

编译原理笔记

陈鸿峥

2020.07*

目录

1	简介	2
2	词法分析	2
2.1	基本定义	2
2.2	正则表达式	3
2.3	有限自动机	5
2.4	正则表达式、DFA与NFA的转换	5
2.5	最小化DFA	11
3	语法分析	12
3.1	上下文无关文法	12
3.2	基础概念与处理	14
3.3	自顶向下分析	16
3.4	自底向上分析	21
3.5	语法制导翻译	30
3.6	总结	32
4	语义分析与中间表示	32
5	运行时系统	36
5.1	存储管理	36
5.2	垃圾回收	36

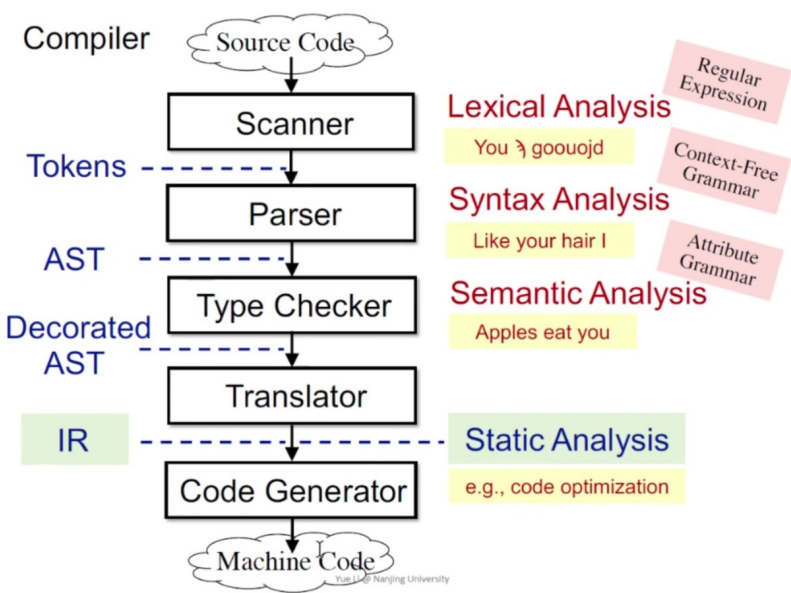
*Build 20200728

6	代码生成及优化	37
6.1	代码生成	37
6.2	代码优化	39

本课程采用书目Alfred V. Aho, Monica S. Lam, Ravi Sethi, Jeffrey D. Ullman, *Compilers: Principles, Techniques & Tools (2nd ed)*, 即大名鼎鼎的龙书。同时也参考了Stanford CS143: Compilers这门课程。

1 简介

编译器的几个阶段如下，前端包括词法(lexical)、语法(syntax)、语义(semantic)分析，中端IR生成、优化，后端代码生成。



编程语言设计的思想：

- 抽象(abstraction)：核心在于信息隐藏(infomation hiding)，只把必要的暴露出来
- 类型(types)：表达抽象、查出常见错误、使程序安全
- 重用(reuse)：开发软件系统中常见的模式（类型参数化、类与继承）

2 词法分析

分离词法分析和语法分析可以简化这两个任务，同时提升编译器的性能与兼容性。

2.1 基本定义

定义 1. 令牌(token)是一个令牌名字与可选属性值构成的对；模式(pattern)描述了每个词素(lexeme)要遵循什么规则；而词素（最小意义单位）则是源程序中一连串满足模式的字母，作为令牌的实例化。

例 1. 考虑C语句

```
printf("Total = %d\n", score);
```

其中`printf`和`score`是匹配(*match*)上令牌`id`模式的词素, 而`"Total = %d\n"`是匹配上字面值`literal`的词素。

简单来讲, 令牌是一个更大的概念, 是同类词素的集合。比如一个令牌`comparison`的样例词素可以有`<=`和`!=`。

定义 2 (字母表与语言). 字母表(*alphabet*) Σ 是有限符号(*symbol*)的集合, 如ASCII就是一个字母表。字符串(*string*) s 是从字母表中抽取的有限符号的序列, $|s|$ 为字符串长度, ϵ 为空串。语言(*language*)是字符串的可数集合。

例 2. 字母表 $\Sigma = \{0, 1\}$, 则 $\{001, 1001\}$ 和 $\{\}$ 都是定义在 Σ 上的语言。

定义 3 (字符串术语). 前缀(*prefix*)和后缀(*suffix*)都可以包括 ϵ 。字串(*substring*)可通过删除任意前缀和任意后缀(包括零个)获得。真(*proper*)字串则不包含 ϵ 。子序列(*subsequence*)是删除零个或多个不一定连续的字母得到的字符串。

语言是一种集合, 故集合运算也适用于语言。

并集(union)	$L \cup M$
连接(concatenation)/交集	LM
柯林闭包(Kleene closure)	$L^* = \cup_{i=0}^{\infty} L^i$
正闭包(positive)	$L^+ = \cup_{i=1}^{\infty} L^i$

2.2 正则表达式

定义 4 (正则表达式(regular expression, regex)). 正则表达式 r 定义了语言 $L(r)$, 以递归形式定义:

1. 奠基:

- ϵ 是正则表达式, 即 $L(\epsilon) = \{\epsilon\}$
- $a \in \Sigma$ 是正则表达式, 即 $L(a) = \{a\}$ (这里用斜体代表符号, 粗体代表符号对应的正则表达式)

2. 推论(*induction*): 若 r 和 s 都是正则表达式给出了语言 $L(r)$ 和 $L(s)$, 则

- $(r)|(s)$ 是正则表达式, 表示 $L(r) \cup L(s)$
- $(r)(s)$ 是正则表达式, 表示 $L(r)L(s)$
- $(r)^*$ 是正则表达式, 表示 $(L(r))^*$
- (r) 是正则表达式, 表示 $L(r)$

正则表达式表示的语言叫做正规集。如果两个正则 r 和 s 定义了相同的正则集, 则记作 $r = s$ 。

正则表达式的拓展¹:

¹更多可参见[Regex101](#)

- r^+ 代表一个或多个
- $r?$ 代表零或一个
- $[a-z]$ 字母类

有以下运算规定：

- 一元运算符 $*$ 有最高优先级，左结合（也包括 $+$ 、 $?$ 等扩展）
- 连接优先级次之，左结合
- $|$ 优先级最低，左结合

等价规则：

- 连接具有分配律： $r(st) = rs|rt$, $(s|t)r = sr|tr$
- ϵ 在闭包里被保证： $r^* = (r|\epsilon)^*$
- 闭包幂等(idempotent): $r^{**} = r^*$

定义 5 (正则定义). $d_i \rightarrow r_i$, 其中 d_i 都是名字, 且各不相同。每个 r_i 是 $\Sigma \cup \{d_1, \dots, d_{i-1}\}$ 中符号上的正则表达式。

例 3. 比如C语言的标识符可记为

$$\begin{aligned} \text{letter_} &\rightarrow A|B|\dots|Z|a|b|\dots|z|_ \\ \text{digit} &\rightarrow 0|1|\dots|9 \\ \text{id} &\rightarrow \text{letter_}(\text{letter_}|\text{digit})^* \end{aligned}$$

更简洁的写法

$$\begin{aligned} \text{letter_} &\rightarrow [A-Za-z] \\ \text{digit} &\rightarrow [0-9] \\ \text{id} &\rightarrow \text{letter_}(\text{letter_}|\text{digit})^* \end{aligned}$$

例 4. 下列正则表达式描述什么语言？

- $a(a|b)^*a$: 首尾是 a 中间任意个（可为0） a 或 b 的字符串
- $(a|b)^*a(a|b)(a|b)$: 倒数第三个字符为 a 仅含 a 或 b 的字符串
- $a^*ba^*ba^*ba^*$: 只含3个 b 且 a 在中间穿插（可没有）的字符串
- $((E|a)b^*)^*$: 空、全 a 全 b 、开头一个 a 紧接多个 b 的重复串
- $b^*(ab^*ab^*)^*$: 所有包含偶数个 a 的由 a 和 b 组成的字符串

注意考虑闭包为空的情况, ϵ 出现也可能导致空串!

例 5. 用正则表达式描述下列语言：

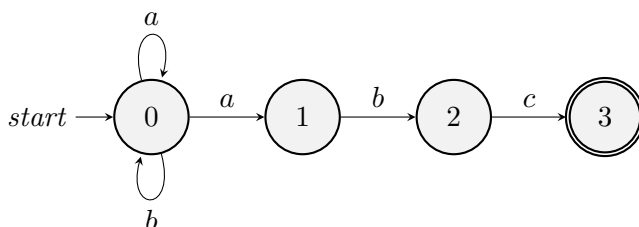
- 所有由按词典递增序排列的小写字母组成的字符串（如 add 、 low 都符合要求，而 zzg 则不符合）²:
 $a^+b^+c^+\dots z^+ \mid a^+b^+c^+\dots z^+ \mid a^+b^+c^+\dots z^+ \mid \dots$
- 不以 ab 开头的只含有字母 a 和 b 的字符串: $(ba|aa|bb)(a|b)^*$

²参见<https://www.zhihu.com/question/28714623/answer/41865697>

2.3 有限自动机

确定有限自动机(DFA)不可对 ϵ 进行移动, 而且对于每一状态 s , 输入符号 a , 只有唯一一条出边标记为 a ; 而非确定性有限自动机(NFA)可能有多种转换路径, 而且有 ϵ 移动。有限状态集 S , 状态 $s_0 \in S$ 为初始状态(start/initial), $F \subset S$ 为终止状态(accepting/final)。

例 6. 识别语言 $L((a|b)^*abb)$, 下面为一个NFA



判别字符串能否被DFA识别很简单, 只需要读入字符按照状态转移表跳转, 判断末态是不是终态即可 (即模拟)。

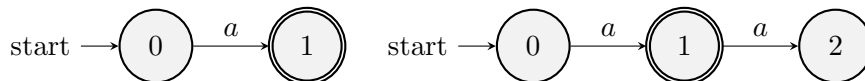
Algorithm 1 基于DFA的识别算法

```

1:  $s = s_0$ 
2:  $c = nextChar()$ 
3: while ( $c \neq eof$ ) do
4:    $s = move(s, c)$ 
5:    $c = nextChar()$ 
6: if  $s \in F$  then
7:   return "yes"
8: elsereturn "no"
  
```

时间复杂度为 $O(|str|)$ 。

对于DFA或NFA求反相当于将所有接受状态改为非接受状态, 非接受状态改为接受状态。注意可能出现DFA无对应符号出边的情况, 如 a , 这时可以添加冗余结点来接受这些非法输入。



定理 1. 对任一正则表达式 R , 一定存在另一正则表达式 R' , 使得 $L(R')$ 是 $L(R)$ 的补集。

分析. 由正则表达式与DFA的等价性, 对于正则表达式 R , 必然存在DFA M 可以识别 $L(R)$, 那么将 M 中的接受状态改为非接受状态, 将非接受状态改为接收状态, 得到新的DFA M' 可以识别 $L(R)$ 的补集, 进而存在 M' 对应的正则表达式 R' , 使得 $L(R')$ 是 $L(R)$ 的补集。

