PUBLICATIONS DE L'INSTITUT MATHÉMATIQUE
Nouvelle série, tome 27 (41), 1980, pp. 51—56

THE ENUMERATION OF BIT-SEQUENCES THAT SATISFY LOKAL CRITERIA

John M. Einbu

Abstract. The problem of enumerating bit-sequences in which every consecutive sub-sequence of a cartain length satisfies a given criterion, is considered. From observation of he problem, the criterion can be identified by means of a symbolic matrix and this suggests that analytical tools from graph-theory can be applied in determining the generating function for the number of bit-sequences; both the length and the content of zero-bits in the sequences are taken into account in the method. By way of illustration an example application is presented.

1. Introduction

Let B_k be the set of distinct sequences of one-bits and zero-bits of length k and let C_k be a criterion applicable to the members of this set. Then C_k defines a partitioning of B_k into two complementary sub-sets: B_k^t and B_k^f, where B_k^t is the sub-set for which C_k is satisfied and B_k^f is the sub-set for which C_k is not satisfied. If C_k is the criterion that the bit-sequence should have at least three one-bits (henceforth referred to as the $3/k$ criterion) then for $k=4$

$$B_k^t=\{1111,\ 1110,\ 1101,\ 1011,\ 0111\}$$

and

$$B_k^f=\{1100,\ 1010,\ 1001,\ 0110,\ 0101,\ 0011,\ 1000,\ 0100,\ 0010,\ 0001,\ 0000\}$$

The criterion C_k is said to be "local" to a bit-sequence of length $j>k$ if the criterion is to be applied to k consecutive bits starting at a certain position in the sequence. Also a bit-sequence of length $j>k$ is said to satisfy the "moving local criterion" C_k if C_k is satisfied for all consecutive sub-sequences of length k in the sequence. (In radar extractor theory the term "moving window" is used to characterize such a criterion). A bit-sequence of length $j>k$ and with i zero-bits which satisfies the moving local criterion C_k will for brevity be called a $C_k|i|j$-sequence. Occasionally, when the values of both i and j are unknown, irrelevant or assumed, such a sequence may be referred to as a C_k-sequence

For illustration the sequence

$$1\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 1\ 0$$

is a $3/4|3|11$-sequence whereas the sequence

$$1\ 0\ 1\ 1\ 1\ \underline{0\ 1\ 1\ 0}\ 1\ 1$$

is not a 3/4-sequence because the underlined sub-sequence of 4 bits carries only two one-bits.

If $s(C_k, i, j)$ is the number of distinct $C_k|i|j$-sequences, the intention of this paper is to derive a closed form function $g(z, t)$ which when expanded into a power series yields

$$g(z, t) = \sum_{j=k}^{\infty} \sum_{i=0}^{j} s(C_k, i, j) z^i t^j$$

2. An adjacency matrix representation of the criterion

Since the set of bit-sequences of length k has 2^k elements there are 2^{2^k} 2-partitions of this set counting the empty set as a proper set. This implies that there exists 2^{2^k} distinct C_k criteria. It is convenient to adopt a standard scheme which allows any of these criteria to be uniquely specified. One approach would be to list all the elements of B_k^t; however a more powerful approach is to label all bit-sequences of length k-1 and to define the admissible sequences of length k by means of a symbolic matrix.

Let the 8 bit-sequences of length 3 be labelled as follows:

1\. 1 1 1, 2. 0 1 1, 3. 1 0 1, 4. 1 1 0

5\. 0 0 1, 6. 0 1 0, 7. 1 0 0, 8. 0 0 0

where the decimal numbers are the labels, then if the number 1 stands

$$\left\|\begin{matrix} 1 & 0 & 0 & z & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}\right\|$$

for a one-bit and the symbol z stands for a zero-bit, the symbolic matrix uniquely specifies the 3/4 criterion. For instance the first row in the matrix carries the information that both a one-bit and a zero-bit are admitted after the sequence 111 and further that by adding a one-bit (zero-bit) to 111 the three last bits will have the label 1(4).

If the matrix is interpreted as an adjacency matrix of a labelled digraph, the associated digraph may take the form shown in Fig. 1 where the "1" or "z" attached to an edge identifies the bit which realises the transition from one sub-sequence to the other.

Since the sub-sequences labelled 5, 6, 7 and 8 are not admitted in any element of B_k^t of the 3/4 criterion rows and columns 5 to 8 of (1) and the vertices 5 to 8 of Fig. 1 may be omitted.

