Information Retrieval I

David Hawking

30 Sep 2010

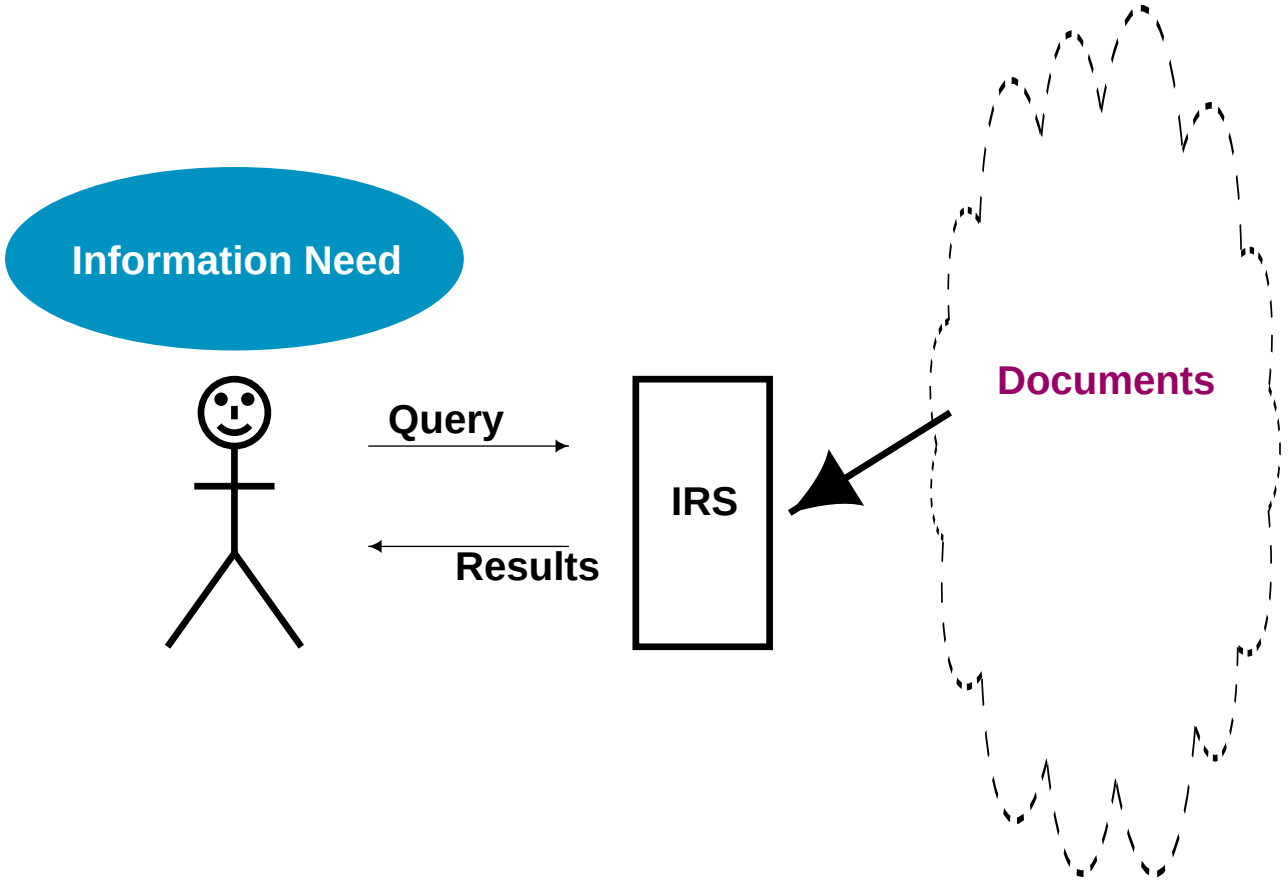
Machine Learning Summer School, ANU

Session Outline

- Ranking documents in response to a query
- Measuring the quality of such rankings
- Case Study: Tuning 40 parameters at the London School of Economics
- *Coffee Break*
- Web SearchEngineering
- Field Work: how do Web search engines really work?
- *Stretch Break*
- Discussion: Other IR problems for machine learning
- Historical context

Start a Machine Learning Run to discuss later

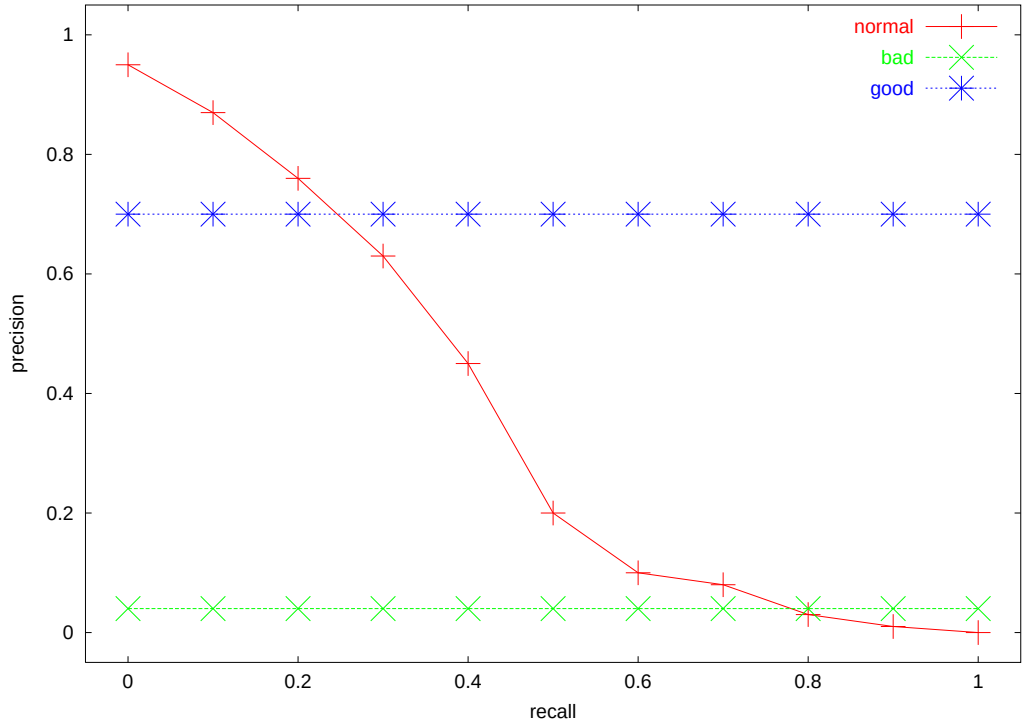
Information Retrieval



- Ranked retrieval $\longrightarrow$ ranked list of results

Measuring/comparing the quality of rankings

Precision - Recall curves



- Mean average precision (MAP) = area under curve.

Normalised Discounted Cumulative Gain

	1																	20
Perfect System	5	5	4	4	3	2	1	1	1	1	1	1	1	1	1	1	1	-	-
Real System A	1	2	3	4	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Real System B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	4	3	2	1

(Relevance judged on a 5 point scale)

$$DCG[r] = \begin{cases} G[1] & \text{if } r = 1 \\ DCG[r-1] + G[r]/\log_b r & \text{otherwise} \end{cases}$$

But where do the utility judgments come from?

