

Chapter IR:V

V. Evaluation

- ❑ Laboratory Experiments
- ❑ Measuring Performance
- ❑ Set Retrieval Effectiveness
- ❑ Ranked Retrieval Effectiveness
- ❑ Training and Testing
- ❑ Logging

Laboratory Experiments

Experiment Scope

Interactive retrieval:

- ❑ Processing a query **depending** on other queries (of the user).
- ❑ The user has a goal or a task that requires many queries and exploration.
- ❑ Dependent variables are result quality, human factors, context, user interface and experience, and the retrieval system's supporting facilities.
- Experiments typically require user studies
- Measurement of retrieval performance depends on the setup

Laboratory Experiments

Experiment Scope

Interactive retrieval:

- ❑ Processing a query **depending** on other queries (of the user).
- ❑ The user has a goal or a task that requires many queries and exploration.
- ❑ Dependent variables are result quality, human factors, context, user interface and experience, and the retrieval system's supporting facilities.
- Experiments typically require user studies
- Measurement of retrieval performance depends on the setup

Ad hoc retrieval:

- ❑ Processing a query **independently** from other queries (of the user).
- Amenable to laboratory environments
- Canonical measurement of retrieval performance
- Reproducibility and scalability

Remarks:

- “ad hoc” (Latin: “for this”) means “concerned with a particular end or purpose” and “formed or used for specific or immediate problems or needs” [\[Merriam Webster\]](#)

Laboratory Experiments

Experimental Setup

A laboratory experiment for ad hoc retrieval requires three items:

1. A document collection (corpus)

- ❑ A representative sample of documents from the “search domain”: web, emails, tweets, ...
- ❑ If representativeness is difficult to achieve, the larger the sample, the better.

2. A set of information needs (topics)

- ❑ Formalized, written descriptions of users’ tasks, goals, or gaps of knowledge.
- ❑ Alternatively, declarative descriptions of desired search results.
- ❑ Often accompanied by specific queries the users (would) have used.

3. A set of relevance judgments (ground truth)

- ❑ Pairs of topics and documents, where each document has been manually assessed with respect to its relevance to the associated topic.
- ❑ Ideally, the users who supplied topics also judge, in practice third parties do so.
- ❑ Judgments may be given in binary form, or on a Likert scale.

Every retrieval system has parameters. Parameter optimization must use an experimental setup (training, validation) different from that used for evaluation (test).

Remarks:

- ❑ This setup is sometimes referred to as an experiment under the Cranfield paradigm, in reference to Cyril Cleverdon's series of projects at the Cranfield University in the 1960s, which first used this evaluation methodology. [\[codalism.com 1\]](#) [\[codalism.com 2\]](#)
- ❑ In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts. They are used to carry out statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. [\[Wikipedia\]](#)

The term has been adopted in various other branches of the human language technologies.
- ❑ The evaluation corpus split between training, validation, and test set should be used in conjunction with k -fold cross-validation, since the variance of performance results is often high. [\[Fuhr 2017\]](#)

Laboratory Experiments

Experimental Setup: Document Collections / Corpora

For ad hoc retrieval, the [Text Retrieval Conference \(TREC\)](#) has organized evaluation tracks since 1992, inviting scientists to compete.

Key document collections used (many more at [ir_datasets](#)):

Collection	Documents	Size	Words/Doc.	Topics	Words/Query	Jdgmts/Query
CACM	3,204	2.2 MB	64	64	13.0	16
AP	242,918	0.7 GB	474	100	4.3	220
GOV2	25 million	426.0 GB	1073	150	3.1	180
ClueWeb09	1 billion	25.0 TB	459	200	2.5	821
ClueWeb12	733 million	27.3 TB	448	200	3.6	793
ClueWeb22B	200 million	11.7 TB	–	–	–	–

- ❑ CACM: titles and abstracts from Communications of the ACM 1958–1979
- ❑ AP: newswire documents from Associated Press 1988–1990
- ❑ GOV2: crawl of .gov domains early 2004
- ❑ ClueWeb: web crawls from 2009, 2012, and 2022 (not in use, yet)

Reusing experimental setups renders previous approaches comparable.

Remarks:

- ❑ TREC is organized by the United States National Institute of Standards and Technology (NIST). The conference has been key to popularize laboratory evaluation of retrieval systems; every year, evaluation tracks on [many different retrieval-related tasks](#) are organized.
- ❑ At TREC, usually 50 topics are provided per edition of a shared task. The ones from previous years can be used for training.
- ❑ Ad hoc retrieval has been studied in the [ad hoc tracks](#), the [terabyte tracks](#), and the [web tracks](#).
- ❑ Several initiatives similar to TREC have formed, namely [CLEF](#), [NTCIR](#), and [FIRE](#).

Laboratory Experiments

Experimental Setup: Topics

```
<topic number="794" type="single">
```

```
<query> pet therapy </query>
```

```
<description>
```

```
How are pets or animals used in therapy for humans and what are the benefits?
```

```
</description>
```

```
<narrative>
```

```
Relevant documents must include details of how pet or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.
```

```
</narrative>
```

```
</topic>
```

Remarks:

- ❑ The description element is a longer version of the query, clarifying it, since the short query itself may be ambiguous.
- ❑ The narrative field is optional. It usually describes the criteria for relevance and is used by assessors to carry out relevance judgments.
- ❑ Another topic type are faceted topics:

```
<topic number="265" type="faceted">
  <query>F5 tornado</query>
  <description>What were the ten worst tornadoes in the USA?</description>
  <subtopic number="1" type="inf">What were the ten worst tornadoes in the USA?</subtopic>
  <subtopic number="2" type="inf">Where is tornado alley?</subtopic>
  <subtopic number="3" type="inf">What damage can an F5 tornado do?</subtopic>
  <subtopic number="4" type="inf">Find information on tornado shelters.</subtopic>
  <subtopic number="5" type="nav">What wind speed defines an F5 tornado?</subtopic>
</topic>
```

Laboratory Experiments

Experimental Setup: Relevance Judgments

```
<topic number="794" type="single">
```

```
<query> pet therapy </query>
```

```
<description>
```

```
How are pets or animals used in therapy for humans and what are the benefits?
```

```
</description>
```

```
<narrative>
```

```
Relevant documents must include details of how pet or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.
```

```
</narrative>
```

```
</topic>
```

Laboratory Experiments

Experimental Setup: Relevance Judgments

<topic number="794" type="single">

<query> pet therapy </query>

<description>

How are pets or animals used in therapy for humans and what are the benefits?

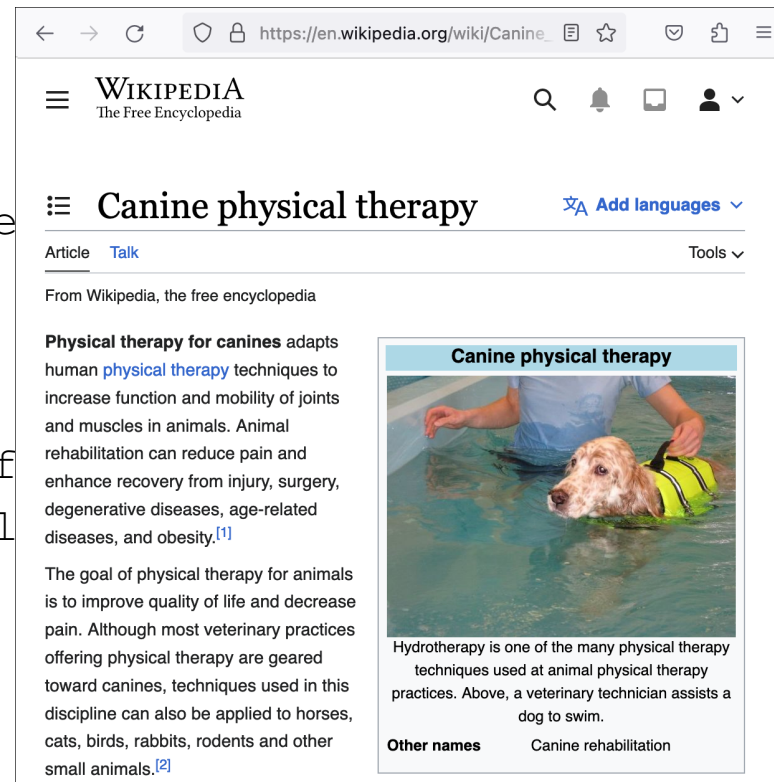
</description>

<narrative>

Relevant documents must include details about animal-assisted therapy is or has details include information about descriptions of the circumstances used, the benefits of this type of success of this therapy, and any 1 governing it.

</narrative>

</topic>



The screenshot shows the Wikipedia article for "Canine physical therapy". The page title is "Canine physical therapy" with a subtitle "The Free Encyclopedia". The article text states: "Physical therapy for canines adapts human physical therapy techniques to increase function and mobility of joints and muscles in animals. Animal rehabilitation can reduce pain and enhance recovery from injury, surgery, degenerative diseases, age-related diseases, and obesity.[1]". Below the text is a photograph of a dog in a pool with a person assisting it. The caption for the photo reads: "Hydrotherapy is one of the many physical therapy techniques used at animal physical therapy practices. Above, a veterinary technician assists a dog to swim." Below the photo, it lists "Other names" as "Canine rehabilitation".

Laboratory Experiments

Experimental Setup: Relevance Judgments

<topic number="794" type="single">

<query> pet therapy </query>

<description>

How are pets or animals used in therapy for humans and what are the benefits?

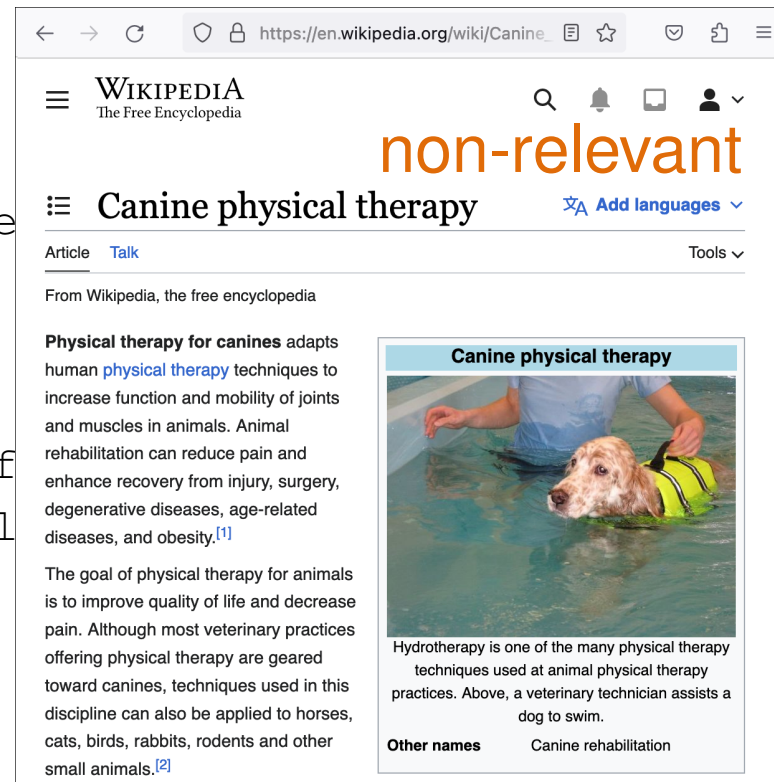
</description>

<narrative>

Relevant documents must include details about **animal-assisted therapy** is or has details include information about descriptions of the circumstances used, the benefits of this type of success of this therapy, and any 1 governing it.

</narrative>

</topic>



The screenshot shows a web browser displaying the Wikipedia article for "Canine physical therapy". The URL in the address bar is "https://en.wikipedia.org/wiki/Canine". The page features the Wikipedia logo and navigation icons. A large orange text "non-relevant" is overlaid on the top right of the article content. The article title "Canine physical therapy" is prominently displayed, with a blue "Add languages" link next to it. Below the title, there are tabs for "Article" and "Talk", and a "Tools" dropdown menu. The main text of the article begins with "Physical therapy for canines adapts human physical therapy techniques to increase function and mobility of joints and muscles in animals. Animal rehabilitation can reduce pain and enhance recovery from injury, surgery, degenerative diseases, age-related diseases, and obesity.[1]". A photograph of a dog in a pool is included, with the caption "Canine physical therapy". Below the photo, there is a paragraph explaining hydrotherapy and its application in animal physical therapy. At the bottom, there is a section for "Other names" which lists "Canine rehabilitation".

Laboratory Experiments

Experimental Setup: Relevance Judgments

<topic number="794" type="single">

<query> pet therapy </query>

<description>

How are pets or animals used in therapy for humans and what are the benefits?

