Chapter IR:VIII

VIII. 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.

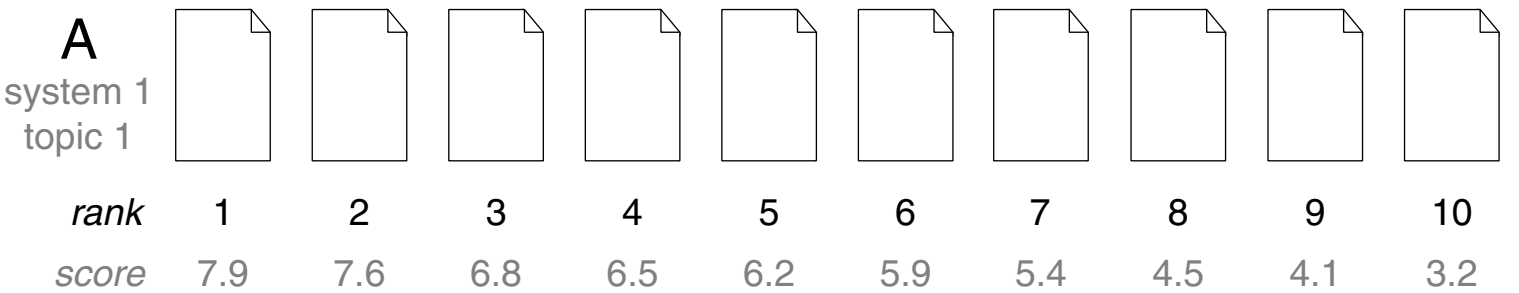
Sufficient justification is achieved by means of measurement, namely "the assignment of a number to a characteristic of an object [a search result], which can be compared with other objects." [Wikipedia]

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:

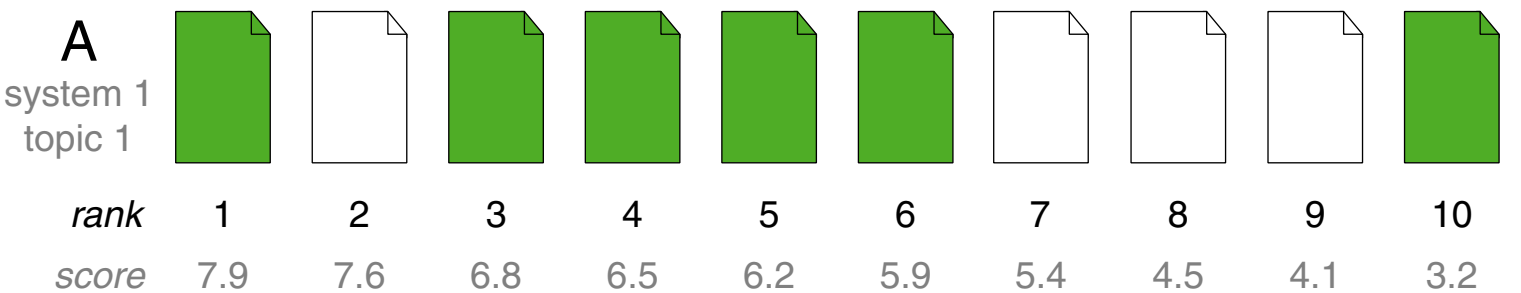


A search result is composed of a list of documents, ordered by the search engine's estimation of relevance, optionally alongside relevance scores for each document.

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.

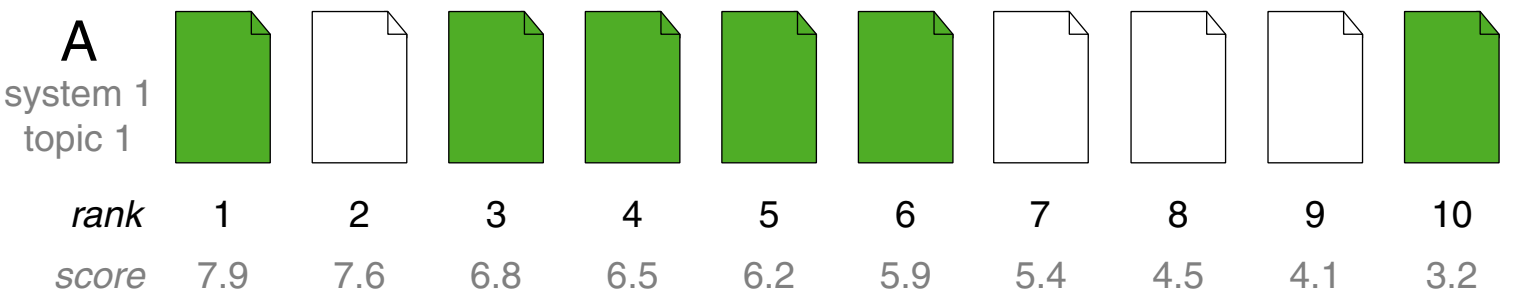
An effectiveness measure maps a given search result and its relevance judgments to the real numbers, rendering rankings from different systems comparable.

The mapping encodes a model of user behavior. Recent measures are based on realistic models; early measures less so, having been defined based on intuition.

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.

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.
- The measures are typically applied in classification scenarios.
- 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

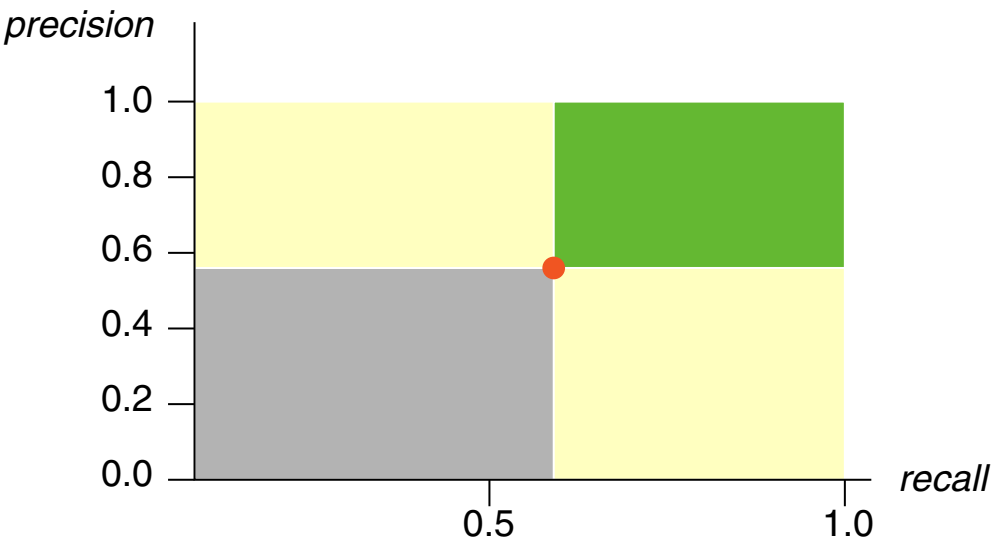
  $$\textit{fallout} = \frac{b}{b + d}$$

  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

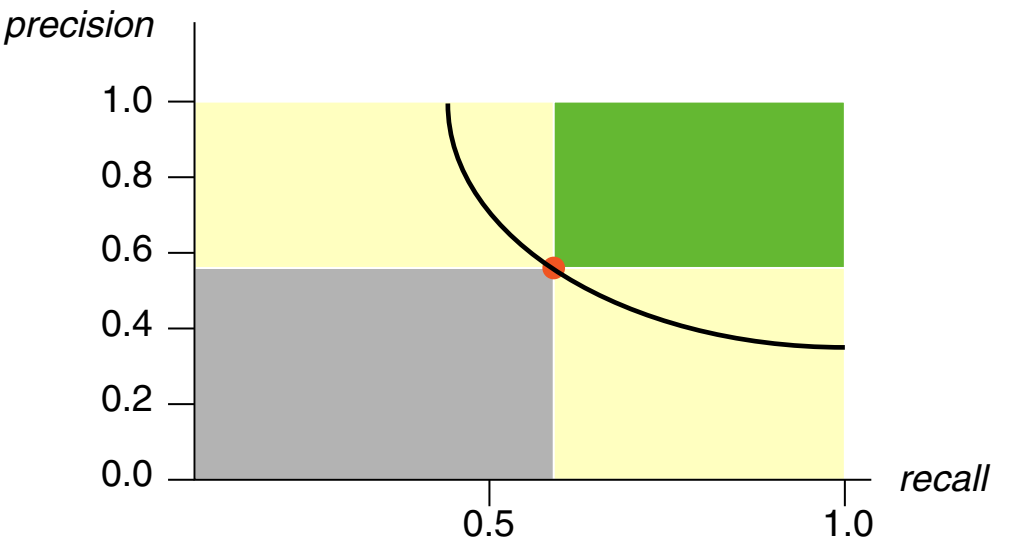
Comparison of retrieval systems:



Performance Measures

Set Retrieval Effectiveness: F-Measure

Comparison of retrieval systems:



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

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

The scores of *precision*, *recall* and F-Measure are in the interval $[0; 1]$.

