О ВОЗМОЖНОСТИ ВЫЧИСЛИТЕЛЬНОГО РАСПОЗНАВАНИЯ КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК

К. В. Сафонов
Красноярский государственный технический университет, Россия
e-mail: safonov@fivt.krasn.ru

A subclass of affine context-free grammars is considered in the class of context-free grammars and a possibility of its computational recognition is investigated.

Введение

Понятия контекстно-свободного языка (кс-языка) и контекстно-свободной грамматики (кс-грамматики), порождающей этот язык, были введены Н. Хомским [1–3] во второй половине 50-х годов прошлого века при попытке построить адекватную математическую модель естественных языков, например английского языка. Вскоре было обнаружено, что некоторые классы языков программирования, в частности класс "языков типа АЛГОЛ", совпадают с классом контекстно-свободных языков. С тех пор исследования, посвященные контекстно-свободным грамматикам и порождаемым ими языкам, получили широкое развитие, и сегодня теория контекстно-свободных грамматик представляет собой центральную составную часть математической лингвистики. Контекстно-свободные грамматики оказались хорошо приспособленными для описания большинства конструкций естественных языков [4–7], хотя, по-видимому, большинство лингвистов считают, что для всеобъемлющего описания структуры естественных языков кс-грамматик все же недостаточно.

Рассмотрим конечное множество $X = \{x_1, \ldots, x_n\}$, состоящее из слов x_i языка, играющее роль словаря и называемое терминальным множеством, а также $Z = \{z_1, \ldots, z_m\}$ — множество вспомогательных символов z_j, необходимых для задания грамматических правил, называемое нетерминальным множеством. Обозначим $(X \cup Z)^*$ свободную полугруппу, порождаемую некоммутативными элементами $x_1, \ldots, x_n, z_1, \ldots, z_n$. Дополним операцию некоммутативного умножения коммутативной операцией формальной суммы "+" мономов из $(X \cup Z)^*$ (вместо суммы можно взять объединение "$\cup$", как в [7]), а также коммутативной операцией умножения мономов на (целые) числа. Таким образом, можно рассматривать не только многочлены, но и формальные степенные ряды с числовыми (как правило, целыми) коэффициентами от некоммутативных переменных.

Под кс-языком понимается [4, 5] первая компонента z_1 решения $(z_1(x), \ldots, z_m(x))$ системы полиномиальных уравнений

$$z_i = p_i(x, z), i = 1, \ldots, m, \tag{1}$$

которое получается методом последовательных приближений:

$$z_i^{(k+1)} = p_i(x, z^{(k)}), z_i^{(0)} = 0, i = 1, \ldots, m,$$

где $z^{(k)} = (z_1^{(k)}, \ldots, z_m^{(k)})$; 0 — нулевой моном, такой, что $0 \cdot u = u \cdot 0 = 0$ для любого монома u. Таким образом, в результате итераций кс-язык представляется формальным степенным рядом

$$z_1 = \sum_i < z_1, w_i > w_i, \tag{2}$$

где $< z_1, w_i >$ — числовой коэффициент, с которым моном w_i от некоммутативных переменных входит в ряд z_1. Мономы w_i являются грамматически правильными предложениями, которые могут быть построены в данном языке из слов $x_1, \ldots, x_n$ этого языка, а весь ряд (2), т. е. формальная сумма всех правильных предложений, и является данным кс-языком (построение системы уравнений (1) по грамматическим правилам языка и описание дополнительных условий, которым она удовлетворяет, приведены ниже в разд. 1).

Основной вопрос настоящей работы состоит в том, как по заданному формальному ряду определить, является ли он кс-языком, т. е. порожден ли он кс-грамматикой — некоторой полиномиальной системой уравнений вида (1). В исследованиях, посвященных алгоритмическим проблемам в классе кс-языков [4, 5], такой вопрос не рассмотрен. Эти исследования были направлены, например, на решение следующих проблем: по заданной кс-грамматике, т. е. по коэффициентам заданной системы уравнений (1), определить, содержит ли порождаемый ею кс-язык заданный моном с ненулевым коэффициентом (алгоритмически разрешимая проблема), порождают ли две заданные грамматики, а фактически две различные системы уравнений вида (1), один и тот же кс-язык (алгоритмически не разрешимая) и др. Несмотря на значительное число результатов, посвященных кс-языкам [4], какие-либо условия, полностью характеризующие эти языки (соответствующие формальные ряды), в настоящее время неизвестны. В случае коммутативных, например комплексных, переменных соответствующий ряд определяет голоморфную в окрестности начала координат алгебраическую функцию от переменных $x_1, \ldots, x_n$ и является ее тейлоровским разложением. Однако и в этой ситуации неизвестно (даже для тривиального с точки зрения математической лингвистики случая $n = 1$), как по коэффициентам ряда установить, является ли его сумма алгебраической функцией. В связи с этим представляют несомненный интерес условия, характеризующие некоторые классы кс-языков, в том числе и необходимые условия.

