Server characterisation and selection for personal metasearch

Paul Thomas[*]
Department of Computer Science
Australian National University
paul.thomas@anu.edu.au

A single search interface to all a person's digital resources, such as email archives, corporate databases, websites, and subscription services, is appealing but a central index of all private, corporate, subscription and web data is impractical. A metasearch approach can instead integrate any number of existing search services over a variety of data.

This thesis advocates and examines *personal metasearch*, or metasearch over a user's entire set of digital resources. Metasearch has been well studied in other environments, but has not before been considered with this range of resources; therefore several aspects are re-examined in this new application. Experiments in document sampling, collection size estimation, language modelling, and server selection, all important subproblems in metasearch, demonstrate that established techniques which work well in traditional settings do not necessarily operate well over the wide range of resources in personal applications.

Many techniques for sampling documents from a collection are biased, especially towards longer documents; other metasearch subproblems often rely on unbiased samples and their performance is adversely affected. A new technique for generating samples is therefore proposed and evaluated, and results indicate improvements in sample quality.

Techniques for collection size estimation, language modelling, and server selection are also investigated in a personal metasearch framework. Several techniques prove inappropriate or have been over-fitted in earlier work, but some appear useful. In each case, performance is improved with better-quality samples of documents as input.

Finally, standard evaluation techniques are a poor fit to the personal metasearch environment, and this thesis proposes a new method based on a functioning search tool inserted into the natural retrieval process. This allows study of real information needs, works with dynamic and/or private

[*] Currently at the CSIRO ICT Centre, Canberra, Australia

collections, and records judgements in their full context. It has been validated in a number of experiments and used with a working personal metasearch tool to compare methods for server selection.

Contributions of this thesis include the first analysis of personal metasearch, from a theoretical basis and from studies of potential users; a new algorithm for document sampling which is better able to operate over the wide variety of data sources found in this application; an evaluation of a number of metasearch algorithms and an analysis of common failures; an evaluation technique suited to personal and dynamic collections; and a platform for further research.

(Available online: `http://es.csiro.au/publications.shtml`.)