Introduction to Information Retrieval

http://informationretrieval.org

IIR 8: Evaluation & Result Summaries

Hinrich Schütze

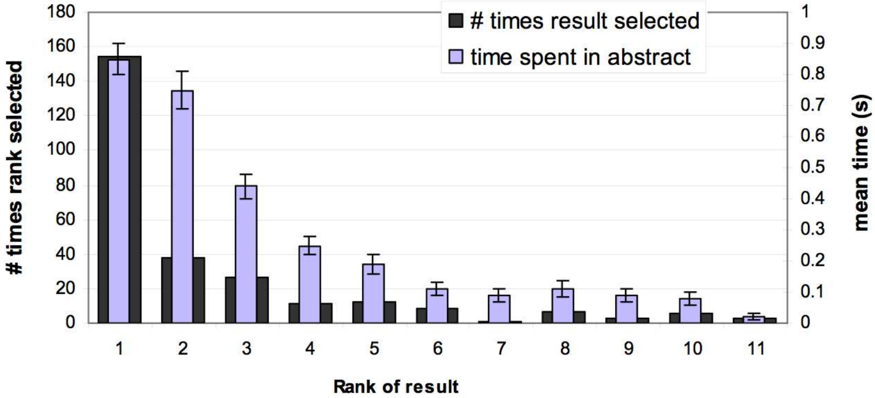
Institute for Natural Language Processing, Universität Stuttgart

2008.06.02

Overview

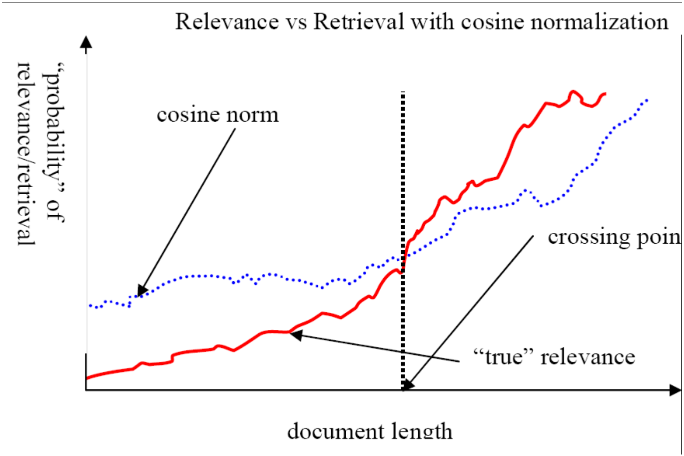
Outline

Looking vs. Clicking



- Users view results one and two more often / thoroughly
- Users click most frequently on result one

Pivot normalization



source: Lillian Lee

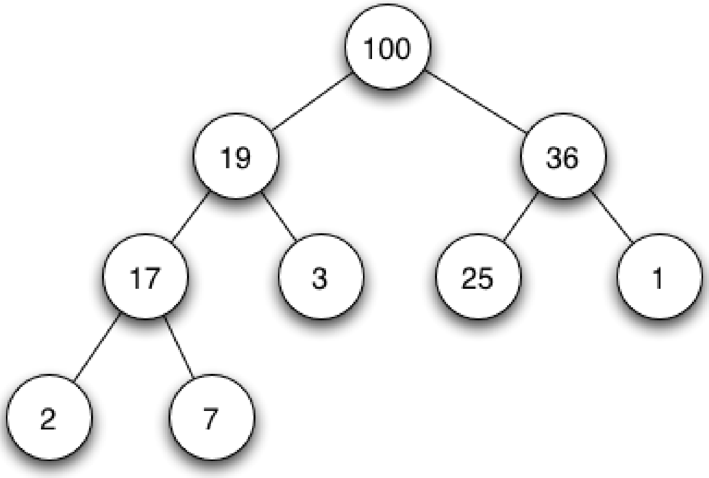
Now we also need term frequencies in the index

BRUTUS	⟶	1,2	7,3	83,1	87,2	...
CAESAR	⟶	1,1	5,1	13,1	17,1	...
CALPURNIA	⟶	7,1	8,2	40,1	97,3	

Use heap for selecting the top k in ranking

- A heap efficiently implements a priority queue.
- Takes $O(N)$ operations to construct (where N is the number of documents) ...
- ... then each of k winners can be read off in $O(k \log k)$ steps.
- Allows to rank in time linear in N (for small k and large N) – as opposed to $O(N \log N)$.

Binary max heap



Outline

Measures for a search engine

- How fast does it index

Measures for a search engine

- How fast does it index
  - Number of documents/bytes per hour

Measures for a search engine

- How fast does it index
  - Number of documents/bytes per hour
- How fast does it search

Measures for a search engine

- How fast does it index
  - Number of documents/bytes per hour
- How fast does it search
  - Latency as a function of index size / queries per second

Measures for a search engine

- How fast does it index
  - Number of documents/bytes per hour
- How fast does it search
  - Latency as a function of index size / queries per second
- What is the cost per query?

