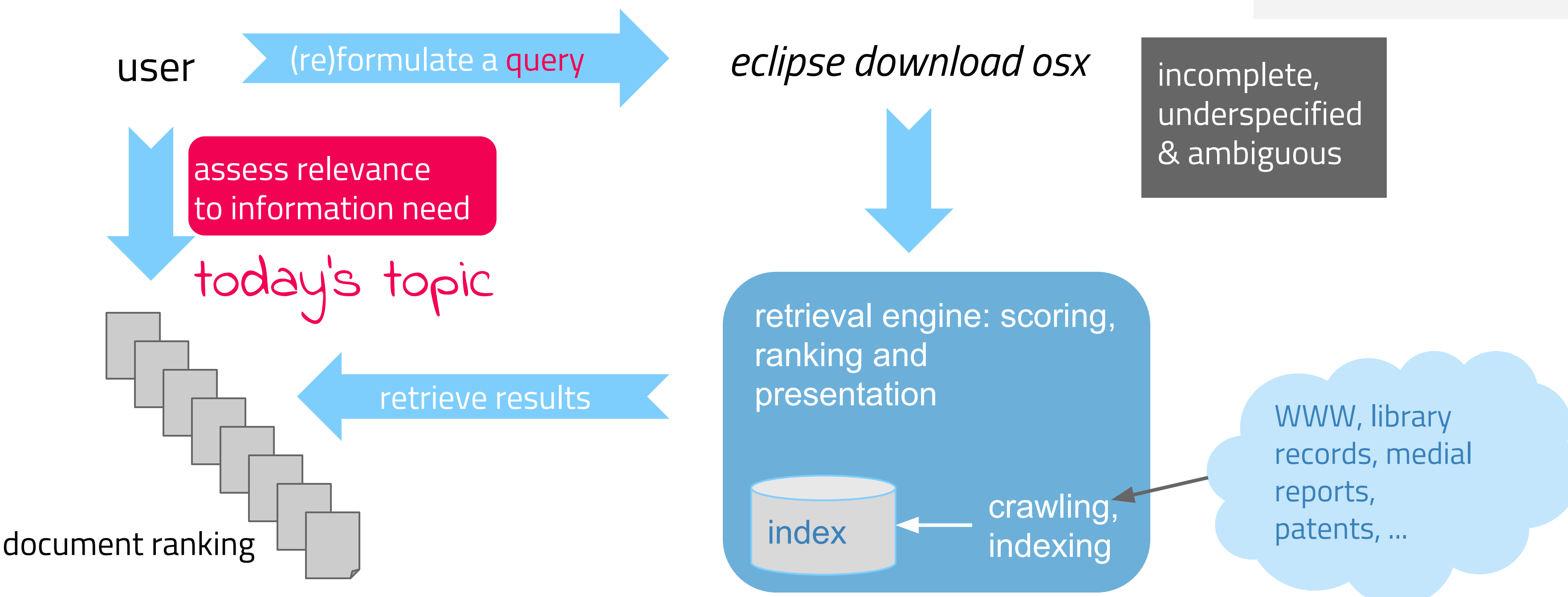
IN4325
Evaluation in IR

Claudia Hauff (WIS, TU Delft)

The big picture

The essence of classic IR

Information need: *Looks like I need Eclipse for this job. Where can I download the latest beta version for macOS Sierra?*



Information need
Topic the user wants to know more about

Query
Translation of need into an input for the search engine

Relevance
A document is relevant if it (partially) provides answers to the information need

Why “classic”?

Classic Web search	Proactive search (zero query search)	Voice search	...
Query = textual input	Query = none	Query = speech input	
Results = ranked list of search result snippets (i.e. “ten blue links”)	Results = a single information card	Results = speech output	
Actions = click, view	Actions = view	Actions = speech input	

Information retrieval is a broad field that deals with a wide range of information access issues.

Connected to information science, NLP, applied machine learning, semantic Web and (in recent years) dialogue systems.

What are we up to as IR community?

Let's quickly look at upcoming benchmark tasks (TREC* 2018)

Complex Answer Retrieval

" The focus … is on developing systems that are capable of answering complex information needs by collating relevant information from an entire corpus. "

Incident Streams

"… to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests … for emergency service operators. "

Precision Medicine

"… building systems that use data (e.g., a patient's past medical history and genomic information) to link oncology patients to clinical trials for new treatments "

News search

"… will foster research that establishes a new sense what relevance means for news search."

* Text REtrieval Conference (1992 - *), changing tracks every year. trec.nist.gov

Benchmarks drive our community

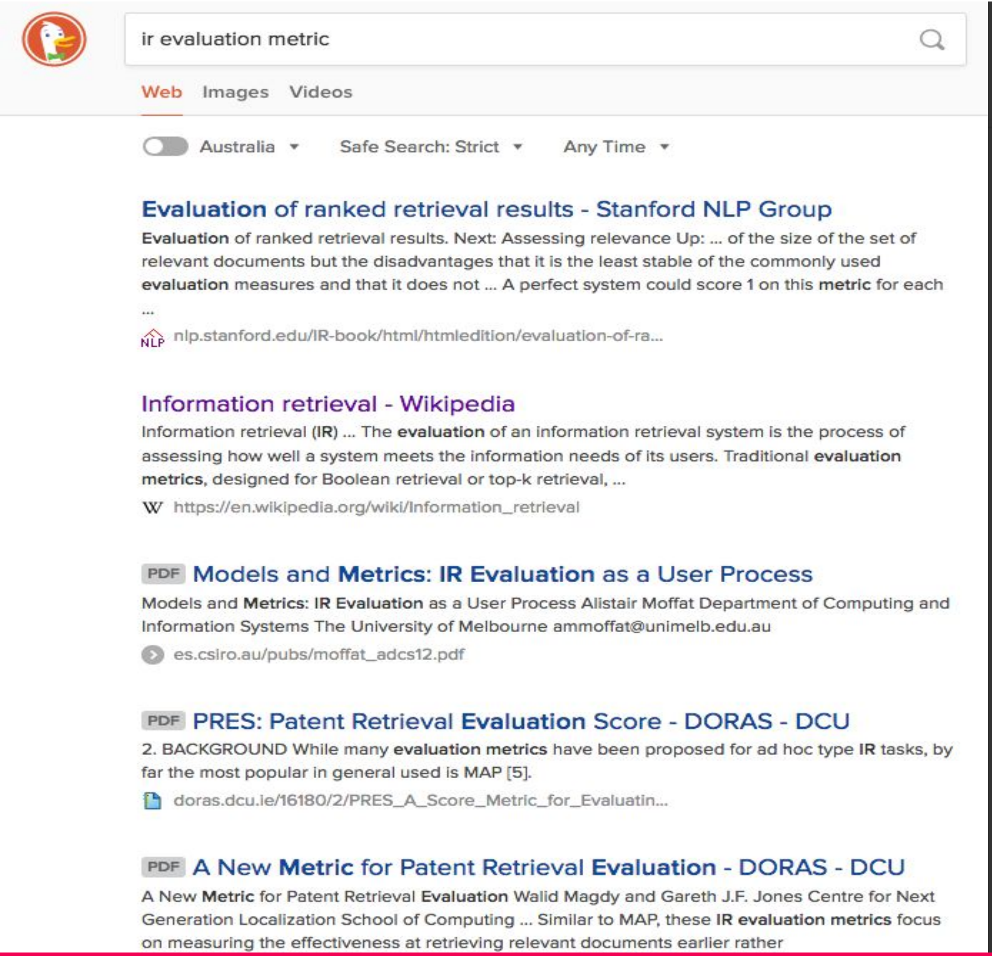
CLEF	MediaEval	NTCIR	FIRE
Conference and Labs of the Evaluation Forum	Benchmarking Initiative for Multimedia Evaluation	NII Test Collection for IR Systems	Forum for Information Retrieval Evaluation
http://www.clef-initiative.eu/	http://www.multimediaeval.org/	http://research.nii.ac.jp/ntcir/index-en.html	http://www.isical.ac.in/~clia/
EUROPE	EUROPE	JAPAN	INDIA

