Stack-based LaTeX File Compression

wcrr51

January 2021

1 Introduction

This report will describe six ideas used in the implementation of a LaTeX file compression scheme. This scheme works under the assumption that only one LaTeX file is being encoded and decoded at a time, however, there is nothing to stop multiple files being independently encoded and decoded.

Additionally, it is assumed that input LaTeX files are encoded using ASCII (7 bits per character plus one padding bit), no assumptions are made regarding the line ending format.

2 Compression Stack

This compression scheme uses a *stack* model. In the same way the OSI model is used in networking to wrap data the more low-level it becomes, this scheme may use many techniques to compress data at each level.

When encoding, this goes from the application-level to the text-level through encoding of the input plaintext LaTeX file, followed by the text-level to the binary level and finally any binary level compression. Decoding simply does these operations in reverse (by sending the input back up the stack).

Standard compression algorithms can be inserted at any point as long as the output type from the previous layer matches the input type to the next (and vice versa). It is worth noting that the same scheme would work for applying encryption. This may involve the use of standardised approaches, context-modified approaches and/or novel context-specific approaches.

An advantage of this is that each layer can add its own metadata.

3 Application-level Encoding

This idea aims to reduce the size of the file using prior knowledge of LaTeX commands and symbols, this could amount to being just shy of generating a parse tree.

LaTeX is, for the most part, very well-structured. This means that known tags and whitespace can be efficiently encoded by a simple combination of run-length encoding and dictionary look-up.

4 Text-level Encoding

As input documents are assumed to be encoded with ASCII, each character can safely have one bit removed, automatically reducing the file size to seven eighths of the original.

5 Binary-level Encoding

References

[1] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.

[2] Albert Einstein. *Zur Elektrodynamik bewegter Körper*. (German) [*On the electrodynamics of moving bodies*]. Annalen der Physik, 322(10):891–921, 1905.

[3] Knuth: Computers and Typesetting