Measures for a search engine

- How fast does it index
  - Number of documents/bytes per hour
- How fast does it search
  - Latency as a function of index size / queries per second
- What is the cost per query?
  - Given certain requirements, e.g., a 20-billion-page index

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
  - Speed of response

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
  - Speed of response
  - Size of index

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- 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

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- 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
  - Most important: relevance

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- 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
  - Most important: relevance
  - (Actually, maybe most important: it's free)

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- 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
  - Most important: relevance
  - (Actually, maybe most important: it's free)
- Note that none of these is sufficient: blindingly fast, but useless answers won't make a user happy.

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- 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
  - Most important: relevance
  - (Actually, maybe most important: it's free)
- Note that none of these is sufficient: blindingly fast, but useless answers won't make a user happy.
- How can we quantify user happiness?

Who is the user?

- Who is the user we are trying to make happy?

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for.

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for.
- Web search engine: advertiser. Do searchers click through to my ads?

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for.
- Web search engine: advertiser. Do searchers click through to my ads?
- Ecommerce: buyer. Buyer buys what she came to the site to buy.

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for.
- Web search engine: advertiser. Do searchers click through to my ads?
- Ecommerce: buyer. Buyer buys what she came to the site to buy.
- Ecommerce: seller. Seller is able to sell her wares (because the search function directed buyers to the correct items).

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for.
- Web search engine: advertiser. Do searchers click through to my ads?
- Ecommerce: buyer. Buyer buys what she came to the site to buy.
- Ecommerce: seller. Seller is able to sell her wares (because the search function directed buyers to the correct items).
- Enterprise: CEO. Employees are more productive because they find right away what they are looking for.

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: advertiser. Do searchers click through to my ads?
- Ecommerce: buyer. Buyer buys what she came to the site to buy.
- Ecommerce: seller. Seller is able to sell her wares (because the search function directed buyers to the correct items).
- Enterprise: CEO. Employees are more productive because they find right away what they are looking for.

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: advertiser. Do searchers click through to my ads? Measure: clickthrough rate
- Ecommerce: buyer. Buyer buys what she came to the site to buy.
- Ecommerce: seller. Seller is able to sell her wares (because the search function directed buyers to the correct items).
- Enterprise: CEO. Employees are more productive because they find right away what they are looking for.

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: advertiser. Do searchers click through to my ads? Measure: clickthrough rate
- Ecommerce: buyer. Buyer buys what she came to the site to buy. Measures: time to purchase, fraction of "conversions" of searchers to buyers
- Ecommerce: seller. Seller is able to sell her wares (because the search function directed buyers to the correct items).
- Enterprise: CEO. Employees are more productive because they find right away what they are looking for.

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: advertiser. Do searchers click through to my ads? Measure: clickthrough rate
- Ecommerce: buyer. Buyer buys what she came to the site to buy. Measures: time to purchase, fraction of "conversions" of searchers to buyers
- Ecommerce: seller. Seller is able to sell her wares (because the search function directed buyers to the correct items). Measure: profit per item sold
- Enterprise: CEO. Employees are more productive because they find right away what they are looking for.

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: advertiser. Do searchers click through to my ads? Measure: clickthrough rate
- Ecommerce: buyer. Buyer buys what she came to the site to buy. Measures: time to purchase, fraction of "conversions" of searchers to buyers
- Ecommerce: seller. Seller is able to sell her wares (because the search function directed buyers to the correct items). Measure: profit per item sold
- Enterprise: CEO. Employees are more productive because they find right away what they are looking for. Measure: profit of the company

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries
  - A binary (or, more rarely, non-binary) assessment of the relevance of each query-document pair

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries
  - A binary (or, more rarely, non-binary) assessment of the relevance of each query-document pair
- This is a type of "canned" evaluation – often criticized as not being realistic enough.

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries
  - A binary (or, more rarely, non-binary) assessment of the relevance of each query-document pair
- This is a type of "canned" evaluation – often criticized as not being realistic enough.
- But has been very successful in IR.

Relevance: query vs. information need

- Relevance to what?

Relevance: query vs. information need

- Relevance to what?
- Take 1: relevance to the query

Relevance: query vs. information need

- Relevance to what?
- Take 1: relevance to the query
- "Relevance to the query" is very problematic.

Relevance: query vs. information need

- Relevance to what?
- Take 1: relevance to the query
- "Relevance to the query" is very problematic.
- Information need i: You are looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.

Relevance: query vs. information need

- Relevance to what?
- Take 1: relevance to the query
- "Relevance to the query" is very problematic.
- Information need i: You are looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.
- This is an information need, not a query.

