ISSN1000-9825 Journal of Software 软件学报 2000,11(8):1030~1034

基于产生式集划分的上下文无关语言句子生成*

王泓皓 董韫美

(中国科学院软件研究所计算机科学开放研究实验室 北京 100080)

E-mail: wanghh@ox.ios.ac.cn

摘要 给出了上下文无关文法(context-free grammar,简称CFG)产生式集的一种划分方法,可将产生式分为两类.使用一类产生式进行推导时,推导过程将无限进行下去;使用另一类进行推导时,推导过程将迅速结束.证明了CFG句子生成过程一定是先使用一类产生式使生成的句型不断变长、变复杂,再使用另一类产生式使句型变成句子.据此,提出了一种可控制的通用句子生成方法.其生成一条句子的时间和空间复杂度是$O(r+n)$,其中n是生成句子的长度或深度限制,r是给定上下文无关文法中产生式的数目.同时,给出了适应不同需要的句子生成策略.

关键词 上下文无关文法,产生式集合,产生式集合划分,句子生成,句子生成策略.

中图法分类号 TP311

对于上下文无关文法(context-free grammar,简称CFG)[1],有确定且有效的方法判断一个字符串是否为该文法的句子,并进行语法分析.在很多情况下还需要有一种确定且有效的方法,生成给定文法的句子实例.它被广泛运用于各种复杂系统的检测和验证,例如,为形式证明系统和超大规模集成电路的功能测试提供测试用例[2].在SAQ(specification acquisition)[3,4]系统中,也需要从文法自动生成若干句子实例,以便检验系统所获取的文法.因此,希望找到一种可控的句子自动生成方法.

以往自动产生给定文法的句子,主要采用随机法和分层迭代法.

随机法中常用的是随机选取产生式法[5].它是从文法开始符号出发,不断地从产生式集中随机选取产生式,对当前句型中相应的非终极符进行替换,直到句型中没有非终极符出现时为止.这种方法实现简单,但无法确定生成句子的长度和推导深度以及生成过程何时可以终结.

另一种随机法是基于概率模型的方法[6],即为文法中每一条产生式赋予一个固定的频度值,根据已使用产生式的频度值计算已生成句子或句型的代价,并确定下一步使用频度值在什么范围内的产生式.这种方法的优点是:按照一定的概率分布选用产生式,提供了一种控制句子生成的手段,产生的句子比较自然.但其效果很大程度上依赖于对产生式频度值的设置和对句子或句型的代价计算,且很难控制所生成句子的长度和推导深度.

Timothy Hickey 和 Jacques Cohen 在1983年也提出一种基于概率的随机生成句子的方法[7].它根据已生成句型的信息和给定的句子长度限制,动态地计算每条产生式被选取的概率,可以等概率地生成给定长度的所有句子.但其算法较复杂,成功生成一条长度为n的句子的时间复杂度一般为$O(n^2(\log n)^2)$.

分层迭代法[8]的思想是,把上下文无关文法(CFG)所对应的上下文无关语言(context-free language,简称CFL)按推导深度分成可数无穷个句子集合,称之为层.CFL即为所有各层中句子的并集.其中第0层为文法开始符号,从第n层中各个句型出发,对其中的每一个非终极符都使用相应的产生式进行替换,从而构造出第$n+1$层.这种方法的时间和空间代价巨大,但可以枚举出文法中所有的句子,并确定每个句子的长度和推导深度.

* 本文研究得到国家自然科学基金(No. 69873042)和国家"九五"重点科技攻关项目基金(No. 96-729-06-02)资助.作者王泓皓,1975年生,硕士生,主要研究领域为软件设计方法.董韫美,1936年生,研究员,博士生导师,中国科学院院士,主要研究领域为软件设计方法.

本文通讯联系人:王泓皓,北京 100080,中国科学院软件研究所计算机科学开放研究实验室

本文2000-01-17收到原稿,2000-04-21收到修改稿

以上各种方法虽然都可以生成给定文法的句子,但均有其不足之处,各适用于不同的用途.

作者通过对句子生成过程以及产生式性质的研究,找到一种产生式划分方法,即将CFG产生式集合划分成两类.使得其中一类产生式可使生成的句型无限变复杂,另一类产生式则在不超过CFG中非终极符个数的有限步内从句型推出句子,并据此提出一种通用的句子生成方法,可以高效地生成符合给定长度或深度限制的句子.本文提出的句子自动生成方法,已经在SAQ系统微机版中实现并得到检验.

本文第1节给出记号约定并定义基本概念.第2节提出产生式划分的方法,证明了有关性质.第3节讨论了这些性质在句子生成中的应用.第4节总结全文.

