Discrete Mathematics 21 (1978) 253–259.

ON STRINGS CONTAINING ALL SUBSETS AS SUBSTRINGS

Witold LIPSKI, Jr.

Institute of Computer Science, Polish Academy of Sciences, P.O. Box 22, 00–901 Warsaw PKiN, Poland

Received 8 March 1977

Let s_n be the length of a shortest sequence of positive integers which contains every $Y \subseteq \{1, \ldots, n\}$ as a subsequence of $|Y|$ consecutive terms. We give the following asymptotic estimation: $(2/\pi n)^{1/2} 2^n \lesssim S_n \lesssim (2/\pi)2^n$. The upper bound is derived constructively.

0. Introduction

The following combinatorial problem has been studied in connection with file organization (see Ghosh [2], Lipski [6]): Given a family $\mathfrak{M}$ of subsets of a finite set X, find a shortest sequence of elements of X containing every M as a subsequence of $|M|$ consecutive terms (by $|M|$ we denote the cardinality of M). Such a sequence will be called an *optimal sequence* for $\mathfrak{M}$. The general problem of constructing an optimal sequence for an arbitrary family $\mathfrak{M}$ seems to be very difficult. By a result of Kou [5], no efficient (i.e. polynomial running time) algorithm for producing an optimal sequence for a given family is likely to exist. However, even the case of a restricted form of $\mathfrak{M}$ can be a source of interesting combinatorial problems. For instance, let $X = \{0, 1\}^n$ and let $\mathfrak{M} = \{M_1, \ldots, M_n\}$ where $M_i = \{\langle b_1, \ldots, b_n \rangle \in X : b_i = 1\}$. Ehrich and Lipski [1] constructed a sequence for $\mathfrak{M}$, of length $l_n = (\frac{2}{3}n + \frac{2}{9})2^{n-1} - \frac{1}{9}(-1)^n$, which has then been proven optimal by Luccio and Preparata [7].

In the present paper we treat the case $\mathfrak{M} = \mathcal{P}(X)$, the family of all subsets of X. Of course, we may assume that X is of the form $\{1, \ldots, n\}$. A sequence of positive integers will be said to have *property* P_n if it contains every $Y \subseteq \{1, \ldots, n\}$ as a subsequence of $|Y|$ consecutive terms. Any shortest sequence with property P_n will be called *optimal* (n will usually be clear from the context), and its length will be denoted by s_n. The following sequences can easily be verified to have properties $P_1, \ldots, P_5$, respectively:

S_1	1
S_2	12
S_3	1231
S_4	12342413
S_5	1234512413524

Throughout the paper we denote by $\lfloor x \rfloor$ the greatest integer not greater than x, and denote by $\lceil x \rceil$ the least integer not less than x. For any two sequences f_n and g_n, $f_n \approx g_n$ means $\lim_{n\to\infty}(f_n/g_n)=1$, and $f_n \lesssim g_n$ (or $g_n \gtrsim f_n$) means $\limsup_{n\to\infty}(f_n/g_n)\leq 1$.

1. The bounds

We begin with the lower bound. Consider a sequence with property P_n. As it contains each of the $\lceil \frac{1}{2}n \rceil$-subsets of $\{1, \ldots, n\}$ as a subsequence of $\lceil \frac{1}{2}n \rceil$ consecutive terms, it must contain at least $\binom{n}{\lceil n/2 \rceil}$ terms as beginnings of these subsequences plus the $\lceil \frac{1}{2}n \rceil - 1$ terms as the remaining elements of the rightmost subsequence. Hence

$$s_n \geq \binom{n}{\lceil n/2 \rceil} + \lceil \tfrac{1}{2}n \rceil - 1 \tag{1}$$

