通过推导树证明 Parikh 定理的最简单方法

Alexander Rubtsov ！ 俄罗斯国立高等经济大学 莫斯科物理技术学院

摘要

Parikh 定理是形式语言理论的一项基础性成果。此前已经发表了许多证明，也有许多论文声称提供了简化的证明，但其中大多数篇幅较长且仍然复杂。我们给出了一个真正简短、简单且揭示这一基础性成果本质的证明。我们采用了与 Parikh 原论文相近的技术；我们的证明类似于 Ryoma Sin'ya 于 2019 年给出的证明，但我们提供了更为详细的阐述，并力求进一步简化。我们通过非构造性方法实现了这种简洁性，从而避免了其他证明所遇到的许多困难。

2012年 ACM 分类 计算理论 → 文法与上下文无关语言

关键词与短语 形式语言、上下文无关语言、Parikh 定理

资助 本文得到俄罗斯基础研究基金会项目 20-01-00645 的资助

致谢 感谢 Dmitry Chistikov 对本文提供了有益的讨论

arXiv:2301.00047v1 [cs.FL] 30 Dec 2022

1 引言

Parikh 定理 [6] 是形式语言理论中的一条基本定理。已有许多证明发表（参见 [1], [4] ，其中综述了不同的证明并对该主题作了详细阐述）。许多论文声称给出了简化证明（[2] 似乎是其中最有名的），但其中大多数篇幅较长，仍然相当复杂。尽管早已有非常简短且简单的证明（例如 [7]），仍不断有论文发表其他证明 [5]。我们给出的证明确实简短、简单，并揭示了这一基本结果的本质。我们的证明与 Parikh 最初的证明 [6]一样，基于推导树。与 Parikh 一样，我们将推导树分解为小树；我们考虑类似的树，即普通推导树和辅助树（其树冠中只有一个与根相同的非终结符）。我们去除辅助树中的重复项，并对其进行泵送。总体而言，我们的方法与原始证明相似，但更简单，因为我们对文法没有限制，也没有其他技术问题（例如，只考虑包含某个固定子集中的非终结符的推导树）。我们构造的简洁性源于非构造性，这使我们能够避开许多技术问题。此外，我们构造中的树（在分解过程中）高度与非终结符的数量呈线性关系，而 Parikh 构造中的树高度则呈二次关系。在我们的证明中，我们将从字到树的推导这一思想加以推广，并将泵引理的思想也转化到树上。这使得我们的构造足够清晰，可以在形式语言与自动机理论基础课程的讲授中向学生解释，因此我们希望，详细版本有助于在学术界推广这一方法。

我们的技巧与 [8]类似。不同于 [8], 中简洁的论述，我们会更详细地阐述，并解释证明的直觉。由于非构造性，我们的证明不需要借助辅助结果。

2 定义

我们用 $\mathbb{N}$ 表示非负整数。设 $\Sigma_k = \{a_1, \ldots, a_k\}$ 是一个字母表，$k \geqslant 1$。我们用 $\Psi: \Sigma_k^* \to \mathbb{N}^k$ 表示 *Parikh* 映射，它将一个字 w 映射为其*Parikh* 像向量（$|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_k}$），其中 $|w|_{a_j}$ 表示 w中字母 a_j 的数量。我们以自然的方式表示语言 $L \subseteq \Sigma_k^*$ 的 Parikh 像:

$$\Psi(L) = \{\Psi(w) \mid w \in L\}.$$

集合 $S \subseteq \mathbb{N}^k$ 若存在向量 v_0、v_1、……、$v_m \in \mathbb{N}^k$ 使得

$$S = \{v_0 + v_1 t_1 + v_2 t_2 + \ldots v_m t_m \mid t_1, \ldots t_m \in \mathbb{N}\}.$$

集合若是有限多个线性集的并，则称其为半线性的。

▶**定理 1**（Parikh）。对于每个上下文无关语言 L，集合 $\Psi(L)$ 是半线性的。

我们对上下文无关文法 [3]使用经典记法。记非终结符的集合为 N，并用大写字母表示非终结符。用字母表末尾的小写字母表示字，希腊字母表示字母表 $\Sigma \cup N$上的字，称为句型。

