研究生课程考试成绩单
（试卷封面）

<table>
<tr><td>院　　系</td><td colspan="2">计算机科学与工程学院</td><td>专业</td><td colspan="3">电子信息</td></tr>
<tr><td>学生姓名</td><td colspan="2">李志成</td><td>学号</td><td colspan="3">232349</td></tr>
<tr><td>课程名称</td><td colspan="6">云计算技术及应用</td></tr>
<tr><td>授课时间</td><td colspan="2">2024 年 2 月至 2024 年 5 月</td><td>周学时</td><td>3</td><td>学分</td><td>2</td></tr>
<tr><td>简
要
评
语</td><td colspan="6"></td></tr>
<tr><td>考核论题</td><td colspan="6">大数据环境下的实体解析研究综述</td></tr>
<tr><td>总评成绩
（含平时成绩）</td><td colspan="6"></td></tr>
<tr><td>备注</td><td colspan="6"></td></tr>
</table>

任课教师签名：_____

日期：

注：1. 以论文或大作业为考核方式的课程必须填此表，综合考试可不填。“简要评语”栏缺填无效。

2. 任课教师填写后与试卷一起送院系研究生秘书处。

3. 学位课总评成绩以百分制计分。

云计算技术及应用
课程作业

标 题	大数据环境下的实体解析研究综述
学 号	232349
姓 名	李志成
方 向	大数据分析处理

东南大学计算机科学与工程学院

二零二四年六月

目录

1 引言

1.1 背景描述

近年来，互联网信息技术和地理位置采集技术不断发展，推进了大规模地理空间数据的快速增长。地理空间数据为地图服务和基于位置的应用提供了重要的数据来源，比如OpenStreetMap、ForeSquare、Yelp 等[1]。这些空间数据大多数为兴趣点（Point-of-interest，POI），通常由地理坐标及其附加属性组成，表示现实世界的一个特定位置或地标，不同数据来源的 POI 存在数据冗余和数据不一致[2]。为了更好地了解城市运行机制和市场需求，实现城市功能分析和商业决策，需要一个海量且细致的空间数据集，由此产生了兴趣点集成技术。兴趣点集成旨在将来自不同数据源的兴趣点数据整合在一起，发现它们之间的对应关系，利用互补的属性更全面地了解某一地点的特征[3]。随着大数据和深度学习等技术的发展，空间数据集成相关研究引起了学界和工业界的广泛关注。

在数据集成相关技术中，实体解析（Entity Resolution，ER）是最关键的一项技术，旨在发现和关联来自不同数据源并指向同一物理世界实体的记录。实体解析的结果是匹配的实体对，然后通过数据融合[4]、数据清洗[5]等操作将多源数据集整合到同一个数据库中[6]。**面向多源空间数据集成的兴趣点实体解析的目标为：比对来自多个不同来源的空间数据集的兴趣点信息，发现并关联指向现实世界中同一兴趣点的记录，旨在建立一致的兴趣点数据库。**近年来，国内外企业和政府部门对高质量兴趣点数据的需求日益增长。例如，以 SafeGraph 和 QUADRANT 为代表的多家数据科技公司开始实施 POI 数据点的整合项目。2023 年《数字中国建设整体布局规划》指出：要夯实数字中国建设基础，畅通数据资源大循环，推动公共数据汇聚利用，建设公共卫生、科技、教育等重要领域国家数据资源库。因此，多源兴趣点实体解析的研究为加强和推动公共数据的汇聚和利用有重要意义。

1.2 研究现状分析

1.2.1 面向关系型数据的实体解析

在现实生活中，大量的数据被存储为关系型数据。然而，这些数据通常分散在彼此孤立的数据库中，从而导致数据孤岛，阻碍数据的关联与共享[7]。关系型数据实体解析旨在识别来自两个不同来源的元组是否指向真实世界中的同一对象(或称两者为正确匹配项)，以打破数据孤岛，实现跨源数据之间的关联互通，从而为数据集成奠定基础。实体解析旨在识别多个异构数据源中引用相同的现实世界实体的数据，通常和重复数据删除，实体匹配等研究工作相关。并已应用于社交网络用户匹配[8][9]、文章刊物作者匹配[10]、家庭族谱关系集成[11]等多个领域。现有的关系型数据实体解析工作主要包含基于规则的方法[12][14]、基于众包技术的方法[15][16]、基于机器学习的方法[17][18]和基于深度学习的方法[19][20][21]。

基于规则的方法主要依靠人工提供的声明性匹配规则来查找匹配的元组对。Fan[12]等人提出了一种形式化的确定性（精确匹配）或非确定性（基于相似度阈值）匹配规则，用于捕获记录之间的相似性，并提出了一种基于图的方法，其中每个记录表示为节点，而匹配规则表示为边，通过对图进行分析，可以找到表示相同实体的记录。但这种方法需要手动编写一组匹配规则，对于复杂异构的多源数据，耗时长且容易出错，且应用到异构数据上存在局限性。

基于众包的方法需要众包工人人工识别两个元组是否指向现实世界中的同一对象。Vesdapunt 等人[15]提出通过分配任务给人工来解决实体解析问题。该方法能够在数据量很大的情况下处理实体解析问题，并且可以自动处理一些错误和噪声。然而，该方法需要大量的

众包工人来完成任务，因此在众包工人数量不足或质量不高的情况下，可能会影响实体解析的准确性和效率。此外，该方法依赖于众包工人提供的结果，因此无法像基于规则的方法那样提供解释性和可追溯性。

基于机器学习模型的方法利用高斯混合模型、支持向量机(简称 SVM)、朴素贝叶斯等技术进行实体解析。Wu 等人[17]提出了一种使用自监督学习方法训练实体解析模型，避免手动标记数据的成本和时间。该方法首先使用启发式规则生成候选实体，然后使用自监督学习方法对候选实体进行训练以提高解析准确性。然而，该方法需要大量的计算资源和时间来训练模型，并且准确性可能受到启发式规则和自监督学习方法的影响。近年来，随着深度学习技术在机器学习领域的快速发展，许多基于深度学习的实体解析方法已被提出，并在取得了显著的性能提升。Li 等人[18]提出了一种基于预测模型和密度聚类的实体解析方法，该方法使用 BERT 模型来提取实体对之间的特征，并使用二分类模型来预测实体对是否匹配，然后采用基于密度聚类的方法将相似的实体对聚类到同一个簇中。该方法还使用了一种自适应阈值的方法来确定实体匹配的阈值，从而提高实体解析的准确性和泛化能力。虽然基于序列特征可以提高实体匹配性能，但是不同属性之间具有序列无关性，将元组视为序列形式可能丢失部分数据特征。

综上所述，本报告分析了现有不同的面向关系型数据的实体解析研究以及其不足之处。可以了解到虽然目前存在很多面向关系型数据实体解析的研究，但是这些研究在不同的层面上都存在一定的不足之处，而且这些研究并未考虑实体的空间信息，因而不能直接将面向关系型数据的方法应用于空间数据的实体解析。

1.2.2 面向空间数据的实体解析

为了获取丰富和准确的兴趣点数据，越来越多的研究机构开始了对空间数据集成的研究。空间数据集成技术包括实体解析、数据融合、数据清洗这几个步骤，其中最重要的步骤是实体解析。目前面向空间数据的实体解析技术研究主要分为以下两个方面：（1）针对空间对象的解析方法[22][24]；（2）针对空间实体的解析方法[2][13][28]。

