arXiv:2601.02824v1 [cs.DB] 6 Jan 2026

实体消解结果比较分析的案例计数指标

John R. Talburt[1],Muzakkiruddin Ahmed Mohammed[1], Mert Can Cakmak [1],
Onais Khan Mohammed[1],Mahboob Khan Mohammed[1], Khizer Syed[1],
Leon Claasssens[21] 实体解析与信息质量高级研究中心（ERIQ）阿肯色大学小石城分校[2] Pilog Group 通讯作者*：Mert Can Cakmak mccakmak@ualr.edu

摘要

本文介绍了一种新的流程和软件系统——案例计数指标系统（CCMS），用于在真实链接（标注）未知的情况下，系统地比较和分析针对同一数据集运行的两种不同实体解析聚类流程的结果。CCMS 会生成一组计数，基于四种可能的变换情形，描述第一个流程生成的簇如何被第二个流程转换。这些变换包括：第一个流程形成的簇保持未变化、合并为一个大型簇、被分区为多个较小的簇，或以其他方式与第二个流程形成的多个簇发生重叠。CCMS 会针对第一个流程形成的每个簇，对上述每种情况进行计数。此外，在分析模式下运行时，CCMS 程序可以通过显示所有变更的详细信息或仅显示特定类型变更的详细信息，帮助用户评估这些变化。本文详细介绍了 CCMS 流程和程序，并给出了 CCMS 在大学和行业研究中的应用示例。

关键词：实体　　解析、记录链接、评估指标、比较分析、聚类评估、数据质量

1 引言

实体解析（ER）对于有效的数据管理至关重要，尤其是在处理大型且复杂的数据集时。实体解析通常被定义为这样一个过程：判断两个信息系统引用（有时称为“提及项”或“观测项”）是否指向现实世界中的同一对象或不同对象（Binette 和 Steorts，2022）。如果两个引用指向同一对象，则称它们等价（Talburt，2011）。

实体解析通常以两种形式之一实施：聚类实体消解或二元实体消解（Binette 和 Steorts，2022）。最基础、历史最悠久的实体解析过程是二元实体消解（Fellegi 和 Sunter，1969），最初用于协调在不同时间对同一人口进行的人口普查。在二元实体消解中，

输入由两个不同的数据集组成，目标是识别并链接这两个数据集中的等价引用。通常假设每个数据集内部都不存在等价引用，因此，第一个数据集中的一条引用至多可以链接到第二个数据集中的一条引用，从而实现一对一的二元实体消解。如果放宽这一条件，则数据集之间的链接可以是一对多的。

然而，本文关注的是聚类实体消解。在聚类实体消解中，输入是一个数据集，目标是将该数据集划分为互不相交的子集（簇），其中每个簇代表数据集中所有引用同一特定实体的引用。Stanford InfoLab 对给定引用集的实体解析作出了正式的数学定义（Benjelloun 等人，2009年）。在聚类实体消解中，同一簇中的所有引用都彼此等价，而不同簇中的引用则不等价。从以引用为节点、以链接为边的图来看，一个簇就是图中的一个极大连通分量。

二元实体消解完全依赖于引用的直接链接，即匹配达到特定相似度阈值的引用；而聚类实体消解还包含两个额外的间接链接过程。第一个间接链接过程是传递闭包。如果接受输入数据集的唯一引用假设（Talburt 和 Zhou，2015年），那么引用等价性在数学意义上就是一种真正的等价关系，即一种幂等、对称且传递的二元集关系（Rotman，2010年）。这样一来，不匹配的引用也可以通过一系列直接匹配链接间接关联为等价引用。例如，如果认为引用 A 和引用 B 等价，是因为 A 与 B 匹配；并且认为引用 B 和引用 C 等价，是因为 B 与 C 匹配，那么根据传递闭包，即使 A 与 C 不匹配，也可得出 A 与 C 等价的结论。加入间接链接有时会使二元实体消解结果与聚类实体消解结果的评估产生混淆。在二元实体消解中，只有直接匹配才会被评估为真阳性或假正链接。然而，在聚类实体消解中，同一簇中的所有引用对（包括直接和间接链接的引用对）都被视为已链接的引用对，并需要接受评估，以判定其为真阳性或假正链接（Ye 和 Talburt，2018年）。

聚类实体消解中的第二种间接链接过程是基于模式进行的。一种常见的间接链接模式是家庭迁移模式，这种模式常见于人口统计数据中：同一家庭的成员被发现居住在两个或多个地址（Fu 等人，2014年）。例如，四方模式中，一位名叫 John Doe 的人被发现先后居住在 Oak Street 和 Elm Street，而另一位名叫 Mary Doe 的人也被发现居住在这两个相同地址，这就提供了证据（提高了这些人可能是同一人的概率）。虽然两条 John Doe 引用和两条 Mary Doe 引用可能尚不足以构成匹配对，但根据这种家庭迁移模式，它们仍可能存在间接关联。尤其是在存在其他证据时，例如年龄相似、姓名较为罕见、家庭成员超过两人，或同一家庭成员还居住过其他地址。

2 问题陈述

主数据管理（MDM）是一种广泛使用的数据整理流程，大多数大型组织都借助它实现大规模、准确的数据集成。聚类实体消解是所有主数据管理系统中的基础流程。从学术角度来看，评估聚类实体消解的典型方法是使用实体解析指标，例如 F 值、精确率和召回率，或这些度量指标的加权版本