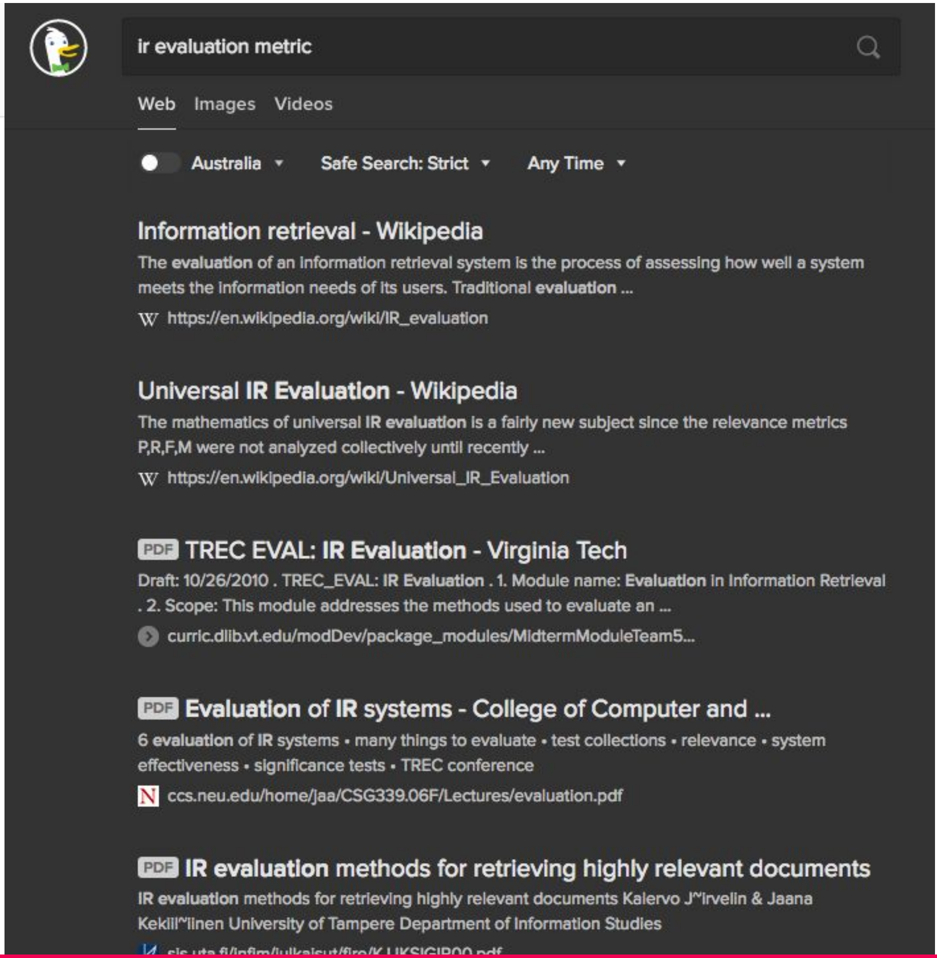
TREC	TRECVID
USA	USA

' … engineers then come up with a ***hypothesis*** *about what signal what data could we integrate into our algorithm we test all these reasonable ideas through* ***rigorous scientific testing*** *… '* Google Inside Search

https://www.youtube.com/watch?v=J5RZOU6vK4Q

Question time





Given two system rankings for a query, how can you decide whether one is better than another?

Relevance

Why are we starting with the evaluation lecture in this course anyway?

Because evaluation is a vital component of ~95% of all published IR papers. No matter your choice of project or survey, you need to understand IR evaluation.

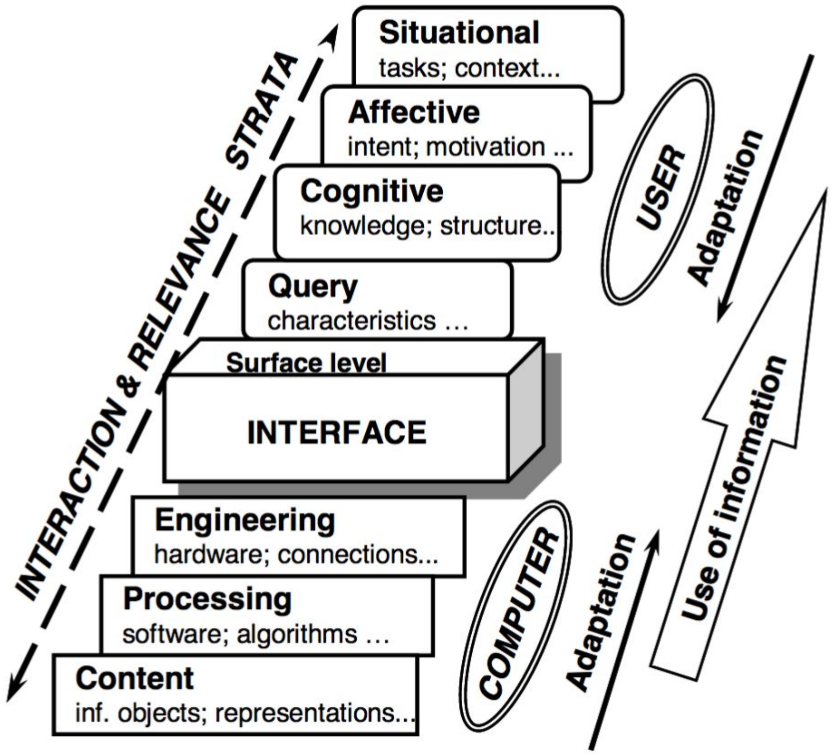
Relevance

- Key notion in information retrieval

- A good retrieval system retrieves all relevant documents but as few non-relevant documents as possible

- Relevance is an intuitive notion for humans
- Retrieval systems create relevance and users derive relevance

- Ongoing debate for the past 40 years (see reviews below)



Stratified model of relevance interactions. (Saracevic, 2007)

[Relevance: A review of and a framework for the thinking on the notion in information science]()
[Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II]()
[Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance]()

Manifestations of relevance (Saracevic, 2007)

- **System relevance:** relation between query and information objects (documents)

- **Topical relevance**: relation between the subject of the topic and the subject of the information objects

- **Cognitive relevance**: relation between the cognitive state of the user and the information objects
  ⟶ cognitive correspondence, novelty, information quality, ...

- **Situational relevance** (utility): relation between the situation and the information objects
  ⟶ appropriateness of information, reduction of uncertainty, ...

- **Affective relevance**: relation between the intent, goals, emotions of the user and information
  ⟶ success, accomplishment, ...

Question time

Goal

Evaluation measures that reflect users' satisfaction with the system

What do you think is part of a user being satisfied with an IR system?

Evaluation is at the heart of IR

Goals

Evaluation measures that reflect users' satisfaction with the system

The perfect metric also allows us to fine-tune the system via machine learning

User satisfaction in terms of

- **Coverage** of the corpus
- **Time lag** between query and retrieved results - even 200ms delays are noticeable to users (1)
- **Presentation** of output
- Required **user effort**
- ...
- Proportion of relevant results retrieved (recall)
- Proportion of retrieved results that is relevant (precision)
- ...

system effectiveness

Assumption: the more effective the system, the more satisfied the user.

(1) https://research.googleblog.com/2009/06/speed-matters.html

Evaluation is difficult

- Which users to evaluate for?
- Which intents to evaluate for?
- How are evaluations be made reusable?
- How can the difference between systems be quantified?

Test Collection Approach *

* Mainstream way of evaluation. Empirical. Another approach is the axiomatic one (found in theoretic research).

Cranfield evaluation paradigm (1960s)

IR evaluation methodology developed by Cyril Cleverdon in the 1960s; Cranfield corpus:

- Test collection of 1,400 documents (1) [scientific abstracts]
- Set of 225 topics (information needs)
- Ad hoc task
- Complete set of binary (0/1) relevance judgments
- Metrics to compare systems with each other
- **i.e. reusable data!**

Example Cranfield corpus topic:
papers applicable to this problem (calculation procedures for laminar incompressible flow with arbitrary pressure gradient)

Paradigm adapted to the modern time

Relevance judgments are **no longer binary**

- Multi-graded decision (somewhat relevant vs. very relevant)
- User-dependent decision (what is relevant for you may not be relevant for me)
- Context-dependent decision (whether something is relevant depends on the time of day, ...)

Topics and queries are not one and the same anymore

TREC 2001 Web ad hoc topic

<top>
<num> Number: 503

<title> Vikings in Scotland?

<desc> Description: What hard evidence proves that the Vikings visited or lived in Scotland?

