第3章 词法分析

3.1 词法分析基本思想

3.1.1 词法分析需求分析

词法分析(lexical analysis)是编译的第一个阶段,它的主要任务是从左至右逐个字符地对源程序进行扫描,产生一个个单词序列,用于语法分析。完成词法分析任务的程序称为词法分析程序,通常也成为词法分析器或扫描器(scanner)。它一般是一个独立的子程序或作为词法分析器的一个辅助子程序。

词法分析程序一般具有如下功能:读入字符串形式的源程序;识别出具有独立意义的最小语法单位——单词,其功能的具体说明参见例3-1。词法分析程序一般将单词变换成带有单词性质且定长的属性字;为方便下一阶段的工作,还可以进行一些简单而力所能及的工作。词法分析程序的功能如图3-1所示。

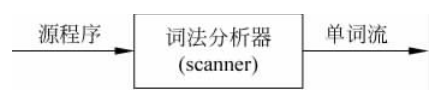


图3-1 词法分析程序的功能

3.1.2 词法分析任务

词法分析程序的主要任务是按照高级语言的词法规则从左到右逐个扫描字符流的源程序,从中识别出各类有独立意义的单词。

除了识别单词外,词法分析程序还应该完成其他一些任务,主要包括以下几个方面:

(1) 消除无用字符。对源程序文本进行处理,过滤掉源程序文本中的注释、空格、换行符以及其他一切与语法分析和代码生成均无关的信息。

(2) 进行内部编码。将长度不一、种类不同的单词变成长度统一、格式规整、分类清晰的内部机器码表示。

(3) 建立各种表格。在词法分析时,可以根据单词特点建立不同表格,例如:

- 名字表(标识符表):源程序中的标识符集中在表中。
- 常数表。
- 数组向量表、过程表等。
- 界限表:包含了保留字、运算符等。

(4) 进行词法检查。为了使编译程序能将发现的错误信息与源程序的出错位置联系起来,词法分析程序负责记录新读入的字符行的行号,以便行号与出错信息相关联;另外,在支持宏处理功能的源语言中,可以由词法分析程序完成其预处理等。

3.1.3 词法分析方式

现阶段词法分析主要有两种分析方式。

1. 将词法分析程序和语法分析程序分开

在多遍扫描的编译程序中,词法分析可以单独作为一遍扫描来完成,此时可将词法分析程序的输出放在一个中间文件上,语法分析程序可以从该文件取得它的输入,如图 3-2 所示。词法分析从语法分析独立出来的原因主要是考虑便于集中进行语法分析,便于建立有效的词法分析技术,将给语法分析提供更多更详细的信息。

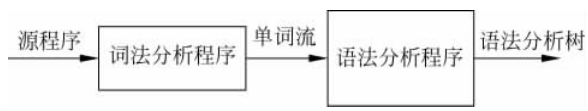


图 3-2 词法分析程序和语法分析程序分开

2. 将词法分析程序编写成一个独立子程序

在一遍扫描的编译程序中,往往将词法分析编写成语法分析的一个子程序,供语法分析时随时调用,每调用一次,则从源程序字符串中读出一个具有独立意义的单词,如图 3-3 所示。这种模式不需要在内存中构造和保留中间文件,所以可以节省内存空间。

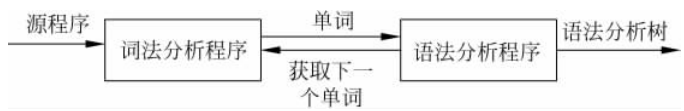


图 3-3 词法分析程序作为语法分析程序的子程序

3.2 单词的描述工具

3.2.1 正规文法和正规式

正规式及正规式所表示的语言——正规集的概念是美国数学家 Kleen 在 20 世纪 50 年代提出来的。这种方法现在已成为处理有限自动机问题的主要数学工具,无论在理论上还是在计算机科学领域的诸多工程实践中都有重要应用。

定义 3-1 正规式与正规集

设 Σ 为有限字母表,在 Σ 上的正规式与正规集可递归定义如下:

(1) ε 和 $\varnothing$ 是 Σ 上的正规式,它们表示的正规集分别为 $\{\varepsilon\}$ 和 $\varnothing$。

(2) 对任何 $a\in\Sigma$,a 是 Σ 上的正规式,它表示的正规集为 $\{a\}$。

(3) 若 r、s 都是正规式，它们表示的正规集分别为 R 和 S，则$(r|s)$(或表示为 $r+s$)、$(r \cdot s)$(或表示为 rs)、$(r)^*$ 也是正规式，它们表示的正规集分别是 $R \cup S$、RS、R^*。

(4) 有限次使用上述 3 条规则构成的表达式称为 Σ 上的正规式，仅由这些正规式表示的集合为正规集。

规定正规式运算的优先级由高到低的次序为 *(闭包)、·(连接)和|(并)，它们的结合性都为左结合。在此规定下，书写正规式时可以省去不致造成混淆的括号。例如$((0 \cdot (1^*))|0)$可写成 $01^*|0$。

【例 3-1】 设字母表 $\Sigma=\{0,1\}$，则 0、1、ε、∅是 Σ 上的正规式，$0|1$、$0 \cdot 1$、$1 \cdot 0$、0^*、1^*是 Σ 上的正规式。

【例 3-2】 设字母表 $\Sigma=\{A,B,0,1\}$。

正规式$(A|B)(A|B|0|1)^*$表示的正规集是以字母 A 或 B 开头后跟任意多个字母 A、B 和数字 0、1 的符号串(标识符)的全体。

正规式$(0|1)(0|1)^*$表示的正规集是二进制数字串。

正规式 r 所表示的正规集 R 是字母表 Σ 上的语言，称为正则语言，用 $L(r)$表示，即 $R=L(r)$。$L(r)$中的元素为字符串(也称为句子)。

若两个正规式 r 和 s 所表示的语言 $L(r)=L(s)$，则称 r、s 等价，记作 $r=s$。

例如，$1(01)^*=(10)^*1$。

正规式的性质参见表 3-1，其中 s、t、r 为正规式。

表 3-1 正规式的代数性质

公 理	描 述	公 理	描 述
$s\|t=t\|s$	并是可交换的	$\varepsilon s=s$	ε 是连接的恒等元素
$s\|(t\|r)=(s\|t)\|r$	并是可结合的	$s\varepsilon=s$	
$(st)r=s(tr)$	连接是可结合的	$s^*=(s\|\varepsilon)^*$	闭包和 ε 间的关系
$s(t\|r)=st\|sr$	连接对并可分配	$a^{**}=a^*$	闭包是幂等的
$(t\|r)s=ts\|rs$			

