

# The Predictive-Substitutional Compression Scheme and its Effective Implementation

Jakub Swacha

Institute of Information Technology in Management, University of Szczecin  
Mickiewicza 64, 71-101 Szczecin, Poland  
e-mail: jakubs@uoo.univ.szczecin.pl

The predictive-substitutional compression scheme combines string-oriented and symbol-oriented models to improve compression speed and effectiveness. It parses the input data into strings and literals, and encodes each string using a string-oriented model, and each literal using a symbol-oriented model; both its models are context-aware.

There are several hybrid compression algorithms known. In this paper, two new one-pass predictive-substitutional algorithms are introduced.

The basic algorithm (PLZ) combines simple switching with a very fast PPM variant. It achieves high compression efficiency to processing time ratio due to a novel technique of switching position encoding, special types of order 0 dictionaries (*cache dictionary* and *distance dictionary* storing recently substituted phrases and offsets regardless of their context), and adaptive data structures (becoming multi-level when necessary).

The improved algorithm (PLZ+) implements fast switching cost evaluation making it capable of doing a substitution only when the substitution identifier's codeword length is shorter than the sum of codeword lengths of the literals that the string is composed of.

For the purpose of testing the algorithms' overall compression performance (combining speed and effectiveness), transmission acceleration (*TA*) measure is proposed:

$$TA = \frac{t_u}{t_c + t_{tc} + t_d},$$

where:  $t_u$  is transmission time of uncompressed data,  $t_c$  is compression time,  $t_{tc}$  is transmission time of compressed data,  $t_d$  is decompression time. *TA* means how many times faster data can be transmitted due to compression.

The table below lists transmission acceleration for 512 kb/s bandwidth measured\* on *Calgary* and *Canterbury* Corpora.

Corpus	PLZ	GRZip -f	LZP3o2	LZOP	Zip -6	PPMd o3	PPMd o4	Bzip2
<i>Calgary</i>	2,910	2,756	1,451	1,944	2,668	2,721	2,744	2,423
<i>Canterbury</i>	4,319	3,959	2,671	2,271	3,250	3,866	3,735	3,012

As the presented experimental results show it, regardless of the recent improvement in predictive algorithms, the predictive-substitutional scheme is one of the best available data compression methods for data transmission purposes in medium bandwidth range.

---

\* Tested on: Intel Celeron 1000 MHz, 256MB RAM SDR, HDD 40GB 7200 rpm, Windows 98 SE. PLZ configuration: order 2, minimum substitution length 4, dictionary sizes order 0(d) - 4, order 2 - 8.