<narr> Narrative: A document that merely states that the Vikings visited or lived in Scotland is not relevant. A relevant document must mention the source of the information, such as relics, sagas, runes or other records from those times.

</top>

flickr@mdpettitt

We are conducting **simulations** of users searching with a retrieval system.

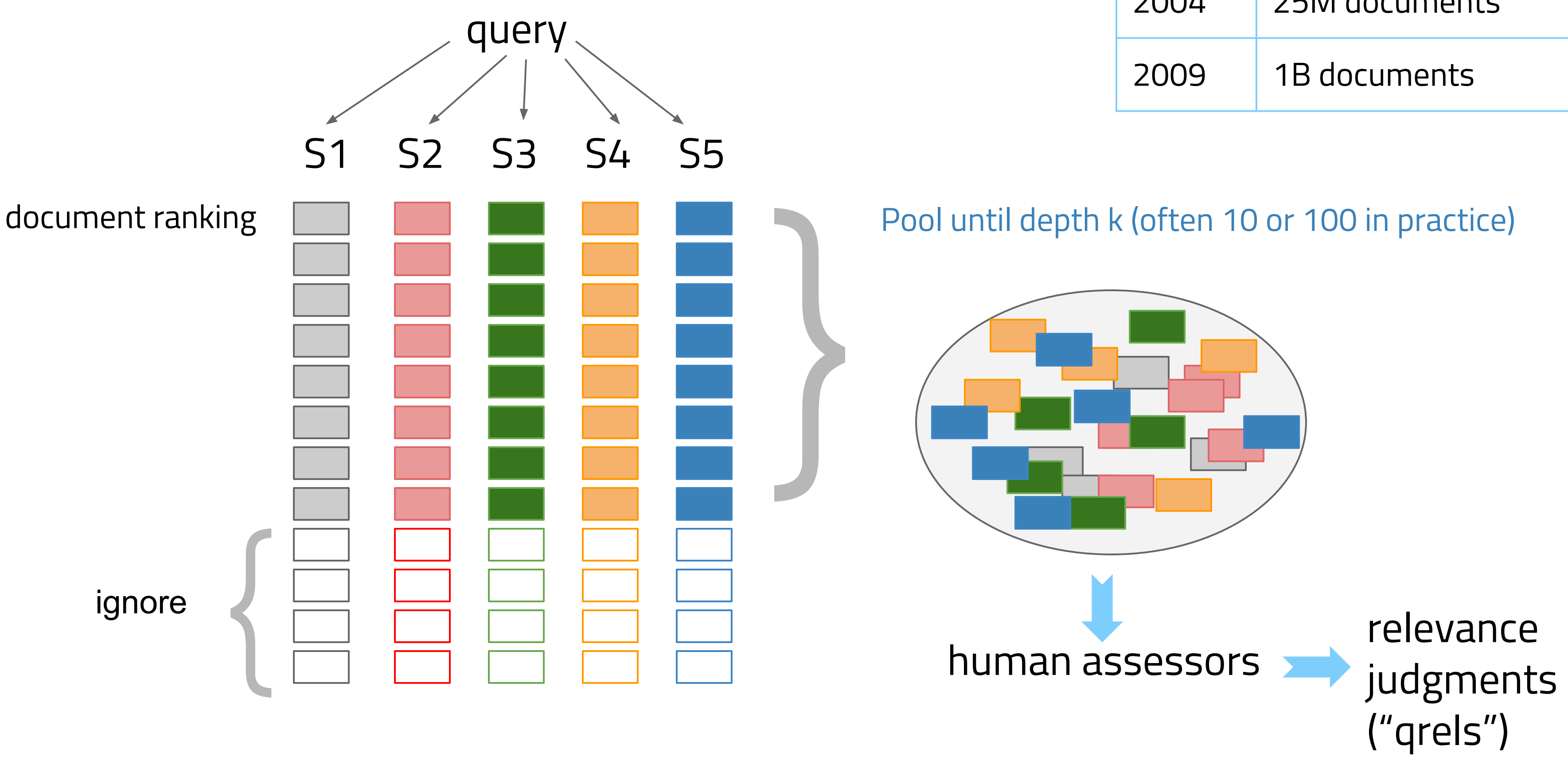
+ Cheaper, easier, reusable, reproducible
- Test collection retrieval effectiveness gains (=simple simulated users) may not translate to operational gains (=real users).

Also known as **batch evaluation** or **offline evaluation.**

What documents to judge: depth pooling

Commonly used today

Year	TREC Web corpus sizes
2001	1.69M documents
2004	25M documents
2009	1B documents



TREC-8 numbers
(ran in 1999)

50

Topics (ad hoc task)

86,830

Pooled documents (k=100)

129

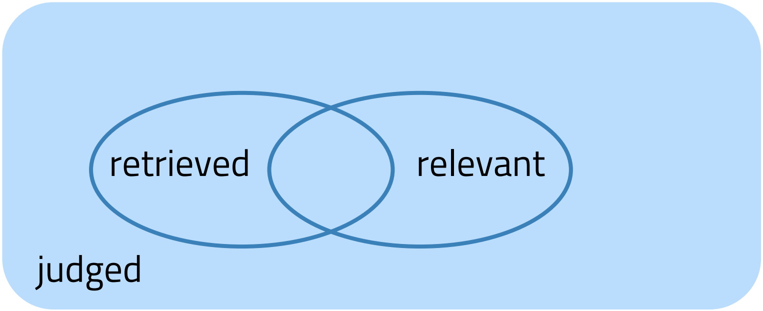
Systems

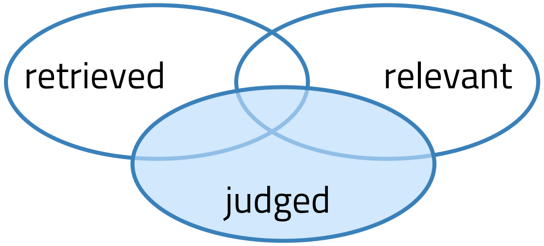
723

Assessor hours

At $20 an hour, that amounts to $14,460. And thus, we are still using the TREC-8 corpus to this day for experiments!

Cranfield vs. TREC depth pooling





Relevant documents not appearing in the pool are missed.

Test collections are vital to ensure continuous algorithmic improvements (one could argue) … however, papers are easier to publish when results are positive.

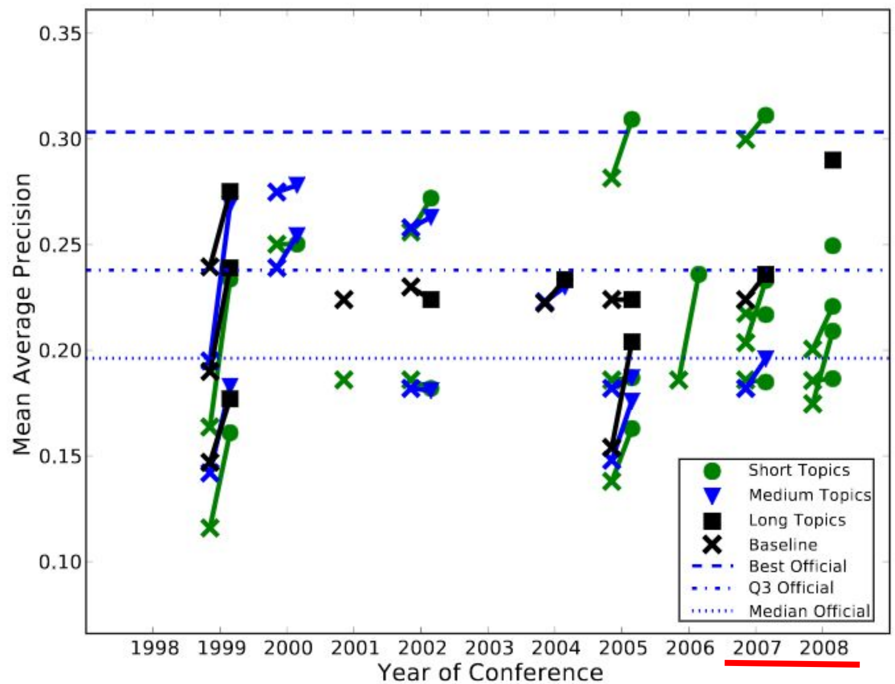


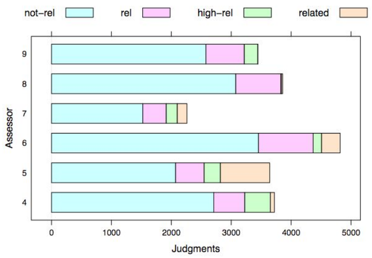
Figure 1: Published MAP scores for the TREC 7 Ad-Hoc collection. The connections show before-after pairs.

