Introduction to

Information Retrieval

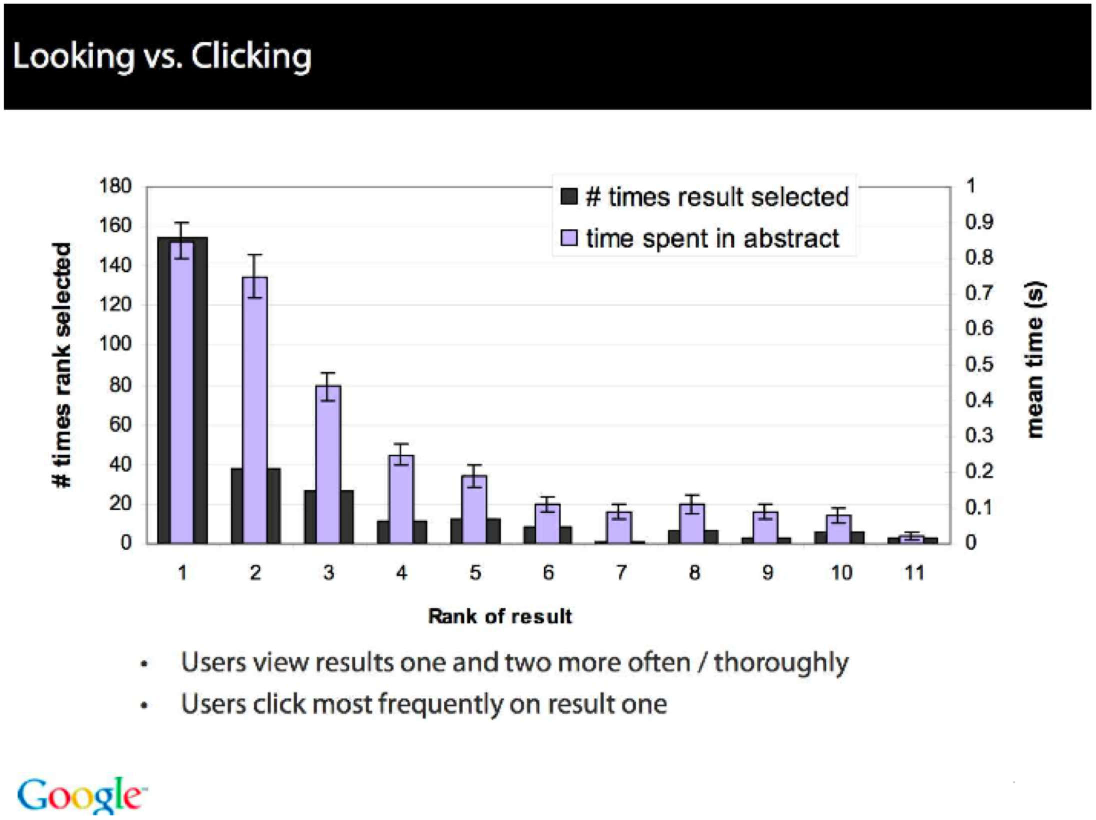
Evaluation & Result Summaries

Evaluation issues

- Presenting the results to users
- Evaluating the quality of results
  - qualitative evaluation
  - quantitative evaluation

Presenting the Information to users

- Results summaries:
  - Making our “good results” usable to a user

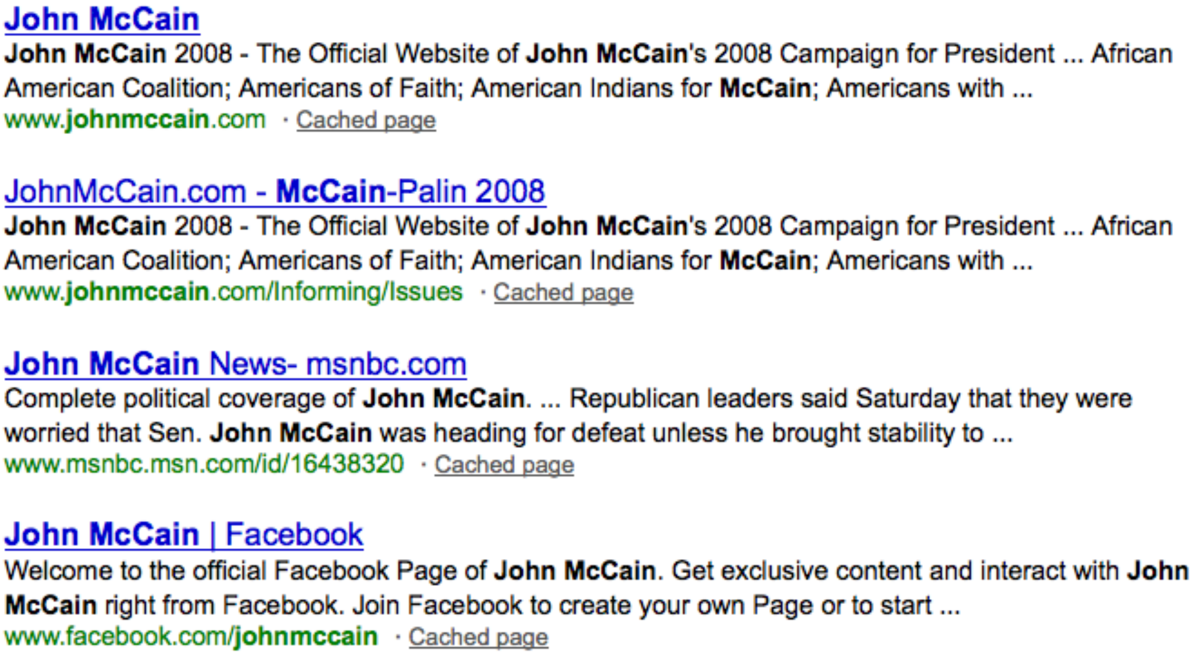


Looking vs. Clicking

- Users view results one and two more often / thoroughly
- Users click most frequently on result one

Result Summaries

- Having ranked the documents matching a query, we wish to present a results list
- Most commonly, a list of the document titles plus a short summary



John McCain
John McCain 2008 - The Official Website of **John McCain**'s 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.**johnmccain**.com · Cached page

JohnMcCain.com - **McCain**-Palin 2008
John McCain 2008 - The Official Website of **John McCain**'s 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.**johnmccain**.com/Informing/Issues · Cached page

John McCain News- msnbc.com
Complete political coverage of **John McCain**. ... Republican leaders said Saturday that they were worried that Sen. **John McCain** was heading for defeat unless he brought stability to ...
www.msnbc.msn.com/id/16438320 · Cached page

John McCain | Facebook
Welcome to the official Facebook Page of **John McCain**. Get exclusive content and interact with **John McCain** right from Facebook. Join Facebook to create your own Page or to start ...
www.facebook.com/**johnmccain** · Cached page

Summaries

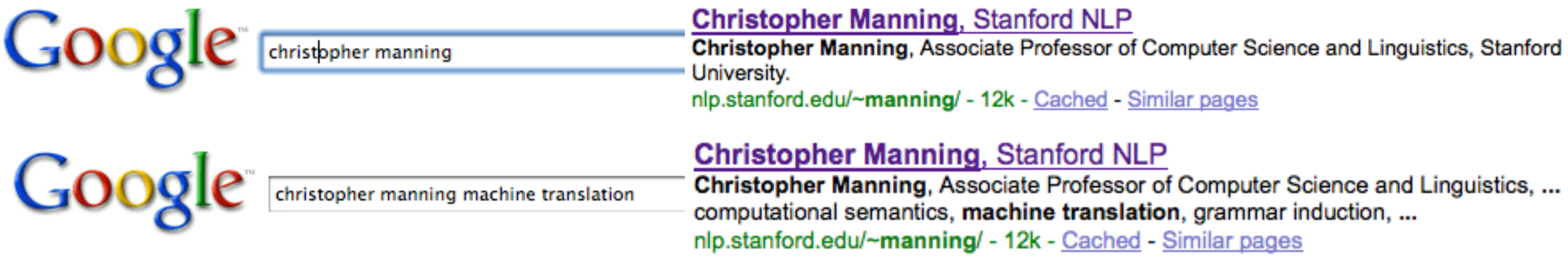
- The title is typically automatically extracted from document metadata. What about the summaries?
  - This description is crucial.
  - User can identify good/relevant hits based on description.
- Two basic kinds:
  - Static
  - Dynamic
- A **static summary** of a document is always the same, regardless of the query that hit the doc
- A **dynamic summary** is a *query-dependent* attempt to explain why the document was retrieved for the query at hand

Static summaries

- In typical systems, the static summary is a **subset of the document**
- Simplest heuristic: the first 50 (or so – this can be varied) words of the document
  - Summary cached at indexing time
- More sophisticated: extract from each document a set of "key" sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
- Most sophisticated: NLP used to synthesize a summary
  - Seldom used in IR; cf. text summarization work

Dynamic summaries

- Present one or more "windows" within the document that contain several of the query terms
  - "KWIC" snippets: Keyword in Context presentation
- Generated in conjunction with scoring
  - If query found as a phrase, all or some occurrences of the phrase in the doc
  - If not, document windows that contain multiple query terms



christopher manning

Christopher Manning, Stanford NLP
Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University.
nlp.stanford.edu/~**manning**/ - 12k - Cached - Similar pages

christopher manning machine translation

Christopher Manning, Stanford NLP
Christopher Manning, Associate Professor of Computer Science and Linguistics, ... computational semantics, **machine translation**, grammar induction, ...
nlp.stanford.edu/~**manning**/ - 12k - Cached - Similar pages

