

Chapter IR:VIII

VIII. Evaluation

- Laboratory Experiments
- Logging
- Effectiveness Measures
- Efficiency Measures
- Training and Testing

Effectiveness Measures

First Principles

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.

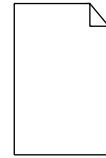
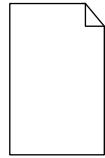
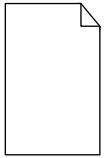
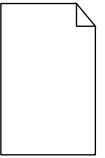
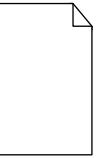
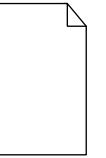
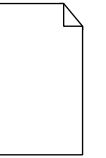
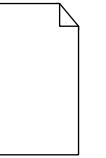
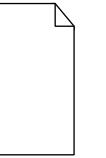
Sufficient justification is achieved by means of measurement, namely “the assignment of a number to a characteristic of an object [a search result], which can be compared with other objects.” [\[Wikipedia\]](#)

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

Effectiveness Measures

First Principles

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

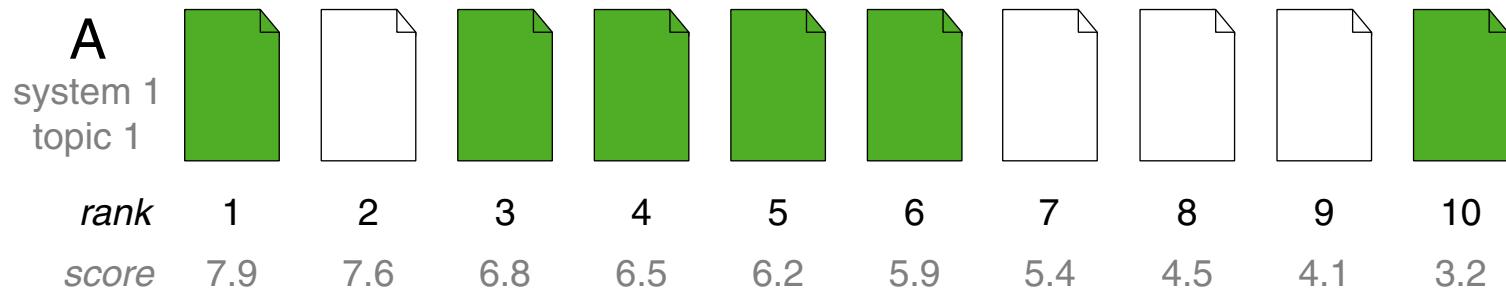
A										
system 1										
topic 1										
rank	1	2	3	4	5	6	7	8	9	10
score	7.9	7.6	6.8	6.5	6.2	5.9	5.4	4.5	4.1	3.2

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.

Effectiveness Measures

First Principles

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.

Remarks:

- ❑ Rankings are typically depicted in a top-to-bottom arrangement, where the item atop is the most important one. For brevity, we arrange them left-to-right.

Effectiveness Measures

Precision and Recall

A contingency table counts successes and failures:

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

with

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

Effectiveness Measures

Precision and Recall

A contingency table counts successes and failures:

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

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

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

with

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

Effectiveness Measures

Precision and Recall

A contingency table counts successes and failures:

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

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

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

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.
- Search results are reduced to sets, and search engines are presumed to return only documents they classify as relevant, i.e., considering relevance to be a binary variable.
- Alternative formulas based on the sets of *Results* and *Relevant* documents:

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

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

- The fraction of non-relevant documents that are retrieved is called

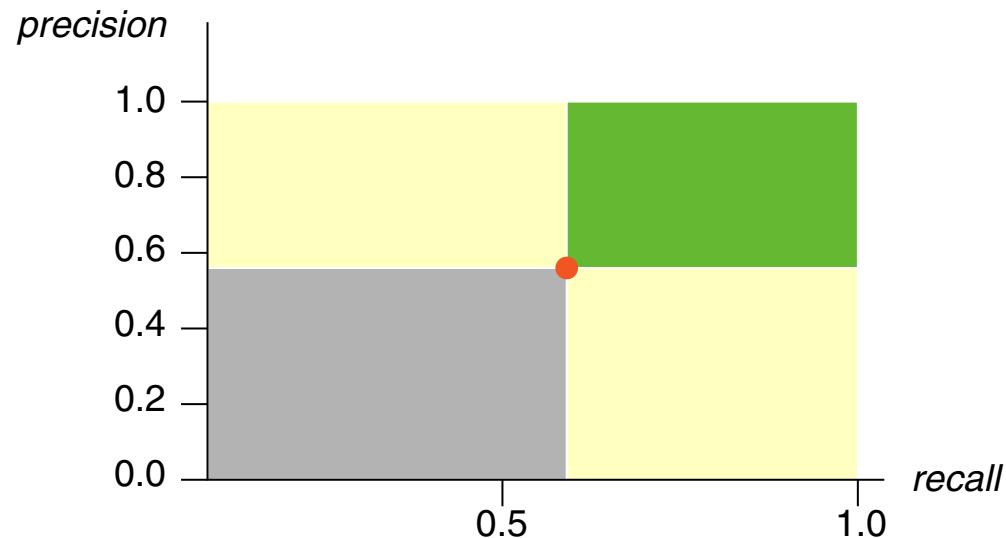
$$fallout = \frac{b}{b + d}$$

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.

Effectiveness Measures

Precision and Recall: *F*-Measure

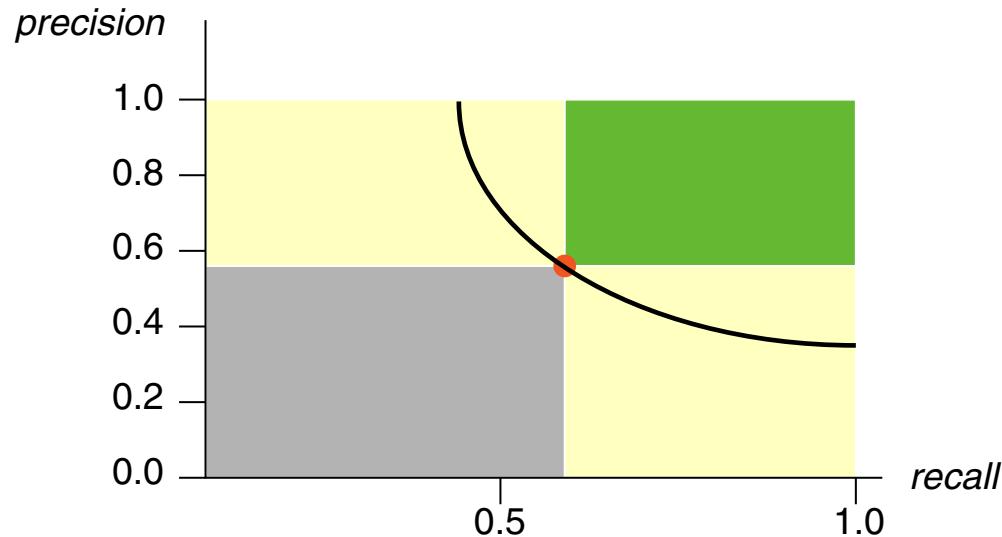
Comparison of retrieval systems:



Effectiveness Measures

Precision and Recall: *F*-Measure

Comparison of retrieval systems:



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

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

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. Its 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:

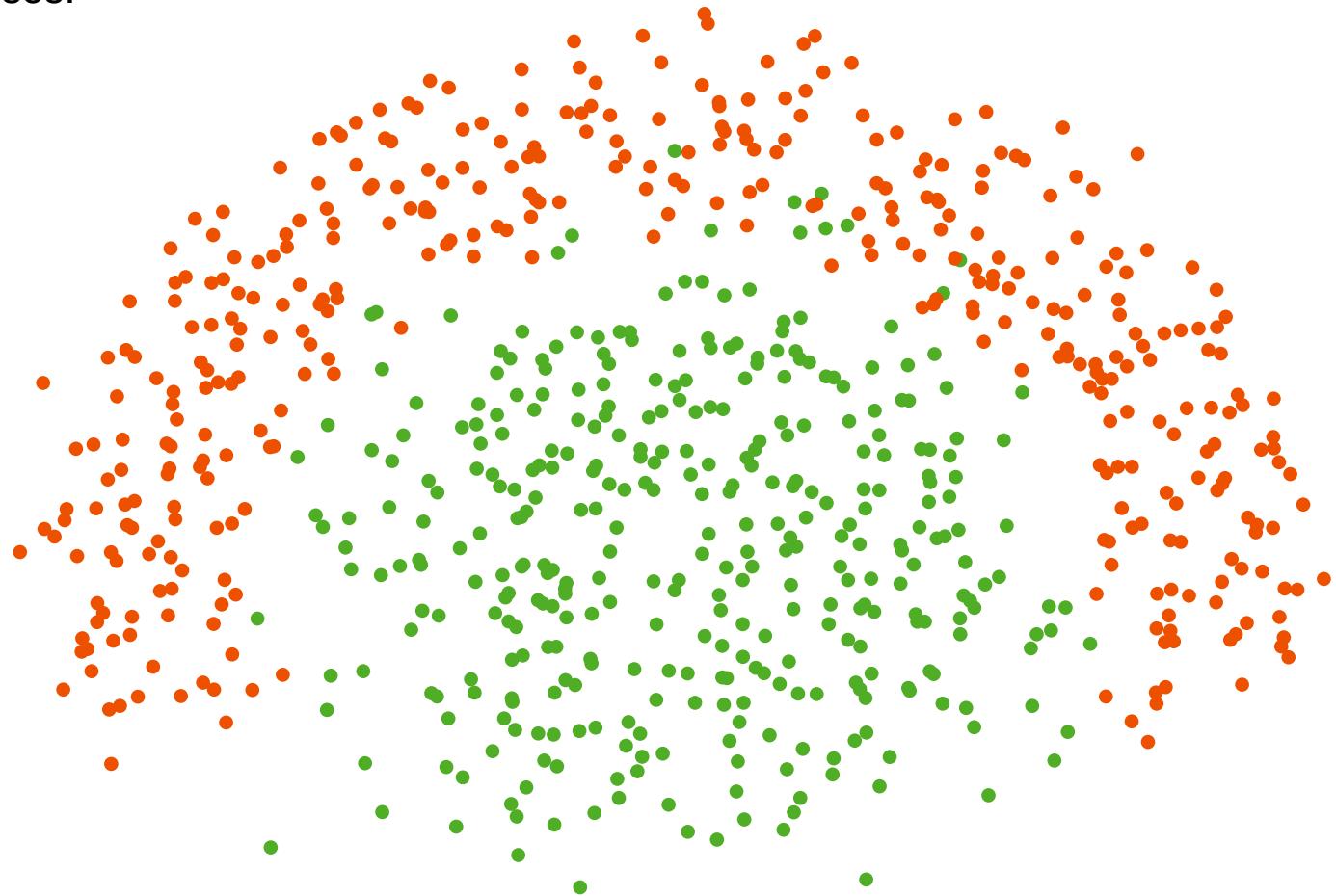
$$F = \frac{1}{\alpha \frac{1}{precision} + (1 - \alpha) \frac{1}{recall}} = \frac{(\beta^2 + 1)precision \cdot recall}{\beta^2 precision + recall}, \text{ where } \beta^2 = \frac{1 - \alpha}{\alpha}.$$

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.

Effectiveness Measures

Precision and Recall: Illustration

Classes: ● ●



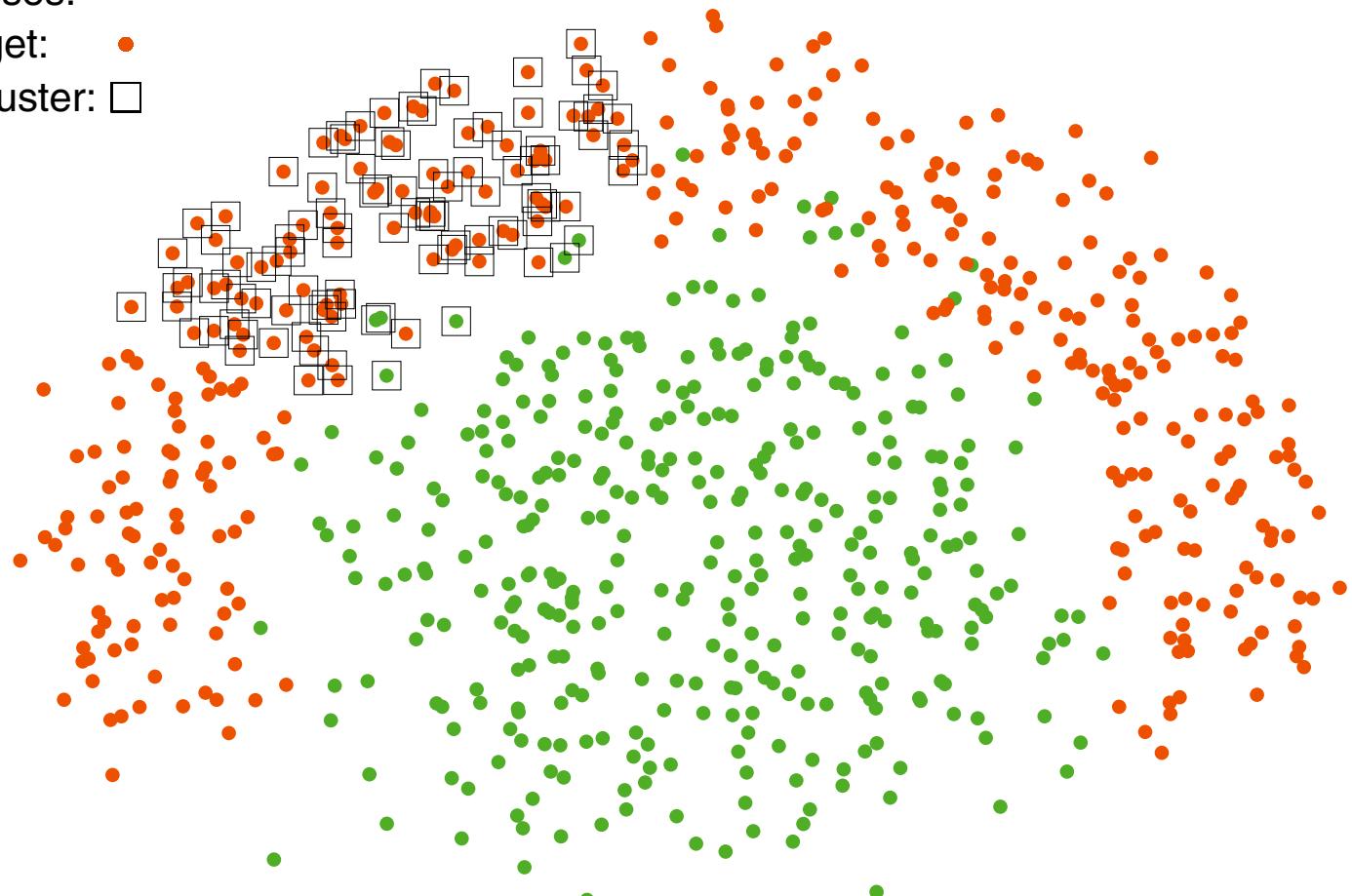
Effectiveness Measures

Precision and Recall: Illustration

Classes: ● ●

Target: ●

In cluster: □



Recall $\frac{\text{□}}{\bullet} = 0.26$ Precision $\frac{\text{□}}{(\text{□} \cup \text{●})} = 0.94$