在针对空间对象的解析的研究中，空间对象完全由坐标和空间形状组成，旨在为不同来源的数据建立统一的表示，并集成到一个数据库中。其中，Balley 等人[22]和 Schäfers 等人[23]在道路数据匹配的问题场景下，分别设计了道路网络的匹配和数据集成方案。前者主要处理栅格数据，不适用于空间实体集成。后者通过比较道路的长度、形状以及道路名称计算道路的相似度，并设计了基于贪婪策略的匹配模型：在每个阶段选择局部最优匹配，在下一轮迭代中如果附近的已确认的有效匹配数量越多，那么当前两个道路对象匹配的可能性越大。另外，Safra 等人[24]在面向多个数据源的位置点匹配的场景下，提出了完全基于位置匹配的两种多源策略：按顺序匹配和整体匹配方法。综上所述，尽管基于空间属性的匹配方法已经有很多研究，但是由于测量和众包录入的误差可能导致经纬度错误，只使用基于坐标和空间形状的方法不能得到准确的结果，因此今年来结合空间数据的文本属性语义信息的匹配方法得到广泛关注。

在针对空间实体解析的研究中，空间实体不仅包含地理位置信息，还具有包含语义信息的文本属性（名称、地址、电话、类别等），现有的空间实体解析工作主要包含基于规则的方法、基于机器学习模型的方法、基于语义模型的方法。基于规则的实体解析方法是通过预先设定的规则和逻辑，对实体对的相似性进行评估，然后给出是否匹配的决策。Isaj 等人[25]提出了一种基于空间四叉树的 QuadFlex 分块算法，在使用文本语义特征比较实体对相似度的基础上，采用帕累托最优的思想划分出天际线，最后通过规则的方法标注出正负类。此外，他们还提出了一种用小样本训练特征偏好函数构建天际线的方法，以得到效率和精度都较高的结果[28]。然而，这种方法没有充分考虑兴趣点的语义信息，难以处理语义相似但表达不同的实体。基于机器学习的实体解析方法结合兴趣点的地理空间信息和文本信息，利用机器学

习算法对其进行分类和标注。shah 等人[2]提出的基于 GeoHash 编码的 Geoprune 分块算法和 SimMetrics 框架可提高匹配精度。该方法支持三种分类器，但数据质量要求高，同时难以处理语义相似但表达不同的实体。大多数方法依然依赖人工定义的规则以及固定的文本相似度度量阈值，无法匹配语义信息。近年来有些工作考虑使用先进的语义模型求解（比如 Transformer，Bert），更好地获取和分析文本属性语义信息。Balsebre 等人[26]开发了一套通用模型来实现地理空间实体解析，该模型同时考虑实例对的文本相似度、经纬度距离和邻域实体的信息来实现匹配预测。这篇文章使用注意力机制比较同名实体的邻域特征，来解决空旷地区的同名 POI 比密集地区的同名 POI 的匹配概率更高的情况。但是该模型单独匹配空间实体对，没有使用之前匹配的结果，且需要标记数据来训练模型。此外，实际上不同来源的 POI 数据集重叠率很低，邻域 POI 节点的数量过少而无法利用邻域特征。

综上所述，目前的空间数据实体解析工作考虑了从多个角度研究分块和匹配算法，具体可以细分为针对空间对象的实体解析和针对空间实体的解析研究，但是由于经纬度误差而无法保证仅使用空间属性匹配的准确性，而现有的空间实体解析方法中，大多数方法依赖于数据集的特定属性和匹配规则，且受人工设置的阈值影响较大。

1.2.3 研究现状总结

综上所述，国内外对面向关系型数据的实体解析和面向空间数据的实体解析均取得了一定的研究成果。但是，这些工作仍然存在一定的不足：

（1）对于面向关系型数据的实体解析的研究，相关研究已经取得很多进展，但是这些方法的研究集中于以人为中心的任务，比如面向出版物和作者的实体解析、社交网络中的实体解析等，且这些方法都没有考虑到空间数据的地理空间特征，因此这些方法并不适用于兴趣点数据。

（2）对于面向空间数据的实体解析的研究，在分块方法上依赖人工定义的经纬度距离阈值和文本相似度阈值，分块过程中的阈值影响匹配过程的计算量和算法的召回率，如何确定合适的阈值具有较大的挑战；在匹配方法上，这些研究大多数依赖人工定义的规则，较少的研究使用了语义相似度特征。另外，现有的研究对空间实体的邻域特征的刻画缺少可解释性，没有考虑到空间邻域周围已匹配数据对待匹配数据的影响。

2 文章选择与逻辑体系

2.1 文章列表

	年份	会议	文章名称
1	2023	SIGMOD	**FlexER**: Flexible Entity Resolution for Multiple Intents
2	2023	SIGKDD	**CampER**: An Effective Framework for Privacy-Aware Deep Entity Resolution
3	2023	SIGKDD	**Unicorn**: A Unified Multi-tasking Model for Supporting Matching Tasks in Data Integration
4	2023	VLDB	Ground Truth Inference for Weakly Supervised Entity Matching
5	2023	VLDB	**Sparkly**: A Simple yet Surprisingly Strong TF/IDF Blocker for Entity Matching
6	2023	VLDB	**PromptEM**: Prompt-tuning for Low-resource Generalized Entity Matching
7	2023	ICDE	**Amalur**: Data Integration Meets Machine Learning
8	2023	ICDE	**Sudowoodo**: Contrastive Self-supervised Learning for Multi-purpose Data Integration and Preparation

9	2024	ICDE	Cost-Effective In-Context Learning for Entity Resolution: A Design Space Exploration
10	2024	ICDE	**MultiEM**: Efficient and Effective Unsupervised Multi-Table Entity Matching

来源会议全称：

SIGMOD：ACM Conference on Management of Data

SIGKDD：ACM Knowledge Discovery and Data Mining

VLDB：International Conference on Very Large Data Bases

ICDE：IEEE International Conference on Data Engineering

2.2 文章选择理由

个人研究方向考量：我隶属于江苏省网络与信息安全重点实验室中的**云计算与大数据研究组（大数据处理方向）**。我的研究重点是**空间实体解析与知识图谱构建**，旨在探索**大数据背景下的数据集成与知识图谱构建技术**。

文章方向考量：在撰写这篇综述的过程中，我调研了 2023 至 2024 年间在 CCF A 类会议和期刊上发表的大数据分析处理方向的文章。我发现，虽然几乎没有文章专门解决空间实体解析的问题，但有不少研究关注于（关系型数据的）实体解析问题。尽管这些研究并没有涉及空间实体解析中特有的空间信息，它们在解决实体解析问题的核心方法上仍对空间实体解析具有重要的参考价值。因此，从不同角度出发，我精选了其中具有代表性的 10 篇文章进行综述。

2.3 文章逻辑体系

这 10 篇关于实体解析（Entity Resolution，ER）的文章构成了一个涵盖从理论探索到具体技术实现的广泛逻辑体系。每篇文章都以不同的方式扩展和深化了实体解析的技术和应用，展示了从基础理论到具体应用的多种发展方向。

1. **技术多样性和定制化**

FlexER 介绍了一个能够处理多种意图的实体解析方法，这种方法可以适应不同应用场景的需求，如推荐系统或电子商务平台，其中实体的意义可能因上下文而异。

Unicorn 则提供了一个统一框架，通过多任务学习支持多种数据匹配任务，强调了模型在多个数据集和任务间的可迁移性和灵活性。

2. **隐私与效率**

CampER 针对的是如何在保持深度学习模型效能的同时，增加数据的隐私保护，通过使用分布式和加密技术来处理敏感数据。

PromptEM 则探讨了在资源受限（低资源）情况下进行高效的实体匹配，通过提示调整技术减少对大量标注数据的依赖。

3. **数据集成与优化**

Amalur 展示了如何利用机器学习与数据集成技术的结合，使用数据集成过程中产生的元数据优化机器学习任务，强调了在分布式环境下数据隐私保护的重要性。

Sudowoodo 则通过对比自监督学习方法，提高了数据集成任务的适应性和性能，特别是在没有或仅有少量标记数据的环境中。

4. **实体解析的应用与扩展**

Sparkly 使用了 TF/IDF 方法进行数据分块，是一个针对实体匹配优化搜索效率的方法，展示了如何在大规模数据集上实现高效的数据预处理。

MultiEM 探讨了如何在多表环境中进行有效的实体匹配，通过自动化属性选择和基于

密度的剪枝策略提升了无监督匹配方法的效率和效果。

5. **经济性与可扩展性**

BATCHER 关注于降低使用大型语言模型进行实体解析的成本，通过批量处理和优化示例选择策略，减少与语言模型的交互次数，从而降低成本。

这些文章共同涵盖了实体解析技术的多个发展方向，包括技术多样化、隐私保护、效率优化、资源管理和成本控制，表明了实体解析领域内对灵活、高效、安全解决方案的持续追求。每篇文章都在其特定领域内推动了理论的进步和技术的实际应用，相互之间展示了从基础研究到具体应用的广泛联系。