(We'll return to this later on.)

Probability Ranking Principle

Maron & Kuhns, JACM, 1960
"... technique called "Probabilistic Indexing", allows a computing machine, given a request for information, to derive a number (called the "relevance number") for each document, which is a measure of the probability that the document will satisfy the given request. The result of the search is an ordered list of those documents which satisfy the request ranked according to their probable relevance."

- Cooper (1977) produced a counter example, based on sub-classes of users with different criteria submitting the same query $\Rightarrow$ need to model diversity.

Modern Ranking Functions

$$RSV = \alpha_0 D_o + \ldots + \alpha_n D_n + \beta_0 S_0 + \ldots + \beta_n S_n \qquad (1)$$

- Machine learned combination of:
  - dynamic scores – probability of relevance given doc and query text
  - static priors, independent of the query

Dynamic factors

Key Concepts

- **Term** — Basic unit of indexing: e.g. a word, a word-stem, a phrase. Could be any discrete feature, not necessarily derived from text.
- **Term Coordination.**
- tf — Term frequency.
- N — Number of documents in the collection.
- V — Vocab – distinct terms in the collection.
- n_i — Number of documents with i-th term present.
- idf — Inverse document frequency. Spärck Jones, J Doc, 1972: $\lceil \log_2 N \rceil - \lceil \log_2 n_i \rceil + 1$.
- **Relevance** — Often modelled as dichotomous variable. $Rel | \overline{Rel}$

Probabilistic Retrieval

(*From Robertson and Zaragoza tutorial, SIGIR 2007*) Starting with the probability ranking principle:

$$P(Rel|d,q) \propto_q \frac{P(Rel|d,q)}{P(\overline{Rel}|d,q)} \quad \text{transform to odds} \quad (2)$$

$$\propto_q \frac{P(d|Rel,q)}{P(d|\overline{Rel},q)} \quad \text{Bayes rule} \quad (3)$$

$$\approx \prod_V \frac{P(tf_i|Rel,q)}{P(tf_i|\overline{Rel},q)} \quad \text{Assume independence} \quad (4)$$

$$\approx \prod_{t \in q} \frac{P(tf_i|Rel,q)}{P(tf_i|\overline{Rel},q)} \quad \text{Restrict to query terms} \quad (5)$$

$$\propto_q \sum_{t \in q} \log \frac{P(tf_i|Rel,q)}{P(tf_i|\overline{Rel},q)} \quad \text{So we can add weights} \quad (6)$$

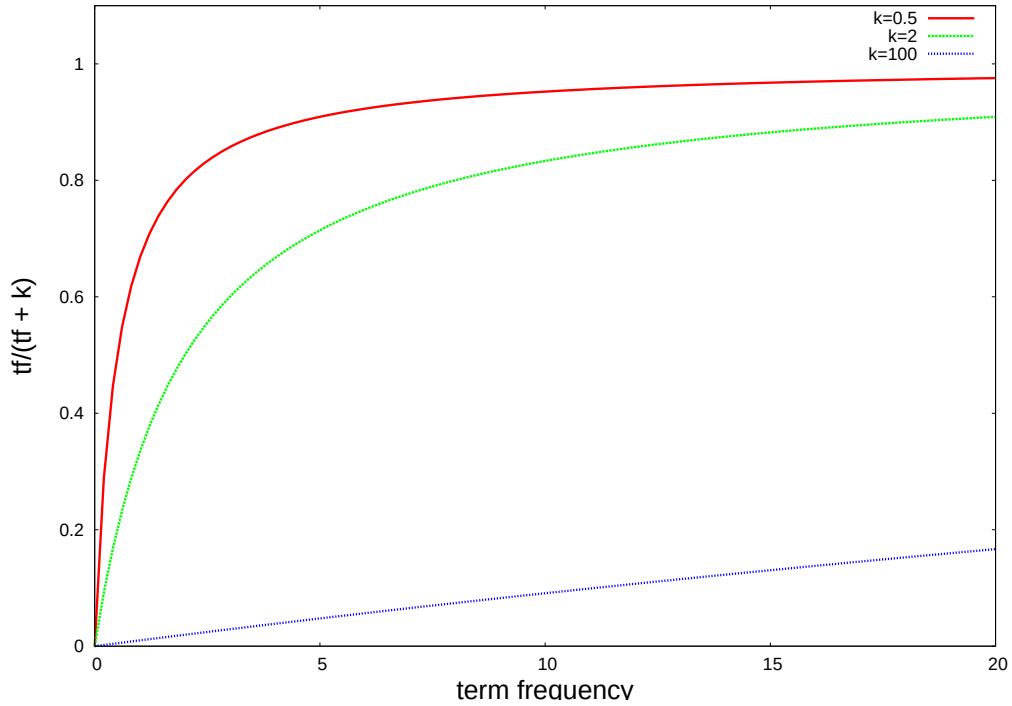
Okapi BM25 (Robertson et al, 1994)

$$w_t = tf_d \times \frac{\log(\frac{N-n+0.5}{n+0.5})}{k_1 \times ((1-b) + b \times \frac{dl}{avdl}) + tf_d} \tag{7}$$

$$S_d = \sum_{t \in q} w_t \tag{8}$$

- S_d is not a probability but should be rank-equivalent to it.

Term saturation



$$\frac{tf}{tf + k} \tag{9}$$

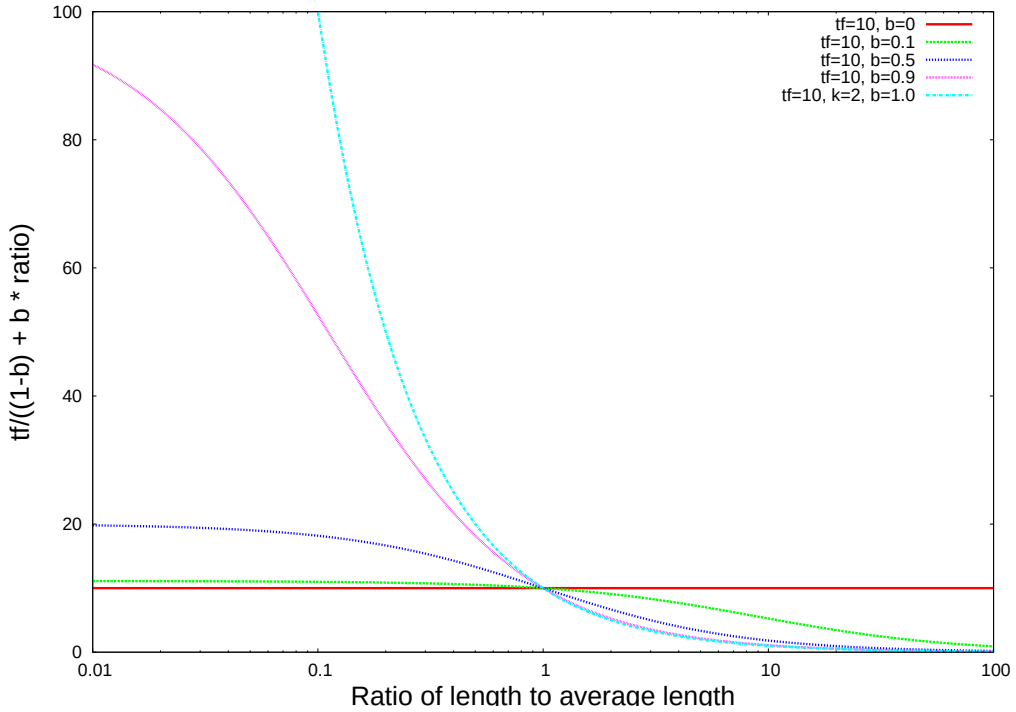
- Modelling saturation is important.

