Periods and Binary Words

Vesa Halava

Turku Centre for Computer Science, TUCS
FIN-20520, Turku, Finland
e-mail: vehalava@cs.utu.fi

Tero Harju

Department of Mathematics, University of Turku
FIN-20014, Turku, Finland
e-mail: harju@utu.fi

Lucian Ilie

Turku Centre for Computer Science, TUCS
FIN-20520, Turku, Finland
e-mail: lucili@utu.fi

Research supported by
the Academy of Finland, Project 137358.

On leave of absence from
Faculty of Mathematics, University of Bucharest,
Str. Academiei 14, R-70109 Bucharest, Romania

Turku Centre for Computer Science
TUCS Technical Report No 213
November 1998

ISBN 952-12-0313-7
ISSN 1239-1891

Abstract

We give an elementary short proof for a well-known theorem of Guibas and Odlyzko stating that the sets of periods of words are independent of the alphabet size. As a consequence of our constructing proof, we give a linear time algorithm which, given a word, computes a binary one with the same periods. We give also a very short proof for the famous Fine and Wilf's periodicity lemma.

Keywords: word, period, binary image, binary word, Fine and Wilf's lemma

TUCS Research Group

Theory Group: Mathematical Structures in Computer Science

1. Introduction and basic definitions

Let A be a finite alphabet of at least two letters and A^* the set of all words over A. For $w \in A^*$, $|w|$ denotes the length of w and w_i its ith letter. An integer $p, 1 \le p \le |w| - 1$ is called a *period* of w if $w_i = w_{i+p}$, for any $1 \le i \le |w| - p$. The set of all periods of w is denoted by $\mathcal{P}(w)$. Notice that $\mathcal{P}(w) = \emptyset$ if and only if w is unbordered.

The notion of period of a word is very central in the theory of combinatorics on words. There are many beautiful results on periods of words. Among them is a well-known theorem of Guibas and Odlyzko which states that the sets of periods of words are independent of the alphabet size. (Unary alphabets are, of course, out of discussion.) Put otherwise, it says that, for every word w, there exists a binary one, say w', such that $\mathcal{P}(w') = \mathcal{P}(w)$; w' will be called a *binary image* of w.

The proof given by [GuOd] to this unexpected result uses properties of the correlation and is very complicated. In this note, we give an elementary short proof for this theorem. As the proof is constructive, we give also a fast algorithm which computes a binary image of a given word. The algorithm runs in linear time, so it is optimal. We shall give also a very short proof (the shortest to our knowledge) for the famous Fine and Wilf's periodicity lemma, cf. [FiWi].

We shall denote by ε the empty word. For two words u, v, we say that u is a *prefix* of v, denoted $u \le v$, if $v = ux$, for some $x \in A^*$. A word u is *primitive* if there is no word v such that $u = v^k$, where $k \ge 2$.

For basic notions and results on words we refer to [ChKa] and [Lo].

2. Properties of words and periods

In this section we give first the announced proof for Fine and Wilf's lemma and then prove some properties of words and periods needed in the proof of the main theorem in the next section.

Lemma 1. *If a word w has periods p and q and $|w| = p + q - \gcd(p, q)$, then w has also period $d = \gcd(p, q)$.*

Proof. By induction on $n = |w|$. The first steps are trivial. Suppose the statement holds for all words shorter than w. Assume $p > q$ and put $w = uv$, $|u| = p - d$. For any $1 \le i \le q - d$, we have $u_i = w_i = w_{i+p} = w_{i+p-q} = u_{i+p-q}$, so u has period $p - q$. Since u has also period q and $\gcd(p - q, q) = d$, the inductive hypothesis shows that u has period d. Thus w has period d, too. ∎

Next lemma gives us the structure of the set of periods. We call the minimum $p \in \mathcal{P}(w)$, the *basic period* of w. For consistency, we take $p = 0$ when $\mathcal{P}(w) = \emptyset$.

