Improved Indexing & Searching Throughput

Matt Crane
Department of Computer Science
University of Otago
Dunedin, New Zealand
mcrane@cs.otago.ac.nz

Abstract

Information retrieval is the process of finding relevant information in large corpora of documents based on user queries. Within the discipline there are a number of open research questions and areas. This thesis presents a systematic study into improving the speed of all aspects of an information retrieval system, without such improvements having an adverse effect on the effectiveness of that system.

Several key areas of the indexing process were investigated: the effect of removing spam and correcting encoding errors at indexing time; the amount of parallelism and further improvements to the indexing process; the methods of vocabulary accumulation and collision resolution within a hash table; and as part of the indexing process, a new family of hash functions for information retrieval which exploit the properties of natural language was proposed.

Search performance was also investigated by examining the effects of the spam removal on search quality. A relationship between the size of a collection and the pre-calculation of retrieval scores was discovered.

Overall results indicate a 30% improvement of indexing throughput. This is accompanied by a 15% increase in search quality, whilst its speed could be increased by 25% without degrading quality. The pre-calculation of retrieval scores further improves retrieval speed by up to 3×.

These results were compared against other open-source indexing systems by ATIRE when participating in the SIGIR 2015 RIGOR Workshop Reproducibility Challenge. The results of this challenge show that ATIRE is the fastest indexing system (taking half the time of the next best system), and the second fastest search system using the discovered relationship.

Supervisors: Dr. Andrew Trotman & Dr. Richard O'Keefe
Available: `https://ourarchive.otago.ac.nz/handle/10523/6223`