Chapter IR:VIII

VIII. Evaluation

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

©HAGEN/POTTHAST/STEIN 2018

Laboratory Experiments

Retrieval Tasks

Ad hoc retrieval:

- One question, one result set.
- ➜ Amenable to laboratory environments
- ➜ Canonical measurement of retrieval performance
- ➜ Reproducibility and scalability

Interactive retrieval / task-based retrieval:

- The user has a goal or a task that requires many queries and exploration, refining the information need along the way.
- Depends not only on result quality, but also on human factors, context, user interface and experience, and the search engine's supporting facilities.
- ➜ Performance is difficult to be measured

©HAGEN/POTTHAST/STEIN 2018

Remarks:

- “ad hoc” (Latin: “for this”) means “for the particular end or case at hand without consideration of wider application” [Merriam Webster]

©HAGEN/POTTHAST/STEIN 2018

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 domain of interest. The sampling method of how documents are drawn from the population determines a corpus's validity. Statistical representativeness may be difficult to achieve, e.g., in case of the web. In that case, the larger a corpus can be built for a given domain, the better.

2. A set of information needs (topics)
   Formalized, written descriptions of users' tasks, goals, or gaps of knowledge. Alternatively, descriptions of desired search results. Often accompanied by specific queries the users (would) have used to search for relevant documents.

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 its associated topic. Ideally, the judgments are obtained from the same users who supplied the topics, but in practice judgments are collected from third parties. Judgments may be given in binary form, or on a Likert scale.

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

©HAGEN/POTTHAST/STEIN 2018

Remarks:

- This setup is sometimes referred to as an experiment under the Cranfield paradigm, in reference to Cleverdon's series of projects at 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 do 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]

©HAGEN/POTTHAST/STEIN 2018

Laboratory Experiments

Experimental Setup: Document Collections / Corpora

For ad hoc retrieval, the Text Retrieval Conference (TREC) has organized evaluation tracks as of 1992, inviting scientists to compete.

Key document collections used:

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	304	200	2.5	821
ClueWeb12	733 million	27.3 TB	n/a	200	3.6	793

- CACM: Communications of the ACM 1958-1979 (only titles and abstracts)
- AP: Associated Press newswire documents 1988–1990
- GOV2: Crawl of .gov domains early 2004
- ClueWeb: Web crawls from 2009 and 2012

Reusing experimental setups renders previous approaches comparable.

©HAGEN/POTTHAST/STEIN 2018

Remarks:

- TREC is organized by the United States National Institute of Standards and Technology (NIST). It has been key to popularize laboratory evaluation of search engines, organizing evaluation tracks on many different retrieval-related tasks every year: trec.nist.gov.
- Ad hoc retrieval has been studied in the ad hoc tracks, the terabyte tracks and the web tracks.
- Several similar initiatives have formed, namely CLEF, NTCIR, and FIRE.

©HAGEN/POTTHAST/STEIN 2018

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>
```


©HAGEN/POTTHAST/STEIN 2018

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 and 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>
```

- At TREC, every year usually 50 topics are provided. The ones from previous years can be used for training.