Length normalisation

Need for normalisation of tf_i depends upon why some documents are longer than others. Make it tunable:

$$tf_i' = tf_i / B \tag{10}$$

$$B = (1 - b) + b\frac{dl}{\overline{dl}} \tag{11}$$



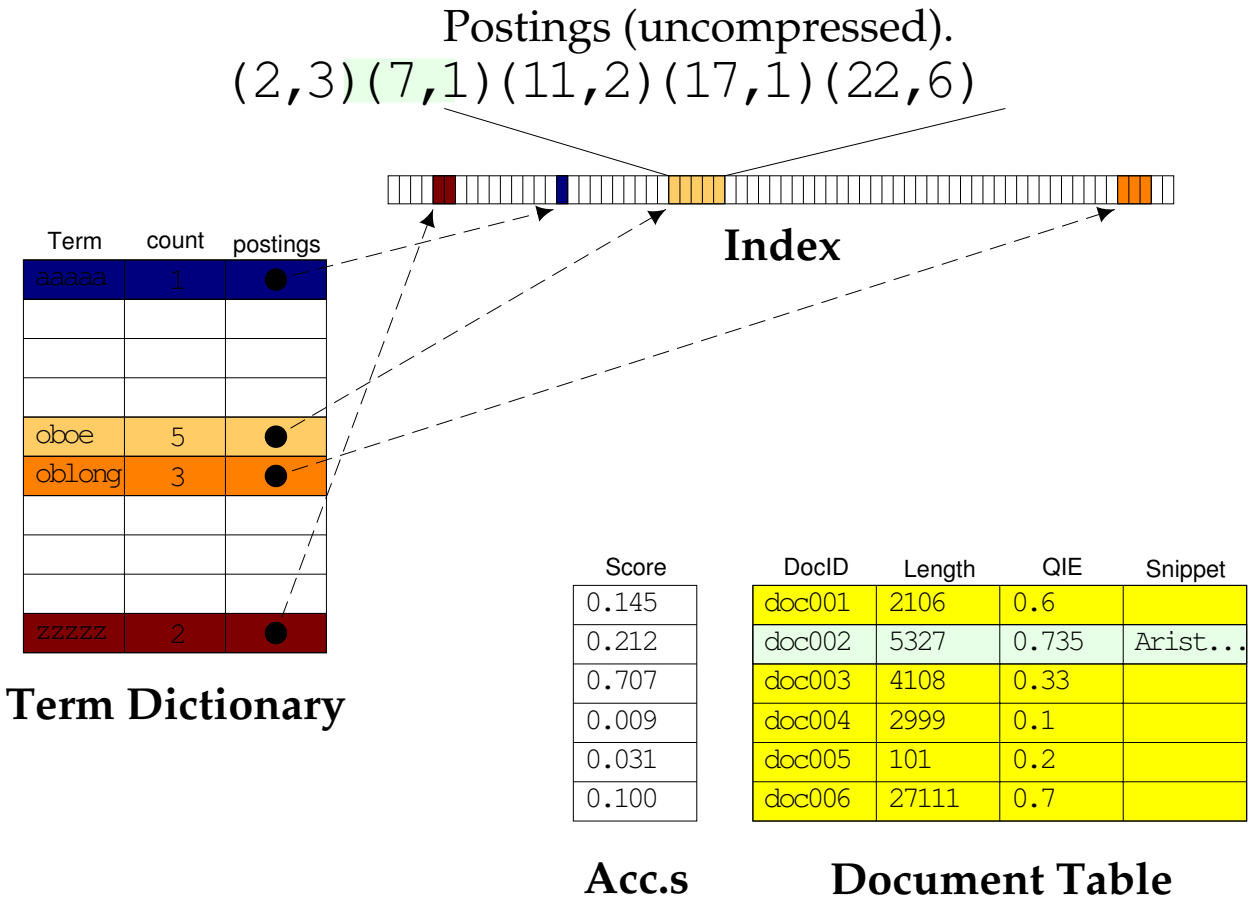
BM25F - Extension to fields

- Weight term frequencies prior to non-linear combination in BM25.
- Robertson, Zaragoza & Taylor, CIKM 2004

Other Retrieval Models

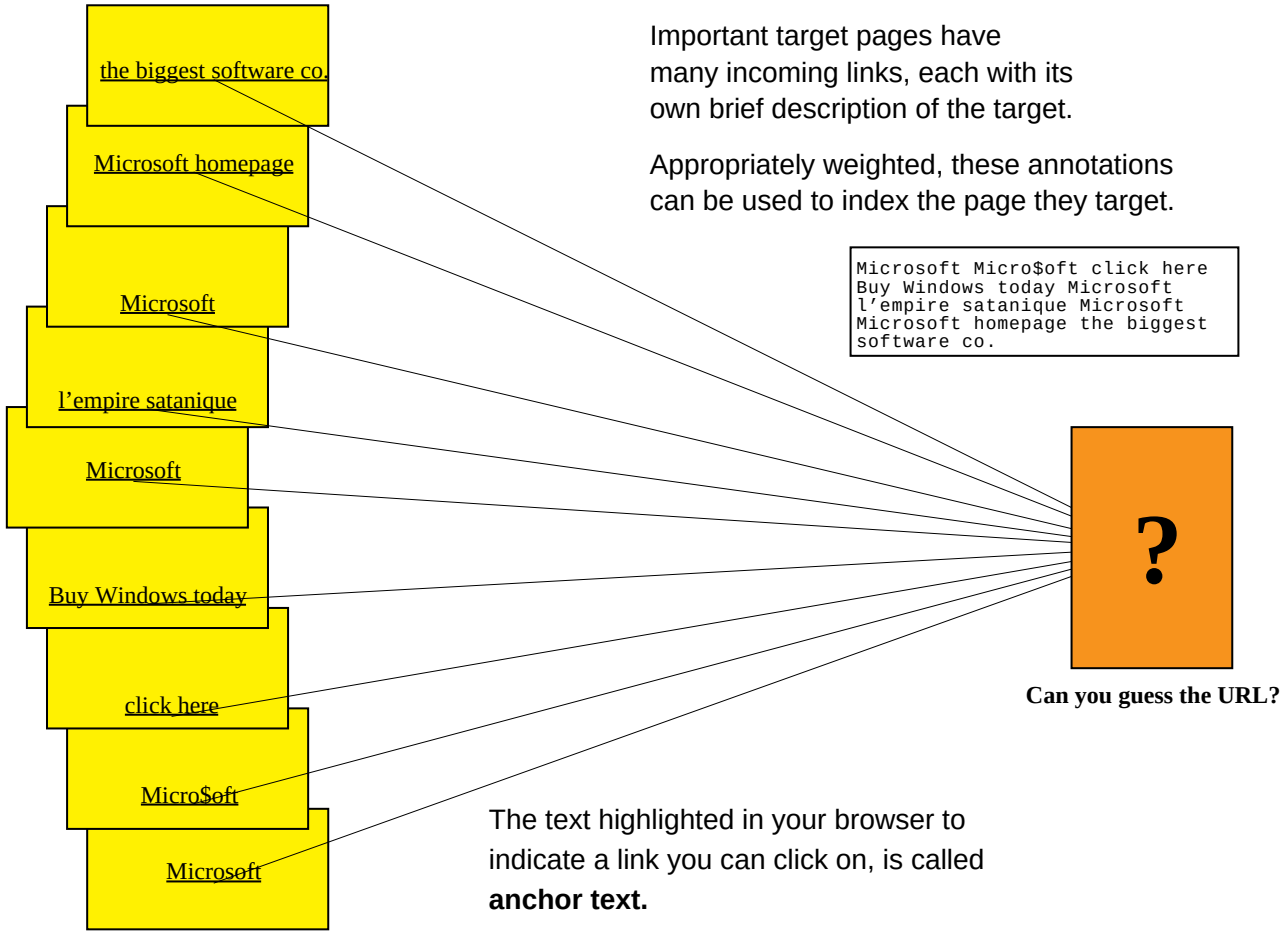
- Vector Space
- Language Models
- Divergence from Randomness (parameter free!)

