Pattern Matching on Simple Collage System Using Finite Automata

Marek Hanuš and Bořivoj Melichar
Prague Stringology Club
Department of Computer Science & Engineering
Czech Technical University in Prague
Karlovo nám. 13, 121 35 Prague 2, Czech Republic
e-mail:{hanusm2 | melichar}@fel.cvut.cz

Abstract: A new approach to the pattern matching on simple collage system, which is a general framework for representing several dictionary-based text compression methods (e.g. LZW, LZ78), is presented. It is based on incremental construction of deterministic finite automaton using the alphabet of the compressed text (the numbers of phrases in the dictionary of the collage system). The input of the algorithm is the pattern matching automaton using the alphabet of the uncompressed text. The algorithm can be used for many pattern matching problems, e.g. for approximate pattern matching or for searching of the set of patterns defined by regular expression.

Keywords: pattern matching, compressed pattern matching, text compression, finite automata, collage systems