F-Measure = 0.40

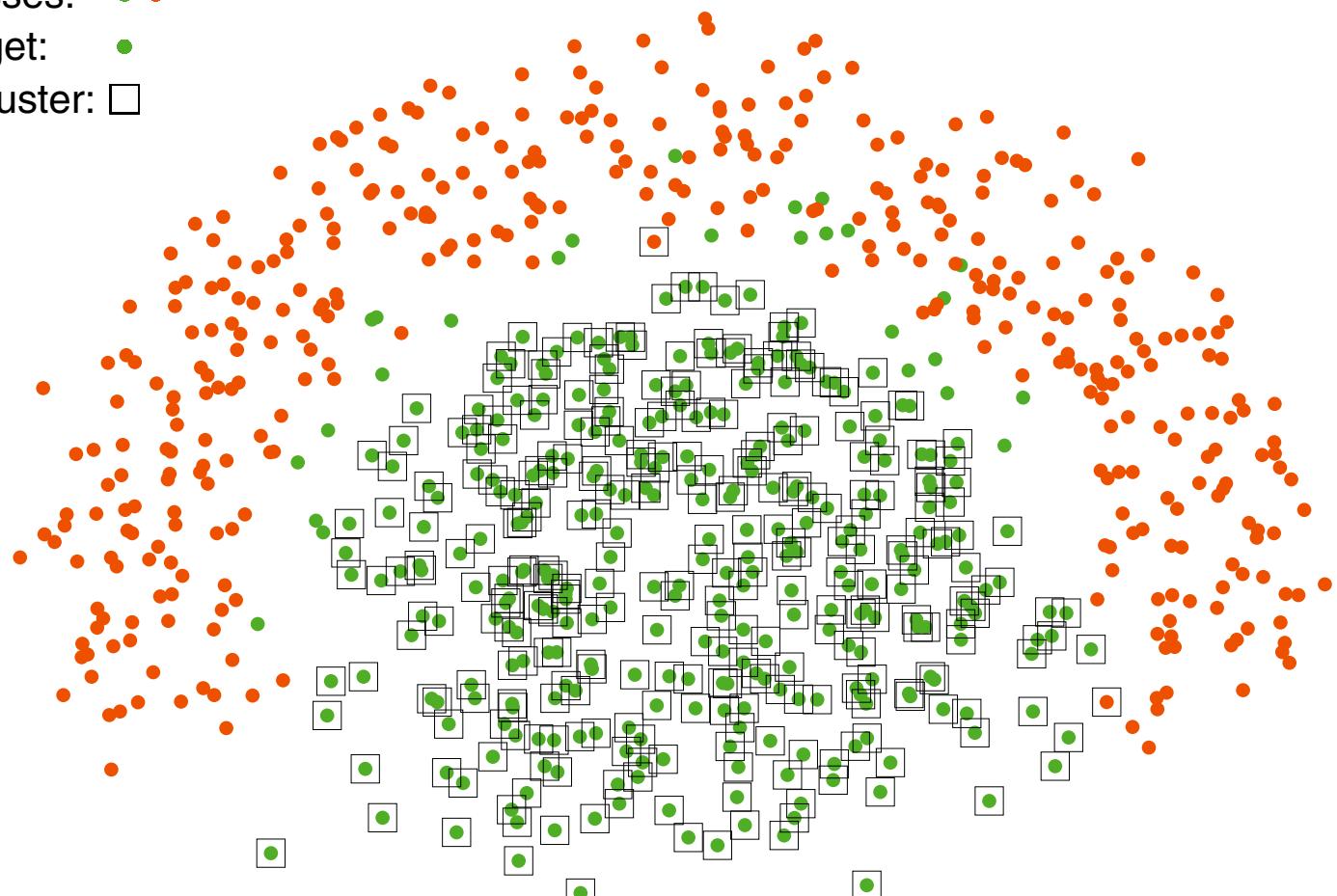
Effectiveness Measures

Precision and Recall: Illustration

Classes: ● ●

Target: ●

In cluster: □



Recall $\frac{\text{□}}{\bullet} = 0.92$ Precision $\frac{\text{□}}{(\text{□} \cup \bullet)} = 0.99$

F-Measure = 0.95

Effectiveness Measures

Precision and Recall: Recall Estimation

The number of relevant documents is neither recognizable by a user, nor can it be estimated with reasonable effort. 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.