3 文章综述

3.1 FlexER: Flexible Entity Resolution for Multiple Intents

3.1.1 文章背景问题

实体解析（Entity Resolution）是数据清洗和整合中的一个长期问题，其目标是识别代表现实世界中同一实体的不同数据记录。传统的实体解析方法将此问题视为单一任务，假设现实世界实体只有一种解释，并专注于寻找匹配的记录。然而，在现实世界的应用场景中，不同的下游应用可能对实体有不同的解释和需求。

3.1.2 文章的研究动机

现有方法主要关注于单一解释的实体解析，这限制了它们在需要多种解释以满足不同用户需求的场景中的应用。例如，在线购物平台需要根据用户的不同偏好提供个性化的购物体验。因此，需要一种灵活的实体解析方法来处理多意图（Multiple Intents）的实体解析问题。

3.1.3 文章的总体解决思路和大致实现方式

文章提出了一种名为 FlexER 的方法来解决多意图实体解析（MIER）问题。FlexER 将实体解析问题视为多标签分类问题，通过构建一个多重图（multiplex graph）来表示元组对之间的意图关系，然后使用图神经网络（GNN）来学习意图表示，并改进对多个解析问题的结果。具体来说，FlexER 利用了现有的通用实体解析解决方案，并采用以下步骤：

1. 构建基于意图的元组对表示；
2. 创建一个多重图，该图作为 GNN 的输入；
3. GNN 学习意图间的潜在关系，以提高实体解析的准确性。

3.1.4 对本篇文章的思考

文章提出的 FlexER 方法在处理多意图实体解析问题上具有创新性，因为它不仅考虑了实体解析的传统需求，还扩展到了能够处理多个不同意图的场景。这种方法特别适用于需要个性化服务的应用，如推荐系统和在线购物平台。此外，通过使用 GNN 来学习意图之间的关系，FlexER 能够提供更加丰富和灵活的实体解析结果。

然而，文章的方法可能面临一些挑战，例如在构建多重图时如何确定不同意图之间的连接，以及 GNN 在处理大规模数据集时的效率和可扩展性。此外，实际应用中可能需要进一步的调整和优化，以适应特定领域的需求和数据特性。总体来说，文章为实体解析领域提供了一个有前景的新方向，并为未来的研究和应用奠定了基础。

3.2 CampER: An Effective Framework for Privacy-Aware Deep Entity Resolution

3.2.1 文章背景问题

实体解析（Entity Resolution， ER）是数据准备中的一项基础问题，它旨在确定来自不同数据源的记录是否指向同一现实世界实体。尽管深度学习方法在ER领域取得了显著成效，但它们通常假设数据能够集中存储，这在实际应用中由于隐私问题很难实现。现有的隐私保护 ER 方法通常基于规则，忽略了复杂的匹配机制，导致效果不佳。

3.2.2 文章的研究动机

标准深度 ER 方法的有效性建立在数据集中存储的假设之上，但在隐私保护方面存在不足。为了解决隐私问题，文章提出了 CampER，一个有效的隐私感知深度实体解析框架。CampER 旨在在不集中存储数据的情况下，实现与现有最佳深度 ER 解决方案相当的有效性，并保护数据隐私。

3.2.3 文章的总体解决思路和大致实现方式

CampER 框架包含两个主要组件：协同匹配感知表示学习（Collaborative Match-Aware Representation Learning， CMRL ）和隐私感知相似性度量（Privacy-Aware Similarity Measurement， PASM）。

协同匹配感知表示学习（CMRL）：通过自我生成训练对策略，使每个组织能够独立构建训练对，并通过协同微调方法学习匹配感知和统一空间的个体元组嵌入，以进行准确的匹配决策。

隐私感知相似性度量（PASM）：首先引入一种密码学安全的方法来确定匹配项，然后提出一种保持顺序的扰动策略，以显著加快匹配计算速度，同时保证 ER 结果的一致性。

CampER 利用预训练语言模型（PLMs）进行微调，并通过差分隐私技术确保在协同微调过程中的隐私保护。此外，CampER 还采用了部分同态加密（PHE）和顺序保持扰动机制来保护隐私。

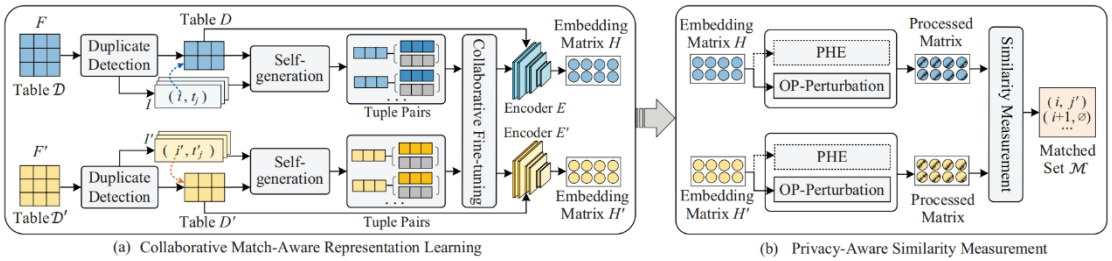

图 3.1 CampER 框架图

3.2.4 对本篇文章的思考

CampER 提出了一种创新的解决方案，通过结合深度学习和隐私保护技术，解决了在保护隐私的同时进行有效实体解析的难题。文章的方法不仅在理论上具有创新性，而且在实验中也显示出了优越的性能和隐私保护能力。

然而，文章的方法可能面临一些挑战，例如在实际应用中的计算效率和可扩展性，以及在不同数据分布和隐私要求下模型的鲁棒性。此外，虽然 CampER 在理论上提供了隐私保护，但在实际部署时可能需要进一步的调整和优化，以满足特定场景的需求。

总体来说，CampER 为隐私感知的深度实体解析提供了一个有前景的新方向，并为未来的研究和应用奠定了基础。随着隐私保护在数据处理中的重要性日益增加，CampER 或类似

框架的发展将对数据科学和人工智能领域产生重要影响。

3.3 Unicorn: A Unified Multi-tasking Model for Supporting Matching Tasks in Data Integration

3.3.1 文章背景问题

数据匹配是数据集成中的关键概念，涉及判断两个数据元素（如字符串、元组、列或知识图实体）是否相同。现有的实践通常构建特定于任务或数据集的解决方案，这些解决方案难以泛化，并且阻碍了跨不同数据集和任务的知识共享。

3.3.2 文章的研究动机

尽管深度学习在各种匹配任务中取得了巨大成功，但现有的深度学习方法通常是任务特定的，甚至特定于数据集，这限制了它们的泛化能力和跨任务的学习能力。此外，这些方法需要针对每个新任务或数据集进行训练或微调，效率低下且成本高昂。

