Test Collections

Temporal Information Retrieval

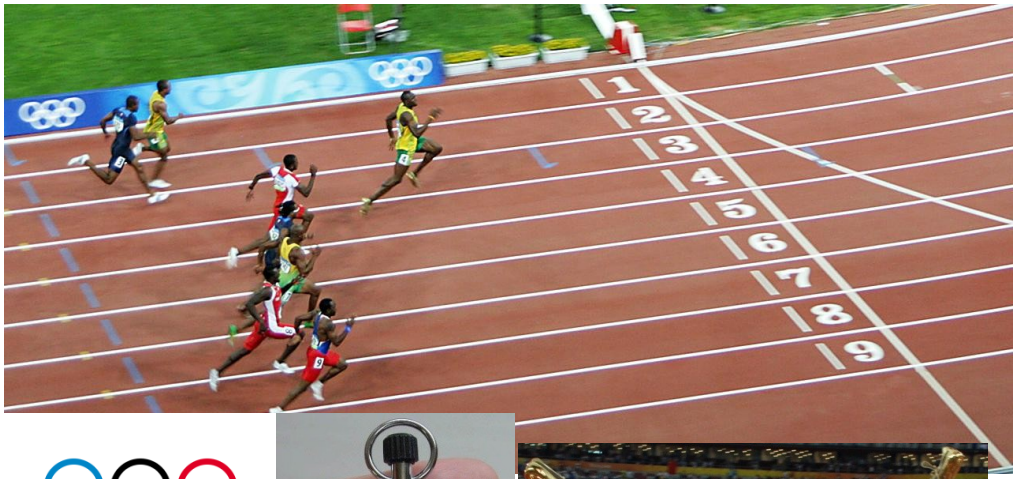
Jaspreet Singh

Outline

- Why do we need evaluation?
- What is a test collection?
  - Defining the task
  - Defining metrics
  - Choosing the right document collection
  - Defining topics and gathering judgements

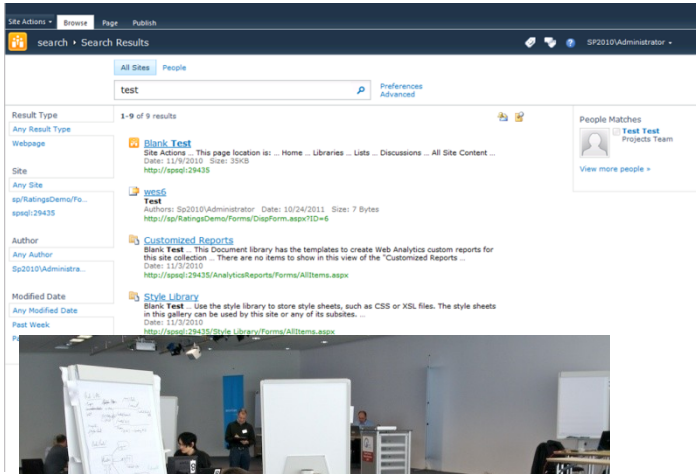
Test collections for Temporal IR

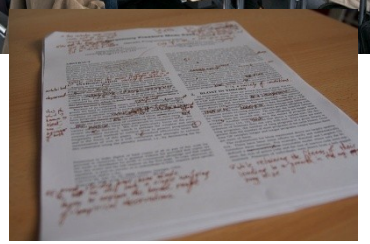
Why do we need evaluation?



NIST National Institute of Standards and Technology U.S. Department of Commerce

TREC





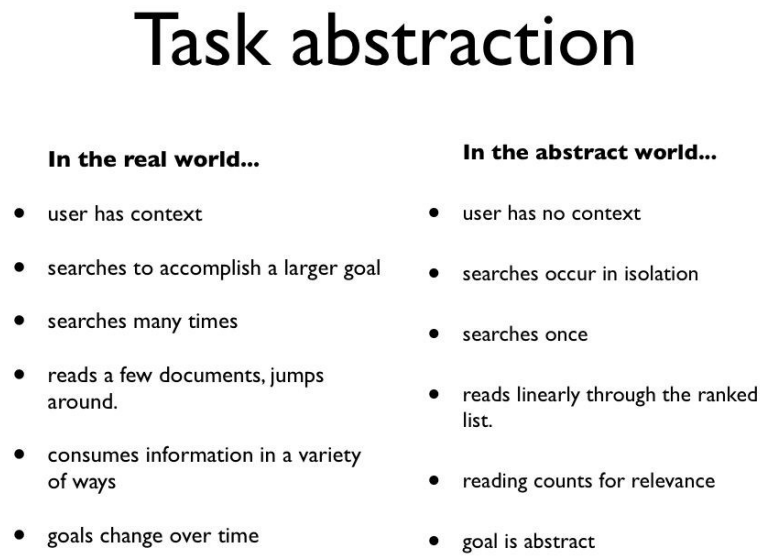
What is a test collection

- Documents + topics + judgements = test collection
- Example of a test collection:
  - TREC diversity web track test collection
    - Web crawl
    - Input queries
    - User relevance judgements

How did they build this collection?

