RESEARCH NOTE

NC-173 Thomas J. Watson Research Center, Yorktown Heights

ON THE MINIMUM NUMBER OF ELEMENTS IN A CUTTING SET OF WORDS

by

M. P. Schutzenberger
11/1/62

Let X be an alphabet of $k < \infty$ letters and F the set of all words in this alphabet. We shall say that a subset K of F is a cutting set iff there exists only a finite number of words f of F which have no factorization of the form $f = f_1 f_2 f_3$ with $f_2 \in K$, $f_1, f_3 \in F$. We shall limit our attention to the cutting sets consisting of words of a fixed length n and we intend to verify that the minimum number of words in such a set is $n^{-1}k^n (1 + \alpha(k, n))$ where $\alpha(k, n)$ tends to zero when $\text{Max}(k, n) \to \infty$.

First inequality

Let us recall that a word $f \in F^+$ (the set of all nonempty words of F) is primitive (or aperiodic) iff $f = f'^p$ implies $p = 1$; every $f \in F^+$ can be written in one and only one manner as f'^p with $p > 0$ and f' primitive.

Two words f and f' are conjugate iff there exist $f_1, f_2 \in F$ such that $f = f_1 f_2$ and $f' = f_2 f_1$; then $fg = gf'$ for all $g = (f_1 f_2)^p f_1$, $p \geq 0$.

Reciprocally, if g is a right and a left factor of some word f'', i.e., if there exist f and f' such that $fg = gf' = f''$, induction on the length $|g|$ of g shows that for some $f_1, f_2 \in F$ and $p \geq 0$ one has $g = (f_1 f_2)^p f_1$, $f = f_1 f_2$, $f' = f_2 f_1$, $f'' = (f_1 f_2)^{p+1} f_1$.

Finally, the number of classes of conjugate primitive words of length n is $\psi_k(n) = n^{-1} \sum_{d|n} k^{n/d} \mu(d)$ where μ denotes the Möbius function [1].

Let f and f' be two primitive words whose lengths divide n and assume that for some positive p the words f^p and f'^p have a common factor f'' of length n. This implies $f'' = (f_2 f_1)^d$ and $f'' = (f'_2 f'_1)^{d'}$ where the words f_1, f_2, f'_1, f'_2 satisfy $f = f_1 f_2$ and $f' = f'_1 f'_2$. Hence, since f and f' are primitive, $d = d'$ and f and f' are conjugate. It follows that the minimum number of words in a cutting set is at least equal to $\sum_{d|n} \psi_k(d) = n^{-1} \sum_{d|n} k^{n/d} \phi(d)$ (where ϕ denotes Euler's function) and, consequently, that $a(k, n) \geq 0$.

Second inequality

We exhibit a cutting set C having exactly $\sum_{m \leq n} \psi_k(m)$ words with the help of the following construction, which has been studied by K. T. Chen, R. H. Fox, and R. C. Lyndon [2].

Let X be totally ordered by $<$ and let $<$ also denote the induced lexicographic order on F. Define the subset H of F^+ by:

$f \in H$ iff, for all $f', f'' \in F^+$, $f = f'f''$ implies $f < f''f'$.

Clearly for each n, the set of all $h \in H$ of length n is a set of representatives of the classes of conjugate primitive words of this length. Further, it has been proved by the authors quoted above that $H = H'$ where $H' \subset F^+$ is defined by the seemingly more restrictive condition:

$f \in H'$ iff, for all $f', f'' \in F^+$, $f = f'f''$ implies $f < f''$.

We recall the proof for the sake of completeness.

Let $f \in H'$; $f = f'f''$; $f', f'' \in F^+$. Since $|f''| < |f|$ (where $|f|$ denotes the length of f), the condition $f < f''$ implies $f < f''f'''$ for all $f''' \in F$. Hence, $H' \subset H$.

In order to show $H \subset H'$ we verify first that H contains no word f such that there exist $f_1, f_2, f_3 \in F^+$ satisfying $f = f_1f_2 = f_2f_3$. Indeed, let $f = f_1f_2 = f_2f_3$; $f < f_2f_1$ and $f < f_3f_2$; either $|f_2| < |f_1|$, or $|f_1| \leq |f_2|$.

In the first case, $|f_2| < |f_1|$, there exists $f_4 \in F^+$ such that $f_1 = f_2f_4, f_3 = f_4f_2$ implying $f = f_2f_4f_2 < f_3f_2 = f_4f_2f_2$ and, consequently, $f_2f_4 < f_4f_2$. Thus $f_2f_2f_4 < f_2f_4f_2 = f$, showing $f \notin H$.

In the second case, $|f_1| \leq |f_2|$, there exists $f_4 \in F$ such that $f_2 = f_1f_4 = f_4f_3$ implying $f = f_1f_4f_3 \leq f_2f_1 = f_1f_4f_1$ and, consequently, $f_3 \leq f_1$. Thus (since $|f_1| = |f_3|$), $f_3f_1f_4 \leq f_1f_4f_3 = f$, showing again $f \notin H$.

Consider now $f \in H$ and any factorization $f = f'f''$ with $f', f'' \in F^+$. By our last remark, f'' can never be a right factor of f. Thus $f < f''f'$ implies $f < f''$. Hence $f \in H'$, and the proof is concluded.

