

词法KAT作用幺半群

- 作者: GaoZheng
- 日期: 2025-09-26
- 版本: v1.0.0

摘要

介绍 Kleene Algebra with Tests (KAT) 与相关闭包/半环结构在本项目中的角色：用以建模可验证控制流、停机点与合规模式。提供从数学结构到工程接口的映射规范，支撑规则检查、代价累积与策略约束的统一表达。

词法KAT作用幺半群 ($\mathbb{M}_{\text{Lex-KAT}} := (\text{End}(\Sigma^*), \circ, \text{id})$)，由左/右乘子、投影族、tests 与闭包算子生成，作用于自由幺半群 $(\Sigma^*, \circ, \varepsilon)$ 。下面给出该 **幺半群视角** 下的算子谱系：把底层串空间视为自由幺半群 $(\Sigma^*, \circ, \varepsilon)$ ，把一切“历史拓扑/预测拓扑/裁剪与匹配”做成该空间上的**端算子** ($\text{End}(\Sigma^*)$)，以“函数合成”作为上层运算。先列**基本算子**，再列**复合算子与可用等式法则**，便于工程侧做融合与优化。

- 自由幺半群**: $M = (\Sigma^*, \circ, \varepsilon)$ (串连接、空串)。
- 端算子幺半群**: $(\text{End}(\Sigma^*), \circ_{\text{func}}, \text{id})$ (以**函数合成**为乘法)。
- 长度与前缀偏序**: $|s|, x \preceq y \iff x$ 是 y 的前缀。
- 参数集合**: 词典 $\mathcal{C} \subset \Sigma^*$, 长度集 $U \subset \mathbb{N}$, 历史步长上限 N , 预测上限 L_p 。

1) 基本算子 ($\text{End}(\Sigma^*)$ 的“元器件”)

1.1 左/右乘子 (来自自由幺半群的左/右作用)

- 左乘子** (历史左扩的母形) : $\mathbf{L}_h(s) = h \circ s, h \in \Sigma^*$ 。

合成律: $\mathbf{L}_{h_1} \circ_{\text{func}} \mathbf{L}_{h_2} = \mathbf{L}_{h_1 \circ h_2}$, 单位元 $\mathbf{L}_\varepsilon = \text{id}$ 。

- 右乘子** (预测延展的母形) : $\mathbf{R}_h(s) = s \circ h$ 。

合成律: $\mathbf{R}_{h_1} \circ_{\text{func}} \mathbf{R}_{h_2} = \mathbf{R}_{h_2 \circ h_1}$ (**反序拼接**)。

这两类算子给出 M 对自身的**左右作用**，是一切“拼接/延长”的代数基元。

1.2 裁剪/投影族 (幂等带)

- **尾裁剪**: $\Pi_L(s) = \text{tail}(s, L)$ 。
幂等: $\Pi_L \circ \text{_func}\Pi_M = \Pi_{\min(L, M)}$ (可交换、可并)。
- **首裁剪**: $\text{Head}_L(s) = \text{prefix}(s, L)$ 。
同上为幂等可交换族。
- **CJK 过滤/清洗**: $\text{CJK}(s)$ (非 CJK 清除/映射)，亦幂等。

这些投影族各自构成**交换幂等半群 (band)**，与 (\mathbb{N}, \min) 同构。

1.3 测试/掩码 (KAT 的 test)

- **后缀命中测试**: $\mathbf{T}_{L,\mathcal{C}}^{\text{Suf}}(s) = \begin{cases} s, & \text{tail}(s, L) \in \mathcal{C} \\ \perp, & \text{否则} \end{cases}$
- **前缀命中测试**: $\mathbf{T}_{L,\mathcal{C}}^{\text{Pref}}$ 同理。
- **合法字符/预算测试**: $\mathbf{T}_{\text{legal}}$ 、 $\mathbf{T}_{\text{budget}}$ (条件成立留 s , 否则 \perp)。

tests 幂等且可交换: $\mathbf{T}_P \circ \text{_func}\mathbf{T}_Q = \mathbf{T}_{P \wedge Q}$ 。

1.4 去重/规范化

- **首位去重**: $\mathbf{D}_{\text{head}}(s)$ (如“辑辑...”→“辑...”)，幂等: $\mathbf{D} \circ \mathbf{D} = \mathbf{D}$ 。

2) 闭包与迭代算子 (“命中即停/直到命中”)

2.1 历史前缀闭包 (左扩直到命中或步尽)

- **定义**:
 $\mathbf{Cl}_{U,N}^{\text{Pref}}(s)$ = 在 $\{\mathbf{L}_h\}$ 可用集合内迭代左乘，遇 $\exists L \in U$: $\mathbf{T}_{L,\mathcal{C}}^{\text{Pref}}$ 通过即停；最多 N 步，否则 \perp 。
- **性质**: 对 (Σ^*, \preceq) 扩张、幂等、单调 (典型闭包算子)。

2.2 预测后缀闭包 (右延直到命中或超上限)

- **定义**:
 $\mathbf{Cl}_{U,L_p}^{\text{Suf}}(s)$ = 按 \mathbf{R} 右延 (逐字符/片段)，按降序 $U \cap [1..]$ 查最长 $L \leq L_p$ 使 $\mathbf{T}_{L,\mathcal{C}}^{\text{Suf}}$ 成立，“命

中即停”；否则回退或上。

- 性质：同为扩张、幂等、单调的闭包算子。

两个闭包与 tests/投影共同给出 **Kleene Algebra with Tests (KAT)** 风格的“while 命中即停”语义。

3) 复合算子（业务可直接调用的“流程件”）

3.1 观测构建器（历史窗口）

$$\mathbf{B}_{L_h}(p, \chi) = \mathbf{R}_{\langle \text{eos} \rangle} \circ \mathbf{R}_\chi \circ \mathbf{R}_{\langle \text{sep} \rangle} \circ \Pi_{L_h}(p) \circ \mathbf{R}_{\langle \text{bos} \rangle}$$