Lemma 2. *Let $w \in A^*$ and $p \in \mathcal{P}(w)$ be the basic period of w. Then, for any $q \in \mathcal{P}(w)$ with $q \leq |w| - p$, q is a multiple of p.*

Proof. As $p + q \leq |w|$, we get by Lemma 1 that $\gcd(p, q) \in \mathcal{P}(w)$. As p is the basic period, we must have $p = \gcd(p, q)$, so $p|q$. ∎

As a corollary we get that if the basic period satisfies $p \leq |w|/2$, then the set of periods can be partitioned into two sets, the first one including the basic period p and all of its multiples and the second one including all the periods $q > |w| - p$.

Lemma 3. *Let $w \in \{0, 1\}^*$. Then there exists $a \in \{0, 1\}$ such that wa is primitive.*

Proof. Assume $w0 = v^k, w1 = u^l$, for some primitive u, v and $k, l \geq 2$. Clearly $|v| \neq |u|$ and assume $|v| < |u|$. Then v^k and u^l have a common prefix of length $l|u| - 1 \geq |u| + |v|$. By Lemma 1, $u = v$, a contradiction. ∎

3. Main theorem

Before the main theorem, we prove two more lemmata.

Lemma 4. *Let $w = (uv)^k u \in A^*$, where $k \geq 2$, $v \neq \varepsilon$, and $p = |uv|$ is the basic period of w. For any q with $|w| - p < q < |w|$, if we put $q = (k-1)p + r$, where $0 < r < p + |u|$, then $q \in \mathcal{P}(w)$ iff $r \in \mathcal{P}(uvu)$.*

Proof. For any $0 < i < |w| - q = p + |u| - r$, we have $w_i = (uvu)_i$ and $w_{i+q} = (uvu)_{i+r}$. Hence $w_i = w_{i+q}$ iff $(uvu)_i = (uvu)_{i+r}$ and we are done. ∎

Lemma 5. *Let $w = (uv)^k u \in A^*$, where $k \geq 1, v \neq \varepsilon$, and $p = |uv|$ is the basic period of w. If $u'v'u'$ is a binary image of uvu, where $|u'v'| = |uv|$, then $w' = (u'v')^k u'$ is a binary image of w.*

Proof. The case $k = 1$ is obvious. Assume $k \geq 2$. We show that $\mathcal{P}(w) = \mathcal{P}(w')$.

For any q with $0 < q \leq |w| - p$, Lemma 2 gives that $q \in \mathcal{P}(w)$ iff $p|q$, in which case also $q \in \mathcal{P}(w')$. If q is not a multiple of p, then $q \notin \mathcal{P}(w')$, as this would imply that the basic period of w' is strictly smaller than p, contradicting the fact that p is the basic period of w.

For any q with $|w| - p < q < |w|$, put $q = (k-1)p + r$, where $0 < r < p + |u|$. Then, by Lemma 4, $q \in \mathcal{P}(w)$ iff $r \in \mathcal{P}(uvu) = \mathcal{P}(u'v'u')$ which, in turn, is equivalent with $q \in \mathcal{P}(w')$. This completes the proof. ∎

Theorem 1. *For any alphabet A and any word $w \in A^*$, there exists a word $w' \in \{0,1\}^*$ such that $\mathcal{P}(w') = \mathcal{P}(w)$.*

Proof. By induction on $|w|$. If $|w| \leq 2$, then w is already binary.

Assume that the claim holds for all words of length less than or equal to $n \geq 2$. Let $w \in A^*$ with $|w| = n+1$ have p as its basic period. Put $w = (uv)^k u$, where $u, v \in A^*$, $k \geq 1$, $v \neq \varepsilon$, and $|uv| = p$. In the case $k \geq 2$ we have $|uvu| \leq n$ and, by the inductive hypothesis, we have $u'v'u'$ is a binary image of uvu. Now, by Lemma 5, $(u'v')^k u'$ is a binary image of w. Consider next the case $k = 1$. As $v \neq \varepsilon$, we have $|u| \leq n$ and thus, by the inductive hypothesis, there exists a binary image of u, say u'.

