A Dictionary-Based Multi-Corpora Text Compression System

Weifeng Sun       Nan Zhang       Amar Mukherjee
Computer Science Department, University of Central Florida, Orlando, Fl.32816
{wsun, nzhang, amar}@cs.ucf.edu

In this paper we introduce StarZip, a multi-corpora text compression system, together with its transform engine StarNT. StarNT achieves a superior compression ratio than almost all the other recent efforts based on BWT and PPM. StarNT is a dictionary-based fast lossless text transform. The main idea is to recode each English word with a representation of no more than three symbols. This transform not only maintains most of the original context information at the word level, but also provides some kind of "artificial" but strong context. The transform also exploits the distribution of lengths of the words and frequency to reduce the size of the transformed text which is provided to a backend compressor. Ternary search tree is used to store the transform dictionary in the transform encoder. This data structure provides a very fast transform encoding with a low storage overhead. Another novel idea of StarNT is to treat the transformed codewords as the offset of words in the transform dictionary. Thus the time complexity of O(1) for searching a word in the dictionary is achieved in the transform decoder.

The transform encoder and transform decoder share the same dictionary, which is prepared off-line according to the following rules: 1) Most frequently used words (i.e. 312 words) are listed in the beginning of the dictionary according to their frequencies. 2) Other words are sorted according to their lengths and frequencies. Words with longer lengths are stored after words with shorter lengths. If two words have the same length, word with higher frequency of occurrence is listed after word with lower frequency. 3) To achieve better compression performance for backend data compressor, only letters [a..Z] are used to represent the codeword.

Experimental results show that the average compression time has improved by orders of magnitude compared to our previous dictionary based transform LIPT and for large files, viz. 400Kbytes or more, the compression time is no worse than those obtained by bzip2 and gzip, and is much faster than PPMD. Meanwhile, the overhead in the decompression phase is negligible. We draw a significant conclusion that bzip2 in conjunction with this transform is better than both gzip and PPMD both in time complexity and compression performance.

One of the key features of StarZip compression system is to develop domain specific dictionaries and provide tools to develop such dictionaries. Results from five corpora show that StarZip achieves an average improvement in compression performance (in terms of BPC) of 13% over bzip2 -9, 19% over gzip -9, and 10% over PPMD.


Acknowledgement
This research is partially supported by NSF grant number: IIS-9977336 and IIS-0207819.