Relative Significance is Insufficient: Baselines Matter Too.

Assessor reliability

- Relevance assessments are collected by assessors: can be highly trained information officers (e.g. retired government officials at TREC) or crowd workers (paid 1-5 cents per label) or graduate students or …

- Assessors differ in their assessments, especially so if they are crowdworkers (1)



- Even well-trained assessors assess differently depending on time of day, emotions, order of the documents to judge, etc.

- At a fixed price, its it better (=more stable systems' ranking) to have more topics (and a shallower pool) than fewer topics (and a deeper pool), i.e. topics are a larger source of variance than missing judgments

(1) http://ir.cis.udel.edu/~carteret/papers/sigir10a.pdf

Task-dependent evaluation - question time

Query:
homepage TU Delft

Query:
TU Delft world-wide university ranking

Query:
TU Delft patents nano-technology

Query:
Successful treatment of Newcastle disease

Q1: Informational vs. navigational
Q2: Number of relevant documents
Q3: How many relevant docs need to be retrieved?

Task-dependent evaluation - question time

Query:
homepage TU Delft

Navigational query

1 relevant entry page

Query:
TU Delft world-wide university ranking

Informational query

N relevant entry pages, retrieving some of those is good enough

Query:
TU Delft patents nano-technology

Informational query

N relevant patents, retrieving all is important

Query:
Successful treatment of Newcastle disease

Informational query

N relevant pages, retrieving all is important

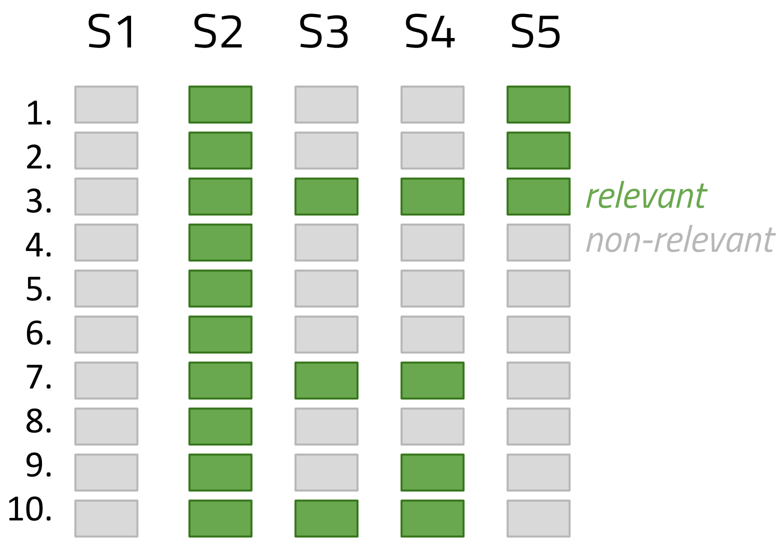
Q1: Informational vs. navigational
Q2: Number of relevant documents
Q3: How many relevant docs need to be retrieved?

Popular Evaluation Measures

There are 60+ published IR metrics, we picked 7 here …. so do not despair by the amount of metrics. It is a tiny part of what is out there.

Precision

One query, five systems



Precision at 10 docs "P@10" 0.0 1.0 0.3 0.4 0.3

Precision measures a system's ability to only retrieve relevant items.

R-precision is P@R where R=number of relevant documents.

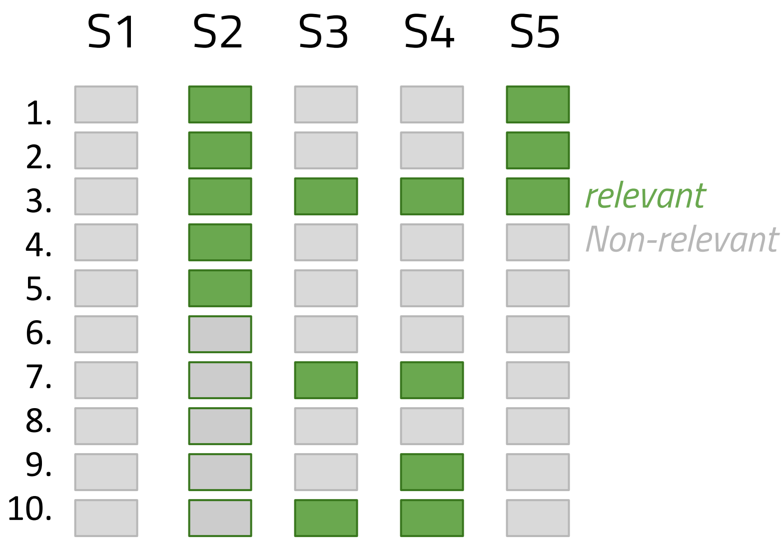
$$precision = \frac{num.\ relevant\ docs\ retrieved}{num.\ docs\ retrieved}$$

Recall

One query, five systems

Recall measures a system's ability to retrieve all R relevant documents.

Recall and Precision are set-based measures. Retrieved are ranked lists.