3.3.3 文章的总体解决思路和大致实现方式

为了解决这些问题，本文提出了 Unicorn，一个统一的多任务模型，用于支持数据集成中的常见匹配任务。Unicorn 通过以下几个关键组件实现：

1. 编码器（Encoder）：将不同类型的数据元素对转换为学习到的表示形式。

2. 专家混合（Mixture-of-Experts，MoE）：通过不同专家的知识组合来增强表示，以适应多种任务的匹配语义。

3. 匹配器（Matcher）：一个二元分类器，用于基于增强的表示来预测数据元素对是否匹配。

Unicorn 模型通过多任务学习，从多个任务和数据集中学习，实现知识共享，并支持零样本（zero-shot）预测，即使在没有标记的匹配/不匹配对的情况下也能对新任务进行预测。

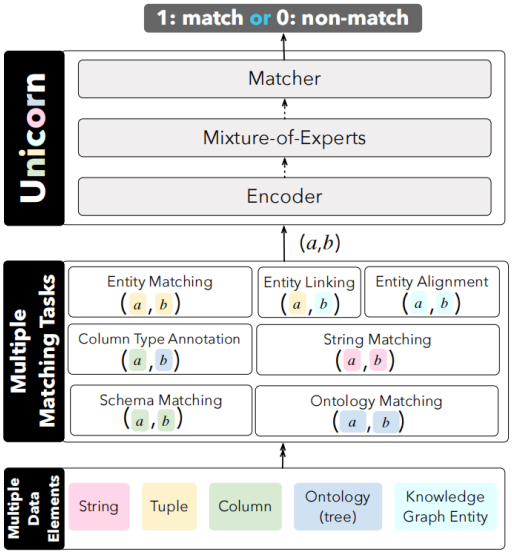

图 3.2 Unicorn 框架图

3.3.4 对本篇文章的思考

Unicorn 提出了一种创新的解决方案，通过统一模型处理多种数据匹配任务，这在数据集成领域是一个重要的进步。它不仅能够提高模型的泛化能力，降低维护复杂性，还能减少模型大小，提高效率。

此外，Unicorn 的零样本学习能力表明，它在面对新任务时具有很好的适应性，这为数

据匹配任务提供了一种更为灵活和经济的解决方案。然而，Unicorn 在实际应用中的表现和对不同类型数据的适应性仍需在更广泛的数据集和任务中进行验证。

文章提出的模型也引发了对多任务学习在数据集成中应用的进一步思考，特别是在处理不同数据类型和任务语义时如何平衡和优化模型性能。未来的工作可能会探索更多的指令技术、数据增强方法，以及将 Unicorn 扩展到更多类型的数据匹配任务和数据模态。

3.4 Ground Truth Inference for Weakly Supervised Entity Matching

3.4.1 文章背景与问题

实体匹配（Entity Matching，EM）是数据集成和知识图谱构建中的一个关键问题，它涉及识别不同数据源中指代同一实体的记录对。尽管监督机器学习模型在实体匹配上取得了最先进的性能，但它们需要大量标记样本，这些样本往往难以获得或成本高昂。因此，研究者们开始探索使用弱监督方法来降低对标记数据的依赖。

3.4.2 研究动机

现有的弱监督方法通常依赖于生成模型来从每个标记函数（Labeling Function，LF）的预测中推断出隐藏的真实标签。这些方法往往需要对生成过程进行简化的假设，以便于建模或求解，但这些假设在实际应用中可能不成立。此外，现有的方法在实现上相对复杂，且性能提升有限。

3.4.3 总体解决思路与实现方式

本文提出了一种新颖的标签模型，用于弱监督实体匹配任务。首先，作者提出了一个简单而强大的通用弱监督任务的标签模型，然后针对实体匹配任务的特定性质——传递性（transitivity），进一步定制了这个模型。具体来说，作者采用了以下技术：

1. 使用随机森林作为标签模型，通过显式限制其容量来避免学习到平凡的解决方案。

2. 引入了一种期望最大化（Expectation-Maximization，EM）算法的变体来学习模型参数，该算法在 M 步中使用交叉验证选择随机森林的超参数，并在 E 步中使用训练得到的模型进行预测。

3. 对于实体匹配任务，作者提出了一种考虑传递性的标签模型，通过在 EM 算法的 E 步中引入约束优化问题来确保预测结果满足传递性。

3.4.4 对文章的思考

文章的创新之处在于它提出了一种简单直观的方法来处理弱监督学习中的标签推断问题，特别是在实体匹配领域。与以往依赖于复杂模型和假设的方法不同，本文的方法基于通用的分类器，并通过交叉验证来显式地限制模型的复杂度，这在实现上更为简洁，且易于理解和应用。

此外，作者针对实体匹配的传递性约束提出了一种有效的解决方案，这在理论上和实践上都是有意义的，因为它确保了匹配结果的一致性。然而，文章的方法可能在处理非常大规模的数据集或在 LFs 非常多样化的情况下的性能和可扩展性还有待进一步验证。

最后，虽然文章的方法在多个数据集上取得了良好的性能，但在实际应用中，如何选择合适的 LFs、如何调整模型参数以及如何处理可能的数据不平衡问题仍然是需要进一步研究的问题。此外，对于传递性约束的处理，虽然本文提出了一种基于优化的方法，但在某些复杂的数据分布情况下，这种方法是否仍然有效也需要进一步探讨。

3.5 Sparkly: A Simple yet Surprisingly Strong TF/IDF Blocker for Entity Matching

3.5.1 文章背景问题

实体匹配（Entity Matching，EM）是数据集成中的一个关键任务，它涉及确定两个数据集中的数据实例是否指向同一现实世界实体。在实体匹配过程中，分块（blocking）步骤至关重要，其目的是快速筛选出不太可能匹配的元组对，以减少匹配步骤的计算量。尽管已有多种分块解决方案被开发出来，但使用众所周知的 TF/IDF 度量进行分块的方法却鲜有关注。

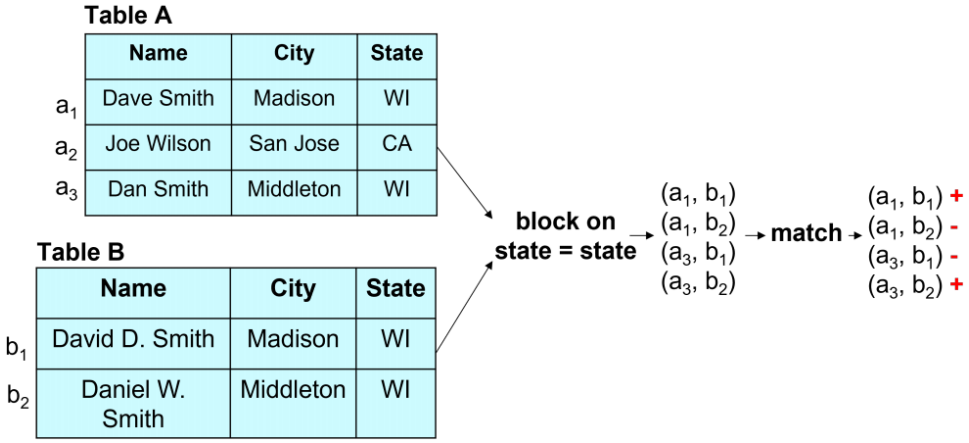

图 3.3 EM 的分块和匹配步骤

3.5.2 文章的研究动机

现有的实体匹配分块方法主要基于特定任务或数据集，难以泛化，且无法实现跨不同数据集和任务的知识共享。此外，现有的方法在处理大规模数据集时面临扩展性、可预测性和可扩展性方面的挑战。TF/IDF 作为一种在信息检索领域广泛使用的度量方法，尚未在实体匹配的分块步骤中得到充分利用。

3.5.3 文章的总体解决思路和大致实现方式

文章提出了一个名为 Sparkly 的新型 TF/IDF 分块器，它利用 Apache Lucene 库在 Spark 集群上以分布式无共享的方式执行 top-k TF/IDF 分块。Sparkly 的主要特点包括：

