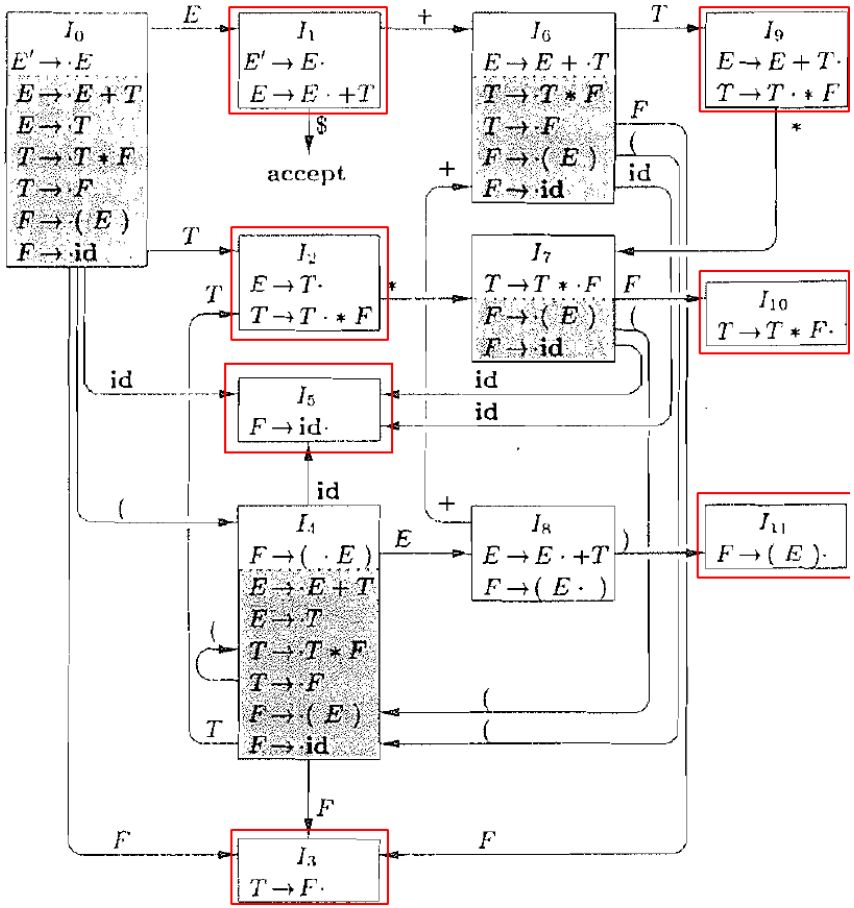
二、语法分析
(6. $LR1$ 语法分析器)

魏恒峰

hfwei@nju.edu.cn

2023 年 05 月 31 日



	ACTION						GOTO		
	id	+	*	(	)	$	E	T	F
0	$s5$			$s4$			$g1$	$g2$	$g3$
1		$s6$				acc			
2	$r2$	$r2$	$s7, r2$	$r2$	$r2$	$r2$			
3	$r4$	$r4$	$r4$	$r4$	$r4$	$r4$			
4	$s5$			$s4$			$g8$	$g2$	$g3$
5	$r6$	$r6$	$r6$	$r6$	$r6$	$r6$			
6	$s5$			$s4$				$g9$	$g3$
7	$s5$			$s4$					$g10$
8		$s6$			$s11$				
9	$r1$	$r1$	$s7, r1$	$r1$	$r1$	$r1$			
10	$r3$	$r3$	$r3$	$r3$	$r3$	$r3$			
11	$r5$	$r5$	$r5$	$r5$	$r5$	$r5$			

$$\mathbf{id} \qquad + \qquad * \qquad (\qquad) \qquad \$$$

$$\begin{array}{|c|}\hline (1)\ E \to E + T \\ \\ (2)\ E \to T \\ \\ (3)\ T \to T * F \\ \\ (4)\ T \to F \\ \\ (5)\ F \to (E) \\ \\ (6)\ F \to \mathbf{id} \\ \hline \end{array}$$

$$\textsc{Follow}(E) = \{+,), \$\}$$

$SLR(1)$ 分析表

状态	ACTION						GOTO		
	id	+	*	(	)	$	E	T	F
0	s5			s4			1	2	3
1		s6				acc			
2		r2	s7		r2	r2			
3		r4	r4		r4	r4			
4	s5			s4			8	2	3
5		r6	r6		r6	r6			
6	s5			s4				9	3
7	s5			s4					10
8		s6			s11				
9		r1	s7		r1	r1			
10		r3	r3		r3	r3			
11		r5	r5		r5	r5			

归约:

(3) $[k : A \to \alpha\cdot] \in I_i \wedge A \neq S' \implies \forall t \in \textsc{Follow}(A).\ \textsc{action}[i, t] = rk$

Definition ($SLR(1)$ 文法)

如果文法 G 的$SLR(1)$ **分析表**是**无冲突**的, 则 G 是 $SLR(1)$ 文法。

无冲突: ACTION 表中每个单元格最多只有一种动作

状态	ACTION						GOTO		
	id	+	*	(	)	$	E	T	F
0	s5			s4			1	2	3
1		s6				acc			
2		r2	s7		r2	r2			
3		r4	r4		r4	r4			
4	s5			s4			8	2	3
5		r6	r6		r6	r6			
6	s5			s4				9	3
7	s5			s4					10
8		s6			s11				
9		r1	s7		r1	r1			
10		r3	r3		r3	r3			
11		r5	r5		r5	r5			

两类可能的冲突: “移入/归约” 冲突、“归约/归约” 冲突

没有移入/移入冲突

非 $SLR(1)$ 文法举例

$$S \longrightarrow L = R \mid R$$

$$L \longrightarrow * R \mid \mathbf{id}$$

$$R \longrightarrow L$$

I_0: $S' \to \cdot S$
$S \to \cdot L = R$
$S \to \cdot R$
$L \to \cdot * R$
$L \to \cdot \mathbf{id}$
$R \to \cdot L$