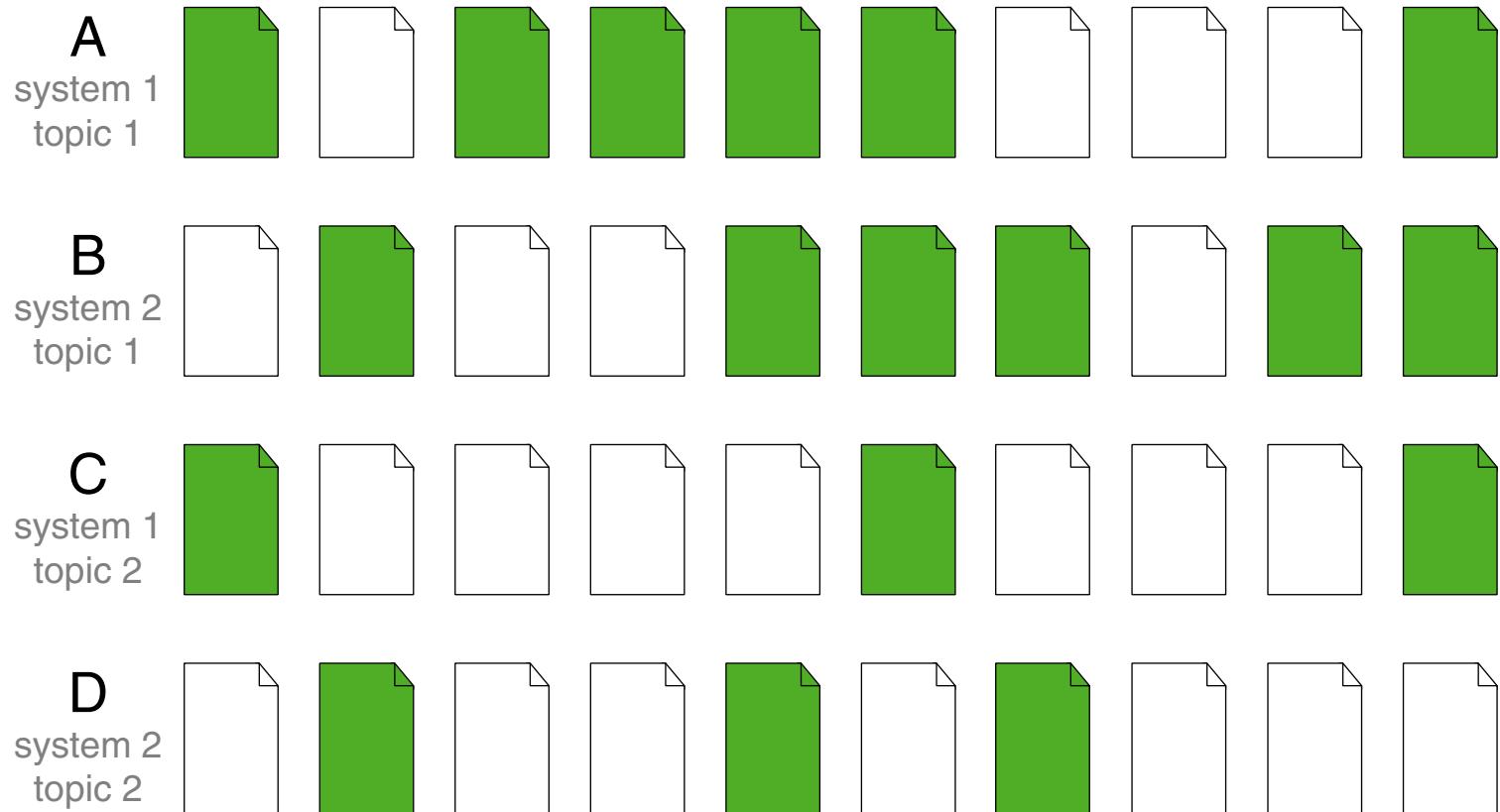
- Query expansion

The search results for a query of a given topic are judge until 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).

Effectiveness Measures

Ranking Effectiveness

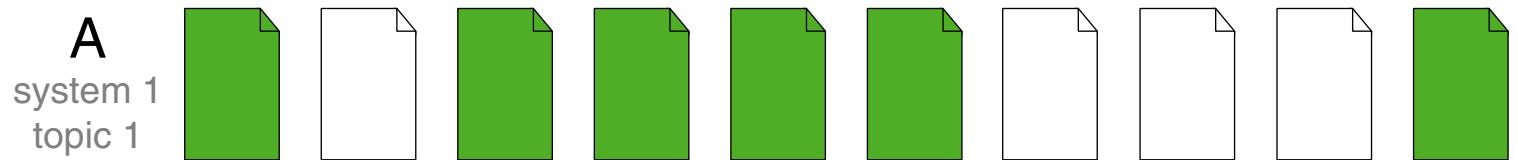


Which ranking on Topic 1 is better? Both have $precision = 0.6$ and $recall = 1.0$.

How good is System 1 compared to System 2 overall?

Effectiveness Measures

Ranking Effectiveness: Precision@k and Recall@k



Assumption:

- Only the top k documents retrieved will be viewed.
- Compute *precision* and *recall* at a rank $k \geq 1$.
- Commonly used ranks are $k \in \{1, 5, 10, 20\}$.

Effectiveness Measures

Ranking Effectiveness: Precision@ k and Recall@ k

A system 1 topic 1	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
<i>precision</i>	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
<i>recall</i>	0.17	0.17	0.33	0.50	0.67	0.83	0.83	0.83	0.83	1.00

Assumption:

- Only the top k documents retrieved will be viewed.
- Compute *precision* and *recall* at a rank $k \geq 1$.
- Commonly used ranks are $k \in \{1, 5, 10, 20\}$.

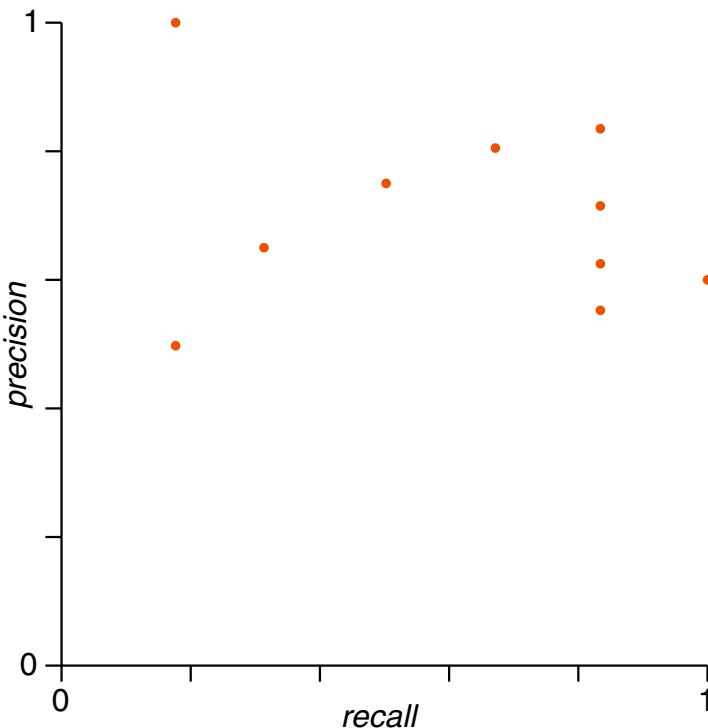
Problems:

- Disregards ranking differences up to rank k .
- Disregards the (estimated) number of relevant documents (e.g., $\ll k$).
- Devise a parameterless effectiveness measure.

Effectiveness Measures

Ranking Effectiveness: Precision-Recall Curves

A		system 1	topic 1	system 1	topic 2	system 1	topic 3	system 1	topic 4	system 1	topic 5	system 1	topic 6	system 1	topic 7	system 1	topic 8	system 1	topic 9	system 1	topic 10
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	



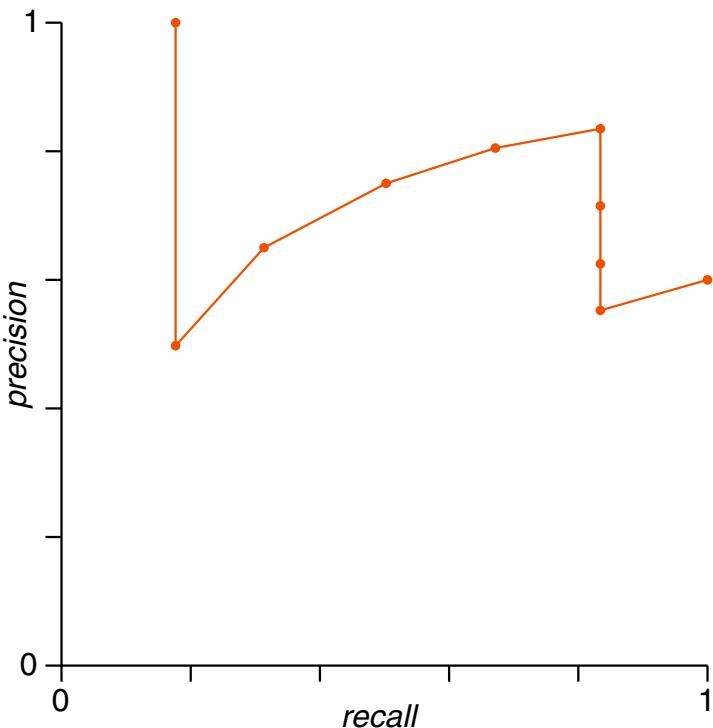
Observations:

- Connecting the dots yields a “curve.”
- The lines have no interpretation.
- The curves capture detailed ranking characteristics: the user experience.
- Given rankings from two systems, we can decide which one is better.
- These observations can be quantified as area under curve.