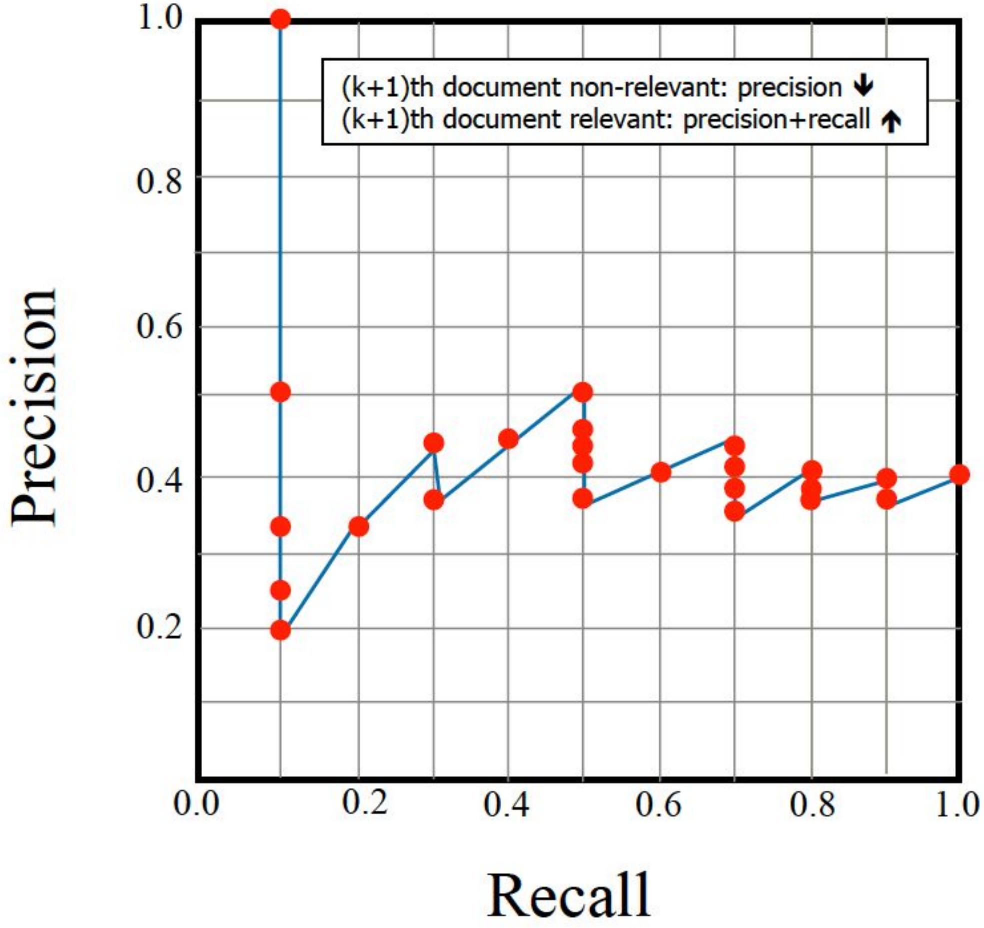


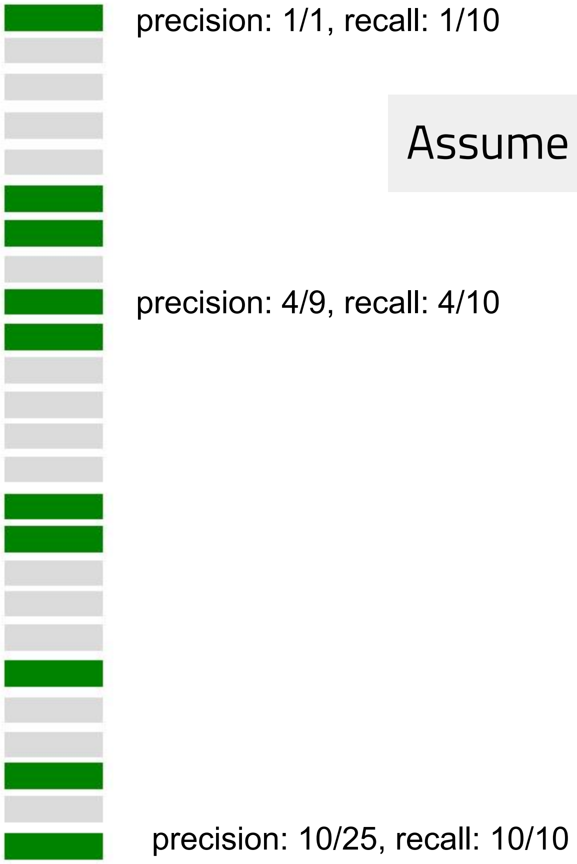
	S1	S2	S3	S4	S5
Recall (assume R=5)	0.0	1.0	0.6	0.8	0.6

$$recall = \frac{num.\ relevant\ docs\ retrieved}{num.\ relevant\ docs\ in\ corpus}$$

Recall-Precision Curve

One query, one system

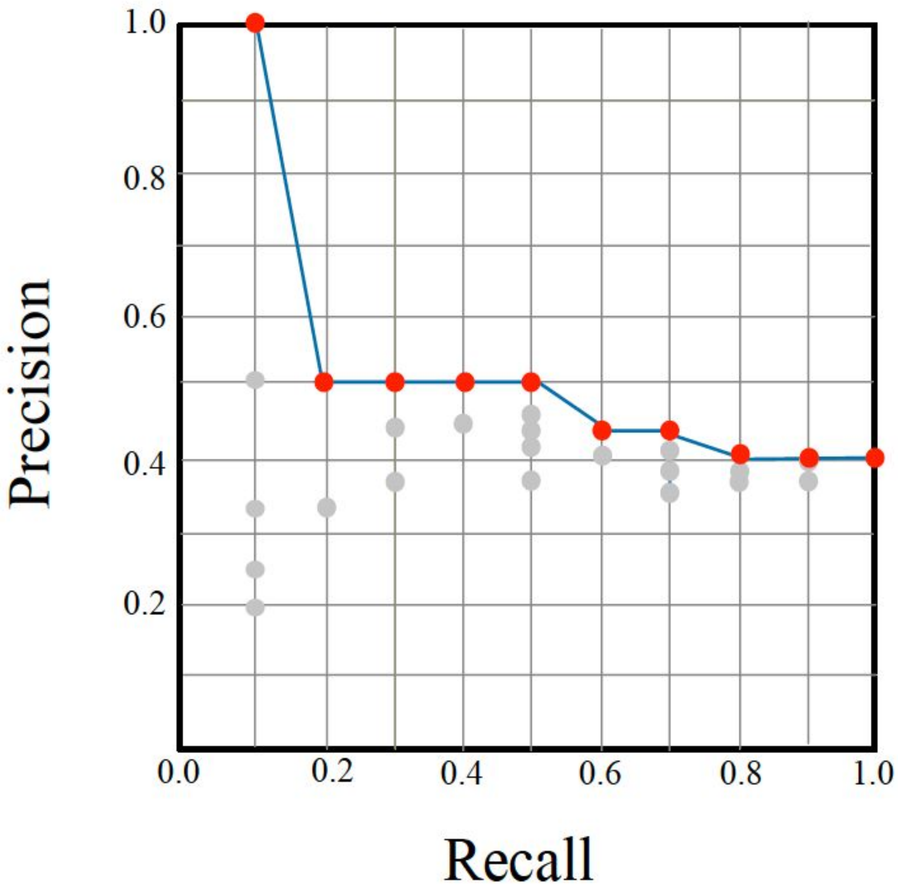




Assume R=10

Recall-Precision Curve

One query, one system

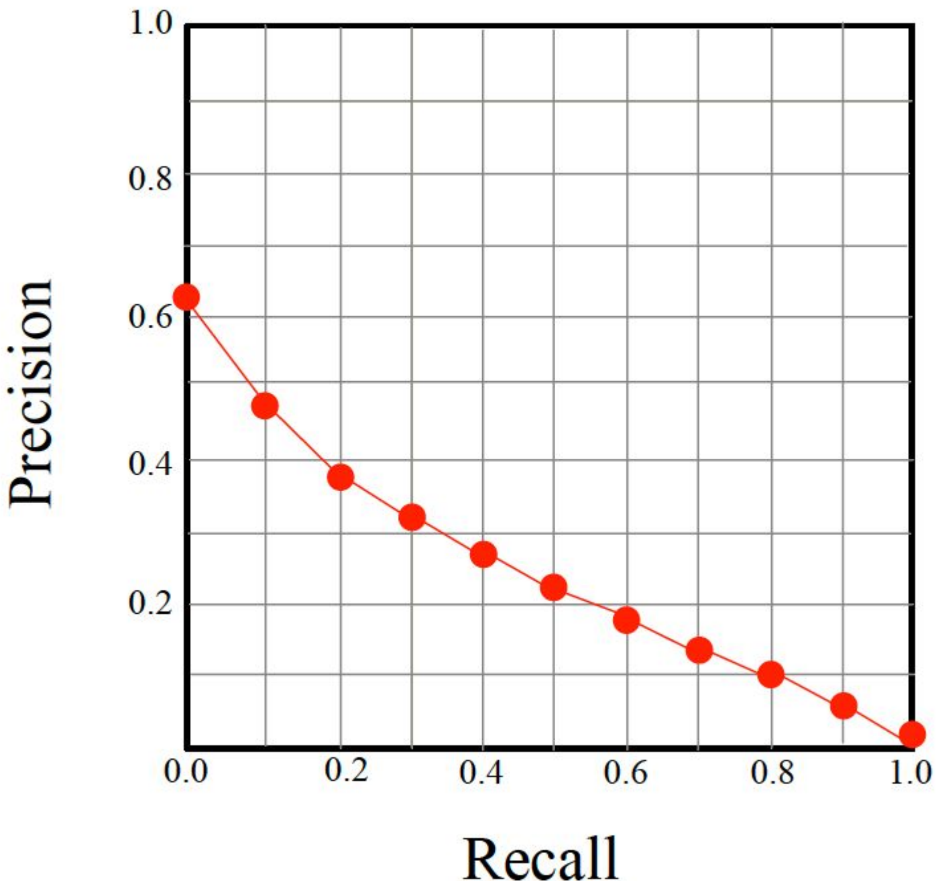


Interpolated precision at recall-level R

$$precision_{\ interp}(r) = max_{r' \geq r}\ precision(r')$$

Recall-Precision Curve

Many queries, one system

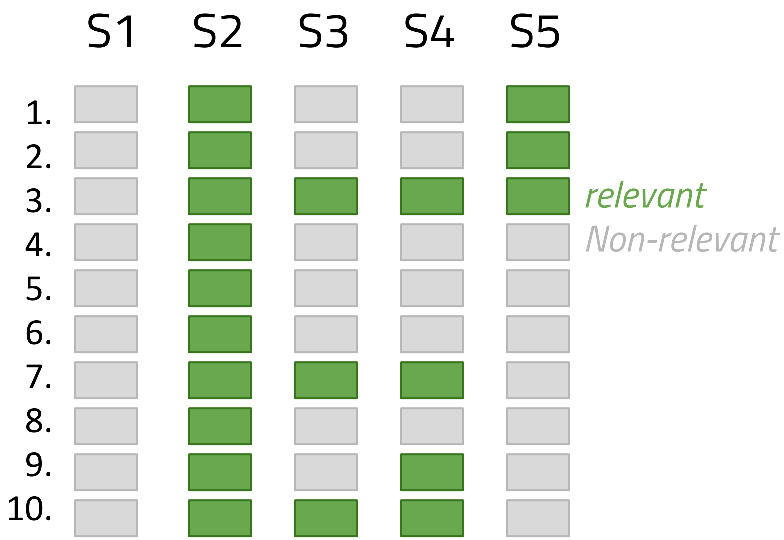


Precision at 11 standard recall values.
Averaged over all queries.