I_1: $S' \to S\cdot$

I_2: $S \to L\cdot = R$
$R \to L\cdot$

I_3: $S \to R\cdot$

I_4: $L \to *\cdot R$
$R \to \cdot L$
$L \to \cdot * R$
$L \to \cdot \mathbf{id}$

I_5: $L \to \mathbf{id}\cdot$

I_6: $S \to L = \cdot R$
$R \to \cdot L$
$L \to \cdot * R$
$L \to \cdot \mathbf{id}$

I_7: $L \to *R\cdot$

I_8: $R \to L\cdot$

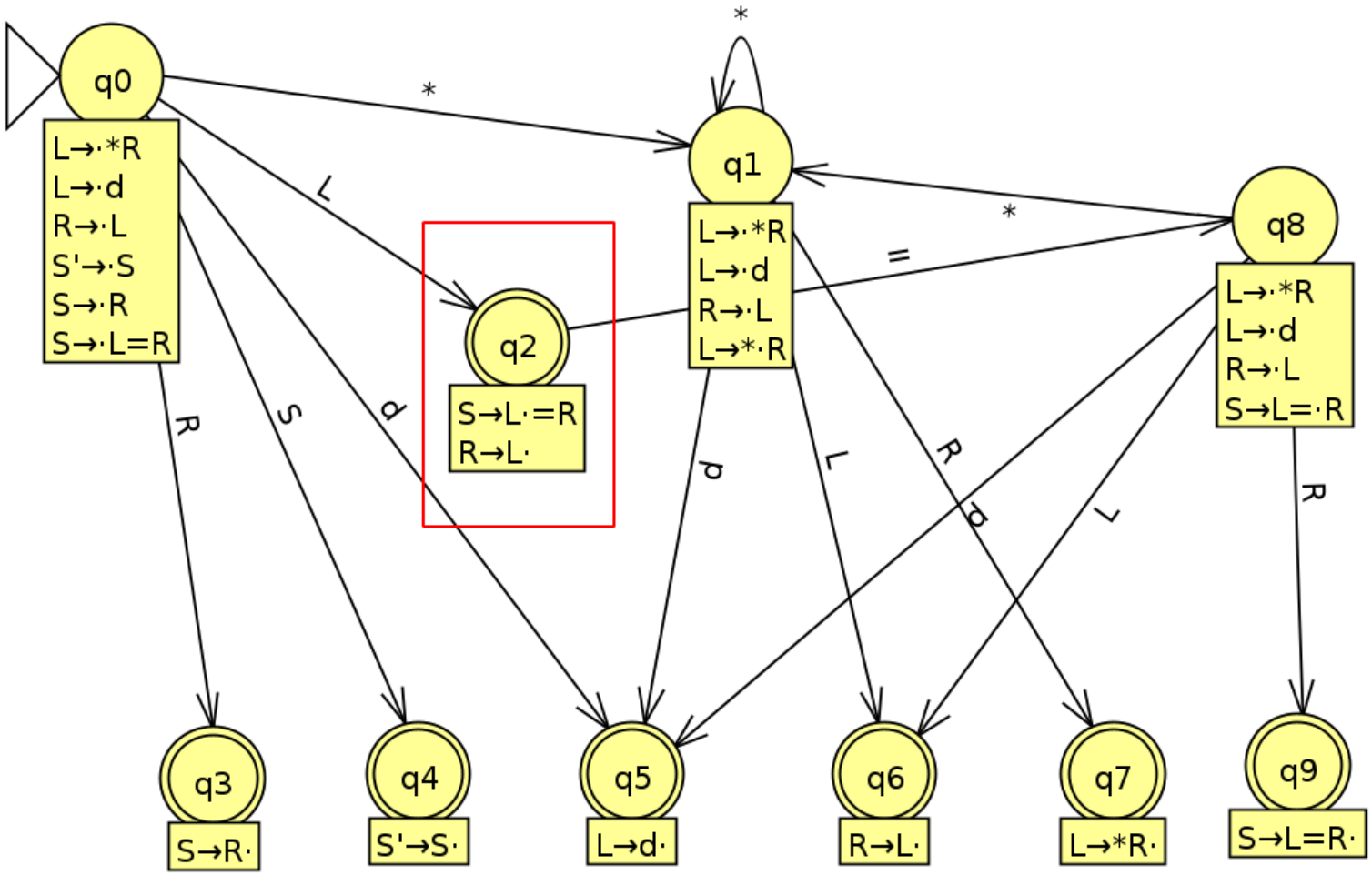
I_9: $S \to L = R\cdot$

$$[S \to L\cdot = R] \in I_2 \implies \text{ACTION}(I_2, =) \leftarrow s6$$

$$= \in \text{FOLLOW}(R) \implies \text{ACTION}(I_2, =) \leftarrow r5$$

即使考虑了$= \in \textsc{Follow}(R)$, 对该文法来说仍然不够

因为, 这仅仅说明 $=$ 可能跟在 R 后面, 是一种 “粗糙” 的估计



实际上, 该文法没有**以 $R = \cdots$ 开头**的最右句型

希望 LR 语法分析器的每个状态能**尽可能精确**地

指明**哪些输入符号可以跟在句柄 $A \to \alpha$ 的后面**

在 $LR(0)$ 自动机中, 某个项集 I_j 中包含 $[A \to \alpha\cdot]$

则在之前的某个项集 I_i 中包含 $[B \to \beta \cdot A\gamma]$ 与 $[A \to \cdot\alpha]$

这表明只有 $a \in \textsc{First}(\gamma)$ 时, 才可以进行 $A \to \alpha$ 归约

但是, 对 I_i 求闭包时, 仅得到 $[A \to \cdot\alpha]$, 丢失了 $\textsc{First}(\gamma)$ 信息

Definition ($LR(1)$ 项 (Item))

$$[A \to \alpha \cdot \beta, a] \qquad (a \in T \cup \{\$\})$$

此处, a 是**向前看符号**, 数量为 **1**.

