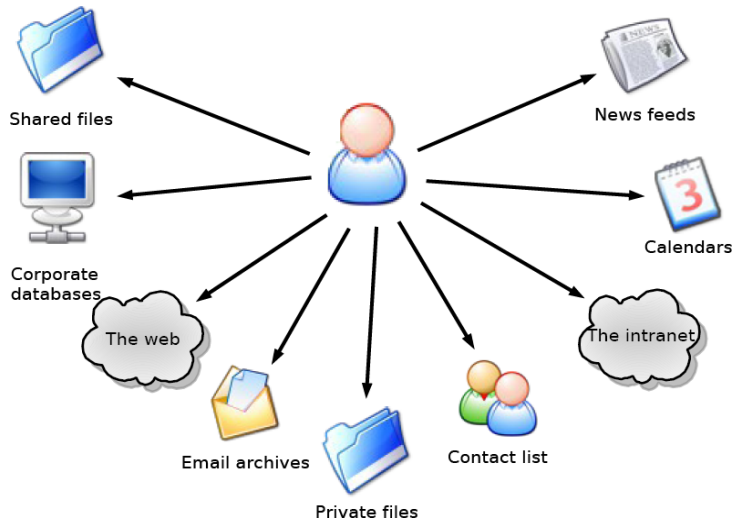


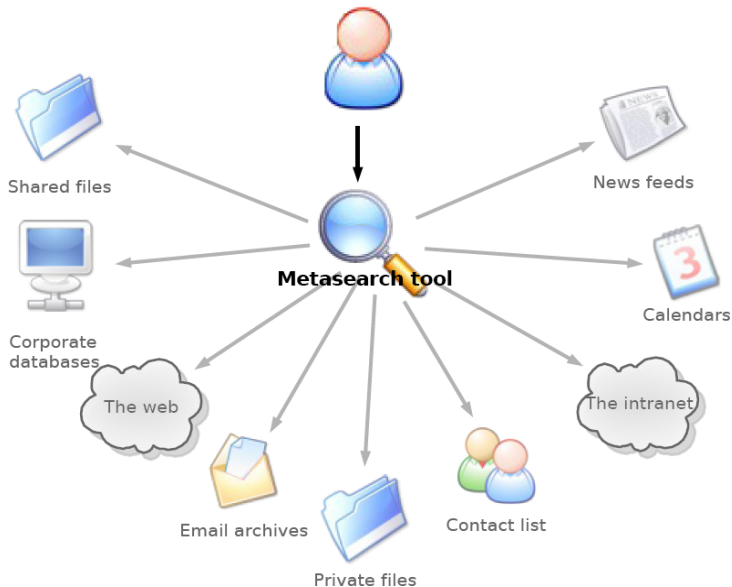
Evaluating information retrieval in context

Paul Thomas (Australian National University)
David Hawking (CSIRO ICT Centre)

Personal metasearch



Personal metasearch



Personal metasearch

Personal metasearch is hard to evaluate with current techniques:

- ▶ **Interactive** systems and multi-stage retrieval
- ▶ **Adaptive** and personalised systems
- ▶ **Private** or semi-private corpora

Many other realistic IR scenarios are also hard.

Approaches to evaluating IR systems

1. Test collections
2. Search log analysis
3. Experimentation in the lab
4. Naturalistic observation

Test collections

Used at Cranfield; now used by TREC, CLEF, INEX, NTCIR,
...

Has a corpus; a set of information needs; and “complete”
relevance judgements over (corpus \times needs).

- ▶ Cheap to run, once it's set up
- ▶ Experiments are reproducible
- ▶ The collection is reusable
- ▶ Produces robust comparisons

Test collections

But how can we handle. . .

- ▶ . . . personal corpora?
- ▶ . . . fast-changing corpora?
- ▶ . . . big corpora?
- ▶ . . . diverse and unarticulated information needs?
- ▶ . . . set-based, contextual judgements?

Search log analysis

Examine log files, and consider a click on a result as an indication of utility.

Natural queries and “judgements” with no burden on experimenters or users.

- ▶ Need to account for trust and quality biases.
- ▶ Most queries have no associated clicks.
- ▶ Can't distinguish brilliant success from abject failure.
- ▶ Logs are maintained by separate systems for separate user populations.