（Christen 等人，2023）。然而，这些度量指标依赖于真值集；真值集能够对每一对输入引用记录作出明确判断，将其归类为真正例、假正例、真负例或假负例，从而准确衡量总体结果。有了真值集，就可以轻松比较两种不同的簇实体解析结果，并判断哪一种更好。然而，在大多数实际应用中，对于规模较大的引用记录样本，其真实链接关系并不明确。虽然可以使用分层抽样技术，估算大型数据集聚类的总体精确率和召回率（Pullen，2017）（Penning，2016），但这些方法既费力又耗时。

除了精确率、召回率和 F 值等度量指标外，还有其他方法可以比较簇实体解析结果。显示假正例和假负例结果的混淆矩阵，对于理解系统所犯错误的类型——假正链接和假负链接——至关重要。Herzog、Scheuren 和 Winkler（Herzog 等人，2007）建议采用该方法，以全面评估并直观比较实体解析系统的准确性；其本质上与精确率和召回率度量指标相同。不过，这种方法也需要有可用的链接真值集。

在不同决策标准下评估和比较实体解析系统的性能时，接收者操作特征（ROC）曲线和曲线下面积（AUC）是重要工具。这些工具尤其有助于调整实体解析系统的灵敏度，以满足特定的运营需求，例如在欺诈检测软件中减少假负例，或提高精准营销活动的准确性（Hand 和 Christen，2018）。同样，这种方法也需要链接真值集。

3 相关工作

聚类实体解析领域目前采用一些无需真值集的替代方法来评估实体消解结果。其中一种方法是基准测试。基准测试使用具有既定且可接受结果的标准数据集。基准数据集提供了受控环境，可用于评估各系统在各种信息问题方面的优缺点，例如不同的数据质量级别、链接逻辑的变更，或比较相互竞争的供应商工具。通过展示各种系统在可比条件下的运行情况，基准测试能够深入揭示其可扩展性和灵活性（Christen，2012）。

另一种方法是使用模拟或合成生成的引用数据。这通常用于被视为敏感信息的个人人口统计数据项，例如姓名和地址信息。例如，合成生成的入住信息 |（Talburt 等人，2009）以及面向人口普查数据的 PseudoPeople 项目（Haddock 等人，2024）。大语言模型在生成合成数据方面也展现出潜力（Meng 等人，2022）。虽然这种方法可以解决真值集问题，但其效用取决于模拟引用数据对现实世界引用数据的仿真程度。

案例研究有助于理解实体解析系统的实际适用性和性能。除了为理论模型提供实证支持外，现实案例研究还凸显了诸如与现有信息技术系统集成、管理实时数据流以及适应变化中的数据环境等实际问题。然而，这些研究大多来自政府或非营利组织（Winkler，2006）。出于竞争力和保密性方面的考虑，商业实体不愿公布此类结果。

在实践中，大多数组织，尤其是商业实体，依靠长期积累的经验和分析来确认其系统已达到基本稳定、准确度合理且结果为用户所接受的状态。由于这些是生产系统，对其进行的更改都十分谨慎且循序渐进。更改（干预）的影响会在基准数据集上进行评估，以观察其总体效果，即比较更改所导致的聚类差异，判断其整体效果是否优于或劣于未变化的基线系统中的簇。然而，这类簇变化评估大多是临时进行的，缺乏系统性。

为解决这一问题，TWI 度量（Talburt 等人，2008）被引入，用于衡量同一底层数据集的两种聚类结果 A 和 B 之间的差异程度。TWI 定义如下：

$$TWI = \frac{\sqrt{|A| \cdot |B|}}{|V|}$$

其中，$|A|$ 表示结果 A中的簇数，$|B|$ 表示结果 B中的簇数，$|V|$ 表示 A 和 B中的簇之间非空交集（重叠）的数量。TWI 的取值范围归一化为 $[0, 1]$ ，因此只有当 A 和 B 为相同结果时，才能达到 1。TWI 的值永远不会为 0；最坏的情况是 $|A| = 1$ 和 $|B| = N$ （其中 N 是底层数据集中的引用数）。在这种情况下，$|V| = N$，TWI 是 N的平方根的倒数；随着 N 增大，这个值会越来越小。尽管 TWI 无需依赖对真实链接的了解，就能提供聚类差异的相对度量，但它无法说明这些差异具体是什么。

4 方法

鉴于现有度量指标和方法固有的局限性，本文提出了一种新方法——簇计数度量系统（CCMS），专门用于在无需真值集的情况下，更细致地分析簇实体解析变化的动态过程。

4.1 簇计数度量系统（CCMS）

令 ER_1 表示针对给定实体引用集合 R 执行的第一个（基准）实体解析过程所生成的簇集合。令 ER_2 表示针对 R 执行的第二个实体解析过程所生成的簇集合。 ER_1 和 ER_2 是 CCMS 的输入，CCMS 的输出是一系列计数。

CCMS 生成的第一组计数提供了 R、 ER_1 和 ER_2 的整体概况，包括：

- $RC =$ R 中引用的数量
- $CC_1 =$ ER_1生成的簇数量
- $SC_1 =$ ER_1 生成的单例簇数量（单例簇是仅包含一个参考文献的簇）

- $CC_2 =$ ER_2生成的簇数量
- $SC_2 =$ ER_2生成的单元素簇数量

不过，CCMS 的主要输出是第二组四项计数，用于描述 ER_1 形成的每个簇如何被 ER_2变换。CCMS 识别出四种彼此不同且互斥的变换，这些变换会将 ER_1 中的一个簇转变为 ER_2中的一个或多个簇。对于给定的簇 A_{ER1}，这些计数的定义如下：