思想: α 在栈顶, 期望剩余输入中**开头**的是可以从 βa 推导出的符号串

$$[A \to \alpha \cdot, a]$$

只有下一个输入符号为 a 时, 才可以按照 $A \to \alpha$ 进行归约

$$[A \to \alpha \cdot B\beta, a] \in I \qquad (a \in T \cup \{\$\})$$

$$\forall b \in \textsc{First}(\beta a).\ [B \to \cdot\gamma, b] \in I$$

$$J = \text{GOTO}(I, X) = \text{CLOSURE}\left(\left\{[A \to \alpha X \cdot \beta, a] \middle| [A \to \alpha \cdot X\beta, a] \in I\right\}\right)$$

$$(X \in N \cup T)$$

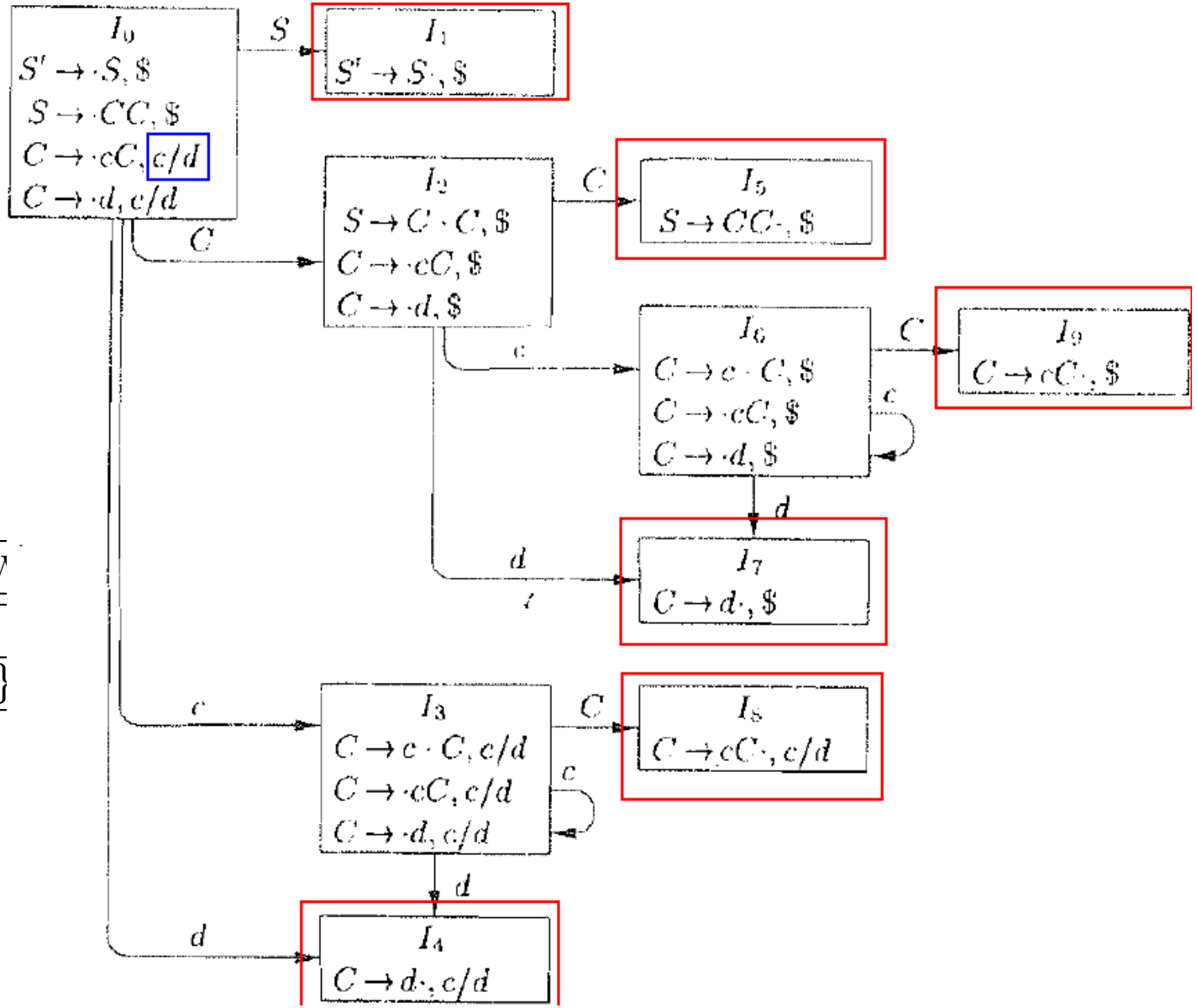
初始状态：$\textsc{closure}([S' \to \cdot S, \$])$

$$S' \longrightarrow S$$

$$S \longrightarrow C\ C$$

$$C \longrightarrow c\ C \mid d$$

	First	Follov
S	$\{c, d\}$	$\$$
C	$\{c, d\}$	$\{c, d, \$\}$



$LR(1)$ 分析表构造规则

(1) $\text{GOTO}(I_i, a) = I_j \wedge a \in T \implies \text{ACTION}[i, a] \leftarrow sj$