Using Stirling's formula ($n! \approx n^n e^{-n}\sqrt{2\pi n}$) we obtain

$$s_n \gtrsim \sqrt{\frac{2}{\pi n}}\, 2^n. \tag{2}$$

Though the above bound was obtained by rather trivial considerations, it is much better than a $2^{n/2}$ bound given by Waksman and Green [9]. From (1) it follows that the sequences S_1, S_2, S_3 are optimal. Now let us notice that every occurrence of an $i \in \{1, \ldots, n\}$ can belong to at most k occurrences of k-subsets containing i. There are $\binom{n-1}{k-1}$ k-subsets containing i, hence i must occur at least

$$\left\lceil \binom{n-1}{k-1} \Big/ k \right\rceil = \left\lceil \binom{n}{k} \Big/ n \right\rceil$$

times in any sequence with property P_n. Taking $k = \lfloor \frac{1}{2}n \rfloor$ we obtain

$$s_n \geq n \left\lceil \binom{n}{\lfloor n/2 \rfloor} \Big/ n \right\rceil \tag{3}$$

which proves the optimality of S_4. By similar methods S_5 can also be proven optimal. We leave it to the reader. Thus we have $s_1 = 1$, $s_2 = 2$, $s_3 = 4$, $s_4 = 8$, $s_5 = 13$.

Now we pass to the upper bound. We shall need some results on decomposing $\mathcal{P}(X)$ into chains ($\mathcal{C} \subseteq \mathcal{P}(X)$ is called a *chain* if $A \subseteq B$ or $B \subseteq A$ for all $A, B \in \mathcal{C}$). From the classical Sperner's and Dilworth's theorems it follows that $\mathcal{P}(X)$ can be partitioned into $\binom{n}{\lfloor n/2 \rfloor}$ chains, where $n = |X|$. It is also well-known how to construct such a partition. Below we shall briefly describe the construction (see e.g. Greene and Kleitman [3]).

A chain is called *symmetric* if it has the form

$$C_{\lfloor n/2 \rfloor - j} \subset C_{\lfloor n/2 \rfloor - j + 1} \subset \cdots \subset C_{\lceil n/2 \rceil + j}$$

where $|C_i|=i$ for $\lfloor\frac{1}{2}n\rfloor-j\leq i\leq\lceil\frac{1}{2}n\rceil+j$ $(n=|X|, 0\leq j\leq\lfloor\frac{1}{2}n\rfloor)$. Since every symmetric chain contains exactly one $\lfloor\frac{1}{2}n\rfloor$-subset of X, it follows that any partition of $\mathcal{P}(X)$ into symmetric chains is composed of exactly $\binom{n}{\lfloor n/2\rfloor}$ chains. We construct such a partition inductively. For $n=1$, $\mathcal{P}(X)$ is itself a symmetric chain. Now assume that we have a partition of $\mathcal{P}(X)$ into symmetric chains, and let $a\notin X$. We replace every chain $A_1\subset A_2\subset\cdots\subset A_k$ of our partition by two chains

$$A_1\subset A_2\subset\cdots\subset A_k\subset A_k\cup\{a\}$$

$$A_1\cup\{a\}\subset A_2\cup\{a\}\subset\cdots\subset A_{k-1}\cup\{a\}$$

(if $k=1$ then we take only the first one). It is easy to see that this procedure produces a partition of $\mathcal{P}(X\cup\{a\})$ into symmetric chains.

A symmetric chain of the form

$$\emptyset=C_0\subset C_1\subset\cdots\subset C_n=X \tag{4}$$

will be called *complete*. Any permutation φ of $\{1,\ldots,n\}$ will be identified with the sequence $\langle\varphi(1),\ldots,\varphi(n)\rangle$. By an *initial* or *final segment* of such permutation we shall mean any set of the form $\{\varphi(1), \varphi(2),\ldots,\varphi(k)\}$, $0\leq k\leq n$, or $\{\varphi(l), \varphi(l+1),\ldots,\varphi(n)\}$, $1\leq l\leq n+1$, respectively. To every permutation there corresponds a complete chain composed of its initial segments, and conversely, any complete chain (4) is the family of initial segments of a unique permutation $\langle a_1,\ldots,a_n\rangle$ where $\{a_i\}=C_i\setminus C_{i-1}$ for $1\leq i\leq n$. Now let

$$\mathcal{P}(X)=\mathcal{C}_1\cup\cdots\cup\mathcal{C}_m,\qquad m=\binom{n}{\lfloor n/2\rfloor}$$

