# Evaluation I

Mark Sanderson

www.rmit.edu.au

RMIT UNIVERSITY

## Who am I?

- Professor at RMIT University, Melbourne
- Before
  - Professor at University of Sheffield
  - Researcher at UMass Amherst
  - Researcher at University of Glasgow
- Online
  - @IR_oldie
  - http://www.seg.rmit.edu.au/mark/



## Where do slides come from?

- Wrote large review of test collection evaluation
  - Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval*, *4*(4), 247-375. doi:10.1561/1500000009
    - http://www.seg.rmit.edu.au/mark/publications/my_papers/FnTIR.pdf
- Couple of slides from ChengXiang Zhai



## Outline

- Why evaluate?
- Evaluation I
  - traditional evaluation, test collections
- Evaluation II
  - Examining test collections
  - Testing by yourself



## Evaluation I

- History of evaluation
  - Brief history of IR
- Test collections
- Evaluation measures
- Exercise



## Evaluation II

- Review exercise
- Statistical significance
- Examining test collection design
- New evaluation measures
- Building your own testing collection
  - Crowd sourcing
- Other evaluation approaches
  - Briefly (if we have time)

# Why evaluate?

www.rmit.edu.au

RMIT UNIVERSITY

## Why evaluate?

- Every researcher defines IR their own way
- For me
  - Underspecified queries



## Can't predict effectiveness

- "*Studies of the software industry indicate that when ideas people thought would succeed are evaluated through controlled experiments, less than 50 percent actually work out.*"
  - http://www.technologyreview.com/printer_friendly_article.aspx?id=32409
- No reason to assume IR is different
  - Evaluate ideas early, find the ones that work.

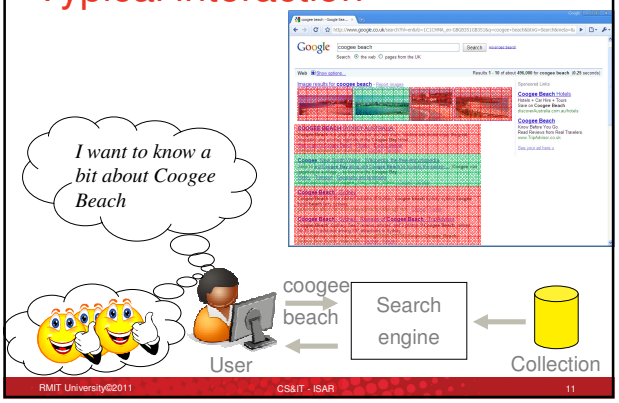


## Combat HiPPOs

- Highest Paid Person's Opinion
  - Often wrong
  - Test test test

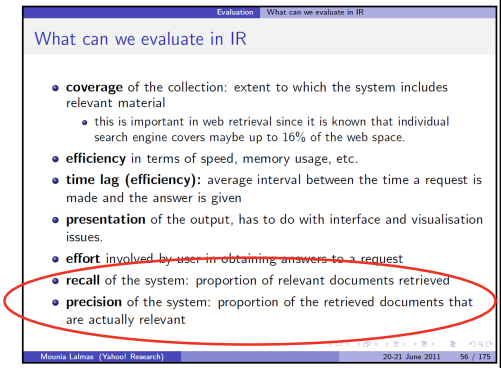


## Typical interaction






## Mounia said…



# Bit of history

www.rmit.edu.au

RMIT UNIVERSITY

## History of evaluation

- Before IR systems, there were libraries
  - The search engine of the day

- Organise information using a subject catalogue
  - Sort cards by author
  - Sort cards by title
  - Sort cards by subject
    - How to do this?




## Not just public libraries

- MIT Masters thesis, Philip Bagley, 1951

To quote Professor Perry: "Recently published statistics relating to chemical publication show that a search of Chemical Abstracts would have been complete in 1920 after considering twelve volumes containing some 184,000 abstracts. But in 1935 there would have been fifteen more volumes to search, and these new volumes alone contain about 382,000 abstracts. By the end of 1950 the forty-four volumes of Chemical Abstracts to be searched contained well over a million abstracts." If the present trend in publication continues, the total abstracts published in this one field by 1960 will be almost 1,800,000.



