R = 20/(20+60) = 1/4$

F: Exercise

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

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

F: Why harmonic mean?

- Why don't we use a different mean of P and R as a measure?

F: Why harmonic mean?

- Why don't we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean $\frac{P+R}{2}$

F: Why harmonic mean?

- Why don't we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean $\frac{P+R}{2}$
- The simple (arithmetic) mean is 50% for "return-everything" search engine, which is too high.

F: Why harmonic mean?

- Why don't we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean $\frac{P+R}{2}$
- The simple (arithmetic) mean is 50% for "return-everything" search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.

F: Why harmonic mean?

- Why don't we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean $\frac{P+R}{2}$
- The simple (arithmetic) mean is 50% for "return-everything" search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
- Taking the minimum achieves this.

F: Why harmonic mean?

- Why don't we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean $\frac{P+R}{2}$
- The simple (arithmetic) mean is 50% for "return-everything" search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
- Taking the minimum achieves this.
- But minimum is not smooth and hard to weight.

F: Why harmonic mean?

- Why don't we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean $\frac{P+R}{2}$
- The simple (arithmetic) mean is 50% for "return-everything" search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
- Taking the minimum achieves this.
- But minimum is not smooth and hard to weight.
- F (harmonic mean) is a kind of smooth minimum.

Difficulties in using precision, recall and F

- We need relevance judgments for information-need-document pairs – but they are expensive to produce.

Difficulties in using precision, recall and F

- We need relevance judgments for information-need-document pairs – but they are expensive to produce.
- For alternatives to using precision/recall and having to produce relevance judgments

Framework for the evaluation of an IR system

- *test collection* consisting of (i) a document collection, (ii) a test suite of information needs and (iii) a set of relevance judgements for each *doc-query* pair
- *gold-standard* judgement of relevance
  $\rightarrow$ classification of a document either as relevant or as irrelevant wrt an information need

Assessing relevance

- How good is an IR system at satisfaying an information need ?
- Needs an agreement between judges

  $\rightarrow$ computable via the **kappa** statistic:

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where:
$P(A)$: the proportion of agreements within the judgements
$P(E)$: what agreement would we get by chance

Assessing relevance: an example

Consider the following judgements (from Manning et al., 2008):

		Judge 2		
		Yes	No	Total
Judge 1	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

$P(A)$ is the proportion of agreements within the judgements
$P(E)$ is the proportion of expected agreements

Assessing relevance: an example

Consider the following judgements (from Manning et al., 2008):

		Judge 2		
		Yes	No	Total
Judge 1	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

- $$P(A) = \frac{370}{400} \qquad P(E) = P(rel)^2 + P(notrel)^2$$
  $$P(rel) = \frac{1}{2}\frac{320}{400} + \frac{1}{2}\frac{310}{400} = \frac{320+310}{800} \qquad P(notrel) = \frac{80+90}{800}$$

$P(A)$ is the proportion of agreements within the judgements
$P(E)$ is the proportion of expected agreements

Assessing relevance: an example

Consider the following judgements (from Manning et al., 2008):

		Judge 2		
		Yes	No	Total
Judge 1	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

- $$P(A) = \frac{370}{400} \qquad P(E) = P(rel)^2 + P(notrel)^2$$
  $$P(rel) = \frac{1}{2}\frac{320}{400} + \frac{1}{2}\frac{310}{400} = \frac{320+310}{800} \qquad P(notrel) = \frac{80+90}{800}$$
- $$kappa = \frac{P(A) - P(E)}{1 - P(E)} \qquad k = 0.776$$

$P(A)$ is the proportion of agreements within the judgements
$P(E)$ is the proportion of expected agreements

Exercise

Consider the following judgements:

		Judge 2		
		Yes	No	Total
Judge 1	Yes	120	30	150
	No	30	20	50
	Total	150	50	200

Exercise

Consider the following judgements:

		Judge 2		
		Yes	No	Total
Judge 1	Yes	120	30	150
	No	30	20	50
	Total	150	50	200

- $$P(A) = \frac{120 + 20}{200} = 0.7$$
  $$P(rel) = \frac{150 + 150}{400} = 0.75 \quad P(notrel) = \frac{50 + 50}{400} = 0.25$$

Exercise

Consider the following judgements:

		Judge 2		
		Yes	No	Total
Judge 1	Yes	120	30	150
	No	30	20	50
	Total	150	50	200

- $$P(A) = \frac{120 + 20}{200} = 0.7$$
  $$P(rel) = \frac{150 + 150}{400} = 0.75 \quad P(notrel) = \frac{50 + 50}{400} = 0.25$$
- $$P(E) = P(rel)^2 + P(notrel)^2 = 0.75^2 + 0.25^2 = 0.625$$
  $$k = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.7 - 0.625}{1 - 0.625} = 0.2$$

Assessing relevance (continued)

- Interpretation of the kappa statistic k:
  - Values of k in the interval [2/3, 1.0] are seen as acceptable.
  - With smaller values: need to redesign relevance assessment methodology used etc.
- Note that the kappa statistic can be negative if the agreements between judgements are worse than random
- In case of large variations between judgements, one can choose an assessor as a gold-standard

  $\rightarrow$ considerable impact on the *absolute* assessment
  $\rightarrow$ little impact on the *relative* assessment

Exercise

Below is a table showing how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you have written an IR system that for this query returns the set of documents $\{4, 5, 6, 7, 8\}$.

- Calculate the kappa measure between the two judges.
- Calculate precision, recall, and F1 of your system if a document is considered relevant only if the two judges agree.
- Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant.

docID	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

Solution

Part a.

- $P(A) = \frac{4}{12}$ $P(rel) = \frac{12}{24}$ $P(notrel) = \frac{12}{24}$
- $P(E) = P(rel)^2 + P(notrel)^2 = \frac{1}{2}$
- $k = \frac{P(A) - P(E)}{1 - P(E)} = -\frac{1}{3}$

Part b.

- Relevant $= \{3,4\}$ Retrieved $= \{4,5,6,7,8\}$
- $P = \frac{1}{5}$ $R = \frac{1}{2}$ $F_1 = \frac{2PR}{P+R} = \frac{2}{7}$

Part c.

- Relevant $= \{3,4,5,6,7,8,9,10,11,12\}$ Retrieved $= \{4,5,6,7,8\}$
- $P = \frac{5}{5} = 1$ $R = \frac{5}{10} = \frac{1}{2}$ $F_1 = \frac{2PR}{P+R} = \frac{2}{3}$

docID	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1