be a partition of $\mathcal{P}(X)$ into symmetric chains. Let us extend every chain $\mathcal{C}_i$ to an arbitrary complete chain $\bar{\mathcal{C}}_i\supseteq\mathcal{C}_i$, and let φ_i be the permutation corresponding to $\bar{\mathcal{C}}_i$. The resulting collection of permutations $\varphi_1,\ldots,\varphi_m$ has the following important property: Every subset of X appears as an initial segment of some φ_i (it is easy to see that $\binom{n}{\lfloor n/2\rfloor}$ is the minimal possible cardinality of a collection with this property). We note in passing that such collections provide a basis for a method of file organization proposed by Lum [8]. While Lum in his paper does not give any general method to obtain $\varphi_1,\ldots,\varphi_m$, from our considerations it should be clear how to construct this collection by a recursive algorithm mimicking the procedure of partitioning $\mathcal{P}(X)$ into symmetric chains (it is convenient to code a symmetric chain $\mathcal{C}$ by a permutation $\langle a_1,\ldots,a_n\rangle$ together with a pair $\langle i,j\rangle$, $0\leq i\leq j\leq n$, such that $\mathcal{C}=\{\{a_1,a_2,\ldots,a_k\}: i\leq k\leq j\}$). Another method to construct $\varphi_1,\ldots,\varphi_m$ has been sketched by Knuth [4, Exercise 1 on p. 567].

By a *special* collection of permutations of X we shall mean any collection $\varphi_1,\ldots,\varphi_r$ with the property that every $Y\subseteq X$ appears as an initial or final segment of some φ_i.

For example,

1234
2413
1423

(commas and brackets omitted) is a special collection of permutations of $\{1, 2, 3, 4\}$.

Lemma 1.1. *For every positive integer n there is a special collection $\varphi_1, \ldots, \varphi_r$ of permutations of $\{1, \ldots, n\}$, where*

$$r = \begin{cases} \dfrac{1}{2}\dbinom{n}{n/2} & \textit{if } n \textit{ even,} \\ \dfrac{1}{2}\left(1+\dfrac{1}{n}\right)\dbinom{n}{\lfloor n/2 \rfloor} & \textit{if } n \textit{ odd.} \end{cases}$$

Proof. Let n be even, and let $\psi_1, \ldots, \psi_m$, $m = \binom{n}{n/2}$, be a collection of permutations of $\{1, \ldots, n\}$ with every $Y \subseteq \{1, \ldots, n\}$ appearing as an initial segment. Since there are $\binom{n}{n/2}$ $\frac{1}{2}n$-subsets of $\{1, \ldots, n\}$, each such subset occurs as an initial segment exactly once. For $1 \leq i \leq m$, let B_i denote the $\frac{1}{2}n$-element initial segment of ψ_i. There are $r = \frac{1}{2}m$ $\frac{1}{2}n$-subsets containing 1, so we may assume that each of $B_1, \ldots, B_r$ contains 1. Now let $1 \leq i \leq r$ and let $\psi_i = \langle a_1, \ldots, a_n \rangle$. There is a unique $j > r$ with $B_j = \{1, \ldots, n\} \setminus B_i$. Let $\psi_j = \langle b_1, \ldots, b_n \rangle$. We define

$$\varphi_i = \langle a_1, a_2, \ldots, a_{n/2}, b_{n/2}, b_{n/2-1}, \ldots, b_1 \rangle.$$

It is easy to see that every $Y \subseteq \{1, \ldots, n\}$ with $|Y| \leq \frac{1}{2}n$ appears in at least one of the permutations $\varphi_1, \ldots, \varphi_r$ as an initial or final segment. Consequently, *every* $Y \subseteq \{1, \ldots, n\}$ appears as an initial or final segment. This follows from the fact that if Y is an initial (final) segment of φ_i then $\{1, \ldots, n\} \setminus Y$ is a final (resp. initial) segment of φ_i. Thus $\varphi_1, \ldots, \varphi_r$ is a special collection of permutations of $\{1, \ldots, n\}$.

