Combinatorics of Periods in Strings

Eric Rivals

Combinatorics of Periods in Strings

Eric Rivals
L.I.R.M.M., U.M.R. 5506
CNRS and Univ. Montpellier II
Montpellier, France
rivals@lirmm.fr
http://www.lirmm.fr/~rivals

Here, we consider a central notion of word combinatorics and string algorithmics: the periods of a string. A *period* is an offset (*i.e.*, a shift) at which a word can overlap itself. A word may have several periods, which we call its *set of periods*, and distinct words of the same length may share the same period set. When denoted by a binary string, a period set is called the *autocorrelation* of a word. In the early 80's, Guibas and Odlyzko provided the first investigation of the structure of period sets [3, 2] and characterized them. Considering the set Γ_n of all period sets of strings of length n over a finite alphabet, they showed that Γ_n is independent of the alphabet (provided the cardinality of $\Sigma \geq 2$).

Pursuing the goal of finding an enumeration algorithm for Γ_n, we study further the properties of Γ_n and exhibit the redundancy in period sets. It enables us to introduce the notion of an *irreducible period set* and to elucidate the structure of both Γ_n and the set of all irreducible period sets, denoted Λ_n. We then propose the first efficient enumeration algorithm for Γ_n. We also exhibit a relation between the number of binary partitions of n and the number of distinct period sets (*i.e.*, the cardinality of Γ_n). It allows us to improve upon the previously known asymptotic lower bounds on the cardinality of Γ_n [3]. Additionally, from these results we derive a new recurrence to compute the population of a period set, as well as an algorithm to sample uniformly irreducible and classical period sets.

All above mentioned results were published in [6, 7]. Related entries of the Encyclopedia of Integer Sequences [8] are A018819 and A000123. This study has been extended to partial words [1]. The enumeration algorithm found applications for the computation of several statistics about the vocabulary of strings, like the number of missing words of length n in a text or the number of common words between two texts [4, 5].

Acknowledgements: Collaboration with S. Rahmann, now at the University of Bielefeld.

References

[1] Francine Blanchet-Sadri, Joshua Gafni, and Kevin Wilson. Correlations of partial words. In W. Thomas and P. Weil, editors, *Theoretical Aspects of Computer Science, 14th Annual Symposium, STACS 2007, Aachen, Germany, 2007*, LNCS Vol. 4393, pages 155–66. Springer, 2007.

[2] L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching and nontransitive games. *J. of Combinatorial Theory series A*, 30:183–208, 1981.

[3] Leo J. Guibas and Andrew M. Odlyzko. Periods in strings. *J. of Combinatorial Theory series A*, 30:19–42, 1981.

[4] Sven Rahmann and Eric Rivals. Exact and Efficient Computation of the Expected Number of Missing and Common Words in Random Texts. In R. Giancarlo and D. Sankoff, editors, *Proc. of the 11th Symposium on Combinatorial Pattern Matching*, volume 1848 of *Lecture Notes in Computer Science*, pages 375–387. Springer-Verlag, Berlin, 2000.

[5] Sven Rahmann and Eric Rivals. The number of missing words in random texts. *Combinatorics, Probability and Computing*, 12:73–87, 2003.

[6] Eric Rivals and Sven Rahmann. Combinatorics of Periods in Strings. In F. Orejas, P. Spirakis, and J. van Leuween, editors, *Proc. of the 28th ICALP*, volume 2076 of *Lecture Notes in Computer Science*, pages 615–626. Springer Verlag, 2001.

[7] Eric Rivals and Sven Rahmann. Combinatorics of Periods in Strings. *J. of Combinatorial Theory series A*, 104(1):95–113, October 2003.

[8] N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences, 2004. Available at http://www.research.att.com/projects/OEIS/.