Remarks:

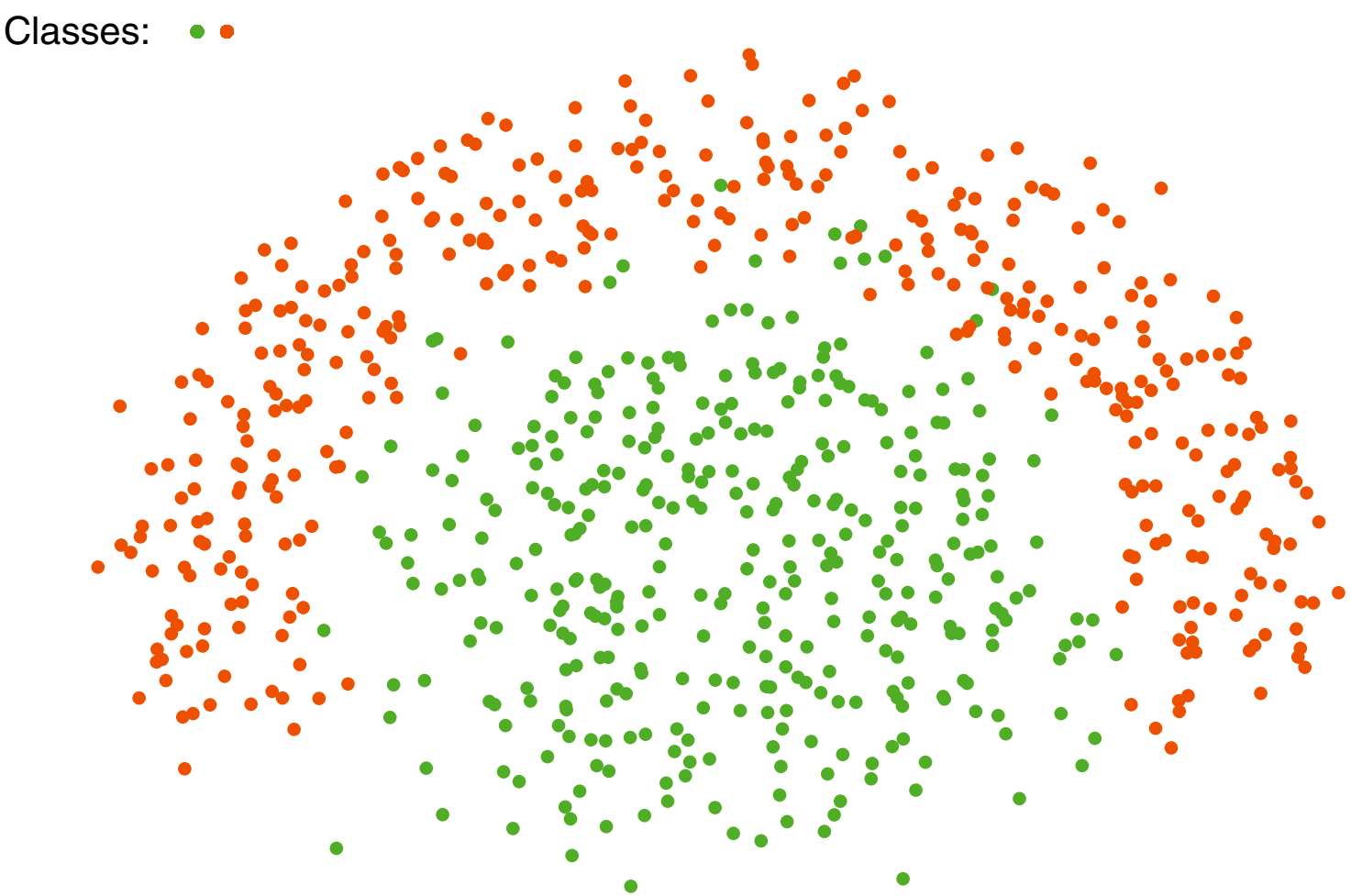
- Precision and recall induce a partial ordering of retrieval systems: systems that perform better in one, but worse in the other measure cannot be ranked with regard to which one is better. The F-Measure calculates a single performance score from precision and recall, inducing a total order.
- The harmonic 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:

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

  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.

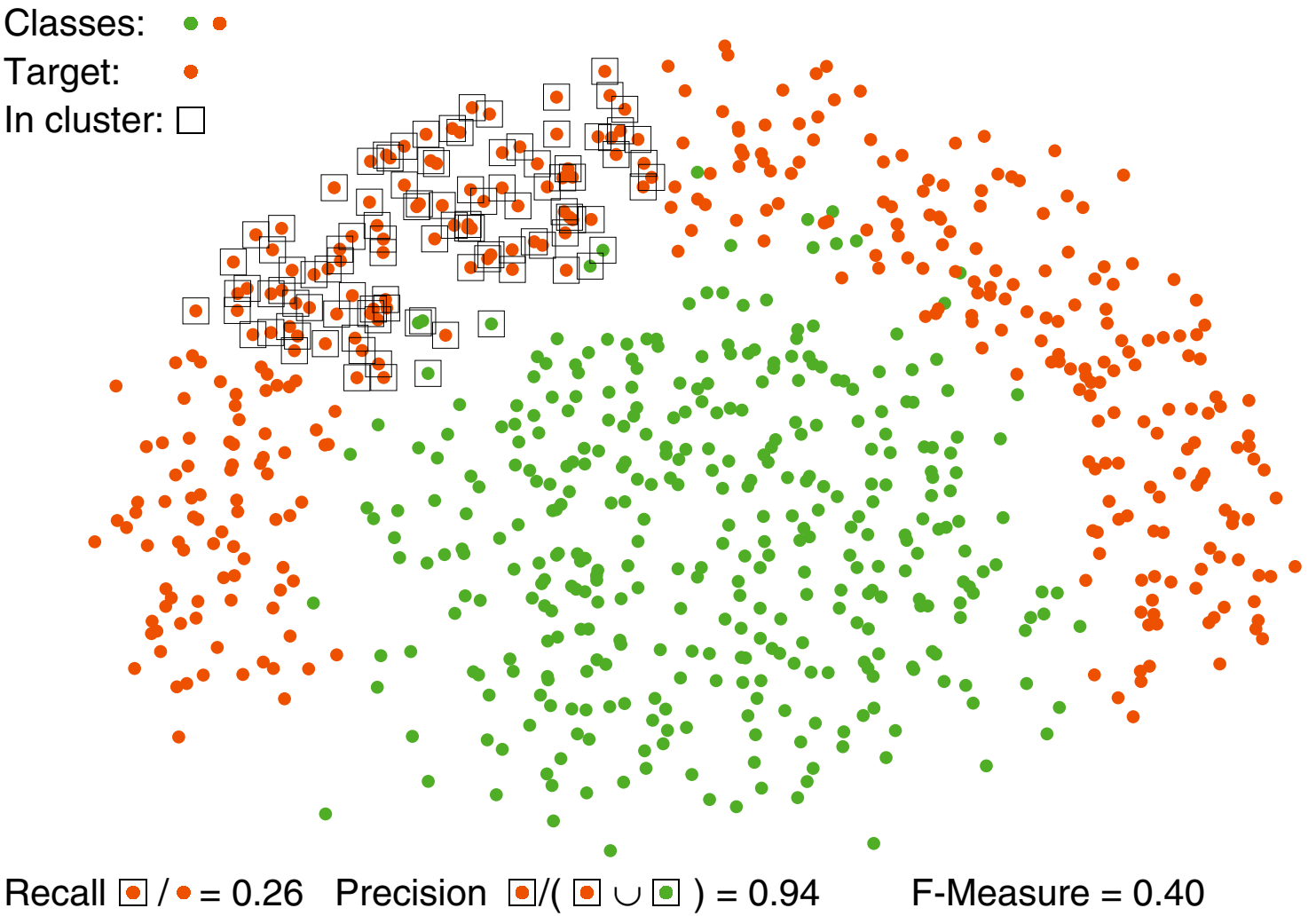
Performance Measures

Set Retrieval Effectiveness: Illustration



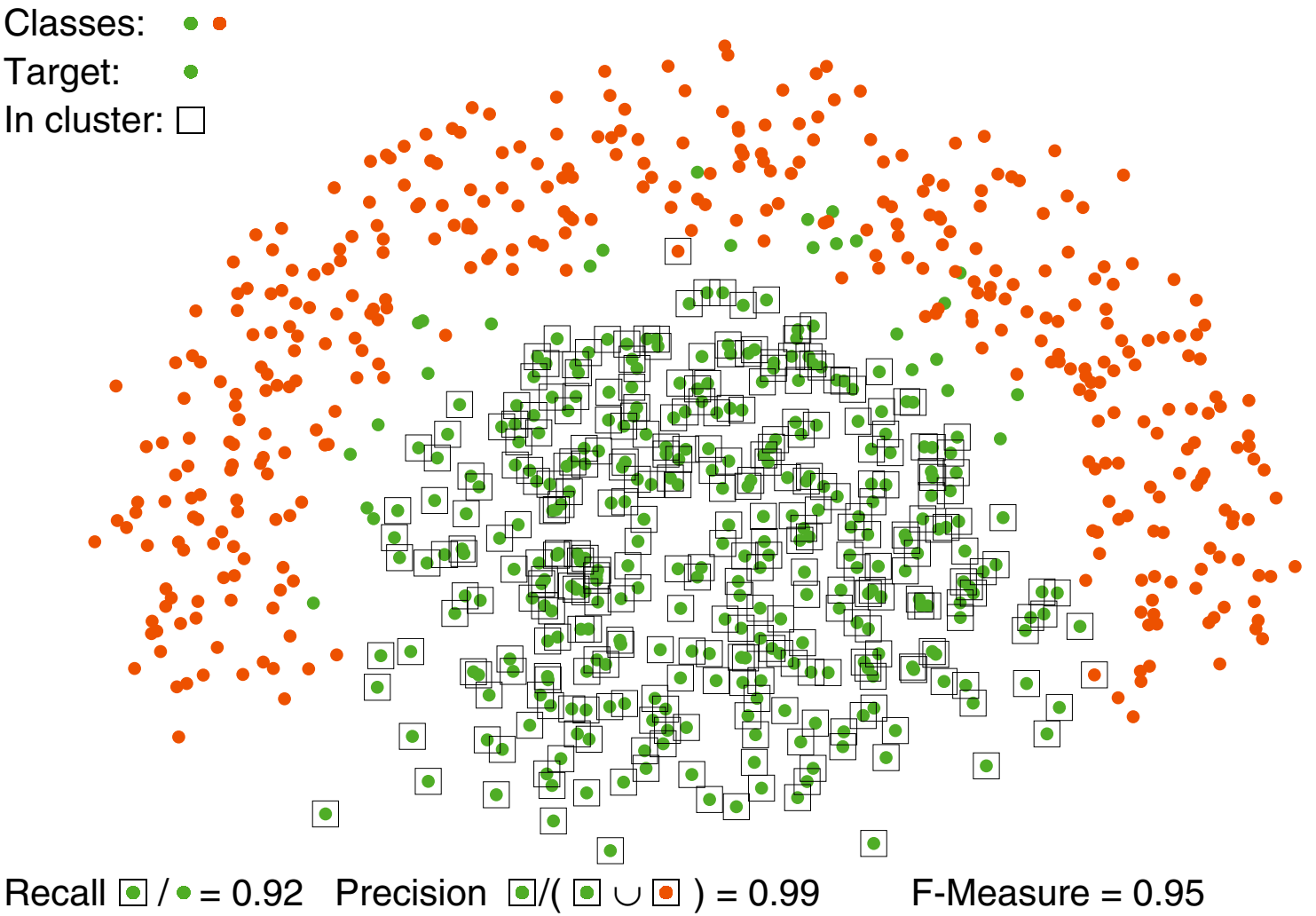
Performance Measures

Set Retrieval Effectiveness: Illustration



Performance Measures

Set Retrieval Effectiveness: Illustration



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.