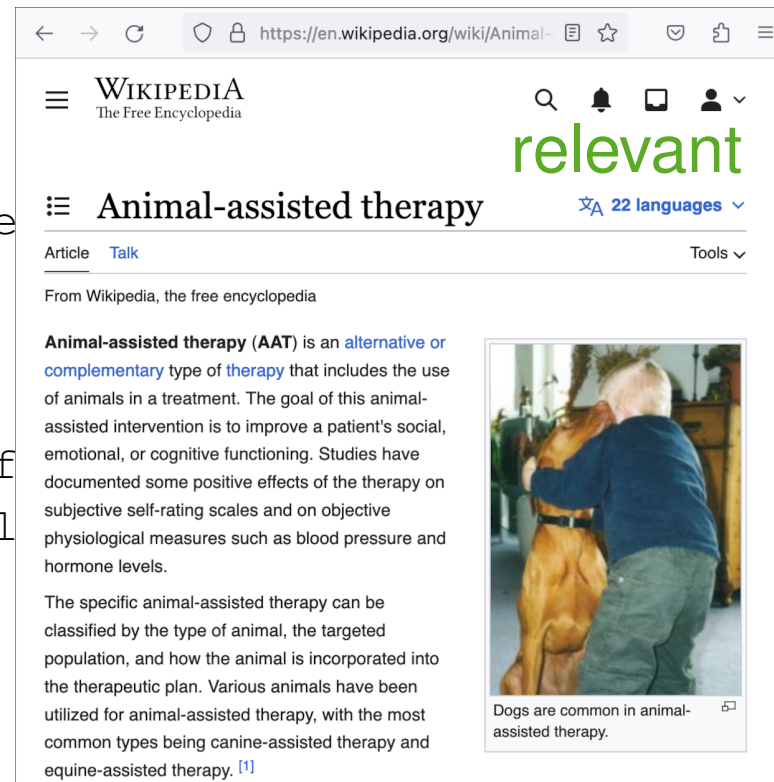
</description>

<narrative>

Relevant documents must include details about **animal-assisted therapy** is or has details include information about descriptions of the circumstances used, the benefits of this type of success of this therapy, and any governing it.

</narrative>

</topic>



The screenshot shows a web browser displaying the Wikipedia article for "Animal-assisted therapy". The browser's address bar shows the URL "https://en.wikipedia.org/wiki/Animal-". The Wikipedia logo and "The Free Encyclopedia" text are visible at the top. A green "relevant" label is overlaid on the right side of the page. The article title "Animal-assisted therapy" is prominently displayed, with a "22 languages" dropdown menu next to it. Below the title, there are links for "Article" and "Talk", and a "Tools" dropdown menu. The main text of the article begins with "From Wikipedia, the free encyclopedia" and defines "Animal-assisted therapy (AAT)" as an alternative or complementary type of therapy that includes the use of animals in a treatment. It states that the goal is to improve a patient's social, emotional, or cognitive functioning. A photograph on the right side of the article shows an elderly man in a blue shirt and grey pants hugging a large brown dog. Below the photo is a caption: "Dogs are common in animal-assisted therapy." A footnote [1] is visible at the bottom of the article text.

Laboratory Experiments

Experimental Setup: Relevance Judgments

A relevance judgment requires the **manual** assessment of whether a document returned by a retrieval system for a given query is relevant for a given topic.

Assessment depth: At what rank k should documents no longer be judged?

- ❑ Assessment does not scale with the number of documents retrieved by retrieval systems.
- ❑ A sampling strategy called pooling is used.

Assessment scale: How many degrees of relevance can be distinguished?

- ❑ Binary scale: relevant and non-relevant
- ❑ n -point Likert scale of degrees of relevance: from non-relevant (0) to highly relevant ($n \leq 5$)

Assessor selection and instruction: Are the assessors sufficiently qualified?

- ❑ The people who had the information needs underlying the topics, if available.
- ❑ Volunteer assessors who receive training and exhaustive topics.

Assessor reliability: Are similar documents judged similar for a topic?

- ❑ Assessors make errors, which affects the objectivity of the results.
- ❑ Multiple assessments can be used to verify the reliability of assessors and assessments.

Remarks:

- At TREC, assessors are recruited from retired NIST staff:



Laboratory Experiments

Experimental Setup: Pooling

Given a set of retrieval systems, each indexing the same corpus, and a set of topics but no relevance judgments. Then pooling selects the documents to be assessed.

For each topic:

1. Collect the top- k results returned by each retrieval system (variant).
2. Merge the results, omitting duplicates, obtaining a “pool” of documents.
3. Present the pool of documents in random order to assessors along the topic.

Caveats:

- ❑ Self-selection bias: Only documents “considered” relevant enough by one of the retrieval systems are assessed.
- ❑ Unknown recall: All documents ranked below the pooling depth are deemed non-relevant by default, regardless the truth.
- ❑ Laborious extensibility: New retrieval systems that are evaluated later may retrieve documents not in the original pool.

Laboratory Experiments

Assessor Reliability

The degree of agreement between assessors and the degree of consistency of the same assessor are quantified using assessor reliability measures. Lack of agreement or consistency indicate flawed setups or insufficient training.

Assessor reliability is measured whenever ambiguous or subjective decisions have to be made. Relevance is a subjective notion.

Several alternative approaches have been proposed:

- ❑ **Joint probability of agreement**
Percentage of times the raters agree. Here, agreement by chance is not taken into account.
- ❑ **Kappa Statistics**
Improvement over joint probability, taking into account agreement by chance.
- ❑ **Correlation coefficient**
Pairwise correlation among assessors on ordered scales. Full rankings are required.

Laboratory Experiments

Assessor Reliability: Kappa Statistics

Given the judgments of two annotators on a given topic, a kappa statistic measures their agreement as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o denotes the proportion of agreement observed, and p_e the expected proportion of agreement by chance.

Properties:

- $\kappa \in (-\infty, 1]$, where 1 indicates perfect agreement, 0 random agreement, and $\kappa < 0$ has no meaningful interpretation [Kvålseth 2015]
- At $p_e = 1$, κ is undefined
- $p_o - p_e$ denotes the agreement **attained** above chance
- $1 - p_e$ denotes the agreement **attainable** above chance

Laboratory Experiments

Assessor Reliability: Kappa Statistics

Given the judgments of two annotators on a given topic, a kappa statistic measures their agreement as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o denotes the proportion of agreement observed, and p_e the expected proportion of agreement by chance.

Suppose A and B are two annotators asked to make n binary relevance judgments. Then a basic kappa statistic can be computed as follows:

		B		Σ
		yes	no	
A	yes	a	b	c
	no	d	e	f
Σ		g	h	n

$$p_o = \frac{a + e}{n}$$

$$p_e = P(\text{yes})^2 + P(\text{no})^2$$

$$P(\text{yes}) = \frac{c + g}{2n}, \quad P(\text{no}) = \frac{f + h}{2n}$$

Laboratory Experiments

Assessor Reliability: Kappa Statistics

Given the judgments of two annotators on a given topic, a kappa statistic measures their agreement as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o denotes the proportion of agreement observed, and p_e the expected proportion of agreement by chance.

Suppose A and B are two annotators who made the following $n = 400$ binary relevance judgments. The basic kappa statistic then yields:

		B		Σ
		yes	no	
A	yes	300	20	320
	no	10	70	80
Σ		310	90	400

$$p_o = \frac{300 + 70}{400}$$

$$p_e = P(\text{yes})^2 + P(\text{no})^2$$

$$P(\text{yes}) = \frac{320 + 310}{2 \cdot 400}, \quad P(\text{no}) = \frac{80 + 90}{2 \cdot 400}$$

$$\kappa = 0.776$$

Remarks:

- Well-known kappa statistics include Cohen’s κ , Scott’s π , and Fleiss’ κ .
- Scott’s π is the one exemplified.
- Fleiss’ κ is a generalization of Scott’s π to arbitrary numbers of annotators and categories. It also does not presume that all cases have been annotated by the same group of people.
- Presuming that annotators A and B work independently, the probability $P(\text{yes})^2$ (and similarly $P(\text{no})^2$) denotes the probability of both voting yes (no) by chance. Another way of computing p_e is to sum the multiplication of the rater-specific probabilities of each rater voting yes (no).
- Some assign the following interpretations to κ values measured (disputed):

κ	Agreement
< 0	poor
0.01 – 0.20	slight
0.21 – 0.40	fair
0.41 – 0.60	moderate
0.61 – 0.80	substantial
0.81 – 1.00	almost perfect

[\[Wikipedia\]](#)

- Within TREC evaluations, typically a “substantial” agreement ($\kappa \approx [0.67, 0.8]$) is achieved.

[Manning 2008]

Chapter IR:V

V. Evaluation

- ❑ Laboratory Experiments
- ❑ Measuring Performance
- ❑ Set Retrieval Effectiveness
- ❑ Ranked Retrieval Effectiveness
- ❑ Training and Testing
- ❑ Logging

Measuring Performance

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
- ❑ index size
- ❑ Query throughput
- ❑ query latency

Measuring Performance

Effectiveness Measures

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

The **desired result** from a retrieval system for a user’s query is relevant documents.

Our goal is to make justifiable claims such as these:

- ❑ This retrieval system is (not) effective.
- ❑ Retrieval system A is (x times) more effective than retrieval system B.
- ❑ This retrieval system achieves the highest effectiveness for its domain.

Measuring Performance

Effectiveness Measures

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

The desired result from a retrieval system for a user’s query is relevant documents.

Our goal is to make **justifiable** claims such as these:

- ❑ This retrieval system is (not) effective.
- ❑ Retrieval system A is (x times) more effective than retrieval system B.
- ❑ This retrieval system achieves the highest effectiveness for its domain.

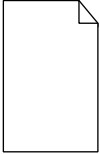
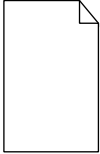





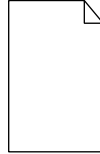
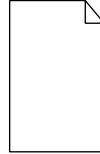

Sufficient justification is achieved by means of measurement, namely “the assignment of a number to a characteristic of an object [a retrieval result], which can be compared with other objects.” [\[Wikipedia\]](#)

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

Measuring Performance

Effectiveness Measures

The object of measurement for a retrieval system's effectiveness are its rankings:

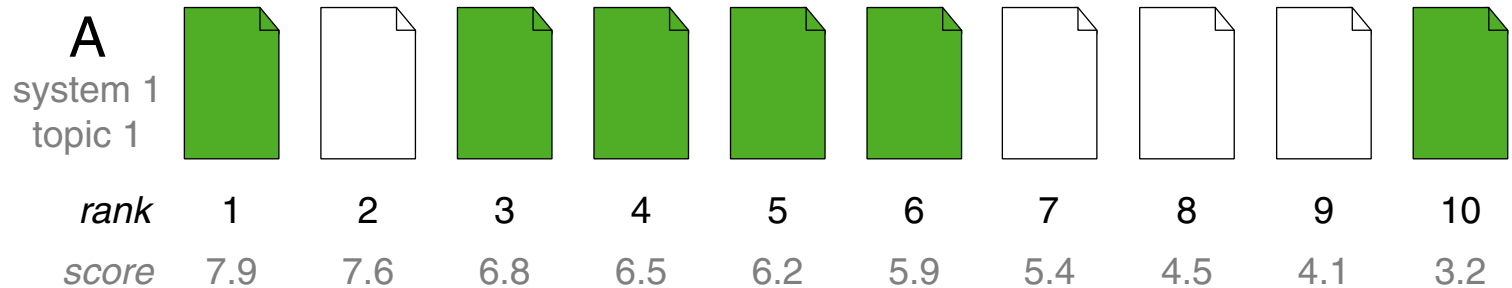
A system 1 topic 1										
<i>rank</i>	1	2	3	4	5	6	7	8	9	10
<i>score</i>	7.9	7.6	6.8	6.5	6.2	5.9	5.4	4.5	4.1	3.2

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

Measuring Performance

Effectiveness Measures

The object of measurement for a retrieval system's effectiveness are its rankings:



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

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

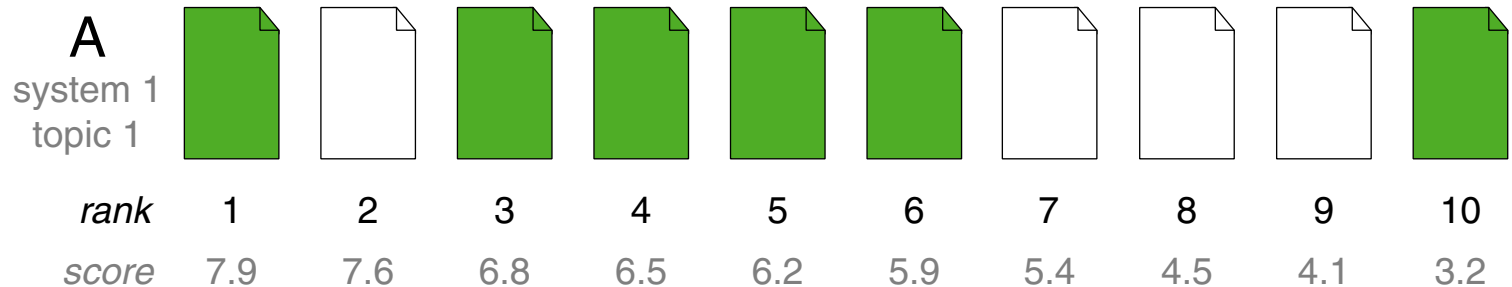
An effectiveness measure maps a given retrieval result and its relevance judgments to the real numbers, rendering rankings from different systems comparable.