未改变计数（UC）：统计 ER_1 簇与 ER_2 簇完全相同的案例数。换言之，通过各自的链接决策， ER_1 和 ER_2 均形成了相同的引用簇。

$$A = B, \quad \text{where } B \in ER_2$$

合并计数（MC）：统计整个 ER_1 簇是真正包含于某个 ER_2簇中的案例数。 ER_1 簇成为更大 ER_2 簇的一部分（合并入其中）。在此情况下， ER_1 簇中的所有引用也都通过 ER_2 过程进行了链接，但 ER_2 过程还对这些引用进行了额外链接，而 ER_1 过程没有进行这些链接。

$$A \subset B \text{ and } A \neq B, \quad \text{where } B \in ER_2$$

分区计数（PC）：统计 ER_1 簇被分解为多个 ER_2簇的案例数。在此情况下， ER_2 过程将 ER_1 簇分区为更小的簇，但未添加任何新引用。与 ER_1簇存在非空交集的每个 ER_2 簇，都是 ER_1 簇的真子集。

$$A = \bigcup_{i=1}^{n} B_i, \quad \text{where } B_i \in ER_2 \text{ and } n > 1$$

重叠计数（OC）：指满足以下条件的情况数量： ER_1 簇与两个或更多 ER_2 簇存在非空交集，且至少有一个相交的 ER_2 不是 ER_1 簇的子集。换言之， ER_1 簇是真子集，属于所有与 ER_1 簇存在非空交集的 ER_2 簇的并集。与分区情况不同，这表示发生了重叠： ER_1 簇被拆分，但产生的 ER_2 簇包含的引用比原始簇更多，表明发生了更复杂的重组。

$$A \subset \bigcup_{i=1}^{n} B_i, \quad A \neq \bigcup_{i=1}^{n} B_i, \quad \text{where } B_i \in ER_2 \text{ and } n > 1$$

由于这些情况互斥且穷尽，因此这四种计数之和也等于 ER_1 簇的总数，即：

$$CC_1 = UC + MC + PC + OC$$

需要注意的是，CCMS 流程并不对称。如果将 ER_2 簇视为基准簇，并将 ER_1 簇视为转换后的簇（ER_2 与 ER_1），那么计数会有所不同。虽然很明显，未改变计数 UC 在两种情况下都会

相同，但对于 ER_1-到-ER_2 和 ER_2-到-ER_1，如果 $CC_1 \neq CC_2$，那么至少有一个其他计数在 ER_1-到-ER_2 和 ER_2-到-ER_1 之间会有所不同。

图1展示了 ER_1-到-ER_2 变换的图形表示，其中 X、Y、Z 和 W 表示 R 中的引用。

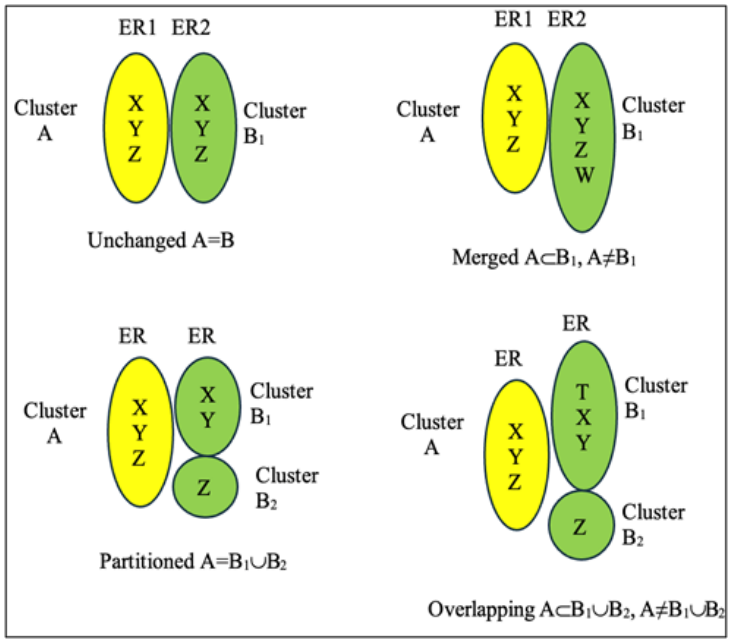


图1. 簇案例的图形表示

表1展示了一个简单示例：对于一个包含 16 条引用的数据集，ER_1 与 ER_2，其中 $RC = 16$、$CC_1 = 7$ 和 $CC_2 = 8$ 展示了全部四种情况。请注意，各行按ER_1 簇标识符的顺序排列。ER_1 簇以交替底色突出显示。

4.2 CCMS 的 Python实现

案例计数指标系统（CCMS）采用 Flask 实现为基于 Python 的 Web 应用程序，使用户能够分析、比较和可视化实体解析（ER）聚类过程的结果。在没有真值集的情况下，该实现提供了一个用于簇分析的交互式平台。下文将详细介绍该实现的设计与功能。

4.2.1 设计与架构

如图2所示，该应用框架使用 Flask 开发，用于管理文件上传、簇分析和结果呈现等服务器端操作。系统集成了使用 Bootstrap 构建的响应式 HTML界面，以提升易用性，并使用 DataTables 高效展示输入数据。系统接受两个 CSV输入文件，其中包含来自两个实体解析（ER）系统的 RecID、ER1 簇 ID 和 ER2 簇 ID。输出包括详细的簇转换分析，其中涵盖簇案例（未变化、合并、分区和重叠）的文本摘要、以条形图和饼图展示这些案例分布情况的可视化内容，以及对 ER_1 和 ER_2 中单例簇的识别。

