广义自动机的 Myhill-Nerode 定理：模式匹配与压缩的应用

Nicola Cotumaccio
GSSI，意大利 达尔豪斯大学，
加拿大

arXiv:2302.06506v2 [cs.FL] 10 Jan 2024

摘要

Eilenberg 于 1974年 引入的广义自动机模型允许通过使边不仅可以标记字符，还可以标记字符串，以比传统自动机更简洁地表示正则语言。Giammaresi 和 Montalbano 为广义自动机引入了确定性的概念 [STACS 1995]。虽然广义确定性自动机保留了传统确定性自动机的许多性质，但最小广义确定性自动机的唯一性不复存在。

在论文的第一部分中，我们通过引入与广义自动机 $\mathcal{A}$ 相关联的集合 $\mathcal{W}(\mathcal{A})$，说明了这种非唯一性产生的原因。如果 $\mathcal{A}$ 是传统自动机，则集合 $\mathcal{W}(\mathcal{A})$ 总是平凡地等于该自动机所识别语言的所有前缀的集合，但对于广义自动机而言则未必如此。通过固定 $\mathcal{W}(\mathcal{A})$，我们首次得以推导出广义自动机的完整 Myhill-Nerode 定理，其中包含传统自动机的教科书式 Myhill-Nerode 定理这一退化情形。

在论文的第二部分，我们证明集合 $\mathcal{W}(\mathcal{A})$ 可应用于模式匹配和数据压缩。Wheeler 自动机 [TCS 2017、SODA 2020] 是一类广受欢迎的自动机，可通过 $e\log\sigma(1+o(1))+O(e)$ 比特进行紧凑存储（e 为边的数量，σ 为字母表的大小），并能在 $\tilde{O}(m)$ 时间内完成模式匹配查询（m 为模式串的长度）。在论文中，我们展示了如何将这些结果推广到广义自动机。更确切地说，Wheeler 广义自动机可通过 $\mathfrak{e}\log\sigma(1+o(1))+O(e+rn)$ 比特进行存储，从而能在 $O(rm)$ 时间内完成模式匹配查询，其中 $\mathfrak{e}$ 为所有边标签的总长度，r 为边标签的最大长度，n 为状态的数量。

2012年 ACM 学科分类 计算理论 → 正则语言；计算理论 → 模式匹配；计算理论 → 数据压缩

关键词与短语 广义自动机、Myhill-Nerode 定理、正则语言、Wheeler 图、FM 索引、Burrows-Wheeler 变换

资助 本研究部分由 Dante Labs 资助。

1 引言

正则语言类可以从非确定有限自动机（*NFA*）出发定义。在关于自动机理论的里程碑式著作中 [17] （可追溯至1974年），Eilenberg 提出了对 NFA 的一种自然推广：边不仅可以用字符标记，也可以用（可能为空的）有限字符串标记，这种自动机称为广义非确定有限自动机（*GNFA*）。尽管传统自动机只是广义自动机的一个特例，但不难看出，广义自动机只能识别正则语言，因为众所周知，ε 转移不会增强表达能力 [31], ，而字符串标签的边可以分解为一条由仅带字符标签的边构成的路径。不过，与传统自动机相比，广义自动机能够以更简洁的方式表示正则语言。衡量正则语言复杂度的一种标准方法，是识别该语言的某个自动机所需的最少状态数，而广义自动机的状态数可能少于传统自动机。对于广义自动机，我们假设状态数和边数都是有限的，但边数无法由状态数和有限字母表大小的某个函数加以界定（因此，边标签的长度可能任意大）。因此，原则上尚不清楚，确定识别给定语言的某个广义自动机的最少状态数这一问题是否可判定。在 [29], Hashiguchi 证明了该问题是可判定的：他证明，必定存在一个状态数最少的广义自动机，其边标签长度可以由一个仅取决于识别该语言的句法幺半群大小的函数加以界定。

如果没有任何状态拥有两条标签相同的不同出边，那么 NFA 就是确定有限自动机（*DFA*）。这一局部确定性概念可以扩展为全局确定性，也就是说，给定字符串 α，从初始状态出发，至多存在一条标签为 α 的路径可供遍历。然而，对于广义自动机来说，情况并非如此（见图 1）。考虑广义自动机时，我们必须增加一项要求：任何状态都不能拥有两条不同的出边，使得其中一条边的标签是另一条边标签的前缀。加入这一要求后，我们便恢复了全局确定性，从而得到广义确定有限自动机（*GDFAs*）。

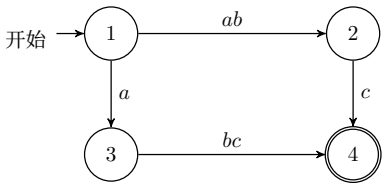


图 1 没有任何状态拥有两条标签相同的不同出边，但从初始状态出发，存在两条标签为 abc 的不同路径。

对于每种正则语言，都存在一个唯一的确定自动机能够识别该语言，且其状态数在所有能够识别该语言的确定自动机中最少；这个自动机称为该语言的最小确定有限自动机。更一般地，自动机理论中的经典教科书结论是*Myhill-Nerode* 定理。令 Pref($\mathcal{L}$) 表示所有至少作为语言 $\mathcal{L}$ 中一个字符串前缀的字符串所组成的集合。我们有以下结论。

▶ **定理 1**（Myhill-Nerode 定理）。 设 $L \subseteq \Sigma^*$。以下各项等价：
1. $\mathcal{L}$ 可由非确定有限自动机识别。2.
Myhill-Nerode 等价关系 $\equiv_{\mathcal{L}}$ 具有有限指标。

3.存在一个右不变等价关系 $\sim$ 定义在 $\mathrm{Pref}(\mathcal{L})$ 上，其指标有限，并且$\mathcal{L}$ 是若干个 $\sim$-类的并集。
4. $\mathcal{L}$ 可由确定有限自动机识别。
此外，如果上述某个陈述成立（因此上述所有陈述都成立），那么存在一个唯一的最小确定有限自动机识别 L（也就是说，所有识别 L 的确定有限自动机中状态数最少的两个确定有限自动机必定同构）。

Giammaresi 和 Montalbano [27, 26] 探讨了广义自动机框架下确定性的概念。可以自然地将同构的概念扩展到广义确定有限自动机（直观来说，两个广义确定有限自动机同构，是指它们仅在状态命名上有所不同）。由此自然产生的问题是，能否类似地定义正则语言的最小广义确定有限自动机。答案是否定的：一般而言，识别给定正则语言的状态数最少的广义确定有限自动机可能有两个或更多个，且彼此不同构。请看图2中的两个不同的广义确定有限自动机。显然，这两个不同构的广义确定有限自动机识别的是同一语言；并且可以证明，不存在状态数少于三个、却能识别同一语言的广义确定有限自动机 [27]。



图2 识别同一正则语言的两个状态数最少的广义确定有限自动机。

状态数最少的广义确定有限自动机并不唯一，这似乎意味着广义自动机与传统自动机的行为存在重大差异，因此看来我们无法在广义自动机模型中推导出类似 Myhill-Nerode 定理的结构性结果。自然会让人思考，这种不唯一性应被视为广义自动机模型的一种弱点，还是某种更深层性质的结果。考虑一个识别语言$\mathcal{L}$的传统自动机$\mathcal{A}$ 。按照自动机理论中的惯例，我们可以假设所有状态都能从初始状态到达，并且每个状态要么是终态，要么可以到达某个终态。于是，从初始状态出发并到达某个状态时所能读取的所有字符串组成的集合$\mathcal{W}(\mathcal{A})$，恰好等于前缀集 Pref($\mathcal{L}$)。然而，在广义自动机模型中，这一结论通常不再成立：通常我们有的不是$\mathcal{W}(\mathcal{A}) = \mathrm{Pref}(\mathcal{L})$，而只有$\mathcal{W}(\mathcal{A}) \subseteq \mathrm{Pref}(\mathcal{L})$。例如，考虑图2。在两个自动机中，我们都有 $a^3 \in$Pref($\mathcal{L}$)，但只有对于左侧的广义确定有限自动机，才有 $a^3 \in \mathcal{W}(\mathcal{A})$。

给定 $\mathcal{W} \subseteq$Pref($\mathcal{L}$)，我们称识别 $\mathcal{L}$ 的 GNFA $\mathcal{A}$ 为 $\mathcal{W}$-GNFA，若 $\mathcal{W}(\mathcal{A}) = \mathcal{W}$。我们将证明，如果 $\mathcal{L}$ 可由 $\mathcal{W}$-GDFA 识别，那么就存在一个识别 $\mathcal{L}$的唯一 状态数最少的$\mathcal{W}$-GDFA。特别地，我们的结果将推出标准 DFA 的最小自动机是唯一的，因为对于 DFA，必然有 $\mathcal{W}$ =Pref($\mathcal{L}$。

实际上，我们将证明更多内容。一旦固定$\mathcal{W}$，非确定性与确定性仍具有相同的表达能力，并且可以用等价关系来刻画。换言之，我们将证明一个广义自动机的 *Myhill-Nerode* 定理。为此，我们将引入局部有界集 的概念（定义8），并利用它证明以下结果。

▶ **定理 2**（广义自动机的 Myhill-Nerode 定理）。 设 $\mathcal{L} \subseteq \Sigma^*$，并设 $\mathcal{W} \subseteq \Sigma^*$是一个局部有界集，使得 $\mathcal{L} \cup \{\epsilon\} \subseteq \mathcal{W} \subseteq \mathrm{Pref}(\mathcal{L})$。以下命题等价：1. $\mathcal{L}$ 可由一个$\mathcal{W}$-*GNFA* 识别。

2.*Myhill-Nerode* 等价关系 $\equiv_{\mathcal{L},\mathcal{W}}$ 具有有限指标。**3.**存在一个定义在 $\sim$ 上的右不变等价关系 $\mathcal{W}$，其指标有限，并且 $\mathcal{L}$ 是若干个 $\sim$-等价类的并集。**4.** $\mathcal{L}$ 可由一个$\mathcal{W}$-*GDFA* 识别。

此外，如果上述某一条陈述成立（因而上述所有陈述都成立），那么就存在唯一的、识别 L 的最小 *W-GDFA*（也就是说，在所有识别 L 的 *W-GDFA* 中，状态数最少的两个 *W-GDFA* 必定同构）。

特别地，我们将证明，假设$\mathcal{W} \subseteq \Sigma^*$是一个局部有界集，且$\mathcal{L} \cup\{\epsilon\} \subseteq \mathcal{W} \subseteq \text{Pref}(\mathcal{L})$并不失一般性，因为这些是存在$\mathcal{W}$-GNFA 的必要条件。由此我们得出结论：我们的 GNFA 的 Myhill-Nerode 定理为广义自动机模型提供了首个结构性结果。

在论文的第二部分，我们展示集合$\mathcal{W}(A)$为标记图中的字符串匹配（*SMLG*）问题带来了新的视角。SMLG 问题有着引人入胜的历史，可追溯到 30 多年前。粗略地说，SMLG 问题可定义如下：给定一个节点带有非空字符串标签的有向图，以及一个模式串，判断是否可以沿图中的一条路径前进并拼接节点标签来读取该模式串。SMLG 问题是文本上的经典模式匹配问题（即判断某个模式串是否出现在文本中）的自然推广，因为文本可以看作由单一路径构成的图。使用 Knuth-Morris-Pratt 算法 [34]，可以在 $O(n+m)$ 时间内高效解决文本上的模式匹配问题（n 为文本长度，m 为模式串长度）。SMLG 问题更具挑战性，其复杂度可能受到所考虑的模式匹配问题具体变体或问题所限制的图类别影响。例如，在允许图中存在错误的（近似）变体中，该问题会变为 NP 困难 [5]，因此通常只允许模式串中出现错误。SMLG 问题在九十年代得到了广泛研究 [35, 1, 39, 5, 41, 37]；Amir 等人展示了如何在 $O(\mathfrak{e}+me)$ 时间内解决任意图上的（精确）SMLG 问题 [5]，其中 e 是图中的边数，m 是模式串长度，$\mathfrak{e}$ 是图中所有标签的总长度。近年来，SMLG 问题再次成为关注焦点。Equi 等人 [20, 18, 19] 证明，对于任意图，对每个 $\epsilon > 0$，除非正交向量假设不成立，否则 SMLG 问题无法在 $O(me^{1-\epsilon})$ 或 $O(m^{1-\epsilon}e)$ 时间内解决。在应用中（尤其是在生物信息学中），我们往往需要更快的算法，因此 SMLG 问题被限制在能够更高效求解的图类别上。例如，*Elastic Founder* 图 可用于表示多序列比对（MSA），这是生物进化的一个核心模型；在若干对模型一般性影响有限的假设下，可以在线性时间内解决 Elastic Founder 图上的 SMLG 问题 [22, 21, 42]。

文本上的模式匹配问题因 Burrows-Wheeler 变换 [10] 和 FM 索引 [23, 24], 的发明而发生了变革；它们能够在压缩 文本上高效地执行模式匹配查询，从而在生物信息学领域确立了一种新范式（海量基因组数据的激增要求开发空间高效算法） [43]。近年来，这些思想被推广到了非确定有限自动机。特别是，*Wheeler* 非确定有限自动机是一类广受关注的自动机，在这类自动机上，SMLG 问题可以在线性时间内求解，同时只需存储 Wheeler 非确定有限自动机的紧凑表示 [25, 3]。de Bruijn 图是 Wheeler 非确定有限自动机的一种特殊情况 [9], ，可用于执行欧拉序列组装 [32, 40, 7]。Wheeler 非确定有限自动机在理论上也具有重要意义：例如，对一个 Wheeler 非确定有限自动机应用幂集构造，得到的等价确定有限自动机的状态数仅线性增长，且该等价确定有限自动机也是 Wheeler [4]；而对于任意非确定有限自动机，

状态数可能呈指数级增长。

缺失的步骤是确定是否可以将 Burrows-Wheeler 变换和 FM 索引推广到广义非确定有限自动机，从而使所得数据结构也能应用于 Elastic Founder 图以及标签可以是任意字符串的其他图类。事实上，在数据压缩中，通常会考虑边标记图，其中将图中的一元路径压缩以节省空间，并用一条标记为所有标签拼接结果的单条边替代该路径。例如，使用这种机制存储的一些常见数据结构包括 Patricia 树、后缀树和泛基因组 [38, 6, 36]。下文中，我们只考虑不含 ϵ-转换的广义非确定有限自动机（即不存在任何边标记为空串 ϵ 的广义非确定有限自动机），因为一般而言，广义非确定有限自动机可能包含由相邻的 ϵ-转换组成的任意长路径，因此 SMLG 问题会变得更加困难，而我们基于 FM 索引的机制也会失效。如果所有边标签的长度至多为 r，我们就称广义非确定有限自动机为 $r\text{-}GNFA$（因此，当且仅当一个广义非确定有限自动机是 1-GNFA 时，它才是传统非确定有限自动机）。令 m、e 和 $\mathfrak{e}$ 如上，令 n 表示状态数，并令 $\sigma = |\Sigma|$。在第 4 节中，我们将把 Wheeler 性质的概念推广到广义非确定有限自动机。关键要素将是我们在 Myhill-Nerode 定理中使用的同一个集合 $\mathcal{W}(\mathcal{A})$：我们将考虑一个偏序 $\preceq_{\mathcal{A}}$，按照 $W(A)$ 中字符串的 *共字典序* 对所有状态集合进行排序（见定义 21）。随后我们将证明以下结果。

▶ **定理 3**（广义自动机的 FM 索引）。设 A 是一个 *Wheeler* r-广义非确定有限自动机。于是，我们可以使用 $\mathfrak{e}\log\sigma(1+o(1)) + O(e+rn)$ 比特对 A 进行编码，使得之后给定一个长度为 m 的模式串 $\alpha \in \Sigma^*$，我们可以在 $O(rm(\log r + \log\log\sigma))$ 时间内解决 A 上的 *SMLG* 问题。在相同的时间界内，我们还可以判定 $\alpha \in L(A)$。

如果 $r = 1$（也就是说，如果 $\mathcal{A}$ 是一个常规的 Wheeler 非确定有限自动机），我们便可得出结论：可以使用 $e\log\sigma(1+o(1)) + O(e)$ 比特对 $\mathcal{A}$进行编码（我们假设每个状态都可从初始状态到达，因此 $e \geq n-1$），从而能在 $O(m\log\log\sigma)$ 时间内解决 SMLG 问题，也就是说，得到了 Wheeler 自动机已有的时间和空间界[25, 14]。如果 $r = O(1)$，我们仍然可以在线性时间内进行模式匹配查询（对于常数字母表），从而突破 Equi 等人给出的下界，同时只存储 $\mathcal{A}$ 的紧凑表示。

由于篇幅所限，部分证明和一些辅助结果可见附录。

2 预备定义

令 Σ 为一个有限字母表。我们用 Σ^* 表示在 Σ上所有有限字符串的集合。我们用ϵ 表示空串，而 $\Sigma^+ = \Sigma^* \setminus \{\epsilon\}$是 Σ上所有非空有限字符串的集合。对于$i \geq 0$，令 $\Sigma^i \subseteq \Sigma^*$ 表示所有长度为 i 的字符串的集合（我们通常将 Σ^i 视为一个新字母表）。如果 L $\subseteq \Sigma^*$，则令 Pref(L) 表示 L 中某个字符串的所有前缀的集合。注意，如果 L $\neq \emptyset$，那么 $\epsilon \in$ Pref(L)。我们称 L $\subseteq \Sigma^*$ 是前缀无关的，如果 L 中没有字符串是 L 中另一个字符串的真前缀。注意，如果 L 是前缀无关的且 $\epsilon \in$ L，那么 L $= \{\epsilon\}$。如果 L $\subseteq \Sigma^*$，则 L 的前缀无关核是集合 K(L)，它由 L 中所有真前缀均不属于 L 的字符串组成。注意，K(L) 始终是前缀无关的，并且当且仅当 L =K(L) 时，L 是前缀无关的。

回顾一下广义确定性有限自动机（GDFA）[26, 27] 的定义。

▶ **定义 4.** 广义非确定有限自动机（GNFA）是一个四元组 $A = (Q, E, s, F)$，其中 Q 是一个有限状态集，$E \subseteq Q \times Q \times \Sigma^*$ 是一个有限的字符串标记边集合，$s \in Q$ 是初始状态，$F \subseteq Q$ 是终态集合。此外，

6 广义自动机的 Myhill-Nerode 定理及其应用

我们假设每个 $u \in Q$ 都可从初始状态到达，且是共可达的，也就是说，它要么是终态，要么能够到达某个终态。

广义确定有限自动机（*GDFA*）是一种广义非确定有限自动机（*GNFA*），满足：对于每个 $u \in Q$，没有任何离开 u 的边标记为 ϵ；离开 u 的不同边具有不同的标签；并且所有标记离开 u 的某条边的字符串构成前缀无关集。

在自动机理论中，假设每个状态都是可达且共可达的，这是标准做法，因为所有不满足此要求的状态都可以移除，而不改变识别语言。传统非确定有限自动机（分别地，确定有限自动机）是指所有边都标记为来自 Σ的字符的广义非确定有限自动机（分别地，广义确定有限自动机）。请注意，我们明确要求广义非确定有限自动机具有有限条边（在传统非确定有限自动机中，状态数有限就自动意味着边数有限，因为字母表是有限的）。如果允许广义非确定有限自动机有无限多条边，那么任何非空语言（可能不是正则语言）都可以由一个只有两个状态的广义非确定有限自动机识别：第一个状态是初始状态，第二个状态是终止状态，所有边都从第一个状态指向第二个状态，且当且仅当某个字符串属于该语言时，该字符串才标记一条边。要求广义非确定有限自动机只有有限条边后，所识别的语言类恰好是正则语言类，因为可以很容易地将广义非确定有限自动机转换为一个识别相同语言、带有 ϵ转换的非确定有限自动机，具体做法如下：对于每条边$(u', u, \rho) \in E$，其中 $\rho = r_1, \ldots, r_{|\rho|} \in \Sigma^+$，且 $r_1, \ldots, r_{|\rho|} \in \Sigma$ 和 $|\rho| \geq 2$，删除边(u', u, ρ)，添加 $|\rho| - 1$ 个新状态 $z_1, \ldots, z_{|\rho|-1}$，然后添加边$(u', z_1, r_1)$、$(z_1, z_2, r_2)$、$\ldots$、$(z_{|\rho|-1}, u, r_{|\rho|})$（这些新状态中没有任何一个会被设为初始状态或终止状态）。

