Chapter IR:VII

VII. Evaluation

- Laboratory Experiments
- Performance Measures
- Training and Testing
- Logging

Performance Measures

Effectiveness and Efficiency

Effectiveness is “the degree to which something is successful in producing a desired result; success”. [Oxford Dictionaries]

Efficiency is “the ratio of the useful work performed by a machine to the total energy expended”. [Oxford Dictionaries]

Effectiveness measures:

- Precision and Recall
- F-Measure
- Precision@k (rank k)
- Mean Average Precision (MAP)
- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (nDCG)

Efficiency measures:

- Indexing time,
  indexing space overhead, and
  index size
- Query throughput and query latency

Performance Measures

Effectiveness Measures

Effectiveness is "the degree to which something is successful in producing a desired result; success". [Oxford Dictionaries]

The desired result from a search engine for a user's query is relevant documents.

Our goal is to make justifiable claims such as these:

- This search engine is (not) effective.
- Search engine A is (ten times) more effective than search engine B.
- This search engine achieves the highest effectiveness for its search domain.

Performance Measures

Effectiveness Measures

Effectiveness is "the degree to which something is successful in producing a desired result; success". [Oxford Dictionaries]

The desired result from a search engine for a user's query is relevant documents.

Our goal is to make justifiable claims such as these:

- This search engine is (not) effective.
- Search engine A is (ten times) more effective than search engine B.
- This search engine achieves the highest effectiveness for its search domain.

Sufficient justification is achieved by means of measurement, namely "the assignment of a number to a characteristic of an object [a search result], which can be compared with other objects." [Wikipedia]

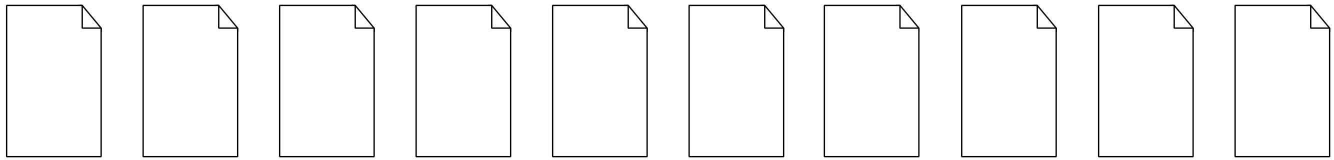
In practice, absolute claims are often difficult to be justified and hence less useful compared to relative claims.

Performance Measures

Effectiveness Measures

The object of measurement for a search engine's effectiveness is its search results:

A
system 1
topic 1



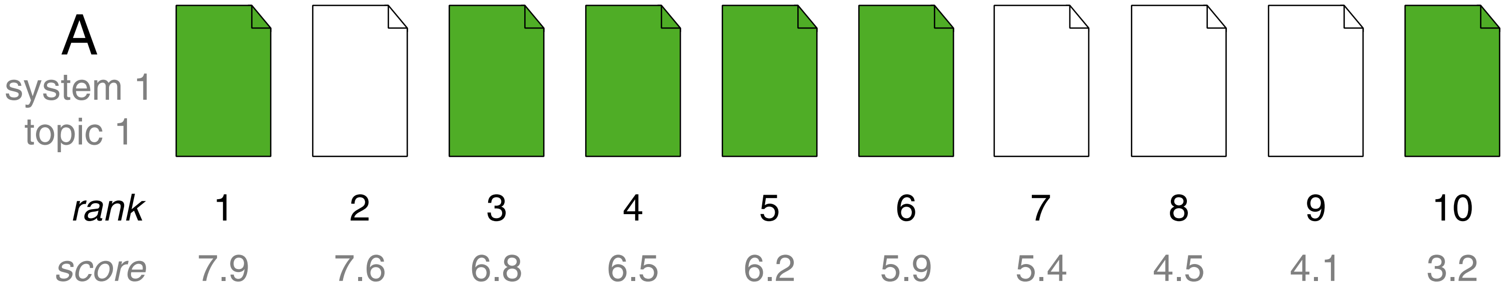
rank	1	2	3	4	5	6	7	8	9	10
score	7.9	7.6	6.8	6.5	6.2	5.9	5.4	4.5	4.1	3.2

A search result is composed of a list of documents, ordered by the search engine's estimation of relevance, optionally alongside relevance scores for each document.

Performance Measures

Effectiveness Measures

The object of measurement for a search engine's effectiveness is its search results:



A search result is composed of a list of documents, ordered by the search engine's estimation of relevance, optionally alongside relevance scores for each document.

The true relevance of each document is supplied (e.g., by relevance judgments).

An effectiveness measure maps a given search result and its relevance judgments to the real numbers, rendering rankings from different systems comparable.

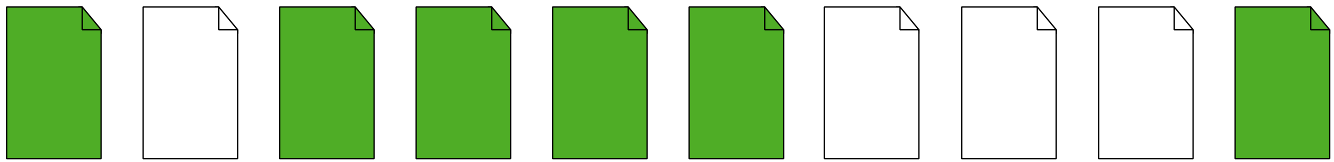
The mapping encodes a model of user behavior. Recent measures are based on realistic models; early measures did less so.

Performance Measures

Effectiveness Measures

The object of measurement for a search engine's effectiveness is its search results:

A
system 1
topic 1



rank	1	2	3	4	5	6	7	8	9	10
score	7.9	7.6	6.8	6.5	6.2	5.9	5.4	4.5	4.1	3.2

A search result is composed of a list of documents, ordered by the search engine's estimation of relevance, optionally alongside relevance scores for each document.

Two fundamental models of user behavior can be distinguished:

1. The user browses the entire result set in no particular order.
   ➜ Set Retrieval

2. The user browses the results in ranking order and eventually decides to stop.
   ➜ Ranked Retrieval

Performance Measures

Set Retrieval Effectiveness: Precision and Recall

The user browses the entire result set returned by the search engine, expecting only the relevant documents. A contingency table counts successes and failures:

	$\in$ *Relevant*	$\notin$ *Relevant*
$\in$ *Results*	a	b
$\notin$ *Results*	c	d

with

- *Results* = set of documents retrieved.
- *Relevant* = set of relevant documents.

Performance Measures

Set Retrieval Effectiveness: Precision and Recall

The user browses the entire result set returned by the search engine, expecting only the relevant documents. A contingency table counts successes and failures:

	$\in$ *Relevant*	$\notin$ *Relevant*
$\in$ *Results*	a	b
$\notin$ *Results*	c	d

➜

$$\textit{precision} = \frac{a}{a+b}$$

$$\textit{recall} = \frac{a}{a+c}$$

with

- *Results* = set of documents retrieved.
- *Relevant* = set of relevant documents.

Performance Measures

Set Retrieval Effectiveness: Precision and Recall

The user browses the entire result set returned by the search engine, expecting only the relevant documents. A contingency table counts successes and failures:

	$\in$ *Relevant*	$\notin$ *Relevant*
$\in$ *Results*	a	b
$\notin$ *Results*	c	d

➜

$$\textit{precision} = \frac{a}{a+b}$$

$$\textit{recall} = \frac{a}{a+c}$$

with

- *Results* = set of documents retrieved.
- *Relevant* = set of relevant documents.

In words:

- *precision* is the fraction of retrieved documents that are relevant.
- *recall* is the fraction of relevant documents that are retrieved.

Remarks:

- A contingency table displays the frequency distribution of two or more variables.
- In machine learning, it is also called confusion matrix. The measures are some of the ones that can be derived from it. [Wikipedia]
- Alternative formulas based on the sets of *Results* and *Relevant* documents:

$$\textit{precision} = \frac{|\textit{Relevant} \cap \textit{Results}|}{|\textit{Results}|}$$

$$\textit{recall} = \frac{|\textit{Relevant} \cap \textit{Results}|}{|\textit{Relevant}|}$$

- Precision and recall values are in the interval $[0, 1]$. Precision is undefined if the result set is empty, recall is undefined if there are no relevant documents.
- It is trivial to maximize recall by returning the entire document collection.
- The fraction of non-relevant documents that are retrieved is called

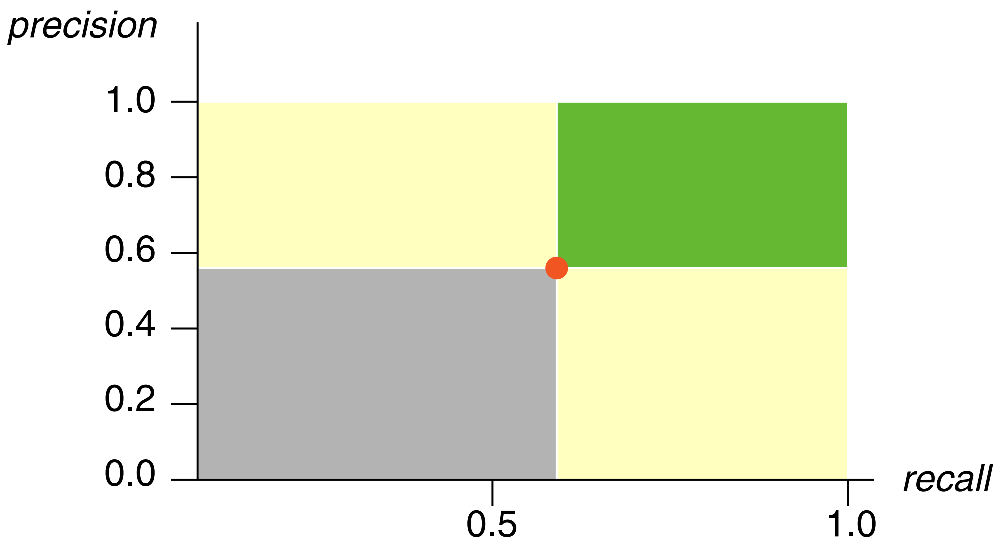
  $$\textit{fallout} = \frac{b}{b+d}$$

  If retrieval were a classification task, *recall* (true positive rate) and *fallout* (false positive rate) would be considered. As it stands, *precision* is more meaningful in ranked retrieval.