Relevance: query vs. information need

- Relevance to what?
- Take 1: relevance to the query
- "Relevance to the query" is very problematic.
- Information need i: You are looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.
- This is an information need, not a query.
- Query q: WINE AND RED AND WHITE AND HEART AND ATTACK

Relevance: query vs. information need

- Relevance to what?
- Take 1: relevance to the query
- "Relevance to the query" is very problematic.
- Information need i: You are looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.
- This is an information need, not a query.
- Query q: WINE AND RED AND WHITE AND HEART AND ATTACK
- Consider document d': *He then launched into the heart of his speech and attacked the wine industry lobby for downplaying the role of red and white wine in drunk driving.*

Relevance: query vs. information need

- Relevance to what?
- Take 1: relevance to the query
- "Relevance to the query" is very problematic.
- Information need i: You are looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.
- This is an information need, not a query.
- Query q: WINE AND RED AND WHITE AND HEART AND ATTACK
- Consider document d': *He then launched into the heart of his speech and attacked the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- d' is relevant to the query q ...

Relevance: query vs. information need

- Relevance to what?
- Take 1: relevance to the query
- "Relevance to the query" is very problematic.
- Information need i: You are looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.
- This is an information need, not a query.
- Query q: WINE AND RED AND WHITE AND HEART AND ATTACK
- Consider document d': *He then launched into the heart of his speech and attacked the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- d' is relevant to the query q ...
- d' is not relevant to the information need i.

Relevance: query vs. information need

- d' is relevant to the query q ...
- d' is not relevant to the information need i.

Relevance: query vs. information need

- d' is relevant to the query q ...
- d' is not relevant to the information need i.
- User happiness can only be measured by relevance to an information need, not by relevance to queries.

Relevance: query vs. information need

- d' is relevant to the query q ...
- d' is not relevant to the information need i.
- User happiness can only be measured by relevance to an information need, not by relevance to queries.
- Our terminology is sloppy in these slides and in IIR: we talk about query-document relevance judgments even though we mean information-need-document relevance judgments.

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)

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

Accuracy

- Why do we use complex measures like precision and recall?

Accuracy

- Why do we use complex measures like precision and recall?
- Why not something simple like accuracy?

Accuracy

- Why do we use complex measures like precision and recall?
- Why not something simple like accuracy?
- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.

Accuracy

- Why do we use complex measures like precision and recall?
- Why not something simple like 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

- Why do we use complex measures like precision and recall?
- Why not something simple like 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?

Why not just use accuracy?

Why not just use accuracy?

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

Why not just use accuracy?

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

Why not just use accuracy?

- 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 not just use accuracy?

- 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.
- Accuracy is not a good measure of user happiness, we'll use precision and recall instead.

Difficulties in using precision/recall

- We should always average over a large set of queries.

Difficulties in using precision/recall

- We should always average over a large set of queries.
  - There is no such thing as a "typical" or "representative" query.

Difficulties in using precision/recall

- We should always average over a large set of queries.
  - There is no such thing as a "typical" or "representative" query.
- We need relevance judgments for information-need-document pairs – but they are expensive to produce.

Difficulties in using precision/recall

- We should always average over a large set of queries.
  - There is no such thing as a "typical" or "representative" query.
- 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 – see end of this lecture.

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.
- Suppose the document with the largest score is relevant. How can we maximize precision?

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: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$
- What value range of β do I choose for weighting recall higher than precision?

F: Example

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

F: Example

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

- precision?

F: Example

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

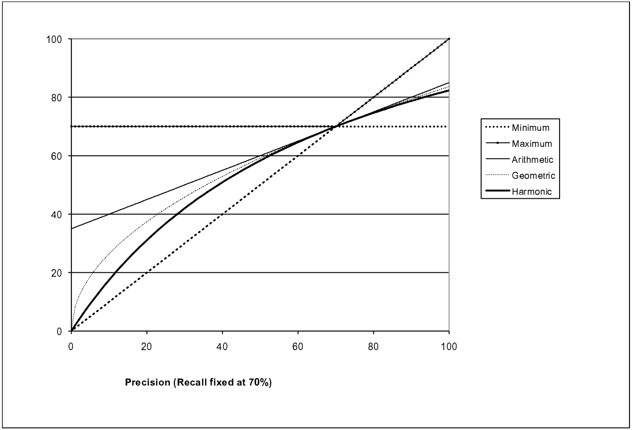
- recall?