1 定义和记号约定

为论述方便,我们约定文中提到的所有文法皆为上下文无关文法.文法G定义为一个四元组

$$G=\{V_N,V_T,\boldsymbol{P},S\},$$

其中V_N为非终极符集合,其元素用大写字母$A,B,C,\ldots$表示;$S\in V_N$,为开始符;V_T为终极符集合.$(V_N\cup V_T)^*$表示由非终极符和终极符的串的集合,其元素用希腊字母$\alpha,\beta,\gamma,\ldots$表示.$V_T^*$表示终极符串的集合.$\varphi$表示空集合.产生式是形如$A\to\alpha$的推导规则,其中$A$称为产生式的左部,$\alpha$称为产生式的右部,且不允许出现形如$A\to A$的产生式.$\boldsymbol{P}$为产生式的有穷集合.

我们用$\alpha\Rightarrow\beta$表示:存在γ,δ,η和A,满足$\alpha=\gamma A\delta,\beta=\gamma\eta\delta$,且$A\to\eta$是$\boldsymbol{P}$中的一个产生式.$\alpha\overset{*}{\Rightarrow}\beta$表示:存在$\alpha_1,\alpha_2,\ldots,\alpha_m(m\geqslant 1)$使得

$$\alpha=\alpha_1\Rightarrow\alpha_2\Rightarrow\ldots\Rightarrow\alpha_m=\beta,$$

并且称序列$\alpha_1,\alpha_2,\ldots,\alpha_m$为从$\alpha$到$\beta$的一个推导.$S\overset{*}{\Rightarrow}\beta$,则称$\beta$为$G$的一个句型.如果$\beta\in V_T^*$,则称$\beta$为句子.文法$G$所定义的语言(上下文无关语言或CFL)即所有句子构成的集合.

下面定义几个概念.

定义1. 如果产生式的右部不含非终极符,则称该产生式为平凡产生式,否则称为非平凡产生式.

定义2. 在文法G中,如果某个非终极符满足以下条件之一,则称其为临界非终极符.

(1) 以该非终极符为左部的产生式皆为平凡产生式.

(2) 如果有非平凡产生式,则该产生式右部中的非终极符必为其他临界非终极符.

根据产生式在文法中的不同作用,可将其分为两类:核心产生式和外围产生式.

定义3. 如果一个产生式满足以下任意一个条件,则称该产生式为外围产生式,否则称为核心产生式.

(1) 该产生式是一个平凡产生式.

(2) 该产生式右部中的非终极符均为临界非终极符.

在文法G中,所有核心产生式组成的集合称为核心产生式集,记为$\boldsymbol{P}^K$;所有外围产生式组成的集合称为外围产生式集,记为$\boldsymbol{P}^M$.

定义4. 在文法G中,若有$X\overset{*}{\Rightarrow}\alpha$,且$X\neq\alpha$(即至少推导一步),则称$X\to\alpha$为文法$G$的一个类产生式.

在推导过程中,使用类产生式一次,相当于使用有推导依赖关系的产生式若干次.因此,如果在CFG产生式集$\boldsymbol{P}$中加入使用该文法生成的类产生式,不会对该文法所对应的CFL产生任何影响,且定义3对产生式的分类方法同样适用于类产生式.

2 分划的性质

根据定义2和定义3,可以很容易找到一个机械方法求出给定文法的$\boldsymbol{P}^K$和$\boldsymbol{P}^M$.

显然,核心产生式集$\boldsymbol{P}^K$和外围产生式集$\boldsymbol{P}^M$构成了对文法产生式集$\boldsymbol{P}$的一个分划,即$\boldsymbol{P}^K\cup\boldsymbol{P}^M=\boldsymbol{P},\boldsymbol{P}^K\cap\boldsymbol{P}^M=\varphi$,并且这两个集合皆有以下性质:

性质1. 从X出发,设若只使用$\boldsymbol{P}^K$中产生式进行推导时,得到$X\overset{*}{\Rightarrow}\alpha$.在只使用$\boldsymbol{P}^M$中产生式进行推导时,得到$X\overset{*}{\Rightarrow}\beta$,则按定义3在对这两个类产生式进行分类时,$X\to\alpha$为核心产生式;$X\to\beta$为外围产生式.即该划分方法

对于产生式推导运算封闭.

证明:分情况证明:

(1) 要证明 $X\to\alpha$ 为核心产生式.

采用反证法.

设在 P^K 中推导得到 $X\dot{\Rightarrow}\mu Y\nu\Rightarrow\mu\beta\nu=\alpha$,且 $X\to\mu\beta\nu$ 为外围产生式,则其右部 $\mu\beta\nu$ 中或不含非终极符,或所有非终极符均为临界非终极符.无论何种情形,均导致 $Y\to\beta$ 为外围产生式.而已知在 $X\dot{\Rightarrow}\alpha$ 的过程中使用的产生式,包括 $Y\to\beta$ 在内,均为核心产生式.矛盾.所以假设不成立.

