Search Engines - A Test Report

Wolfgang Dalitz[1] <dalitz@zib.de>

At the beginning of this year, Google has released the beta version of Google Scholar. Google plans to scan and make searchable scientific publications on a large scale.
Do scientists need specific search engines at all?
A closer look, however, reveals that Google as well as the other big Internet search engines have dark spots. With concrete test sites one can prove that the hit lists of the big search engines by no means retrieve all documents, that the hits displayed refer to outdated documents (lying only in the search engine cache), there is no evaluation of meta data, limitation of the maximum hit number, no transparency of ranking strategies, no possibility of adapting the output of hits to one's own needs. Also, the firms running the currently leading search engines meantime have become commercial enterprises so that the search engines will not necessarily be available free of charge to academic institutions in the long run.
We consider it essential that scientists continue operating and using own search engines. In the field of public domain software various programs are suitable, e.g. Harvest, Swish-e, or Lucene/Nutch. These search engines together with others have been implemented and tested at ZIB.
Search engines are always put together from different components, i.e., the gatherer, the indexer, the ranker, and the user interface. The talk presents the different search engines and analyzes their quality, reliability, and configurability. We will outline how these technologies may help improve existing and future services.

[1]ZIB Berlin