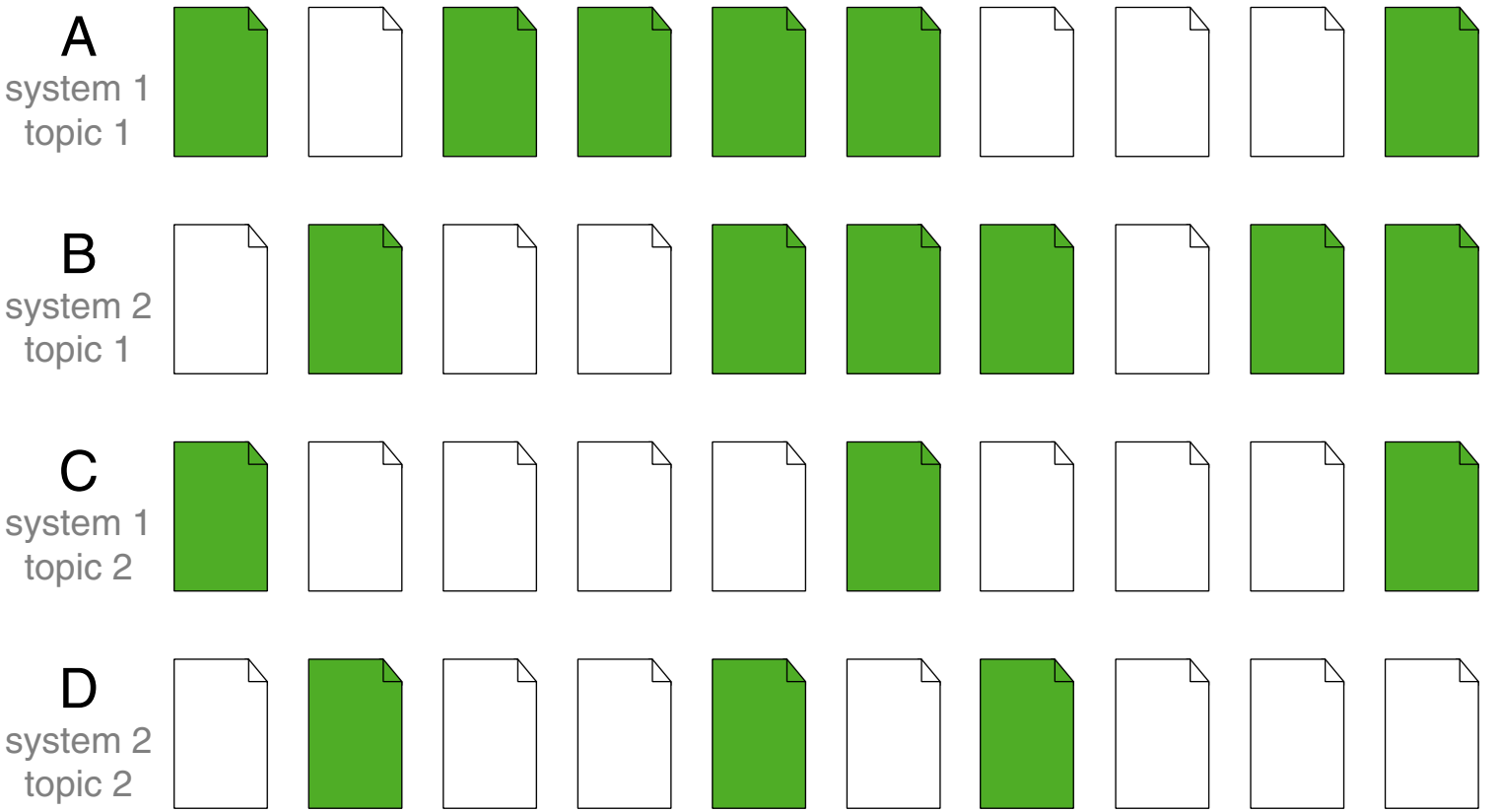
- Sample analysis
  Problem: Often the number of relevant documents is only a small fraction of the entire document collection. A representative sample would therefore encompass a large portion of documents from the collection.

- 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

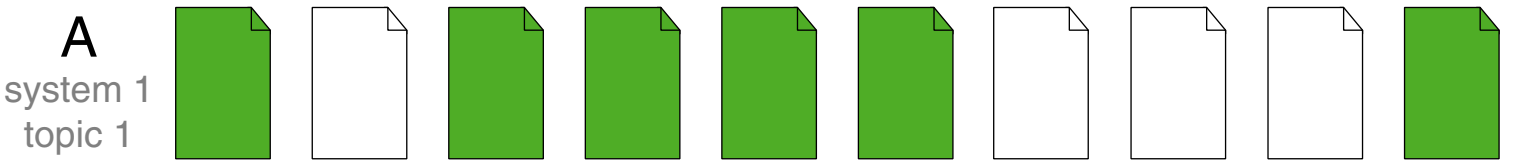


Which ranking on Topic 1 is better? Both have *precision* $= 0.6$ and *recall* $= 1.0$.

How good is System 1 compared to System 2 overall?

Performance Measures

Ranking Effectiveness: Precision@k and Recall@k



Assumption:

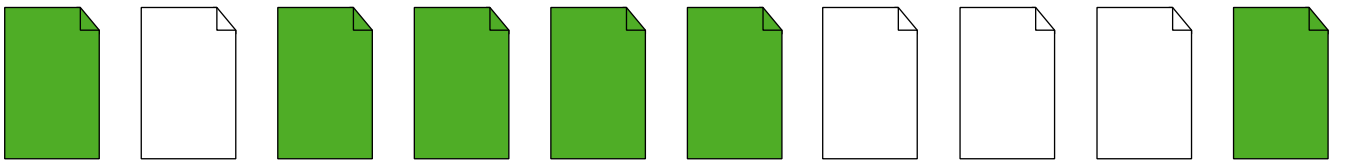
- The user browses all documents up to rank k.

➜ Compute *precision* and *recall* at a rank $k \geq 1$.

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

Performance Measures

Ranking Effectiveness: Precision@k and Recall@k



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:

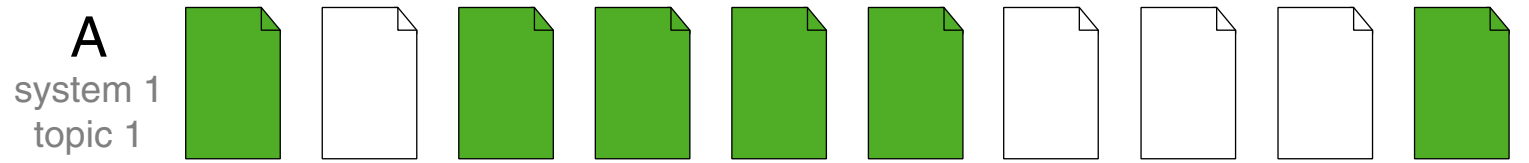
- The user browses all documents up to rank k.
- ➜ Compute *precision* and *recall* at a rank $k \geq 1$.
- Commonly used ranks are $k \in \{1, 5, 10, 20\}$.

Problems:

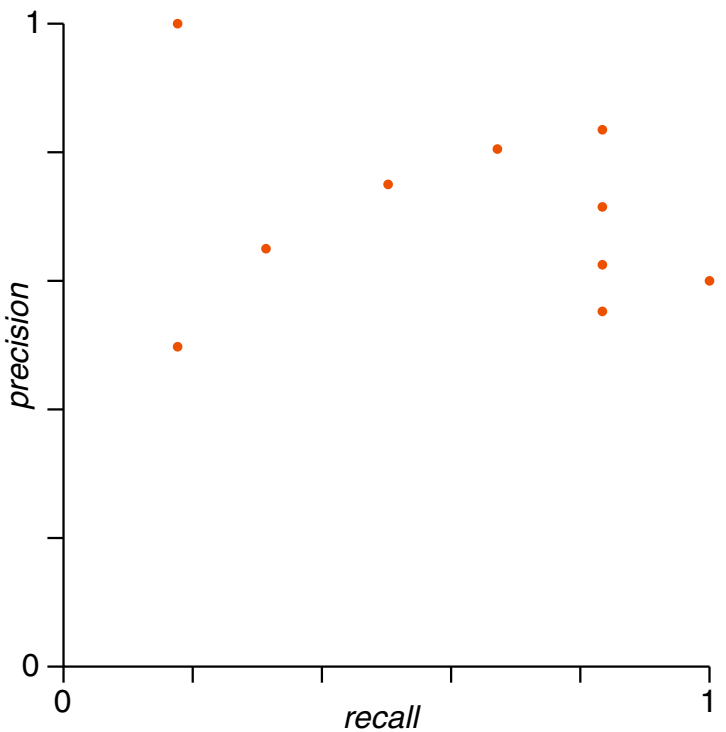
- Disregards ranking differences up to rank k.
- Disregards the (estimated) number of relevant documents (e.g., $\ll k$).
- Based on binary relevance judgments.