Generating dynamic summaries

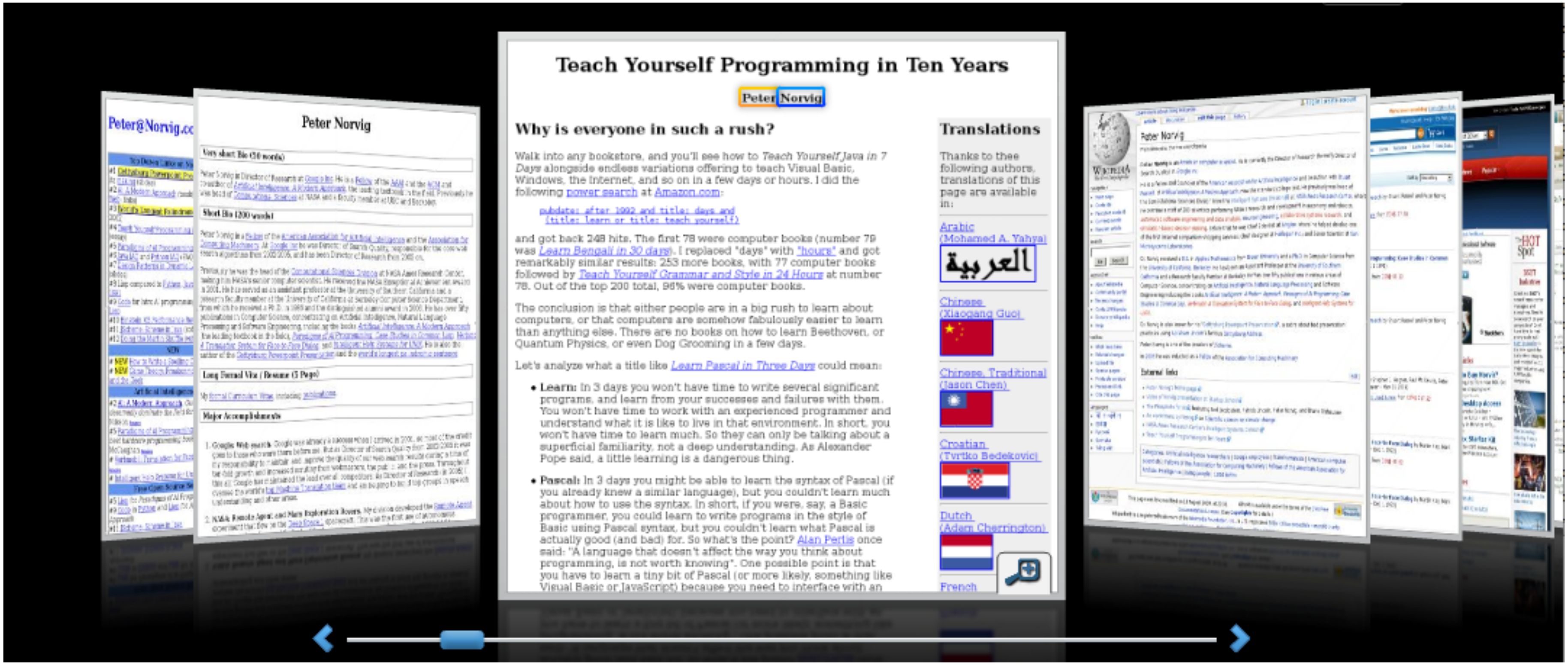
- If we have only a positional index, we cannot (easily) reconstruct context window surrounding "hits"
- Remember **positional index**: for each term in the vocabulary, we store postings of the form docID: <position1, position2, ...>
- If we *cache the documents* at index time, can find windows in it, cueing from hits found in the positional index
  - E.g., positional index says "the query is a phrase/term in position 4378" so we go to this position in the cached document and stream out the content
- Most often, cache only a fixed-size prefix of the doc
  - Note: Cached copy can be outdated!

Dynamic summaries

- Producing good dynamic summaries is a tricky optimization problem
  - The real estate for the summary is normally **small and fixed**
  - Want short items, to show as many matches as possible in first page, and perhaps other things like title
  - Want snippets to be long enough to be useful
  - Want linguistically well-formed snippets: users prefer snippets that contain complete phrases
  - Want snippets maximally informative about doc
- But users really like snippets, even if they complicate IR system design

Alternative results presentations?

- An active area of HCI research



Peter Norvig

Very short Bio (50 words)

Short Bio (200 words)

Long Formal Vita / Resume (3 Page)

Major Accomplishments

Teach Yourself Programming in Ten Years

Peter Norvig

Why is everyone in such a rush?

Walk into any bookstore, and you'll see how to *Teach Yourself Java in 7 Days* alongside endless variations offering to teach Visual Basic, Windows, the Internet, and so on in a few days or hours. I did the following power search at Amazon.com:

```
pubdate: after 1992 and title: days and
 (title: learn or title: teach yourself)
```

and got back 248 hits. The first 78 were computer books (number 79 was *Learn Bengali in 30 days*). I replaced "days" with "hours" and got remarkably similar results: 253 more books, with 77 computer books followed by *Teach Yourself Grammar and Style in 24 Hours* at number 78. Out of the top 200 total, 96% were computer books.

The conclusion is that either people are in a big rush to learn about computers, or that computers are somehow fabulously easier to learn than anything else. There are no books on how to learn Beethoven, or Quantum Physics, or even Dog Grooming in a few days.

Let's analyze what a title like *Learn Pascal in Three Days* could mean:

- **Learn:** In 3 days you won't have time to write several significant programs, and learn from your successes and failures with them. You won't have time to work with an experienced programmer and understand what it is like to live in that environment. In short, you won't have time to learn much. So they can only be talking about a superficial familiarity, not a deep understanding. As Alexander Pope said, a little learning is a dangerous thing.
- **Pascal:** In 3 days you might be able to learn the syntax of Pascal (if you already knew a similar language), but you couldn't learn much about how to use the syntax. In short, if you were, say, a Basic programmer, you could learn to write programs in the style of Basic using Pascal syntax, but you couldn't learn what Pascal is actually good (and bad) for. So what's the point? Alan Perlis once said: "A language that doesn't affect the way you think about programming, is not worth knowing". One possible point is that you have to learn a tiny bit of Pascal (or more likely, something like Visual Basic or JavaScript) because you need to interface with an

Translations

Thanks to thee following authors, translations of this page are available in:

Arabic (Mohamed A. Yahya)

العربية

Chinese (Xiaogang Guo)

Chinese, Traditional (Jason Chen)

Croatian (Tvrtko Bedekovic)

Dutch (Adam Cherrington)

French

Evaluating the quality of results

1. Unranked evaluation (a.k.o. evaluation that does not take into account the rank of a document)
3. Ranked evaluation
4. Evaluation benchmarks
5. Result summaries

Measures for a search engine

- How fast does it index
  - e.g., number of bytes per hour
- How fast does it search
  - e.g., latency as a function of queries per second
- What is the cost per query?
  - In currency (dollars/euros)

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
  - Speed of response
  - Size of index
  - User Interface
  - Most important: relevance
- (Note that none of these is sufficient: blindingly fast, but **useless answers won't make a user happy**.
- How can we quantify user happiness?

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Success: Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: advertiser. Success: Searcher clicks on ad. Measure: clickthrough rate
- Ecommerce: buyer. Success: Buyer buys something. Measures: time to purchase, fraction of "conversions" of searchers to buyers
- Ecommerce: seller. Success: Seller sells something. Measure: profit per item sold
- Enterprise: CEO (chief executive). Success: Employees are more productive (because of effective search). Measure: profit of the company

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries
  - An assessment of the relevance of each query-document pair

Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
- "Relevance to the query" is very problematic.
- Information need *i* : "I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine."
- **This is an information need, not a query.** (Remember: state of the art search engines are now focusing on information needs!)
- With "standard" keyword approach: *q*= [red wine white wine heart attack]
- Consider document *d*′: *At heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving*.
- *d*′ is an excellent match for query *q* . . .
- *d*′ is not relevant to the information need *i* .

Relevance: query vs. information need

- User happiness can only be measured by **relevance to an information need**, not by relevance to queries.
- Difficult to evaluate, since whatever processing we apply on the user’s input, we always come out with some query
- E.g., for the query: “*pizza hut calories per slice*” we assume that the the user is on a diet and the most relevant web site is a calory counter (even if “slice” is absent)
- “Big players” use query logs, click trough, etc. (ex-post evaluation)
- Alternatively, **human judgement** by a team of users on a pre-defined set of “relevant” queries