Using an inverted file to generate dynamic scores

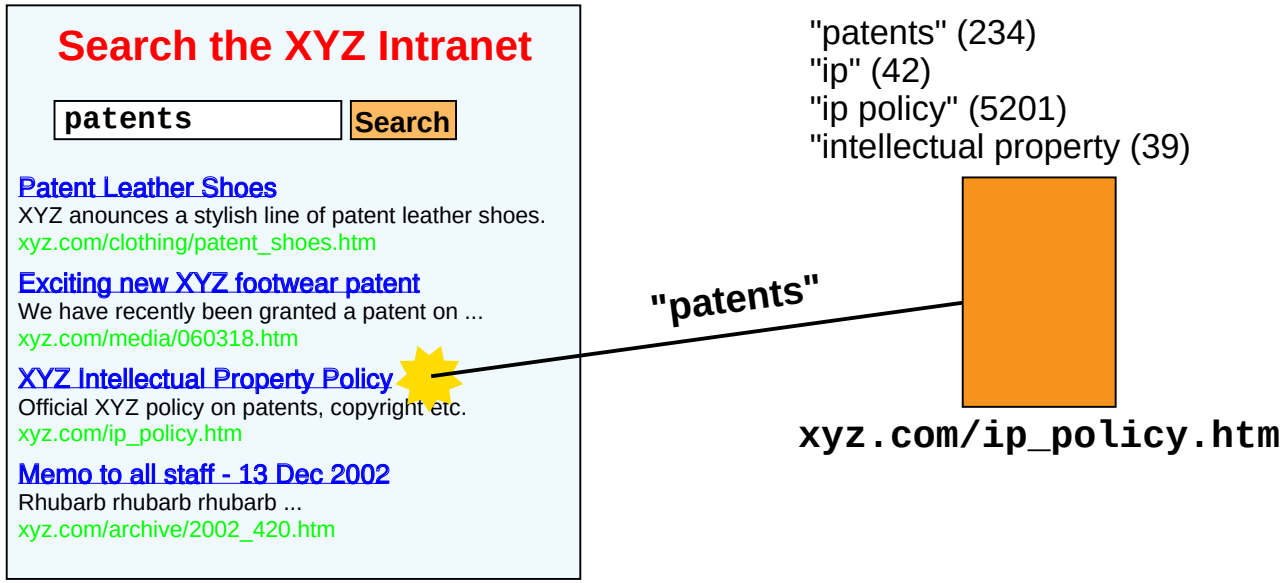


External Textual Evidence

Anchor text

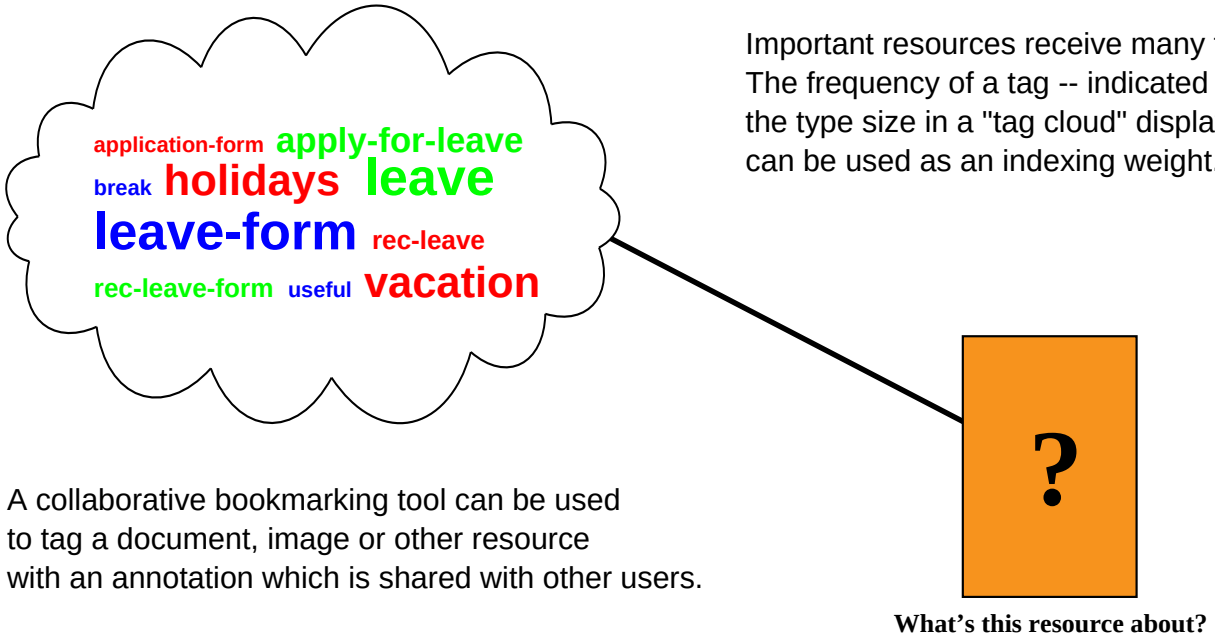

the biggest software co.
Microsoft homepage
Microsoft
l’empire satanique
Microsoft
Buy Windows today
click here
Micro$oft
Microsoft
Important target pages have many incoming links, each with its own brief description of the target.
Appropriately weighted, these annotations can be used to index the page they target.
Microsoft Micro$oft click here Buy Windows today Microsoft l’empire satanique Microsoft Microsoft homepage the biggest software co.
?
Can you guess the URL?
The text highlighted in your browser to indicate a link you can click on, is called **anchor text.**