Performance Measures

Set Retrieval Effectiveness: F-Measure

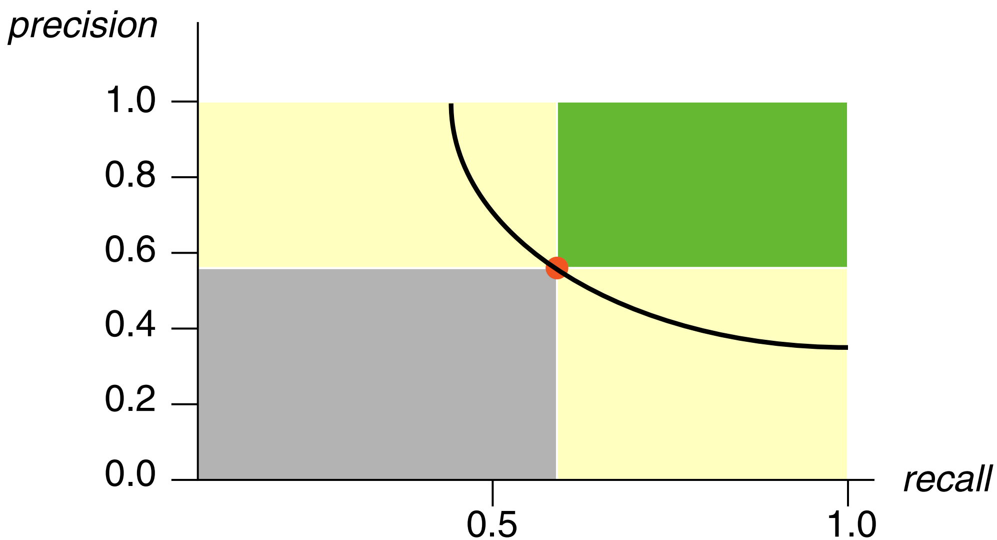
Comparison of retrieval systems: [plot]



Performance Measures

Set Retrieval Effectiveness: F-Measure

Comparison of retrieval systems: [plot]



The F-Measure is the harmonic mean of *precision* and *recall*:

$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{\textit{precision}} + \frac{1}{\textit{recall}}\right)} = \frac{2\,\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

Remarks:

- The scores of the F-Measure are in the interval $[0, 1]$.
- Precision and recall induce a partial ordering of retrieval systems: systems that perform better in one, but worse in the other measure cannot be ranked with regard to which one is better. The F-Measure calculates a single performance score from precision and recall, inducing a total order.
- The harmonic mean is employed, since it penalizes extreme values more than the arithmetic mean. It's equivalence curves also better resemble trade-offs human users might be willing to take when trading recall for precision, or vice versa.
- Precision and recall are not equally important in all retrieval tasks. Examples: Web search (high precision) vs. criminal suspect search (high recall). A weighted F-Measure computes as follows:

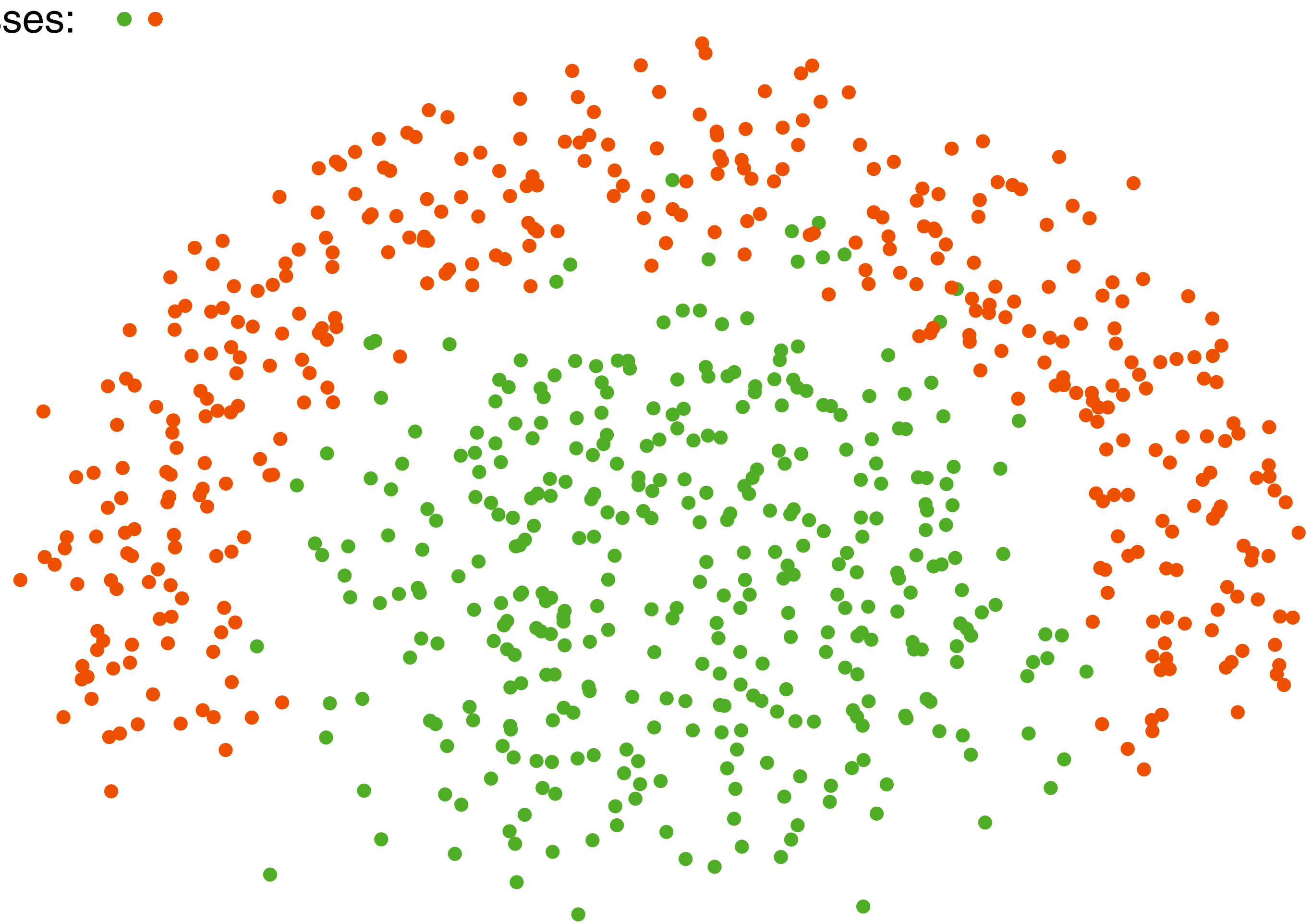
$$F = \frac{1}{\alpha \frac{1}{\textit{precision}} + (1-\alpha)\frac{1}{\textit{recall}}} = \frac{(\beta^2+1)\textit{precision} \cdot \textit{recall}}{\beta^2 \textit{precision} + \textit{recall}}, \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}.$$

  Values of $\beta > 1$ emphasize recall, values of $\beta < 1$ emphasize precision. The default F-Measure used is $F_{\beta=1}$, or F_1 for short.

Performance Measures

Set Retrieval Effectiveness: Illustration

Classes:



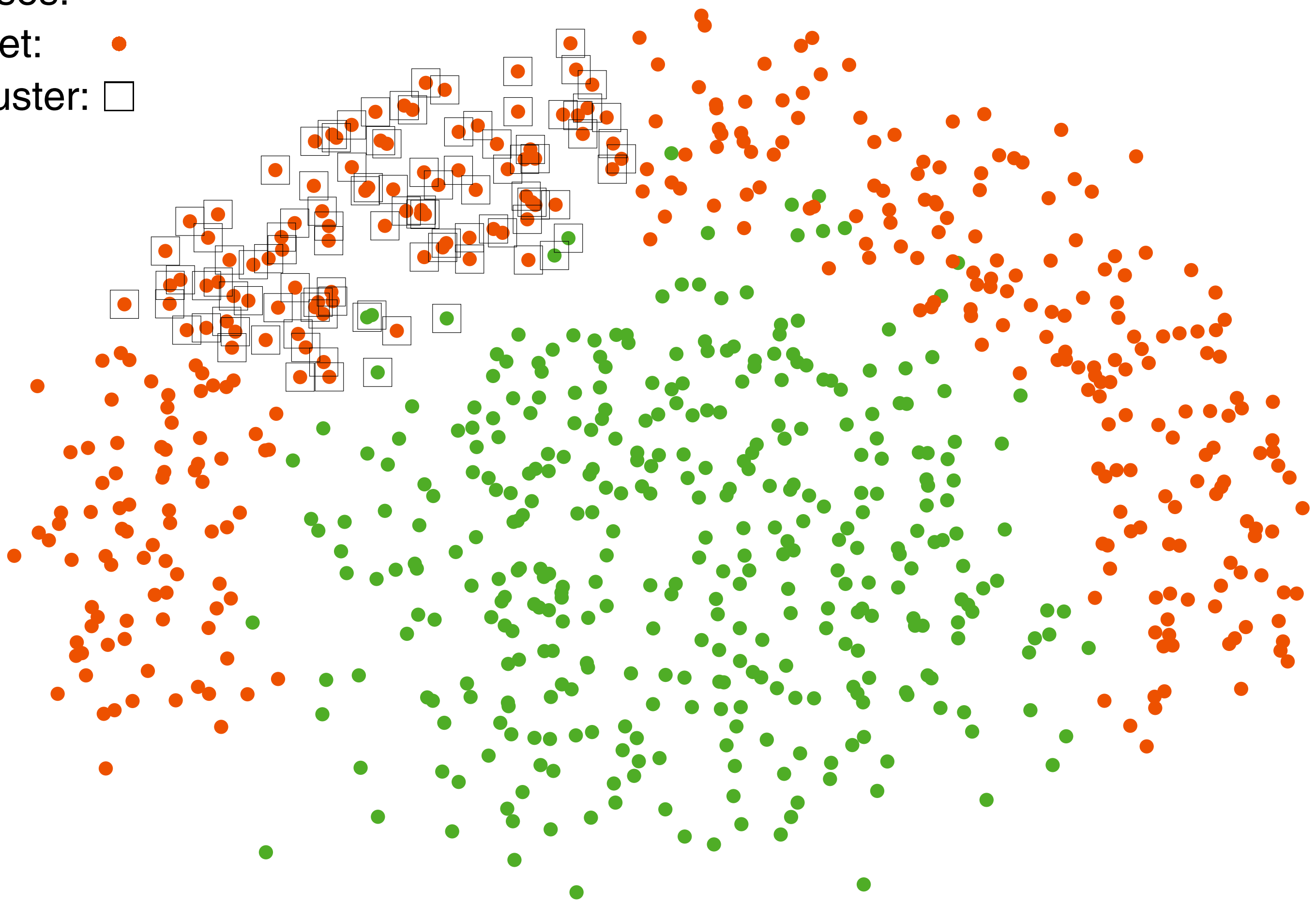
Performance Measures

Set Retrieval Effectiveness: Illustration

Classes:

Target:

In cluster:



Recall / = 0.26 Precision /(∪) = 0.94 F-Measure = 0.40

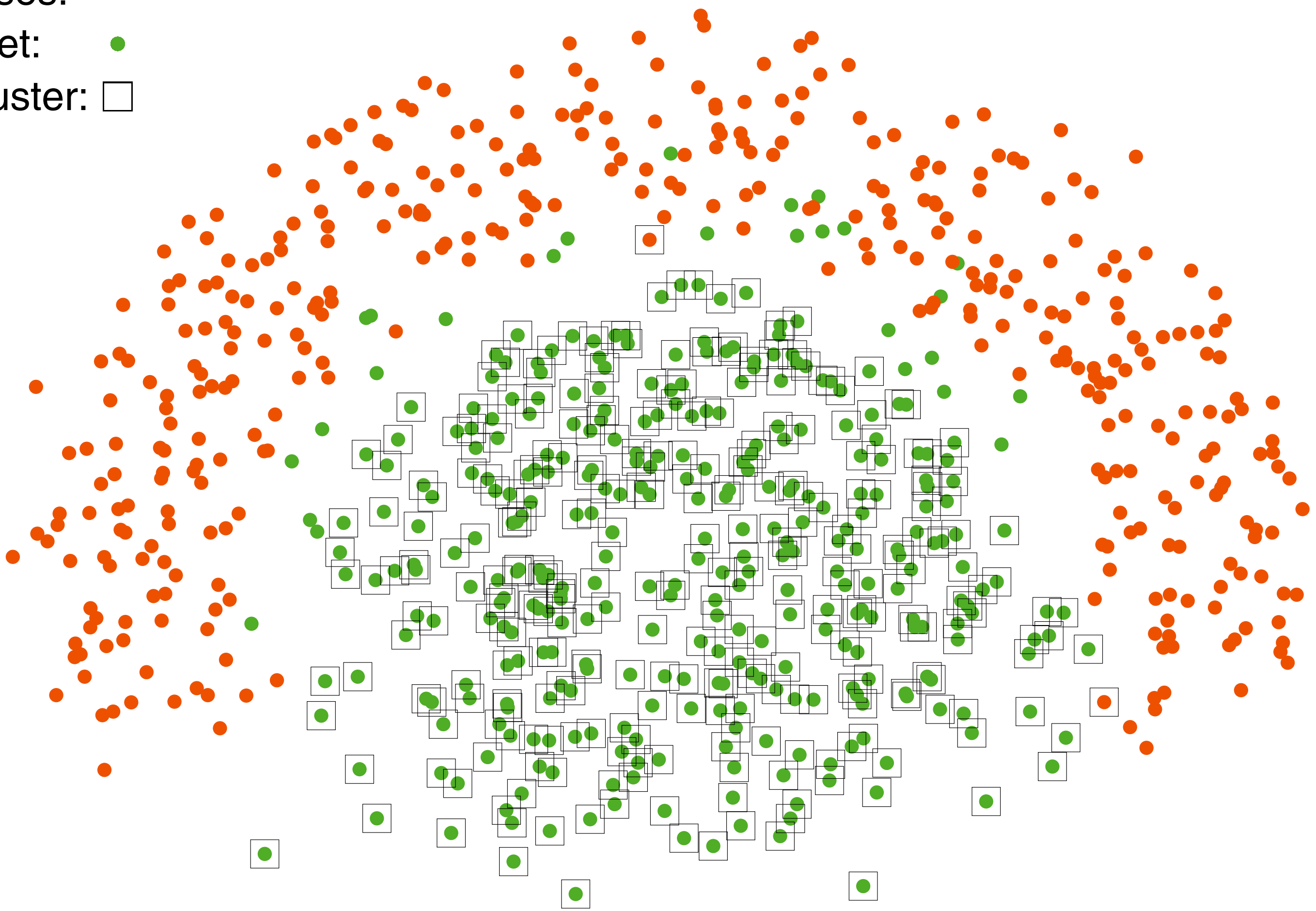
Performance Measures

Set Retrieval Effectiveness: Illustration

Classes: ● ●

Target: ●

In cluster: □



Recall ▣ / ● = 0.92 Precision ▣/(▣ ∪ ▣) = 0.99 F-Measure = 0.95

Performance Measures

Set Retrieval Effectiveness: Precision and Recall Averaging

To obtain a reliable estimate of a search engine's performance, its precision and recall scores must be based on a set of topics Q instead of just one topic q.

Macro-averaging: (user-oriented)

$$\textit{precision}_{\text{macro}} = \frac{1}{|Q|} \sum_{q \in Q} \textit{precision}_q \qquad \textit{recall}_{\text{macro}} = \frac{1}{|Q|} \sum_{q \in Q} \textit{recall}_q$$

Macro-averaging gives equal importance to each topic.

Performance Measures

Set Retrieval Effectiveness: Precision and Recall Averaging

To obtain a reliable estimate of a search engine's performance, its precision and recall scores must be based on a set of topics Q instead of just one topic q.

Macro-averaging: (user-oriented)

$$\textit{precision}_{\text{macro}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{a_q}{a_q + b_q} \qquad \textit{recall}_{\text{macro}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{a_q}{a_q + c_q}$$

Macro-averaging gives equal importance to each topic.

Performance Measures

Set Retrieval Effectiveness: Precision and Recall Averaging

To obtain a reliable estimate of a search engine's performance, its precision and recall scores must be based on a set of topics Q instead of just one topic q.

Macro-averaging: (user-oriented)

$$\textit{precision}_{\text{macro}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{a_q}{a_q + b_q} \qquad \textit{recall}_{\text{macro}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{a_q}{a_q + c_q}$$

Macro-averaging gives equal importance to each topic.

Micro-averaging: (system-oriented)

$$\textit{precision}_{\text{micro}} = \frac{\sum_{q \in Q} a_q}{\sum_{q \in Q} a_q + b_q} \qquad \textit{recall}_{\text{micro}} = \frac{\sum_{q \in Q} a_q}{\sum_{q \in Q} a_q + c_q}$$

In micro-averaging, a topic's importance depends on its number of relevant documents compared to that of other topics.

Remarks:

- Illustration: Consider a university that offers 10 classes, 5 with 1 student each, 5 with 99 students each. The macro-average number of students per class is 50. The micro-average number of students per class is 98.02, since almost all students are in classes with 98 other students. [Salton 1983]

- Macro-averaging is user-oriented in that it ensures that users have a consistently good search experience across topics.

- Micro-averaging is system-oriented in that it allows engineers to focus on topics for which the search engine is capable of finding lots of relevant documents, while mostly neglecting topics whose underlying information need is difficult or expensive to be satisfied. For example, if the majority of users cares only about topics of the former kind, investing the effort to solve the latter properly may not be economical, or may even degrade the search experience for the majority, presuming that the search engine's parameters are set globally.

- Macro-averaging, the user-oriented view, is preferred for most search domains.

Performance Measures

Set Retrieval Effectiveness: Recall Estimation

The set of relevant documents in a large collection usually cannot be obtained with reasonable effort, nor can its size be estimated easily. Heuristic approximations:

- Pooling with or without large-scale relevance judgments

  Problems: Requires at least a number of paradigmatically different search engines that have been tuned by experts. Without relevance judgments, the search engines' results are considered as votes on document relevance.

- Sample analysis

  Problem: Often the number of relevant documents is only a small fraction of the entire document collection. A representative sample would therefore encompass a large portion of documents from the collection.

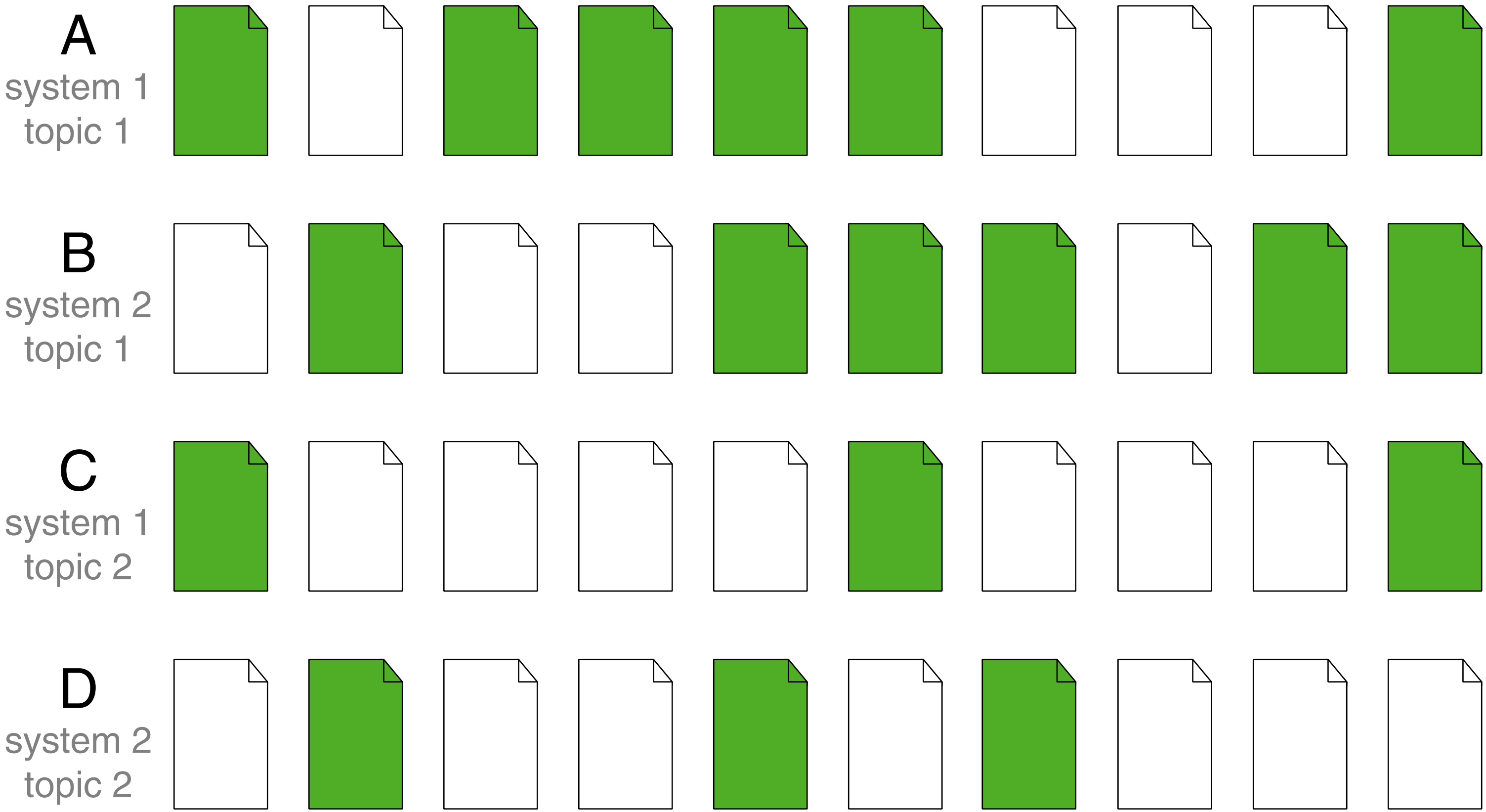
- Query expansion

  The search results for a query of a given topic are judged down to depth k, and then the query is expanded or rephrased, repeating the judgment of new documents found. This may increase the chances of finding more relevant documents.