1. Define the task

- Think of a real world scenario and abstract it.



- Operationalize the task – Break it down into simple steps
- Drives the rest of the test collection

The 100m sprint

- Intention and task? Who is the worlds‘ fastest man?
- Measure? Time taken
- Documents? The standard 100m track
- Competitors? The contestants
- Judgements? At the olympics using a watch and then entered in the record books

In Information Retrieval ...

- TREC Diversity Track
  - „ …a diversity task whose goal was to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list“

Which algorithm produces the best ranking?

2. Determine the metric

- How do you measure the performance of a retrieval model?
- Precision– What fraction of retrieved results for a query are relevant?

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

- Recall – What fraction of the relevant results for a query are returned?

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

- set-based measures

Precision – Recall curve

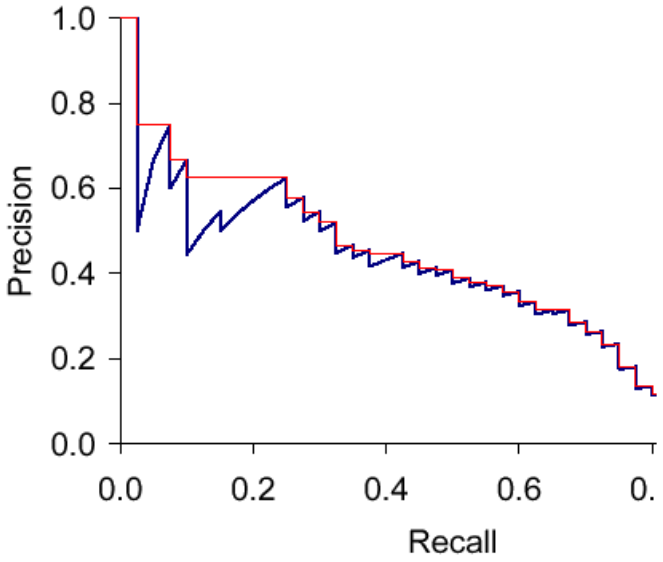


Figure 8.2: Precision/recall graph.

- "The justification is that almost anyone would be prepared to look at a few more documents if it would increase the percentage of the viewed set that were relevant (that is, if the precision of the larger set is higher)"

User centric metrics

- nDCG – Normalised Discounted Cumulative Gain

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i \qquad \mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)} \qquad \mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

$$G[k] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}}.$$

Alpha – nDCG (takes diversity and novelty into account)

Captures the users seeking behavior

- ERR – estimated reciprocal rank
- http://plg.uwaterloo.ca/~gvcormac/novelty.pdf

Diversity track measures

- Intent Aware metrics for diversity
  - IA-Precision @ k
  - Subtopic Recall @ k
  - MAP (Mean Avg. Precision)

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \mathrm{Precision}(R_{jk})$$

http://olivier.chapelle.cc/pub/diversity-inrt.pdf

3. Document Collection

- Tasks sometimes imply the collection.

- Tweet filtering? – Twitter dataset
- Diversity? – which dataset?

- The documents affect how the systems will search. Changing the documents will change the performance of systems.
- Collections can opportunistic (email conversations), constructed (tweets for a hashtag) or naturalistic (large web crawl)

4. Find Topics

- Try to put yourself in the user‘s shoes
- Find the intent and not just the query

- How do „you“ build these topics?
  - Manual
  - Query log driven
  - Explore the collection
  - Observe real users

  Topics are hard to find! More topics you have the lesser the variability in your results.

```
{
  "topic": 1,
  "type": "person",
  "query": "rudolph giuliani",
  "description": "I want to know the history of rudolph giuliani the american politician between 1987-2007",
  "subtopics": [
    {
      "subtopic":1,
      "type": "span",
      "description": "Giuliani the litigator. Life as a lawyer in New York."
    },
    {
      "subtopic":2,
      "type": "span",
      "description": "Mayoral Campaigns - 1989 (losing to Dinkins) 1993 (improving police protection, beating dinkins) 1997( first Republican to win a secon
    },
    {
      "subtopic":3,
      "type": "span",
      "description": "Mayoralty - mayor of New York City from 1994 through 2001. The major obstacles he had to overcome during his time as mayor. (Law enfor
    },
    {
      "subtopic":4,
      "type": "span",
      "description": "2000 U.S. Senate campaign. His main opponent being Hilary Clinton."
    },
    {
      "subtopic":5,
      "type": "burst",
      "description": "September 11 terrorist attacks. Giulianis work for helping New York recover"
    },
    {
      "subtopic":6,
      "type": "span",
      "description": "Post-mayoralty- what did Giuliani do in the political scene after leaving his post as mayor (after 2001) (running for president for 20
    },
    {
      "subtopic":7,
      "type": "span",
      "description": "His personal life - knighthood, time person of the year, cancer, affair, divorce"
    }
  ]
}
```