The more relevant documents are retrieved (recall ↗), the more non-relevant documents are retrieved (precision ↙)

Problem: this is a graph, not a single number ... how do systems compare with different precision-recall curves?

Average Precision



AvP 0.0 1.0 0.09 0.13 0.3
(assume R=10)

$$AvP = \frac{1/3 + 2/7 + 3/9 + 4/10}{10}$$

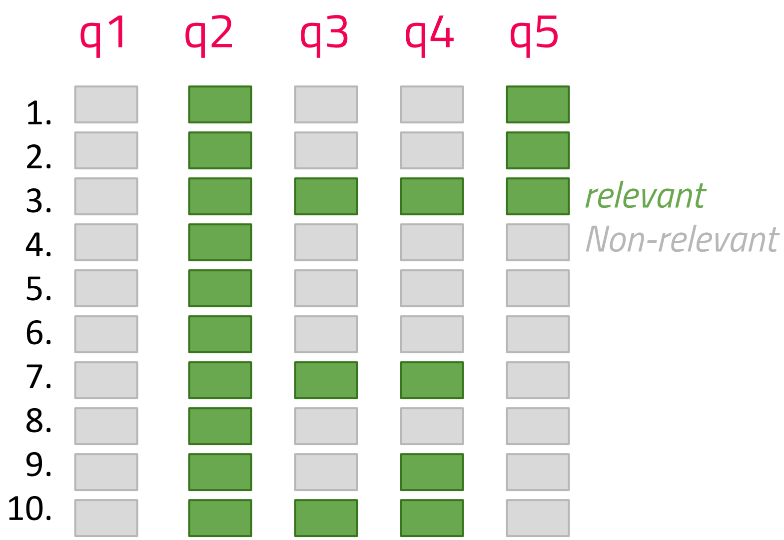
One query, five systems

Average precision takes the order (ranking) of the relevant and non-relevant documents into account

Average precision takes the number R of relevant documents into account.

$$AvP = \frac{\sum_{k=1}^{s} P@k \times rel(k)}{R}$$

Mean Average Precision



AvP 0.0 1.0 0.09 0.13 0.3
(assume R=10)

MAP = 0.364

One system, five queries

Given a set of queries, the average effectiveness is the mean over AvP.

MAP remains one of the most commonly employed retrieval evaluation measure to this day.

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{k=1}^{s} P@k \times rel(k)}{R}$$

Geometric Mean Average Precision

A measure designed to highlight improvements for low-performing topics

Geometric mean of per-topic average precision values (n is the num. topics):

$$GMAP = \sqrt[n]{\prod_n AP_n}$$

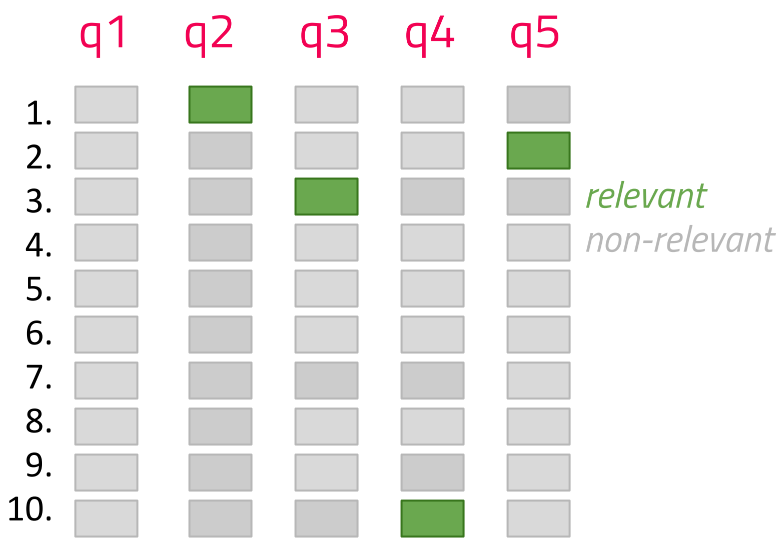
$$= exp \frac{1}{n} \sum_n log\ AP_n$$

Two systems, five queries

	S1	S2
q1	0.60	0.58
q2	0.20	0.18
q3	0.01	0.03
q4	0.04	0.06
q5	0.90	0.90
	MAP = 0.350	MAP = 0.350
	GMAP = 0.134	GMAP = 0.176

S2 performs better on the **worst topics**! Can we have a measure that prefers systems that do well on the worst topics?

Mean Reciprocal Rank



RR 0.0 1.0 0.33 0.01 0.5

MRR=0.369

One system, five queries

One relevant document per query

Reciprocal rank averaged over all queries.

$$RR = \frac{1}{\textit{rank of relevant doc}}$$

Turning away from binary qrels: Normalized Discounted Cumulative Gain (NDCG)

- Standard Web search queries are short (2-3 terms), e.g. "cheap internet", "dinosaurs", "solar panels"
- Graded relevance scales needed (e.g. 0-3)
- NDCG measures the "gain" of documents
- Assumptions:
  - Highly relevant documents are more valuable than marginally relevant documents
  - The greater the ranked position of a relevant document, the less valuable it is for the user
    - Few users go further than the first 10 blue links
    - Probability of reaching the document is lower
    - Users have limited time
    - Users may have seen the information in the document already

Instead of just giving you the end-result, let's look at how the metric was developed.

Turning away from binary qrels: Normalized Discounted Cumulative Gain (NDCG)

Direct cumulative gain can be defined iteratively

$$G' =< 3,2,3,0,0,1,2,2,3,0, \ldots >$$

$$CG_i = \begin{cases} G_i, \ \mathit{if} \ i = 1 \\ CG_{i-1} + G_i, \ \mathit{otherwise} \end{cases}$$

$$CG' =< 3,5,8,8,8,9,11,13,16,16, \ldots >$$

Turning away from binary qrels: Normalized Discounted Cumulative Gain (NDCG)

Discounted cumulative gain: reduce the document score as its rank increases (but not too steeply)

- Divide the document score by the log of its rank
- Base of the logarithm determines discount factor

$$DCG_i = \begin{cases} CG_i, \; if \; i < b \\ CG_{i-1} + G_i / log_b i, \; if \; i \geq b \end{cases}$$

assume b=2

$$CG' \; =< 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \ldots >$$
$$DCG =< 3, 5, 6.9, 6.9, 7.3, 8, 8.7, 9.6, 9.6, \ldots >$$

Turning away from binary qrels: Normalized Discounted Cumulative Gain (NDCG)

Normalized discounted cumulative gain: compare DCG to the theoretically best possible

- Ideal vector sorts the document relevance judgments in decreasing order of relevance

I' is based on the search topic, not the retrieval result!

$$I' =< 3,3,3,2,2,2,1,1,1,1,0,0,0,... >$$
$$CG_{I'} =< 3,6,9,11,13,15,16,17,18,19,19,19,19,... >$$
$$DCG_{I'} =< 3,6,7.9,8.9,9.8,10.5,10.9,11.2,11.5,11.8,... >$$

- The DCG vectors are divided component-wise by the corresponding ideal DCG vectors

Normalization so that a perfect ranking at k for query j is 1

Relevance score assessors gave D at query j

- NDCG for queries Q at rank k:

$$NDCG(Q,k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{log_2(1+m)}$$

trec_eval

- T.H.E. standard tool to evaluate a system's ranking (given a set of qrels)

- Maintained by the TREC community, with 60+ measures, many of which can be parameterized (e.g. P@10)