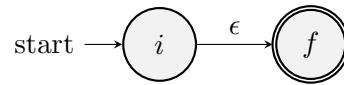
2.4 正则表达式、DFA与NFA的转换

2.4.1 正则表达式转NFA

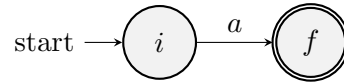
McNaughton-Yamada-Thompson算法

1. 奠基

- 对于表达式 ϵ ，构建NFA

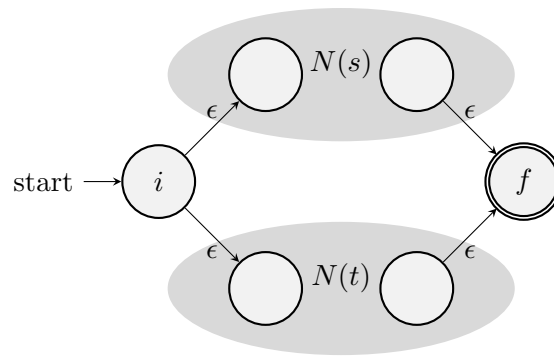


- 对于任意子表达式 $a \in \Sigma$ ，构建NFA

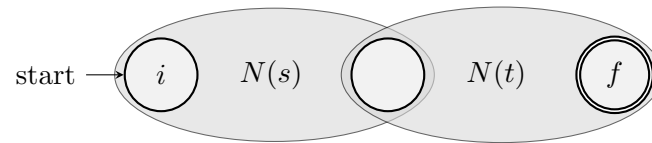


2. 推论

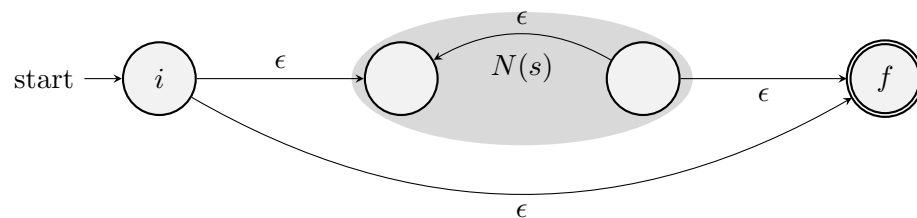
- $r = s|t$ ，取并集



- $r = st$ ，取连接



- $r = s^*$ ，Kleene闭包



其他拓展符号可通过上述基本符号得到，如

- R^+ 等价于 RR^*
- $R?$ 等价于 $\epsilon|R$

2.4.2 NFA转DFA

定义 6 (ϵ 闭包及 $move$). ϵ 闭包是可通过NFA的 ϵ 边转换的状态 (包括自己)。 $move(T, a)$ 为状态 $s \in T$ 通过输入符号 a 可到达的新的状态。

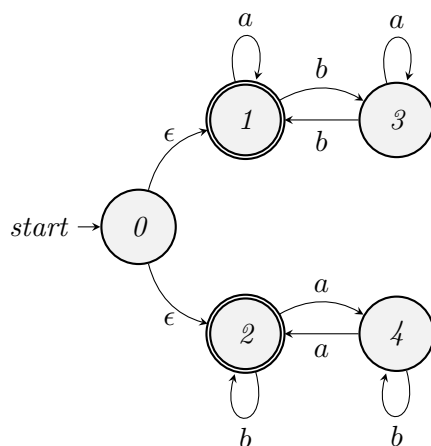
Algorithm 2 子集构造 (NFA转DFA)

Require: NFA N **Ensure:** DFA D (与 N 接受相同的语言)

- 1: ϵ -closure(s_0)是 $Dstates$ 的唯一状态, 且未被标记(unmarked)
 - 2: **while** 在 $Dstates$ 中还有未被标记的状态 T **do**
 - 3: 标记 T
 - 4: **for** 每一个输入符号 a **do**
 - 5: $U = \epsilon$ -closure(move(T, a))
 - 6: **if** $U \notin Dstates$ **then**
 - 7: 将 U 作为未标记的状态加入 $Dstates$
 - 8: $Dtran[T, a] = U$
-

思路即先求出初态的 ϵ 闭包, 然后对每个输入符号做转移后再求 ϵ 闭包, 看是否产生新的子集状态。注意这里的输入符号转移一定得转, 即不能留在原状态。

例 7. 考虑以下NFA:



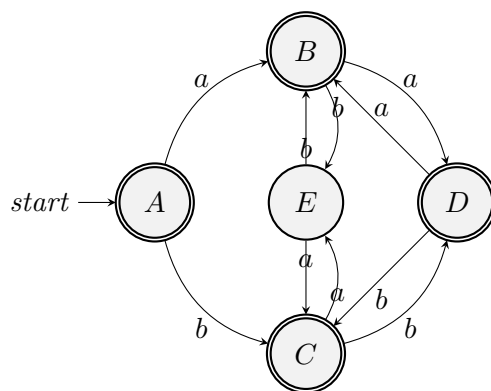
1. 这一NFA接受什么语言 (用自然语言描述)?

2. 构造接受同一语言的DFA.

分析. 1. 含有偶数个 a 或偶数个 b 的由 a 、 b 构成的字符串, 或者全是 a 或全是 b

2. 由subset construction算法构造如下

NFA	DFA	a	b
$\{0, \underline{1}, 2\}$	A	$\{1, 4\}$	$\{2, 3\}$
$\{\underline{1}, 4\}$	B	$\{1, 2\}$	$\{3, 4\}$
$\{2, 3\}$	C	$\{3, 4\}$	$\{1, 2\}$
$\{\underline{1}, 2\}$	D	$\{1, 4\}$	$\{2, 3\}$
$\{3, 4\}$	E	$\{2, 3\}$	$\{1, 4\}$



直接用NFA识别语言算法如下，需要每次算所有当前可能状态执行动作 c 后的 ϵ 闭包。

Algorithm 3 用NFA识别语言

```

1:  $S = \epsilon\text{-closure}(s_0)$ 
2:  $c = \text{nextChar}()$ 
3: while  $c \neq \text{eof}$  do
4:    $S = \epsilon\text{-closure}(\text{move}(S, c))$ 
5:    $c = \text{nextChar}()$ 
6: if  $S \cap F \neq \emptyset$  then
7:   return “yes”
8: else
9:   return “no”

```

定理 2. DFA ， NFA 和正则表达式三者的描述能力是一样的。

但从NFA转为DFA可能导致状态数的指数增长。

例 8. $L_n = (a \mid b)^* a (a \mid b)^{n-1}$ ，与此 NFA 等价的 DFA 状态数必不少于 2^n 。

分析. 反证法。假设存在一个 DFA D 接受语言 L_n ，且状态数少于 2^n 。构造 2^n 个长度为 n 的字符串

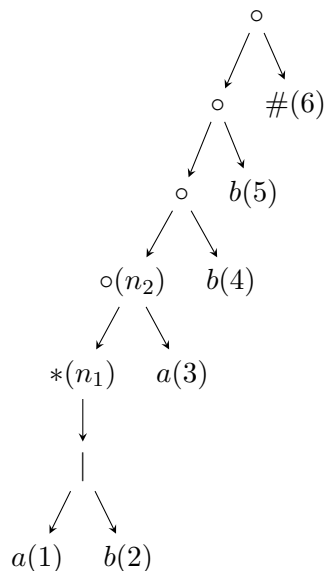
$aa \cdots a$
 $aa \cdots 1$
 \dots
 $bb \cdots a$
 $bb \cdots b$

由于 D 的状态数少于 2^n ，故上面必存在两个不同的字符串 s 和 t ，它们在 DFA 上会走到同一状态。因为 s 和 t 不等，因此总存在 i ，使得 $s[i] \neq t[i]$ 。不妨设 $s[i] = 0$ ， $t[i] = 1$ ，令 $s' = s + (n-1)$ 个 a ， $t' = t + (n-1)$ 个 a 。由 L_n 的表达式， s' 应该走到接受状态，而 t' 应该走到非接受状态。但由于 s 和 t 走到同一状态，那么它们再走 $(n-1)$ 个 a 也应该到达同一状态，但这个状态既是接受状态又是非接受状态，因此矛盾。

2.4.3 正则表达式转DFA

可以由正则表达式通过NFA转为DFA，本节则讲述直接由正则表达式转为DFA。

构造正则表达式的语法树，以#结尾。



*为star, |为or, o为cat

- $nullable(n)$: ϵ 包含在子树中则为真
- $firstpos(n)$: 符合regex子树的字符串中第一个字符可能出现的位置
- $lastpos(n)$: 符合要求字符串最后一个字符可能出现的位置
- $followpos(p)$: 紧跟 p 可能的位置

例 9. 结点 n_1 代表 $(a|b)^*$, 结点 n_2 代表 $(a|b)^*a$

- $nullable(n_1) = true$ $nullable(n_2) = false$
- $firstpos(n_2) = \{1, 2, 3\}$
- $lastpos(n_2) = \{3\}$
- $followpos(1) = \{1, 2, 3\}$

计算上述函数的方式:

结点 n	$nullable(n)$	$firstpos(n)$	$lastpos(n)$
标记为 ϵ 的叶子	true	\emptyset	\emptyset
位置为 i 的叶子	false	$\{i\}$	$\{i\}$
or结点 $n = c_1 c_2$	$nullable(c_1)$ or $nullable(c_2)$	$firstpos(c_1) \cup firstpos(c_2)$	$lastpos(c_1) \cup lastpos(c_2)$
cat结点 $n = c_1c_2$	$nullable(c_1)$ and $nullable(c_2)$	if ($nullable(c_1)$) $firstpos(c_1) \cup firstpos(c_2)$ else $firstpos(c_1)$	if ($nullable(c_2)$) $lastpos(c_1) \cup lastpos(c_2)$ else $lastpos(c_2)$
star结点 $n = c_1^*$	true	$firstpos(c_1)$	$followpos(c_1)$

- 若 n 为cat结点, 则对于左子树 c_1 的所有 $i \in lastpos(c_1)$, 有右子树 $firstpos(c_2) \in followpos(i)$
- 若 n 为star结点, 则对于所有 $i \in lastpos(n)$, 有 $firstpos(n) \in followpos(i)$

构建 $followpos$ 的过程实际上是深搜的过程, 可构造出一个有向图表示状态迁移。

Algorithm 4 Regex转DFA

Require: 正则表达式 r **Ensure:** DFA D 可识别 $L(r)$

- 1: 语法树 T 的根节点为 n_0 , $firstpos(n_0)$ 是 $Dstates$ 的唯一状态, 且未被标记(unmarked)
 - 2: **while** 在 $Dstates$ 中还有未被标记的状态 S **do**
 - 3: 标记 S
 - 4: **for** 每一个输入符号 a **do**
 - 5: $U = \bigcup_{\text{对应 } a \text{ 的位置 } p \in S} followpos(p)$
 - 6: **if** $U \notin Dstates$ **then**
 - 7: 将 U 作为未标记的状态加入 $Dstates$
 - 8: $Dtran[S, a] = U$
-

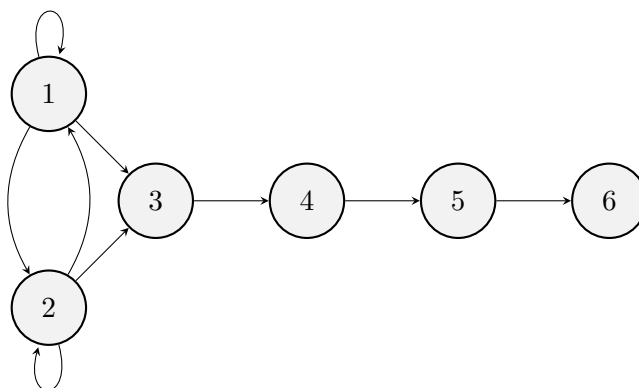
例 10. 考虑正则表达式 $(a \mid b)^* a b b \#$

$\begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix}$

分析. 由表达式其实可以直接得到 $followpos$ 函数

$position$	$followpos(i)$
1	$\{1, 2, 3\}$
2	$\{1, 2, 3\}$
3	$\{4\}$
4	$\{5\}$
5	$\{6\}$
6	\emptyset

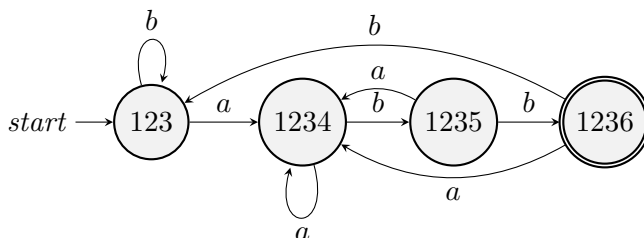
进而构造出一个有向图 (加上标号可变成 NFA)



$firstpos(n_0)$ 为 $\{1, 2, 3\}$, 由正则表达式转 DFA 的算法可得以下状态转移表。比如对于 $S = \{1, 2, 3\}$, a 的位置是 1 和 3, 那么取 $followpos$ 的并为 $\{1, 2, 3\} \cup \{4\}$ 。

	a	b
$\{1, 2, 3\}$	$\{1, 2, 3, 4\}$	$\{1, 2, 3\}$
$\{1, 2, 3, 4\}$	$\{1, 2, 3, 4\}$	$\{1, 2, 3, 5\}$
$\{1, 2, 3, 5\}$	$\{1, 2, 3, 4\}$	$\{1, 2, 3, 6\}$
$\{1, 2, 3, \underline{6}\}$	$\{1, 2, 3, 4\}$	$\{1, 2, 3\}$