Diversity

- Jaguar
  - Car
  - Guitar
  - Animal

Jaguar
www.**jaguar**.com/ ▾ Jaguar Cars ▾
Official worldwide web site of **Jaguar** Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

Jaguar UK | How alive are you?
www.**jaguar**.co.uk/ ▾ Jaguar Cars ▾
Jaguar has always believed that a car is the closest thing you can create to something that is alive. Explore our range of luxury models, XF, XJ, XK and the new ...

Jaguar Cars - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**_Cars ▾ Wikipedia ▾
Jaguar Cars is a brand of **Jaguar** Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by Tata Motors since ...

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar** ▾ Wikipedia ▾
The **jaguar** Panthera onca, is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas. The **jaguar** is the third-largest ...

Jaguar | Facebook
https://www.facebook.com/**Jaguar** ▾
Jaguar. 6598288 likes · 114518 talking about this. **Jaguar**, as Alive as you are.

Images for jaguar Report images

More images for jaguar

Jaguar | Basic Facts About Jaguars | Defenders of Wildlife
www.defenders.org/**jaguar**/basic-facts ▾ Defenders of Wildlife ▾
The **jaguar** is the largest cat in the Americas. The **jaguar** has a compact body, a broad head and powerful jaws. Its coat is normally yellow and tan, but the color ...

Jaguar Land Rover Careers – Excellence In Motion
www.**jaguar**landrovercareers.com/ ▾
Welcome to **Jaguar** Land Rover careers. With an ever-evolving history and exhilerating future, this is where you'll put your excellence in motion.

Latest models View 4+ more


2015 Jaguar XF


2015 Jaguar XJ


2015 Jaguar F-TYPE


2015 Jaguar XK


2008 Jaguar X-TYPE

People also search for View 15+ more

Maserati


BMW

Aston Martin


Porsche


Mercedes...

Feedback

See results about

Jaguar (Animal)
Mass: 100 kg on average (Adult, Brazilian Pantanal region)
Scientific name: Panthera onca

5. Judging relevance

- In Ad-hoc retrieval, relevance is defined minimally as:

  „A document is relevant if any part of the document is relevant, even a single sentence“

- And independently:

  „A document is relevant independent of all other documents the user has already seen.“

Eliciting relevance

- Having this simple definition of relevance makes it easier for the assesor
- It is very important to have clear (and minimal) instructions for the assesors otherwise the users have to take decisions „is it relevant enough?“
- Leads to inconsistencies -> bad test collection
- How do you judge large scale collections for a ranking task? – **Pooling**
- All competitors are trying to solve the same problem. Each competitor submits „runs“ which are pooled together and evaluated.

Other things to consider ...

- Is the test collection stable?
  - Standard stability tests : Buckley and Voorhess (SIGIR 2000)
- Do you need expert assessors?
  - TREC employs ex-CIA officials to judge relevance (may not be true but they do not use students certainly)
- Do the assesors agree with each other?
  - Inter rater agreement

Summary

- Define the task first. It influences everything else.
- Select the metric to evaluate performance in the task
- Choose the document collection.
- Choose topics which abstract the user‘s real world need.
- Judge documents for a given topic independently and with a minimum requirement.

Borrowed heavily from:

- https://isoboroff.github.io/Test-Colls-Tutorial/Tutorial-slides/assets/fallback/index.html
- http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html

- More on the latest TREC web track : http://www-personal.umich.edu/~kevynct/trec-web-2014/

Historical Search

- Task: I want to know the history of .....
- Collection: News archive
- Measure: Time aware Subtopic Recall
- Topics and judgements?

- http://pharos.l3s.uni-hannover.de:7080/ArchiveSearch/starterkit/relevance.html

IR experiments look a bit like ths....

	T-SBR	TIA-NDCG	TIA-PREC.	TIA-MAP	TIA-ERR	TIA-SBR
LM	0.302	0.209	0.01	0.01	0.023	0.453
TIA-SELECT	0.325	0.213	0.01	0.012	0.028	0.456
T-PM2	0.182	0.107	**0.011**	0.012	0.022	0.322
IA-SELECT	0.258	0.161	0.008	0.009	0.02	0.376
PM2	0.295	0.192	0.011	0.011	0.025	0.444
MDIV	0.309	0.209	0.009	0.011	0.025	0.454
ONLYTIME	0.344	0.22	0.007	0.01	0.024	0.482
HISTDIV	**0.351**	**0.275**	0.01	**0.012**	**0.031**	**0.519**

Table 2: Retrieval Effectiveness ($k = 10$)