Discrete Applied Mathematics 157 (2009) 890–900

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Unordered partial words☆

F. Blanchet-Sadri[a,*], C.D. Davis[a], Joel Dodge[b], Robert Mercaş[c,d], Margaret Moorefield[a]

[a] Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402–6170, USA
[b] Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, Dept 0112, LaJolla, CA 92093–0112, USA
[c] GRLMC, Universitat Rovira i Virgili, Plaça Imperial Tárraco, 1, Tarragona, 43005, Spain
[d] MOCALC Research Group, Faculty of Mathematics and Computer Science, University of Bucharest, Academiei, 14, 010014, Bucharest, Romania

ARTICLE INFO

Article history:
Received 24 March 2006
Received in revised form 31 March 2008
Accepted 4 April 2008
Available online 21 May 2008

Keywords:
Words
Partial words
Unbordered words
Unbordered partial words

ABSTRACT

An *unbordered* word is a string over a finite alphabet such that none of its proper prefixes is one of its suffixes. In this paper, we extend the results on unbordered words to unbordered partial words. Partial words are strings that may have a number of "do not know" symbols. We extend a result of Ehrenfeucht and Silberger which states that if a word u can be written as a concatenation of nonempty prefixes of a word v, then u can be written as a unique concatenation of nonempty unbordered prefixes of v. We study the properties of the longest unbordered prefix of a partial word, investigate the relationship between the minimal *weak* period of a partial word and the maximal length of its unbordered factors, and also investigate some of the properties of an unbordered partial word and how they relate to its critical factorizations (if any).

1. Introduction

Periodicity and *borderedness* are two fundamental properties of words that play a role in several research areas including string searching algorithms [9–11,14], data compression [16], theory of codes [3], sequence assembly [15] and superstrings [7] in computational biology, and serial data communication systems [8]. It is well known that these two word properties do not exist independently from each other.

Let A be a nonempty finite set, also called an *alphabet*. Consider a nonempty word $u = a_0a_1 \dots a_{n-1}$ with $a_i \in A$. Then a *period* of u is a positive integer p such that $a_i = a_{i+p}$ for $0 \leq i < n-p$. The word u is called *bordered* if one of its proper prefixes is one of its suffixes. The length of the longest such prefix (also called longest border) is the length of u minus the length of its shortest period. The word u is called *unbordered* otherwise. In other words, it is *unbordered* if it has no proper period. For example, *abaabb* is unbordered while *abaab* is bordered. Unbordered words turn out to be primitive, that is, they cannot be written as a power of another word. Unborderedness has the following important property: Different occurrences of an unbordered factor u in a word v never overlap. A related property is that no primitive word u can be an inside factor of uu. Fast algorithms for testing primitivity of words can be based on this property [10].

The study of unbordered partial words was initiated in [4]. Partial words are strings that may have a number of "do not know" symbols. In this paper, we pursue this study by extending some more results on unbordered words to unbordered partial words. We begin in Section 2 by reviewing basic concepts on words and partial words. In Section 3, we recall a result of Ehrenfeucht and Silberger [13] which states that if a word u can be written as a concatenation of nonempty prefixes of a

☆ This material is based upon work supported by the National Science Foundation under Grant Nos. CCF–0207673 and DMS–0452020. We thank the referees of a preliminary version of this paper for their very valuable comments and suggestions.

* Corresponding author.
E-mail address: blanchet@uncg.edu (F. Blanchet-Sadri).

0166-218X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2008.04.004

word v, then u can be written as a unique concatenation of nonempty unbordered prefixes of v, and we extend this result to partial words. In Section 4, we give more results on concatenations of prefixes. In particular, we study the properties of the longest unbordered prefix of a partial word. We also investigate the relationship between the minimal *weak* period of a partial word and the maximal length of its unbordered factors. In Section 5, we investigate some of the properties of an unbordered partial word and how they relate to its critical factorizations (if any). Blanchet-Sadri and Wetzler extended the well-known critical factorization theorem to partial words and their result states that the minimal weak period of a *nonspecial* partial word can be locally determined in at least one position [6]. Finally, we prove that, with regard to Chomsky hierarchy, the set of all partial words over an arbitrary nonunary fixed finite alphabet having a critical factorization is a context sensitive language that is not context-free.

2. Preliminaries

Fixing an alphabet A, we first review the basic concepts on words and partial words over A.

2.1. Words

A *string* or *word* u over A is a finite concatenation of symbols or letters from A. The number of symbols in u, or *length* of u, is denoted by $|u|$. For any word u, $u[i..j-1]$ is the *factor* of u that starts at position i and ends at position $j-1$ (it is called *proper* if $0 \leq i < j \leq |u|$ and ($i > 0$ or $j < |u|$)). In particular, $u[0..j-1]$ is the *prefix* of u that ends at position $j-1$ and $u[i..|u|-1]$ is the *suffix* of u that begins at position i. The factor $u[i..j-1]$ is the empty word if $i \geq j$ (the empty word is denoted by ε). The set of all finite length words over A (length greater than or equal to zero) is denoted by A^*. It is a monoid under the associative operation of concatenation or product of words where ε serves as the identity, and it is referred to as the *free monoid* generated by A. Similarly, the set of all nonempty words over A is denoted by A^+. It is a semigroup under the operation of concatenation of words and is referred to as the *free semigroup* generated by A.

