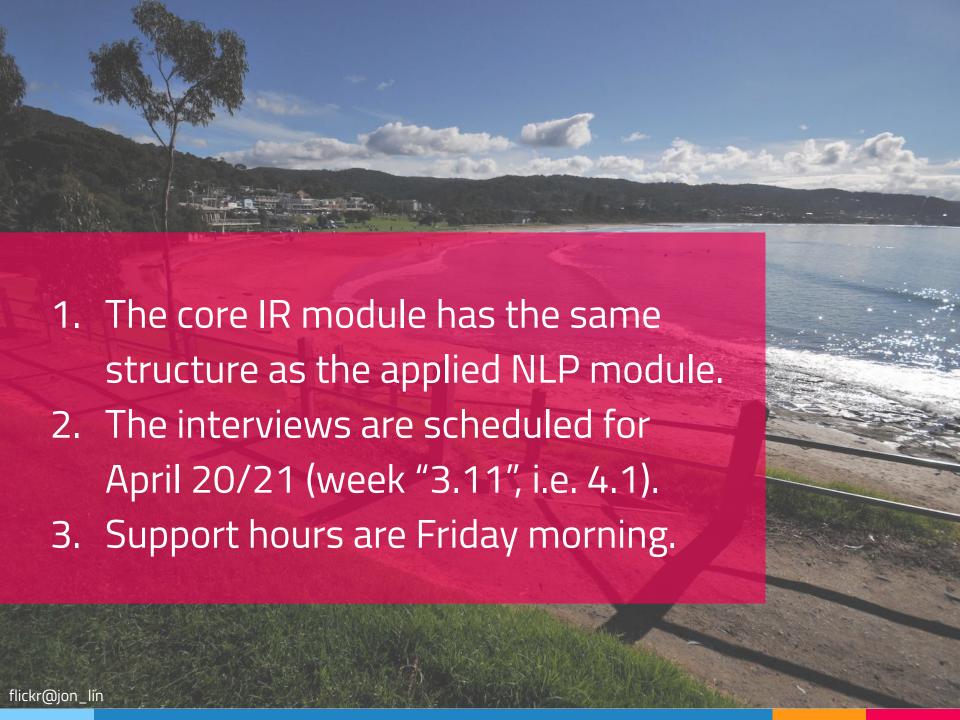
IN4325 Evaluation in IR

Claudia Hauff (WIS, TU Delft)



Core IR project

Reproducing an IR paper of your choice from a given list (19 papers)

- a. IR papers with user studies
- b. IR papers without user studies

Investigating/reproducing a research idea of your own choice

Moving the lectures online or not?

Continue as-is.

Pre-record lectures and use the lecture slot as Q&A session via zoom (or Skype or ...). Won't be ready by tomorrow but next week.

Moving the lectures online or not?

Continue as-is.

Pre-record lectures and use the lecture slot as Q&A session via zoom (or Skype or ...). Won't be ready by tomorrow but next week.

Support hours are offered online via Skype (email me your Skype IDs) and f2f.

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

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

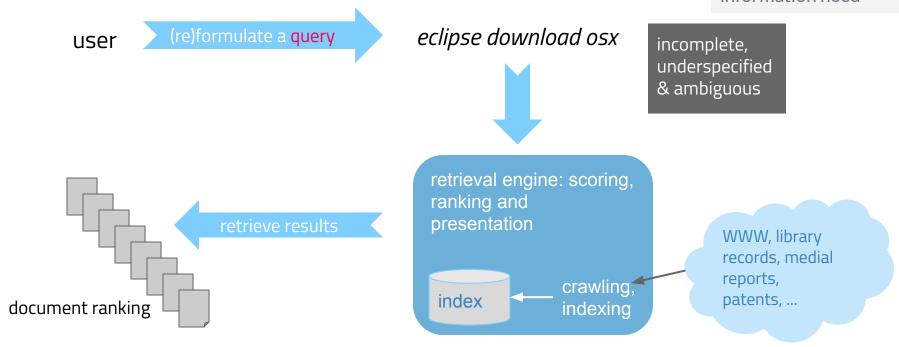
Topic the user wants to know more about

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



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?

