角色反演算法*

白 硕[1,2], 张 浩[2+]

[1](国家计算机与网络信息安全管理中心,北京 100031)
[2](中国科学院 计算技术研究所,北京 100080)

A Role Inverse Algorithm

BAI Shuo[1,2], ZHANG Hao[2+]

[1](National Administrative Center for Network and Information Security, Beijing 100031, China)
[2](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)
+ Corresponding author: Phn: 86-10-62587953, E-mail: zhanghao@software.ict.ac.cn
http://www.ict.ac.cn
Received 2001-10-19; Accepted 2002-04-10

Bai S, Zhang H. A role inverse algorithm. *Journal of Software*, 2003,14(3):328~333.

Abstract: A computational mechanism for context-free language parsing is proposed in this paper, which is called role inverse algorithm. The mechanism is based on the assignment of appropriate roles to categories according to their contexts. A kind of effectual "look ahead" function is introduced into chart parsing by this mechanism at acceptable cost. As a result, lots of useless arcs in a chart can be avoided, so that a parsing process is accelerated. This mechanism can be used in many applications, such as natural language processing.
Key words: parsing; CFG (context-free grammar); algorithm; NLP (natural language processing)

摘 要: 给出了面向上下文无关语言的句法分析的一种计算机制:角色反演算法.这种机制通过引入句法范畴的"角色"这一概念以及相应的角色反演操作,用较小的空间代价在 Chart 算法中实现了较强的"预读"(look ahead)功能.这使其能节约大量的无用边,从而加速分析过程的推进.这种机制可以用于自然语言处理等多种应用领域.
关键词: 句法分析;上下文无关文法;算法;自然语言处理
中图法分类号: TP18 **文献标识码**: A

上下文无关语言的句法分析最早产生于编译技术的需要[1,2].广义 LR 算法(即 Tomita 算法及其变种)[3]和 Chart 算法[4,5]是目前采用得最广泛的两种算法.下面,我们对比它们的优缺点.

广义 LR 算法的时间效率比较高.空间效率受到分析表体积(=状态数*(终结符数+非终结符数))的影响而比较耗费[3].而且,广义 LR 算法在分析表构造这一环节上不具有增量化的性质.也就是说,在增加一条规则的时候,原有 LR 分析表的内容无法使用,必须重新生成.

* Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1998030510 (国家重点基础研究发展规划(973))

第一作者简介: 白硕(1956－),男,辽宁辽阳人,博士,研究员,博士生导师,主要研究领域为自然语言处理,网络安全.

Chart 算法的空间效率比较高.它几乎没有像 LR 分析表那样的静态数据结构.但是由于没有预读的机制,它的时间效率偏低.Chart 算法有很好的增量化性质,规则的改变几乎不会引起额外的计算量.

把二者的优点结合起来,克服二者的缺点,是新的高效率的分析算法追求的目标.

因此,我们提出在这个方向上的一个新的计算机制:角色反演算法.

1 基本定义

设 G 是一个无空产生式的上下文无关文法. G 的规则从 1 开始按自然数顺序编号.用 $x.y$ 表示"编号为 x 的规则的右部第 y 个成分"这样一个角色.下面的函数在后面的算法中会反复用到:

$Length(x)$:编号为 x 的规则的右部的长度. $Left(x)$:编号为 x 的规则的左部非终结符句法范畴. $Cat(x.y)$:角色 $x.y$ 的句法范畴(终结符或非终结符).

显然,范畴和角色之间是一对多的关系.

定义 1. G 的一个范畴 C (终结符或非终结符)在预读字符(终结符) t 的条件下的角色反演函数 $I(C,t)$ 为

$$I(C,t)=\{x.y \mid Cat(x.y)=C \wedge (y<Length(x) \rightarrow t \in FIRST(Cat(x.y+1))) \wedge (y=Length(x) \rightarrow t \in FOLLOW(Left(x)))\}.$$

定义 2. G 的一个非终结符范畴 C 在预读字符(终结符) t 下的开放规则函数 $Start(C,t)$ 为

$$Start(C,t)=\{x \mid Left(x)=C \wedge t \in FIRST(Cat(x.1)\}.$$

角色反演算法的思想是,在 Chart 算法中引入预读(look ahead)机制,利用角色反演函数计算和开放规则函数计算出一套类似 LR 分析表那样的角色反演分析表(但所占空间要小得多),利用这套分析表来指导 Chart 算法中边的选择和构造.

2 分析表的构造

"角色反演分析表"由 I 表和 Start 表构成.我们首先来描述角色反演分析表的构造算法.该算法称为双图(twins-graph)算法.