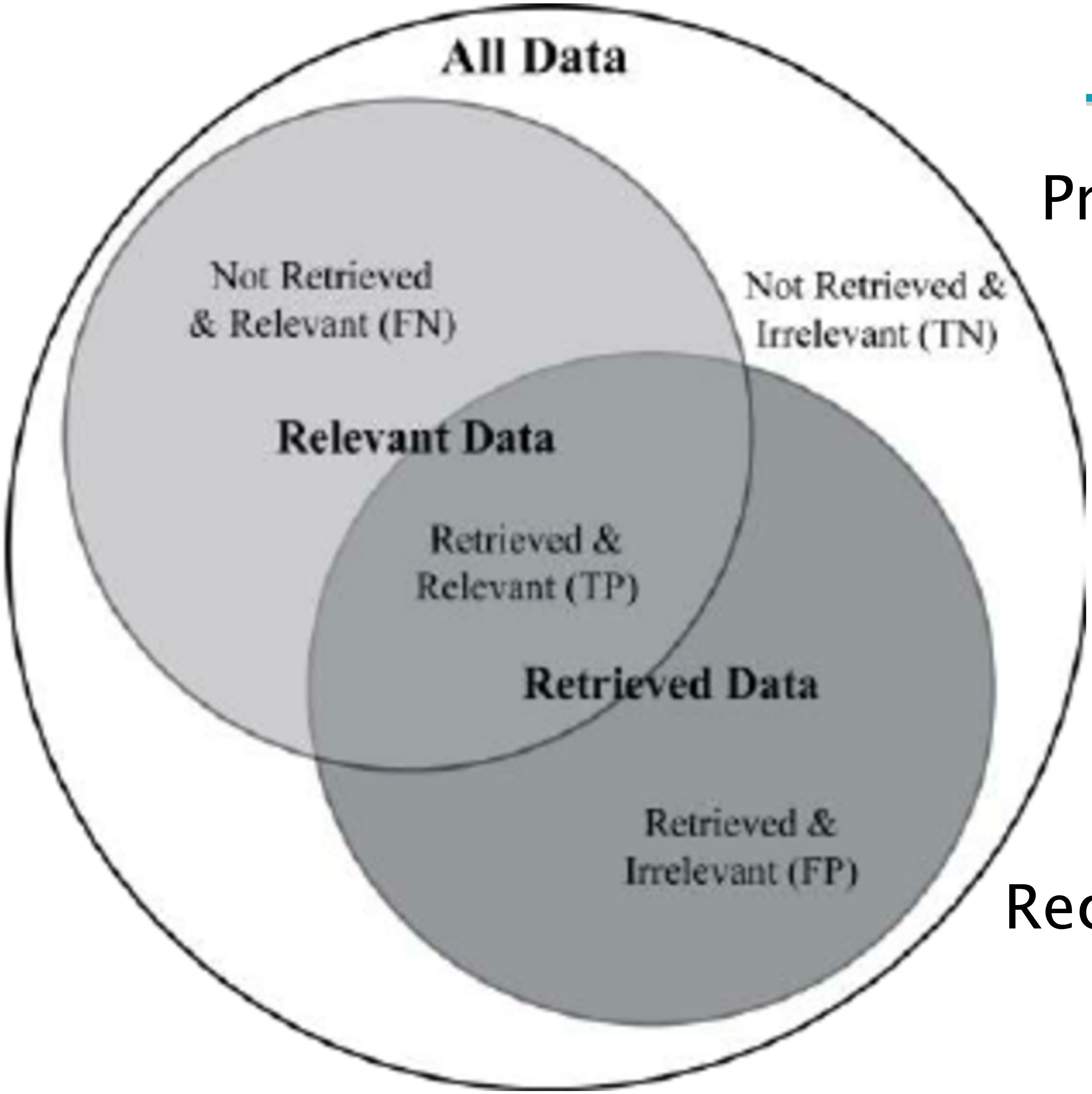
Standard measures of relevance: Precision and Recall

- Precision (*P*) is the fraction of retrieved documents that are relevant

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- Recall (*R*) is the fraction of relevant documents that are retrieved

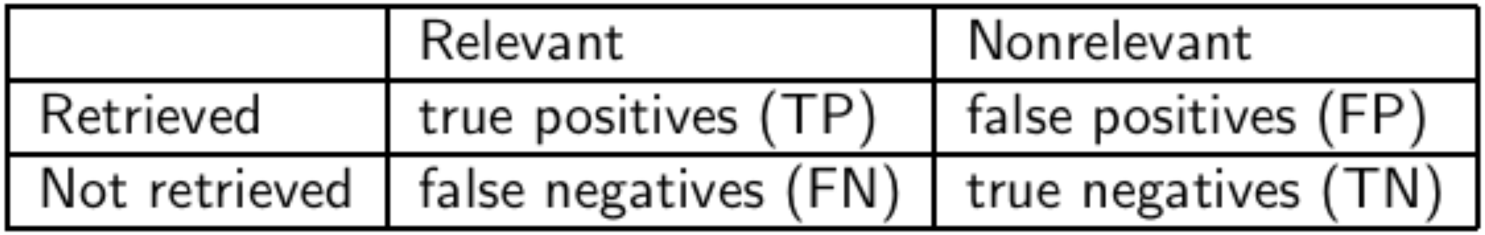
$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$


All Data
Not Retrieved & Relevant (FN)
Not Retrieved & Irrelevant (TN)
Relevant Data
Retrieved & Relevant (TP)
Retrieved Data
Retrieved & Irrelevant (FP)

$$\text{Precision}=\frac{retrieved\ relevant}{retrieved\ data}$$

$$\text{Recall}=\frac{retrieved\ relevant}{Not\ retrieved\ \&\ relevant}$$

Coinfusion Matrix



	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$P = TP / (TP + FP)$$

$$R = TP / (TP + FN)$$

Precision/recall tradeoff

- We can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs in the collection has 100% recall!
- The converse is also true (usually): It’s easy to get high precision for very low recall.

A combined measure: *F*

- *F-measure* allows us to trade off precision against recall.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

- $\alpha \in [0, 1]$ and thus $\beta^2 \in [0,\infty]$
- Most frequently used: balanced *F* with $\beta = 1$ or $\alpha = 0.5$
  - This is the harmonic mean of *P* and *R*: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$

F: Example (confusion matrix)

	relevant	**not relevant**	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

- P = 20/(20 + 40) = 1/3
- R = 20/(20 + 60) = 1/4
- $$F_1 = 2\frac{1}{\frac{1}{\frac{1}{3}}+\frac{1}{\frac{1}{4}}} = 2/7$$

Accuracy

- Why do we use complex measures like precision, recall, and *F*?
- Why not something simple like accuracy?
- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table above,

  accuracy = $(TP + TN)/(TP + FP + FN + TN)$.

TP= true positive (relevant documents that the system judged relevant); TN= true negative FP=false positive FN=false negative

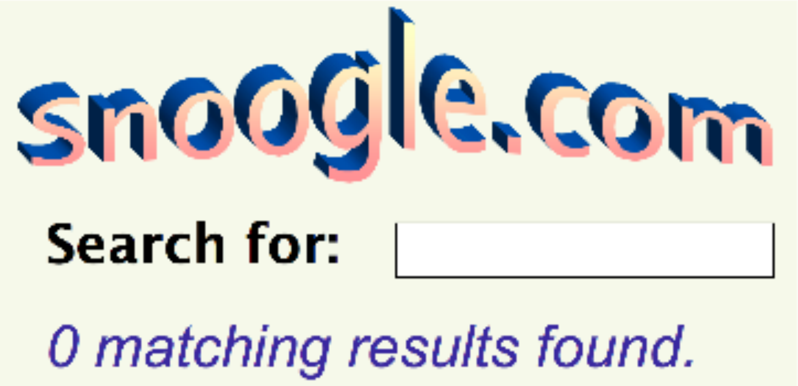
- Why is accuracy not a useful measure for web information retrieval?

Example

- Compute precision, recall and F_1 for this result set:

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

- The snoogle search engine below always returns 0 results (“0 matching results found”), regardless of the query. Why does snoogle demonstrate that accuracy is not a useful measure in IR?



0 matching results found.

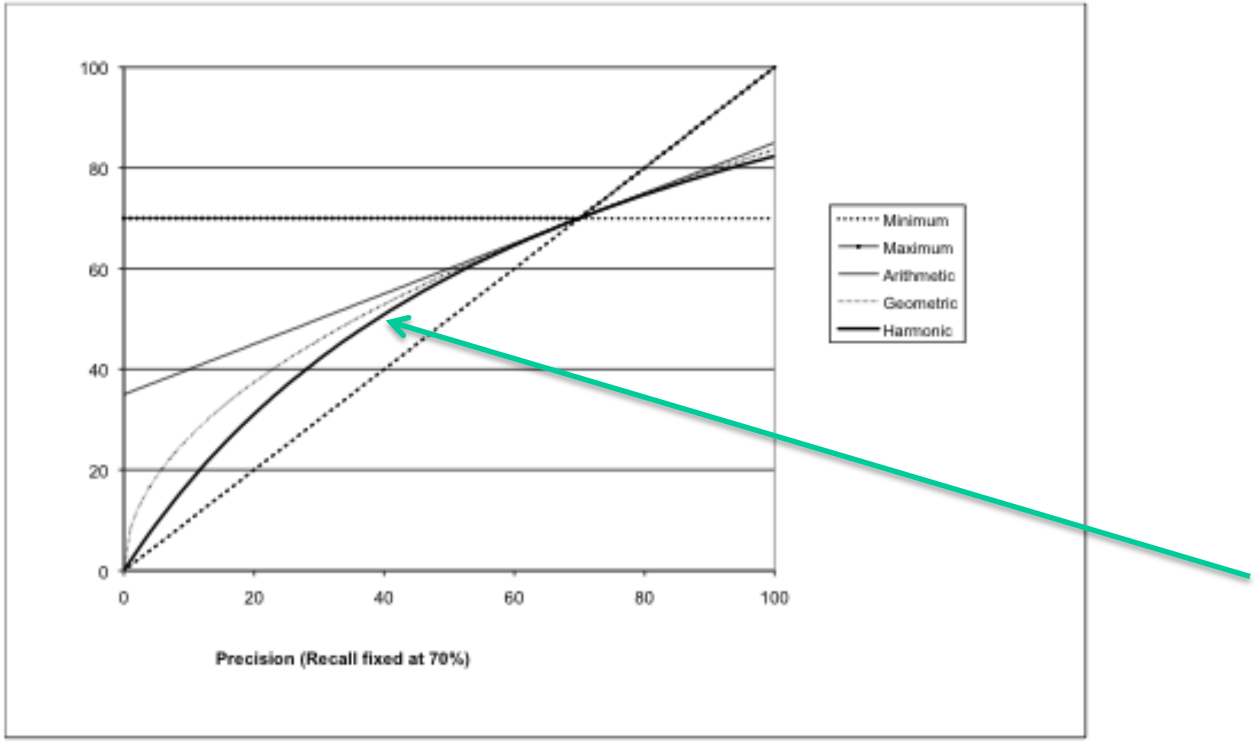
Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say “nothing found” and return nothing
- You then get 99.99% accuracy on most queries.
- Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk.
- It’s better to return some bad hits as long as you return something.
- →We use precision, recall, and *F* for evaluation, **not accuracy**.

