Lyndon word

From Wikipedia, the free encyclopedia

In mathematics, in the areas of combinatorics and computer science, a **Lyndon word** is a nonempty string that is strictly smaller in lexicographic order than all of its rotations. Lyndon words are named after mathematician Roger Lyndon, who investigated them in 1954, calling them **standard lexicographic sequences**.[1] Anatoly Shirshov introduced Lyndon words in 1953 calling them **regular words**. [2]

Contents

Definitions

Several equivalent definitions are possible.

A k-ary Lyndon word of length $n > 0$ is an n-character string over an alphabet of size k, and which is the unique minimum element in the lexicographical ordering of all its rotations. Being the singularly smallest rotation implies that a Lyndon word differs from any of its non-trivial rotations, and is therefore aperiodic.[3]

Alternately, a Lyndon word has the property that it is nonempty and, whenever it is split into two nonempty substrings, the left substring is always lexicographically less than the right substring. That is, if w is a Lyndon word, and $w = uv$ is any factorization into two substrings, with u and v understood to be non-empty, then $u < v$. This definition implies that a string w of length ≥ 2 is a Lyndon word if and only if there exist Lyndon words u and v such that $u < v$ and $w = uv$.[4] Although there may be more than one choice of u and v with this property, there is a particular choice, called the *standard factorization*, in which v is as long as possible.[5]

Enumeration

The Lyndon words over the two-symbol binary alphabet {0,1}, sorted by length and then lexicographically within each length class, form an infinite sequence that begins

0, 1, 01, 001, 011, 0001, 0011, 0111, 00001, 00011, 00101, 00111, 01011, 01111, ...

The first string that does not belong to this sequence, "00", is omitted because it is periodic (it consists of two repetitions of the substring "0"); the second omitted string, "10", is aperiodic but is not minimal in its permutation class as it can be cyclically permuted to the smaller string "01".

The numbers of binary Lyndon words of each length, starting with length zero, form the integer sequence

0, 2, 1, 2, 3, 6, 9, 18, 30, 56, 99, 186, 335, ... ((sequence A001037 in the OEIS), from $k = 1$ on).

Lyndon words correspond to aperiodic necklace class representatives and can thus be counted with Moreau's necklace-counting function.[3][6]

Generation

Duval (1988) provides an efficient algorithm for listing the Lyndon words of length at most n with a given alphabet size s in lexicographic order. If w is one of the words in the sequence, then the next word after w can be found by the following steps:

1. Repeat the symbols from w to form a new word x of length exactly n, where the ith symbol of x is the same as the symbol at position (i mod length(w)) of w.
2. As long as the final symbol of x is the last symbol in the sorted ordering of the alphabet, remove it, producing a shorter word.
3. Replace the final remaining symbol of x by its successor in the sorted ordering of the alphabet.

The worst-case time to generate the successor of a word w by this procedure is $O(n)$. However, if the words being generated are stored in an array of length n, and the construction of x from w is performed by adding symbols onto the end of w rather than by making a new copy of w, then the average time to generate the successor of w (assuming each word is equally likely) is constant. Therefore, the sequence of all Lyndon words of length at most n can be generated in time proportional to the length of the sequence.[7] A constant fraction of the words in this sequence have length exactly n, so the same procedure can be used to efficiently generate words of length exactly n or words whose length divides n, by filtering out the generated words that do not fit these criteria.

Standard factorization

According to the Chen–Fox–Lyndon theorem, every string may be formed in a unique way by concatenating a sequence of Lyndon words, in such a way that the words in the sequence are nonincreasing lexicographically.[8] The final Lyndon word in this sequence is the lexicographically smallest suffix of the given string.[9] A factorization into a nonincreasing sequence of Lyndon words (the so-called Lyndon factorization) can be constructed in linear time.[9] Lyndon factorizations may be used as part of a bijective variant of the Burrows–Wheeler transform for data compression,[10] and in algorithms for digital geometry.[11]

Duval (1983) developed an algorithm for finding the standard factorization that runs in linear time and constant space. It iterates over a string trying to find as long a Lyndon word as possible. When it finds one, it adds it to the result list and proceeds to search the remaining part of the string. The resulting list of strings is the standard factorization of the given string. More formal description of the algorithm follows.