1. Top-k Blocking：与阈值分块相比，top-k 分块可以提高召回率，同时保持输出大小的可控性。

2. 分布式无共享架构：通过在 Spark 集群上并行执行，提高了可扩展性和效率。

3. 自动识别属性和分词器：Sparkly 自动选择最佳的属性和分词器进行分块，提高了模型的自动化程度。

3.5.4 对本篇文章的思考

Sparkly 的研究展示了 TF/IDF 在实体匹配分块步骤中的潜力，其简单而有效的方法在多个数据集上超越了现有的最先进方法。文章的创新之处在于将 TF/IDF 与分布式计算框架结合，提出了一种新的分块解决方案，这为未来在数据集成领域提供了新的思路。

然而，Sparkly 在处理特定类型的数据（如数值型数据）时可能存在局限性，这表明未来工作需要进一步探索如何改进 Sparkly 以处理不同类型的数据。此外，尽管 Sparkly 在大规模数据集上表现出良好的扩展性，但其在实际应用中的性能和成本效益仍需进一步评估。

文章还提出了一些未来的研究方向，包括深入研究 TF/IDF 分块、开发更大的基准数据集、探索数据清洗和标准化方法、改进现有分块解决方案以及开发新的分块技术。这些方向

为数据集成领域的研究者提供了有价值的指导。

总的来说，Sparkly 的研究为实体匹配的分块步骤提供了一种新的、有效的解决方案，其创新性和实验结果表明，TF/IDF 分块值得在未来的研究中得到更多的关注。

3.6 PromptEM: Prompt-tuning for Low-resource Generalized Entity Matching

3.6.1 文章背景问题

实体匹配（Entity Matching，EM）是数据管理中的一个基础性问题，其目标是识别来自两个关系表的实体记录是否指向同一个现实世界中的实体。传统的 EM 通常假设两个表具有相同的结构和对齐的模式，但在实际应用中，实体记录可能以不同格式存在（例如，关系型、半结构化或文本类型），这使得统一它们的模式变得不切实际。为了支持不同格式实体记录的匹配，提出了广义实体匹配（Generalized Entity Matching，GEM）。

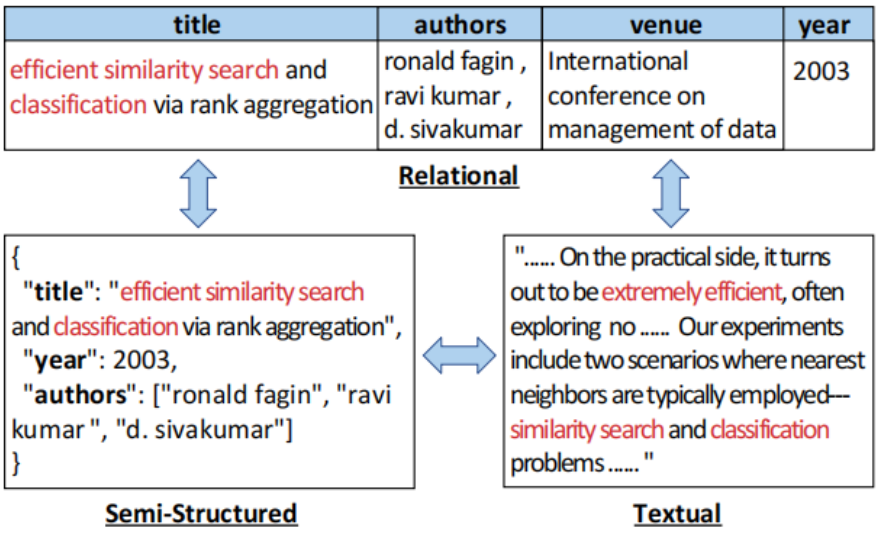

图 3.4 GSM 的例子

3.6.2 文章的研究动机

现有的 GEM 方法通常采用监督学习方法，这依赖于大量高质量的标记样本。然而，标记过程极其耗费人力，限制了 GEM 的应用。低资源 GEM，即只需要少量标记样本的 GEM，成为了迫切需求。此外，现有的基于预训练语言模型（PLMs）和微调范式的方法虽然取得了一定的性能，但仍然需要一定数量的高质量标记样本。

3.6.3 文章的总体解决思路和大致实现方式

本文提出了一种名为 PromptEM 的低资源 GEM 方法，首次关注低资源 GEM 问题。PromptEM 通过以下三个方面来解决低资源 GEM 的挑战：

1. GEM 特定的提示调整（Prompt-tuning）：通过设计 GEM 特定的提示模板，将 GEM 问题转化为填空式任务，与预训练中的目标形式一致，从而更好地利用 LMs 中的知识。

2. 提高伪标签质量：开发了一种轻量级的不确定性感知自训练方法，使用贝叶斯深度学习来估计教师模型的不确定性，以选择高质量的伪标签。

3. 高效的自训练：提出了一种动态数据修剪方法（MC-EL2N），在自训练过程中动态地剪枝无用的训练数据，使自训练过程更加轻量级和高效。

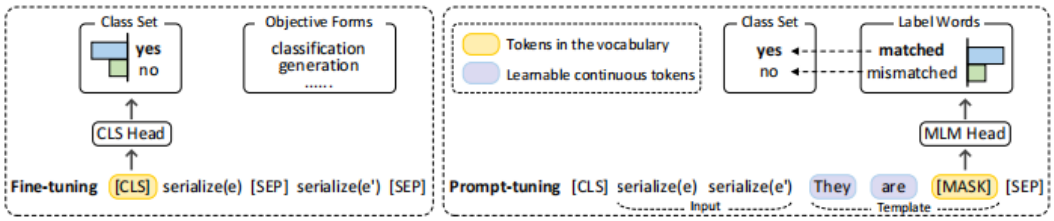

图 3.5 fine-tuning 和 prompt-tuning 的例子

3.6.4 对本篇文章的思考

PromptEM 的提出是对低资源情况下实体匹配问题的一种创新性解决方案。文章通过引入提示调整（prompt-tuning）这一新兴的自然语言处理范式，有效地弥合了预训练和微调之间的目标形式差异，这在利用预训练语言模型的知识方面具有显著的优势。此外，通过不确定性感知的自训练和动态数据修剪，PromptEM 进一步提高了在低资源环境下的性能，同时保持了训练过程的高效性。

文章的方法在多个真实世界的基准数据集上进行了广泛的实验验证，展示了其在有效性和效率方面的优势。这表明 PromptEM 不仅能够处理传统的实体匹配任务，还能够应对更广泛的实际应用场景，包括不同格式的实体记录匹配。

然而，尽管 PromptEM 在低资源 GEM 问题上取得了显著的成果，但在实际应用中可能还需要考虑其他因素，如模型的可解释性、对不同类型噪声的鲁棒性，以及在更大规模数据集上的性能等。未来的工作可以探索将 PromptEM 扩展到更多的数据管理任务，并进一步优化其在不同环境下的泛化能力和实用性。

3.7 Amalur: Data Integration Meets Machine Learning

3.7.1 文章背景问题

在机器学习（ML）应用中，训练数据往往分散在不同的数据孤岛（data silos）中。这些数据孤岛间的信息隔离给数据集成和模型训练带来了挑战。传统上，数据集成（DI）系统用于解决数据孤岛间的数据互操作性问题，但它们并不直接支持 ML 应用。此外，数据隐私的约束要求模型训练以去中心化的方式进行，这进一步增加了在数据孤岛上进行有效 ML 模型训练的难度。

3.7.2 文章的研究动机

现有的 ML 系统通常将数据集成视为一个预处理步骤，而专注于模型的学习部分。然而，这种方法没有充分利用数据集成过程中产生的元数据，这些元数据可以用于改善 ML 模型训练的有效性、效率和隐私保护。本文提出了一种新的方法，将数据集成技术与现代机器学习系统的需求相结合，以提高在数据孤岛上进行 ML 训练的效率和效果。