利用正规式的代数性质，可以对正规式进行等价变换及化简。

有些语言不能用正规式表达，说明了正规式的描述能力有限。例如，正规式不能描述配对或嵌套的结构。

3.2.2 正规式与有限自动机的等价性

从前面的讨论得知，有限自动机接收的语言等价于正规文法产生的正规语言。而正规式或正规集定义的语言也是正规语言。那么，正规式与有限自动机在描述语言上应该等价。

定理 3-1

(1) 字母表是 Σ 的确定的有限自动机 M 所接受的语言 $L(M)$是 Σ 上的一个正规集。

(2) 对于 Σ 上的每一个正规式 r，存在一个字母表是 Σ 的非确定有限自动机 M，使得 $L(M)=L(r)$。

定理 3-1 表明，正规式所表示的语言即正规集与有限自动机所识别的语言是完全等价

的，只是表示形式不同。也就是说，从描述语言的角度，没有必要对非确定的有限自动机、确定的有限自动机及它们所识别的语言（正规集）加以区分。同一个语言，既可以用有限自动机描述，也可以用正规式描述。有兴趣的读者可以对定理 3-1 进行证明。

根据上面的等价性证明，构造出等价的转换算法。

对于字母表 Σ 上任意一个正规式 r，一定可以构造一个非确定的有限自动机 M，使得 $L(M)=L(r)$。

首先构造非确定的有限自动机 M 的一个广义的状态图，也是该非确定的有限自动机 M 的初始状态图。其中，只有一个开始状态 q_S和一个终止状态 q_Z，连接 q_S和 q_Z的有向弧上的标记是正规式 r。然后，按照图 3-4 的替换规则 1 对正规式 r 依次进行分解，分解的过程是一个不断加入结点和弧的过程，直到转换图上的所有弧标记上都是 Σ 上的元素或 ε 为止。

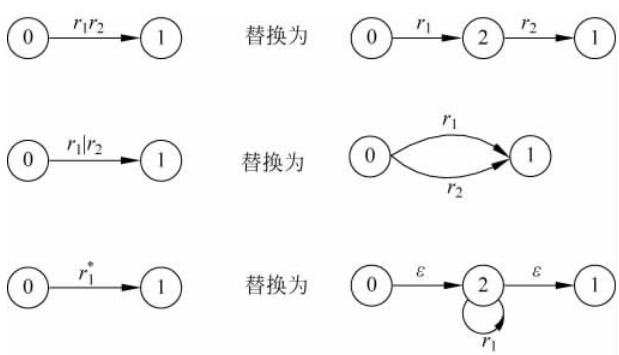


图 3-4 替换规则 1

【例 3-3】 设 $\Sigma=\{x,y\}$，Σ 上的正规式 $r=xy^*(xy|yx)x^*$，构造一个非确定的有限自动机 M，使 $L(M)=L(r)$。

第一步，构造 M 的初始状态图，得如图 3-5(a)所示的状态图。

第二步，将 $r=xy^*(xy|yx)x^*$ 拆成 4 个正规式 x、y^*、$xy|yx$、x^* 的连接，得如图 3-5(b)所示的状态图。

第三步，将 y^*、$xy|yx$、x^* 分别拆成正规式 y 的闭包、xy 与 yx 的并、x 的闭包，得如图 3-5(c)所示的状态图。

第四步，将 xy、yx 分别拆成正规式 x 与 y 的连接和 y 与 x 的连接，得如图 3-5(d)所示的状态图。

所有弧上的标记都属于 $\Sigma\cup\{\varepsilon\}$，构造完毕。

上面讨论了由正规式到非确定的有限自动机的转换，同样，对于一个字母表 Σ 上的非确定的有限自动机 M，也可以在 Σ 上构造相应的正规式 r，使 $L(r)=L(M)$。

根据前面的定理，构造之前，首先对非确定的有限自动机 M 对应的有限自动机进行拓广，加进两个状态，一个为唯一初态 q_S，一个为唯一终态 q_Z。具体用状态图描述的构造步骤如下：

(1) 在非确定的有限自动机 M 的状态转换图中加进两个结点，一个为 q_S结点，另一个为 q_Z结点。其中 q_S是唯一的开始状态，q_Z是唯一的终止状态。

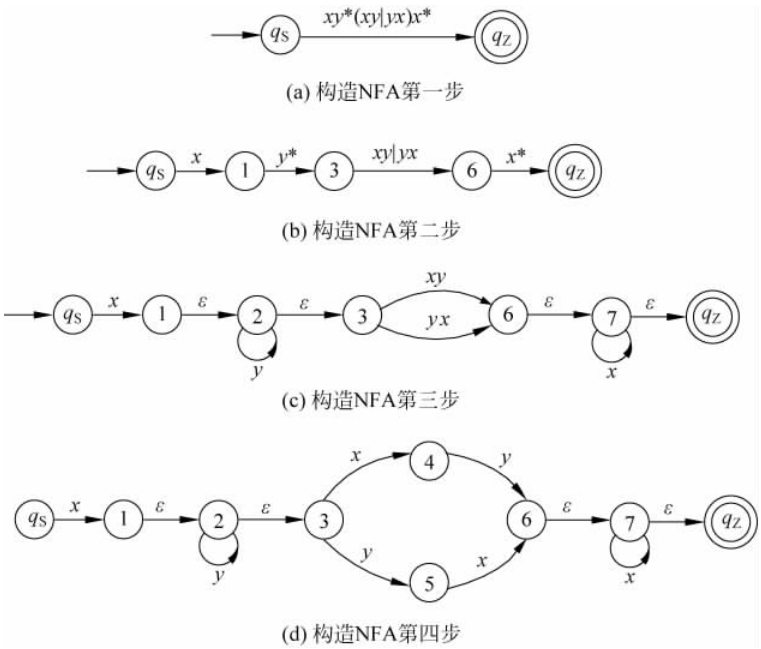


图 3-5　正规式 $r=xy^*(xy|yx)x^*$ 的分解

（2）从 q_S 用 ε 弧连接到 M 的初态结点，从 M 的所有终态结点用 ε 弧连接到 q_Z 结点，形成一个与 M 等价的 M'。M' 有一个没有射入弧的初态结点 q_S 和一个没有射出弧的终态结点 q_Z。

（3）对新的非确定的有限自动机按照图 3-6 所示的替换规则 2 进行替换，这个过程实际上是正规式的合成过程，即对 M' 不断消去结点和弧的过程。直到状态图中只剩下状态结点 q_S 和 q_Z 为止。当状态图中只有状态 q_S 和 q_Z 时，在 q_S 到 q_Z 的弧上标记的正规式即所求结果。