Now let n be odd. We produce a special collection $\vartheta_1, \ldots, \vartheta_q$, $q = \frac{1}{2}\binom{n-1}{(n-1)/2}$, of permutations of $\{1, \ldots, n-1\}$, and then we replace every $\vartheta_i = \langle a_1, \ldots, a_{n-1} \rangle$ by the two permutations $\langle n, a_1, \ldots, a_{n-1} \rangle$ and $\langle a_1, \ldots, a_{n-1}, n \rangle$. The resulting collection $\varphi_1, \ldots, \varphi_r$ is easily seen to be a special collection of permutations of $\{1, \ldots, n\}$, and

$$r = 2q = \binom{n-1}{(n-1)/2} = \frac{(n-1)/2+1}{n}\binom{n}{(n-1)/2+1}$$

$$= \frac{1}{2}\left(1 + \frac{1}{n}\right)\binom{n}{\lfloor n/2 \rfloor}.$$

Obviously, for n even the value of r given by Lemma 1.1 is the minimal possible. It is not the case for n odd. For instance,

12345
23514
34215
13425
24135

is a special collection of permutations of $\{1, \ldots, 5\}$, whereas $\frac{1}{2}(1+\frac{1}{5})\binom{5}{2}=6$. It would be interesting to know whether or not for every n there is a special collection of $\lceil\frac{1}{2}\binom{n}{\lfloor n/2\rfloor}\rceil$ permutations of $\{1, \ldots, n\}$.

Now we are ready to present a construction of a sequence with property P_n which has length of order $2^{n+1}/\pi$. For any sequences $T_1, T_2, \ldots, T_p$ we shall denote their concatenation by $T_1T_2 \cdots T_p$. We begin with the case $n = 2k$. Let $\varphi_1, \ldots, \varphi_t$ be a special collection of permutations of $\{1, \ldots, k\}$ where

$$t=\begin{cases}\dfrac{1}{2}\dbinom{k}{k/2} & \text{if } k \text{ even}\\[2ex] \dfrac{1}{2}\left(1+\dfrac{1}{k}\right)\dbinom{k}{\lfloor k/2\rfloor} & \text{if } k \text{ odd}\end{cases} \tag{5}$$

Let $\psi_1, \ldots, \psi_t$ be a special collection of permutations of $\{k+1, \ldots, 2k\}$ (we may put $\psi_i = \langle a_1+k, \ldots, a_k+k\rangle$ for every $\varphi_i = \langle a_1, \ldots a_k\rangle$). For every $\psi_i = \langle b_1, \ldots, b_k\rangle$, let us denote $\bar{\psi}_i = \langle b_k, b_{k-1}, \ldots, b_1\rangle$. First we define the sequences

$$A_1 = \varphi_1\psi_1\varphi_2\psi_2 \cdots \varphi_{t-1}\psi_{t-1}\varphi_t\psi_t$$
$$A_2 = \varphi_1\psi_2\varphi_2\psi_3 \cdots \varphi_{t-1}\psi_t\varphi_t\psi_1$$
$$\vdots$$
$$A_i = \varphi_1\psi_i\varphi_2\psi_{i+1} \cdots \varphi_{t-1}\psi_{i-2}\varphi_t\psi_{i-1}$$
$$\vdots$$
$$A_t = \varphi_1\psi_t\varphi_2\psi_1 \cdots \varphi_{t-1}\psi_{t-2}\varphi_t\psi_{t-1}$$
$$B_1 = \varphi_1\bar{\psi}_1\varphi_2\bar{\psi}_2 \cdots \varphi_{t-1}\bar{\psi}_{t-1}\varphi_t\bar{\psi}_t$$
$$\vdots$$
$$B_t = \varphi_1\bar{\psi}_t\varphi_2\bar{\psi}_1 \cdots \varphi_{t-1}\bar{\psi}_{t-2}\varphi_t\bar{\psi}_{t-1}.$$

(Strictly speaking, any subscript s should be understood as $(s-1)(\text{mod } t)+1$.) A_{i+1} may be thought of as resulting from A_i by a cyclic shift of the ψ's to the left. B_i

differs from A_i only in that every ψ_j is replaced by $\bar{\psi}_j$. We define our sequence as

$$L_{2k} = A_1 A_2 \cdots A_t B_1 B_2 \cdots B_t \varphi_1.$$

We shall prove that L_{2k} has property P_{2k}. To this end, let us notice that any $Y \subseteq \{1, \ldots, 2k\}$ can be written as $Y = P \cup Q$ where $P \subseteq \{1, \ldots, k\}$ and $Q \subseteq \{k+1, \ldots, 2k\}$. Let us assume that P appears as a final segment of φ_i, and Q as an initial segment of ψ_j. Then Y occurs as a subsequence of $|Y|$ consecutive terms of A_p where $p = (j-i)(\text{mod } t) + 1$. Indeed, φ_i and ψ_j appear consecutively in A_p. The remaining three cases (P initial, Q final; P final, Q final; P initial, Q final) are similar. We leave them to the reader.

