

www.trebleclef.eu



Evaluation for (Cross-Language) Information Retrieval

*Mark Sanderson, Department of Information Studies, University of Sheffield,
United Kingdom*

TrebleCLEF Summer School on
Multilingual Information Access
Hotel Santa Croce in Fossabanda, Pisa, Italy
15-19 June 2009

Evaluation in IR

Mark Sanderson

Aims

- To enable you to understand the practicalities of evaluation in IR

Objectives

- At the end of the lecture, you'll be able to...
 - ... calculate precision;
 - ... conduct your own testing
 - ... determine if Google is actually any good.

Schedule

- 09:00 – 10:00 Evaluation lecture
- 10:00 – 10:15 Introduction to exercise
- 10:15 – 11:00 Do exercise
- 11:00 – 11:30 Coffee
- 11:30 – 12:15 Reading
- 12:15 – 12:45 Group reporting
- 12:45 – 13:00 Exercise results

Why?

- You've heard all week about IR systems and improvements
 - but how do we know they are any good?
- Need to evaluate

Evaluation – big topic

- Many different potential approaches
 - One main approach – *test collections*
- Let's look at a bit of history

History of evaluation

- Before rival search engines, rival library catalogue systems
 - Cleverdon and Thorne, 1953
 - *“the author has found the need for a ‘yardstick’ to assist in assessing a particular system’s merits ... the arguments of librarians would be more fertile if there were quantitative assessments of efficiency of various cataloguing systems in various libraries”*



Solution?

- *“Suppose the questions put to the catalogue [from users] are entered in a log, and 100 test questions are prepared which are believed to represent typically such a log. If the test questions are based on material known to be included in the collection, they can then be used to assess the catalogue’s probability of success.”.*



Testing

- Find a document in library
 - construct information request from it
 - challenge librarian to find document using on catalogue or the other
 - *“The pressure distributions over the nose of a body of revolution of fineness ratio 6 for angles of attack 0° to 8° at high subsonic Mach number ($RN > 4 \times 10^6$),”*



Others and followers

- Other library catalogues
 - Gull, 1953
- Then computers
 - Cleverdon, late 1950s, early 1960s
 - Cranfield test collection



At the core of this approach

- Create a *test collection*
 - documents
 - topics
 - documents relevant to topics (*relevance judgments, qrels*)
- Run
 - topics on system
 - compare relevant set with what IR system returned
- Count number of relevant documents.
 - System oriented/laboratory approach

Another approach

- NASA/RECON citation search engine
 - 270,000 citations
- Meister and Sullivan (1967)
 - Examined search logs
 - Number of queries submitted
 - “clicks” on search results
 - Conducted questionnaires with users



Results

- System was popular with users
 - >6,000 searches in 6 weeks
 - “click rate” 35%-46%
- Users liked system
 - Much quicker than manual library catalogue
 - Didn’t have to go to library
- Users wished system searched faster

At the core of this approach

- Focus on user and user actions
 - Let users define what they wish to see
- Later
 - How searching fits into their work
- User oriented approach

Two approaches?

- Test collections
 - Good
 - If IR system being tested changes
 - Just re-run new system on existing test collection
 - Bad
 - User’s opinions & context ignored



Two approaches?

- User oriented approach
 - Bad
 - If IR system being tested changes
 - Need to re-run evaluation
 - Good
 - User’s opinions & context *the focus*



What can you evaluate?

Most IR research

- Focus on relevance
 - What is relevance?
- Translates to
 - lots of relevant documents in top 10
- Our focus in this part of the morning

What is relevance?

- Relevance depends on context of system you build
 - Mizarro, S. How Many Relevances in Information Retrieval? Interacting With Computers, 10(3):305–322, 1998.

What is relevance?

- Does document satisfy information need
- Others
 - Recency
 - Authoritative
 - Cheapest
 - Sense
 - Aspectual
 - Plagiarism
 - Readable

Assume relevance defined

- How to measure?
- Classic approach

$$\text{Precision} = \frac{\text{Relevant and Retrieved}}{\text{Retrieved}}$$

Calculating for one query

- Precision at ?

Rank	Doc ID	Rel?
1	20	
2	7	Relevant
3	18	
4	10	Relevant
5	2	
6	12	
7	16	
8	6	Relevant
9	17	
10	3	

Evaluate a system

- Compute precision at fixed rank for each query
 - 10, 20, 100?
- Average across the queries
- We're all happy right?

What's missing?

- How many documents did we not get?

$$\text{Recall} = \frac{\text{Relevant and Retrieved}}{\text{Total relevant}}$$

Total relevant?

- Early test collections
 - Set of documents (few hundred)
 - Set of queries (50-400)
 - Set of relevance judgements
 - Humans check all documents!