然后可得DFA



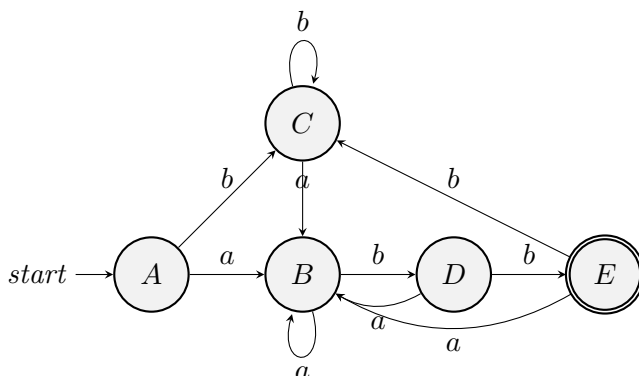
2.5 最小化DFA

定义 7 (区别(distinguish)). 字符串 w 区别状态 s 和 t , 如果DFA M 从状态 s 出发, 对输入串 w 进行状态转换, 最后停在某个接受状态; 从 t 出发, 对输入串 w 进行状态转换, 停在一个非接受状态; 反之亦然。

定理 3. 每一个正则集都可以唯一由一个状态数最少的DFA识别。

分析. 反证法。设算法得到的DFA为 D , 假设存在另一个DFA D' , D' 和 D 接受同一语言, 并且 D' 的状态数比 D 更少。设 D 的起始状态为 S , D' 的起始状态为 S' , 则 S 与 S' 不可区分。如果对于 S 和输入符号 a , 在 D 中迁移到状态 A ; 对于 S' 和输入符号 a , 在 D' 中迁移到状态 A' , 则 A 与 A' 不可区分。依此类推可知对于 D 中的任一状态 T , 在 D' 中都有一个状态 T' 与 T 不可区分。又由于 D 的状态数多于 D' 的状态数, 所以 D 中至少存在两个状态 T_1 和 T_2 , 使得 D' 中的一个状态 T 与它们均不可区分。因此 T_1 和 T_2 也不可区分, 于是矛盾。

例 11. 如下状态转移图



分析. 初始划分 Π 包括两个组(group): 接受状态组(E)和非接受状态组($ABCD$)³。构造 Π_{new} , 先考虑(E), 仅一个状态, 不可划分, 仍将(E)放回 Π_{new} 。然后考虑($ABCD$), 对于输入 a , 这些状态都转换到 B , 分

³准确来说是接受状态的补, 如果接受状态组为全集, 那么另一组将为空。另外需要考虑转移到外部结点的情况, 即DFA中无对应跳转符号。

组 $(ABCD)$ 不变；但对于输入 b ， A 、 B 和 C 都转换到状态组 $(ABCD)$ 的一个成员，而 D 转换到另一组成员 E 。因此，在 Π_{new} 中，状态组 $(ABCD)$ 需要分裂为两个新组 (ABC) 和 D ， $\Pi_{new} = (ABC)(D)(E)$ 。继续执行下一轮操作，最终得到 $\Pi_{final} = (AC)(B)(D)(E)$ 。因此选择 A 作为 (AC) 的代表，其他不变，可得到简化的自动机。

	a	b
A	B	A
B	B	D
D	B	E
E	B	A

3 语法分析

3.1 上下文无关文法

语法分析需要解决：从词法分析中获得的每个属性字(token)在语句中承担什么角色，同时检查语句是否符合程序语言的语法。

很多语言并非正则的，比如匹配的括号串 $\{(^i)^i \mid i \geq 0\}$ ，原因是FA不能记住其访问某一状态的次数，因此需要有更加强大的语言。

3.1.1 基本定义

定义 8 (上下文无关文法(context-free grammar, CFG)). 包括四部分

- 终端符号(*terminal*)的集合 T (即 $token$ 名字)
- 非终端符号的集合 N
- **唯一的**开始符号 $S \in N$
- 若干以下形式的产生式(*production*)

$$X \rightarrow Y_1 Y_2 \dots Y_n$$

其中 $X \in N$ 且 $Y_i \in T \cup N \cup \{\epsilon\}$ 。多个**左侧相同**的产生式右侧可用 $|$ 合并。

定义 9 (推导(derivation)). 从开始符号开始，每一步推导就是用一个产生式的右方取代左端的非终端符号。

CFG定义语言的能力比正则表达式强很大原因是它引入了**递归**的因素。

例 12. 用上下文无关文法定义下列语言：

- $L = \{0^n 1^n \mid n \geq 1\}$: $E \rightarrow 0E1 \mid 01$
- 只含有0和1的回文串: $S \rightarrow 0S0 \mid 1S1 \mid 0 \mid 1 \mid \epsilon$
- 只含有(和)的匹配括号串: $E \rightarrow (E) \mid EE \mid \epsilon$

- 最左推导(left-most): 每步推导都替换最左侧的非终端符号

$$E \xRightarrow{lm} -E \xRightarrow{lm} -(E) \xRightarrow{lm} -(E + E) \xRightarrow{lm} -(id + E) \xRightarrow{lm} -(id + id)$$

- 最右推导(right-most): 每步推导都替换最右侧的非终端符号

$$E \xRightarrow{rm} -E \xRightarrow{rm} -(E) \xRightarrow{rm} -(E + E) \xRightarrow{rm} -(E + id) \xRightarrow{rm} -(id + id)$$

事实上逆向的最右推导即自底向上分析法。

定义 10 (二义性). 如果对于一个文法, 存在一个句子, 对这个句子可以构造两棵不同的分析树, 那么我们称这个文法为二义的。

看语法分析树的叶子结点能不能连成句子。

例 13. 对于文法 $E \rightarrow E + E \mid E * E \mid -E \mid (E) \mid id$ 及句子 $id + id * id$, 有以下两种推导:

$ \begin{aligned} E &\Rightarrow E + E \\ &\Rightarrow id + E \\ &\Rightarrow id + E * E \\ &\Rightarrow id + id * E \\ &\Rightarrow id + id * id \end{aligned} $	$ \begin{aligned} E &\Rightarrow E * E \\ &\Rightarrow E + E * E \\ &\Rightarrow id + E * E \\ &\Rightarrow id + id * E \\ &\Rightarrow id + id * id \end{aligned} $
---	--

文法二义性可通过引入更多的产生式来消除。

例 14. $E \rightarrow E + E \mid E * E \mid (E) \mid id$ 是有二义的, 因为不知道应该先算加法还是乘法。可将其改为

$$\begin{aligned}
 E &\rightarrow E + T \mid T \\
 T &\rightarrow T * F \mid F \\
 F &\rightarrow (E) \mid id
 \end{aligned}$$

其中 E 为 *Expression*, T 为 *Term*, F 为 *Factor*, 即可消除二义性 (必然得先算乘法)。相当于先算 F , 再算 T , 最后算 E , 强行添加了括号/优先级。

例 15 (悬挂的if-else). `if E1 then if E2 then E3 else E4`, 可以令 `else` 匹配最近的 `then`。

$ \begin{aligned} E &\rightarrow MIF \\ &\mid UIF \\ MIF &\rightarrow \text{if } E \text{ then } MIF \text{ else } MIF \\ &\mid OTHER \\ UIF &\rightarrow \text{if } E \text{ then } E \\ &\mid \text{if } E \text{ then } MIF \text{ else } UIF \end{aligned} $	$ \begin{aligned} &// \text{所有的 } then \text{ 都被匹配} \\ &// \text{仅有一些 } then \end{aligned} $
---	---

并不是所有上下文无关文法都可以做到无二义, 也无法判断一个上下文无关文法是否是二义的。

3.1.2 NFA转CFG

1. 对于NFA的每一状态 i ，创建非终态 A_i
2. 若状态 i 在输入 a 上有转换边到状态 j ，则添加生成式 $A_i \rightarrow aA_j$ ；若状态 i 在输入 ϵ 上转换到状态 j ，则添加生成式 $A_i \rightarrow A_j$
3. 若 i 是接受状态，则添加 $A_i \rightarrow \epsilon$
4. 若 i 是初始状态，则令 A_i 为语法的初始符号

定义 11 (右线性文法). 如果每个产生式都属于下列形式之一

$$A \rightarrow aB \quad A \rightarrow a \quad A \rightarrow \epsilon$$

则这样的文法称为右线性文法

定义 12 (左线性文法). 如果每个产生式都属于下列形式之一

$$A \rightarrow Ba \quad A \rightarrow a \quad A \rightarrow \epsilon$$

则这样的文法称为左线性文法

定理 4. 正则表达式/NFA/DFA与左/右线性文法得表达能力是等价的

在处理程序时，上下文无法文法存在局限性，无法解决诸如以下问题：

- 变量先声明，再使用
- 调用函数时，实参个数和形参个数一致

都得留到语义分析阶段才解决。

3.2 基础概念与处理

对于CFG的预处理包括消除左递归和提取左因子。

3.2.1 消除左递归

定义 13 (左递归). 对于非终端符号 A 有生成式 $A \rightarrow A\alpha$ ，则该文法是左递归的。

消除左递归的方法：先把单元素拎出来放左侧，然后把所有递归移至右侧

$$A \rightarrow A\alpha_1 \mid \cdots \mid A\alpha_m \mid \beta_1 \mid \cdots \mid \beta_n \implies \begin{aligned} A &\rightarrow \beta_1 A' \mid \cdots \mid \beta_n A' \\ A' &\rightarrow \alpha_1 A' \mid \cdots \mid \alpha_m A' \mid \epsilon \end{aligned}$$

例 16. 如果是多级左递归，则需要先将上级生成式代入到中间级生成式中，再做消除。如下例

$$\begin{aligned} S &\rightarrow Aa \mid b \\ A &\rightarrow Ac \mid Sd \mid \epsilon \end{aligned}$$

将下式改写为 $A \rightarrow Ac \mid Aad \mid bd \mid \epsilon$ ，进而可消除左递归

$$\begin{aligned} S &\rightarrow Aa \mid b \\ A &\rightarrow bdA' \mid A' \\ A' &\rightarrow cA' \mid adA' \mid \epsilon \end{aligned}$$

3.2.2 提取左因子

定义 14 (提取左因子(left-factoring)). 将生成式右侧左部相同的因子部分提取出来，找最长前缀

$$A \rightarrow \alpha\beta_1 \mid \alpha\beta_2 \mid \cdots \alpha\beta_n \mid \gamma$$

γ 代表所有不以 α 开始的生成式，提取左因子则得到（将 α 拿出来）

$$\begin{aligned} A &\rightarrow \alpha A' \mid \gamma \\ A' &\rightarrow \beta_1 \mid \beta_2 \mid \cdots \beta_n \end{aligned}$$

直至没有生成式有相同前缀

3.2.3 FIRST集与FOLLOW集

定义 15 (FIRST集与FOLLOW集). $FIRST(\alpha)$ 集为从 α 中推导出来的字符串第一个终端符号的集合，若 $\alpha \rightarrow \epsilon$ ，则 $\epsilon \in FIRST(\alpha)$ ；若 $A \rightarrow c\gamma$ ，则 $c \in FIRST(A)$ 。 $FOLLOW(A)$ 集为可以出现在 A 右侧的终端符号的集合。若 A 是最右端的符号，则字符串结束符号 $\$ \in FOLLOW(A)$ 。

算法 1. 计算 $FIRST(X)$ 集

1. 如果 X 是终端符号，则 $FIRST(X) = \{X\}$
2. 如果 X 是非终端符号，且 $X \rightarrow Y_1Y_2 \cdots Y_k$ 。
 - 若 $Y_1 \cdots Y_{i-1} \rightarrow \epsilon$ ，则将 $a \in FIRST(Y_i)$ 放入 $FIRST(X)$ 。
 - 若 $\epsilon \in FIRST(Y_j)$, $j = 1, 2, \dots, k$ ，则将 ϵ 放入 $FIRST(X)$ 中。
3. 若 $X \rightarrow \epsilon$ 是生成式，将 ϵ 放入 $FIRST(X)$ 中

算法 2. 计算 $FOLLOW(A)$ 集

1. 将 $\$$ 放入 $FOLLOW(S)$ ，其中 S 是开始符号
2. 如果有生成式 $A \rightarrow \alpha B\beta$ ，那么 $\forall a \in FIRST(\beta), a \neq \epsilon : a \in FOLLOW(B)$
3. 如果有生成式 $A \rightarrow \alpha B$ ，或生成式 $A \rightarrow \alpha B\beta$ 且 $\epsilon \in FIRST(\beta)$ ，则 $\forall a \in FOLLOW(A) : a \in FOLLOW(B)$
(注意这里的因果关系)

