Discrete Applied Mathematics 171 (2014) 141–146

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Note

Abelian borders in binary words

Manolis Christodoulakis [a], Michalis Christou [b,*], Maxime Crochemore [b,c], Costas S. Iliopoulos [b,d]

[a] University of Cyprus, Cyprus
[b] King's College London, UK
[c] Université Paris-Est, France
[d] Curtin University, Digital Ecosystems & Business Intelligence Institute, Center for Stringology & Applications, Australia

ARTICLE INFO

Article history:
Received 22 October 2012
Received in revised form 3 February 2014
Accepted 14 February 2014
Available online 6 March 2014

Keywords:
Strings
Abelian
Borders
Periods

ABSTRACT

In this article we study the appearance of abelian borders in binary words, a notion closely related to the abelian period of a word. We show how many binary words have shortest border of a given length by identifying relations with Dyck words. Furthermore, we give some bounds on the number of abelian border-free words of a given length and on the number of abelian words of a given length that have at least one abelian border. Finally, using some techniques employed in a recent paper by Christodoulakis et al. (2013), we show that there exists an algorithm that finds the shortest abelian border of a binary word that is not abelian border-free in $\Theta(\sqrt{n})$ time on average.

1. Introduction

Abelian periodicity has been extensively studied over the last years. Abelian periods are more flexible than classical ones and are defined in terms of Parikh vectors as in [9]. The Parikh vector of a string x, denoted by $\mathcal{P}_x$, enumerates the number of occurrences of each letter of Σ in x.

In 2006 Constantinescu and Ilie [9] proved a variant of Fine and Wilf's theorem for abelian periods of strings, later extended for abelian periods in partial words [2]. Early efficient algorithms for abelian pattern matching were given in [10,11] and later some linear-time algorithms have been designed in [4,5,8]. Recently, Fici et al. [12] gave five algorithms for the computation of all abelian periods of a string. They have proposed two offline algorithms, a brute force algorithm and one that uses a select array, that run in time $O(|x|^2|\Sigma|)$, and three online algorithms, where the first two run in time $O(|x|^3|\Sigma|)$ and the other one runs in time $O(|x|^3 \log(|x|)|\Sigma|)$. Christou et al. [7] gave two $O(|x|^2)$ time algorithms for the computation of all abelian periods of a string x by mapping factors of the string to a unique number depending on the letters that compose it. They have also defined weak abelian periods on strings and gave a $O(|x| \log(|x|))$ time algorithm for their computation.

In this article, we study the appearance of abelian borders in binary words. First, we investigate the number of binary words whose shortest border has a given length, by identifying relations with Dyck words. Next, we give some bounds on the number of abelian border-free words of a given length and on the number of abelian words of a given length that have at least one abelian border. Finally, using some techniques employed by Christodoulakis et al. in [6], we provide an algorithm that finds the shortest abelian border of a non-abelian-border-free binary word in time $\Theta(\sqrt{n})$ on average. We would like to

* Corresponding author. Tel.: +44 35799547450; fax: +44 35725373719.
E-mail addresses: christodoulakis.manolis@ucy.ac.cy (M. Christodoulakis), michalis.christou@kcl.ac.uk (M. Christou), Maxime.Crochemore@kcl.ac.uk (M. Crochemore), csi@dcs.kcl.ac.uk (C.S. Iliopoulos).

http://dx.doi.org/10.1016/j.dam.2014.02.012
0166-218X/

mention that while our paper was under review the work of Rampersad et al. [14] was published. They show the connection of abelian unbordered words with irreducible symmetric Motzkin paths and give expressions for their number in a different manner than us. Furthermore, they also comment on the lengths of the abelian unbordered factors of the Thue–Morse word.

2. Definitions

We define an *alphabet* Σ as a finite, non-empty set of symbols. An ordering can be defined via a bijection $\phi : \Sigma \to \{1, 2, \ldots, \sigma\}$, where $|\Sigma| = \sigma$. Throughout this article we consider a word x composed by letters drawn from an *alphabet* $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$. It is represented as $x[1..n]$. A string w is a *factor* of x if $x = uwv$ for two strings u and v. It is a *prefix* of x if u is empty and a *suffix* of x if v is empty. A string u is a *border* of x if u is both a proper prefix and a suffix of x. A *proper* factor of x is a factor which is not equal to x itself; *proper* prefixes, suffixes and borders are defined similarly. A string u is a *period* of x, if both u is a prefix of x and x is a prefix of u^e for some positive integer e (i.e. x is a prefix of ux). *The period* of x, denoted by *Period*(x), is the length of the shortest period of x.