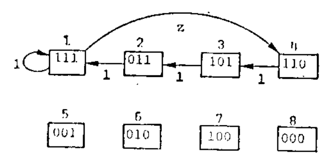

Fig. 1 Digraph associated with the 3/4-criterion.

3. The relation between $g(z, t)$ and powers of the adjacency matrix

The advantage of specifying the criterion C_k by means of an adjacency matrix is that this permits analytic tools from graph theory to be applied in the derivation of the generating function $g(z, t)$. For indeed we have

Theorem 1: *If A is the adjacency matrix associated with C_k, and $\boldsymbol{u}$ is a vector having as its n-th component the monomial $z^m t^{k-1}$ (where m is the number of zero-bits in the sub-sequence of length $k-1$ labelled n), and $\boldsymbol{v}$ is a vector with the same number of components as $\boldsymbol{u}$ and with all components equal to 1, then $\boldsymbol{u}^T A \boldsymbol{v} t$ will yield the counting series of all bit-sequences of length k satisfying C_k. And in general,*

$$g(z, t) = \sum_{j=1}^{\infty} \boldsymbol{u}^T A^j \boldsymbol{v} t^j. \tag{2}$$

If C_k is the 3/4-criterion then

$$A = \left\| \begin{matrix} 1 & 0 & 0 & z \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{matrix} \right\|, \quad u = \left\| \begin{matrix} 1 \\ z \\ z \\ z \end{matrix} \right\| t^3, \quad v = \left\| \begin{matrix} 1 \\ 1 \\ 1 \\ 1 \end{matrix} \right\|$$

where A, $\boldsymbol{u}$ and $\boldsymbol{v}$ have been reduced as much as possible according to the remark made at the end of Section 2. It follows that

$$\boldsymbol{u}^T A \boldsymbol{v} t = (1 + 4z) t^4$$

which indeed is the counting series for the elements of B_4^t for the 3/4-ctiterion.

Equation (2) suggests that $g(z, t)$ may be derived from the generating function for powers of A. This latter generating function will be established in the following section.

4. The generating function for powers of A

Let A be a square symbolic matrix of order n and let $P_A(\lambda)$ be its characteristic polynomial. Let furthermore $P_A(\lambda)$ in its expanded form be

$$P_A(\lambda) = \sum_{i=0}^{n} c_i \lambda^i$$

where c_i, $i=0, \ldots, n$ are algebraic functions of the symbols of A and $c_n=1$. Then, by the Hamilton-Cayley theorem

$$P_A(A)=\sum_{i=0}^{n} c_i A^i=0.$$

The generating function for powers of A is by definition a closed form function, $G(t)$, of a formal parameter, t, which can be expanded into the power series

$$G(t)=I+At+A^2t^2+\cdots+A^jt^j+\cdots=\sum_{j=0}^{\infty} A^jt^j \tag{3}$$

where I is the identity matrix.

Theorem 2: *The rational function*

$$G(t)=(a_0+a_1t+\cdots+a_{n-1}t^{n-1})/(1+c_{n-1}t+\cdots+c_0t^n) \tag{4}$$

where

$$\begin{aligned} a_0&=I,\\ a_1&=c_{n-1}I+A,\\ &\vdots\\ a_j&=c_{n-j}I+c_{n-j+1}A+\cdots+A^j,\\ &\vdots\\ a_{n-1}&=c_1I+c_2A+\cdots+A^{n-1} \end{aligned}$$

is a generating function for the powers of the matrix A.

Proof: The theorem is easily proved by noting that formally

$$(I-tA)^{-1}=\sum_{j=0}^{\infty} A^jt^j. \tag{5}$$

The function $G(t)$ is then obtained from the lefthand side of Eq. (5) by the multiplication by the matrix $a_0+a_1t+\cdots+a_{n-1}t^{n-1}$ and its inverse.