В настоящей работе развивается подход, при котором коэффициенты степенных рядов алгебраических функций располагаются на "диагональной" гиперплоскости в многомерной матрице тейлоровских коэффициентов соответствующих рациональных функций. Кратные степенные ряды рациональных функций значительно легче поддаются исследованию. В частности, возможно применение критерия Кронекера по каждой переменной кратного ряда: для рациональности суммы ряда

$$\sum_{k \geq 0} a_k x^k$$

необходимо и достаточно, чтобы, начиная с некоторого номера, все ганкелевы определители

$$H_n = \det(a_{i+j-2})_{i,j=1}^{n+1}$$

были равны нулю — для кратных рядов соответствующие условия в виде равенства нулю сумм определителей, составленных из коэффициентов кратного ряда, в принципе могут быть выписаны в явном виде. В случае некоммутативных переменных кратным степенным рядам рациональных функций соответствуют формальные степенные ряды, представляющие линейные языки [3–5], вследствие чего вычислительное распознавание таких языков принципиально возможно.

Таким образом, оказалось, что кс-языки тесно связаны с линейными языками, что, в свою очередь, дает принципиальную возможность для распознавания важного класса кс-языков. Для установления этой связи понадобилась информация о бесконечноудаленных корнях системы уравнений (1): в зависимости от наличия таких корней кс-языки делятся на два класса — аффинные и неаффинные кс-языки, имеющие различные свойства коэффициентов.

1. Грамматики и собственные системы уравнений

При моделировании структуры естественных языков общепринят следующий подход, предложенный Н. Хомским [3]. Среди вспомогательных символов выделяется начальный символ предложения, например z_1. Этот символ заменяется, согласно одному из заданных правил подстановки, мономом, который может состоять как из терминальных, так и из нетерминальных символов. Пусть, например, это подстановка $z_1 \to z_2 z_3$, где z_2 означает "группу подлежащего", а z_3 — "группу сказуемого". Далее применяются другие правила, например подстановки $z_2 \to z_4 z_5$, где z_4 означает "прилагательное", а z_5 — "существительное", и $z_4 \to x_i, z_5 \to x_j$, где x_i, x_j — некоторые терминальные символы — выбранные прилагательное и существительное из словаря. Аналогичные подстановки применяются к символу z_3. В результате применения правил подстановки получатся мономы только от переменных (слов) $x_1, \ldots, x_n$, являющиеся предложениями данного языка. Формальная сумма всех полученных мономов (предложений) и является данным формальным языком, который определен совокупностью правил подстановки, называемой грамматикой данного языка.

Заметим, что грамматика в приведенном выше примере характеризуется тем свойством, что в левой части любого правила подстановки стоит лишь один нетерминальный символ, т. е. правило действует независимо от окружения этого символа, другими словами, независимо от контекста (это объясняет название кс-грамматик).

Для построения грамматики, порождающей бесконечное множество предложений, необходимо ввести рекурсию, например, добавив к правилам подстановки правило $z_5 \to x_k z_1$, где z_5 — "существительное", x_k означает слово "что" или "который", а z_1 — начальный символ предложения. Очевидно, что теперь можно получать сколь угодно "длинные" мономы (сложноподчиненные предложения).

Далее правилам подстановки ставится в соответствие [3] система полиномиальных уравнений: каждому вспомогательному символу z_i, содержащемуся в левой части правил подстановки $z_i \to f_{i1}(x,z), \ldots, z_i \to f_{iq}(x,z)$, сопоставляется уравнение $z_i = p_i(x,z)$, где $p_i(x,z) = f_{i1}(x,z) + \ldots + f_{iq}(x,z)$. Таким образом, грамматике соответствует система полиномиальных уравнений (1), которая решается методом последовательных приближений в виде формального ряда (2).