- Check with external source (e.g., by questioning experts).

Performance Measures

Ranking Effectiveness

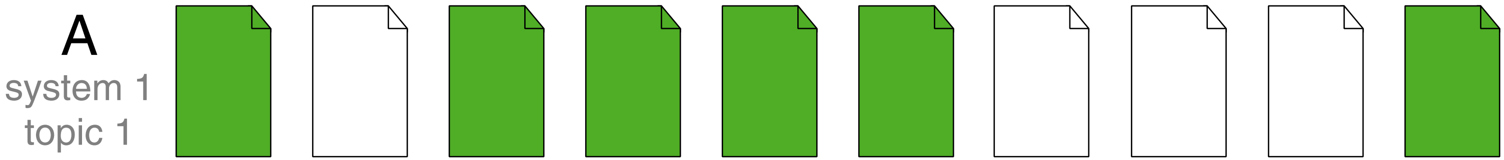


Which ranking is better? Both, Topics 1 and 2, have equal *precision* and *recall*.

How good is System 1 compared to System 2 overall?

Performance Measures

Ranking Effectiveness: Precision@k and Recall@k



Assumption:

- The user browses all documents up to rank $k \geq 1$.

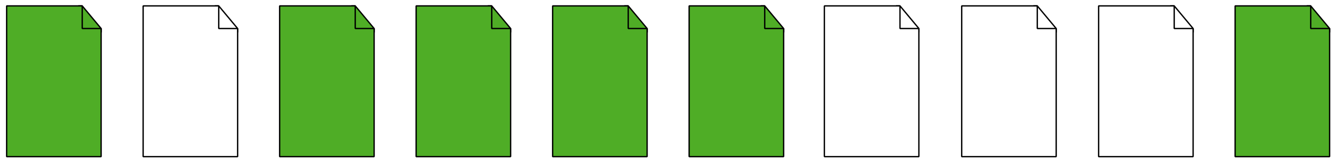
➜ Compute *precision* and *recall* at a rank k.

- Commonly used ranks are $k \in \{1, 5, 10, 20\}$.

Performance Measures

Ranking Effectiveness: Precision@k and Recall@k

A
system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00

Assumption:

- The user browses all documents up to rank $k \geq 1$.

➜ Compute *precision* and *recall* at a rank k.

- Commonly used ranks are $k \in \{1, 5, 10, 20\}$.

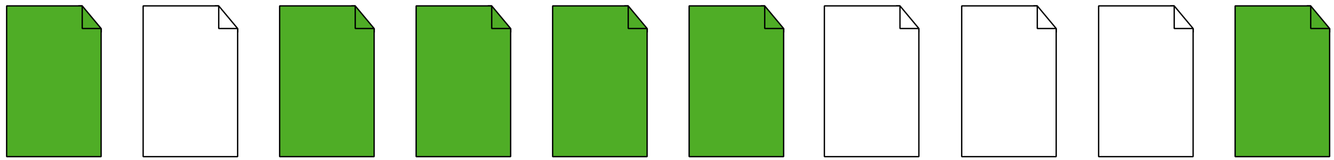
Caveats:

- Disregards ranking differences up to rank k.
- Disregards the (estimated) number of relevant documents (e.g., $\ll k$).
- Based on binary relevance judgments.

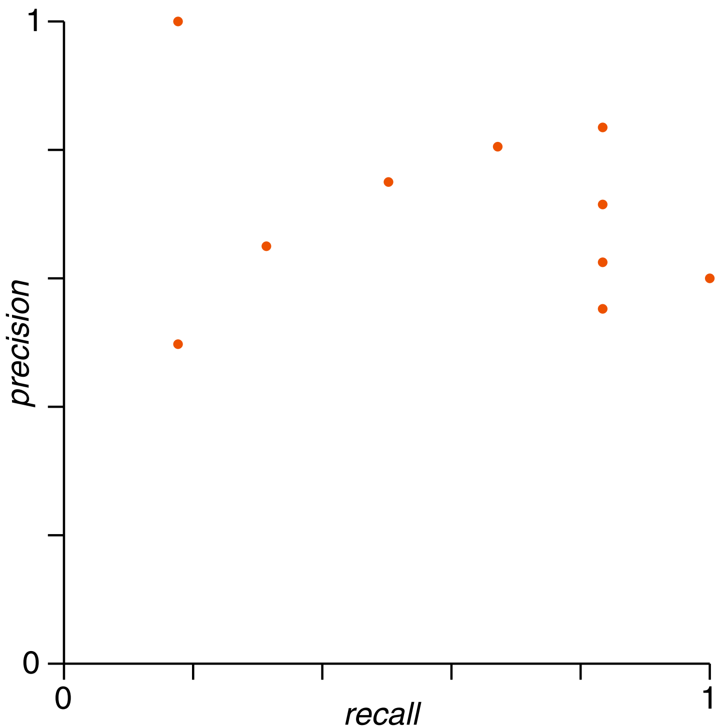
Performance Measures

Ranking Effectiveness: Precision-Recall Curves

A
system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00



Observations:

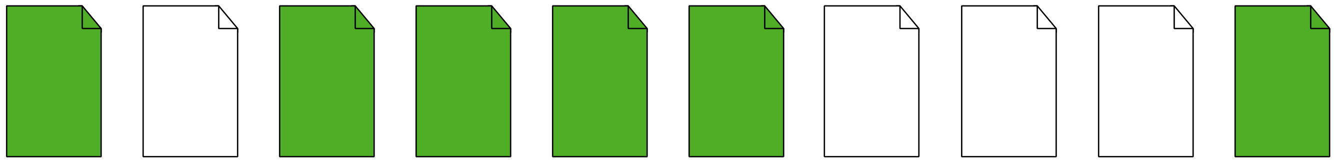
- Connecting the dots yields a “curve.”
- The curve captures detailed ranking characteristics: the user experience.
- Points on a curve other than the original ones lack interpretation.
- Given rankings from two systems, we can decide which one is better.

➜ These observations can be quantified as area under curve.

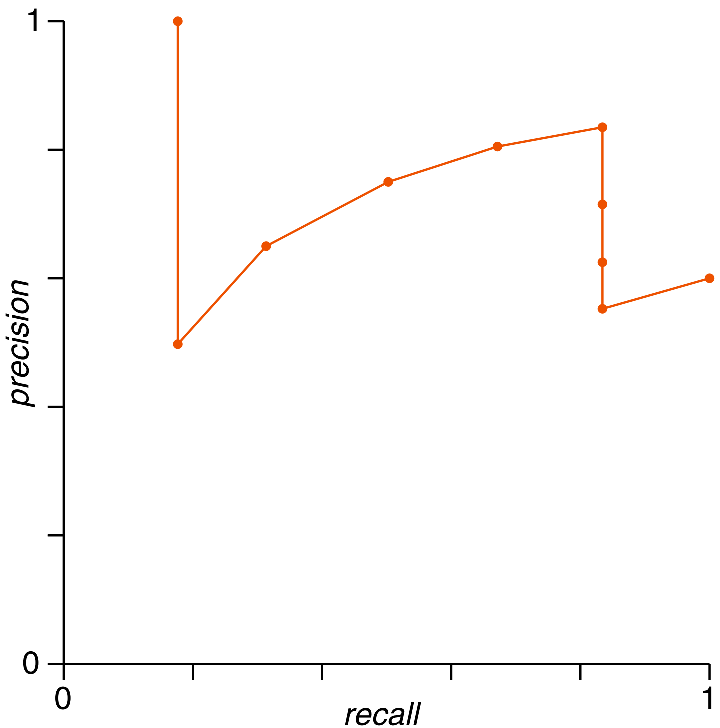
Performance Measures

Ranking Effectiveness: Precision-Recall Curves

A
system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00



Observations:

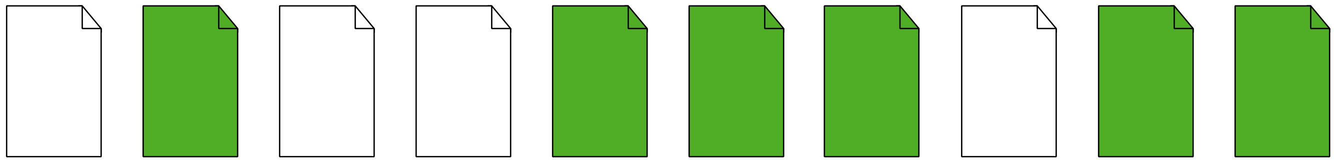
- Connecting the dots yields a “curve.”
- The curve captures detailed ranking characteristics: the user experience.
- Points on a curve other than the original ones lack interpretation.
- Given rankings from two systems, we can decide which one is better.

→ These observations can be quantified as area under curve.

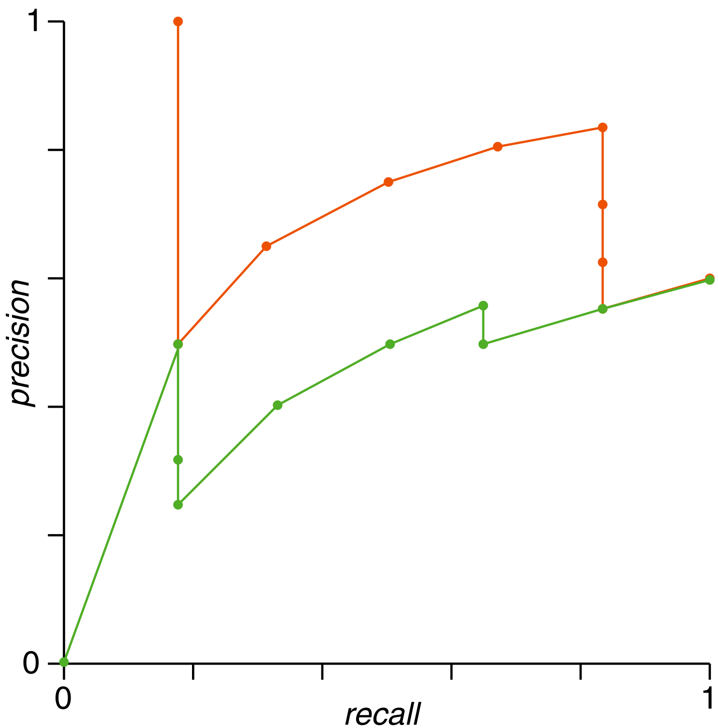
Performance Measures

Ranking Effectiveness: Precision-Recall Curves

B
system 2
topic 1



precision	0.00	0.50	0.33	0.25	0.40	0.50	0.57	0.50	0.56	0.60
recall	0.00	0.17	0.17	0.17	0.33	0.50	0.67	0.67	0.83	1.00



Observations:

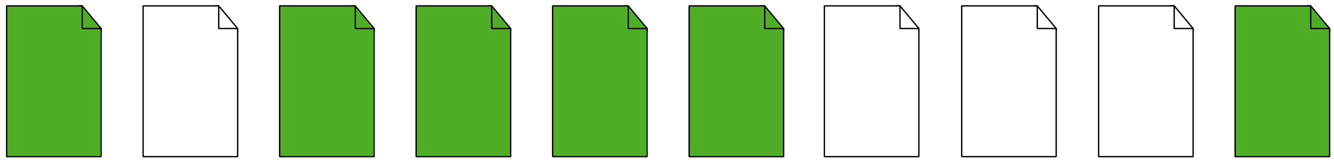
- Connecting the dots yields a "curve."
- The curve captures detailed ranking characteristics: the user experience.
- Points on a curve other than the original ones lack interpretation.
- Given rankings from two systems, we can decide which one is better.