簇比较 ER_1 到 ER_2						
参考标识符	ER_1 聚类 ID	ER_2 聚类 ID	未变化	合并	分区	重叠
1	a	x	1			
2	b	y	1			
3	b	y				
4	c	z				
5	c	z		1		
6	c	z				
7	d	z		1		
8	e	w				
9	e	w			1	
10	e	t				
11	f	u				
12	f	u				1
13	f	v				
14	g	u				
15	g	v				1
16	g	s				
总计	7	8	2	2	1	2
总计 ER_1 簇 (CC_1) $= 7$ {a, b, c, d, e, f, g}，总计 ER_2 簇 (CC_2) $= 8$ {x, y, z, s, t, u, v, w}						
ER_1 单例簇（SC_1）$= 2$ {a, d}，ER_2 单例簇（SC_2）$= 3$ {x, t, s}						
ER_1 与 ER_2 簇之间的非空交集 $= 11$						
$TWI = \frac{\sqrt{7 \times 8}}{11} = 0.68$						

表1. 聚类计数示例

4.2.2 关键功能组件

聚类分析与分类

来自以下簇　ER_1 和 ER_2 用于分析，以将变换归类为四种不同情形　　：

- 未变化：在　ER_1 和　ER_2中完全相同的簇。
- 合并：来自 ER_1 的簇，它们是　ER_2中较大簇的子集。
- 分区：来自　ER_1 的簇，在　ER_2中被拆分为多个更小的簇。
- 重叠：来自　ER_1 的簇，与　ER_2中的多个簇重叠，从而形成复杂的关系。

案例判定逻辑示例

以下 Python 代码展示了用于识别相同簇案例和分区簇案例的逻辑条件：

```
defis_相同(簇):returnlen(cluster['er2_clusters']) == 1和\cluster['

定义是_分区（集群）：返回长度（集群[的2_集群]）　> 1和\集群[大小_e
```

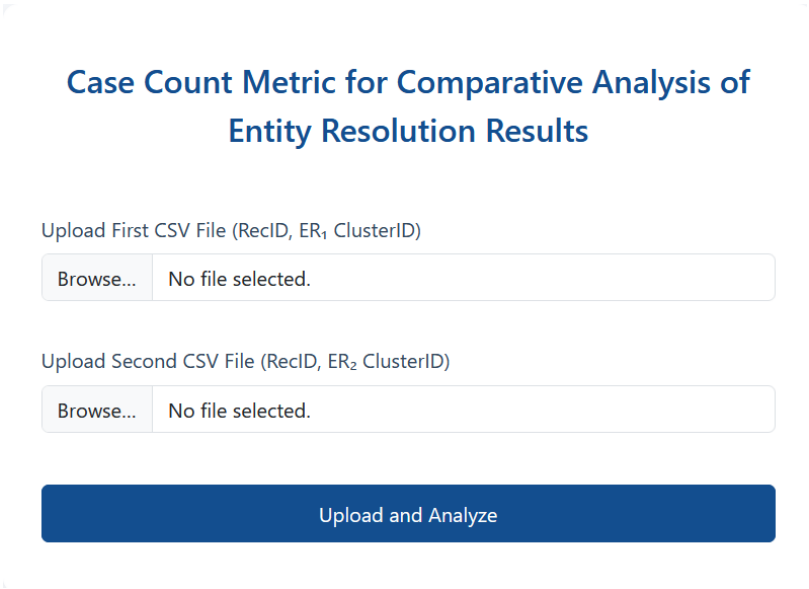


图2. 案例计数指标系统的用户输入界面。该图展示了应用程序的上传界面，用户可在此提交分别对应 ER_1 和 ER_2聚类输出的两个 CSV 文件。该界面便于提交数据，以进行实体解析比较分析，并支持可视化和指标计算。

