

1. who I am; what this is (a quick overview of what personal MS is and why we should care; a closer look at a couple of problems and some work I've done on these problems); and a quick look at some software which may shed more light on the area

Personal metasearch

└ Why should we care?



1. these are some of the information sources I use from day to day — only some! Also have: ANU library, other libraries, book reviews and bookshops, newspapers and similar, ACM digital library, etc etc etc

Personal metasearch

└ Why should we care?



1. so to look for X I have to use Y, which has one syntax; for X' I have to use Y', which has another; etc etc
2. so what I have to do for any bit of information is: (1) decide where it's most likely to be; (2) start the appropriate tool, go to the right web page, or whatever; (3) convert my information need into the appropriate query; (4) search through the results. If I can't find it, I need to refine or re-word my query. If I think I'm looking in the wrong place, I start again: decide where else it might be, start another tool, etc. If I miss a possible source, too bad
3. would be nice if I didn't have to decide where to look. Would be nice if I didn't have to remember a different query language or technique for each place something might be. Would be nice if, when I thought I knew where something was, I didn't run the risk of missing something else; maybe even something I never knew existed

Personal metasearch

└ Compared with "desktop search" (1)

Why is this different from existing "desktop search" tools?



1. describe current desktop search tools and what they provide

Personal metasearch

└ Compared with "desktop search" (2)

- Must cover a **diverse range of data**.
- Needs to handle **access restrictions** on data.
- Must handle a mixture of **structured, semi-structured, and unstructured data**.
- Some data is **rapidly changing** or otherwise hard to index.
- Can make use of **knowledge about the user** to inform the search and/or the presentation of results.
- Not clear how to measure success; traditional metrics like *P*, *R*, *SIR* don't always apply.

1. diverse range: must consider attributes as well as content: reliability, specificity, ...
2. ... must be able to make decisions about ranking and presentation from very different sources
3. ... current desktop search tools only consider local files plus maybe the web; in particular, they don't search subscription sites, don't distinguish local websites from any others, etc etc
4. access restrictions: for most search engines, everything they cover is public; not the case here since we have eg subscription sites, file shares, etc
5. ... results for the same query could differ from user to user
6. ... tuning algorithms based on one user may not improve things for another
7. ... evaluation is more complicated
8. structured, semi-structured, unstructured: need to do this if we consider e.g. corporate databases at the same time as email
9. knowledge: physical location, past preferences, languages, organisational structure, friends and colleagues, ...
10. success: personal metasearch is interactive; may be an aid to navigation rather than replacing navigation; etc.
11. ... what do I mean by an aid to navigation? Orienteering; getting close enough with a tool then searching "manually" or looking for landmarks. We often evaluate IR as if it were a one-shot process where the goal is to get straight to the best document in one step — this might not be how people would work even if they could

2006-08-01

Personal metasearch

└ Compared with personal information management

- Like PIM tools, we are interested in integrating different sources of information;
- but we're focusing on **search**;
- and considering not just known but **unknown items**.

1. search: not, for example, organising information; connecting information; capturing information in the first place
2. unknown items: eg the new version of a policy doc which is on the intranet but which we didn't know about; or news about a company we're interested in
3. not interested in building one tool to do everything, in the manner of MyLifeBits or Haystack, but this work could well inform the search capabilities of such tools

2006-08-01

Personal metasearch

└ Some outstanding problems

We're focusing on **things that are different**:

- ▶ Discovering and selecting very different data sources;
- ▶ Indexing and searching very different data sources and types;
- ▶ Combining results;
- ▶ Designing a user interface;
- ▶ Making use of personal or task context;
- ▶ Measuring success or failure;
- ▶ ...

things that are different in the PMS case from the "normal", better-studied case

If we're going to build a tool, it'd be nice to know how well it works.

Standard IR measures (P , R , $50n$, etc) may not be appropriate, and even if they are they're not feasible to gather.

Problems include:

- ▶ Collecting a suitable corpus;
- ▶ Accounting for interactive searching;
- ▶ Accounting for context;
- ▶ Getting realistic tasks;
- ▶ Deciding on a measure, or measures.

Personal metasearch

└ Measuring success or failure

2006-08-01

1. problems include: problems with the standard measures . . .
2. collecting a corpus: it should really include data which we can't make public (or possibly even see ourselves), such as personal email archives and corporate databases. Scale may well be an issue too — there is a LOT of stuff I have access to
3. interactive: I don't see this as a problem in offline retrieval for later analysis, as (say) TREC ad-hoc and related tasks, but as a highly interactive process. Are there measures which are easy to calculate, and easy to repeat, but which still capture what's useful for an interactive process?
4. accounting: searches over personal and corporate information stores will presumably be loaded with contextual constraints and cues. How can we account for something as personal and ephemeral as search context in an evaluation? How can we repeat experiments?
5. tasks: what sort of things do people really look for? We don't have the equivalent of web search engine query logs for e.g. email search, desktop search. Should we even be worrying about this? Maybe we should be asking what people would be looking for if they weren't constrained by their current tools.
6. end: so we need some way to account for these problems when we try to evaluate a personal IR tool

└ Another evaluation scheme (1)

Use a real search tool instead of TREC-style batch evaluations.