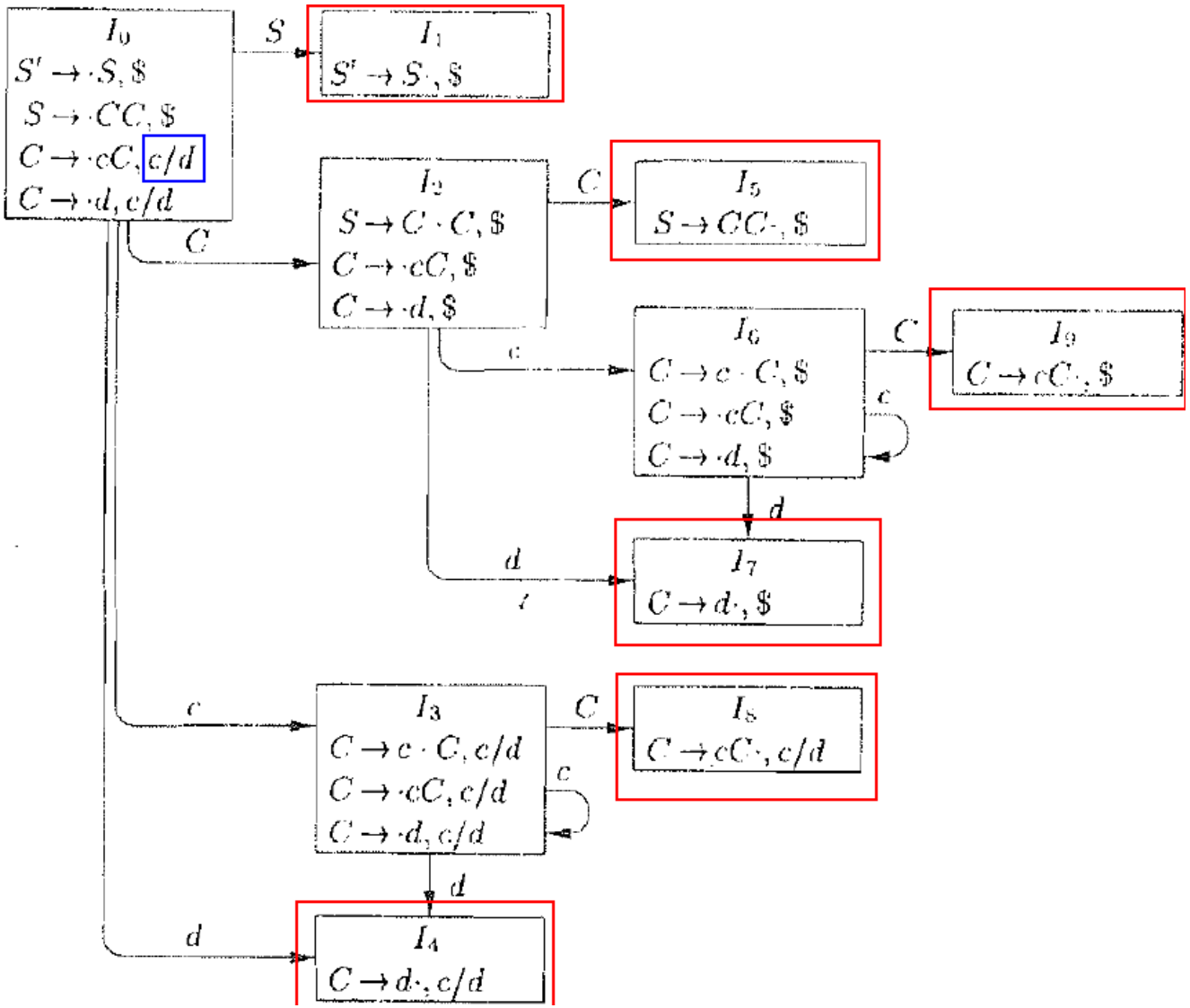
(2) $\text{GOTO}(I_i, A) = I_j \wedge A \in N \implies \text{GOTO}[i, A] \leftarrow gj$

(3) $[k : A \to \alpha\cdot, a] \in I_i \wedge A \neq S' \implies \text{ACTION}[i, a] = rk$

(4) $[S' \to S\cdot, \$] \in I_i \implies \text{ACTION}[i, \$] \leftarrow acc$

Definition ($LR(1)$ 文法)