简而言之， $FOLLOW$ 集看下一符号的 $FIRST$ ，如果 ϵ 在下一符号的 $FIRST$ 集中，则看生成式左端的 $FOLLOW$ 集。通常下面的生成式的 $FOLLOW$ 集都会包含上面生成式的 $FOLLOW$ 集。

另一种方式：

1. $\$ \in FOLLOW(S)$
2. $\forall A \rightarrow \alpha X \beta : FIRST(\beta) - \{\epsilon\} \subset FOLLOW(X)$
3. $\forall A \rightarrow \alpha X \beta, \epsilon \in FIRST(\beta) : FOLLOW(A) \subset FOLLOW(X)$

- Recall the grammar

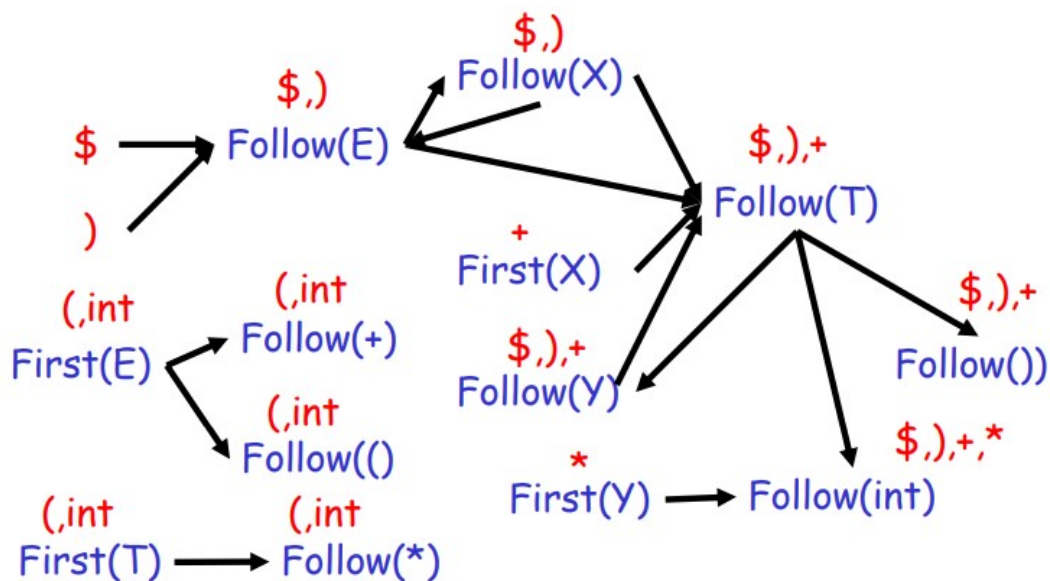
$E \rightarrow T X$

$T \rightarrow (E) \mid \text{int } Y$

$X \rightarrow + E \mid \epsilon$

$Y \rightarrow * T \mid \epsilon$

- $\$ \in \text{Follow}(E)$
- $\text{First}(X) \subseteq \text{Follow}(T)$
- $\text{Follow}(E) \subseteq \text{Follow}(X)$
- $\text{Follow}(E) \subseteq \text{Follow}(T)$
- $) \in \text{Follow}(E)$
- $\text{Follow}(T) \subseteq \text{Follow}(Y)$
- $\text{Follow}(X) \subseteq \text{Follow}(E)$
- $\text{Follow}(Y) \subseteq \text{Follow}(T)$



29

3.3 自顶向下分析

3.3.1 递归下降

递归下降语法翻译即从顶层的非终端符号 E 开始，顺序尝试 E 的所有规则，不断回溯遍历，完整例子可见 [此文档](#)。

Algorithm 5 递归下降(top-down parser)

```
1: 选择 $A$ 生成式 $A \rightarrow X_1X_2 \cdots X_k$ 
2: for  $i = 1$  to  $k$  do
3:   if  $X_i$ 是非终端符号 then
4:     调用 $X_i()$ 
5:   else
6:     if  $X_i$ 等于当前的输入符号 $a$  then
7:       读取下一输入符号
8:     else
9:        $error()$ 
```

3.3.2 LL(1)文法

定义 16 (LL(1)文法⁴). 语法 G 是LL(1)文法当且仅当对于任意 $A \rightarrow \alpha \mid \beta$ 为 G 两个不同的生成式, 满足

1. α 和 β 不会同时推导出由同一终端符号 a 开始的字符串
2. α 和 β 中至多一个能获得空字符串
3. 若 $\beta \rightarrow \epsilon$, 则 α 不能推出任何以 $FOLLOW(A)$ 中终端符号开始的字符串; 同样地, 若 $\alpha \rightarrow \epsilon$, 则 β 不能推出任何以 $FOLLOW(A)$ 中终端符号开始的字符串

前两个条件等价于 $FIRST(\alpha) \cap FIRST(\beta) = \emptyset$, 第三个条件等价于若 $\epsilon \in FIRST(\beta)$, 则 $FIRST(\alpha) \cap FOLLOW(A) = \emptyset$, 反之同理。这三个条件的作用都是避免在输入符号上冲突。

通常没有左递归、无歧义的语法可以是LL(1)。

递归下降在每一步都会有多种生成式的选择, 这会导致大量的回溯。而在LL(1)文法中, 每一步都只有一种生成式的选择, 避免了回溯。

例 17. 经典的二义if-else语法 (已提取左因子)

$$\begin{aligned} S &\rightarrow iEtSS' \mid a \\ S' &\rightarrow eS \mid \epsilon \\ E &\rightarrow b \end{aligned}$$

可以求得

$$\begin{aligned} FIRST(S) &= \{i, a\} \\ FIRST(S') &= \{\epsilon, e\} \\ FIRST(E) &= b \\ FOLLOW(S) &= \{\$ \} + FIRST(S') - \{\epsilon\} = \{\$, e\} \\ FOLLOW(S') &= \{\$ \} + FOLLOW(S) = \{\$, e\} \\ FOLLOW(E) &= \{\$, t\} \end{aligned}$$

⁴第一个L代表输入字符串从左边开始扫描, 第二个L代表得到的推导是最左推导, (1)代表向前看1个输入符号 (或单词)

因为 $\epsilon \in FIRST(S' \rightarrow \epsilon)$ 里, 而 $FIRST(S' \rightarrow eS) \cap FOLLOW(S') = \{e\} \neq \emptyset$, 所以不是 $LL(1)$ 文法。

算法 3 (构造 $LL(1)$ 预测语法表). 对于每一生成式 $A \rightarrow \alpha$,

1. 对于每一终端符号 $a \in FIRST(\alpha)$, 将 $A \rightarrow \alpha$ 添加到 $M[A, a]$ 中。
2. 若 $\epsilon \in FIRST(\alpha)$, 则对 $b \in FOLLOW(A)$, 将 $A \rightarrow \alpha$ 添加到 $M[A, b]$ 中 (β 可以为 $\$$)。

注意是生成式右端。

例 18. 考虑以下文法

$$\begin{aligned} E &\rightarrow T + E \mid T \\ T &\rightarrow int \mid int * T \mid (E) \end{aligned}$$

提取左因子后即得

$$\begin{aligned} E &\rightarrow TX \\ X &\rightarrow +E \mid \epsilon \\ T &\rightarrow int Y \mid (E) \\ Y &\rightarrow *T \mid \epsilon \end{aligned}$$

$FIRST$ 集和 $FOLLOW$ 集的推导可见上面的图, 注意 $FOLLOW(T)$ 和 $FOLLOW(Y)$ 很容易推漏

$$\begin{aligned} FIRST(Y) &= \{*, \epsilon\} & FOLLOW(E) &= \{), \$\} \\ FIRST(T) &= \{int, (\} & FOLLOW(X) &= \{), \$\} \\ FIRST(X) &= \{+, \epsilon\} & FOLLOW(T) &= \{+,), \$\} \\ FIRST(E) &= \{int, (\} & FOLLOW(Y) &= \{+,), \$\} \end{aligned}$$

有 $LL(1)$ 语法表, 其中最左列为最左非终端符号, 最上行为下一输入符号, 表格内容为使用的右端生成式。

	int	$*$	$+$	$($	$)$	$\$$
E	TX			TX		
X			$+E$		ϵ	ϵ
T	$int Y$			(E)		
Y		$*T$	ϵ		ϵ	ϵ

基于表的预测语法分析, 用栈实现。

Algorithm 6 Table-Driven Predictive Parsing

```
1: ip=0
2: X=stack.top()
3: while X ≠ $ do
4:   if X == w[ip] then
5:     stack.pop(); ip++;
6:   else
7:     if X is a terminal or M[X,a]=∅ then
8:       Error()
9:     else
10:      Output production M[X,a]=  $X \rightarrow Y_1Y_2 \cdots Y_k$ 
11:      stack.pop()
12:      push  $Y_k, Y_{k-1}, \dots, Y_1$  onto the stack
13:   X=stack.top()
14: if w[ip] != '$' then
15:   Error()
```

例 19. 考虑下面语法

- (1) $E \rightarrow TE'$
- (2) $E' \rightarrow +TE' \mid \epsilon$
- (3) $T \rightarrow FT'$
- (4) $T' \rightarrow *FT' \mid \epsilon$
- (5) $F \rightarrow (E) \mid id$

计算FIRST集

- 从终端符号多的开始(5), $FIRST(F) = \{ (, id \}$
- 向上找含F的生成式(3), $FIRST(T) = FIRST(F)$
- 向上找含T的生成式(1), $FIRST(E) = FIRST(T)$
- $FIRST(E') = \{ +, \epsilon \}$
- $FIRST(T') = \{ *, \epsilon \}$

计算FOLLOW集

- 从起始符号开始(1), 注意(5)也有E, 故 $FOLLOW(E) = \{), \$ \}$
- E' 只出现在E的生成式末尾, 因此 $FOLLOW(E') = FOLLOW(E) = \{), \$ \}$
- $FOLLOW(T) \subset FIRST(E') = \{ +, \epsilon \}$, 由于 $\epsilon \in FIRST(E')$, 故 $FOLLOW(T) \subset FOLLOW(E) = \{), \$ \}$, 即 $FOLLOW(T) = \{ +,), \$ \}$
- $FOLLOW(F) \subset FIRST(T') = \{ *, \epsilon \}$, 由于 $\epsilon \in FIRST(T')$, 故 $FOLLOW(F) \subset FOLLOW(T) = \{ +,), \$ \}$, 即 $FOLLOW(F) = \{ *, +,), \$ \}$

- T' 只出现在 T 的生成式末尾，因此 $FOLLOW(T') = FOLLOW(T) = \{*, +,), \$\}$

对应语法预测表

	id	$+$	$*$	$($	$)$	$\$$
E	$E \rightarrow TE'$			$E \rightarrow TE'$		
E'		$E' \rightarrow +TE'$			$E' \rightarrow \epsilon$	$E' \rightarrow \epsilon$
T	$T \rightarrow FT'$			$T \rightarrow FT'$		
T'		$T' \rightarrow \epsilon$	$T' \rightarrow *FT'$		$T' \rightarrow \epsilon$	$T' \rightarrow \epsilon$
F	$F \rightarrow id$			$F \rightarrow (E)$		

注意上述标红的项为通过第2条规则推导出来的项 ($\epsilon \in FIRST(\alpha)$)。

例 20. 考虑以下文法：

$$S \rightarrow (L) \mid a$$

$$L \rightarrow L, S \mid S$$

1. 消除文法的左递归.
2. 构造文法的 $LL(1)$ 分析表.
3. 对于句子 $(a, (a, a))$ ，给出语法分析的详细过程（参照课本228页的图4.21）.

分析. 1. 如下

$$S \rightarrow (L) \mid a$$

$$L \rightarrow SL'$$

$$L' \rightarrow , SL' \mid \epsilon$$

2. 先求出 $FIRST$ 集和 $FOLLOW$ 集（由于文法中存在逗号，故将字符用引号括起来以示区分）

$$FIRST(S) = \{(' , a')\} \quad FOLLOW(S) = \{\$, ', '\}$$

$$FIRST(L) = \{(' , a')\} \quad FOLLOW(L) = \{')'\}$$

$$FIRST(L') = \{', \epsilon\} \quad FOLLOW(L') = \{')'\}$$

$LL(1)$ 分析表如下，其中第一列为非终端符号，第一行为输入符号.