所谓双图,就是要把一部上下文无关文法的所有终结符和非终结符的每个符号对应到两个顶点(整个图可以想象为上下两层,每层的顶点对应于一套终结符+非终结符).双图中的边有如下 3 种类型:(1) 在上层顶点间,如果符号 C_1 出现在某规则的左部,符号 C_2 出现在该规则右部的末尾,则连接一条从 C_2 到 C_1 的有向边.(2) 下层顶点间,如果符号 C_1 出现在某规则的左部,符号 C_2 出现在该规则右部的开头,则连接一条从 C_1 到 C_2 的有向边.(3) 上下层顶点之间,如符号 C_1 和符号 C_2 在某个规则的右部相左-右邻接,则连接一条从上层的 C_1 到下层的 C_2 之间的有向边.最后,为处理句结束符\$,我们在下层增设\$节点,并规定上层的 S 与下层的\$之间有一条有向边.

容易发现,每条从上层节点出发的有向边对应于边起点范畴的一个角色;每条从下层节点出发的有向边对应于边起点范畴的一条规则.特别地, $\langle S,\$\rangle$ 对应于假想的 0 号规则 $S' \rightarrow S\$$ 为 S 带来的 0.1 角色(该角色在分析中的出现标志着对句子的分析宣告成功).这样,我们把就把角色和规则标记到了对应的边上.

在构造了相对于一部文法的双图之后,我们就可以发现 FIRST 和 FOLLOW 函数的直观意义及其计算方法,进而构造出 I 表和 Start 表.

明显地,有如下的两个引理:

引理 1. 从上层的 C_1 顶点到下层的 C_2 (终结符)顶点之间存在一条路径,当且仅当 C_2 属于 $FOLLOW(C_1)$.

引理 2. 从下层的 C_1 顶点到下层的 C_2 (终结符)顶点之间存在一条路径,当且仅当 C_2 属于 $FIRST(C_1)$.

通过对路径长度应用数学归纳法可以证明必要性,通过对派生长度应用数学归纳法可以证明充分性.

根据上面两个引理,可以比较容易地得到 I 表和 Start 表.

定理 1. $I(C,t)$ 等于所有从上层的 C 顶点到下层的 t 顶点之间的路径上的第 1 条边的标记.

定理 2. $Start(C,t)$ 等于所有从下层的 C 顶点到下层的 t 顶点之间的路径上的第 1 条边的标记.

由定义直接可得.

此外,由于语法规则的添加只能导致在双图上添加边而不会减少边,所以上述分析表(I 表和 Start 表)有很好

的增量性质.

下面是一个语法例子:

词性标注规则:

N→我|县长

V→是|派|来

句法规则:

(1) *S*→*NP VP*;

(2) *NP*→*N*;

(3) *NP*→*S*ϕ的;

(4) *VP*→*V NP*;

(5) *S*ϕ→*NP VP*ϕ;

(6) *VP*ϕ→*V V*.

构造双图如图 1 所示.

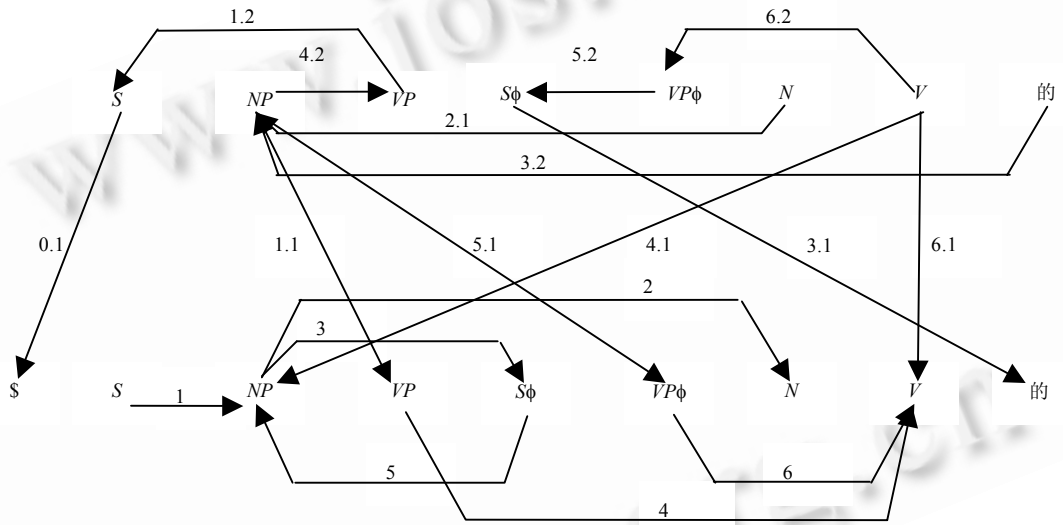

Fig.1 An example of twins-graph

图 1 双图示例

例如,从上层 *NP* 到下层 *V* 的路径只有两条:5.1-6 和 1.1-4,则 $I(NP,V)$=5.1/1.1.

