arXiv:2303.06629v1 [cs.AI] 12 Mar 2023

另　一种实体解析的通用设置：基础理论

Xiuzhan Guo、Arthur Berrill、Ajinkya Kulkarni、Kostya Belezko、Min Lu　o

摘要

Benjelloun 等人 [2] 将实体解析（ER）问题视为这样一个通用过程：匹配并合并被判定为代表同一现实世界对象的实体记录。他们将用于匹配和合并实体记录的函数视为黑盒，并提出了四项重要性质（称为幂等性、交换性、结合性和代表性），从而能够实现高效的通用实体解析算法。

偏群胚是一个非空集合 P，带有一个偏运算 $\circ$，该偏运算是一个偏函数$P \times P \to P$。实体解析系统（$\mathbb{E}$、$\approx$、$\langle\rangle$）与偏群胚（P、D、$\circ$）相同，其中二元运算 $\circ$ 由合并函数 $\langle\rangle$ 给出，而运算的偏性 D 由匹配函数 $\approx$规定。给定一个实例 I，I 的合并闭包最终构成有限生成偏子群胚 $[I]$。

本文将研究匹配函数与合并函数共有的性质，基于匹配函数和合并函数所满足的性质，在偏群胚中对 ER 的匹配与合并黑盒进行建模，并说明偏群胚为 ER 提供了另一种通用框架。

当偏群胚满足幂等性和链式结合性时，可以在其上定义自然偏序。给定偏群胚上的一个偏序，若偏运算的定义域是自反的，则最小上界性质和相容性（LU_{pg} 和 CP_{pg}）性质等价于幂等性、交换性、结合性和代表性，并且该偏序必须是我们所定义的自然偏序。偏群胚的偏性可以通过其定义域图的连通分支和团覆盖来降低；如果非交换偏群胚具有类似偏幂等半群的结构，则可以将其同态地映射为交换偏群胚。

在一个有限生成偏群胚（$P, D, \circ$）中，若不要求任何条件，我们所关注的 ER 是 P 中的全元素。若（$P, D, \circ$）满足幂等性和链式结合性，则 ER 是 P 中的极大元；这些极大元都是全元素，并构成 [2] 中定义的 ER。此外，在这种情况下，由于存在传递的二元序关系，我们将 ER 视为“对有限生成偏群胚中的元素进行排序、选择和查询”。

1 引言

实体解析（ER，也称为记录链接、去重或合并清洗）旨在判定数据库或数据流中的两个或多个实体记录是否表示同一个现实世界对象。众所周知，ER 是整合来自不同来源的数据的一项关键任务 [6]，也是影响信息系统和流数据分析中数据质量的主要障碍之一 [22, 7]。

Benjelloun 等人认为，在实体解析问题中，被确定为代表同一现实世界实体的记录会被依次定位并合并 [2]。他们将用于匹配和合并记录的函数视为黑盒，从而对通用实体解析问题进行了形式化，这

使得实体解析解决方案具有更强的表达能力和可扩展性。他们提出了四个重要性质：(I)、(C)、(A) 和 (R)（幂等性、交换性、结合性和代表性，定义见下文）；如果匹配函数和合并函数满足这些性质，就能实现更高效的通用实体解析算法。

假设 $\mathbb{E}$ 是实体记录的集合。在 Benjelloun 等人 [2], 的实体解析模型中， $\approx$ 上的匹配函数 $\mathbb{E}$ 是一个布尔函数 $\approx$: $\mathbb{E} \times \mathbb{E} \to \{0,1\}$，其中 $\approx(e_1, e_2) = 1$ 表示 e_1 与 e_2 匹配，记作 $e_1 \approx e_2$。否则， $\approx (e_1, e_2) = 0$ 表示 e_1 与 e_2 不匹配，记作 $e_1 \not\approx e_2$。相对于匹配函数 $\approx$ 的合并函数 $\langle\,\rangle$ 是一个偏函数 $\langle\,\rangle$: $\mathbb{E} \times \mathbb{E} \to \mathbb{E}$，它仅定义在相匹配的实体记录对上，即， $\langle\,\rangle$ 定义在 $D \subseteq \mathbb{E} \times \mathbb{E}$ 上，记作 domain($\langle\,\rangle$) $= D$，其中 $(e_1, e_2) \in D$ 当且仅当 $e_1 \approx e_2$，在此情况下，值 $\langle\,\rangle(e_1, e_2)$ 记作 $\langle e_1, e_2\rangle$。实体解析系统 $(\mathbb{E}, \approx, \langle\,\rangle)$ 由实体记录的非空集合 $\mathbb{E}$、定义在 $\mathbb{E}$ 上且用于建模匹配函数的二元关系 $\approx$，以及定义在 $\mathbb{E}$ 上且用于建模合并函数的偏运算 $\langle\,\rangle$ 组成。

合并函数旨在将匹配实体记录的信息合并在一起，从而生成包含更多信息的新记录。正如 [2], 中所强调的，合并可能会产生新的匹配，并且可能存在这样的情形：最终匹配只有在经过多次匹配与合并后才会发生。

实例 I 是一个有限子集 $I \subseteq \mathbb{E}$。实例 I的合并闭包 $\overline{I}$是集合 R 中最小的实体记录集合：它识别 I 中的所有匹配记录对并将其合并，且满足

(i) $I \subseteq R$, 并且

(ii) R 是合并闭合的：对于所有 e_1、$e_2 \in R$， $e_1 \approx e_2$ 蕴含 $\langle e_1$、$e_2\rangle \in R$。

由于合并闭合子集 $\subseteq \mathbb{E}$ 的交集同样是合并闭合的，因此有

$$\overline{I} = \bigcap \Big\{ R | I \subseteq R, R \text{ is merge closed} \Big\}.$$

给定一个实例 I， [2] 中的实体解析（ER）目标是从合并闭包 $\overline{I}$ 中的实体记录里选择最小子集 ER（I），使得 ER（I）包含的信息不少于 $\overline{I}$。为此， [2]中引入了偏序 $\leq$：

给定两个实体记录 e_1、$e_2 \in \mathbb{E}$、$e_1 \leq e_2$ 当且仅当 $e_1 \approx e_2$ 且 $\langle e_1$、$e_2\rangle = e_2$.

给定两个实体记录集合 I_1、$I_2 \subseteq \mathbb{E}$、$I_1 \leq I_2$, 称为 I_1 当且仅当对于每个 $e \in I_1$，存在 $e' \in I_2$ 使得 $e \leq e'$, 其被 I_2支配。

在 [2], 中，实例 I 的实体解析被定义为：

定义 1.1（定义 2.3 [2]）。给定一个实例 $I \subseteq \mathbb{E}$, 实例 I 的实体解析 ER（I）是

(ER) $ER(I) = set\ I'\ of\ entity\ records\ such\ that:$

(i) $I' \subseteq \overline{I}$,

(ii) $\overline{I} \leq I'$,

(*iii*) I' 在满足条件（i）和（ii）的范围内是最小的： I' 的任何真子集都不满足条件（i）和（ii）。

在 [2], Benjelloun 等人证明了实例 I 的实体解析存在且唯一。 [2] 中 (I)、(C)、(A) 和 (R) 的定义如下：

幂等性：

(I) $$\text{For all } e \in \mathbb{E}, e \approx e \text{ and } \langle e, e \rangle = e.$$

交换性：

(C) 对于所有 e_1、 $e_2 \in \mathbb{E}$、 $e_1 \approx e_2$，当且仅当 $e_2 \approx e_1$；如果 $e_1 \approx e_2$，则 $\langle e_1, e_2 \rangle = \langle e_2, e_1 \rangle$。

结合性：

(A) For all $e_1, e_2, e_3 \in \mathbb{E}$, if $\langle e_1, \langle e_2, e_3 \rangle \rangle$ and $\langle \langle e_1, e_2 \rangle, e_3 \rangle$ exist, then $\langle e_1, \langle e_2, e_3 \rangle \rangle = \langle \langle e_1, e_2 \rangle, e_3 \rangle$.

代表性：

(R) $$\text{If } e_3 = \langle e_1, e_2 \rangle, \text{ then for any } e_4 \text{ such that } e_1 \approx e_4,\ e_3 \approx e_4.$$

在 [2], Benjelloun 等人开发了三种高效的实体解析算法： R-Swoosh 利用 (I)、(C)、(A) 和 (R) 性质，并以记录为粒度执行比较（算法 3，命题 3.3 [2]）； F-Swoosh 利用 (I)、(C)、(A) 和 (R) 性质，采用特征级比较，通过避免重复的特征比较，其效率可以高于 R-Swoosh（算法 4，命题 4.1 [2]）；以及最通用的实体解析算法 G-Swoosh，该算法不要求具备 (I)、(C)、(A) 和 (R) 性质（算法 1 和 2，命题 3.2 [2] 以及定理 3.1 [2]）。感兴趣的读者可以查阅 [2]以了解算法的详细信息。

然而，实体记录不一定是客户，也可以是其他事物，例如数据集、本体、文档、图等。合并函数也可能是其他相关操作，例如分别对应数据集连接、本体合并、文档合并、图合并等。匹配与合并过程非常普遍，不仅存在于客户记录链接搜索中，也存在于数据集关系检测、本体对齐与合并、文档匹配与合并、图相似性搜索与合并等场景中。首先研究匹配与合并操作所共有的核心性质，然后探讨是否存在一种灵活的设置，使上述常见实践能够在不涉及任何内部细节的情况下统一建模，这是很自然的做法。

代数结构，如群胚、幺半群、半群、群、环等，都具有一个对象集合以及一个或多个运算。相关实体记录构成一个非空集合，而合并函数可以视为该非空集合上的二元运算。现在我们回顾群胚和半群的概念。

群胚是一个非空集合 P，其上定义了一个二元运算，该运算是从笛卡尔积 $P \times P$到 P的函数。半群是另一种代数结构，它是一个非空集合 T，并配备了 $\circ$上的结合二元运算；也就是说，无论如何对 T中的对象进行分组或放置括号，都可以执行 $\circ$。半群在

20世纪50年代以来的计算机科学中一直具有特别重要的地位，例如自动半群（配备了多个字母表上的正则语言、用于表示生成集 [5]的有限生成半群）、图逆半群 [1, 19], 等。

非空集合 P 上的某些二元运算仅定义在 $P \times P$的一个子集上。例如，结构化查询语言中的左连接仅定义在满足特定连接条件的表上，而矩阵乘法 $M \cdot N$ 仅在 M 的列数等于 N的行数时才有定义。带有偏二元运算（偏函数 $\circ\colon\ P \times P \to P$）的非空集合 P 称为偏群胚；带有结合的偏二元运算的非空集合称为偏半群。

实体记录 $\mathbb{E}$ 的集合，连同匹配函数 $\approx$ 和合并函数 $\langle\rangle$，可以视为一个偏群胚或半群（如果 $\langle\rangle$ 满足结合律）$(\mathbb{E}, \approx, \langle\rangle)$，其中 $\langle\rangle$ 是 $\mathbb{E}$ 上的偏二元运算，而 $\approx$ 规定了 $\langle\rangle$的偏定义域。因此，实体解析系统中的匹配函数和合并函数可以在偏群胚中结合起来研究，从而能够将相关的代数结果（例如半群上的偏序和有限性条件）应用于实体解析系统。

为了比较实体记录中的信息， [2]中引入了偏序。众所周知，每个半群都可以自然地定义偏序 [20]。

给定实体记录的有限子集 I ，将匹配函数和合并函数应用于 I ，以生成所有可能的新实体记录。结果表明， $\overline{I}$ 的合并闭包 I 是 $\mathbb{E}$中由 I生成的 $[I]$偏子群胚，因此， I 的实体解析最终就是在有限生成偏群胚 $[I]$中查找特定元素。

为了有效地计算匹配和合并，希望匹配和合并的步骤是有限的，或者每个有限实例 I 的合并闭包 $\overline{I}$ 是有限的，这意味着由 I 生成的偏子群胚 $[I]$ 是有限的。为了使 $[I]$ 或 $\overline{I}$ 成为有限的，需要满足一些条件。若满足结合性，则记录的运算顺序不会影响最终结果，从而可以显著减少由 I 生成的实体记录数量。若偏群胚满足结合性，则它就成为偏半群。

本文的目标是

- 将实体解析系统建模为偏群胚，并研究偏群胚上匹配与合并操作所共有的核心性质，
- 研究偏群胚上的偏序及其与偏群胚运算之间的相互作用，
- 提供降低合并操作的偏态程度的方法，并通过偏群胚同态，将满足某些合理条件的非交换偏群胚映射为交换偏群胚，以及
- 说明满足某些合理条件的偏群胚，为建模匹配与合并操作以及计算实体解析系统提供了另一种通用框架。

T该理论的应用将在本文的后续篇章中予以总结 r.