	$($	$)$	a	$,$	$\$$
S	$S \rightarrow (L)$		$S \rightarrow a$		
L	$L \rightarrow SL'$		$L \rightarrow SL'$		
L'		$L' \rightarrow \epsilon$		$L' \rightarrow , SL'$	

3. 语法分析过程如下

<i>Matched</i>	<i>Stack</i>	<i>Input</i>	<i>Action</i>
	$S\$$	$(a, (a, a))\$$	
	$(L)\$$	$(a, (a, a))\$$	<i>output</i> $S \rightarrow (L)$
$($	$L)\$$	$a, (a, a))\$$	
$($	$SL')\$$	$a, (a, a))\$$	<i>output</i> $L \rightarrow SL'$
$($	$aL')\$$	$a, (a, a))\$$	<i>output</i> $S \rightarrow a$
$(a$	$L')\$$	$, (a, a))\$$	
$(a$	$, SL')\$$	$, (a, a))\$$	<i>output</i> $L' \rightarrow, SL'$
$(a,$	$SL')\$$	$(a, a))\$$	
$(a,$	$(L)L')\$$	$(a, a))\$$	<i>output</i> $S \rightarrow (L)$
$(a, ($	$L)L')\$$	$a, a))\$$	
$(a, ($	$SL')L')\$$	$a, a))\$$	<i>output</i> $L \rightarrow SL'$
$(a, ($	$aL')L')\$$	$a, a))\$$	<i>output</i> $S \rightarrow a$
$(a, (a$	$L')L')\$$	$, a))\$$	
$(a, (a$	$, SL')L')\$$	$, a))\$$	<i>output</i> $L' \rightarrow, SL'$
$(a, (a,$	$SL')L')\$$	$a))\$$	
$(a, (a,$	$aL')L')\$$	$a))\$$	<i>output</i> $S \rightarrow a$
$(a, (a, a$	$L')L')\$$	$))\$$	
$(a, (a, a$	$)L')\$$	$))\$$	<i>output</i> $L' \rightarrow \epsilon$
$(a, (a, a)$	$L')\$$	$)\$$	
$(a, (a, a)$	$)\$$	$)\$$	<i>output</i> $L' \rightarrow \epsilon$
$(a, (a, a))$	$\$$	$\$$	

3.4 自底向上分析

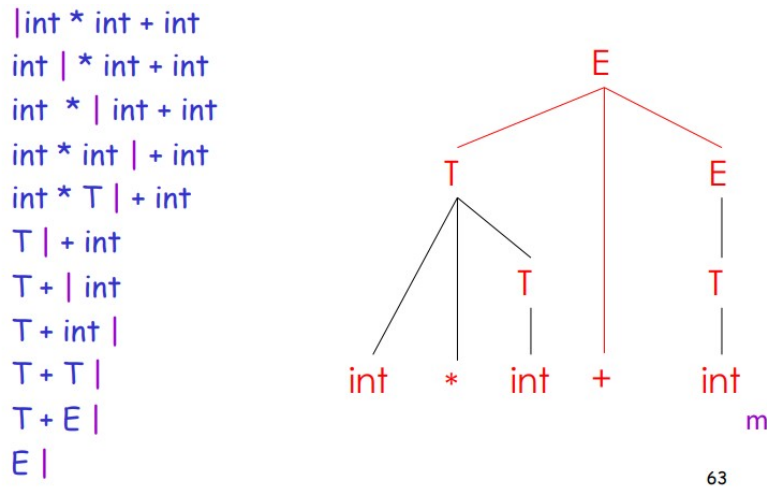
自底向上的语法分析采用两种动作：

- 移进(shift)：将 $|$ 向右移动一格

$$ABC \mid xyz \Longrightarrow ABCx \mid yz$$

- 规约(reduce)：在字符串右侧逆向应用生成式

$$C'bx y \mid ijk \Longrightarrow CbA \mid ijk$$



移进将终端符号移入栈中，规约将生成式的右端符号弹出，将生成式的左端非终端符号推入。

定义 17 (句柄(handle)). *A handle is a string that can be **reduced** and also allows further reductions back to the start symbol.* 可以理解为当前正在处理的token，用于**规约**（生成式的右端）而不是移进。

定义 18 (活前缀(viable prefix)). α 是活前缀若存在 ω 使得 $\alpha \mid \omega$ 是移进-规约语法分析器的状态。

例 21. 考虑以下CFG⁵

$$\begin{aligned}
 S &\rightarrow XX \\
 X &\rightarrow aX \mid b
 \end{aligned}$$

对aabb做最右推导有

a	abb
aa	bb
aab	b
aa X	b
aX	b
X	b
Xb	
XX	
S	

上面加粗被代替的字符串称为句柄，而句柄左侧（含句柄本身）的子字符串称为活前缀，包括

$$\{a, aa, aab\} \cup \{a, aa, aaX\} \cup \{a, aX\} \cup \{X, Xb\} \cup \{X, XX\}$$

LR分析是最通用的**非回溯**移进-规约语法解析方法，难点在于构建分析表太过麻烦，但Yacc等工具可辅助构建。

⁵例子来源于<https://gatecse.in/viable-prefixes-and-handle-in-lr-parsing/>

3.4.1 LR(0)语法

为构建规范(canoical)LR(0)项, 定义增量语法 G' 有生成式 $S' \rightarrow S$, 其中 S 为原语法 G 的开始符号。这个生成式用于告知parser接受(accept)输入并停止解析, 即接受仅发生在要对 $S' \rightarrow S$ 进行规约的时候。

定义 19 (CLOSURE). 若 I 为语法 G 项的集合, 则 $CLOSURE(I)$ 为从 I 中构造出项的集合:

1. 初始时, 在 I 中的每一项都会被加到 $CLOSURE(I)$ 中
2. 若 $A \rightarrow \alpha \cdot B\beta$ 在 $CLOSURE(I)$ 中且 $B \rightarrow \beta$ 是一个生成式, 则将项 $B \rightarrow \cdot\gamma$ 加入到 $CLOSURE(I)$ 中; 重复使用此规则, 直至没有新项可以被加入

定义 20 (GOTO). I 是项的集合, X 为输入符号, $GOTO(I, X)$ 为所有项 $[A \rightarrow \alpha X \cdot \beta]$ 的闭包使得 $[A \rightarrow \alpha \cdot X\beta]$ 在 I 中

定义 21 (核项(kernel item)). 初始项 $S' \rightarrow \cdot S$ 及所有 \cdot 不在左端的项称为核项, 除了初始项外所有 \cdot 在左端的项称为非核项 (即那些新加入闭包的项)

定义 22 (LR(0)语法). 栈包含 α , 下一输入是 t , DFA在输入 α 上终止在状态 s

- 当 s 包含 $X \rightarrow \beta$ 的项时进行规约 (没有得移进就规约, 自动机无对应符号出边)
- 当 s 包含 $X \rightarrow \beta \cdot t\omega$ 的项时移进

LR(0)可能存在以下两种冲突:

- 规约-规约冲突: $X \rightarrow \beta$ 且 $Y \rightarrow \omega$.
- 移进-规约冲突: $X \rightarrow \beta$ 且 $Y \rightarrow \omega \cdot t\delta$

注意接受-移进不是冲突! LR(0)可以直接处理左递归。

3.4.2 SLR分析

SLR(simple left-to-right scan)⁶用启发式算法提升了LR(0)移进规约的效率, 减少冲突。

定义 23 (SLR(1)语法). 栈包含 α , 下一输入是 t , DFA在输入 α 上停在状态 s

- 当 s 包含 $X \rightarrow \beta$ 的项且 $t \in FOLLOW(X)$ 时进行 $X \rightarrow \beta$ 规约
- 当 s 包含 $X \rightarrow \beta \cdot t\omega$ 的项时移进

算法 4 (SLR(1)分析表). 构造 $C = \{I_0, I_1, \dots, I_n\}$ 为LR(0)项的集合 G'

1. 若 $[A \rightarrow \alpha \cdot a\beta] \in I_i$ 且 $GOTO(I_i, a) = I_j$, 则设 $ACTION[i, a]$ 为移进 j ,
2. 若 $[A \rightarrow \alpha \cdot] \in I_i$, 则 $\forall a \in FOLLOW(A)$, 设 $ACTION[i, a]$ 为规约 $A \rightarrow \alpha$
3. 若 $[S' \rightarrow S \cdot] \in I_i$, 则设 $ACTION[i, \$]$ 为接受(ACC)

若上述有冲突的动作, 则该文法不是SLR(1)的。

依照分析表可以得到语法分析的算法

- 若 $ACTION[s, a]$ 为移进 t , 则将 t 推入栈中

⁶或SLR(1)分析, 通常省略(1)。用的是LR(0)项, 但是在语法分析时才前看1个输入符号。

- 若 $ACTION[s, a]$ 为规约 $A \rightarrow \beta$ ，则将 $|\beta|$ 个符号从栈顶弹出，令 t 为栈顶符号，将 $GOTO[t, A]$ 推入栈中，输出规约 $A \rightarrow \beta$
- 若 $ACTION[s, a] = acc$ ，则语法解析结束

例 22. 考虑以下文法：

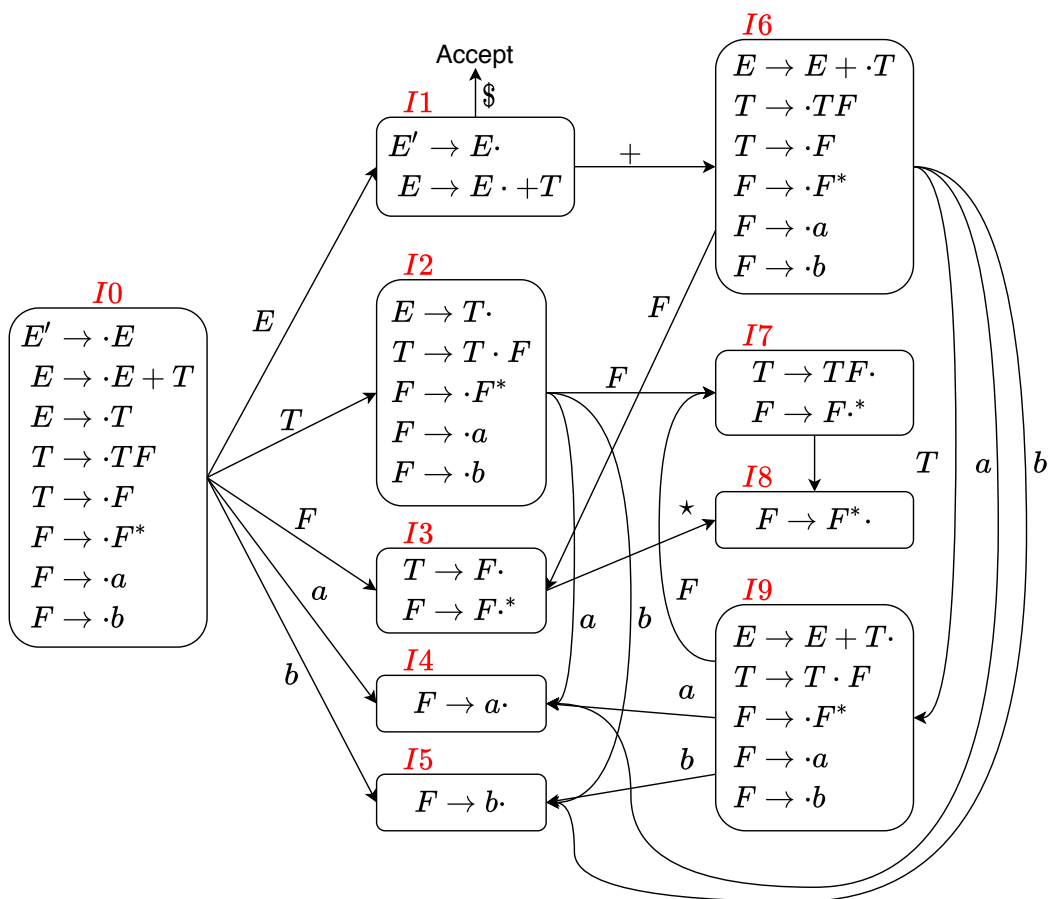
- (1) $E \rightarrow E + T$
- (2) $E \rightarrow T$
- (3) $T \rightarrow TF$
- (4) $T \rightarrow F$
- (5) $F \rightarrow F^*$
- (6) $F \rightarrow a$
- (7) $F \rightarrow b$

1. 写出每个非终端符号的 $FIRST$ 集和 $FOLLOW$ 集.
2. 构造识别这一文法所有活前缀(*viable prefixes*)的 $LR(0)$ 自动机 (参照课本4.6.2节图4.31) .
3. 构造这一文法的 SLR 分析表 (参照课本4.6.3节图4.37) .
4. 给出 SLR 分析器识别输入串 $a + ab^*$ 的过程 (参照课本4.6.4节图4.38)