例如,从下层 *NP* 到下层 *N* 的路径包括:2 和 3-5-2,3-5-3-5-2,…,则 Start(NP, N) =2/3.

利用双图算法生成的角色反演表见表 1 和表 2.

Table 1 The inverted role list

表 1 I 表(反演角色表)

	N	*V*	的	$
S				0.1(success)
NP		1.1/5.1		4.2
VP				1.2
*S*ϕ			3.1	
*VP*ϕ			5.2	
N		2.1		2.1
V	4.1	6.1	6.2	
的		3.2		3.2

Table 2 The table of starting rules

表 2 Start 表(开放规则表)

	N	*V*	的
S	1		
NP	2/3		
VP		4	
*S*ϕ	5		
*VP*ϕ		6	

3 角色反演分析算法

有了上面这些准备,我们就可以介绍角色反演算法了.角色反演算法与Chart算法同样要求把字符间隔作为节点编号,同样有 init,pre,scan,comp 这 4 种分析动作[4,5].不同的是,角色反演算法通过划分角色对添加的边进行了更加严格的筛选,使得无用边的数目得到了有效的控制.

Chart 算法中的边表示为 $[i,j,A\to\alpha\cdot B\beta]$,若 $A\to\alpha B\beta$ 的规则编号为 x,且 $|\alpha|=y$,我们可以将其编码为 $[i,j,x.y]$,这就是角色编码(注意:我们用 $[i,i,x.0]$ 来表示开放规则 x,下面可以看到,这样统一表达有计算上的便利).

算法. 角色反演分析算法.

- init 动作:
  添边 $[0,0,S'\to\cdot S]$ ($[0,0,0.0]$);
- 用以下动作反复添边,直至无边可添:
  对于边 $[i,j,A\to\alpha\cdot B\beta]$ ($[i,j,x.y]$),
  - pre 动作:
    如果 z: $B\to\gamma$ 满足 $z\in\text{Start}(B,\text{string}[j+1])$,则添边 $[j,j,B\to\cdot\gamma]$ ($[j,j,z.0]$);
  - scan 动作:
    如果 string $[j+1]$ 属于范畴 B,并且 $x.y+1\in I(B,\text{string}[j+2])$,则添边 $[i,j+1,A\to\alpha B\cdot\beta]$ ($[i,j+1,x.y+1]$);
  - comp 动作:
    如果有边 $[j,k,B\to F\cdot]$,并且 $x.y+1\in I(B,\text{string}[k+1])$,则添边 $[i,k,A\to\alpha B\cdot\beta]$ ($[i,k,x.y+1]$);
- 如果有边 $[0,n,S'\to S\cdot]$ ($[0,n,0.1]$),则分析成功.

4 运行实例

仍采用第 2 节中引入的语法例子.以输入中文句子"我是县长派来的"为背景,输入字符串"*N V N V V* 的",分析器得到如表 3 所示的分析过程和如图 2 所示的分析结果.

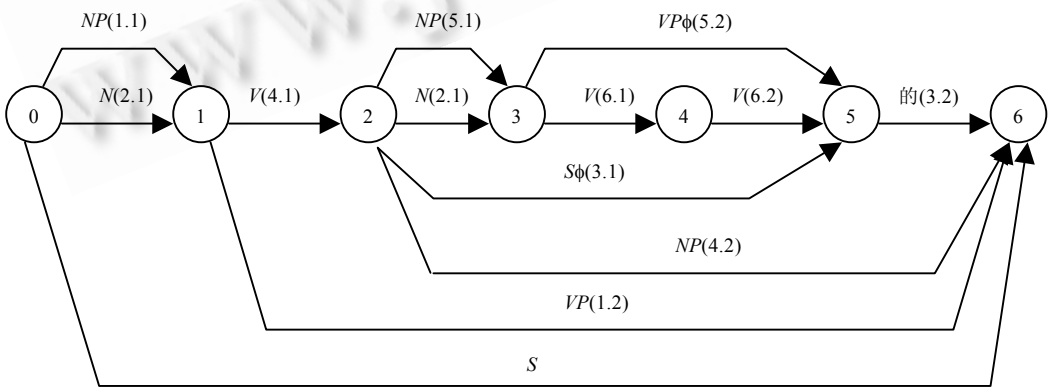

Fig.2 The result of parsing

图 2 句法分析结果

Table 3 An instance of execution

表 3 运行实例

Action	Starting point	Ending point	Preread	Prediction	To start	Category	Role
init	0	0	*N*	*S*	1		
pre				*NP*	2/3		
pre				*S*ϕ	5		
scan	0	1	*V*			*N*	2.1*
comp	0					*NP*	1.1
comp	0					*NP*	5.1
pre	1			*VP*	4		
pre				*VP*ϕ	6		
scan	1	2	*N*			*V*	4.1
pre	2			*NP*	2/3		
pre				*S*ϕ	5		
scan	2	3	*V*			*N*	2.1*
comp	2					*NP*	5.1
pre	3			*VP*ϕ	6		
scan	3	4	*V*			*V*	6.1
scan	3	5	的			*V*	6.2*
comp	2					*VP*ϕ	5.2*
comp	2					*S*ϕ	3.1
scan	2	6	$			的	3.2*
comp	1					*NP*	4.2*
comp	0					*VP*	1.2*
comp	0					*S*	**success**