我们的证明使用推导树，并借鉴了经典泵引理证明中的类似思路。我们不仅用推导树表示字，也用它表示一种特殊的句型。我们将对应于形如 $A \overset{*}{\Longrightarrow} uAv$ a 的推导的树称为块树。我们将字的推导树称为基本树。如果一棵树 T（基本树或块树）没有以块树为子树，也就是说，T 不具有 $(S \overset{*}{\Longrightarrow} xAz \overset{*}{\Longrightarrow} xuAvz \overset{*}{\Longrightarrow} xu\beta vz)$ 的形式，我们就称它为极小。下文中，我们将用对应于该树的任意推导来描述这棵树。

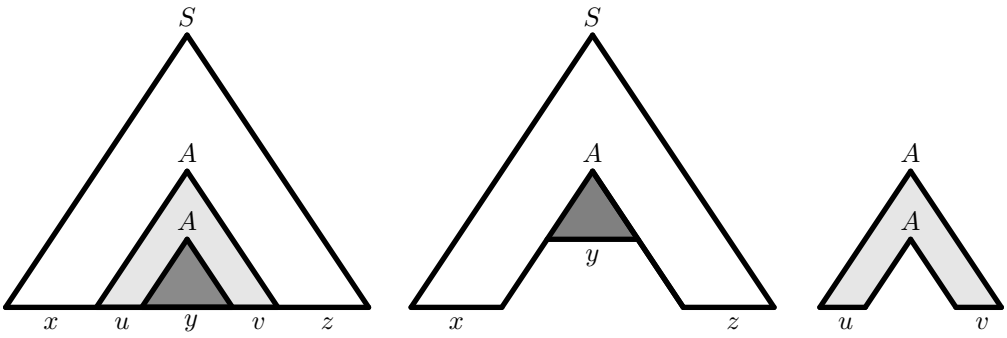


图 1 将树 T 分解为一对树 (T_1, T_2)

非极小基树具有 $S \overset{*}{\Longrightarrow} xAz \overset{*}{\Longrightarrow} xuAvz \overset{*}{\Longrightarrow} xuyvz$ 的形式。它可以分解 为基树 $S \overset{*}{\Longrightarrow} xAz \overset{*}{\Longrightarrow} xyz$ 和块树 $A \overset{*}{\Longrightarrow} uAv$，见图1。形式上，如果存在某个 $B \in N$：$T = A \overset{*}{\Longrightarrow} xBz \overset{*}{\Longrightarrow} xuBvz \overset{*}{\Longrightarrow} xu\beta vz$, $T_1 = A \overset{*}{\Longrightarrow} xBz \overset{*}{\Longrightarrow} x\beta z$ 和 $T_2 = B \overset{*}{\Longrightarrow} uBv$, 使得 T 分解为树对 (T_1, T_2)，我们就称之为分解；这里 β 要么是 A（如果 T 是块树），要么是 $\beta \in \Sigma^*$（如果 T 是基树，此时 A 是公理）。我们称 T 可以由 T_1 和 T_2 组合而成。注意，如果 T_1 有多个 B 节点，T_1和 T_2 可以通过多种方式组合。我们记作

$$T_1 \circ T_2 = \{T \mid T \text{ can be decomposed into } (T_1, T_2)\}$$

除了词的推导之外，我们还考虑如下树的推导 。

▶**定义 2.** 树的推导从一棵极小基树开始，每个推导步骤也会得到一棵基树。在一个推导步骤中，我们选择一个节点 A 和一棵极小块树 $A \overset{*}{\Longrightarrow} xAy$，用这棵块树替换所选节点；将所选节点的子树粘接到块树树冠中的 A 上。

因此，在树的推导中，极小基树起到公理的作用（可以有多棵），而替换 $A \to T_i$ ，其中 $T_i =$（$A \overset{*}{\Longrightarrow} xAz$），起到产生式规则的作用。

▶**定义 3.** 设 S 是一个树的多重集。若 S' 可由 S 导出，即 S' 可由 S 将两棵树 T_1 和 T_2 替换为一棵树 $T \in T_1 \circ T_2$而得到，我们称之为 导出 。我们将其记为 $S \text{‘} S'$。若存在某棵基树 T，使得 $S \overset{*}{\text{‘}} \{T\}$，则称树的多重集 S 是良构 的。