## Competing catalogue schemes

- Librarians argued over which was the best subject catalogue to use
  - "*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*"
  - "*Suppose the questions put to the catalogue 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*"

– Thorne, R. G. (1955). The efficiency of subject catalogues and the cost of information searches. Journal of documentation, 11, 130-148.



## Created test collections

- Collection of documents
  - Everything in the library
- Topics
  - Typical queries users would have
- Judgements on what comes back




## Test



# Invented twice – 1953

- Thorne and Cleverdon
  - Cranfield, UK
- Gull
  - USA
    - Gull, C. D. (1956). Seven years of work on the organization of materials in the special library. *American Documentation, 7*(4), 320-329. doi:10.1002/asi.5090070408
- Relatively small projects
  - Each made mistakes

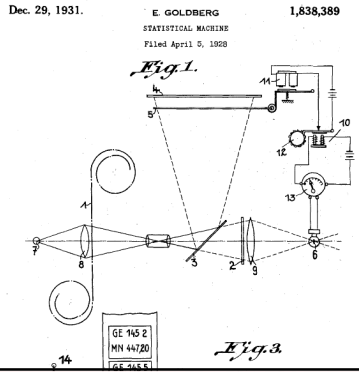


# At the same time…

- While librarians were coping with the information explosion
  - Could machines help?
  - Could computers help?
- Very brief history of machines and computers for search

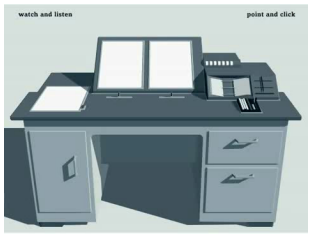


# Machines doing IR






# As we may think – Bush 1945






# Computers doing IR

- Holmstrom 1948

Then there is also in America a machine called the Univac which has a typewriter keyboard connected to a device whereby letters and figures are coded as a pattern of magnetic spots on a long steel tape. By this means the text of a document, preceded by its subject code symbol, can be recorded on the tape by any typist. For searching, the tape is run through the machine which thereupon automatically selects and types out those references which have been coded in any desired way at a rate of 120 words a minute —complete with small and capital letters, spacing, paragraphing, indentations and so on. (If the tape is run through the other way, it obediently types out the text backwards at the same rate !)



# Information Retrieval

- Calvin Mooers, 1950

The problem under discussion here is machine searching and retrieval of information from storage according to specification by subject. An example is the library problem of selection of technical abstracts from a listing of such abstracts. It should not be necessary to dwell upon the importance of information retrieval before a scientific group such as this, for all of us have known frustration from the operation of our libraries -- all libraries, without exception.

# 1950s IR research

- Kent, A., Berry, M. M., Luehrs Jr, F. U., & Perry, J. W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation*, *6*(2), 93-101. doi:10.1002/asi.5090060209
- Maron, M. E., Kuhns, J. L., & Ray, L. C. (1959). *Probabilistic indexing: A statistical technique for document identification and retrieval* (Technical Memorandum No. 3) (p. 91). Data Systems Project Office: Thompson Ramo Wooldridge Inc, Los Angeles, California.
- Mooers, C. N. (1959). *The Intensive Sample Test for the Objective Evaluation of the Performance of Information Retrieval System* ( No. ZTB-132) (p. 20). Cambridge, Massachusetts: Zator Corporation.



# Back to evaluation

- Testing ideas started with librarians
  - Subject catalogues
- At same time, computers being used for search
  - Initially searching catalogue metadata
  - Soon searching words
  - How to test them?



