第三章 词法分析 →∪Σε

3.4 词法单元识别

*digit*  🡪 [0-9]

*digits* 🡪 *digit*+

*number* 🡪 *digits*(.*digits*)?(E[+-]?*digits*)?

*letter* 🡪 [A-Za-z]

*id* 🡪 *letter* (*letter* | *digit*)\*

*if* 🡪 if

*then*  🡪 then

*else* 🡪 else

*relop* 🡪 < | > | <= | >= | = | <>

s = s1s2s3s4..sn

if (a > 0) {

a = 100;

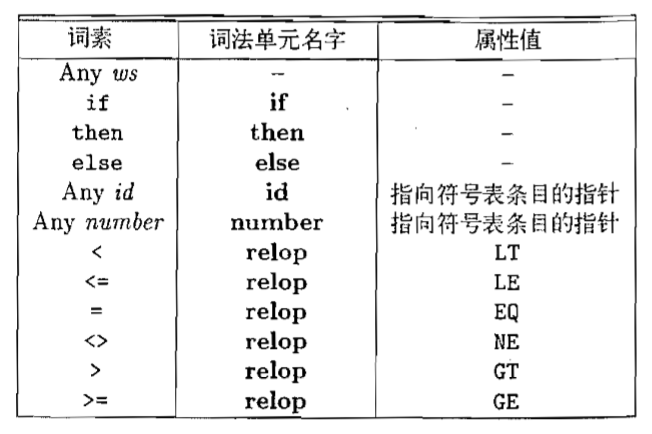
}

识别字符串s中的词法单元.

词法分析器负责消除空白符，需要另一个词法单元ws (whitespace),

ws → ( blank | tab | newline ) +

(注意: 词法分析器不会将ws词法单元返回给语法分析器), 制定ws只是为了过滤程序中的空白字符.

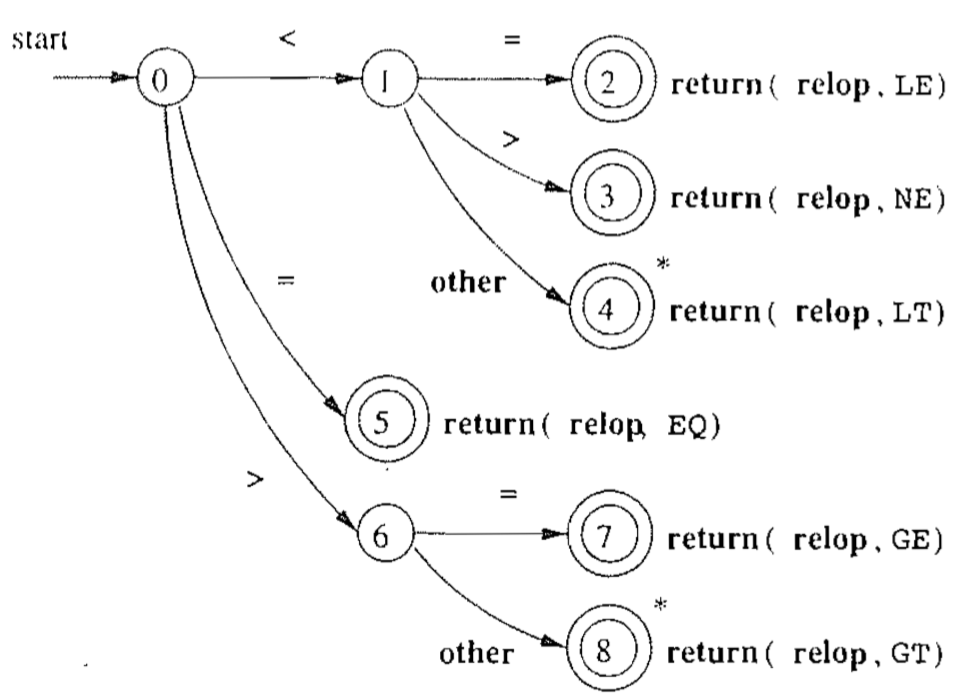


3.4.1 状态识别图

可以更好地呈现 模式怎么来匹配字符串.

