

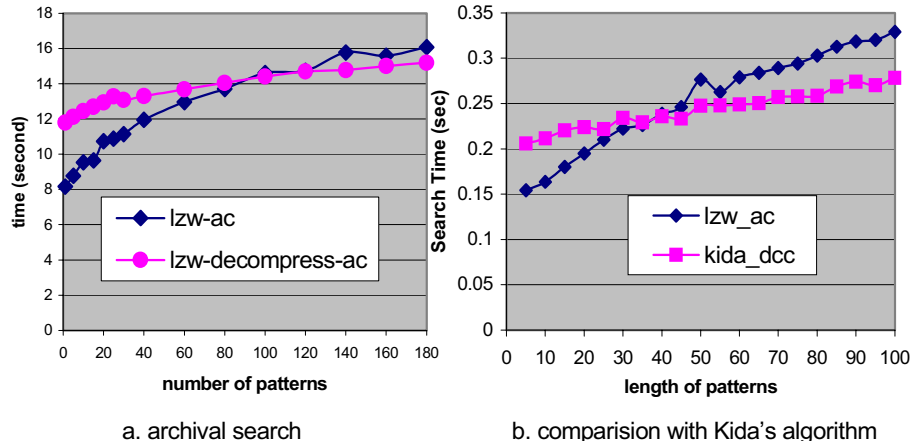
## Multiple-Pattern Matching In LZW Compressed Files Using Aho-Corasick Algorithm

Tao Tao, Amar Mukherjee

School of Computer Science, University of Central Florida

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

A novel multiple-pattern matching algorithm for LZW compressed files using Aho-Corasick algorithm [1] is reported. Since the LZW trie can be reconstructed without explicit decompression [2], a state transition table is then built from the LZW trie and the AC automaton and is used in our algorithm for fast multiple-pattern searching. The algorithm takes  $O(mt+n+r)$  time with  $O(mt)$  extra space, where  $n$  is the size of the compressed file,  $m$  is the total size of all patterns,  $t$  is the size of the LZW trie and  $r$  is the number of occurrences of the patterns. The algorithm is compared with a similar algorithm developed by Kida [3]. Extensive experiments have been conducted to test the performance of the algorithm and to compare with other well-known compressed pattern matching algorithms, particularly Kida's algorithm. The results show that the proposed algorithm is practically the fastest among all approaches when the number of patterns is not very large. Therefore, our algorithm is preferable for general string matching applications. The proposed algorithm is efficient for large files and it is particularly efficient when being applied on archival search if the archives are compressed with a common LZW trie.



A full version of this manuscript is available on: <http://vlsi.cs.ucf.edu>.

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

### References:

1. A.V. Aho and M.J. Corasick, "Efficient string matching", Commun. ACM 18. 6, June 1975, pp.333-340.
2. 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.
3. Kida, M. Takeda, M. Miyazaki, A. Shinohara and S. Arikawa, "Multiple Pattern Matching in LZW Compressed Text", Journal of Discrete Algorithm, Vol. 1(1). 2000.