3.7.3 文章的总体解决思路和大致实现方式

文章提出了一个名为 Amalur 的新型 ML 系统，该系统利用数据集成过程中获得的元数据来提升 ML 模型训练和推理的性能。Amalur 系统的设计考虑了以下几个关键方面：

系统设计：提出了 Amalur 系统的设计，该系统通过元数据目录来存储和管理数据、ML 模型和硬件设置的元数据。

异构元数据：展示了不同类型的元数据及其在提升 ML 任务效率、效果和隐私保护方面的作用。

元数据表示：提出了基于矩阵的元数据表示方法，包括映射矩阵、指示矩阵和冗余矩阵，以捕获数据源之间的列匹配、行匹配和数据冗余。

计算效率：探讨了如何使用 DI 元数据来提高在数据孤岛上进行 ML 模型训练的时间效率。

隐私保护：讨论了使用 DI 元数据改进垂直联邦学习中的隐私保护问题。

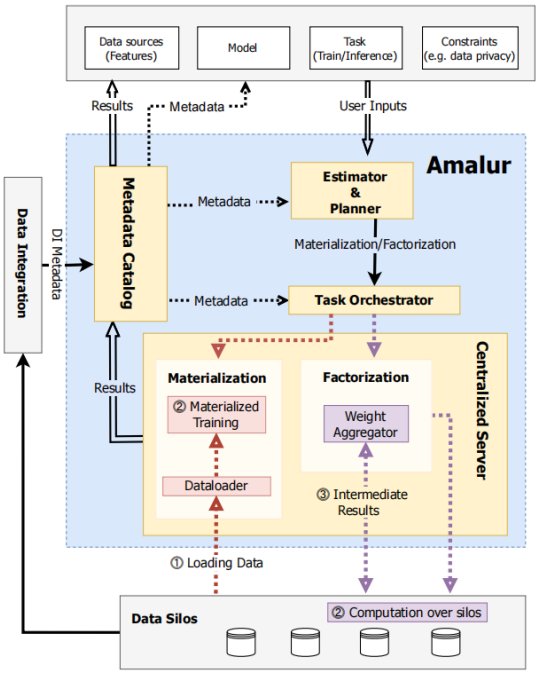

图 3.6 Amalur 框架图

3.7.4 对本篇文章的思考

文章提出的 Amalur 系统在数据集成与机器学习结合方面具有创新性，特别是在处理来自不同数据孤岛的数据时，能够有效地利用元数据来优化 ML 模型的训练过程。这种方法不仅提高了模型训练的效率，还增强了数据的隐私保护，这对于当前越来越重视数据隐私的环境下尤为重要。

然而，文章的方法可能面临一些挑战，例如在处理更复杂的数据集成场景时，如何有效地表示和利用 DI 元数据，以及如何在不同的硬件架构上实现高效的成本估算。此外，文章的方法在实际应用中的可扩展性和鲁棒性也需要进一步验证。

总体来说，Amalur 系统为数据集成和机器学习领域提供了一个新的视角，特别是在联邦学习等分布式学习场景下，展示了 DI 元数据的巨大潜力。未来的工作可能会集中在解决现有挑战，以及探索更多 DI 元数据在 ML 任务中的应用。

3.8 Sudowoodo: Contrastive Self-supervised Learning for Multi-purpose Data Integration and Preparation

3.8.1 文章背景问题

机器学习（ML）在数据管理任务中扮演着越来越重要的角色，尤其是在数据整合和准备（DI&P）方面。然而，ML 方法的成功在很大程度上依赖于大规模、高质量的标记数据集。获取这些数据集需要大量的人工努力和计算资源。此外，由于数据隐私的限制，数据往往不能离开数据孤岛（data silos）的场所，因此模型训练应在分散的方式下进行。

3.8.2 文章的研究动机

现有的 ML 方法在处理数据孤岛问题时面临两大挑战：标签需求和任务多样性。为了提

高 ML 模型的性能，通常需要大量高质量的标注数据对，这在新任务和新领域中是不切实际的。此外，DI&P 任务的广泛范围和多样性要求为每个任务定制专门的 ML 解决方案，这导致了模型工程的成本增加。为了克服这些挑战，本文提出了 Sudowoodo，一个基于对比表示学习的多用途 DI&P 框架。

3.8.3 文章的总体解决思路和大致实现方式

Sudowoodo 框架通过对比学习来学习数据表示，无需使用任何标签即可捕捉数据项之间的相似性。这种学习到的表示可以直接用于或通过少量标签进行微调，以支持不同的 DI&P 任务。Sudowoodo 采用统一的匹配问题定义，可以捕捉包括数据整合中的实体匹配（EM）、数据清洗中的错误更正、数据发现中的语义类型检测等多种 DI&P 任务。此外，Sudowoodo 还引入了几种优化方法，包括基于 cutoff 的数据增强技术、基于聚类的负样本采样，以及结合对比学习和 Barlow Twins 的冗余正则化。

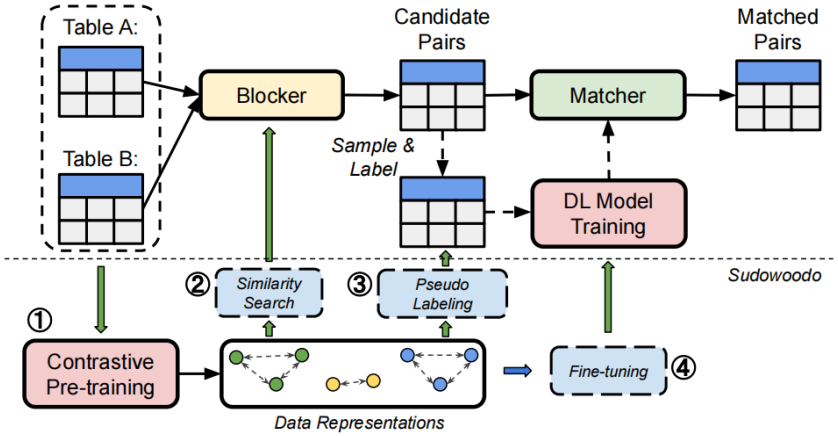

图 3.7 Sudowoodo 框架图

3.8.4 对本篇文章的思考

Sudowoodo 提出了一种创新的解决方案，通过对比学习来克服传统 ML 方法在数据孤岛问题上的局限。这种方法不仅减少了对大量标注数据的依赖，还通过统一的问题定义和优化技术，提高了在多种 DI&P 任务上的适用性和性能。文章的实验结果表明，Sudowoodo 在不同监督水平下都取得了新的最佳结果，显示出其在低资源环境中的优越性能和多样性。

然而，Sudowoodo 作为一种新方法，可能还需要在更广泛的数据集和更复杂的场景中进一步验证其有效性和鲁棒性。此外，对比学习在 DI&P 任务中的应用仍处于起步阶段，未来可能需要探索更多的数据增强和负样本采样技术，以及如何更好地整合到现有的数据管理和机器学习工作流程中。总的来说，Sudowoodo 为数据整合和准备领域提供了一个有前景的新方向。

3.9 Cost-Effective In-Context Learning for Entity Resolution: A Design Space Exploration

3.9.1 文章背景问题

实体解析（Entity Resolution， ER）是数据集成中的一项关键任务，它涉及识别并合并指代同一实体的不同记录。随着机器学习（ML）技术的发展，基于深度学习的 ER 方法已经取得了显著的进展。然而，现有的基于预训练语言模型（PLMs）的方法需要大量的标记数据进行微调，这不仅成本高昂，而且限制了这些方法在新任务和新领域中的应用。