【例 3-4】 设非确定的有限自动机 M 的状态转换图如图 3-7 所示。在 $\{x,y\}$ 上构造一个正规式 r，使 $L(M)=L(r)$。

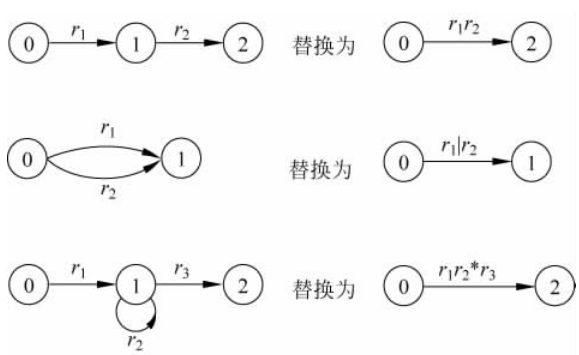


图 3-6　替换规则 2

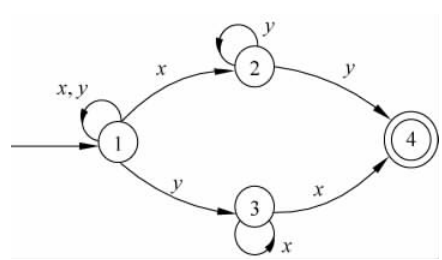


图 3-7　NFA M 的状态图

第一步，拓广状态图，得如图 3-8(a)所示的状态图。

第二步，消去状态 2 和 3，得如图 3-8(b)所示的状态图。

第三步，消去状态 4，得如图 3-8(c)所示的状态图。

第四步，消去状态 1，得如图 3-8(d)所示的状态图。

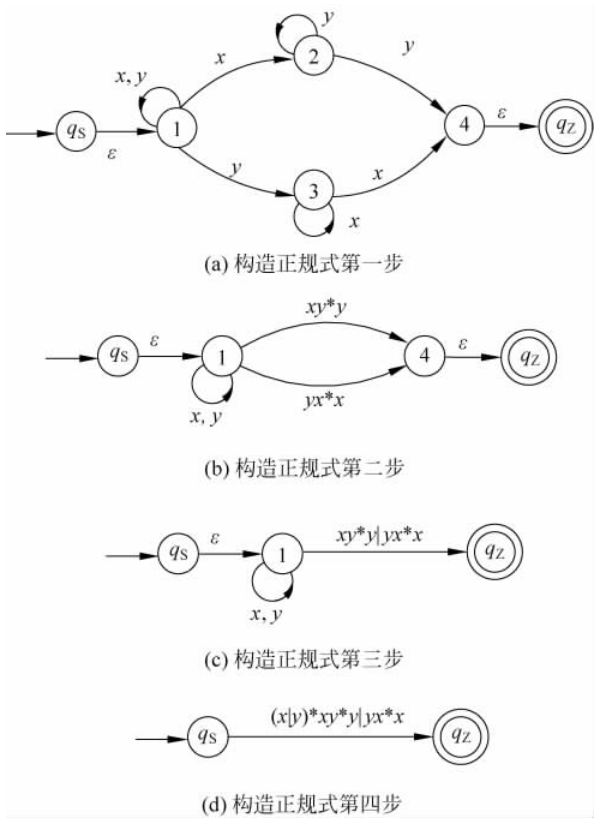


图 3-8　NFA M 到正规式的转换

此时状态转换图中只有状态 q_S 和 q_Z，所以非确定的优先自动机 M 对应的正规式为 $r=(x|y)^*(xy^*y|yx^*x)$。

【例 3-5】 非确定的有限自动机 M 的状态图如图 3-9 所示，写出与其语言等价的正规式 r。

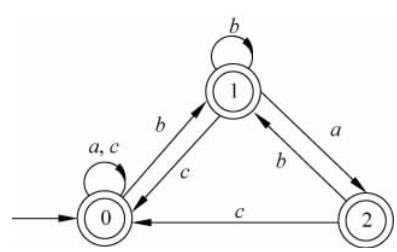


图 3-9　NFA M 的状态图

第一步，拓广文法，得如图 3-10(a)所示的状态图。

第二步，消去状态 2，得如图 3-10(b)所示的状态图。

第三步，化简状态 1 到状态 0、状态 1 到状态 1、状态 1 到状态 q_Z 间的弧，使之不存在并行的弧，得如图 3-10(c)所示的状态图。

第四步，消去状态 1，得如图 3-10(d)所示的状态图。

第五步，化简状态 0 到状态 0、状态 0 到状态 q_Z 间的弧，使之不存在并行的弧，得如图 3-10(e)所示的状态图。

第六步，消去状态 0，得如图 3-10(f)所示的状态图。

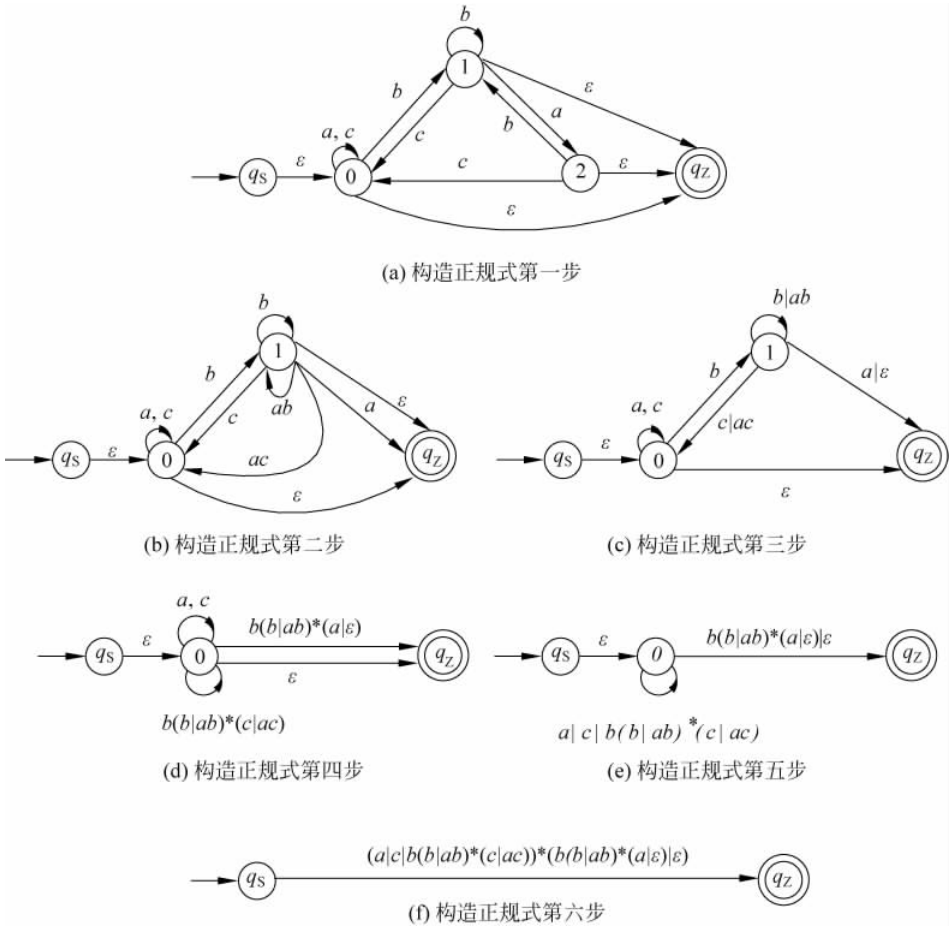


图 3-10　NFA M 的正规式的转换

状态图中只有状态 q_S 和 q_Z，所以正规式 $r=(a|c|b(b|ab)^*(c|ac))^*(b(b|ab)^*(a|\varepsilon)|\varepsilon)$。

