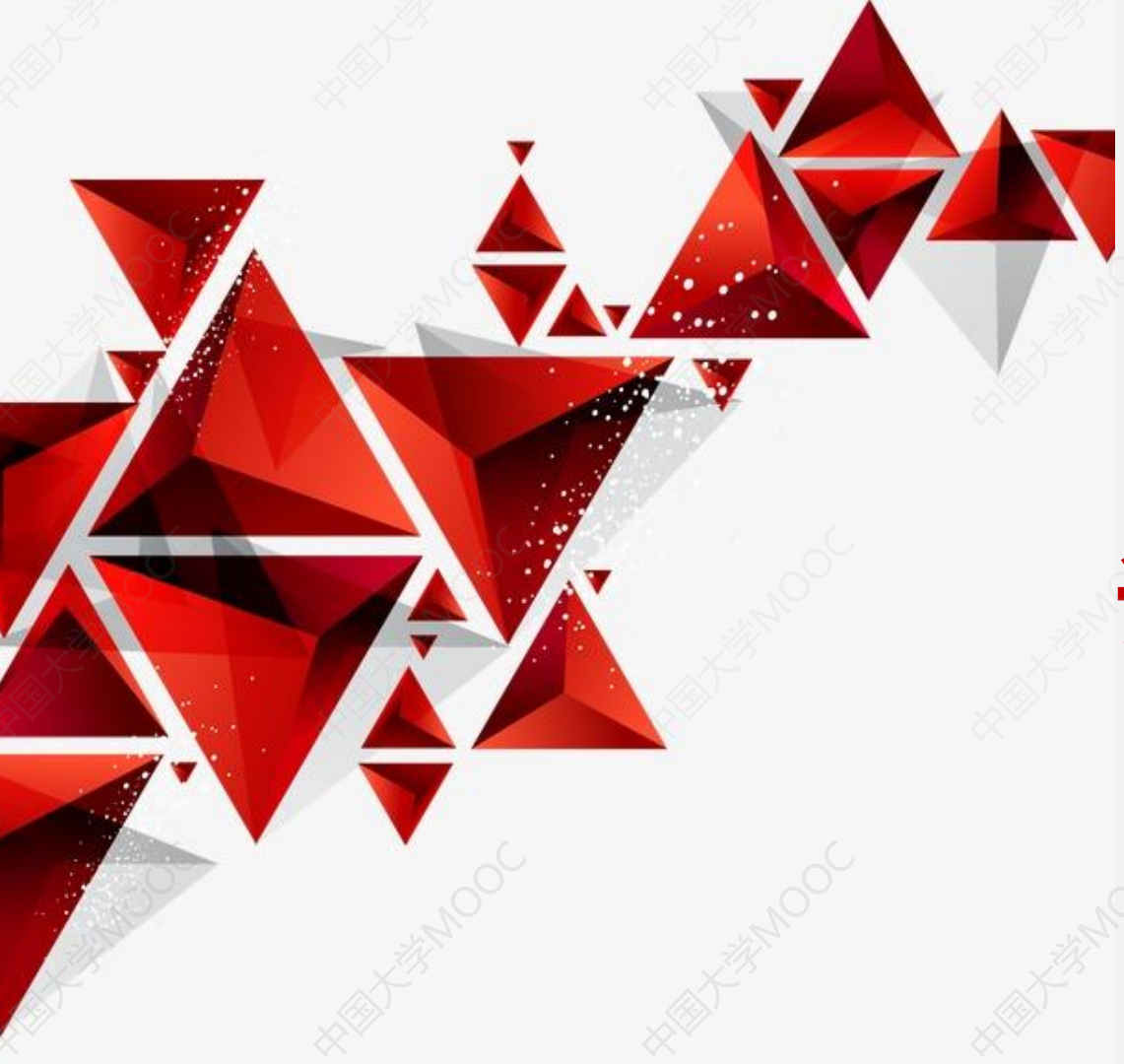


编译技术

词法分析

大连理工大学软件学院

本讲纲要

词法模式的表示方法，是词法记号描述的核心

01 正规式的定义和运算

02 正规式和状态转换机之间的关系

正规式

- 正规式：按照一组定义规则，由较简单的正规式构成的，每个正规式 r 表示一个语言 L(r)。
- 定义规则说明 L(r) 是怎样以各种方式从 r 的子正规式所表示的语言组合而成。
- 正规式用来表示简单的语言，叫做正规集。