Recall gives...

Rank	Doc ID	Rel?	Recall	Precision	ReIs	Total Rel
			0	0	8	3
1	8	Relevant	0.33	1.00	4	
2	17		0.33	0.50	10	
3	18		0.33	0.33		
4	1		0.33	0.25		
5	9		0.33	0.20		
6	13		0.33	0.17		
7	11		0.33	0.14		
8	16		0.33	0.13		
9	19		0.33	0.11		
10	20		0.33	0.10		

...another perspective

Rank	Doc ID	Rel?	Recall	Precision	ReIs	Total Rel
			0.00	0.00	4	70
1	20		0.00	0.00	7	
2	7	Relevant	0.01	0.50	6	
3	18		0.01	0.33	...	
4	10		0.01	0.25		
5	2		0.01	0.20		
6	12		0.01	0.17		
7	16		0.01	0.14		
8	6	Relevant	0.03	0.25		
9	17		0.03	0.22		
10	3		0.03	0.20		

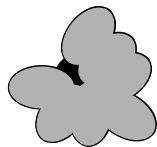
Test collections

- Test collections got bigger
 - Set of documents (few thousand-few million)
 - Humans check all documents?
- Use pooling
 - Target a subset (described in literature)
 - Manually assess these only.
 - Query pooling
 - System pooling

Query pooling

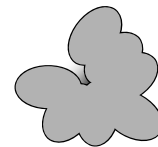
1. Nuclear waste dumping
2. Radioactive waste
3. Radioactive waste storage
4. Hazardous waste
5. Nuclear waste storage
6. Utah nuclear waste
7. Waste dump

Collection



System pooling

All documents



Creating test collections...

- ...is time consuming
- Ad hoc sharing
- Centralised creation
 - TREC, CLEF, NTCIR
 - Many, many others

Test collection references

- System pooling
 - Any of the TREC/CLEF overview papers
- Query pooling
 - Cormack, G.V., Palmer, R.P., Clarke, C.L.A. (1998): Efficient Constructions of Large Test Collections, in *Proceedings of the 21st annual international ACM-SIGIR conference on Research and development in information retrieval*: 282-289
 - With relevance feedback
 - Soboroff, I, Robertson, S. (2003) Building a filtering test collection for TREC 2002, in *Proceedings of the ACM SIGIR conference*.

Other ways of finding relevant?

- Web site structure?
 - Harmandas, V., Sanderson, M. and Dunlop, M.D. (1997) Image retrieval by hypertext links, in *the proceedings of the 20th ACM Conference of the Special Interest Group in Information Retrieval (SIGIR)*, 296-303
- Sitemaps
 - Hawking, D. (2004) Challenges in Enterprise Search, in *Proceedings of the Australasian Database Conference (ADC2004)*
- Topic hierarchies
 - Use groupings of documents in Open Directory to locate related documents
 - Haveliwala, T., Gionis, A., Klein, D. and Indyk, P. (2002) Evaluating Strategies for Similarity Search on the Web in *Proc. of the 11th Int. WWW Conference*

More ways?

- References?
 - Ritchie, A., Teufel, S., Robertson, S. (2006) Creating a Test Collection for Citation-based IR Experiments, in *Proc of NAAACL/HLT conference*
- Temporal clues?
 - Sheridan, Wechsler, Schäuble (1997) Cross-Language Speech Retrieval: Establishing a Baseline Performance, in *Proc. Of ACM SIGIR*

Even more ways?

- Display time?
 - Kelly, D., Belkin, N.J. (2004) Display Time as Implicit Feedback: Understanding Task Effects, in *Proceedings ACM SIGIR*
- Clickthroughs
 - Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T. (2005) Evaluating implicit measures to improve web search, *ACM Transactions on Information Systems*, Vol. 23, No. 2, 147-168
- Tagging and bookmarks
 - Xu, S., Bao, S., FeiB., Su, Z. and Yu, Y. (2008) Exploring Folksonomy for Personalized Search, in *Proceedings ACM SIGIR*

Test collections popular

- Underpins majority of research in IR
- Validation of pooling
 - Zobel, J. (1998) How Reliable Are the Results of Large-Scale Information Retrieval Experiments?, in *Proceedings of the 21st ACM SIGIR conference*
- Validation of relevance judgement variability
 - Voorhees, E. (1998): Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness, in *Proceedings of the 21st annual international ACM-SIGIR conference on Research and development in information retrieval*: 315-323