F: Why harmonic mean?

- Why don’t we use a different mean of *P* and *R* as a measure?
  - e.g., the arithmetic mean
- The simple (arithmetic) mean is 50% for “return-everything” search engine, which is still too high.
- Desideratum: Punish really bad performance on either precision or recall.
- *F* (harmonic mean) is a kind of *smooth minimum*.

F_1 and other averages



- We can view the harmonic mean as a kind of soft minimum

Difficulties in using precision, recall and *F*

- We need relevance judgments for information-need-document pairs – but **they are expensive to produce**. Mostly not available, **need to hire experts** to tag documents as relevant/not relevant

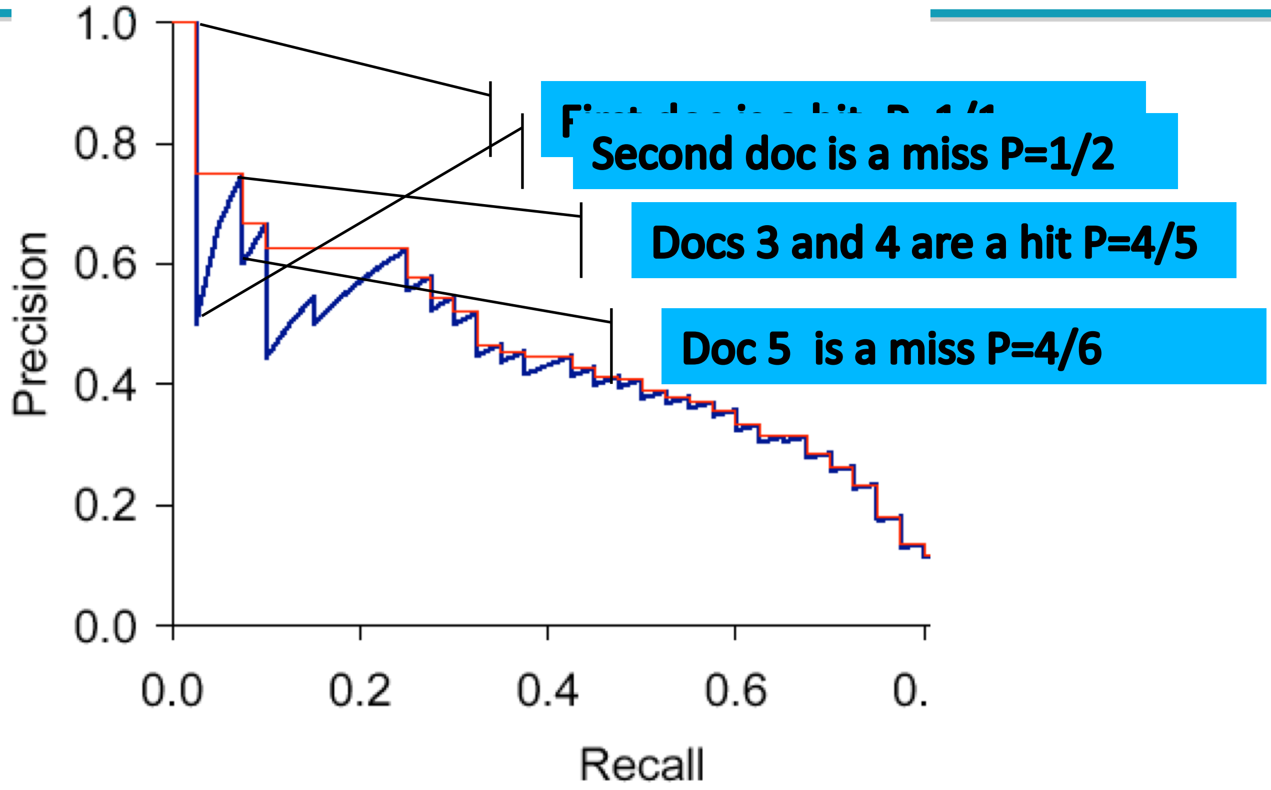
Outline

1. Unranked evaluation
2. Ranked evaluation (what if we consider ranking?)
3. Evaluation benchmarks
4. Result summaries

Precision-recall curve

- Precision, recall, and the F measure are set-based measures. They are computed using **unordered sets of documents**.
- Count of P, R, F performed on the full set of retrieved docs, regardless of their ranking
- We need to extend these measures (or to define new measures) if we are to evaluate the **ranked retrieval** results.
- We can easily turn set measures into measures of ranked lists.
- Just compute the set measure for each "prefix": the top 1, top 2, top 3, top 4 etc results (i.e. : precision /recall/F for top 1 result, for top 3 results, for top k results)
- Doing this for precision and recall gives you a precision-recall curve.

Jig-saw precision/recall curve



First doc is a hit P=1/1

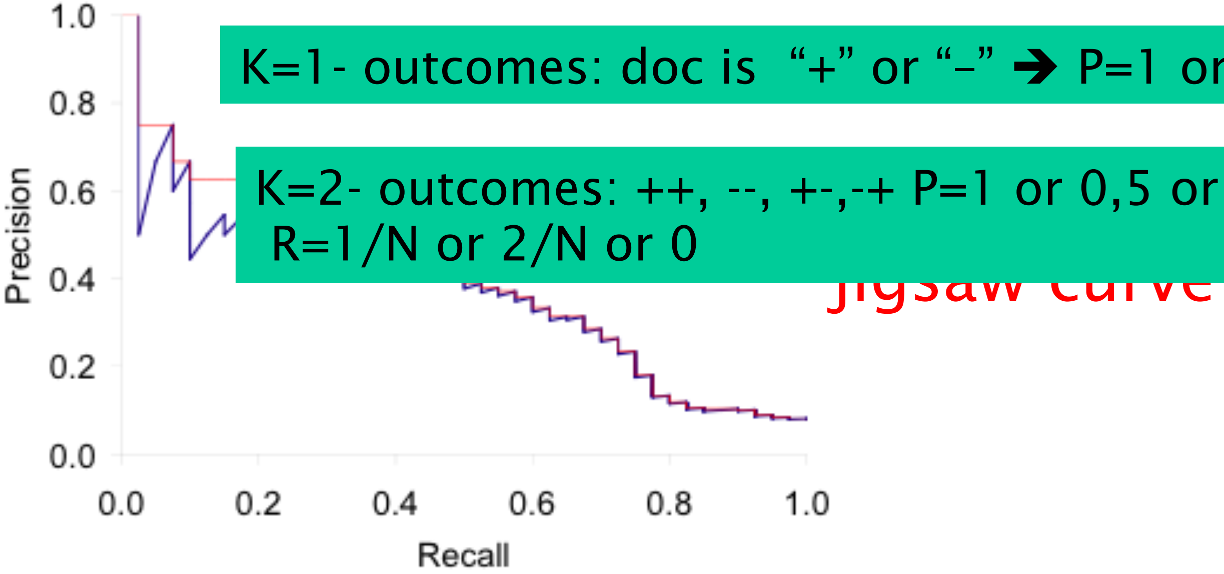
Second doc is a miss P=1/2

Docs 3 and 4 are a hit P=4/5

Doc 5 is a miss P=4/6

Precision-recall curves have a distinctive saw-tooth shape:
if the (k+1th) document retrieved is nonrelevant then recall is the same as for the top k documents, but precision has dropped.
If it is relevant, then both precision and recall increase,
and the curve jags up and to the right.

Jig-saw precision-recall curve



K=1- outcomes: doc is “+” or “-” ➔ P=1 or 0, R=1/N or 0

K=2- outcomes: ++, --, +-,-+ P=1 or 0,5 or 0, R=1/N or 2/N or 0

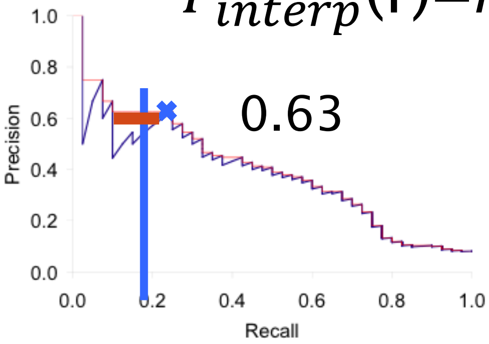
- Each point corresponds to a result for the top k ranked hits (*k* = 1, 2, 3, 4, . . .).
- Interpolation (in red)

11-point interpolated average precision

Recall	Interpolated Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

It is often useful to remove these jiggles and the standard way to do this is with an interpolated precision: the *interpolated precision* at a certain recall level r is defined as the **highest precision** found for any recall level r’>r

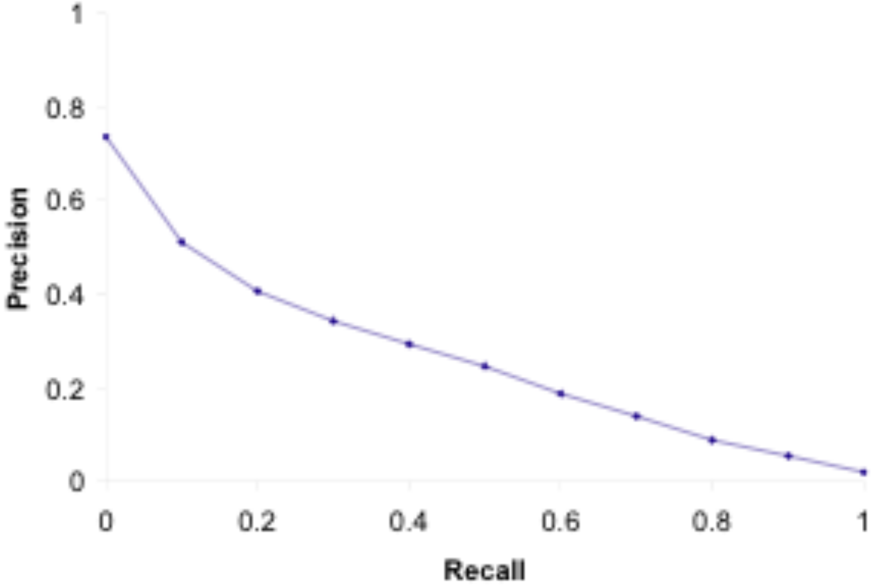
$$P_{interp}(\mathrm{r}) = max_{\geq r} P(r')$$



0.63

11-point average: ≈ 0.425

Averaged 11-point precision/recall graph



- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, . . .
- Do this for each of the queries in the evaluation benchmark
- Average over queries
- This measure measures performance at all recall levels.
- Note that (in this example) performance is not very good!