可视化

ECharts 用于生成交互式柱状图和饼图，有效展示簇变换的结果。这些可视化清晰地呈现了以下方面的对比概览：

- 不同转换类型的案例计数分布。
- ER_1 和 ER_2之间的聚类总数占比。

交互式图表使研究人员能够探索聚类结果之间的关系、评估转换频率，并直观地识别实体解析流程中的离群行为。图 3 展示了一个示例。

单例检测

单例检测用于在 ER_1 和 ER_2 中识别孤立聚类，这些聚类通常代表唯一记录或未关联实体。以下函数展示了识别单例簇的一种实现方式：

清单1. 用于检测单元素簇的 Python 函数示例defdetermine_singleton

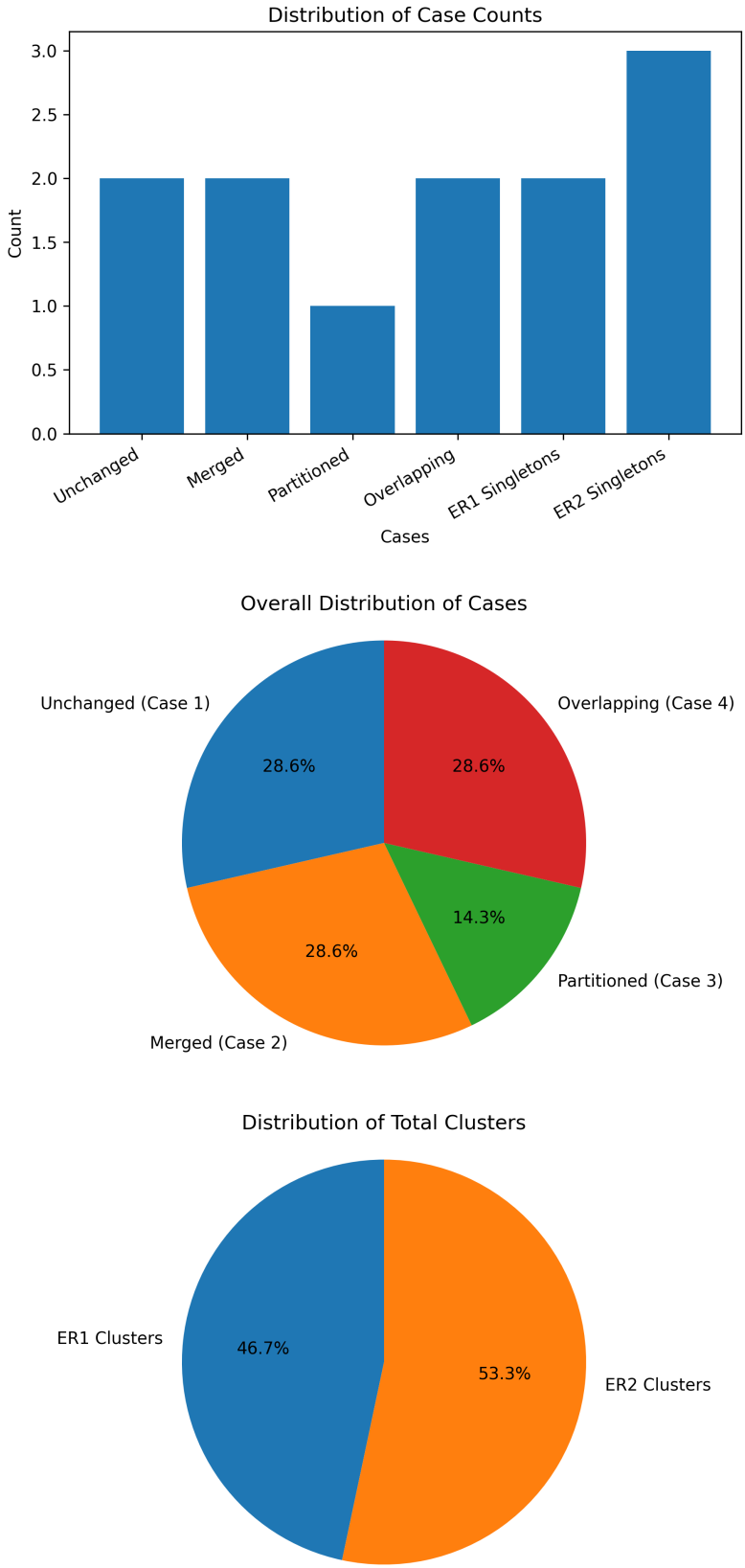


图 3. 案例计数和簇比例的比较可视化。顶部：案例计数的条形图。中部：案例比例的饼图。底部：ER_1 与 ER_2 簇总数。

随后对单元素簇进行摘要，以帮助理解实体解析结果的粒度和完整性。

摘要报告生成

摘要报告汇总所有计算指标，以文本形式概述 ER_1 和 ER_2之间的聚类结果。报告包括案例计数、单元素簇识别，以及两个比较方向的簇总数摘要。

输出示例（来自应用程序）：

详细摘要报告 ER1 为主、ER2 为辅：
未变化（案例 1）：2 合并（案例 2）：
2 分区（案例 3）：1 重叠（案例 4）：
2 ER1 簇：7

ER2 为主、ER1 为辅：未变化（案例
5）：2 合并（案例 6）：3 分区（案
例 7）：1 重叠（案例 8）：2 ER2 簇：
8

聚类总数：15

单例：ER1 单例：
2 ER2 单例：3

示例用例：

输入：两个示例 CSV 文件，包含以下数据字段：

- 记录 ID，ER1 簇 ID
- 记录 ID，ER2 簇 ID

输出：

- 可视化（柱状图和饼图），展示未变化、合并、分区和重叠簇案例的分布。

- 详细的文本报告，总结案例计数、单例簇以及整体簇变换。

基于 Python 实现的簇计数度量系统（CCMS）提供了一个实用且易于使用的平台，可全面分析和可视化两个实体解析流程之间的簇差异。该实现弥合了理论

实体解析指标与其在真实数据集中的应用之间的差距，即使没有真值集，也能提供易于理解的见解。

5 CCMS 的示例应用

这里的示例展示了 CCMS 的两种不同用途。第一个示例使用一个真值集已知的合成的人口统计数据集。在这一应用中，目标是了解聚类实体解析系统对某些连续值控制参数变化的敏感程度。在这种情况下，输入数据集和实体解析系统均保持不变，干预措施是改变参数值。

在第二个示例中，真值集未知，干预措施是更换系统。在第二个示例中，目标是通过分析簇的变化，确定哪个系统能给出更好的结果。

5.1 大学研究：实体解析参数敏感性分析

第一个示例涉及用于控制名为数据洗衣机（DWM）的无监督聚类实体解析系统的参数效果与敏感性（Talburt 等人，2020；Al Sarkhi 和 Talburt，2021；Anderson 等人，2023）。DWM 由 29 个参数控制。虽然其中许多参数涉及输入位置和输出格式等基本问题，但有 14 个参数用于调节 DWM 的无监督实体解析功能，包括阻塞、链接和簇评估。

