Engineering top-k document retrieval systems based on succinct data structures

Simon Gog and Peter Sanders

Inst. Theoretical Informatics, Karlsruhe Institute of Technology, Germany

1 Praxis der Forschung 2016/2017 – Project Description

Given a set of documents, the goal of document retrieval is to efficiently answer queries like „which documents contain all query words", or „what are the k most relevant documents to a given query". For large sets of documents, e.g. web crawls like gov2, clueweb, or CommonCrawl, time-efficient solutions rely on precomputed index data structures. For collections of natural language text or any other collection which can be split into well defined words, the predominant structure is the Inverted-Index. The Inverted-Index stores for every word a sorted list of occurrences of this token in the collection. Queries are answered by scanning and/or intersecting the lists of the query words. However the functionality of Inverted-Indexes are limited by its design. In application domains where there is no clear notion of a word – e.g. in Bioinformatics where DNA sequences are handled or IT security where binaries are investigated – can not be directly applied. On one hand splitting the input in fixed size words will keep the index size small but degrade the performance of queries. On the other hand indexing all substrings is not possible as the index size will be quadratic with respect to the input.

In this project we will engineer document retrieval systems which are based on succinct and compressed data structures[1]. We will first study the existing theoretical proposals and compare their functionality to Inverted-Index based frameworks. In a second step, we move from theory to practice by investigating the performance of a state-of-the-art Inverted-Index-based system: Lucence[2] In a third phase we will design and implement our own system which is based on succinct and compressed data structures[3]. This phase includes both theoretical and also engineering challenges in different areas (construction speed, index sizes, and query performance). The developed system should be general enough to work out-of-the-box in different application domains which makes it unparalleled.

2 Group size

We can take 3-4 students.

References

1. Simon Gog, Timo Beller, Alistair Moffat, and Matthias Petri. From theory to practice: Plug and play with succinct data structures. In *Proc. SEA*, pages 326–337, 2014.

[1] Examples are described in [3, 6, 5]

[2] Lin et al. [4] recently observed that Lucene is a reasonable baseline. It is also by far the most popular Inverted-Index based framework in industry. Other well know search solutions like Elasticsearch (`https://www.elastic.co/products/elasticsearch`) are based on it.

[3] We can make used of already developed prototypes published in [1, 2].

2. Simon Gog and Gonzalo Navarro. Improved single-term top-k document retrieval. In *Proc. ALENEX*, pages 24–32, 2015.
3. W.-K. Hon, R. Shah, and J. S. Vitter. Space-efficient framework for top-k string retrieval problems. In *Proc. FOCS*, pages 713–722, 2009.
4. Jimmy J. Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig MacDonald, and Sebastiano Vigna. Toward reproducible baselines: The open-source IR reproducibility challenge. In *Proc. of ECIR*, pages 408–420, 2016.
5. G. Navarro and Y. Nekrich. Top-k document retrieval in optimal time and linear space. In *Proc. SODA*, pages 1066–1078, 2012.
6. M. Patil, S. V. Thankachan, R. Shah, W.-K. Hon, J. S. Vitter, and S. Chandrasekaran. Inverted indexes for phrases and strings. In *Proc. SIGIR*, pages 555–564, 2011.