(re)formulate a query eclipse download osx user incomplete, underspecified & ambiguous assess relevance to information need retrieval engine: scoring, ranking and presentation retrieve results WWW, library records, medial reports, crawling, index patents, ... document ranking indexing

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

Query = textual input

Results = ranked list of search result snippets (i.e. "ten blue links")

Actions = click, view

Proactive search

(zero query search)

Query = none

<u>Results</u> = a single information card

Actions = view

Voice search

Query = speech input

Results = speech output

<u>Actions</u> = 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 <u>link</u> oncology patients to clinical trials for new treatments"

News search

"... will foster research that establishes a new sense what <u>relevance</u> <u>means for news</u> search."

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

Benchmarks drive our community

CLEF

Conference and Labs of the Evaluation Forum

http://www.clef-initiative.eu/

EUROPE

TREC

USA

MediaEval

Benchmarking Initiative for Multimedia Evaluation

http://www.multimediaeval.org/

EUROPE

TRECVID

USA

NTCIR

NII Test Collection for IR Systems

http://research.nii.ac.j p/ntcir/index-en.html

JAPAN

FIRE

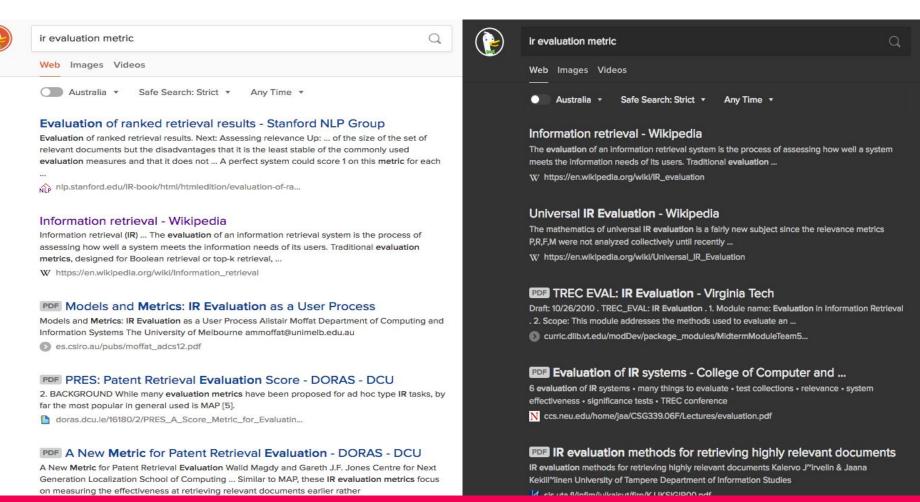
Forum for Information Retrieval Evaluation

http://www.isical.ac.in/~clia/

INDIA

'... 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

Question time



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

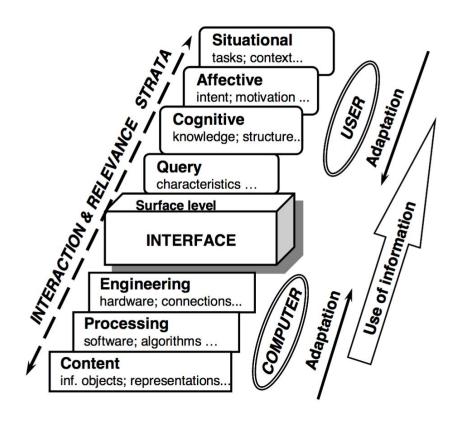
Relevance

Why are we starting with the evaluation lecture in the core IR module 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)

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
 - \rightarrow 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.

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 * (system-centered)

^{*} Mainstream way of evaluation. Empirical. Another approach is the axiomatic one (found in more 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>



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

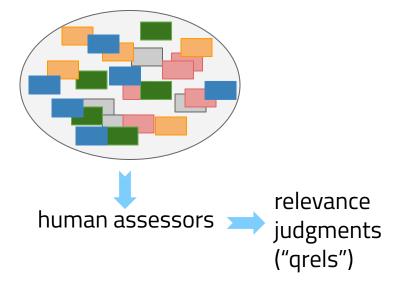
, query .

Commonly used today

		//	/ ·		<u></u>	
	s1	S2	S3	S4	S5	
document ranking						1
ignore						

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







Topics (ad hoc task)

▶86,830

Pooled documents (k=100)

129

Systems

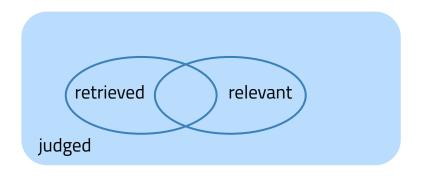
723

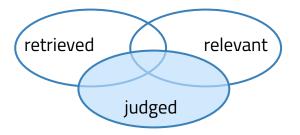
Assessor hours

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

TREC-8 numbers (ran in 1999)

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.