正规式是用于说明词法单元如何对应到词法记号的模式。
与非形式化的描述相比，正规式更具形式化，更加精确。

正规式

- 正规式　定义的语言　　备注

ε	{ε}	
a	$\{a\}$	$a \in \Sigma$
$(r) \mid (s)$	$L(r) \cup L(s)$	r 和 s 是正规式
$(r)(s)$	$L(r)L(s)$	r 和 s 是正规式
$(r)^*$	$(L(r))^*$	r 是正规式
(r)	$L(r)$	r 是正规式

定义字母表Σ上正规式的规则

- 运算符的优先级：

* > 连接运算 > |

((a) $(b)^*$)| (c)可以写成ab^*| c

正规式

正规式的例子 $\Sigma = \{a, b\}$

$a \mid b$	$\{a, b\}$
$(a \mid b)(a \mid b)$	$\{aa, ab, ba, bb\}$
$aa \mid ab \mid ba \mid bb$	$\{aa, ab, ba, bb\}$
a^*	由字母 a 构成的所有串集
$(a \mid b)^*$	a 和 b 构成的所有串集

复杂的例子

$(\ 00 \ \mid \ 11 \mid \ ((01 \mid 10)(00 \mid 11)^* (01 \mid 10) \ ) \ )^*$

句子：0100110100001000001011100 1

正规定义

- 对正规式命名，使表示简洁。
  - $d_1 \rightarrow r_1$
  - $d_2 \rightarrow r_2$
  - ...
  - $d_n \rightarrow r_n$

保证：每个名字对应的正规式中使用的各种符号已经在前面定义了，从而可以避免递归定义的情况。

- 各个d_i的名字都不同
- 每个r_i都是$\Sigma \cup \{d_1, d_2, \ldots, d_{i-1}\}$上的正规式

Pascal里面的标识符模式

正规式表示

letter → A | B | … | Z | a | b | … | z

digit → 0 | 1 | … | 9

id → letter(letter|digit)*

怎么用语言来描述Pascal的标识符模式?

- Pascal标识符模式的自然语言描述:
  - 首字符必须是字母，由字母或数字组成的字符串。

C语言的标识符模式

- 模式的非形式描述

  首字符必须是_或者字母，由_、字母或数字组成的字符串。

- 请仿照Pascal标识符的例子，写出C语言的标识符的正规式表示

C语言的标识符模式

正规定义的例子

- C语言的标识符是字母、数字和下划线组成的串。

letter_ $\rightarrow A \mid B \mid \ldots \mid Z \mid a \mid b \mid \ldots \mid z \mid \_$

digit $\rightarrow 0 \mid 1 \mid \ldots \mid 9$

id $\rightarrow$ letter_(letter_ |digit)*

正规定义的例子

Pascal无符号数集合，例1946,11.28,63.6E8,1.99E–6

digit → 0 | 1 | … | 9

digits → digit digit*

optional_fraction → .digits|ε

optional_exponent → (E (+ | – | ε) d

num→digits optional_fraction optional_exponent

简化规则：

(1) $r^+ = rr^*$

(2) r?=r| ε

(3) [a-z]=a|b|c|…|z

简化表示

num → $digit^+$ (.$digit^+$)? (E(+|–)? $digit^+$)?

正规定义的例子

$\text{while} \rightarrow \text{while}$

$\text{do} \rightarrow \text{do}$

$\text{relop} \rightarrow < \mid <= \mid = \mid <> \mid > \mid >=$

$\text{id} \rightarrow \text{letter}\ (\text{letter} \mid \text{digit})^*$

$\text{num} \rightarrow \text{digit}^+\ (.\text{digit}^+)?\ (E\ (+ \mid -)?\ \text{digit}^+)?$

