Discrete Mathematics 64 (1987) 313–315
North-Holland

NOTE

ENUMERATING TOTALLY CLEAN WORDS

Doron ZEILBERGER
Department of Mathematics, Drexel University, Philadelphia, PA 19104, U.S.A.

Received 7 May 1985
Revised 19 November 1985

Let A be a finite alphabet and let D be a finite set of words in A^* labelled dirty. We give a recursive procedure for computing the generating function for the number of words not containing any subsequences that belong to D and having a specified number of each letter. We show that this generating function is always a rational function.

Let A be a finite alphabet and let $D \subset A^*$ be a finite set of words to be labelled "dirty". Let $\{\chi_a : a \in A\}$ be commuting indeterminates. To every letter $a \in A$ we assign the weight χ_a and the weight of a word is the product of the weights of its letters. For example, weight $(13221) = \chi_1\chi_3\chi_2\chi_2\chi_1 = \chi_1^2\chi_2^2\chi_3$. Given any set S of words we let weight (S) be the sum of the weights of the members of S. For example, weight $\{1, 12, 213, 2113\} = \chi_1 + \chi_1\chi_2 + \chi_1\chi_2\chi_3 + \chi_1^2\chi_2\chi_3$. The significance of the formal power series weight (S) is that the coefficient of a typical term $\prod_{a\in A} \chi_a^{\alpha_a}$ tells us the number of words in S that have α_a occurrences of the letter a, $a \in A$. It is well known and easy to see that weight $(A^*) = (1 - \sum_{a\in A} \chi_a)^{-1}$. (Recall that A^* is the set of all words (strings) that can be formed with the letters of A).

There are three standards of cleanliness that words can have.

First if we define "clean" as non-dirty, then the weight enumerator is of course

$$\text{weight}\,(A^*) - \text{weight}\,(D) = \left(1 - \sum_{a\in A} \chi_a\right)^{-1} - \text{weight}\,(D),$$

which is a rational function since weight (D) is a polynomial.

However, you may decide to be more proper and forbid words (like ESSEX) that contain a consecutive substring that is dirty. Formally $w_1 \dots w_f$ is not clean if there exists a substring of *consecutive* letters $w_i w_{i+1} w_{i+2} \dots w_j$ that belongs to D. The weight enumerator of clean words was considered in [2] and with great erudition in Goulden and Jackson's magnum opus [1] where it is shown that it is always a rational function.

But if you are really prim and proper you will even forbid words (like SCHMIDT) that contain a subsequence of letters that consistitutes a dirty word.

Thus, given a finite alphabet A and a finite set of words D let $\mathcal{W}(A; D)$ be the

0012-365X/87/$3.50

set of words in A^*, $w_1 w_2 \ldots w_f$ such that you can *not* find any subsequence $w_{i_1} w_{i_2} \ldots w_{i_r}$ $(1 \leq i_1 < i_2 < \cdots < i_r \leq f)$ that belongs to D. Let $W(A; D)$ be the weight of $\mathcal{W}(A; D)$. Before stating the theorem we need just one more piece of notation. For any set of words D and any letter $a \in A$ we denote by $D \backslash a$ the set of words obtained from D by chopping the last letter from those words that end in a and leaving the other words intact.

Thus, if $D = \{\text{DORON, MORON, PIG}\}$, $D \backslash N = \{\text{DORO, MORO, PIG}\}$, $D \backslash G = \{\text{DORON, MORON, PI}\}$ and $D \backslash A = D \backslash B = \{\text{DORON, MORON, PIG}\}$.

Having set up all the notation, the following theorem is almost trivial.

Theorem.

$$W(A; D) = 1 + \sum_{a \in A} \chi_a W(A; D \backslash a) \tag{$*$}$$

Proof. Any word in $\mathcal{W}(A; D)$ (or for that matter any word in A^*) is either the empty word or ends with one of the letters $a \in A$. If you chop the last letter a you get a typical word in $\mathcal{W}(A; D \backslash a)$. The χ_a factor in the right hand side of $(*)$ corresponds to the chopped letter a. $\square$

Formula $(*)$ enables us to compute $W(A; D)$ recursively, for every conceivable finite A and D. Let A' be the letters of A such that $D \backslash a = D$, i.e., those letters that are at the end of no dirty word. Then $(*)$ can be rewritten as

$$\left(1 - \sum_{a' \in A} \chi_{a'}\right) W(A; D) = \sum_{a \notin A'} \chi_a W(A; D \backslash a). \tag{$**$}$$

The right hand side of $(**)$ has $W(A; D')$ with a shorter list D' of dirty words. Repeated use of $(**)$ will eventually reduce to computing $W(A; D)$ where at least one of the words of D consists of just one letter, say b. Then of course $W(A; D) = W(A/b; D/b)$, that is, since the letter b by itself is a taboo we may just as well throw it out of our alphabet. Further down the line we will get $W(A'; \emptyset)$, that is *no* dirty words, and this is just the weight of $(A')^*$, $(1 - \sum_{a \in A'} \chi_a)^{-1}$. Since these bottom of the liners are rational it follows from $(**)$ and by induction that $W(A; D)$ is always rational (since the language is regular).

Examples.

$$A = \{1, 2, 3\}, \qquad D = \{123\},$$

$$W(1, 2, 3; 123) = 1 + \chi_1 W(1, 2, 3; 123) + \chi_2 W(1, 2, 3; 123) + \chi_3 W(1, 2, 3; 12).$$

Thus

$$(1 - \chi_1 - \chi_2) W(1, 2, 3; 123) = 1 + \chi_3 W(1, 2, 3; 12).$$

Now

$$W(1, 2, 3; 12) = 1 + \chi_1 W(1, 2, 3, 12) + \chi_2 W(1, 2, 3; 1) + \chi_3 W(1, 2, 3; 12).$$

Thus

$$(1-\chi_1-\chi_3)W(1, 2, 3; 12) = 1+\chi_2 W(1, 2, 3; 1).$$

But

$$W(1, 2, 3; 1) = W(2, 3; \emptyset) = \{2, 3\}^* = (1-\chi_2-\chi_3)^{-1}.$$

Thus

$$W(1, 2, 3; 12) = (1-\chi_1-\chi_3)^{-1}[1+\chi_2(1-\chi_2-\chi_3)^{-1}].$$

and

$$W(1, 2, 3; 123) = (1-\chi_1-\chi_2)^{-1}[1+\chi_3(1-\chi_1-\chi_3)^{-1}(1+\chi_2(1-\chi_2-\chi_3))^{-1}].$$

References

[1] I. Goulden and D. Jackson, Combinatorial Enumeration (Wiley, New York, 1983).
[2] D. Zeilberger, Enumerating words by their number of mistakes, Discrete Math. 34 (1981) 89–91.