(2) 同理,易证 $X\to\beta$ 为外围产生式. □

推论 1. 对于性质 1 中的 $X\dot{\Rightarrow}\alpha$,α 中必定包含非终极符.证略.

推论 2. 从 X 出发,仅使用 P^M 中产生式进行推导,则经过有数步推导后得到 $X\dot{\Rightarrow}\beta$,且有 $\beta\in V_T^*$.我们称 $X\dot{\Rightarrow}\beta$ 所对应语法树的高度为该推导的深度,则其推导深度不超过文法 G 中临界非终极符的个数.证略.

性质 1 及其推论说明,只使用 P^K 中的产生式将会使推导无限地进行下去,而使用 P^M 中的产生式,推导必在有数步内终止.

可以猜测,CFG 生成句子的一般过程是:首先使用 P^K 中的产生式,使语法树从文法开始符 S 不断生长,相应地得到越来越复杂的句型.然后使用 P^M 中的产生式,使语法树生出叶节点,相应地句子生成过程终结.

对于任意文法 G 的任何非空句子语法树 T,从树根到叶子节点的节点序列称为该叶节点(相对于树 T)的路径,记为 $Y_1,\ldots,Y_n$.其中 $Y_1,\ldots,Y_{n-1}$ 为非终极符,Y_1 是根节点,Y_n 是叶节点.对于树 T,$Y_1,\ldots,Y_{n-1}$ 中的每一个,都对应于一个产生式,构成路径产生式序列,记为 $P_1,\ldots,P_{n-1}$.n 称为路径长度.显然,语法树 T 中所有叶节点的路径长度的极大值为树 T 的高度.下面证明任意 CFG 的句子生成过程都具有前述猜测的性质.

性质 2. 对于任何句子的任一叶节点 Y_n,其路径产生式序列 $P_1,\ldots,P_{n-1}$ 都可以分成两段,分别可以为空.第 1 段全部由 P^K 中的产生式组成,第 2 段全部由 P^M 中的产生式组成.即存在 $1\leqslant j\leqslant n$,使得产生式序列 $P_1,\ldots,P_{j-1}\in P^K$,余下的产生式序列 $P_j,\ldots,P_{n-1}\in P^M$.

证明:施归纳于语法树 T 的高度.

基始. 当高度为 2 时,该命题显然成立.此时,路径产生式序列只包含一个产生式 P_1,且为 $Y\to\alpha,\alpha\in V_T^*$.包括 α 是空串的退化情形.

归纳. 设对于所有高度不超过 $n(n\geqslant 2)$ 的语法树,性质 2 为真.则对于高度为 $n+1$ 的语法树 T,剪去根节点及由它直接指向的叶节点(对于它们,路径产生式序列只有一个产生式,故性质 2 为真).余下的子树高度均不超过 n,根据归纳假设,对于这些子树中任一路径产生式序列,只需考察以 P^K 中的产生式开始(即 $P_2\in P^K$)的情形.可以肯定,此时树 T 的根节点所对应的产生式 $P_1\in P^K$,否则,如果有 $P_1\in P^M$,则根据定义 3 和定义 2,必定有 $P_2\in P^M$,矛盾. □

3 在句子生成中的应用

利用本文证明的产生式性质,可以得到一种新的句子生成方法.

其主要思想是将句子的生成过程分为两个步骤:首先从开始符号出发,仅使用核心产生式集 P^K 进行推导,此时生成句型的过程不会终止,且句型不断变长变复杂;任何时候改用 P^M 中的产生式进行推导,句子生成的过程均在有数步内终止.

根据上述性质 1 的推论 2,可以确定文法中每一个非终极符在使用 P^M 中的产生式进行推导时所生成句子的长度和推导深度,即每个非终极符的终极化代价.对于在 P^M 中没有产生式的非终极符,则求其最小终极化代价.因此,在新的生成句子过程中,可以随时计算出已生成句型所对应的一组句子的长度和推导深度,从而根据需要对句子的生成过程加以控制.

具体步骤如下:

步骤 1. 根据定义 2 和定义 3,求出给定文法的 P^K 和 P^M.

步骤 2. 求出文法中每个非终极符的终极化代价或最小终极化代价.

步骤 3. 从开始符号出发,使用核心产生式集 P^K 进行推导,同时计算当前句型所对应句子的长度和推导深度,直到满足要求时为止.