Performance Measures

Ranking Effectiveness: Precision-Recall Curves



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



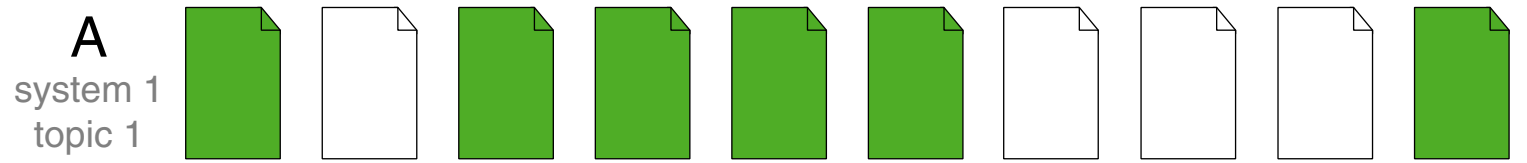
Observations:

- Connecting the dots yields a “curve.”
- The curves capture 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.

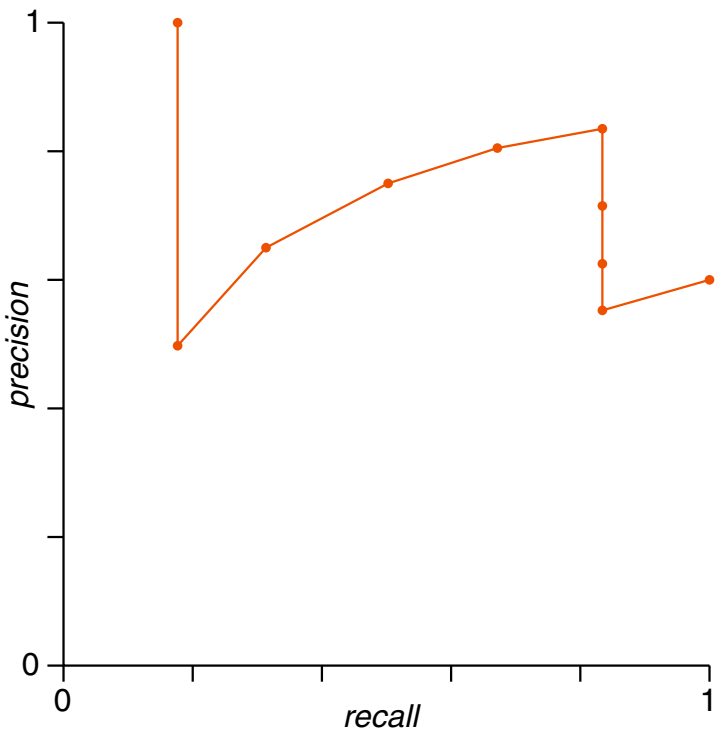
➜ These observations can be quantified as area under curve.

Performance Measures

Ranking Effectiveness: Precision-Recall Curves



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:

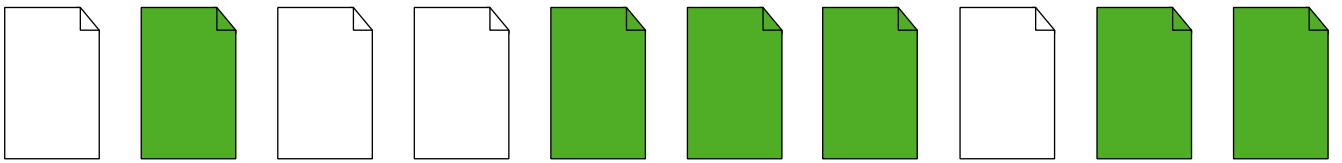
- Connecting the dots yields a “curve.”
- The curves capture detailed ranking characteristics: the user experience.
- Points on a curve other than the original ones lack interpretation.
- Given rankings from two systems, we can decide which one is better.

➜ These observations can be quantified as area under curve.

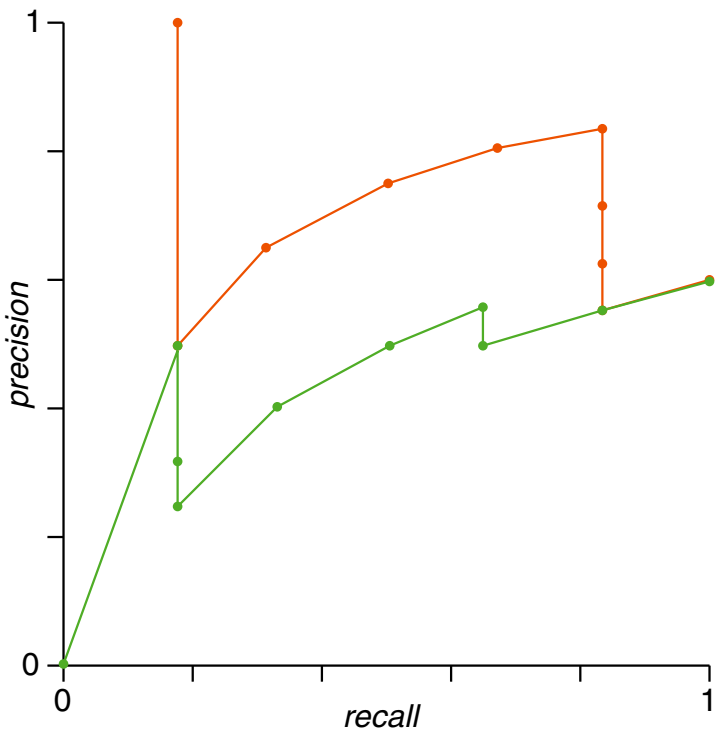
Performance Measures

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



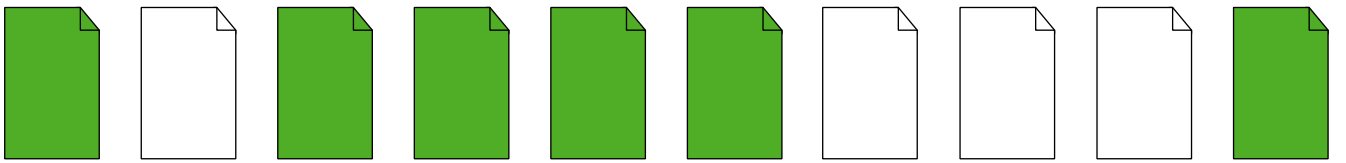
Observations:

- Connecting the dots yields a “curve.”
- The curves capture 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.

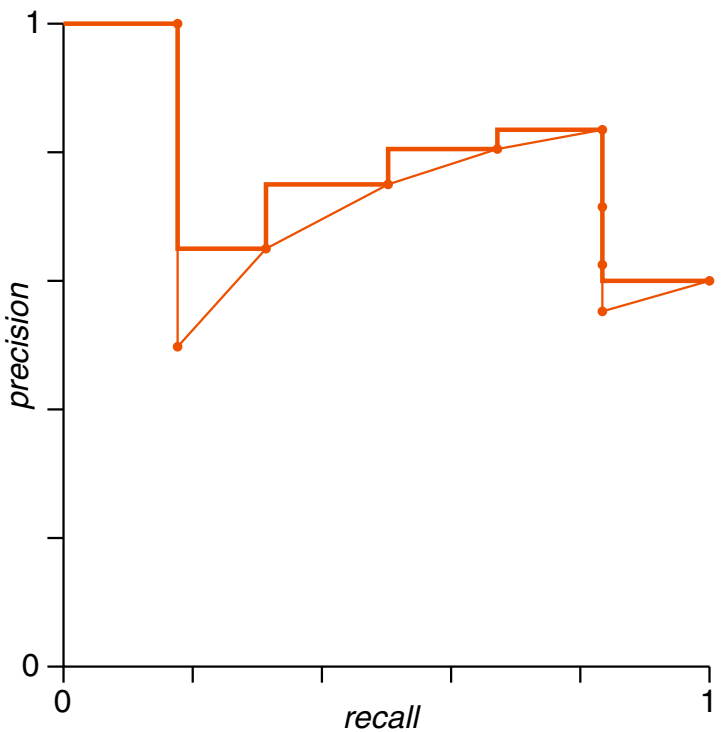
➜ These observations can be quantified as area under curve.

Performance Measures

Ranking Effectiveness: Average Precision



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



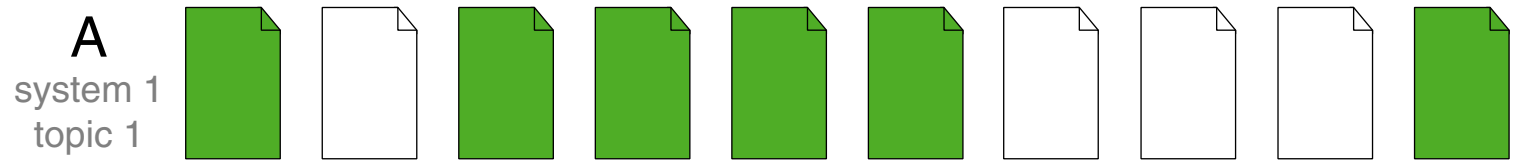
Average precision approximates the area under the precision-recall curve.