For a word u, the powers of u are defined inductively by $u^0 = \varepsilon$ and, for any $n \geq 1$, $u^n = uu^{n-1}$. If u is nonempty, then v is a *root* of u if $u = v^n$ for some positive integer n. The shortest root of u, denoted by $\sqrt{u}$, is called the *primitive root* of u, and u is itself called *primitive* if $\sqrt{u} = u$. If $u = (\sqrt{u})^n$, then $\sqrt{u}$ is the unique primitive word v and n is the unique positive integer such that $u = v^n$. All positive powers of u have the same primitive root.

A word of length n over A can be defined by a total function $u : \{0, \ldots, n-1\} \to A$ and is usually represented as $u = a_0a_1 \ldots a_{n-1}$ with $a_i \in A$. A positive integer p is a *period* of u if for all $0 \leq i < n-p$ we have $a_i = a_{i+p}$. This can be equivalently formulated, for $p \leq n$, by $u = xv = wx$ for some words x, v, w satisfying $|v| = |w| = p$. For a word u, there exists a *minimal period* which is denoted by $p(u)$. A nonempty word u is *unbordered* if $p(u) = |u|$. Otherwise, it is *bordered*. A nonempty word x is a *border* of a word u if $u = xv = wx$ for some nonempty words v and w. Unbordered words turn out to be primitive.

2.2. Partial words

A *partial word* u of length n over A is a partial function $u : \{0, \ldots, n-1\} \to A$. For $0 \leq i < n$, if $u(i)$ is defined, then we say that i belongs to the *domain* of u, denoted by $i \in D(u)$, otherwise we say that i belongs to the *set of holes* of u, denoted by $i \in H(u)$. A (*full*) word over A is a partial word over A with an empty set of holes.

For convenience, we will refer to a partial word over A as a word over the enlarged alphabet $A_\diamond = A \cup \{\diamond\}$, where $\diamond$ represents a "do not know" symbol. So a partial word u of length n over A can be viewed as a total function $u : \{0, \ldots, n-1\} \to A \cup \{\diamond\}$ where $u(i) = \diamond$ whenever $i \in H(u)$. For example, $u = a \diamond bbc \diamond cb$ is a partial word of length 8 where $D(u) = \{0, 2, 3, 4, 6, 7\}$ and $H(u) = \{1, 5\}$. We can thus define for partial words concepts such as concatenation, powers, etc. in a trivial way.

The length of a partial word u over A is denoted by $|u|$, while the set of distinct letters in A occurring in u is denoted by $\alpha(u)$. For the set of all partial words over A with an arbitrary number of holes we write $A_\diamond^*$. The set $A_\diamond^*$ is a monoid under the operation of concatenation where ε serves as the identity element. If $X \subset A_\diamond^*$, then the *cardinality* of X is denoted by $\| X \|$.

For partial words, we use the same notions of prefix, suffix and factor, as for full ones. The unique *maximal common prefix* of u and v will be denoted by $\mathrm{pre}(u, v)$. Now, if $u \in A_\diamond^*$ and $0 \leq i < j \leq |u|$, then $u[i..j-1]$ denotes the factor $u(i) \ldots u(j-1)$. For a subset X of $A_\diamond^*$, we denote by $P(X)$ the set of prefixes of elements in X and by $S(X)$ the set of suffixes of elements in X. If X is the singleton $\{u\}$, then $P(X)$ (respectively, $S(X)$) will be abbreviated by $P(u)$ (respectively, $S(u)$).

A *factorization* of a partial word u is any tuple $(u_0, u_1, \ldots, u_{i-1})$ of partial words such that $u = u_0u_1 \ldots u_{i-1}$. For a subset X of $A_\diamond^*$ and an integer $i \geq 0$, the set

$$\{u_0u_1 \ldots u_{i-1} \mid u_0, \ldots, u_{i-1} \in X\}$$

is denoted by X^i. The submonoid of $A_\diamond^*$ generated by X will be denoted by X^* where $X^* = \bigcup_{i \geq 0} X^i$ and $X^0 = \{\varepsilon\}$. The subsemigroup of $A_\diamond^*$ generated by X is denoted by X^+ where $X^+ = \bigcup_{i>0} X^i$. By definition, each partial word u in X^* admits at least one factorization $(u_0, u_1, \ldots, u_{i-1})$ whose elements are all in X. Such a factorization is called an *X-factorization*.