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

Measuring Performance

Effectiveness Measures

The object of measurement for a retrieval system's effectiveness are its rankings:



A retrieval result is composed of a list of documents, ordered by the system'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

Set Retrieval Effectiveness

Precision and Recall

The user browses the entire result set returned by the retrieval system, but expects only relevant documents. A contingency table counts successes and failures:

	\in <i>Relevant</i>	\notin <i>Relevant</i>
\in <i>Results</i>	a	b
\notin <i>Results</i>	c	d

with

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

Set Retrieval Effectiveness

Precision and Recall

The user browses the entire result set returned by the retrieval system, but expects only relevant documents. A contingency table counts successes and failures:

	$\in \textit{Relevant}$	$\notin \textit{Relevant}$	
$\in \textit{Results}$	a	b	\rightarrow
$\notin \textit{Results}$	c	d	

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

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

with

- $\textit{Results}$ = set of documents retrieved.
- $\textit{Relevant}$ = 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:

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

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

- ❑ 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 simply returning the entire document collection—not that helpful, though.
- ❑ The fraction of non-relevant documents that are retrieved is called

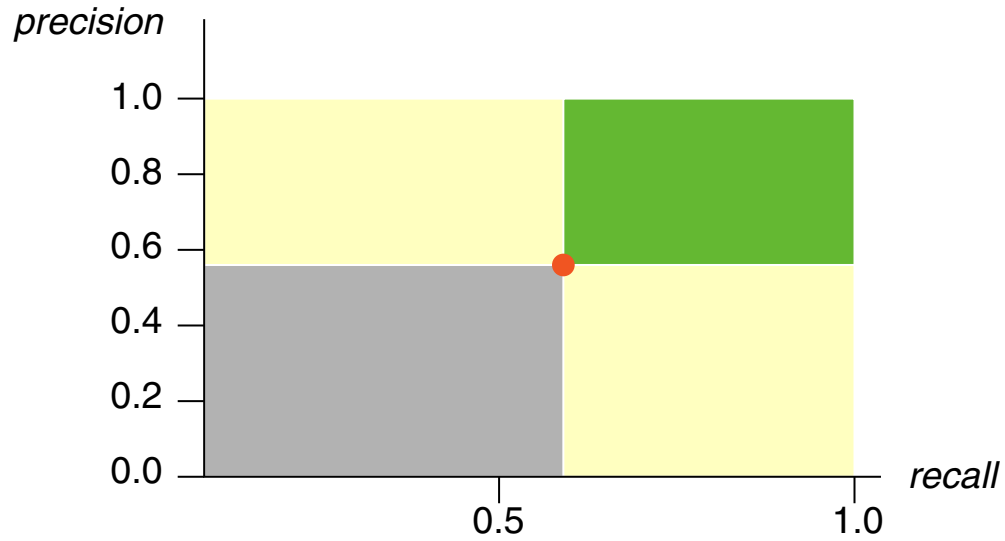
$$fallout = \frac{b}{b + d}.$$

If retrieval were a classification task, *recall* would be considered the true positive rate and *fallout* the false positive rate.

Set Retrieval Effectiveness

F -Measure

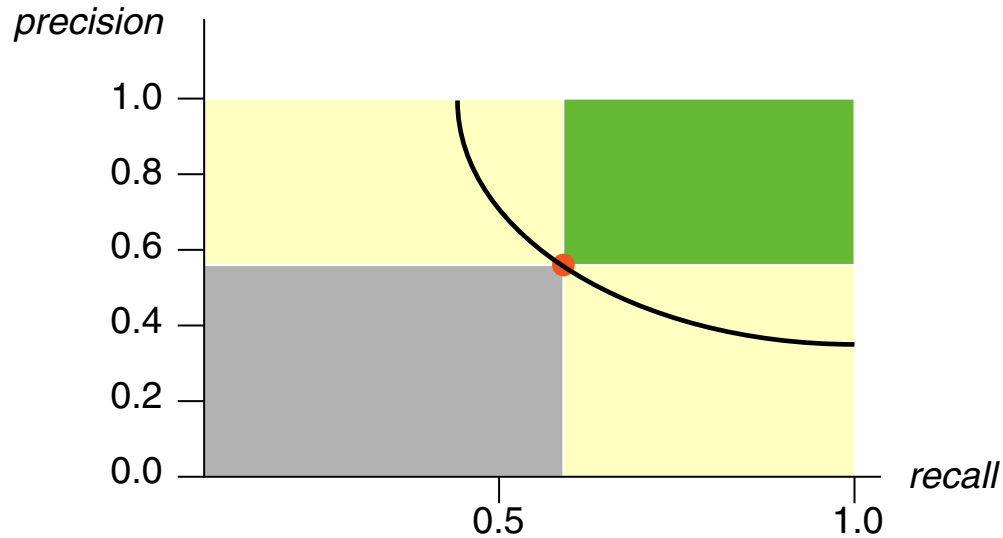
Comparison of retrieval systems: [\[plot\]](#)



Set Retrieval Effectiveness

F-Measure

Comparison of retrieval systems: [\[plot\]](#)



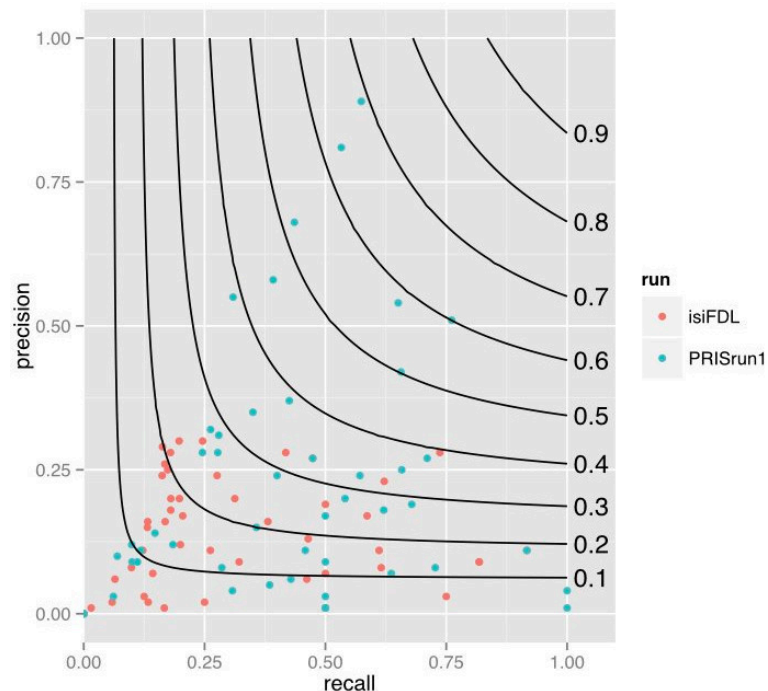
The *F*-Measure is the harmonic mean of *precision* and *recall*:

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

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 effectiveness 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 “isocurves” (points with same value) also better resemble trade-offs human users might be willing to take when trading recall for precision, or vice versa.

When two systems have similar F -Measure scores (e.g., is a 0.29 system really better than a 0.27 system?) also the per-topic precision and recall values in a scatterplot with the 0.1, 0.2, 0.3, ... F -Measure isocurves and the retrieval task actually are important comparison parameters. [\[Soboroff 2019\]](#)



Remarks (ctd.):

- Precision and recall are not equally important in all retrieval tasks. Examples: Web search (high precision) vs. intellectual property search (high recall). A weighted F -Measure computes as follows:

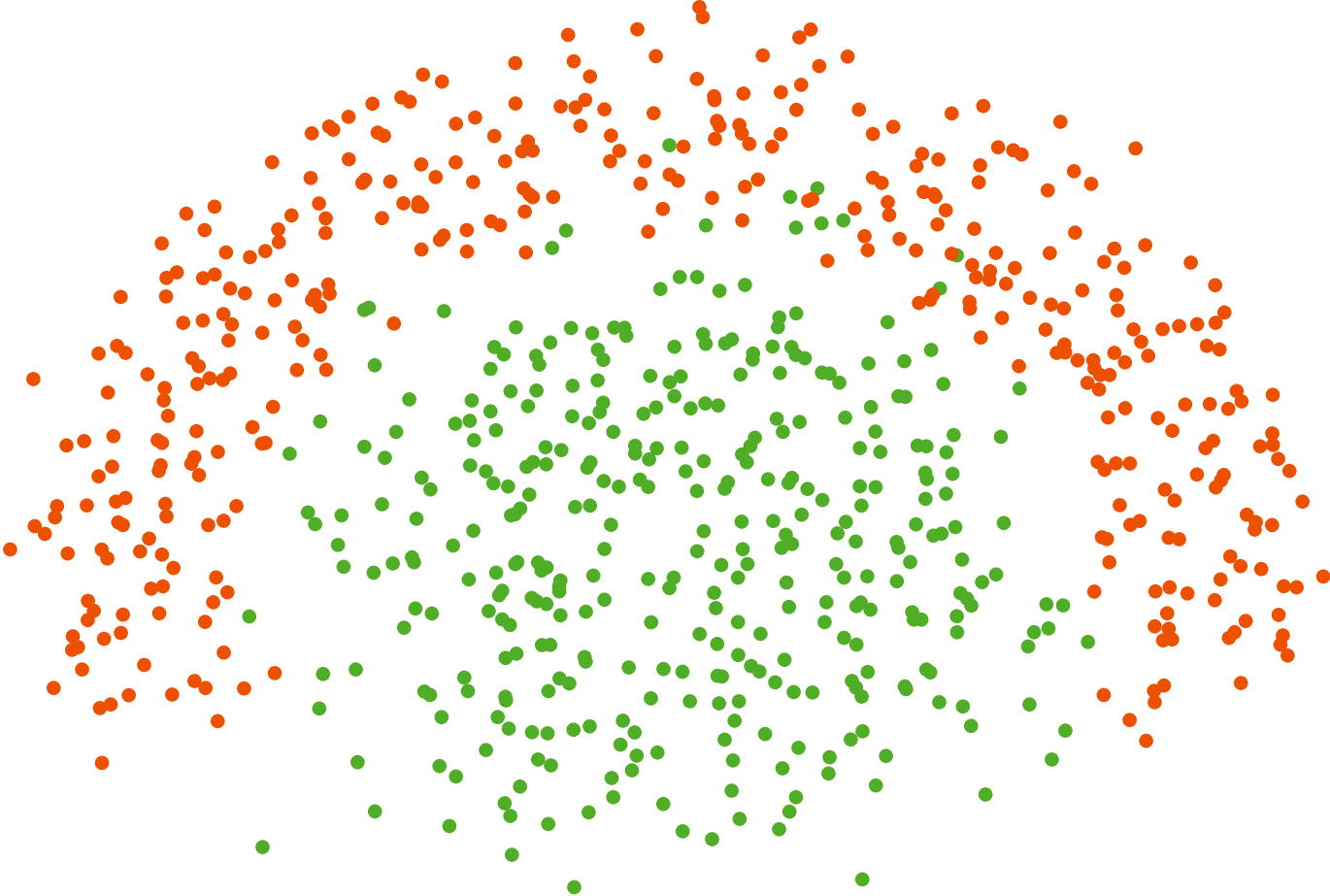
$$F = \frac{1}{\alpha \frac{1}{precision} + (1 - \alpha) \frac{1}{recall}} = \frac{(\beta^2 + 1)precision \cdot recall}{\beta^2 precision + recall}, \text{ where } \beta^2 = \frac{1 - \alpha}{\alpha}.$$

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.