下面介绍一些对本文有用的记号。

▶ **定义 5.** 设 $A = (Q, E, s, F)$ 是一个广义非确定有限自动机。

1. 对于每个 $\alpha \in \Sigma^*$，令 I_α 为这样一组状态：从初始状态出发，沿着标签串联后得到 α的边即可到达这些状态。换言之，对于某个 $t \geq 0$，存在边 (s, u_1, α_1)、(u_1, u_2, α_2)、(u_2, u_3, α_3)……、(u_{t-1}, u_t, α_t)，使得 $\alpha = \alpha_1\alpha_2\alpha_3 \ldots \alpha_t$。*2.* 令 *L(A)* 为 *A* 识别的语言，即 *L(A)* $= \{\alpha \in \Sigma^* \mid I_\alpha \cap F \neq \emptyset\}$。*3.* 对于每个 $u \in Q$，令 I_u 为从初始状态到达 u时，串联边标签所能读出的所有字符串的集合，即 $I_u = \{\alpha \in \Sigma^* \mid u \in I_\alpha\}$。注意，对于每个 $u \in Q$，都有 $\emptyset \subsetneq I_u \subseteq Pref(L(A))$，因为每个状态都是可达且共可达的。*Wh*

en $\mathcal{A}$ is not clear from th 最终 *text,* 我们可以 *ite $I_\alpha^{\mathcal{A}}$ and $I_u^{\mathcal{A}}$.*

最后，在介绍完论文之后，我们可以自然地定义 m

广义非确定有限自动机的

▶ **定义 6.** 设 *A* 是一个广义非确定有限自动机。广义非确定有限自动机的标记图中的字符串匹配（SMLG）问题定义如下：构建一个编码 *A* 的数据结构，使得给定一个字符串 α，我们能够高效地计算所有通过以 α 为后缀的路径到达的状态的集合。

3 广义自动机的 Myhill-Nerode 定理

传统自动机的 Myhill-Nerode 定理（定理 1）给出了 Pref($\mathcal{L}$) 必须满足的一些代数性质，才能使 $\mathcal{L} \subseteq \Sigma^*$ 成为正则语言。直观而言，Pref($\mathcal{L}$) 能够刻画正则语言的正则性（也就是说，正则语言的代数刻画与自动机刻画之间的联系），其原因在于：给定一个识别 $\mathcal{L}$ 的非确定有限自动机 $\mathcal{A} = (Q, E, s, F)$，我们有 $\bigcup_{u \in Q} I_u = \text{Pref}(\mathcal{L})$，因为如果$\alpha \in$Pref($\mathcal{L}$)，就可以从初始状态开始在 $\mathcal{A}$ 上读取 α。然而，更一般地说

如果 $\mathcal{A} = (Q, E, s, F)$ 是一个识别 $\mathcal{L}$ 的广义非确定有限自动机，那么我们只能得到 $\bigcup_{u \in Q} I_u \subseteq \text{Pref}(\mathcal{L})$，因为如果 $\alpha \in \text{Pref}(\mathcal{L})$，就可以从初始状态开始在 $\mathcal{A}$ 上读取 α；但如果最后一条边的标签是长度至少为 2 的字符串，就可能出现我们只读完了该标签的一个真前缀的情况。

下面给出如下定义。

▶ **定义7.** 设 $A = (Q, E, s, F)$ 是一个广义非确定有限自动机。定义：

$$\mathcal{W}(\mathcal{A}) = \bigcup_{u \in Q} I_u.$$

我们称 A 为 $W(A)$-广义非确定有限自动机。

注意，对于每个 $\alpha \in \Sigma^*$，$I_\alpha \neq \emptyset$ 当且仅当 $\alpha \in \text{W(A)}$。此外，$\text{L(A)} \cup \{\epsilon\} \subseteq \text{W(A)} \subseteq \text{Pref(L(A))}$，因为 (i) 若 $\alpha \in \text{L(A)}$，则 $I_\alpha \cap F \neq \emptyset$，特别是$\alpha \in \text{W(A)}$；(ii) $\epsilon \in I_s$；(iii) 对每个 $u \in Q$，$I_u \subseteq \text{Pref(L(A))}$。下面证明 W(A) 的一个附加性质。取 $\alpha \in \text{W(A)}$，并考虑集合 $T_\alpha = \{\rho \in \Sigma^+ \mid \alpha\rho \in \text{W(A)}\}$。考虑其前缀无关核 $\text{K}(T_\alpha)$。若 $\rho \in \text{K}(T_\alpha)$，则 $I_{\alpha\rho} \neq \emptyset$；但对于每个作为 ρ 的严格非空前缀的 $\rho' \in \Sigma^+$，有 $I_{\alpha\rho'} = \emptyset$。这意味着 $|\rho| \leq r$，其中 r 是边标签长度的最大值。由此可知，$\text{K}(T_\alpha)$ 必为有限集，因为Σ 是有限的。由此引出以下定义。

▶ 定义*8*。*1.* 设 $W \subseteq \Sigma^*$。如果对于每个 $\alpha \in W$，都有 $K(T_\alpha)$ 是有限的，其中 $T_\alpha = \{\rho \in \Sigma^+ \mid \alpha\rho \in W\}$，则称 W 是局部有界的。*2.* 设 $A = (Q, E, s, F)$ 是一个广义非确定有限自动机，并设 $W \subseteq \Sigma^*$ 是一个局部有界集，使得 $L(A) \cup \{\epsilon\} \subseteq W \subseteq Pref(L(A))$。如果 $W(A) = W$，则称 A 是一个 W-广义非确定有限自动机。

▶ 注 9。设 $\text{A} = (Q, E, s, F)$ 是一个广义确定有限自动机。设 $\alpha \in \text{W(A)}$。如果 $\alpha = \epsilon$，则 $I_\epsilon = \{s\}$，因为没有边标记为 ϵ。如果 $|\alpha| > 1$，则存在一个 α 的前缀 $\alpha_1 \in \Sigma^*$ 和$u_1 \in Q$，使得 $(s, u_1, \alpha_1) \in E$；由于 A 是一个广义确定有限自动机，我们有 $\alpha_1 \in \Sigma^+$，并且 α_1和 u_1 都是唯一的。特别地，$\alpha_1 \in \text{W(A)}$。如果 α_1 是 α的真前缀，我们可以从 u_1开始重复上述论证。由此可知，对于每个 $\alpha \in \text{W(A)}$，都有 $|I_\alpha| = 1$。因此，如果 u和$v \in Q$ 不同，则 $I_u \cap I_v = \emptyset$。下文中，如果 A 是一个广义确定有限自动机且$\alpha \in \text{W(A)}$，我们将把 I_α 与作为其唯一元素的状态视为同一对象。

此外，我们的论证表明，如果$\mathcal{A}$ 是一个广义确定有限自动机，那么对于每个 $\alpha \in \mathcal{W}(\mathcal{A})$ 使得 $|\alpha| > 0$，α 在$\mathcal{W}(\mathcal{A})$中的最长真前缀，是$\text{W}(\mathcal{A})$中 α 的唯一真前缀 α'，并且，令 $\rho \in \Sigma^+$ 为满足 $\alpha = \alpha'\rho$的字符串，我们有 $(I_{\alpha'}, I_\alpha, \rho) \in E$。这意味着，如果$\mathcal{A}$ 是一个广义确定有限自动机且 $\alpha \in \mathcal{W}(\mathcal{A})$，那么$\mathcal{K}(T_\alpha) = \{\rho \in \Sigma^+ \mid \alpha\rho \in \mathcal{W}(\mathcal{A}), (I_\alpha, I_{\alpha\rho}, \rho) \in E\}$。

在经典 Myhill-Nerode 理论中，我们考虑定义在 $\text{Pref}(\mathcal{L})$上的等价关系。在我们的设定中，需要在 $\text{Pref}(\mathcal{L})$的子集上定义等价关系。这引出了以下一般性定义。

▶ **定义 10.** 设 $W \subseteq \Sigma^*$，且 $\sim$ 是 W 上的一个等价关系。我们称 $\sim$ 是右不变 的，如果：

$$(\forall \alpha, \beta \in \mathcal{W})(\forall \phi \in \Sigma^*)(\alpha \sim \beta \to ((\alpha\phi \in \mathcal{W} \iff \beta\phi \in \mathcal{W}) \wedge (\alpha\phi \in \mathcal{W} \implies \alpha\phi \sim \beta\phi)).$$

▶ 注 11. 请注意，如果 ϕ 是空串，那么定义右不变等价关系的性质显然成立，因此可以改述如下：

$$(\forall \alpha, \beta \in \mathcal{W})(\forall \phi \in \Sigma^+)(\alpha \sim \beta \to ((\phi \in T_\alpha \iff \phi \in T_\beta) \wedge (\phi \in T_\alpha \implies \alpha\phi \sim \beta\phi)).$$

下面证明，$\sim$ 是右不变的，当且仅当：

$$(\forall \alpha, \beta \in \mathcal{W})(\forall \phi \in \Sigma^+)(\alpha \sim \beta \to ((\phi \in \mathcal{K}(T_\alpha) \iff \phi \in \mathcal{K}(T_\beta)) \wedge (\phi \in \mathcal{K}(T_\alpha) \implies \alpha\phi \sim \beta\phi)).$$

($\Rightarrow$) 令 α，$\beta \in \mathcal{W}$ 使得 $\alpha \sim \beta$，并令 $\phi \in \mathcal{K}(T_\alpha)$。我们必须证明 $\phi \in \mathcal{K}(T_\beta)$ 且 $\alpha\phi \sim \beta\phi$。由于 $\phi \in \mathcal{K}(T_\alpha) \subseteq T_\alpha$，我们立即得到 $\phi \in T_\beta$ 和 $\alpha\phi \sim \beta\phi$，因此只需证明 $\phi \in \mathcal{K}(T_\beta)$。由于对于每个 $\phi' \in \Sigma^+$，$\phi' \in T_\alpha$ 当且仅当 $\phi' \in T_\beta$，于是 $T_\alpha = T_\beta$，因此 $\mathcal{K}(T_\alpha) = \mathcal{K}(T_\beta)$。由此，根据 $\phi \in \mathcal{K}(T_\alpha)$，我们得出 $\phi \in \mathcal{K}(T_\beta)$。

($\Leftarrow$) 令 $\alpha, \beta \in \mathrm{W}$ 使得 $\alpha \sim \beta$，并令 $\phi \in T_\alpha$。我们必须证明 $\phi \in T_\beta$ 和$\alpha\phi \sim \beta\phi$。设 $\phi_1, \ldots, \phi_s$ 为 ϕ 的所有前缀，且对每个 $1 \le i \le s$，均有 $\alpha\phi_i \in \mathrm{W}$，其中 ϕ_i 对于每个 $1 \le i \le s-1$ 都是 ϕ_{i+1} 的真前缀。注意，$s \ge 2, \phi_1 = \epsilon$ 和 $\phi_s = \phi$。对于每个 $1 \le i \le s-1$，令 $\psi_i \in \Sigma^+$ 使得 $\phi_{i+1} = \phi_i\psi_i$。注意，根据定义，对于每个 $1 \le i \le s-1$，我们都有$\psi_i \in \mathrm{K}(T_{\alpha\phi_i})$。由于 $\alpha, \beta \in \mathrm{W}$，$\alpha \sim \beta$ 且 $\psi_1 \in \mathrm{K}(T_{\alpha\phi_1}) = \mathrm{K}(T_\alpha)$，我们得到 $\psi_1 \in \mathrm{K}(T_\beta)$ 和 $\alpha\phi_2 = \alpha\psi_1 \sim \beta\psi_1 = \beta\phi_2$。由于 $\alpha\phi_2, \beta\phi_2 \in \mathrm{W}$，$\alpha\phi_2 \sim \beta\phi_2$ 且$\psi_2 \in \mathrm{K}(T_{\alpha\phi_2})$，我们得到 $\psi_2 \in \mathrm{K}(T_{\beta\phi_2})$ 和 $\alpha\phi_3 = \alpha\phi_2\psi_2 \sim \beta\phi_2\psi_2 = \beta\phi_3$。如此继续下去，我们得出 $\phi \in T_\beta$ 和 $\alpha\phi = \alpha\phi_s \sim \beta\phi_s = \beta\phi$。

一般来说，等价关系并非右不变。下面说明如何从任意 等价关系出发，定义一个典范的右不变等价关系。

▶ 引理 *12. </style}*设 $W \subseteq \Sigma^*$，且 $\sim$ 是 W 上的一个等价关系。对于任意 $\alpha, \beta \in W$，定义：

$$\alpha \sim_r \beta \iff (\forall \phi \in \Sigma^*)((\alpha\phi \in \mathcal{W} \iff \beta\phi \in \mathcal{W}) \wedge (\alpha\phi \in \mathcal{W} \implies \alpha\phi \sim \beta\phi)).$$

那么，$\sim_r$ 是 W 上的一个等价关系，且它是右不变的，也是细化 $\sim$的最粗右不变等价关系。我们称 $\sim_r$ 为 $\sim$的右不变细化。

Myhill-Nerode 等价关系在经典 Myhill-Nerode 定理中发挥着重要作用。下面说明当 $\mathcal{W}$ 不一定等于 Pref($\mathcal{L}$) 时，如何对其进行推广。

▶ **定义 13. </style}设 L, W** 。**L和**W 上的Myhill-Nerode 等价关系是 W 上的等价关系$\equiv_{\mathcal{L},\mathcal{W}}$，定义为$\sim_{\mathcal{L},\mathcal{W}}$的右不变细化，其中$\sim_{\mathcal{L},\mathcal{W}}$ 是 W 上的等价关系，满足对于任意 $\alpha, \beta \in W$：

$$\alpha \sim_{\mathcal{L},\mathcal{W}} \beta \iff (\alpha \in \mathcal{L} \iff \beta \in \mathcal{L}).$$

若 $\mathcal{W}$ =前缀集($\mathcal{L}$)，则可得到 $\mathcal{L}$ 的经典 Myhill-Nerode 等价关系。下面描述 $\equiv_{\mathcal{L},\mathcal{W}}$ 的一些基本性质。

▶ **引理 14. </style}设 L,W** 。**则 ,**$\mathcal{W}$ 是右不变的，且 L 是若干个 $\equiv_{\mathcal{L},\mathcal{W}}$-等价类的并集。

证明. 首先，$\equiv_{\mathcal{L},\mathcal{W}}$ 是右不变的，因为根据定义，它是一个右不变细化。此外，$\mathcal{L}$ 是若干个 $\sim_{\mathcal{L},\mathcal{W}}$-等价类的并集，因此也必然是若干个 $\equiv_{\mathcal{L},\mathcal{W}}$-等价类的并集，因为 $\equiv_{\mathcal{L},\mathcal{W}}$ 细化了 $\sim_{\mathcal{L},\mathcal{W}}$。◀

设 $\mathcal{A}$ 是一个传统非确定有限自动机，并在前缀集 $\sim_{\mathcal{A}}$ Pref($\mathcal{L}(\mathcal{A})$) 上定义等价关系如下：对任意 $\alpha, \beta \in$ Pref($\mathcal{L}(\mathcal{A})$)，当且仅当 $I_\alpha = I_\beta$ 时，$\alpha \sim_{\mathcal{A}} \beta$。这个等价关系是传统自动机的 Myhill-Nerode 定理中的一个中间工具；通过考虑等价关系，也可以将其推广到广义自动机 $\mathcal{A}$：

$\sim_{\mathcal{A}}$ 在 $\mathcal{W}(\mathcal{A})$ 上满足：对任意 $\alpha, \beta \in \mathcal{W}(\mathcal{A})$，当且仅当 $I_\alpha = I_\beta$ 时，有 $\alpha \sim_{\mathcal{A}} \beta$。如果$\mathcal{A}$ 是一个非确定有限自动机（或带有 ϵ-转换的非确定有限自动机），那么 $\sim_{\mathcal{A}}$ 是右不变的，因为对任意$\alpha \in \mathrm{Pref}(\mathcal{L}(\mathcal{A}))$以及 α 的任意前缀 α'，从初始状态到 I_α 中某个节点的任何路径都必须经过 $I_{\alpha'}$ 中的某个节点。然而，一般来说，如果 $\mathcal{A}$ 是广义非确定有限自动机，这一性质便不成立，因此当 $\mathcal{A}$ 是广义非确定有限自动机时，$\sim_{\mathcal{A}}$ 不一定是右不变的（见图 3）。由于右不变性对 Myhill-Nerode 定理至关重要，我们将考虑 $\sim_{\mathcal{A}}$ 的右不变细化。

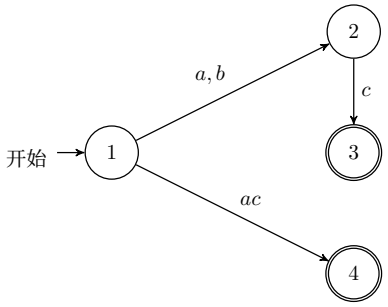


图 3 一个广义非确定有限自动机$\mathcal{A}$，使得 $\sim_{\mathcal{A}}$ 不是右不变的。事实上，a，b，ac，$bc \in \mathcal{W}(\mathcal{A})$ 和 $a \sim_{\mathcal{A}} b$，但 $ac \not\sim_{\mathcal{A}} bc$。

▶ **定义 15**。设 $A = (Q, E, s, F)$ 是一个广义非确定有限自动机。设 $\equiv_{\mathcal{A}}$ 是 $\sim_{\mathcal{A}}$的右不变细化，其中 $\sim_{\mathcal{A}}$ 是 $W(A)$ 上的等价关系，满足对任意 $\alpha, \beta \in W(A)$：

$$\alpha \sim_{\mathcal{A}} \beta \iff I_\alpha = I_\beta.$$

▶ 注 16。我们来证明，若$\mathcal{A}$ 是广义确定有限自动机，则 $\sim_{\mathcal{A}}$ 是右不变的。设 α，$\beta \in \mathcal{W}$满足 $I_\alpha = I_\beta$，并设 $\phi \in \mathcal{K}(T_\alpha)$。根据注 11，我们只需证明 $\phi \in \mathcal{K}(T_\beta)$ 和 $I_{\alpha\phi} = I_{\beta\phi}$。根据注 9，有 $\alpha\phi \in \mathcal{W}(\mathcal{A})$ 和 $(I_\alpha, I_{\alpha\phi}, \phi) \in E$。因此 $(I_\beta, I_{\alpha\phi}, \phi) \in E$，由此得到 $I_{\alpha\phi} = I_{\beta\phi}$，再根据注 9，得出 $\phi \in \mathcal{K}(T_\beta)$。请注意，事实上，图 3 中的广义自动机$\mathcal{A}$ 并非广义确定有限自动机。

由于 $\equiv_{\mathcal{A}}$ 是 $\sim_{\mathcal{A}}$的右不变细化，我们得出：如果 $\mathcal{A}$ 是广义确定有限自动机，那么$\equiv_{\mathcal{A}}$ 和 $\sim_{\mathcal{A}}$ 是同一个等价关系。

让我们研究 $\equiv_{\mathcal{A}}$ 的性质。

▶ **引理 17**。设 $A = (Q, E, s, F)$ 是一个广义非确定有限自动机。那么，$\equiv_{\mathcal{A}}$ 是右不变的，它细化了 $\equiv_{\mathcal{L}(\mathcal{A}),\mathcal{W}(\mathcal{A})}$，具有有限指数，并且 $L(A)$ 是若干个 $\equiv_{\mathcal{A}}$-等价类的并集。

以下引理对于推导广义自动机的 Myhill-Nerode 定理至关重要。

▶ **引理 18**。设 $L \subseteq \Sigma^*$，且 $W \subseteq \Sigma^*$ 是一个局部有界集，满足 $L \cup \{\epsilon\} \subseteq W \subseteq$前缀集$(L)$。假设 L 是 W 上某个有限指数右不变等价关系 $\sim$ 的若干等价类的并集。那么，L 可由一个 W-GDFA $A_\sim = (Q_\sim, E_\sim, s_\sim, F_\sim)$ 识别，使得：

1. $|Q_\sim|$ 等于 $\sim$的指标。2. $\equiv_{\mathcal{A}_\sim}$ 和 $\sim$ 是相同的等价关系。此外，如果 $\mathcal{B}$ 是一个 $\mathcal{W}$-广义确定有限自动机，识别 $\mathcal{L}$，那么 $\mathcal{A}_{\equiv_{\mathcal{B}}}$ 与 $\mathcal{B}$同构。

证明。(概要) 所需的 $\mathcal{W}$-广义确定有限自动机 $\mathcal{A}_\sim = (Q_\sim, E_\sim, s_\sim, F_\sim)$ 定义如下。$Q_\sim = \{[\alpha]_\sim \mid \alpha \in \mathcal{W}\}$。$s_\sim = [\epsilon]_\sim$。

- $E_\sim = \{([\alpha]_\sim, [\alpha\rho]_\sim, \rho) \mid \alpha \in W \ \ \rho \in K(T_\alpha)\}$.
- $F_\sim = \{[\alpha]_\sim \mid \alpha \in L\}$.