如果文法 G 的$LR(1)$ **分析表**是**无冲突**的, 则 G 是 $LR(1)$ 文法。



状态	ACTION			GOTO	
	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		

$LR(1)$ 通过**不同的向前看符号**, 区分了状态对 $(3,6)$, $(4,7)$ 与 $(8,9)$

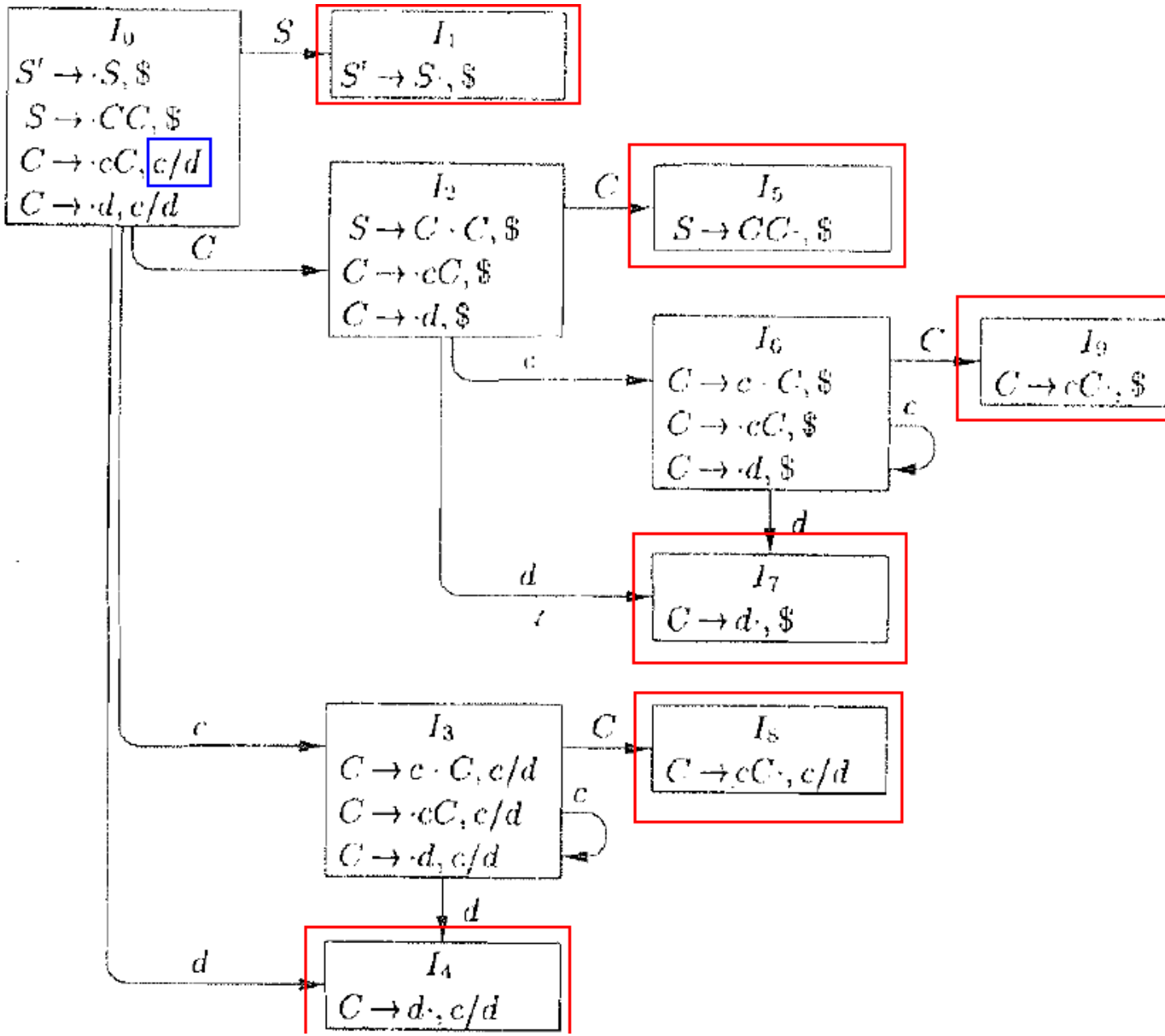
$w = ccdcd\$$

$S' \rightarrow S$

$S \rightarrow C\ C$

$C \rightarrow c\ C \mid d$

$L(G) = c^*dc^*d$



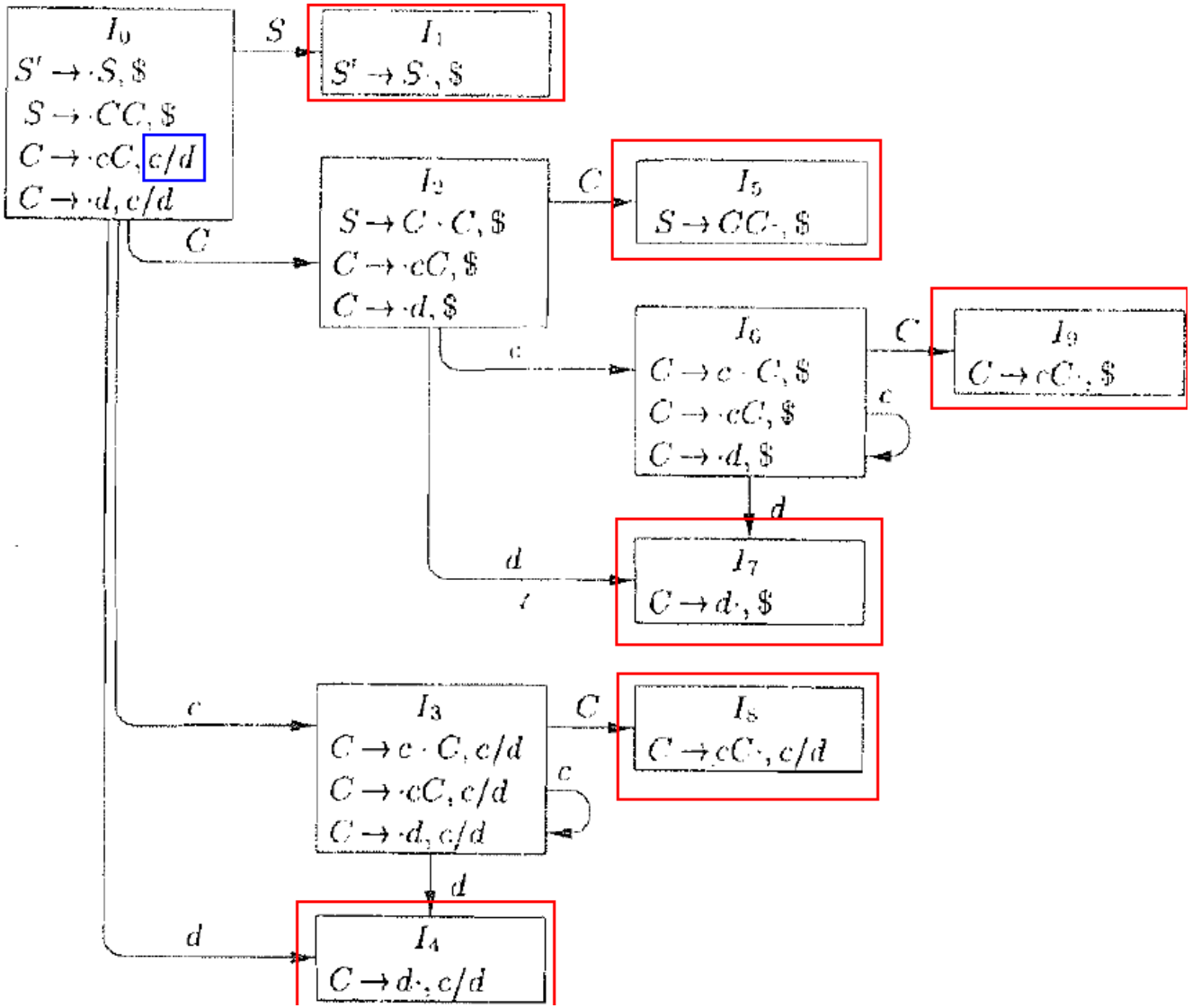
总结: $LR(0)$、$SLR(1)$、$LR(1)$ 的**归约**条件

$$[k : A \to \alpha\cdot] \in I_i \wedge A \neq S' \implies \forall t \in T \cup \{\$\}.\ \textsc{action}[i, t] = rk$$

$$[k : A \to \alpha\cdot] \in I_i \wedge A \neq S' \implies \forall t \in \textsc{Follow}(A).\ \textsc{action}[i, t] = rk$$

$$[k : A \to \alpha\cdot, a] \in I_i \wedge A \neq S' \implies \textsc{action}[i, a] = rk$$