Set Retrieval Effectiveness

Illustration

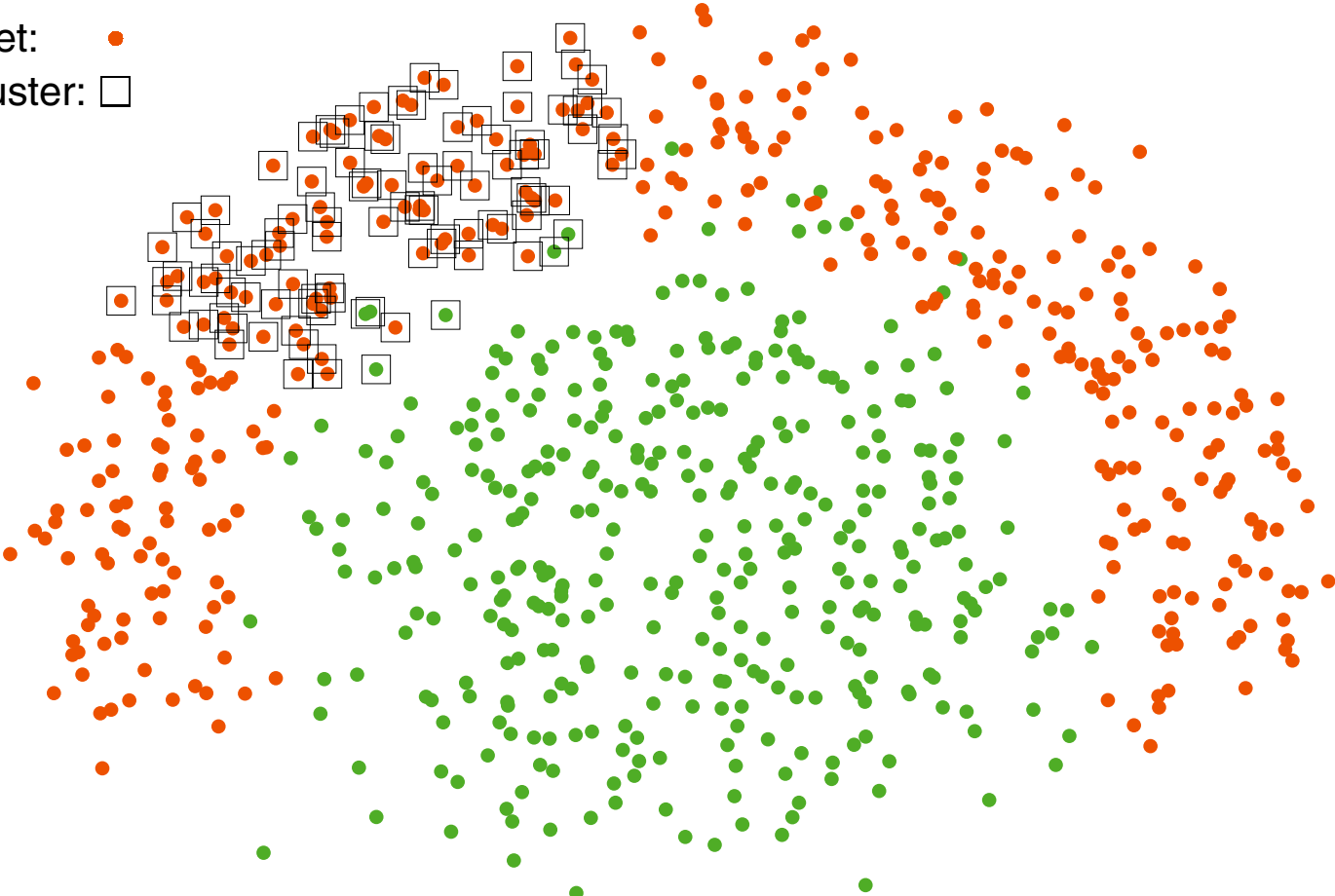
Classes: ● ●



Set Retrieval Effectiveness

Illustration

Classes: ● ●
Target: ●
In cluster: □



Recall $\frac{\text{orange squares}}{\text{orange circles}} = 0.26$

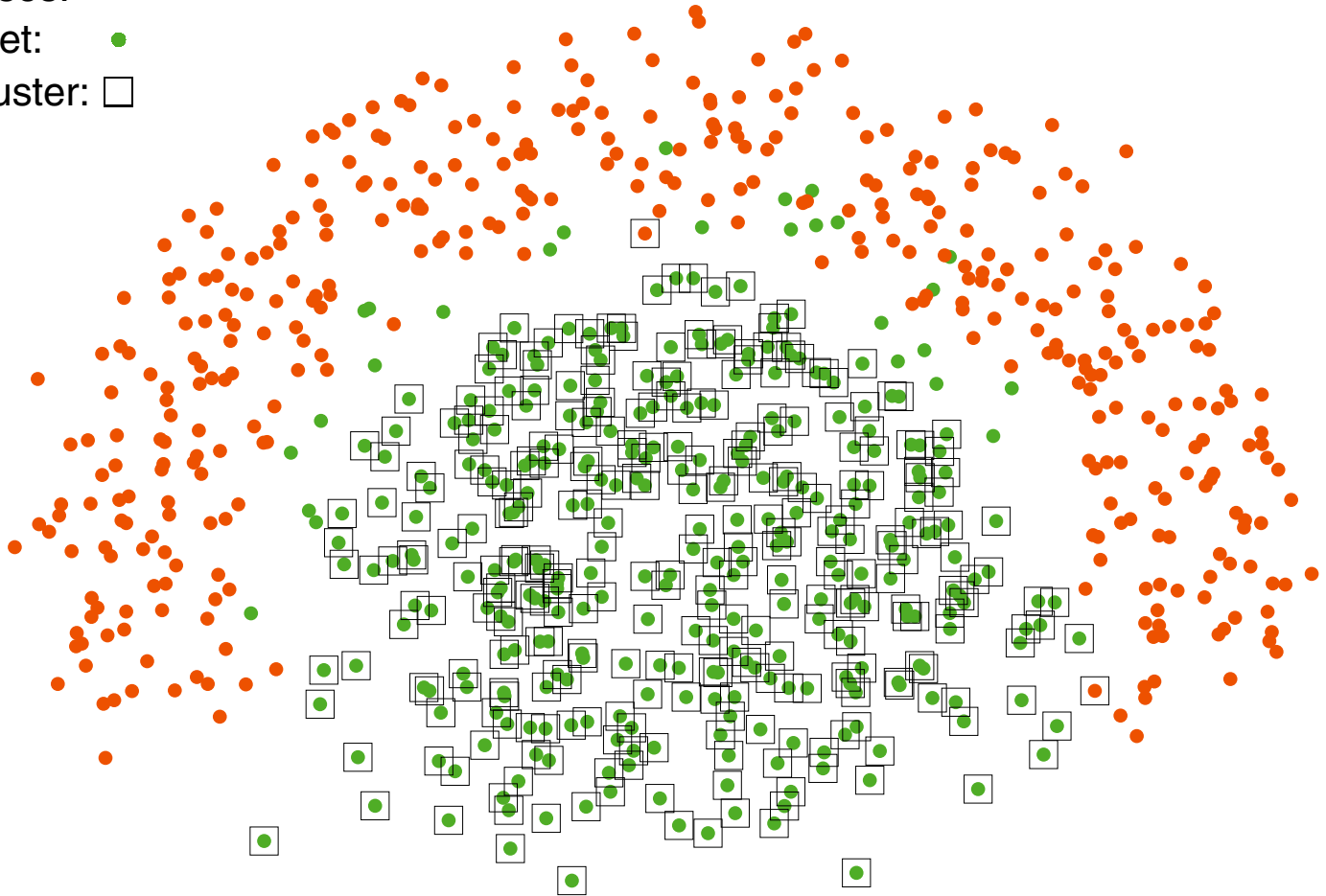
Precision $\frac{\text{orange squares}}{(\text{orange squares} \cup \text{green squares})} = 0.94$

F-Measure = 0.40

Set Retrieval Effectiveness

Illustration

Classes: ● ●
Target: ●
In cluster: □



Recall $\frac{\text{In cluster} \cap \text{Target}}{\text{Target}} = 0.92$

Precision $\frac{\text{In cluster} \cap \text{Target}}{\text{In cluster}} = 0.99$

F-Measure = 0.95

Set Retrieval Effectiveness

Precision and Recall Averaging

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

Macro-averaging: (user-oriented)

$$precision_{macro} = \frac{1}{|Q|} \sum_{q \in Q} precision_q$$

$$recall_{macro} = \frac{1}{|Q|} \sum_{q \in Q} recall_q$$

Macro-averaging gives equal importance to each topic.

Set Retrieval Effectiveness

Precision and Recall Averaging

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

Macro-averaging: (user-oriented)

$$precision_{macro} = \frac{1}{|Q|} \sum_{q \in Q} \frac{a_q}{a_q + b_q}$$

$$recall_{macro} = \frac{1}{|Q|} \sum_{q \in Q} \frac{a_q}{a_q + c_q}$$

Macro-averaging gives equal importance to each topic.

Set Retrieval Effectiveness

Precision and Recall Averaging

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

Macro-averaging: (user-oriented)

$$precision_{macro} = \frac{1}{|Q|} \sum_{q \in Q} \frac{a_q}{a_q + b_q}$$

$$recall_{macro} = \frac{1}{|Q|} \sum_{q \in Q} \frac{a_q}{a_q + c_q}$$

Macro-averaging gives equal importance to each topic.

Micro-averaging: (system-oriented)

$$precision_{micro} = \frac{\sum_{q \in Q} a_q}{\sum_{q \in Q} a_q + b_q}$$

$$recall_{micro} = \frac{\sum_{q \in Q} a_q}{\sum_{q \in Q} a_q + c_q}$$

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 (class-level) number of students per class is

$$50 = \frac{1 + 1 + 1 + 1 + 1 + 99 + 99 + 99 + 99 + 99}{10} .$$

- The micro-average (student-level) number of students per class is

$$98.02 = \frac{1 + 1 + 1 + 1 + 1 + 99 \cdot 5 \cdot 99}{500} ,$$

since almost all of the 500 (not necessarily distinct) student “instances” are in classes with 99 students (in these 5 courses, 99 students “see” a course with 99 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 retrieval system 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 retrieval system’s parameters are set globally.
- Macro-averaging, the user-oriented view, is preferred for most search domains.

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

- ❑ Execution of a set of paradigmatically different retrieval systems tuned by experts.
- ❑ Pooling of the systems' top- k ranked documents, followed by optional relevance judgment.
- ❑ Without judgments, documents retrieved by more than m systems are pseudo-relevant.

Sample analysis

- ❑ High class imbalance: Typically, only a small fraction of documents are relevant.
- ❑ Drawing a representative sample from a small subpopulation requires a large sample size.

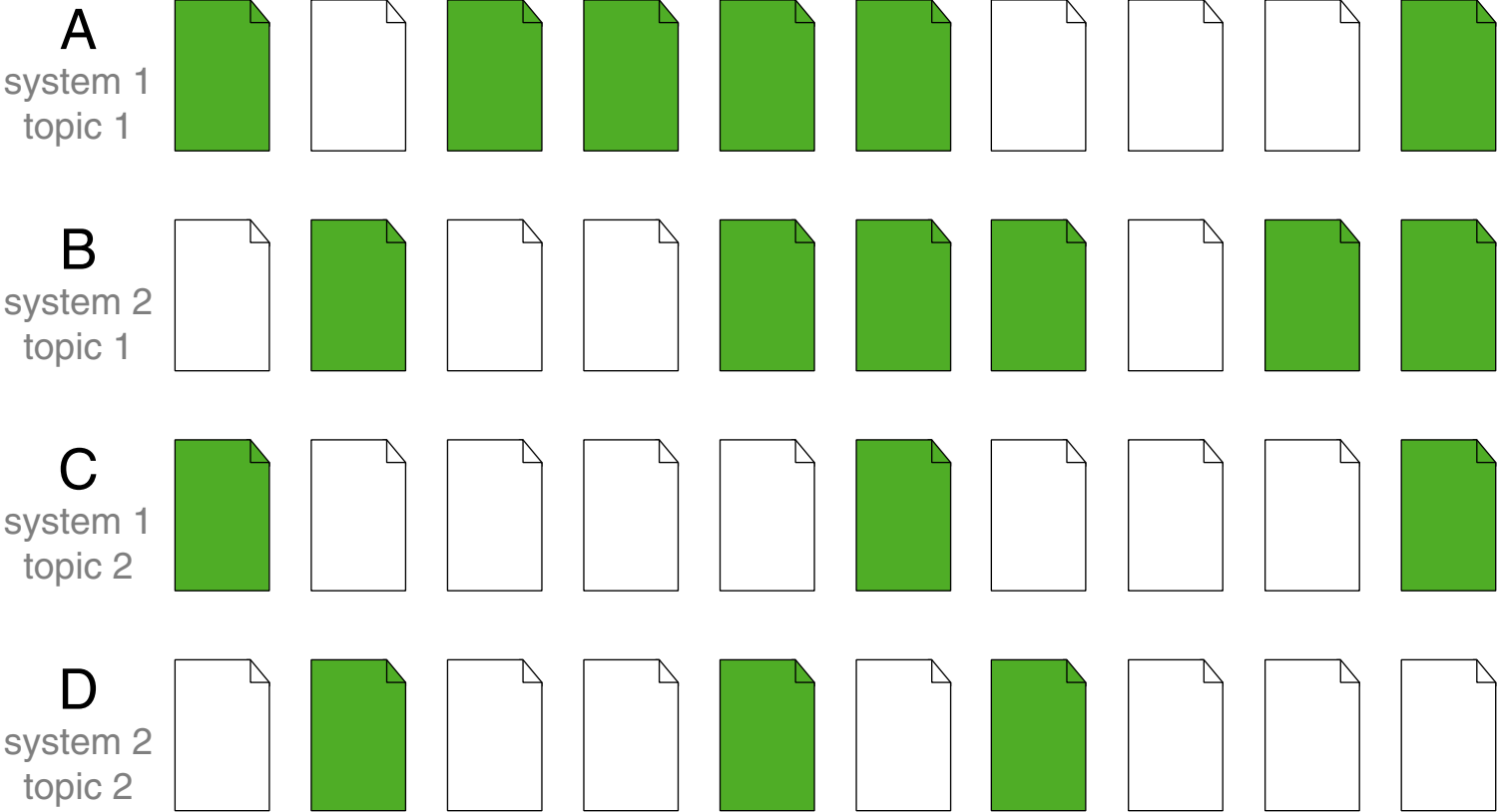
Query rewriting via relevance feedback

- ❑ Collection of relevance judgments down to rank k .
- ❑ Iterative query rewriting based on relevant documents to find more to be judged.