# Cleverdon

- Observed mistakes in earlier testing
- Proposed larger project
  - Initially for library catalogues
    - Funded by the NSF (US government agency)
  - Then for computers
- Cranfield collections



# Legacy of Cranfield Tests

*"What, then, is the Cranfield legacy? … First, and most specifically, it has been very difficult to undermine the major result of Cleverdon's work… Second, methodologically, Cranfield 2, whatever its particular defects, clearly indicated what experimental standards ought to be sought. Third, our whole view of information retrieval systems and how we should study them has been manifestly influenced, almost entirely for the good, by Cranfield."* (Spärck Jones, 1981)

Cleverdon received the ACM SIGIR Salton Award in 1991

http://www.sigir.org/awards/awards.html

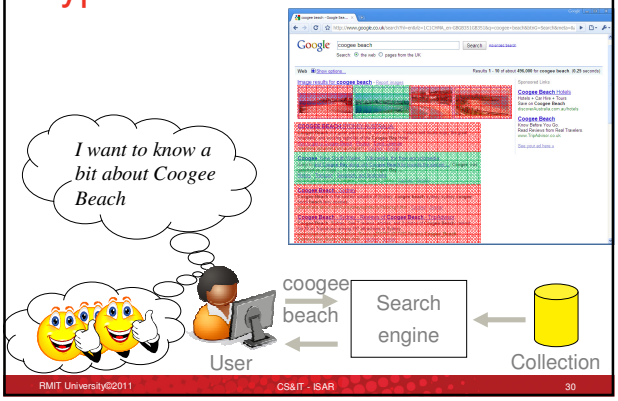


# Cranfield model

- Test collection
  - Collection of documents
  - Topics
    - Typical queries users would enter
  - QRELS
    - List of documents relevant to each query
  - Measure



# Typical interaction



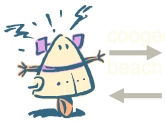
## Simulation of real searching

| Id | Topic | QRELS |
|---|---|---|
| 1 | Coogee beach | 7623, 9256 |
| 2 | Melbourne zoo | 5425, 7654, 9582 |
| 3 | The Ghan | 3417, 6589 |
| 4 | Healsville sanctuary | 6539, 8042 |
| 5 | Kings canyon | 4375, 5290 |
| 6 | Great ocean road | 9301, 7392 |

Doc 3452
Doc 7623
Doc 4652
Doc 8635

Evaluation measure

| Id | Score |
|---|---|
| 1 | 25% |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |

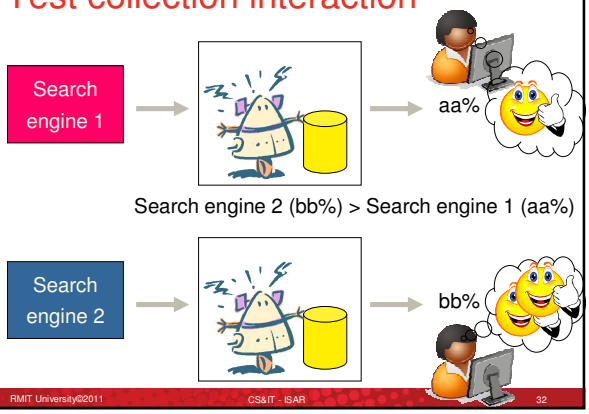



Search engine

Collection



## Test collection interaction






## Cleverdon's Cranfield Tests

Cyril Cleverdon
(Cranfield Inst. of Tech, UK)

- 1957-1960: Cranfield I
  - Comparison of cataloguing methods
  - Controversial results (lots of criticisms)
- 1960-1966: Cranfield II
  - More rigorous evaluation methodology
  - Introduced precision & recall
  - Decomposed study of each component in an indexing method
  - Still lots of criticisms, but….

*Slide from ChengXiang Zhai's presentation*



## Cleverdon's major result?

- Searching based on words was as good as searching the subject catalogues
  - Implication
    - May not need librarians to classify document
- Controversial
  - Stood up because testing done well.