Effectiveness Measures

Ranking Effectiveness: Precision-Recall Curves

A system 1 topic 1	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60
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



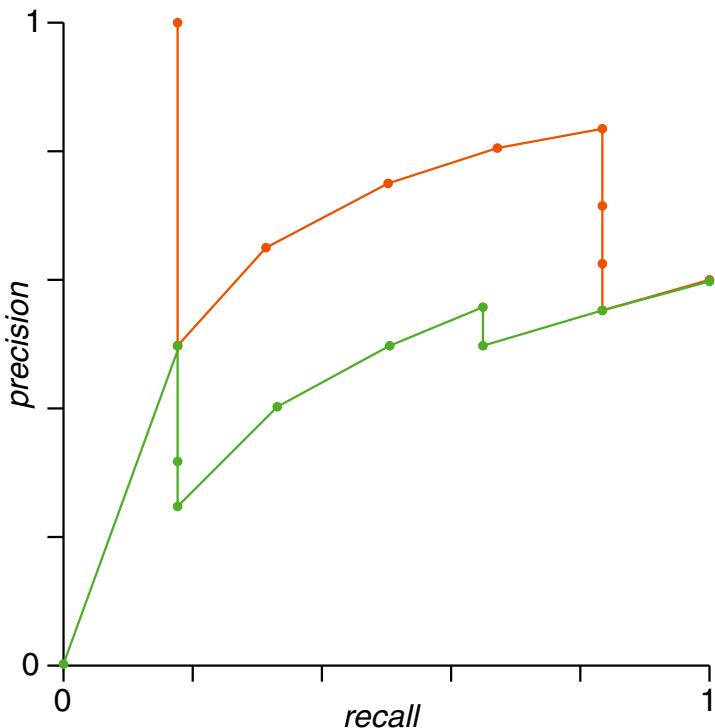
Observations:

- Connecting the dots yields a “curve.”
- The lines have no interpretation.
- The curves capture detailed ranking characteristics: the user experience.
- Given rankings from two systems, we can decide which one is better.
- These observations can be quantified as area under curve.

Effectiveness 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



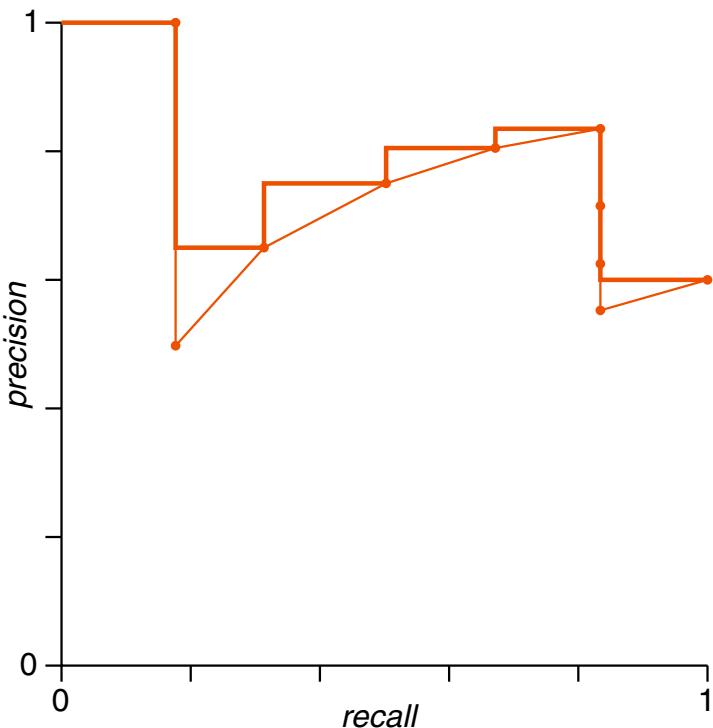
Observations:

- Connecting the dots yields a “curve.”
- The lines have no interpretation.
- The curves capture detailed ranking characteristics: the user experience.
- Given rankings from two systems, we can decide which one is better.
- These observations can be quantified as area under curve.

Effectiveness Measures

Ranking Effectiveness: Average Precision

A system 1 topic 1	precision	recall								
	1.00	0.17	0.50	0.17	0.67	0.33	0.75	0.50	0.80	0.67
										0.83



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

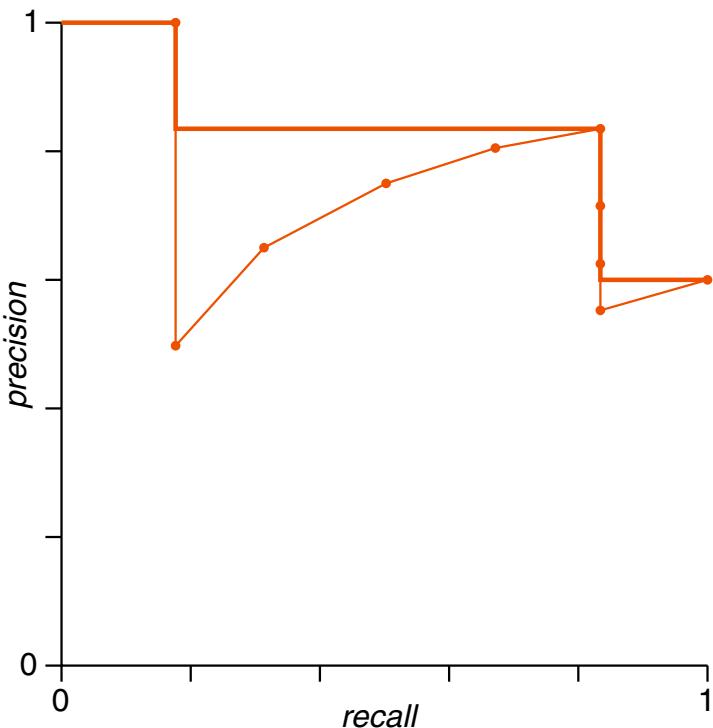
Interpolation alternatives:

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.