Пусть вообще $< r, w >$ обозначает коэффициент, с которым моном w (от некоммутативных переменных) входит в формальный ряд или многочлен r. Порождающая способность

грамматики не уменьшится, если из правил подстановки исключить правила вида $z_i \to z_j$ и $z_i \to e$, где e — пустая цепочка (играющая роль единицы относительно умножения мономов: $ex = xe = x$), в связи с чем общепринято [5] рассматривать только так называемые собственные системы (1), правые части которых удовлетворяют условиям

$$< p_i, e >= 0, < p_i, z_j >= 0, i, j = 1, \ldots, m. \tag{3}$$

Пример. Рассмотрим собственную систему

$$z_1 = x_1 z_1 + x_1 z_2,$$
$$z_2 = x_2 z_2 x_3 + x_2 x_3,$$

для которой метод последовательных приближений дает решение $z^{(0)} = (0, 0)$, $z^{(1)} = (0; x_2 x_3)$, $z^{(2)} = (x_1 x_2 x_3; x_2^2 x_3^2 + x_2 x_3)$, $z^{(3)} = (x_1^2 x_2 x_3 + x_1 x_2^2 x_2^2 + x_1 x_2 x_3; x_2^2 x_3^2 + x_2^3 x_3^3), \ldots$ Нетрудно проверить, что первая компонента решения представляется рядом

$$r = \sum_{i,j \geq 1} x_1^i x_2^j x_3^j,$$

таким образом, данный кс-язык (над словарем $\{x_1, x_2, x_3\}$) состоит из предложений вида $x_1^i x_2^j x_3^j, i, j \geq 1$.

В случае комплексных переменных $z \in \mathbb{C}^m, x \in \mathbb{C}^n$ собственная система (1) имеет в окрестности нуля $(0, 0) \in \mathbb{C}^{m+n}$ единственное голоморфное решение $(z_1(x), \ldots, z_m(x))$, где $z_j(x)$ — алгебраические функции. Действительно, записывая систему (1) в виде

$$q_i(x, z) = z_i - p_i(x, z) = 0, i = 1, \ldots, m, \tag{4}$$

видим, что условия (3) равносильны тому, что

$$q_i(0, 0) = 0, i = 1, \ldots, m, \frac{\partial q_i(0, 0)}{\partial z_j} = \delta_{ij},$$

где δ_{ij} — символ Кронекера, следовательно, собственная система удовлетворяет условию теоремы о неявном отображении (метод последовательных приближений дает в этом случае тейлоровские разложения голоморфных функций $z_i(x), i = 1, \ldots, m$).

2. Аффинные кс-грамматики. Основной результат

Формальному степенному ряду (или многочлену) поставим в соответствие ряд (многочлен) с комплексными переменными, задав отображение терминальных x_i и нетерминальных z_j символов из множества $X \cup Z$ в множество комплексных переменных, за которыми оставляем прежние обозначения x_i и z_j соответственно, тогда $(x, z) \in \mathbb{C}_{x,z}^{n+m}$. Таким образом, получаем фиксированный гомоморфизм, который ставит в соответствие формальному ряду (многочлену)

$$r = \sum_i < r, w_i >$$

степенной ряд (многочлен) от комплексных переменных

$$ci(r) = \sum_k a_k x^k,$$

называемый его *коммутативным образом*, где $k = (k_1, \ldots, k_n)$, $a_k x^k = a_{k_1,\ldots,k_n} x_1^{k_1} \ldots x_1^{k_n}$, при этом

$$a_k = \sum_{\#x_1(w_i)=k_1,\ldots,\#x_n(w_i)=k_n} < r, w_i >,$$

где $\#\alpha(\beta)$ — число вхождений символа α в моном β. Заметим, что коммутативный образ кс-языка является функцией, голоморфной в непустой окрестности нуля, поскольку коэффициент при каждом мономе исходного формального ряда, представляющего кс-язык, равен числу выводов этого монома с помощью грамматических правил языка [3], а оценки этих чисел показывают, что радиусы поликруга сходимости его коммутативного образа положительны [8].

Рассмотрим систему полиномиальных уравнений (4), а также ее подсистему

$$q_i(x, z) = 0, i = 2, \ldots, m,$$

считая переменные x и z_1 параметрами, а корни этой системы $z[1] = (z_2, \ldots, z_m) = z[1](x, z_1)$ зависящими от этих параметров. При этом удобно рассматривать корни $z[1](x, z_1)$ в проективном пространстве $\mathbb{CP}^{m-1}_{z[1]}$, а параметры x, z_1 — в пространстве $\mathbb{C}^{n+1}_{x,z_1}$. В таких координатах система имеет вид

$$\xi_0^{\nu_i} q_i \left(x, z_1, \frac{\xi_2}{\xi_0}, \ldots, \frac{\xi_m}{\xi_0} \right) = 0, \ i = 2, \ldots, m, \tag{5}$$