3.2.3　正规文法和有限自动机的等价性

定理 3-2　如果对于某个正规文法 G 和某个有限自动机 M，有 $L(G)=L(M)$（即正规文法 G 所产生的语言和某个有限自动机所识别的语言相同），则称 G 和 M 是等价的。

定理 3-3　对每一个右线性正规文法或左线性正规文法 G，都存在一个有限自动机 M，使 $L(M)=L(G)$。

定理 3-4　对于每一个有限自动机 M，都存在一个右线性正规文法 G 和一个左线性正规文法 G'，使 $L(M)=L(G)=L(G')$。

本节主要考虑右线性文法与有限自动机之间的等价转换，对于左线性文法与有限自动机的相互转换，有兴趣的读者可以自行考虑分析。

首先考虑有限自动机转换为等价的右线性文法，算法描述如下：

(1) 对转换函数 $f(A,t)=B$，可写成一个产生式 $A\rightarrow tB$。

(2) 对可接受状态 Z,增加一个产生式 $Z\rightarrow\varepsilon$。

(3) 有限自动机的初态对应于文法的开始符号,有限自动机的字母表为文法的终结符号集。

【例 3-6】 给出与如图 3-11 所示的有限自动机等价的正规文法 G。

与该有限状态机等价的文法为 $G=(\{A,B,C,D\},\{a,b\},A,P)$,其中 P: $A\rightarrow aB$,$A\rightarrow bD$,$B\rightarrow bC$, $C\rightarrow aA$,$C\rightarrow bD$,$C\rightarrow\varepsilon$,$D\rightarrow aB$,$D\rightarrow bD$,$D\rightarrow\varepsilon$。

接着考虑右线性文法转换为等价的有限状态机,算法描述如下:

(1) 字母表与 G 的终结符号相同。

(2) 为 G 中的每个非终结符生成有限状态机 M 的一个状态,G 的开始符号 S 是开始状态 S。

(3) 增加一个新状态 Z,作为 M 的终态。

(4) 对 G 中的形如 $A\rightarrow tB$,其中 t 为终结符或 ε,A 和 B 为非终结符的产生式,构造 M 的一个转换函数 $f(A,t)=B$。

(5) 对 G 中的形如 $A\rightarrow t$ 的产生式,构造 M 的一个转换函数 $f(A,t)=Z$。

【例 3-7】 设有以下文法

$$G[S]:S\rightarrow aA\mid bB\mid\varepsilon$$
$$A\rightarrow aB\mid bA$$
$$B\rightarrow aS\mid bA\mid\varepsilon$$

求与文法 $G[S]$等价的非确定的有限状态机。

求得等价的非确定的有限状态机如图 3-12 所示。

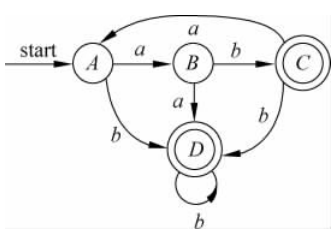


图 3-11 有限状态机状态转换图

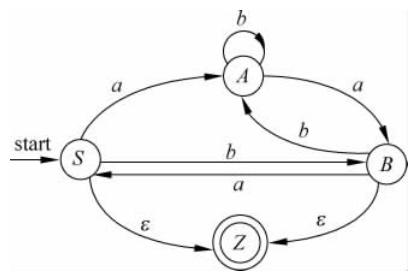


图 3-12 等价的 NFA

3.3 单词的识别

3.3.1 单词分类

逐个读入源程序字符并按照构词规则切分成一系列单词。单词是语言中具有独立意义的最小单位,包括保留字、标识符、运算符、标点符号和常量等。例如,在 C 语言中,一般的表达式不一定是单词,因为它们虽然具有独立的意义,但不一定是最小的语法单位。如表达式 $a+b$,它的单词是 a、+、b。C 语言中的字母和数组也不一定是单词,因为它们虽然是最小的语法单位,但常常不具有独立意义,如程序中定义一个变量 $a10$,a 和 10 若单独作为

一个变量名和一个常数存在于程序中，则是合法的单词，但作为对变量标识的定义，只有 $a10$ 作为一个整体才是合法的单词。到底哪些语法符号是语言中具有独立意义的最小语法单位呢？这与具体语言的词法规则有关。一般常用的程序设计语言的单词可分为这样几类：

(1) 保留字，也称关键字。保留字一般是语言系统本身定义的，通常是由字母组成的字符串。例如，C 语言中的 int 、float、while、for、break 和 switch 等。

(2) 常数。程序设计语言中包含各种类型的常数，如整型常数、实型常数、不同进制的常数、布尔常数、字符及字符串常数等。

(3) 标识符。用来表示各类名字的标识，如常量名、变量名、数组名、结构名、函数名、类名、对象名和文件名等。

(4) 运算符。表示程序中算术运算、逻辑运算、字符及位串等运算的字符(或串)，如各类语言中较通用的＋、－、＊、/、＜＝、＞＝、＜和＞等。还有一些语言包含特有的运算符，如 C 语言中的＋＋、?:、||、＋＝等。

(5) 界限符。如逗号、分号、括号、单引号和双引号等。

词法分析程序所输出的单词符号常常采用以下二元式表示：(单词种类，单词值)。其中，单词种类表示单词的类别，用来刻画和区分单词的特性或特征，通常用整数编码来表示。例如，可以将 C 语言中的关键字设计为一类，标识符为一类。运算符既可以按优先级分为不同的类，也可以作为一类。单词值则是编译其他阶段需要的信息，可以省略。例如，在 C 语言的语句 int i=10, j=1; 中的单词 i 和 j 的种类都是整型，单词的值为 10 和 1 对于代码生成来说是必不可少的。有时，对某些单词来说，不仅需要它的值，还需要其他一些信息以便编译的进行。比如，对于标识符来说，还需要记载它的类别、层次和其他属性，如果这些属性统统收集在符号表中，那么可以将单词的二元式表示设计成如下形式：(标识符，指向该标识符所在符号表中位置的指针)。