Interpolation alternatives:

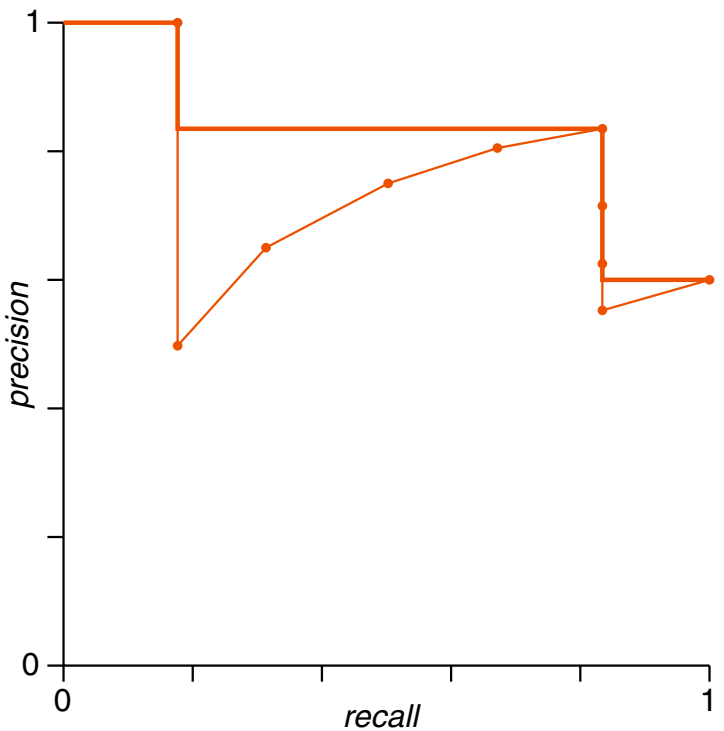
1. Integral of the step function visiting the maximum precision at every recall point.
2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

Performance Measures

Ranking Effectiveness: Average Precision



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



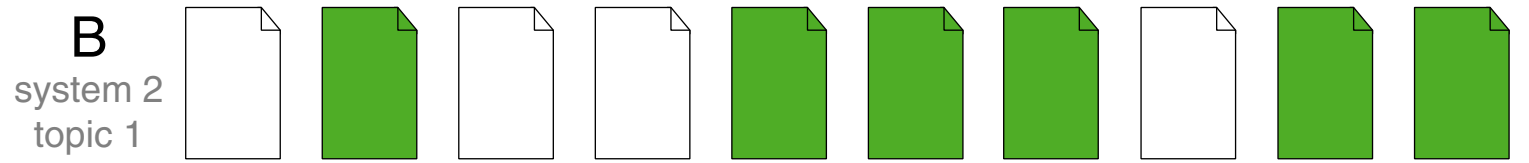
Average precision approximates the area under the precision-recall curve.

Interpolation alternatives:

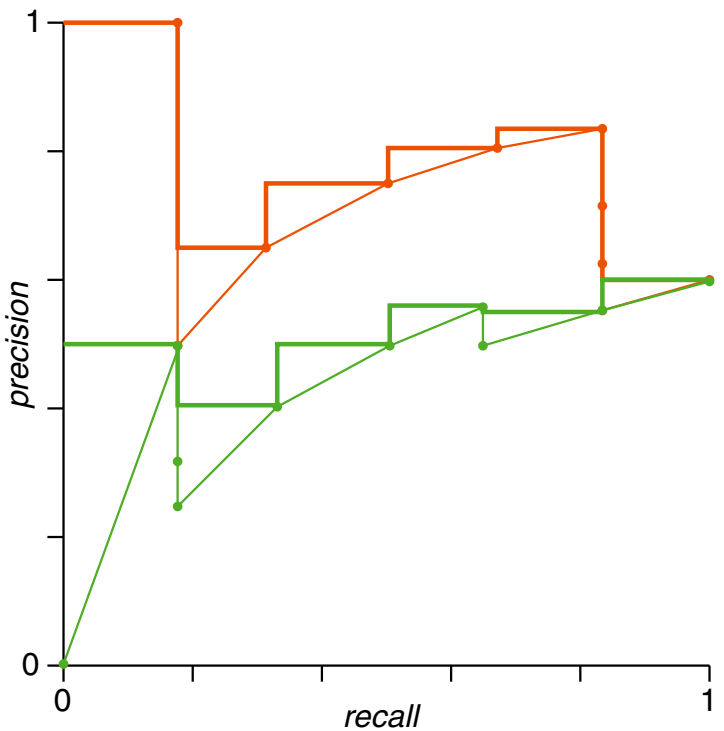
1. Integral of the step function visiting the maximum precision at every recall point.
2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

Performance Measures

Ranking Effectiveness: Average Precision



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



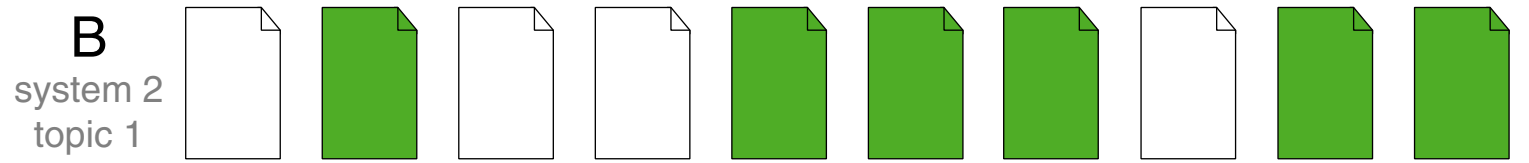
Average precision approximates the area under the precision-recall curve.

Interpolation alternatives:

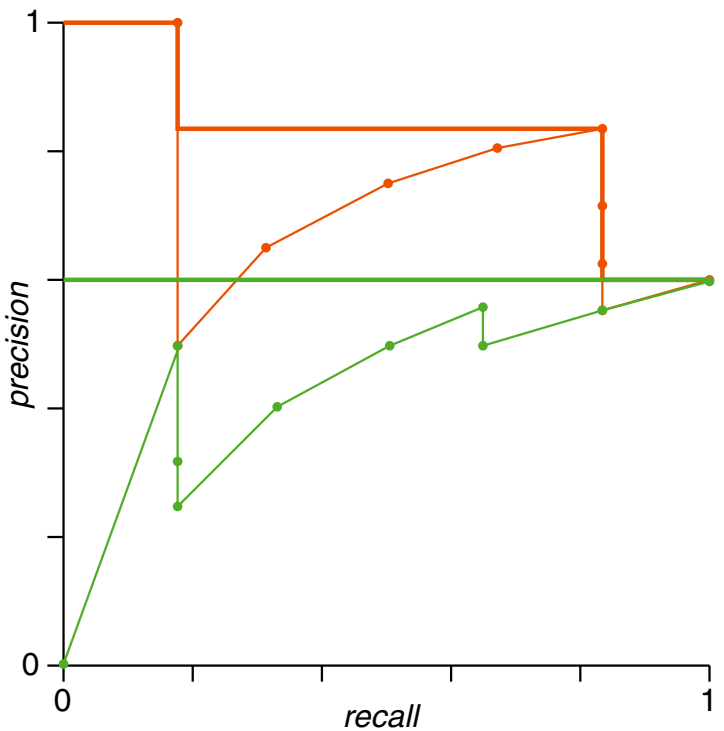
1. Integral of the step function visiting the maximum precision at every recall point.
2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

Performance Measures

Ranking Effectiveness: Average Precision



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



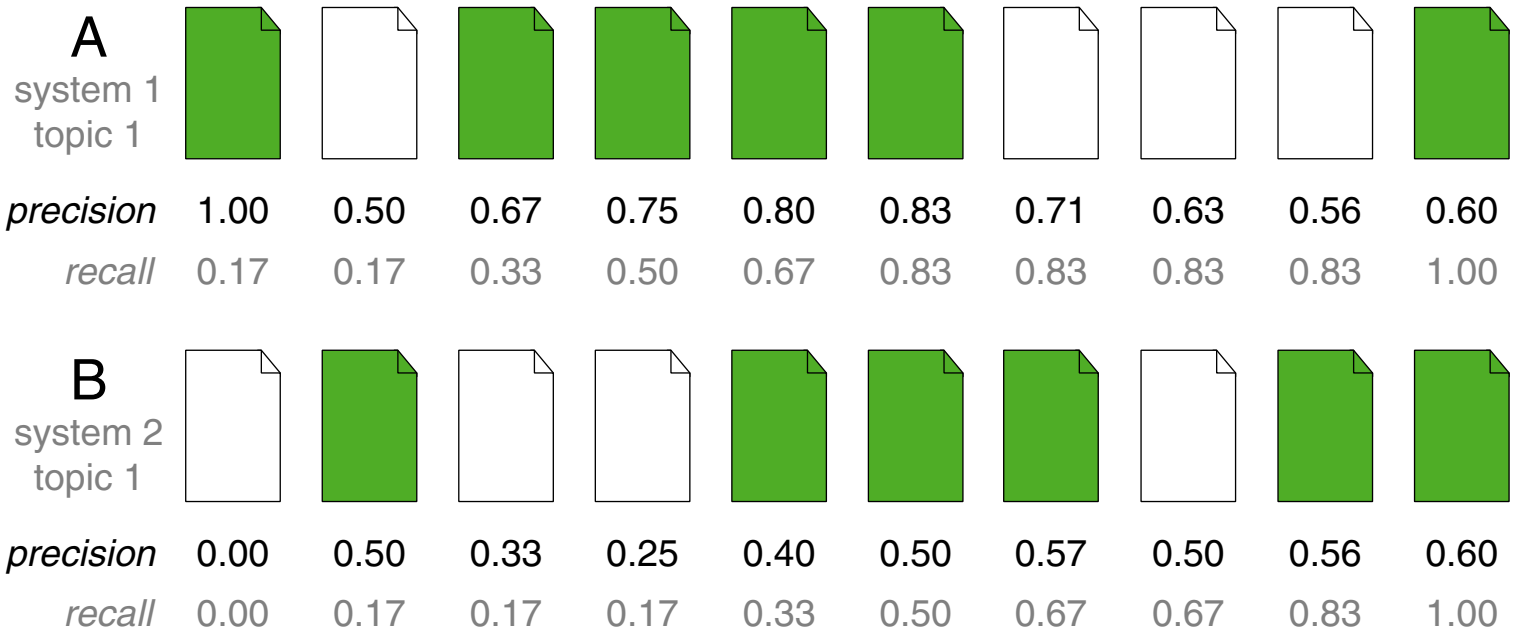
Average precision approximates the area under the precision-recall curve.