Click-associated queries



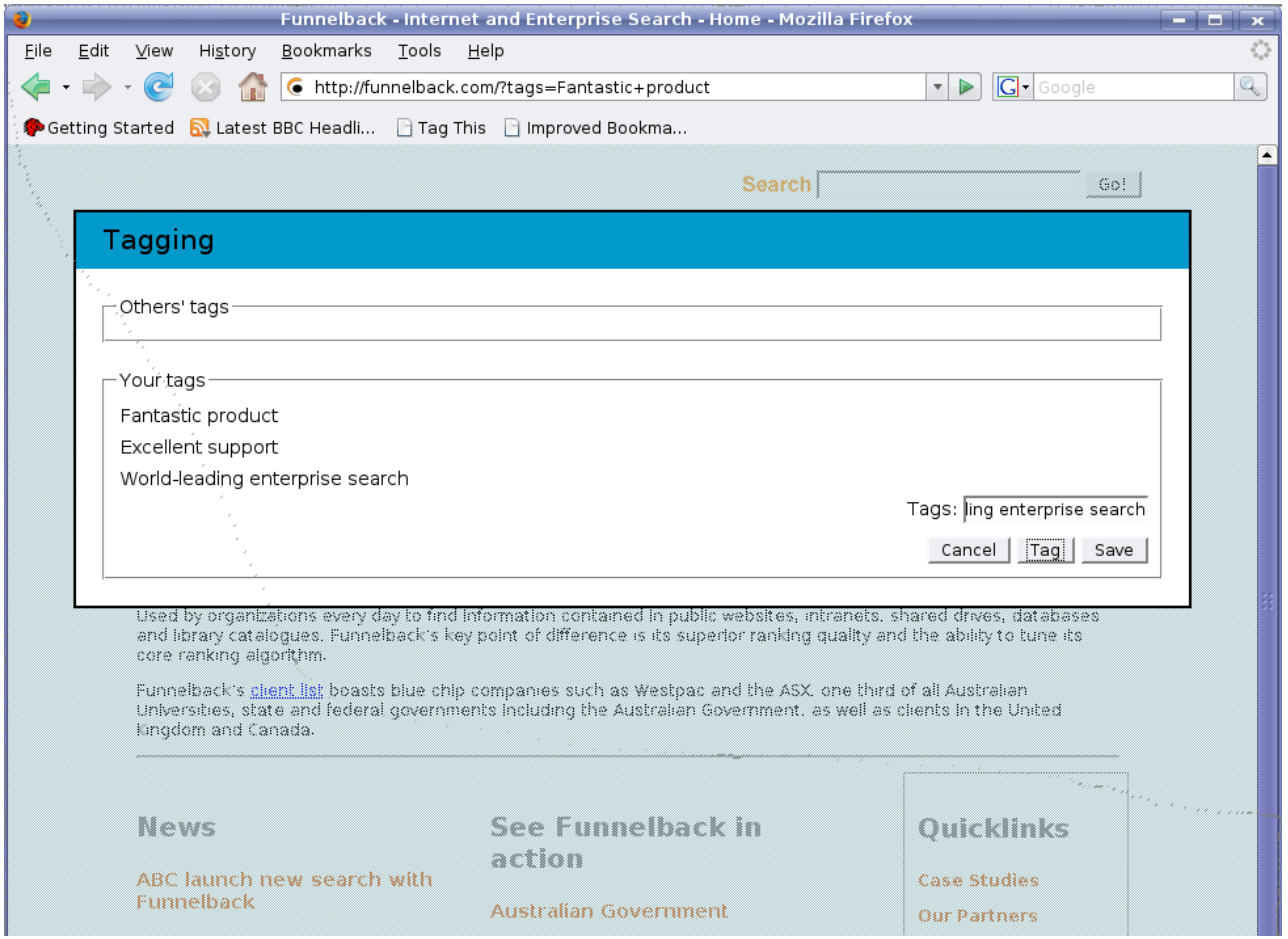
When a searcher enters a query and clicks on a document, we can associate that query with the document. Associated queries can be weighted by click frequency and used in indexing and retrieval.

Folksonomy tags



- See e.g. Dubinko et al, WWW 2006

Collecting tags



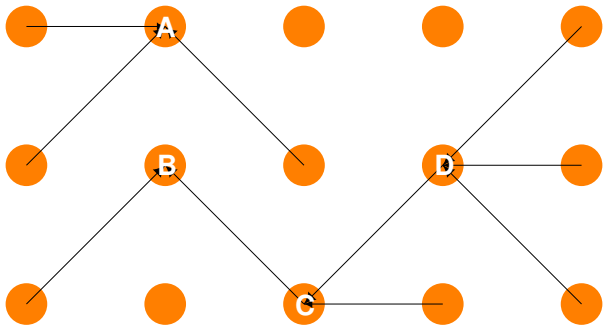
Should these external texts be treated as document fields?

Static factors

Adapted from Richardson, Prakash and Brill, WWW 2006

- Incoming hyperlinks
  - e.g. raw count, PageRank, Kleinberg Hub/Authority
- Searcher behaviour data
  - e.g. Frequency of visits to page (from toolbars, or proxy logs); Frequency of clicks when this page appears in search results; Average dwell time on the page
- Query independent use of anchor text
  - Amount of referring anchor text; Size of anchor text vocabulary
- Page properties
  - e.g. Word count; Frequency of most common term;
- URL properties (Kraiij & Westerveld, SIGIR 2002)
  - e.g. Length, depth of URL; type (root, subroot, page, dynamic)
- Domain properties
  - e.g. Average outlink count for pages in this domain.
- Spam rating.
  - e.g. Presence of AdSense ads!
- Adult content score.