Variance of measures like precision/recall

- For a test collection, it is usual that a system does badly on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and really well on others (e.g., $P = 0.95$ at $R = 0.1$).
- Indeed, it is usually the case that the variance of the same system across queries is much greater than the variance of different systems on the same query.
- That is, there are **easy information needs and hard ones**.

Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The marginal relevance of a document presented **at position *k* in the result list** of retrieved documents, is the additional information it contributes over and above the information that was contained in previous documents $d_1 \ldots d_{k-1}$.

Example

System A: Ranked pages (C=correct W=Wrong)

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	....	Qn
A1	W	W	C	W	C	W	W	W	....	W
A2	W	W	W	W	W	W	W	W	....	W
A3	W	W	W	W	W	W	W	W	....	W
A4	W	W	W	W	W	W	W	W	....	W
A5	W	C	W	W	W	C	W	W	....	W

System B: Ranked pages

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	....	Qn
A1	W	W	W	W	C	W	W	W	....	W
A2	C	W	C	W	W	C	C	W	....	C
A3	W	C	W	W	W	W	W	W	....	W
A4	W	W	W	C	W	W	W	W	....	W
A5	W	W	W	W	W	W	W	W	....	W

Mean Reciprocal Rank (MRR)

- Score for an individual query:
  - The reciprocal of the rank at which the first correct answer is returned
  - 0 if no correct response is returned
- The score for a full run of a set of queries:
  - **Mean over the set of questions in the test**

MRR in action

System A: MRR = (.2+1+1+.2)/10 = 0.24

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	....	Qn
A1	W	W	C	W	C	W	W	W	....	W
A2	W	W	W	W	W	W	W	W	....	W
A3	W	W	W	W	W	W	W	W	....	W
A4	W	W	W	W	W	W	W	W	....	W
A5	W	C	W	W	W	C	W	W	....	W

System B: MRR = (.5+.33+.5+.25+1+.5+.5+.5)/10=0.42

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	....	Qn
A1	W	W	W	W	C	W	W	W	....	W
A2	C	W	C	W	W	C	C	W	....	C
A3	W	C	W	W	W	W	W	W	....	W
A4	W	W	W	C	W	W	W	W	....	W
A5	W	W	W	W	W	W	W	W	....	W

Outline

1. Unranked evaluation
2. Ranked evaluation
3. Evaluation benchmarks
4. Result summaries

What we need for a benchmark

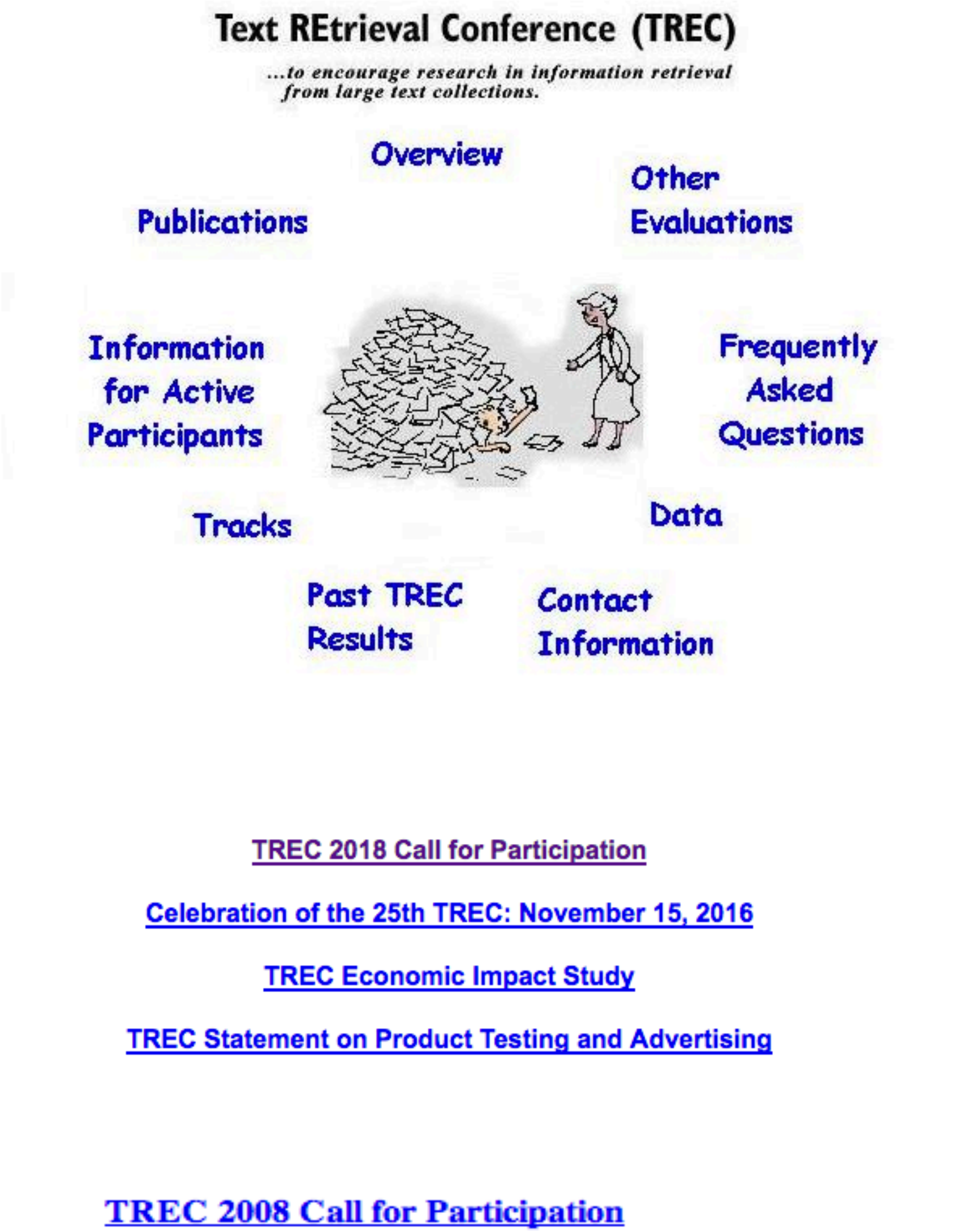
- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
  - . . .which we will often refer to as queries, even though we already said that there is indeed a difference. In these evaluations, “test queries” are actually QUESTIONS in natural language.
  - Information needs must be representative of the information needs we expect to see in reality (e.g. analyze user search behavior in a real setting) Ex: frequent search query types in TripAdvisor
- Human relevance assessments
  - We need to hire/pay “judges” or assessors to do this.
  - Expensive, time-consuming
  - Judges must be representative of the users we expect to see in reality.

Standard relevance benchmark: Cranfield

- Pioneering: first test-bed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK
- 1398 abstracts of aerodynamics journal articles, a set of 225 queries, exhaustive relevance judgments of all query-document-pairs
- Too small, too untypical for serious IR evaluation today

Standard relevance benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST)
- TREC is actually a set of several different relevance benchmarks.
- No exhaustive relevance judgments – too expensive
- Rather, NIST assessors’ relevance judgments **are available only for the documents that were among the top k returned** for some system which was entered in the TREC evaluation for which the information need was developed.



Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

Overview

Publications

Other Evaluations

Information for Active Participants

Frequently Asked Questions

Tracks

Data

Past TREC Results

Contact Information

TREC 2018 Call for Participation

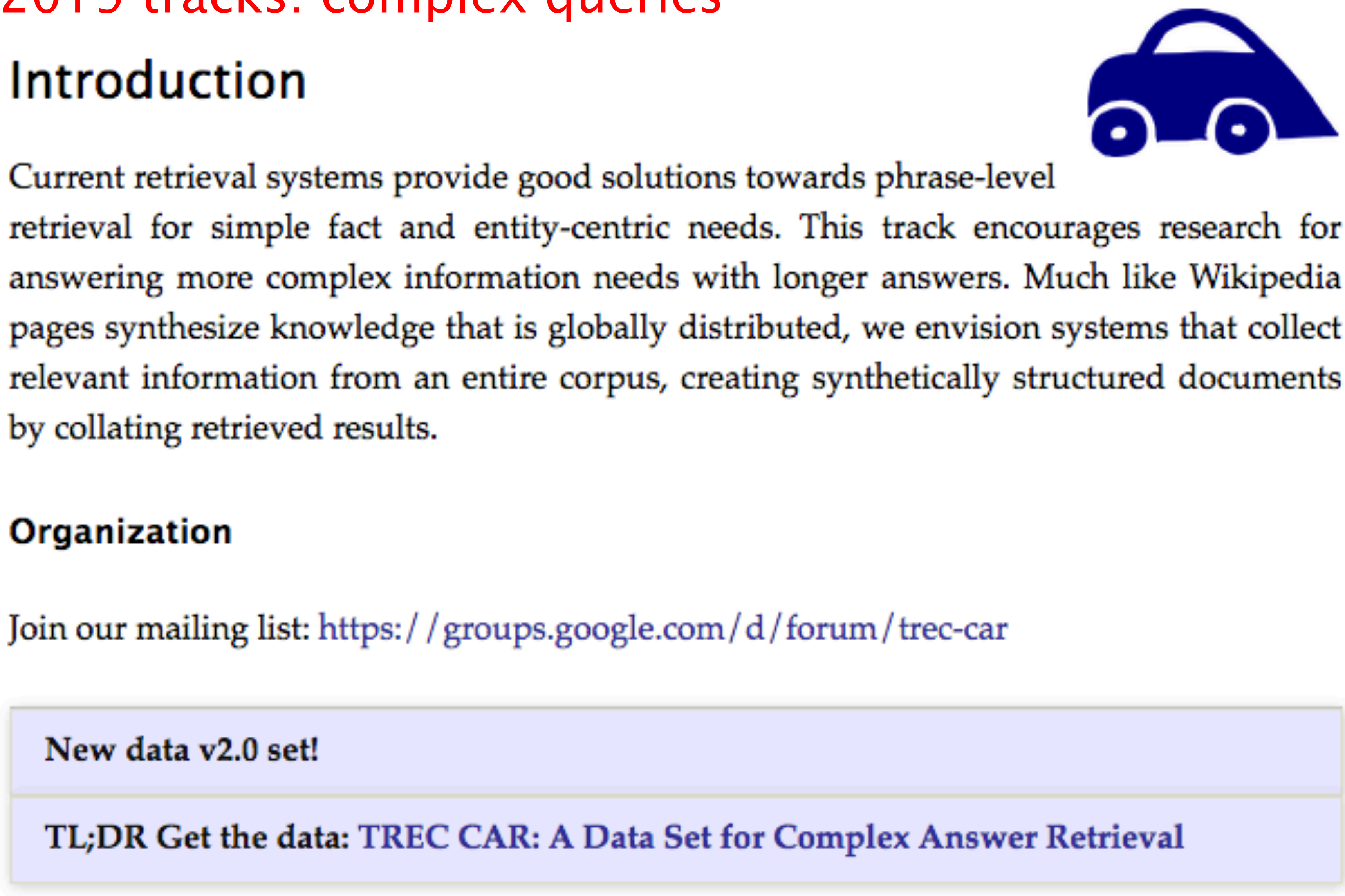
Celebration of the 25th TREC: November 15, 2016

TREC Economic Impact Study

TREC Statement on Product Testing and Advertising

TREC 2008 Call for Participation

2019 tracks: complex queries



Introduction

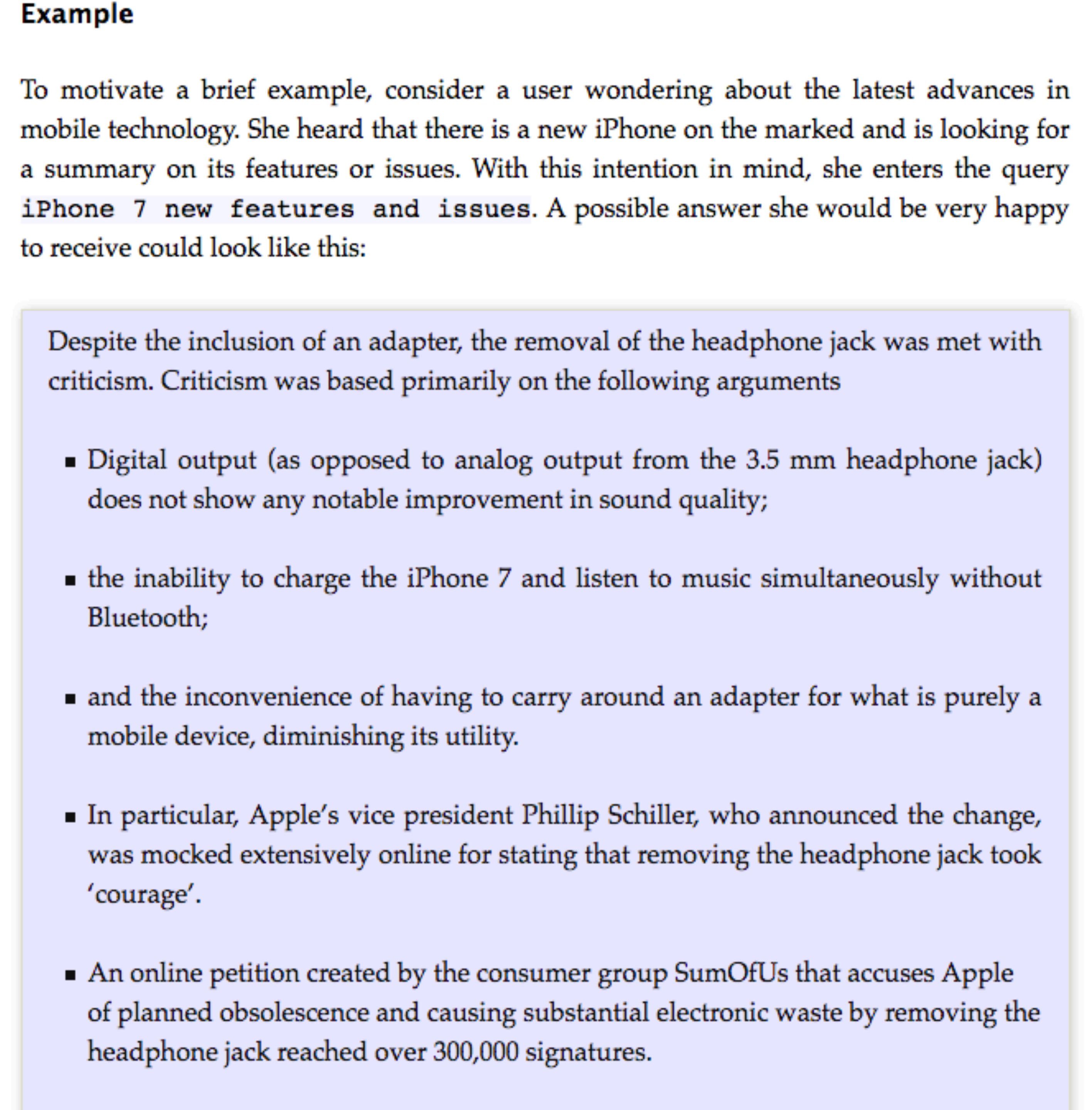
Current retrieval systems provide good solutions towards phrase-level retrieval for simple fact and entity-centric needs. This track encourages research for answering more complex information needs with longer answers. Much like Wikipedia pages synthesize knowledge that is globally distributed, we envision systems that collect relevant information from an entire corpus, creating synthetically structured documents by collating retrieved results.

Organization

Join our mailing list: https://groups.google.com/d/forum/trec-car

New data v2.0 set!

