Journal of Discrete Algorithms 16 (2012) 113–128

Contents lists available at SciVerse ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Periods in partial words: An algorithm ☆

F. Blanchet-Sadri [a,∗], Travis Mandel [b], Gautam Sisodia [c]

[a] Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402-6170, USA
[b] Department of Mathematics, The University of Texas at Austin, 1 University Station, C1200, Austin, TX 78712, USA
[c] Department of Mathematics, University of Washington, P.O. Box 354350, Seattle, WA 98195-4350, USA

ARTICLE INFO

Article history:
Available online 3 April 2012

Keywords:
Automata and formal languages
Combinatorics on words
Partial words
Fine and Wilf's theorem
Strong periods
Graph connectivity
Optimal lengths

ABSTRACT

Partial words are finite sequences over a finite alphabet that may contain some holes. A variant of the celebrated Fine–Wilf theorem shows the existence of a bound $L = L(h, p, q)$ such that if a partial word of length at least L with h holes has periods p and q, then it also has period $\gcd(p, q)$. In this paper, we associate a graph with each p- and q-periodic word, and study two types of vertex connectivity on such a graph: modified degree connectivity and r-set connectivity where $r = q \bmod p$. As a result, we give an algorithm for computing $L(h, p, q)$ in the general case and show how to use it to derive the closed formulas.

1. Introduction

The problem of computing periods in *words*, or finite sequences of symbols from a finite alphabet, has important applications in several areas including data compression, coding, computational biology, string searching and pattern matching algorithms. Repeated patterns and related phenomena in words have played over the years a central role in the development of combinatorics on words [5], and have been highly valuable tools for the design and analysis of algorithms. In many practical applications, such as DNA sequence analysis, repetitions admit a certain variation between copies of the repeated pattern because of errors due to mutation, experiments, etc. Approximate repeated patterns, or repetitions where errors are allowed, are playing a central role in different variants of string searching and pattern matching problems [13]. *Partial words*, or finite sequences that may contain some holes, have acquired importance in this context. A (*strong*) *period* of a partial word u over an alphabet A is a positive integer p such that $u(i) = u(j)$ whenever $u(i), u(j) \in A$ and $i \equiv j \bmod p$ (in such a case, we call u p-*periodic*). In other words, p is a period of u if for all positions i and j congruent modulo p, the letters in these positions are the same or at least one of these positions is a hole. For example, the word *aabaabaa* has period 3 but not 4, while the partial word $a\diamond\diamond aabaa$, with holes at positions 1 and 2, has periods 3 and 4 (note that our words are starting at position 0 rather than 1).

There are many fundamental results on periods of words. Among them is the well-known periodicity result of Fine and Wilf [8], which determines how long a p- and q-periodic word needs to be in order to also be $\gcd(p, q)$-periodic. More precisely, any word having two periods p, q and length at least $p + q - \gcd(p, q)$ has also $\gcd(p, q)$ as a period. Moreover, the length $p + q - \gcd(p, q)$ is optimal since counterexamples can be provided for shorter lengths, that is, there exists an *optimal* word of length $p + q - \gcd(p, q) - 1$ having p and q as periods but not having $\gcd(p, q)$ as period [5]. Extensions

☆ This material is based upon work supported by the National Science Foundation under Grant No. DMS-0452020. Part of this paper was presented at IWOCA 2011 [4]. 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).

1570-8667/$ – see front matter
http://dx.doi.org/10.1016/j.jda.2012.04.001

of Fine and Wilf's result to more than two periods have been given. For instance, in [6], Constantinescu and Ilie give an extension for an arbitrary number of periods and prove that their lengths are optimal.

Fine and Wilf's result has been generalized to partial words [1–3,10–12,14]. Some of these papers are concerned with *weak* periodicity, a notion not discussed in this paper (a *weak period* of a partial word u over an alphabet A is a positive integer p such that $u(i) = u(i+p)$ whenever $u(i), u(i+p) \in A$). The papers that are concerned with strong periodicity refer to the basic fact, proved by Shur and Konovalova (Gamzova) in [12], that for positive integers h, p and q, there exists a positive integer l such that a partial word u with h holes, two periods p and q, and length at least l has period $\gcd(p, q)$. The smallest such integer is called the optimal length and it will be denoted by $L(h, p, q)$. They gave a closed formula for the case where $h = 2$ (the cases $h = 0$ or $h = 1$ are implied by the results in [8,1]), while in [11], they gave a formula in the case where $p = 2$ as well as an optimal asymptotic bound for $L(h, p, q)$ in the case where h is "large." In [3], Blanchet-Sadri et al. gave closed formulas for the optimal lengths when q is "large," whose proofs are based on connectivity in the so-called (p, q)-periodic graphs. The *(p, q)-periodic graph of size l* is the graph $G = (V, E)$, with $V = \{0, 1, \ldots, l-1\}$, such that $\{i, j\} \in E$ if and only if $i \equiv j \bmod p$ or $i \equiv j \bmod q$.