$LR(1)$ 虽然强大, 但是生成的 $LR(1)$ 分析表可能过大, 状态过多



状态	ACTION			GOTO	
	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		

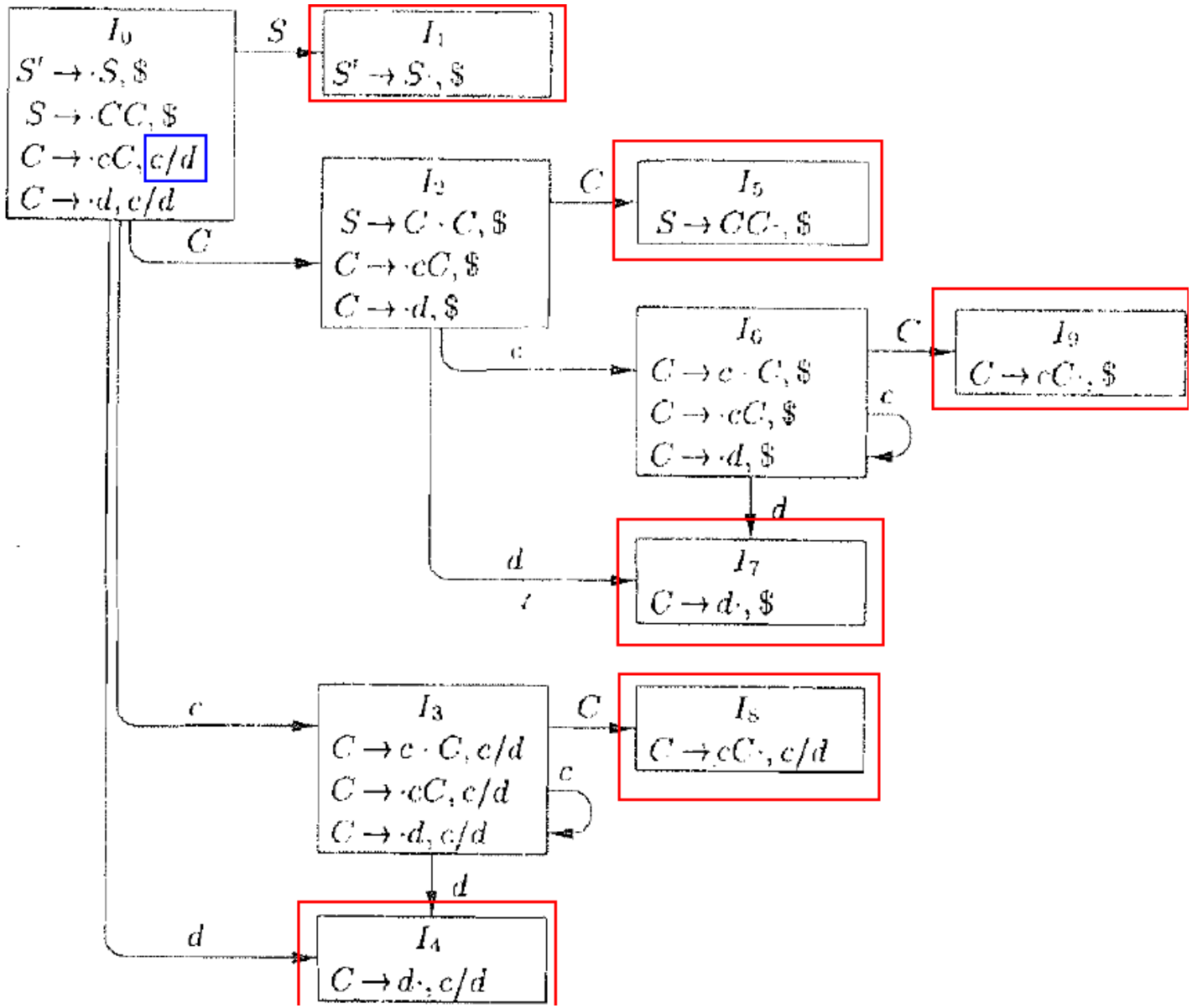
$LALR(1)$: 合并具有相同**核心** $LR(0)$项的状态 (忽略不同的向前看符号)

$S' \rightarrow S$

$S \rightarrow C\ C$

$C \rightarrow c\ C \mid d$

$L(G) = c^*dc^*d$



合并 I_4 与 I_7 为 I_{47} $\{[C \rightarrow d\cdot, c/d/\$]\}$, 继续合并 (I_8, I_9) 以及 (I_3, I_6)

状态	ACTION			GOTO	
	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		

状态	ACTION			GOTO	
	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		

Q：对于 $LR(1)$ 文法，合并得到的 $LALR(1)$ 分析表是否会引入冲突？

Theorem

$LALR(1)$ 分析表**不会**引入移入/归约冲突。

反证法

假设合并后出现 $[A \to \alpha\cdot, a]$ 与 $[B \to \beta\cdot a\gamma, b]$

则在 $LR(1)$ 自动机中，

存在某状态同时包含 $[A \to \alpha\cdot, a]$ 与 $[B \to \beta\cdot a\gamma, c]$ $(c \neq b)$

矛盾!