本文其余部分的组织如下：首先，在第 2 节中，我们回顾群胚的基本定义和符号，包括偏群胚、半群和偏半群、子集乘积、偏群胚同态、在偏群胚同态下保持的偏群胚性质、不可约生成集，并

从不同领域提供一些示例，以说明匹配与合并操作非常普遍。

在第 3 节中，我们从偏群胚性质（I_{pg}）、（SC_{pg}）、（A_{pg}）、（Rl_{pg}）、（Rr_{pg}）和（R_{pg}）开始。要定义偏半群，必须从众多结合性性质中指定一种结合性形式。在本文中，我们关注以下三种结合性性质：（A_{pg}）、（CA_{pg}）和（SA_{pg}）。然后，我们讨论这些性质之间的关系，例如，（CA_{pg}）等价于（A_{pg}）和（R_{pg}）（命题 3.1）；（SA_{pg}）或（CA_{pg}）蕴含（A_{pg}）（命题 3.2）；这些性质在偏群胚同态下保持不变（命题 3.4）。

偏序是自反的、反对称且传递的二元关系。半群上的自然偏序已得到广泛研究 [20]。沿用半群上的研究方法，在第 4 节中，我们分别定义偏群胚上的自然偏序 $\leq_l$、$\leq_r$ 和 $\leq$，以及相对于这些偏序的极大元。极大元构成给定实例 I 的ER（I），其定义见 [2]。给定偏群胚上的一个偏序，为研究偏序与偏运算之间的相互作用，我们考察相对于该偏序的性质（LU_{pg}）和（CP_{pg}），并证明当偏运算的定义域是自反的时，（LU_{pg}）和（CP_{pg}）等价于（I_{pg}）、（C_{pg}）、（A_{pg}）和（R_{pg}），且该偏序就是我们定义的自然偏序（定理 4.12）。

一般而言，实体解析中的匹配与不匹配标签高度不平衡，这被称为类别不平衡问题 [12, 24]。因此，给定一个实体解析系统（$\mathbb{E}$、$\approx$、$\langle\rangle$）和一条实体记录 $e \in \mathbb{E}$，e 只能与集合 $\mathbb{E}$ 的一小部分进行组合，这体现了 $\langle\rangle$ 的偏性。从代数视角来看，给定偏群胚（P、D、$\circ$）的偏二元运算 $\circ$ 的偏性，可以通过其定义域 D 来度量，而该定义域可以表示为有向图 $\mathcal{G}_P$。在第 5 节中，我们利用（P、D、$\circ$）的连通分量和团覆盖，降低偏群胚（P、D、$\circ$）的偏性。

交换性（C）在 [2] 中非常重要。在第 6 节中，我们展示了每个具有幂等半群型结构的偏群胚，都可以通过同态方式转换为交换偏群胚。

由于给定实例 I 的合并闭包 $\overline{I}$ 可视为一个有限生成群胚P，在讨论偏群胚的核心性质和偏序后，我们就可以介绍给定有限生成偏群胚 P上的实体解析。为保持一般性，第 7 节研究基于有限生成偏群胚所满足性质的实体解析。我们从任意偏群胚开始。在这种情况下，实体解析（ER_f）可视为偏群胚中的满元。如果给定的偏群胚满足（I_{pg}）和（CA_{pg}），那么它具有自然序 $\leq$，其实体解析（ER_m）可视为极大元，这些极大元构成在 [2]中定义的（ER）。此外，如果有限生成偏群胚具有传递序且为有限的，那么根据 [8]中关于传递偏序的结果，其实体解析（ER_{pg}）可视为对有限生成偏群胚中的元素进行排序、选择和查询。最后，我们给出总结性评述。

2 预备知识

设 P为一个非空集合。集合 P上的二元运算 $\circ$是一个函数 $\circ\colon P\times P\to P$。集合 P上的偏二元运算 $\circ$是一个偏函数 $\circ\colon P\times P\to P$，仅在子集 $D\subseteq P\times P$上定义。 D是偏运算 $\circ$的定义域，记作 $\mathrm{domain}(\circ)=D$。如果（x， y）$\in D$，则称 $\circ$（x， y）有定义，或称 x与 y可复合，或称 $x\circ y$存在。 $\circ$（x， y）的值，或 x与 y的复合，记作 $x\circ y$。如果（x， y）$\notin D$，则称 $\circ$（x， y）未定义，记作 $x\circ y=\uparrow$。

定义域 D是自反的，如果 $(p,p)\in D$对所有 $p\in P$均成立。 D是对称的，如果

$$(S)\qquad (x,y)\in D\Leftrightarrow (y,x)\in D.$$

一个非空集合 P，连同（偏）二元运算 $\circ$（定义在 P 上），称为一个（偏）群胚。为强调运算 $\circ$的偏性，我们将偏群胚写作三元组 $(P,D,\circ)$ 或 $(P,\sim,\circ)$，其中 P是非空集合， $\circ$是定义在 P上的偏二元运算， D是 $\circ$的定义域，而$\sim$ 是 P 上的二元关系，用于规定 D：$(x,y)\in D$ 当且仅当 $x\sim y$。偏群胚也称为 Magma [15]。

回顾一下，无向图（或图）由一对 (V,E) 组成，其中 V是一个（有限）对象集合，称为顶点或节点，而 E是无序对 $\{v_1,v_2\}$的集合，称为边， $v_1v_2\in V$。有向图是一对 (V,E)，其中 V是一个（有限）对象集合，称为顶点或节点，而 E是有序对 (v_1,v_2) 的集合，称为边， $v_1v_2\in V$。在有向图中，一条边 $e=(v_1,v_2)$ 可视为一支箭头，从 v_1 出发并终止于 v_2，记作 $d_0(e)=v_1$ 和 $d_1(e)=v_2$。

众所周知，在有向图与无向图之间存在经典的转换方法。

在本文中，$(P,D,\circ)$ 是一个偏群胚。我们可以将“$(P,D,\circ)$”简写为 P 或 $(P,\circ)$，并在“$x\circ y$”中省略 $\circ$；当 D 和 $\circ$ 根据上下文可以明确时，将其写作“xy”。

例 2.1. 下面给出一些群胚和偏群胚的例子。

1. 设 a、b、c 是互不相同的元素，并且 $P_1=\{a,b,c\}$。定义

$$a\cdot a=a,\quad b\cdot b=b,\quad c\cdot c=c,\quad a\cdot b=b,\quad b\cdot c=c,$$

且不定义其他复合。于是 $(P_1,\cdot)$ 是一个偏群胚。

2. 设 $\mathbb{N}$ 为所有自然数组成的集合。定义

$$m\vee n=\max(m,n)\text{ and }m\wedge n=\min(m,n).$$

那么，$(\mathbb{N},\vee)$ 和 $(\mathbb{N},\wedge)$ 都是群胚。

3. 每个实体解析系统 $(\mathbb{E},\approx,\langle\ \rangle)$ 显然都是偏群胚，其中

$$\langle\ \rangle(x,y)=\begin{cases}\langle x,y\rangle, & \text{whenever } x\approx y\\ \uparrow, & \text{otherwise.}\end{cases}$$

4. 设 $A=\{a_1,\cdots,a_n,\cdots\}$ 为一个包含无限多个实体对象的集合。对所有正整数 i 定义 $a_i*a_{i+1}=a_{i+2}$，不定义其他复合。则（A， $*$）是一个偏群胚。

5. 给定一个有向图 $G = (V(G)\ E(G))$， G 中的路径是按同一方向指向的一系列边，它连接一系列顶点，并且其中的边和顶点均互不相同。更正式地说， G 中的路径由 $[e_1, \cdots, e_m]$给出，其中 e_1、$\cdots$、$e_m \in E(G)$，满足对于 $i = 1$、$\cdots$、$m-1$、$|\{e_1, \cdots, e_m\}| = m$， $d_1(e_i) = d_0(e_{i+1})$，且 $|\{d_1(e_1)\ \cdots$、$d_1(e_m)\}| = m$，其中对于每条边 e： $v_1 \to v_2$， $d_1(e) = v_2$，并且 $d_0(e) = v_1$。

给定两条路径 $[e_1, \cdots, e_m]$、$[f_1, \cdots, f_n]$，当且仅当存在一个最小正整数 $i \leq \min\{m, n\}$ 使得 $[e_1, \cdots, e_m] \sim_p [f_1, \cdots, f_n]$

$$m - i + 1 \leq n,\ e_i = f_1, \cdots, e_m = f_{m-i+1}.$$

令 $\mathcal{P}ath(G)$ 表示 G 中所有路径的集合。路径 $[e_1, \cdots, e_m]$ 和 $[f_1, \cdots, f_n]$的复合定义为

$$[e_1, \cdots, e_m] \between [f_1, \cdots, f_n] = \begin{cases} [e_1, \cdots, e_i, f_{i+1}, \cdots, f_n], & \text{if } [e_1, \cdots, e_m] \sim_p [f_1, \cdots, f_n] \\ \uparrow, & \text{otherwise.} \end{cases}$$

那么（$\mathcal{P}ath(G)/\sim_p$， $\between$）是一个偏群胚。

例如，令 $W = \{p_1, \cdots, p_n\}$ 为网页的集合。点击图 $\mathcal{C}_W$ 是一个以 W 为顶点的图。如果访问者在查看 p_i 后点击了 p_j，则从 p_i 到 p_j 存在一条边。那么（$\mathcal{P}ath(\mathcal{C}_W)/\sim_p$， $\between$）是一个偏群胚，其对象是点击流或点击路径。

6. 令 T为关系数据库中的表集合，令 x为选定的一列。对于所有 t_1、$t_2 \in T$，定义

$$t_1 \bowtie_l t_2 = \begin{cases} t_1\ left_join\ t_2\ on\ t_1 \cdot x = t_2 \cdot x, & \text{if } x \text{ exists in both } t_1 \text{ and } t_2 \\ \uparrow, & \text{otherwise.} \end{cases}$$

and

$$t_1\ \&\ t_2 = \begin{cases} row\ bind\ of\ t_1\ and\ t_2, & \text{if } t_1 \text{ and } t_2 \text{ are matched by their columns} \\ \uparrow, & \text{otherwise.} \end{cases}$$

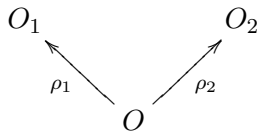
则 $(T, \bowtie_l)$ 和 $(T, \&)$ 都是偏群胚。

7. 给定一个数据库 $\mathcal{D}$ 及其数据库操作集合 $\mathcal{T}$，例如连接 $\bowtie$、 σ、 α， $\mathcal{D}$ 中的一个数据血缘元素由三元组 $\langle I$、T、$O\rangle$组成，其中 $T \subseteq \mathcal{T}$是数据库操作序列， I是 T的输入集合，而 O 是将 T 中的操作依次应用于 I后得到的输出。令 $\mathcal{L}$ 为 $\mathcal{D}$中的数据血缘元素集合。定义

$$\langle I, T, O\rangle c \langle I', T', O'\rangle = \begin{cases} (I, T +\!\!+ T', O'), & \text{if } O = I' \\ \uparrow, & \text{otherwise.} \end{cases}$$

这里 $+\!\!+$ 是序列连接操作。于是（$\mathcal{L}$、$\rightsquigarrow$、c）是一个偏群胚，其中（I、T、O）（I'、T'、O'）当且仅当 $O = I'$。

8. 本体在多个定义域中都有应用，用于表示类及其实例之间的关系。本体对齐表达了不同本体之间的关系 [16]。本体合并通过连接现有本体来创建一个新本体，并可由推出构造 [26, 14] 捕获。设 $\mathbb{O}$ 为本体及其间本体同态所构成的范畴。一个 V- 对齐是 $\mathbb{O}$ 中的一个跨越:



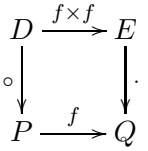
其中， O、 O_1、 O_2 是本体， $\rho_1 : O \to O_1$ 和 $\rho_2 : O \to O_2$ 是本体同态。 V-对齐的推出 $\sqcup$ 是对本体合并 [26], 的一种简明描述；本体合并可视为余切片范畴 $O/\mathbb{O}$ ($O/\mathbb{O}$, $\vee$ $\sqcup$)中的余积。 $O_1 \vee O_2$ 是一个偏群胚，当且仅当存在一对本体同态 $\rho_1 : O \to O_1$ 和 $\rho_2 : O \to O_2$。