©HAGEN/POTTHAST/STEIN 2018

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>
```


©HAGEN/POTTHAST/STEIN 2018

Laboratory Experiments

Experimental Setup: Relevance Judgments

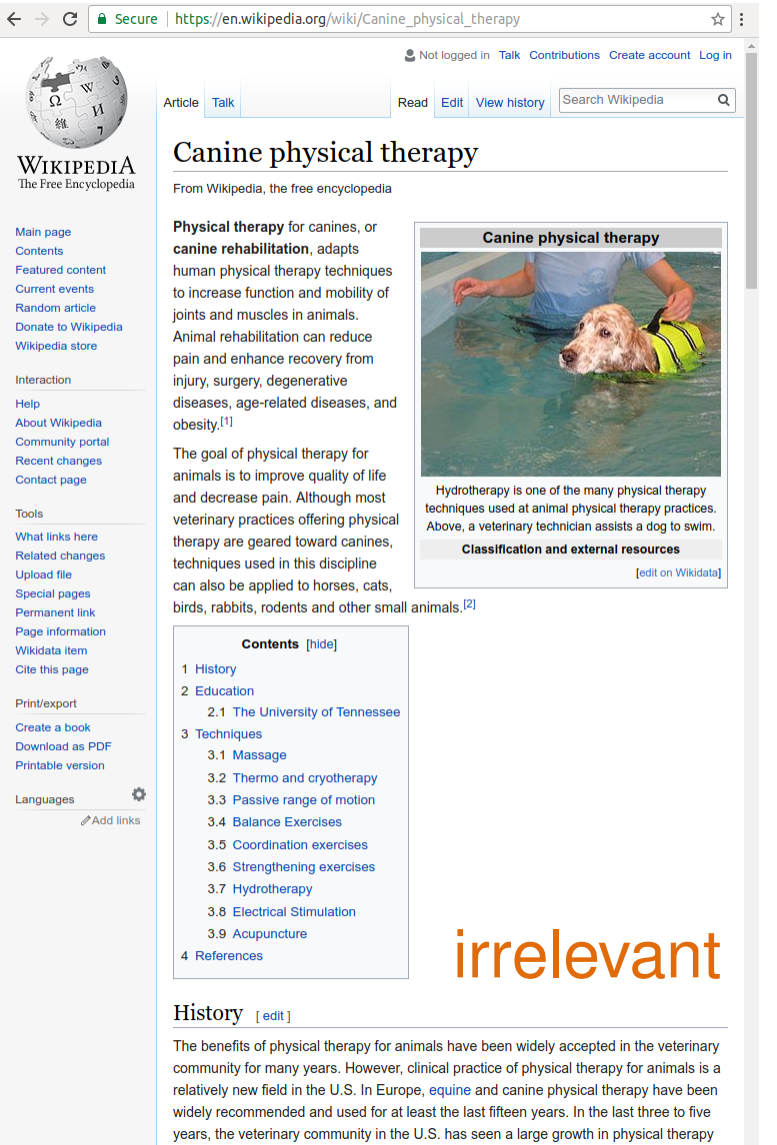
```
<topic number="794" type="single">

<query> pet therapy </query>

<description>
How are pets or animals used in th
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>
```



©HAGEN/POTTHAST/STEIN 2018

Laboratory Experiments

Experimental Setup: Relevance Judgments

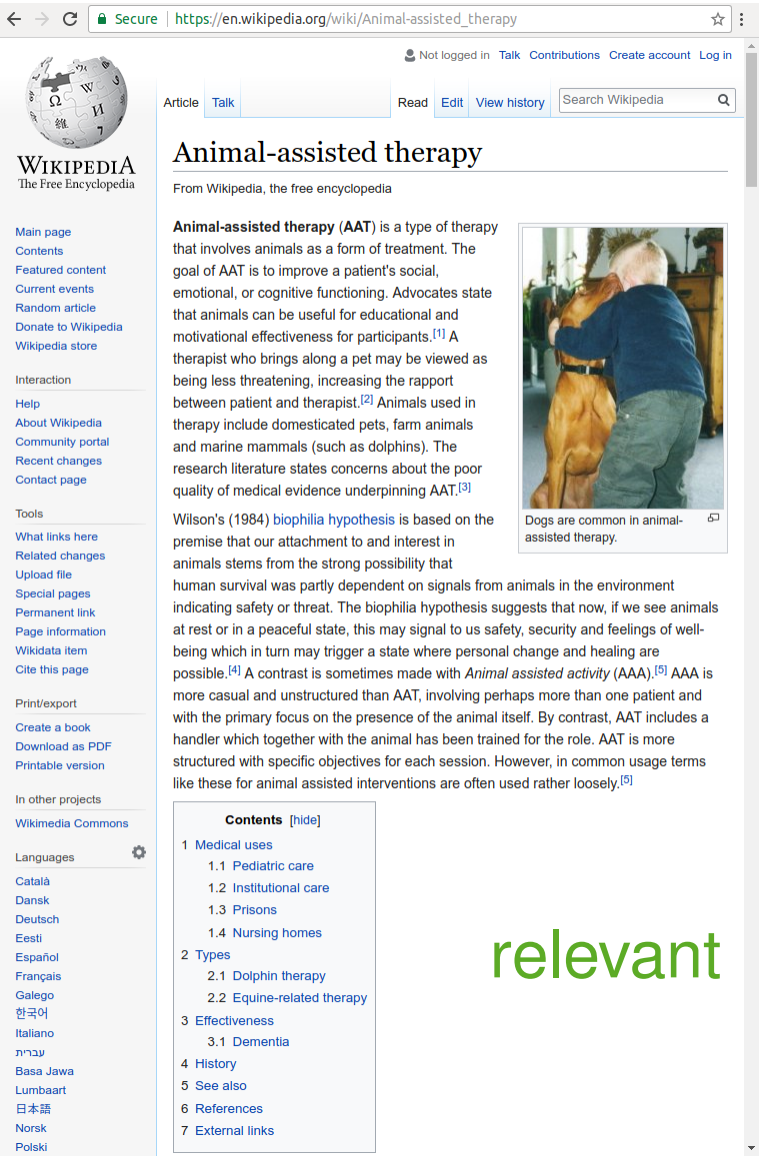
```
<topic number="794" type="single">