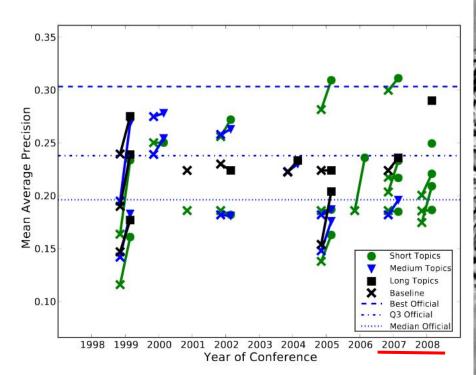


Figure 1: Published MAP scores for the TREC 7 Ad-Hoc collection. The connections show before-after pairs.

This is an ongoing issue

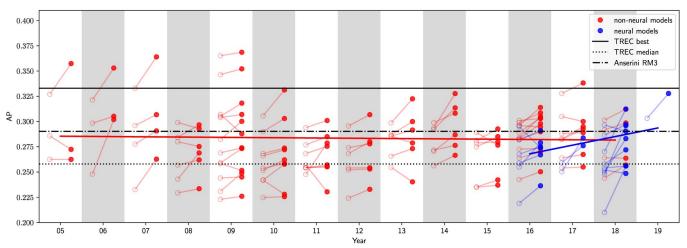


Figure 1: Visualization results on Robust04, where baseline and best AP scores are represented by empty and filled circles.



This is an ongoing issue

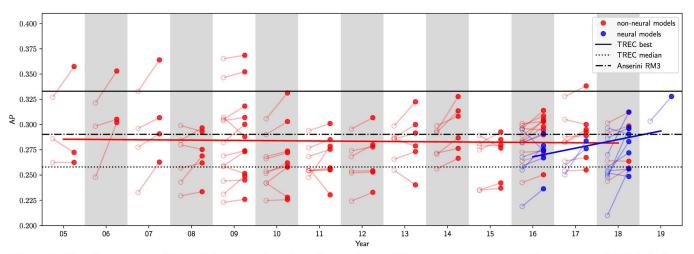


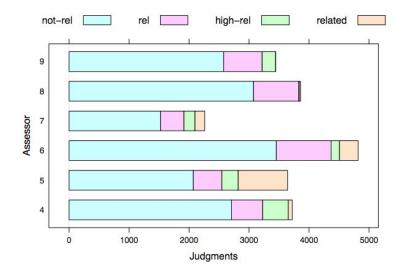
Figure 1: Visualization results on Robust04, where baseline and best AP scores are represented by empty and filled circles.

2020: BERT finally seems to beat those old baselines.



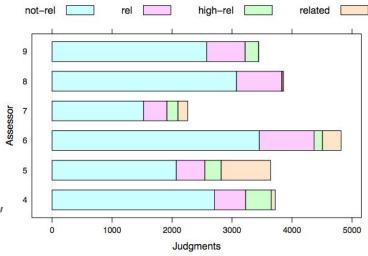
Assessor reliability

- 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)



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

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

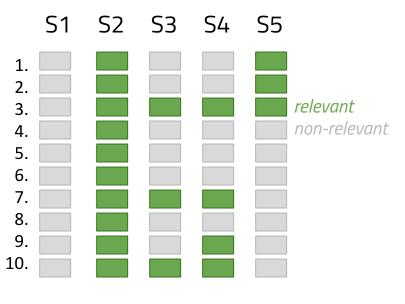
Q2: Number of relevant documents

Q3: How many relevant docs need to be

retrieved?