→ These observations can be quantified as area under curve.

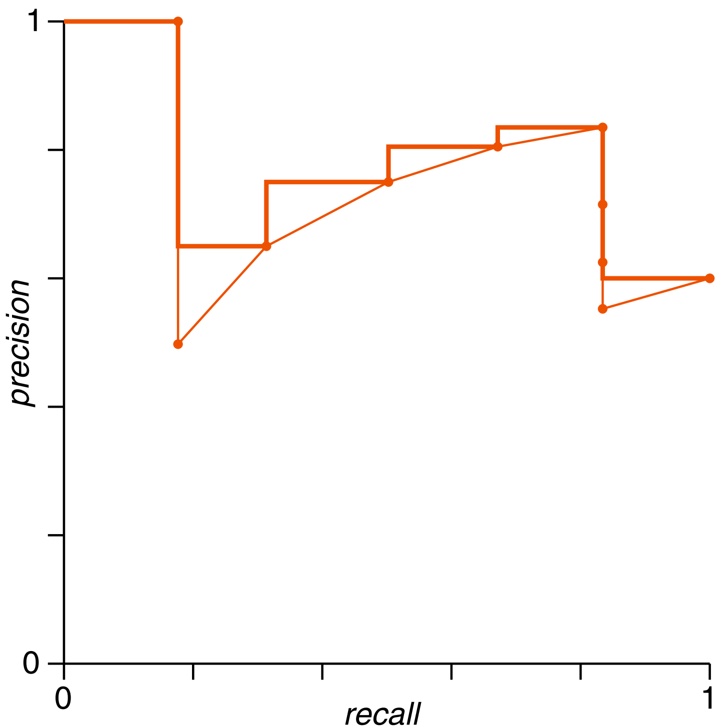
Performance Measures

Ranking Effectiveness: Average Precision

A
system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00



Average precision approximates the area under the precision-recall curve.

Interpolation alternatives:

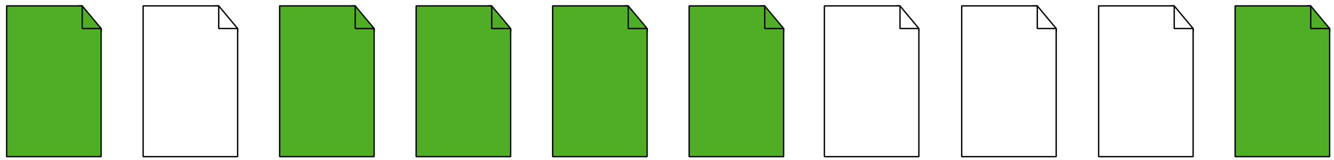
1. Integral of the step function visiting the maximum precision at every recall point.

2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

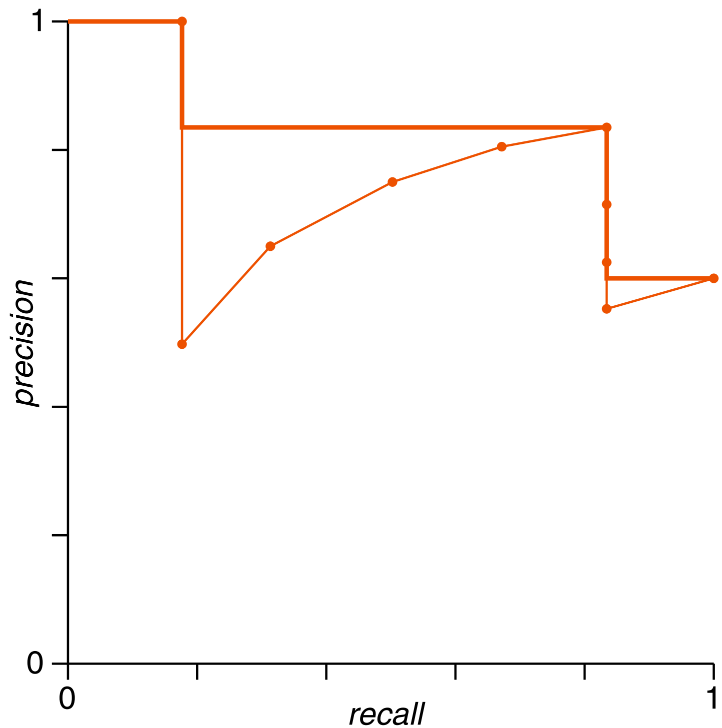
Performance Measures

Ranking Effectiveness: Average Precision

A
system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00



Average precision approximates the area under the precision-recall curve.

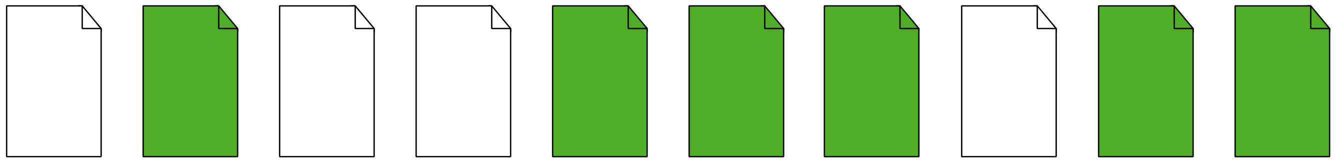
Interpolation alternatives:

1. Integral of the step function visiting the maximum precision at every recall point.
2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

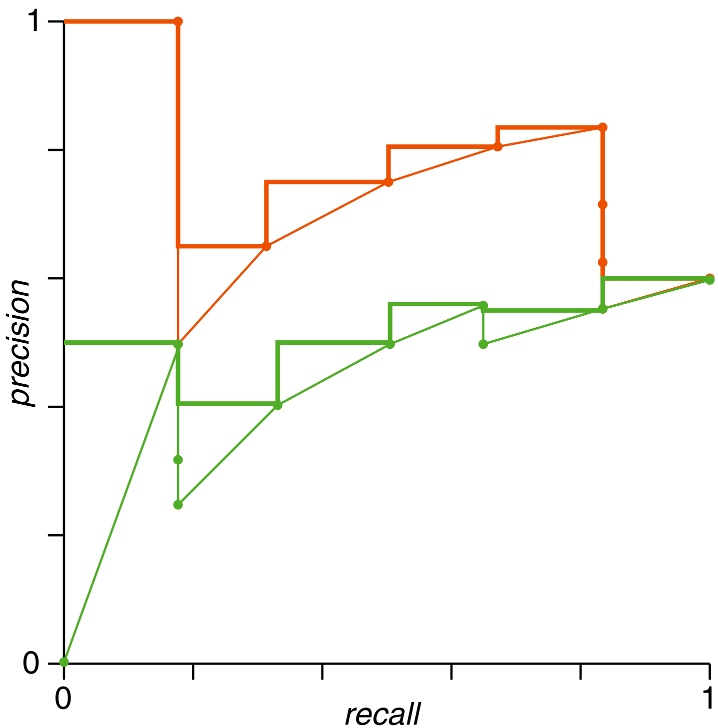
Performance Measures

Ranking Effectiveness: Average Precision

B
system 2
topic 1



precision	0.00	0.50	0.33	0.25	0.40	0.50	0.57	0.50	0.56	0.60
recall	0.00	0.17	0.17	0.17	0.33	0.50	0.67	0.67	0.83	1.00



Average precision approximates the area under the precision-recall curve.

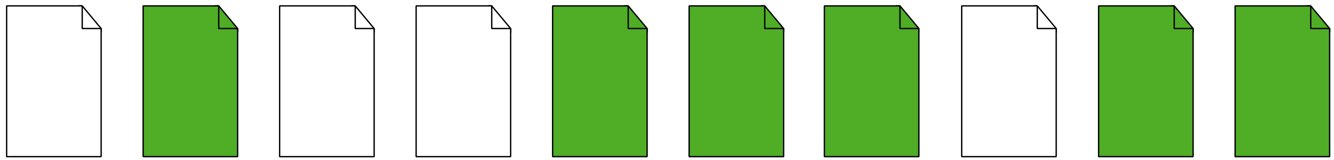
Interpolation alternatives:

1. Integral of the step function visiting the maximum precision at every recall point.
2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

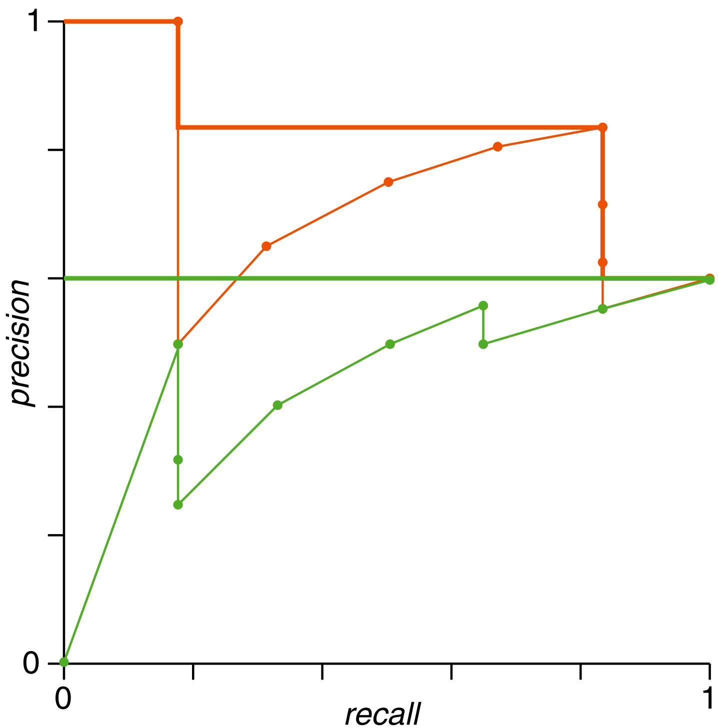
Performance Measures

Ranking Effectiveness: Average Precision

B
system 2
topic 1



precision	0.00	0.50	0.33	0.25	0.40	0.50	0.57	0.50	0.56	0.60
recall	0.00	0.17	0.17	0.17	0.33	0.50	0.67	0.67	0.83	1.00



Average precision approximates the area under the precision-recall curve.