<query> pet therapy </query>

<description>
How are pets or animals used in th
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>
```



relevant

©HAGEN/POTTHAST/STEIN 2018

Laboratory Experiments

Experimental Setup: Relevance Judgments

A relevance judgment requires the manual assessment whether a document returned by a search engine for a given query is relevant under a given topic.

- **Assessment depth**
  Assessing every document's relevance for every topic quickly becomes infeasible with growing collection size, numbers of topics, and (variants of) search engines to be evaluated. Partial assessments are made based on a sampling strategy called pooling.
- **Assessment scale**
  Typically, binary assessments are made, judging documents as relevant or irrelevant. Less often, degrees of relevance are distinguished (e.g., relevant and highly relevant).
- **Assessor selection and instruction**
  Ideally, the persons who supplied topics also assess relevance, presuming they genuinely perceived the underlying information need. In practice, this usually cannot be achieved. Hence, a topic's description must be sufficiently exhaustive to serve as instruction.
- **Assessor reliability**
  Humans are unreliable judges, their judgments depending on many outside influences. One cannot expect objective results from just one assessment per document for a given topic. Multi-assessments yield more reliable judgments; assessor reliability can be measured.

©HAGEN/POTTHAST/STEIN 2018

Remarks:

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

©HAGEN/POTTHAST/STEIN 2018

Laboratory Experiments

Experimental Setup: Pooling

Pooling is a sampling strategy for documents retrieved by to-be-evaluated search engines for a given set of topics.

For each topic:

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

Caveats:

- Presuming a certain correlation between search engines’ results on a topic, only documents considered relevant by search engines are analyzed.
- New retrieval algorithms that are evaluated later may retrieve unjudged documents, requiring new judgments, probably from different assessors.
- All documents ranked below the threshold are deemed irrelevant by default, regardless the truth.

©HAGEN/POTTHAST/STEIN 2018

Laboratory Experiments

Assessor Reliability

Assessor reliability measures the degree of agreement between assessors, and the degree of consistency of the same assessor. 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 time the raters agree. Agreement by chance it 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.

©HAGEN/POTTHAST/STEIN 2018

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.
- $1 - p_e$ denotes the agreement attainable above chance
- $p_o - p_e$ denotes the agreement attained above chance

©HAGEN/POTTHAST/STEIN 2018

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 relevance judgments. Then a simple kappa statistic can be computed as follows:

		B		$\sum$
		yes	no	
A	yes	a	b	c
	no	d	e	f
$\sum$		g	h	n

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

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

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

©HAGEN/POTTHAST/STEIN 2018

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 $n = 400$ relevance judgments. Then a simple kappa statistic can be computed as follows:

		B		$\sum$
		yes	no	
A	yes	300	20	320
	no	10	70	80
$\sum$		310	90	400

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

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

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

$$\kappa = 0.776$$

©HAGEN/POTTHAST/STEIN 2018

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 raters A and B work independently, the probability $P(\text{yes})^2$ ($P(\text{no})^2$) denotes the probability of both voting yes (no) by chance. Another way of computing p_e is to sum the mutplication 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]

©HAGEN/POTTHAST/STEIN 2018