PageRank

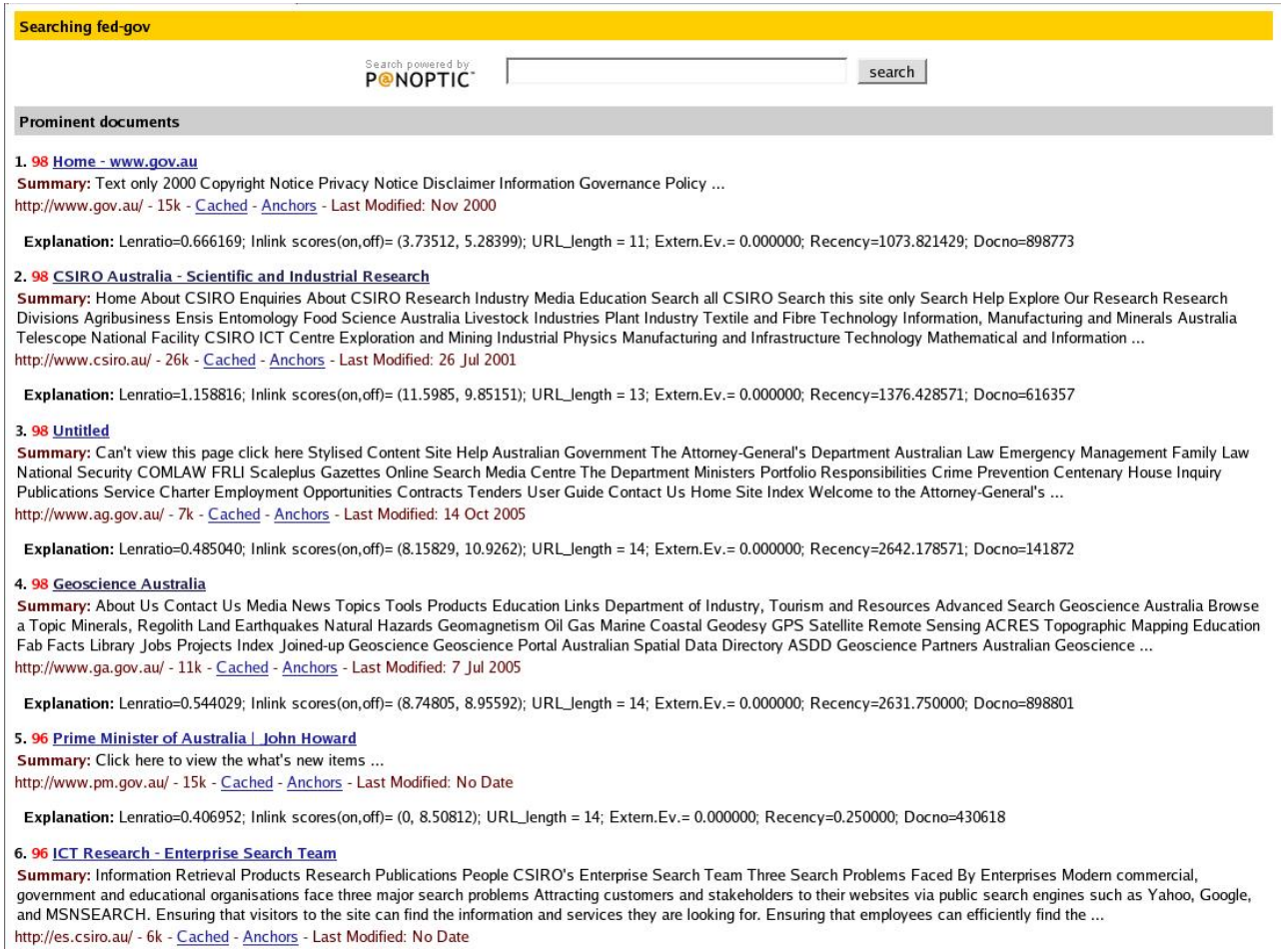


Initial PR value for all 15 nodes: 1/15

After convergence, which of A,B,C,D has highest PR?

- random surfer
- start with equal probability for all bookmarked pages
- follow outgoing links with equal probability
- teleport to a bookmark with probability d

Query independent evidence in the Australian Government



Searching fed-gov

Search powered by POPTIC search

Prominent documents

1. 98 Home - www.gov.au
Summary: Text only 2000 Copyright Notice Privacy Notice Disclaimer Information Governance Policy ...
http://www.gov.au/ - 15k - Cached - Anchors - Last Modified: Nov 2000

Explanation: Lenratio=0.666169; Inlink scores(on,off)= (3.73512, 5.28399); URL_length = 11; Extern.Ev.= 0.000000; Recency=1073.821429; Docno=898773

2. 98 CSIRO Australia - Scientific and Industrial Research
Summary: Home About CSIRO Enquiries About CSIRO Research Industry Media Education Search all CSIRO Search this site only Search Help Explore Our Research Research Divisions Agribusiness Ensis Entomology Food Science Australia Livestock Industries Plant Industry Textile and Fibre Technology Information, Manufacturing and Minerals Australia Telescope National Facility CSIRO ICT Centre Exploration and Mining Industrial Physics Manufacturing and Infrastructure Technology Mathematical and Information ...
http://www.csiro.au/ - 26k - Cached - Anchors - Last Modified: 26 Jul 2001

Explanation: Lenratio=1.158816; Inlink scores(on,off)= (11.5985, 9.85151); URL_length = 13; Extern.Ev.= 0.000000; Recency=1376.428571; Docno=616357

3. 98 Untitled
Summary: Can't view this page click here Stylised Content Site Help Australian Government The Attorney-General's Department Australian Law Emergency Management Family Law National Security COMLAW FRLI Scaleplus Gazettes Online Search Media Centre The Department Ministers Portfolio Responsibilities Crime Prevention Centenary House Inquiry Publications Service Charter Employment Opportunities Contracts Tenders User Guide Contact Us Home Site Index Welcome to the Attorney-General's ...
http://www.ag.gov.au/ - 7k - Cached - Anchors - Last Modified: 14 Oct 2005

Explanation: Lenratio=0.485040; Inlink scores(on,off)= (8.15829, 10.9262); URL_length = 14; Extern.Ev.= 0.000000; Recency=2642.178571; Docno=141872

4. 98 Geoscience Australia
Summary: About Us Contact Us Media News Topics Tools Products Education Links Department of Industry, Tourism and Resources Advanced Search Geoscience Australia Browse a Topic Minerals, Regolith Land Earthquakes Natural Hazards Geomagnetism Oil Gas Marine Coastal Geodesy GPS Satellite Remote Sensing ACRES Topographic Mapping Education Fab Facts Library Jobs Projects Index Joined-up Geoscience Geoscience Portal Australian Spatial Data Directory ASDD Geoscience Partners Australian Geoscience ...
http://www.ga.gov.au/ - 11k - Cached - Anchors - Last Modified: 7 Jul 2005

Explanation: Lenratio=0.544029; Inlink scores(on,off)= (8.74805, 8.95592); URL_length = 14; Extern.Ev.= 0.000000; Recency=2631.750000; Docno=898801

5. 96 Prime Minister of Australia | John Howard
Summary: Click here to view the what's new items ...
http://www.pm.gov.au/ - 15k - Cached - Anchors - Last Modified: No Date

Explanation: Lenratio=0.406952; Inlink scores(on,off)= (0, 8.50812); URL_length = 14; Extern.Ev.= 0.000000; Recency=0.250000; Docno=430618

6. 96 ICT Research - Enterprise Search Team
Summary: Information Retrieval Products Research Publications People CSIRO's Enterprise Search Team Three Search Problems Faced By Enterprises Modern commercial, government and educational organisations face three major search problems Attracting customers and stakeholders to their websites via public search engines such as Yahoo, Google, and MSNSEARCH. Ensuring that visitors to the site can find the information and services they are looking for. Ensuring that employees can efficiently find the ...
http://es.csiro.au/ - 6k - Cached - Anchors - Last Modified: No Date

There's no query!

Machine-Learning the Overall Ranking Function

$$RSV = \alpha_0 D_o + \ldots + \alpha_n D_n + \beta_0 S_0 + \ldots + \beta_n S_n \qquad (12)$$

- We need to be able to compute ranking quality for gezillions of combinations of the αs and βs.
- Ranking quality is highly dependent upon the query so at each point we need to run very large numbers of queries and measure the quality of results.

Thoughts on a loss function

(Except for nerds like me, people don't actually search for the fun of it. They do it in order to complete a task.)

