Theoretical Computer Science 522 (2014) 77–84

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Linear computation of unbordered conjugate on unordered alphabet

J.-P. Duval, T. Lecroq, A. Lefebvre *

Normandie Université, LITIS EA4108, Université de Rouen, 76821 Mont-Saint-Aignan Cedex, France

ARTICLE INFO

Article history:
Received 11 January 2013
Received in revised form 26 September 2013
Accepted 10 December 2013
Communicated by M. Crochemore

Keywords:
Combinatorics on words
Border
Conjugate
Unbordered word
Unordered alphabet

ABSTRACT

We present an algorithm that, given a word w of length n on an unordered alphabet, computes one of its unbordered conjugates. If such a conjugate does not exist, the algorithm computes one of its conjugates that is a power of an unbordered word. The time complexity of the algorithm is $O(n)$: the number of comparisons between letters of w is bounded by $4n$.

1. Introduction

A *word* is a sequence of zero or more symbols (also called letters) from an alphabet A; the word with zero symbols is denoted by ε. The set of all finite words over the alphabet A is denoted by A^*. A word w of length n on an alphabet A is denoted by $w[1..n]$ where $w[i] \in A$ for $1 \leqslant i \leqslant n$. A word u is a *prefix* of w if $w = ux$ for $w \in A^*$. Similarly, u is a *suffix* of w if $w = xu$ for $x \in A^*$.

An integer p is a *period* of a word w of length n if and only if $w[i] = w[i + p]$ for all $1 \leqslant i \leqslant n - p + 1$. A word u is called a *border* of a word w if u is both a prefix and a suffix of w. Periodicity and borderness are dual notions that play a crucial role in many applications such as string matching [6], code theory [3] and computational biology [14,5].

A word without any border, except the empty word, is called *unbordered*. Periodicity and unbordered factors have already been investigated [9,7,12]. Two words uv and vu are called *conjugates* of each other. Results are known on unbordered conjugates of words [16].

In this article we are interested in the computation of an unbordered conjugate of a given word w on an unordered alphabet. If such a conjugate does not exist, we are interested in computing one conjugate of w that is a power of an unbordered word. A word is called *primitive* if it is not an integer power of another word. It is known that every primitive word has at least one unbordered conjugate. Unbordered conjugates of a word are closely related to the critical points of this word [13].

The problem has been extensively studied when the alphabet is ordered. In this case, symbols can be compared not only using equality ($=$) or inequality ($\neq$) comparators but also using ordering comparators ($<, \leqslant, >, \geqslant$). This order naturally extends to the lexicographic order on words defined as follows: $u < v$ for $u, v \in A^*$ if u is a proper prefix of v or if

* Corresponding author.
E-mail addresses: jean-pierre.duval@univ-rouen.fr (J.-P. Duval), thierry.lecroq@univ-rouen.fr (T. Lecroq), arnaud.lefebvre@univ-rouen.fr (A. Lefebvre).

0304-3975/$ – see front matter
http://dx.doi.org/10.1016/j.tcs.2013.12.008

$u = xay$ and $v = xbz$ with $a, b \in A$, $x, y, z \in A^*$ and $a < b$. In this setting, the first linear algorithm that computes the least conjugate of a word is due to Booth [4]. Several refinements were proposed (see [17,8,2]). These results are linked to the factorization of a given word in Lyndon words (see [11,8]). Lyndon words are words that are lexicographically less than all their conjugates, and thus that are unbordered.

In this article we consider words build over an unordered alphabet meaning that symbols can only be compared using equality ($=$) or inequality ($\neq$) comparators and that the lexicographic order on words cannot be used. Unordered alphabets arise when symbols are objects such as ideograms for instance. In this model, no linear time algorithm is known for computing either an unbordered conjugate of a primitive word or a conjugate that is a power of an unbordered word. In this paper we propose such an algorithm.