Q : 对于 $LR(1)$ 文法, 合并得到的 $LALR(1)$ 分析表是否会引入冲突?

Theorem

$LALR(1)$ 分析表**可能会**引入归约/归约冲突。

$$L(G) = \{acd, ace, bcd, bce\}$$

$$S' \longrightarrow S$$

$$S \longrightarrow a\ A\ d \mid b\ B\ d \mid a\ B\ e \mid b\ A\ e$$

$$A \longrightarrow c$$

$$B \longrightarrow c$$

$$\{[A \to c\cdot, d], [B \to c\cdot, e]\}$$
$$\{[A \to c\cdot, e], [B \to c\cdot, d]\}$$

$$\{[A \to c\cdot, d/e], [B \to c\cdot, d/e]\}$$

$LALR(1)$ 语法分析器的优点

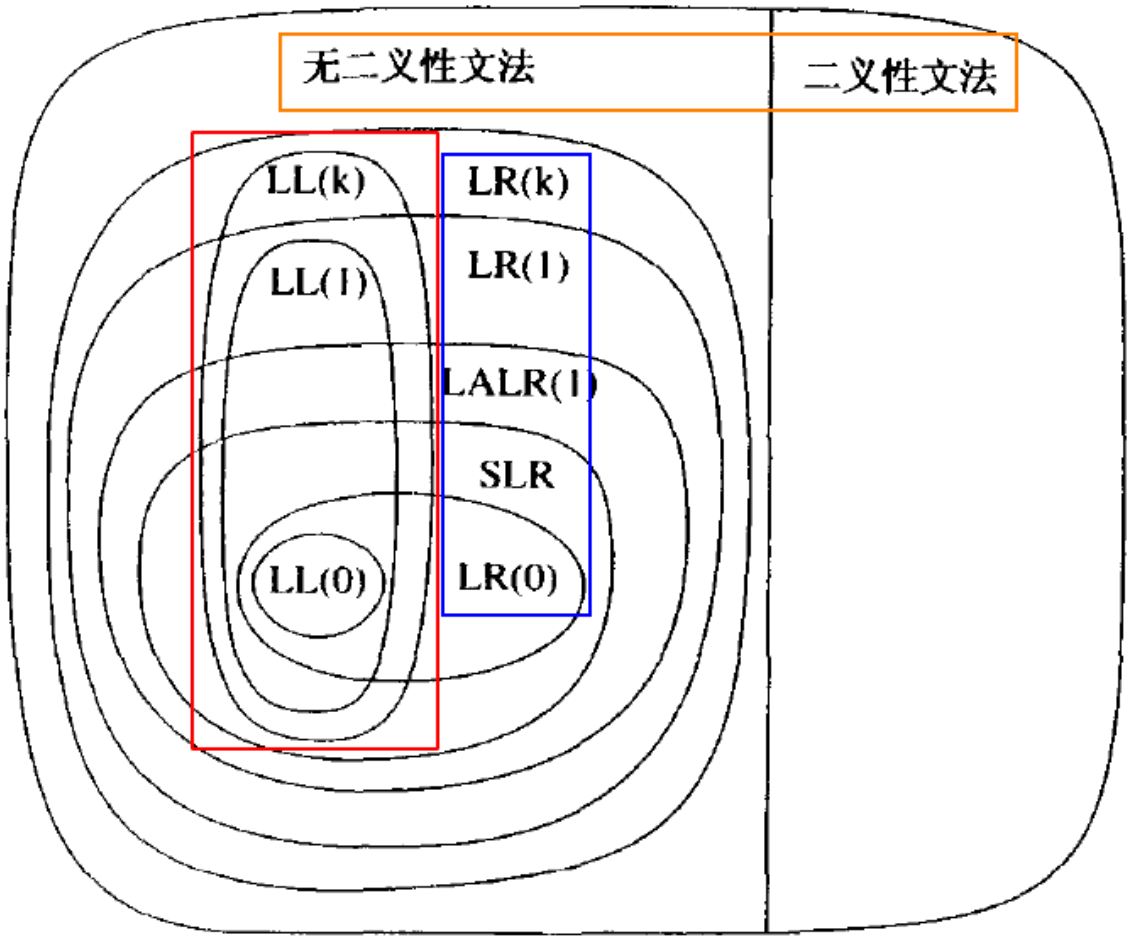
状态数量与 $SLR(1)$ 语法分析器的状态**数量相同**

($LALR(1)$ 与 $SLR(1)$ 都使用相同的 $LR(0)$ 核心项)

对于 $LR(1)$ 文法, **不会**产生移入/归约冲突

Q: 但是你却通过 $LR(1)$ 自动机构造 $LALR(1)$ 项集族?

第 **4.7.5** 节 (本科教学版): 高效构造 $LALR(1)$ 语法分析表的方法



“除 $LR(0)$ 外, 以上各种 LR 类**文法**对应的**语言**是等价的”

List of Parser Generators

Name	Parsing algorithm	Input grammar notation	Output languages	Grammar, code	Lexer	Development platform	IDE	License

Bison/Yacc/Lark

PEP 617 – New PEG parser for CPython

Author:	Guido van Rossum <guido at python.org>, Pablo Galindo <pablogsal at python.org>, Lysandros Nikolaou <lisandrosnik at gmail.com>
Discussions-To:	Python-Dev list
Status:	Accepted
Type:	Standards Track
Created:	24-Mar-2020
Python-Version:	3.9
Post-History:	02-Apr-2020

Rejected Alternatives

We did not seriously consider alternative ways to implement the new parser, but here's a brief discussion of LALR(1).