容易看出，如果 S 仅包含极小树且 $S \overset{*}{\vdash} \{T\}$，那么存在 T 的一个推导，使得 $S = \{T_0, T_1, \ldots, T_n\}$，其中 T_0 是基树，而 $T_i, i \geqslant 1$ 是在第 i 个推导步骤中选取的块树。

我们以自然的方式定义树的 Parikh 像。若 $T = (S \overset{*}{\Longrightarrow} w)$ 且 $\Psi(T)$，则 $\Psi(T) = \Psi(w)$; 若 $T = (A \overset{*}{\Longrightarrow} xAz)$，则 $= \Psi(xz)$。下面的引理可直接由定义得出。

▶**引理 4**。

- 若 $T, T' \in T_1 \circ T_2$，则 $\Psi(T) = \Psi(T') = \Psi(T_1) + \Psi($
- $T_2)$。若 $S \overset{*}{\text{‘}} \{T\}$ 且 $S \overset{*}{\text{‘}} \{T'\}$，则 $\Psi(T) = \Psi(T')$。

引理 4 意味着我们可以将 Parikh 映射的定义扩展到良构多重集： $\Psi(S) = \sum_{T \in S} \Psi(T)$.

根据鸽巢原理，每棵极小树的深度至多为 $|N| - 1$（否则，从根到叶节点的最长路径上就会有非终结符重复）。由于（对于固定的文法）树的每个节点的度数都有上界，因此极小树只有有限棵。我们将它们依次列出，并用 m 表示极小树的数量。因此，每个极小树的多重集 S 都有一个对应向量 $\vec{v} \in \mathbb{N}^m$ ，其中 $\vec{v}_i$ 是 S中 T_i 出现的次数。

3 证明

我们先介绍证明思路。由文法 G 推导出的任意字 w 都有某棵推导树 T_w ，并且 Ψ（T_w）$= \Psi$（w）。我们将 T_w 分解为由极小树组成的多重集 S_w （$S_w \overset{*}{\vdash} T_w$）。将 S_w 中的基树记为 T_g 。通过去除重复项，从 S_w 得到集合 S'_w 。注意， $S'_w \subseteq M_G$ ，其中 M_G 是文法 G 的极小树有限集合。因此，我们定义了一个具有有限陪域 2^{M_G} 的映射 $w \mapsto S'_w$ 。对于通过重复非基树从 S'_w 得到的任意多重集 S ，存在一棵推导树 T （ G 的），使得 $S \overset{*}{\vdash} T$（S 是良构的）。因此， Ψ（T） $\in \Psi$（L（G））。换言之，

$$\left(\Psi(S'_w) + \sum_{T' \in S'_w \setminus \{T_g\}} t_{T'} \cdot \Psi(T') \right) \in \Psi(L(G))$$

无论我们为每个 T' 选择哪个 $t_{T'} \in \mathbb{N}$。设 S 是一个带有基树 $T_g \in S$ 的良构多重集。记

$$\mathrm{Lin}(S) = \left\{ \Psi(S) + \sum_{T'_i \in S \setminus \{T_g\}} t_i \cdot \Psi(T'_i) \;\middle|\; t_i \in \mathbb{N} \right\}. \tag{1}$$

根据定义，集合 $\mathrm{Lin}(S)$ 是线性集。由于不同的集合 S'_w 只有有限多个，集合 $\Psi(L(G))$ 是有限个线性集 $\mathrm{Lin}(S'_w)$ 的并，因此根据定义它是半线性的。证明思路如图 2 所示

