Repetition-Based Text Indexes

Juha Kärkkäinen

Department of Computer Science
P.O. Box 26, FIN-00014 University of Helsinki, Finland
Juha.Karkkainen@cs.Helsinki.FI
http://www.cs.Helsinki.FI/Juha.Karkkainen/

PhD Thesis, Series of Publications A, Report A-1999-4
Helsinki, November 1999, 106 pages
ISSN 1238-8645, ISBN 951-45-8917-3

Abstract

Repetition-based indexing is a new scheme for preprocessing a text to support fast pattern matching queries. The scheme provides a general framework for representing information about repetitions, i.e., multiple occurrences of the same string in the text, and for using the information in pattern matching. Well-known text indexes, such as suffix trees, suffix arrays, DAWGs and their variations, which we collectively call *suffix indexes*, can be seen as instances of the scheme.

Based on the scheme, we introduce the *Lempel–Ziv index*, a new text index for string matching. It uses the repetition information in a *Lempel–Ziv parse*, which is a division of the text into non-overlapping substrings with earlier occurrences, and which is also used in the Ziv–Lempel family of text compression methods. The Lempel–Ziv index offers a possibility for a space–time tradeoff. The space requirement can be smaller than for suffix indexes by up to a logarithmic factor, while the query time is larger but still sublinear in the length of the text. The only previous text index offering a space–time tradeoff is the sparse suffix tree. The Lempel–Ziv index improves on the results of the sparse suffix tree in many cases.

Text indexes for q-gram matching, i.e., for matching string patterns of length q, are used in some approximate string matching algorithms. We introduce a new repetition-based q-gram index, the *Lempel–Ziv index for q-grams*, that has asymptotically optimal space requirement and query time provided that q is a constant or grows slowly enough with respect to the length of the text. Queries are as fast as with traditional q-gram indexes, but the space requirement can be smaller by a logarithmic factor.

Some additional novel data structures are developed for subproblems arising in the Lempel–Ziv indexing methods. These include a new variation of the suffix tree with a faster query time, a variation of a data structure for two-dimensional range searching with new possibilities for space–time tradeoffs, and a new data structure, called the *nesting leveled list*, for the range containment problem.

Computing Reviews (1998) Categories and Subject Descriptors:

F.2.2 [Analysis of Algorithms and Problem Complexity]: Nonnumerical Algorithms and Problems—pattern matching, sorting and searching, geometrical problems and computations

E.1 Data Structures—trees

General Terms:

Algorithms, Theory

Additional Key Words and Phrases:

String matching, text indexes, Lempel–Ziv parsing, q-grams, range searching