Thirty years ago the first author decided to go his own way with Python's parser rather than using LALR(1), which was the industry standard at the time (e.g. Bison and Yacc). The reasons were primarily emotional (gut feelings, intuition), based on past experience using Yacc in other projects, where grammar development took more effort than anticipated (in part due to shift-reduce conflicts). A specific criticism of Bison and Yacc that still holds is that their meta-grammar (the notation used to feed the grammar into the parser generator) does not support EBNF conveniences like `[optional_clause]` or `(repeated_clause)*`. Using a custom parser generator, a syntax tree matching the structure of the grammar could be generated automatically, and with EBNF that tree could match the "human-friendly" structure of the grammar.

Other variants of LR were not considered, nor was LL (e.g. ANTLR). PEG was selected because it was easy to understand given a basic understanding of recursive-descent parsing.

好消息： 善用 LR 语法分析器，处理**二义性**文法

坏消息： 二义性表现为“冲突”，需要根据某种规则进行消解

表达式文法

$$E \to E + E \mid E * E \mid (E) \mid \mathbf{id}$$

$$E \to E + T \mid T$$

$$T \to T * F \mid F$$

$$F \to (E) \mid \mathbf{id}$$

$$E \to TE'$$

$$E' \to +\, TE' \mid \epsilon$$

$$T \to FT'$$

$$T' \to *\, FT' \mid \epsilon$$

$$F \to (E) \mid \mathbf{id}$$

表达式文法: 使用 $SLR(1)$ 语法分析方法

$$E \to E + E \mid E * E \mid (E) \mid \mathbf{id}$$

I_0: $E' \to \cdot E$; $E \to \cdot E + E$; $E \to \cdot E * E$; $E \to \cdot(E)$; $E \to \cdot\mathbf{id}$

I_1: $E' \to E\cdot$; $E \to E\cdot + E$; $E \to E\cdot * E$

I_2: $E \to (\cdot E)$; $E \to \cdot E + E$; $E \to \cdot E * E$; $E \to \cdot(E)$; $E \to \cdot\mathbf{id}$

I_3: $E \to \mathbf{id}\cdot$

I_4: $E \to E + \cdot E$; $E \to \cdot E + E$; $E \to \cdot E * E$; $E \to \cdot(E)$; $E \to \cdot\mathbf{id}$

I_5: $E \to E * \cdot E$; $E \to \cdot E + E$; $E \to \cdot E * E$; $E \to \cdot(E)$; $E \to \cdot\mathbf{id}$

I_6: $E \to (E\cdot)$; $E \to E\cdot + E$; $E \to E\cdot * E$

I_7: $E \to E + E\cdot$; $E \to E\cdot + E$; $E \to E\cdot * E$

I_8: $E \to E * E\cdot$; $E \to E\cdot + E$; $E \to E\cdot * E$

I_9: $E \to (E)\cdot$

$$\{+, *\} \subseteq \textsc{Follow}(E)$$

考虑到**结合性与优先级**:

状态	ACTION						GOTO
	id	+	*	(	)	$	E
0	s3			s2			1
1		s4	s5			acc	
2	s3			s2			6
3		r4	r4		r4	r4	
4	s3			s2			7
5	s3			s2			8
6		s4	s5		s9		
7		r1	s5		r1	r1	
8		r2	r2		r2	r2	
9		r3	r3		r3	r3	

$\mathbf{id} + \mathbf{id} * \mathbf{id}$ $\mathbf{id} + \mathbf{id} + \mathbf{id}$

条件语句文法

stmt → **if** *expr* **then** *stmt*

| **if** *expr* **then** *stmt* **else** *stmt*

| **other**

$$S' \rightarrow S$$

$$S \rightarrow i\ S\ e\ S \mid i\ S \mid a$$

条件语句文法: 使用 $SLR(1)$ 语法分析方法

$$S' \rightarrow S$$

$$S \rightarrow i\ S\ e\ S \mid i\ S \mid a$$

I_0: $S' \rightarrow \cdot S$
$S \rightarrow \cdot iSeS$
$S \rightarrow \cdot iS$
$S \rightarrow \cdot a$

I_1: $S' \rightarrow S\cdot$

I_2: $S \rightarrow i\cdot SeS$
$S \rightarrow i\cdot S$
$S \rightarrow \cdot iSeS$
$S \rightarrow \cdot iS$
$S \rightarrow \cdot a$

I_3: $S \rightarrow a\cdot$

I_4: $S \rightarrow iS\cdot eS$
$S \rightarrow iS$

I_5: $S \rightarrow iSe\cdot S$
$S \rightarrow \cdot iSeS$
$S \rightarrow \cdot iS$
$S \rightarrow \cdot a$

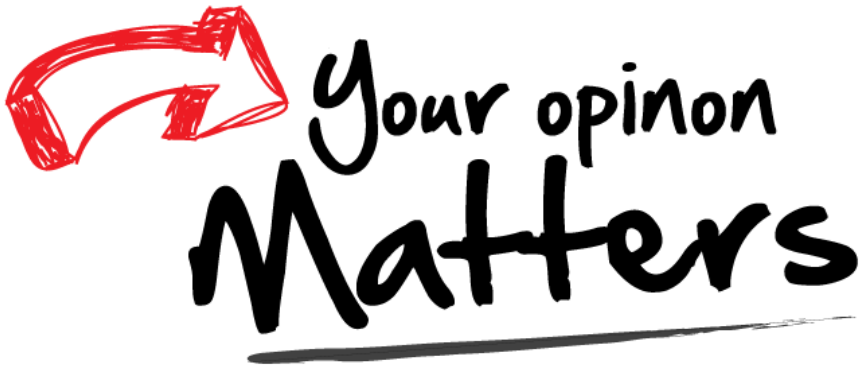
I_6: $S \rightarrow iSeS\cdot$

状态	ACTION				GOTO
	i	e	a	\$	S
0	s2		s3		1
1				acc	
2	s2		s3		4
3		r3		r3	
4		s5		r2	
5	s2		s3		6
6		r1		r1	

$e \in \text{FOLLOW}(S)$

$\text{ACTION}[4, e] = s5$


Thank
You!



Office 926

hfwei@nju.edu.cn