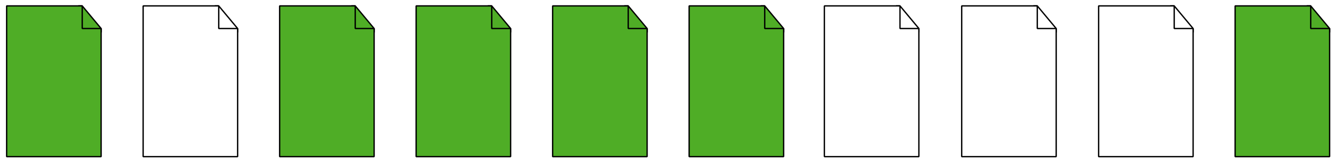
Interpolation alternatives:

1. Integral of the step function visiting the maximum precision at every recall point.
2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

Performance Measures

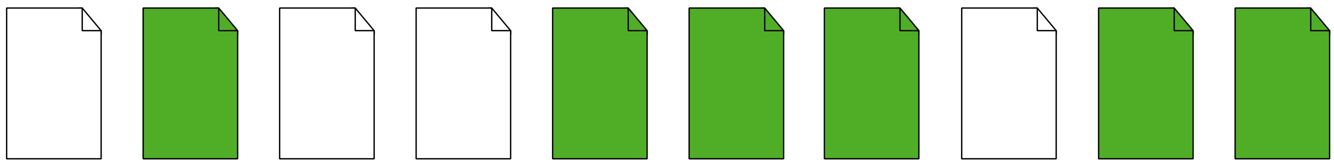
Ranking Effectiveness: Average Precision (Alternative 1)

A
system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00

B
system 2
topic 1



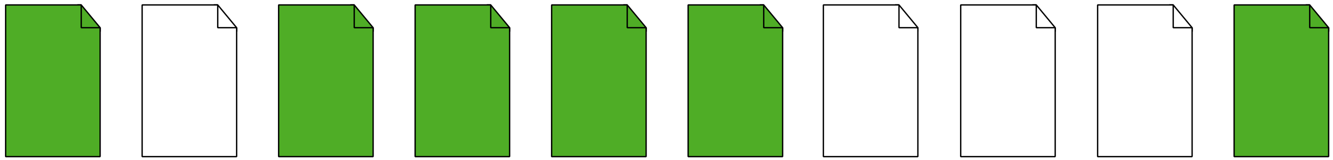
precision	0.00	0.50	0.33	0.25	0.40	0.50	0.57	0.50	0.56	0.60
recall	0.00	0.17	0.17	0.17	0.33	0.50	0.67	0.67	0.83	1.00

- Sum of Precision@k at ranks with relevant documents, divided by the expected number of relevant documents.
- Ranking A: $(1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$
  Ranking B: $(0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$
- If a relevant document is not found, it gets 0.0 precision.

Performance Measures

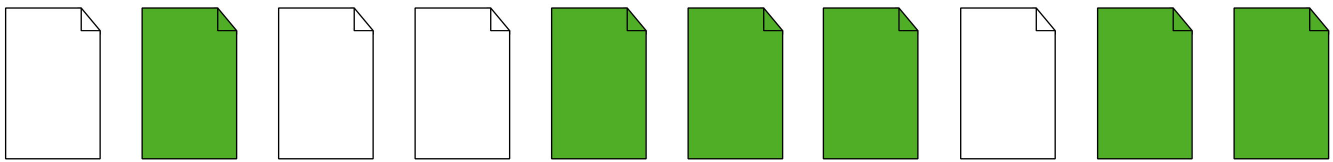
Ranking Effectiveness: Average Precision (Alternative 2)

A
system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00

B
system 2
topic 1



precision	0.00	0.50	0.33	0.25	0.40	0.50	0.57	0.50	0.56	0.60
recall	0.00	0.17	0.17	0.17	0.33	0.50	0.67	0.67	0.83	1.00

- Average of interpolated precision values at 11 recall points: $0, 0.1, \ldots, 0.9, 1$.
- Ranking A: $(2 \cdot 1.0 + 7 \cdot 0.83 + 2 \cdot 0.6)/11 = 0.82$
  Ranking B: $(11 \cdot 0.6)/11 = 0.6$
- Also called: Eleven-Point Interpolated Average Precision

Performance Measures

Ranking Effectiveness: Average Precision

Let $R = (d_1, \ldots, d_{|D|})$ denote a ranking of the documents D for a given query $q \in Q$ according to a search engine.

Let $r : Q \times D \to \{0, 1\}$ denote the relevance function which maps pairs of queries and documents to a Boolean value indicating the latter's relevance to the former.

Then the two alternatives of average precision are computed as follows:

$$\textit{AP}_1@k(q, R) = \frac{1}{\min(k, \sum_{d \in D} r(q, d))} \cdot \sum_{i=1}^{k} \Big(r(q, d_i) \cdot \textit{precision}@i(R) \Big)$$

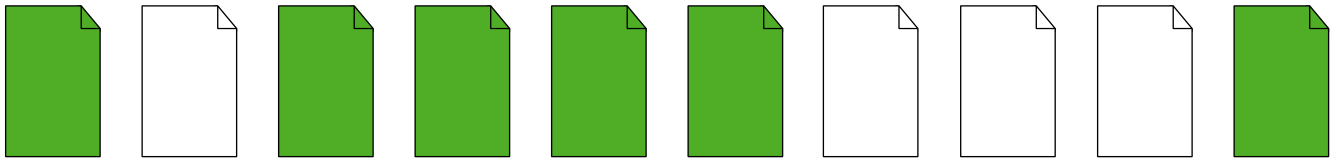
$$\textit{AP}_2(q, R) = \frac{1}{11} \cdot \sum_{i \in \{0, 0.1, \ldots, 1\}} \Big(\max_{j:\ \textit{recall}@j(R) \geq i} \textit{precision}@j(R) \Big)$$

Performance Measures

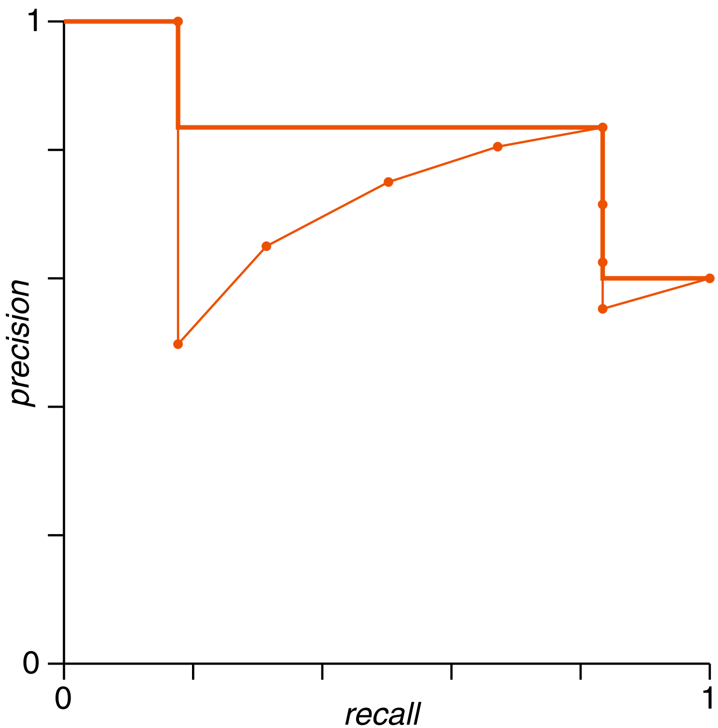
Ranking Effectiveness: Averaging Precision-Recall Curves

A

system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00



Problem:

- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

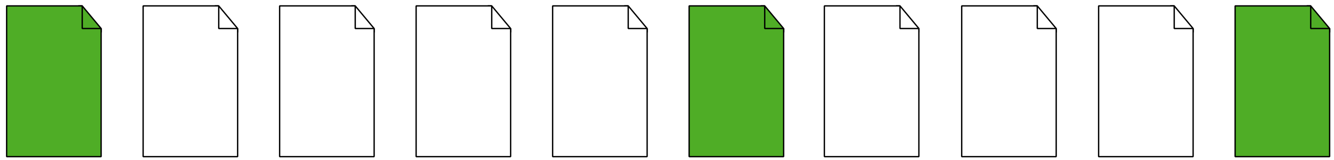
Solution:

- Compute averages across 11 recall points at 0.1 steps.

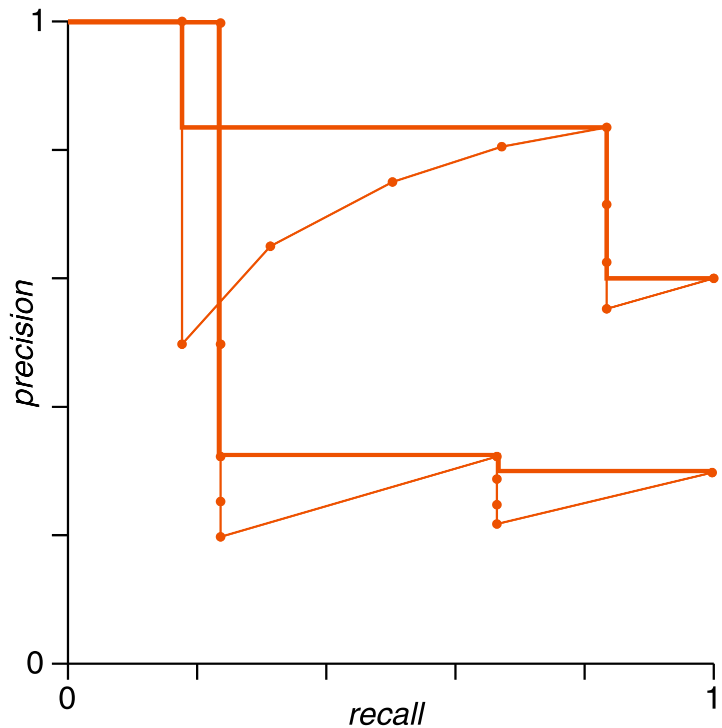
Performance Measures

Ranking Effectiveness: Averaging Precision-Recall Curves

C
system 1
topic 2



precision	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
recall	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00



Problem:

- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

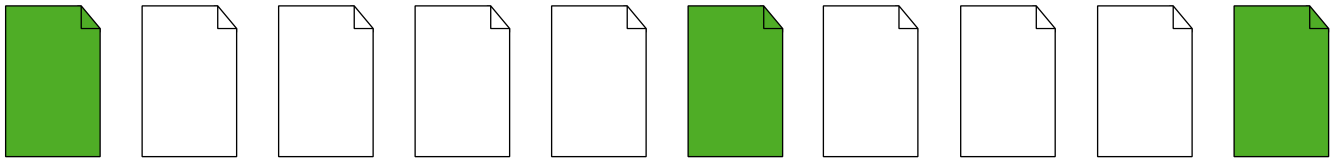
Solution:

- Compute averages across 11 recall points at 0.1 steps.

