

DISCRETE MATHEMATICS

W W L CHEN

© W W L Chen, 1991, 2003.

This chapter is available free to all individuals, on the understanding that it is not to be used for financial gains, and may be downloaded and/or photocopied, with or without permission from the author.

However, this document may not be kept on any information storage and retrieval system without permission from the author, unless such system is not accessible to any individuals other than its owners.

Chapter 5

LANGUAGES

5.1. Introduction

Most modern computer programmes are represented by finite sequences of characters. We therefore need to develop an algebraic way for handling such finite sequences. In this chapter, we shall study the concept of languages in a systematic way.

However, before we do so, let us look at ordinary languages, and let us concentrate on the languages using the ordinary alphabet. We start with the 26 characters A to Z (forget umlauts, *etc.*), and string them together to form words. A language will therefore consist of a collection of such words. A different language may consist of a different such collection. We also put words together to form sentences.

Let A be a non-empty finite set (usually known as the alphabet). By a string of A (word), we mean a finite sequence of elements of A juxtaposed together; in other words, $a_1 \dots a_n$, where $a_1, \dots, a_n \in A$.

DEFINITION. The length of the string $w = a_1 \dots a_n$ is defined to be n , and is denoted by $\|w\| = n$.

DEFINITION. The null string is the unique string of length 0 and is denoted by λ .

DEFINITION. Suppose that $w = a_1 \dots a_n$ and $v = b_1 \dots b_m$ are strings of A . By the product of the two strings, we mean the string $wv = a_1 \dots a_n b_1 \dots b_m$, and this operation is known as concatenation.

The following results are simple.

PROPOSITION 5A. *Suppose that w , v and u are strings of A . Then*

- (a) $(wv)u = w(vu)$;
- (b) $w\lambda = \lambda w = w$; and
- (c) $\|wv\| = \|w\| + \|v\|$.

DEFINITION. For every $n \in \mathbb{N}$, we write $A^n = \{a_1 \dots a_n : a_1, \dots, a_n \in A\}$; in other words, A^n denotes the set of all strings of A of length n . We also write $A^0 = \{\lambda\}$. Furthermore, we write

$$A^+ = \bigcup_{n=1}^{\infty} A^n \quad \text{and} \quad A^* = A^+ \cup A^0 = \bigcup_{n=0}^{\infty} A^n.$$

DEFINITION. By a language L on the set A , we mean a (finite or infinite) subset of the set A^* . In other words, a language on A is a (finite or infinite) set of strings of A .

DEFINITION. Suppose that $L \subseteq A^*$ is a language on A . We write $L^0 = \{\lambda\}$. For every $n \in \mathbb{N}$, we write $L^n = \{w_1 \dots w_n : w_1, \dots, w_n \in L\}$. Furthermore, we write

$$L^+ = \bigcup_{n=1}^{\infty} L^n \quad \text{and} \quad L^* = L^+ \cup L^0 = \bigcup_{n=0}^{\infty} L^n.$$

The set L^+ is known as the positive closure of L , while the set L^* is known as the Kleene closure of L .

DEFINITION. Suppose that $L, M \subseteq A^*$ are languages on A . We denote by $L \cap M$ their intersection, and denote by $L + M$ their union. Furthermore, we write $LM = \{wv : w \in L \text{ and } v \in M\}$.

EXAMPLE 5.1.1. Let $A = \{1, 2, 3\}$.

- $A^2 = \{11, 12, 13, 21, 22, 23, 31, 32, 33\}$ and A^4 has $3^4 = 81$ elements. A^+ is the collection of all strings of length at least 1 and consisting of 1's, 2's and 3's.
- The set $L = \{21, 213, 1, 2222, 3\}$ is a language on A . The element $213222213 \in L^4 \cap L^5$, since $213, 2222, 1, 3 \in L$ and $21, 3, 2222, 1, 3 \in L$. On the other hand, $213222213 \in A^9$.
- The set $M = \{2, 13, 222\}$ is also a language on A . Note that $213222213 \in L^4 \cap L^5 \cap M^5 \cap M^7$. Note also that $L \cap M = \emptyset$.

PROPOSITION 5B. Suppose that $L, M \subseteq A^*$ are languages on A . Then

- (a) $L^* + M^* \subseteq (L + M)^*$;
- (b) $(L \cap M)^* \subseteq L^* \cap M^*$;
- (c) $LL^* = L^*L = L^+$; and
- (d) $L^* = LL^* + \{\lambda\}$.