Now let C be the set of all the left factors of length n of all the words of the form h^p with $p > 0$, $h \in H$, $|h| \leq n$.

We verify that C is a cutting set, i.e., that any infinite sequence $s = x_{i_1} x_{i_2} \ldots x_{i_j} \ldots$ of letters of X has at least one factor in C. Since $X \subset H$ we can assume $n > 1$, and since Card $X < \infty$ we can also assume that the left factor $f = x_{i_1} x_{i_2} \ldots x_{i_n}$ of length n of s is $\leq$ any other factor $x_{i_j} x_{i_{j+1}} \ldots x_{i_{j+n}}$ of the same sequence s. Thus any factor f' of f satisfies $x_{i_1} x_{i_2} \ldots x_{i_{|f'|}} \leq f'$. This shows directly that $f \in H' \subset C$ when f admits no word of F^+ as a right and a left factor.

In the remaining cases, let $g = (f_1 f_2)^p f_1$ be the word of maximal length $< |f|$ which is a right and a left factor of f. We verify that $f_1 f_2 \in H$. Indeed, because of the maximality of $|g|$, the word $f_1 f_2$ is primitive, and we can define g_1 and g_2 by the conditions $g_1 g_2 = f_1 f_2$ and $g_2 g_1 \in H$.

If $|g_1| \leq |f_1|$ or if $p > 0$, the word $g_2 g_1$ itself is a factor of f since $f = (f_1 f_2)^{p+1} f_1$. Hence, $f_1 f_2 \leq g_2 g_1$ and $g_2 g_1 \leq f_1 f_2$ $(= g_1 g_2)$, showing that $f_1 f_2 = g_2 g_1 \in H$.

If $p = 0$ and $|f_1| \leq |g_1|$, there exists f_3 such that $g_1 = f_1 f_3$, $f_2 = f_3 g_2$, and the left factor $g_2 f_1$ of $g_2 g_1$ is a factor of $f = f_1 f_2 f_1 = f_1 f_3 f_2 f_1$. Hence, since $g_2 g_1 \leq g_1 g_2 = f_1 f_2$, the word $g_2 f_1$ is equal to a left factor of f, that is, $|g_2| = 0$, since by hypothesis $f_1 = g$ is the longest word to be a right and a left factor of f. It follows that $p = 0$ and $|f_1| \leq |g_1|$ imply $f_1 = g_1$ and the verification is concluded.

Now, $\underline{\text{Card}}\ C = n^{-1} k^n (1 + \alpha'(k,n)) = \sum_{0<m\leq n} \psi_k(m) \leq \sum_{0<m\leq n} m^{-1} k^m$ $= n^{-1} k^n \sum_{0\leq j<n} n(n-j)^{-1} k^{-j}$. Thus for each $\varepsilon > 0$ there exists a finite number k_ε such that, for all $k > k_\varepsilon$ and n, one has $\alpha(k,n) \leq \alpha'(k,n) < \varepsilon$.

Finally, let the set C' consist of all the words x^{n+1} $(x \in X)$ and of all the words x'f where $x' \in X$, $f \in C$, and where the first letter $x'' \in X$ of f satisfies $x'' < x'$. C' is a cutting set because it contains $x_{i_1} x_{i_2} \ldots x_{i_{n+1}}$ if $x_{i_2} x_{i_3} \ldots x_{i_{n+1}}$ is $\leq$ any other factor of length n of the infinite sequence $x_{i_1} x_{i_2} \ldots$. Now, $\underline{\text{Card}}\ C' = (n+1)^{-1} k^{n+1} (1 + \alpha''(k, n+1)) \leq k + (k-1)\ \underline{\text{Card}}\ C$ $\leq k + (k-1) \sum_{0<m\leq n} m^{-1} k^m = (n+1)^{-1} k^{n+1} (1 + (n+1)k^{-n} + (n.n-1.k)^{-1}$ $+ \ldots + (2.1.k^n)^{-1} - k^{-n-1})$. Thus, for each $\varepsilon > 0$ and k, there exists a finite number $n_{k,\varepsilon}$ such that for all $n > n_{k,\varepsilon}$ one has $\alpha(k,n) \leq \alpha''(k,n) < \varepsilon$ and the property is entirely verified.

Remark. For $n = 1, 2$, or 3, one has Card $C' = \sum_{d|n} \psi_k(d)$. For $n = 4$, the same bound is attained by the set C'' consisting of all words $xx'x''x'''$ where x, x', x'', and x''' satisfy one of the following mutually exclusive conditions:

i) $x = x' = x'' = x'''$;
ii) $x' < x$ and $x' \leq x'' \leq x'''$;
iii) $x' < x$, $x''' < x''$ and $x'' < x$;
iv) $x' < x$, $x''' < x''$, $x'' = x$ and $x''' \leq x'$.

For $n = 5$ and $k = 2$ the minimum number of elements in a cutting set is $9 = 1 + \psi_2(1) + \psi_2(5)$.

REFERENCES

[1] C. Moreau, in E. Lucas, Théorie des nombres, Paris, 1891, pp. 501-503.

[2] K. T. Chen, R. H. Fox, and R. C. Lyndon, "Free differential calculus IV," Annals of Mathematics 68, 82-86 (1958).