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

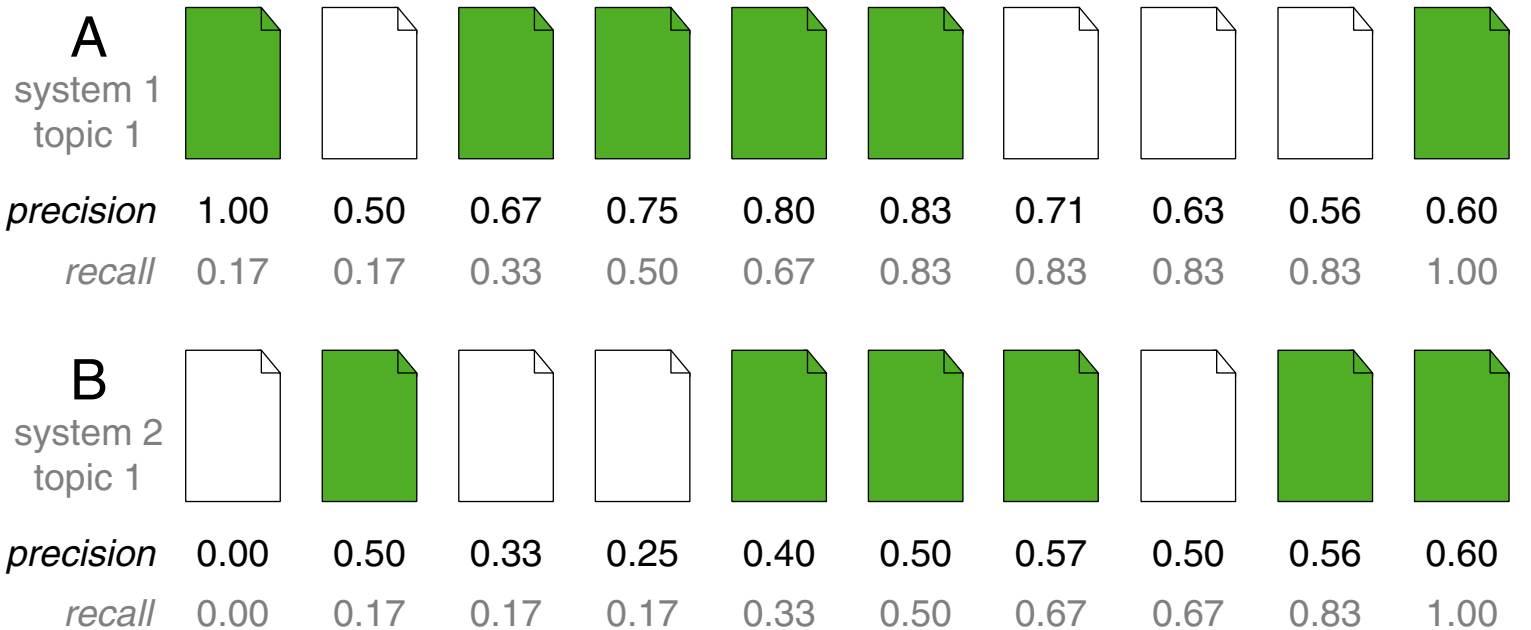
Ranking Effectiveness: Average Precision (Alternative 1)



- 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)



- 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.

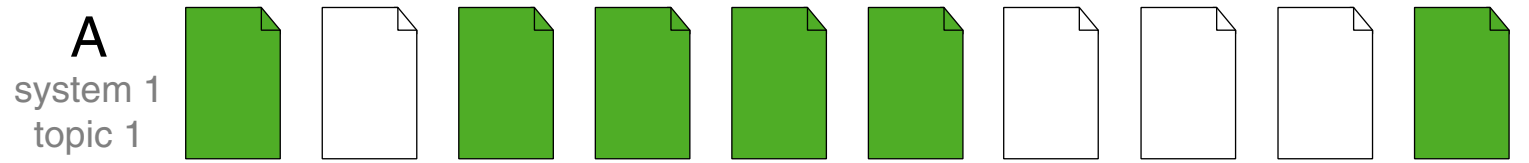
Then the two alternatives of average precision are computed as follows:

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

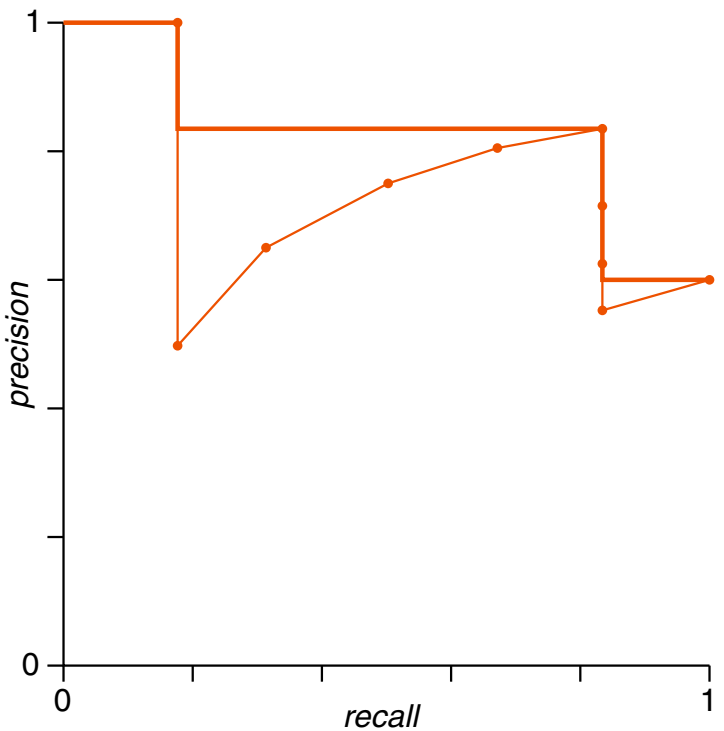
$$\mathit{AP}_2(q, R) = \frac{1}{11} \cdot \sum_{i \in \{0, 0.1, \ldots, 1\}} \Big(\max_{j:\ \mathit{recall}@j(R) \geq i} \mathit{precision}@j(R) \Big)$$

Performance Measures

Ranking Effectiveness: Average Precision-Recall Curves



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:

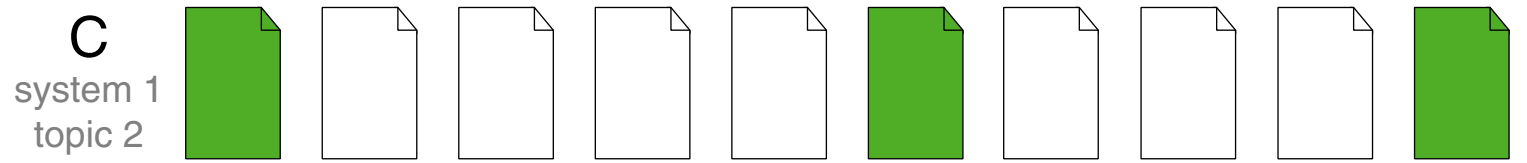
- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

Solution:

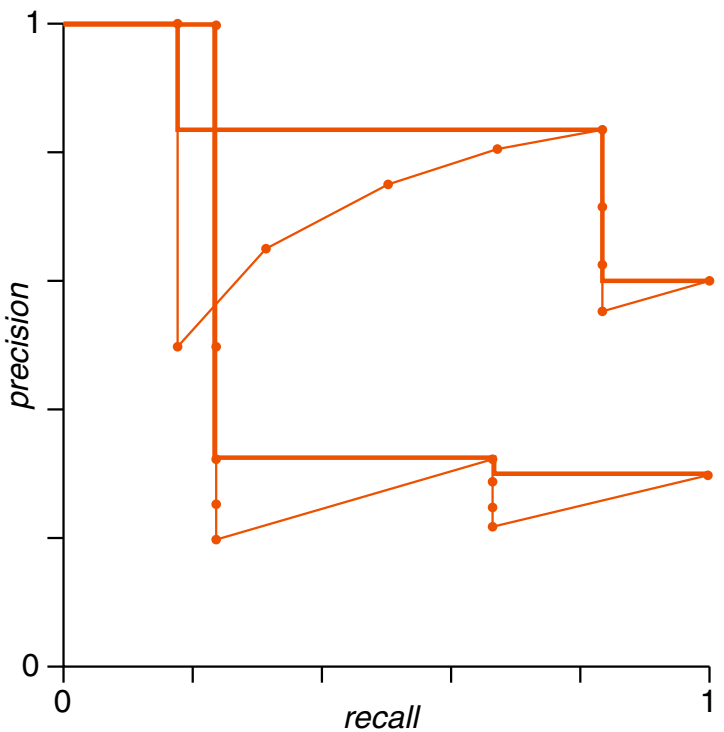
- Compute averages across 11 recall points at 0.1 steps.

Performance Measures

Ranking Effectiveness: Average Precision-Recall Curves



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:

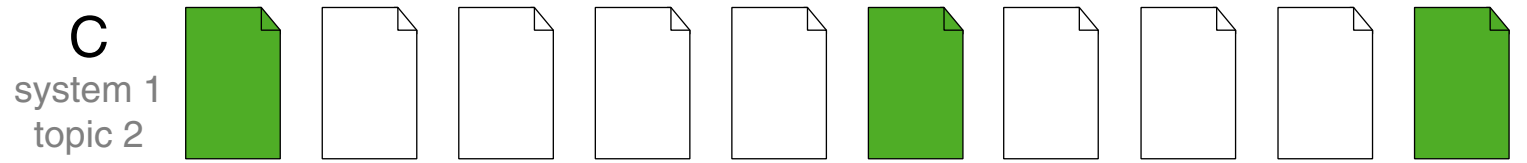
- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

Solution:

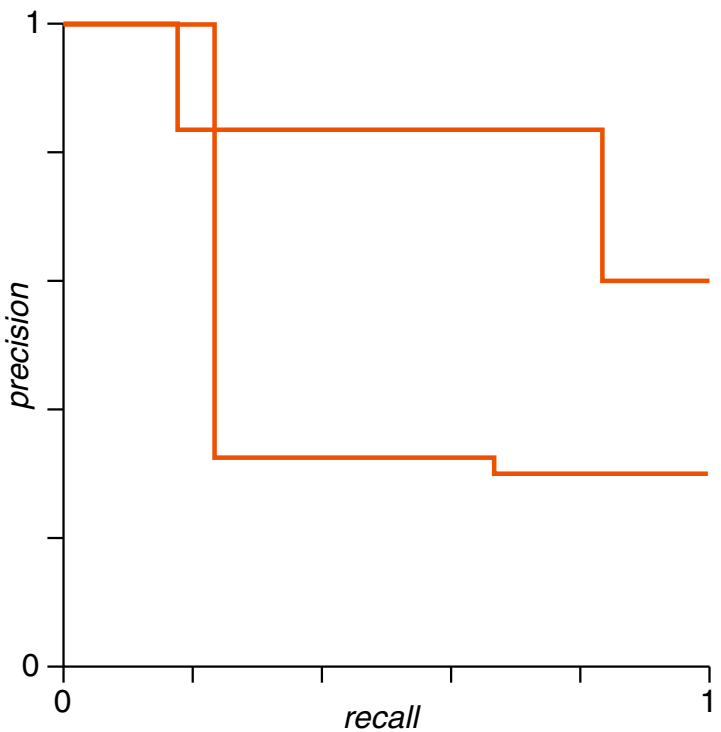
- Compute averages across 11 recall points at 0.1 steps.