9. 给定一个由实体对象组成的集合 E，例如数据集、知识图谱、文档等，对于 e_1、 $e_2 \in E$，定义 $e_1 \smile e_2$ 当且仅当 e_1 与 e_2 之间存在重叠部分； $e_1 \curlywedge e_2$ 是 e_1和 e_2的并集，通过识别重叠部分并保留其中一份副本得到。于是，(E, $\smile$, $\curlywedge$)是一个偏群胚。

10. 给定一个群胚（P， $\cdot$）， S的每个非空子集都是一个偏群胚: 对于所有 s_1、 $s_2 \in S$，如果$s_1 \cdot s_2 \notin S$，则 $s_1 \circ s_2$ 未定义。否则， $s_1 \circ s_2 = s_1 \cdot s_2$。 （$S$， $\circ$）是一个偏群胚。

一个非空集合 P可以关联一组偏运算。例如，数据库中的一组数据表 D 具有连接、组合、插入、更新、笛卡尔积等运算。

设（P， D， $\circ$）和（Q， E， $\cdot$）是两个偏群胚。偏群胚同态 f: （P， D， $\circ$） $\to$（Q， E， $\cdot$）是一个函数 f: $P \to Q$，满足



交换，即对于所有 x、 $y \in P$，如果 $x \circ y$有定义，则 f（x） $\cdot f$（y）有定义，并且 f（$x \circ y$） $=f$（x） $\cdot f$（y）。

偏群胚同态 f: ($P, D, \circ$) $\to$($Q, E, \cdot$) 的像，记作 im(f)，定义为偏群胚 ($f(P)$ $f(D)$ $\cdot$)，其中 $f(P) = \{f(p) \mid p \in P\}$，且 $f(D) = \{(f(p)\ f(q)) \mid (p, q) \in D\}$。如果对于每个偏群胚同态 f: ($P, D, \circ$) $\to$($Q, E, \cdot$)，偏群胚 im(f) 都满足该性质，则称偏群胚 ($P, D, \circ$) 上的一个性质在偏群胚同态下保持不变或具有不变性。

定义 2.2. 给定一个偏群胚（$P, \circ$）和一个正整数 n， P 的子集I_1、 $\cdots$、 I_n的子集乘积定义如下: $I_1 \cdots I_n \subseteq P$归纳地定义:

- *If* $n = 1$, $I_1 \cdots I_n = I_1$.

• 如果 $n > 1$,

$$x \in I_1 \cdots I_n \text{ if and only if } \left\{ \begin{array}{l} \exists \text{ postive integer } k : 1 \leq k < n \\ \exists\, y \in P : y \in I_1 \cdots I_k \\ \exists\, z \in P : z \in I_{k+1} \cdots I_n \end{array} \right\} x = yz$$

显然，乘积 $I_1 \cdots I_n$ 是所有可能值的集合

$$x_1 \cdots (x_i x_{i+1}) \cdots x_n,$$

其中， $x_1 \cdots x_i \cdots x_n$ 通过二元分组，并插入圆括号“(”和“)”， $x_i \in I_i$、 $i = 1$、 $\cdots$、 n。

$I_1 \cdots I_n$ 可以是单元素集合 $\{\uparrow\}$，如果将“(”和“)”以所有可能的方式插入 $x_1 \cdots x_i \cdots x_n$ 后，对于所有 $x_i \in I_i$、 $i = 1$、 $\cdots$、 n，始终得到 $\uparrow$。

如果 $I_1 = \cdots = I_n = I$，记作

$$I_1 \cdots I_n = I^n.$$

如果每个 $I_i = \{x_i\}$，记作

$$I_1 \cdots I_n = x_1 \cdots x_n.$$

在本文中，对于 x_1、 $\cdots$、 $x_n \in P$、 $x_1 \cdots x_n$，表示 $\{x_1\} \cdots \{x_n\}$ 的乘积。

偏群胚（P， $\circ$）的任意非空子集 T，连同运算 $\circ' = \circ|_T$（即将 $\circ$限制在集合 T上），都是一个偏群胚，称为（P， $\circ$）的偏子群胚。如果 T关于 $\circ$是闭的，即对于所有正整数 n都有 $T^n \subseteq T$，则偏子群胚（T， $\circ'$）称为闭偏子群胚。

设 $M \subseteq P$和 T是（P， $\circ$）的一个闭偏子群胚。如果 $M \subseteq T$且 T是包含 M的（P， $\circ$）的最小偏子群胚，则（T， $\circ$）由 M生成， M是 T的生成集，记为 $T = [M]$。显然，

$$[M] = \bigcup_{i=1}^{+\infty} M^i.$$

偏群胚（P， $\circ$）的生成集 M 是不可约的，如果不存在 M的真子集，同时它又是（P， $\circ$）的生成集。

定理 2.3 (1.4.9 [17])。如果一个偏群胚有一个有限生成集 M，那么每个生成集都包含一个不可约生成集 M。

根据定理 2.3，对于给定的有限生成偏群胚，我们可以假设其生成集是不可约的。

3 偏群胚性质

根据例 2.1.3，每个实体解析系统（$\mathbb{E}$、 $\approx$、 $\langle\ \rangle$）都是一个偏群胚。因此， [2], 中介绍的实体解析系统的性质，也可以在偏群胚上进行考察。我们考察偏群胚（P、 D、 $\circ$）的以下性质:

幂等性：对于所有 $p \in P$,

(I_{pg}) $(p,p) \in D$ and $p \circ p = p$.

每个元素始终都可以与自身复合，并且每个元素与自身的复合结果仍为该元素。

交换性：对于所有 p_1、$p_2 \in P$，如果 $(p_1, p_2) \in D$ 且 $(p_2, p_1) \in D$，那么

(C_{pg}) $p_1 \circ p_2 = p_2 \circ p_1$.

如果复合存在，则复合不取决于参与复合的元素顺序。

强交换性：对于所有 p_1、$p_2 \in P$,

(SC_{pg}) $(p_1,p_2) \in D \Leftrightarrow (p_2,p_1) \in D$ and if $(p_1,p_2) \in D$ then $p_1 \circ p_2 = p_2 \circ p_1$.

D 具有对称性，并且如果复合存在，则复合不取决于元素的顺序。

左表示性：令 p_1、$p_2 \in P$ 满足 $(p_1, p_2) \in D$。对于所有满足 $(p, p_1) \in D$ 的 $p \in P$,

(Rl_{pg}) $(p, p_1 \circ p_2) \in D$.

右代表性：令 p_1、$p_2 \in P$ 满足 $(p_1, p_2) \in D$。对于所有满足 $(p_2, p) \in D$ 的 $p \in P$,

(Rr_{pg}) $(p_1 \circ p_2, p) \in D$.

代表性:

(R_{pg}) Rl_{pg} and Rr_{pg}.

由 p_1 和 p_2 复合得到的元素 p_3 代表原始的 p_1 和 p_2：能够与 p_1 复合的元素 p 也能够从左侧与 p_3 复合，而能够与 p_2 复合的元素 p 也能够从右侧与 p_3 复合。

要定义一个偏半群（T、D、$\circ$），必须指定偏二元运算 $\circ$ 所满足的某一种结合性属性，以保证对于所有 x、$\cdots$、y、$\cdots$、$z \in T$，在表达式 $x \circ \cdots y \circ \cdots z$ 中重新排列括号都会得到相同的结果。令人惊讶的是，结合性属性有许多不同的分支。在 [17], 中，讨论了弱结合性、链式结合性、中间结合性、强结合性、强链式结合性、弱链式结合性，以及其中若干属性的 n 版本。本文关注给定偏群胚（P、D、$\circ$）上的以下三种结合性属性（A_{pg}）、（CA_{pg}）和（SA_{pg}）。

[2] 中的结合性：对于所有 p_1、p_2、$p_3 \in P$，如果（p_1、p_2）$\in D$，（p_2、p_3）$\in D$，（p_1、$p_2 \circ p_3$）$\in D$，并且（$p_1 \circ p_2$、p_3）$\in D$，则

$$(A_{pg}) \qquad p_1 \circ (p_2 \circ p_3) = (p_1 \circ p_2) \circ p_3 \neq \uparrow .$$

当 $p_1 \circ (p_2 \circ p_3)$ 和 $(p_1 \circ p_2) \circ p_3$ 均存在时，无论将 p_1、p_2、p_3 以何种顺序进行二元分组，对三个元素 p_1、p_2、p_3 进行组合所得的结果都相同。

链式结合性：对于所有 p_1、p_2、$p_3 \in P$，如果 $(p_1$、$p_2) \in D$ 且 $(p_2$、$p_3) \in D$，则

$$(CA_{pg}) \qquad p_1 \circ (p_2 \circ p_3) = (p_1 \circ p_2) \circ p_3 \neq \uparrow .$$

当 $p_1 \circ p_2$ 和 $p_2 \circ p_3$ 均存在时，无论将 p_1、p_2、p_3 以何种顺序进行二元分组，对三个元素 p_1、p_2、p_3 进行组合所得的结果都相同。

强结合性：对于所有 p_1、p_2、$p_3 \in P$,

$$(SA_{pg}) \qquad p_1 \circ (p_2 \circ p_3) = (p_1 \circ p_2) \circ p_3$$

其含义是，等式两边的值相同，或者二者同时未定义。这等价于以下任一情况：$(p_1 \circ p_2) \circ p_3$，且 $p_1 \circ (p_2 \circ p_3)$ 有定义；或者反之，二者均有定义且相等。

上述部分群胚性质之间的关系由以下两个命题给出。

命题 3.1. 对于给定的偏群胚（P、D、$\circ$），

（i）（P、$D_\circ$）满足（C_{pg}），且当且仅当（P、D、$\circ$）满足（SC_{pg}）时，D是对称的;

（ii）当且仅当（P、D、$\circ$）满足（A_{pg}）且（R_{pg}）时，（P、D、$\circ$）满足（CA_{pg}）.

证明。（i）显然。

（ii）“$\Rightarrow$”：对于所有 p_1、p_2、$p_3 \in P$，如果（p_1、p_2）、（p_2、p_3）$\in D$，那么根据（CA_{pg}），

$$(p_1 \circ p_2) \circ p_3 = p_1 \circ (p_2 \circ p_3) \neq \uparrow$$

因此（p_1）$p_2 \circ p_3$且（$p_1 \circ p_2$、p_3）$\in D$，从而（R_{pg}）成立。显然，（CA_{pg}）$\Rightarrow$（A_{pg}）。因此，（CA_{pg}）蕴含（A_{pg}）和（R_{pg}）。

“$\Leftarrow$”：假设 $p_1 \circ p_2$ 和 $p_2 \circ p_3$ 均存在。那么根据（R_{pg}），$(p_1 \circ p_2) \circ p_3$ 和$p_1 \circ (p_2 \circ p_3)$ 均存在，因此根据（A_{pg}），$(p_1 \circ p_2) \circ p_3 = p_1 \circ (p_2 \circ p_3) \neq \uparrow$。因此，（$CA_{pg}$）成立。

□

命题 3.2。对于偏群胚（P、D、$\circ$），

(1) 强结合性（SA_{pg}）蕴含结合性（A_{pg}）；

（2）链式结合性（CA_{pg}）蕴含结合性（A_{pg}）。

证明。（1）显然。

（2）对于所有 x、y、$z \in P$，若 $(x \circ y) \circ z \neq \uparrow$且 $x \circ (y \circ z) \neq \uparrow$，则 $x \circ y \neq \uparrow$且 $y \circ z \neq \uparrow$。因此，根据 (CA_{pg})，$(x \circ y) \circ z = x \circ (y \circ z) \neq \uparrow$，所以 (A_{pg}) 成立。

□

然而，偏群胚中的链式结合性（CA_{pg}）与强结合性（SA_{pg}）彼此独立，正如 [17] (1.5.7) 中的例子所示；命题 3.2 中各陈述的逆命题也不成立，以下例 3.3.1对此进行了说明。

例 3.3。（1）上面例 2.1.1 中定义的偏群胚（P_1、$\cdot$）显然满足（I_{pg}）。它也满足结合性（A_{pg}），因为

$$(a \cdot b) \cdot c = b \cdot c = c \text{ and } a \cdot (b \cdot c) = a \cdot c = \uparrow .$$

但是（$P_1, \cdot$）不满足强结合性（SA_{pg}）和链式结合性（CA_{pg}）。因此，命题 3.2 中各陈述的逆命题并不成立。

（2）由于识别并保留重叠部分的一个副本的行为类似于集合并，例 2.1.5 中的（$\mathcal{P}ath$（G）$\sim_{\mathcal{P}}$◊）和（$\mathcal{P}ath$（$\mathcal{C}_W$）$\sim_{\mathcal{P}}$◊）、例 2.1.8 中的（$O/\mathbb{O}$, ∨ ⊔），以及例 2.1.9 中的（$E, \smile, \mathbb{M}$）满足（I_{pg}）、（SC_{pg}）、（A_{pg}）、（R_{pg}）、（CA_{pg}）和（SA_{pg}）。

(3) （$\mathbb{N}, \vee$）和（ $\mathbb{N}, \wedge$）在例 2.1.2 中满足（I_{pg}）、（SC_{pg}）、（A_{pg}）、（R_{pg}）、（CA_{pg}）和（SA_{pg}）.