对于一个语言来说，如何对其中的单词进行分类，分成几类，怎样编码，单词属性部分能包含多少信息等，并没有一个原则性的规定，要视具体情况而定，主要取决于处理上的方便。一般来说，一个单词为一类，处理较为方便，但对于标识符却是不可行的。可以将具有一定共性的单词视为一类，统一给出一个属性，但属性部分信息包含越多，实现起来越复杂。

属性字中的单词值部分要直接或间接给出单词在机内存储的内码表示。如对于某个标识符或某个常数，常把指向存放它的有关信息的符号或常数表入口的指针作为它的单词值。

3.3.2 单词的内部表示

单词的内部表示 Token 的结构一般由两部分组成：单词类别和语义信息。单词类别用来区分单词的不同种类，通常可以用整数编码来表示。单词的语义信息也取决于今后处理上的方便。对于常量和标识符还可以单独构造常量表和标识符名字表，此时，单词的语义信息的值就是指向常量表或标识符名字表中相应位置的指针。

3.3.3 单词的形式化描述

描述程序设计语言中单词的工具主要有以下 3 种：正则表达式、自动机和正则文法。

它们的功能彼此相当。对于一个一般的程序设计语言，各类单词的正则表达式可能如下：

(1) 标识符：$L(L|D)^*$，其中 $L=[a-z, A-Z]$，$D=[0-9]$

(2) 整数：$D1D^*$，其中 $D1=[1-9]$

(3) 特殊符号：+|；|：|:=|<|<=|…

(4) 保留字：begin|end|while|…

构造识别单词的有限自动机的方法与步骤如下：

(1) 根据构成规则对程序语言的单词按类构造出相应的状态转换图。

(2) 合并各类单词的状态转换图，构成一个能识别该语言的所有单词的状态转换图。合并方法为如下：

① 将各类单词的状态转换图的初始状态合并为一个唯一的初始状态。

② 化简调整状态冲突和对冲突状态重新编号。

③ 如果有必要，增加出错状态。

3.4 词法分析程序的设计及实现

词法分析程序的设计步骤如下：

(1) 确定词法分析器的接口，即确定词法分析器是作为语法分析的一个子程序还是作为独立的一遍。

(2) 确定单词分类和单词结构。

(3) 构造每一类单词的正规式以及对应的 NFA。

(4) 对各类单词对应的 NFA 进行合并，构成一个能识别该语言所有单词的等价 DFA。

(5) 编程实现第(4)步得到的 DFA。

3.4.1 词法分析程序的输入

词法分析器工作的第一步是源程序以字符流形式输入。根据内存空间情况有两种选择：

(1) 如果内存比较大，可以直接输入到内存的一个源程序区，然后词法分析程序从源程序区依次读入字符进行扫描和处理。这种方式可以大大节省源程序输入的时间，提高词法分析器的效率。但在一个有限的内存空间内要满足各种规模的源程序的一次输入是困难的，这样的系统开销也过大的。

(2) 如果内存不足，将源程序以文件的形式存储在外部介质上，如磁盘、磁带等。可以先在内存中开辟一个大小适宜的缓冲区，这个缓冲区称为输入缓冲区。词法分析程序工作时，先从外部介质上将输入符号串分批读入缓冲区，单词符号的识别可在这个缓冲区中进行。

3.4.2 扫描缓冲区及其预处理

扫描缓冲区就是在内存中开辟一部分单元，供识别单词用。注意：扫描缓冲区和 3.4.1 节提到的输入缓冲区是不同的，输入缓冲区是从外存上读入部分字符，而扫描缓冲区仅供识别单词用。按照上述的输入缓冲区和预处理子程序的概念，可以给出一个带有两个缓冲区

的词法分析器，其结构如图 3-13 所示。

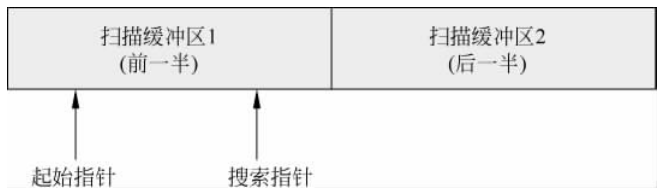


图 3-13　扫描缓冲区结构示意图

在扫描缓冲区中一般设两个指针，一个指向当前正在识别的单词的开始位置，另一个用于向前搜索寻找单词的终点。不论扫描缓冲区定为多大，都不能保证单词符号不会被它的边界所打断，因此，扫描缓冲区最好使用如图 3-13 所示的一分为二的区域。

开始时两个扫描缓冲区均为空。首先填满扫描缓冲区 1，并将两个指针指向扫描缓冲区一的第一个字符。将搜索指针向后移动，当识别出一个单词之后，搜索指针已指向下一个单词的第一个字符，然后再将起始指示器移到搜索指示器指字符，接着搜索指针又开始扫描第二个单词。当搜索指针越界时，说明扫描缓冲区 1 中的字符不足一个单词，这时填满扫描缓冲区 2，再将搜索指示器扫描缓冲区 2 第 1 个字符。这样，两个扫描缓冲区交替工作。一般，扫描缓冲区的长度可以存放最长的一个单词即可正常工作，否则就不能保证单词符号不会被扫描缓冲区边界所打断。

扫描缓冲区两个半区互补功能的算法描述如下(设 F 为搜索指针)：

```
if F at end of first half
{
    reload second half;
    F++;
}
else if F at end of second half
{
    reload first half;
    move F to beginning of first half
}
else F++;
```

实际的词法分析程序往往带有预处理子程序，因此它真正接受的输入是经过预处理后的源程序串。这是由于在源程序中，特别是非自由格式书写的源程序中，往往有大量的空白符、回车换行符及注释等，这是为增加程序可读性及程序编辑的方便而设置的，对程序本身无实际意义。另外，像 C 语言有宏定义、文件包含、条件编译等语言特性，为了减轻词法分析器实质性处理的负担，在源程序从输入缓冲区进入词法分析器之前，要先对源程序进行预处理，预处理子程序一般完成的主要功能如下：