分析. 1. $FIRST$ 集和 $FOLLOW$ 集如下

$$\begin{aligned} FIRST(E) &= \{a, b\} & FOLLOW(E) &= \{\$, +\} \\ FIRST(T) &= \{a, b\} & FOLLOW(T) &= \{\$, +, a, b\} \\ FIRST(F) &= \{a, b\} & FOLLOW(F) &= \{\$, +, *, a, b\} \end{aligned}$$

2. 构造增广语法 $E' \rightarrow E$ ，并得到 $LR(0)$ 自动机如下



3. 依据上述两问结果，可构造SLR分析表如下（s后面的数字为DFA状态编号，r后面的数字为生成式的编号）

STATE	ACTION					GOTO		
	a	b	+	*	\$	E	T	F
0	s4	s5				1	2	3
1			s6		ACC			
2	s4	s5	r2		r2			7
3	r4	r4	r4	s8	r4			
4	r6	r6	r6	r6	r6			
5	r7	r7	r7	r7	r7			
6	s4	s5					9	3
7	r3	r3	r3	s8	r3			
8	r5	r5	r5	r5	r5			
9	s4	s5	r1		r1			7

4. 依上述ACTION-GOTO表，可得以下过程（注意弹出后要跳转，红色标注的为栈顶弹出多项的情况）

	STACK	SYMBOLS	INPUT	ACTION
(1)	0		$a + ab*\$$	$[0, a]s4$
(2)	04	a	$+ab*\$$	$[4, a]r6 F \rightarrow a, [0, F]s3$
(3)	03	F	$+ab*\$$	$[3, +]r4 T \rightarrow F, [0, T]s2$
(4)	02	T	$+ab*\$$	$[2, +]r2 E \rightarrow T, [0, E]s1$
(5)	01	E	$+ab*\$$	$[1, +]s6$
(6)	016	$E+$	$ab*\$$	$[6, a]s4$
(7)	0164	$E + a$	$b*\$$	$[4, b]r6 F \rightarrow a, [6, F]s3$
(8)	0163	$E + F$	$b*\$$	$[3, b]r4 T \rightarrow F, [6, T]s9$
(9)	0169	$E + T$	$b*\$$	$[9, b]s5$
(10)	01695	$E + Tb$	$*\$$	$[5, *]r7 F \rightarrow b, [9, F]s7$
(11)	01697	$E + TF$	$*\$$	$[7, *]s8$
(12)	016978	$E + TF^*$	$\$$	$[8, \$]r5 F \rightarrow F^*, [9, F]s7$
(13)	01697	$E + TF$	$\$$	$[7, \$]r3 T \rightarrow TF, [6, T]s9$
(14)	0169	$E + T$	$\$$	$[9, \$]r1 E \rightarrow E + T, [0, E]s1$
(15)	01	E	$\$$	$[1, \$]ACC$

例 23. 下面的文法并非 $SLR(1)$

$$S \rightarrow L = R \mid R$$

$$L \rightarrow *R \mid id$$

$$R \rightarrow L$$

对于 I_2 项:

$$S \rightarrow L \cdot = R$$

$$R \rightarrow L \cdot$$

有 $ACTION[2, =]$ 是移进, 但 $FOLLOW(R) = \{ \$, = \}$ 又会导致在 $=$ 上进行规约, 故移进-规约冲突, 该文法不是 $SLR(1)$ 的

3.4.3 LR(1)语法

定义 24 (LR(1)项). $[A \rightarrow \alpha \cdot \beta, a]$, 其中 $A \rightarrow \alpha \cdot \beta$ 为该项的核(core), $A \rightarrow \alpha\beta$ 为生成式, a 是终端符号或 $\$$, 1 指项中第二个元素的长度, a 也被称为前看(lookahead)。只有当LR(1)项有 $[A \rightarrow \alpha \cdot, a]$ 的形式, 且下一输入符号为 a 时, 才会用 $A \rightarrow \alpha$ 进行规约 (这在构造解析表时会用到)。 a 总是 $FOLLOW(A)$ 的子集, 但往往是真子集。

算法 5 (计算 $CLOSURE(I)$). 对于每一 $[A \rightarrow \alpha \cdot B\beta, a] \in I$, 每一 G' 中的生成式 $B \rightarrow \gamma$, $\forall b \in FIRST(\beta a)$, 将 $[B \rightarrow \cdot \gamma, b]$ 加入 I 中。

初始化 $C = \{CLOSURE(\{[S' \rightarrow \cdot S, \$]\})\}$ 。

例 24. 考虑下面的增量语法

- (1) $S' \rightarrow S$
- (2) $S \rightarrow CC$
- (3) $C \rightarrow cC \mid d$

有 $FIRST(C) = \{c, d\}$, 可以构造得下面的 $LR(1)$ 自动机。以 $[S \rightarrow \cdot CC, \$]$ 为例, 考虑 $FIRST(C\$) = \{c, d\}$, 故闭包会新增四项 $[C \rightarrow \cdot cC, c]$ 、 $[C \rightarrow \cdot cC, d]$ 、 $[C \rightarrow \cdot d, c]$ 、 $[C \rightarrow \cdot d, d]$ 。

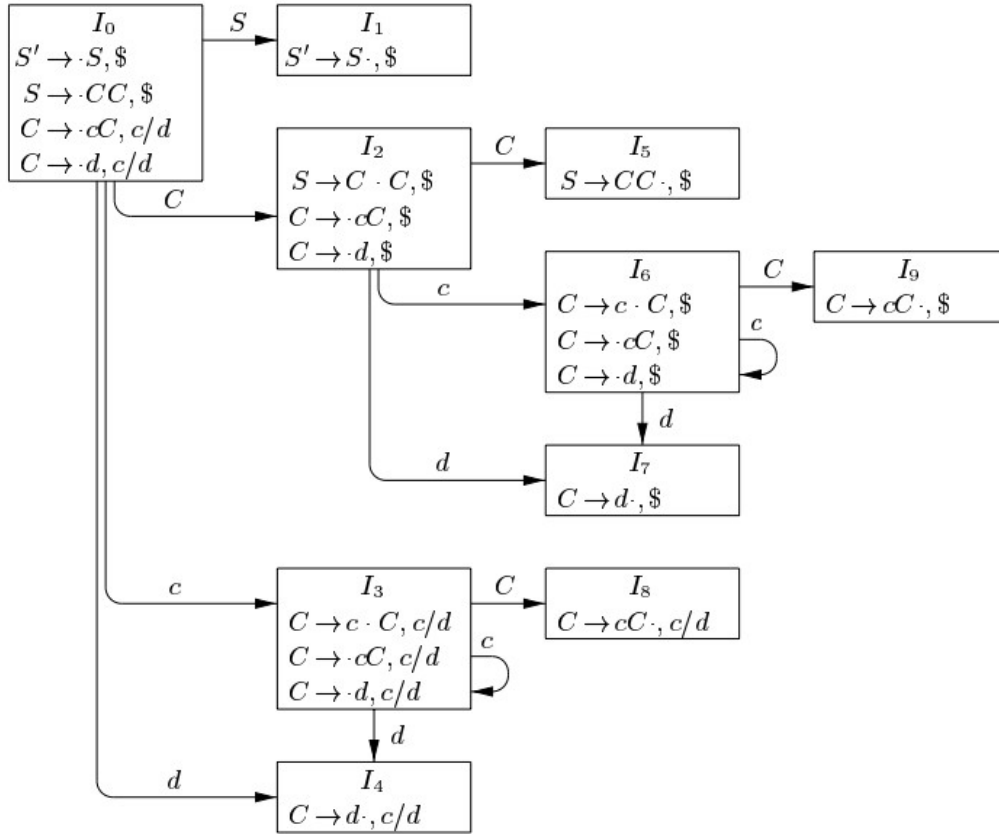


Figure 4.41: The GOTO graph for grammar (4.55)

类似地有规范 $LR(1)$ 解析表

	c	d	$\$$	S	C
0	$s3$	$s4$		1	2
1			acc		
2	$s6$	$s7$			5
3	$s3$	$s4$			8
4	$r3$	$r3$			
5			$r1$		
6	$s6$	$s7$			9
7			$r3$		
8	$r2$	$r2$			
9			$r2$		

LR(0)=SLR(1)的表项较少，但LR(1)的表项就指数级上涨。

3.4.4 LALR分析

LALR(Lookahead LR)在实践中很常用，表大小通常远小于规范LR表。核心思想是将LR(1)中具有相同核的表项合并。

定义 25 (核(core)). $LR(1)$ 项中第一部分的集合，如例24中 I_0 的核为 $\{S' \rightarrow \cdot S, S \rightarrow \cdot CC, C \rightarrow \cdot cC, C \rightarrow \cdot d\}$

例 25. 将上面例子中的 I_3 和 I_6 合并得到

$$C \rightarrow c \cdot C, c/d/\$$$

$$C \rightarrow \cdot cC, c/d/\$$$

$$C \rightarrow \cdot d, c/d/\$$$

I_4 和 I_7 合并得到

$$C \rightarrow d \cdot, c/d/\$$$

I_8 和 I_9 合并得到

$$C \rightarrow cC \cdot, c/d/\$$$

进而有LALR解析表

	c	d	$\$$	S	C
0	$s36$	$s47$		1	2
1			acc		
2	$s36$	$s47$			5
36	$s36$	$s47$			89
47	$r3$	$r3$	$r3$		
5			$r1$		
89	$r2$	$r2$	$r2$		

合并LR(1)项不会导致新的移进-规约冲突，但可能会产生新的规约-规约冲突。

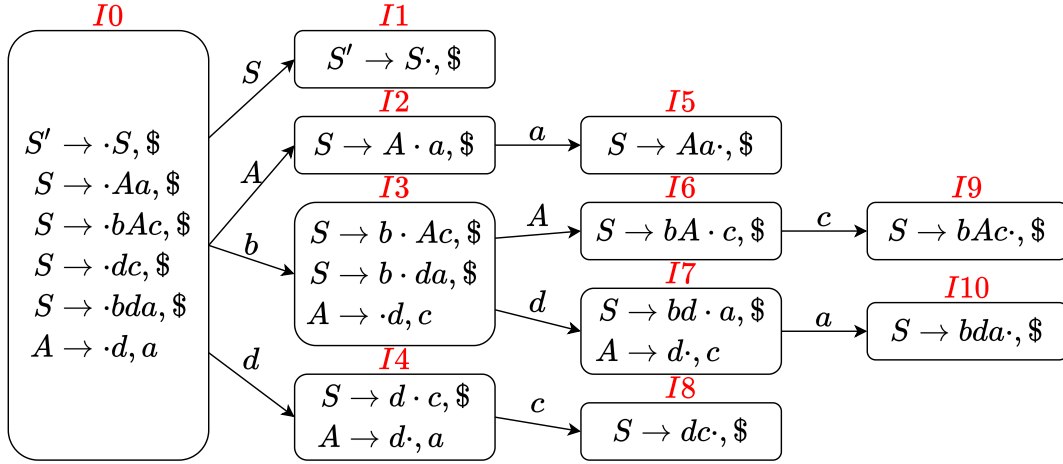
例 26. 证明下列文法

$$S \rightarrow Aa \mid bAc \mid dc \mid bda$$

$$A \rightarrow d$$

是LALR(1)文法但不是SLR(1)文法.

分析. 构造增广文法 $S' \rightarrow \cdot S$, $FIRST(\epsilon\$) = \$$, $FIRST(a\$) = a$, 可以得到 I_0 .