无论是手动还是自动调整参数值，了解这些调整对 DWM 最终聚类结果的影响，以及系统对参数值变化的敏感程度，都很有帮助。这正是 CCMS 能发挥作用的地方。举个简单的例子，考虑 DWM 参数“mu”。该参数设定匹配级别阈值，且必须设为 $[0, 1]$ 区间内的某个值。

表 2 展示了 CCMS 的输出，其中输入数据集（S8P）包含由 SOG 生成的 1,000 条合成引用（Talburt 等人，2009）。在本例中，ER 参数设置为被认为是最佳的配置（包括 mu 值为 0.67），由此生成了 ER1 输出。在此参数配置下，ER1 生成了 255 个簇。在 ER1 参数保持不变的情况下，表 3 的每一行都表示一次 DWM聚类（ER2），其参数与 ER1 相同，唯一的变化是 mu 值。在本例中，mu 值的基线值为 0.67；干预操作是以 0.10 为增量，在基线值基础上递增或递减 mu。

Run	μ	CC1	$C1	UC	MC	PC	OC	CC2	$C2		
1	0.17	255	94	8	247	0	0	9	7
2	0.27	255	94	8	247	0	0	9	7
3	0.37	255	94	10	245	0	0	11	8
4	0.47	255	94	26	229	0	0	36	13
5	0.57	255	94	76	179	0	0	108	35
6	0.67	255	94	255	0	0	0	255	94
7	0.77	255	94	150	0	105	0	451	244
8	0.87	255	94	119	0	136	0	629	408
9	0.97	255	94	126	0	129	0	697	496

表 2. S8P 在基线值为 $= 0.67$ 时 Mu 参数变化的影响

表 2 第 6 行展示了基线输出，此时 ER1 和 ER2 相同，所有簇均未变化（UC $= 255$）。不出所料，降低 mu 会减少簇的数量，而提高 mu 会增加簇的数量。不过，表 2 确实揭示了几个有趣的发现。

首先，表 2 显示，对于 S8P，增大或减小 mu 的影响是可以预测的。如前所述，增大 mu 只会使 ER1 簇分区，而减小 mu 只会使 ER1 簇合并。mu 的变化都不会导致与 ER2 簇重叠。

表 2 的第二个观察结果是，对于 S8P，聚类对 mu 的变化有一定敏感性。仅将 mu 从 0.67 增大到 0.77，就会使簇的数量从 255 大幅增加到 451。mu 增加 15%，导致 41% 的 ER1 簇被分区，簇总数增加 176%。这表明，如果认为 mu $= 0.67$ 接近在其他参数设置不变时 S8P 获得最佳 DWM 聚类结果所对应的值，那么分析应着重考察 mu 在 0.67 附近以更小幅度（例如 0.01 而非 0.10）变化的影响。

再次需要注意的是，在没有真值集的情况下，CCMS 只会显示变化。变化究竟带来更好还是更差的结果，还需要进一步分析。不过，如第二个示例所示，CCMS 也能在这方面提供帮助。

表3展示了更改 Epsilon 参数值所产生的影响。Epsilon 是保留一个簇所需的最低质量度量。簇的质量通过修正的香农熵计算来衡量。未达到 Epsilon阈值的簇不会被保留，其引用会在 Mu 递增后返回输入，以便在下一次迭代中重新清洗。清洗周期会持续进行，直到所有簇的质量得分都高于 Epsilon，或 Mu 达到 1.00。对于 S8P 数据集，最佳 Epsilon 值相当小，将其降低 0.05 不会改变结果。另一方面，较大的 Epsilon 值会对簇的质量提出更高要求，并产生更小的簇，如拆分案例数量增加、单例簇增多所示。

Epsilon	ER1 聚类	ER1 单例	未变化	合并	分区	重叠	ER2 聚类	ER2 单例
0.10	255	94	255	0	0	0	255	94
0.15	255	94	255	0	0	0	255	94
0.20	255	94	253	0	2	0	257	98
0.25	255	94	250	0	5	0	261	105
0.30	255	94	244	0	11	0	272	121
0.35	255	94	230	0	25	0	305	158
0.40	255	94	219	0	36	0	330	181
0.45	255	94	193	0	62	0	414	250
0.50	255	94	164	0	91	0	506	231
0.55	255	94	141	0	114	0	570	366
0.60	255	94	126	0	129	0	614	397

表3. Epsilon 参数变化（最优 $= 0.15$）

表4展示了改变 Beta 参数值所产生的影响。Beta 控制 DWM 阻塞过程，表示可用作阻塞词元的词元的最低频率。通过汇集共享同一阻塞词元（或同一对阻塞词元）的引用来形成数据块。对于 S8P 数据集，Beta 值的敏感性较低。将其值提高至 31 不会影响最终生成的簇，但会减慢处理速度，因为

这会产生更多阻塞词元，进而需要处理更多数据块。将 Beta 降至 17 只会导致一个 ER1 簇被划分为两个 ER2 簇，其中一个是单元素簇。若采用更大的 Beta 值范围，可能会出现更多变化。

