Journal of Discrete Algorithms 46–47 (2017) 1–15

Contents lists available at ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Searching of gapped repeats and subrepetitions in a word

Roman Kolpakov[a,b,*], Mikhail Podolskiy[a], Mikhail Posypkin[b], Nickolay Khrapov[c]

[a] Lomonosov Moscow State University, Leninskie Gory, Moscow, 119992, Russia
[b] Dorodnicyn Computing Centre, FRC CSC RAS, Vavilov st., Moscow, 119333, Russia
[c] Institute for Information Transmission Problems, Bolshoy Karetny per., Moscow, 127994, Russia

ARTICLE INFO

Article history:
Received 2 June 2015
Received in revised form 27 September 2017
Accepted 14 October 2017
Available online 19 October 2017

Keywords:
Combinatorics on words
Combinatorial algorithms
Time complexity
Repetitions

ABSTRACT

A gapped repeat is a factor of the form uvu where u and v are nonempty words. The period of the gapped repeat is defined as $|u|+|v|$. The gapped repeat is maximal if it cannot be extended to the left or to the right by at least one letter with preserving its period. The gapped repeat is called α-gapped if its period is not greater than $\alpha|u|$. A δ-subrepetition is a factor whose exponent is less than 2 but is not less than $1+\delta$ (the exponent of the factor is the quotient of the length and the minimal period of the factor). The δ-subrepetition is maximal if it cannot be extended to the left or to the right by at least one letter with preserving its minimal period. We reveal a close relation between maximal gapped repeats and maximal subrepetitions. Moreover, we show that in a word of length n the number of maximal α-gapped repeats is bounded by $O(\alpha^2 n)$ and the number of maximal δ-subrepetitions is bounded by $O(n/\delta^2)$. Using the obtained upper bounds, we propose algorithms for finding all maximal α-gapped repeats and all maximal δ-subrepetitions in a word of length n (in assumption that the alphabet of the word is integer). The algorithm for finding all maximal α-gapped repeats has $O(\alpha^2 n)$ time complexity. For finding all maximal δ-subrepetitions we propose two algorithms. The first algorithm has $O(\frac{n \log\log n}{\delta^2})$ time complexity. The second algorithm has $O(n\log n + \frac{n}{\delta^2}\log\frac{1}{\delta})$ expected time complexity.

Introduction

Let $w = w[1]w[2]\dots w[n]$ be an arbitrary word. The length n of w is denoted by $|w|$. 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 for factors we have two different notions of equality: factors can be equal as the same fragment of the original word or as the same word. To avoid this ambiguity, we will use two different notations: if two factors u and v are the same word (the same fragment of the original word), 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. The factor v *covers* a letter $w[k]$ if $\mathrm{beg}(v) \le k \le \mathrm{end}(v)$. For any two factors u, v of w the factor u *is contained* (*is strictly contained*) in v if $\mathrm{beg}(v) \le \mathrm{beg}(u)$ and $\mathrm{end}(u) \le \mathrm{end}(v)$ (if $\mathrm{beg}(v) < \mathrm{beg}(u)$ and $\mathrm{end}(u) < \mathrm{end}(v)$). Let u, v be two factors of w such that $\mathrm{beg}(v) = \mathrm{end}(u) + 1$. In

[*] Corresponding author at: Lomonosov Moscow State University, Leninskie Gory, Moscow, 119992, Russia.
E-mail addresses: foroman@mail.ru (R. Kolpakov), mpodolskii@inbox.ru (M. Podolskiy), mposypkin@gmail.com (M. Posypkin), nkhrapov@gmail.com (N. Khrapov).

https://doi.org/10.1016/j.jda.2017.10.004
1570-8667/© 2017 Elsevier B.V. All rights reserved.

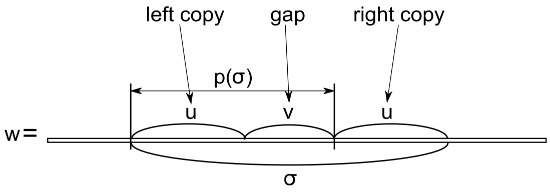

Fig. 1. A gapped repeat σ in w.

this case we say that v *follows* u. The number end(u) is called *the frontier* between the factors u and v. A factor v *contains* a frontier j if beg$(v) - 1 \le j \le$ end(v). If some word u is equal to a factor v of w then v is called *an occurrence* of u in w.

A positive integer p is called a *period* of w if $w[i] = w[i+p]$ for each $i = 1, \dots, n-p$. We denote by $p(w)$ the minimal period of 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. By repetition in a word we mean any factor of exponent greater than or equal to 2. Repetitions are fundamental objects, due to their primary importance in word combinatorics [23] as well as in various applications, such as string matching algorithms [12,4], molecular biology [14], or text compression [24]. 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*. We call the first (second) factor u of the square uu *the left (right) root* of this square. Avoiding ambiguity,[1] by *the period* of a square we mean the length of its roots. A square is called *primitive* if its roots are primitive. The questions concerned to squares are well studied in the literature. In particular, it is known (see, e.g., [4]) that a word of length n contains no more than $n\log_\varphi n$ primitive squares. In [3] an $O(n\log n)$-time algorithm for finding of all primitive squares in a word of length n is proposed. In [15] an algorithm for finding of all primitive squares in a word of length n with time complexity $O(n+S)$ where S is the size of output is proposed for the case of constant alphabet size.

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:

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, [9]). For any word w we will denote by $\mathcal{R}(w)$ the set of all maximal repetitions in w and by $\mathrm{E}(w)$ the sum of exponents of all maximal repetitions in w. The following facts are proved in [18].

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

Corollary 1. $|\mathcal{R}(w)| = O(n)$ *for any word* w *of length* n.

Moreover, in [18] an $O(n)$ time algorithm for finding of all runs in a word of length n is proposed for the case of constant alphabet size (in the case of arbitrary alphabet size all runs in a word of length n can be found in $O(n\log n)$ time). Further many papers were devoted to obtaining more precise upper bounds on $\mathrm{E}(w)$ and $|\mathcal{R}(w)|$ (see, e.g., [7,8]). Recently, a remarkable result is obtained in [1] where the so-called "runs conjecture" $|\mathcal{R}(w)| < n$ is proved. To our knowledge, at present time the best upper bound $|\mathcal{R}(w)| \le \frac{22}{23}n$ for a binary word w is obtained in [11].

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''. Note that gapped repeats with distinct periods can be the same factor, i.e. can have the same both start and end positions in the word. In this case, for convenience, we will consider these repeats as different ones, i.e. a gapped repeat is not determined uniquely by its start and end positions in the word because this information is not sufficient for determining the both copies and the gap of the repeat. For any real $\alpha > 1$ a gapped repeat σ is called *α-gapped* if $p(\sigma) \le \alpha c(\sigma)$. Analogously to repetitions, we can introduce the notion of maximality for gapped repeats. 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 $w[j'+1] \neq w[j''+1]$.

[1] Note that the period of a square is not necessarily the minimal period of this word.