由上图知，没有相同核心(core)的状态，因此不需要合并，从而LALR分析表不冲突，该文法是LALR(1)文法。

又有 $FOLLOW(A) = \{a, c\}$ ，考虑图中的状态I4，当输入符号为c时， $c \in FOLLOW(A)$ ，既有移进 $S \rightarrow d \cdot c$ ，又有归约 $S \rightarrow d \cdot$ ，因此SLR分析表有冲突，该文法不是SLR(1)文法。

例 27. 证明下列文法

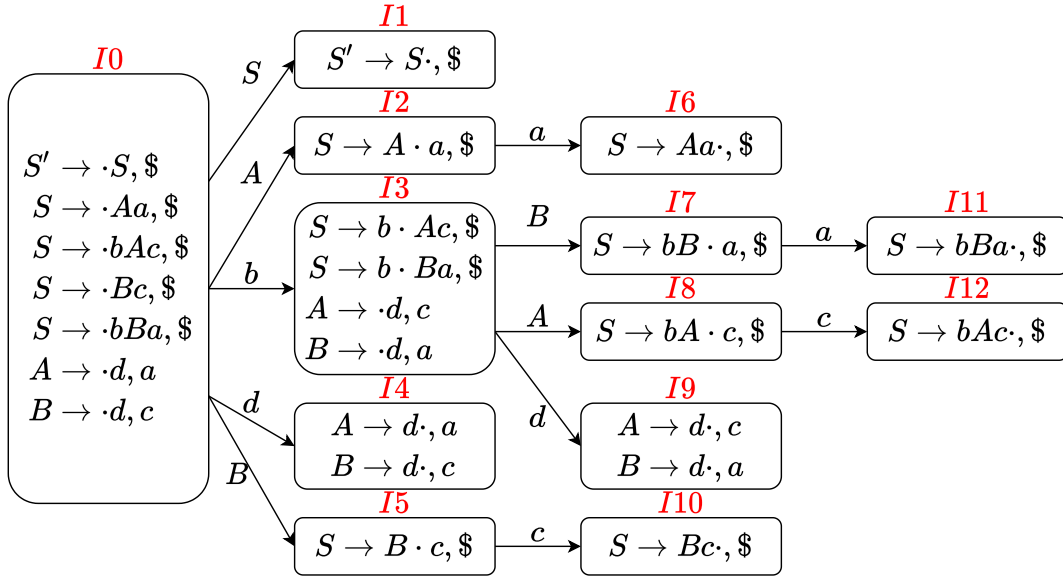
$$S \rightarrow Aa \mid bAc \mid Bc \mid bBa$$

$$A \rightarrow d$$

$$B \rightarrow d$$

是LR(1)文法但不是LALR(1)文法.

分析. 构造增广文法 $S' \rightarrow \cdot S$, $FIRST(a\$) = a$, $FIRST(c\$) = c$, 可以得到状态 I_0 。



由上面的DFA知LR分析表没有冲突，因此该文法是LR(1)文法。

但如果将图中相同核心的状态I4和I9合并，会有

$$A \rightarrow d., a/c$$

$$B \rightarrow d., a/c$$

即出现了规约-规约冲突，因此该文法不是LALR(1)文法。

可以在解析表(parsing table)层面来解决语法的二义性，即选择移进/规约的特定操作。

3.5 语法制导翻译

抽象语法树(Abstract Syntax Trees, AST)是将原本语法树中冗余的成分给去除，比如左右括号原本都是各自一个结点，但在AST中不会呈现。

语法制导翻译(syntax-directed translation)给语法符号提供了属性(attribute)，给生成式提供了动作(action)。

例 28. 对下列语法进行求值

$$E \rightarrow int \mid E + E \mid (E)$$

有语法制导定义

$$E \rightarrow int \quad E.val = int.val$$

$$E \rightarrow E_1 + E_2 \quad E.val = E_1.val + E_2.val$$

$$E \rightarrow (E_1) \quad E.val = E_1.val$$

可以在AST上标注(annotate)两种属性：

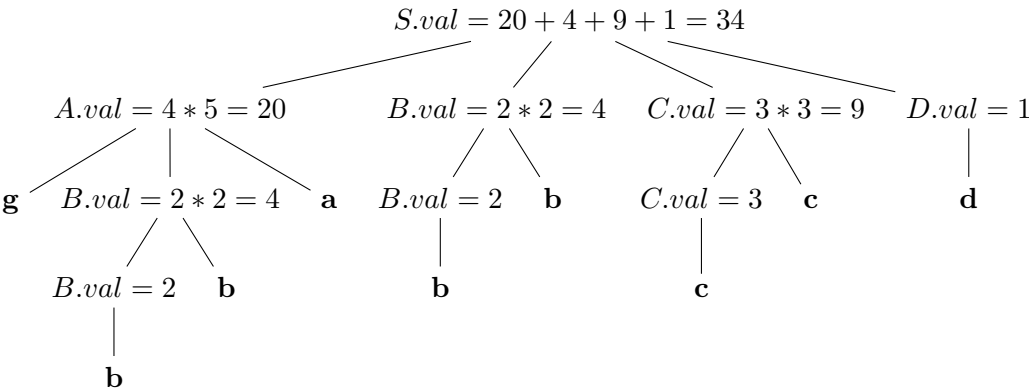
- 继承属性(inherited)：从语法树的父亲或兄弟中计算得到
- 综合属性(synthesized)：从后代计算得到

例 29. 考虑以下语法制导定义：

语法规则	语义规则
$S \rightarrow ABCD$	$S.val = A.val + B.val + C.val + D.val$
$A \rightarrow gBa$	$A.val = B.val * 5$
$B \rightarrow B_1b$	$B.val = B_1.val * 2$
$B \rightarrow b$	$B.val = 2$
$C \rightarrow C_1c$	$C.val = C_1.val * 3$
$C \rightarrow c$	$C.val = 3$
$D \rightarrow d$	$D.val = 1$

对于输入串 *gbbabbccd* 构造带注释的分析树 (annotated parse tree).

分析. 带注释的分析树如下



例 30. 下列文法定义了二进制整数的语法规则

$$\begin{aligned}
 N &\rightarrow SL \\
 L &\rightarrow LB \mid B \\
 S &\rightarrow + \mid - \\
 B &\rightarrow 0 \mid 1
 \end{aligned}$$

给出语法制导定义，求出该二进制整数的十进制值，存在综合属性 *val* 中

序号	生成式	语义规则
1	$N \rightarrow SL$	$N.val = S.sign * L.val$
2	$L \rightarrow L_1B$	$L.val = L_1.val * 2 + B.val$
3	$L \rightarrow B$	$L.val = B.val$
4	$S \rightarrow +$	$S.sign = 1$
5	$S \rightarrow -$	$S.sign = -1$
6	$B \rightarrow 0$	$B.val = 0$
7	$B \rightarrow 1$	$B.val = 1$

扩展文法:

$$\begin{aligned} \text{expr} &\rightarrow \text{expr}_1 + \text{term} \quad \{\text{print}(' + ')\} \\ \text{expr} &\rightarrow \text{expr}_1 - \text{term} \quad \{\text{print}(' - ')\} \\ \text{expr} &\rightarrow \text{term} \\ \text{term} &\rightarrow 0 \quad \{\text{print}(' 0 ')\} \\ \text{term} &\rightarrow 1 \quad \{\text{print}(' 1 ')\} \\ &\vdots \\ \text{term} &\rightarrow 9 \quad \{\text{print}(' 9 ')\} \end{aligned}$$

大括号中的语句称为动作(action)，这一系列产生式称为翻译模式(translation scheme)。

将动作视为产生式右端的一部分，则可得到扩展的语法树。对语法分析树做先序遍历，则可以得到后缀表达式。

3.6 总结

语法分析分为自顶向下和自底向上两大类，自顶向下是从开始符号出发，根据产生式规则**推导**给定的句子；自底向上则是由给定的句子**规约**到文法的开始符号。

其中LL(1)是自顶向下的分析法，LR(0)、SLR(1)、LR(1)、LALR(1)是自底向上的分析法。

文法	文法要求	构造分析表
LL(1)	生成式首符号不能相同	将生成式加到首符号项上
LR(0)	LR(0)自动机无冲突	
SLR(1)	下一输入 $a \in FOLLOW(X)$ 时进行规约	ACTION-GOTO表同左
LR(1)	用 $FIRST(\beta a)$ 确定前看项无冲突	下一输入为 a 才规约 $[A \rightarrow \alpha \cdot, a]$
LALR(1)	将LR(1)相同核项合并无冲突	同上

判断是否符合文法，LL(1)看下一输入符号是否冲突，自底向上则看是否有移进/规约冲突。

文法处理能力：LR(0) < SLR(1) < LALR(1) < LR(1)

即是右边的文法不一定是左边的文法。

4 语义分析与中间表示

- 高层中间表示：语法树、有向无环图(DAG)，用于静态类型检查
- 低层中间表示：三地址码，适合机器相关的任务（寄存器分配、指令选择）

```
x = y op z // arithmetic and logical
x = op y // negation and conversion
x = y // copy
goto L // unconditional jump
if x goto L // conditional jump
if False x goto L // conditional jump
if x op y goto L // relational operation
```



```

param x1 // parameter passing
param x2
...
param xn
call p, n // procedure call
y = call p, n // function call
return y // return a value
x = y[i] // indexed copy, i is the offset
x[i] = y
x = &y // address and pointer assignment
x = *y
*x = y

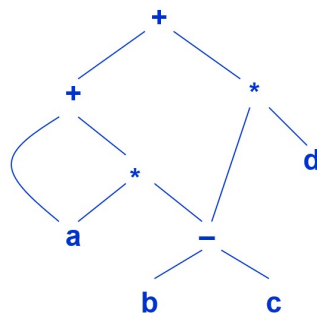
```

`top`指代当前的符号表，`gen`代表生成中间代码，`||`代表代码的连接，`next`定义执行完 S 之后紧跟的指令。

$S \rightarrow id = E$	<code>S.code = E.code gen(top.get(id.lexeme) '=' E.addr)</code>
$E \rightarrow E_1 + E_2$	<code>E.addr = new Temp() E.code = E1.code E2.code gen(E.addr '=' E1.addr '+' E2.addr)</code>
$E \rightarrow -E_1$	<code>E.addr = new Temp() E.code = E1.code gen(E.addr '=' minus E1.addr)</code>
$E \rightarrow (E_1)$	<code>E.addr = E1.addr E.code = E1.code</code>
$L \rightarrow L_1[E]$	<code>L.array = L1.array L.type = L1.type.element t = new Temp() L.addr = new Temp() gen(t '=' E.addr '*' L.type.width) gen(L.addr '=' L1.addr '+' t)</code>
$S \rightarrow \text{if } (B) \text{ } S1 \text{ else } S2$	<code>B.true = new Label() B.false = new Label() S1.next = S2.next = S.next S.code = B.code label(B.true) S1.code gen('goto' S.next) label(B.false) S2.code</code>

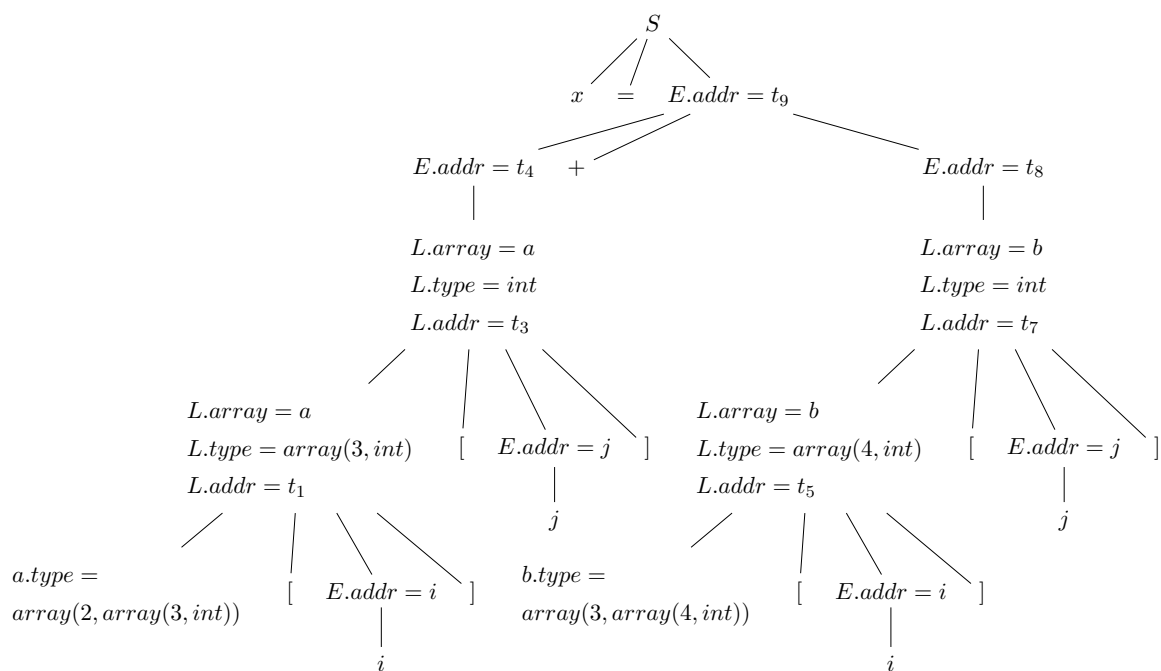
三地址码与对应的DAG如下图所示，注意这里将相同终端符号结点都给合并了。

$$\begin{aligned}
t_1 &= b - c \\
t_2 &= a * t_1 \\
t_3 &= a + t_2 \\
t_4 &= t_1 * d \\
t_5 &= t_3 + t_4
\end{aligned}$$



例 31. 令a表示一个 2×3 的整型数组, b表示一个 3×4 的整型数组. 假定一个整数的宽度为4. 试使用课本图6.22的翻译模式, 翻译赋值语句 $x=a[i][j]+b[i][j]$. 提示: 参考课本例6.12.