5. The generating function $g(z, t)$

A reformulation of Eq. (2) leads to

$$g(z, t)=u^T\left(\sum_{j=1}^{\infty} A^jt^j\right)v$$

which when combined with Eq. (3) yields

$$g(z, t)=u^T(G(t)-I)v. \tag{6}$$

Thus,

Theorem 3: *The rational function formed when the righthand side of Eq. (4) is substituted for $G(t)$ in Eq. (6) and all the matrix/vector multiplications are carried out, is the generating function for the number of bit-sequences satisfying the given local criterion C_k.*

If probabilities instead of absolute numbers are of concern the following theorem applies:

Theorem 4: *If the probabilities of apperance of a one-bit and a zero--bit in a bit-sequence are p and q respectively $(p+q=1)$, then the qenerating function for the probabilities of appearance of bit-sequences satisfying C_k is given by $g(q/p, pt)$.*

The characeristic polymonial for A of the 3/4-criterion turns out to be $\lambda^4-\lambda^3-z$, and this yields

$$a_1=\begin{Vmatrix}0 & 0 & 0 & z\\ 1 & -1 & 0 & 0\\ 0 & 1 & -1 & 0\\ 0 & 0 & 1 & -1\end{Vmatrix},\quad a_2=\begin{Vmatrix}0 & 0 & z & 0\\ 0 & 0 & 0 & z\\ 1 & -1 & 0 & 0\\ 0 & 1 & -1 & 0\end{Vmatrix},\quad a_3=\begin{Vmatrix}0 & z & 0 & 0\\ 0 & 0 & z & 0\\ 0 & 0 & 0 & z\\ 1 & -1 & 0 & 0\end{Vmatrix}.$$

From Eq. (6) it then follows

$$g(z, t)=[(1+4z)t^4+(z+z^2)t^5+(z+2z^2)t^6+(z+3z^2)t^7]/(1-t-zt^4)$$

and the first few terms of the counting series of 3/4-sequences are

$$g(z, t)=(1+4z)t^4+(1+5z+z^2)t^5+(1+6z+3z^2)t^6+(1+7z+6z^2)t^7+\cdots.$$

If the number of zero-bits in a sequence is not a discriminating attribute $z=1$ can be substituted in A and u. For the 3/4-criterion one then obtains

$$g(1, t)=(5t^4+2t^5+3t^6+4t^7)/(1-t-t^4). \tag{7}$$

In this case $g(1, t)$ can be established more directly by means of existing theory. For if $\bar{A}$ denotes the matrix obtained from A by interchanging 0 and 1, it follows from [1] that

$$g(1, t)=[(-1)^n P_{\bar{A}}(-1/t)/P_A(1/t)-1]t^2-4t^3. \tag{8}$$

It can be verified that if A is the matrix associated with the 3/4 criterion Eq. (7) follows from Eq. (8).

Also in the case of a non-symbolic adjacency matrix $s(C_k, i, j)$ can, as shown in [1], be expressed as an algebraic function of the form

$$s(C_k, u, j)=\sum_{r-1}^{m} c_r \lambda_r^{j-k+1}$$

where $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of the main part of the spectrum of A and c_r are as defined in [2].

6. Extension of method

The method in this paper can be extended logically in at least one direction: it can be used to enumerate ternary or ternary-like sequences subject to certain local con traints, or more generally the same for m-ary sequences with $m>3$. An artificial problem of this kind is to find the number of words or codes that can be formed by the alphabet a, r, s where all words, except those containing the sequences aa, rr, ss and sr and those starting with rs, are permissible. The method will provide both the number of words of various length and the number off letters of each kind in the counted words.

The moving local criterion, denoted by W_2, has in this case a minimum window length of 2 letter positions. Let the letters a, r and s be labelled as follows

$$1.\ a, \quad 2.\ r, \quad 3.\ s,$$

then the procedure of Section 2 leads to the following adjacency matrix for W_2

$$A_W = \begin{Vmatrix} 0 & r & s \\ a & 0 & s \\ a & 0 & 0 \end{Vmatrix}.$$

Let U_W^T be defined by

$$U_W^T = \| \, at, \ rt, \ st - rst^2 \, \|$$

then it is readily seen that the generating function for the number of permissible words including single-letter words is

$$g(a, r, s, t) = U_W^T G(t) v$$

where in the computation of $G(t)$ A_W, is substituted for A. This is equivalent to

$$g(a, r, s, t) = [(a + r + s)\, t + 2\,(ar + as)\, t^2 + arst^3]/[1 - (ar + as)\, t^2 - arst^3]$$

from which the desired counting series can be obtained.

REFERENCES

1. D. M. Cvetković, *The generating function for variations with restrictions and paths of the graph and self-complementary graphs*, Univ. Beograd, Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No. 320— No. 328 (1970), 27—34.

2. F. Harary, A. J. Schwenk, *The spectral approach to determining the number of walks in a graph*, Pasific J. Math. (in press).

John M. Einbu
Ostadalsveien 68
Oslo 7
Norway