例1, 关系运算符的状态识别图

假设正则表达式 relop → < | > | <= | >= | = | <>



状态转换图 (transition diagram):

* 圆圈表示状态, 开始状态由一条没有出发节点、标号为“开始/start”的边指明.
* 双层圆圈表示接受状态, 表示已识别一个词法单元.
* 有向边表示从一个状态到另一状态.
* 每条边的标号包含一个或多个符号，若离开状态s的某边上标号为other，则它表示离开s的其他边所指示的字符以外的任意字符.

分析下面两个符号串的识别过程

<=a

<a

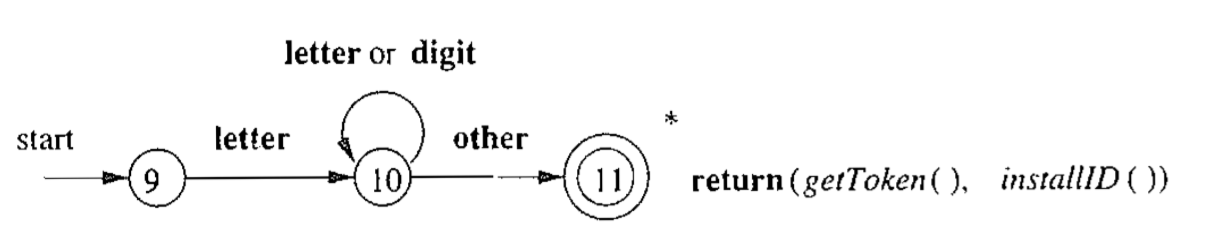
例2 标识符的状态转换图

letter → [A-Za-z\_]

digit → [0-9]

id → letter(letter|digit)\*

画出它的状态转换图

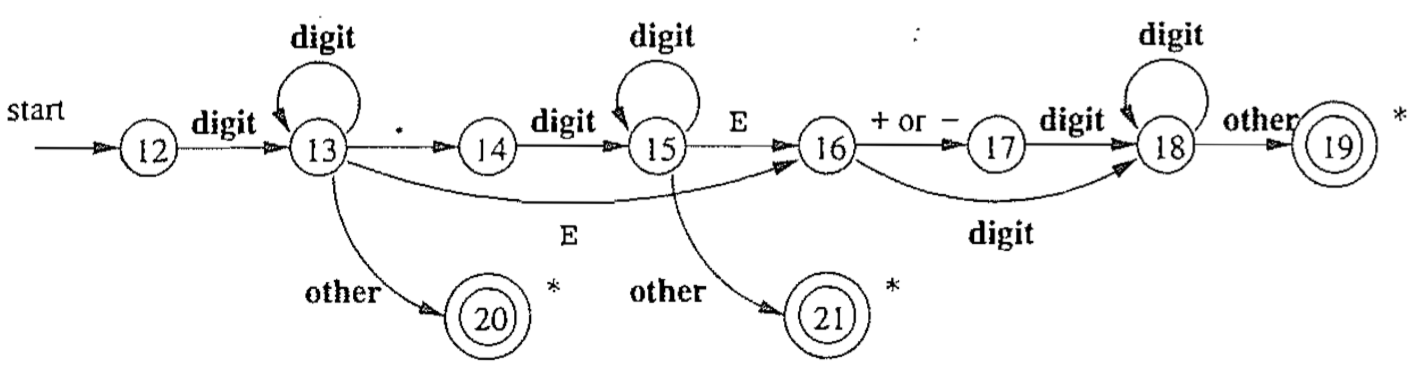


分析ab+3的识别过程

问(1) 如何区分关键字和标识符??

例3 无符号数的状态转换图

**number** → **digit+ (.digit+)? (E (+ | -)? digit+)?**

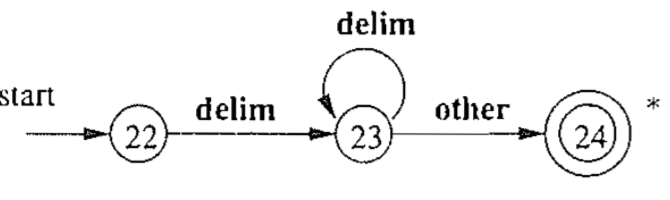


分析 1.93E-3+a 的识别过程

例4 空白符的状态转换图

delim → blank | tab | newline

ws → delim+



问(2) 如何将前面几个转换图集成在一块?

开始

练习1, 画出下面各正则表达式的转换图:

1) a(a|b)\*a

2) ((ε|a)b\*)\*

3) (a|b)\*a(a|b)(a|b)

4) a\*ba\*ba\*ba\*