## Test collection is

- Simulating your operational setting
- Results from test collection are predicting how users will behave



## Advantages

- Batch processing
- Great for ranking
  - Different systems
  - Versions of systems

# Sharing

- IR community recognised importance of sharing test beds
  - One of the very first CS disciplines to do this.
- My first trip to another IR group



# Early test collections

- 1950s
  - Cleverdon and Thorne
  - Gull
- 1960s
  - Cleverdon - Cranfield
  - Salton – SMART
  - Many others



# Examples

| Name | Docs. | Qrys | Year | Size, Mb | Source document |
|---|---|---|---|---|---|
| Cranfield 2 | 1,400 | 225 | 1962 | 1.6 | Title, authors, source, abstract of scientific papers from the aeronautic research field, largely ranging from 1945-1963. |
| ADI | 82 | 35 | 1968 | 0.04 | A set of short papers from the 1963 Annual Meeting of the American Documentation Institute. |
| IRE-3 | 780 | 34 | 1968 | - | A set of abstracts of computer science documents, published in 1959-1961. |
| NPL | 11,571 | 93 | 1970 | 3.1 | Title, abstract of journal papers |
| MEDLARS | 450 | 29 | 1973 | - | The first page of a set of MEDLARS documents copied at the National Library of Medicine. |
| Time | 425 | 83 | 1973 | 1.5 | Full text articles from the 1963 edition of Time magazine. |

http://ir.dcs.gla.ac.uk/resources/test_collections/



# QRELS

- List of documents relevant to each query?
  - Most early collections small enough to check all documents
  - More on this later.



# Other problems




# Evaluating early IR system

- Many early IR systems Boolean
  - Split collection in two: documents that
    - Match the query (Retrieved)
    - Don't match the query (Not retrieved)
  - Test collection: those documents that are
    - Relevant
    - Not Relevant

# Measuring Boolean output

- Contingency table

| | Relevant | Not-relevant | |
|---|---|---|---|
| Retrieved | a | b | a+b |
| Not retrieved | c | d | c+d |
| | a+c | b+d | a+b+c+d |

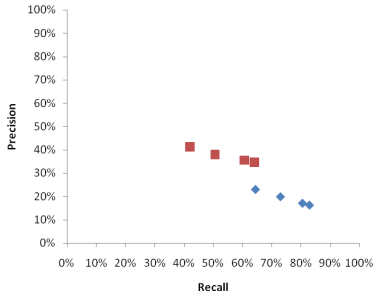
$$\text{Precision} = \frac{a}{a+b} \qquad \text{Recall} = \frac{a}{a+c}$$

$$\text{Fallout} = \frac{b}{b+d}$$



# Precision/Recall

- Inverse relationship






# Summarising the two

- Isn't one measure better than two?
  - Van Rijsbergen's f: weighted harmonic mean

$$f = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$



# Aggregate across topics

- Compute score for each topic
  - Take the mean
- Simple for Boolean
  - Can be harder for other IR systems



# Review where we are

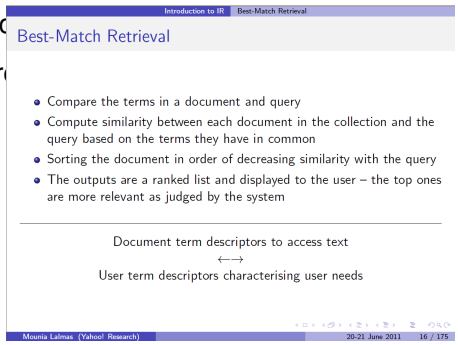
- Cleverdon's Cranfield model of evaluation
  - Test collection
    - Collection
    - Topics
    - QRELS