Popular batch evaluation measures

Precision



One query, five systems

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

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

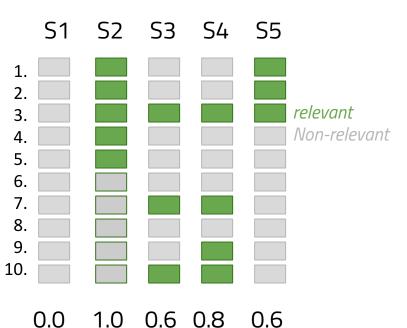
Precision 0.0 1.0 0.3 0.4 0.3 at 10 docs "P@10"

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

Recall

Recall

(assume R=5)



One query, five systems

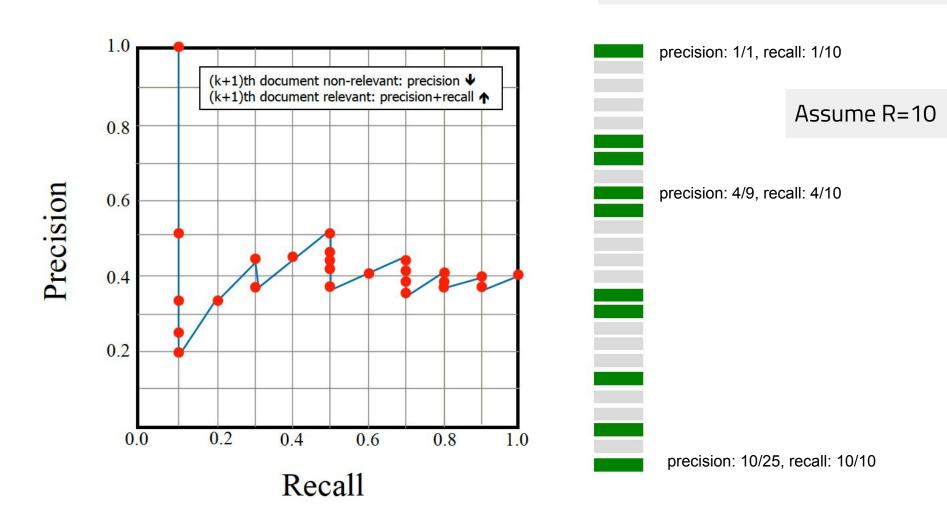
Recall measures a system's ability to retrieve all R relevant documents.

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

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

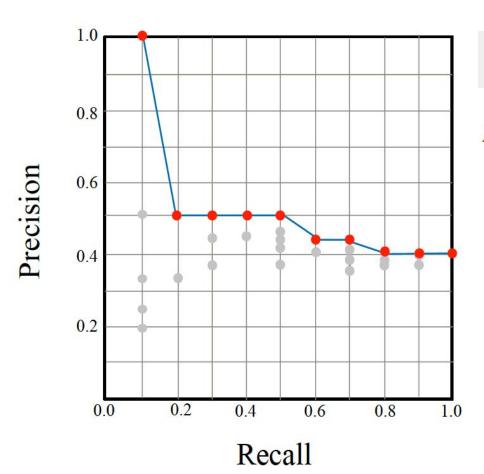
Recall-Precision Curve

One query, one system



Recall-Precision Curve

One query, one system

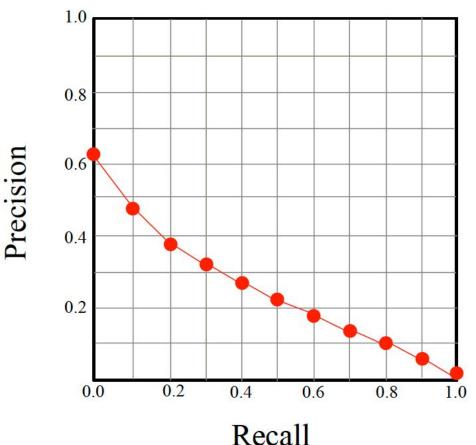


Interpolated precision at recall-level R

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

Recall-Precision Curve

Many queries, one system

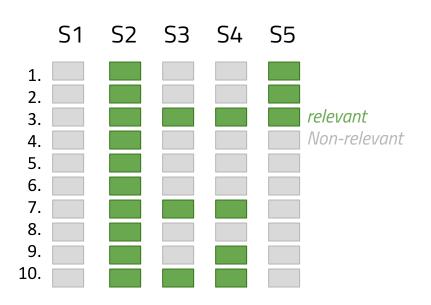


Precision at 11 standard recall values. Averaged over all queries.

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

Average Precision

1.0



0.09 0.13 0.3

AvP 0.0 (assume R=10)

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

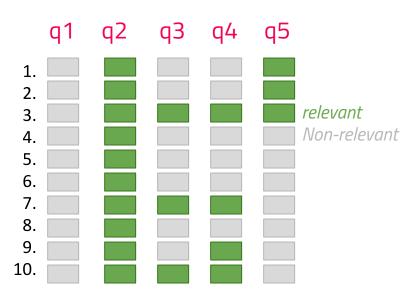
One query, five systems

Average precision takes the order (ranking) of the relevant and non-relevant documents into account. We evaluate up to rank s.

