ISCAS 中国科学院软件研究所学术年会暨重点实验室科技活动周 2025 第十届

工具系统

智能查重一体化工具

王成纬，鲍越，刘李仕皓，常宏芳，贾佳妮，王丹丹，胡渊喆*

智能博弈重点实验室

＊ 通讯地址:yuanzhe@iscas.ac.cn

系统简介

- **智能查重一体化工具**作为系统建设的重要辅助工具，主要支撑代码、文档、功能三类研发产物的查重需求，服务于成果评审与合规性审查等业务环节，依托单位内部科研网络部署运行，按照“统标准、统接口、统服务”的总体要求，立足科研产物质量管理主线，聚焦多源研发数据的结构建模与比对，开展统一规范、高效准确的重复性检测能力建设。
- **代码查重子工具：**基于轻量级指纹索引算法，支持多语言环境下的代码快速查重。系统可对大规模项目中的源码文件进行高速预处理与索引比对，定位逻辑雷同或复制粘贴片段，具备高吞吐、低资源消耗的工程实用性优势。
- **文档查重子工具：**支持文本、表格与图片等多模态内容的深度比对，融合 OCR 技术、表格结构解析与自然语言对齐算法，能够准确识别不同格式文档中的内容重复。输出详细的重复片段高亮与相似度分析报告。
- **功能查重子工具：**以自然语言理解为核心，通过分析调用路径与需求功能等信息，评估功能实现之间的潜在重合度。该模块可辅助识别伪创新、功能复现及研发过程中的冗余设计，提升项目功能层级的结构合理性。

问题挑战

➢ **数据割裂，视角不统一**

- 代码、文档、功能数据分散存储
- 缺乏统一的查重平台

➢ **定义混乱，难以对齐**

- 模块命名不统一，难以识别重复逻辑
- 文档内容结构差异大，描述方式主观性强

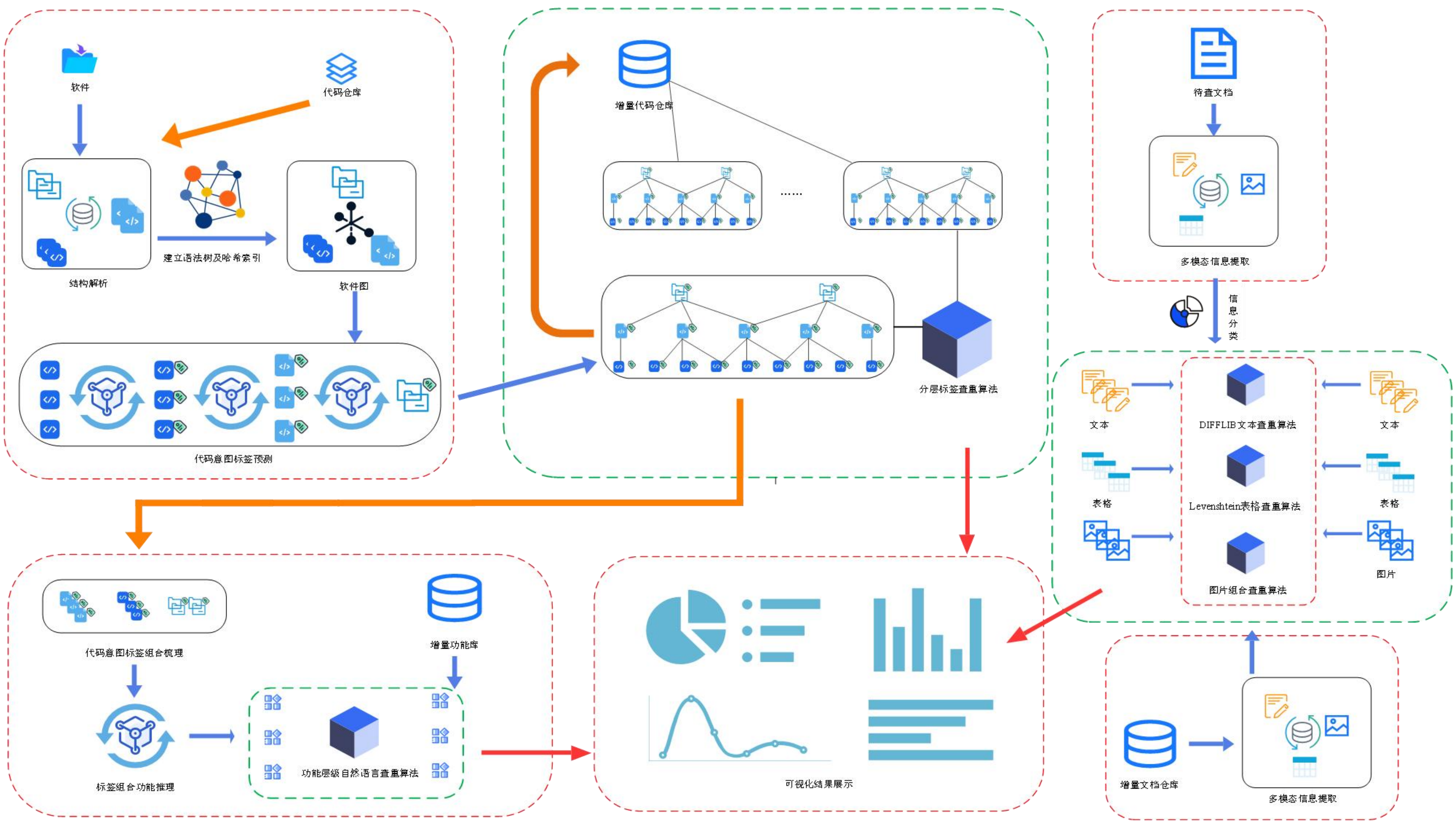
➢ **多模态难融合，匹配易偏差**

- 自然语言、代码之间缺乏有效对齐机制
- 模型无法充分理解上下文，导致语义查重不准确

➢ **多模态难融合，匹配易偏差**

- 仅靠文本或语法比对易误判
- 判重结果可解释性弱，人工复核成本高

设计实现



应用方向

科研项目管理

辅助项目产物查重，保障成果唯一性与规范性

成果转化评估

对比已有软件、专利、文档，提升成果创新性认定精度。

代码资产审查

为大型代码平台提供模块相似性检测服务。

研发流程质量控制

自动嵌入流程，降低重复，提高研发效率与质量。

数据驱动的绩效评估

量化查重结果，支撑科研评估与立项决策

工具赋能科研数字化转型

模块化集成查重，赋能科研平台与工具链建设