- 滤掉源程序中的注释。
- 剔除源程序中的无用字符。
- 进行宏替换。
- 实现文件包含的嵌入和条件编译的嵌入等。

3.4.3 由词法规则画出状态转换图

设C语言子的一个集由下列单词符号构成，以正规式的形式表示如下。

关键字：int,if,for

标识符：字母(字母|数字)*

无符号整常数：数字(数字)*

运算符或分界符：=,*,+,++,+=,{,}

根据3.2.2节正规式等价转化成NFA的方法，可以得到如图3-14所示的单词状态转换图。

将各类状态转换图合并，构成一个能识别语言所有单词的状态转换图。然后化简并调整冲突和状态编号。

图3-14给出的各类单词的状态转换图经合并和调整后，得到识别标识符号和正整数的状态转换图，如图3-15所示。

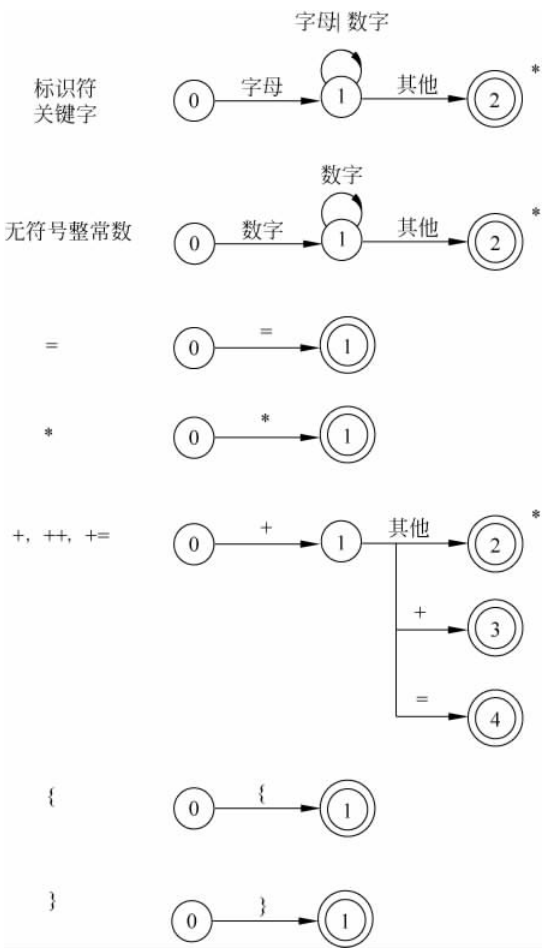


图3-14 C语言子集各类单词状态转换图

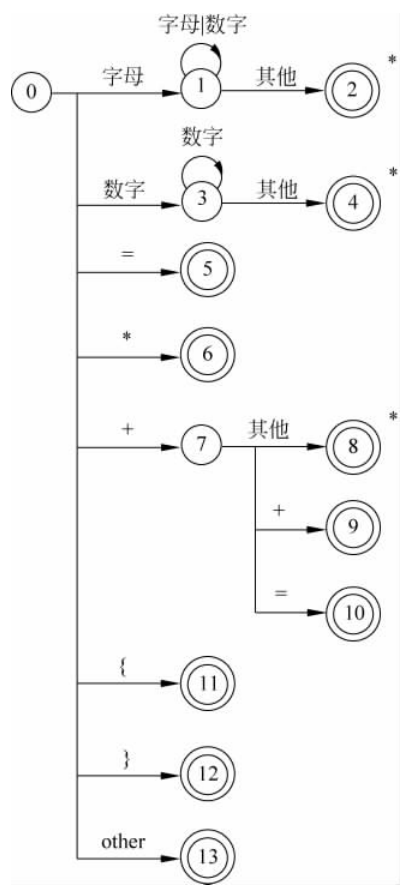


图3-15 C语言子集单词状态转换图

3.4.4 单词对应状态转换图的实现

根据语言的词法规则构造出识别其单词的状态转换图，仅仅是理论上的词法分析器，是一个数学模型。那么，如何将状态转换图变为一个可行的词法分析器呢？最常用的状态转换图的实现方法称为程序中心法，即把状态转换图看成一个流程图，从状态转换图的初态开始，对它的每一个状态结点编一段相应的程序。

为此，给出图 3-15 所示的状态转换图实现的 C 语言伪代码如下。

```
int state = 0;
enum letter('a'… 'z');
enum letter('0'… '9');
char char1;
  {
   char1 = nextchar();
   switch(state)
   {
     case 0:switch(char1)
          {
            case 'a'… 'z':state = 1;break;
            case '0'… '9':state = 3;break;
            case ' = ':state = 5;break;
            case ' * ':state = 6;break;
            case ' + ':state = 7;break;
            case '{':state = 11;break;
            case '}':state = 12;break;
            default:state = 13;
          }
          break;
     case1:while(char1 == letter||number)
          char1 = nextchar();
          state = 2;
          break;
     case2:untread();
        return(02,value)or return(01,value);
        break;
     /* 函数 untread()回退一个已读进的字符,属性 01 表示关键字,属性 02 表示标识符 */
     case3:while(char1 == number)char1 = nextchar();
       state = 4;
       break;
     case4:untread();return(03,value);break;          /* 属性 03 表示无符号整常数 */
     case5:return(04, );break;                        /* 属性 04 表示 = */
     case6:return(05, );break;                        /* 属性 05 表示 * */
     case7:if(char1 == ' + ')state = 9;
           else if(char1 == " = ")state = 10;
                else state = 8;
           break;
```

```
    case8:untread();return(08, );break;          /* 属性 08 表示 + */
    case9:return(09, );break;                    /* 属性 09 表示++ */
    case10:return(12, );break;                   /* 属性 12 表示 += */
    case11:return(10, );break;                   /* 属性 10 表示{ */
    case12:return(11, );break;                   /* 属性 11 表示} */
    case13:error();                              /* error 是语法错处理函数 */
    }
}
```

3.5 词法分析程序的自动实现

本节介绍一个著名的词法分析器自动生成工具 Lex。它是以有限自动机理论为基础而设计的,本质上是一个可以识别所有模式的确定有限自动机。

3.5.1 Lex 介绍

Lex 是一个词法分析程序的自动生成器(lexical analyzer generator)。1972 年贝尔实验室在 UNIX 上首次实现了 Lex。1984 年,GNU 工程推出 Flex(Fast Lexical Analyzer Generator),它是对 Lex 的扩充。Lex/Flex 支持使用正则表达式来描述各个单词的模式,由此给出一个词法分析器的规约。

Lex 工具的输入表示方法称为 Lex 语言,而工具本身称为 Lex 编译器。图 3-16 表示了 Lex 编译器的主要组成部分。从图中可以看出,Lex 编译器接收用正规式表示的单词规则,然后利用算法 X 从此正规式出发构造能识别正规式描述的单词集(正规集)的非确定有限自动机 M',再用算法 Y(即子集法)将 M'确定化,得到与之等价的确定有限自动机 M',最后还可用算法 Z(即划分算法)对 M'进行化简,得到确定的有限自动机 M,则这个有限自动机 M 即是理论上的扫描器。

