PATTERNS, AUTOMATA, AND STIRLING NUMBERS OF THE SECOND KIND

Curtis Cooper and Robert E. Kennedy
Department of Mathematics, Central Missouri State University
Warrensburg, MO 64093

1. **Introduction.** While investigating the distribution of substrings within the collection of character strings [1], we became interested in the concept of (linear) "patterns." Thus, realizing that the character strings

$$ABAAC,$$

$$XZXXP,$$

and $$\%*\%\%\$$$

are examples of the same "pattern," we set out to determine the number of "patterns" of a given length.

After developing a definition for the term "pattern," we arrived at what was (at least to us) a surprising and interesting relationship between the number of patterns of a given length and Stirling Numbers of the Second Kind.

2. **Notation and Definitions.** One way to determine whether or not character strings, such as those given above, are examples of the same pattern is to use a "coding" function, f, that defines a correspondence between a character string and a sequence of non–negative integers by identifying each character of the string with its location and whether or

not it has occurred previously. More specifically, for the character string

$$S = c_1 c_2 c_3 \cdots c_n \ ,$$

of length n, we define

$$f(S) = \{a_i\}_{i=1}^{\infty} \ ,$$

where $a_i = \ \min \{1 \leq k \leq n : c_i = c_k\}$ for $1 \leq i \leq n$ and $a_i = 0$ for $i > n$. For example,

$$f(a * b\# * *a) = \{1, 2, 3, 4, 2, 2, 1, 0, 0, \ldots\} \ .$$

With this in mind, we make the following definition.

Definition. A sequence $\{a_i\}_{i=1}^{\infty}$ of integers is called an "n–pattern sequence" if

$$(1) \quad a_i = 0 \ \text{ for } \ i > n \ ,$$

and for $i \leq n$,

$$a_i = i \ \text{ or } \ a_i = a_j \ \text{ for some } \ j < i \ .$$

Thus, an n–pattern sequence emphasizes the location of an integer in the sequence, and we see that the above coding function, f, defines a correspondence between each character string and a unique n–pattern sequence. Hence, the determination of the number of patterns of length n, is equivalent to the determination of the number of $\mathrm{n} - \mathrm{pattern}$ sequences. Some specific examples are:

$\{1, 0, 0, 0, 0, 0, \ldots\}$ is the only $1-$pattern sequence .

$\{1, 1, 0, 0, 0, 0, \ldots\}$ and

$\{1, 2, 0, 0, 0, 0, \ldots\}$ are the only $2-$pattern sequences .

$\{1, 1, 1, 0, 0, 0, \ldots\}$,

$\{1, 1, 3, 0, 0, 0, \ldots\}$,

$\{1, 2, 1, 0, 0, 0, \ldots\}$,

$\{1, 2, 2, 0, 0, 0, \ldots\}$, and

$\{1, 2, 3, 0, 0, 0, \ldots\}$ are the only $3-$pattern sequences.

After further investigation it can be found that, in addition to the above, there are

$$\begin{aligned} 15 \quad & 4-\text{pattern sequences,} \\ 52 \quad & 5-\text{pattern sequences,} \\ 203 \quad & 6-\text{pattern sequences,} \\ 877 \quad & 7-\text{pattern sequences,} \end{aligned}$$

and so forth. This raises the question as to how the sequence

$$1,\ 2,\ 5,\ 15,\ 52,\ 203,\ 877,\ \ldots$$

is generated.

3. **Automata.** To count the number of n–pattern sequences (which is the number of patterns of length n over an alphabet with an infinite number of symbols) we define an automaton which, after n moves, produces all n–pattern sequences. For a general discussion of various automata, see [2]. Our automaton will have an infinite number of states, i.e.