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

Ranked Retrieval Effectiveness

Example

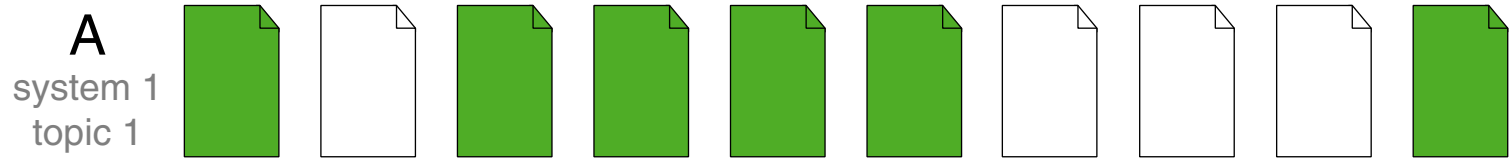


Which system is better? They achieve equal *precision* and *recall* for Topics 1 and 2.

How good is System 1 compared to System 2 overall?

Ranked Retrieval Effectiveness

Precision@k and Recall@k

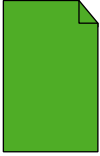
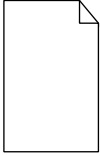
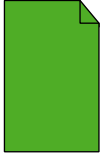




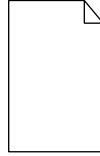
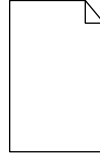



Assumption:

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

Ranked Retrieval Effectiveness

Precision@k and Recall@k

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

Assumption:

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

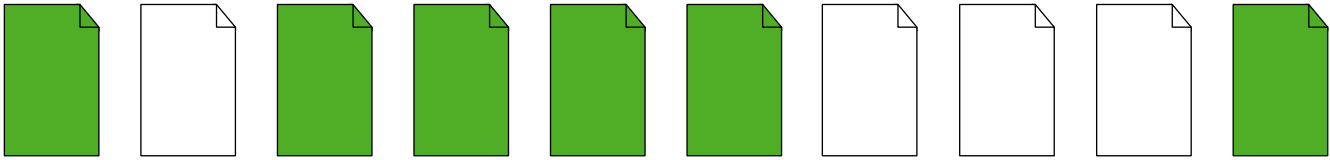
Caveats:

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

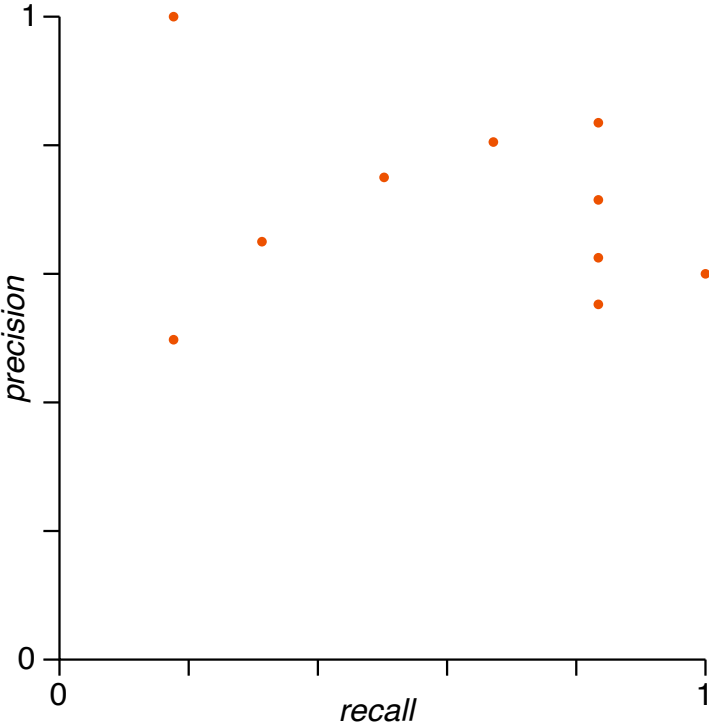
Ranked Retrieval Effectiveness

Precision-Recall Curves

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



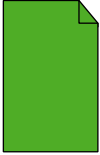
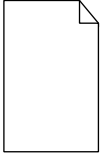
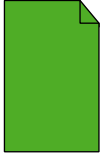




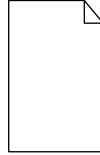
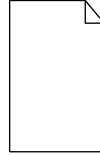

Observations:

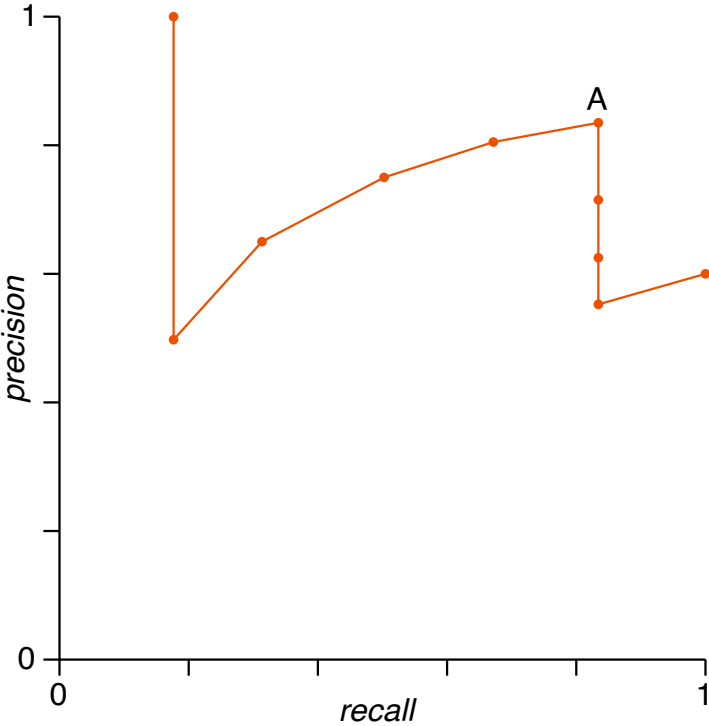
- Connecting the dots yields a “curve.”
- The curve captures detailed ranking characteristics: the user experience.
- Points on a curve other than the original ones lack interpretation.
- Given rankings from two systems, we can decide which one is better.
- These observations can be quantified as area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

Ranked Retrieval Effectiveness

Precision-Recall Curves

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



Observations:

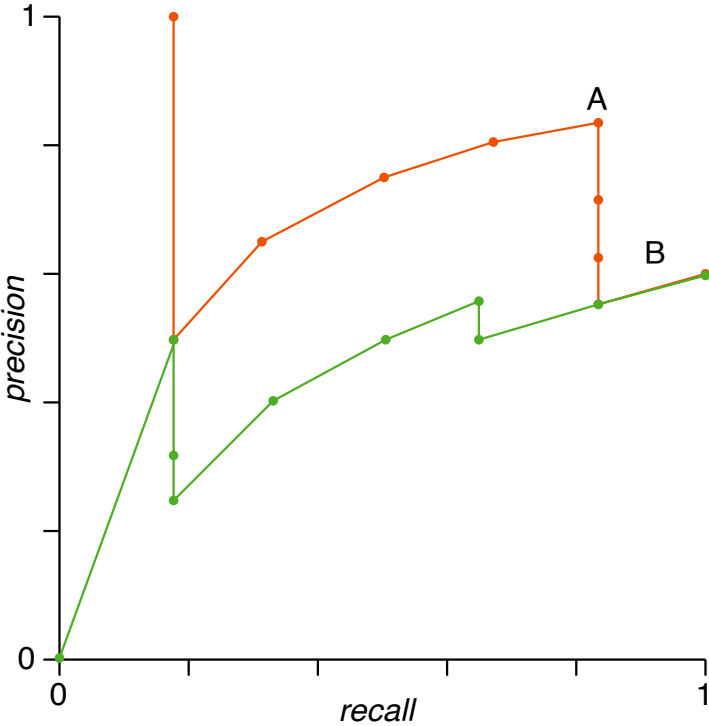
- Connecting the dots yields a “curve.”
- The curve captures detailed ranking characteristics: the user experience.
- Points on a curve other than the original ones lack interpretation.
- Given rankings from two systems, we can decide which one is better.
- These observations can be quantified as area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

Ranked Retrieval Effectiveness

Precision-Recall Curves

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



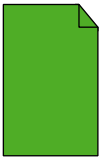
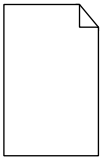
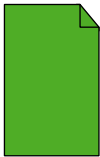




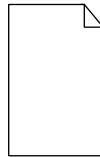
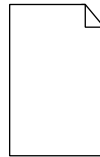

Observations:

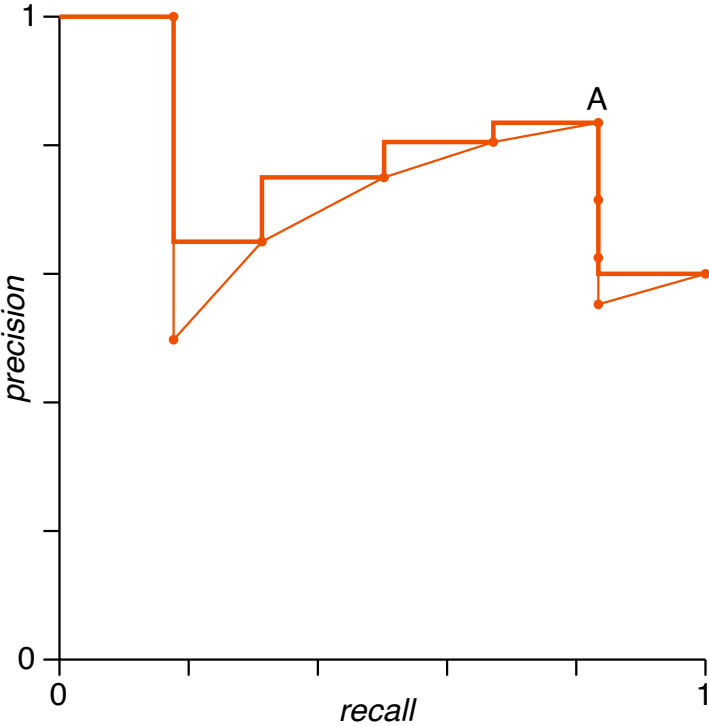
- Connecting the dots yields a “curve.”
- The curve captures detailed ranking characteristics: the user experience.
- Points on a curve other than the original ones lack interpretation.
- Given rankings from two systems, we can decide which one is better.
- These observations can be quantified as area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

Ranked Retrieval Effectiveness

Average Precision

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



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

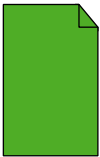
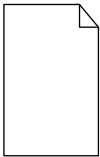
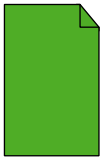




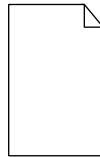
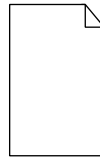

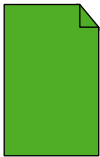
Interpolation alternatives:

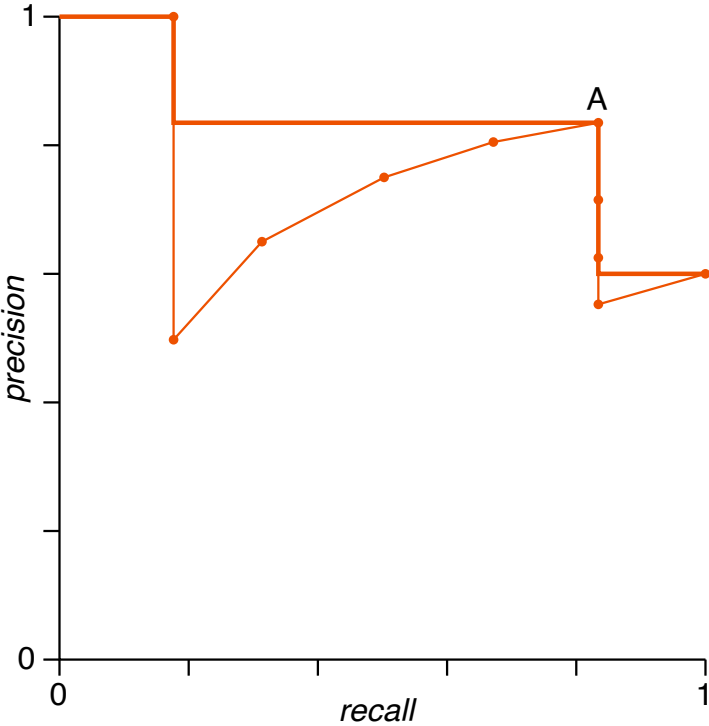
1. Integral of the step function visiting the maximum precision at every recall point.
2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

Ranked Retrieval Effectiveness

Average Precision

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	0.83	1.00



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

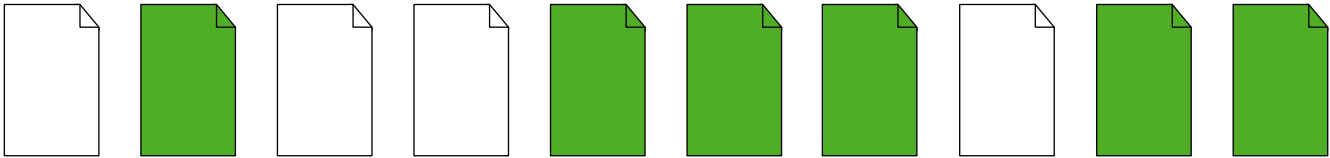
Interpolation alternatives:

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.