步骤 4. 使用 P^M 中的产生式,将步骤 3 生成的句型终极成句子.

其中,步骤 1 和步骤 2 的时间和空间复杂度是一个与给定文法有关的常数,一般为 $O(r)$,其中 r 为文法中产生式的数目.由于文法中不允许出现自推出的产生式,所以步骤 3 在成功生成一个句型时的时间和空间复杂度为 $O(n)$,其中 n 为对句子长度或推导深度的限制.在步骤 4 中,生成每一个句子的时间和空间复杂度也是一个常数.因此在生成一条句子时,整个方法的时间和空间复杂度为 $O(r+n)$.对比 Timothy Hickey 和 Jacques Cohen 提出的方法,该方法以较小的时间、空间代价,生成满足长度或推导深度限制的句子.

在 SAQ 系统的概念(即非终极符)获取过程中,需要及时将所获取概念的实例反馈给用户,以检验其正确性.当获取全部概念后,还可能要求产生若干句子实例来进行检验.根据 SAQ 系统中句子生成的不同要求,还可以在产生式使用上采取不同策略,以满足不同需要.

一般而言,对生成句子的要求可以归结为以下 3 种情况:

(1) 生成尽可能简短的句子.策略:仅使用 P^M 中的产生式进行推导.

(2) 追求语法现象的多样性,生成包括各种不同语法结构的复杂句子.策略:可以根据对句子长度或深度的限制,使用 P^K 中的产生式推导出满足条件的各种句型,然后使用 P^M 中的产生式对每个句型生成句子.

(3) 生成结构简单的长句.策略:循环使用 P^K 中的一条或几条产生式,生成结构简单的句型,再改用 P^M 中的产生式将句型变成句子.

同时,以上方法也可以是基于概率模型的,即如果对核心产生式赋予频度分布值,并据此选用核心产生式.那么所生成的句子既是长度和推导深度可控的,句子的结构也会更接近自然.

4 小 结

本文从 CFG 产生式的性质出发,研究了产生式在产生句子时的不同作用,给出了核心产生式、外围产生式以及类产生式等概念的定义.以此为基础提出一种对产生式集的划分方法,并证明其相关性质.利用这些性质得到一种新的句子生成方法,使得该方法具有时间和空间代价小,而且可以控制生成句子的长度和推导深度的特点.并且由与其通用性,该方法以及针对不同情况的策略还可以于其他方法相结合,从而很好地适应了生成句子时的不同需要.

参考文献

1 Solomaa A. Formal Languages. London: Academic Press, 1973

2 Maurer P M. Generating test data with enhanced context-free grammars. IEEE Software, 1990,7(4):50~55

3 Dong Yun-mei. Collection of SAQ Report No. 1~7. Technical Report, ISCAS-LCS-95-09. Laboratory of Computer Science, Institute of Software, The Chinese Academy of Sciences, 1995

4 Dong Yun-mei, Chen Hai-ming, Zhang Rui-ling. Collection of SAQ Report No. 8~16. Technical Report, ISCAS-LCS-96-1. Laboratory of Computer Science, Institute of Software, The Chinese Academy of Sciences, 1996

5 Zhang Rui-ling. Research on an interactive model of concept acquisition [Ph. D. Thesis]. Institute of Software, The Chinese Academy of Sciences, 1999

6 Wetherell C S. Probabilistic languages: a review and some open questions. ACM Computing Surveys, 1980,12(4):361~379

7 Hickey T, Cohen J. Uniform random generation of strings in a context-free language. SIAM Journal on Computing, 1983, 12(4):64~655

8 Dong Yun-mei. Recursive functions with define on CFL (I). SAQ Report No. 22, Technical Report, ISCAS-LCS-98-14, Laboratory of Computer Science, Institute of Software, The Chinese Academy of Sciences, 1998

Generating Sentences of CFL Based on Partition of CFG Production Set

WANG Hong-hao DONG Yun-mei

(*Laboratory of Computer Science Institute of Software The Chinese Academy of Sciences Beijing* 100080)

Abstract In this paper, a method is presented to partition productions of CFG (context-free grammar). It divides production set into two parts. The derivation with productions in one part will never terminate, while it must terminate rapidly with productions in the other part. It is proved that the procedure of generating sentences of CFL (context-free language) is using productions in one part to make the sentential form longer and more complex first, and then using productions in the other part to terminate the procedure. A general controllable method is attained for generating sentences of CFL with restricted length or depth. The time and space complexity for generating one sentence is $O(r+n)$, where n is the restricted length or depth of sentences and r is the number of productions in given CFG. The generating strategies for different conditions are also discussed.

Key words CFG (context-free grammar), production set, production set partition, sentence generation, sentence generating strategy.