◀

▶ 注 19. 在引理 18 的陈述中，我们不能删去 W 局部有界这一假设，因为我们已经证明，若 A 是广义非确定有限自动机，则 W(A) 是局部有界的。然而，若 W 不是局部有界的，那么 $A_\sim$ 仍是一个定义良好的自动机，且状态数有限，但它有无限多条边。例如，$W = \{\epsilon\} \cup a^*b$ 不是局部有界的，因为 $T_\epsilon = a^*b$ 和 $K(T_\epsilon) = a^*b$ 是一个无限集合。若 $L = a^*b$，则 $L \cup \{\epsilon\} \subseteq W \subseteq Pref(L)$。此外，$\equiv_{\mathcal{L},\mathcal{W}}$ 的指标有限（等价类为 $\{\epsilon\}$ 和 a^*b），根据引理 14，我们知道 $\equiv_{\mathcal{L},\mathcal{W}}$ 是右不变的，并且 L 是若干个 $\equiv_{\mathcal{L},\mathcal{W}}$-类的并集。因此我们得出结论：$A_{\equiv_{\mathcal{L},\mathcal{W}}}$ 是定义良好的，但它有无限多条边（见图 4）。

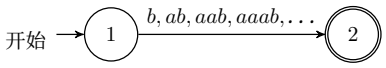


图4 一个 $\mathcal{W}$ 不是局部有界集的例子。

现在我们可以陈述广义自动机的 Myhill-Nerode 定理。

▶**定理20** (广义自动机的 Myhill-Nerode 定理)。 设 $L \subseteq \Sigma^*$，且 $W \subseteq \Sigma^*$是一个局部有界集，满足 $L \cup \{\epsilon\} \subseteq W \subseteq$ 前缀集(L)。以下条件等价：
1. $\mathcal{L}$ 可由一个$\mathcal{W}$-广义非确定有限自动机识别。**2.***Myhill-Nerode* 等价关系 $\equiv_{\mathcal{L},\mathcal{W}}$ 具有有限指标。**3.**存在一个定义在 $\sim$ 上的右不变等价关系 $\mathcal{W}$，其指标有限，且 $\mathcal{L}$ 是若干个 $\sim$等价类的并。**4.** $\mathcal{L}$ 可由一个$\mathcal{W}$-广义确定有限自动机识别。

此外，如果上述某个陈述成立（因而上述所有陈述都成立），那么存在唯一的识别 L 的最小 *W-GDFA*（也就是说，在所有识别 L 的 *W-GDFA* 中，状态数最少的两个 *W-GDFA* 必须同构）。

证明。 (1) → (2) 设 $\mathcal{A}$ 是识别 $\mathcal{L}$ 的 $\mathcal{W}$-GDFA。由引理 17 可知，$\equiv_{\mathcal{A}}$ 具有有限指数，且细化了 $\equiv_{\mathcal{L},\mathcal{W}}$，因此 $\equiv_{\mathcal{L},\mathcal{W}}$ 也具有有限指数。
(2) → (3) 根据引理 14，所需的等价关系是 $\equiv_{\mathcal{L},\mathcal{W}}$。(3) → (4) 由引理 18 可得。(4) → (1) 每个 $\mathcal{W}$-GDFA 都是一个 $\mathcal{W}$-GNFA。

现在，我们来证明最小自动机是引理 18 中定义的 $A_{\equiv_{\mathcal{L},\mathcal{W}}}$。首先，$A_{\equiv_{\mathcal{L},\mathcal{W}}}$ 定义良好，因为$\equiv_{\mathcal{L},\mathcal{W}}$ 是右不变的，而且根据引理 14，L 是若干个$\equiv_{\mathcal{L},\mathcal{W}}$-等价类的并；此外，根据我们假定成立的某个陈述，$\equiv_{\mathcal{L},\mathcal{W}}$ 具有有限指数。现在，由引理 18，$A_{\equiv_{\mathcal{L},\mathcal{W}}}$ 的状态数等于$\equiv_{\mathcal{L},\mathcal{W}}$ 的指标，或者等价地，等于$\equiv_{\mathcal{A}_{\equiv_{\mathcal{L},\mathcal{W}}}}$的指标。另一方面，设 B 是任意一个识别 L 且不同构于 $A_{\equiv_{\mathcal{L},\mathcal{W}}}$ 的 W-GDFA。根据引理 17，$\equiv_{\mathcal{B}}$ 是 $\equiv_{\mathcal{L},\mathcal{W}}$ 的细化，并且它必定是 $\equiv_{\mathcal{L},\mathcal{W}}$ 的严格细化；否则，$A_{\equiv_{\mathcal{L},\mathcal{W}}}$将等于 $A_{\equiv_{\mathcal{B}}}$，而根据引理 18，后者与 B 同构，这就产生了矛盾。由此可知，$\equiv_{\mathcal{L},\mathcal{W}}$ 的指标小于 $\equiv_{\mathcal{B}}$的指标，因此再次根据引理 18，$A_{\equiv_{\mathcal{L},\mathcal{W}}}$ 的状态数小于 $A_{\equiv_{\mathcal{B}}}$ 的状态数，因而也小于 B 的状态数。◀

传统自动机的 Myhill-Nerode 定理（定理 1）是定理20的特例，因为如果 $\mathcal{A}$ 是一个非确定有限自动机，那么 $\mathcal{W}(\mathcal{A}) = Pref(\mathcal{L}(\mathcal{A}))$。此外，定理20

与图2中的示例一致，因为将图2中的两个广义确定有限自动机分别记为 $\mathcal{A}_1$ 和 $\mathcal{A}_2$ 后，我们已经证明了 $\mathcal{W}(\mathcal{A}_1) \neq \mathcal{W}(\mathcal{A}_2)$。

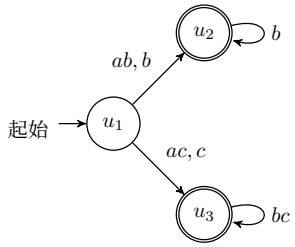
4 广义自动机的 FM 索引

在本节中，我们证明定理 3。为了阐述主要思路，只需考虑广义确定有限自动机。对于不含 ϵ-转换的广义非确定有限自动机这一更一般的情形，只需进行少量技术性修改，我们将在附录C中讨论。

设 V 是一个集合。我们称 V 上的（二元）关系 $\leq$ 为偏序，如果 $\leq$ 具有自反性、反对称性和传递性。若对任意 $u,\ v \in V$，都有 $(u \leq v) \vee (v \leq u)$，则称偏序 $\leq$ 为全序。若对任意 $u,\ v,\ z \in V$，只要 $u \leq v \leq z$ 且 $u,\ z \in U$，就有 $v \in U$，则称 $U \subseteq V$ 是 $\leq$-凸的。对于任意 $u,\ v \in V$，若 $(u \leq v) \wedge (u \neq v)$，则记作 $u < v$。

我们来定义 Wheeler 广义确定有限自动机。按照 Wheeler 自动机相关文献中的惯例 [3, 16],，我们假设存在一个固定的全序 $\preceq$，定义在 Σ 上（在我们的示例中，我们始终假设 $a \prec b \prec c \prec \ldots$），并将 $\preceq$ 扩展为反字典序 Σ^*（也就是说，对于每个 $\alpha,\ \beta \in \Sigma^*$，若反转字符串 α^R 的字典序小于反转字符串 β^R，则有 $\alpha \prec \beta$）。设 $\mathcal{A} = (Q,\ E,\ s,\ F)$ 是一个广义确定有限自动机。设 $\preceq_{\mathcal{A}}$ 为 Q 上的自反关系，使得对于每个 $u,\ v \in Q$，若 $u \neq v$，则当且仅当 $(\forall \alpha \in I_u)(\forall \beta \in I_v)(\alpha \prec \beta)$ 时有 $u \prec_{\mathcal{A}} v$。由于每个 I_u 都非空，因此不难看出 $\preceq_{\mathcal{A}}$ 是一个偏序，但通常不是全序。于是我们可以给出以下定义（示例见图 5）。

▶ **定义 21**。设 $A = (Q, E, s, F)$ 是一个广义确定有限自动机。若 $\preceq_{\mathcal{A}}$ 是全序，则称 A 为Wheeler。



- $\texttt{OUT}_1 = 001011$
- $\texttt{OUT}_2 = 001101$
- $\texttt{IN}_1 = 100101$
- $\texttt{IN}_2 = 101001$
- $\texttt{LAB}_1 = (b)(c)(b)$
- $\texttt{LAB}_2 = (ab)(ac)(bc)$
- $\texttt{FIN} = 011$

图 5 左图：一个 Wheeler 广义确定有限自动机 $\mathcal{A}$。状态按照全序 $\preceq_{\mathcal{A}}$编号。右图：$\mathcal{A}$ 的 Burrows-Wheeler 变换（BWT）（见定义 29）。

若 A 是传统确定有限自动机，则 Definition 21 是否等价于文献中常用的 Wheeler 确定有限自动机局部定义，并非一目了然 [3, 11, 15]。按照局部定义，若存在一个关于 Q 的全序 $\leq$，使得 (i) s 在全序中排在最前，(ii) 对每个 (u', u, a)、$(v', v, b) \in E$，若 $u < v$，则 $a \preceq b$；以及 (iii) 对每个 (u', u, a)、$(v', v, a) \in E$，若 $u < v$，则 $u' < v'$，则称确定有限自动机 A $= (Q, E, s, F)$ 为 Wheeler 确定有限自动机。Alanko 等人 [3,推论 3.1] 证明了，若这样的全序 $\leq$ 存在，则它是唯一的，且等于 $\preceq_{\mathcal{A}}$。因此，我们只需证明：若 $\preceq_{\mathcal{A}}$ 是一个全序，则它满足性质 (i)、(ii) 和 (iii)。这可由以下引理推出。

▶ **引理 22**。设 $A = (Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机。那么：

1. s 在全序 $\preceq_{\mathcal{A}}$ 中排在最前。*2.* 对每个 (u', u, ρ)、$(v', v, \rho') \in E$，若 $u \prec_{\mathcal{A}} v$ 且 ρ' 不是 ρ 的真后缀，则 $\rho \preceq \rho'$。*3.* 对每个 (u', u, ρ)、$(v', v, \rho) \in E$，若 $u \prec_{\mathcal{A}} v$，则 $u' \prec_{\mathcal{A}} v'$。

在引理 22 的情形 2 中，我们不能去掉 ρ' 不是 ρ 的真前缀这一假设；因此，我们无法利用引理 22 给出 Wheeler 广义确定有限自动机的等价局部定义（见图 6）。Wheeler 确定有限自动机的局部定义很容易推出，判断给定的确定有限自动机是否为 Wheeler 确定有限自动机这一问题可在多项式时间内解决[3], 但由于我们没有 Wheeler 广义确定有限自动机的局部定义，因此尚不清楚相应的广义确定有限自动机判定问题是否也能在多项式时间内解决（而且我们在引言中看到，广义自动机上的计算问题通常很难）。不过，我们可以证明该问题仍然是易解的：将其归约为计算与给定广义确定有限自动机等价的传统确定有限自动机的部分序 $\preceq_{\mathcal{A}^*}$。$\mathcal{A}^*$ $\mathcal{A}$。

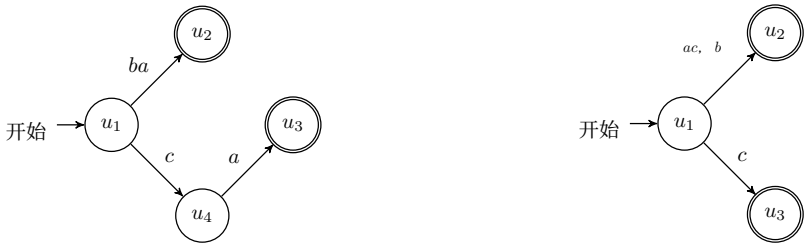


图 6 左：Wheeler 广义确定有限自动机 A $=(Q, E, s, F)$。状态按照全序 $\preceq_{\mathcal{A}}$ 编号。注意，(u_1, u_2, ba)、(u_4, u_3, a) $\in E$, $u_2 \prec_{\mathcal{A}} u_3$, a 是 ba 和 $a \prec ba$ 的真后缀。右：广义确定有限自动机 A $=(Q, E, s, F)$，其状态按照全序 $\leq$ 编号，并满足：(i) s 在全序 $\leq$ 中位于最前；(ii) 对于每个 (u', u, ρ)、(v', v, ρ') $\in E$，若 $u < v$ 且 ρ' 不是 ρ 的真后缀，则 $\rho \preceq \rho'$；以及 (iii) 对于每个 (u', u, ρ)、(v', v, ρ) $\in E$，若 $u < v$，则 $u' < v'$。注意，A 不是 Wheeler 广义自动机，因为 u_2 和 u_3 不具有 $\preceq_{\mathcal{A}}$-可比较性，这是因为 $b \prec c \prec ac$、$c \in I_{u_3}$ 以及 b、$ac \in I_{u_2}$。

▶ **引理 23.** 设 A $=(Q, E, s, F)$ 是一个广义确定有限自动机，令 $\mathfrak{e}$ 为所有边标签的总长度。我们可以在 $O(\mathfrak{e}\log \mathfrak{e})$ 时间内判定 A 是否为 *Wheeler* 自动机；若是，则可以计算出 $\preceq_{\mathcal{A}}$。

证明。我们从 A 出发构造一个确定有限自动机 A*（示例见图 7）。一般而言，如果 $C \subseteq \Sigma^+$，我们可以构造一棵字典树，使得从根出发沿路径可读出的所有字符串，恰好构成至少有一个元素以其为前缀的字符串集合 C。如果C 是前缀无关的，那么当且仅当一个字符串可以从根读到叶节点时，该字符串才属于 C；我们称这样的叶节点拼出所讨论的字符串。对于每个 $u \in Q$，令 C_u 为离开 u的某条边所标记的所有字符串组成的前缀无关集。我们通过选取 A、删除所有边，并为每个非空的 C_u 构造一棵字典树，以此构造自动机 A$^* = (Q^*, E^*, s^*, F^*)$，其中 (a) 字典树的根为 u，(b) 拼出 ρ 的叶节点是唯一满足 (u, v, ρ) $\in E$ 的 $v \in Q$，(c) 字典树的每个内部状态都是一个新状态。注意，$Q \subseteq Q^*$；我们定义 $s^* = s$和 $F^* = F$。根据构造，A*是一个确定有限自动机，并且对每个 $u \in Q$，都有 $I_u^{\mathcal{A}} = I_u^{\mathcal{A}^*}$（因此 L(A*) =L(A)）。由此可知，A 是 Wheeler 自动机，当且仅当偏序 $\preceq_{\mathcal{A}^*}$ 在 Q 上的限制是全序。由于 A* 是传统确定有限自动机，我们可以在多项式时间内计算偏序 $\preceq_{\mathcal{A}^*}$ [16, 33, 13, 8]。例如，[8] 中的算法可在 $O(|E^*|\log |Q^*|)$ 时间内运行；所述时间界由此得证，因为$|Q^*| - 1 \leq |E^*| = \mathfrak{e}$。◀

本节余下部分将用于证明，SMLG 问题可以在 Wheeler 广义确定性有限自动机上高效求解。如果 $\mathcal{A}$ $=(Q,$ $E,$ $s,$ $F)$ 是一个 Wheeler 广义确定有限自动机，我们记作$Q = \{Q[1],$ $Q[2],$ $\ldots,$ $Q[n]\}$，其中 $Q[1] \prec_{\mathcal{A}} Q[2] \prec_{\mathcal{A}} \cdots \prec_{\mathcal{A}} Q[n]$；若 $1 \leq i \leq j \leq n$，则记作 $Q[i, j] = \{Q[i],$ $Q[i+1],$ $\ldots,$ $Q[j-1],$ $Q[j]\}$。若 $\alpha,$ $\beta \in \Sigma^*$，我们记作 $\alpha \dashv \beta$，当且仅当 α 是 β 的后缀。

▶ **定义 24。** 设 A $=(Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机，且令 $\alpha \in \Sigma^*$。定义：

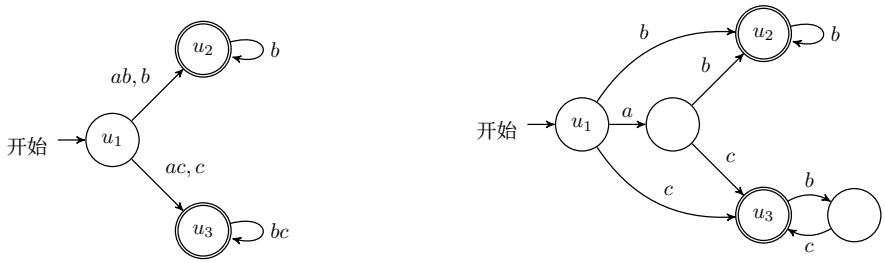


图 7 左：图 5 中的广义确定有限自动机 $\mathcal{A}$。右：证明引理 23 时从 $\mathcal{A}$ 开始构造的确定有限自动机 A^*。

$G^{\prec}(\alpha) = \{u \in Q \mid (\forall \beta \in I_u)(\beta \prec \alpha)\}$; $G_{\dashv}(\alpha) = \{u \in Q \mid (\exists \beta \in I_u)(\alpha \dashv \beta)\}$; $G^{\preceq}_{\dashv}(\alpha) = G^{\prec}(\alpha) \cup G_{\dashv}(\alpha) = \{u \in Q \mid (\forall \beta \in I_u)(\beta \prec \alpha) \vee (\exists \beta \in I_u)(\alpha \dashv \beta)\}$。

直观地说，集合 $G_{\dashv}(\alpha)$ 是 SMLG 问题在输入 α 时必须返回的状态集合，而 $G^{\prec}(\alpha)$ 是所有仅由小于 α 的字符串到达的状态的集合。下面的引理表明，与常规 Wheeler 自动机的情形一样，$G^{\prec}(\alpha)$ 和 $G_{\dashv}(\alpha)$ 都是关于全序 $\preceq_{\mathcal{A}}$ 的区间，并且在 $G^{\prec}(\alpha)$ 和 $G_{\dashv}(\alpha)$ 之间不存在任何状态。

▶ **引理 25**。设 $A = (Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机，并设 $\alpha \in \Sigma^*$。那么：
1. $G^{\prec}(\alpha) \cap G_{\dashv}(\alpha) = \emptyset$。*2.* $G_{\dashv}(\alpha)$ 是 $\preceq_{\mathcal{A}}$-凸的。*3.* 如果 u、$v \in Q$ 满足 $u \prec_{\mathcal{A}} v$ 且 $v \in G^{\prec}(\alpha)$，那么 $u \in G^{\prec}(\alpha)$。换句话说，$G^{\prec}(\alpha) = Q[1, |G^{\prec}(\alpha)|]$。*4.* 如果 u、$v \in Q$ 满足 $u \prec_{\mathcal{A}} v$ 且 $v \in G^{\preceq}_{\dashv}(\alpha)$，那么 $u \in G^{\preceq}_{\dashv}(\alpha)$。换句话说，$G^{\preceq}_{\dashv}(\alpha) = Q[1, |G^{\preceq}_{\dashv}(\alpha)|]$。

为了计算 $G_{\dashv}(\alpha)$，只需计算 $G^{\prec}(\alpha)$ 和 $G^{\preceq}_{\dashv}(\alpha)$。下面说明如何计算 $G^{\prec}(\alpha)$ 和 $G^{\preceq}_{\dashv}(\alpha)$；为此，我们将不断使用引理 22 中的性质 3，该性质对于传统的 Wheeler 自动机也至关重要（它是 FM 索引 [24] 中 LF 映射的推广）。首先，令 $G^*(\alpha)$ 表示所有可通过标签为以后缀 α 的字符串的边到达的状态的集合。形式化地说：

$$G^*(\alpha) = \{v \in Q \mid (\exists v' \in Q)(\exists \rho \in \Sigma^*)((v', v, \rho) \in E \wedge (\alpha \dashv \rho))\}.$$

显然，$G^*(\alpha) \subseteq G_{\dashv}(\alpha)$。现在，我们给出如下定义。

▶ **定义 26**。设 $A = (Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机。设 $U \subseteq Q$ 和 $\rho \in \Sigma^+$。我们用 $\mathtt{out}(U, \rho)$ 表示离开 U 中状态且标签为 ρ 的边数，并用 $\mathtt{in}(U, \rho)$ 表示进入 U 中状态且标签为 ρ 的边数。

若 $\alpha \in \Sigma^*$ 和 $0 \leq k \leq |\alpha|$，则令 $p(\alpha, k)$ 和 $s(\alpha, k)$ 分别表示 α 的长度为 k 的前缀和后缀。若 $\mathcal{A} = (Q, E, s, F)$ 是一个广义确定有限自动机且 $u \in Q$，令 $\lambda(u)$ 表示 Σ^+ 中为一条到达 u 的边提供标签的所有字符串的集合；我们分别用 $\min^{\lambda(u)}$ 和 $\max^{\lambda(u)}$ 表示 $\lambda(u)$ 中（按反字典序）最小和最大的字符串。