- Some measures are obsolete by now

- https://github.com/usnistgov/trec_eval/

```
extern TREC_MEAS te_meas_runid;
extern TREC_MEAS te_meas_num_q;
extern TREC_MEAS te_meas_num_ret;
extern TREC_MEAS te_meas_num_rel;
extern TREC_MEAS te_meas_num_rel_ret;
extern TREC_MEAS te_meas_map;
extern TREC_MEAS te_meas_gm_map;
extern TREC_MEAS te_meas_Rprec;
extern TREC_MEAS te_meas_bpref;
extern TREC_MEAS te_meas_recip_rank;
extern TREC_MEAS te_meas_iprec_at_recall;
extern TREC_MEAS te_meas_P;
extern TREC_MEAS te_meas_relstring;
extern TREC_MEAS te_meas_recall;
extern TREC_MEAS te_meas_infAP;
extern TREC_MEAS te_meas_gm_bpref;
extern TREC_MEAS te_meas_Rprec_mult;
extern TREC_MEAS te_meas_utility;
extern TREC_MEAS te_meas_11pt_avg;
extern TREC_MEAS te_meas_binG;
extern TREC_MEAS te_meas_G;
extern TREC_MEAS te_meas_ndcg;
extern TREC_MEAS te_meas_ndcg_rel;
extern TREC_MEAS te_meas_Rndcg;
extern TREC_MEAS te_meas_ndcg_cut;
extern TREC_MEAS te_meas_map_cut;
extern TREC_MEAS te_meas_relative_P;
extern TREC_MEAS te_meas_success;
extern TREC_MEAS te_meas_set_P;
extern TREC_MEAS te_meas_set_relative_P;
extern TREC_MEAS te_meas_set_recall;
extern TREC_MEAS te_meas_set_map;
extern TREC_MEAS te_meas_set_F;
extern TREC_MEAS te_meas_num_nonrel_judged_ret;
extern TREC_MEAS te_meas_prefs_num_prefs_poss;
extern TREC_MEAS te_meas_prefs_num_prefs_ful;
extern TREC_MEAS te_meas_prefs_num_prefs_ful_ret;
extern TREC_MEAS te_meas_prefs_simp;
extern TREC_MEAS te_meas_prefs_pair;
extern TREC_MEAS te_meas_prefs_avgjg;
extern TREC_MEAS te_meas_prefs_avgjg_Rnonrel;
extern TREC_MEAS te_meas_prefs_simp_ret;
extern TREC_MEAS te_meas_prefs_pair_ret;
extern TREC_MEAS te_meas_prefs_avgjg_ret;
extern TREC_MEAS te_meas_prefs_avgjg_Rnonrel_ret;
extern TREC_MEAS te_meas_prefs_simp_imp;
extern TREC_MEAS te_meas_prefs_pair_imp;
extern TREC_MEAS te_meas_prefs_avgjg_imp;
extern TREC_MEAS te_meas_map_avgjg;
extern TREC_MEAS te_meas_P_avgjg;
extern TREC_MEAS te_meas_Rprec_mult_avgjg;
extern TREC_MEAS te_meas_yaap;
```

Statistical significance tests

Significance tests

- Given the results from a number of queries, how can we conclude that ranking **algorithm A** is better than **algorithm B**?
- Significance tests enable us to reject the **null hypothesis** (no difference) in favor of the **alternative hypothesis** (B is better than A)
- (`trec_eval` does not come with those)

Significance tests

1. Compute the effectiveness measure for every query for both rankings.
2. Compute a *test statistic* based on a comparison of the effectiveness measures for each query. The test statistic depends on the significance test, and is simply a quantity calculated from the sample data that is used to decide whether or not the null hypothesis should be rejected.
3. The test statistic is used to compute a *P-value*, which is the probability that a test statistic value at least that extreme could be observed if the null hypothesis were true. Small P-values suggest that the null hypothesis may be false.
4. The null hypothesis (no difference) is rejected in favor of the alternate hypothesis (i.e., B is more effective than A) if the P-value is $\leq \alpha$, the *significance level*. Values for α are small, typically .05 and .1, to reduce the chance of a Type I error.

Paired t-Test

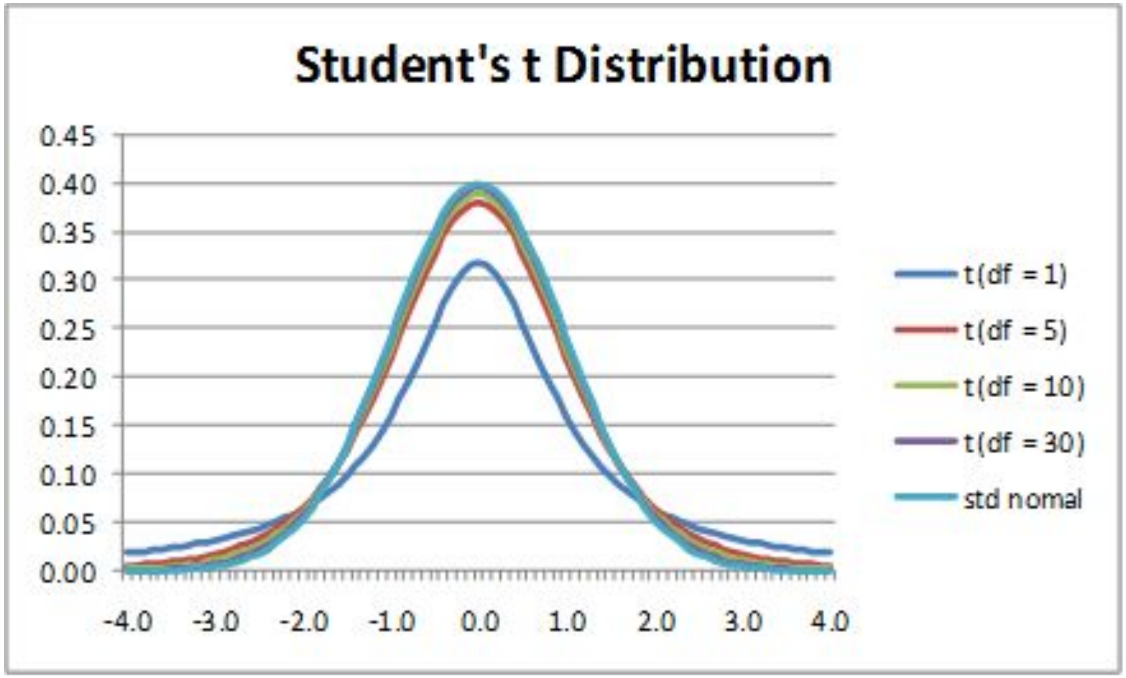
- Assumption is that the difference between the effectiveness values is a sample from a normal distribution

- Null hypothesis is that the mean of the distribution of differences is zero

- Test statistic

$$t = \frac{\overline{B-A}}{\sigma_{B-A}} . \sqrt{N}$$

Student's t distribution



In Python:

```
from scipy import stats
stats.ttest_rel(A,B)
```

https://towardsdatascience.com/inferential-statistics-series-t-test-using-numpy-2718f8f9bf2f
https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.ttest_rel.html

$$t = \frac{\overline{B-A}}{\sigma_{B-A}} . \sqrt{N}$$

Example

Queries	A	B	C	D
1.	0.1	0.2	0.101	0.15
2.	0.2	0.1	0.201	0.20
3.	0.9	0.5	0.901	0.99
4.	0.5	0.9	0.501	0.65
5.	0.5	0.5	0.501	0.55
6.	0.1	0.1	0.101	0.60
7.	0.1	0.1	0.101	0.15
8.	0.5	0.5	0.501	0.50
9.	0.9	0.9	0.900	0.95
10.	0.3	0.3	0.301	0.45
MAP	0.409	0.409	0.411	0.519
p-value		1.0	0.0000	0.043

“A statistically significant result is one that is unlikely to be the result of chance. But a practically significant result is meaningful in the real world. It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically non significant and important."

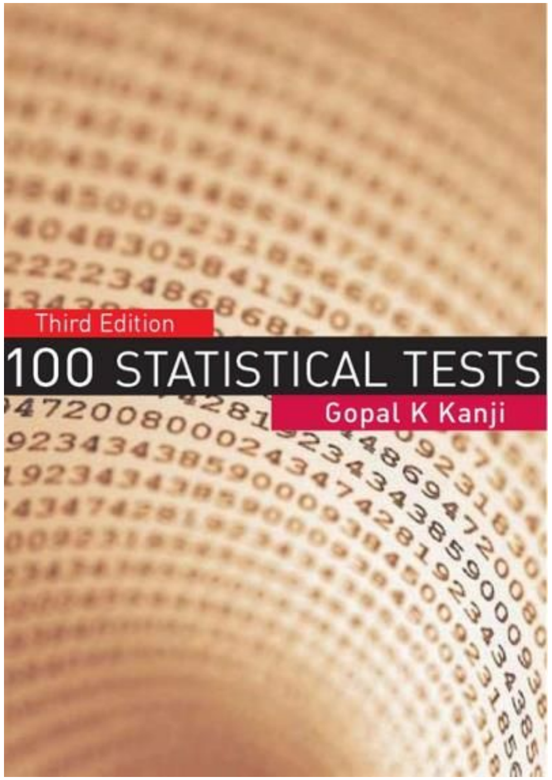
Ellis, Paul D. The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. Cambridge University Press, 2010.

