Chapter IR:V

V. Evaluation

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

Laboratory Experiments

Experiment Scope

Ad hoc retrieval

- Assumption: a searcher's task requires just one query

➜ Retrieval experiments can be run in laboratory environments

➜ Canonical measurement of retrieval performance

➜ Reproducibility and scalability

Laboratory Experiments

Experiment Scope

Ad hoc retrieval

- Assumption: a searcher's task requires just one query

➜ Retrieval experiments can be run in laboratory environments

➜ Canonical measurement of retrieval performance

➜ Reproducibility and scalability

Interactive retrieval

- Assumption: a searcher's task requires multiple queries and exploration
- Dependent variables: result quality, human factors, context, user interface and experience, the retrieval system's supporting facilities, etc.

➜ Retrieval experiments typically require user studies

➜ Measurement of retrieval performance depends on the setup

➜ Difficult to reproduce and scale

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 a “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 searchers’ tasks, goals, or gaps of knowledge.
   - Alternatively, declarative descriptions of desired search results.
   - Often accompanied by specific queries the searchers (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 searchers 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 analyses 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.
- For an evaluation corpus with a fixed split of training, validation, and test set, still k-fold cross-validation should be used as the variance of performance results is often high. [Fuhr 2017]
- Likert is pronounced "lick-ert" (not "lai-kert"), as it is named after the American social psychologist Rensis Likert. [Wikipedia]

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 de
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 l
governing it.
</narrative>

</topic>
```

Laboratory Experiments

Experimental Setup: Relevance Judgments

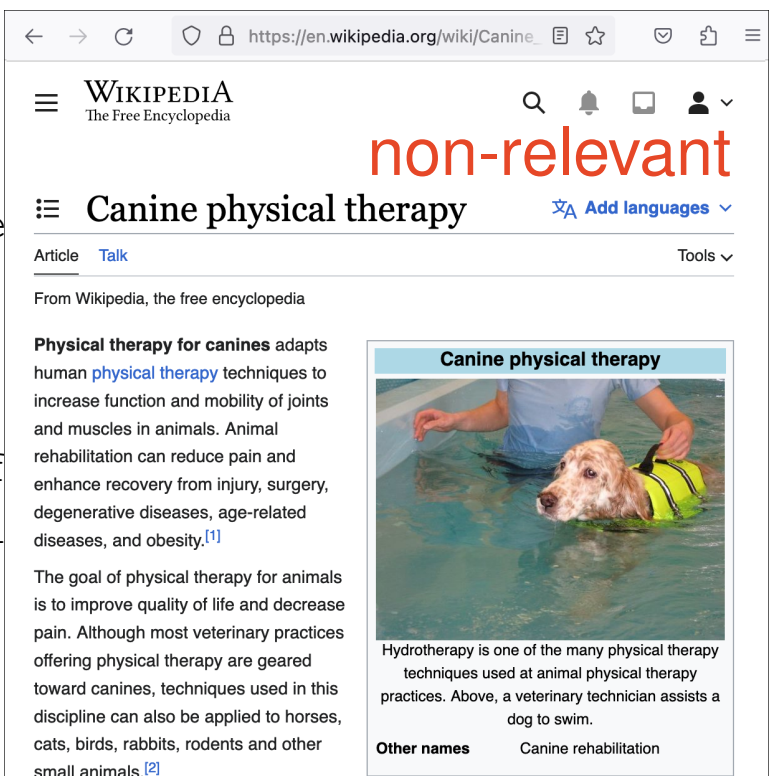
```
<topic number="794" type="single">

<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 de
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 l
governing it.
</narrative>

</topic>
```



Laboratory Experiments

Experimental Setup: Relevance Judgments

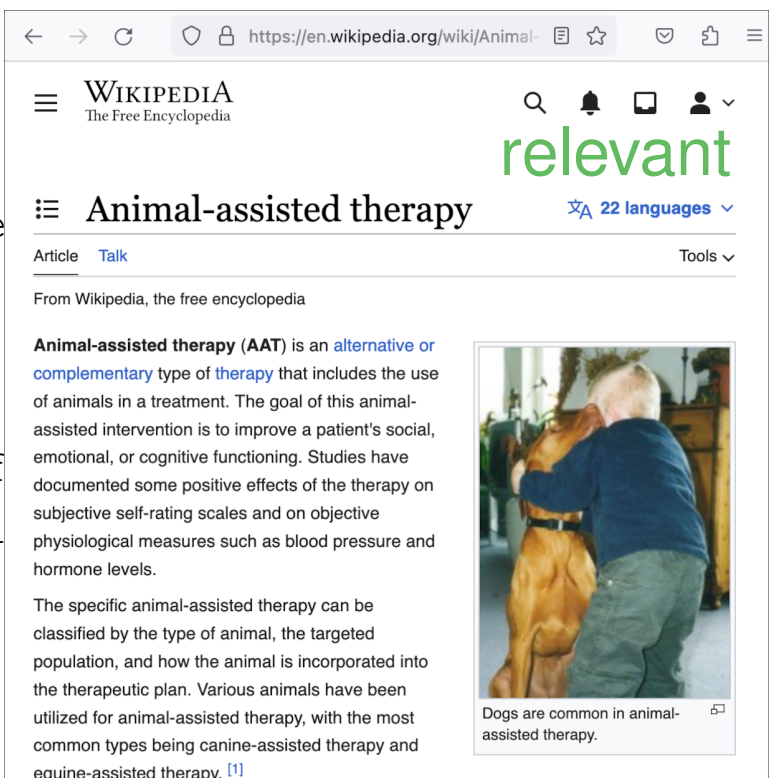
```
<topic number="794" type="single">

<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 de
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 l
governing it.
</narrative>

</topic>
```



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 (Up to what rank k should documents 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 should 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 exhaustively formulated 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 without relevance judgments, pooling selects the documents to be assessed.

For each topic . . .

1. Collect the top-k results of each retrieval system (variant).
2. Merge the results, omitting duplicates, to obtain the document "pool".
3. Let assessors judge the documents from the pool in random order.

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		$\sum$
		yes	no	
A	yes	a	b	c
	no	d	e	f
$\sum$		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		$\sum$
		yes	no	
A	yes	300	20	320
	no	10	70	80
$\sum$		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 measured κ values (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]