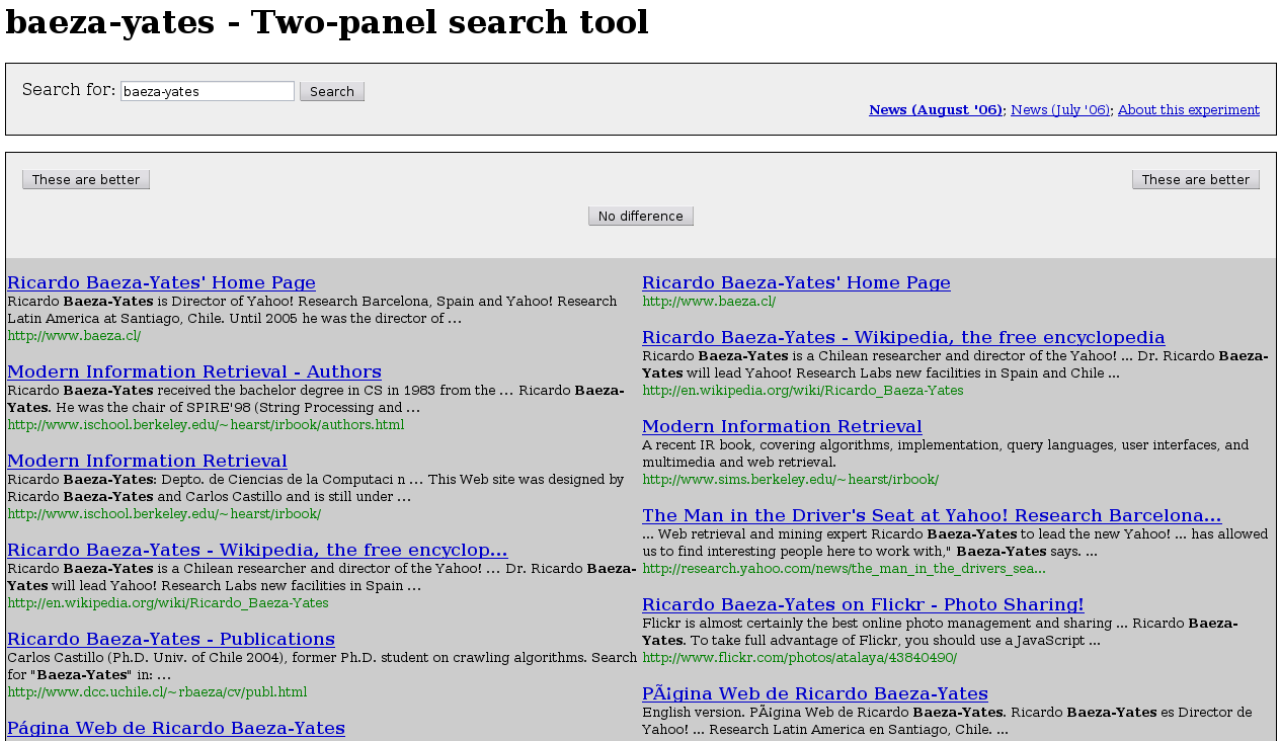
- What we really want to optimise:
  - Proportion of search-facilitated tasks that people complete
  - How satisfactorily they complete them
  - How fast they complete them
- That's very difficult. How can we do it?

User Studies

- Bring large numbers of human subjects into a laboratory and ask them to do search tasks.
- Measure their task performances.
- But:
  - Expensive
  - Not a real task – do the subjects do it properly?
  - Huge sources of variance to be controlled
    - individual differences
    - order effects
    - interactions
  - Results are set level – not reusable

In-Situ Studies

- Ask representative human subjects to use a two-panel search tool instead of their normal search engine.
- Controls for many of the problems
- Still not re-usable
- Explicit or implicit judgments.



- Results are still set level – not reusable

Observing natural user behaviour

- Via search engine or browser logs
- Where do people click?
  - Trust bias
  - Interpreting **no-click**
  - Increased frequency of clicks before and after page boundaries and "the fold"
- Can get preference judgments:
  - If someone skips over $\mathcal{D}_n$ and clicks on $\mathcal{D}_{n+1}$ we have evidence that they prefer $\mathcal{D}_{n+1}$ to $\mathcal{D}D_n$ for
- That could be input into a machine learning system.

Manipulating Rankings

- Reordering results
- Interleaving results
- Inserting results
- Observe behavioural differences
  - Flights and Buckets
- GYB do lots of this.

TREC

- Cranfield? TREC? Huh?

```
<top>

<num> Number: 151

<title> Topic: Coping with overcrowded prisons

<desc> Description:

The document will provide information on jail and
prison overcrowding and how inmates are forced to
cope with those conditions; or it will reveal plans
to relieve the overcrowded condition.

<narr> Narrative:

A relevant document will describe scenes of overcrowding
that have become all too common in jails and prisons
around the country. The document will identify how
inmates are forced to cope with those overcrowded
conditions, and/or what the Correctional System is
doing, or planning to do, to alleviate the crowded
condition.

</top>
```

Cranfield / TREC Style Judging

- Employ judges to assign relevance / utility scores to all documents (or for a large pool of documents which might possibly be relevant to the query).
- TREC pools– Union of top 100 docs for participating systems
- Results in re-usable test sets, modulo:
  - completeness
  - judging errors and disagreements
- TREC studies of stability of rankings across strict/lenient judging
- GYB have large budgets for this.
  - Bing: 5 point scale, Gains are 2^n

Issues with TREC style test sets

- Of what population are the TREC topics a representative sample?
- No penalty for duplicates – they're very common
- No reward for diversification
- Solution: es.csiro.au/C-TEST
  - Interpretations
  - Differential utilities
  - Equivalence sets

C-TEST: CSIRO TOOLKIT FOR THE EVALUATION OF SEARCH TECHNOLOGY

David Hawking and Tom Rowlands

(draft four; not yet complete)

C-TEST implements a simple but generalised test-collection framework for evaluating the effectiveness of search systems. It is designed primarily for the evaluation of web and enterprise web search but is more generally applicable. It comprises:

- two XML file formats specifying:
  - queries, interpretations, weighted answers
  - results of a `run'
- a collection of tools to evaluate based on various metrics, sample from existing files, pool results for evaluation and use a web based tool for evaluation
- a library, written in Perl, to work with the files from your own tools

It is capable of appropriately evaluating the Homepage Finding, Named Page Finding and Topic Distillation tasks from the TREC Web Track. The framework effectively penalises the return of duplicate results and has a prominent recognition of the one query text having many potential, variably-weighted, interpretations.

C-TEST is distributed free under the Mozilla Public Licence. CSIRO welcome your suggestions, criticism and patches. We do, however, ask that if you are to use C-TEST in the development of an academic publication, you cite our C-TEST paper (BiBTeX):

> David Hawking, Tom Rowlands, Paul Thomas. *C-TEST: Supporting novelty and diversity in testfiles for search evaluation. Proceedings of the SIGIR workshop on redundancy, diversity and interdependent document relevance, Boston 2009.*

REQUIREMENTS AND INSTALLATION

C-TEST is tested on Ubuntu, Mac OS X and Windows XP systems. It almost certainly runs without modification on other systems as well. Packages that are required are Perl, LibXML::XML, Statistics::Distributions and File::Slurp. Directions for installing the latter two are included in the download.