下一个引理将以下直觉形式化：要计算 $G^{\prec}(\alpha)$，我们必须考虑所有入边标签小于 α 的状态；此外，如果标签为 $s(\alpha,\ k)$（对于某个 k），那么起始状态必须属于 $G^{\prec}(p(\alpha,\ |\alpha| - k))$。

14 广义自动机的 Myhill-Nerode 定理及其应用

▶ 引理 *27.* 设 $A = (Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机，且 $\alpha \in \Sigma^*$。对于 $0 < k < |\alpha|$，令 $f_k = \mathtt{out}(Q[1, |G^{\prec}(p(\alpha, |\alpha| - k))|], s(\alpha, k))$。那么，$|G^{\prec}(\alpha)|$ 是满足以下条件的最大整数 $0 \le j \le |Q|$: *(i)* 对每个 $0 < k < |\alpha|$，$\mathtt{in}(Q[1, j], s(\alpha, k)) \le f_k$；且 *(ii)* 若 $j \ge 1$，则 $max_{\lambda(Q[j])} \prec \alpha$。

以下关键引理表明，为了计算 $|G^{\preceq}_{\dashv}(\alpha)|$，我们只需考虑 $|G^{\prec}(\alpha)|$，即 $G^*(\alpha)$ 中按 $\preceq_{\mathcal{A}}$ 意义下最大的状态，以及 $G^{\preceq}_{\dashv}(\alpha) \setminus G^*(\alpha)$ 中的状态。

▶ 引理 *28.* 设 $A = (Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机，且 $\alpha \in \Sigma^*$。对于 $0 < k < |\alpha|$，令 $f_k = \mathtt{out}(Q[1, |G^{\prec}(p(\alpha, |\alpha| - k))|], s(\alpha, k))$ 和 $g_k = \mathtt{out}(Q[1, |G^{\preceq}_{\dashv}(p(\alpha, |\alpha| - k))|], s(\alpha, k))$。那么，$g_k \ge f_k$ 对于每个 $0 < k < |\alpha|$。此外，$|G^{\preceq}_{\dashv}(\alpha)|$ 等于以下各项中的最大值：
1. $|G^{\prec}(\alpha)|$；**2.** 满足以下条件的最大整数 $0 \le i \le |Q|$：如果 $i \ge 1$，则 $Q[i] \in G^*(\alpha)$；**3.** 满足以下条件的最小整数 $0 \le j \le |Q|$：对于每个 $0 < k < |\alpha|$ 使得 $g_k > f_k$，都有 $\mathtt{in}(Q[1, j], s(\alpha, k)) \ge g_k$。

现在，我们准备将 Burrows-Wheeler 变换和 FM 索引推广到 Wheeler 广义确定性有限自动机。设 $A = (Q, E, s, F)$ 是一个广义确定有限自动机。若所有边标签的长度至多为 r，则称 A 为 r-广义确定有限自动机。固定 $1 \le i \le r$。令 $E(i) = \{(u', u, \rho) \in E \mid \rho \in \Sigma^i\}$ 表示标记为长度为 i 的字符串的边的数量，并令 $\Sigma(i) = \{\rho \in \Sigma^i \mid (\exists u', u \in Q)((u', u, \rho) \in E(i)\}$ 表示为某条边添加标签的所有长度为 i 的字符串的集合。令 $e_i = |E(i)|$ 和 $\sigma_i = |\Sigma(i)|$；则有 $\sigma_i \le \min\{\sigma^i, e_i\}$。一个状态的 i-出度（分别为 i-入度）等于 $E(i)$ 中离开（分别到达）该状态的边数。所有状态的 i-出度之和以及所有状态的 i-入度之和都等于 e_i。最后，$\sum_{i=1}^{r} e_i = e$，且所有边标签的总长度为 $\mathfrak{e} = \sum_{i=1}^{r} e_i i$。

下面定义 Wheeler 广义确定有限自动机的 Burrows-Wheeler 变换（示例见图 5）。

▶ **定义 29** (Wheeler 广义确定有限自动机的 Burrows-Wheeler 变换)。设 $A = (Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机。A 的Burrows-Wheeler 变换 $BWT(A)$ 由以下字符串组成。

1. 对于每个 $1 \le i \le r$，位串 $\mathtt{OUT}_i \in \{0, 1\}^{e_i + n}$ 以一元形式存储 i-出度。更准确地说，(i) $\mathtt{OUT}_i$ 恰好包含 n 个值为 1 的字符，(ii) $\mathtt{OUT}_i$ 恰好包含e_i 个值为 0 的字符，并且 (iii) 第 $(\ell - 1)$ 个值为 1 的字符与第 ℓ 个值为 1 的字符之间的零的数量（如果 $\ell = 1$，则从序列开头算起）给出 $Q[\ell]$ 的 i-出度。2. 对于每个 $1 \le i \le r$，位串 $\mathtt{IN}_i \in \{0, 1\}^{e_i + n}$ 以一元形式存储 i-入度。更准确地说，(i) $\mathtt{IN}_i$ 恰好包含 n 个值为 1 的字符，(ii) $\mathtt{IN}_i$ 恰好包含e_i 个值为 0 的字符，并且 (iii) 第 $(\ell - 1)$ 个值为 1 的字符与第 ℓ 个值为 1 的字符之间的零的数量（如果 $\ell = 1$，则从序列开头算起）给出 $Q[\ell]$ 的 i-入度。3. 对于每个 $1 \le i \le r$，字符串 $\mathtt{LAB}_i \in (\Sigma^i)^{e_i}$ 存储长度为 i 的边标签（及其重数）。更准确地说，我们按照起始状态的指标（相对于 $\preceq_{\mathcal{A}}$）对 $E(i)$ 中的所有边进行排序。起始状态相同的边按标签排序。最后，按此边序连接所有边的标签，得到 $\mathtt{LAB}_i$。4. 位串 $\mathtt{FIN} \in \{0, 1\}^n$ 用于标记终态，也就是说，当且仅当 $Q[i] \in F$ 时，$\mathtt{FIN}$ 的第 i 位等于 1。

现在我们可以证明，Wheeler 广义确定有限自动机$\mathcal{A}$ 的 BWT 是$\mathcal{A}$的有效编码，就像字符串的 BWT 是该字符串的有效编码一样。

▶ **定理 30**。 设 $A = (Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机。若我们只知道 $BWT(A)$，那么就可以还原 A（直至同构）。换言之， $BWT(A)$ 是 A 的一种编码。

证明。考虑状态 $Q[1]$、……、$Q[n]$。由引理 22 可得 $s = Q[1]$，并且利用 $\texttt{FIN}$，我们可以检索所有终态的集合 F。我们只需证明可以检索 E。换言之，对于每个 $1 \leq i'$、$i \leq n$ 以及每个 $\rho \in \Sigma^+$，我们必须判断 $(Q[i'], Q[i], \rho) \in E$ 是否成立。只要对每个 $1 \leq i \leq r$ 检索集合 $E(i)$ 即可，因为 E 是各个 $E(i)$ 的（不相交）并集。利用 $\texttt{LAB}_i$，我们可以检索 $E(i)$ 中所有边的标签及其重数。利用 $\texttt{IN}_i$，我们可以检索每个 $Q[\ell]$ 的 i-入度。由引理 22，对于每个标记某条边并到达某个节点 $Q[\ell]$ 的 $\rho \in \Sigma^i$，以及每个标记某条边并到达 $Q[\ell+1]$ 的 $\rho' \in \Sigma^i$，必有 $\rho \preceq \rho'$。由于我们知道 $E(i)$ 中所有边的标签及其重数，也知道 i-入度，因此可以检索进入每个 $Q[\ell]$ 的所有边的标签及其重数。利用 $\texttt{OUT}_i$，我们可以检索每个$Q[\ell]$的 i-出度，而 $\texttt{LAB}_i$ 定义中采用的顺序意味着我们可以检索离开每个 $Q[\ell]$ 的所有边的标签。由于我们知道进入每个 $Q[\ell]$ 的所有边的标签，也知道离开每个 $Q[\ell]$ 的所有边的标签，因此对于每个 $\rho \in \Sigma^i$，我们知道由标记为 ρ 的边到达的所有状态的集合（含重数），以及具有标记为 ρ 的出边的所有状态的集合。由引理 22，对于每个 $1 \leq j < k \leq n$，如果 $Q[j]$ 是由一条离开状态 $Q[j']$、标签为 ρ 的边到达的，而 $Q[k]$ 是由一条离开状态 $Q[k']$、标签为 ρ 的边到达的，那么必有 $j' < k'$。因此，对于每个 $\rho \in \Sigma^i$，我们可以检索所有标签为 ρ 的边的集合 $E(i)_\rho$，进而可以检索 $E(i)$，因为 $E(i)$ 是各个 $E(i)_\rho$ 的（不相交）并集。◀

现在我们可以证明本节的主定理（完整证明见附录）。回顾一下， n 表示顶点数， e 表示边数， $\mathfrak{e}$ 表示所有边标签的总长度。

▶ **定理 31**（Wheeler 广义确定性有限自动机的 FM 索引）。 设 A 是一个 *Wheeler* r-广义确定有限自动机。则我们可以使用 $\mathfrak{e}\log\sigma\,(1+o\,(1))\,+O\,(e+rn)$ 比特对 A 进行编码，使得之后给定一个长度为 m 的模式串 $\alpha \in \Sigma^*$，我们可以在 $O\,(rm\,(\log r + \log\log\sigma)$ 时间内解决 A 上的 *SMLG* 问题。在相同的时间界内，我们还可以判定 $\alpha \in \mathcal{L}\,(A)$。

证明。（概要）根据引理 25，我们只需计算 $|G^{\prec}\,(\alpha)|$ 和 $|G_{\dashv}^{\preceq}\,(\alpha)|$。在 $O\,(m)$ 步内，我们递归地为每个 $0 \leq k \leq |\alpha|$计算 $|G^{\prec}\,(p\,(\alpha,\ k))|$和 $|G_{\dashv}\,(p\,(\alpha,\ k))|$。根据引理 27 和引理 28，我们可以将该问题归约为求解一系列基本查询，例如为每个 $0 < k < |\alpha|$计算 $f_k = \texttt{out}\,(Q[1, |G^{\prec}\,(p\,(\alpha,\ |\alpha|-k)|],\ s\,(\alpha,\ k))$。但我们只需为 $0 < k < r+1$ 计算 f_k，因为否则 $|s\,(\alpha,\ k)| > r$，并且所有边标签的长度都至多为 r。一般而言，每个基本查询至多需要 $O\,(r)$ 时间。通过为 A 的 Burrows–Wheeler 变换配备用于秩/选择查询的紧凑数据结构和简洁字典，我们可以在 $O\,(\log r + \log\log\sigma)$ 时间内解决每个基本查询。◀

5 结论与未来工作

本文研究了广义自动机模型，并引入了集合$\mathcal{W}(\mathcal{A})$。我们证明，$\mathcal{W}(\mathcal{A})$ 在传统非确定有限自动机中发挥着与 Pref($\mathcal{L}(\mathcal{A})$) 相同的作用：集合$\mathcal{W}(\mathcal{A})$ 可用于推导 Myhill-Nerode 定理，也是将 FM 索引扩展到广义自动机的起点。

后续研究方向包括将 Burrows-Wheeler 变换和 FM 索引推广到任意广义非确定有限自动机。事实上，Burrows-Wheeler 变换和 FM 索引

16 广义自动机的 Myhill-Nerode 定理及其应用

最近已借助所谓的*co-lex* 序 [14, 16]和 *co-lex* 关系 [12]，从 Wheeler 非确定有限自动机推广到任意非确定有限自动机。然而，我们指出，由于引言中提到的 Equi 等人给出的（条件性）下界，本文推导出的 SMLG 问题高效时间复杂度界不可能适用于任意广义非确定有限自动机。

Giammaresi 和 Montalbano 描述了一种有效过程，可用于计算与给定 GDFA 等价的最小广义确定有限自动机 [27, 26], ，但我们尚不清楚是否存在能够对 GDFA 进行最小化的高效算法。广义自动机的 Myhill-Nerode 定理表明，对于每个最小广义确定有限自动机 $\mathcal{A}$ ，都存在一个集合$\mathcal{W} \subseteq \Sigma^*$ ，使得$\mathcal{A}$ 与识别$\mathcal{L}(\mathcal{A})$的最小$\mathcal{W}$-GDFA 同构。同时，给定一个识别$\mathcal{L}$的$\mathcal{W}$-GDFA，我们可以将 Hopcroft 算法[30] 扩展到 GDFA，从而构造出识别$\mathcal{L}$ 的最小$\mathcal{W}$-GDFA。如果我们能够证明，对于每个可容许的$\mathcal{W} \subseteq \Sigma^*$，识别$\mathcal{L}$ 的最小$\mathcal{W}$-GDFA 的状态数与识别$\mathcal{L}$的最小广义确定有限自动机的状态数相当，那么我们就能得到一种快速算法，在不改变识别语言的前提下显著减少 GDFA 的状态数。

本文留下了许多值得从理论角度探讨的问题。*Wheeler 语言* 是所有可由某个 Wheeler 非确定有限自动机识别的正则语言组成的类[4]。Wheeler 语言具有若干性质：例如，可以用凸 等价关系来刻画它们 [4]。此外，每个 Wheeler 语言也都可由某个确定有限自动机识别；特别地，对于给定的 Wheeler 语言，存在唯一的状态数最少的确定有限自动机来识别它 [2]。Wheeler 语言的主要局限在于，它们只代表正则语言中的一个很小的子类：例如，一元语言（即字母表大小为一的语言）是 Wheeler 语言，当且仅当它是有限语言或余有限语言 [4]。大多数正则语言并非 Wheeler 语言，直观原因在于，Wheeler 非确定有限自动机$\mathcal{A}$ 的定义对 Pref($\mathcal{L}(\mathcal{A})$) 的集合提出了严格要求（这也引出了以凸等价关系为基础的刻画）。然而，当我们转向广义非确定有限自动机时，发挥 Pref($\mathcal{L}(\mathcal{A})$) 作用的是$\mathcal{W}(\mathcal{A})$，并且可能有$\mathcal{W}(\mathcal{A}) \subsetneq$Pref($\mathcal{L}(\mathcal{A})$)；因此，同样的要求如今只适用于一个更小的子集。一个自然的问题是：Wheeler 广义非确定有限自动机是否扩展了 Wheeler 语言这一类？答案是肯定的：存在一个正则$\mathcal{L}$ 语言，它不是 Wheeler 语言（也就是说，没有 Wheeler 非确定有限自动机能够识别$\mathcal{L}$），但可以由 Wheeler 广义确定有限自动机识别。定义$\mathcal{L} = \{ a^{2n} \mid n \geq 0\}$。那么，$\mathcal{L}$不是 Wheeler 语言 [4], ，但$\mathcal{L}$ 可以由一个仅含单个状态的广义确定有限自动机识别，该状态既是初始状态也是终止状态，并带有一条标记为 aa的自环。因此，所有可由 Wheeler 广义非确定有限自动机识别的语言构成的类，严格大于 Wheeler 语言这一类。我们可以将这个新类中的语言称为广义 *Wheeler 语言*：下一步是了解 Wheeler 语言的哪些性质仍然成立，以及如何刻画这个新类。

参考文献

1 Tatsuya Akutsu. 字符串与树之间的线性时间模式匹配算法。载于 Alberto Apostolico、Maxime Crochemore、Zvi Galil 和 Udi Manber 编，组合模式匹配，第1–10页，柏林，海德堡，1993年。Springer Berlin Heidelberg。2 Jarno Alanko、Nicola Cotumaccio 和 Nicola Prezza。Wheeler 确定有限自动机的线性时间最小化。载于 2022年数据压缩大会（DCC），第53–62页，2022年。 doi:10.1109/DCC52660.2022.00013。3 Jarno Alanko、Giovanna D'Agostino、Alberto Policriti 和 Nicola Prezza。正则语言遇上前缀排序，第911–930页。网址：https://epubs.siam.org/doi/abs/10.1137/1.9781611975994.55, arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611975994.55, doi:10.1137/1.9781611975994.55。4 Jarno Alanko、Giovanna D'Agostino、Alberto Policriti 和 Nicola Prezza。Wheeler 语言。《信息与计算》，281:104820，2021年。网址：https://www.sciencedirect.com/science/article/pii/S0890540121001504, doi:10.1016/j.ic.2021.104820。5 Amihood Amir、Moshe Lewenstein 和 Noa Lewenstein。超文本中的模式匹配。《算法杂志》，35(1):82–99，2000年。网址：https://www.sciencedirect.com/science/article/pii/S0196677499910635, doi:10.1006/jagm.1999.1063。6 Jasmijn A. Baaijens、Paola Bonizzoni、Christina Boucher、Gianluca Della Vedova、Yuri Pirola、Raffaella Rizzi 和 Jouni Sirén。计算图泛基因组学：数据结构及其应用教程。《自然计算》，21(1):81–108，2022年。 doi:10.1007/s11047-022-09882-6。7 Anton Bankevich、Sergey Nurk、Dmitry Antipov、Alexey A. Gurevich、Mikhail Dvorkin、Alexander S. Kulikov、Valery M. Lesin、Sergey I. Nikolenko、Son Pham、Andrey D. Prjibelski、Alexey V. Pyshkin、Alexander V. Sirotkin、Nikolay Vyahhi、Glenn Tesler、Max A. Alekseyev 和 Pavel A. Pevzner。SPAdes：一种新的基因组组装算法及其在单细胞测序中的应用。《计算生物学杂志》，19(5):455–477，2012年。PMID: 22506599。arXiv:https://doi.org/10.1089/cmb.2012.0021, doi:10.1089/cmb.2012.0021。8 Ruben Becker、Manuel Cáceres、Davide Cenzato、Sung-Hwan Kim、Bojana Kodric、Francisco Olivares 和 Nicola Prezza。通过划分细化对有限自动机进行排序。载于 Inge Li Gørtz、Martin Farach-Colton、Simon J. Puglisi 和 Grzegorz Herman 编，第31届欧洲算法年会（ESA 2023），莱布尼茨国际信息学论文集（LIPIcs）第274卷，第15:1–15:15页，达格施图尔，德国，2023年。Schloss Dagstuhl – Leibniz-Zentrum für Informatik。网址：https://drops.dagstuhl.de/opus/volltexte/2023/18668, doi:10.4230/LIPIcs.ESA.2023.15。9 Alexander Bowe、Taku Onodera、Kunihiko Sadakane 和 Tetsuo Shibuya。简洁 de Bruijn 图。载于 Ben Raphael 和 Jijun Tang 编，《生物信息学算法》，第225–235页，柏林，海德堡，2012年。Springer Berlin Heidelberg。10 M. Burrows 和 D. J. Wheeler。一种块排序无损数据压缩算法。技术报告，1994年。11 Alessio Conte、Nicola Cotumaccio、Travis Gagie、Giovanni Manzini、Nicola Prezza 和 Marinella Sciortino。在 Wheeler 确定有限自动机上计算匹配统计。载于 2023年数据压缩大会（DCC），第150–159页，2023年。 doi:10.1109/DCC55655.2023.00023。12 Nicola Cotumaccio。图可以用简洁方式建立索引，以便在 $o(|e|^2 + |v|^{5/2})$时间内进行模式匹配。载于 2022年数据压缩大会（DCC），第272–281页，2022年。 doi:10.1109/DCC52660.2022.00035。13 Nicola Cotumaccio。确定有限自动机的前缀排序：一种递归算法。载于 Satoru Iwata 和 Naonori Kakimura 编，第 34 届国际算法与计算研讨会（ ISAAC 2023 ），莱布尼茨国际信息学论文集（LIPIcs）第283卷，第22:1–22:15页，达格施图尔，德国，2023年。Schloss Dagstuhl – Leibniz-Zentrum für Informatik。网址：https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ISAAC.2023.22, doi:10.4230/LIPIcs.ISAAC.2023.22。