(4) 示例 2.1.4 中的 $(A, *)$ 满足 (A_{pg})，但不满足 (I_{pg})、(SC_{pg})、(R_{pg})、(CA_{pg}) 和 (SA_{pg}) 中的任何一个。

(5) 示例 2.1.6 中的 $(T, \bowtie_l)$ 满足 (I_{pg})、(A_{pg})、(R_{pg})、(CA_{pg}) 和 (SA_{pg})，但不满足 (SC_{pg})。示例 2.1.6 中的 $(T,\&)$ 满足 (SC_{pg})、(A_{pg})、(R_{pg})、(CA_{pg}) 和 (SA_{pg})，但不满足 (I_{pg})。

(6) 示例 2.1.7 中的 $(\mathcal{L}, \rightsquigarrow, c)$ 满足 (A_{pg})、(R_{pg})、(CA_{pg}) 和 (SA_{pg})，但不满足 (I_{pg}) 和 (SC_{pg})。

这些性质在下述命题所示的偏群胚同态下保持不变。

命题3.4. 给定一个偏群胚（P、D、$\circ$），以下性质：(S)、(I_{pg})、(C_{pg})、(SC_{pg})、(R_{pg})、(A_{pg})、(CA_{pg}) 和 (SA_{pg}) 在偏群胚同态下保持不变。

证明。设 f: $(P$、D、$\circ)$ $\to (Q$、E、$\cdot)$ 是一个偏群胚同态。

若 D是对称的，显然 $f(D) = \{(f(x)\ f(y))|(x$、$y) \in D\}$ 也是对称的。

如果（P，D，$\circ$）满足（I_{pg}），那么，对于每个 $p \in P$，（p，p）$\in P$和 $p \circ p = p$。因此，对于每个f（p）$\in f$（P），$p \in P$,

$$(f(p), f(p)) = f(p, p) \in f(D)$$

and

$$f(p)\cdot f(p)=f(p\cdot p)=f(p),$$

因此 $\mathrm{im}(f)=(f(P)\ f(D)\ \cdot)$ 满足（I_{pg}）。

假设（P，D，$\circ$）满足（C_{pg}）。对于所有 f（p） f（q） $\in f$（P），有 f（p） $\cdot f$（q） $=f$（$p\circ q$）$=f$（$q\circ p$） $=f$（q） $\cdot f$（p）。因此，im（f） $=$（f（P） f（D） $\cdot$）满足（C_{pg}）。

若（P，D，$\circ$）满足（A_{pg}），则对于所有 f（p） f（q） f（r） $\in$im（f），若 f（p） $\cdot$（f（q） $\cdot f$（r））且（f（p） $\cdot f$（q）） $\cdot f$（r）存在，则有

$$f(p)\cdot(f(q)\cdot f(r))=f(p\circ(q\circ r))=f((p\circ q)\circ r)=(f(p)\cdot f(q))\cdot f(r).$$

因此，im（f）满足（A_{pg}）。

假设（P，D，$\circ$）满足（R_{pg}）。对于所有 f（p） f（q） f（r） $\in f$（P），使得

$$(f(p),f(q)),(f(q),f(r))\in f(D),$$

我们有 (p,q)、$(q,r)\in D$。因此 $(p\circ q,r)\in D$，从而 $(f(p)\circ f(q)\ f(r))\in f(D)$。因此，$\mathrm{im}(f)$ 满足 (Rr_{pg})。类似地，在偏群胚同态下，(Rl_{pg}) 得以保持，(R_{pg}) 也如此。

如果 $(P,D,\circ)$ 满足 (SA_{pg})，那么，对于所有 $f(p)\ f(q)\ f(r)\in f(P)$，

$$f(p)\cdot(f(q)\cdot f(r))=f(p\circ(q\circ r))=f((p\circ q)\circ r)=(f(p)\cdot f(q))\cdot f(r).$$

因此 $\mathrm{im}(f)$ 满足 (SA_{pg})。 □

4 偏群胚中的自然偏序

实体记录合并旨在将实体记录组合起来，以获取更多信息。为了比较实体记录的信息，需要在实体记录上定义一个偏序。

回顾一下，集合 P 上的偏序是 P 上的自反、反对称且传递的二元关系。众所周知，对于每个半群 T， T 中所有幂等元构成的集合 E_T 在 $\leq_i$ 下构成偏序：

$$e\leq_i f \text{ if and only if } e=ef=fe.$$

半群上的偏序若由半群运算 [20] 定义，则称为自然偏序。我们在偏群胚上定义 $\leq_l$、 $\leq_r$ 和 $\leq$ 如下。

定义 4.1. 设 $(P$、 D、 $\circ)$ 为偏群胚，且 p、 $q\in P$。定义

(i) 当且仅当 $pq=q$ 时， $p\leq_r q$；

(ii) 当且仅当 $qp=q$ 时， $p\leq_l q$；（

$iii)$ 当且仅当 $pq=qp=q$ 时， $p\leq q$。

显然，将其限制在所有幂等元组成的集合上时， $\leq$ 是 $\leq_i$的对偶。根据以下引理，这些定义是有意义的。

引理 4.2。给定一个偏群胚（P、 D、 $\circ$）

(*i*) $\leq$是反对称的;

(*ii*) 如果（P、 D、 $\circ$）满足（I_{pg}），那么 $\leq_l$、 $\leq_r$和 $\leq$是自反的;

(*iii*) 如果（P、 D、 $\circ$）满足（CA_{pg}），那么 $\leq_l$、 $\leq_r$和 $\leq$是传递的。

证明。(*i*) 如果 $p \leq q$ 和 $q \leq p$，那么

$$p = pq = qp \text{ and } q = pq = qp$$

因此 $p = pq = q$。所以 $\leq$ 也是反对称的。

(*ii*) 由于（P、 D、 $\circ$）满足（I_{pg}），对于所有 $p \in P$， $\leq_l$、 $\leq_r$和 $\leq$都是自反的，因为 $p = pp$。

(*iii*) 若 $p \leq_l q$ 和 $q \leq_l r$，则 $qp = q$ 和 $rq = r$。因此

$$\begin{aligned} r &= r(qp) \\ &= (rq)p \quad (\text{by } (CA_{pg})) \\ &= rp, \end{aligned}$$

并且因此 $p \leq_l r$。于是， $\leq_l$ 是传递的。同理， $\leq_r$ 和 $\leq$ 都是传递的 .

□

注记 4.3. 上述引理 4.2 中的条件“结合性 (CA_{pg})”是必要的，不能用结合性 (A_{pg}) 代替。

例如，设 $(P_1, \cdot)$ 是例 2.1.1 中定义的幂等偏群胚。则 $(P_1, \cdot)$ 满足 (A_{pg})，因为 $(a \cdot b) \cdot c = c$ 和 $a \cdot (b \cdot c) = \uparrow$。显然， $a \leq_r b$ 和 $b \leq_r c$，但 $a \not\leq_r c$。也就是说， $\leq_r$ 不是传递的。

复合包含更多信息，如下列引理所示。

引理 4.4. 在满足 (I_{pg}) 和 (CA_{pg}) 的偏群胚中，如果 pq 存在，则 $q \leq_l pq$ 和$p \leq_r pq$。

证明。由（I_{pg}）和（CA_{pg}），我们有

$$(pq)q = p(qq) = pq$$

and

$$p(pq) = (pp)q = pq.$$

因此 $q \leq_l pq$ 和 $p \leq_r pq$。 □

极大元定义如下:

定义 4.5。在偏群胚中，如果 $m \leq_r n$（$m \leq_l n$或 $m \leq n$）蕴含 $n \leq_r m$（$n \leq_l m$或 $n \leq m$），则称 m 关于 $\leq_r$ （$\leq_l$ 或 $\leq$）是极大元。

二元关系 $\leq_l$、 $\leq_r$ 和 $\leq$ 具有相同的极大元，如下述引理 4.6 和命题 4.7 所示。

引理 4.6。在满足 (I_{pg})、(S) 和 (CA_{pg}) 的偏群胚中，以下命题等价：

(i) m关于 $\leq_r$ 是极大的；

(ii) m关于 $\leq_l$ 是极大的；

(iii) m关于 $\leq$ 是极大的。

证明。$(i) \Rightarrow (iii)$：假设 m 关于 $\leq_r$ 是极大的，且 $m \leq n$。则 $n = mn = nm$，从而 $m \leq_r n$。因此 $n \leq_r m$，进而 $n = nm = m$。故 m 关于 $\leq$ 是极大的。

$(iii) \Rightarrow (i)$：假设 m 关于 $\leq$ 是极大的，且 $m \leq_r n$。则 $n = mn$，因此根据 (S)，nm 存在。于是，

$$\begin{aligned} m(nm) &= (mn)m \quad (\text{by } (CA_{pg})) \\ &= nm \\ &= (nm)m. \end{aligned}$$

因此 $m \leq nm$，从而 $m = nm$，因为 m 关于 $\leq$ 是极大的。因此，$n \leq_r m$。所以 m 关于 $\leq_r$ 是极大的。因此，$(iii) \Rightarrow (i)$。

$(iii) \Leftrightarrow (ii)$ 与 $(iii) \Leftrightarrow (i)$ 相似。 □

Let

$$M_l(P) = \{\text{all left maximal elements in partial groupoid } P\},$$

$$M_r(P) = \{\text{all right maximal elements in partial groupoid } P\},$$

and

$$M(P) = \{\text{all maximal elements in partial groupoid } P\}.$$

根据引理 4.6，我们有：

命题 4.7。在满足 (S)、(I_{pg}) 和 (CA_{pg}) 的偏群胚 $(P, D, \circ)$ 中，$M_l(P) = M_r(P) = M(P)$。

极大元构成了 [2] 中定义的实体解析。

定理 4.8。给定一个满足 (I_{pg}) 和 (CA_{pg}) 的实体解析系统 $(\mathbb{E}, \approx, \langle\rangle)$，以及一个实例 $I \subseteq \mathbb{E}$，使得 $\overline{I}$ 是有限的，则 $ER(I) = M(\overline{I})$。

证明。对于每个 $r \in \overline{I}$，如果相对于 $\leq$，r 在 $\overline{I}$ 中是极大的，那么 $r \in M(\overline{I})$，因此存在 $r \in M(\overline{I})$ 使得 $r \leq r$。否则，存在 $s_1 \in \overline{I}$ 使得 $r \leq s_1$ 且 $s_1 \neq r$。

如果相对于 $\leq$ 而言，s_1 在 $\overline{I}$ 中是极大的，那么 $s_1 \in M(\overline{I})$，因此存在 $s_1 \in M(\overline{I})$ 使得 $r \leq s_1$。否则，存在 $s_2 \in \overline{I}$ 使得 $s_1 \leq s_2$ 且 $s_2 \neq s_1$。

以此方式继续，我们在 $\overline{I}$ 中得到如下升链：

$$r < s_1 < s_2 < \cdots$$

而这必然会在 $\overline{I}$ 是有限的情况下终止。因此，存在 $s_n \in M(\overline{I})$使得 $r \leq s_n$ ，从而$\overline{I} \leq M(\overline{I})$。显然，$M(\overline{I}) \subseteq \overline{I}$，并且 $M(\overline{I})$相对于 $ER(\overline{I})$定义中的性质而言是最小的。因此，$M(\overline{I}) = ER(I)$。□

注 4.9。我们不仅关注极大元，也关注某些使用场景中的最小元或其他特殊元，例如，在实体对象的偏群胚中寻找具有最少信息的原子元或不可约生成元。类似地，我们可以分别在偏群胚中定义相对于 $\leq_l$、 $\leq_r$和 $\leq$的最小元。

设（P、 D、 $\circ$）为一个偏群胚，且 $\preceq$是 P上的一个给定偏序。为研究 $\preceq$ 与 $\circ$之间的相互作用，我们考虑相对于 $\preceq$的以下性质：

复合给出最小上界：对于所有 p_1、 $p_2 \in P$，使得 (p_1)、$(p_2) \in D$,

(LU_{pg}) $p_1 \circ p_2$ is the least upper bound of p_1 and p_2 with respect to $\preceq$.

