Hosting search services for the Australian Government

George Ferizis
CSIRO ICT Centre
ACT 2601 Australia
George.Ferizis@csiro.au

David Hawking
CSIRO ICT Centre
ACT 2601 Australia
David.Hawking@csiro.au

Aim: CSIRO's information retrieval group currently host several search related services for the Australian government (accessible, for example, through *http://www.australia.gov.au*), using the P@noptic search engine (*http://www.panopticsearch.com*). We aim to address issues such as the reliability and cost effectiveness of the services and the quality of the search results.

Methods: The reliability of the service has been increased by the introduction of fault tolerance. Several servers serve queries for each service so that if one server is offline the other servers will continue serving queries.

We make measurements on the quality of the search results using various evidence in the document, including *metadata*, *document title*, *anchor text*, *document contents* and *click data*, to determine what evidence returns the most relevant results. We also measure the benefits of an additional, smaller daily crawl to supplement a weekly large whole of government crawl. These are measured by obtaining the mean rank of the homepage of government agencies when queries containing the name of the agency are submitted.

We also assess the bandwidth costs associated with a large whole of government crawl. To reduce bandwidth costs we use a technique named *"incremental crawling"*, which compares the content length present in the HTTP header returned by a web server and the length of the document from previous crawls. If the length has not changed the document is not downloaded again. We measure the reduction in the data downloaded using this method.

Results: Figure 1, shows the effects of using different evidence for ranking. The mean reciprocal rank(MRR) that is shown is defined as the average reciprocal rank of the "correct" result to a query over several different queries. The results show that using evidence provided by a reader of the document (eg. anchor text, or search user clicks) gives more relevant results thatn using any evidence the document provides.

The network traffic reduction that results from the use of incremental crawling show that it is possible to obtain a content length header for approximately 30% of the web pages crawled, and approximately 90% of these have not changed from the previous crawl. On a recent crawl of federal government web sites it was found that incremental crawling reduced the amount of data downloaded from 140 gigabytes to 60 gigabytes.

Conclusions: Improving accessibility of government services and information to the public has provided a great opportunity to deliver impact from our research. It has also posed an interesting and diverse set of new engineering and scientific challenges. By studying and addressing customer requirements in the areas of search quality, functionality, coverage, freshness, efficiency, robustness, and cost-effectiveness, we have improved both our technology and our understanding.

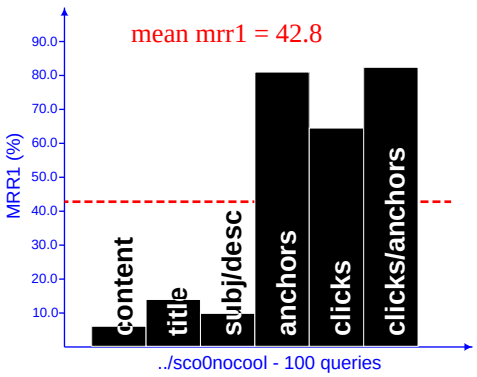

Figure 1: Effects of different evidence on ranking quality

Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.