Information Retrieval manuscript No.
(will be inserted by the editor)

Efficiency in Information Retrieval: Introduction to Special Issue

David Hawking · Alistair Moffat · Andrew Trotman

March 2017

The efficiency of information retrieval (IR) algorithms has always been of interest to researchers at the computer science end of the IR field, and index compression techniques, intersection and ranking algorithms, and pruning mechanisms have been a constant feature of IR conferences and journals over many years. Efficiency is also of serious economic concern to operators of commercial web search engines, where a cluster of a thousand or more computers might participate in processing a single query, and where such clusters of machines might be replicated hundreds of times to handle the query load (Dean 2009). In this environment even relatively small improvements in query processing efficiency could potentially save tens of millions of dollars per year in terms of hardware and energy costs, and at the same time significantly reduce greenhouse gas emissions.

In commercial data centres, query processing is by no means the only big IR consumer of server processing cycles. Crawling, indexing, format conversion, PageRank calculation, ranker training, deep learning, knowledge graph generation and processing, social network analysis, query classification, natural language processing, speech processing, question answering, query autocompletion, related search mechanisms, navigation systems and ad targeting are also computationally expensive, and potentially capable of being made more efficient. Data centers running such services are replicated across the world, and their operations provide every-day input to the lives of billions of

David Hawking
Microsoft Bing,
Canberra, Australia

Alistair Moffat
The University of Melbourne,
Melbourne, Australia

Andrew Trotman
University of Otago,
Dunedin, New Zealand

people. Information retrieval algorithms also run at large scale in cloud-based services and in social media sites such as Facebook and Twitter.

Efficiency in indexing and searching email and documents in a multi-tenant cloud is important, and difficult to achieve. Even when the individual enterprise search applications are small in scale, the investment of programmer time to achieve gains in efficiency can soon pay for itself in reduced server hosting costs.

Ultimately it is the end-users who benefit most from efficient IR algorithms, through more up-to-date information, through "instant" query response, and through the deployment of sophisticated algorithms which would not be practical without efficient implementations.

While it is clear that industry has a strong motivation to work on efficient IR algorithms, academic research also continues to have an important role to play, communicating new efficiency ideas and presenting careful analyses of methods which may be known in industry but not thoroughly explored. In this context we present here five academic papers under the heading of "Efficiency in Information Retrieval". These five papers were selected from a pool of twenty-one that were submitted in response to a Call for Papers that was circulated in December 2015, with submissions closing in April 2016. Expert referees were then invited to read those papers, with all paper authors "blinded" to the identity of their reviewers. Iterations of revision and further review then led to the five papers selected to appear in this issue. In the remainder of this introduction, we briefly introduce the five papers.

Tolosa et al (2017) introduce the integrated cache, a static cache that stores both individual postings lists and bigram intersections of postings lists. They provide an algorithm to re-write queries to take advantage of this cache and show that this provides a speed improvement over previous strategies. Compressed and uncompressed caches are examined, and several strategies for populating the static cache are tested.

Lin and Trotman (2017) provide an investigation into the effect of compression in a Score-at-a-Time search engine. Their analysis of current CPU architectures suggests that (even though the postings lists are likely to be longer) a small speed improvement is seen if no compression is used. They test against several codecs including those coded with SIMD instructions, and those with and without difference encoding.

Kim et al (2017) present a simulation-based investigation of selective search (topical sharding) under a number of different hardware configurations. They examine the Rank-S and Taily algorithms as well as allocation of resources to machines using random distribution and query-log based approaches. Their work, using large query logs, provides new insights into the relative efficiency of selective search compared to exhaustive random sharding, how to distribute those shards across machines, and yields details of trade-offs possible between throughput and latency constraints.

Gagie et al (2017) examine indexing for repetitive collections. Their work includes effective compression techniques, methods for top-k retrieval and identifying the number of documents containing a given string. They also show

how to perform ranked conjunctive and ranked disjunctive search using $tf \times idf$ relevance ranking.

Finally, Daoud et al (2017) introduce Waves, a new approach for storing and intersecting postings lists in order to compute top-k results. They use a multi-tier index and examine documents in a given tier before moving on to the next. Waves outperforms several variants of the BlockMaxWAND both with and without document id reassignment.

We hope that you will enjoy reading these five papers, and join with us in agreeing that efficiency in IR is a critically important topic which continues to offer the prospect of both exciting research opportunities, and also widespread commercial and social benefit.

References

Daoud CM, de Moura ES, Fernandes D, da Silva AS, Rossi C, Carvalho A (2017) Waves: a fast multi-tier top-k query processing algorithm. Information Retrieval Journal pp 1–25, URL http://dx.doi.org/10.1007/s10791-017-9298-6

Dean J (2009) Challenges in building large-scale information retrieval systems. In: Proc. Second ACM Int. Conf. on Web Search and Data Mining (WSDM 2009), p 1, URL http://doi.acm.org/10.1145/1498759.1498761, Keynote presentation

Gagie T, Hartikainen A, Karhu K, Kärkkäinen J, Navarro G, Puglisi SJ, Sirén J (2017) Document retrieval on repetitive string collections. Information Retrieval Journal pp 1–39, URL http://dx.doi.org/10.1007/s10791-017-9297-7

Kim Y, Callan J, Culpepper JS, Moffat A (2017) Efficient distributed selective search. Information Retrieval Journal pp 1–32, URL http://dx.doi.org/10.1007/s10791-016-9290-6

Lin J, Trotman A (2017) The role of index compression in score-at-a-time query evaluation. Information Retrieval Journal pp 1–22, URL http://dx.doi.org/10.1007/s10791-016-9291-5

Tolosa G, Feuerstein E, Becchetti L, Marchetti-Spaccamela A (2017) Performance improvements for search systems using an integrated cache of lists+intersections. Information Retrieval Journal pp 1–27, URL http://dx.doi.org/10.1007/s10791-017-9299-5