DOWNLOAD

C-TEST pre-release **Not for re-distribution (yet).**

DTDS

C-TEST Example

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<testfile name="Airline␣Homepage␣Finding">
  <query id="1" text="Qantas␣airways" weight="1.0" depth
      ="10">
    <interpretation comment="homepage␣finding" weight="
        0.9">
      <eset util="10" comment="qantas␣homepage">
        <docid>qantas.com
        </docid>
        <docid>www.qantas.com.au
        </docid>
        <docid>www.qantas.com/index.html
        </docid>
      </eset>
    </interpretation>
    <interpretation comment="share␣price" weight="0.1">
      <eset util="10" comment="qantas␣stock␣exchange␣
          listing">
        <docid>asx.com.au/companylisting=QAN
        </docid>
      </eset>
      <eset util="3" comment="QANTAS'␣share␣information␣
          page">
        <docid>qantas.com.au/info/about/investors/
            shareholderInfo
        </docid>
      </eset>
    </interpretation>
  </query>
```

C-TEST Example Outfile

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<results label="Airways␣Homepage␣Finding,␣run␣with␣alpha
    =0,␣beta=2">
  <query id="1" text="Qantas␣airways" weight="1.0" depth
      ="10">
    <docid rank="1">www.qantas.com.au/</docid>
    <docid rank="2">www.qantas.com.au/regions/dyn/</
        docid>
    <docid rank="3">www.anzac.com/qantas/qantas.htm</
        docid>
    <docid rank="4">www.qantas.com/index.html</docid>
    <docid rank="5">www.quantas.com.au/</docid>
    <docid rank="6">en.wikipedia.org/wiki/Qantas</docid>
    <docid rank="7">www.airlinequality.com/Forum/qantas.
        htm</docid>
    <docid rank="8">www.travelmood.com/qantas.asp</docid>

    <docid rank="9">www.oneworld.com/ow/member-airlines/
        qantas</docid>
    <docid rank="10">www.webjet.com.au/airlines/qantas.
        htm</docid>
  </query>
</results>
```

C-TEST: Tools for

- Creating testfiles
  - From a spreadsheet
  - From TREC topics
  - By searching and browsing
  - By sampling a query log and judging
- Computing measures and significance testing of differences

LSE Case Study

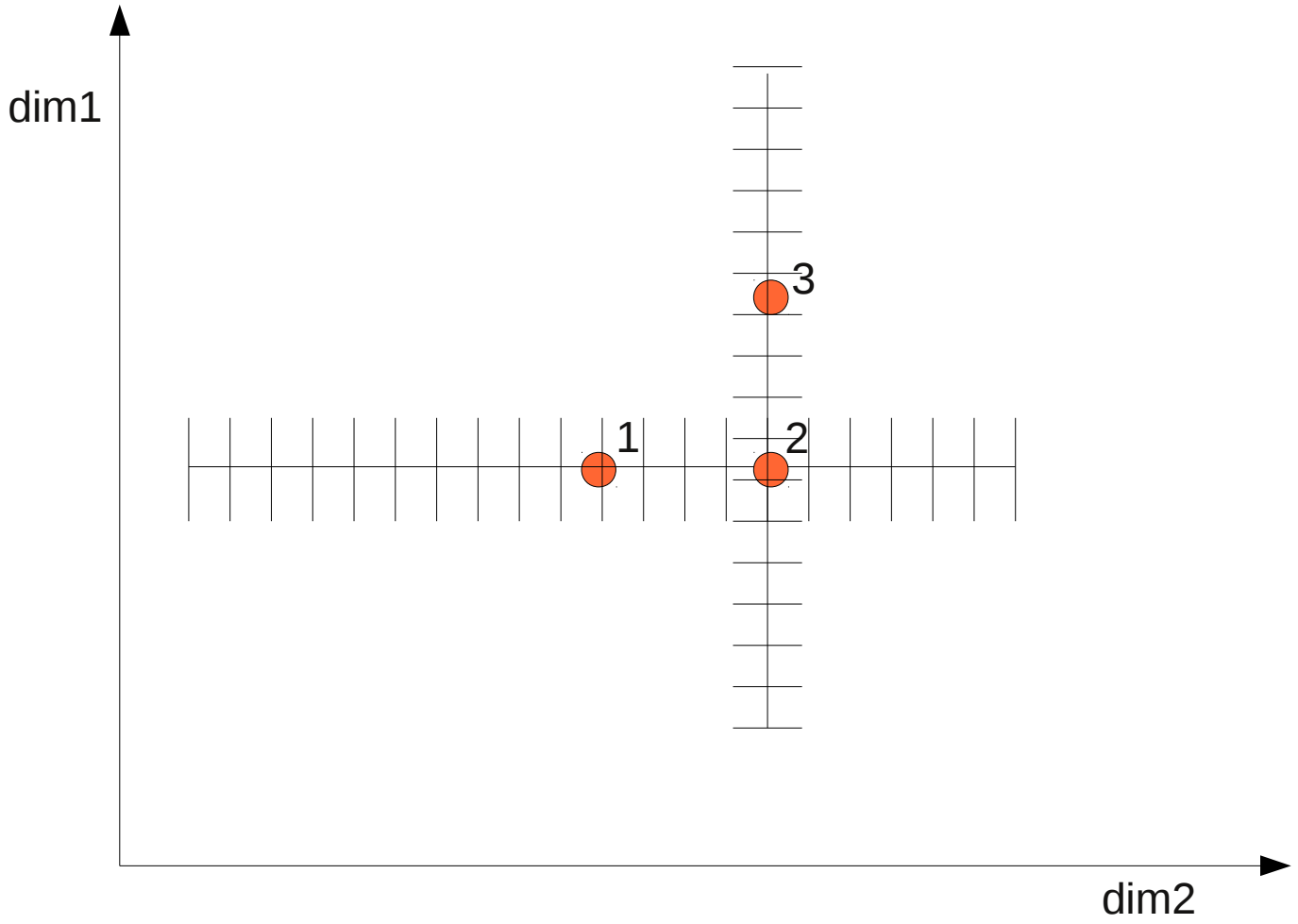
Sources of testfiles at LSE

- A-Z Sitemap (¿500 entries)
  - Biased toward anchortext
- Keymatches file (¿500 entries)
  - Pessimistic
- Click data (¿250 queries with ¿t clicks)
  - Biased toward clicks - can achieve 100% success
- Random sample of workload, post-judged
- Popular/Critical queries (134 manually judged)
  - Optimising for searchers or for publishers

Tuning problem

- Approximately 40 parameters, some continuous, some binary, some integer
- Not much idea about the shape of the function.
  - Pretty sure that there are multiple points of inflection.
- Some combinations make no sense
- Obviously brute-force grid search is impossible
- Even so, millions of query executions are needed.

Dimension at a time tuning



Where have we got with our tuning run?

LSE Tuning results (failure rates)

- Out-of-the-box: 24.63%
- As configured: 22.39%
- After tuning (DAAT mode): 8.21%

On the flipside of coffee ...

- Web SearchEngineering
- Field Work: how do Web search engines really work?
- *Stretch Break*
- Discussion: Other IR problems for machine learning
- Historical context