Repetitions in words – a brief survey

Marcin Piątkowski*

Faculty of Mathematics and Computer Science
Nicolaus Copernicus University
martinp@mat.umk.pl

The investigation of the structure of repetitions in words allows us to create new and improve already existing data compression methods. The most interesting are maximal repetitions (runs), which represent in compact way the set of all repetitions in a word, and squares – the simplest forms of repetitions.

The maximal repetition (the run, in short) in a word $s$ is a nonempty subword $s[i..j] = w^k v$ ($k \geq 2$), where $w$ is of the minimal length and $v$ is a proper prefix (possibly empty) of $w$, that can not be extended (neither $s[i-1..j]$ nor $s[i..j+1]$ is a run with period $|u|$). Let $\rho(n)$ be the maximal number of runs in the word of the length $n$. There are known close estimations of the $\rho(n)$ but the calculation of the exact value is still the open problem.

The square in a word $s$ is a subword of the form $ww$, where $w$ is nonempty. Let $sq(n)$ be the maximal number of distinct squares in the word of the length $n$. The best known results related to the value of the $sq(n)$ are: $n - o(n) \leq sq(n) \leq 2n - O(\log n)$.

The more precise estimation we can be achieved by the reduction of the set of studied words. For the standard Sturmian words there are known the exact formulas for the number of squares and the number of runs, and their asymptotical behaviour.

The standard Sturmian words are aperiodic highly compressible words defined over a binary alphabet $\Sigma = \{a, b\}$ and described by the recurrences (or grammar-based representation) corresponding to so called directive sequences. For integer sequences of the form: $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{n-1})$, where: $\gamma_0 \geq 0$ and $\gamma_i > 0$ for $0 < i < n$, the $n$-th standard Sturmian word is given by:

$$
x_{i-1} = b
x_0 = a
x_i = x_{i-1}x_{i-2} \text{ for } 0 < i < n
$$

The special case of the standard Sturmian words are the Fibonacci words given by the directive sequences of the form: $\gamma = (1, 1, \ldots, 1)$.

References


*The research supported by Ministry of Science and Higher Education of Poland, grant N N206 580335