$$S = \{S_0, S_1, S_2, \ldots\} \ .$$

To construct all n–pattern sequences, start with the 0–pattern sequence

$$\{0, 0, 0, 0, \ldots\}$$

in state S_0. A $(k+1)$–pattern sequence can be constructed from a k–pattern sequence by replacing the first zero in the k–pattern sequence with either $k+1$ or a previous term of the k–pattern sequence. We say that this construction process has reached state S_i if the $(k+1)$–pattern sequence has i distinct non–zero members. For example, the 5–pattern sequence

$$\{1, 2, 1, 4, 2, 0, 0, 0, \ldots\}$$

is in state S_3 since the 5–pattern sequence has 3 distinct non–zero members. From this sequence we can construct the 6–pattern sequences

$$\{1, 2, 1, 4, 2, 6, 0, 0, \ldots\}$$

in state S_4, or

$$\{1, 2, 1, 4, 2, 1, 0, 0, \ldots\},$$

$$\{1, 2, 1, 4, 2, 2, 0, 0, \ldots\},$$

$$\{1, 2, 1, 4, 2, 4, 0, 0, \ldots\},$$

all in state S_3.

A pictorial representation of the number of possible transitions from state to state in the automaton is,

where the numbers above the arrows denote the number of possible transitions from one state to another state. For example, if the construction process is in state S_k, there is one possible transition to S_{k+1} while there are k possible transitions back into state S_k. Hence, the transition matrix defined by the above automaton is,

$$T = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 2 & 1 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 3 & 1 & 0 & 0 & \ldots \\ & & & . & . & . & & \ldots \\ 0 & 0 & 0 & & & k & 1 & \ldots \\ & & & . & . & . & & \ldots \end{pmatrix},$$

where the entry in the i^{th} row and the j^{th} column is the number of ways that a transition from state S_{i-1} to S_{j-1} is possible.

Now the sum of the first row of T^n is the number of possible n–pattern sequences. In fact, the entries in the first row of T^n are the number of n–pattern sequences in states S_0, S_1, S_2, ... This follows by induction on n, the structure of T, and the fact that $T^{k+1} = T \cdot T^k$.

Thus, we only have to determine the sum of the first row of T^n to find the number of patterns of length n which can be constructed using an infinite alphabet.

4. **Stirling Numbers of the Second Kind.** Consider the following examples of the first row of various powers of T:

The first row of T is : 0 1 0 0 0 0 ...

The first row of T^2 is : 0 1 1 0 0 0 ...

The first row of T^3 is : 0 1 3 1 0 0 ...

The first row of T^4 is : 0 1 7 6 1 0 ...

The first row of T^5 is : 0 1 15 25 10 1 ...

No discernible pattern for the entries in the first row of powers of T seem to present itself at first. However, one of the authors thought that the first row of T^4 looked familiar, and upon some thought realized that he had seen such a sequence before. He was correct! In [3; page 66], a table listing the Stirling Numbers of the Second Kind is given, and the entries of the first row of the various powers of T match exactly with the rows of this table. It might be recalled that Stirling Numbers of the Second Kind are used to convert a power to an expression involving binomial

coefficients and can be defined recursively by

$$\left\{ {n \atop m} \right\} = m \left\{ {n-1 \atop m} \right\} + \left\{ {n-1 \atop m-1} \right\} ,$$

where

$$\left\{ {n \atop n} \right\} = 1, \quad \left\{ {n \atop 1} \right\} = 1, \quad \text{and} \quad \left\{ {n \atop 0} \right\} = 0 \ .$$

We would like to recommend Knuth [3] for a more detailed discussion of the Stirling Numbers.

Since the first row of T is

$$0, 1, 0, 0, 0, \ldots \ ,$$

the first row of T^{k+1} is the second row of T^k. Using this and the above recurrence relation, it follows immediately by mathematical induction that,

Theorem. The number of patterns of length n which can be constructed using an infinite alphabet is

$$\sum_{k=1}^{n} \left\{ {n \atop k} \right\} .$$

The reader might note that the sum given in the above theorem is known as the "nth Bell number," which is the number of partitions of a set with n elements. Here, we have shown the connection between the number of patterns of length n over an infinite alphabet and the nth Bell number.

References.

1. R.E. Kennedy and C. Cooper, "Substring Statistics," (Submitted).

2. H.R. Lewis and C.H. Papadimitriou, *Elements of the Theory of Computation*, Prentice–Hall, Inc., (1981).

3. D.E. Knuth, *The Art of Computer Programming*, Addison–Wesley Publishing Co., (1969), vol. 1, 65–68.