TL;DR Get the data: TREC CAR: A Data Set for Complex Answer Retrieval

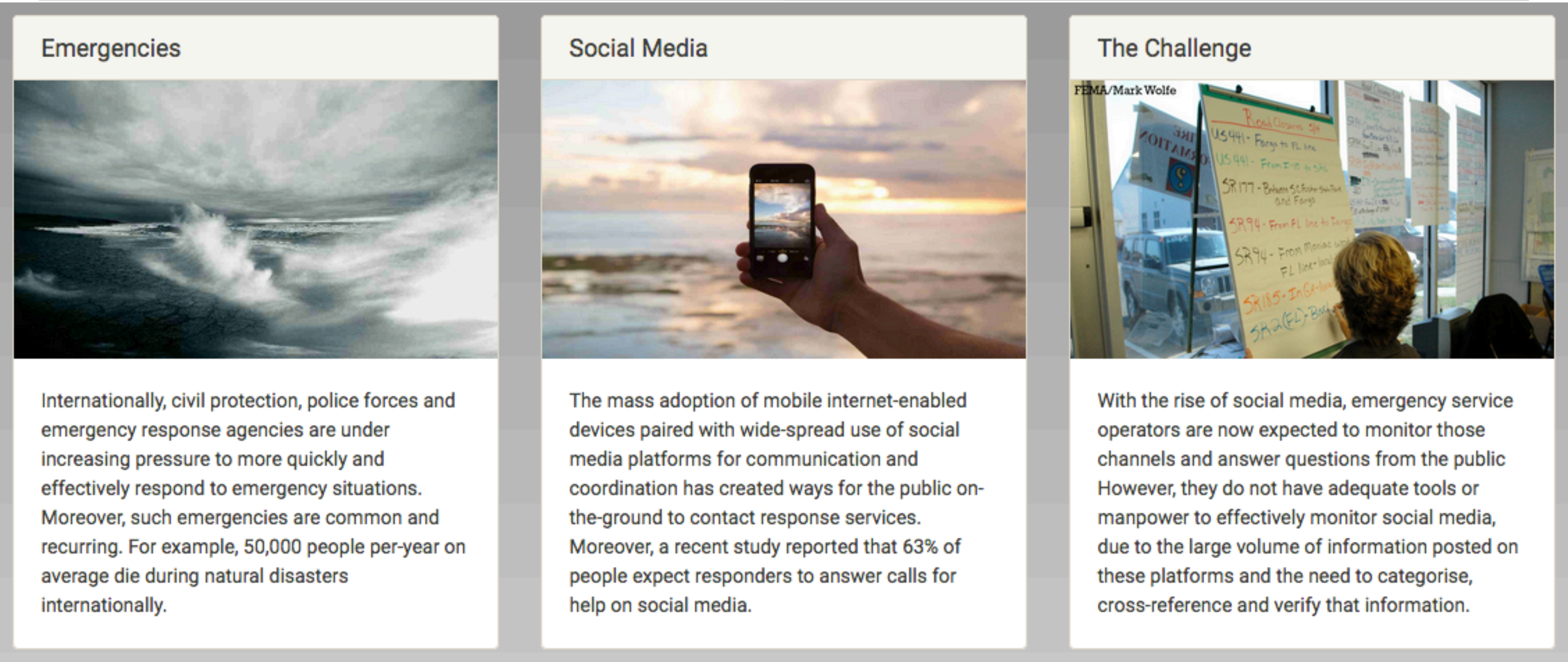


Example

To motivate a brief example, consider a user wondering about the latest advances in mobile technology. She heard that there is a new iPhone on the marked and is looking for a summary on its features or issues. With this intention in mind, she enters the query `iPhone 7 new features and issues`. A possible answer she would be very happy to receive could look like this:

> Despite the inclusion of an adapter, the removal of the headphone jack was met with criticism. Criticism was based primarily on the following arguments
>
> - Digital output (as opposed to analog output from the 3.5 mm headphone jack) does not show any notable improvement in sound quality;
> - the inability to charge the iPhone 7 and listen to music simultaneously without Bluetooth;
> - and the inconvenience of having to carry around an adapter for what is purely a mobile device, diminishing its utility.
> - In particular, Apple's vice president Phillip Schiller, who announced the change, was mocked extensively online for stating that removing the headphone jack took 'courage'.
> - An online petition created by the consumer group SumOfUs that accuses Apple of planned obsolescence and causing substantial electronic waste by removing the headphone jack reached over 300,000 signatures.

Trec 2019: Incident Streams



Emergencies

Internationally, civil protection, police forces and emergency response agencies are under increasing pressure to more quickly and effectively respond to emergency situations. Moreover, such emergencies are common and recurring. For example, 50,000 people per-year on average die during natural disasters internationally.

Social Media

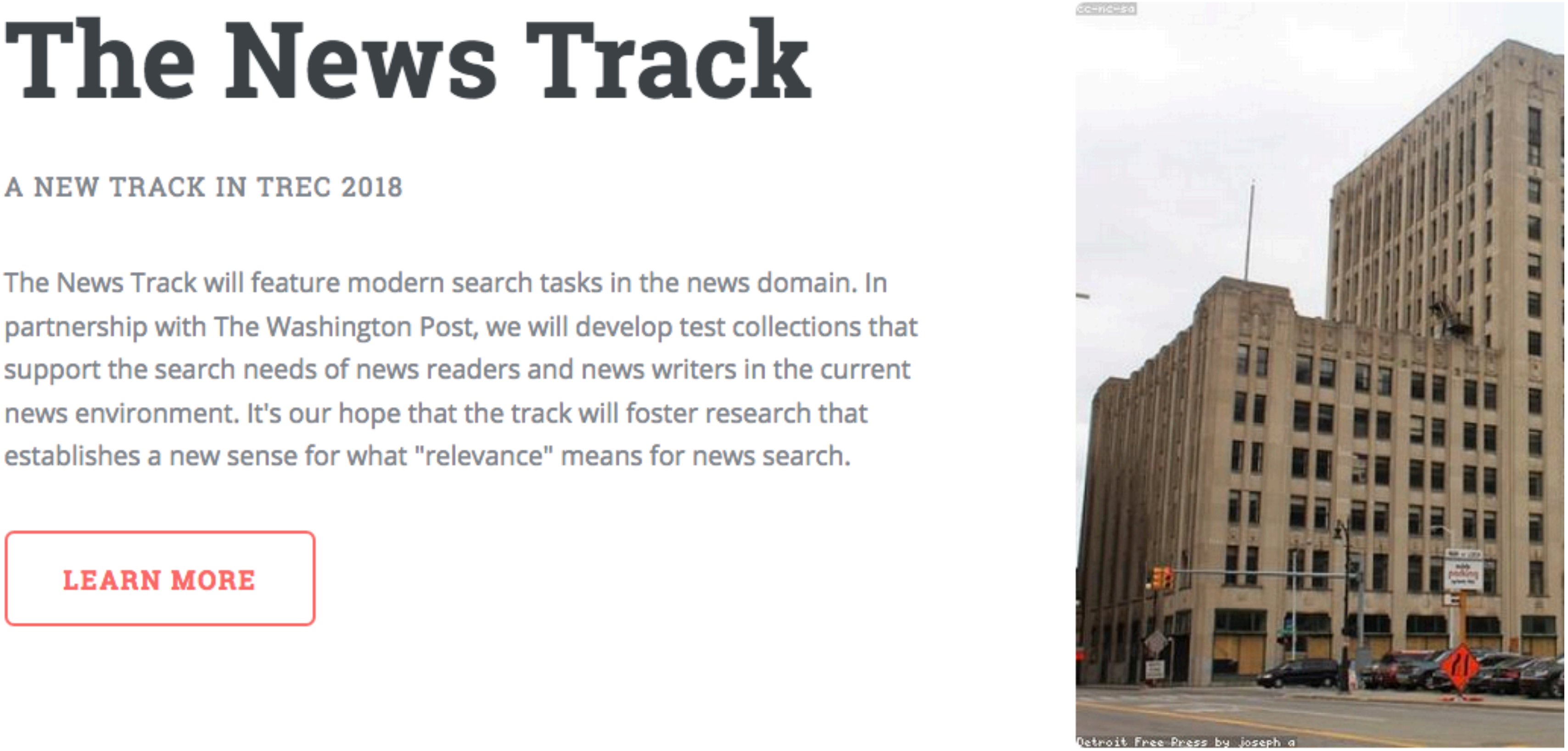
The mass adoption of mobile internet-enabled devices paired with wide-spread use of social media platforms for communication and coordination has created ways for the public on-the-ground to contact response services. Moreover, a recent study reported that 63% of people expect responders to answer calls for help on social media.

The Challenge

With the rise of social media, emergency service operators are now expected to monitor those channels and answer questions from the public However, they do not have adequate tools or manpower to effectively monitor social media, due to the large volume of information posted on these platforms and the need to categorise, cross-reference and verify that information.

- Monitoring social media to answer questions/help requests based on real-time emergence.

Trec 2019: news track

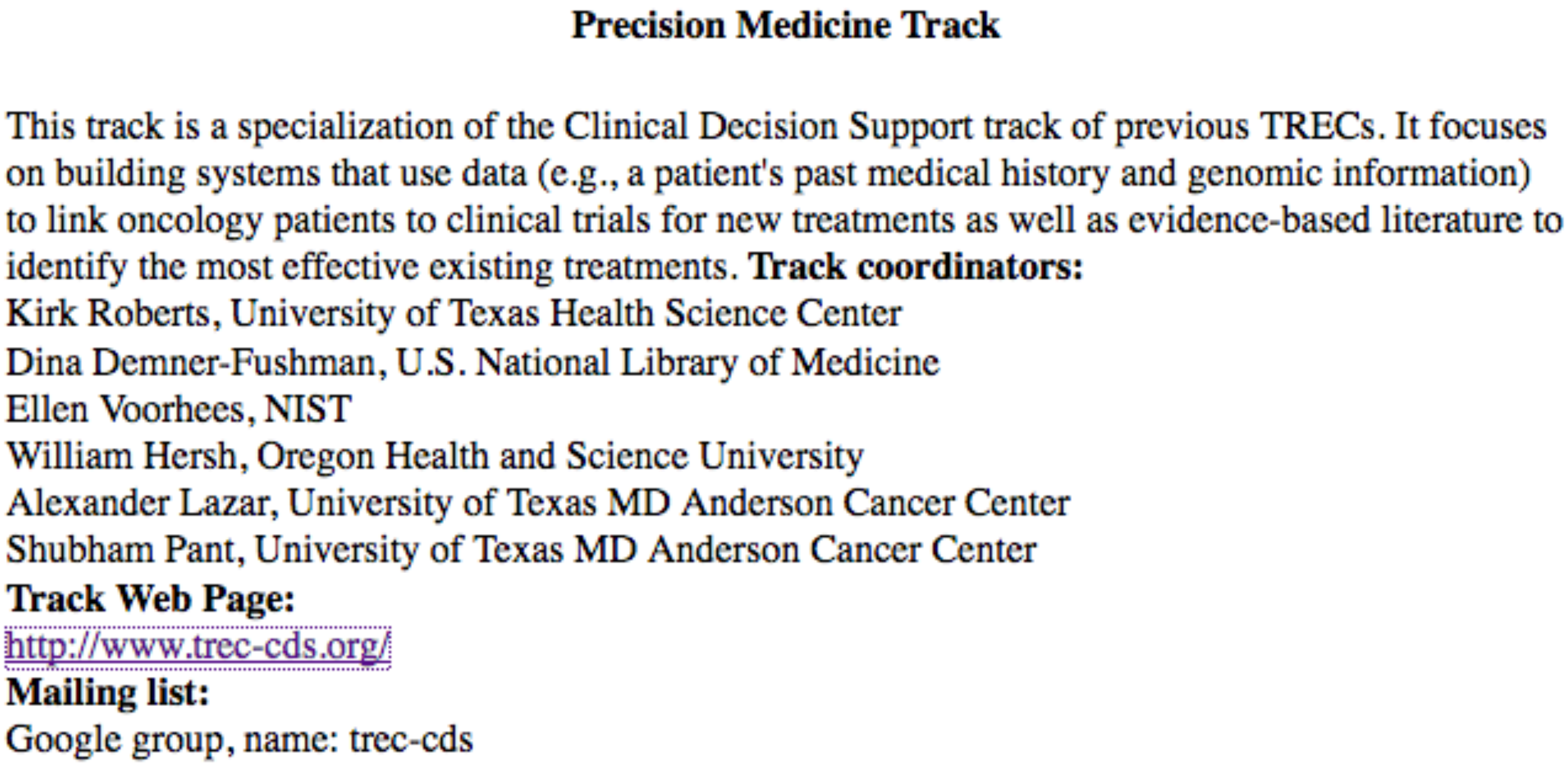


The News Track

A NEW TRACK IN TREC 2018

The News Track will feature modern search tasks in the news domain. In partnership with The Washington Post, we will develop test collections that support the search needs of news readers and news writers in the current news environment. It's our hope that the track will foster research that establishes a new sense for what "relevance" means for news search.