# Measuring and scaling

# Evaluation of ranked retrieval

- Mo
- Pr





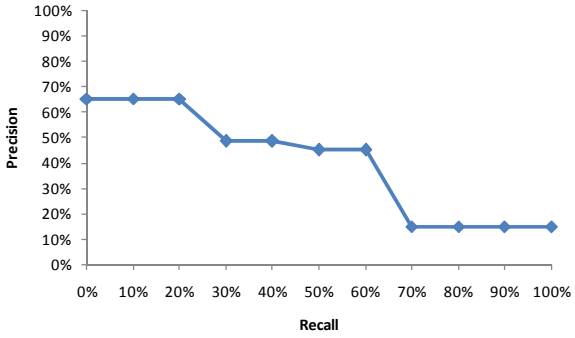

# Precision down ranking

Topic 1

| Rank |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |



# Graph two topics






# Produce single number?

- Measure area under graph
  - In old papers often called
    - average precision
    - interpolated average precision



# Finding everything?

- Cooper's Expected Search Length (ESL) – 1968
  - *"most measures do not take into account a crucial variable: the amount of material relevant to [the user's] query which the user actually needs"*
  - *"the importance of including user needs as a variable in a performance measure seems to have been largely overlooked"*
  - ESL measured what user had to see in order to get to what they wanted to see.
    - Rarely used, but highly influential



# Problems with scale

| Name | Docs. | Qrys | Year | Size, Mb | Source document |
|---|---|---|---|---|---|
| Cranfield 2 | 1,400 | 225 | 1962 | 1.6 | Title, authors, source, abstract of scientific |
| ADI | 82 | 35 | 1968 | 0.04 | A set of short papers from the 1963 Annual |
| INE-3 | 780 | 34 | | | |
| NPL | 11,571 | 93 | 1970 | | |
| MEDLARS | 450 | 29 | 1973 | | The first page of a set of MEDLARS |
| Time | 425 | 83 | 1973 | 1.5 | |

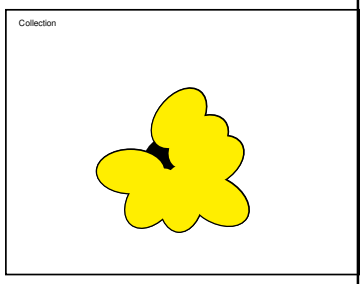
**By mid 1970s, commercial IR systems searched hundreds of thousands of documents**

# Test collections

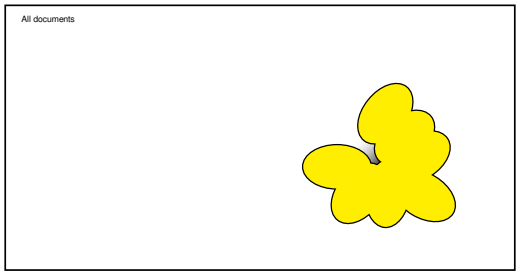
- Test collections got bigger
  - Set of documents (few thousand-few million)
    - Humans check all documents?
- Use pooling
  - Target a subset (described in literature)
  - Manually assess these only.
    - Spärck Jones, K., & van Rijsbergen, C. J. (1975). *Report on the need for and the provision of an "ideal" information retrieval test collection* (British Library Research and Development Report No. 5266) (p. 43). Computer Laboratory, University of Cambridge.
  - 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




# System pooling




# Slightly bigger collections

| Name | Docs. | Qrys. | Year | Size, Mb | Source document |
|---|---|---|---|---|---|
| INSPEC | 12,684 | 77 | 1981 | - | Title, abstract, and indexing information from Sep-Dec 1979 issues of |
| CACM | 3,204 | 64 | 1983 | 2.2 | Titles and abstracts from articles of |
| CISI | 1,460 | 112 | 1988 | 2.2 | Author, title/abstract, and citation data for the and manually-indexed from publication, 1969-1979 |
| LISA | 6,004 | 35 | 1988 | 3.1 | Ta Science Abstracts database. |