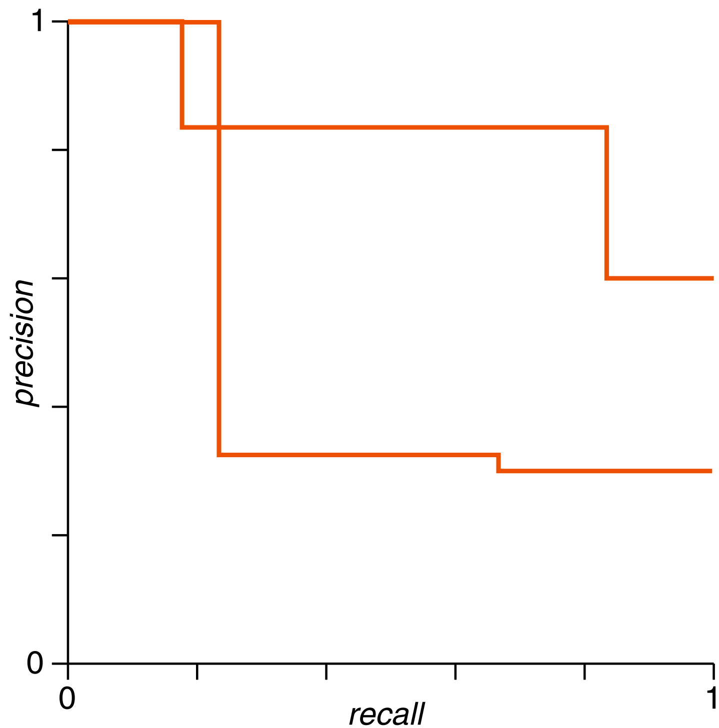
Performance Measures

Ranking Effectiveness: Averaging Precision-Recall Curves

C
system 1
topic 2



precision	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
recall	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00



Problem:

- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

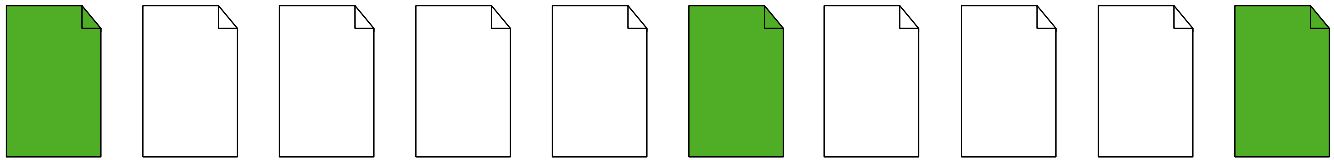
Solution:

- Compute averages across 11 recall points at 0.1 steps.

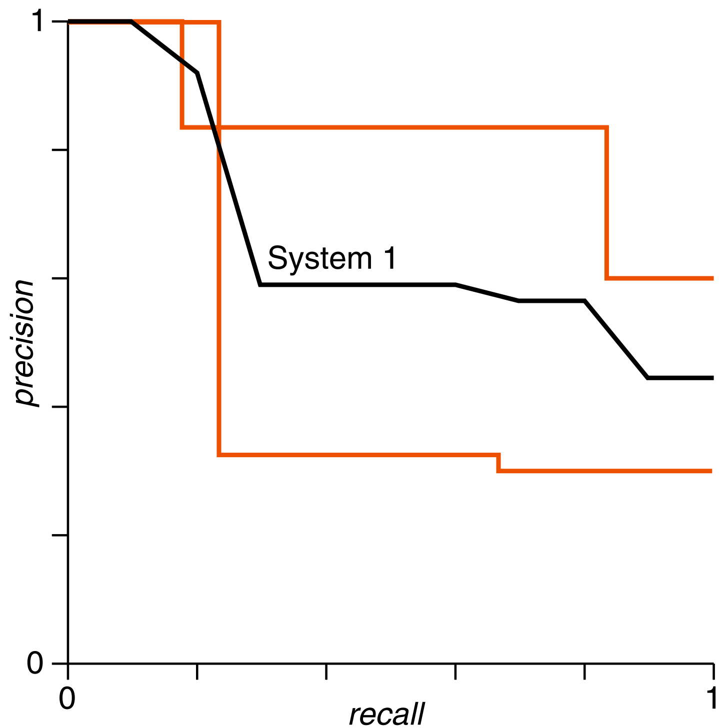
Performance Measures

Ranking Effectiveness: Averaging Precision-Recall Curves

C
system 1
topic 2



precision	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
recall	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00



Problem:

- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

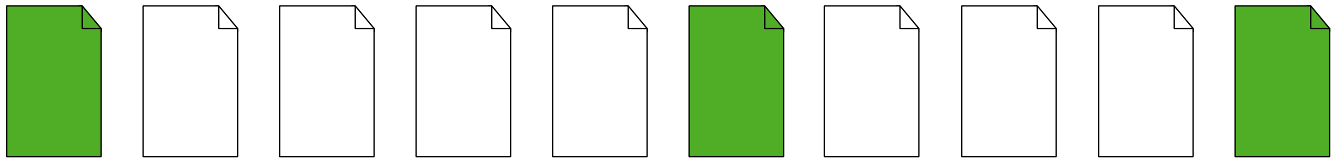
Solution:

- Compute averages across 11 recall points at 0.1 steps.

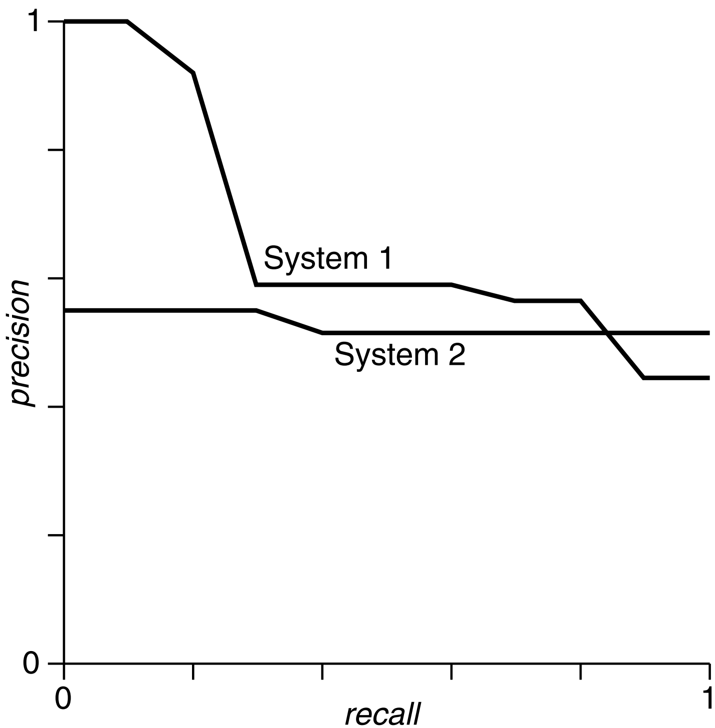
Performance Measures

Ranking Effectiveness: Averaging Precision-Recall Curves

C
system 1
topic 2



precision	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
recall	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00



Interpretation:

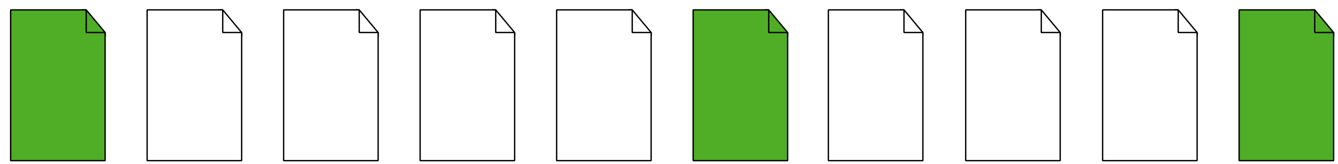
- Judging a system at various operating points.
- System 1 delivers very good precision at high ranks.
- System 2 delivers slightly better recall across rankings.
- Neither system dominates the other.

Curves are a lot smoother for 50 topics.

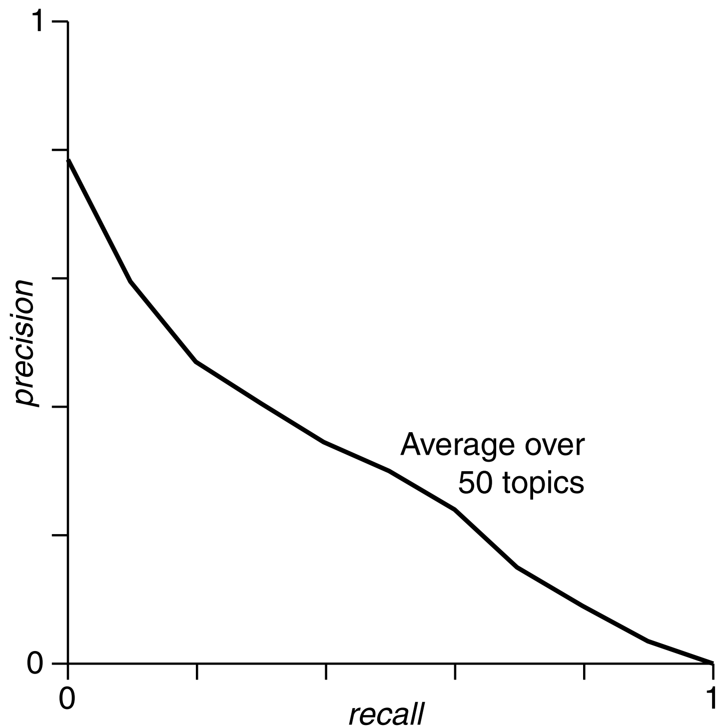
Performance Measures

Ranking Effectiveness: Averaging Precision-Recall Curves

C
system 1
topic 2



precision	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
recall	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00



