## Abstract

A fundamental problem in Information Retrieval is to determine the $k$ most relevant documents of a collection for a given query word or phrase $P$. In a recent result, Navarro and Nekrich [SODA 2012] showed that this problem can be solved in optimal time complexity of $O(|P| + k)$ with a precomputed linear-space index. The size of this optimal-time index was estimated to be 80 times the collection size, rendering it not to be practical. In subsequent work, Navarro and Konow [DCC 2013] and Gog and Navarro [ALENEX 2015] created a practical version with slightly worse query time guarantees but reduced the space to $2.5-3$ times the collection size. The index is conceptually simple and is divided in five components. In this paper we show how the $n \log N$ bits required by the usually largest component – the so called *repetition array* – can be reduced to $n \log\log n + O(n)$, where $n$ is the size of the collection and $N$ the number of documents. As the overall query time complexity matches the one of the old index, we achieve a theoretically superior time-space trade-off. We explore the practical properties of the improved index in a detailed experimental study and compare to the previously established baseline. Index sizes are now between $1.5 - 2$ times the collection size while query speed is comparable to the larger indexes. We also show that the new approach automatically adapts to highly repetitive text collections, which are for instance produced by version control systems.