In this paper, we study two types of vertex connectivity in the (p, q)-periodic graphs: the modified degree connectivity and r-set connectivity where $r = q \bmod p$. Although the graph-theoretical approach is not completely new, our paper gives insights into periodicity in partial words and provides an algorithm for determining $L(h, p, q)$ in *all* cases. Our paper also shows how the closed formulas can be derived from our methods.

We end this section by reviewing basic concepts on partial words. Fixing a nonempty finite set of letters or an *alphabet* A, finite sequences of letters from A are called *(full) words* over A. The number of letters in a word u, or *length* of u, is denoted by $|u|$. The unique word of length 0, denoted by ε, is called the *empty* word. 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 $u = a_0 a_1 \ldots a_{n-1}$ with $a_i \in A$. The set of all words over A of finite length (greater than or equal to zero) is denoted by A^*. A *partial word* u of length n over A is a partial function $u : \{0, \ldots, n-1\} \to A$. For $0 \leqslant i < n$, if $u(i)$ is defined, then i belongs to the *domain* of u, denoted by $i \in D(u)$, otherwise i belongs to the *set of holes* of u, denoted by $i \in H(u)$. The set of distinct letters of A occurring in u is denoted by $\alpha(u)$. For convenience, we will refer to a partial word over A as a word over the enlarged alphabet $A_\diamond = A \cup \{\diamond\}$, where $\diamond \notin A$ represents a "do not know" symbol or hole. So a partial word u of length n over A can be viewed as a total function $u : \{0, \ldots, n-1\} \to A_\diamond$ where $u(i) = \diamond$ whenever $i \in H(u)$.

2. (p, q)-Periodic graphs

In this section, we discuss the fundamental property of periodicity, our goal, and some initial results. We can restrict our attention to the case where p and q are coprime, that is $\gcd(p, q) = 1$, since it is well known that the general case can be reduced to the coprime case (see, for example, [1,11]). Also, we assume without loss of generality that $1 < p < q$.

Fine and Wilf show that $L(0, p, q) = p + q - \gcd(p, q)$ [8], Berstel and Boasson that $L(1, p, q) = p + q$ [1], and Shur and Konovalova prove $L(2, p, q)$ to be $2p + q - \gcd(p, q)$ [12]. Other results include the following.

Theorem 1. *(See [3,11].) Let $q > 2$ be an integer satisfying $\gcd(2, q) = 1$. Then*

$$L(h, 2, q) = h + q\left(1 + \left\lfloor \frac{h}{q} \right\rfloor\right) + 1.$$

Theorem 2. *(See [3].) Let p and q be integers satisfying $1 < p < q$ and $\gcd(p, q) = 1$. If $q > p\lfloor \frac{h+1}{2} \rfloor$, then*

$$L(h, p, q) = \begin{cases} p(\frac{h+2}{2}) + q - 1, & \text{if } h \text{ is even;} \\ p(\frac{h+1}{2}) + q, & \text{if } h \text{ is odd.} \end{cases}$$

The problem of finding $L(h, p, q)$ is equivalent to a problem involving the vertex connectivity of certain graphs, as described in [3], which we now discuss.

Definition 1. Let p and q be integers satisfying $1 < p < q$ and $\gcd(p, q) = 1$. The *(p, q)-periodic graph* of size l is the graph $G = (V, E)$ where $V = \{0, 1, \ldots, l-1\}$ and for $i, j \in V$, the pair $\{i, j\} \in E$ if and only if $i \equiv j \bmod p$ or $i \equiv j \bmod q$.

The *p-class of vertex i* is $\{j \in V \mid j \equiv i \bmod p\}$. A *$p$-connection* (or *$p$-edge*) is an edge $\{i, j\} \in E$ such that $i \equiv j \bmod p$. If an edge $\{i, j\}$ is a p-connection, then i and j are *p-connected*. Similar statements hold for q-classes, q-connections and pq-classes, pq-connections.

Fig. 1 illustrates a (p, q)-periodic graph.

The (p, q)-periodic graph G of size l can be thought to represent a full word u of length l with periods p and q as well as a partial word w with h holes of length l with periods p and q. Key observations are:

- Positions in u correspond to vertices in G.