18 广义自动机的 Myhill-Nerode 定理及其应用

14 Nicola Cotumaccio, Giovanna D'Agostino, Alberto Policriti 和 Nicola Prezza。自动机与正则语言的共字典序排列——第一部分。《ACM杂志》，70(4)，2023年8月。 doi:10.1145/3607471。 15 Nicola Cotumaccio、Travis Gagie、Dominik Köppl 和 Nicola Prezza。Wheeler 确定有限自动机最长公共前缀数组的时空权衡。载于 Franco Maria Nardini、Nadia Pisanti 和 Rossano Venturini 编，字符串处理与信息检索，第143–156页，尚，2023年。施普林格·自然瑞士。 16 Nicola Cotumaccio 和 Nicola Prezza。有限自动机的索引与压缩。载于《第三十二届 ACM-SIAM 离散算法年会论文集》，SODA '21，第2585–2599页，美国，2021年。工业与应用数学学会。 17 Samuel Eilenberg。《自动机、语言与机器》。Academic Press, Inc.，美国，1974年。 18 Massimo Equi、Roberto Grossi、Veli Mäkinen 和 Alexandru I. Tomescu。图的字符串匹配复杂度。载于 Christel Baier、Ioannis Chatzigiannakis、Paola Flocchini 和 Stefano Leonardi 编，《第46届自动机、语言与编程国际研讨会》，ICALP 2019，2019年7月9日至12日，希腊帕特雷，第132卷 LIPIcs，第55:1–55:15页。Schloss Dagstuhl - Leibniz-Zentrum für Informatik，2019年。 doi:10.4230/LIPIcs.ICALP.2019.55。 19 Massimo Equi、Veli Mäkinen 和 Alexandru I. Tomescu。除非强指数时间假设不成立，否则无法在多项式时间内为实现次二次时间字符串匹配的图建立索引。载于 Tomáš Bureš、Riccardo Dondi、Johann Gamper、Giovanna Guerrini、Tomasz Jurdziński、Claus Pahl、Florian Sikora 和 Prudence W.H. Wong 编，《SOFSEM 2021：计算机科学理论与实践》，第608–622页，尚，2021年。Springer International Publishing。 20 Massimo Equi、Veli Mäkinen、Alexandru I. Tomescu 和 Roberto Grossi。图的字符串匹配复杂度。《算法学汇刊》，19(3)，2023年4月。 doi:10.1145/3588334。 21 Massimo Equi、Tuukka Norri、Jarno Alanko、Bastien Cazaux、Alexandru I. Tomescu 和 Veli Mäkinen。弹性创始图索引的算法与复杂度。载于 Hee-Kap Ahn 和 Kunihiko Sadakane 编，《第32届算法与计算国际研讨会》，ISAAC 2021，2021年12月6日至8日，日本福冈，第212卷 LIPIcs，第20:1–20:18页。Schloss Dagstuhl - Leibniz-Zentrum für Informatik，2021年。 doi:10.4230/LIPIcs.ISAAC.2021.20。 22 Massimo Equi、Tuukka Norri、Jarno Alanko、Bastien Cazaux、Alexandru I. Tomescu 和 Veli Mäkinen。创始图索引的算法与复杂度。《算法学》，85(6):1586–1623，2023年。 doi:10.1007/s00453-022-01007-w。 23 P. Ferragina 和 G. Manzini。机会型数据结构及其应用。载于《第41届计算机科学基础年会论文集》（FOCS'00），第390–398页，2000年。doi:10.1109/SFCS.2000.892127。 24 Paolo Ferragina 和 Giovanni Manzini。压缩文本的索引。《ACM杂志》，52(4):552–581，2005年7月。 doi:10.1145/1082036.1082039。 25 Travis Gagie、Giovanni Manzini 和 Jouni Sirén。Wheeler 图：基于 BWT 的数据结构框架。《理论计算机科学》，698:67–78，2017年。《大数据时代的算法、字符串与理论方法》（献给 Raffaele Giancarlo 教授60岁生日）。URL: https://www.sciencedirect.com/science/article/pii/S0304397517305285, doi:10.1016/j.tcs.2017.06.016。 26 Dora Giammarresi 和 Rosa Montalbano。确定性广义自动机。载于 Ernst W. Mayr 和 Claude Puech 编，《STACS 95，第12届计算机科学理论方面年会》，德国慕尼黑，1995年3月2日至4日，论文集，第900卷《计算机科学讲义》，第325–336页。Springer，1995年。 doi:10.1007/3-540-59042-0_84。 27 Dora Giammarresi 和 Rosa Montalbano。确定性广义自动机。《理论计算机科学》，215(1-2):191–208，1999年。 doi:10.1016/S0304-3975(97)00166-7。 28 Daniel Gibney 和 Sharma V. Thankachan。Wheeler 图识别的复杂度。《算法学》，84(3):784–814，2022年3月。 doi:10.1007/s00453-021-00917-5。

29 Kosaburo Hashiguchi。确定生成（接受）正则语言所需的最少非终结符（状态）数的算法。载于 Javier Leach Albert、Burkhard Monien 和 Mario Rodríguez Artalejo 编，《自动机、语言与编程》，第641–648页，柏林、海德堡，1991年。Springer Berlin Heidelberg。 30 John Hopcroft。用于最小化有限自动机状态数的 n log n 算法。载于 Zvi Kohavi 和 Azaria Paz 编，《机器与计算理论》，第189–196页。Academic Press，1971年。URL: https://www.sciencedirect.com/science/article/pii/B9780124177505500221, doi:10.1016/B978-0-12-417750-5.50022-1。 31 John E. Hopcroft、Rajeev Motwani 和 Jeffrey D. Ullman。《自动机理论、语言与计算导论》（第3版）。Addison-Wesley Longman Publishing Co., Inc.，美国，2006年。 32 Ramana M. Idury 和 Michael S. Waterman。一种新的 DNA 序列组装算法。《计算生物学杂志：计算分子细胞生物学杂志》，2 2:291–306，1995年。 33 Sung-Hwan Kim、Francisco Olivares 和 Nicola Prezza。确定有限自动机的更快前缀排序算法。载于 Laurent Bulteau 和 Zsuzsanna Lipták 编，《第34届组合模式匹配年度研讨会》，CPM 2023，2023年6月26日至28日，法国马恩拉瓦莱，第259卷 LIPIcs，第16:1–16:16页。Schloss Dagstuhl - Leibniz-Zentrum für Informatik，2023年。 doi:10.4230/LIPIcs.CPM.2023.16。 34 Donald E. Knuth、James H. Morris, Jr. 和 Vaughan R. Pratt。字符串的快速模式匹配。SIAM计算杂志，6(2):323–350，1977年。 arXiv:https://doi.org/10.1137/0206024, doi:10.1137/0206024。 35 Udi Manber 和 Sun Wu。针对文本与超文本、代价任意的近似字符串匹配。载于《结构与句法模式识别进展》，第22–33页。World Scientific，1992年。 36 Veli Mäkinen、Djamal Belazzougui、Fabio Cunial 和 Alexandru I. Tomescu。《基因组规模算法设计：高通量测序时代的生物信息学》。剑桥大学出版社，第2版，2023年。 37 Gonzalo Navarro。改进的超文本近似模式匹配。《理论计算机科学》，237(1–2):455–463，2000年4月。 doi:10.1016/S0304-3975(99)00333-3。 38 Gonzalo Navarro。《紧凑数据结构：实用方法》。剑桥大学出版社，2016年。 doi:10.1017/CBO9781316588284。 39 Kunsoo Park 和 Dong Kyue Kim。超文本中的字符串匹配。载于 Zvi Galil 和 Esko Ukkonen 编，《组合模式匹配》，第318–329页，柏林、海德堡，1995年。Springer Berlin Heidelberg。 40 Pavel A. Pevzner、Haixu Tang 和 Michael S. Waterman。一种用于 DNA 片段组装的欧拉路径方法。《美国国家科学院院刊》，98(17):9748–9753，2001年。URL: https://www.pnas.org/doi/abs/10.1073/pnas.171285098, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.171285098, doi:10.1073/pnas.171285098。 41 Mikko Rautiainen 和 Tobias Marschall。以 o(v + me) 时间将序列比对到一般图。bioRxiv，2017年。URL: https://www.biorxiv.org/content/early/2017/11/08/216127, arXiv:https://www.biorxiv.org/content/early/2017/11/08/216127.full.pdf, doi:10.1101/216127。 42 Nicola Rizzo 和 Veli Mäkinen。可索引弹性创始图的线性时间构建。载于 Cristina Bazgan 和 Henning Fernau 编，《组合算法》，第480–493页，尚，2022年。Springer International Publishing。 43 Jared T. Simpson 和 Richard Durbin。使用 FM 索引高效构建组装字符串图。《生物信息学》，26(12):367–373，2010年。 doi:10.1093/bioinformatics/btq217。

第 3 节的证明

引理 12 的陈述。设 $\mathcal{W} \subseteq \Sigma^*$，并令 $\sim$ 为$\mathcal{W}$上的一个等价关系。对于每个 $\alpha, \beta \in \mathcal{W}$，令:

$$\alpha \sim_r \beta \iff (\forall \phi \in \Sigma^*)((\alpha\phi \in \mathcal{W} \iff \beta\phi \in \mathcal{W}) \wedge (\alpha\phi \in \mathcal{W} \Longrightarrow \alpha\phi \sim \beta\phi)).$$

那么， $\sim_r$ 是$\mathcal{W}$上的一个等价关系，它是右不变的，并且是细化 $\sim$的$\mathcal{W}$ 上最粗的右不变等价关系。我们称 $\sim_r$ 为 $\sim$的右不变细化。

证明。显然， $\sim_r$ 是一个等价关系。我们来证明 $\sim_r$ 是右不变的。设 α、 $\beta \in$ W 和 $\phi \in \Sigma^*$ 满足 $\alpha \sim_r \beta$ 且 $\alpha\phi \in$ W。我们必须证明 $\beta\phi \in$ W 且 $\alpha\phi \sim_r \beta\phi$。现在， $\beta\phi \in$ W 可由 $\alpha \sim_r \beta$和 $\alpha\phi \in$ W 立即推出。此外， $\alpha\phi \sim \beta\phi$ 也可由 $\alpha \sim_r \beta$ 和 $\alpha\phi \in$ W 推出。接下来，固定一个满足 $\alpha\phi\psi \in$W 的 $\psi \in \Sigma^*$。我们必须证明 $\beta\phi\psi \in$W 且 $\alpha\phi\psi \sim \beta\phi\psi$。这可由 $\alpha \sim_r \beta$ 和 $\alpha\phi\psi \in$ W 推出。现在， $\sim_r$ 细化 $\sim$，因为对于任意 α、 $\beta \in$ W，若 $\alpha \sim_r \beta$，令 ϕ 为空串，便可得出 $\alpha \sim \beta$。最后，我们要证明 $\sim_r$ 是 W 上细化 $\sim$的最粗右不变等价关系。设 $\sim_*$ 是 W 上细化 $\sim$的一个右不变等价关系。假设对于某些 α、 $\beta \in$ W，有 $\alpha \sim_* \beta$。我们必须证明 $\alpha \sim_r \beta$。设 $\phi \in \Sigma^*$ 满足 $\alpha\phi \in$ W。我们必须证明 $\beta\phi \in$ W 且 $\alpha\phi \sim \beta\phi$。由于 $\sim_*$ 是右不变的，由 $\alpha \sim_* \beta$ 和 $\alpha\phi \in$W 可得 $\beta\phi \in$W 且 $\alpha\phi \sim_* \beta\phi$；这意味着 $\alpha\phi \sim \beta\phi$，因为 $\sim_*$ 细化 $\sim$。◀

引理 17 的陈述。设 $\mathcal{A} = (Q, E, s, F)$ 是一个广义非确定有限自动机。那么， $\equiv_{\mathcal{A}}$是右不变的，它细化了$\equiv_{\mathcal{L}(\mathcal{A}),\ \mathcal{W}(\mathcal{A})}$，它具有有限指数，并且 $\mathcal{L}(\mathcal{A})$ 是某些 $\equiv_{\mathcal{A}}$-等价类的并集。

证明。 首先， $\equiv_{\mathcal{A}}$ 是右不变的，因为根据定义，它是一个右不变细化。

我们来证明 $\mathcal{L}(\mathcal{A})$ 是某些 $\equiv_{\mathcal{A}}$-等价类的并集。只需证明 $\mathcal{L}(\mathcal{A})$ 是某些 $\sim_{\mathcal{A}}$-等价类的并集，因为 $\equiv_{\mathcal{A}}$ 是 $\sim_{\mathcal{A}}$ 的细化。设 α， $\beta \in \mathcal{W}(\mathcal{A})$ 满足 $\alpha \sim_{\mathcal{A}} \beta$，即 $I_\alpha = I_\beta$。我们必须证明 $\alpha \in \mathcal{L}(\mathcal{A})$ 当且仅当 $\beta \in \mathcal{L}(\mathcal{A})$。我们有 $\alpha \in \mathcal{L}(\mathcal{A})$ 当且仅当 $I_\alpha \cap F \neq \emptyset$，当且仅当 $I_\beta \cap F \neq \emptyset$，当且仅当 $\beta \in \mathcal{L}(\mathcal{A})$。

让我们证明 $\equiv_{\mathcal{A}}$ 细化 $\equiv_{\mathcal{L}(\mathcal{A}),\ \mathcal{W}(\mathcal{A})}$。由于 $\mathcal{L}(\mathcal{A})$ 是若干个 $\equiv_{\mathcal{A}}$-等价类的并集，因此 $\equiv_{\mathcal{A}}$ 细化 $\sim_{\mathcal{L}(\mathcal{A}),\ \mathcal{W}(\mathcal{A})}$。于是， $\equiv_{\mathcal{A}}$ 也细化 $\equiv_{\mathcal{L}(\mathcal{A}),\ \mathcal{W}(\mathcal{A})}$，因为 $\equiv_{\mathcal{A}}$ 是右不变的，并且 $\equiv_{\mathcal{L}(\mathcal{A}),\ \mathcal{W}(\mathcal{A})}$ 是细化 $\sim_{\mathcal{L}(\mathcal{A}),\ \mathcal{W}(\mathcal{A})}$的最粗右不变等价关系。

最后，我们证明 $\sim_{\mathcal{A}}$ 具有有限指标。设 A 是我们在第 2 节中构造的、与 A 等价的带有 ϵ-转移的非确定有限自动机。设 $\alpha, \beta \in$ W(A)，使得 $\alpha \sim$ A β。我们来证明它必定是 $\alpha \equiv_{\mathcal{A}} \beta$。根据右不变细化的定义，我们必须证明，对于每个 $\phi \in \Sigma^*$， $\alpha\phi \in$ W(A) 当且仅当 $\beta\phi \in$ W(A)；并且，如果$\alpha\phi \in$ W(A)，那么 $\alpha\phi \sim_{\mathcal{A}} \beta\phi$。由于 A 是带有 ϵ-转移的非确定有限自动机，我们知道 $\sim_{\bar{\mathcal{A}}}$ 是右不变的（参见定义 15 之前的讨论），因此，由于 $\alpha\phi \in$ W(A) $\subseteq$ Pref(L(A))，我们得到 $\alpha\phi \sim$ A $\beta\phi$，这意味着 $I_{\alpha\phi}$ A $= I_{\beta\phi}$ A；而这又（根据 A 的构造）意味着 $I^{\mathcal{A}}_{\alpha\phi} = I_{\alpha\phi}\text{A} \cap Q = I_{\beta\phi}\text{A} \cap Q = I^{\mathcal{A}}_{\beta\phi}$。由此可得， $\alpha\phi \in$ W(A) 当且仅当 $I^{\mathcal{A}}_{\alpha\phi} \neq \emptyset$，当且仅当 $I^{\mathcal{A}}_{\beta\phi} \neq \emptyset$，当且仅当 $\beta\phi \in$ W(A)；并且，如果 $\alpha\phi \in$ W(A)，那么 $\alpha\phi \sim_{\mathcal{A}} \beta\phi$。由于对于每个 α， $\beta \in$ W(A)，我们都有 $\alpha \sim$ A β 蕴含 $\alpha \equiv_{\mathcal{A}} \beta$，要证明 $\equiv_{\mathcal{A}}$ 具有有限指标，只需证明 $\sim_{\bar{\mathcal{A}}}$ 具有有限指标。这可由以下观察得出：每个 $\sim$ A 等价类都可以对应到 A 中所有状态构成的集合 Q 的一个互不相同的非空子集（通过良定义映射 $[\alpha]_{\sim \text{A}} \mapsto I_{\alpha}\text{A}$），因此指标至多为 $2^{|Q|} - 1$。

◀

引理 18 的陈述。设 $\mathcal{L} \subseteq \Sigma^*$，且设 $\mathcal{W} \subseteq \Sigma^*$ 是一个局部有界集，使得 $\mathcal{L} \cup \{\epsilon\} \subseteq \mathcal{W} \subseteq \mathrm{Pref}(\mathcal{L})$。假设 $\mathcal{L}$ 是 $\mathcal{W}$ 上的某个有限指数右不变等价关系 $\sim$ 的若干等价类的并。那么，$\mathcal{L}$ 可由一个 $\mathcal{W}$-广义确定有限自动机 $\mathcal{A}_\sim = (Q_\sim, E_\sim, s_\sim, F_\sim)$ 识别，使得：

1. $|Q_\sim|$ 等于 $\sim$ 的指标。2. $\equiv_{\mathcal{A}_\sim}$ 和 $\sim$ 是相同的等价关系。此外，如果 $\mathcal{B}$ 是一个 $\mathcal{W}$-广义确定有限自动机，且识别 $\mathcal{L}$，那么 $\mathcal{A}_{\equiv_\mathcal{B}}$ 与同构。

证明。定义 $\mathrm{A}_\sim = (Q_\sim, E_\sim, s_\sim, F_\sim)$ 如下。$Q_\sim = \{[\alpha]_\sim \mid \alpha \in \mathrm{W}\}$。$s_\sim = [\epsilon]_\sim$。$E_\sim = \{([\alpha]_\sim, [\alpha\rho]_\sim, \rho) \mid \alpha \in \mathrm{W},\ \rho \in \mathrm{K}(T_\alpha)\}$。$F_\sim = \{[\alpha]_\sim \mid \alpha \in \mathrm{L}\}$。

我们来证明 $\mathrm{A}_\sim$ 是一个定义良好的广义确定有限自动机。首先，状态的数量有限，因为 $\sim$ 的指标有限。其次，$\epsilon \in \mathrm{W}$，因此 $s_\sim$ 定义良好。集合 $F_\sim$ 定义良好，因为 (i) 若 $\alpha \in \mathrm{L}$，则 $\alpha \in \mathrm{W}$，因此 $[\alpha]_\sim$ 定义良好；且 (ii) L 是若干个 $\sim$-等价类的并，因此若对某个 α，$\alpha' \in \mathrm{W}$ 有 $\alpha \sim \alpha'$，则 $\alpha \in \mathrm{L}$ 当且仅当 $\alpha' \in \mathrm{L}$。我们来证明 $E_\sim$ 定义良好。首先，注意到 $\rho \in T_\alpha$，因此 $\alpha\rho \in \mathrm{W}$，且 $[\alpha\rho]_\sim$ 定义良好。此外，若 $\alpha' \in \mathrm{W}$ 满足 $\alpha \sim \alpha'$，则 $\alpha'\rho \in \mathrm{W}$ 且 $\alpha\rho \sim \alpha'\rho$，因为 $\sim$ 右不变。特别地，我们有 $T_\alpha = T'\alpha$，因此 $\mathrm{K}(T_\alpha) = \mathrm{K}(T'_\alpha)$，这说明 $E_\sim$ 定义良好。此外，$E_\sim$ 是有限集，因为 W 局部有界，所以对每个 $\alpha \in \mathrm{W}$，$\mathrm{K}(T_\alpha)$ 都是有限的。我们对 $|\alpha|$ 进行归纳，证明对于每个 $\alpha \in \mathrm{W}$ 状态 $[\alpha]_\sim$ 都可从初始状态到达。若 $|\alpha| = 0$，则 $\alpha = \epsilon$，因此 $[\alpha]_\sim$ 是初始状态。现在假设 $|\alpha| > 0$。令 α' 为 α 的最长真前缀，使得 $\alpha' \in \mathrm{W}$；它必定存在，因为 ϵ 是 α 的真前缀，这是因为 $|\alpha| > 0$，且 $\epsilon \in \mathrm{W}$。令 $\rho \in \Sigma^+$ 满足 $\alpha = \alpha'\rho$。我们来证明 $([\alpha']_\sim, [\alpha]_\sim, \rho) \in E_\sim$。由于 $\alpha = \alpha'\rho$，我们只需证明 $\rho \in \mathrm{K}(T_{\alpha'})$。我们有 $\rho \in T_{\alpha'}$，因为 $\alpha = \alpha'\rho \in \mathrm{W}$；并且我们有 $\rho \in \mathrm{K}(T_{\alpha'})$，因为若 $\rho' \in \Sigma^+$ 是 ρ 的真前缀，则 $\rho' \notin T_\alpha$，否则就会有 $\alpha'\rho \in \mathrm{W}$，而 $\alpha'\rho'$ 会是比 α' 更长的 α 的真前缀，这与 α' 的极大性矛盾。根据归纳假设，$[\alpha']_\sim$ 可从初始状态到达，因此 $[\alpha]_\sim$ 也可从初始状态到达，因为 $([\alpha']_\sim, [\alpha]_\sim, \rho) \in E_\sim$。我们来证明每个状态都是共可达的。固定 $\alpha \in \mathrm{W}$。我们将证明 $[\alpha]_\sim$ 要么是终态，要么可以到达某个终态。特别地，$\alpha \in$ 属于 $\mathrm{Pref(L)}$，因此存在 $\rho \in \Sigma^*$ 使得 $\alpha\rho \in \mathrm{L}$，从而 $\alpha\rho \in \mathrm{W}$。令 $\rho_1, \ldots, \rho_s$ 为 ρ 的所有不同前缀，且对每个 $1 \leq i \leq s$ 都有 $\alpha\rho_i \in \mathrm{W}$，其中对每个 $1 \leq i \leq s-1$，ρ_i 都是 ρ_{i+1} 的真前缀。注意，$s \geq 1$，$\rho_1 = \epsilon$ 且 $\rho_s = \rho$。用于证明可达性的同一论证表明 $([\alpha\rho_i]_\sim, [\alpha\rho_{i+1}]_\sim, \tau_i) \in E_\sim$，其中 $\tau_i \in \Sigma^+$ 满足对每个 $1 \leq i \leq s-1$ 都有 $\rho_{i+1} = \rho_i\tau_i$。由于 $\alpha\rho_s = \alpha\rho \in \mathrm{L}$，因此 $[\alpha\rho_s]_\sim$ 是终态，所以对每个 $1 \leq i \leq s$，$[\alpha\rho_i]_\sim$ 要么是终态，要么可以到达某个终态；结论由此得出，因为 $\alpha = \alpha\rho_1$。最后，对每个 $\alpha \in \mathrm{W}$，每个 $\rho \in \Sigma^+$ 至多标记一条离开 $[\alpha]_\sim$ 的边，因为我们已经证明，$E_\sim$ 的定义不依赖于等价类中代表元的选择；并且所有标记离开 $[\alpha]_\sim$ 的边的字符串构成的集合是 $\mathrm{K}(T_\alpha)$，它是前缀无关的，因为它是一个前缀无关核。