где $\nu_i = \deg_{z[1]} q_i(x, z)$, а $[\xi_0 : \xi_2 : \ldots : \xi_m]$ — координаты проективного пространства $\mathbb{CP}^{m-1}$. При $\xi_0 = 1$ корни системы (5) расположены в пространстве $\mathbb{C}^{m-1}$ (аффинной части $\mathbb{CP}^{m-1}$), а при $\xi_0 = 0$ — на бесконечноудаленной гиперплоскости пространства $\mathbb{CP}^{m-1}$. Бесконечноудаленные корни системы (5) (при $\xi_0 = 0$) совпадают с корнями системы

$$q_i^*(x, z_1, \xi) = 0, i = 2, \ldots, m, \tag{6}$$

где $q_i^*(x, z)$ — старшая однородная составляющая многочлена $q_i(x, z)$ по переменной $z[1]$ с коэффициентами, зависящими от x и z_1, а $\xi = (\xi_2, \ldots, \xi_m)$.

Если многочлены последней системы не зависят от x и z_1 и она имеет только единственный корень $\xi_2 = \ldots = \xi_m = 0$ (при $\xi_0 = 0$ не определяющий точки в проективном пространстве), то будем называть *аффинным* кс-язык, определяемый системой уравнений (4), поскольку в этом случае соответствующая ей подсистема уравнений (5) не имеет бесконечноудаленных корней, следовательно, ее корни могут быть расположены лишь в аффинной части проективного пространства.

Нам необходимо установить, когда многочлен $P(x, z_1)$, определяющий алгебраическую функцию $z_1(x)$ — первую компоненту решения $(z_1(x), \ldots, z_m(x))$ системы (4), $P(x, z_1(x)) \equiv 0, P(0, 0) = 0$, обладает свойством $P'_{z_1}(0, 0) \neq 0$. Имеет место следующая

Лемма 1. *Пусть система* (6) *не зависит от* x *и* z_1 *и имеет единственный корень* $\xi = 0$. *Тогда можно выбрать многочлен* $P(x, z_1)$, *определяющий алгебраическую функцию* $z_1 = z_1(x)$, *такой, что*

$$P(0, 0) = 0, \frac{\partial P(0, 0)}{\partial z_1} \neq 0.$$

Доказательство. В условиях леммы система уравнений $q_i(x, z) = 0, i = 2, \ldots, m$, относительно $z[1]$ — собственная, т. е. при всех значениях параметров x, z_1 число корней системы $z^{(k)}[1] = z^{(k)}[1](x, z_1)$ с учетом кратностей μ_k — одно и то же (никакие из корней

“не уходят в бесконечность” при изменении x и z_1 на любом компакте), что дает возможность корректно определить результант этой системы относительно уравнения $q_1(x,z)$:

$$P(x,z_1)=\prod_k q_1^{\mu_k}(x,z_1,z^{(k)}[1](x,z_1)),$$

где в произведение входят все корни системы. В силу симметричной зависимости от корней, которые являются алгебраическими функциями, результант также представляет собой симметрическую алгебраическую функцию, следовательно, $P(x,z)$ — многочлен от x и z_1. Согласно одной теореме А.К. Циха [9, с. 238], кратность корня $(0,0)$ системы (4), равная 1, совпадает с кратностью корня $z_1=0$ результанта $P(x,z_1)$ при $x=0$. Лемма 1 доказана. □

В случае комплексных переменных все компоненты $z_i(x), i=1,\dots,m$, решения системы (1) являются алгебраическими функциями и каждая из них может быть задана с помощью одного полиномиального уравнения, тогда как в некоммутативном случае задать формальный ряд, например, для $z_1(x)$ с помощью одного полиномиального уравнения, вообще говоря, невозможно, поскольку в этом случае не всегда возможно исключить неизвестные из системы уравнений.

Пусть дан ряд

$$R(x_0,x)=\sum_{k_0,k} b_{k_0,k}x_0^{k_0}x^k$$

от $n+1$ комплексных переменных $x_0\in\mathbb{C}^1, x\in\mathbb{C}^m$, где $k=(k_1,\dots,k_m)$, тогда его диагональю назовем ряд

$$\Delta_{x_0,x_1}(R(x_0,x))=\sum_k b_{k_1,k}x^k.$$

Имеет место следующая

Лемма 2. *Для любого аффинного кс-языка $z_1(x)$ над терминальным множеством X существует линейный язык $L(x_0,x)$ над терминальным множеством $\{x_0\}\cup X$, такой, что*

$$ci(z_1(x))=\Delta_{x_0,x_1}(ci(L(x_0,x))).$$

Доказательство. Для алгебраической функции $ci(z_1(x))$ существует [10] рациональная функция $R(x_0,x)$, такая, что выполнено равенство