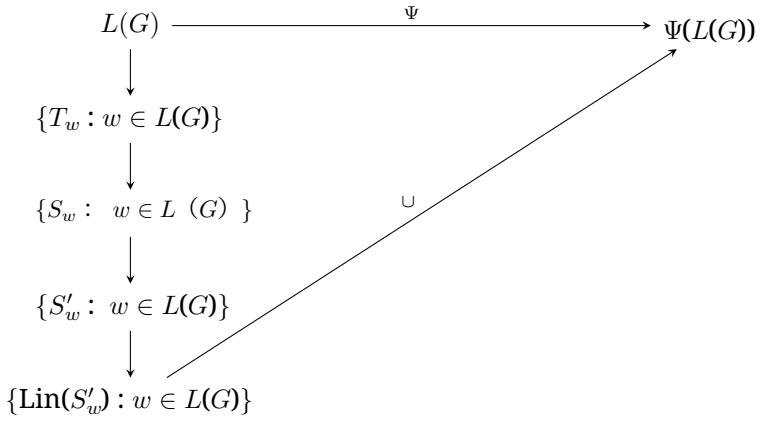


图 2 证明方案

3.1 辅助引理

▶**引理 5**。 设 S 是由极小树组成的良构多重集， S'' 是一个多重集。分别用 $\vec{v}$ 和 $\vec{v}''$ 表示 S 和 S''所对应的向量。假设 $\vec{v}_i > 0$ 当且仅当 $\vec{v}''_i > 0$；则 S'' 也是一个良构多重集。

证明。 该陈述等价于以下两个断言的合取。第一个断言是：如果 $\vec{v}_i > 0$并且 $\vec{v}''_i = \vec{v}_i + 1$ 而 $\vec{v}''_j = \vec{v}_j$ 对于 $j \neq i$，则 S'' 是良构的。第二个断言是：如果 $\vec{v}_i > 1$并且 $\vec{v}''_i = \vec{v}_i - 1$ 而 $\vec{v}''_j = \vec{v}_j$ 对于 $j \neq i$，则 S'' 是良构的。从向量 $\vec{v}$ 开始，随后将其分量增加或减少 1，我们可以得到满足该引理条件的任意向量 $\vec{v}''$。由于在每一步中两个断言之一的条件都成立，因此可知最终结果满足该引理的条件。

回顾关于树和多重集推导的定义 2 和定义 3。用 $T_i = (A \overset{*}{\Longrightarrow} uAv)$ 表示第 i 个块树。固定一棵推导树 T，使得 $S \overset{*}{\vdash} \{T\}$，并且树 T 也有一个推导。

第一个断言的证明。由于 $T_i \in S$ 的形式，树 T 包含非终结符 A。因此，我们可以将 T_i 拼接到非终结符 A 的某个出现位置，得到树 T'。根据构造， S'' 是良构的，因为 T' 是 G 的一棵推导树。

第二个论断的证明。如果 T_i 是 T 的一棵子树，那么可以将其从 T 中移除（如图1所示），所得的树 T' 也将是 G 的一棵推导树。否则，考虑 T 的固定推导过程。我们按如下方式修改这一推导过程。回顾 $\vec{v}_i > 1$，在多重集 S 中固定树 T_i 的两个副本 T_i^1 和 T_i^2，使得 T_i^1 被使用

早于推导 T 中的 T_i^2。当 T_i^2 必须与基树组合时，我们跳过这一步。如果在某一步中，一棵树 $T_j = (B \overset{*}{\Longrightarrow} u'Bv')$ 被粘合到 T_i^2 的一个节点 B 上，那么
我们将它粘接到 T_i^1 中相应的节点 B 上（T_i^1 已在基树中）。因此，这一修改产生了一个推导 $S \overset{*}{\vdash} \{T', T_i^2\}$，其中 T' 是某个基树。但随后，我们可以从 S'' 出发做同样的操作，并得到 $S'' \overset{*}{\vdash} \{T'\}$。因此，$S''$ 是良构的。◀

▶ **推论 6.** 对于任意 $w \in L(G)$: $\mathrm{Lin}(S'_w) \subseteq \Psi(L(G))$。

▶ **引理 7.** 对于任意 $w \in L(G)$:

$$\Psi(w) \in \mathrm{Lin}(S'_w).$$

证明。根据定义可知，$\Psi(w) = \Psi(S_w) \in \mathrm{Lin}(S_w)$。我们证明 $\mathrm{Lin}(S_w) \subseteq \mathrm{Lin}(S'_w)$。注意，$\Psi(S_w) = \Psi(S'_w) + \sum_{i=1}^{m}(\vec{v}_i - 1)\Psi(T_i)$，其中 $\vec{v}$ 是 S_w 对应的向量，而 T_i 是第 i 个最小树（按上述枚举顺序）。根据定义，向量 $\vec{u} \in \mathrm{Lin}(S_w)$具有如下形式

$$\vec{u} = \Psi(S_w) + \sum_{T'_j \in S_w \setminus \{T_g\}} t'_j \cdot \Psi(T'_j) = \Psi(S_w) + \sum_{i=1}^{n} t_i \cdot \Psi(T_i),$$