分析. 带注释的分析树如下



生成的三地址码如下

$$\begin{aligned}
t_1 &= i * 12 \\
t_2 &= j * 4 \\
t_3 &= t_1 + t_2 \\
t_4 &= a[t_3] \\
t_5 &= i * 16 \\
t_6 &= j * 4 \\
t_7 &= t_5 + t_6 \\
t_8 &= b[t_7] \\
t_9 &= t_4 + t_8 \\
x &= t_9
\end{aligned}$$

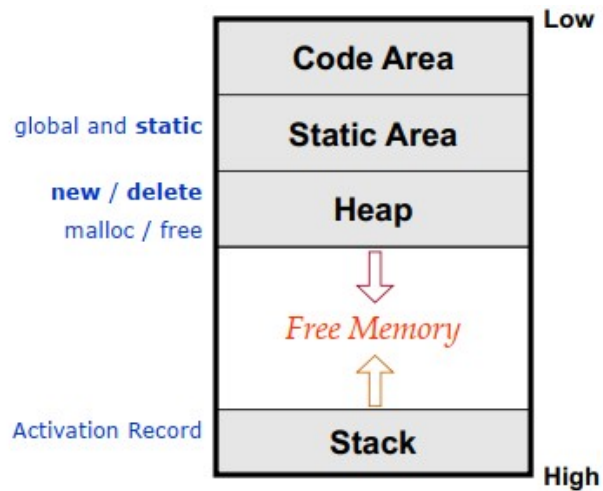
注意区别语法分析树(parse tree)和抽象语法树(AST)，前者从CFG得到，后者直接由源码得到。下面例子源自Wiki。

```
while b != 0
  if a > b
    a = a - b
  else
    b = b - a
return a
```



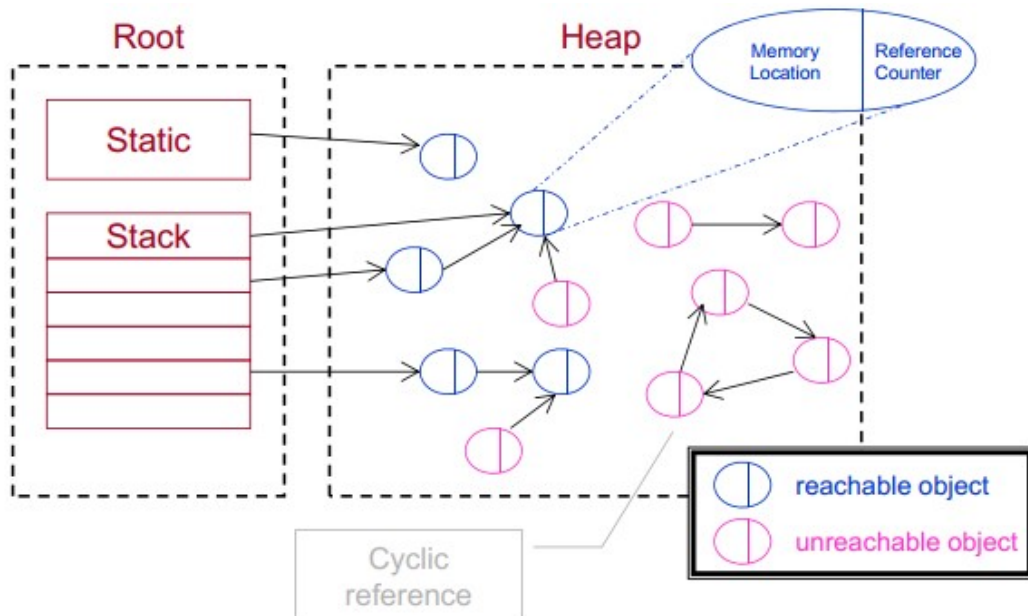
5 运行时系统

5.1 存储管理



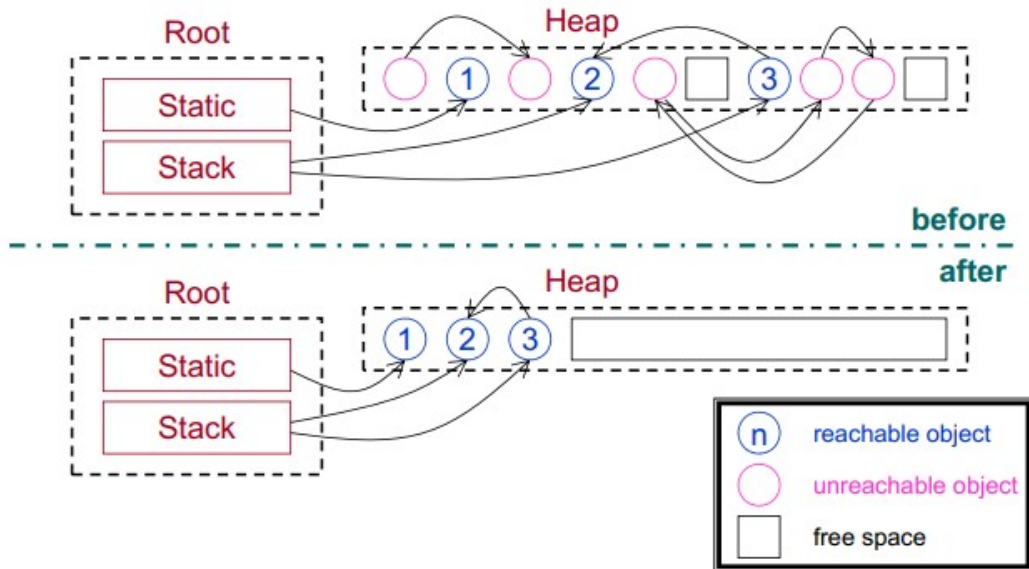
5.2 垃圾回收

- 引用计数(reference counting): 创建加1, 删除减1
 - 简单、立即增量式回收
 - 不能回收循环引用的示例

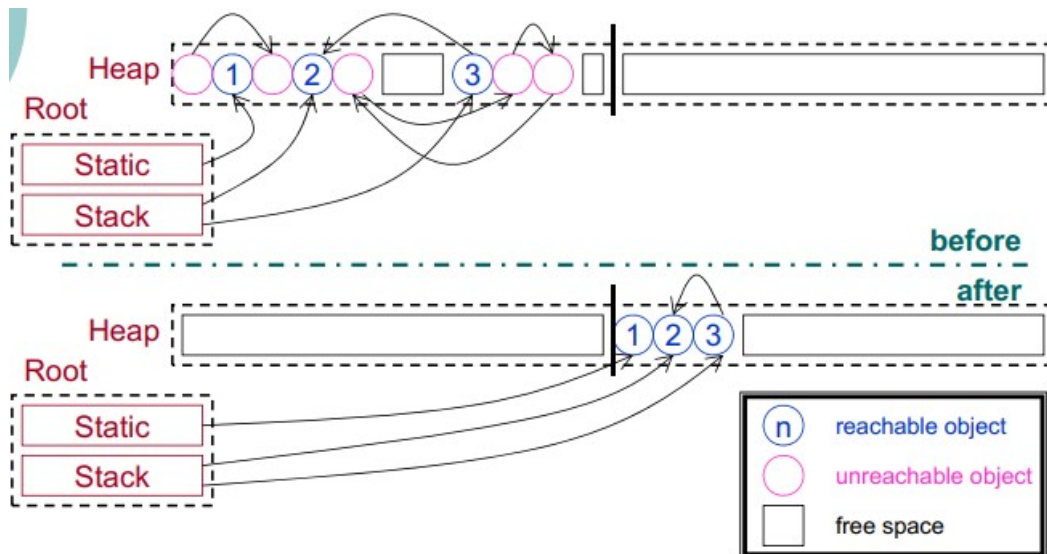


- 标记扫除(mark and sweep): 做图深搜找连通块
 - 有办法清除循环引用

- 大量垃圾时效率低，无法满足实时应用
- 标记压缩(mark and compact): 标记，计算新地址，拷贝对象到新地址并更新引用



- 拷贝收集(copying collector): 堆被划分为两个区域，可达对象一旦被发现就会立即被移动，但不可达对象不做改动



JVM采用了两代(young & old)的方式，对于年轻的对象采用拷贝收集，对于老的对象则采用标记压缩。

6 代码生成及优化

6.1 代码生成

- 指令选择: 选择最适合目标机器的指令来实现IR

- 寄存器分配和指派
- 指令调度

定义 26 (基本块(basic block)). 单一入口单一出口。成为 $leader$ 的指令:

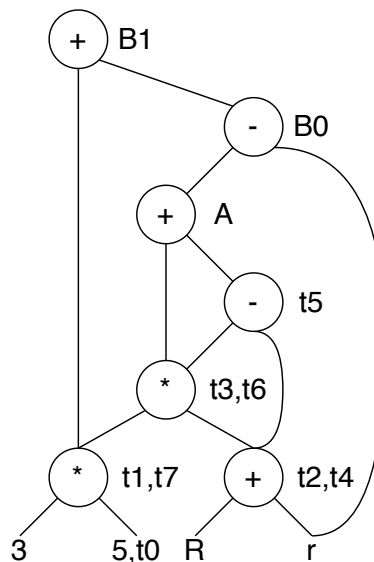
1. 第一条三地址指令
2. 条件或无条件跳转指令的目标
3. 条件或无条件跳转指令的下一指令

例 32. 考虑以下基本块:

$$\begin{aligned} t_0 &= 5 \\ t_1 &= 3 * t_0 \\ t_2 &= R + r \\ t_3 &= t_1 * t_2 \\ t_4 &= t_2 \\ t_5 &= t_3 - t_4 \\ t_6 &= t_1 * t_2 \\ A &= t_6 + t_5 \\ B &= A - r \\ t_7 &= t_1 \\ B &= t_7 + B \end{aligned}$$

1. 构造这一基本块的 DAG .
2. 假设只有 A 和 B 在基本块后面还要被引用, 产生优化后的三地址代码.

分析. 1. 基本块的 DAG 如下图所示



2. 由于只有 A 和 B 在基本块后面还要被引用，因此最终只需保留 A 和 B 的结果即可。无依赖关系的代码都可以作为死代码优化删除。最终优化后的三地址代码如下

$$\begin{aligned}t_2 &= R + r \\t_3 &= 15 * t_2 \\t_5 &= t_3 - t_2 \\A &= t_3 + t_5 \\B &= A - r \\B &= 15 + B\end{aligned}$$

6.2 代码优化

- 窥孔优化(peephole): 基于滑动窗口，最小粒度

```
x = x + 0 // eliminated
x = x * 1 // eliminated
y = x * 2 // y = x << 1
LD R0, a
ST a, R0 // eliminated
```

- 局部优化：在基本块内的优化
 - 公共子表达式删除
 - 常量/拷贝传递
 - 冗余操作消除
- 循环优化：在循环内的优化
- 全局优化：最粗粒度的优化

定义 27 (循环(loop)). 只有唯一入口/头的强连通子图

例 33. 考虑下列代码片段：

```
(1) m := 0
(2) v := 0
(3) if v >= n goto (19)
(4) r := v
(5) s := 0
(6) if r < n goto (9)
(7) v := v + 1
(8) goto (3)
```

```
(9) s := v + r
(10) y := 0 * x
(11) z := v - y
(12) x := z + r
(13) r := m - x
(14) if s <= m goto (17)
(15) m := s
(16) s := s + r
(17) r := r+1
(18) goto (6)
(19) return m
```

为这段代码划分基本块(*Basic Block*), 并画出控制流图(*Control Flow Graph*). 在答案中你可以直接画出控制流图, 但对图中的每个结点, 请用 $m \sim n$ 表示相应的基本块由第 m 至第 n 条语句组成.

分析. 如下图所示, 共9个基本块, 每个基本块包含的语句已在图中标出。