REMARKS. (1) Equality does not hold in (a). To see this, let $A = \{a, b, c\}$, $L = \{a\}$ and $M = \{b\}$. Then $L + M = \{a, b\}$. Note now that $aab \in (L + M)^*$, but clearly $aab \notin L^*$ and $aab \notin M^*$.

(2) Equality does not hold in (b) either. Again let $A = \{a, b, c\}$. Now let $L = \{a, b\}$ and $M = \{ab\}$. Then $L \cap M = \emptyset$, so $(L \cap M)^* = \{\lambda\}$. On the other hand, we clearly have $ababab \in L^*$ and $ababab \in M^*$.

PROOF OF PROPOSITION 5B. (a) Clearly $L \subseteq L + M$, so that $L^n \subseteq (L + M)^n$ for every $n \in \mathbb{N} \cup \{0\}$. Hence

$$L^* = \bigcup_{n=0}^{\infty} L^n \subseteq \bigcup_{n=0}^{\infty} (L + M)^n = (L + M)^*.$$

Similarly $M^* \subseteq (L + M)^*$. It follows that $L^* + M^* \subseteq (L + M)^*$.

(b) Clearly $L \cap M \subseteq L$, so that $(L \cap M)^* \subseteq L^*$. Similarly, $(L \cap M)^* \subseteq M^*$. It follows that $(L \cap M)^* \subseteq L^* \cap M^*$.

(c) Suppose that $x \in LL^*$. Then $x = w_0y$, where $w_0 \in L$ and $y \in L^*$. It follows that either $y \in L^0$ or $y = w_1 \dots w_n \in L^n$ for some $w_1, \dots, w_n \in L$ and where $n \in \mathbb{N}$. Therefore $x = w_0$ or $x = w_0w_1 \dots w_n$, so that $x \in L^+$. Hence $LL^* \subseteq L^+$. On the other hand, if $x \in L^+$, then $x = w_1 \dots w_n$ for some $n \in \mathbb{N}$

and $w_1, \dots, w_n \in L$. If $n = 1$, then $x = w_1y$, where $w_1 \in L$ and $y \in L^0$. If $n > 1$, then $x = w_1y$, where $w_1 \in L$ and $y \in L^{n-1}$. Clearly $x \in LL^*$. Hence $L^+ \subseteq LL^*$. It follows that $LL^* = L^+$. A similar argument gives $L^*L = L^+$.

(d) By definition, $L^* = L^+ \cup L^0$. The result now follows from (c). \circ

5.2. Regular Languages

Let A be a non-empty finite set. We say that a language $L \subseteq A^*$ on A is regular if it is empty or can be built up from elements of A by using only concatenation and the operations $+$ and $*$.

One of the main computing problems for any given language is to find a programme which can decide whether a given string belongs to the language. Regular languages are precisely those for which it is possible to write such a programme where the memory required during calculation is independent of the length of the string. We shall study this question in Chapter 7.

We shall look at a few examples of regular languages. It is convenient to abuse notation by omitting the set symbols $\{$ and $\}$.

EXAMPLE 5.2.1. Binary natural numbers can be represented by $1(0+1)^*$. Note that any such number must start with a 1, followed by a string of 0's and 1's which may be null.

EXAMPLE 5.2.2. Let m and p denote respectively the minus sign and the decimal point. If we write $d = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9$, then decimal numbers can be represented by

$$0 + (\lambda + m)d(0 + d)^* + (\lambda + m)(0 + (d(0 + d)^*))p(0 + d)^*d.$$

Note that any integer must belong to $0 + (\lambda + m)d(0 + d)^*$. On the other hand, any non-integer may start with a minus sign, must have an integral part, followed by a decimal point, and some extra digits, the last of which must be non-zero.

EXAMPLE 5.2.3. Binary strings in which every consecutive block of 1's has even length can be represented by $(0 + 11)^*$.

EXAMPLE 5.2.4. Binary strings containing the substring 1011 can be represented by $(0+1)^*1011(0+1)^*$.

EXAMPLE 5.2.5. Binary strings containing the substring 1011 and in which every consecutive block of 1's has even length can be represented by $(0 + 11)^*11011(0 + 11)^*$. Note that 1011 must be immediately preceded by an odd number of 1's.

Note that the language in Example 5.2.5 is the intersection of the languages in Examples 5.2.3 and 5.2.4. In fact, the regularity is a special case of the following result which is far from obvious.

PROPOSITION 5C. *Suppose that L and M are regular languages on a set A . Then $L \cap M$ is also regular.*

PROPOSITION 5D. *Suppose that L is a regular language on a set A . Then the complement language \bar{L} is also regular.*

This is again far from obvious. However, the following is obvious.