Effectiveness Measures

Ranking Effectiveness: Average Precision

A system 1 topic 1	precision	recall								
	1.00	0.17	0.50	0.17	0.67	0.33	0.75	0.50	0.80	0.67
										0.83



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

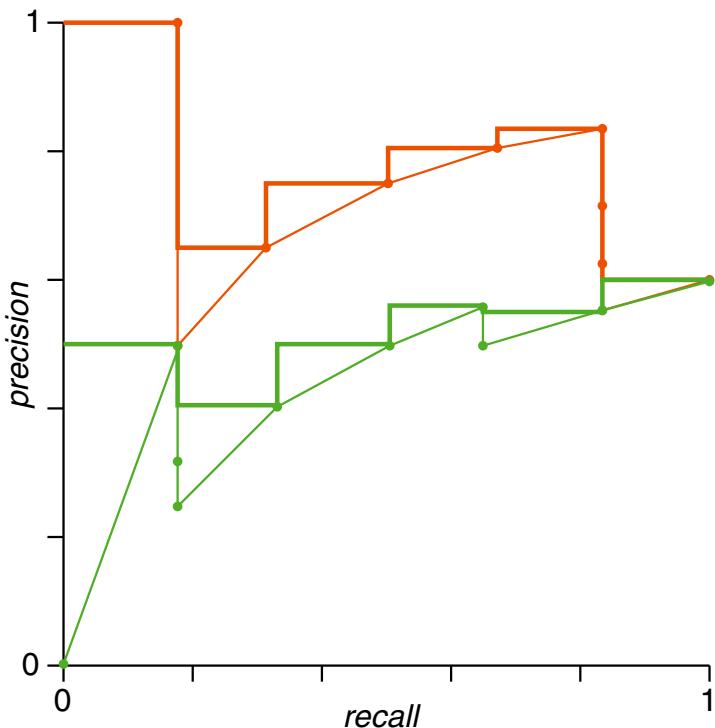
Interpolation alternatives:

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.

Effectiveness 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.

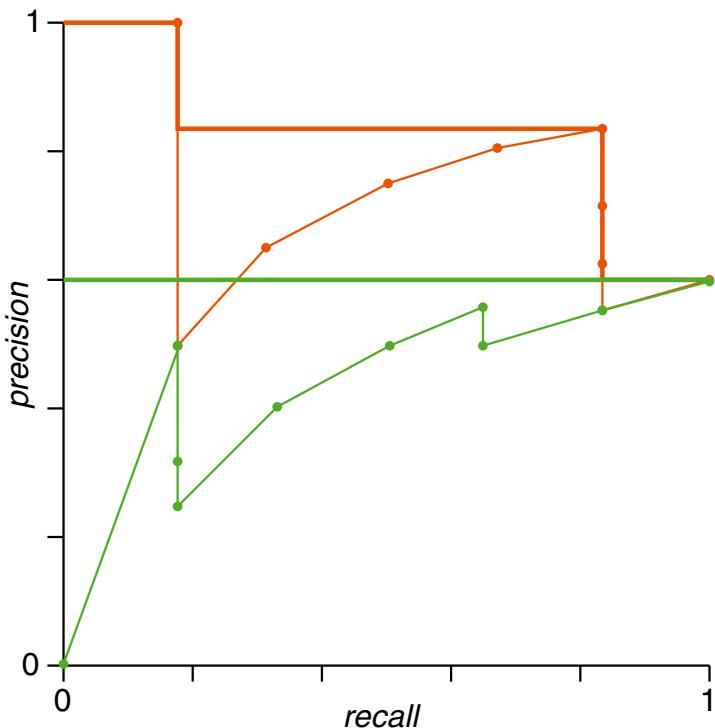
Interpolation alternatives:

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.

Effectiveness 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.

Effectiveness Measures

Ranking Effectiveness: Average Precision (Alternative 1)

A		system 1 topic 1	precision	recall	precision	recall	precision	recall	precision	recall	
			1.00	0.17	0.50	0.17	0.67	0.33	0.75	0.50	
					0.80	0.67	0.83	0.83	0.71	0.83	
B		system 2 topic 1	precision	recall	precision	recall	precision	recall	precision	recall	
			0.00	0.00	0.50	0.17	0.33	0.17	0.25	0.17	
					0.40	0.33	0.50	0.50	0.57	0.67	
						0.50	0.50	0.56	0.60	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.

Effectiveness Measures

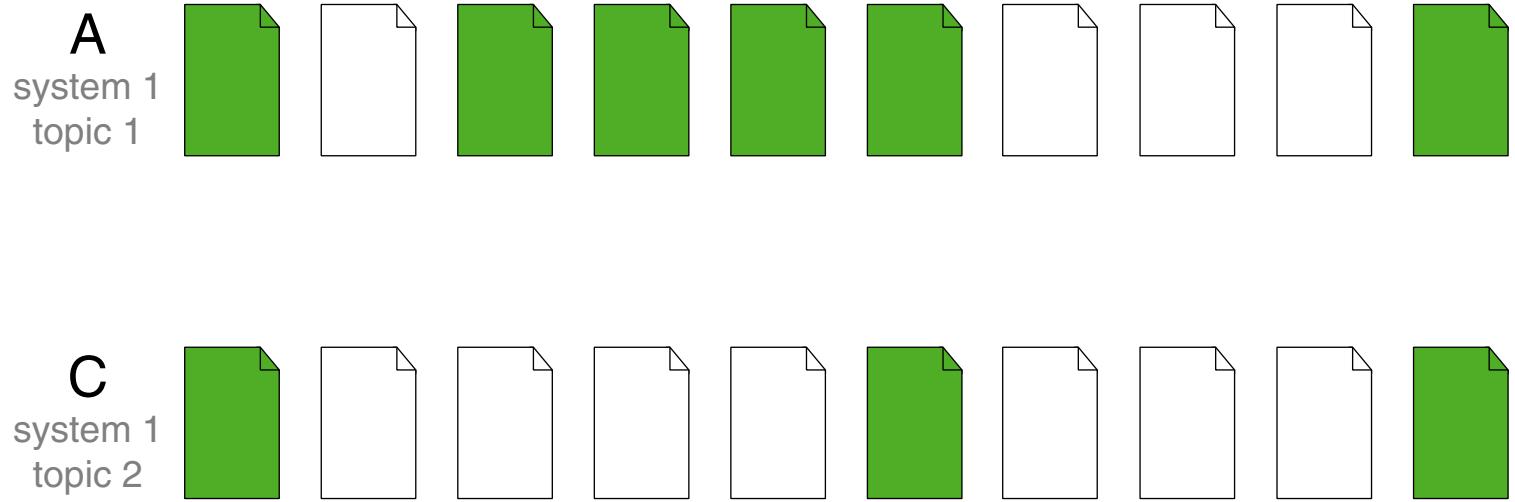
Ranking Effectiveness: Average Precision (Alternative 2)

A											
system 1	topic 1										
<i>precision</i>	1.00	0.50	0.67	0.75	0.80	0.83	0.71	0.63	0.56	0.60	
<i>recall</i>	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										
<i>precision</i>	0.00	0.50	0.33	0.25	0.40	0.50	0.57	0.50	0.56	0.60	
<i>recall</i>	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: $0, 0.1, \dots, 0.9, 1.$
- 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

Effectiveness Measures

Ranking Effectiveness: Mean Average Precision (MAP)



- ❑ Meaningful system evaluation requires **many topics**.

