Less Redundant Codes for Variable Size Dictionaries

Zhen Yao Nasir Rajpoot

Dept. of Computer Science, University of Warwick
Coventry CV4 7AL, United Kingdom
{csvhe, nasir}@dcs.warwick.ac.uk

Abstract

In this paper, we report work on a family of variable-length codes with less redundancy than the flat code used in most of the variable size dictionary-based compression methods. The length of codes belonging to this family is still bounded above by $\lceil \log_2 |D| \rceil$ where $|D|$ denotes the dictionary size. We describe three of these codes, namely, the balanced code [2], the Phase-in-Binary code (PB)[1] and the Depth-Span code (DS). As the name implies, the balanced code is constructed by a height balanced tree, so it has the shortest average codeword length. The corresponding coding tree for the PB code has an interesting property that it is made of full binary phases, and thus the code can be computed efficiently using simple binary shifting operations. The DS coding tree is maintained in such a way that the coder always finds the longest extendable codeword and extends it until it reaches the maximum length. It is optimal with respect to the code-length contrast. The PB and balanced codes have almost similar improvements, around 3% to 7% which is very close to the relative redundancy in flat code. The DS code is particularly good in dealing with files with a large amount of redundancy, such as a running sequence of one symbol. We also did some empirical study on the codeword distribution in the LZW dictionary and proposed a scheme called *Dynamic Block Shifting* (DBS) to further improve the codes' performance. Experiments suggest that the DBS is helpful in compressing random sequences. From an application point of view, PB code with DBS is recommended for general practical usage.

References

[1] T. Acharya and J.F. Já Já. Enhancing Lempel-Ziv codes using an on-line variable-length binary encoding. In *Proceedings IEEE Data Compression Conference (DCC'96)*, March 1996.

[2] R. Gallager and D.V. Voorhis. Optimal source codes for geometrically distributed integer alphabets. *IEEE Trans. on Information Theory*, IT-21:228–230, March 1975.