Beta	F 值	ER1 簇	ER1 单元素簇	未变化	合并	分区	重叠	ER2 簇	ER2 单元素簇
17	0.8079	255	94	254	0	1	0	256	95
19	0.8079	255	94	254	0	1	0	256	95
21	0.8079	255	94	254	0	1	0	256	95
23	0.8082	255	94	255	0	0	0	255	94
25	0.8082	255	94	255	0	0	0	255	94
27	0.8082	255	94	255	0	0	0	255	94
29	0.8082	255	94	255	0	0	0	255	94
31	0.8082	255	94	255	0	0	0	255	94

表4. Beta 参数变化（最优 $= 23$）

5.2 行业研究：材料分类

第二个示例比较了行业应用中的两个不同实体解析系统。第一个实体解析系统是前文所述的数据洗衣机（DWM）。第二个实体解析系统是 PiLog Group（Group，2024 年）研发团队对 DWM 的改编版本（记作 PDWM），旨在帮助对工程零件和材料进行分类。

尽管 PiLog 团队开展了广泛的研究和测试，但这里只介绍两个数据集的结果，以说明 CCMS 的实用性。这两个数据集分别为测试集 100 和测试集 5000，包含 100 条和 5,000 条记录，分别涉及零件和材料的引用，其中物品描述包括电机和滚珠轴承，特征描述包括电压、尺寸、颜色及其他技术细节。PiLog 研究的目标是确定：将其特有的材料描述知识（标注）添加到 PDWM 中，是否能够提升其性能，使其优于通用的开源版本 DWM，并有可能使 PDWM 成为适用于其内部处理的实用工具。

通过这些实验，我们旨在全面比较簇，并根据四种不同的案例计数进行区分，从而详细了解簇如何重新配置或保持稳定，为两个聚类结果之间的变化或一致性提供有价值的见解。

我们的方法包括整理数据集，并根据记录标识符（RecID）以及 ER1 和 ER2 生成的聚类 ID 进行细致排序。这种结构化排序为全面分析奠定了基础；在分析中，我们将簇分为不同的案例——未变化、合并、分区和重叠。

除了簇比较之外，我们还将编制一份详细报告，列出案例计数、ER1 总数、ER2 总数和单例数量。这份综合报告将帮助我们全面了解每种数据洗衣机的聚类结果和性能指标，并为类似情境下的数据管理策略提供有价值的指导。

通过仔细审视这些簇，我们深入了解了两台数据洗衣机如何重新配置簇或维持簇的稳定性，从而进一步洞察数据集状态在聚类结果之间的变化或一致性。

此外，我们还评估了辅助指标，例如 ER1 和 ER2 识别出的簇总数，以及两种输出中单例簇的数量。这些辅助指标进一步揭示了各系统的聚类行为和性能。通过实验与分析，我们旨在阐明每台数据洗衣机的能力与局限性，为类似情境下的数据管理策略提供有价值的指导。

我们在表5所示的两个数据集（测试集 100 和测试集 5000）上进行了实验，以评估大学数据清洗机（UALR DWM）与 Pilog 数据清洗机（Pilog DWM）的分类和非分类变体相比的性能。结果表明，这些系统在实体解析处理方面存在显著差异，具体取决于所用的数据集和变体。

指标	分类		非分类	
	测试集 100	测试集 5000	测试集 100	测试集 5000
ER1 簇	80	3918	80	3918
ER1 单例	68	3423	68	3423
未变化	65	2517	45	2464
合并	10	918	23	993
分区	5	236	6	334
重叠	0	247	6	127
ER2 簇	81	4205	80	4207
ER2 单例	67	3621	66	3625

表5. 分类与非分类条件下 UALR DWM 和 PiLog DWM 的案例计数结果

对于测试集 100，UALR DWM 共识别出 80 个簇，其中包括 68 个单例，在两种比较中均表现出一致的结果。然而，Pilog DWM 的分类与非分类变体表现出不同的簇转换行为。与分类 Pilog DWM 相比，65 个簇未发生变化，10 个簇合并为更大的簇，5 个簇被分区为更小的簇；而 ER2 识别出 81 个簇，其中包括 67 个单例。另一方面，非分类 Pilog DWM 表现出更大的差异，其中有 45 个未变化簇、23 个合并簇、6 个分区簇和 6 个重叠簇，最终 ER2 中有 80 个簇和 66 个单例。

测试集 5000 的结果进一步凸显了这些差异。UALR DWM 识别出 3,918 个簇，其中包括 3,423 个单例，在各项比较中结果一致。然而，分类 Pilog DWM 显示出显著变化：2,517 个簇未发生变化，918 个簇合并，236 个簇被分区，247 个簇发生重叠；ER2 识别出 4,205 个簇和 3,621 个单例。相比之下，非分类 Pilog DWM 产生了略有不同的结果，其中有 2,464 个未变化簇、993 个合并簇、334 个分区簇和 127 个重叠簇，最终 ER2 中有 4,207 个簇和 3,625 个单例。

这些实验凸显了数据清洗机的性能可能因数据集以及是否具备分类能力而出现显著差异。对于测试集 100 这类较小的数据集，UALR DWM 生成的簇计数较为稳定，但 Pilog DWM 的不同变体所采用的变换差异显著。对于测试集 5000 这类较大的数据集，

差异则更加明显，反映出分类对实体解析结果的影响。结果还揭示了实体解析任务的复杂性，尤其是涉及合并、分区和重叠簇等因素时。这些发现强调了针对不同数据集制定专门策略的必要性，并表明所提出的指标有助于识别实体解析系统中可优化和改进的方面。

