Compressed and Searchable Indexes for Highly Similar Strings*

Kunsoo Park

Dept. of Computer Science and Engineering, Seoul National University, Korea
kpark@theory.snu.ac.kr

Abstract

The collection indexing problem is defined as follows: Given a collection of highly similar strings, build a compressed index for the collection of strings, and when a pattern is given, find all occurrences of the pattern in the given strings. Since the index is compressed, we also need a separate operation which retrieves a specified substring of one of the given strings.

Such a collection of highly similar strings can be found in genome sequences of a species and in documents stored in a version control system. Many indexes for the collection indexing problem have been developed, most of which use classical compression schemes such as run-length encoding and Lempel-Ziv compressions to exploit the similarity of the given strings.

We introduce a new index for highly similar strings, called FM index of alignment. We start by finding common regions and non-common regions of highly similar strings. We need not find a multiple alignment of non-common regions. Finding common and non-common regions is much easier and simpler than finding a multiple alignment. Then we make a transformed alignment of the given strings, where gaps in a non-common region are put together into one gap. We define a suffix array of alignment on the transformed alignment, and the FM index of alignment is an FM index of this suffix array of alignment. The FM index of alignment supports the LF mapping and backward search, the key functionalities of the FM index. The FM index of alignment takes less space than other indexes and its pattern search is also fast.

1998 ACM Subject Classification E.1 [Data Structures] Arrays, Tables, F.2 Analysis of Algorithms and Problem Complexity, F.2.2 [Nonnumerical Algorithms and Problems] Pattern matching

Keywords and phrases Index for similar strings, FM index, Suffix array, Alignment

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2016.2

Category Invited Talk

* This research was supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (NRF-2014M3C9A3063541).