Which test to use depends on the setting ...

- Commonly used in IR papers:
  - Mann-Whitney-Wilcoxon test (Wilcoxon Rank-Sum test)
  - Wilcoxon signed rank test (paired)

- Software packages exist in R, Python, SPSS, etc. that help you test

- A good book to find the right test for a given scenario



User-centered system evaluation

What if we are no longer happy to consider the toy Eiffel tower only?

Lets evaluate real systems with **real users** (=people using the system) at small scale.

Instead of "is the system any good?" we are now interested in "can users use the system to retrieve any good results?".

Relevant Factors in Interactive IR (or IIR)

- **Physical, cognitive and affective**: satisfaction with the system, difficulty of use (cognitive load), feelings after usage, etc.

- **Interactions between users and systems**: number of clicks, number of queries issued, query length, etc.

- **Interactions between users and information**: dwell time on a document, terms extracted from a snippet and used in a query, etc.

IIR approaches are diverse

- An evaluation measures the quality of a system, interface widget, etc. while an experiment compares at least two items (usually a baseline and an experimental system) with each other

- Lab (lots of control but artificial), online (some control, still artificial) vs. naturalistic (little control) studies

- Longitudinal studies: require an extended period of time (e.g. *investigate how students interact across 10 weeks with search engine X during their literature survey*)

- Wizard of Oz study: participants interact with a system they believe to be automated (in reality it is operated by a human)

Variables

Independent variables:
the causes

E.g. investigate how young an old people use an experimental and baseline IR system.

→ age is the independent variable

Dependent variables:
the effects

E.g. satisfaction with the search systems

Confounding variables

Affect the independent and dependent variables, but have not been controlled by the experimenter.

E.g. older people are not as familiar with the experimental device as young people.

The experimental design in IIR examines the relationship between 2 or more systems (independent variable) on some set of outcome measures (dependent variables).

Measurements

- "Query logs" or "transaction logs" are usually analyzed
- What can and should be measured depends on the research questions and the setup of the experiment (in a lab or online?)
- Logging clicks is insufficient as user studies have usually few participants (in contrast to Google/Bing with billions of clicks per day)
- Client-side logging is often necessary to track mouse hovers, document dwell time, eye movements (can be done via the Webcam), user activities in other browser tabs/windows, rephrasing of queries, ...

Online evaluation: “large-scale” A/B testing

Online evaluation

- A/B tests commonly require large amounts of users and are typically employed by large-scale Web portals (e.g. Bing runs hundreds of A/B tests concurrently [1])

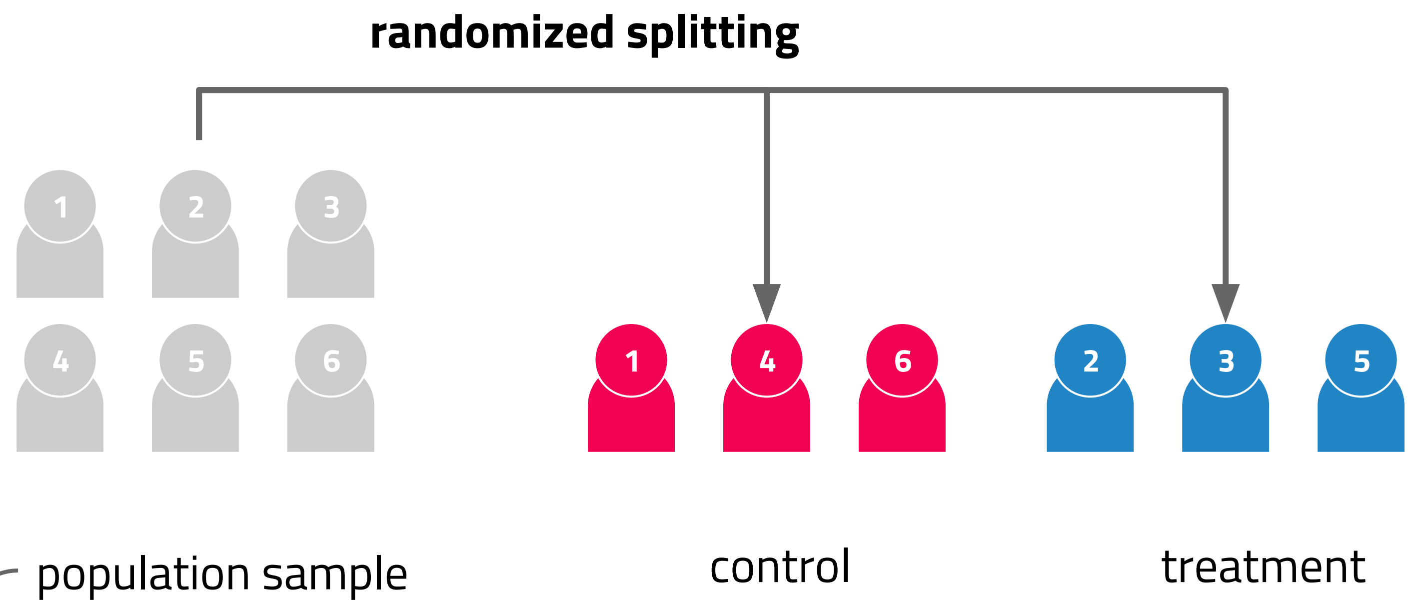
- Focus on **implicit user feedback** (instead of explicit feedback - e.g. by answering “Is this relevant to you?”)
  - Derived from observable user activity
  - Captured during natural interactions

- Implicit signals with various levels of noise
  - *Clicks, dwell-times, purchase decisions*

[1] http://www.exp-platform.com/Documents/2013%20controlledExperimentsAtScale.pdf

Between-subject experiments

Each user is exposed to a single variant (note that a user can participate in multiple experiments at once)



In practice, initially a small sample of all users (e.g., 5%) may be in the exp.

A/B test process

Also known as: <u>Flights</u> (Microsoft), <u>1% tests</u> (Google), <u>bucket tests</u> (Yahoo!), <u>randomized clinical trials</u> (medicine)

- Randomly split traffic between two or more versions
  - A (**control**, typically the existing system)
  - B (**treatment 1**)
  - C (**treatment 2**)
  - ...

May also be called "experimental" group

- Collect metrics of interest (e.g. ad revenue, retention, product conversion)

- Determine impact on the previously identified metrics

- Important: due to the large size of the sample, stat. sig. differences are easy to achieve (effect size becomes much more important [1])

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/

Most common online evaluation metrics

- **Document-level**
    - Click rate, click models
- **Ranking-level**
    - Reciprocal rank, CTR@k, time-to-click, abandonment
- **Session-level**
    - Queries per session, session length, time to first click

Lecture Summary

- Evaluation is not straightforward
- The task is paramount to the correct choice of evaluation measure
- Still researched today - every few months or so a new metric is being proposed
- The most widely used offline eval. metrics today are MAP and NDCG
- A very accessible survey on evaluation: "*Test Collection Based Evaluation of Information Retrieval Systems*" by Mark Sanderson [1]
- A great survey on interactive IR evaluation: "Methods for Evaluating Interactive Information Retrieval Systems with Users" by Diane Kelly [2]
- A tutorial on A/B testing: "A/B Testing at Scale Tutorial" by Pavel Dmitriev et al. [3]

[1] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.646.8451&rep=rep1&type=pdf
[2] https://pdfs.semanticscholar.org/328c/7b4ce5a0d81326ee2a3befa0f2dd630a48c1.pdf
[3] http://exp-platform.com/2017abtestingtutorial/

That’s it!

Don’t forget that milestone M1 (IR vs NLP) is coming up next week!

Slack: in4325.slack.com

Email: in4325-ewi@tudelft.nl