其中在第二个等式中，我们将所有作为同一棵树 T_i 的副本的 T'_j 合并在一起。注意，每个 T'_j 都是极小块树，因此对于某个 i，有 $T'_j = T_i$。于是，

$$\vec{u} = \left(\Psi(S'_w) + \sum_{i:\vec{v}_i > 0} (t_i + \vec{v}_i - 1)\Psi(T_i) \right) \in \mathrm{Lin}(S'_w) \quad ◀$$

3.2 Parikh 定理的证明

设 G 是生成 L 的上下文无关文法。对于一个字 $w \in L$，记 $T(w)$ 为所有推导树 T_w 的集合。注意，$T(w) \neq \varnothing$；如果 G 中有 ε-规则，它甚至可能是无限集。记 $S(w) = \{S_w \mid \exists T_w \in T(w) : S_w \overset{*}{\vdash} \{T_w\}\}$，其中 S_w 是一个只由极小树组成的多重集（如前所述）。最后，$S'(w) = \{S'_w \mid S_w \in S(w)\}$；回忆一下，$S'_w$ 是从 S_w 中删除重复项后得到的集合。现在我们证明，对于每个 w，$S'(w)$ 都是有限集，而且并集 $\cup_{w \in L} S'(w) = S'(L)$ 也是有限集。回忆一下，极小树只有有限棵，我们将它们记为 T_1，……，T_m。因此，每个 S_w 都对应一个 m 维向量 $\vec{v}$，使得

$$\Psi(w) = \Psi(S_w) = \sum_{i=1}^{m} \vec{v}_i \cdot \Psi(T_i).$$

$\vec{v}'$ 的 S'_w 对应的向量是一个 0-1 向量，满足：$\vec{v}'_i = 1$ 当且仅当 $\vec{v}_i > 0$。因此，由于长度为 m 的 0-1 向量只有有限多个，并且每个 $S'_w \in S'(L)$都有一个对应的 0-1 向量 $\vec{v}'$，所以集合 $S'(L)$也是有限的。

综合以上各点，根据引理 7

$$\Psi(L) \subseteq \bigcup_{S' \in S'(L)} \mathrm{Lin}(S')$$

并根据推论 6

$$\bigcup_{S' \in S'(L)} \mathrm{Lin}(S') \subseteq \Psi(L).$$

由于集合 $S'(L)$ 是有限的，且根据等式 (1)，每个集合 $\mathrm{Lin}(S')$ 都是线性的，因此等式

$$\Psi(L) = \bigcup_{S' \in S'(L)} \mathrm{Lin}(S')$$

证明了 Parikh 定理。 ◀

参考文献

1 Javier Esparza、Pierre Ganty、Stefan Kiefer 和 Michael Luttenberger。Parikh 定理：一种简单直接的自动机构造。《信息处理快报》，111(12):614–619，2011。网址：https://www.sciencedirect.com/science/article/pii/S0020019011000822, doi:https://doi.org/10.1016/j.ipl.2011.03.019。2 Jonathan Goldstine。Parikh 定理的简化证明。《离散数学》，19(3):235–239，1977。3 John E Hopcroft、Rajeev Motwani 和 Jeffrey D Ullman。自动机理论、语言与计算导论。ACM SIGACT News，32(1):60–65，2001。4 Caleb Koch。Parikh 定理友好导览。2018。网址：http://cakoch10.github.io/papers/Kleene_Algebra.pdf。5 Toshihiro Koga。通过狄克森引理证明 Parikh 定理。《国际计算机科学基础杂志》，32(02):163–173，2021。doi:10.1142/S012905412150009X。6 Rohit J. Parikh。论上下文无关语言。《ACM杂志》，13(4):570–581，1966年10月。doi:10.1145/321356.321364。7 Jeffrey O. Shallit。《形式语言与自动机理论进阶》。剑桥大学出版社，2008。8 Ryoma Sin'ya。Takahashi 风格的 Parikh 定理简明证明。CoRR，abs/1909.09393，2019。