3.9.2 文章的研究动机

尽管大型语言模型（LLMs）如 GPT-4 展示了无需调整模型参数即可执行多项任务的能力，即上下文学习（In-Context Learning，ICL），但现有的基于 ICL 的 ER 方法在调用 LLMs 时存在经济成本问题。这些方法需要为每对实体提供任务描述和示例，导致成本较高。为了解决这一问题，本文提出了一种成本效益高的批量提示（Batch Prompting）方法，用于 ER 任务。

3.9.3 文章的总体解决思路和大致实现方式

文章提出了一个名为 BATCHER 的框架，该框架包括示例选择（demonstration selection）和问题批量化（question batching）两个主要模块。BATCHER 框架的目标是在保持高匹配精度的同时，减少与 LLMs 交互的经济成本。具体技术包括：

示例选择：开发了一种基于覆盖的示例选择策略，通过选择最少数量的示例来覆盖所有问题，从而有效平衡匹配精度和经济成本。

问题批量化：通过不同的策略将问题分组，包括基于相似度、多样性和随机的批量化策略。

特征提取器：使用基于语义和结构感知的特征提取器，将问题转换为特征向量，以便在向量空间中测量它们的相关性。

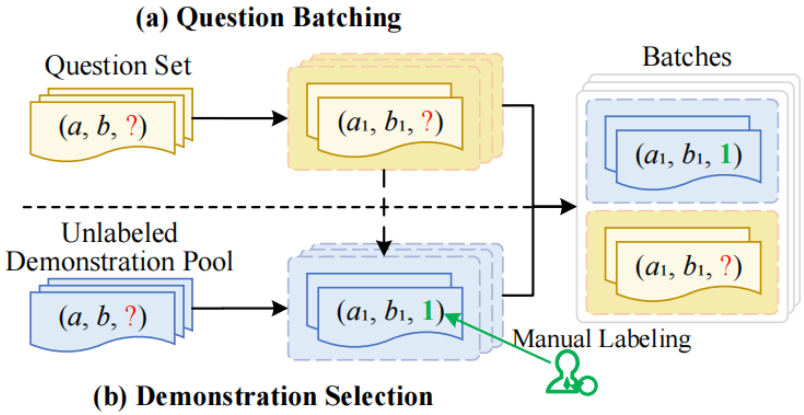

图 3.8 BATCHER 框架图

3.9.4 对本篇文章的思考

本文提出的 BATCHER 框架在 ER 任务中展示了成本效益高的优势，特别是在与需要大量标记数据的 PLM 方法和需要手动设计提示的 LLM 方法相比时。通过广泛的实验，作者展示了不同设计选择对精度和成本的影响，并提供了选择合适设计的指导。

文章的创新之处在于：

1. 提出了一种新的批量提示方法，通过在问题之间共享信息来提高 LLMs 在 ER 任务中的性能。

2. 开发了一种基于覆盖的示例选择策略，有效减少了标注成本，同时保持了高精度。

3. 对于不同的问题批量化和示例选择策略进行了系统性的设计空间探索，提供了实证见解。

然而，文章的方法可能依赖于特定的 LLM 架构和参数设置，其泛化能力和在不同数据分布上的表现需要进一步研究。此外，尽管文章提出了结构感知的特征提取器，但如何进一步优化特征提取以适应更广泛的 ER 任务仍然是一个开放的问题。最后，文章的方法在实际部署时可能需要考虑实时性能和可扩展性，以适应大规模数据集成需求。

3.10 MultiEM: Efficient and Effective Unsupervised Multi-Table Entity Matching

3.10.1 文章背景问题

实体匹配（Entity Matching，EM）是数据管理中的一个基础任务，目的是从多个关系表中识别出指向相同现实世界实体的记录对。传统的实体匹配方法通常假设所有实体来自两个表，但在实际应用中，经常需要从多个表中进行实体匹配，即多表实体匹配（multi-table EM）。现有的无监督多表实体匹配方法在效率和效果上表现不佳，且在处理大规模数据时面临挑战。

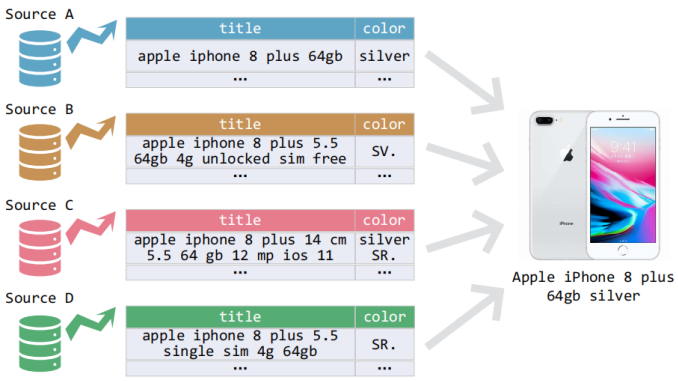

图 3.9 multi-table EM 的例子

3.10.2 文章的研究动机

现有的无监督多表实体匹配方法主要基于聚类，存在效率低下和可扩展性差的问题。此外，实体表示的有限能力以及实体匹配中的传递性冲突也影响了匹配的有效性。为了解决这些问题，本文提出了一种新的无监督多表实体匹配方法，称为 MultiEM，旨在提高匹配的效率和效果。

3.10.3 文章的总体解决思路和大致实现方式

MultiEM 方法将多表实体匹配问题形式化为两个步骤：合并（merging）和剪枝（pruning）。在合并阶段，提出了一种可并行化的逐表层次合并算法，以提高匹配效率。在剪枝阶段，通过增强实体表示质量和处理传递性冲突来提高匹配效果。此外，开发了基于密度的剪枝策略来消除异常值，进一步提高匹配效果。具体技术包括：

1. 利用 Sentence-BERT 模型进行实体表示。
2. 自动属性选择策略来增强实体表示。
3. 基于近似最近邻搜索（ANNS）的合并策略。
4. 基于密度的剪枝策略来识别和移除异常实体。

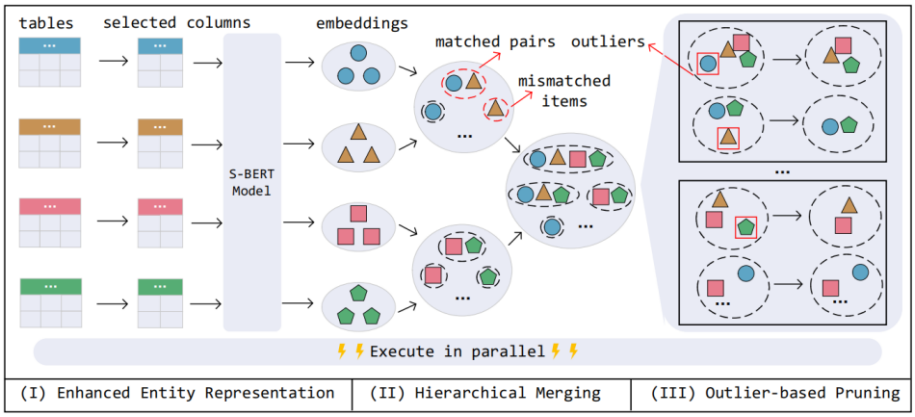

图 3.10 MultiEM 框架图

3.10.4 对本篇文章的思考

MultiEM 方法在无监督多表实体匹配问题上提供了一种新颖的解决方案，通过合并和剪枝两个阶段有效提高了匹配的效率和效果。特别是，该方法通过自动属性选择和基于密度的剪枝策略，解决了现有方法在实体表示和传递性冲突处理方面的局限性。

然而，MultiEM 方法也有其局限性。首先，为了确保效率，该方法主要关注基于表示的实体匹配技术，而没有探索更多基于交互的技术，这可能限制了其在更复杂数据集上的应用。其次，该方法在层次合并过程中没有考虑每个数据项的合并路径，这可能对后续的剪枝步骤有帮助。