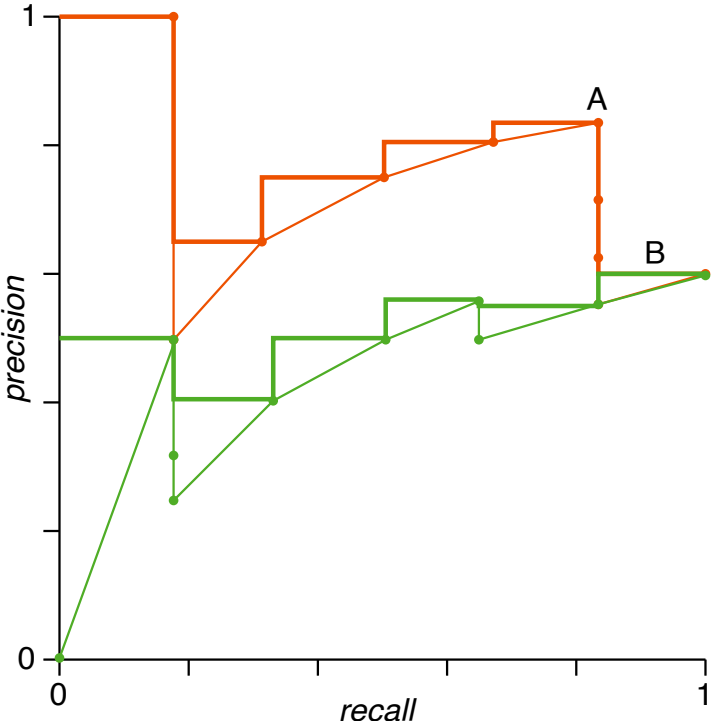
Ranked Retrieval Effectiveness

Average Precision

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 precision approximates the area under the precision-recall curve.

Interpolation alternatives:

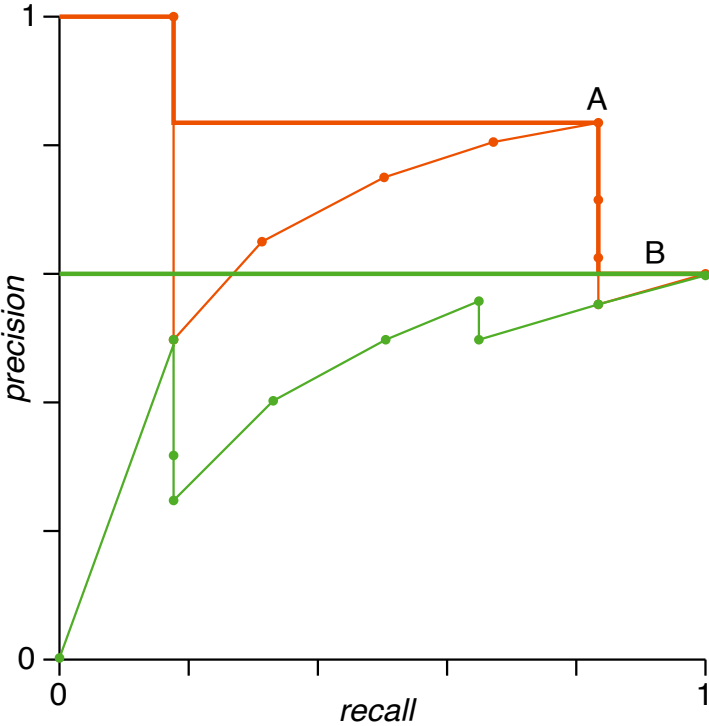
1. Integral of the step function visiting the maximum precision at every recall point.
2. Integral of the monotone step function visiting the maximum precision at any subsequent recall point.

Ranked Retrieval Effectiveness

Average Precision

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



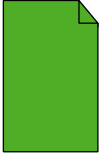
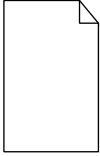
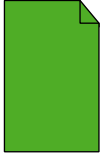




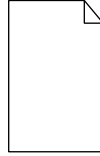
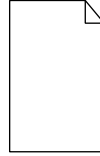

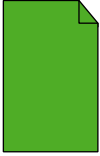
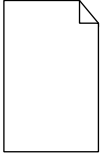
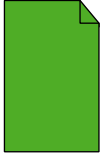




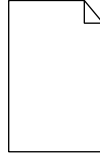
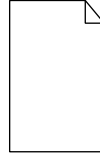

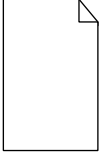
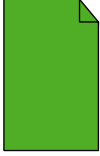





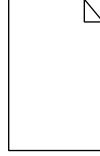


Average precision approximates the area under the precision-recall curve.

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.

Ranked Retrieval Effectiveness

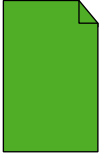
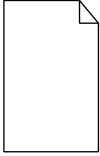
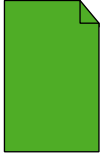




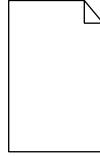
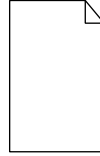

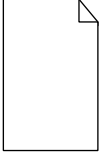
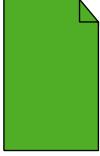





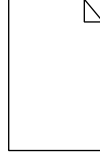


Average Precision (Alternative 1)

										
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

- 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.

Ranked Retrieval 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, ..., 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

Ranked Retrieval Effectiveness

Average Precision

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

Let $r : Q \times D \rightarrow \{0, 1\}$ 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:

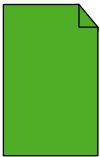
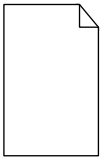
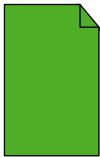




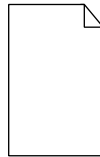
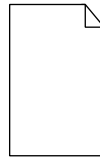

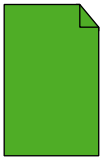
$$AP_{1@k}(q, R) = \frac{1}{\min(k, \sum_{d \in D} r(q, d))} \cdot \sum_{i=1}^k \left(r(q, d_i) \cdot \mathit{precision}@i(R) \right)$$

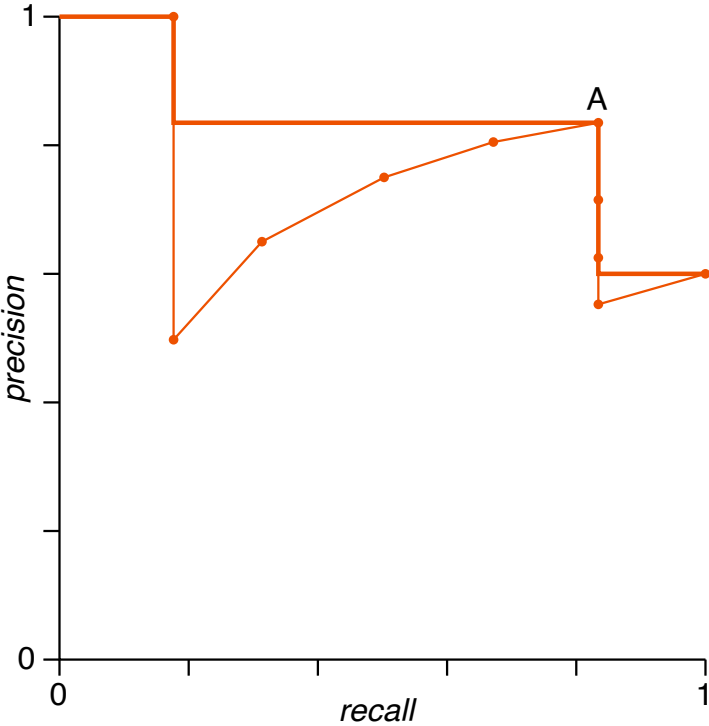
$$AP_2(q, R) = \frac{1}{11} \cdot \sum_{i \in \{0, 0.1, \dots, 1\}} \left(\max_{j: \mathit{recall}@j(R) \geq i} \mathit{precision}@j(R) \right)$$

Ranked Retrieval Effectiveness

Averaging Precision-Recall Curves

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	0.83	1.00



Problem:

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

Solution:

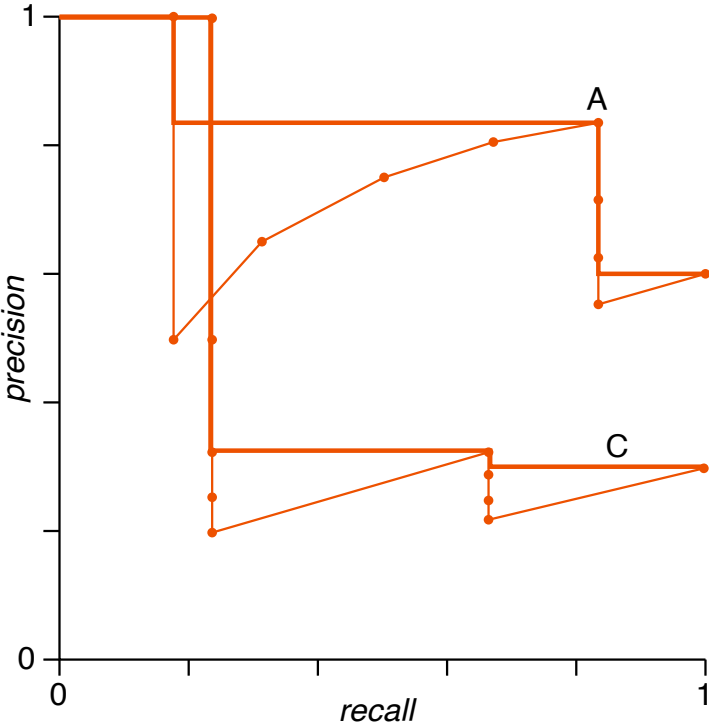
- ❑ Compute averages across 11 recall points at 0.1 steps.

Ranked Retrieval Effectiveness

Averaging Precision-Recall Curves

C
system 1
topic 2

<i>precision</i>	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.22	0.30
<i>recall</i>	0.33	0.33	0.33	0.33	0.33	0.66	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.

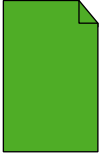
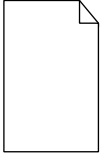
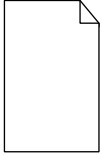







Solution:

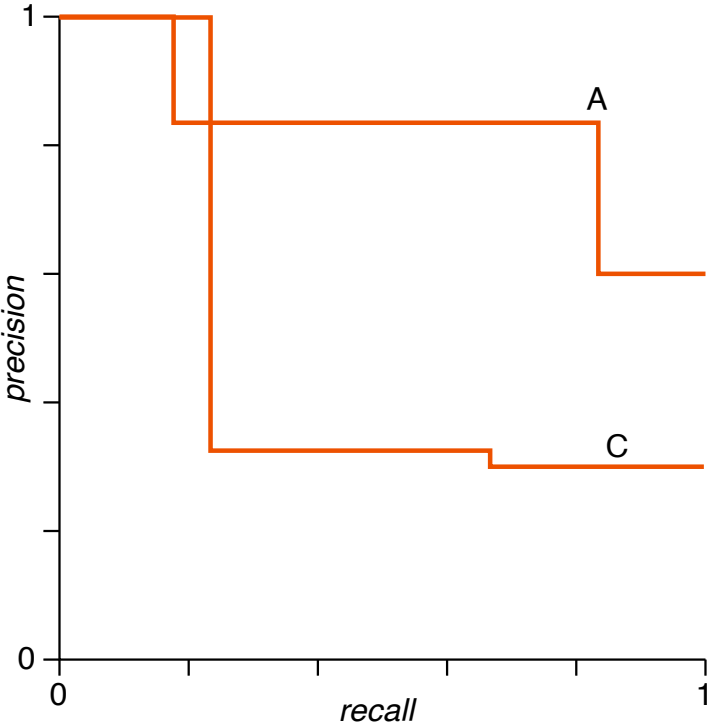
- ❑ Compute averages across 11 recall points at 0.1 steps.

