

An extension of the Burrows Wheeler Transform to k words

Sabrina Mantaci Antonio Restivo Marinella Sciortino

University of Palermo, Dipartimento di Matematica ed Applicazioni

Via Archirafi 34, 90123 Palermo, ITALY

E-mail: {sabrina,restivo,mari}@math.unipa.it

We introduce an extension E of the Burrows-Wheeler Transform to a multiset of k primitive words. Primitiveness is not actually a restrictive hypothesis, since in practice almost all the processed texts are primitive (or become primitive by adding an end-of-string symbol). Let \preceq_ω be a new order relation such that, given two primitive words u and v , $u \preceq_\omega v$ if u^ω is lexicographically smaller than v^ω , where u^ω and v^ω are the infinite words obtained by infinitely iterating u and v , respectively. Let $S = \{u_1, \dots, u_k\}$ be a multiset of k primitive words of A^* . Let w_1, w_2, \dots, w_m be the sequence of conjugates of elements of S , sorted according to \preceq_ω , that is $w_i \preceq_\omega w_j$ for $1 \leq i < j \leq m$. We denote by \mathcal{I} the set of indices representing the positions in the sequence $\{w_i\}_{i=1}^m$ of the original words in S . We define $L = L[1]L[2] \dots L[m]$, where $L[i]$ denotes the last character of the word w_i , for $i = 1, \dots, m$. The output is the couple (L, \mathcal{I}) . For example, if $S = \{abac, cbab, bca, cba\}$, then $E(S) = (L, \mathcal{I})$ where $L = ccbbbcacaaabba$ and $\mathcal{I} = \{1, 9, 13, 14\}$. We prove that such a transformation, as the BWT , is reversible. Differently from the BWT , E is surjective, that is, for any word $L \in A^*$, there exists a multiset S of primitive words and a set \mathcal{I} of indices such that $E(S) = (L, \mathcal{I})$.

We show how to use the transformation E as a preprocessing for the simultaneous compression of k different texts. We denote by \mathcal{C} a compressor that uses the transformation E as preprocessing. If $S = \{x_1, \dots, x_k\}$ is a multiset of words, we denote by $\mathcal{C}(S)$ the word obtained by applying the compressor \mathcal{C} to $E(S)$. Under general conditions, it is possible to verify that if X and Y are two multisets of words, then

$$|\mathcal{C}(X \cup Y)| \leq |\mathcal{C}(X)| + |\mathcal{C}(Y)|. \quad (1)$$

Previous inequality shows that the simultaneous compression of a sequence of k words $\{x_1, \dots, x_k\}$ by using the transformation E as preprocessing, is better than compressing each word x_i separately and concatenating the outputs of the compressed words. Such a result has several interesting application. For example, most BWT -based compressors process the input file x of length n by parsing it in k blocks x_1, \dots, x_k of size n/k . A single block is read, compressed and written to the output file before the next one is considered. So, by using the extended transformation, one can derive a method of text compression that is a good trade-off, in terms of compression ratio, memory requirement and time complexity, between compressing, by using a BWT -based compressor, the whole text and the text divided into blocks.

With reference to Inequality 1, we observe that the more similar two texts x and y are, the smaller $|\mathcal{C}(x, y)|$ is with respect to $|\mathcal{C}(x)| + |\mathcal{C}(y)|$. So, we can define the following distance measure, that inherits from E the symmetry property:

$$\delta(x, y) = \frac{|\mathcal{C}(x, y)| - \min\{|\mathcal{C}(x)|, |\mathcal{C}(y)|\}}{\max\{|\mathcal{C}(x)|, |\mathcal{C}(y)|\}}$$