One attempt of a compression algorithm using the BWT

Bernhard Balkenhol[*]
Fakultät für Mathematik
Universität Bielefeld
Postfach 100 131
33501 Bielefeld
Germany

Yuri M. Shtarkov[†]
Institute for Problems on
Information Transmission
100 447, Moscow, GSP-4
B. Karetnyi per. 19
Russia

1 Introduction

In 1994 Burrows and Wheeler [5] described a universal data compression algorithm (BW-algorithm, for short) which achieved compression rates that were close to the best known compression rates. Due to it's simplicity, the algorithm can be implemented with relatively low complexity. Fenwick [8] described ideas to improve the efficiency (i.e. the compression rate) and complexity of the BW-algorithm. He also discusses relationships of the algorithm with other compression methods. Schindler [14] proposed a Burrows and Wheeler Transformation (BWT, for short) that is based on a limited ordering. This speeds up the algorithm for compression, but slows it down for decompression and slightly decreases the efficiency. Larsson [10] describes the relationship of the BWT with suffix and context trees. Sadakane [13] suggests a method to compute the BWT faster, and compares it to other methods. Recently Balkenhol and Kurtz [2] gave a thorough analysis of the BWT from an information theoretic point of view. They described implementation techniques for data compression algorithms based on the BWT, and developed a program with a better compression rate. In [4] these previous results on the BW-algorithm are improved. Based on the context tree model, the authors consider the specific statistical properties of the data at the output of the BWT. They describe six important properties. These considerations lead to modifications of the coding method, which in turn improve the coding efficiency. Further improvements related to the sorting are presented in [3, 11, 12].

2 Context Tree Models of Sources

Let A be the discrete alphabet of α symbols, $\alpha \geq 2$; x^k be the first k symbols of the message x^n with $x_i \in A$; $p(x^k|\omega)$ be the probability of the occurrence of x^k at the output of the source ω and $\varphi = \{\varphi(x^n),\ x^n \in A^n\}$ be a uniquely decodable (arithmetic) code for sequences of arbitrary (in particular, unknown beforehand) length n. The codeword lengths satisfy the inequality $|\varphi(x^n)| = -\log q(x^n|\varphi) \leq$

[*]Email: bernhard@mathematik.uni-bielefeld.de
[†]Email: shtarkov@iitp.ru, Work done while visiting the Fakultät für Mathematik, Universität Bielefeld

$-\sum_{k=1}^{n} \log \vartheta(x_k|x^{k-1}, \omega) + 1$, where $|z|$ is the length of the sequence z or the cardinality of the set z, $\log(.) = \log_2(.)$, $\{q(x^n)\}$ is a coding probability distribution, which is described by conditional probabilities $\{\vartheta(x_k|x^{k-1}, \omega),\, x^k \in A^k, k = 1, 2, \ldots\}$. The choice $q(x^n) = p(x^n|\omega)$ is the optimal one. But if the statistic of the source is unknown, then the universal coding approach for a given source model or for some set of source models can be used. Below we shall refer to the important set of *Context Tree* (CT) models.

A finite memory CT-source ω is defined by a proper and complete set S of contexts (sequences over the alphabet A) of length $d = |s| \leq D$, by the set of conditional probability distributions $\{\Theta_s, s \in S\} = \{\{\theta(a|s), a \in A\}, s \in S\}$, and by the probability distribution for the first D symbols of the message. Completeness and properness of a set S mean that exactly one context exists for any $x^k \in A^k, k \geq D$, i.e. the equality $x_k \ldots x_{k-d+1} = s_k \in S$ holds for exactly one $d \leq D$. Then the probability $p(x^n|\omega)$, divided by the probability of the first D symbols, is equal to the product of the conditional probabilities $\theta(x_{k+1}|s_k)$, $k \geq D$, or (in other words) to the product of the probabilities $p(x^k(s)|\Theta_s)$ over all $s \in S$, where $x^k(s)$ is the subsequence of *independent* symbols from x^k occurring after context s. Thus S is a model of the CT-source which can be represented as a set of leaves (contexts) of a complete and proper α-ary tree T_S (and vice versa).

Any CT-source can be described by a Markov chain of order $D' \geq D$ with a *larger number* K of “free” parameters (values of conditional probabilities). Since the cumulative (per message) redundancy of universal (relative to the values of free parameters) coding is proportional to $K \log n$, decreasing K is important for decreasing the coding redundancy under the condition that *it does not decrease* the accuracy of the source description. CT-models satisfy the last condition: their structure allows to exploit the fact that the “actual” lengths of the contexts are *different.*

Usually the number of different symbols that occur after a context, decreases with the length of the context. Consider for example english text: if x_k is a blank and x_{k-1} is a period, then with high probability the next symbol will be *any* capital letter independent of x_{k-2}, x_{k-3}, etc. (the actual length of the context is two). Of course, exceptions exist: for example, if $x_k = q$, then almost surely $x_{k+1} = u$ (although the actual length of the context q is one).

Thus some paradoxical situations occur: the shorter the contexts, the smaller the number K of free parameters and the coding redundancy (see above), and the larger the uncertainty relative to the next symbol and the coding rate. In fact, only rather short contexts influence the coding rate, and increasing D above some threshold does not improve the coding rate. Therefore we shall use conventional terms “bad” and “good” for short and long contexts, respectively, although contexts of almost all lengths exist, as well as exceptions of the kind mentioned above.

3 The BW-algorithm and CT-Models

The BW-algorithm [5] consists of three phases:

1. The Burrows and Wheeler transformation (BWT) $x^n \to y^n$ is the reordering