Ranked Retrieval Effectiveness

Averaging Precision-Recall Curves

C
system 1
topic 2

										
<i>precision</i>	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
<i>recall</i>	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00



Problem:

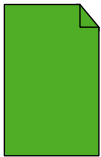
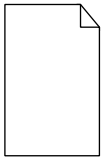
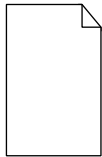

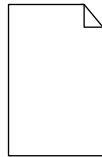

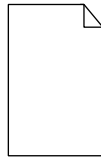
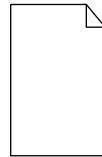
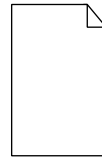

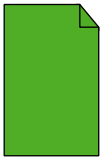
- Precision-recall curves do not necessarily share recall points.
- This renders averaging the curves across topics difficult.

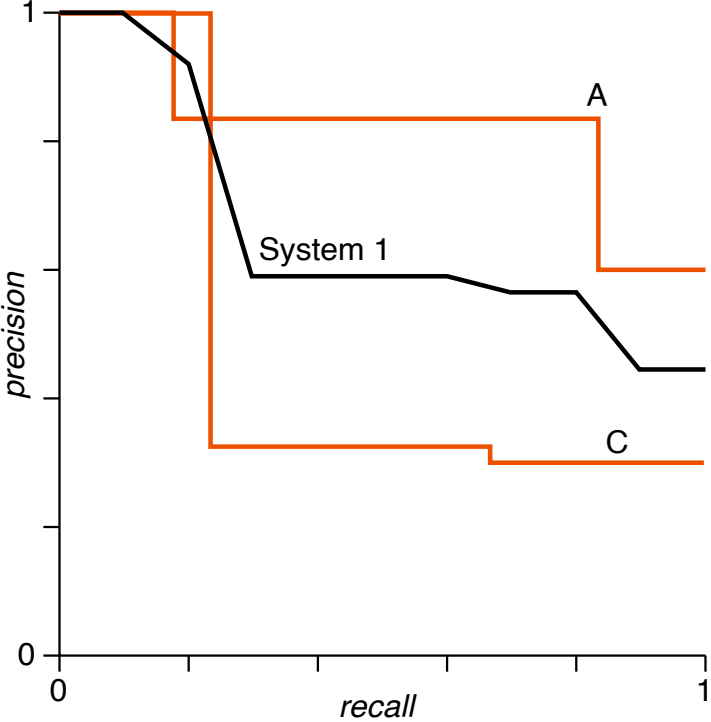
Solution:

- Compute averages across 11 recall points at 0.1 steps.

Ranked Retrieval Effectiveness

Averaging Precision-Recall Curves

C system 1 topic 2											
	<i>precision</i>	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
<i>recall</i>	0.33	0.33	0.33	0.33	0.33	0.66	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.

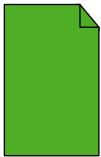
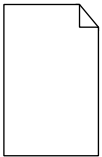
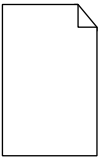
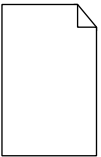
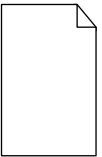
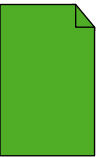
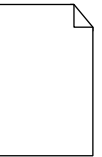
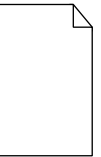
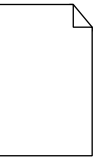

Solution:

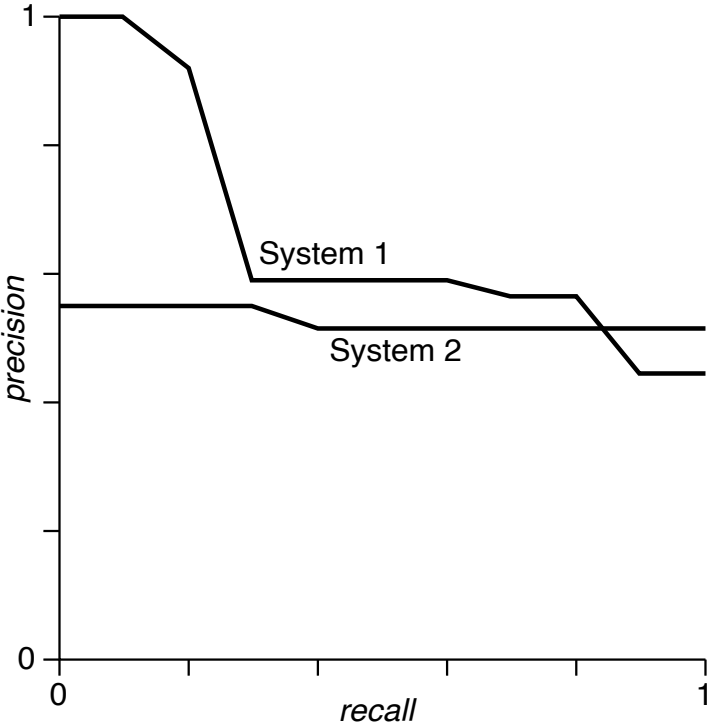
- ❑ Compute averages across 11 recall points at 0.1 steps.

Ranked Retrieval Effectiveness

Averaging Precision-Recall Curves

C
system 1
topic 2

										
<i>precision</i>	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
<i>recall</i>	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00



Interpretation:

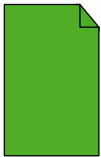
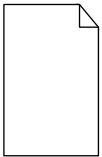
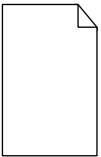
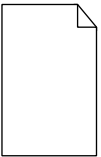
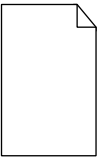
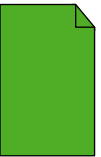
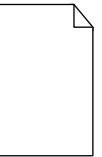
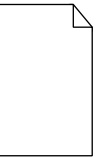
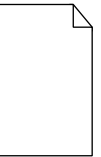

- Judging a system at various operating points.
- System 1 delivers very good average precision at high ranks.
- System 2 delivers slightly better average precision at low ranks.
- Neither system dominates the other.

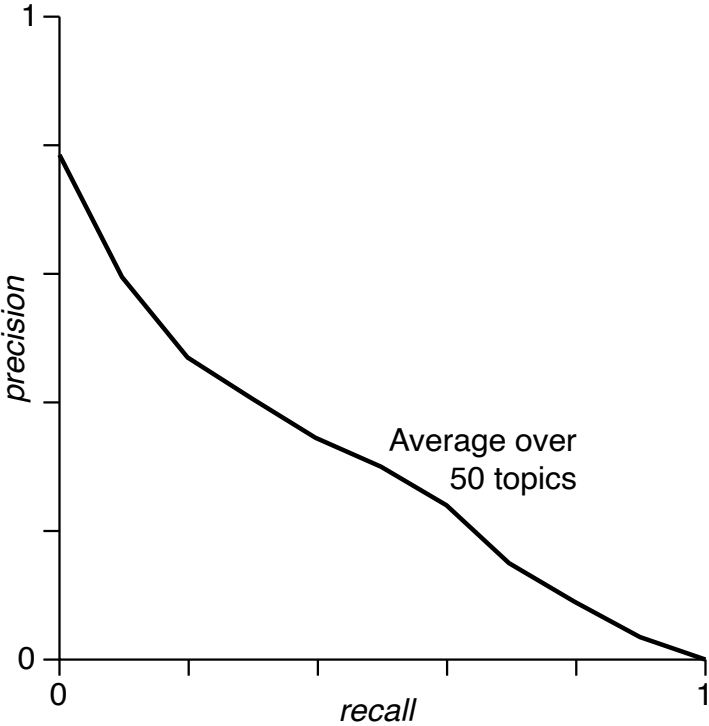
Curves are a lot smoother for 50 topics.

Ranked Retrieval Effectiveness

Averaging Precision-Recall Curves

C
system 1
topic 2

										
<i>precision</i>	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
<i>recall</i>	0.33	0.33	0.33	0.33	0.33	0.66	0.66	0.66	0.66	1.00



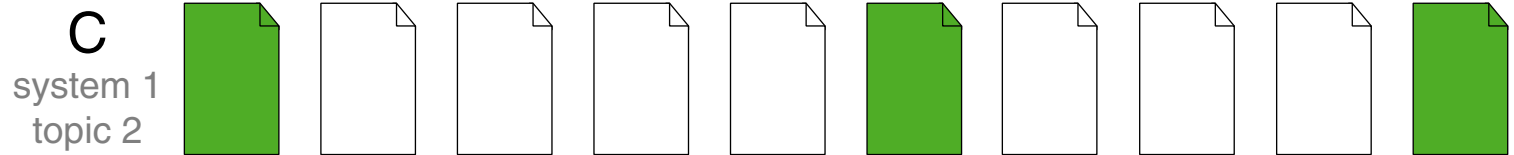
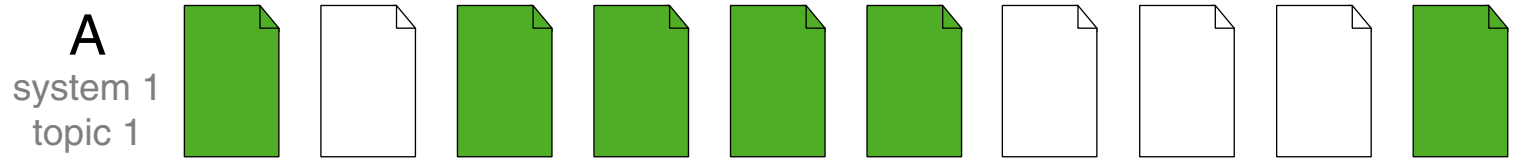
Interpretation:

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

Curves are a lot smoother for 50 topics.

Ranked Retrieval Effectiveness

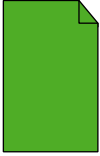
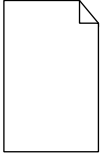
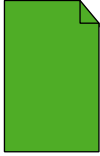




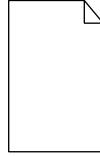
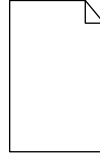

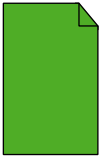
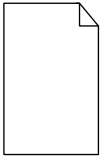





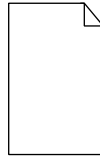
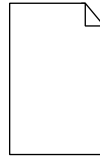

Mean Average Precision (MAP)



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

Ranked Retrieval Effectiveness

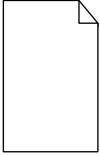
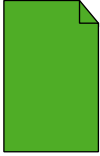





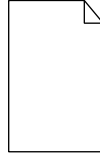


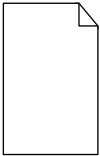






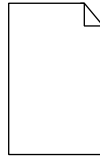
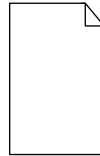

Mean Average Precision (MAP)

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
C system 1 topic 2											
	<i>precision</i>	1.00	0.50	0.33	0.25	0.20	0.33	0.29	0.25	0.22	0.30
	<i>recall</i>	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 $(0.78 + 0.54)/2 = 0.66$.
(A: $(1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$ and C: $(1.0 + 0.33 + 0.3)/3 = 0.54$)

Ranked Retrieval 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 $(0.78 + 0.54)/2 = 0.66$.
- 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: $(0.5 + 0.4 + 0.43)/3 = 0.44$)

Ranked Retrieval 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)

Ranked Retrieval 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:

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

where r is the rank of the first relevant document (i.e., RR is kind of Precision@ k but with a “variable” k across rankings). The mean reciprocal rank (MRR) is the average of the reciprocal ranks across many topics:

$$MRR@k = \sum_{i=1}^k RR@k$$

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

Remarks:

- ❑ MRR is disputed among IR researchers.
- ❑ 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 effectiveness. [\[Fuhr 2017\]](#)
- ❑ MRR can produce unintuitive scores: Assume that 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 $1/3 \cdot (1/1 + 1/2 + 1/4) = 0.58$, and System 2 has an MRR of $1/3 \cdot (3 \cdot 1/2) = 0.5$. Compared to the average ranks of the relevant documents, where System 1 has 2.3 and System 2 has 2, this is contradictory. [\[Fuhr 2017\]](#)
- ❑ Fuhr's criticism have sparked a academic dispute which was followed up by [\[Sakai 2021\]](#) (pro), [\[Ferrante et al. 2021\]](#) (con), [\[Moffat 2022\]](#) (pro), and [\[Ferrante et al. 2022\]](#) (con).