$\text{delim} \rightarrow \text{blank} \mid \text{tab} \mid \text{newline}$

$\text{ws} \rightarrow \text{delim}^+$

前面所提到的词法记号，
实际上就是正规式的名字！

词法记号的识别

词法记号的识别

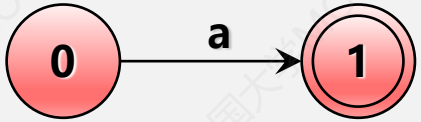
- 等同于对字符串的匹配过程

这个匹配过程可以基于状态转换图来完成

状态转换图==有限状态机

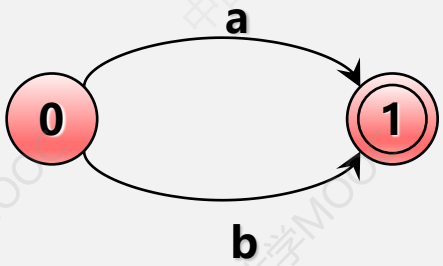
状态转换图

- 简单的正规式d->a



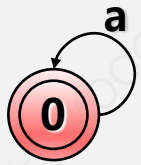
- 正规式d->ab

- 正规式d->a|b



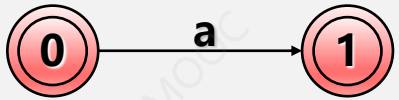
状态转换图

正规式d->a*

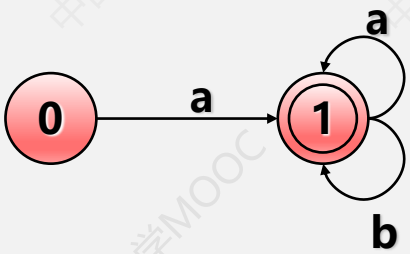


正规式d->a?