We first give some basic notions in Section 2. Then, in Section 3, we briefly present our main result that is based on a canonical decomposition presented in Section 4. Finally, in Section 5, we present the algorithm together with its correctness and complexity proofs.

2. Preliminaries

Let A be a set of letters and A^* be the set of all finite words over A. The length of word $w = w[1..n]$ is denoted by $|w| = n$, and its ith letter by $w[i]$ for all $1 \leqslant i \leqslant n$. The word of length 0, called empty word, is denoted by ε. A word of the form $w[1..i]$, for all $1 \leqslant i \leqslant n$, is called a *prefix* of w. A word of the form $w[i..n]$, for all $1 \leqslant i \leqslant n$, is called a *suffix* of w. A prefix (resp. suffix) different from ε and w is called a *proper* prefix (resp. suffix) of w. The concatenation of two words x and y is denoted by $x \cdot y$. Given a word w of length n, the infinite *circular* word $circ(w)$ is defined by $circ(w)[i] = w[i]$, for $1 \leqslant i \leqslant n$, and $circ(w)[i] = w[i \bmod n]$, for $i > n$. One can remark that any factor of length n of $circ(w)$ is a conjugate of w.

Let $\mathtt{Pref}(w)$ denote the set of proper prefixes of word w. $\mathtt{Pref}^*(w)$ is the set of words obtained by concatenations of elements of $\mathtt{Pref}(w)$ and $\mathtt{Pref}^+(w) = \mathtt{Pref}^*(w) \backslash \{\varepsilon\}$. We denote $y \in \mathtt{Pref}(w)$ by $y \prec_{pref} w$. Let $\mathtt{Suff}(w)$ be the set of proper suffixes of w. Any word in $\mathtt{Pref}(w) \cap \mathtt{Suff}(w)$ is called a *border* of w.

Definition 1. A word w is *unbordered* iff $\mathtt{Pref}(w) \cap \mathtt{Suff}(w) = \emptyset$.

Definition 2 *(Purity).* A word is periodically pure (pure for short) iff it is unbordered or it is the integer power of an unbordered word.

Let us now introduce a relation between words.

Definition 3. Given two words x and y, $y \ll x$ iff the following conditions hold:

- y is unbordered;
- $|y| \leqslant |x|$;
- the prefix of x of length $|y|$ is a concatenation of prefixes of y.

This definition leads to the following propositions.

Proposition 1. *Given three words $x, y, z \in A^*$, conditions $y \ll x$ and $x \prec_{pref} z$ imply $y \ll z$.*

Proof. The prefix of z of length $|y|$ is a prefix of x. Since $y \ll z$, this prefix is a concatenation of proper prefixes of y, then $y \ll z$. □

Proposition 2. *Relation $\ll$ is transitive.*

Proof. Let $z \ll y$ and $y \ll x$. The prefix of y of length $|z|$ is a concatenation of prefixes of z. The prefix of x of length $|y|$ is a concatenation of prefixes of y. Then the prefix of x of length $|z|$ is a concatenation of prefixes of z. □

Proposition 3. *Given u an unbordered word and $v \in \mathtt{Pref}^+(u)$. The following properties hold*:

- *$v \cdot u$ is an unbordered word*;
- *$u \ll v \cdot u$.*

Proof. Let $v \cdot u \in \mathtt{Pref}^+(u) \cdot u$. A proper prefix y of $v \cdot u$ is in $\mathtt{Pref}^+(u)$, and $y = y' \cdot v'$ with $v' \in \mathtt{Pref}(u)$. Since u is unbordered, v' cannot be a suffix of u and $y' \cdot v'$ cannot be a border of $v \cdot u$. Then $v \cdot u$ is unbordered.

The prefix of $v \cdot u$ of length $|u|$ is in $\mathtt{Pref}^+(u)$ then $u \ll v \cdot u$. □

Corollary 1. *Given an unbordered word u and $v \in \mathtt{Pref}^+(u)$, for all integer $q \geqslant 1$, $v \cdot u^q$ is unbordered.*