Search users' real information sources.

Gather real queries and real judgments.

But! not repeatable in the ways we'd like.

1. real search tool: something that actually works, at least well enough that people might use it and contribute to our experiments in so doing
2. real information sources: even if they're possibly private (like email) or otherwise irreproducible. Means giving up on the idea of building a public corpus other researchers can use — even means giving up on building a corpus we can use ourselves for later experiments
3. real queries: not the made-up queries that dominate TREC. We will however wind up capturing queries that match people's mental model of the search system, its syntax and capabilities; not real descriptions of information needs
4. get several things for free: (1) this accounts for context and interactivity; (2) this has realistic tasks; (3) this has realistic corpora
5. but! can't repeat experiments, at least not in the same way as before — can't change a formula and re-run to get answers straight away. Much harder to compare with work from other researchers. Can't make much of the tests public, but maybe instead of sharing a testbed we can share a methodology

Personal metasearch

└ Another evaluation scheme (2)

Another evaluation scheme (2)

Present two sets of results side-by-side and record which set of results people find better.

Do it for long enough with enough users and the results should be fairly reliable.



1. present two sets: lets us compare presentation issues such as combining results, other presentation issues, source selection, etc etc

Personal metasearch

└ Another evaluation scheme (3)

Experiments to validate the scheme:

- Users are able to distinguish between high- and low-quality result sets;
- Users can still distinguish between high- and fairly high-quality result sets, with overlap;
- Clicks on results are a good predictor of overall preference;
- We can still obtain preference information without using any explicit feedback.

1. distinguish 1: google 1-10, 21-30
2. distinguish 1: google 1-10, 15-24
3. clicks: a variety of predictors (first click, last click, etc) each get around 80% right
4. still obtain preference: so there's no extra overhead for test users

Personal metasearch

└ Server selection (1)

Given a query and a set of servers which may include relevant documents, which servers should we forward the query to?

Well-studied, but not in environments like that of our personal metasearcher. Do standard techniques (CORI, KL divergence, CVV, RdDDE, etc) still work?

not in environments: VERY different sources; eg personal calendar vs the whole web

└ Server selection (2)

Experiments with up to six "servers": personal email archives, two mailing lists, calendar, news website, and the entire public web:

CORI (Callan, Zu, & Croft) seems to be thrown by difference in vocabulary:

KL divergence (Xu & Croft) does fairly well;

CVV (Yusono & Lee) is thrown in the same way as CORI;

REDDE (Si & Callan) is thrown by small samples of very large servers;

Stemming and stopping makes little difference.

So how can we choose between these vastly different servers?

1. tried variants of these, both from the literature and from my own analysis, which attempt to compensate for big differences in server size and sample size – but we have 7 orders of magnitude difference here!
2. cori: vocab in email is much more focused, fewer words repeated more often
3. kl: doesn't often get it right first time, but best server is typically ranked in the top two
4. redde: weights at one point by est size / sample size; for e.g. the web, this is huge and completely dominates

Personal metasearch

└─What's really out there?

What's really out there?

We are planning a survey of "volume searchers" (librarians, lawyers, researchers, . . .) to find out:

- ▶ What data sources they use;
- ▶ What characterises these sources;
- ▶ What makes a search a success (or a failure);
- ▶ What typical search behaviour is;
- ▶ How much use a metasearcher can be;
- ▶ . . .

meanwhile, in the real world. . .

2006-08-01

Personal metasearch

└ PIS, a prototype Personal Information Searcher



1. putting these ideas together, we've got a prototype search system which is a testbed for further experiments
2. quick look at what there is: query, list of results in a fairly standard way except they come from many different sources: can see ANU, email, references, a few others there. Opened one result — it knows what apps to use to view documents and doesn't bother trying to view them itself
3. Searches Google, BibTeX, email, contacts, calendar, as well as anything with a web interface (Wikipedia, ACM digital library, . . .); full-text plus some simple metadata, might do more on metadata later
4. fairly straightforward to write more search interfaces
5. two or more views, but screen space runs out pretty fast. Randomised each time.
6. Records queries, result selection, relevance judgements, . . .

└ Questions

Questions

Questions for the future:

- What sort of data sources do people use day to day? Which are useful, and for what type of need?
- What sort of queries do people issue? What relation is there between information need and query?
- How can we decide where to look for results?
- How can we merge (or rank) very different results when we get them?
- What use can we make of knowledge about the context of a query? How can we get this knowledge in the first place?
- ...

1. merge: or cluster, or ...
2. also questions of method: how do we answer these questions?
3. the PIS tool and side-by-side evaluation help — but that's slow, are there shortcuts?
4. if we use PIS, how do we manage the experiments? Don't want to spend all our time developing bulletproof software, handling user questions, etc