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

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

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.

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

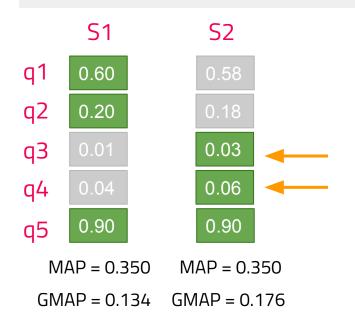
Geometric Mean Average Precision

A measure designed to highlight improvements for low-performing topics

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

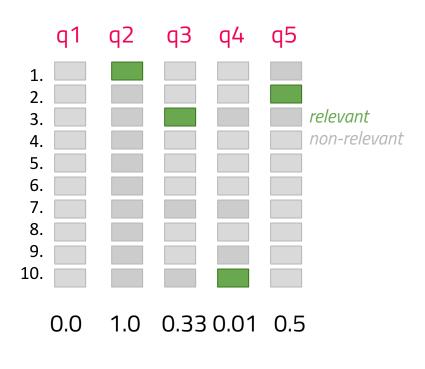
$$GMAP = \sqrt[n]{\prod_{n} AP_{n}}$$
$$= exp \frac{1}{n} \sum_{n} log AP_{n}$$

Two systems, five queries



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

MRR=0.369

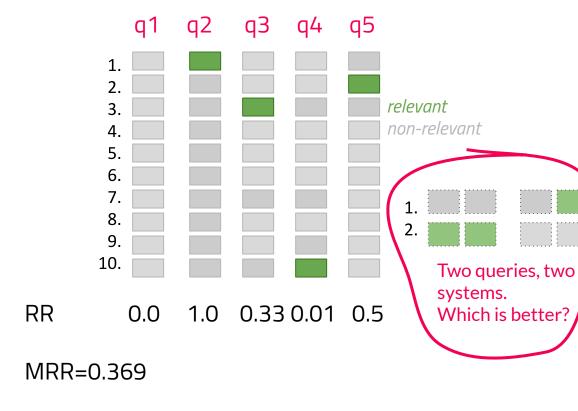
One system, five queries

One relevant document per query.

Reciprocal rank (RR) averaged over all queries.

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

Mean Reciprocal Rank



One system, five queries

One relevant document per query.

Reciprocal rank (RR) averaged over all queries.

Heavily used (easy to compute and understand), but also heavily criticized (1).

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

- 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

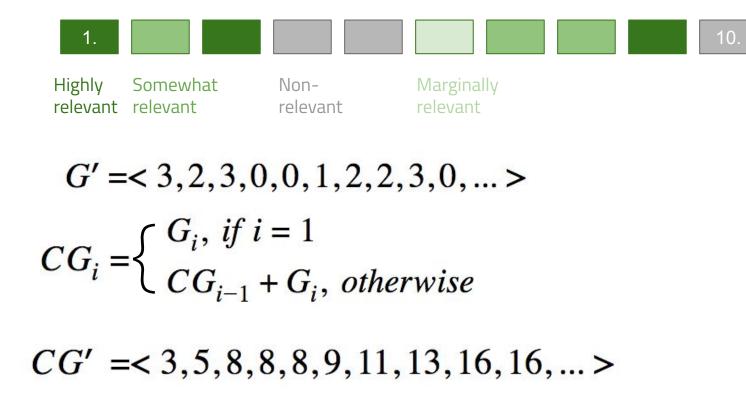
Instead of just giving you the end-result, let's look at how the metric was developed.

- 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

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

- 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

Direct cumulative gain (CG) can be defined iteratively:



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

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

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

assume b=2

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

 $DCG = <3,5,6.9,6.9,7.3,8,8.7,9.6,9.6,...>$

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

 Ideal vector I' sorts the document relevance judgments in decreasing order of relevance

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

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

 NDCG: the DCG vectors are divided component-wise by the corresponding ideal DCG vectors (and then averaged across topics)

trec eval

- T.H.E. standard tool to evaluate a system's ranking (given a set of topics and grels)
- Maintained by the TREC 189 extern TREC_MEAS te_neas_nap_avgjg; community, with 60+ 110 extern TREC_MEAS te_meas_P_avgjg; 111 extern TREC_MEAS te_meas_Rprec_mult_avgjg; 112 extern TREC_MEAS te_meas_yaap; 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 num ret: TREC_MEAS te_meas_num_rel_ret;