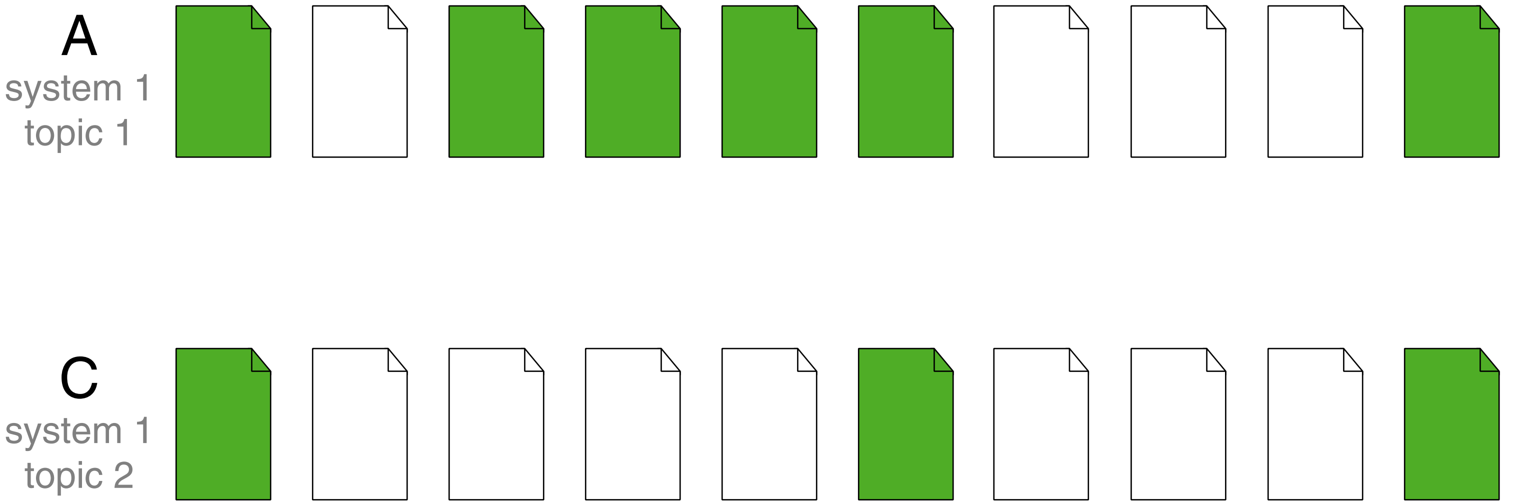
Interpretation:

- Judging a system at various operating points.
- System 1 delivers very good precision at high ranks.
- System 2 delivers slightly better recall across rankings.
- Neither system dominates the other.

Curves are a lot smoother for 50 topics.

Performance Measures

Ranking Effectiveness: Mean Average Precision (MAP)

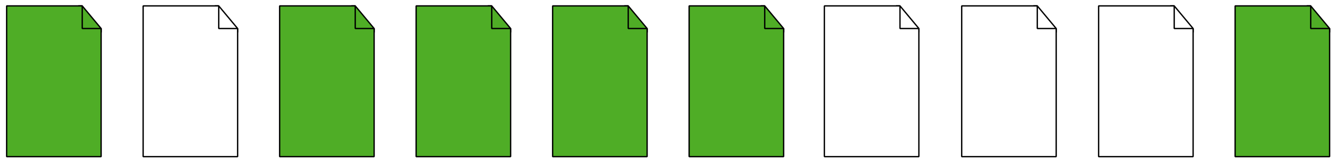


- Meaningful system evaluation requires many topics.

Performance Measures

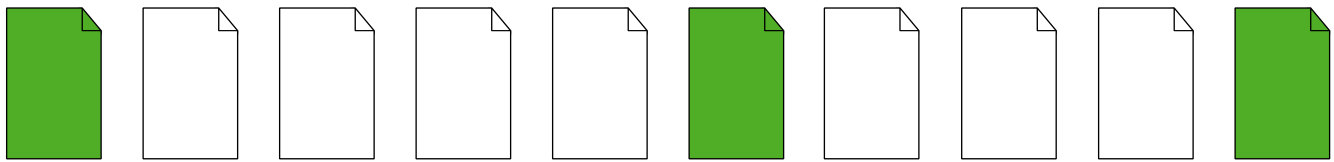
Ranking Effectiveness: Mean Average Precision (MAP)

A
system 1
topic 1



precision	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
recall	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00

C
system 1
topic 2



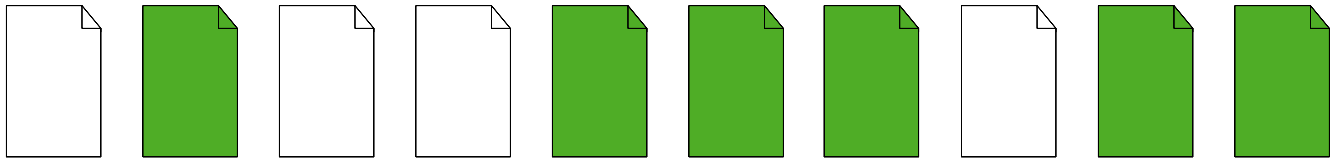
precision	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
recall	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00

- Meaningful system evaluation requires many topics.
- Averaging average precision over topics gives us mean average precision.
- The MAP for System 1, Rankings A and C is $(0.78 + 0.44)/2 = 0.66$.
  (A: $(1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$ and C: $(1.0 + 0.33 + 0.3)/3 = 0.54$)

Performance Measures

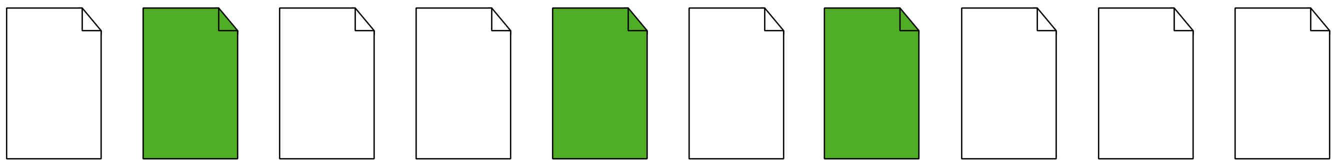
Ranking Effectiveness: Mean Average Precision (MAP)

B
system 2
topic 1



precision	0.00	0.50	0.33	0.25	0.40	0.50	0.57	0.50	0.56	0.60
recall	0.00	0.17	0.17	0.17	0.33	0.50	0.67	0.67	0.83	1.00

D
system 2
topic 2



precision	0.00	0.50	0.33	0.25	0.40	0.33	0.43	0.38	0.33	0.30
recall	0.00	0.33	0.33	0.33	0.67	0.67	1.00	1.00	1.00	1.00

- Meaningful system evaluation requires many topics.
- Averaging average precision over topics gives us mean average precision.
- The MAP for System 1, Rankings A and C is $(0.78 + 0.44)/2 = 0.66$.
- The MAP for System 2, Rankings B and D is $(0.52 + 0.44)/2 = 0.48$.
  (B: $(0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$ and D: $(0.5 + 0.4 + 0.43)/3 = 0.44$)

Performance Measures

Ranking Effectiveness: Mean Average Precision (MAP)

Is (mean) average precision a good measure?

User model: [Robertson 2008]

1. The user stops browsing only after a relevant document.
2. The probability of stopping is the same for all relevant documents.

Problems:

- Assumption 1 is true in some applications.
  But the user does not know which is the last relevant document. Users who do not decide to stop browsing at the last relevant document are doomed to explore the entire ranking.
- Assumption 2 is unrealistic: Most users will stop earlier rather than later.

Solution:

- Assume users decide to stop with increasing probability at any given rank.

➜ (Normalized) Discounted Cumulative Gain (nDCG)

Performance Measures

Ranking Effectiveness: Mean Reciprocal Rank (MRR)

User model:

- The user stops browsing at the first relevant document encountered.

The rank of the first relevant document determines the quality of a ranking:

$$\text{RR} = \frac{1}{r},$$

where r denotes the rank of the relevant document. The mean reciprocal rank (MRR) is the average of the reciprocal rank across many topics.

Example:

Rank	1	2	3	4	5	6	7	8	9	10
Reciprocal rank	1	0.50	0.33	0.25	0.20	0.17	0.14	0.13	0.11	0.10

MRR has several caveats, rendering it untrustworthy in practice.

Remarks: [Fuhr 2017]

- MRR scores form an ordinal scale, not an interval scale. This is evidenced by the fact that the distance between first and second rank is as large as that between second rank and the infinite rank. For ordinal scales, averages cannot be computed, but only medians. Using the median, however, would yield many ties, which defeats the purpose of comparing system performance.

- MRR can produce unintuitive scores: For three topics, System 1 achieves $r_1 = 1$, $r_2 = 2$, and $r_3 = 4$, whereas System 2 achieves $r_1 = r_2 = r_3 = 2$. System 1 has an MRR of $1/3(1/1 + 1/2 + 1/4) = 0.58$, and System 2 has an MRR of $1/3(3 \cdot 1/2) = 0.5$. Compared to the average ranks of relevant documents, where System 1 has 2.3 and System 2 has 2, this is contradictory.

Performance Measures

Ranking Effectiveness: Discounted Cumulative Gain (DCG)

User model:

- Every document has a gain when read by the user.

  Gain is operationalized in terms of graded relevance assessment: $r : D \times Q \rightarrow \{0, 1, 2, 3, 4, 5\}$, where 0 indicates no relevance, and 5 top relevance.

- While browsing the ranking, the gain cumulates.

  Gain cumulation is computed similar to $\sum_i^k r(d_i, q)$, where k denotes a rank, d_i denotes the document $d \in D$ at rank i, and q denotes the query.

- The lower a document is ranked, the less likely it is examined; its gain must be discounted.

  For this, a variant of the reciprocal rank measure is used.

Altogether, the discounted cumulative gain measure is defined as follows:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{r(d_i,q)} - 1}{\log_2(1+i)},$$

where k is the maximum rank, the logarithm ensures smooth reduction, and $2^{r(d_i,q)}$ emphasizes highly relevant documents.

Performance Measures

Ranking Effectiveness: Normalized Discounted Cumulative Gain (nDCG)

DCG values are normalized with DCG* scores obtained for an ideal ranking, sorting the documents in the ranking by decreased relevance grades.

This yields the normalized discounted cumulative gain measure.

Example:

Rank k	1	2	3	4	5	6	7	8	9	10
Gain $r(d_i, q)$	3	2	3	0	0	1	2	2	3	0
$DCG@k$	7.00	8.89	12.39	12.39	12.39	12.75	13.75	14.70	16.80	16.80
Ideal $r^*(d_i, q)$	3	3	3	2	2	2	1	0	0	0
$DCG^*@k$	7.00	11.42	14.92	16.21	17.37	18.44	18.77	18.77	18.77	18.77
$nDCG@k$	1.00	0.78	0.83	0.76	0.71	0.69	0.73	0.78	0.90	0.90

Remarks:

- When comparing more than one system, the ideal ranking is usually formed by the joint relevance assessments for all systems.

