Theoretical Computer Science 723 (2018) 11–22

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the number of gapped repeats with arbitrary gap

Roman Kolpakov [a,b,*]

[a] Moscow State University, Leninskie Gory, 119992 Moscow, Russia
[b] Dorodnicyn Computing Centre, FRC CSC RAS, Vavilov st. 40, 119333 Moscow, Russia

ARTICLE INFO

Article history:
Received 18 May 2017
Received in revised form 18 November 2017
Accepted 2 March 2018
Available online 15 March 2018
Communicated by J. Karhumäki

Keywords:
Combinatorics on words
Maximal number of repeats
Gapped repeats

ABSTRACT

For any functions $f(x)$, $g(x)$ from $\mathbb{N}$ to $\mathbb{R}$ we call repeats uvu such that $g(|u|) \le |v| \le f(|u|)$ as f, g-*gapped repeats*. We study the possible number of f, g-gapped repeats in words of fixed length n. For quite weak conditions on $f(x)$, $g(x)$ we obtain an upper bound on this number which is linear in n.

1. Introduction

Let $w = w[1]w[2]\dots w[n]$ be an arbitrary word of length $|w| = n$. A fragment $w[i]\cdots w[j]$ of w, where $1 \le i \le j \le n$, is called a *factor* of w and is denoted by $w[i..j]$. Note that this factor can be considered either as a word itself or as the fragment $w[i]\dots w[j]$ of w. So for factors we have two different notions of equality: factors can be equal as the same fragment of the word w or as the same word. To avoid this ambiguity, we use two different notations: if two factors u and v of w are the same word (the same fragment of w) we will write $u = v$ ($u \equiv v$). For any $i = 1, \dots, n$ the factor $w[1..i]$ ($w[i..n]$) is called a *prefix* (a *suffix*) of w. By positions in w we mean the order numbers $1, 2, \dots, n$ of letters of the word w. For any factor $v \equiv w[i..j]$ of w the positions i and j are called *start position* of v and *end position* of v and denoted by $\mathrm{beg}(v)$ and $\mathrm{end}(v)$ respectively. For any two factors u, v of w the factor u *is contained* in v if $\mathrm{beg}(v) \le \mathrm{beg}(u)$ and $\mathrm{end}(u) \le \mathrm{end}(v)$. If some word u is equal to a factor v of w then v is called *an occurrence* of u in w.

We denote by $p(w)$ the minimal period of a word w and by $e(w)$ the ratio $|w|/p(w)$ which is called the *exponent* of w. A word is called *primitive* if its exponent is not an integer greater than 1. A word is called *periodic* if its exponent is greater than or equal to 2. Occurrences of periodic words are called *repetitions*. Repetitions are fundamental objects, due to their primary importance in word combinatorics [22] as well as in various applications, such as string matching algorithms [13,5], molecular biology [14], or text compression [23]. The simplest and best known example of repetitions is factors of the form uu, where u is a nonempty word. Such repetitions are called *squares*. A square uu is called *primitive* if u is primitive. The questions on the number of squares and effective searching of squares in words are well studied in the literature (see, e.g., [5,4,15]).

A repetition in a word is called *maximal* if this repetition cannot be extended to the left or to the right in the word by at least one letter with preserving its minimal period. More precisely, a repetition $r \equiv w[i..j]$ in w is called *maximal* if it satisfies the following conditions:

* Correspondence to: Moscow State University, Leninskie Gory, 119992 Moscow, Russia.
E-mail address: foroman@mail.ru.

https://doi.org/10.1016/j.tcs.2018.03.007
0304-3975/

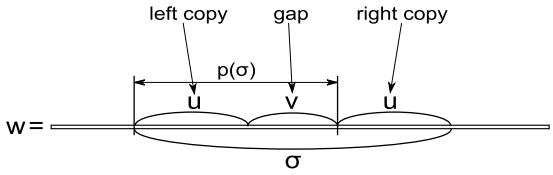

Fig. 1. A gapped repeat σ in w.

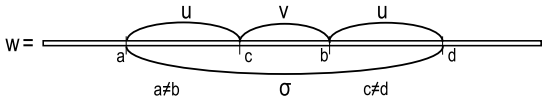

Fig. 2. A maximal gapped repeat σ in w.

