Abstract

Indexed pattern search in text has been studied for many decades. For small alphabets, the FM-Index provides unmatched performance for Count operations, in terms of both space required and search speed. For large alphabets – for example, when the tokens are words – the situation is more complex, and FM-Index representations are compact, but potentially slow. In this paper we apply recent innovations from the field of inverted indexing and document retrieval to compressed pattern search, including for alphabets into the millions. Commencing with the practical compressed suffix array structure developed by Sadakane, we show that the Elias-Fano code-based approach to document indexing can be adapted to provide new trade- off options in indexed pattern search, and offers significantly faster pattern processing compared to previous implementations, as well as reduced space requirements. We report a detailed experimental evaluation that demonstrates the relative advantages of the new approach, using the standard Pizza&Chili methodology and files, as well as applied use-cases derived from large-scale data compression, and from natural language processing. For large alphabets, the new structure gives rise to space requirements that are close to those of the most highly-compressed FM-Index variants, in conjunction with unparalleled Count throughput rates.