

Real-time traversal in grammar-based compressed files*

Leszek Gąsieniec Roman Kolpakov Igor Potapov Paul Sant

Department of Computer Science
University of Liverpool
Liverpool L69 7ZF, UK
{leszek, roman, igor, pauls}@csc.liv.ac.uk

Historically, one of the main aims of text compression was to reduce the size of data stored for future analysis and processing, see, e.g., [2, 3]. Unfortunately, further processing of the compressed data is usually preceded by complete decompression. This standard approach may cause problems in certain applications, where the original (uncompressed) file is very large and we face running out of space available for the computation. Thus, it is important to be able to process compressed data without requiring (complete) decompression. Moreover, in some applications (e.g., pattern matching problems), the information represented by the compressed file can be processed in relatively small chunks (length of a pattern). In this context it is crucial to study compression methods that allow time/space efficient access to any fragment of a compressed file without being forced to perform complete decompression.

We study here real-time recovery of consecutive symbols from compressed files, in the context of grammar-based compression, see, e.g., [1]. In this setting, a compressed text is represented as a small (a few kilobytes) dictionary D (containing a set of code words), and a very long (a few megabytes) string based on symbols drawn from the dictionary D . The space efficiency of this kind of compression, is comparable with standard compression methods based on the Lempel-Ziv approach [3]. We show, that one can visit consecutive symbols of the original text, moving from one symbol to another in constant time and extra $O(|D|)$ space. This algorithm is an improvement of the on-line linear (amortised) time algorithm presented in [1].

References

- [1] L. Gąsieniec and I. Potapov. Time/Space Efficient Compressed Pattern Matching. In Proc. of 13th International Symposium on Fundamentals of Computation Theory, LNCS, Volume 2138, pp 138-152, Springer-Verlag, 2001.
- [2] A. Lempel and J. Ziv On the complexity of finite sequences, *IEEE Transactions on Information Theory*, pp. 22:75–81, 1976.
- [3] J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Transactions on Information Theory*, pp. IT-23(3):337–343, 1977.

*This work is supported by the EPSRC grant GR/R84917/01.