$\preceq$与定义域和复合左相容：对于所有 p_1、 p_2、 $p \in P$，使得 $p_1 \preceq p_2$,

(lCP_{pg}) if $(p, p_1) \in D$ then $(p, p_2) \in D$ and $p \circ p_1 \preceq p \circ p_2$.

$\preceq$与定义域和复合右相容：对于所有 p_1、 p_2、 $p \in P$，使得 $p_1 \preceq p_2$,

(rCP_{pg}) if $(p_1, p) \in D$ then $(p_2, p) \in D$ and $p_1 \circ p \preceq p_2 \circ p$.

$\preceq$ 与定义域和复合相容：如果 $\preceq$ 是

(CP_{pg}) both left and right compatible with domain and composition.

以下 命题表明，(I_{pg})、(C_{pg})、(A_{pg}) 和 (R_{pg}) 蕴含 (LU_{pg}) 和 (CP_{pg}) .

命题 4.10。设 $(P, D, \circ)$ 为偏群胚。若 $(P, D, \circ)$ 满足 (I_{pg})、(C_{pg})、(A_{pg}) 和 (R_{pg})，则相对于 $\leq$，$(P, D, \circ)$ 满足 (LU_{pg}) 和 (CP_{pg})。

证明。设 $p_1, p_2 \in P$ 使得 $(p_1, p_2) \in D$。由于

$$p_1(p_1p_2) = (p_1p_1)p_2 = p_1p_2$$

and

$$(p_1p_2)p_1 = p_1(p_2p_1) = p_1(p_1p_2) = p_1p_2$$

根据（I_{pg}）、（A_{pg}）和（R_{pg}），我们有 $p_1 \leq p_1p_2$。同样地，我们 $p_2 \leq p_1p_2$。那么 p_1p_2 是 p_1 和 p_2 的上界。

如果 $p_1 \leq x$ 和 $p_2 \leq x$，那么

$$p_1x = xp_1 = x \text{ and } p_2x = xp_2 = x.$$

因此

$$(p_1p_2)x = p_1(p_2x) = p_1x = x$$

and

$$x(p_1p_2) = (xp_1)p_2 = xp_2 = x$$

因此 p_1 $p_2 \leq x$。因此， p_1p_2 是 p_1 和 p_2的最小上界。所以（LU_{pg}）成立 e.
如果 $p_1 \leq p_2$ 且（p， p_1）$\in D$，那么

$$p_2 = p_1p_2 = p_2p_1 \text{ and } pp_1 \text{ exists}$$

因此，根据（A_{pg}）和（R_{pg}），

$$pp_2 = p(p_1p_2) = (pp_1)p_2 \neq \uparrow .$$

因此，（p， p_2）$\in D$。计算

$$\begin{aligned}
(pp_1)(pp_2) &= (pp_2)(pp_1) \quad (\text{by } (SC_{pg})) \\
&= ((pp_2)p)p_1 \quad (\text{by } (CA_{pg})) \\
&= (p(pp_2))p_1 \quad (\text{by } (CA_{pg})) \\
&= p(p_2p_1) \quad (\text{by } (I_{pg}) \text{ and } (CA_{pg})) \\
&= pp_2.
\end{aligned}$$

于是 $pp_1 \leq pp_2$，因此（lCP_{pg}）成立。同样，（rCP_{pg}）也得到满足。因此（CP_{pg}）成立。 □

另一方面，当定义域是自反的时，（LU_{pg}）和（CP_{pg}）蕴含（I_{pg}）、（C_{pg}）、（A_{pg}）以及（R_{pg}）。

命题 4.11. 设（P、 D、 ∘）是一个偏群胚， D是自反的，且 $\preceq$是 P上的偏序。如果（P、 D、 ∘）关于 $\preceq$满足（LU_{pg}）和（CP_{pg}），那么（P、 D、 ∘）满足（I_{pg}）、（C_{pg}）、（A_{pg}）和（R_{pg}），并且 $\preceq = \leq$。

证明。假设（P、 D、 ∘）关于 $\preceq$满足（LU_{pg}）和（CP_{pg}）。

幂等性：对于所有 $p \in P$，由于 D 是自反的， pp 存在。根据（LU_{pg}），

$$p \leq pp \text{ and } pp \leq p$$

因此 $pp = p$。（I_{pg}）成立。

交换性：对于所有 p_1、 $p_2 \in P$，如果 p_1p_2和 p_2p_1存在，那么根据（LU_{pg}），

$$p_1p_2 \leq p_2p_1 \text{ and } p_2p_1 \leq p_1p_2$$

因此满足 $p_1p_2 = p_2p_1$。（C_{pg}）

结合性：对于所有满足 p_1、p_2、$p_3 \in P$ 的情况，由于 p_1（p_2p_3）和（p_1p_2）p_3 存在，且 $p_1 \leq p_1p_2$ 和（p_1、p_2p_3）$\in D$,

$$p_1(p_2p_3) \leq (p_1p_2)(p_2p_3).$$

由于根据（LU_{pg}）有 $p_1p_2 \leq p_1(p_2p_3)$ 和 $p_2p_3 \leq p_1(p_2p_3)$，我们得到

$$(p_1p_2)(p_2p_3) \leq p_1(p_2p_3).$$

因此

$$p_1(p_2p_3) = (p_1p_2)(p_2p_3).$$

类似地，（p_1p_2）$p_3 = (p_1p_2)(p_2p_3)$。因此，$p_1(p_2p_3) = (p_1p_2)p_3$，正如所需。
（A_{pg}）成立。

代表性：如果 p_1p_2 存在且（p, p_1）$\in D$，那么根据（lCP_{pg}），（p, p_1p_2）$\in D$，因此（Rl_{pg}）成立。类似地，我们有（Rr_{pg}）。于是，（R_{pg}）成立。

如果 $p_1 \preceq p_2$，那么根据（CP_{pg}）和（I_{pg}），

$$p_1p_2 \preceq p_2p_2 = p_2 \text{ and } p_2p_1 \preceq p_2p_2 = p_2$$

因此 $p_2p_1 = p_1p_2 = p_2$ 由于 $p_2 \preceq p_1p_2$ 且 $p_2 \preceq p_2p_1$ 根据（LU_{pg}）。所以 $p_1 \leq p_2$。

另一方面，如果 $p_1 \leq p_2$，那么 $p_1p_2 = p_2p_1 = p_2$，因此 $p_1 \preceq p_2$ 根据（LU_{pg}）。所以，$\preceq = \leq$。□

在（P，D，$\circ$）上的偏序中，如果 D 是自反的，则该偏序必为自然偏序 $\leq$，并且（I_{pg}）、（C_{pg}）、（A_{pg}）和（R_{pg}）成立，当且仅当结合命题 4.10 和 4.11 后，（LU_{pg}）和（CP_{pg}）得到满足。

定理 4.12。设（P，D，$\circ$）为偏群胚，D 是自反的，且 $\preceq$ 是 P 上的偏序。那么，（P，D，$\circ$）关于 $\preceq$ 满足（LU_{pg}）和（CP_{pg}），当且仅当（P，D，$\circ$）满足（I_{pg}）、（C_{pg}）、（A_{pg}）和（R_{pg}），并且 $\preceq = \leq$。

注 4.13。在具有传递关系或自反且传递关系的非空集合中研究了排序和选择 [8, 11]。因此，我们可以在具有传递序的偏群胚（P、D、$\circ$）中对元素进行排序、选择和查询。

5 约化偏性

众所周知，一般而言，实体解析的匹配与不匹配标签高度不平衡（类别不平衡问题）。因此，给定一个实体解析系统（$\mathbb{E}$、$\approx$、$\langle\rangle$）和一个实体记录 $e \in \mathbb{E}$、e 只能与集合 $\mathbb{E}$ 的一小部分进行组合。也就是说，$\langle e, -\rangle$ 和 $\langle -, e\rangle$ 仅定义在集合 $\mathbb{E}$ 的非常小的一部分上。

偏群胚（P、D、$\circ$）的偏二元运算 $\circ$ 的偏性 D 由其定义域 D 度量。由于 D 是 $P \times P$ 的子集，D 是 P 上的二元关系，因此可以表示为有向图。

定义 5.1。有限偏群胚（P、 D、 $\circ$）的定义域图记为 $\mathcal{G}_P$，定义为具有以下结构的有向图：

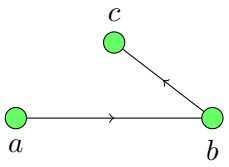
节点： P的所有元素，

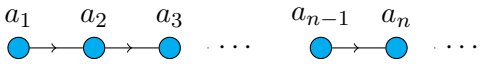
边：如果定义了 $p_1 \circ p_2$ ，则存在一条从节点 p_1 到节点 p_2 的边。

偏群胚（$P, D, \circ$）是连通的，当且仅当其定义域图 $\mathcal{G}_P$是连通图。

如果 D 是对称的，那么 $\mathcal{G}_P$ 可以视为简单无向图。

例 5.2. 1. 例 2.1.1 和 2.1.4 中定义的偏群胚的定义域图分别如下所示：





回顾一下，无向图中的路径是由一系列边组成的序列，这些边依次连接一系列互不相同的顶点。有向图中的有向路径是由方向相同的一系列边组成的序列，这些边依次连接一系列互不相同的顶点。无向图中的一对顶点 u 和 v 是连通的，当且仅当这两个顶点之间存在一条路径。有向图中的一对顶点 u 和 v 是强连通的，当且仅当存在一条从 u 到 v 的有向路径，以及一条从 v 到 u的有向路径。

如果每一对不同的顶点都通过唯一一条边（每个方向各一条唯一的边）相连，则无向图（有向图）是完全的。

命题 5.3。设 $(P, D, \circ)$ 是有限偏群胚，且 D 对称。则 $(P, D, \circ)$ 是全的，当且仅当其定义域图 $\mathcal{G}_P$ 是完全图。

证明。偏半群 $(P, D, \circ)$ 是全的，当且仅当对于所有 p_1， $p_2 \in P$, $p_1 \circ p_2$ 有定义；当且仅当 $\mathcal{G}_P$ 中每一对不同的节点都通过其唯一的边相连；当且仅当其定义域图 $\mathcal{G}_P$ 是完全图。□

无向图的一个连通分支是一个连通子图，并且不被恰当地包含在任何其他连通子图中。众所周知，每个有限无向图都可以在线性时间内划分为不相交的连通分量。设 $\{C_1, \cdots, C_n\}$ 为 $\mathcal{G}_P$ 的所有连通分量。则 $P = \bigcup_{i=1}^{n} P_i$，其中每个 $P_i = P|_{C_i}$ 都是将 P 限制于 C_i 的子群胚，$P_i \cap P_j = \emptyset$，并且对于每个 $s \in P_i$，当 $i \neq j$ 时，$t \in P_j$、 $s \circ t$ 和 $t \circ s$ 都未定义。每个 P_i 都称为 P 的连通分支偏子群胚。

对全群胚（半群）的研究比对偏群胚的研究更为广泛。因此，最好能用一些全群胚（半群）覆盖一个偏群胚（半群）。

G 的一个团是 G 的完全子图，而 G 的一个极大团则是不包含于任何其他团中的团。 G 的一个团覆盖是 G的一个团集合，其并集为G。

已有用于求给定无向（有向）图的极大团和团覆盖的算法 [21, 23, 25] ，因此可以将偏群胚 P 分解为与 $\mathcal{G}_P$ 的团覆盖相对应的全子群胚的并集。

消除偏群胚（半群） P 的偏性的另一种方法，是将 P 扩张为全群胚（半群） T，使得 P 可以成为 T 的一部分。

命题 5.4. 设 $T^* = T \cup \{\uparrow\}$ 是偏半群 T 的零扩张：对于所有 $s \in T$、$s\circ\uparrow = \uparrow\circ s = \uparrow$ 和 $s \circ t = \uparrow$，若 $s \circ t$ 未定义，则 T^* 是一个（幂等元）半群，当且仅当 T 是满足（SA_{pg}）的一个（幂等元）偏群胚。

证明。“⇐”：对于所有 x、y、$z \in T^*$，若 x、y、z 中有一个是 $\uparrow$，则（xy）$z = \uparrow = x$（yz）.