**By 1990s, commercial IR systems searched millions of documents**



# Individual groups

- Weren't able to produce test collections at sufficient scale
- Someone needed to coalesce the research community
  - TREC
    - Donna Harman

http://www.itl.nist.gov/iad/photos/trec2001.gif



# TREC 1992

- create test collections for a set of retrieval tasks;
- promote as widely as possible research in those tasks;
- organize a conference for participating researchers to meet and disseminate their research work using TREC collections.


http://trec.nist.gov/images/paper_3.jpg

# TREC approach

- TREC
  - Gets a large collection
  - Forms topics
- Participating groups
  - Get collection, run topics on their IR system
  - Return to TREC top ranked documents for each topic (run)
    - Used to build the pool
- TREC judges the pool
- TREC holds a conference
  - Calculates and publishes results



# TREC collections

- Adhoc
  - Newspaper and government documents
- Spoken document
- Cross language
- Confusion (OCR data)
- Question answering
- Medical data
- Etc, etc

- Collections became standard



# TREC approach successful

- Many spin off exercises
  - NTCIR
  - CLEF
  - INEX
  - FIRE
  - Etc, etc



# TREC evaluation measures

- TREC defined many standard evaluation measures
  - Mean Average Precision

$$AP = \frac{\sum_{rn=1}^{N} (P(rn) \times rel(rn))}{R}$$

  - *N* is the number of documents retrieved
  - *rn* is the rank number
  - *rel(rn)* returns either 1 or 0 depending on the relevance of the document at *rn*
  - *P(rn)* is the precision measured at rank *rn*
  - *R* is the total number of relevant documents for this particular topic



# **Mean** average precision?

- Calculate AP for each topic in the test collections
- Take the mean of those AP scores

- Mean Average Precision
  - Average Average Precision
    - Would have been silly.
  - Sometimes called
    - non-interpolated average precision



# Precision at fixed rank

- Existed before TREC
  - Popularised around TREC

$$P(n) = \frac{r(n)}{n}$$

- Variant
  - R-Precision

$$P(R)$$

- What do these measure ignore?

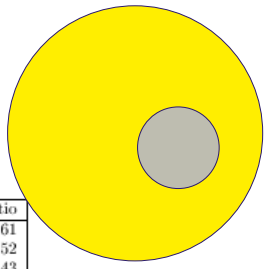
# Property of R-precision

- At the point *R*,
  - #relevant documents ranked below R
  - Equals
  - #non-relevant documents ranked above *R*,
- Some call R
  - *equivalence number*
- Calling R-precision
  - *missed@equivalent*



# P(10) behaviour

- TREC VLC – 1997
  - VLC - 20Gb
  - Baseline - 2Gb



| Group | Baseline | VLC | Ratio |
|---|---|---|---|
| City | 0.320 | 0.515 | 1.61 |
| ATT | 0.348 | 0.530 | 1.52 |
| ANU | 0.356 | 0.509 | 1.43 |
| UMass | 0.387 | 0.505 | 1.31 |
| IBMg(Brown) | 0.275 | 0.361* | 1.31* |
| U Waterloo | 0.408 | 0.643 | 1.29 |
| IBMs(Franz) | 0.271 | 0.348 | 1.28 |



# Why is this happening?

- This effect happens for
  - P(10)
- But not for
  - P(R)
  - MAP
- Why?



# Measuring one document

- Known item search
  - Thorne's 1955 test collection
- Mean Reciprocal Rank (MRR)