F: Example

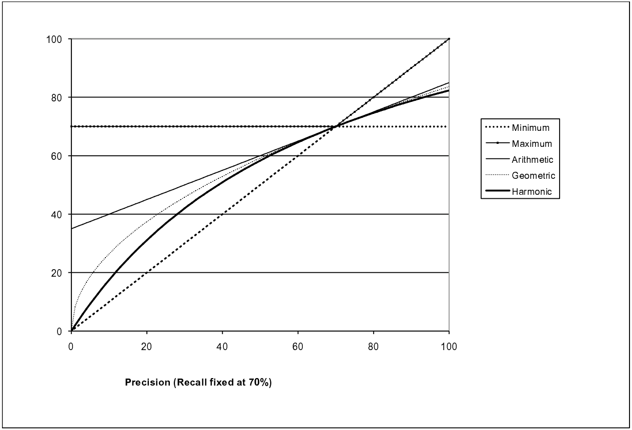
	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

- F_1?

F_1 and other averages



F_1 and other averages



- We can view the harmonic mean as a kind of soft minimum

F: Why harmonic mean?

- The simple (arithmetic) mean is 50% for "return-everything" search engine, which is too high.

F: Why harmonic mean?

- 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?

- 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?

- 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?

- 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.

Outline

Precision-recall curve

- Precision/recall/F are measures for unranked sets.

Precision-recall curve

- Precision/recall/F are measures for unranked sets.
- We can easily turn set measures into measures of ranked lists.

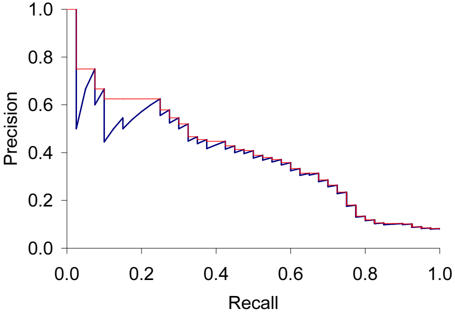
Precision-recall curve

- 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

Precision-recall curve

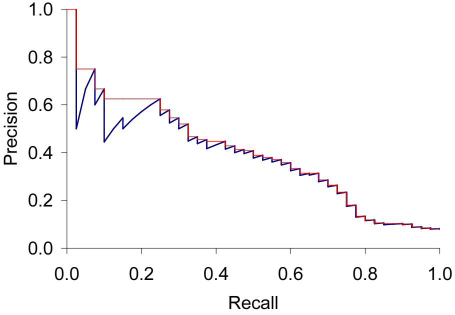
- Precision/recall/F are measures for unranked sets.
- We can easily turn set measures into measures of ranked lists.
- Just compute the set measure for each "prefix": the top 1, top 2, top 3, top 4 etc results
- Doing this for precision and recall gives you a precision-recall curve.

A precision-recall curve



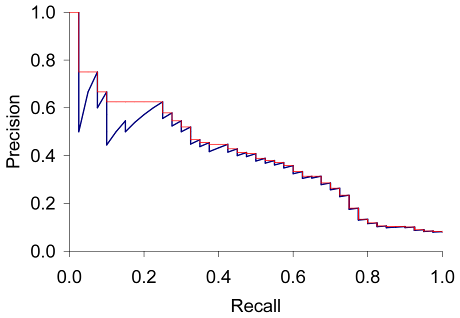
.

A precision-recall curve



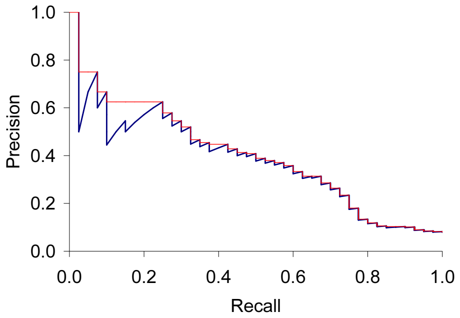
- Each point corresponds to a result for the top k ranked hits ($k = 1, 2, 3, 4, \ldots$).

A precision-recall curve



- 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

A precision-recall curve



- 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 more stuff if both precision and recall get better.

11-point interpolated average precision

Recall	Interpolated Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

11-point interpolated average precision

Recall	Interpolated Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

11-point average: $\approx$ 0.425

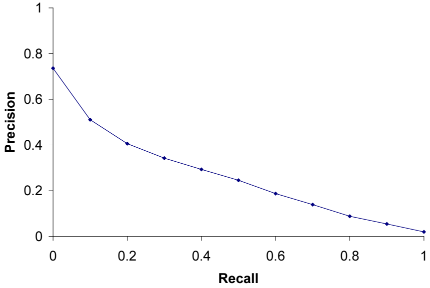
11-point interpolated average precision

Recall	Interpolated Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

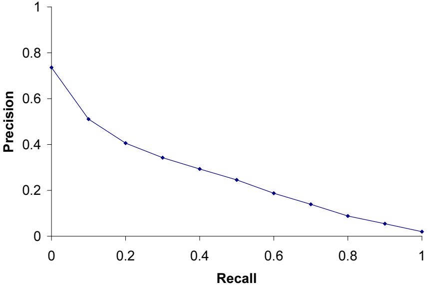
11-point average: $\approx$ 0.425

How can precision at 0.0 be > 0?