Ranked Retrieval 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 $\sum_{i=1}^k r(d_i, q)$, 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 depth to which DCG should be computed, the logarithm ensures smooth reduction, and $2^{r(d_i, q)}$ emphasizes highly relevant documents.

Ranked Retrieval Effectiveness

Normalized Discounted Cumulative Gain (nDCG)

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

This yields the normalized discounted cumulative gain measure:

$$nDCG@k = \frac{DCG@k}{DCG^*@k}$$

Example (if no other documents outside the top-10 were relevant):

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:

- Note that when comparing more than one system, the ideal ranking is usually formed by the joint relevance assessments for all systems (i.e., some documents in the ideal ranking may not have been retrieved by some of the systems but only by others).

Chapter IR:V

V. Evaluation

- Laboratory Experiments
- Measuring Performance
- Set Retrieval Effectiveness
- Ranked Retrieval Effectiveness
- User Models**
- Training and Testing
- Logging

User Models

Defining User Models

“All models are wrong, but some are useful” [[George Box](#)]

User Model: [[Moffat et al. 2017](#)]

- ❑ Formal description of the actions a universe of users scanning a ranking.
- ❑ Derive a probabilistic effectiveness metric.

User Models

Defining User Models

“All models are wrong, but some are useful” [\[George Box\]](#)

User Model: [\[Moffat et al. 2017\]](#)

- Formal description of the actions a universe of users scanning a ranking.
- Derive a probabilistic effectiveness metric.

Users look at results starting from the top until they are satisfied.

User 1:

1. Document 1
2. Document 2
3. Document 3
4. **Stop looking**

User 2:

1. Document 1
2. **Stop looking**

User 3:

1. Document 1
2. Document 2
3. Document 3
4. Document 4
5. **Stop looking**

User Models

Defining User Models

“All models are wrong, but some are useful” [\[George Box\]](#)

User Model: [\[Moffat et al. 2017\]](#)

- Formal description of the actions a universe of users scanning a ranking.
- Derive a probabilistic effectiveness metric.

Users look at results starting from the top until they are satisfied.

User 1:

1. Document 1
2. Document 2
3. Document 3
4. **Stop looking**

User 2:

1. Document 1
2. **Stop looking**

User 3:

1. Document 1
2. Document 2
3. Document 3
4. Document 4
5. **Stop looking**

Aggregated over millions of users, everyone looks at the first document, fewer look at the second, even fewer look at the third, and so on.

User Models

Defining User Models

Continuation, **Weight**, and **Last Rank**: defines **what** is viewed.

- Each defines a distribution modelling different aspects of user behaviour.

User Models

Defining User Models

Continuation, **Weight**, and **Last Rank**: defines **what** is viewed.

- Each defines a distribution modelling different aspects of user behaviour.

Continuation: Conditional continuation probability at depth i .

$$C(i) = \frac{W(i+1)}{W(i)}$$

User Models

Defining User Models

Continuation, **Weight**, and **Last Rank**: defines **what** is viewed.

- Each defines a distribution modelling different aspects of user behaviour.

Continuation: Conditional continuation probability at depth i .

$$C(i) = \frac{W(i+1)}{W(i)}$$

Weight: Fraction of user attention at depth i , such that $\sum_i^\infty W(i) = 1$.

$$W(i) = W(1) \cdot \prod_{j=1}^{i-1} C(j)$$

User Models

Defining User Models

Continuation, **Weight**, and **Last Rank**: defines **what** is viewed.

- Each defines a distribution modelling different aspects of user behaviour.

Continuation: Conditional continuation probability at depth i .

$$C(i) = \frac{W(i+1)}{W(i)}$$

Weight: Fraction of user attention at depth i , such that $\sum_i^\infty W(i) = 1$.

$$W(i) = W(1) \cdot \prod_{j=1}^{i-1} C(j)$$

Last Rank: Fraction of users that exist upon viewing depth i .

$$L(i) = \frac{W(i) - W(i+1)}{W(1)}$$

User Models

Defining User Models

Continuation, **Weight**, and **Last Rank**: defines **what** is viewed.

- Each defines a distribution modelling different aspects of user behaviour.

Continuation: Conditional continuation probability at depth i .

$$C(i) = \frac{W(i+1)}{W(i)}$$

Weight: Fraction of user attention at depth i , such that $\sum_i^\infty W(i) = 1$.

$$W(i) = W(1) \cdot \prod_{j=1}^{i-1} C(j)$$

Last Rank: Fraction of users that exist upon viewing depth i .

$$L(i) = \frac{W(i) - W(i+1)}{W(1)}$$

C/W/L is tightly coupled: specifying any function fixes the other two. [\[Moffat et al. 2013\]](#)

User Models

Analysing Effectiveness Measures

Modelling user behaviour with $C(i)$.

1. $i \leftarrow 1$
2. Look at document i
3. Gain r_i benefit from document
4. With probability $C(i)$, $i \leftarrow i + 1$ and go to step 2, otherwise stop looking with probability $1 - C(i)$

User Models

Analysing Effectiveness Measures

Modelling user behaviour with $C(i)$.

1. $i \leftarrow 1$
2. Look at document i
3. Gain r_i benefit from document
4. With probability $C(i)$, $i \leftarrow i + 1$ and go to step 2, otherwise stop looking with probability $1 - C(i)$

C can be used to analyse whether evaluation measures represent a realistic user model of browsing behaviour.

- ❑ Any “weighted precision” measures (e.g., Average Precision, RR, nDCG, etc.) can be derived entirely from C . [[Azzopardi et al. 2018](#)]
- ❑ The W function only tells use how much attention the user gave the document.
- ❑ The user browsing model for all set-based evaluation measures (e.g., precision, recall, etc.) can be described entirely in terms of L .

User Models

Analysing User Models (AP)

Average Precision:

$$AP = \frac{\sum_{i=1}^k \mathit{rel}(i)/k}{|\mathit{relevant}|}$$

User Models

Analysing User Models (AP)

Average Precision:

$$AP = \frac{\sum_{i=1}^k \mathit{rel}(i)/k}{|\mathit{relevant}|}$$

Conditional continuation probability at depth i for AP . [[Moffat et al. 2013](#)]

$$C_{AP}(i) = \begin{cases} \frac{\sum_{j=i+1}^k \mathit{rel}(j)/j}{\sum_{j=1}^k \mathit{rel}(j)/j} & \text{if } \sum_{j=i+1}^k \mathit{rel}(j)/j > 0 \\ 0 & \text{otherwise.} \end{cases}$$

User Models

Analysing User Models (AP)

Average Precision:

$$AP = \frac{\sum_{i=1}^k \mathit{rel}(i)/k}{|\mathit{relevant}|}$$

Conditional continuation probability at depth i for AP . [[Moffat et al. 2013](#)]

$$C_{AP}(i) = \begin{cases} \frac{\sum_{j=i+1}^k \mathit{rel}(j)/j}{\sum_{j=1}^k \mathit{rel}(j)/j} & \text{if } \sum_{j=i+1}^k \mathit{rel}(j)/j > 0 \\ 0 & \text{otherwise.} \end{cases}$$

What does C_{AP} tell us about the user model for AP ?

User Models

Analysing User Models (AP)

Average Precision:

$$AP = \frac{\sum_{i=1}^k \mathit{rel}(i)/k}{|\mathit{relevant}|}$$

Conditional continuation probability at depth i for AP . [[Moffat et al. 2013](#)]

$$C_{AP}(i) = \begin{cases} \frac{\sum_{j=i+1}^k \mathit{rel}(j)/j}{\sum_{j=1}^k \mathit{rel}(j)/j} & \text{if } \sum_{j=i+1}^k \mathit{rel}(j)/j > 0 \\ 0 & \text{otherwise.} \end{cases}$$

What does C_{AP} tell us about the user model for AP ?

- Users continue looking until all relevant documents have been looked at.
- C_{AP} reveals that Average Precision represents an impossible user model.

User Models

Analysing User Models (RR)

Reciprocal Rank:

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

User Models

Analysing User Models (RR)

Reciprocal Rank:

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

Conditional continuation probability at depth i for RR .

$$C_{RR}(i) = \begin{cases} 1 & \text{if } rel(i) = 0 \\ 0 & \text{if } rel(i) = 1. \end{cases}$$

User Models

Analysing User Models (RR)

Reciprocal Rank:

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

Conditional continuation probability at depth i for RR .

$$C_{RR}(i) = \begin{cases} 1 & \text{if } rel(i) = 0 \\ 0 & \text{if } rel(i) = 1. \end{cases}$$

What does C_{RR} tell us about the user model for RR ?

User Models

Analysing User Models (RR)

Reciprocal Rank:

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

Conditional continuation probability at depth i for RR .

$$C_{RR}(i) = \begin{cases} 1 & \text{if } rel(i) = 0 \\ 0 & \text{if } rel(i) = 1. \end{cases}$$

What does C_{RR} tell us about the user model for RR ?

- Users continue looking until a relevant document has been found.
- C_{RR} reveals that Reciprocal Rank has a realistic user model.

User Models

Analysing User Models (DCG)

Reciprocal Rank:

$$DCG = \sum_i^k \frac{rel(i)}{\log_2(i + 1)}$$

User Models

Analysing User Models (DCG)

Reciprocal Rank:

$$DCG = \sum_i^k \frac{rel(i)}{\log_2(i+1)}$$

Conditional continuation probability at depth i for DCG . [\[Moffat et al. 2013\]](#)

$$C_{DCG}(i) = \begin{cases} \frac{\log_2(i+1)}{\log_2(i+2)} & \text{when } 1 \leq i < k \\ 0 & \text{otherwise.} \end{cases}$$

User Models

Analysing User Models (DCG)

Reciprocal Rank:

$$DCG = \sum_i^k \frac{rel(i)}{\log_2(i+1)}$$

Conditional continuation probability at depth i for DCG . [\[Moffat et al. 2013\]](#)

$$C_{DCG}(i) = \begin{cases} \frac{\log_2(i+1)}{\log_2(i+2)} & \text{when } 1 \leq i < k \\ 0 & \text{otherwise.} \end{cases}$$

What does C_{DCG} tell us about the user model for DCG ?

User Models

Analysing User Models (DCG)

Reciprocal Rank:

$$DCG = \sum_i^k \frac{rel(i)}{\log_2(i+1)}$$

Conditional continuation probability at depth i for DCG . [\[Moffat et al. 2013\]](#)

$$C_{DCG}(i) = \begin{cases} \frac{\log_2(i+1)}{\log_2(i+2)} & \text{when } 1 \leq i < k \\ 0 & \text{otherwise.} \end{cases}$$

What does C_{DCG} tell us about the user model for DCG ?

- Users continue looking with smoothly decreasing probability until all (or k) documents have been looked at.
- C_{DCG} reveals that DCG has a realistic user model.

User Models

Unjudged Documents

Until now, we have assumed we have complete relevance assessments.

Is this a reasonable assumption to make?

What would be do if we retrieved “unjudged” documents?

User Models

Unjudged Documents

Until now, we have assumed we have complete relevance assessments.

Is this a reasonable assumption to make?

What would be do if we retrieved “unjudged” documents?

- ❑ Was once possible to assess all documents in a collection.
- ❑ Quickly became impossible (millions of documents to assess per topic).
- ❑ Solution: assume unjudged documents are non-relevant.

User Models

Unjudged Documents

Until now, we have assumed we have complete relevance assessments.

Is this a reasonable assumption to make?

What would be do if we retrieved “unjudged” documents?

- ❑ Was once possible to assess all documents in a collection.
- ❑ Quickly became impossible (millions of documents to assess per topic).
- ❑ Solution: assume unjudged documents are non-relevant.

What are some problems when we assume unjudged documents are non-relevant?

User Models

Unjudged Documents

Until now, we have assumed we have complete relevance assessments.

Is this a reasonable assumption to make?

What would be do if we retrieved “unjudged” documents?

- ❑ Was once possible to assess all documents in a collection.
- ❑ Quickly became impossible (millions of documents to assess per topic).
- ❑ Solution: assume unjudged documents are non-relevant.

What are some problems when we assume unjudged documents are non-relevant?

- ❑ Newer systems that are more effective than those used to pool assessments are at a disadvantage.
- ❑ Some measures like Average Precision are unstable when unjudged documents are discovered to be relevant. [[Moffat and Zobel 2008](#)]
- ❑ Measures like Rank-biased Precision (RBP) allow us to compute upper and lower bounds on effectiveness.