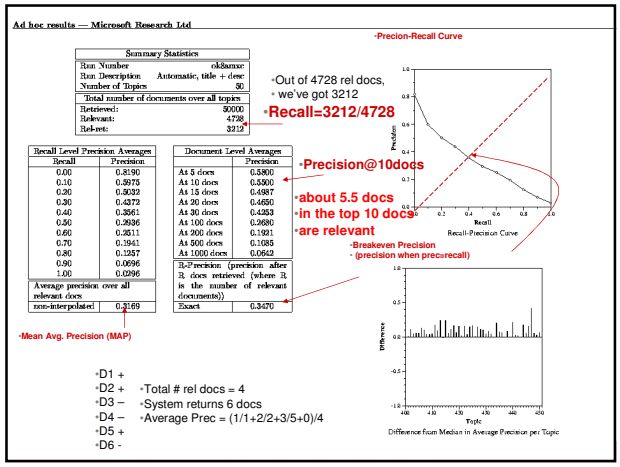
| MRR=1 | | MRR=0.5 | | MRR=0.5 | | MRR=0 | |
|---|---|---|---|---|---|---|---|
| **Rank** | **Rel** | **Rank** | **Rel** | **Rank** | **Rel** | **Rank** | **Rel** |
| 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 2 | 0 | 2 | 1 | 2 | 1 | 2 | 0 |
| 3 | 0 | 3 | 1 | 3 | 0 | 3 | 0 |
| 4 | 0 | 4 | 0 | 4 | 0 | 4 | 0 |
| 5 | 0 | 5 | 1 | 5 | 0 | 5 | 0 |



# trec_eval

- TREC standardised evaluation code
  - http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz
- Given output from IR system searching over test collection
  - Produces all measures (and many more)
  - Most researchers use trec_eval to save time, avoid introducing bugs.
    - Look at some output…





Ad hoc results — Microsoft Research Ltd

Summary Statistics

| | |
|---|---|
| Run Number | ok8amxc |
| Run Description | Automatic, title + desc |
| Number of Topics | 50 |
| Total number of documents over all topics | |
| Retrieved: | 50000 |
| Relevant: | 4728 |
| Rel-ret: | 3212 |

Recall Level Precision Averages

| Recall | Precision |
|---|---|
| 0.00 | 0.8190 |
| 0.10 | 0.5975 |
| 0.20 | 0.5032 |
| 0.30 | 0.4372 |
| 0.40 | 0.3661 |
| 0.50 | 0.2936 |
| 0.60 | 0.2611 |
| 0.70 | 0.1941 |
| 0.80 | 0.1257 |
| 0.90 | 0.0696 |
| 1.00 | 0.0296 |
| Average precision over all relevant docs | |
| non-interpolated | 0.3169 |

Document Level Averages

| | Precision |
|---|---|
| At 5 docs | 0.5800 |
| At 10 docs | 0.5500 |
| At 15 docs | 0.4987 |
| At 20 docs | 0.4650 |
| At 30 docs | 0.4353 |
| At 100 docs | 0.3680 |
| At 200 docs | 0.1921 |
| At 500 docs | 0.1085 |
| At 1000 docs | 0.0642 |
| R-Precision (precision after R docs retrieved (where R is the number of relevant documents)) | |
| Exact | 0.3470 |

Precion-Recall Curve

Out of 4728 rel docs,
we've got 3212

Recall=3212/4728

Precision@10docs

about 5.5 docs
in the top 10 docs
are relevant

Breakeven Precision
(precision when prec=recall)

Mean Avg. Precision (MAP)

D1 +
D2 +
D3 –
D4 –
D5 +
D6 -

Total # rel docs = 4
System returns 6 docs
Average Prec = (1/1+2/2+3/5+0)/4

Recall-Precision Curve

Difference from Median in Average Precision per Topic

# TREC lessons

- Highly successful, but some issues
  - Collections
  - Topics
  - Relevance



# TREC collections

- Early collections
  - Largely articles (news, journals, government)
  - Long time to try web search
    - Assumption web wasn't different
      - Very wrong
    - Fixed now



# TREC topics

- Not criticising the form
  - Though many do

```
<top>

<num> Number:  200

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description:  Document must report on how the importation of foreign
textiles or textile products has influenced or impacted on the U.S. textile
industry.

<narr>  Narrative:  The impact can be positive or negative or qualitative.
It may include the expansion or shrinkage of markets or manufacturing volume
or an influence on the methods or strategies of the U.S. textile industry.
"Textile industry" includes the production or purchase of raw materials;
basic processing techniques such as dyeing, spinning, knitting, or weaving;
the manufacture and marketing of finished goods; and also research in the
textile field.

</top>
```