Averaged 11-point precision/recall graph

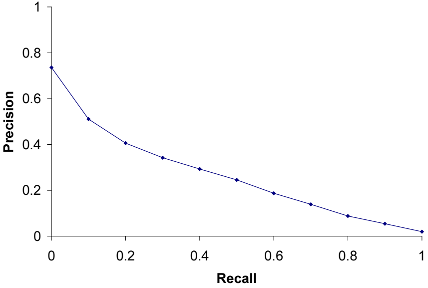


Averaged 11-point precision/recall graph



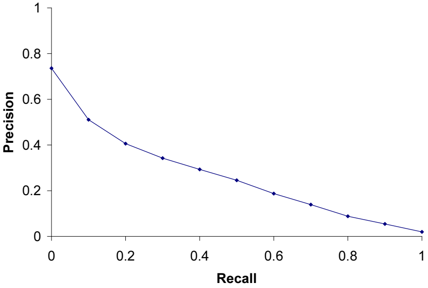
- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, . . .

Averaged 11-point precision/recall graph



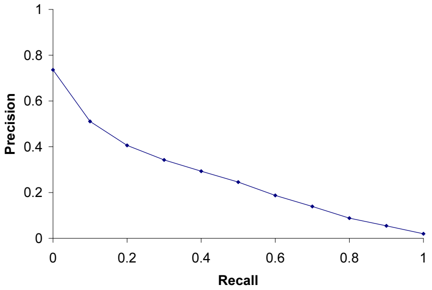
- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, . . .
- Do this for each of the queries in the evaluation benchmark

Averaged 11-point precision/recall graph



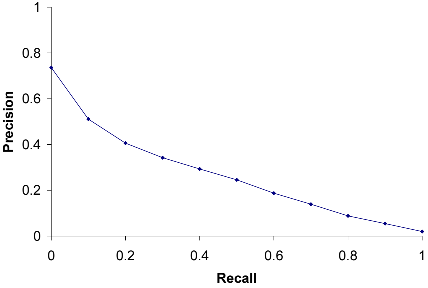
- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, . . .
- Do this for each of the queries in the evaluation benchmark
- Average over queries

Averaged 11-point precision/recall graph



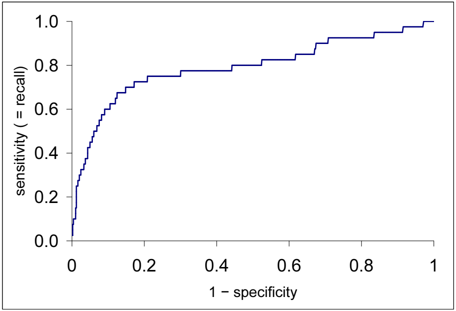
- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, ...
- Do this for each of the queries in the evaluation benchmark
- Average over queries
- This measure measures performance at all recall levels.

Averaged 11-point precision/recall graph

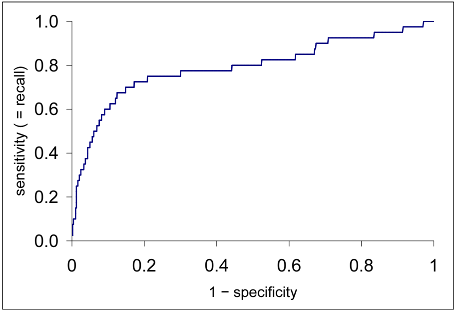


- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, . . .
- Do this for each of the queries in the evaluation benchmark
- Average over queries
- This measure measures performance at all recall levels.
- The curve is typical of performance levels at TREC.

ROC curve

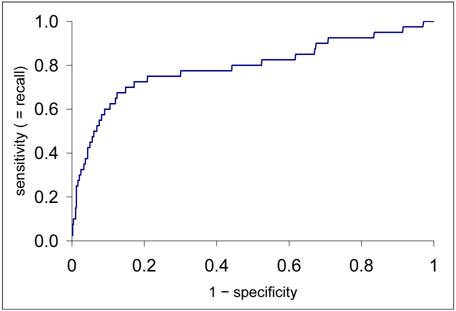

1.0
0.8
0.6
0.4
0.2
0.0
sensitivity (= recall)
0
0.2
0.4
0.6
0.8
1
1 − specificity

ROC curve



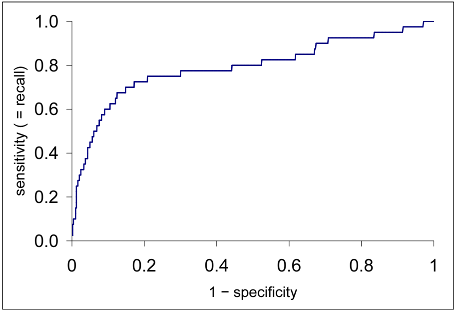
- Similar to precision-recall graph

ROC curve



- Similar to precision-recall graph
- But we are only interested in the small area in the lower left corner.

