華中科技大學

研究生（参加校内外公开学术活动）报告

题 目：**大型语言模型中数据库实体匹配会议报告**

学　　号　**M202273662**

姓　　名　金灿果

专　　业　计算机科学与技术

指 导 教 师　左琼

院（系、所）　计算机学院

华中科技大学研究生院制

填 表 注 意 事 项

一、本表适用于攻读硕士学位研究生选题报告、学术报告，攻读专业硕士学位研究生实践环节报告，攻读博士学位研究生文献综述、选题报告、论文中期进展报告、学术报告等。

二、以上各报告内容及要求由相关院（系、所）做具体要求。

三、以上各报告均须存入研究生个人学籍档案。

四、本表填写要求文句通顺、内容明确、字迹工整。

报告内容

2023 年 9 月 19 日，香港科技大学周晓方教授进行计算机学院计算机科学与技术专业建设 50 周年学术论坛——使用大型模型进行数据库实体匹配研究的讲座报告。

周晓方教授率阐述了数据科学的三个核心构成部分，即数据质量管理、大数据系统以及高级查询处理。介绍了 JC STEM 实验室的发展规划蓝图，以及现阶段香港生成人工智能研发中心（HKGAI）开展的开源大语言模型（LLM）项目。

重点聚焦于数据质量管理中的数据整合领域，深入探究实体匹配在大语言模型（LLM）情境下的发展态势与改进路径。周晓方教授在阐释工作整体知识架构之后，详细讲述了多种技术应用于实体匹配的演进历程并剖析其优缺点，涵盖基于规则的方法、基于机器学习的方法、基于深度学习的方法、微调语言模型方法以及最新的 LLM 方法。早期技术所依托的模型尽管结构较为简易且资源消耗低，然而其在文本语义与上下文理解方面存在明显不足；与之相较，运用 LLM 能够凭借极少的用例输入实现良好的训练成效，极大地削减了人力成本投入。

在 MFL、Diversity、VoteK 以及 VoteK + 这四种采样方法之中，VoteK + 方法因灵活运用近邻图而展现出相对更优的效果表现。然而，鉴于不同数据集环境所具有的独特性与差异性，各方法在实际应用中所达成的最终效果均存在一定程度的浮动变化空间。当前，关于如何在多样化的情境下合理地抉择适宜的方法尚未形成具有唯一性与确定性的定论，仍有待于后续更为深入且系统的研究探索以及针对性的改进优化工作予以推进。此外，针对缺失值处理等相关方面所开展的研究工作对于实体匹配精度的提升具有极为关键的意义与价值，其能够为整体研究的深入发展提供重要的支撑与助力。

大语言模型（LLM）于语言翻译、文本分类以及问答等自然语言处理任务中展现出卓越的先进性能，其具备捕获语义与上下文信息之能力，由此引发了针对将 LLM 应用于实体链接任务的探索。实体链接作为一种跨数据集开展实体匹配（如涉及人员、组织或产品等实体）的关键任务，隶属于结构化数据（即数据库表中的记录）范畴内的数据库研究课题。

系统回顾了实体匹配研究的发展脉络，涵盖早期基于规则或统计方法的实践探索，近期基于深度学习与预训练模型的微调方法应用，以及当下基于基础模型的研究进展。通过深入剖析，旨在明确这一新兴领域的研究趋势与现存差距，并详细阐述在运用 LLM 进行数据库实体匹配研究方面所开展的相关工作，以期为该领域的后续研究提供参考与借鉴，推动实体匹配技术在数据库研究中的进一步发展与创新。

学习感悟

在现代数据科学的宏大架构中，每一个组成部分都相互关联、相互影响，数据质量管理、大数据系统以及高级查询处理的协同运作，是推动诸如实体匹配这类具体任务研究进展的基石。科研工作在实际落地过程中的系统性与复杂性。

从早期基于规则的方法，其以简单直接的逻辑判断为基础，虽然易于理解和实施，但在面对复杂多变的自然语言文本时，其局限性逐渐显现，对于语义和上下文信息的把握能力不足成为制约其发展的关键因素。

随着深度学习技术的兴起，基于深度学习的方法为实体匹配注入了新的活力，它能够自动学习数据中的深层次特征和模式，在一定程度上提升了匹配的准确性和泛化能力。而预训练模型的微调方法进一步利用了大规模语料库的预训练优势，减少了对大量标注数据的依赖，加快了模型的训练速度和效果提升。然而，这些方法在面对实体匹配任务时，仍存在一些难以突破的瓶颈，如对语义理解的深度有限、模型训练的资源消耗较大等问题。

技术的创新与研究需要调研清楚前沿完整的学术成果，深刻了解并分析其不足，例如在进行 LLM 研究前需要对不同采样方法如 MFL、Diversity、VoteK 和 VoteK + 进行使用场景、创新点、不足之处的完整调研。在提出新的研究方法时要明确其改进点与适用场景，有明确的对比分析条件。

在技术研究中，即使是一个看似微小的环节 —— 输入用例的选择，也蕴含着丰富的研究内涵和不断探索的空间。VoteK + 方法虽在当前表现出相对较好的效果，但不同数据集环境下各方法效果的不确定性也表明，在技术研究领域，永远不存在一劳永逸的解决方案。需要持续地深入研究、不断地尝试和优化，才能在复杂多变的数据世界中找到更为精准和高效的方法。

研　究　生　签　字_____

指　导　教　师　签　字_____

院(系、所)领导签字_____

年　　月　　日