字符a出现一次或者0次



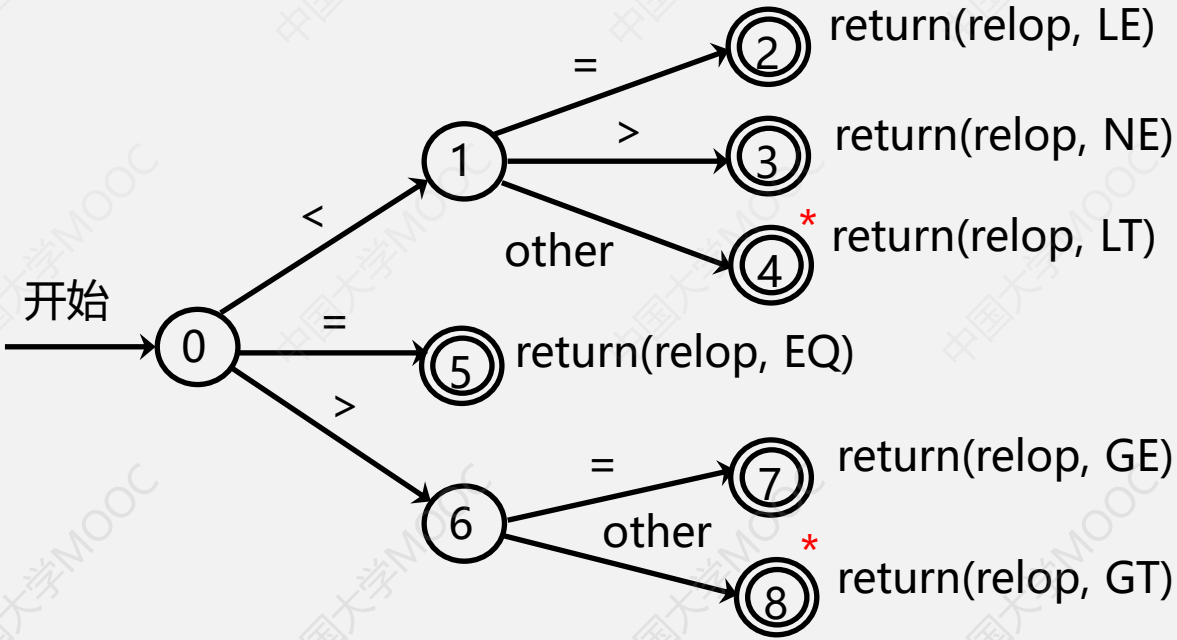
正规式d->a(a|b)*



状态转换图

关系算符的转换图

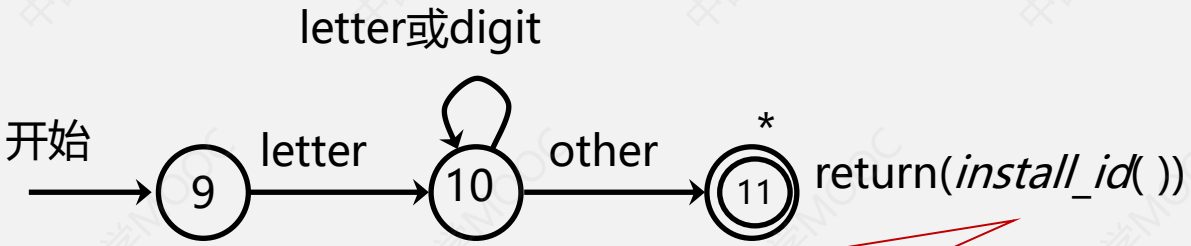
relop → < | < = | = | < > | > | > =



状态转换图

标识符和关键字的转换图

$$id \rightarrow letter\ (letter \mid digit)^*$$

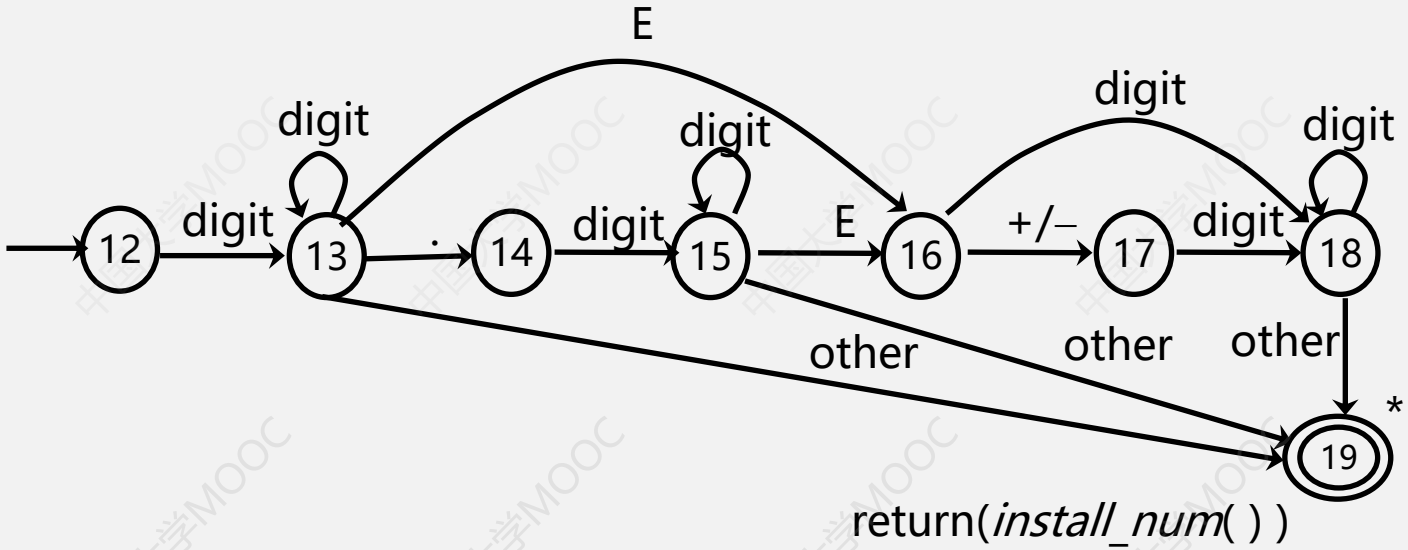


1、检查关键字表，如果在表中发现该词法单元则返回相应的记号并退出，否则转向2
2、该词法单元是标识符，在符号表中查找，若找到该词法单元则返回该条目的指针并退出，否则执行3
3、在符号表中建立一个新的条目，把该词法单元填入，并返回此新条目的指针

状态转换图

无符号数的转换图

num → digit⁺ (.digit⁺)? (E (+ | –)? digit⁺)?