# TREC topics

- Criticising topic formation
  - Test collection simulates operational setting
    - Topics need to be typical topics
  - Early TREC collections
    - Searched collection for potential topics
    - Removed topics that returned too many
    - Removed topics that returned too few
    - Removed topics that appeared ambiguous
    - Discuss



# TREC relevance

- TREC documents judged either
  - Relevant
    - Even if just a single sentence was relevant
  - Not relevant



# TREC lessons

- Criticisms apply to early TREC collections
  - More recent TREC collections
    - Collections from wide range of sources
      - Web, Blogs, Twitter, etc
    - Topics sampled from query logs
    - Multiple degrees of relevance
  - However, early TREC model copied by others
    - So need to be cautious.

## Review where we are

- Measures for Boolean retrieval
  - Precision, Recall, and F
- Early ranking measures
  - Interpolated AP
- New test collections built
  - Failed to keep up with commercial scale
- Pooling invented
  - Researchers gave up on knowing all relevant documents



## Review where we are

- TREC collections formed
  - Gave researchers
    - large test collections
    - Forum to meet and share research
- Newer evaluation measures defined
  - MAP, P(n), P(R), MRR



# Search engine comparison

www.rmit.edu.au

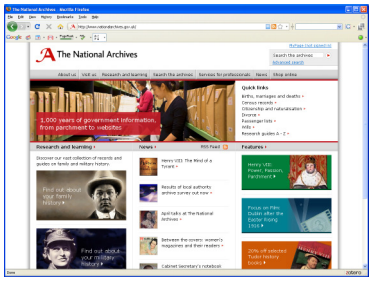
RMIT UNIVERSITY

## 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

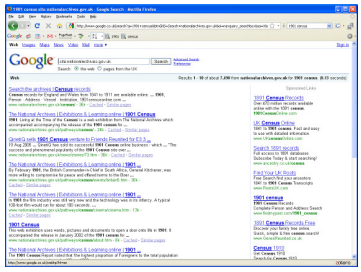
## Search Engine 1

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



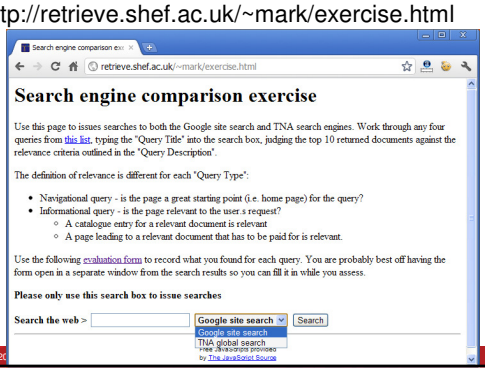

## Search engine 2

- Google site search



# Use this page

- http://retrieve.shef.ac.uk/~mark/exercise.html




# Two types of relevance

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

# 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.

# From the list

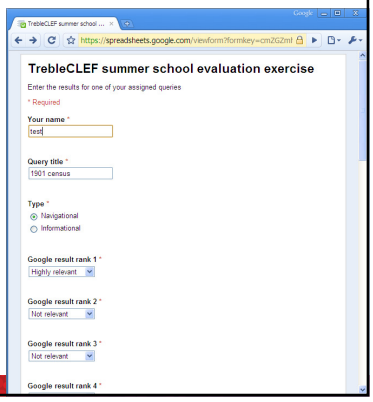
- 4 queries each
  - 2 Navigational
  - 2 Informational
- Enter the query (the initial query)
  - In each search engine
  - Use the description to judge relevance of retrieved documents
  - Judge the top 10 results
    - Record URLs of relevant

# What to judge

- First 10 results only
  - Ignore Google adverts
  - Ignore National Archive documents beyond top 10

# tinyurl.com/trebleclef

- Fill in this online form



## I will collate a set of results

- For the next evaluation lecture.