对于所有 x、y、$x \in T^*$，如果 $(xy)z$ 和 $x(yz)$ 都有定义，则 $(xy)z = x(yz)$。如果 $(xy)z$ 和 $x(yz)$ 都未定义，则 $(xy)z = \uparrow = x(yz)$。因此， T^* 是半群。

“⇒”：对于所有 x、y、$z \in T$，由于 T^* 是半群，我们在 T^* 中有 $x(yz) = (xy)z$。因此，$x(yz) = (xy)z$ 在 T 中成立，从而 T 满足 (SA_{pg})。

显然， T^* 是幂等元，当且仅当 T 是幂等元。 □

6 消除非交换性

交换性 (C) 在 [2] 中十分重要。本节将证明，每个满足 (NR_{pg}) 和 (S) 的偏群胚，都有一个交换偏群胚作为其同态像。我们考虑偏群胚 $(P, D, \circ)$ 上的以下条件：

$$(NR_{pg}) \qquad r_1 \cdots r_k r_1 \cdots r_k = r_1 \cdots r_k$$

对于所有 r_1、$\cdots$、$r_k \in P$，其中 $k \geq 1$ 是一个整数

给定一个幂等半群 T，结合性意味着，在表达式 $r_1 \cdots r_k r_1 \cdots r_k$ 中重新排列括号，对于所有 r_1、$\cdots$、$r_k \in T$，所得结果都相同。因此

$$r_1 \cdots r_k r_1 \cdots r_k = (r_1 \cdots r_k)(r_1 \cdots r_k) = r_1 \cdots r_k$$

因此，每个幂等半群都满足（NR_{pg}）和（S）。

显然，如果偏群胚（P、D、$\circ$）满足（NR_{pg}），那么它满足（I_{pg}） .

(NR_{pg}) 在偏群胚同态下保持不变：

引理 6.1。设 f：$(P, D, \circ) \to (Q, E, \cdot)$ 是一个偏群胚同态。则

(*i*) 如果 $(P, D, \circ)$ 满足 (NR_{pg})，那么 im(f) 也满足该条件；

(*ii*) 如果 $(P, D, \circ)$ 是有限生成的，那么 im(f) 也是有限生成的。

证明。（i）显然。

（ii）如果（P， D， $\circ$）是由 $\{p_1, \cdots, p_n\}$ 有限生成的偏群胚，那么 im（f） =（f（P）.f（D） ·）是由 $\{f(p_1)$ $\cdots$、 $f(p_n)\}$ 有限生成的偏群胚。

□

定义 6.2。定义 $\sim_c$ 上的关系为

$$p \sim_c q \text{ if and only if } pqp = p \text{ and } qpq = q.$$

显然，如果 $p \sim_c q$ ，那么 pq 和 qp 都存在。

引理 6.3。在给定的偏群胚（P、 D、 $\circ$）中，如果

$$x_1 \cdots x_i = y_1 \cdots y_j \text{ and } p = p_1 \cdots p_s x_1 \cdots x_i q_1 \cdots q_t,$$

then

$$p = p_1 \cdots p_s x_1 \cdots x_i q_1 \cdots q_t = p_1 \cdots p_s y_1 \cdots y_j q_1 \cdots q_t$$

对于所有 x_1、 $\cdots$、 x_i、 $y_1 \cdots$、 y_j、 p、 p_1、 $\cdots$、 p_s、 q_1、 $\cdots$、 q_t、 r_1、 $\cdots$、 $r_k \in P$，其中 $s \geq 0$、 $t \geq 0$、 $i \geq 1$ 和 $j \geq 1$ 是整数。

证明。证明通过对 j 使用归纳法进行。

如果 $j = 1$，那么 $x_1 \cdots x_i = y_1$ 和 $p = p_1 \cdots p_s x_1 \cdots x_i q_1 \cdots q_t$。由于二者都

$$x_1 \cdots x_i \text{ and } p_1 \cdots p_s x_1 \cdots x_i q_1 \cdots q_t$$

结果为单个元素，因此无论表达式中的元素如何分组或括号放在何处，都可以执行运算 $\circ$。因此

$$p = p_1 \cdots p_s x_1 \cdots x_i q_1 \cdots q_t = p_1 \cdots p_s y_1 q_1 \cdots q_t.$$

假设 引理中的结论在 $j = h$ 时成立。对于 $h + 1$ 的情形，假设 t

$$x_1 \cdots x_i = y_1 \cdots y_{h+1} \text{ and } p = p_1 \cdots p_s x_1 \cdots x_i q_1 \cdots q_t.$$

Then

$$x_1 \cdots x_i = y_1 \cdots y_{h-1}(y_h y_{h+1}) \text{ and } p = p_1 \cdots p_s x_1 \cdots x_i q_1 \cdots q_t.$$

根据归纳假设，

$$\begin{aligned} p &= p_1 \cdots p_s x_1 \cdots x_i q_1 \cdots q_t \\ &= p_1 \cdots p_s y_1 \cdots y_{h-1}(y_h y_{h+1}) q_1 \cdots q_t \quad \text{(by induction hypothesis)} \\ &= p_1 \cdots p_s y_1 \cdots y_h y_{h+1} q_1 \cdots q_t. \quad \text{(since the expression results in a single element)} \end{aligned}$$

因此，引理对 $j = h + 1$ 成立，正如所需。 □

对于 p，如果 $q \in P$ 被 q 左可除（右可除），或 q 是 p 的左因子（右因子），则记为 $q|_l p$（$q|_r p$），其中存在 $x \in P$ 使得 $p = qx$ （$p = xq$）。

引理 6.4. 给定一个满足 (NR_{pg}) 的偏群胚 $(P,\ D,\ \circ)$ 以及 p、q、$r \in P$，当相关运算存在时，

(i) $pq \sim_c qp$；(ii) 如果 $p \sim_c q$ 且 $p \sim_c r$，
那么 $pqrp = prqp = p$；(iii) $\sim_c$ 是等价关系；(
iv) 如果 $p|_l q$ 且 $q|_l p$，那么 $pr \sim_c qr$ 且 $rp \sim_c rq$；
(v) 如果 $p|_r q$ 且 $q|_r p$，那么 $pr \sim_c qr$ 且 $rp \sim_c rq$；
(vi) 如果 $p \sim_c q$，那么 $pr \sim_c qr$ 且 $rp \sim_c rq$。

证明。(i) 由 (NR_{pg}) 可知 $pq = pq\,pq = pq\,qp\,pq$ 且 $qp = qp\,qp = qp\,pq\,qp$，因此有 $pq \sim_c qp$。

(ii) 由于 $p \sim_c q$ 且 $p \sim_c r$，根据引理 6.3，我们有

$$\begin{aligned} p &= pqp \quad (\text{since } p \sim_c q) \\ &= prp\,qp \quad (\text{since } prp = p) \\ &= p\,rpqrpq\,p \quad (\text{since } rpq = rpqrpq) \\ &= pqrp \quad (\text{since } p = prp = pqp) \end{aligned}$$

and

$$\begin{aligned} p &= pqp\,prp \quad (\text{since } p = pp) \\ &= p\,qpr\,p \quad (\text{since } qppr = qpr) \\ &= p\,qprqpr\,p \quad (\text{since } qprqpr = qpr) \\ &= pqp\,rq\,prp \\ &= prqp. \end{aligned}$$

(iii) 对于每个 $p \in P$，根据 (NR_{pg})，$ppp = p$，因此 $p \sim_c p$。所以 $\sim_c$是自反的。

如果 $p \sim_c q$，那么 $pqp = p$ 和 $qpq = q$，从而 $qpq = q$ 和 $pqp = p$。因此，$q \sim_c p$。所以，$\sim_c$ 是对称的。

假设 $p \sim_c q$ 和 $q \sim_c r$。那么，根据（ii），$qprq = qrpq = q$，因此，根据引理 6.3，

$$\begin{aligned} p &= pqqp \quad (\text{since } p = pqp \text{ and } q = qq) \\ &= pqp\,rqqr\,pqp \quad (\text{since } qprq = qrpq = q) \\ &= p\,rqr\,p \quad (\text{since } p = pqp, q = qq) \\ &= prp \quad (\text{since } r = rqr) \end{aligned}$$

and

$$\begin{aligned} r &= rqr \quad (\text{since } r = rqr) \\ &= r\ qprq\ r \quad (\text{since } q = qprq) \\ &= rqp\ r \quad (\text{since } rqr = r) \\ &= r\ qrpq\ pr \quad (\text{since } q = qrpq) \\ &= rpr. \quad (\text{since } rqr = r \text{ and } pqp = p) \end{aligned}$$

因此， $p \sim_c r$ ，从而 $\sim_c$ 具有传递性。

(iv) 假设对于某些 x、$y \in P$， $p = qx$ 且 $q = py$。则

$$\begin{aligned} pr\ qr\ pr &= prqr\ qx\ r \\ &= prqxr \quad (\text{since } rqrq = rq) \\ &= pr \end{aligned}$$

and

$$\begin{aligned} qr\ pr\ qr &= qrpr\ py\ r \\ &= qrpyr \quad (\text{since } rprp = rp) \\ &= qr. \end{aligned}$$

因此 $pr \sim_c qr$。类似地，

$$rp\ rq\ rp = rprq\ r\ qx = rprqx = rp$$

and

$$rq\ rp\ rq = rqrp\ r\ py = rqrpy = rq.$$

因此 $rp \sim_c rq$。

(v) 与（iv）类似。

(vi) 如果 $p \sim_c q$，则 $p|_l pq$、 $pq|_l q$ 以及 $pq|_r q$、 $q|_r pq$。因此，由（iv）和（v）可得，
$pr \sim_c pqr \sim_c qr$。由（i），我们有 $rp \sim_c rq$。

□

$\sim_c$ 与复合相容，如下所示：

引理 6.5. 给定一个满足 (S) 和 (R_{pg}) 的偏群胚 $(P, D, \circ)$，若 $p_1 \sim_c q_1$ 且 $p_2 \sim_c q_2$，则 $(p_1, p_2) \in D$ 当且仅当 $(q_1, q_2) \in D$。

证明。由于 $p_i \sim_c q_i$, $i = 1,2$，我们有 $p_i = p_i q_i p_i$ 且 $q_i = q_i p_i q_i$, $i = 1, 2$。

如果 $(p_1, p_2) \in D$，那么根据 (S) 和 (R_{pg})，$(p_1, q_2 p_2 q_2) \in D$。也就是说，$(p_1, q_2) \in D$。同样地，根据 (S) 和 (R_{pg})，$(q_1 p_1 q_1, q_2) \in D$，因此 $(q_1, q_2) \in D$。

由于 $\sim_c$ 是等价关系，如果 $(q_1, q_2) \in D$，那么对称地，$(p_1, p_2) \in D$。因此，当且仅当 $(q_1, q_2) \in D$ 时，$(p_1, p_2) \in D$。□

现在，根据以下引理，$\sim_c$ 是同余关系。

引理 6.6. 对于满足（NR_{pg}）的给定偏群胚（P、D、$\circ$），$\sim_c$是同余关系：$p \sim_c p'$和 $q \sim_c q'$ 推出 $pq \sim_c p'q'$。

证明。根据引理 6.3 和 6.4（vi），$pq \sim_c p'q \sim_c p'q'$。 □

给定一个偏群胚（P、D、$\circ$），根据引理 6.6，我们有一个商偏群胚（$P/\sim_c D/\sim_c$、$\star$），记作 $\mathcal{Q}$（P、D、$\circ$），其中$\overline{x}$是 x的等价类，相对于 $\sim_c$，$\overline{x} \star \overline{y} = \overline{x \circ y}$、$D/\sim_c = \{(\overline{x}、\overline{y}) \mid (x、y) \in D\}$；并且有一个偏群胚同态（$P$、$D$、$\circ$）$\to \mathcal{Q}$（$P$、$D$、$\circ$），将 x 发送到 $\overline{x}$。根据引理 6.4、6.5 和 6.6，我们得到以下命题：

命题 6.7。设 $(P, D, \circ)$ 是一个满足 (S) 和 (NR_{pg}) 的偏群胚。那么存在一个交换偏群胚 $(Q, E, \cdot)$ 以及一个满射偏群胚同态

$$f : (P, D, \circ) \to (Q, E, \cdot).$$

将命题 6.7 应用于一个有限生成偏群胚，我们有：

推论 6.8。设 $P = [p_1, \cdots, p_n]$ 是一个满足 (S) 和 (NR_{pg}) 的有限生成偏群胚。那么存在一个交换有限生成偏群胚 Q 以及一个满射偏群胚同态 ϕ: $P \to Q$。

证明。根据命题 6.7，存在一个交换偏群胚 Q 以及一个满射偏群胚同态 ϕ：$P \to Q$。由于 ϕ 是满射的，$Q = [\phi(p_1) \cdots, \phi(p_n)]$ 是有限生成的。□

E如下命题表明，每个包含 $\sim_c$ 的等价类都是半群 .