ROC curve



- Similar to precision-recall graph
- But we are only interested in the small area in the lower left corner.
- Precision-recall graph "blows up" this area.

Variance of measures like precision/recall

- For a test collection, it is usual that a system does crummily on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and excellently on others (e.g., $P = 0.95$ at $R = 0.1$).

Variance of measures like precision/recall

- For a test collection, it is usual that a system does crummily on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and excellently 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.

Variance of measures like precision/recall

- For a test collection, it is usual that a system does crummily on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and excellently 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.

Outline

1. Recap
2. Unranked evaluation
3. Ranked evaluation
4. Evaluation benchmarks
5. Result summaries

What we need for a benchmark

- A collection of documents

What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.

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

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
  - ...which we will often incorrectly refer to as queries

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
  - . . . which we will often incorrectly refer to as queries
  - Information needs must be representative of the information needs we expect to see in reality.

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
  - ...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

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
  - ...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
  - We need to hire/pay "judges" or assessors to do this.

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
  - ...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
  - We need to hire/pay "judges" or assessors to do this.
  - Expensive, time-consuming

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
  - ...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
  - 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.

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
  - ...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
  - 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.
  - Relevance assessments are only usable if they are consistent.

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
  - . . . 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
  - 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.
  - Relevance assessments are only usable if they are consistent.
  - How can we measure this consistency or agreement among judges?

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
  - . . . 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
  - 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.
  - Relevance assessments are only usable if they are consistent.
  - How can we measure this consistency or agreement among judges?

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
  - ...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
  - 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.
  - Relevance assessments are only usable if they are consistent.
  - How can we measure this consistency or agreement among judges? Kappa measure in a few slides

Standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness

Standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK

Standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK
- 1398 abstracts of aerodynamics journal articles, a set of 225 queries, exhaustive relevance judgments of all query-document-pairs

Standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK
- 1398 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

Standard relevance benchmark: TREC

- TREC = Text Retrieval Conference (TREC)

Standard relevance benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST)

Standard 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.

Standard 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

Standard 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

Standard 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

Standard 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.

Standard relevance benchmarks: Others

- GOV2

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR
  - East Asian language and cross-language information retrieval

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR
  - East Asian language and cross-language information retrieval
- Cross Language Evaluation Forum (CLEF)

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR
  - East Asian language and cross-language information retrieval
- Cross Language Evaluation Forum (CLEF)
  - This evaluation series has concentrated on European languages and cross-language information retrieval.

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR
  - East Asian language and cross-language information retrieval
- Cross Language Evaluation Forum (CLEF)
  - This evaluation series has concentrated on European languages and cross-language information retrieval.
- Many others

Kappa measure

- Kappa is measure of how much judges agree or disagree.

Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments

Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments
- Corrects for chance agreement

Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments
- Corrects for chance agreement
- $P(A) =$ proportion of time judges agree

Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments
- Corrects for chance agreement
- $P(A) =$ proportion of time judges agree
- $P(E) =$ what agreement would we get by chance

Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments
- Corrects for chance agreement
- $P(A) =$ proportion of time judges agree
- $P(E) =$ what agreement would we get by chance
- $$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments
- Corrects for chance agreement
- $P(A)$ = proportion of time judges agree
- $P(E)$ = what agreement would we get by chance
- $$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$
- $\kappa =$? for (i) chance agreement (ii) total agreement

Kappa measure (2)

- Values of κ in the interval $[2/3, 1.0]$ are seen as acceptable.

Kappa measure (2)

- Values of κ in the interval $[2/3, 1.0]$ are seen as acceptable.
- With smaller values: need to redesign relevance assessment methodology used etc.

Calculating the kappa statistic

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

Observed proportion of the times the judges agreed

$P(A) = (300 + 70)/400 = 370/400 = 0.925$

Pooled marginals

$P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

Probability that the two judges agreed by chance

$P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2 = 0.665$

Kappa statistic

$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$

Calculating the kappa statistic

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

Observed proportion of the times the judges agreed

$P(A) = (300 + 70)/400 = 370/400 = 0.925$

Pooled marginals

$P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

Probability that the two judges agreed by chance

$P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2 = 0.665$

Kappa statistic

$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$

(still in acceptable range)

Interjudge agreement at TREC

information need	number of docs judged	disagreements	NR	R
51	211	6	4	2
62	400	157	149	8
67	400	68	37	31
95	400	110	108	2
127	400	106	12	94

Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?

Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.

Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.
- Large impact on absolute performance numbers

Impact of interjudge 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

Impact of interjudge 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
- Supposes we want to know if algorithm A is better than algorithm B

Impact of interjudge 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
- Supposes 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.

Critique of pure relevance

- We've defined relevance for an isolated query-document pair.

Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance

Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The marginal relevance of a document in a result list is the additional information it contributes.

Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The marginal relevance of a document in a result list is the additional information it contributes.
- Example: a duplicate can be highly relevant, but it has zero marginal relevance.

Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The marginal relevance of a document in a result list is the additional information it contributes.
- Example: a duplicate can be highly relevant, but it has zero marginal relevance.
- Marginal relevance is a better measure of user happiness.

Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The marginal relevance of a document in a result list is the additional information it contributes.
- Example: a duplicate can be highly relevant, but it has zero marginal relevance.
- Marginal relevance is a better measure of user happiness.
- But it is virtually impossible to run information retrieval experiments based on marginal relevance.

Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The marginal relevance of a document in a result list is the additional information it contributes.
- Example: a duplicate can be highly relevant, but it has zero marginal relevance.
- Marginal relevance is a better measure of user happiness.
- But it is virtually impossible to run information retrieval experiments based on marginal relevance.
- Why?

Evaluation at large search engines

- Recall is difficult to measure on the web

Evaluation at large search engines

- Recall is difficult to measure on the web
- Search engines often use precision at top k, e.g., $k = 10$ ...

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 measures that reward you more for getting rank 1 right than for getting rank 10 right.

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 measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures.

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 measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures.
  - Example 1: clickthrough on first result

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 measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures.
  - Example 1: clickthrough on first 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 nonrelevant) ...

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 measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures.
    - Example 1: clickthrough on first 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 nonrelevant) ...
    - ... but pretty reliable in the aggregate.

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 measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures.
  - Example 1: clickthrough on first 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 nonrelevant) ...
  - ... but pretty reliable in the aggregate.
  - Example 2: Ongoing studies of user behavior in the lab – recall last lecture

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 measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures.
  - Example 1: clickthrough on first 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 nonrelevant) ...
  - ... but pretty reliable in the aggregate.
  - Example 2: Ongoing studies of user behavior in the lab – recall last lecture
  - Example 3: A/B testing

A/B testing

- Purpose: Test a single innovation

A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.

A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system

A/B testing

- Purpose: Test a single innovation
- Prerequisite: 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

A/B testing

- Purpose: Test a single innovation
- Prerequisite: 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

A/B testing

- Purpose: Test a single innovation
- Prerequisite: 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.

A/B testing

- Purpose: Test a single innovation
- Prerequisite: 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

A/B testing

- Purpose: Test a single innovation
- Prerequisite: 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
- Variant: Give users the option to switch to new algorithm/interface

Outline

How do we present results to the user?

- Most often: as a list – aka "10 blue links"

How do we present results to the user?

- Most often: as a list – aka "10 blue links"
- How should each document in the list be described?

How do we present results to the user?

- Most often: as a list – aka "10 blue links"
- How should each document in the list be described?
- This description is crucial.

How do we present results to the user?

- Most often: as a list – aka “10 blue links”
- How should each document in the list be described?
- This description is crucial.
- User can identify good hits (= relevant hits) based on description.

How do we present results to the user?

- Most often: as a list – aka "10 blue links"
- How should each document in the list be described?
- This description is crucial.
- User can identify good hits (= relevant hits) based on description.
- No need to "click" on all documents sequentially

Doc description in result list

- Most commonly: doc title, url, some metadata . . .

Doc description in result list

- Most commonly: doc title, url, some metadata ...
- ...and a summary

Doc description in result list

- Most commonly: doc title, url, some metadata . . .
- . . . and a summary
- How do we “compute” the summary?

Summaries

- Two basic kinds: (i) static (ii) dynamic

Summaries

- Two basic kinds: (i) static (ii) dynamic
- A static summary of a document is always the same, regardless of the query that hit the document.

Summaries

- Two basic kinds: (i) static (ii) dynamic
- A static summary of a document is always the same, regardless of the query that hit the document.
- Dynamic summaries are query-dependent. They attempt to explain why the document was retrieved for the query at hand.

Static summaries

- In typical systems, the static summary is a subset of the document.

Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document

Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of “key” sentences

Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of "key" sentences
  - Simple NLP heuristics to score each sentence

Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of "key" sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.

Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of "key" sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
  - Machine learning approach: see IIR 13

Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of "key" sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
  - Machine learning approach: see IIR 13
- Most sophisticated: complex NLP to synthesize/generate a summary

Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of "key" sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
  - Machine learning approach: see IIR 13
- Most sophisticated: complex NLP to synthesize/generate a summary
  - For most IR applications: not quite ready for prime time yet

Dynamic summaries

- Present one or more “windows” or snippets within the document that contain several of the query terms.

Dynamic summaries

- Present one or more "windows" or snippets within the document that contain several of the query terms.
- Generated in conjunction with scoring

Dynamic summaries

- Present one or more “windows” or snippets within the document that contain several of the query terms.
- Generated in conjunction with scoring
- Prefer snippets in which query terms occurred as a phrase

Dynamic summaries

- Present one or more “windows” or snippets within the document that contain several of the query terms.
- Generated in conjunction with scoring
- Prefer snippets in which query terms occurred as a phrase
- Prefer snippets in which query terms occurred jointly in a small window

Dynamic summaries

- Present one or more "windows" or snippets within the document that contain several of the query terms.
- Generated in conjunction with scoring
- Prefer snippets in which query terms occurred as a phrase
- Prefer snippets in which query terms occurred jointly in a small window
- The summary that is computed this way gives the entire content of the window – all terms, not just the query terms.

A dynamic summary

Query: "new guinea economic development"

Snippets (in bold) that were extracted from a document: ... **In recent years, Papua New Guinea has faced severe economic difficulties and** economic growth has slowed, partly as a result of weak governance and civil war, and partly as a result of external factors such as the Bougainville civil war which led to the closure in 1989 of the Panguna mine (at that time the most important foreign exchange earner and contributor to Government finances), the Asian financial crisis, a decline in the prices of gold and copper, and a fall in the production of oil. **PNG's economic development record over the past few years is evidence that** governance issues underly many of the country's problems. Good governance, which may be defined as the transparent and accountable management of human, natural, economic and financial resources for the purposes of equitable and sustainable development, flows from proper public sector management, efficient fiscal and accounting mechanisms, and a willingness to make service delivery a priority in practice. ...

Google examples for dynamic summaries

Generating dynamic summaries

- Where do we get these other terms in the snippet from?

Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.

Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.

Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.

Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.
- Byte offset or word offset?

Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.
- Byte offset or word offset?
- Note that the cached copy can be outdated

Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.
- Byte offset or word offset?
- Note that the cached copy can be outdated
- Don't cache very long documents – just cache a short prefix

Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .

Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .
- . . . but snippets must be long enough to be meaningful.

Dynamic summaries

- Real estate on the search result page is limited → snippets must be short ...
- ... but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.

Dynamic summaries

- Real estate on the search result page is limited → snippets must be short ...
- ...but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets

Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .
- . . . but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don't have to look at the document.

Dynamic summaries

- Real estate on the search result page is limited → snippets must be short ...
- ... but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don't have to look at the document.
- Dynamic summaries are a big part of user happiness because

Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .
- . . . but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don't have to look at the document.
- Dynamic summaries are a big part of user happiness because
  - We can quickly scan them to find the relevant document we then click on.

Dynamic summaries

- Real estate on the search result page is limited → snippets must be short ...
- ... but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don't have to look at the document.
- Dynamic summaries are a big part of user happiness because
  - We can quickly scan them to find the relevant document we then click on.
  - In many cases, we don't have to click at all and save time.

Resources

- Chapter 8 of IIR

Resources

- Chapter 8 of IIR
- Resources at `http://ifnlp.org/ir`

Resources

- Chapter 8 of IIR
- Resources at `http://ifnlp.org/ir`
- The TREC home page – TREC had a huge impact on information retrieval evaluation.

Resources

- Chapter 8 of IIR
- Resources at `http://ifnlp.org/ir`
- The TREC home page – TREC had a huge impact on information retrieval evaluation.
- Originator of *F*-measure: Keith van Rijsbergen

Resources

- Chapter 8 of IIR
- Resources at `http://ifnlp.org/ir`
- The TREC home page – TREC had a huge impact on information retrieval evaluation.
- Originator of *F*-measure: Keith van Rijsbergen
- More on A/B testing

Resources

- Chapter 8 of IIR
- Resources at `http://ifnlp.org/ir`
- The TREC home page – TREC had a huge impact on information retrieval evaluation.
- Originator of F-measure: Keith van Rijsbergen
- More on A/B testing
- Tombros & Sanderson 1998: one of the first papers on dynamic summaries

Resources

- Chapter 8 of IIR
- Resources at `http://ifnlp.org/ir`
- The TREC home page – TREC had a huge impact on information retrieval evaluation.
- Originator of F-measure: Keith van Rijsbergen
- More on A/B testing
- Tombros & Sanderson 1998: one of the first papers on dynamic summaries
- Google VP of Engineering on search quality evaluation at Google