If $u = \varepsilon$ then $v' = 01^{|v|-1}$ is clearly a binary image of $v = w$. Otherwise, assume that u begins with the letter 0 and take $w' = u'1^{|v|-1}au'$, $a \in \{0,1\}$, such that $u'1^{|v|-1}a$ is primitive. Such an a can be found by Lemma 3. We shall prove that $\mathcal{P}(w') = \mathcal{P}(w)$. Clearly all periods of w are periods of w', since u' is a binary image of u. Assume that there is $q \in \mathcal{P}(w') - \mathcal{P}(w)$ and also that q is minimal with this property. Clearly, either $q < |u'|$ or $|u'| + |v| - 1 \leq q < |w|$, since u' does not begin with 1. If $q < |u'|$, then, by Lemma 2, the minimality of q implies $q|p$, and so $u'1^{|v|-1}a$ is not primitive, a contradiction.

It is possible that $q = |u'| + |v| - 1$ only if $a = 0$, in which case we get $u'1 = 0u'$, which is impossible. Therefore $q > p = |uv|$. Put $q = p + r, r > 0$. Then, clearly, r is a period of u' an hence of u. Lemma 4 implies $q \in \mathcal{P}(w)$, a contradiction. The theorem is proved. ∎

4. Algorithm

From the proof of Theorem 1, we get a recursive algorithm for constructing a binary image of a given word w, denoted below $Bin(w)$.

Bin(w)

1. Find the basic period p of w. If $p = 0$, then output $Bin(w) = 01^{|w|-1}$.
2. Find $u, v \in A^*$ and $k \geq 1$ such that $w = (uv)^k u$, where $v \neq \varepsilon$ and $|uv| = p$.
3. If $k \geq 2$ and $Bin(uvu) = u'v'u'$, $|u'v'| = |uv|$, then output $Bin(w) = (u'v')^k u'$.
4. Find $a \in \{0,1\}$ such that the word $Bin(u)1^{|v|-1}a$ is primitive and then output $Bin(w) = Bin(u)1^{|v|-1}aBin(u)$.

The correctness follows from the proof of Theorem 1. We finally consider the complexity of the algorithm. It is recursive, so let us compute the complexity of a single call of the procedure Bin. Assume that the length of the current word for this call, say x, is n. For Step 1, a pattern matching algorithm can be easily adapted to computing the basic period of x; just find the leftmost occurrence of x as a factor of $x\#^{|x|-1}$, where $\#$ is a symbol that passes all tests $\# \stackrel{?}{=} a$, $a \in A$. If there is no such occurrence, then $p = 0$. Thus Step 1 can be performed in time $\mathcal{O}(n)$. The same is obvious for Steps 2 and 3. At Step 4, it is known that the primitivity can be tested in linear time (x is primitive iff x is not a proper factor of x^2), so we have again $\mathcal{O}(n)$. Therefore, the complexity for one call is linear in terms of the length of the current word. But the length of the current word decreases from one call to another at least as fast as the function $\left(\frac{3}{4}\right)^n$. Thus, when we sum up the complexities for all the calls of Bin needed to compute $Bin(w)$ (logarithmically many), we get that the whole complexity of $Bin(w)$ is $\mathcal{O}(|w|)$. We have therefore proved

Theorem 2. *The algorithm $Bin(w)$ runs in linear time and therefore is optimal.*

References

[ChKa] C. Choffrut, J. Karhumäki, Combinatorics of Words, in G. Rozenberg, A. Salomaa, eds., *Handbook of Formal Languages, Vol. 1* (Springer-Verlag, Berlin, Heidelberg, 1997) 329 – 438.

[FiWi] N. J. Fine, H. S. Wilf, Uniqueness theorem for periodic functions, *Proc. Amer. Math. Soc.* **16** (1965) 109 – 114.

[GuOd] L. J. Guibas, A. M. Odlyzko, Periods in strings, *J. Combin. Theory, Ser A*, **30**(1) (1981) 19 – 42.

[Lo] M. Lothaire, *Combinatorics on Words* (Addison-Wesley, Reading, MA., 1983).

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

University of Turku
- **Department of Mathematical Sciences**

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**