命题 6.9. 设 $(P, D, \circ)$ 是满足 (NR_{pg}) 和 $\overline{p} \in P/\sim_c$ 的偏群胚。则

(i) 对所有 x_1、$\cdots$、$x_k \in \overline{p}$，$x_1 \cdots x_k = x_1 x_k$ 和 $k \geq 2$ 都是整数；

(ii) 关于 $\circ$，$\overline{p}$ 是一个半群；

(iii) P 是某些半群的互不相交并集：$P = \cup_{x \in P} \overline{x}$，其中 $\overline{x}$ 是关于 $\sim_c$ 的 x 的等价类。

证明。（i）当 $k = 2$ 时显然成立。若 $k = 3$，则

$$\begin{aligned} x_1 x_2 x_3 &= x_1 x_3 x_1 \, x_2 x_3 \\ &= x_1 x_3 \quad (x_3 x_1 x_2 x_3 = x_3 \text{ by Lemma } 6.4(ii)). \end{aligned}$$

假设对于满足 $2 \leq i \leq k-1$ 的任意整数 i，都有 $x_1 \cdots x_i = x_1 x_i$。则

$$\begin{aligned} x_1 \cdots x_k &= x_1 \cdots x_{k-1} x_k \\ &= x_1 x_{k-1} x_k \quad \text{(By induction hypothesis)} \\ &= x_1 x_k, \end{aligned}$$

如所愿。

（ii）对于所有 x、$y \in \overline{p}$、$p \sim_c x$ 和 $p \sim_c y$。根据引理 6.4（ii），$pxyp = p$。根据（i），$xypxy = xy$。于是 $xy \sim_c p$，因此 $xy \in \overline{p}$。

根据（i），对于所有 x、y、$z \in \overline{p}$，$xyz = xz$ 都是唯一确定的值。因此，P 在 $\overline{p}$ 中的二元运算满足结合律，从而 $\overline{p}$ 关于 P 的二元运算构成一个半群。

（iii）由于 $\sim_c$ 是 P 上的等价关系，$P = \cup_{x \in P} \overline{x}$ 为互不相交的并集。根据（ii），每个 $\overline{x}$ 都是一个半群。

□

商运算 $\mathcal{Q}$ 是幂等的。

命题 6.10。对于满足（NR_{pg}）的所有偏群胚（P、D、$\circ$），$\mathcal{Q}^2$（P、D、$\circ$）$= \mathcal{Q}$（P、D、$\circ$）。

证明。如果（P、D、$\circ$）满足（NR_{pg}），则显然 $\mathcal{Q}$（P）也满足（NR_{pg}）。由于 $\mathcal{Q}^2$（P、D、$\circ$）$= \mathcal{Q}$（P、D、$\circ$），因为

$$\dfrac{\dfrac{\dfrac{\overline{x} \sim_c \overline{y} \ \ in \ \ \mathcal{Q}(P)}{\overline{x}\,\overline{y}\,\overline{x} = \overline{x} \ \ and \ \ \overline{y}\,\overline{x}\,\overline{y} = \overline{y} \ \ in \ \ \mathcal{Q}(P)} \Leftrightarrow}{xyx \sim_c x \ \ and \ \ yxy \sim_c y \ \ in \ \ P} \Leftrightarrow}{\dfrac{xyx\ x\ xyx = xyx, x\ xyx\ x = x, yxy\ y\ yxy = yxy, y\ yxy\ y = y \ \ in \ \ P}{xyx = xyx, xyx = x, yxy = yxy, yxy = y \ \ in \ \ P} \Leftrightarrow} \Leftrightarrow$$

$$\dfrac{}{x \sim_c y \ \ in \ \ P} \Leftrightarrow$$

□

7 有限生成偏群胚上的实体解析

回顾一下，每个实体解析系统（$\mathbb{E}$、$\approx$、$\langle\,\rangle$）都会产生一个偏群胚（$\mathbb{E}$、D、$\langle\,\rangle$），其中 D 是 $\langle\rangle$ 的定义域，由 $\approx$ 指定。给定一个实例 $I \subseteq \mathbb{E}$，在 [2], 中，Benjelloun 等人将实体解析 ER（I）定义为满足以下条件的最小子集 $I' \subseteq \overline{I}$：$\overline{I}$ 被 I' 支配。ER（I）存在且唯一（命题 2.1 [2]）；当（I）、（C）、（A）、（R）成立时，可以高效地计算它（命题 3.2、3.3 和 4.1[2]）。显然，$\overline{I}$ 是 [I], 在（$\mathbb{E}$、D、$\langle\,\rangle$）中由 I 有限生成。

在第 3 节中，我们讨论了性质（I_{pg}）、（C_{pg}）、（SC_{pg}）、（R_{pg}）、（A_{pg}）、（CA_{pg}）和（SA_{pg}），这些性质在偏群胚同态下保持不变。在第 4 节中，我们研究了给定偏群胚上的自然偏序 $\leq$。本节将说明，有限生成偏群胚为实体解析提供了另一种通用设置。

有些二元运算不一定满足结合律或交换律，例如 Prolog 中的中缀运算“ : –”、除法和平均值不满足结合律，而左连接、函数复合和

除法不满足交换律。我们将从一个任意的有限生成偏群胚（P、 D、 $\circ$）开始，不作任何性质上的假设。然后，我们将研究具有某些性质的情形，这些性质意味着自然偏序的存在以及该结构的有限性。在本节中，$(P, D, \circ)$ 是由 p_1、 $\cdots$、 p_n 有限生成的偏群胚，其中

$$P = [p_1, \cdots, p_n] = \bigcup_{i=1}^{+\infty} \{p_1, \cdots, p_n\}^i$$

不失一般性，根据定理 2.3，我们假设生成集 $\{p_1, \cdots, p_n\}$ 是不可约的，并且 P 是连通的。每个 $p \in P$ 都可以写成

$$p = x_1 \cdots x_i \cdots x_l$$

其中“(”和“)”以二叉方式插入，且 $x_i \in \{p_1, \cdots, p_l\}$、 $i = 1$、 $\cdots l$。

定义 7.1。在偏群胚 P 中，如果对于所有 $x \in P$， xp 的存在蕴含 $xp = p$，则称元素 p 为左满（右满）。 (如果 px 的存在蕴含 $px = p$.) p，则称其为满的；也就是说，它既是左满的，也是右满的。

Let

$$F_l(P, D, \circ) = \{p \in P \,|\, p \text{ is left full}\},$$
$$F_r(P, D, \circ) = \{p \in P \,|\, p \text{ is right full}\},$$

and

$$F(P, D, \circ) = \{p \in P \,|\, p \text{ is full}\}.$$

显然,

$$F(P, D, \circ) = F_l(P, D, \circ) \cap F_r(P, D, \circ).$$

对于有限生成偏群胚（P、 D、 $\circ$），在不要求任何其他条件的情况下，我们定义

(ER_f) $$ER_f(P, D, \circ) = \{\text{full elements in } P\}.$$

对于满足（I_{pg}）和（CA_{pg}）的有限生成偏群胚（P、 D、 $\circ$），根据引理 4.2， P具有偏序关系。我们定义

(ER_m) $$ER_m(P, D, \circ) = \{\text{maximal elements in } P\}.$$

根据定理 4.8，对于给定的偏群胚（P、 D、 $\circ$）以及满足 $\overline{I}$ 为有限集且满足（I_{pg}）和（CA_{pg}）的实例 I,

$$ER(I) = M(\overline{I}) = ER_m(\overline{I}),$$

其中， ER（I）定义于 [2] 中（定义 2.3）。

满元 与下述命题中所示的极大元相同。

命题 7.2. 对于满足 (I_{pg}) 和 (CA_{pg}) 的偏群胚 $(P, D, \circ)$,

$$F_l(P, D, \circ) = M_l(P, D, \circ), F_r(P, D, \circ) = M_r(P, D, \circ), \text{ and } F(P, D, \circ) = M(P, D, \circ).$$

证明。对于每个 $p \in F_l(P,D,\circ)$，若 $p \leq_l x$，则 $xp = x$。由于 p 是左满的且 xp 存在，因此 $p = xp = x$。所以 $p = px$，从而 $x \leq_l p$。因此，$p \in M_l(P,D,\circ)$。所以 $F_l(P,D,\circ) \subseteq M_l(P,D,\circ)$。

反之，对于所有 $p \in M_l(P,D,\circ)$，假设 xp 存在。由于 $p \leq_l xp$ 根据引理 4.4 且 p 是左极大的，$xp \leq_l p$。因此 $p = p(xp) = (px)p = xp$，从而 p 是左满。因此，$M_l(P,D,\circ) \subseteq F_l(P,D,\circ)$。因此 $F_l(P,D,\circ) = M_l(P,D,\circ)$，如所要求。

同样地，我们有 $F_r(P,D,\circ) = M_r(P,D,$**{v8)**。

最后，根据命题 4.7，$M_l(P,D,\circ) = M_r(P,D,\circ) = M(P,D,\circ)$。因此

$$M(P,D,\circ) = M_l(P,D,\circ) \cap M_r(P,D,\circ) = F_l(P,D,\circ) \cap F_r(P,D,\circ) = F(P,D,\circ).$$

□

给定一个实体解析系统（$\mathbb{E}, \approx, \langle\rangle$）和一个有限实例 $I \subseteq \mathbb{E}$，使用 $\approx$ 和 $\langle\rangle$ [2]可以从 I 生成无限多个元素（属于合并闭包 $\overline{I}$）。为高效计算I 的实体解析，Benjelloun 等人 [2] 提出了四个重要性质 (I)、(C)、(A) 和 (R)，以保证 $\overline{I}$ 的有限性以及 $\overline{I}$中自然偏序的存在。

包含来自 I 的所有可能复合的合并闭包 $\overline{I}$ 是由 I 在偏群胚（$\mathbb{E}$、$\approx$、$\langle\rangle$）中生成的偏子群胚 $[I]$。因此 $\overline{I} = [I] = \bigcup_{i=1}^{+\infty} I^i$. 一般而言，集合链 I、I^2、$\cdots$、I^n、$\cdots$ 可能既不会终止，也不会发生周期性变化，因此 $[I]$ 是无限的，如以下示例所示。

Ex例 7.3. 1. 设 $U = \{a_1, \cdots, a_n, \cdots\}$ 为一个无限的对象集。定义 e

$$a_i a_i = a_i \text{ and } a_i a_{i+1} = a_{i+2},$$

对所有正整数 i，且不定义其他复合。则 U 是一个偏群胚。令 $I = \{a_1, a_2\}$。则

$$\begin{aligned} I^2 &= \{a_1, a_2, a_3\}, \\ I^3 &= \{a_1, a_2, a_3, a_4\}, \\ &\vdots \\ I^n &= \{a_1, \cdots, a_{n+1}\}, \\ &\vdots \end{aligned}$$

因此

$$\overline{I} = [I] = \bigcup_{i=1}^{+\infty} I^i = \left\{a_1, a_2, \cdots, a_n, \cdots\right\} = U,$$

这不是有限的。

2. 设 T 为关系数据库中的表集，& 为例 2.1.6 中定义的行绑定。则 $(T, \&)$ 是一个偏群胚。设 t 为 T 中的一个非空表，并且

$$t^n = \overbrace{t \,\&\, \cdots \,\&\, t}^{n\ times}.$$

因此

$$[t] = \{t, t^2, \cdots, t^n, \cdots\}$$

是由 t生成的无限子群胚。

为了有效计算 $\overline{I}$ （从而计算 $\overline{I}$中的全元素或极大元），链

$$I, I^2, \cdots, I^n, \cdots$$

必须终止，且$\overline{I}$（$= [I]$）必须是有限的。因此，对于给定实例 I，需要一些条件来保证$[I]$ 的有限性。在这些条件中，结合性是一项重要条件。