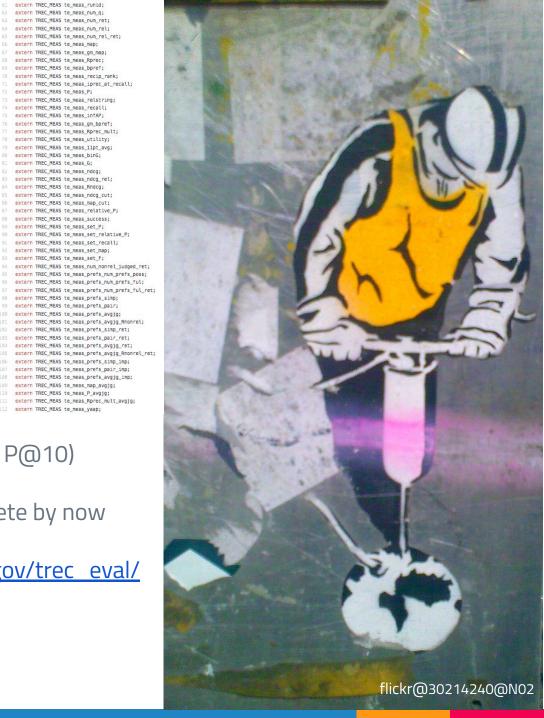
extern TREC MEAS te meas boref; extern TREC_MEAS te_meas_recip_rank; TREC_MEAS te_meas_iprec_at_recall;

extern TREC_MEAS te_neas_qn_bpref;

extern TREC MEAS te meas ndcg rel; extern TREC MEAS te neas map cut;

89 extern TREC MEAS to meas set P:

extern TREC_MEAS te_meas_P; TREC MEAS te meas relstring; extern TREC_MEAS te_meas_infAP;



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

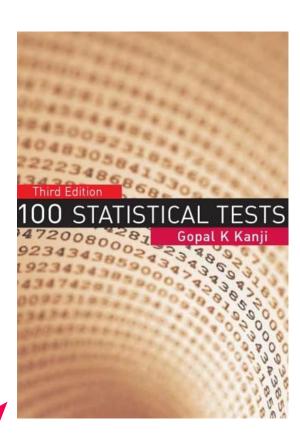
- Compute the effectiveness measure for every query for both rankings.
- 2. Compute a test statistic based on a comparison of the effectiveness measures for each query. The test statistic depends on the significance test, and is simply a quantity calculated from the sample data that is used to decide 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 ≤ α, the significance level. Values for α are small, typically .05 and .1, to reduce the chance of a Type I error.

66

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

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 evaluation

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

Let's 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 <u>users use</u> 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

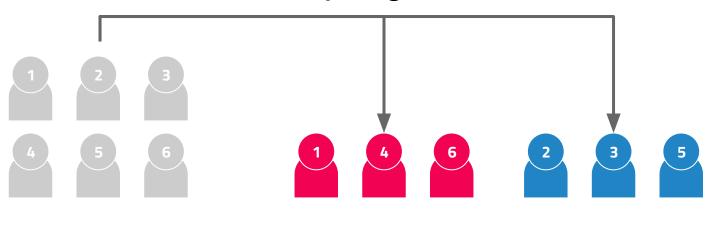
Online evaluation

- 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

Between-subject experiments

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

randomized splitting



population sample

control

treatment

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



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])

Common online evaluation metrics

- Document-level

- Click rate
- Click models (modeling of users' clicks during search)

- Ranking-level

- Mean reciprocal rank
- Click through rate CTR@k (ratio of clicks vs. skips)
- Time to click
- Abandonment (tricky: positive abandonment ...)

- Session-level

- Queries per session
- Session length
- Time to first click

Lecture Summary

- Evaluation is **not** straightforward
- The <u>task</u> is paramount to the choice of evaluation measure
- Remains heavily researched today: every few months or so a new metric is being proposed
- The most widely used offline evaluation metrics today are MAP and NDCG (though MRR is making a comeback!)

Additional materials

- 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]

That's it!

Don't forget that you need to sign up for an IR project group!

Slack: in4325.slack.com

Email: in4325-ewi@tudelft.nl