Effectiveness Measures

Ranking Effectiveness: Mean Average Precision (MAP)

A		system 1 topic 1	precision	recall	precision	recall	precision	recall	precision	recall
			1.00	0.17	0.50	0.17	0.67	0.33	0.75	0.50
			0.80	0.67	0.83	0.83	0.75	0.50	0.80	0.67
C		system 1 topic 2								
			1.00	0.33	0.50	0.33	0.33	0.33	0.25	0.33
			0.20	0.33	0.25	0.33	0.20	0.33	0.33	0.20
			0.33	0.66	0.33	0.66	0.29	0.66	0.25	0.66
			0.66	0.66	0.22	0.66	0.25	0.66	0.30	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: $(1.0 + 0.33 + 0.3)/3 = 0.54$)

Effectiveness Measures

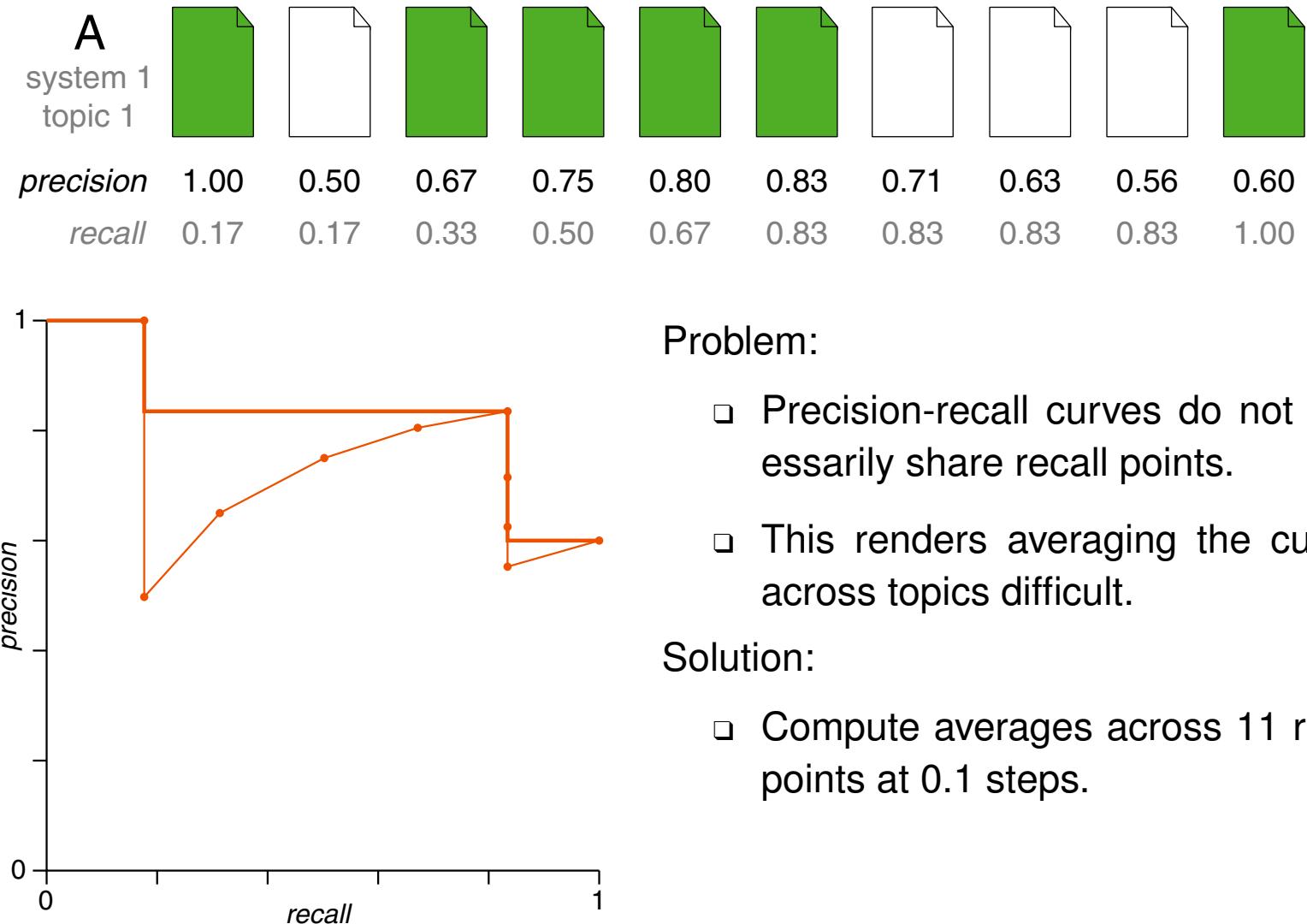
Ranking Effectiveness: Mean Average Precision (MAP)

B										
system 2 topic 1										
<i>precision</i>	0.00	0.50	0.33	0.25	0.40	0.50	0.57	0.50	0.56	0.60
<i>recall</i>	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										
<i>precision</i>	0.00	0.50	0.33	0.25	0.40	0.33	0.43	0.38	0.33	0.30
<i>recall</i>	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 $(0.52 + 0.44)/2 = 0.48$.
 $(B: (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52 \text{ and } D: (0.5 + 0.4 + 0.43)/3 = 0.44)$

Effectiveness Measures

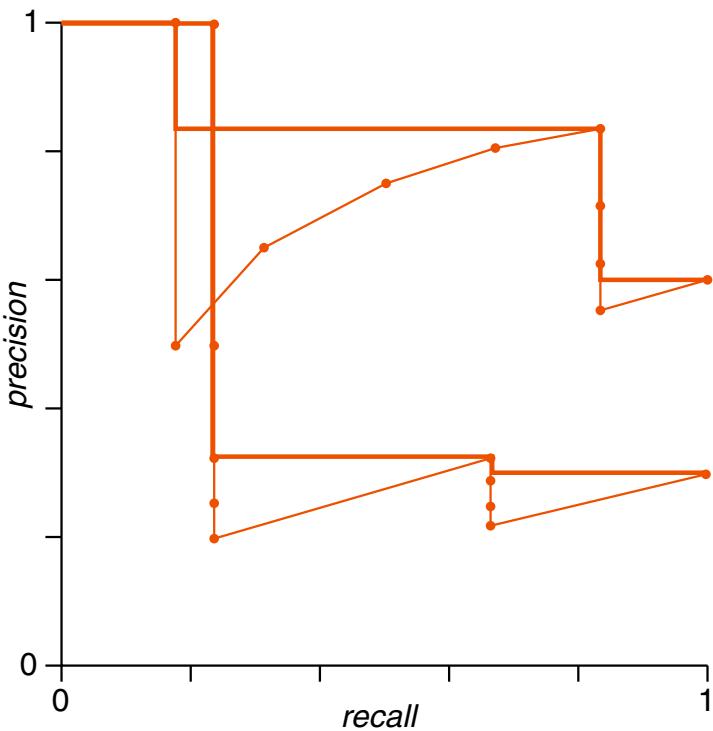
Ranking Effectiveness: Average Precision-Recall Curves



Effectiveness Measures

Ranking Effectiveness: Average Precision-Recall Curves

C	system 1 topic 2	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
precision											
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.

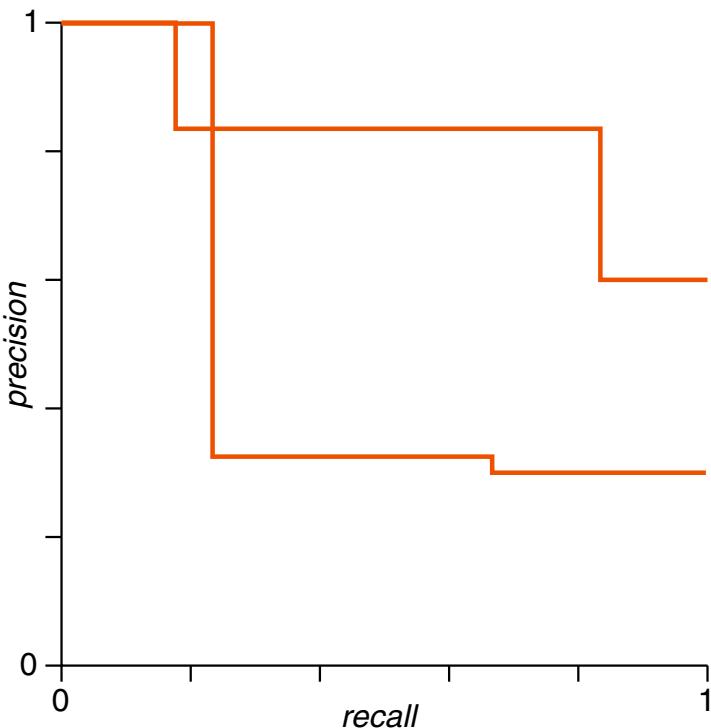
Solution:

- Compute averages across 11 recall points at 0.1 steps.