Let's do an evaluation

- Today

National archives exercise

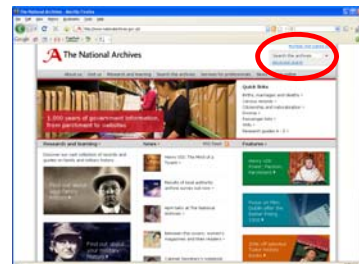
Search engine comparison

Aim

- To compare two search engines searching over The National Archives (TNA)
 1. TNA's in-house search engine
 2. Google site search
- Use precision as well as your impression of the two search engines as your means of comparison

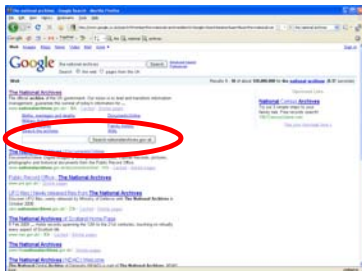
Search Engine 1

- <http://www.nationalarchives.gov.uk/>



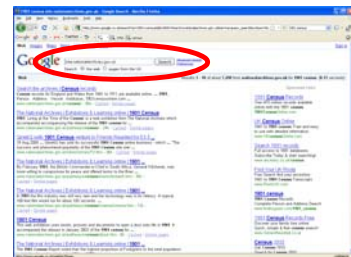
Search Engine 2

- <http://www.google.co.uk/>



Search engine 2

- Use the "site text" in the query



Two types of relevance

- On the web queries
 - Informational – almost all test collections
 - Your classic IR query
 - Navigational
 - I want a home page
 - Transactional
 - I want a service

References

- Broder, A. (2002) A taxonomy of web search, *SIGIR Forum*, 36(2), 3-10.

Judging for relevance

- The question to ask is different for each type
 - Navigational query
 - Is the page a great starting point (i.e. home page) for the query
 - Informational query
 - Is the page relevant to the user's request?
 - A catalogue entry for a relevant document is relevant
 - A page leading to a relevant document that has to be paid for is relevant.

For each engine

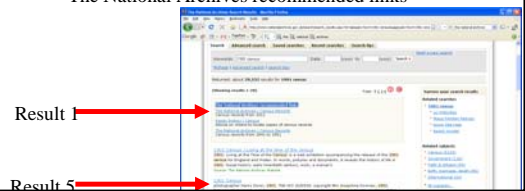
- Calculate the precision at 10 for the initial query
$$\text{Precision} = \frac{\text{Relevant and Retrieved}}{\text{Retrieved}}$$
- E.g. find 3 relevant in the top 10
 - Precision = 3/10 (0.3)

You will be given

- 4 queries each
 - 2 Navigational
 - 2 Informational
- Type in the query title (the initial query)
 - In each search engine
 - Use the description to judge relevance of retrieved documents
 - Judge the top 10 results
 - Record URLs of relevant

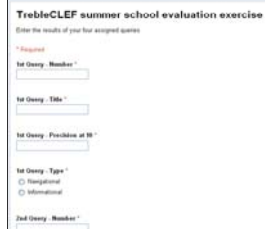
What to judge

- First 10 results only
 - Ignore Google adverts
 - Start National Archive links at
 - “The National Archives recommended links”



Fill in your answers at...

- This live form
 - <http://tinyurl.com/trebleclef>



TrebleCLEF summer school evaluation exercise

Enter the results of your four assigned queries

1st Query - Number *

1st Query - Title *

1st Query - Position at IR *

1st Query - Type *

Informational

Informational

2nd Query - Number *

I will collate a set of results

- Before the end of the session.

Sources for further reading?

- Foundations and Trends® in Information Retrieval
 - *Methods for Evaluating Interactive Information Retrieval Systems with Users*, By Diane Kelly (University of North Carolina)
 - *Mining Query Logs: Turning Search Usage Data into Knowledge*, By Fabrizio Silvestri (National Council of Research, IT)
 - *Test Collection Evaluation of Ad-hoc Retrieval Systems*, By Mark Sanderson (University of Sheffield)

Sources for further reading?

- *TREC: Experiment and Evaluation in Information Retrieval*, By Ellen M. Voorhees, Donna K. Harman
- *The Turn: Integration of Information Seeking and Retrieval in Context*, By Peter Ingwersen, Kalervo Järvelin



Reading

- Five papers
 - Joachims
 - Agichtein
 - Smith
 - Kelly
 - Järvelin

In groups

- Skim read one paper,
- In 40 minutes
 - Elect a representative to tell us all about one aspect of the paper
 - 25 minutes to read paper
 - 15 minutes to discuss aspect
- 5 minute (report back) talk
 - Just talk, no PPT

Aspect

- Joachims
 - Describe methodology one important result
- Agichtein
 - Describe “click prediction” methodology
- Smith
 - Describe methodology, main result
- Kelly
 - Describe methodology used
- Järvelin
 - Describe the two measures presented