我们注意标注深颜色的一行记录.这里,*NP* 在预读到 *V* 的情况下,根据反演表,本来有 1.1/5.1 两个选择,但是当前节点开放的规则只有 2/3/5,可匹配的预测只有 5.1,于是排除了 1.1 这个选择.另外,按照预测,这个 *NP* 本来也可以充当角色 4.2,但是根据反演表,*NP* 遇到 *V* 只能在 1.1/5.1 中选择,于是也排除了 4.2 这个角色选择.这意味着"县长"既不能作整句的主语,也不能作整句的宾语,而只能作定语从句 *S*ϕ的主语.这正是依靠了角色反演分析表,实现了上下文配合双重筛选.

5 结论与讨论

角色反演算法的前期准备工作——双图算法的复杂度与填入分析表中的角色(规则)总数(一个角色/规则可能出现在表中多次)呈线形关系,即不超过 *O*(所有规则长度之和*终结符总数),而且得到的分析表的大小只是((终结符个数+非终结符个数)*终结符个数).可见,开销是比较小的.而我们在句子分析中得到的优化是可观的.简单看来,我们只是得到了基于对角色的筛选细致程度大小的常数比的时间复杂度缩减,而实际上边表的规模缩小了,而边表是要频繁访问的,我们从访问边表环节得到的效率提升要更大.在针对一个有 432 条规则、3 000 词条的词典的英文句法,对 3 700 多个高度歧义的句子进行分析并生成压缩树林[4,5]的实验中,我们得到的实验结果是:在 PII266 的机器上只需 40 秒(和我们挑战的同类优化过的 Chart 分析器需要 1 分钟以上).

本文工作表明:角色反演算法是一种综合广义 LR 算法与 Chart 算法各自的优点、具有更高时空效率的算法.这种算法可以有效地应用于自然语言处理的实际问题当中.当然,本文工作只是提出了一个一般的计算机制.在处理一门具体的自然语言(比如英语、汉语)的时候,还需要做大量的工作,包括:

词性的标注.理想的上下文无关语言的终结符集合是确定的,但自然语言中词汇的词性标注却是非确定的,存在着大量的兼类现象.我们必须针对这一特点,首先设计出一套程序内部使用的确定性的词性标注体系,然后再通过某种映射规则(也是上下文无关语法规则的一部分),把内部的词性标注符号翻译成有语言学意义的词性标注符号.值得说明的是,这一步翻译也是句法分析过程中的一部分,翻译中的不确定性可以通过句法分析的全过程得到一定程度的消解.

规则库的建立.规则库是一项不可缺少的基本建设任务.面向汉语语法分析的规则库的建立,因为缺少相应的积累和借鉴,有更大的难度.由于后续工作(比如语义分析以及语义信息反过来指导句法分析)的需要,实用化

的规则库中一般都不可能只是纯粹的上下文无关产生式规则,其中非产生式规则的部分如何纳入角色反演算法的机制,值得进一步深入研究.

分析结果的表示.在实际工作中,我们使用了共享子树森林这种缩合的数据结构.这种结构有利于节约空间和时间.由于这部分工作基本上与Chart分析法中的分析结果表示机制是一致的,所以在本文中没有花费太多笔墨.事实上,如何在考虑到与后续的语义分析机制相衔接的情况下优化地表示分析结果,远不是一个已经解决的问题.

致谢 感谢中国科学院计算技术研究所的程学旗、刘群二位先生.作者从与他们的交流讨论中得到许多启发,这对形成本文的结果有重要的帮助.感谢参加信息科学技术讨论班"句法分析技术"专题研讨的各位同行就本文工作提出的宝贵意见.

References:

[1] Aho AV, Ullman JD. The Theory of Parsing, Translation and Compiling. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[2] Aho AV, Ullman JD. Principles of Compiler Design. Reading, MA: Addison-Wesley, 1977.

[3] Tomita M, Ng SK. The Generalized LR Parsing Algorithm. Boston: Kluwer, 1991. 1~16.

[4] Russell SJ, Norvig P. Artificial Intelligence: a Modern Approach. Prentice-Hall, 1995. 696~703.

[5] Graham SJ, Harrison MA, Ruzzo WL. An improved context-free recognizer. ACM Transactions on Programming Languages and Systems, 1980,2(3):415~462.