

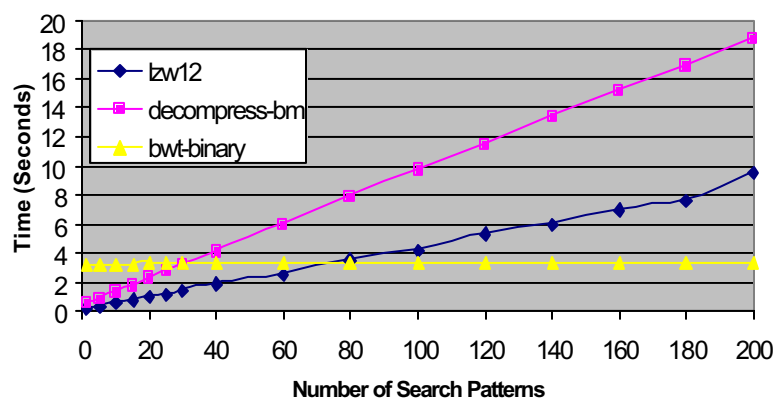
## LZW Based Compressed Pattern Matching

Tao Tao, Amar Mukherjee

School of Computer Science, University of Central Florida

Email: (ttao+amar)@cs.ucf.edu

Compressed pattern matching is an emerging research area that addresses the following problem: given a file in compressed format and a pattern, report the occurrence(s) of the pattern in the file with minimal (or no) decompression. In this paper, we report our work on compressed pattern matching in LZW compressed files. The reported work is based on Amir's well-known "almost-optimal" algorithm [1] but has been improved to search not only the first occurrence of the pattern but also all other occurrences. The improvements also include the multi-pattern matching and a faster implementation for so-called "simple pattern", which is defined as "a pattern with no symbol appearing more than once". Extensive experiments have been conducted to test the search performance and to compare with not only the "decompress-then-search" approach but also the best available compressed pattern matching algorithms, particularly the BWT-based algorithms [2, 3]. The results showed that our method is competitive among the best algorithms.



LZW is one of the most efficient and popular compression algorithms used extensively and our method requires no modification on the compression algorithm.

The full paper is available at <http://vlsi.cs.ucf.edu>.

The work has been partially supported by National Science Foundation grants IIS-0312724 and IIS-0207819.

### Reference:

- [1] A. Amir, G. Benson and M. Farach, "Let sleeping files lie: Pattern matching in Z-compressed file", Journal of System Sciences, 52: 299-307, 1996.
- [2] T. Bell, M. Powell, A. Mukherjee and D. Adjero, "Searching BWT compressed text with the Boyer-Moore algorithm and binary search", Proc. IEEE DCC, March 2002.
- [3] Andrew Firth, "A comparison of BWT approaches to compressed-domain pattern matching", Honours report at the University of Canterbury, New Zealand, 2002