Definitions relative to Parikh vectors are as in [9,12]. The Parikh vector of a string x, denoted by $\mathcal{P}_x$, enumerates the number of times each letter of Σ occurs in x. That is $\mathcal{P}_x[i]$ is the number of occurrences of a_i in x, where $1 \le i \le \sigma$. The sum of the components of a Parikh vector is denoted by $|\mathcal{P}|$. Given two Parikh vectors $\mathcal{P}, \mathcal{Q}$ we write $\mathcal{P} \subseteq \mathcal{Q}$ if $\mathcal{P}[i] \le \mathcal{Q}[i]$, for every $1 \le i \le \sigma$ and $|\mathcal{P}| \le |\mathcal{Q}|$.

The string x is said to have an *abelian period* (h, p) if $x = u_0u_1 \ldots u_{k-1}u_k$ such that: $\mathcal{P}_{u_0} \subseteq \mathcal{P}_{u_1} = \cdots = \mathcal{P}_{u_{k-1}} \supseteq \mathcal{P}_{u_k}$, $|\mathcal{P}_{u_0}| = h$ and $|\mathcal{P}_{u_1}| = p$.

Factors u_0 and u_k are called the *head* and the *tail* of the abelian period respectively. Moreover, x is said to have a *weak abelian period* p if $|\mathcal{P}_{u_0}| = |\mathcal{P}_{u_1}| = p$.

A string u of length $|u| = m < n$ is an *abelian border* of x if $\mathcal{P}_y = \mathcal{P}_{x[1..m]} = \mathcal{P}_{x[n-m+1..n]}$. A string that has only the empty abelian border is called an *abelian border-free* string.

A *Dyck* word of length $2n$ is a binary string consisting of n zeros and n ones such that no prefix of the string has more ones than zeros. It is known that Catalan numbers enumerate Dyck words [13]. The *nth Catalan number* is given in terms of binomial coefficients:

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!\,n!} = \prod_{k=2}^{n} \frac{n+k}{k} \quad \text{for } n \ge 0.$$

3. Abelian borders in binary words

Let W_n denote the set of binary words of length n, and S_n denote the subset of W_n having no abelian borders. For small values of n, the sets S_n can be easily identified as:

$$S_1 = \{0, 1\}, \qquad S_2 = \{01, 10\}, \qquad S_3 = \{001, 011, 100, 110\},$$
$$S_4 = \{0001, 0011, 0111, 1000, 1100, 1110\}.$$

Similarly, we denote by S'_n the complementary set of S_n, the set of binary words of length n having at least one abelian border. The first 3 sets are:

$$S'_2 = \{00, 11\}, \qquad S'_3 = \{000, 010, 101, 111\},$$
$$S'_4 = \{0000, 0010, 0100, 0110, 1001, 1011, 1101, 1111, 0101, 1010\}.$$

The following lemma implies some elementary properties of abelian borders, such as that the shortest abelian border has length at most $\lfloor \frac{n}{2} \rfloor$ and that the longest abelian border has length at least $\lceil \frac{n}{2} \rceil$.

Lemma 1 ([6]). *For every abelian border u of a word $x[1..n]$, of length $|u| \neq \frac{n}{2}$, there exists one more abelian border u' of x of length $n - |u|$.*

In the following lemma, we establish the relation of abelian borders to Dyck words. We will need the following definition; given a binary word x of length $n > 2$, the ternary word y_x, $1 \le |y_x| \le \lfloor \frac{n}{2} \rfloor$ is defined as:

$$y_x[i] = \begin{cases} a, & \text{if } x[i] = x[n+1-i] \\ b, & \text{if } x[i] = 0 \text{ and } x[n+1-i] = 1 \\ c, & \text{if } x[i] = 1 \text{ and } x[n+1-i] = 0. \end{cases}$$

Lemma 2. *A binary word x of length n has a shortest abelian border of length k, $2 \le k \le \lfloor \frac{n}{2} \rfloor$, iff $y_x[1..k]$ is the shortest prefix of y_x that contains a Dyck word (or its bitwise negation) of length $0 < 2h \le k$ as a subsequence.*