Experimentation in the lab

Observe test users in a laboratory setting, with simulated information needs; can also use questionnaires and post-hoc judgements for more feedback.

Design lets us control for searcher, topic differences; and also for presentation order.

- ▶ Experimental design can be complex.
- ▶ Uses fixed corpora and artificial information needs.

Naturalistic observation

Place an experimenter in the field to observe day-to-day information seeking behaviour.

- ▶ Expensive!
- ▶ Risk of experimenter effect.

Alternatively, use instrumented PCs or search software to do the “observation” for us.

Embedded comparisons

Provide two (or more) panels, each presenting a different IR system, in a single search interface. Log interactions with each panel, and optionally ask for explicit judgements of preference.

Uses real corpora; real information needs; and online set-based judgements.

Experiments are cheap and easy to run.

Embedded comparisons

metasearch - Two-panel search tool

Search for:

[News \(August '06\)](#); [News \(July '06\)](#); [About this experiment](#)

<p>Metasearch.com - The Original & Best Since 1995! The original way to search the search engines. Images, Video, and Audio from ... 10 20 40 100 hits. Copyright © 1995-2004 - Scott Banister ... http://metasearch.com/</p>	<p>Metasearch.com - The Original & Best Since 1995! The original way to search the search engines. Images, Video, and Audio from Google, Yahoo, AltaVista, ... http://metasearch.com/</p>
<p>Search AllinOne MetaSearch feature: PanamaLaw.org - Panam... Meta search engine which displays search results from multiple search engines, organized by relevancy. http://www.searchallinone.com/</p>	<p>Mamma Metasearch - The Mother of All Search Engines Metasearch tool for finding web pages, news, pictures or MP3s. http://www.mamma.com/</p>
<p>Metasearch.biz Metasearch.biz The Fast Meta Search. ... Web. Directory. Make Metasearch Your Homepage Link To Us. Contact Us Privacy Policy ... Make Metasearch Your Homepage ... http://www.metasearch.biz/</p>	<p>Ixquick Metasearch Ixquick submits your search to the major search engines and finds sites that are universally ranked in the top ten! http://www.ixquick.com/</p>
<p>CNET Search.com Searches dozens of leading search engines to bring you the best results</p>	<p>Starting Point Welcome to Starting Point! ... Directory Add a Site. © 2006 Stpt.com. http://www.stpt.com/</p>

Embedded comparisons

But...

- ▶ There's still a risk of experimenter effect.
- ▶ Experiments aren't repeatable in some ways we'd like.
- ▶ Hard to say **why** something is better.

Experiment 1 — Popular queries

Popular queries;

Compared results from Google ranks 1–10 and 21–30.

Clickthrough attributes were a good predictor of final vote.

Significant preference for ranks 1–10, no significant preference for either side.

Experiment 2 — Natural queries

Natural queries;

Compared results from Google ranks 1–10 and 21–30.

Clickthrough attributes were a good predictor of final vote.

Significant preference for ranks 1–10, no significant preference for either side.

Experiment 3 — Overlapping result sets

Natural queries;

Compared results from Google ranks 1–10 and 6–15.

Clickthrough attributes were a good predictor of final vote.

Significant preference for ranks 1–10, no significant preference for either side.

Experiment 4 — Implicit feedback

Natural queries, **no explicit judgement of result sets**;

Compared results from Google ranks 1–10 and 21–30.

Clickthrough attributes were a good predictor of **quality**.

Significant preference for ranks 1–10, no significant preference for either side.

A case study

Compared search engines “A” and “B” (via their public APIs) for whole-of-web search.

49 users had no significant overall preference.

(Still no left-right bias; clickthrough attributes still a good predictor of final judgement.)

Records user satisfaction directly, not the proxy of e.g. a precision score.

Continuing work

- ▶ Personal metasearch;
- ▶ External results in intranet search;
- ▶ Branding.

Other applications

- ▶ Merging vs segmenting vs clustering vs ... ;
- ▶ Source selection;
- ▶ Personalisation;
- ▶ Query-biased summaries;
- ▶ Result ranking;
- ▶ ...

Embedded comparisons

There are some IR scenarios in which standard evaluation techniques don't work well.

Direct side-by-side comparison of two working systems is an alternative;

It works;

There are many applications.