把“prev 的尾窗 + 分隔 + 当前目标符”拼成单步观测串。

3.2 历史拓扑器（左扩对齐）

$$\mathbf{H}_{U,N} = \mathbf{Cl}_{U,N}^{\text{Pref}} \circ \mathbf{B}_{L_h}$$

对 \mathbf{B} 的输出做“前缀最长可用命中”的闭包（控制步数 N ）。

3.3 预测拓扑器（命中即停）

$$\mathbf{P}_{U,L_p} = \mathbf{Cl}_{U,L_p}^{\text{Suf}} \circ \mathbf{D}_{\text{head}}$$

首位去重后，做“后缀最长可用命中（上限 L_p ）”。

3.4 bigram 拓扑器（前向组合）

$$\mathbf{Bi}_{U,L_p}(\chi, \cdot) = \mathbf{Cl}_{U,L_p}^{\text{Suf}} \circ \mathbf{R}_\chi$$

先右乘当前字符，再做后缀闭包，用于 bigram 奖励与注记。

3.5 法规/医疗等域的强约束管线（KAT 形式）

$$\mathbf{F} = \mathbf{T}_{\text{legal}} \circ \mathbf{H}_{U,N} \circ \mathbf{T}_{\text{budget}} \circ \mathbf{P}_{U,L_p} \circ \mathbf{T}_{\text{clean}}$$

tests（门控）—闭包—tests—闭包—清洗 的标准序列；每段均可热更。

4) 加权评分 (半环语义, 非必须)

把“隶属度×语义阈×IDF”视为权重半环 S 上的评分：

$$w(s) = \mu(\text{seg}) \otimes \underbrace{\max(0, \text{sim} - \tau)}_{\text{门控}} \otimes \text{idf}(\text{seg}), \quad (\oplus = \max, \otimes = \times)$$

与上面的闭包/tests 结合即形成**带权 Kleene 代数**；用于择优“最长且可信”的命中路径。

5) 关键等式法则 (优化/融合用)

投影融合 (幂等/可交换)

- $\Pi_L \circ \Pi_M = \Pi_{\min(L, M)}$; $\text{Head}_L \circ \text{Head}_M = \text{Head}_{\min(L, M)}$ 。

乘子融合

- $\mathbf{L}_{h_1} \circ \mathbf{L}_{h_2} = \mathbf{L}_{h_1 \circ h_2}$;
 $\mathbf{R}_{h_1} \circ \mathbf{R}_{h_2} = \mathbf{R}_{h_2 \circ h_1}$ 。

tests 结合

- $\mathbf{T}_P \circ \mathbf{T}_Q = \mathbf{T}_{P \wedge Q}$ (幂等、交换)。

闭包幂等

- $\mathbf{Cl} \circ \mathbf{Cl} = \mathbf{Cl}$ (前缀/后缀两类皆然)。

条件交换 (工程可判)

- 若 Π_{L_h} 不改变 \mathbf{Cl}^{Suf} 的停点，则
 $\mathbf{Cl}_{U, L_p}^{\text{Suf}} \circ \Pi_{L_h} = \Pi_{L_h} \circ \mathbf{Cl}_{U, L_p}^{\text{Suf}}$ 。
 - 同理对 $\mathbf{Cl}^{\text{Pref}}$ 与 \mathbf{L}_h 存在可判定的可交换前提 (裁剪不破坏前缀命中路径)。
-

6) 快速总表 (命名对照)

类别	名称	记号	核心性质
左乘	左乘子	\mathbf{L}_h	幺半群作用；可合并
右乘	右乘子	\mathbf{R}_h	反序合并
裁剪	尾裁剪	Π_L	幂等、交换、可与 min 同构
裁剪	首裁剪	\mathbf{Head}_L	同上
清洗	CJK 过滤	\mathbf{CJK}	幂等
测试	后缀/前缀命中	$\mathbf{T}_{L,C}^{\text{Suf/Pref}}$	幂等、交换
规范	首位去重	\mathbf{D}_{head}	幂等
闭包	前缀闭包	$\mathbf{Cl}_{U,N}^{\text{Pref}}$	扩张、幂等、单调
闭包	后缀闭包	$\mathbf{Cl}_{U,L_p}^{\text{Suf}}$	扩张、幂等、单调
复合	观测构建	\mathbf{B}_{L_h}	右乘链 + 尾裁剪
复合	历史拓扑器	$\mathbf{H}_{U,N}$	$\mathbf{Cl}^{\text{Pref}} \circ \mathbf{B}$
复合	预测拓扑器	\mathbf{P}_{U,L_p}	$\mathbf{Cl}^{\text{Suf}} \circ \mathbf{D}$
复合	bigram 拓扑器	\mathbf{Bi}_{U,L_p}	$\mathbf{Cl}^{\text{Suf}} \circ \mathbf{R}_\chi$
复合	合规管线	\mathbf{F}	tests–闭包–tests–闭包–清洗

一句话落地：把所有流程都归约为“**左/右乘**（拼接）、**投影**（裁剪）、**tests**（门控）与“**闭包**（直到命中/命中即停）”四类元算子；它们在 $\text{End}(\Sigma^*)$ 的**函数合成幺半群**中封闭，且具备可复用的幂等/交换/合并等式，工程侧据此即可做**规则融合、步数削减、可判交换与形式验证**。

许可声明 (License)

Copyright (C) 2025 GaoZheng

本文档采用[知识共享-署名-非商业性使用-禁止演绎 4.0 国际许可协议 \(CC BY-NC-ND 4.0\)](#)进行许可。