未来的工作计划包括探索更高效的合并策略以支持更大规模的数据，以及研究交互技术如自监督学习来增强匹配效果。这些努力将进一步推动实体匹配技术的发展，并使其能够处理现实世界中更大规模和更复杂的数据。

4 总结

本文综述了最近两年在 CCF A 类会议和期刊上发布的关于实体解析（Entity Resolution, ER）的十篇代表性研究。这些研究涵盖了实体解析技术的各个方面，从基础理论的探索到具体技术的实现，体现了该领域内的技术进步和多样化应用。文章的选择和分析主要围绕实体解析的核心问题，包括多意图实体解析、隐私保护、多任务处理、弱监督学习、成本效益以及数据集成与机器学习的结合等方面。

在实体解析的应用中，这些研究不仅提高了匹配的准确性和效率，而且还着重考虑了隐私保护和低资源环境下的应用需求，显示了实体解析技术在处理复杂和敏感数据中的关键作用。此外，通过引入先进的机器学习模型和算法，如图神经网络、自监督学习和对比学习，研究人员能够更有效地解决传统实体解析方法无法应对的多源和多类型数据集成问题。

从技术创新角度来看，这些研究不仅推动了实体解析技术的发展，还为数据科学和人工智能领域的相关研究提供了新的思路和工具。例如，通过结合深度学习与数据隐私技术，解决了在保护隐私的同时进行有效实体解析的难题；而多任务学习模型的应用则显著提高了模型的泛化能力和应用范围。

综合来看，虽然当前实体解析技术已取得显著进展，但仍面临一些挑战，如算法的可扩展性、高效性以及在特定应用场景下的适应性等问题。未来的研究可朝向优化现有算法、提高算法在复杂环境下的鲁棒性，以及探索更多跨学科的应用场景，以进一步推动实体解析技术的深入发展和广泛应用。

参考文献

[1] A.Morana, T.Morel, B.Berjawi, et al. Geobench: A geospatial integration tool for building a spatial entity matching benchmark[C]. In proceedings of the 22nd International Conference on Advances in Geographic Information Systems (SIGSPATIAL/GIS), 2014: 533–536.

[2] S.Shah, V.Meduri, and M.Sarwat. GEM: An efficient entity matching framework for geospatial data[C]. In proceedings of the 29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL/GIS), 2021: 346–349.

[3] G.McKenzie, K.Janowicz, and B.Adams. Weighted multi-attribute matching of user-generated points of interest[C]. In proceedings of the 21st International Conference on Advances in Geographic Information Systems (SIGSPATIAL/GIS), 2013: 440–443.

[4] X.Meng, and Z.Du. Research on the big data fusion: Issues and challenges[J]. Journal of Computer Research and Development, 2016, 53(2): 231–246.

[5] X.Chu, I F.Ilyas, and P.Papotti. Holistic data cleaning: Putting violations into context[C]. In proceedings of the 29th IEEE International Conference on Data Engineering (ICDE), 2013: 458–469.

[6] 高云君,.葛丛丛,.郭宇翔, 等. 面向关系型数据与知识图谱的数据集成技术综述[J]. 软件学报, 2022: 0–0.

[7] R.Cappuzzo, P.Papotti, and S.Thirumuruganathan. Creating embeddings of heterogeneous relational datasets for data integration tasks[C]. In proceedings of the 2020 International Conference on Management of Data (SIGMOD), 2020: 1335–1349.

[8] M.Bilgic, L.Licamele, L.Getoor, et al. D-dupe: An interactive tool for entity resolution in social networks[C]. In proceedings of the 1st IEEE Symposium On Visual Analytics Science And Technology (IEEE VAST), 2006: 43–50.

[9] K.Qian, L.Popa, and P.Sen. Active learning for large-scale entity resolution[C]. In proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM), 2017: 1379–1388.

[10] V.Rastogi, N.Dalvi, and M.Garofalakis. Large-scale collective entity matching[J]. arXiv preprint arXiv:1103.2410, 2011.

[11] P.Kouki, J.Pujara, C.Marcum, et al. Collective entity resolution in multi-relational familial networks[J]. Knowledge and Information Systems, Springer, 2019, 61: 1547–1581.

[12] W.Fan, X.Jia, J.Li, et al. Reasoning about record matching rules[J]. Proceedings of the VLDB Endowment, VLDB Endowment, 2009, 2(1): 407–418.

[13] R.Singh, V V.Meduri, A.Elmagarmid, et al. Synthesizing entity matching rules by examples[J]. Proceedings of the VLDB Endowment, VLDB Endowment, 2017, 11(2): 189–202.

[14] R.Singh, V.Meduri, A.Elmagarmid, et al. Generating concise entity matching rules[C]. In proceedings of the 2017 International Conference on Management of Data (SIGMOD), 2017: 1635–1638.

[15] N.Vesdapunt, K.Bellare, and N.Dalvi. Crowdsourcing algorithms for entity resolution[J]. Proceedings of the VLDB Endowment, VLDB Endowment, 2014, 7(12): 1071–1082.

[16] C.Chai, G.Li, J.Li, et al. Cost-effective crowdsourced entity resolution: A partial-order approach[C]. In proceedings of the 2016 International Conference on Management of Data (SIGMOD), 2016: 969–984.

[17] R.Wu, S.Chaba, S.Sawlani, et al. Zeroer: Entity resolution using zero labeled examples[C], In

proceedings of the 2020 International Conference on Management of Data (SIGMOD), 2020: 1149–1164.

[18] B.Li, Y.Miao, Y.Wang, et al. Improving the efficiency and effectiveness for BERT-based entity resolution[C]. In Proceedings of Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI), 2021, 35: 13226–13233.

[19] M.Ebraheem, S.Thirumuruganathan, S.Joty, et al. Distributed representations of tuples for entity resolution[J]. Proceedings of the VLDB Endowment, VLDB Endowment, 2018, 11(11): 1454–1467.

[20] S.Mudgal, H.Li, T.Rekatsinas, et al. Deep learning for entity matching: A design space exploration[C]. In proceedings of the 2018 International Conference on Management of Data (SIGMOD), 2018: 19–34.

[21] Y.Li, J.Li, Y.Suhara, et al. Deep entity matching with pre-trained language models[J]. arXiv preprint arXiv:2004.00584, 2020.

[22] S.Balley, C.Parent, and S.Spaccapietra. Modelling geographic data with multiple representations[J]. International Journal of Geographical Information Science, Taylor & Francis, 2004, 18(4): 327–352.

[23] M.Schäfers, and U W.Lipeck. SimMatching: Adaptable road network matching for efficient and scalable spatial data integration[C]. In proceedings of the 1st ACM SIGSPATIAL PhD workshop (SIGSPATIAL), 2014: 1–5.

[24] E.Safra, Y.Kanza, Y.Sagiv, et al. Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets[J]. International journal of geographical information science, Taylor & Francis, 2010, 24(1): 69–106.

[25] S.Isaj, E.Zimányi, and T B.Pedersen. Multi-source spatial entity linkage[C]. In proceedings of the 16th International Symposium on Spatial and Temporal Databases (SSTD), 2019: 1-10.

[26] P.Balsebre, D.Yao, G.Cong, et al. Geospatial entity resolution[C]. In proceedings of the ACM Web Conference 2022 (WWW), 2022: 3061–3070.

[27] V.Sehgal, L.Getoor, and P D.Viechnicki. Entity resolution in geospatial data integration[C]. In proceedings of the 14th ACM International Symposium on Geographic Information Systems (GIS), 2006: 83–90.

[28] S.Isaj, V.Kaffes, T B.Pedersen, et al. A supervised skyline-based algorithm for spatial entity linkage[C]. In proceedings of the 25th International Conference on Extending Database Technology (EDBT), 2022: 2–220.