Performance Measures

Ranking Effectiveness: Average Precision-Recall Curves



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:

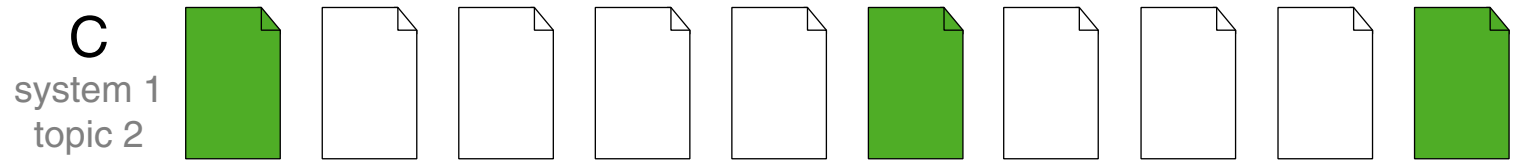
- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

Solution:

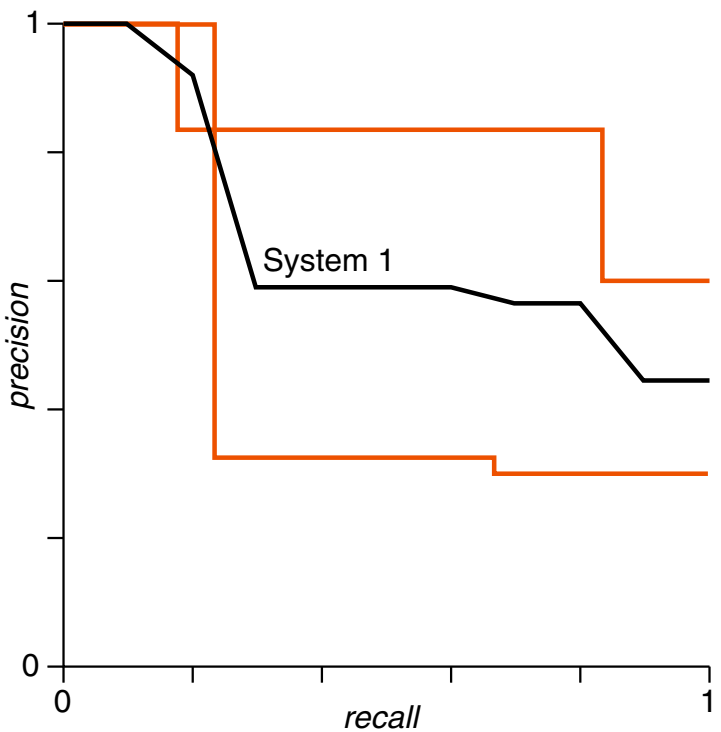
- Compute averages across 11 recall points at 0.1 steps.

Performance Measures

Ranking Effectiveness: Average Precision-Recall Curves



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:

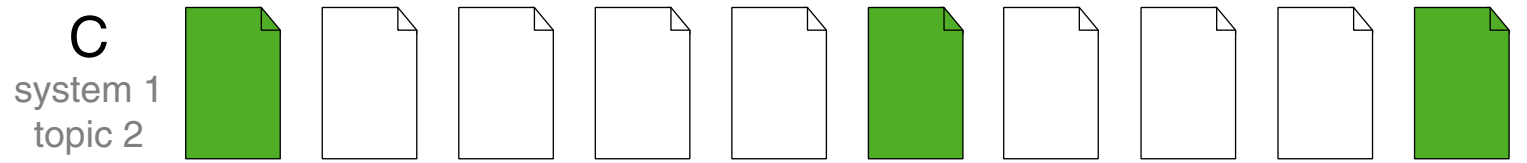
- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

Solution:

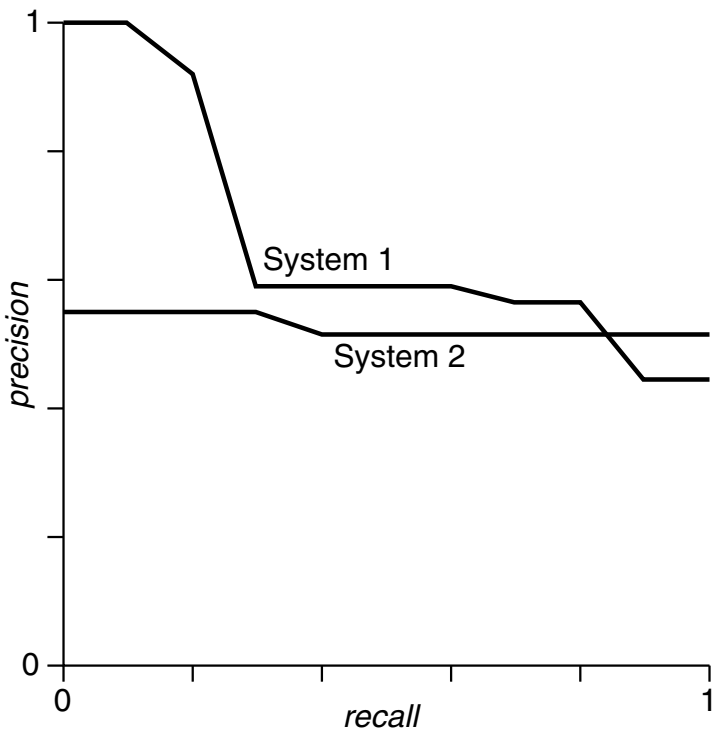
- Compute averages across 11 recall points at 0.1 steps.

Performance Measures

Ranking Effectiveness: Average Precision-Recall Curves



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



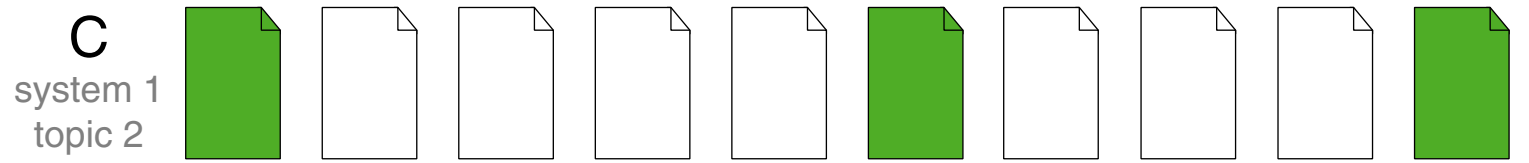
Interpretation:

- 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.

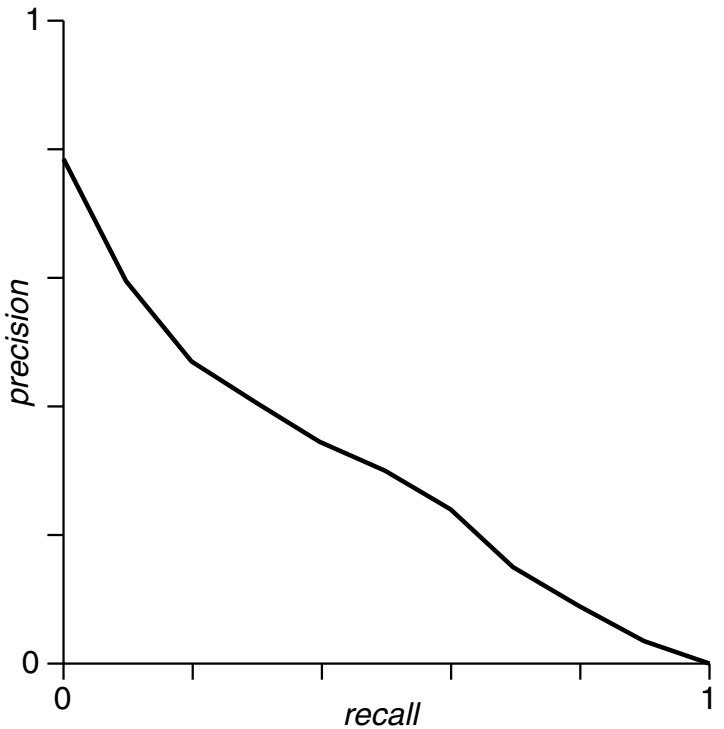
Curves are a lot smoother for 50 topics.