1. if $i > 1$, then $w[i-1] \neq w[i-1+p(r)]$,
2. if $j < n$, then $w[j+1-p(r)] \neq w[j+1]$.

Maximal repetitions are usually called *runs* in the literature. Since runs contain all the other repetitions in a word, the set of all runs can be considered as a compact encoding of all repetitions in the word which has many useful applications (see, for example, [7]). For any word w we will denote by $\mathrm{E}(w)$ the sum of exponents of all maximal repetitions in w. The following bound for $\mathrm{E}(w)$ is proved in [16].

Theorem 1. $\mathrm{E}(w) = O(n)$ *for any* w.

More precise upper bounds on $\mathrm{E}(w)$ were obtained in [6,8,2].

A natural generalization of squares is factors of the form uvu where u and v are nonempty words. We call such factors *gapped repeats*. In the gapped repeat uvu the first (second) factor u is called *the left (right) copy*, and v is called *the gap*. By *the period* of this gapped repeat we will mean the value $|u| + |v|$. For a gapped repeat σ we denote the length of copies of σ by $c(\sigma)$ and the period of σ by $p(\sigma)$ (see Fig. 1). By (u', u'') we will denote the gapped repeat with the left copy u' and the right copy u''. Not that gapped repeats may form the same segment but have different periods. Such repeats are considered as distinct, i.e. a gapped repeat is not specified by its start and end positions in the word because these positions are not sufficient for determining the both copies and the gap of the repeat. Analogously to repetitions, a gapped repeat $(w[i'..j'], w[i''..j''])$ in w is called *maximal* if it satisfies the following conditions:

1. if $i' > 1$, then $w[i'-1] \neq w[i''-1]$,
2. if $j'' < n$, then $w[j'+1] \neq w[j''+1]$.

In other words, a gapped repeat in a word is maximal if its copies cannot be extended to the left or to the right in the word by at least one letter with preserving its period (see Fig. 2).

Let f, g be functions from $\mathbb{N}$ to $\mathbb{R}$ such that $0 < g(x) \leq f(x)$ for all $x \in \mathbb{N}$. We call a gapped repeat uvu *f, g-gapped repeat* if $g(|u|) \leq |v| \leq f(|u|)$. To our knowledge, maximal f, g-gapped repeats were firstly investigated in [3] where it was shown that for computed in constant time functions f, g all maximal f, g-gapped can be found in a word of length n with time complexity $O(n \log n + S)$ where S is the size of output. An algorithm for finding in a word all gapped repeats with a fixed gap length in time $O(n \log d + S)$ where d is the gap length, n is the word length, and S is the size of output was proposed in [17]. The f,g-gapped repeats naturally generalize gapped repeats σ such that $p(\sigma) \leq \alpha c(\sigma)$ for some $\alpha > 1$. Such gapped repeats, which can be considered as a particular case of f, g-gapped repeats for $f(x) = (\alpha - 1)x$ and $g(x) = \min\{1, \alpha - 1\}$, are called *$\alpha$-gapped repeats*. The notion of α-gapped repeats was introduced in [20] where it was proved that the number of maximal α-gapped repeats in a word of length n is bounded by $O(\alpha^2 n)$ and all maximal α-gapped repeats can be found in $O(\alpha^2 n)$ time for the case of integer alphabet. A new approach to computing α-gapped repeats was proposed in [11] where it was shown that the longest α-gapped repeat in a word of length n over an integer alphabet can be found in $O(\alpha n)$ time. This approach was further developed in [10] for computing of gapped repeats of various types in a word. In [24] an algorithm using an approach previously introduced in [1] is proposed for finding all maximal α-gapped repeats in $O(\alpha n + S)$ time where S is the output size, for a constant-size alphabet. Finally, in [9,12] an asymptotically tight $O(\alpha n)$ bound on the number of maximal α-gapped repeats in a word of length n was independently proved and, moreover, algorithms for finding of all maximal α-gapped repeats in $O(\alpha n)$ time were proposed.

For any real x denote

$$|x|^+ = \begin{cases} x, & \text{if } x > 0; \\ 0, & \text{otherwise;} \end{cases} \qquad |x|^- = \begin{cases} -x, & \text{if } x < 0; \\ 0, & \text{otherwise;} \end{cases}$$