Let $\bar{s}_n$ denote the length of L_n. We have

$$\bar{s}_{2k} = (2t \cdot 2t + 1)k \approx \binom{k}{\lfloor k/2 \rfloor}\binom{k}{\lfloor k/2 \rfloor} k \approx \left(\sqrt{\frac{2}{\pi k}}\, 2^k\right)^2 k$$
$$= \frac{2}{\pi} 2^{2k}. \tag{6}$$

Now consider the case $n = 2k+1$. To this end, let

$$A_i^* = \varphi_1 n \psi_i n \varphi_2 n \psi_{i+1} n \cdots n \varphi_t n \psi_{i-1},$$
$$B_i^* = \varphi_1 n \bar{\psi}_i n \varphi_2 n \bar{\psi}_{i+1} n \cdots n \varphi_t n \bar{\psi}_{i-1}$$

for $1 \leq i \leq t$, where t is given by (5) and $\varphi_1, \ldots, \varphi_t, \psi_1, \ldots, \psi_t$ are the same as before. We define

$$L_{2k+1} = A_1 A_2 \cdots A_t B_1 B_2 \cdots B_t A_1^* A_2^* \cdots A_t^* B_1^* B_2^* \cdots B_t^* n \varphi_1.$$

It is easily seen that L_{2k+1} has property P_{2k+1}: the first half contains all subsets not containing $2k+1$, whereas all subsets which do contain $2k+1$ appear in the second half. Moreover, we have

$$\bar{s}_{2k+1} \approx 2\bar{s}_{2k} \approx 2\frac{2}{\pi}2^{2k} = \frac{2}{\pi}2^{2k+1}. \tag{7}$$

From (6) and (7) it follows that $\bar{s}_n \approx \frac{2}{\pi} 2^n$. Hence

$$s_n \lesssim \frac{2}{\pi} 2^n. \tag{8}$$

Comparing (2) and (8) we see that there is still much room for improvement of (at least one of) these bounds.

Apart from the problem of determining the exact order of growth of s_n, one may also ask for the behaviour of s_n^k defined to be the length of an optimal sequence for $\mathcal{P}_k(X)$, the family of all k-subsets of $X = \{1, \ldots, n\}$. A plausible conjecture is that $s_n \approx s_n^{\lfloor n/2 \rfloor}$.

References

[1] H.-D. Ehrich and W. Lipski, On the storage space requirement of consecutive retrieval with redundancy, Information Processing Lett. 4 (1976) 101–104.
[2] S.P. Ghosh, Consecutive storage of relevant records with redundancy, Comm. ACM 18 (1975) 464–471.
[3] C. Greene and D. Kleitman, Strong versions of Sperner's theorem, J. Combinatorial Theory Ser. A 20 (1976) 80–88.
[4] D.E. Knuth, The Art of Computer Programming, Vol. III: Sorting and Searching (Addison-Wesley, Reading, MA, 1973).
[5] L.T. Kou, Polynomial complete consecutive information retrieval problems, Technical Report TR 74–193, Dept. of Computer Science, Cornell University, Ithaca, N.Y. (1974); SIAM J. Comput. (March 1977).
[6] W. Lipski, Information storage and retrieval — mathematical foundations II (Combinatorial problems), Theoret. Comput. Sci. 3 (1976) 183–212.
[7] F. Luccio and F.P. Preparata, Storage for consecutive retrieval, Information Processing Lett. 5 (1976) 68–71.
[8] V.Y. Lum, Multi-attribute retrieval with combined indexes, Comm. ACM 13 (1970) 660–665.
[9] A. Waksman and M.W. Green, On the consecutive retrieval property in file organization, IEEE Trans. Comput. 23 (1974) 173–174.