接下来，我们要证明对于每个 $\alpha \in \Sigma^*$ 以及每个 $\beta \in \mathcal{W}$，都有：

$$(\alpha \in \mathcal{W}(\mathcal{A}_\sim)) \wedge (I_\alpha = [\beta]_\sim) \iff (\alpha \in \mathcal{W}) \wedge (\alpha \sim \beta). \tag{1}$$

我们对 $|\alpha|$ 使用归纳法。若 $|\alpha| = 0$，则 $\alpha = \epsilon$。注意，$\epsilon \in \mathcal{W}(\mathcal{A}_\sim)$，因为 $\epsilon \in I_{s_\sim}$，且 $\epsilon \in \mathcal{W}$。此外，对于每个 $\beta \in \mathcal{W}$，都有 $I_\epsilon = [\beta]_\sim \iff [\beta]_\sim = s_\sim \iff \epsilon \sim \beta$。

现在，假设 $|\alpha| > 0$。注意，归纳假设意味着，对于每个满足 $|\alpha'| < |\alpha|$ 的 $\alpha' \in \Sigma^*$，都有 $\alpha' \in \mathcal{W}(\mathcal{A}_\sim) \iff \alpha' \in \mathcal{W}$；并且，如果 $\alpha' \in \mathcal{W}(\mathcal{A}_\sim)$，那么 $I_{\alpha'} = [\alpha']_\sim$。

事实上，(⇒) 可通过在公式 1 中任取满足 $I_{\alpha'} = [\beta]_\sim$ 的 $\beta \in \mathcal{W}$ 得出（这样的 $\beta \in \mathcal{W}$ 存在，因为$\alpha' \in \mathcal{W}(\mathcal{A}_\sim)$）；(⇐) 和 $I_{\alpha'} = [\alpha']_\sim$ 则可通过在公式 1 中取 $\beta = \alpha'$ 得出。

令 α' 为 α 的最长严格前缀，使得 $\alpha' \in \mathcal{W}(\mathcal{A}_\sim)$；该前缀必定存在，因为 ϵ 是 α 的真前缀，即 $|\alpha| > 0$，且 $\epsilon \in \mathcal{W}(\mathcal{A}_\sim)$。由于我们已证明，长度小于 α 的字符串属于 $\mathcal{W}(\mathcal{A}_\sim)$ 当且仅当它属于 $\mathcal{W}$，因此 α' 也是 α 的最长严格前缀，使得 $\alpha' \in \mathcal{W}$。此外，它必定是 $I_{\alpha'} = [\alpha']_\sim$。记 $\alpha = \alpha'\rho$，其中 $\rho \in \Sigma^+$。

1. 假设 $(\alpha \in \mathrm{W}(\mathrm{A}_\sim)) \wedge (I_\alpha = [\beta]_\sim)$。我们必须证明 $(\alpha \in \mathrm{W}) \wedge (\alpha \sim \beta)$。由于 α'，$\alpha \in \mathrm{W}(\mathrm{A}_\sim)$，$\alpha'$ 是 $\mathrm{W}(\mathrm{A}_\sim)$ 中 α 的最长严格前缀，$\alpha = \alpha'\rho$，且 $I_{\alpha'} = [\alpha']_\sim$，因此根据注 9，我们得到 $([\alpha']_\sim, [\beta]_\sim, \rho) \in E_\sim$。$E_\sim$ 的定义意味着 $\alpha = \alpha'\rho \in \mathrm{W}$ 且 $\alpha \sim \beta$。

2. 假设 $(\alpha \in \mathrm{W}) \wedge (\alpha \sim \beta)$。我们必须证明 $(\alpha \in \mathrm{W}(\mathrm{A}_\sim)) \wedge (I_\alpha = [\beta]_\sim)$。证明 $([\alpha']_\sim, [\alpha]_\sim, \rho) \in E_\sim$。由于 α'，$\alpha \in \mathrm{W}$，我们只需证明 $\rho \in \mathrm{K}(T_{\alpha'})$。由于 $\alpha = \alpha'\rho$，我们有 $\rho \in T_{\alpha'}$。如果 $\rho' \in \Sigma^+$ 是 ρ的真前缀，那么根据 α'的极大性，$\alpha'\rho' \in \mathrm{W}$ 不可能成立，因此 $\rho \in \mathrm{K}(T_{\alpha'})$。由 $I_{\alpha'} = [\alpha']_\sim$，$([\alpha']_\sim, [\alpha]_\sim, \rho) \in E_\sim$

and $\alpha = \alpha'\rho$ 我们btain $\alpha \in W(A_\sim)$ and $I_\alpha = [\alpha]_\sim$. Si一次 $\alpha \sim \beta$, 我们得出结论 lude $I_\alpha = [\beta]_\sim$。至此，公式 1 的证明完成。特别是，利用归纳步骤开头所用的同一论证，我们得到 $\mathcal{W}(\mathcal{A}_\sim) = \mathcal{W}$，这证明$\mathcal{A}_\sim$ 是一个$\mathcal{W}$-广义确定有限自动机，并且对于每个 $\alpha \in \mathcal{W}(\mathcal{A}_\sim) = \mathcal{W}$:

$$I_\alpha = [\alpha]_\sim. \tag{2}$$

根据 $F_\sim$ 的定义和公式 1，我们得到:

$$\mathcal{L}(\mathcal{A}_\sim) = \{\alpha \in \mathcal{W}(\mathcal{A}_\sim) \mid I_\alpha = [\beta]_\sim \text{ for some } \beta \in \mathcal{L}\} = \{\alpha \in \mathcal{W} \mid \alpha \sim \beta \text{ for some } \beta \in \mathcal{L}\} = \mathcal{L}$$

其中，在最后一个等式中，我们有 (⊆)，因为$\mathcal{L}$ 是若干个 $\sim$-等价类的并；并且有 (⊇)，因为 $\mathcal{L} \subseteq \mathcal{W}$。这证明$\mathcal{A}_\sim$ 识别$\mathcal{L}$。此外:

1. A 的状态数 $\mathcal{A}_\sim$ 显然等于 $\sim$的指标。2. 根据公式 2，对每个 α，$\beta \in \mathcal{W} = \mathcal{W}(\mathcal{A}_\sim)$ 我们有 $\alpha \equiv_{A_\sim} \beta \iff I_\alpha = I_\beta \iff [\alpha]_\sim = [\beta]_\sim \iff \alpha \sim \beta$，因此 $\equiv_{A_\sim}$ 和 $\sim$ 是同一个等价关系。

最后，假设 $\mathrm{B} = (Q_\mathcal{B}, E_\mathcal{B}, s_\mathcal{B}, F_\mathcal{B})$ 是一个识别 L 的 W-GDFA。注意，根据引理 17，我们有 $\equiv_\mathcal{B}$ 是右不变的，它具有有限指数，并且 L 是若干个 $\equiv_\mathcal{B}$-等价类的并集，因此 $\mathrm{A}_{\equiv_\mathcal{B}}$ 有良好定义，且 $\mathrm{W}(\mathrm{A}_{\equiv_\mathcal{B}}) = \mathrm{W} = \mathrm{W}(\mathrm{B})$。令 $\phi : Q_{\equiv_\mathcal{B}} \to Q_\mathcal{B}$ 为一个函数，它将 $Q_{\equiv_\mathcal{B}}$ 的状态 $[\alpha]_{\equiv_\mathcal{B}}$ 映射到 $Q_\mathcal{B}$ 的状态 $I^\mathcal{B}_\alpha$，也就是说，对每个 $\alpha \in \mathrm{W}$，$\phi([\alpha]_{\equiv_\mathcal{B}}) = I^\mathcal{B}_\alpha$。注意，$\phi$ 有良好定义且为单射，因为根据 $\equiv_\mathcal{B}$ 的定义，对于每个 $\alpha, \beta \in \mathrm{W}$，我们有 $[\alpha]_{\equiv_\mathcal{B}} = [\beta]_{\equiv_\mathcal{B}} \iff \alpha \equiv_\mathcal{B} \beta \iff I^\mathcal{B}_\alpha = I^\mathcal{B}_\beta$。下面证明 ϕ 确定了 $\mathrm{A}_{\equiv_\mathcal{B}}$ 与 B 之间的一个同构。首先，ϕ 是满射，因为 B 的每个状态都可以从初始状态到达。接下来，$\phi(s_{\equiv_\mathcal{B}}) = \phi([\epsilon]_{\equiv_\mathcal{B}}) = I^\mathcal{B}_\epsilon = s_\mathcal{B}$。此外，对于每个 $\alpha \in \mathrm{W}$，我们有 $[\alpha]_{\equiv_\mathcal{B}} \in F_{\equiv_\mathcal{B}} \iff \alpha \in \mathrm{L} \iff I^\mathcal{B}_\alpha \in F_\mathcal{B}$。最后，根据注 9，对于每个$\alpha, \beta \in \mathrm{W}$ 以及每个 $\rho \in \Sigma^+$:

$$\begin{aligned}([\alpha]_{\equiv_\mathcal{B}}, [\beta]_{\equiv_\mathcal{B}}, \rho) \in E_{\equiv_\mathcal{B}} &\iff \rho \in \mathcal{K}(T_\alpha) \wedge \alpha\rho \equiv_\mathcal{B} \beta \iff \\ &\iff \alpha\rho \in \mathcal{W} \wedge (I^\mathcal{B}_\alpha, I^\mathcal{B}_{\alpha\rho}, \rho) \in E_\mathcal{B} \wedge I^\mathcal{B}_{\alpha\rho} = I^\mathcal{B}_\beta \iff \\ &\iff (I^\mathcal{B}_\alpha, I^\mathcal{B}_\beta, \rho) \in E_\mathcal{B} \iff (\phi([\alpha]_{\equiv_\mathcal{B}}), \phi([\beta]_{\equiv_\mathcal{B}}), \rho) \in E_\mathcal{B}\end{aligned}$$

由此我们得出结论：$\mathcal{A}_{\equiv_\mathcal{B}}$ 和$\mathcal{B}$ 同构。 ◀

B 第 4 节的证明

引理 22 的陈述。设 $\mathcal{A} = (Q, E, s, F)$ 是一个 Wheeler 广义确定有限自动机。则:

1. s 在全序 $\preceq_{\mathcal{A}}$ 中排在最前。2. 对于每个 (u', u, ρ)、$(v', v, \rho') \in E$，如果 $u \prec_{\mathcal{A}} v$ 且 ρ' 不是 ρ 的真后缀，则 $\rho \preceq \rho'$。3. 对于每个 (u', u, ρ)、$(v', v, \rho) \in E$，如果 $u \prec_{\mathcal{A}} v$，则 $u' \prec_{\mathcal{A}} v'$。

证明。1. 设 $u \in Q \setminus \{s\}$。我们必须证明 $s \prec_{\mathcal{A}} u$。由于 $\preceq_{\mathcal{A}}$ 是全序，我们只需证明：若 s 和 u 是 $\preceq_{\mathcal{A}}$-可比较的，则必有 $s \prec_{\mathcal{A}} u$。这是因为对于每个 $\alpha \in I_u$，都有 $\epsilon \prec \alpha$，且 $\epsilon \in I_s$。2. 如果 $\rho = \rho'$ 或 ρ 是 ρ' 的真后缀，则结论显然成立，因此我们可以假设 $\rho \neq \rho'$，ρ 不是 ρ' 的真后缀，且 ρ' 不是 ρ 的真后缀。设 $\alpha' \in I_{u'}$ 和$\beta' \in I_{v'}$。于是，$\alpha'\rho \in I_u$ 且 $\beta'\rho' \in I_v$。由于 $u \prec_{\mathcal{A}} v$，因此 $\alpha'\rho \prec \beta'\rho'$。由于 $\rho \neq \rho'$，ρ 不是 ρ' 的真后缀，且 ρ' 不是 ρ 的真后缀，因此我们得出 $\rho \prec \rho'$。3. 设 $\alpha' \in I_{u'}$ 且 $\beta' \in I_{v'}$。我们必须证明 $\alpha' \prec \beta'$。由 (u', u, ρ)、$(v', v, \rho) \in E$，可得 $\alpha'\rho \in I_u$ 和 $\beta'\rho \in I_v$，因此由 $u \prec_{\mathcal{A}} v$ 可得 $\alpha'\rho \prec \beta'\rho$，进而得到 $\alpha' \prec \beta'$。

◀

引理25的陈述。设 $\mathcal{A} = (Q, E, s, F)$ 是一个 Wheeler 广义确定有限自动机，且 $\alpha \in \Sigma^*$。那么：

1. $G^{\prec}(\alpha) \cap G_{\dashv}(\alpha) = \emptyset$。 2. $G_{\dashv}(\alpha)$ 是 $\preceq_{\mathcal{A}}$-凸的。 3. 如果 $u, v \in Q$ 满足 $u \prec_{\mathcal{A}} v$ 且 $v \in G^{\prec}(\alpha)$，那么 $u \in G^{\prec}(\alpha)$。换句话说，$G^{\prec}(\alpha) = Q[1, |G^{\prec}(\alpha)|]$。 4. 如果 $u, v \in Q$ 满足 $u \prec_{\mathcal{A}} v$ 且 $v \in G^{\preceq}_{\dashv}(\alpha)$，那么 $u \in G^{\preceq}_{\dashv}(\alpha)$。换句话说，$G^{\preceq}_{\dashv}(\alpha) = Q[1, |G^{\preceq}_{\dashv}(\alpha)|]$。

证明。1. 若 $u \in G_{\dashv}(\alpha)$，则存在 $\beta \in I_u$ 使得 $\alpha \dashv \beta$。特别是，$\alpha \preceq \beta$，因此$u \notin G^{\prec}(\alpha)$。

2. 假设 $u, v, z \in Q$ 满足 $u \prec_{\mathcal{A}} v \prec_{\mathcal{A}} z$和 $u, z \in G_{\dashv}(\alpha)$。我们必须证明 $v \in G_{\dashv}(\alpha)$。由于 $u, z \in G_{\dashv}(\alpha)$，因此存在 $\beta \in I_u$ 和 $\delta \in I_z$，使得 $\alpha \dashv \beta$和 $\alpha \dashv \delta$。任取 $\gamma \in I_v$；我们只需证明 $\alpha \dashv \gamma$。由 $u \prec_{\mathcal{A}} v \prec_{\mathcal{A}} z$ 可得 $\beta \prec \gamma \prec \delta$。因此，由 $\alpha \dashv \beta$ 和 $\alpha \dashv \delta$ 可得 $\alpha \dashv \gamma$。

3. 设 $\beta \in I_u$。我们必须证明 $\beta \prec \alpha$。任取 $\gamma \in I_v$。由于 $u \prec_{\mathcal{A}} v$，因此有 $\beta \prec \gamma$。由 $v \in G^{\prec}(\alpha)$ 可得 $\gamma \prec \alpha$，因此我们得出 $\beta \prec \alpha$。
4. 由于 $v \in G^{\preceq}_{\dashv}(\alpha)$，因此 $v \in G^{\prec}(\alpha)$ 或 $v \in G_{\dashv}(\alpha)$ 至少有一个成立。若 $v \in G^{\prec}(\alpha)$，则由前一点可得 $u \in G^{\prec}(\alpha)$，从而 $u \in G^{\preceq}_{\dashv}(\alpha)$。现在假设 $v \in G_{\dashv}(\alpha)$。若 $u \in G_{\dashv}(\alpha)$，则 $u \in G^{\preceq}_{\dashv}(\alpha)$，证明即告完成，因此我们可以假设 $u \notin G_{\dashv}(\alpha)$。我们来证明必有 $u \in G^{\prec}(\alpha)$，这同样意味着 $u \in G^{\preceq}_{\dashv}(\alpha)$。任取 $\beta \in I_u$；我们必须证明$\beta \prec \alpha$。由于 $v \in G_{\dashv}(\alpha)$，存在 $\gamma \in \Sigma^*$ 使得 $\gamma\alpha \in I_v$。由 $u \prec_{\mathcal{A}} v$ 可得 $\beta \prec \gamma\alpha$。由于 $u \notin G_{\dashv}(\alpha)$ 蕴含 $\neg(\alpha \dashv \beta)$，因此由 $\beta \prec \gamma\alpha$ 可得 $\beta \prec \alpha$。

◀

▶ **引理 32**。 设 $A = (Q, E, s, F)$ 是一个 *Wheeler* 广义确定有限自动机，且 $\alpha \in \Sigma^*$。那么，$u \in G^{\prec}(\alpha)$当且仅当满足：*(i)* $max_{\lambda(u)} \prec \alpha$；*(ii)* 若 $u' \in Q$ 满足：对于某个 $0 < k < |\alpha|$，$(u', u, s(\alpha, k)) \in E$，则 $u' \in G^{\prec}(p(\alpha, |\alpha| - k))$。

证明。($\Rightarrow$)。我们来证明 (i)。取 $\rho \in \lambda(u)$。于是，存在 $u' \in Q$，使得 $(u', u, \rho) \in E$。我们必须证明 $\rho \prec \alpha$。令 $\beta' \in I_{u'}$；由 $(u', u, \rho) \in E$可得$\beta'\rho \in I_u$，因此由 $u \in G^{\prec}(\alpha)$ 可得 $\beta'\rho \prec \alpha$，这意味着 $\rho \prec \alpha$。我们来证明 (ii)。假设 $u' \in Q$ 满足：对于某个 $0 < k < |\alpha|$，$(u', u, s(\alpha, k)) \in E$。我们必须证明 $u' \in G^{\prec}(p(\alpha, |\alpha| - k))$。取 $\gamma' \in I_{u'}$。我们必须证明 $\gamma' \prec p(\alpha, |\alpha| - k)$。由于 $(u', u, s(\alpha, k)) \in E$，因此有 $\gamma' s(\alpha, k) \in I_u$；于是由 $u \in G^{\prec}(\alpha)$ 可得 $\gamma' s(\alpha, k) \prec \alpha$，这意味着 $\gamma' \prec p(\alpha, |\alpha| - k)$。

24 广义自动机的 Myhill-Nerode 定理及其应用

(⇐) 令 $\beta \in I_u$。我们必须证明 $\beta \prec \alpha$。根据性质 (i)，我们有 $\epsilon \preceq \max_{\lambda(u)} \prec \alpha$，因此若 $\beta = \epsilon$，结论显然成立。现在，假设 $\beta \neq \epsilon$。这意味着存在 $u' \in Q$，使得对于某个 $0 < k \leq |\beta|$，有 $(u', u, s(\beta, k)) \in E$ 且 $p(\beta, |\beta| - k) \in I_{u'}$。根据性质 (i)，我们知道 $s(\beta, k) \prec \alpha$。如果 $\neg(s(\beta, k) \dashv \beta)$，那么 $\beta \prec \alpha$，证明即告完成。否则，有 $s(\beta, k) = s(\alpha, k)$，并且 $0 < k < |\alpha|$（否则 $k = |\alpha|$，且 $\alpha = s(\alpha, k) = s(\beta, k) \prec \alpha$）。因此，根据性质 (ii)，我们得到 $u' \in G^{\prec}(p(\alpha, |\alpha| - k))$。由于$p(\beta, |\beta| - k) \in I_{u'}$，我们得出 $p(\beta, |\beta| - k) \prec p(\alpha, |\alpha| - k)$，这意味着 $\beta \prec \alpha$。◀

