

Two-level directory based compression

Przemysław Skibiński

University of Wrocław, Institute of Computer Science, Przesmyckiego 20,

51-151 Wrocław, Poland, e-mail: inikep@ii.uni.wroc.pl

The basic idea of preprocessing is to transform the text into some intermediate form, which can be used as input of any existing general-purpose compressor and compressed more efficiently. Dictionary-based preprocessing is based on the notion of replacing whole words with shorter codes. The dictionary of words is usually static (for a given language) and given in advance.

In this paper¹ we present dictionary-based preprocessing technique and its implementation called TWRT (Two-level Word Replacing Transformation). Our preprocessor uses several dictionaries and divides files into various kinds. The first level dictionaries (small dictionaries) are specific for some kind of data (e.g., programming language, references). The second level dictionaries (large dictionaries) are specific for natural language (e.g., English, Russian, French). Before ordinary preprocessing TWRT uses “faster dictionaries detection” mechanism, which chooses the best combination of the dictionaries, one small and one large. This technique gives additional opportunities, as TWRT can decide if binary data filter, surrounding words with spaces, EOL coding, and record preprocessing should be used in ordinary preprocessing.

On the Calgary corpus, TWRT improves the compression performance of bzip2 by over 7% and PPMonstr by about 6% on average. Even for the top compressor nowadays, PAQ6, the gain is significant – 5%. On the multilingual text files TWRT improves the compression performance of bzip2, PPMonstr, and PAQ6 by about 8%. Moreover, TWRT improves the compression speed with PAQ6 and on larger files with PPMonstr.

| | bzip2 | TWRT (BWT) + bzip2 | PAQ6 | TWRT (PAQ) + PAQ6 | PPMonstr | TWRT (PPM) + PPMonstr | Durilca | RKC | Dictionaries used by TWRT |
|---------|-------|--------------------------|--------------|-------------------------|----------|-----------------------------|--------------|-------|------------------------------|
| bib | 1.975 | 1.724 | 1.617 | 1.476 | 1.663 | 1.501 | 1.542 | 1.647 | Ref, ENG |
| book1 | 2.420 | 2.130 | 2.090 | 1.900 | 2.119 | 1.913 | 1.853 | 2.048 | ENG |
| book2 | 2.062 | 1.876 | 1.691 | 1.597 | 1.742 | 1.614 | 1.649 | 1.723 | CompSc, ENG |
| geo | 4.447 | 4.178 | 3.536 | 3.542 | 3.869 | 3.831 | 3.870 | 3.674 | |
| news | 2.516 | 2.333 | 2.034 | 1.919 | 2.084 | 1.937 | 1.956 | 2.086 | ENG |
| obj1 | 4.013 | 4.048 | 3.047 | 3.041 | 3.345 | 3.330 | 3.349 | 3.234 | CompSc, ENG |
| obj2 | 2.478 | 2.475 | 1.703 | 1.703 | 1.898 | 1.898 | 1.899 | 1.803 | |
| paper1 | 2.492 | 2.105 | 2.052 | 1.770 | 2.122 | 1.842 | 1.883 | 2.087 | CompSc, ENG |
| paper2 | 2.437 | 2.069 | 2.056 | 1.788 | 2.112 | 1.836 | 1.772 | 2.062 | ENG |
| pic | 0.776 | 0.529 | 0.456 | 0.448 | 0.693 | 0.455 | 0.693 | 0.683 | |
| progC | 2.533 | 2.379 | 2.031 | 1.919 | 2.106 | 1.989 | 2.022 | 2.119 | C++, ENG |
| progl | 1.740 | 1.629 | 1.314 | 1.246 | 1.352 | 1.283 | 1.382 | 1.378 | Lisp, ENG |
| progp | 1.735 | 1.707 | 1.312 | 1.283 | 1.360 | 1.327 | 1.409 | 1.462 | Lisp, ENG |
| trans | 1.528 | 1.531 | 1.126 | 1.104 | 1.151 | 1.156 | 1.067 | 1.155 | Ref, ENG |
| average | 2.368 | 2.194 | 1.861 | 1.766 | 1.972 | 1.851 | 1.881 | 1.940 | |
| ctime | 4.07 | 6.04 | 488.6 | 389.4 | 25.0 | 26.7 | 26.6 | 76.2 | |

Results of the experiments on the Calgary corpus. Results are given in bits per character (bpc).

The compression times (ctime) are given in seconds.

¹ Full version of this paper is available at <http://www.ii.uni.wroc.pl/~inikep/papers/05-TwoLevelDict.pdf>