A fast and efficient post BWT-stage for the Burrows-Wheeler Compression Algorithm

Jürgen Abel
University of Duisburg-Essen, Department "Communications Systems",
Faculty of Engineering Sciences, Bismarckstrasse 81, 47057 Duisburg, Germany.
Phone: +49 - 2137 - 999333, E-mail: juergen.abel@data-compression.info.

A new stage for the Burrows-Wheeler Compression Algorithm (BWCA) is presented, called Incremental Frequency Count (IFC), which is together with a Run Length Encoding (RLE) stage located between the Burrows-Wheeler Transform (BWT) and the Entropy Coding (EC) stage of the algorithm. The IFC stage offers a high throughput similar to a Move To Front (MTF) stage combined with good compression rates, similar to the strong but slow Weighted Frequency Count (WFC) stage. A BWCA based on a IFC stage and a corresponding RLE stage achieves compression times double as fast as based on a WFC stage while the compression rates are under the top of the BWT based compression algorithms.

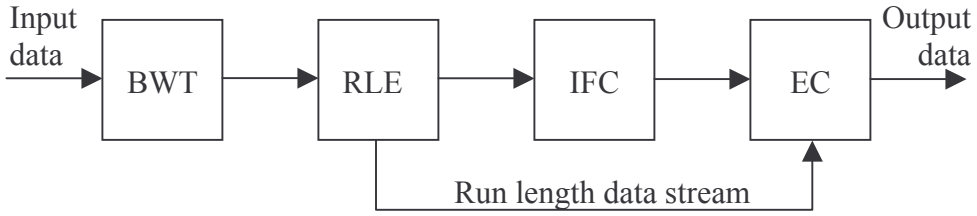

Figure 1: The BWCA with an RLE and IFC stage

The main idea behind the IFC stage is the use of a raising increment. The increment is depending on the average index value of the near past. Each symbol has a corresponding counter and the counter of the current symbol is increased by the current increment. Since only one counter is changed for each symbol processed, only one element needs to be resorted inside the counter list, which leads to an implementation with much less computational complexity than the WFC stage. Similar to arithmetic coding, the increment and counters are halved if a counter exceeds a fixed threshold.

Since a RLE stage is usually many times faster than a ranking scheme and in order to decrease the pressure of runs inside the ranking scheme, an RLE stage is used in front of the IFC stage. The RLE stage replaces all runs of repeated symbols, which have a length of two or more symbols, by a run consisting of exactly two symbols. The length of a run is transmitted into a separate data stream as shown in Figure 1. This way, the length information does not disturb the context of the main data stream. Table I and II present the compression rates and times of IFC based and other compression schemes.

Scheme	GZIP93	BW94	F96	BS99	D02	MTF04	**IFC04**	WFC04
Avg./bps	**2.697**	**2.40**	**2.34**	**2.26**	**2.249**	**2.276**	**2.239**	**2.231**

TABLE 1: COMPRESSION RATES FOR THE CALGARY CORPUS

Scheme	compr. t. GZIP93	decom. t. GZIP93	compr. t. MTF04	decom. t. MTF04	**compr.t. IFC04**	**decom.t. IFC04**	compr. t. WFC04	decom. t. WFC04
Sum/sec	**2.91**	**1.13**	**2.73**	**2.26**	**3.34**	**2.86**	**6.70**	**6.19**

TABLE 2: COMPRESSION TIMES FOR THE CALGARY CORPUS