引理 27 的陈述。设 $\mathcal{A} = (Q, E, s, F)$ 是一个 Wheeler 广义确定有限自动机，且令 $\alpha \in \Sigma^*$。对于$0 < k < |\alpha|$，令 $f_k = \mathtt{out}(Q[1, |G^{\prec}(p(\alpha, |\alpha| - k))|], s(\alpha, k))$。那么，$|G^{\prec}(\alpha)|$ 是满足以下条件的最大整数 $0 \leq j \leq |Q|$：(i) 对每个 $0 < k < |\alpha|$，$\mathtt{in}(Q[1, j], s(\alpha, k)) \leq f_k$；且 (ii) 若$j \geq 1$，则 $\max_{\lambda(Q[j])} \prec \alpha$。

证明。首先，注意到根据引理 25，我们有 $G^{\prec}(\alpha) = Q[1, |G^{\prec}(\alpha)|]$，并且对于每个 $0 < k < \alpha$，都有 $G^{\prec}(p(\alpha, |\alpha| - k)) = Q[1, |G^{\prec}(p(\alpha, |\alpha| - k)|]$。令 j^* 为满足以下条件的最大整数$0 \leq j \leq |Q|$：(i) 对每个 $0 < k < |\alpha|$，$\mathtt{in}(Q[1, j], s(\alpha, k)) \leq f_k$；且 (ii) 若 $j \geq 1$，则 $\max_{\lambda(Q[j])} \prec \alpha$。我们要证明 $|G^{\prec}(\alpha)| = j^*$。

(≤) 若 $|G^{\prec}(\alpha)| = 0$，结论显然成立，因此我们可以假设 $|G^{\prec}(\alpha)| \geq 1$。只需证明，对于每个 $0 < k < |\alpha|$，都有 $\mathtt{in}(Q[1, |G^{\prec}(\alpha)|], s(\alpha, k)) \leq f_k$，且 $\max_{\lambda(Q[|G^{\prec}(\alpha)|])} \prec \alpha$。这由引理 32 和 f_k 的定义可得。

(≥) 我们只需证明：若 $|G^{\prec}(\alpha)| + 1 \leq j \leq |Q|$，则要么存在某个 $0 < k < |\alpha|$，使得 $\mathtt{in}(Q[1, j], s(\alpha, k)) > f_k$；要么 $a \preceq \max^{\lambda(Q[j])}$。假设 $\max_{\lambda(Q[j])} \prec \alpha$。由于 $Q[j] \notin G^{\prec}(\alpha)$，根据引理 32，存在某个 $0 < k < |\alpha|$，使得 $(v', Q[j], s(\alpha, k)) \in E$ 且 $v' \notin Q[1, |G^{\prec}(p(\alpha, |\alpha| - k))|]$。若证明如下命题，我们便可得出 $\mathtt{in}(Q[1, j], s(\alpha, k)) > f_k$：若 $u', u \in Q$ 满足 $u' \in Q[1, |G^{\prec}(p(\alpha, |\alpha| - k))|]$ 且 $(u', u, s(\alpha, k)) \in E$，则$u \in Q[1, j]$。反设 $u \notin Q[1, j]$。这意味着 $Q[j] \prec_{\mathcal{A}} u$，因此由 $(v', Q[j], s(\alpha, k))$、$(u', u, s(\alpha, k)) \in E$ 以及引理 22 可得 $v' \prec_{\mathcal{A}} u'$。由于$u' \in Q[1, |G^{\prec}(p(\alpha, |\alpha| - k))|]$，我们得出 $v' \in Q[1, |G^{\prec}(p(\alpha, |\alpha| - k))|]$，矛盾。◀

引理 28 的陈述。设 $\mathrm{A} = (Q, E, s, F)$ 是一个 Wheeler 广义确定有限自动机，且 $\alpha \in \Sigma^*$。对于 $0 < k < |\alpha|$，设 $f_k = \mathtt{out}(Q[1, |G^{\prec}(p(\alpha, |\alpha| - k))|], s(\alpha, k))$ 和 $g_k = \mathtt{out}(Q[1, |G^{\preceq}_{\dashv}(p(\alpha, |\alpha| - k))|], s(\alpha, k))$。那么，对于每个 $0 < k < |\alpha|$，$g_k \geq f_k$。此外，$|G^{\preceq}_{\dashv}(\alpha)|$ 等于以下各项中的最大值：

1. $|G^{\prec}(\alpha)|$；**2.** 满足以下条件的最大整数 $0 \leq i \leq |Q|$：若 $i \geq 1$，则 $Q[i] \in G^*(\alpha)$；**3.** 满足以下条件的最小整数 $0 \leq j \leq |Q|$：对于每个 $0 < k < |\alpha|$，若 $g_k > f_k$，则有 $\mathtt{in}(Q[1, j], s(\alpha, k)) \geq g_k$。

证明。首先注意，根据引理 25，我们有 $G^{\prec}(\alpha) = Q[1, |G^{\prec}(\alpha)|], G^{\preceq}_{\dashv}(\alpha) = Q[1, |G^{\preceq}_{\dashv}(\alpha)|]$，并且对于每个 $0 < k < \alpha$，都有 $G^{\prec}(p(\alpha, |\alpha| - k)) = Q[1, |G^{\prec}(p(\alpha, |\alpha| - k)|]$ 且$G^{\preceq}_{\dashv}(p(\alpha, |\alpha| - k)) = Q[1, |G^{\preceq}_{\dashv}(p(\alpha, |\alpha| - k)|]$。对于每个 $0 < k < \alpha$，都有 $g_k \geq f_k$，因为 $G^{\prec}(p(\alpha, |\alpha| - k) \subseteq G^{\preceq}_{\dashv}(p(\alpha, |\alpha| - k)$。令 i^* 为满足以下条件的最大整数：若 $i \geq 1$，则 $Q[i] \in G^*(\alpha)$；再令 j^* 为满足以下条件的最小整数：对于每个满足 $g_k > f_k$的 $0 < k < |\alpha|$，都有$\mathtt{in}(Q[1, j], s(\alpha, k)) \geq g_k$。令 $l = \max(|G^{\prec}(\alpha)|, i^*, j^*)$。我们要证明 $l = |G^{\preceq}_{\dashv}(\alpha)|$。

(≤) 我们需要证明 $|G^{\prec}(\alpha)| \leq |G^{\preceq}_{\dashv}(\alpha)|$，$i^* \leq |G^{\preceq}_{\dashv}(\alpha)|$ 以及 $j^* \leq |G^{\preceq}_{\dashv}(\alpha)|$。不等式 $|G^{\prec}(\alpha)| \leq |G^{\preceq}_{\dashv}(\alpha)|$ 由 $G^{\prec}(\alpha) \subseteq G^{\preceq}_{\dashv}(\alpha)$ 得出。我们来证明 $i^* \leq |G^{\preceq}_{\dashv}(\alpha)|$。若 $i^* = 0$，则证明完成，因此可以假设 $i^* \geq 1$。我们知道 $Q[i^*] \in G^*(\alpha)$，所以 $Q[i^*] \in G_{\dashv}(\alpha)$ 且 $Q[i^*] \in G^{\preceq}_{\dashv}(\alpha)$，这意味着 $i^* \leq |G^{\preceq}_{\dashv}(\alpha)|$。我们来证明 $j^* \leq |G^{\preceq}_{\dashv}(\alpha)|$。若 $j^* = 0$，则证明完成，因此可以假设 $j^* \geq 1$。我们知道存在 $0 < k < |\alpha|$，使得

$g_k > f_k$，使得 $\texttt{in}(Q[1, j^* - 1], s(\alpha, k)) < g_k$，但 $\texttt{in}(Q[1, j^*], s(\alpha, k)) \geq g_k$。我们只需证明存在 $u' \in G_\dashv(p(\alpha, |\alpha| - k))$，使得 $(u', Q[j^*], s(\alpha, k)) \in E$，因为这将推出 $Q[j^*] \in G_\dashv(\alpha)$，从而 $Q[j^*] \in G_\dashv^\prec(\alpha)$ 且 $j^* \leq |G_\dashv^\prec(\alpha)|$。反设对每个满足 $(u', Q[j^*], s(\alpha, k)) \in E$ 的 $u' \in Q$，都有 $u' \notin G_\dashv(p(\alpha, |\alpha| - k))$。首先，我们来证明不可能存在 $u'_1 \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$ 和 $u'_2 \in Q[|G_\dashv^\prec(p(\alpha, |\alpha| - k))| + 1, |Q|]$，使得 $(u'_1, Q[j^*], s(\alpha, k)) \in E$ 且 $(u'_2, Q[j^*], s(\alpha, k)) \in E$。反设这样的 u'_1 和 u'_2 存在。我们将通过证明不存在 $u'_3 \in G_\dashv(p(\alpha, |\alpha| - k))$，使得对某个 $u_3 \in Q$ 有 $(u'_3, u_3, s(\alpha, k)) \in E$，来导出矛盾，因为这将意味着 $g_k = f_k$，与 $g_k > f_k$ 矛盾。事实上，(i) 这不可能是 $u_3 = Q[j]$，因为我们知道，对每个满足 $(u', Q[j^*], s(\alpha, k)) \in E$ 的 $u' \in Q$，都有 $u' \notin G_\dashv(p(\alpha, |\alpha| - k))$；(ii) 这不可能是 $u_3 \prec_{\mathcal{A}} Q[j]$，因为根据引理 22，我们将得到 $u'_3 \prec_{\mathcal{A}} u'_1$，且由于 $u'_1 \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$，我们将得出 $u'_3 \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$；(iii) 这不可能是 $Q[j] \prec_{\mathcal{A}} u_3$，因为根据引理 22，我们将得到 $u'_2 \prec_{\mathcal{A}} u'_3$，且由于 $u'_2 \in Q[|G_\dashv^\prec(p(\alpha, |\alpha| - k))| + 1, |Q|]$，我们将得出 $u'_3 \in Q[|G_\dashv^\prec(p(\alpha, |\alpha| - k))| + 1, |Q|]$。因此，以下两种情况必有一种成立：

1. 对于每个 $u'_1 \in Q$，只要 $(u'_1, Q[j^*], s(\alpha, k)) \in E$，我们就有 $u'_1 \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$。下面证明，若 $v', v \in Q$ 满足 $v \in Q[1, j^*]$，且 $(v', v, s(\alpha, k)) \in E$，则 $v' \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$。由此可推出 $\texttt{in}(Q[1, j^*], s(\alpha, k)) \leq f_k < g_k$，这正是我们所需的矛盾。若 $v = Q[j^*]$，结论便成立，因为我们知道，对每个 $u'_1 \in Q$，只要 $(u'_1, Q[j^*], s(\alpha, k)) \in E$，就有 $u'_1 \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$。现在假设 $v \prec_{\mathcal{A}} Q[j]$。由于 $\texttt{in}(Q[1, j^* - 1], s(\alpha, k)) < g_k$，但 $\texttt{in}(Q[1, j^*], s(\alpha, k)) \geq g_k$，因此存在 $u_1^* \in Q$，使得 $(u_1^*, Q[j^*], s(\alpha, k)) \in E$；并且我们知道，它必定满足 $u_1^* \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$。由 $v \prec_{\mathcal{A}} Q[j]$ 和引理 22 可得 $v' \prec_{\mathcal{A}} u_1^*$；又因为 $u_1^* \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$，所以我们得出 $v' \in Q[1, |G^\prec(p(\alpha, |\alpha| - k))|]$。

2. 对于每个 $u'_2 \in Q$，只要满足 $(u'_2, Q[j^*], s(\alpha, k)) \in E$，我们就有 $u'_2 \in Q[|G_\dashv^\prec(p(\alpha, |\alpha| - k))| + 1, |Q|]$。下面证明，若 $v', v \in Q$ 满足 $v' \in Q[1, |G_\dashv^\prec((p(\alpha, |\alpha| - k))|]$ 且 $(v', v, s(\alpha, k)) \in E$，则 $v \in Q[1, j^* - 1]$。这将推出 $\texttt{in}(Q[1, j^* - 1], s(\alpha, k)) \geq g_k$，从而得到所需的矛盾。$v = Q[j^*]$ 不可能成立，因为这将推出 $v' \in Q[|G_\dashv^\prec(p(\alpha, |\alpha| - k))| + 1, |Q|]$。现在，为了导出矛盾，假设 $Q[j^*] \prec_{\mathcal{A}} v$。由于 $\texttt{in}(Q[1, j^* - 1], s(\alpha, k)) < g_k$，但 $\texttt{in}(Q[1, j^*], s(\alpha, k)) \geq g_k$，因此存在 $u_2^* \in Q$，使得 $(u_2^*, Q[j^*], s(\alpha, k)) \in E$，并且我们知道必有 $u_2^* \in Q[|G_\dashv^\prec(p(\alpha, |\alpha| - k))| + 1, |Q|]$。由 $Q[j^*] \prec_{\mathcal{A}} v$ 和引理 22 可得 $u_2^* \prec_{\mathcal{A}} v'$；又因为 $u_2^* \in Q[|G_\dashv^\prec(p(\alpha, |\alpha| - k))| + 1, |Q|]$，所以我们推出 $v' \in Q[|G_\dashv^\prec(p(\alpha, |\alpha| - k))| + 1, |Q|]$，矛盾。

($\geq$) 我们只需证明，如果 $l + 1 \leq j \leq |Q|$，那么 $Q[j] \notin G_\dashv^\prec(\alpha)$。由于 $j \geq l + 1 > |G^\prec(\alpha)|$，我们有 $Q[j] \notin G^\prec(\alpha)$，因此只需证明 $Q[j] \notin G_\dashv(\alpha)$。反设 $Q[j] \in G_\dashv(\alpha)$。不可能是 $Q[j] \in G^*(\alpha)$，否则$j \leq i^* \leq l$，这就产生了矛盾。因此，$Q[j] \in G_\dashv(\alpha) \setminus G^*(\alpha)$。这意味着，对于某个$0 < k < |\alpha|$，存在 $u' \in Q$ 使得 $(u', Q[j], s(\alpha, k)) \in E$ 且 $u' \in G_\dashv(p(\alpha, |\alpha| - k))$，由此可得 $g_k > f_k$。下面证明 $\texttt{in}(Q[1, j - 1], s(\alpha, k)) < g_k$。假设 $(v', v, s(\alpha, k)) \in E$ 满足 $v \in Q[1, j - 1]$。由于 $v \prec_{\mathcal{A}} Q[j]$，根据公理 2，我们得到 $v' \prec_{\mathcal{A}} u'$；又由于 $u' \in G_\dashv(p(\alpha, |\alpha| - k))$，进而有 $u' \in Q[1, |G_\dashv^\prec(p(\alpha, |\alpha| - k))|]$，因此我们得到$v' \in Q[1, |G_\dashv^\prec(p(\alpha, |\alpha| - k))|]$。这意味着 $\texttt{in}(Q[1, j - 1], s(\alpha, k)) \leq g_k$，并且必有$\texttt{in}(Q[1, j - 1], s(\alpha, k)) < g_k$，因为 $(u', Q[j], s(\alpha, k)) \in E$ 且 $u' \in G_\dashv(p(\alpha, |\alpha| - k))$。由此我们得到 $j \leq j^* \leq l$，矛盾。◀

26 广义自动机的 Myhill-Nerode 定理及其应用

现在，我们要说明如何构建 Wheeler 广义确定有限自动机的 FM 索引（定理 31）。首先，我们将给出一个引理，说明在引理 27 和引理 28 中，当我们遇到诸如“对于每个 k.....”这样的陈述时，如果$\mathcal{A}$ 是一个 r-广义确定有限自动机，我们总可以假设 k 至多取 r 个值。

▶ 引理 *33.* 设 $\mathcal{A} = (Q, E, s, F)$ 是一个 *Wheeler* r-广义确定有限自动机，且 $\alpha \in \Sigma^*$。对于 $0 < k < \min\{r+1, |\alpha|\}$，令 $f_k = \mathtt{out}(Q[1, |G^{\prec}(p(\alpha, |\alpha|-k))|], s(\alpha, k))$ 和 $g_k = \mathtt{out}(Q[1, |G^{\preceq}_{\dashv}(p(\alpha, |\alpha|-k))|], s(\alpha, k))$。那么，对于每个 $0 < k < \min\{r+1, |\alpha|\}$，都有 $g_k \geq f_k$。此外：

1. $|G^{\prec}(\alpha)|$ 是满足以下条件的最大整数 $0 \leq j \leq |Q|$：*(i)* 对每个$0 < k < \min\{r+1, |\alpha|\}$，都有 $\mathtt{in}(Q[1, j], s(\alpha, k)) \leq f_k$；*(ii)* 对于每个满足存在 $1 \leq \ell \leq j$ 且$\rho' \in \lambda(Q[\ell])$ 的 $\rho' \in \Sigma^k$，都有 $\rho' \preceq w_{\alpha,k}$，其中 $w_{\alpha,k}$ 是对每个 $0 < k < r+1$ 而言小于 α 的 Σ^k 中最大的字符串。

2. $|G^{\preceq}_{\dashv}(\alpha)|$ 等于以下各项中的最大值：
   a. $|G^{\prec}(\alpha)|$; *b.* 满足以下条件的最大整数 $0 \leq i \leq |Q|$：若 $i \geq 1$，则对于某个 $0 < k < r+1$，存在 $\rho' \in \Sigma^k \cap \lambda(Q[h])$，使得 $\alpha \dashv \rho'$。*c.* 满足以下条件的最小整数 $0 \leq j \leq |Q|$：对于每个 $0 < k <$，我们有满足 $g_k > f_k$ 的 $\min\{r+1, |\alpha|\}$，使得 $\mathtt{in}(Q[1, j], s(\alpha, k)) \geq g_k$。

证明。**1.** 我们要证明，推论的陈述等价于引理 27 的陈述。对于每个 $\min\{r+1, |\alpha|\} \leq k < |\alpha|$，我们有 $|s(\alpha, k)| > r$，因此 $f_k = 0$并且 $\mathtt{in}(Q, s(\alpha, k)) = 0$因为 $\mathcal{A}$是一个 r-广义确定有限自动机。这意味着，陈述中的条件 (i) 正确地取代了引理 27 中的条件 (i)。我们只需证明，陈述中的条件 (ii) 正确地取代了引理 27 中的条件 (ii)。这可由引理 22 直接得出。

2. 我们来证明，陈述中的三个条件等价于引理 28 中的三个条件。

a. 第一个条件相同。**b.** 第二个条件等价，因为 $\mathcal{A}$ 是一个 r-广义确定有限自动机。**c.** 第三个条件等价，因为如果 $\min\{r+1, |\alpha|\} \leq k < |\alpha|$，那么 $|s(\alpha, k)| > r$，因此 $g_k = f_k = 0$。

◀

我们回顾一下压缩数据结构中的以下引理（我们采用 [14] 中的方法）。回想一下，长度为 n、定义在字母表Σ 上的序列 $S \in \Sigma^n$ 的零阶熵为 $H_0(S) = \sum_{c \in \Sigma} \frac{|S|_c}{n} \log_2 \frac{n}{|S|_c}$，其中 $|S|_c$ 表示字符c 在S中的出现次数。

▶**引理 34** (秩/选择，参见 [14])。设 $S \in \Sigma^n$ 是一个定义在大小为 $\sigma \leq n$ 的整数字母表 $\Sigma = [0, \sigma - 1]$ 上的字符串。那么，存在一个占用 $nH_0(S)(1 + o(1)) + O(n)$ 比特的数据结构，能够在 $O(\log\log \sigma)$ 时间内支持以下操作：

- 访问： $S[i]$，对于任意 $1 \leq i \leq n$。秩： $S.rank(i, c) = |\{j \in \{1, \ldots, i\} \mid S[j] = c\}|$，
- 对于任意 $1 \leq i \leq n$ 和 $c \in \Sigma$。选择： $S.select(i, c)$ 等于满足 $S[j] = c$ 且 $S.rank(j, c) = i$ 的整数 j，对于任意$1 \leq i \leq S.rank(n, c)$ 以及 $c \in \Sigma$。

▶**引理 35** (完全可索引简洁字典，参见 [14])。集合 $A = \{x_1, \ldots, x_n\} \subseteq [1, u]$，其基数为 n，可以用一个占用 $n\log(u/n) + O(n)$ 比特的数据结构表示，从而可以在 $O(\log\log(u/n))$ 时间内实现以下操作：

- 秩： $A.rank(x) = |\{y \in A \mid y \leq x\}|$，对于任意 $1 \leq x \leq u$。选择： $A.select(i) = x$ 使得 $x \in A$ 且 $A.rank(x) = i$，对于任意 $1 \leq i \leq |A|$。