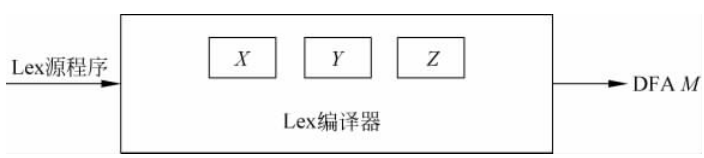


图 3-16 词法分析器自动构造的思想与实现

使用 Lex 一般分为 3 个步骤,如图 3-17 所示。

(1) 用 Lex 语言编写一个输入文件,描述将要生成的词法分析器。在图 3-17 中这个输入文件称为 lex.l。

(2) Lex 编译器将 lex.l 转换成 C 语言程序,生成的文件称为 lex.yy.c。

(3) lex.yy.c 被 C 编译器编译为一个名为 a.out 的文件。C 编译器的输出就是一个读取输入字符流并生成单词流的可运行的词法分析器。

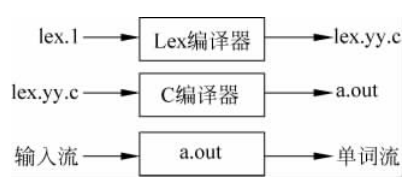


图 3-17 用 Lex 生成一个词法分析器

3.5.2 Lex 语法基础

Lex 语言是对表示语言单词集的正规式的描述，以解决正规式规则输入问题。一个 Lex 程序具有如下形式：

```
说明部分                    /* 包含模式宏定义和 C 语言的说明信息 */
%%
规则部分                    /* 转换规则 */
%%
辅助函数                    /* 规则动作部分所需的辅助过程的 C 语言代码 */
```

其中，第一部分“说明部分”包括 C 语言代码、模式宏定义等。模式宏定义实际是对识别规则中出现的正规式的辅助定义。如语言中的字母可定义为

$$\text{Letter}[A \mid B \mid \cdots \mid Z \mid a \mid b \mid \cdots \mid z]$$

数字可定义为

$$\text{digital}[0 \mid 1 \mid 2 \mid \cdots \mid 9]$$

除宏定义外，定义部分的其余代码须用符号%{和%}括起来。另外，Lex 源程序所使用的 C 语言库文件和外部变量也应分别用 #include 及 extern 予以说明，并置于%{和%}之内。例如，有 C 语言说明和 Lex 宏定义示例如下：

```
%{
#include <stdio.h>
#include <ctype.h>
      /* definitions of manifest constants LT,LE,EQ,GT,GE,IF,THEN,ELSE,ID */
%}
      /* regular expression */
      delim     [\t\n]
      ws        {delim}+
      letter    [A-Za-z]
      digit     [0-9]
      id        {letter}({letter}|{digit})*
%%
```

在 Lex 源程序中，起标识作用的符号%%、%{和%}都必须处于所在行的最左字符位置。另外，在其中也可以随意添加 C 语言形式的注释。

第二部分“规则部分”是 Lex 程序的主体部分。其一般形式如下：

```
模式 1 {动作 1}
模式 2 {动作 2}
⋮
模式 n{动作 n}
```

其中模式是对单词的描述，用正规式表示。动作是与匹配的模式对应的，用 C 语言代码表示对模式处理的动作，表示当识别出某个模式所表示的单词后，词法分析器需要做的处理工作，即应执行动作的程序。通常使用的 Lex 模式定义如表 3-2 所示。

表 3-2 Lex 模式定义

模 式	说 明
x	匹配单个字符 x
[abc]	匹配 a 或 b 或 c
[^abcde]	匹配除 a～e 以外的任意字符(可写为[^a－e])
\	转义符定义同 ANSI C
·	匹配除换行符之外的任意字符
r*	r 是正规式,r* 匹配 0 个或多个 r
r＋	r 是正规式,r＋匹配 1 个或多个 r
r?	r 是正规式,r? 匹配 0 个或 1 个 r
r{2,5}	r 同上,匹配 2～5 个 r
r{2,}	r 同上,匹配 2 个或更多个 r
r{2}	r 同上,匹配 2 个 r
{name}	name 是在定义部分出现的模式宏名
"text"	匹配字符串"text"
r|s	匹配正规式 r 或 s
Rs	匹配正规式 r 和 s 的连接

第三部分“辅助函数”定义对模式进行处理的 C 语言函数、主函数等。作为辅助过程,它是支持规则的动作部分所需要的处理过程,是对规则部分中动作的补充。这些过程若不是 C 语言的库函数,需给出具体定义,然后分别编译且与生成的词法分析器装配在一起。

需要说明的是,说明部分和辅助函数部分是任选的,规则部分是必需的。

3.5.3 词法分析器自动构造

词法分析程序主要由两部分组成:一张转换矩阵表和一个控制程序。转换矩阵表由 Lex 对某语言单词集描述的 Lex 源程序变换得来。基于 Lex 源程序,Lex 编译器的实现步骤如下:

(1) 对 Lex 源程序识别规则中的每个P_i构造一个相应的非确定的有限自动机M_i。引入唯一初态 X,从初态 X 通过 ε 弧将所有非确定的有限自动机$M_i(i=1,2,\cdots,n)$连接成新的确定的有限自动机M'。

(2) 对非确定的有限自动机M_i确定化,产生确定的有限自动机 M。

(3) 化简确定的有限自动机 M。

(4) 给出总控程序。尽管程序设计语言不同,但是总控程序却是一样的。总控程序的作用是激活有限自动机,即控制输入字符串 α 在有限自动机上运行,一旦到达有限自动机的终态,即识别出 Lex 源程序模式描述的某类单词,即转去调用相应的动作部分$\{A_i\}$的处理程序完成所识别单词的处理工作。总控程序的实现流程如图 3-18 所示。