Given a string S of length N, one should proceed with the following steps:

1. Let m be the index of the symbol-candidate to be appended to the already collected symbols. Initially, $m = 1$ (indices of symbols in a string start from zero).
2. Let k be the index of the symbol we would compare others to. Initially, $k = 0$.
3. While k and m are less than N, compare $S[k]$ (the k-th symbol of the string S) to $S[m]$. There are three possible outcomes:
    1. $S[k]$ is equal to $S[m]$: append $S[m]$ to the current collected symbols. Increment k and m.
    2. $S[k]$ is less than $S[m]$: if we append $S[m]$ to the current collected symbols, we'll get a Lyndon word. But we can't add it to the result list yet because it may be just a part of a larger Lyndon word. Thus, just increment m and set k to 0 so the next symbol would be compared to the first one in the string.
    3. $S[k]$ is greater than $S[m]$: if we append $S[m]$ to the current collected symbols, it will be neither a Lyndon word nor a possible beginning of one. Thus, add the first m-k collected symbols to the result list, remove them from the string, set m to 1 and k to 0 so that they point to the second and the first symbol of the string respectively.
4. When $m > N$, it is essentially the same as encountering minus infinity, thus, add the first m-k collected symbols to the result list after removing them from the string, set m to 1 and k to 0, and return to the previous step.
5. Add S to the result list.

Connection to de Bruijn sequences

If one concatenates together, in lexicographic order, all the Lyndon words that have length dividing a given number n, the result is a de Bruijn sequence, a circular sequence of symbols such that each possible length-n sequence appears exactly once as one of its contiguous subsequences. For example, the concatenation of the binary Lyndon words whose length divides four is

0 0001 0011 01 0111 1

This construction, together with the efficient generation of Lyndon words, provides an efficient method for constructing a particular de Bruijn sequence in linear time and logarithmic space.[12]

Additional properties and applications

Lyndon words have an application to the description of free Lie algebras, in constructing a basis for the homogeneous part of a given degree; this was Lyndon's original motivation for introducing these words.[4] Lyndon words may be understood as a special case of Hall sets.[4]

A theorem of Radford states that a shuffle algebra over a field of characteristic 0 can be viewed as a polynomial algebra over the Lyndon words. More precisely, let A be an alphabet, let k be a field of characteristic 0 (or, more general, a commutative $\mathbb{Q}$-algebra), and let R be the free noncommutative k-algebra $k \langle x_a \mid a \in A \rangle$. The words over A can then be identified with the "noncommutative monomials" (i.e., products of the x_a) in R; namely, we identify a word $(a_1,a_2,...,a_n)$ with the monomial $x_{a_1}x_{a_2}...x_{a_n}$. Thus, the words over A form a k-vector space basis of R. Then, a *shuffle product* is defined on R; this is a k-bilinear, associative and commutative product, which is denoted by ш, and which on the words can be recursively defined by

1 ш $v = v$ for any word v;
u ш $1 = u$ for any word u;
ua ш $vb = (u$ ш $vb)a + (ua$ ш $v)b$ for any $a,b \in$ A and any words u and v.

The *shuffle algebra* on the alphabet A is defined to be the additive group R endowed with ш as multiplication. Radford's theorem[13] now states that the Lyndon words are algebraically independent elements of this shuffle algebra, and generate it; thus, the shuffle algebra is isomorphic to a polynomial ring over k, with the indeterminates corresponding to the Lyndon words.[13]

See also

- Lexicographically minimal string rotation

Notes

1. Lyndon (1954); Berstel & Perrin (2007); Melançon (2001).
2. Shirshov (1953).
3. Berstel & Perrin (2007); Melançon (2001).
4. Melançon (2001).
5. Berstel & Perrin (2007).
6. Ruskey (2003) provides details of these counts for Lyndon words and several related concepts.
7. Berstel & Pocchiola (1994).
8. Melançon (2001). Berstel & Perrin (2007) write that although this is commonly credited to Chen, Fox & Lyndon (1958), and follows from results in that paper, it was not stated explicitly until Schützenberger (1965).
9. Duval (1983).
10. Gil & Scott (2009); Kufleitner (2009).
11. Brlek et al. (2009).
12. According to Berstel & Perrin (2007), the sequence generated in this way was first described (with a different generation method) by Martin (1934), and the connection between it and Lyndon words was observed by Fredricksen & Maiorana (1978).
13. Radford (1979)