特别地，在 $O(\log\log(u/n))$ 时间内，我们还可以处理以下查询：

- 前驱： A 中小于或等于 x的最大元素（如果存在）。对于任意$1 \le x \le u$，如果 $A.rank(x) > 0$，则 $A.pred(x) = A.select(A.rank(x))$；否则为 $A.pred(x) = \perp$。严格后继： A 中严格大于 x的最小元素（如果存在）。对于任意$1 \le x \le u$，如果 $A.rank(x) < |A|$，则 $A.succ(x) = A.select(A.rank(x)+1)$；否则为 $A.succ(x) = \perp$。成员查询：对于任意 $1 \le x \le u$，当且仅当 $x = A.pred(x)$ 时， $x \in A$ 。

定理 31 的陈述。设 $\mathcal{A}$ 是一个 Wheeler r-广义确定有限自动机。于是，我们可以使用 $\log \sigma(1+o(1)) + O(e+rn)$ 比特对 $\mathcal{A}$ 进行编码；之后，给定长度为 m 的模式串 $\alpha \in \Sigma^*$，我们可以在 $O(rm(\log r + \log\log\sigma))$ 时间内解决 $\mathcal{A}$ 上的 SMLG 问题。在相同的时间界内，我们还可以判定 $\alpha \in \mathcal{L}(\mathcal{A})$ 是否成立。

证明。 我们将使用字符串熵 $H_0(S)$ 的一些众所周知的性质 $S \in \Sigma^h$：数量 $hH_0(S)$ 是对 S 进行任何编码时长度的下界，其中每个字符都通过字母表的前缀码独立于其他字符进行编码 Σ，特别地$H_0(S) \le \log_2 \sigma$，其中 $\sigma = |\Sigma|$。我们存储以下数据结构。

- 在每个 $\Sigma(i)$ 上使用一个全可索引简洁字典（引理 35）。所需比特总数为 $\sum_{i=1}^{r}(\sigma_i \log(\sigma^i/\sigma_i) + O(\sigma_i)) \le \sum_{i=1}^{r}(e_i \log(\sigma^i/\sigma_i) + O(e_i)) = \log \sigma - (\sum_{i=1}^{r} e_i \log \sigma_i) + O(e)$。因此，我们可以在每个字典上用 $O(\log\log(\sigma^i/\sigma_i)) \subseteq O(\log\log\sigma^r) = O(\log r + \log\log\sigma)$ 时间执行秩、选择、前驱、严格后继和成员查询。由引理 34 的数据结构表示定义 29 的位串 $\texttt{FIN} \in \{0,1\}^n$ 。所需比特数为 $nH_0(\texttt{FIN})(1+o(1)) + O(n) \subseteq O(n)$。对于每个 $1 \le i \le r$，由引理 34 的数据结构表示定义 29 的位串 $\texttt{OUT}_i \in \{0,1\}^{e_i+n}$ 。所需比特总数为 $\sum_{i=1}^{r}((e_i+n)H_0(\texttt{OUT}_i)(1+o(1)) + (e_i+n)\cdot O(1)) = O(e+nr)$。因此，我们可以在每个 $\texttt{OUT}_i$ 上用 $O(1)$ 时间执行秩和选择查询。对于每个 $1 \le i \le r$，由引理 34 的数据结构表示定义 29 的位串 $\texttt{IN}_i \in \{0,1\}^{e_i+n}$ 。所需比特总数再次为 $O(e+nr)$。因此，我们可以在每个 $\texttt{IN}_i$ 上用 $O(1)$ 时间执行秩和选择查询。对于每个 $1 \le i \le r$，由引理 34 的数据结构表示定义 29 的字符串 $\texttt{LAB}_i \in (\Sigma^i)^{e_i}$ 。所需比特总数为 $\sum_{i=1}^{r}(e_i H_0(\texttt{LAB}_i)(1+o(1)) + O(e_i)) \le (\sum_{i=1}^{r}(e_i \log \sigma_i)(1+o(1))) + O(e) \le (\sum_{i=1}^{r} e_i \log \sigma_i) + \log \sigma \cdot o(1) + O(e)$。因此，我们可以在每个 $\texttt{LAB}_i$ 上用 $O(\log\log\sigma^i) \subseteq O(\log\log\sigma^r) = O(\log r + \log\log\sigma)$ 时间执行秩和选择查询。对于每个 $1 \le i \le r$，由引理 34 的数据结构表示一个位串 $\texttt{AUX}_i \in \{0,1\}^{e_i}$ ，并按如下方式存储辅助信息。按照终止状态对 $E(i)$ 中的所有边进行排序；如果终止状态相同，则按标签排序。通过连接排序后各条边的所有标签，构造字符串 $\texttt{AUX}'_i \in (\Sigma^i)^{e_i}$ 。注意，根据引理 22，如果 $1 \le k \le e_i - 1$，则 $\texttt{AUX}'_i[k] \preceq \texttt{AUX}'_i[k+1]$。最后， $\texttt{AUX}_i \in \{0,1\}^{e_i}$ 是满足以下条件的位串：对于每个 $1 \le k \le e_i$，当且仅当 $\texttt{AUX}[k] = 1$ 或 $\texttt{AUX}'_i[k] \ne \texttt{AUX}'_i[k-1]$ 时， $k = 1$ 。所需比特总数为 $O(e)$。因此，我们可以在每个$\texttt{AUX}_i$ 上用 $O(1)$时间执行秩和选择查询。

将所有组件所占的空间相加（注意，字典 $\Sigma(i)$ 中的 $-\sum_{i=1}^{r} e_i \log \sigma_i$ 项与序列 $\texttt{LAB}_i$ 中的 $\sum_{i=1}^{r} e_i \log \sigma_i$ 项相互抵消），我们得出数据结构所占空间至多为 $\log \sigma(1+o(1)) + O(e+nr)$ 比特。特别是，我们正在

存储 $\mathcal{A}$ 的 Burrows-Wheeler 变换（其增强变体），因此根据定理 30，我们存储的是 $\mathcal{A}$ 的一种编码。

为了解决输入 $\alpha \in \Sigma^m$ 上的 SMLG 问题，我们需要计算 $G^{\prec}_{\dashv}(\alpha)$。根据引理 25，我们只需计算 $|G^{\prec}(\alpha)|$ 和 $|G_{\dashv}(\alpha)|$。经过 $O(m)$ 步，我们将递归地为每个 $0 \le k \le |\alpha|$ 计算 $|G^{\prec}(p(\alpha,k))|$ 和 $|G_{\dashv}(p(\alpha,k))|$。注意，如果 $k=0$，那么 $|G^{\prec}(\epsilon)|=0$ 和 $|G_{\dashv}(\epsilon)|=|Q|=n$。现在，假设我们已经为每个 $0 \le k < |\alpha|$ 计算出了 $|G^{\prec}(p(\alpha,k))|$ 和 $|G_{\dashv}(p(\alpha,k))|$，下面说明如何计算 $|G^{\prec}(\alpha)|$ 和 $|G_{\dashv}(\alpha)|$。根据引理 33，我们只需对以下每个查询求解至多 $O(r)$ 次：

- （op1）对于任意 $1 \le k \le |Q|$、每个 $1 \le i \le r$ 以及任意 $\rho \in \Sigma^i$，计算 $\mathtt{out}(Q[1,k],\rho)$；
- （op2）对于任意 $1 \le i \le r$、任意 $\rho \in \Sigma^i$ 以及任意 $h \ge 0$，求满足 $\mathtt{in}(Q[1,k],\rho) \le h$ 的最大整数 $0 \le k \le |Q|$；（op3）对于任意 $1 \le i \le r$、任意 $\rho \in \Sigma^i$ 以及任意 $z \ge 1$，若存在，求满足 $\mathtt{in}(Q[1,t],\rho) \ge z$ 的最小整数 $1 \le t \le |Q|$；否则报告该整数不存在。（op4）对于任意 $1 \le i \le r$ 和任意 $\rho \in \Sigma^i$，求满足以下条件的最大整数 $0 \le h \le |Q|$：对于每个满足存在 $1 \le j \le h$ 使得 $\rho' \in \lambda(Q[j])$ 的 $\rho' \in \Sigma^i$，都有 $\rho' \preceq \rho$。（op5）对于任意 $\rho \in \Sigma^*$，在 $|\rho| \le r$ 的条件下，并对于任意 $|\rho| \le i \le r$，求满足以下条件的最大整数 $0 \le h \le |Q|$：若 $h \ge 1$，则存在 $\rho' \in \Sigma^i \cap \lambda(Q[h])$ 使得 $\rho \dashv \rho'$。

因此，我们只需证明能够在 $O(\log r + \log\log\sigma)$ 时间内处理这些查询中的每一个。

（op1）。从 $Q[1,k]$ 中某个状态出发的 $E(i)$ 中边的数量，可以通过计算 $d = \mathtt{OUT}_i.rank(\mathtt{OUT}_i.select(k,1),0)$ 得到。因此，我们有 $\mathtt{out}(Q[1,k],\rho) = \mathtt{LAB}_i.rank(d,\rho)$。

（op2）。标记为 ρ 的边的数量为 $t = \mathtt{LAB}_i.rank(e_i,\rho)$。如果 $t \le h$，那么满足所需性质的最大k 是 $k = |Q|$。现在假设 $t > h$。在 $\Sigma(i)$ 中出现的、小于或等于 ρ 的字符串数量为 $\Sigma(i).rank(\rho)$，并且在 $\mathtt{AUX}'_i$ 给定的顺序中，第 $h+1$ 条标记为 ρ 的边是第 g 条边，其中 $g = \mathtt{AUX}.rank(\Sigma(i).rank(\rho),1) + h$。这条边到达状态 $Q[f]$，此处 $f = \mathtt{IN}_i.rank(\mathtt{IN}_i.select(g,0),1) + 1$，因此满足所需性质的最大$k$ 是 $f-1$。

（op3）。我们只需使用运算 (op2)，计算满足 $\mathtt{in}(Q[1,k],\ \rho) \le z-1$ 的最大整数 $0 \le k \le |Q_i|$。如果 $k = |Q_i|$，则所需的整数不存在；否则，它等于 $k+1$。

（op4）。我们首先检查 $\Sigma(i).succ(\rho)$ 是否有定义。如果没有定义，那么具有所需性质的最大 h 是 $|Q|$。现在假设 $\Sigma(i).succ(\rho)$ 有定义。那么根据引理 22，具有所需性质的最大整数就是满足 $\mathtt{in}(Q[1,k],\ \Sigma(i).succ(\rho)) \le 0$ 的最大整数 $0 \le k \le |Q_i|$，可使用 (op2) 计算得出。

（op5）。设 $c \in \Sigma$ 是 $(\Sigma,\ \preceq)$ 中最大的字符。设 $a = \Sigma(i).pred(c^{i-|\rho|}\rho)$。如果 a 不以 ρ 为后缀，或者它没有定义，那么具有所需性质的最大 h 是 $h=0$。否则，根据引理 22，我们有 h 是满足 $a \in \lambda(Q[h])$ 的最大整数。标签为 a 的边的数量是 $d = \mathtt{LAB}_i.rank(e_i,\ \rho)$，因此 h 是满足 $in(Q[1,t],\ \rho) \ge d$ 的最小整数 t，可使用 (op3) 计算得出。

现在，我们证明可以在 $O(rm(\log r + \log\log\sigma))$ 时间内判断$\alpha \in \mathcal{L}(\mathcal{A})$ 是否成立。我们声称，只要人为地以$|G^{\prec}(\epsilon)|=0$ 和 $|G_{\dashv}(\epsilon)|=1$（而不是 $|G^{\prec}(\epsilon)|=0$ 和 $|G_{\dashv}(\epsilon)|=n$）作为起点，同一算法就能奏效。事实上，考虑由 A 加入一个新的初始状态 s'，并加入边 $(s',s,\#)$ 得到的广义确定有限自动机 $\mathcal{A}$'；其中 s 是原来的初始状态，# 是一个小于 Σ 中所有字符的新字符。

那么，$\mathcal{A}'$ 也是 Wheeler（注意，$\preceq_{\mathcal{A}}$ 包含于 $\preceq'_{\mathcal{A}}$），而判断 $\alpha \in \mathcal{L}(\mathcal{A})$ 是否成立，等价于在输入 $\mathcal{A}$' 和 $\#\alpha$ 时求解 SMLG 问题（并检查唯一到达的状态是否属于 F，若该状态存在）。经过一步后可得到 $G^{\prec}(\#) = \{s', s\}$ 和$G_{\dashv}(\#) = \{s\}$，因此只需使用 FIN 检查唯一到达的状态是否属于F，即可得出结论。◀

C FM 索引：从广义确定有限自动机到不含 ϵ-转移的广义非确定有限自动机

本节旨在说明，定理 31 对不含 ϵ-转移的广义非确定有限自动机同样成立，从而得到定理 3。为此，我们将把 Wheeler 性质的概念扩展到不含 ϵ-转移的广义非确定有限自动机。

我们回顾一下 Wheeler 非确定有限自动机的定义。若存在一个关于 Q 的全序 $\leq$，使得 (i) s 在该全序中排在最前，(ii) 对任意 (u', u, a)、$(v', v, b) \in E$，若 $u < v$，则 $a \preceq b$，且 (iii) 对任意 (u', u, a)、$(v', v, a) \in E$，若 $u < v$，则 $u' \leq v'$，则称非确定有限自动机 A 具有 Wheeler 性质。Alanko 等人 [3, 引理 2.3] 证明了，若 $u < v$ 在该全序中，则 $(\forall \alpha \in I_u)(\forall \beta \in I_v)((\{\alpha, \beta\} \not\subseteq I_u \cap I_v) \rightarrow (\alpha \prec \beta))$。下面来看如何将定义 21 扩展到不含 ϵ-转移的广义非确定有限自动机。我们首先借鉴 Alanko 等人的结果，将 $\preceq_{\mathcal{A}}$ 的定义从广义确定有限自动机推广到不含 ϵ-转移的广义非确定有限自动机。

设 $\mathcal{A} = (Q, E, s, F)$ 是一个不含 ϵ-转移的广义非确定有限自动机。设 $\preceq_{\mathcal{A}}$ 是 Q 上的关系，使得对于每个 $u, v \in Q$，当且仅当 $(\forall \alpha \in I_u)(\forall \beta \in I_v)((\{\alpha, \beta\} \not\subseteq I_u \cap I_v) \rightarrow (\alpha \prec \beta))$。如果 $\mathcal{A}$ 是广义确定有限自动机，那么 $\preceq_{\mathcal{A}}$ 就简化为第 4 节给出的定义，因为对于每个满足 $u \neq v$ 的 $u, v \in Q$，我们有 $I_u \cap I_v = \emptyset$（见注 9）。我们已经看到，如果 $\mathcal{A}$ 是广义确定有限自动机，那么 $\preceq_{\mathcal{A}}$ 是一个偏序。如果 $\mathcal{A}$ 是任意一个不含ϵ-转移的广义非确定有限自动机，那么一般而言 $\preceq_{\mathcal{A}}$ 只是一个预序，也就是说，它是自反且传递的关系，但不一定是反对称的（见图 8）。

现在可以给出如下定义。

▶ **定义 36**。 设 $A = (Q, E, s, F)$ 是一个不含 ϵ-转移的广义非确定有限自动机。若存在关于 Q 的一个全序 $\leq$，使得：则称 A 为Wheeler 自动机。
1. 对于每个 $u, v \in Q$，若 $u \leq v$，则 $u \preceq_{\mathcal{A}} v$。 *2.* 对于每个 (u', u, ρ)、$(v', v, \rho') \in E$，若 $u < v$ 且 ρ' 不是 ρ 的真后缀，则 $\rho \preceq \rho'$。 *3.* 对于每个 (u', u, ρ)、$(v', v, \rho) \in E$，若 $u < v$，则 $u' \leq v'$。我们称 $\leq$ 为 A 上的 *Wheeler* 顺序。

注意，若 $\leq$ 是 Wheeler 顺序，则根据性质 1，s 排在最前面，因为 $\epsilon \in I_s$，并且对于每个 $u \in Q \setminus \{s\}$，都有 $\epsilon \notin I_u$。因此，引理 22 对不含 ϵ-转移的广义非确定有限自动机也成立（若在引理 22 的陈述中，将 $\preceq_{\mathcal{A}}$ 替换为 Wheeler 顺序$\leq$）。

如果$\mathcal{A}$ 是一个非确定有限自动机，那么定义 36 就退化为 Wheeler 非确定有限自动机的定义，因为根据 Alanko 等人的结果，性质 1 可由性质 2、性质 3 以及 s 在 Wheeler 顺序中排在最前面的事实推出。如果$\mathcal{A}$是一个广义确定有限自动机，那么根据引理 22，定义 36 就退化为 Wheeler 广义确定有限自动机的定义（定义 21）。

我们先给出一些预备性说明（见图 8）。一般而言，GNFA 上的 Wheeler 顺序并不唯一（这对非确定有限自动机已经成立）。此外，定义 36 中的性质 2、性质 3 并不能由性质 1 推出（这对非确定有限自动机也已经成立）。最后，性质 1 不能由性质 2、性质 3 以及要求 s 必须排在最前面推出（尽管我们刚刚看到，如果$\mathcal{A}$ 是一个非确定有限自动机，那么性质 1 可由性质 2、性质 3 以及要求 s必须排在最前面推出）。

判断一个广义确定有限自动机是否为 Wheeler 是可处理的（见引理 23），但对于广义非确定有限自动机，这个问题变成了 NP 困难，因为对于非确定有限自动机，它已经是 NP 困难的 [28]。定义广义非确定有限自动机上的

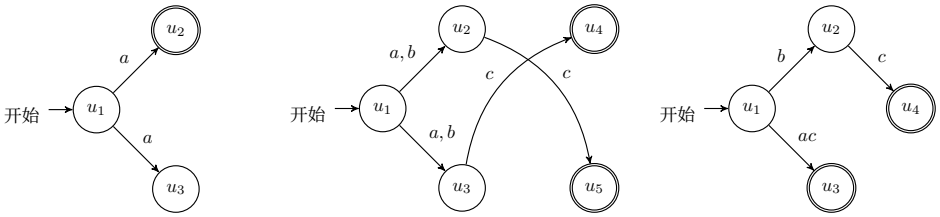


图 8 左：广义非确定有限自动机 $\mathcal{A}$ 满足 $\preceq_{\mathcal{A}}$ 不是反对称的，并且在使 u_1 排在首位的两个全序中，二者都是 Wheeler 顺序（特别地，Wheeler 顺序不一定唯一）。中：由 $u_1 < u_2 < u_3 < u_4 < u_5$ 给出的全序 $\leq$ 使得 $\leq$ 满足定义 36 中的性质 1，但不满足性质 2 和性质 3。右：由 $u_1 < u_2 < u_3 < u_4$ 给出的全序 $\leq$ 使得 $\leq$ 满足定义 36 中的性质 2 和性质 3，且初始状态排在首位，但不满足性质 1。

Wheeler 顺序有自然的方法。例如，容易验证，对于不含 ϵ-转换的广义非确定有限自动机，其上的任意 Wheeler 顺序都会诱导出第 2 节中定义的等价非确定有限自动机（不含 ϵ-转换）上的一个 Wheeler 顺序。

现在我们将说明，如何将我们的结果从 Wheeler 广义确定有限自动机推广到没有 ϵ-转移的 Wheeler 广义非确定有限自动机。若将 $\preceq_{\mathcal{A}}$ 替换为广义非确定有限自动机上的任意全序 $\leq$，引理 25 仍然成立。这是由 $\preceq_{\mathcal{A}}$ 在广义非确定有限自动机上的定义方式所决定的。例如，我们来证明引理 25 中的性质 3。我们知道，$u, v \in Q$ 满足 $u < v$ 和 $v \in G^{\prec}(\alpha)$，并且必须证明 $u \in G^{\prec}(\alpha)$。由于 $u < v$ 且 $\leq$ 是 Wheeler 顺序，因此 $u \preceq_{\mathcal{A}} v$。令$\beta \in I_u$；我们必须证明 $\beta \prec \alpha$。若 $\beta \in I_v$，则由 $v \in G^{\prec}(\alpha)$ 可立即得出$\beta \prec \alpha$。现在假设 $\beta \notin I_v$，并任取 $\gamma \in I_v$。由 $u \preceq_{\mathcal{A}} v, \beta \in I_u \setminus I_v$ 和 $\gamma \in I_v$可得 $\beta \prec \gamma$。由 $\gamma \in I_v$ 和 $v \in G^{\prec}(\alpha)$ 可得 $\gamma \prec \alpha$，因此得出 $\beta \prec \alpha$，证毕。最后，很容易验证，第 4 节中其余所有结果的证明仍然成立（只需将所有地方的 $\preceq_{\mathcal{A}}$ 替换为 Wheeler 顺序 $\leq$），因为我们已经证明，引理 22 对于没有 ϵ-转移的广义非确定有限自动机也成立。