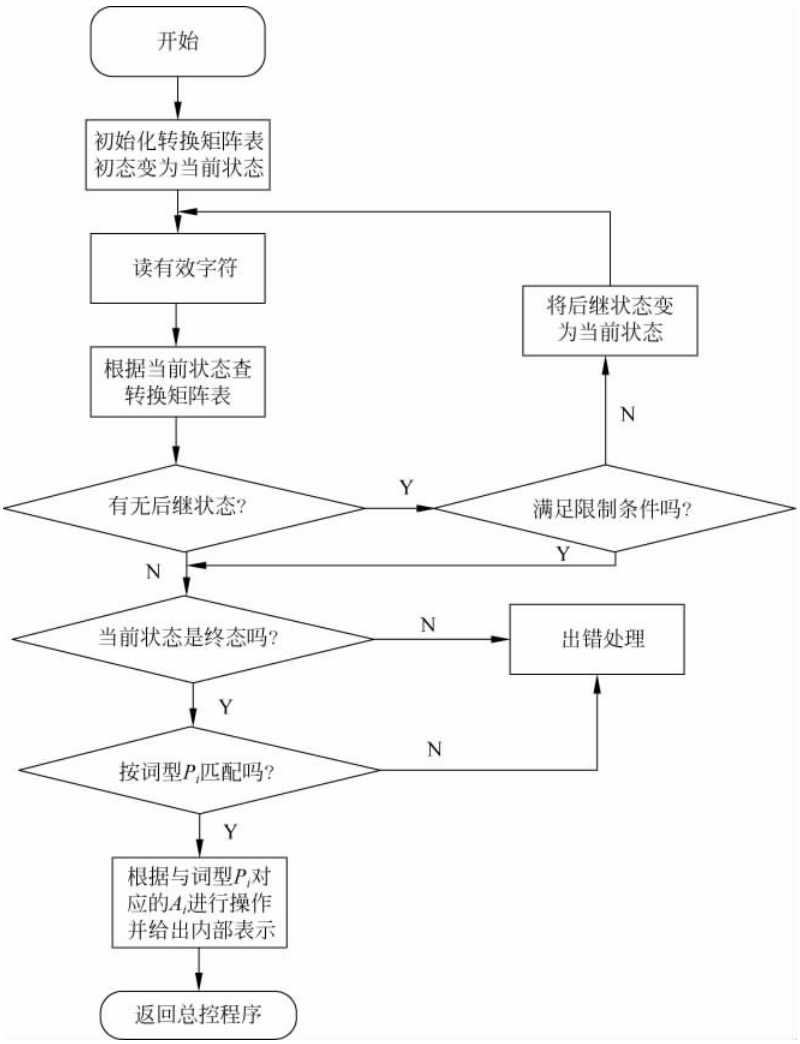


图 3-18　Lex 总控程序流程图

【例 3-8】 表 3-3 给出某语言的单词的词法分析结果。

表 3-3　某语言的单词及词法分析结果

单词的正规式表示	属性	单词值	单词的正规式表示	属性	单词值
ws			<	relop	LT
if	if		<=	relop	LE
then	then		=	relop	EQ
else	else		<>	relop	NE
id	id	名表指针	>	relop	GT
num	num	常数表指针	>=	relop	GE

该语言的 Lex 源程序如下：

```
% {
```

```
# include < stdio. h >
# include < ctype. h >
# include < string. h >
# define IF                    1
# define THEN                  2
# define ELSE                  3
# define ID                    4
# define LT                    5
# define LE                    6
# define EQ                    7
# define NE                    8
# define GT                    9
# define GE                    10
% }
/ * 正规式模式宏定义 * /
digit                          [0 - 9]
alpha                          [a - zA - Z]
alnum                          [a - zA - Z0 - 9]
delim                          [ \t\n]
ws                             {delim]} +
id                             {letter}({letter}|{digit}) *
number                         {digit} + (\.{digit} + )?(E[ + \ - ]?{digit} + )?
% %
{ws}                           {/ * no action and no return * /}
If                             {return (IF);}
then                           {return (THEN);}
else                           {return (ELSE);}
{id}                           {yyIval = install_id(); return(ID);}
{number}                       {yyIval = install_num(); return(NUMBER);}
"<"                            {yyIval = LT; return(RELOP);}
"< = "                         {yyIval = LE; return(RELOP);}
" = "                          {yyIval = EQ; return(RELOP);}
"<>"                           {yyIval = NE; return(RELOP);}
">"                            {yyIval = GT; return(RELOP);}
"> = "                         {yyIval = GE; return(RELOP);}
int main()
{
    yyparse();
    return 0;
}
int yyerror(char * msg)
{
    printf("Error
    encountered: % s \n", msg);
}
```

程序开始是内部表示常数的定义，即用字符%{和%}括起来的部分。接下来是说明部分的模式宏定义。正规式规则定义部分出现的第一个正规名 delim，它对字符属性的描述是无意义的，仅表示 3 个字符，即空白、回车标记(用\t 表示)、换行标记(用\n 表示)。可以注意到正规式允许递归定义。

程序中第一次出现的字符%%标记识别规则部分，表示当某个模式被识别，则执行其后的处理动作。例如，当关键字 if 被识别，则产生执行语句 return(IF)，该单词属性为 IF，此时，单词属性字的值部分为空。在识别规则中，对关系运算符“<”的识别，其动作部分由两个语句来完成，一是给出单词“<”的属性 RELOP，由语句 return(RELOP)完成；二是给出单词的内部值，由语句 yylval=LT 完成；而对标识符的识别，其动作部分也是由两个语句完成的，但在给出标识符内部值的语句中，则是转去调用辅助过程部分的一个过程 install_id，即由该过程给出相应的标识符的内部值。

Lex 程序的第三部分是辅助过程，定义了两个用 C 语言编写的函数，省略了具体代码，仅给出了函数的功能说明，其功能是完成对标识符及数值型常数的内部值的处理。

习题

1. 写出表示下列语言的正规式。

(1) $\{x \mid x \in \{0,1\}^*$ 且 x 中不含形如 01 的子串$\}$

(2) $\{x \mid x \in \{0,1\}^+$ 且 x 中至少含两个 1$\}$

(3) $\{x \mid x \in \{0,1\}^+$ 且 x 中最多只含两个 1 的 0、1 串构成的语言$\}$

(4) $\{x \mid x \in \{0,1\}^+$ 且 x 以 1 开头和以 1 结尾的 0、1 串构成的语言$\}$

2. 构造下列正规式的等价 DFA。

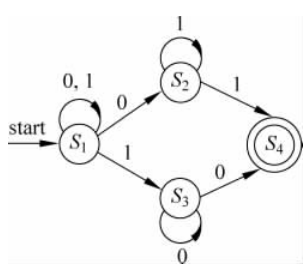
(1) $(a|b)^*$

(2) $(a^*|b^*)^*$

(3) $(0|1)*(0*|1)*(0|1)$

(4) $((\varepsilon|a)b^*)^*$

3. 设有 NFA M 如图所示，在 $\Sigma=\{0,1\}$ 上构造一个正规式 R，使 $L(M)=L(R)$。



4. 已知文法 $G=(\{S,A,B\},\{a,b,c\},P,S)$，其中：

P：$S \rightarrow aS \mid aB$

$B \rightarrow bB \mid bA$

$A \rightarrow cA \mid c$

构造一个与文法 G 等价的正规式 R。

5. 假设有两类单词：①关键字 if；②标识符，它表示除 if 之外的所有字母组成的串。

(1) 给出识别这两类单词的 NFA。

(2) 给出识别这两类单词的 DFA。

6. 给出 C、C++和 Java 语言的全部单词或单词类。

7. C++有各种整型常量：十进制数的简单序列总为整型常量，0x 为前缀的十六进制数序列为整型常量，以 0 为前缀的八进制数序列为整型常量，以 L 或 l 为后缀的整型常量表示的类型为 long int。编写一个描述用于识别 C++中整型常量的正规式，并构造相应的 NFA 和 DFA。

8. 编写一个 Lex 程序，该程序将程序中的关键字 int 的每个实例转换成 float。

9. 编写描述 C++的单词符号的 Lex 程序。