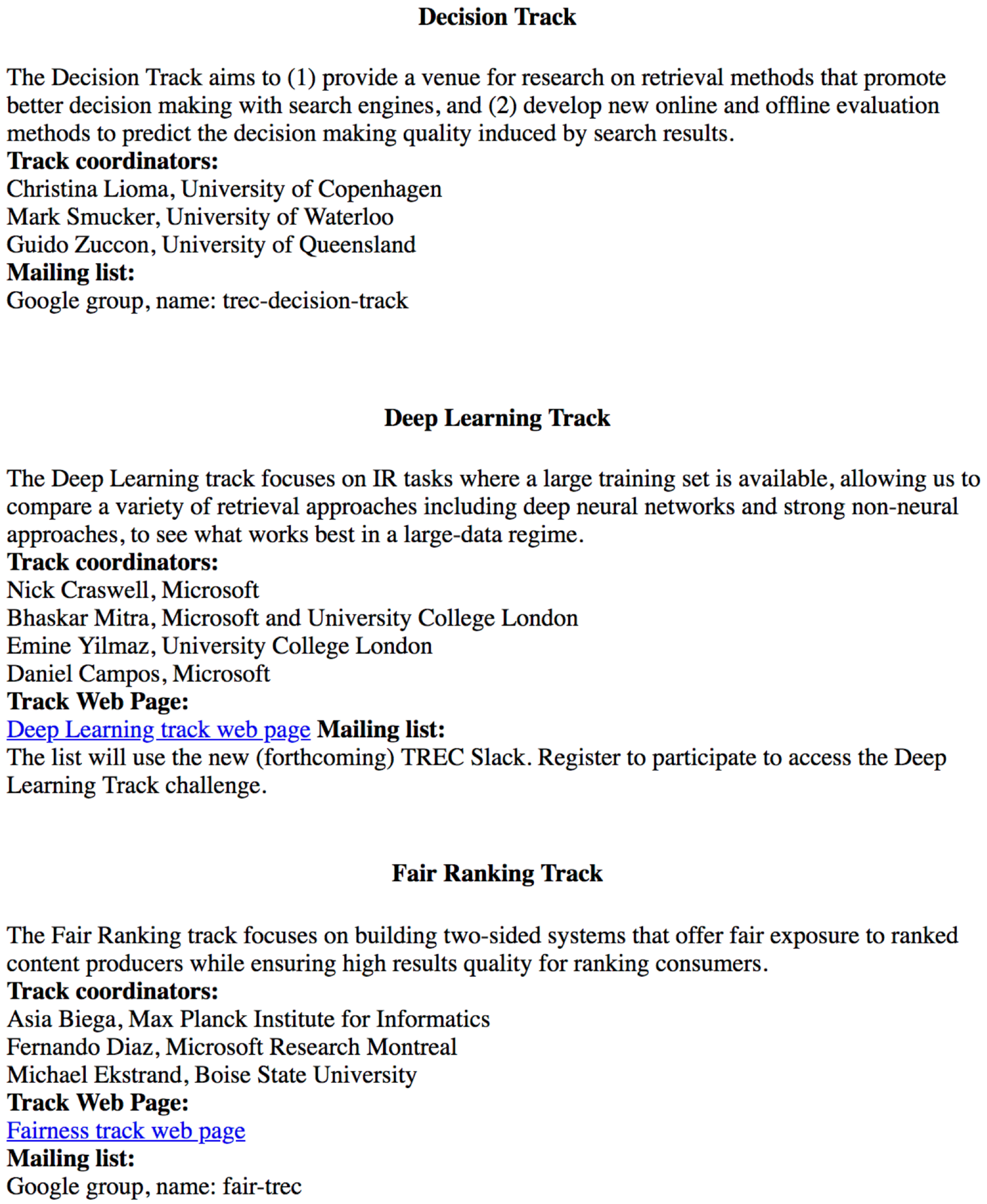
TREC 2017-18-19 Precision Medicine



Precision Medicine Track

This track is a specialization of the Clinical Decision Support track of previous TRECs. It focuses on 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 as well as evidence-based literature to identify the most effective existing treatments. **Track coordinators:**
Kirk Roberts, University of Texas Health Science Center
Dina Demner-Fushman, U.S. National Library of Medicine
Ellen Voorhees, NIST
William Hersh, Oregon Health and Science University
Alexander Lazar, University of Texas MD Anderson Cancer Center
Shubham Pant, University of Texas MD Anderson Cancer Center
Track Web Page:
http://www.trec-cds.org/
Mailing list:
Google group, name: trec-cds

Other 2019 TREC tracks



Decision Track

The Decision Track aims to (1) provide a venue for research on retrieval methods that promote better decision making with search engines, and (2) develop new online and offline evaluation methods to predict the decision making quality induced by search results.
Track coordinators:
Christina Lioma, University of Copenhagen
Mark Smucker, University of Waterloo
Guido Zuccon, University of Queensland
Mailing list:
Google group, name: trec-decision-track

Deep Learning Track

The Deep Learning track focuses on IR tasks where a large training set is available, allowing us to compare a variety of retrieval approaches including deep neural networks and strong non-neural approaches, to see what works best in a large-data regime.
Track coordinators:
Nick Craswell, Microsoft
Bhaskar Mitra, Microsoft and University College London
Emine Yilmaz, University College London
Daniel Campos, Microsoft
Track Web Page:
Deep Learning track web page **Mailing list:**
The list will use the new (forthcoming) TREC Slack. Register to participate to access the Deep Learning Track challenge.

Fair Ranking Track

The Fair Ranking track focuses on building two-sided systems that offer fair exposure to ranked content producers while ensuring high results quality for ranking consumers.
Track coordinators:
Asia Biega, Max Planck Institute for Informatics
Fernando Diaz, Microsoft Research Montreal
Michael Ekstrand, Boise State University
Track Web Page:
Fairness track web page
Mailing list:
Google group, name: fair-trec

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Used to be largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR
  - East Asian language and cross-language information retrieval
- Cross Language Evaluation Forum (CLEF)
  - This evaluation series has concentrated on European languages and cross-language information retrieval.
- Many others

Validity of relevance assessments

- Relevance assessments are only usable if they are consistent (e.g. judges agree on results).
- If they are not consistent, then there is no “truth” and experiments are not repeatable.
- How can we measure this consistency or agreement among judges?
- → Fleiss‘ Kappa measure

Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments
- Corrects for „chance“ agreement
- *P*(*A*) = proportion of time judges agree
- *P*(*E*) = what agreement would we obtained by pure chance

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Agreement by chance: probability that random judges take the same random decision

Kappa measure (2)

- Values of *k* in the interval [2/3, 1.0] are seen as acceptable.
- With smaller values: need to redesign relevance assessment methodology used etc.

Calculating the kappa statistic

		Judge 2 Relevance		
		Yes	No	Total
Judge 1 Relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

Observed proportion of the times the judges agreed

$P(A) = (300 + 70)/400 = 370/400 = 0.925$

Pooled marginals

$P(nonrelevant)$ = (80(J1) + 90(J2))/(400 + 400) = 170/800 = 0.2125

$P(relevant)$ = (320(J1) + 310(J2))/(400 + 400) = 630/800 = 0.7878

Probability that the two judges agreed by chance $P(E)$ =

$P(nonrelevant)^2 + P(\mathrm{relevant})^2 = 0.2125^2 + 0.7878^2 = 0.665$

Kappa statistic $\kappa = (P(A) - P(E))/(1 - P(E))$ =

$(0.925 - 0.665)/(1 - 0.665) = 0.776$ (still in acceptable range)

Inter-judge agreement at TREC

Information need	number of docs judged	disagreements
51	211	6
62	400	157
67	400	68
95	400	110
127	400	106

Evaluation in large search engines: clicktrough

- Recall is difficult to measure on the web
- Search engines often use precision at top *k*, e.g., *k* = 10 . . .
- . . . or use measures that reward you more for getting rank 1 right than for getting rank 10 right (MRR).
- Search engines also use **non-relevance-based measures**.
  - Example 1: *clickthrough* on first result
  - Not very reliable if you look at a single clickthrough (you may realize after clicking that the summary was misleading and the document is nonrelevant) . . .
  - . . . but **pretty reliable** in the aggregate.

Evaluation in large search engines: A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- **Probably the evaluation methodology that large search engines trust most (e.g. Google)**