Effectiveness Measures

Ranking Effectiveness: Average Precision-Recall Curves

C system 1 topic 2	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
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.

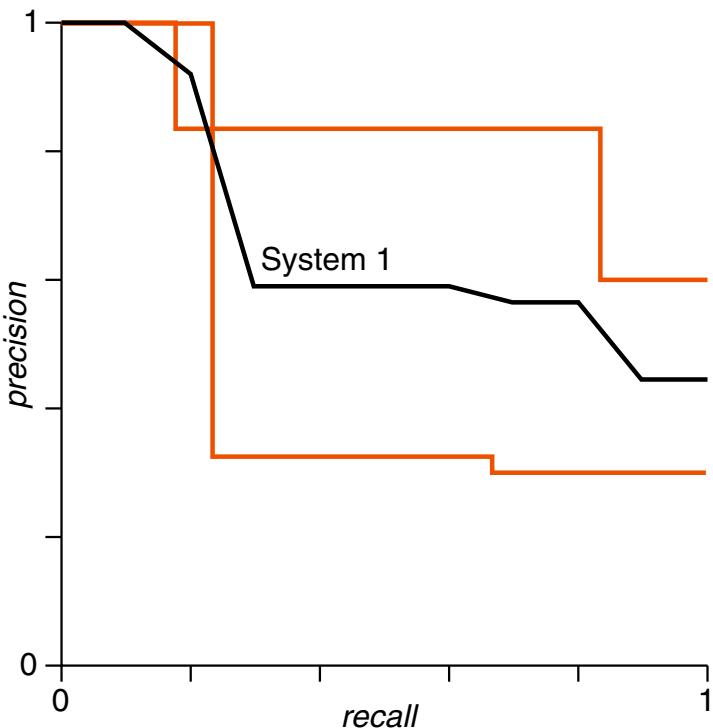
Solution:

- Compute averages across 11 recall points at 0.1 steps.

Effectiveness Measures

Ranking Effectiveness: Average Precision-Recall Curves

C	system 1 topic 2	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
precision											
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.

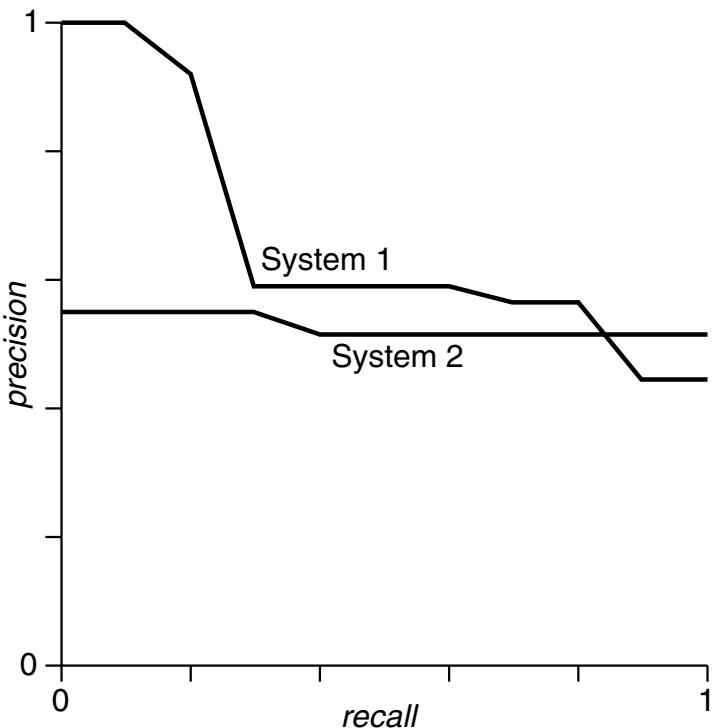
Solution:

- Compute averages across 11 recall points at 0.1 steps.

Effectiveness Measures

Ranking Effectiveness: Average Precision-Recall Curves

C	system 1 topic 2	precision	recall								
		1.00	0.33	0.50	0.33	0.33	0.33	0.25	0.25	0.20	0.22
											0.30



Interpretation:

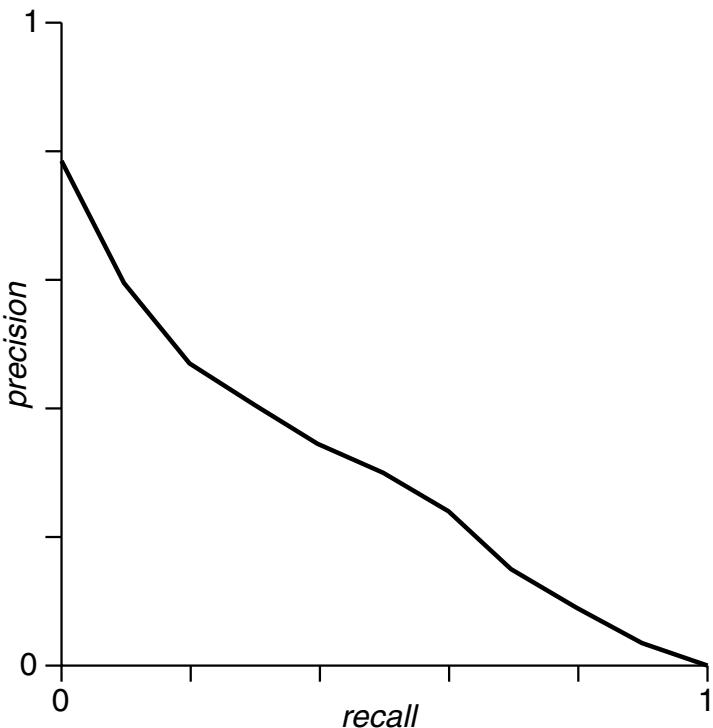
- Judging a system at various operating points.
- System 1 delivers very good precision at high ranks.
- System 2 delivers slightly better recall across rankings.
- Neither system dominates the other.

Curves are a lot smoother for 50 topics.

Effectiveness Measures

Ranking Effectiveness: Average 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.

Effectiveness 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.

Effectiveness 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 $r : D \times Q \rightarrow \{0, 1, 2, 3, 4, 5\}$ 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.

Effectiveness 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	10.37	17.87	18.44	19.05	20.70	24.36	28.05	35.75	36.46
Ideal $r^*(d_i, q)$	3	3	3	2	2	2	1	0	0	0
$DCG^*@k$	7.00	14.37	21.87	25.44	29.05	32.70	34.36	35.05	35.75	36.46
$NDCG@k$	1.00	0.72	0.82	0.72	0.66	0.63	0.71	0.80	1.00	1.00