$$ci(z_1(x))=\Delta_{x_0,x_1}(R(x_0,x)),$$

например, можно взять

$$R(x_0,x)=\frac{z_1P'_{z_1}(z_1,x)}{P(z_1,x)}|_{z_1=x_0,x_1=x_0x_1}.$$

Из условия $P'_{z_1}(0,0)\neq 0$ следует, что $P(z_1,x)=cz_1+P_1$, тогда все мономы многочлена $P(x_0,x_0x_1,x_2,\dots,x_m)$ содержат x_0 в качестве множителя, поэтому, сокращая на него, получим, что R задается равенством $(c+P_1)R=x_0P'_{z_1}(x_0,x_0x_1,x_2,\dots,x_m)$ или равенством $R=\frac{1}{c}(-P_1R+x_0P')$, являющимся линейной грамматикой. Следовательно, выполняя последовательные приближения по линейной рекуррентной формуле

$$R_{n+1}=\frac{1}{c}(-P_1R_n+x_0P'),$$

получим формальный ряд, который в случае некоммутативных переменных определяет линейный язык $L(x_0, x)$, причем диагональю его коммутативного образа является коммутативный образ языка $z_1(x)$.

Если записать в явном виде коэффициенты степенных рядов голоморфных функций $ci(z_1(x))$ и $ci(L(x_0, x))$, то получится основной результат данной работы — следующая

Теорема. *Для любого аффинного кс-языка (1) над терминальным множеством X существует линейный язык*

$$L(x_0, x) = \sum_j < L, u_j > u_j$$

над терминальным множеством $\{x_0\} \cup X$, такой, что при всех $k_1, \ldots, k_m \geq 0$ выполнены равенства

$$\sum_{\#x_1(w_i)=k_1,\ldots,\#x_m(w_i)=k_m} < z_1, w_i >= \sum_{\#x_1(u_j)=k_1,\ldots,\#x_m(u_j)=k_m,\#x_0(u_j)=\#x_1(u_j)} < L, u_j > . \quad (7)$$

Теорема дает необходимое условие того, что данный формальный ряд $z_1(x)$ является аффинным кс-языком: если возможно показать, что ни для какой рациональной $L(x_0, x)$ функции равенство (7) или равенство

$$ci(z_1(x)) = \Delta_{x_0,x_1}(L(x_0, x))$$

не выполняется, то этот ряд не принадлежит классу аффинных кс-языков.

При выполнении равенств (7) естественно называть ряд $z_1(x)$ *коммутативной* (слабой) диагональю ряда $L(x_0, x)$. Если же выполнено равенство

$$\sum_i < z_1, w_i > w_i = \sum_{\#x_0(u_j)=\#x_1(u_j)} < L, u_j > u_j|_{x_0=e},$$

то назовем этот ряд *некоммутативной* (сильной) диагональю ряда $L(x_0, x)$. В заключении поставим в качестве нерешенного вопроса следующий: будет ли аффинным кс-языком всякий формальный ряд, который является некоммутативной (сильной) диагональю линейного языка?

Список литературы

[1] Хомский Н. Три модели для описания языка // Кибернетический сборник: Сб. перевед. статей. М.: Иностр. лит., 1961. Вып. 2. С. 237–266.

[2] Хомский Н. О некоторых формальных свойствах грамматик // Кибернетический сборник: Сб. перевед. статей. М: Иностр. лит., 1962. Вып. 5. С. 279–311.

[3] Хомский Н., Щютценберже М.П. Алгебраическая теория контекстно-свободных языков // Кибернетический сборник, нов. серия: Сб. перевед. статей. М.: Мир, 1966. Вып. 2. С. 121–230.

[4] Гинзбург С. Математическая теория контекстно-свободных языков. М.: Мир, 1970.

[5] Salomaa A., Soittola M. Automata-Theoretics Aspects of Formal Power Series. N.Y.: Springer-Verlag, 1978.

[6] Гладкий А.В. Формальные грамматики и языки. М.: Наука, 1973.

[7] Глушков В.М., Цейтлин Г.Е., Ющенко Е.Л. Алгебра, языки, программирование. Киев: Наук. думка, 1974.

[8] Семенов А.Л. Алгоритмические проблемы для степенных рядов и контекстно-свободных грамматик // Докл. АН СССР. 1973. Т. 212. С. 50–52.

[9] Айзенберг Л.А., Южаков А.П. Интегральные представления и вычеты в многомерном комплексном анализе. Новосибирск: Наука, 1979.

[10] Safonov K.V. On power series of algebraic and rational functions in C^n // J. of Math. Anal. and Appl. 2000. Vol. 243. P. 261–277.

Поступила в редакцию 30 января 2005 г.,
в переработанном виде — 1 апреля 2005 г.