References

- Berstel, Jean; Perrin, Dominique (2007), "The origins of combinatorics on words" (PDF), *European Journal of Combinatorics*, **28** (3): 996–1022, doi:10.1016/j.ejc.2005.07.019, MR 2300777.
- Berstel, J.; Pocchiola, M. (1994), "Average cost of Duval's algorithm for generating Lyndon words" (PDF), *Theoretical Computer Science*, **132** (1-2): 415–425, doi:10.1016/0304-3975(94)00013-1, MR 1290554.
- Brlek, S.; Lachaud, J.-O.; Provençal, X.; Reutenauer, C. (2009), "Lyndon + Christoffel = digitally convex" (PDF), *Pattern Recognition*, **42** (10): 2239–2246, doi:10.1016/j.patcog.2008.11.010.
- Chen, K.-T.; Fox, R. H.; Lyndon, R. C. (1958), "Free differential calculus. IV. The quotient groups of the lower central series", *Annals of Mathematics*, 2nd Ser., **68** (1): 81–95, doi:10.2307/1970044, MR 0102539.
- Duval, Jean-Pierre (1983), "Factorizing words over an ordered alphabet", *Journal of Algorithms*, **4** (4): 363–381, doi:10.1016/0196-6774(83)90017-2.
- Duval, Jean-Pierre (1988), "Génération d'une section des classes de conjugaison et arbre des mots de Lyndon de longueur bornée", *Theoretical Computer Science* (in French), **60** (3): 255–283, doi:10.1016/0304-3975(88)90113-2, MR 979464.
- Fredricksen, Harold; Maiorana, James (1978), "Necklaces of beads in k colors and k-ary de Bruijn sequences", *Discrete Mathematics*, **23** (3): 207–210, doi:10.1016/0012-365X(78)90002-X, MR 523071.
- Gil, J.; Scott, D. A. (2009), *A bijective string sorting transform* (PDF).
- Kufleitner, Manfred (2009), "On bijective variants of the Burrows-Wheeler transform", in Holub, Jan; Žďárek, Jan, *Prague Stringology Conference*, pp. 65–69, arXiv:0908.0239.
- Lothaire, M. (1983), *Combinatorics on words*, Encyclopedia of Mathematics and its Applications, **17**, Addison-Wesley Publishing Co., Reading, Mass., ISBN 978-0-201-13516-9, MR 675953
- Lyndon, R. C. (1954), "On Burnside's problem", *Transactions of the American Mathematical Society*, **77**: 202–215, doi:10.2307/1990868, MR 0064049.
- Martin, M. H. (1934), "A problem in arrangements", *Bulletin of the American Mathematical Society*, **40** (12): 859–864, doi:10.1090/S0002-9904-1934-05988-3, MR 1562989.
- Melançon, G. (2001), "Lyndon word", in Hazewinkel, Michiel, *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Ruskey, Frank (2003), *Info on necklaces, Lyndon words, De Bruijn sequences*.
- Schützenberger, M. P. (1965), "On a factorisation of free monoids", *Proceedings of the American Mathematical Society*, **16**: 21–24, doi:10.2307/2033993, MR 0170971.
- Shirshov, A. I. (1953), "Subalgebras of free Lie algebras", *Mat. Sbornik N.S.*, **33** (75): 441–452, MR 0059892
- Radford, David E. (1979), "A natural ring basis for the shuffle algebra and an application to group schemes", *Journal of Algebra*, **58**: 432–454, doi:10.1016/0021-8693(79)90171-6.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Lyndon_word&oldid=773766636"

Categories: Combinatorics on words

- This page was last modified on 4 April 2017, at 07:33.
- Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.