对于一个ER系统（$\mathbb{E}$, $\approx$, $\langle\rangle$）和一个有限实例 I，Benjelloun 等人 [2] 证明，如果$\overline{I}$ 满足（I）、（C）、（A）和（R），那么 $\overline{I}$是有限的。将其转换到偏群胚的框架下，可得：

命题7.4. 每个满足（I_{pg}）、（SC_{pg}）、（A_{pg}）和（R_{pg}）的有限生成偏群胚都是有限的。

半群有限性条件的研究旨在给出能够推出半群有限性的若干条件。在这些条件中，通常要求半群是有限生成的。半群的有限性已得到广泛研究，例如 [3, 9, 10]。

半群（群胚或群）中的元素 s 若 $[s]$，则称其为周期的；其中，由 s 生成的子半群（子群胚、子群）是有限的。若半群（群胚或群）的所有元素都是周期的，则称该半群（群胚或群）是周期的。William Burnside 在1902 [4], 提出的Burnside问题，询问有限生成的周期群是否为有限群。Golod [13] 为Burnside问题提供了一个反例。也就是说，存在无限的、有限生成的周期群。

随后，Burnside问题被推广到半群。对有限生成的周期半群是否为有限半群这一问题的研究，被称为半群的Burnside问题。Morse和Hedlund对半群的Burnside问题给出了否定答案 [18, 10]。

偏群胚的Burnside问题询问：有限生成的偏周期群胚$P = [p_1, \cdots, p_n]$ 何时是有限的。

如果 $n = 1$，那么 $P = \cup_{i=1}^{+\infty} p_1^i = \{p_1, \cdots, p_1^k, \cdots\}$ 的有限性等价于 p_1 的周期性。为简化起见，我们要求每个生成元 p_i 都是幂等元。

设 $([p_1, \cdots, p_n], D, \circ)$ 是满足 (S)、(CA_{pg}) 和 (NR_{pg}) 的有限生成偏群胚。根据命题 6.7，存在一个满足 (CA_{pg})、(SC_{pg}) 和 (NR_{pg}) 的有限生成偏群胚 $([q_1, \cdots, q_n], E, \cdot)$，以及一个满射偏群胚同态

$$f : ([p_1, \cdots, p_n], D, \circ) \to ([q_1, \cdots, q_n], E, \cdot).$$

由于 (CA_{pg})、(SC_{pg}) 和 (NR_{pg}) 蕴含 (I_{pg})、(SC_{pg})、(A_{pg}) 及 (R_{pg})，根据命题7.4，$([q_1, \cdots, q_n], E, \cdot)$ 是有限的。

对于每个 $p \in [p_1, \cdots, p_n]$，根据命题 6.9，在第 6 节中定义的等价类 $\overline{p}$ 关于 $\circ$和 $x_1 \cdots x_k = x_1 x_k$构成半群，其中对所有 x_1、$\cdots$、$x_k \in \overline{p}$均成立。由于对于所有 x、$y \in \overline{p}$、$xpy = xy$ 和 $pxyp = p$，对每个 $x \in \overline{p}$，x的长度，即满足以下条件的最小 k

$$x = x_1 \cdots x_k$$

其中“(”和“)”以二元方式插入，且 $x_i \in \{p_1, \cdots, p_n\}$ 必须有界。因此我们得到：

命题 7.5。每个满足（S）、（CA_{pg}）和（NR_{pg}）的有限生成偏群胚 $[p_1, \cdots, p_n]$ 都是有限的。

一般而言，用于实体解析的匹配与不匹配标签高度不平衡，因此有限生成偏群胚中的连通分支并不大。本文不关注有限生成偏群胚的有限性条件；这些条件本身具有独立的研究价值，且超出了本文的范围。我们将在另一篇论文中研究这一主题。

(I_{pg})、(SC_{pg})、(A_{pg}) 和 (R_{pg}) 是偏群胚有限性的充分条件，但并非必要条件，以下例子对此进行了说明。

例 7.6。令 $Q_2 = \{a, b, c\}$ 为三个不同元素的集合，并定义

$$a \star b = c, \quad b \star c = b, \quad c \star c = b,$$

且不定义其他复合。于是，($Q_2, \star$) 是一个不满足 (I_{pg})、(SC_{pg})、(A_{pg}) 和 (R_{pg}) 中任何一项的有限偏群胚。

证明。(I_{pg}) 由于 $c \star c = b$、$a \star a$ 和 $b \star b$ 未定义，性质 (I_{pg}) 不成立。

(SC_{pg}) 由于 $a \star b = c$ 和 $b \star a$ 未定义，($Q_2, \star$) 不满足性质 (SC_{pg})。

(A_{pg}) 注意

$$(a \star b) \star c = c \star c = b \text{ and } a \star (b \star c) = a \star b = c.$$

性质 (A_{pg}) 不成立。

(R_{pg}) 由于 $a \star b = c$, $a \sim b$。注意，由于 $c \star b$ 未定义，$a \star b = c \nsim b$。因此 ($Q_2, \star$) 不满足性质 (R_{pg})。

□

为了匹配来自多个信息孤岛的客户，我们可能希望识别合并闭包中具有最大信息量的客户记录。然而，给定一个包含某些操作（例如选择、排序、连接）的数据库，我们可能希望找出数据库中最小的表集合，这些表是该数据库的不可约生成元。

一旦有限生成偏群胚或半群是有限的并且具有传递序，就可以利用 [8] 中关于传递偏序的结果，对该偏群胚进行排序、选择和查询。因此，我们将实体解析视为

(ER_{pg})　　S在有限生成偏群胚上对元素进行排序、选择和查询　　id

该偏群胚是有限的，并且带有偏序。

在现实生活中，实体并非彼此孤立，而是通过它们之间的关系相互连接；例如，客户会根据其社交媒体关系进行分组，机器学习模型建立在与之相关的某些数据集之上，而机器学习模型则会输出数据。在合并匹配的实体时，最好保留实体对象之间检测到的关系。

实体解析系统可以建模为具有某些性质的有限生成偏群胚（$[p_1, \cdots, p_n], D, \circ$），这些性质可确保 $[p_1, \cdots, p_n]$ 的有限性以及自然

偏序的存在。假设存在一组初始关系，由有向图G_0表示，它是以 $[p_1, \cdots, p_n]$中的节点构成的完全图的子图。随后，可以将保留这些关系的实体解析建模为（$[p_1, \cdots, p_n]$，（G_i） $D, *$），其中 G_i 是该关系有向图从 G_0开始的一系列序列， $*$是扩展 $\circ$，其中包括合并 $[p_1, \cdots, p_n]$ 中的元素，并更新关系图 G_{i-1} 以得到 G_i，从而保留 G_{i-1} 中的既有关系。

8 结论

一个实体解析系统（$\mathbb{E}$， $\approx$, $\langle\rangle$）等同于一个偏群胚（P， D， $\circ$），其中二元运算 $\circ$ 由合并函数 $\langle\rangle$ 给出，而运算 $\circ$ 的部分性 D 由匹配函数 $\approx$给出。给定一个实例 I， I 的合并闭包结果为有限生成偏子群胚 $[I]$。

研究了偏群胚上的性质：(I_{pg})、(C_{pg})、(A_{pg})、(R_{pg})、(LU_{pg})和(CP_{pg})，以及匹配函数与合并函数的共同性质及其相互关系。当偏群胚满足(I_{pg})和(CA_{pg})时，引入了偏群胚上的自然偏序。给定偏群胚上的一个偏序，(LU_{pg})和(CP_{pg})等价于(I_{pg})、(C_{pg})、(A_{pg})和(R_{pg})；当偏运算的定义域是自反的时，该偏序必须是我们定义的自然偏序。

通过使用其定义域图的连通分量和团覆盖，降低了偏群胚的偏性；当非交换偏群胚具有类似偏半群的结构时，则将其同态地映射为交换偏群胚。

在无需任何条件的有限生成偏群胚（P、 D、 $\circ$）中，我们所关注的实体解析（ER_f）被定义为 P中的全元素。如果（P、 D、 $\circ$）满足（I_{pg}）和（CA_{pg}），那么实体解析（ER_m）就是 P中的极大元；这些极大元是全元素，并构成定义于 [2]中的实体解析 ER（P）。此外，在该情形下，我们将实体解析视为（ER_{pg}）：“对有限生成偏群胚中的元素进行排序、选择和查询。”

参考文献

[1] C. J. Ash 和 T. E. Hall，图上的逆半群，Semigroup Forum，11(1)(1975)，140-145。[2] O. Benjelloun、D. Garcia-Molina、D. Menestrina、Q. Su、S.E. Wang 和 J. Widom，Swoosh：一种实体解析的通用方法，VLDB J.，18(1)(2009)，255-276。[3] T.C. Brown 和 Earl Lazerson，关于有限生成的幂等半群，Semigroup Forum，78(2009)，183-186。[4] W. Burnside，关于不连续群理论中的一个未决问题，Quart. J. Math.，33(1902)，230-238。[5] Colin M. Campbell、Edmund F. Robertson、Nik Ruskuc 和 Richard M. Thomas，自动半群，理论计算机科学，250(1-2)(2001)，365-391。

[6] Victor Christen、Peter Christen 和 Erhard Rahm，面向实体解析的基于信息量的主动学习，载于 ECML PKDD，第 1168 卷（2019），125-141。[7] V. Christophides、V. Efthymiou、T. Palpanas、G. Papadakis 和 K. Stefanidis，面向大数据的端到端实体解析概述，ACM Comput. Surv.，53(6)(2020)，1-42。[8] C. Daskalakis、R.M. Karp、E. Mossel、S.J. Riesenfeld 和 E. Verbin，偏序集中的排序与选择，SIAM Journal on Computing，40(3)(2011)，392-401。[9] Aldo de Luca 和 Stefano Varricchio，半群的有限性与迭代条件，理论计算机科学，87(1991)，315-327。[10] Aldo de Luca 和 Stefano Varricchio，半群与形式语言中的有限性和正则性，《理论计算机科学专著》，Springer-Verlag，1999。[11] U. Faigle 和 Gy. Tur´an，有序集的排序与识别问题，SIAM J. Comput.，17(1)(1988)，100-113。[12] J. Fisher、P. Christen 和 Q. Wang，使用马尔可夫逻辑的基于主动学习的实体解析，载于 PAKDD，338-349，Springer，2016。[13] E.S. Golod，关于幂零代数和有限可逼近的 p 群，苏联科学院学报，数学辑，28(1964)，273-276。[14] Xiuzhan Guo、Arthur Berrill、Ajinkya Kulkarni、Kostya Belezko 和 Min Luo，代数化地合并本体，https://arxiv.org/abs/2208.08715。[15] M. Hazewinkel，Magma，载于 Hazewinkel, Michiel（编），《数学百科全书》，Springer Science+Business Media B.V. / Kluwer Academic Publisher，2001。[16] Pascal Hitzler、Markus Kr¨otzsch、Marc Ehrig 和 York Sure，本体合并是什么？——使用推出的范畴论视角，载于第一届语境与本体国际研讨会：理论、实践与应用（C&O）论文集，104-107。AAAI Press，2005年7月。[17] E.S. Ljapin 和 A.E. Evseev，《部分代数运算理论》，Springer Netherlands，1997。[18] M. Morse 和 G. Hedlund，无尽棋局、符号动力系统与半群中的一个问题，Duke Math. J.，11(1944)，1-7。[19] Zachary Mesyan 和 J. D. Mitchell，图逆半群的结构，Semigroup Forum，93(1)(2016)，111-130。[20] H. Mitsch，半群的一个自然偏序，Proc. AMS，97(3)(1986)，384-388。[21] N. J. Pullman，图的团覆盖——综述，载于《组合数学 X》，Springer，1983，72-85。

[22] Klaus-Dieter Schewe 和 Qing Wang，《基于知识的实体解析理论框架》，《理论计算机科学》，549(11)(2014)，101-126。[23]S. Schwartz，《图覆盖与划分概述》，《离散数学》，345(8):112884，2022。[24] J. Shao、Q. Wang 和 Y. Lin，《用于实体解析的 Skyblocking》，《信息系统》，85(2019)，30-43。[25] Q. Wu 和 J. Hao，《最大团问题算法综述》，《欧洲运筹学杂志》，242(3)(2015)，693-709。[26] Antoine Zimmermann、Markus Kr¨otzsch、J´erˆome Euzenat 和 Pascal Hitzler，《用范畴论形式化本体对齐及其操作》，《信息系统中的形式本体》，第四届国际会议论文集，FOIS 2006，277-288。