6 结论

本文介绍的参数敏感性和材料分类示例表明，在没有真值集来自动计算和比较精确率、召回率、F 值及其他传统实体解析指标的情况下，案例计数指标系统（CCMS）可作为比较聚类实体消解结果的有用工具。与单值 TWI 度量相比，它也有了显著改进，能够更全面地揭示簇变更的性质，并展示和分析示例之间的差异。

CCMS 将簇变换分为互斥的四类：未变化、合并、分区和重叠，从而更深入、更具可操作性地理解簇如何响应实体解析系统内部或系统之间的变化而发生变换。这种细粒度分析有助于判断对聚类实体消解系统的特定更改总体上是有利还是不利，因此可用于在复杂数据环境中微调性能。

尽管 CCMS 在研究环境中已被证明稳健且富有洞见，但其在噪声更大、异质性更高和规模更大的数据集上的表现仍需进一步研究。还需要进一步研究和验证 CCMS 在医疗保健、金融和供应链管理等更多领域中的应用。

作者贡献

JRT负责研究构思、主要方法的制定以及论文撰写工作。MAM参与了数据分析和验证工作。MCC协助完成论文审阅、编辑及出版格式编排。MKM、OKM和KS参与团队讨论，提供关键反馈，并支持实验设计和结果解读的完善。LC担任PiLog数据代表，协调获取本研究所使用的数据集。所有作者均审阅并批准了最终稿件，并同意对本文所呈现的工作负责。

资金支持

本研究部分得到了美国国家科学基金会EPSCoR奖项编号OIA-1946391和PiLog Group的支持。

引用

Al Sarkhi, A. K. 和 Talburt, J. R.（2021）。用于估算未标准化参考文献实体解析中评分矩阵最优参数值的模型。载于《未来信息与通信会议》（Springer），16–33

Anderson, K. E.、Talburt, J. R.、Hagan, N. K.、Zimmerman, T. J. 和 Hagan, D.（2023）。数据洗衣机中无监督数据聚类与清洗的最优初始参数。载于《未来技术会议论文集》（Springer），106–125 Benjelloun, O.、Garcia-Molina, H.、Menestrina, D.、Su, Q.、Whang, S. E. 和 Widom, J.（2009）。Swoosh：一种通用的实体解析方法。《VLDB期刊》18，255–276 Binette, O. 和 Steorts, R. C.（2022）。关于实体解析的（几乎）全部内容。《科学进展》8，eabi8021 Christen, P.（2012）。数据匹配流程。载于《数据匹配：记录链接、实体解析与重复检测的概念和技术》（Springer）。23–35 Christen, P.、Hand, D. J. 和 Kirielle, N.（2023）。F值综述：历史、属性、批评与替代方案。《ACM计算评论》56，1–24 Fellegi, I. P. 和 Sunter, A. B.（1969）。记录链接理论。《美国统计学会会刊》64，1183–1210 Fu, Z.、Christen, P. 和 Zhou, J.（2014）。一种用于历史人口普查家庭记录链接的图匹配方法。载于亚太知识发现与数据挖掘会议（Springer），485–496[数据集] Group, P.（2024）。主页。https://www.piloggroup.com/。访问日期：2024-11-17 Haddock, B.、Pletcher, A.、Blair-Stahn, N.、Keyes, O.、Kappel, M.、Bachmeier, S. 等（2024）。不受隐私限制、用于人口普查规模实体解析研究的模拟数据：通过个体建模生成的大型数据集。《盖茨开放研究》8，36 Hand, D. 和 Christen, P.（2018）。关于使用 F 值评估记录链接算法的说明。《统计与计算》28，539–547 Herzog, T. N.、Scheuren, F. J. 和 Winkler, W. E.（2007）。《数据质量与记录链接技术》，第1卷（Springer）Meng, Y.、Huang, J.、Zhang, Y. 和 Han, J.（2022）。使用语言模型生成训练数据：迈向零样本语言理解。《神经信息处理系统进展》35，462–477 Penning, M.（2016）。实体解析的推断错误率。博士论文，小石城阿肯色大学 Pullen, D. L.（2017）。《用于分层抽样实体解析结果、以尽量减少人工审核工作量来评估并提升准确性的系统》。博士论文，小石城阿肯色大学 Rotman, J. J.（2010）。《高等现代代数》，第114卷（American Mathematical Soc.）Talburt, J.、Wang, R.、Hess, K. 和 Kuo, E.（2008）。一种针对大型数据集实体解析的数据质量度量指标的代数方法。载于《数据仓储与挖掘：概念、方法、工具与应用》（IGI Global Scientific Publishing）。3067–3084 Talburt, J. R.（2011）。《实体解析与信息质量》（爱思唯尔）Talburt, J. R.、Pullen, D.、Claassens, L.、Wang, R. 等（2020）。一种迭代式、自我评估的实体解析系统：迈向数据洗衣机的第一步。《国际高级计算机科学与应用期刊》11 Talburt, J. R. 和 Zhou, Y.（2015）。《大数据的实体信息生命周期：主数据管理与信息集成》（Morgan Kaufmann）Talburt, J. R.、Zhou, Y. 和 Shivaiah, S. Y.（2009）。SOG：用于支持实体解析教学与研究的合成占用生成器。ICIQ 9，91–105

Winkler, W. E.（2006）。数据质量：自动化编辑/插补和记录链接。《统计学》，7 Ye, Y. 和 Talburt, J.（2018）。传递闭包对实体解析逻辑回归校准的影响。《信息技术管理杂志》10, 1–11