PROPOSITION 5E. *Suppose that L is a language on a set A , and that L is finite. Then L is regular.*

PROBLEMS FOR CHAPTER 5

1. Suppose that $A = \{1, 2, 3, 4\}$. Suppose further that $L, M \subseteq A^*$ are given by $L = \{12, 4\}$ and $M = \{\lambda, 3\}$. Determine each of the following:

a) LM	b) ML	c) L^2M	d) L^2M^2
e) LM^2L	f) M^+	g) LM^+	h) M^*
2. Let $A = \{1, 2, 3, 4\}$.
 - a) How many elements does A^3 have?
 - b) How many elements does A^0 have?
 - c) How many elements of A^4 start with the substring 22?
3. Let $A = \{1, 2, 3, 4, 5\}$.
 - a) How many elements does A^4 have?
 - b) How many elements does $A^0 + A^2$ have?
 - c) How many elements of A^6 start with the substring 22 and end with the substring 1?
4. Let $A = \{1, 2, 3, 4\}$.
 - a) What is A^2 ?
 - b) Let $n \in \mathbb{N}$. How many elements does the set A^n have?
 - c) How many strings are there in A^+ with length at most 5?
 - d) How many strings are there in A^* with length at most 5?
 - e) How many strings are there in A^* starting with 12 and with length at most 5?
5. Let A be a finite set with 7 elements, and let L be a finite language on A with 9 elements such that $\lambda \in L$.
 - a) How many elements does A^3 have?
 - b) Explain why L^2 has at most 73 elements.
6. Let $A = \{0, 1, 2\}$. For each of the following, decide whether the string belongs to the language $(0+1)^*2(0+1)^*22(0+1)^*$:

a) 120120210	b) 1222011	c) 22210101
--------------	------------	-------------
7. Let $A = \{0, 1, 2, 3\}$. For each of the following, decide whether the string belongs to the language $(0+1)^*2(0+1)^+22(0+1+3)^*$:

a) 120122	b) 1222011	c) 3202210101
-----------	------------	---------------
8. Let $A = \{0, 1, 2\}$. For each of the following, decide whether the string belongs to the language $(0+1)^*(102)(1+2)^*$:

a) 012102112	b) 1110221212	c) 10211111
d) 102	e) 001102102	
9. Suppose that $A = \{0, 1, 2, 3\}$. Describe each of the following languages in A^* :
 - a) All strings in A containing the digit 2 once.
 - b) All strings in A containing the digit 2 three times.
 - c) All strings in A containing the digit 2.
 - d) All strings in A containing the substring 2112.
 - e) All strings in A in which every block of 2's has length a multiple of 3.
 - f) All strings in A containing the substring 2112 and in which every block of 2's has length a multiple of 3.
 - g) All strings in A containing the substring 2112 and in which every block of 2's has length a multiple of 3 and every block of 1's has length a multiple of 2.
 - h) All strings in A containing the substring 2112 and in which every block of 2's has length a multiple of 3 and every block of 1's has length a multiple of 3.

-
10. Suppose that $A = \{0, 1\}$. Describe each of the following languages in A^* :
- All strings containing at least two 1's.
 - All strings containing an even number of 0's.
 - All strings containing at least two 1's and an even number of 0's.
 - All strings starting with 1101 and having an even number of 1's and exactly three 0's.
 - All strings starting with 111 and not containing the substring 000.
11. Let A be an alphabet, and let $L \subseteq A^*$ be a non-empty language such that $L^2 = L$.
- Suppose that L has exactly one element. Prove that this element is λ .
 - Suppose that L has more than one element. Show that there is an element $w \in L$ such that
 - $w \neq \lambda$; and
 - every element $v \in L$ such that $v \neq \lambda$ satisfies $\|v\| \geq \|w\|$.Deduce that $\lambda \in L$.
 - Prove that $L^* = L$.
12. Let L be an infinite language on the alphabet A such that the following two conditions are satisfied:
- Whenever $\alpha = \beta\delta \in L$, then $\beta \in L$. In other words, L contains all the initial segments of its elements.
 - $\alpha\beta = \beta\alpha$ for all $\alpha, \beta \in L$.
- Prove that $L = \omega^*$ for some $\omega \in A$ by following the steps below:
- Explain why there is a unique element $\omega \in L$ with shortest positive length 1 (so that $\omega \in A$).
 - Prove by induction on n that every element of L of length at most n must be of the form ω^k for some non-negative integer $k \leq n$.
 - Prove the converse, that $\omega^n \in L$ for every non-negative integer n .