Performance Measures

Ranking Effectiveness: Average Precision-Recall Curves



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



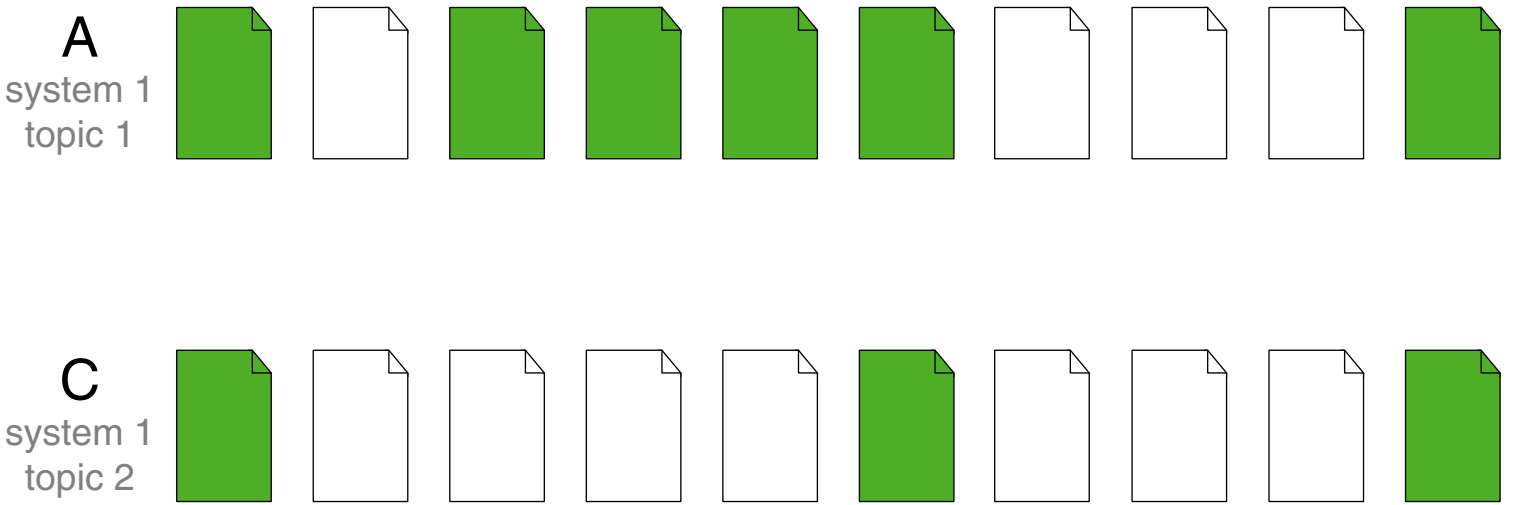
Interpretation:

- 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

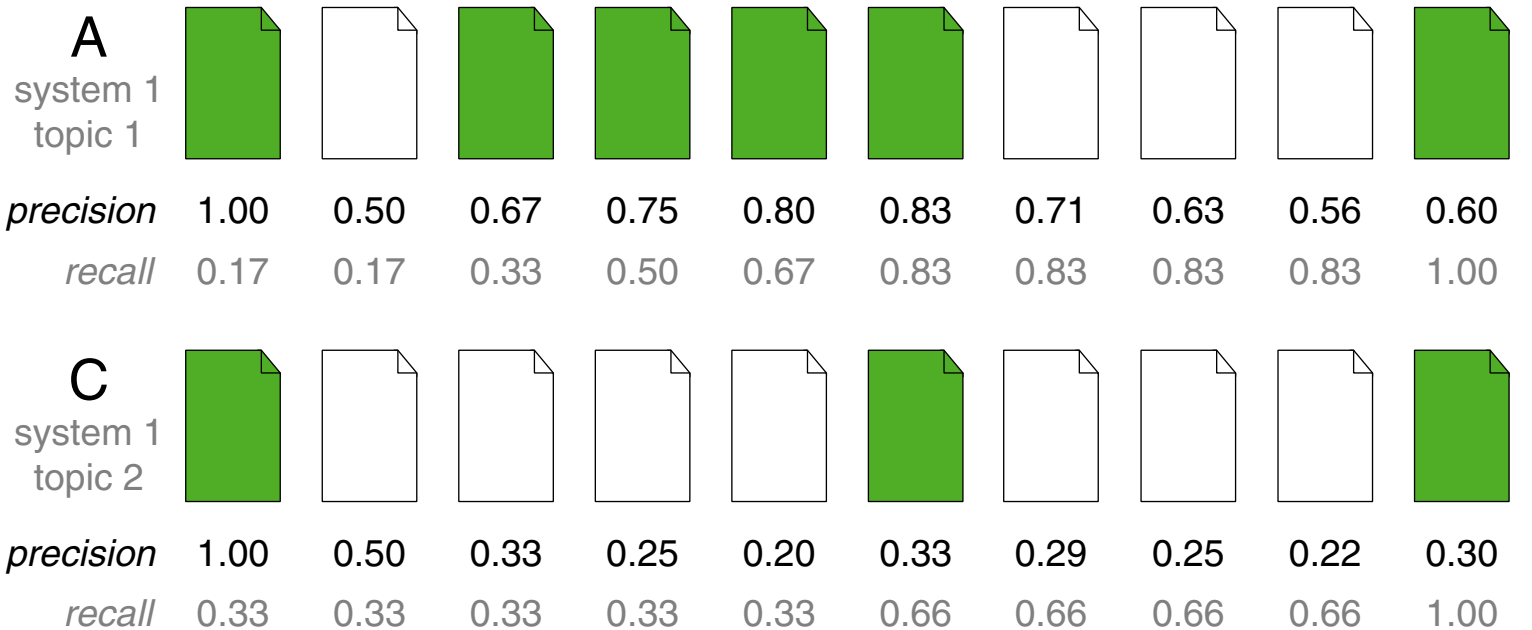
Ranking Effectiveness: Mean Average Precision (MAP)



- Meaningful system evaluation requires many topics.

Performance Measures

Ranking Effectiveness: Mean Average Precision (MAP)

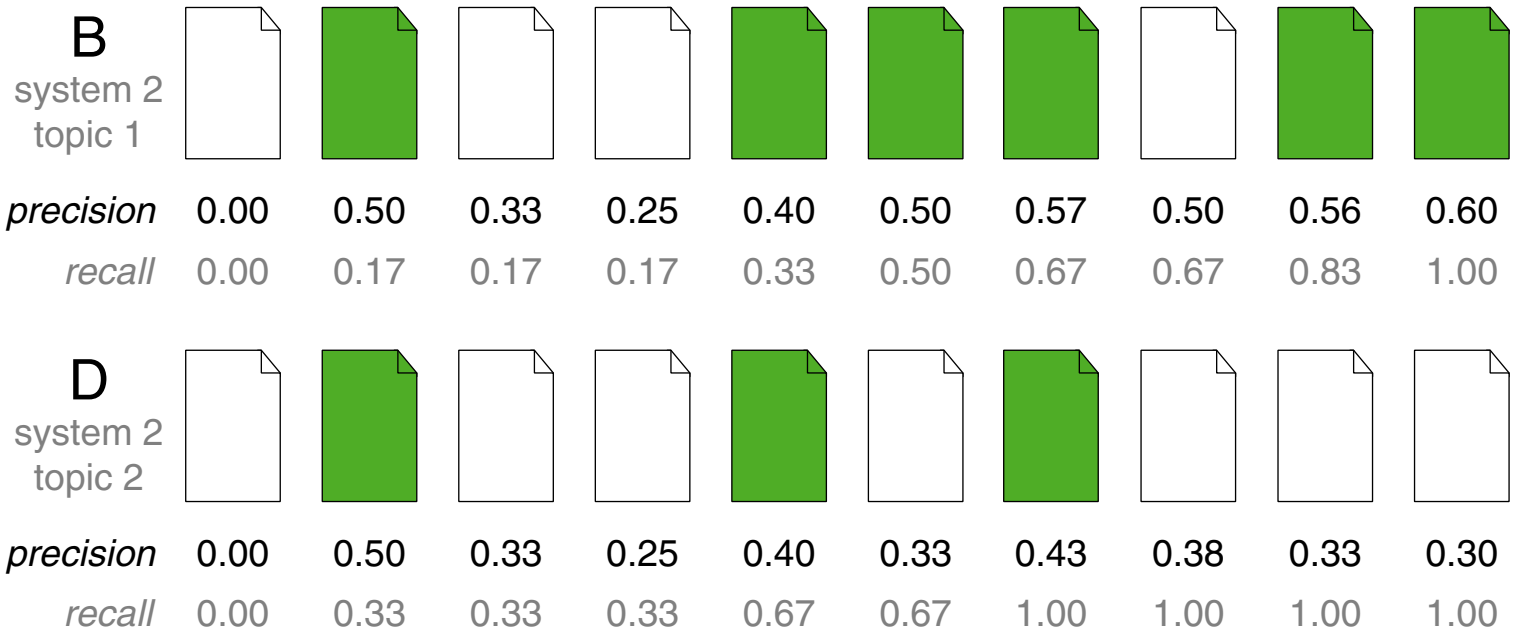


- Meaningful system evaluation requires many topics.
- Averaging average precision over topics gives us mean average precision.
- The MAP for System 1, Rankings A and C is $(0.78 + 0.44)/2 = 0.66$.

  (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)



- 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.

Solutions:

- Assume that users may decide to stop at any given rank.
- Assume that the stopping probability rises with rank depth.

Performance Measures

Ranking Effectiveness: Mean Reciprocal Rank (MRR)

For some topics only one relevant document is sought. In that case, the rank of the relevant document determines the quality of a ranking.

This can be measured by the reciprocal $1/r$ of r, which 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

The reciprocal rank strongly penalizes when relevant documents are not on top ranks.

MRR has several caveats, rendering it untrustworthy in practice. [Fuhr 2017]

Performance Measures

Ranking Effectiveness: Discounted Cumulative Gain (DCG)

When relevance is graded on a Likert scale, every relevant document has a gain determined by its relevance score. For example, let $r : D \times Q \to \{0, 1, 2, 3, 4, 5\}$ denote a relevance function, where D denotes the document collection and Q denotes the topics' queries.

When reading a ranked list top to bottom, the gain cumulates: $\sum_i^k r(d_i, q)$, where k denotes a rank, d_i denotes the document $d \in D$ at rank i, and q denotes the query.

However, the lower a document is ranked, the less likely it is actually examined, diminishing its potential gain in practice. The gain may be discounted, e.g., with reciprocal ranks.

Altogether, the discounted cumulative gain computes as follow:

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

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

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