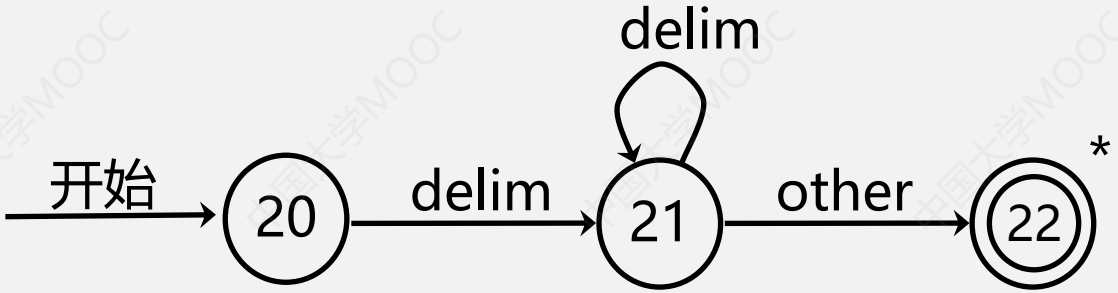


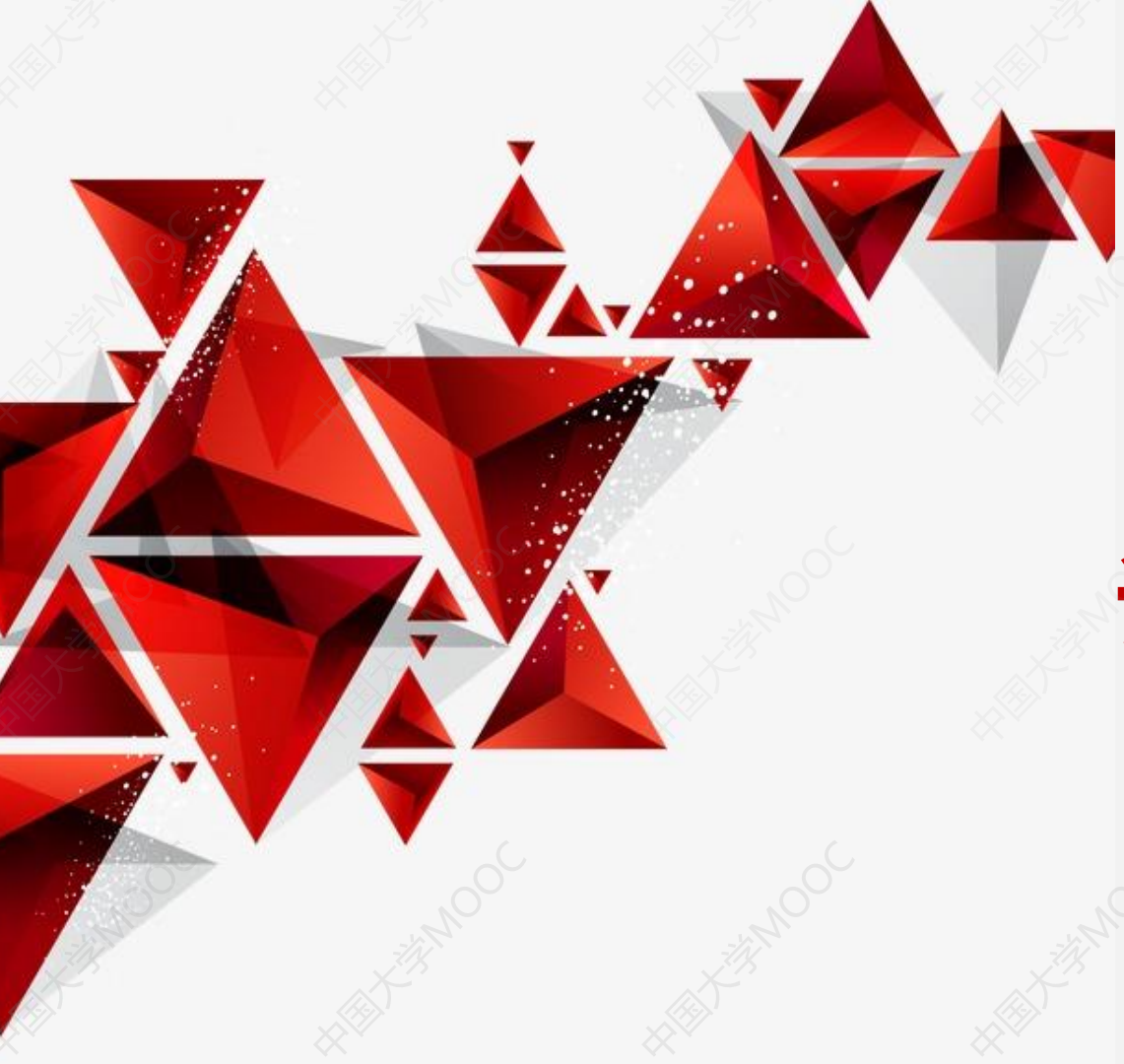
状态转换图

空白的转换图

delim → blank | tab | newline

ws → delim+





编译技术

词法分析

大连理工大学软件学院