从知识图谱到数据中台: 华谱系统

吴信东[1, 2, 3, 4] 盛绍静[1, 2, 3] 蒋婷婷[1, 2, 3] 卜晨阳[1, 2, 3] 吴明辉[4, 5]

摘 要 针对碎片化的各姓氏家谱数据, 华谱系统通过构建家谱知识图谱的数据中台, 能够解决数据孤岛、烟囱式开发等问题. “数据中台”是一个源自国内的新近技术概念, 在华谱系统建设中, 我们通过家谱知识图谱的构建和应用, 对这个概念进行了正式定义. 基于这个定义和对应的 7 项核心功能, 本文提出一种用于家谱数据分析的数据中台建设架构 Huapu-CP (华谱系统), 并通过该架构详细介绍面向家谱领域的数据中台核心技术, 分析数据中台构建的关键问题.

关键词 家谱建设, 数据中台, 数据治理, 知识图谱

引用格式 吴信东, 盛绍静, 蒋婷婷, 卜晨阳, 吴明辉. 从知识图谱到数据中台: 华谱系统. 自动化学报, 2020, **46**(10): 2045−2059

DOI 10.16383/j.aas.c200502

Huapu-CP: From Knowledge Graphs to a Data Central-Platform

WU Xin-Dong[1, 2, 3, 4] SHENG Shao-Jing[1, 2, 3] JIANG Ting-Ting[1, 2, 3] BU Chen-Yang[1, 2, 3] WU Ming-Hui[4, 5]

Abstract With fragmented family tree data, Huapu aims to solve the problems of data islands and chimney development by building a data central-platform based on genealogical knowledge graphs. Data central-platform is a new technology that has recently emerged in China. During the construction of the Huapu system, we started with knowledge graph construction and analytics, and then formulated a formal definition for this technology concept. Based on this definition and its corresponding 7 key components, this paper proposes a Huapu-CP (or Huapu for short) framework for genealogical data analytics and presents the core techniques and key challenges in building such a genealogical data platform.

Key words Genealogy, data central-platform, data governance, knowledge graph

Citation Wu Xin-Dong, Sheng Shao-Jing, Jiang Ting-Ting, Bu Chen-Yang, Wu Ming-Hui. Huapu-CP: from knowledge graphs to a data central-platform. *Acta Automatica Sinica*, 2020, **46**(10): 2045−2059

家谱数据是典型的碎片化数据, 具有海量、多源、异构、自治等典型的大数据特征. 家谱起源最少可追溯至先秦时代[1], 家谱中不仅记录了族人最基本的世系状况、族人的姓氏源流、族规家训等内容, 还蕴含了丰富的历史、经济等复杂信息, 因此为人类历史、经济和文化研究提供了重要基础, 还为遗传病研究[2]、人类寿命长短分析[3] 提供了宝贵资料, 被认为与正史、地方志并列为记录和研究历史的三大基石[4]. 因而, 家谱数据系统的建设不仅需要解决家谱数据的存储问题, 还应该为家谱知识挖掘与推理的研究提供数据支撑, 实现大数据技术与人文社会科学研究的 “双赢”[5−6], 为广大民众寻根问祖提供家谱应用支撑.

目前, 家谱数据系统建设还面临许多难题. 1) 数据汇聚困难, 传统的家谱以纸质化家谱为主, 家谱数据类型多样、数据多源异构使得家谱数据电子化仍旧处于手工保存阶段, 需要非常巨大的处理成本和转换成本. 2) 数据融合、治理困难, 多源家谱数据往往具有不确定性和不一致性, 传统数据融合、治理技术在家谱数据中并没有很好的适用性. 3) 如何实现个性化家谱应用, 现有家谱修建网站仍旧停留在家谱数据的存储机制开发上, 存在平台同质化、功能单一化等问题, 未能根据家谱数据需求开发相应的家谱应用如同名人物分析、跨姓氏关联分析等. 为提高跨姓家谱大数据的挖掘和分析利用, 我们从 2016 年开始, 利用大数据知识工程项目的基础理论和关键技术[7−8], 建设了一个面向所有华人姓氏的家谱系统—华谱系统 (https://www.zhonghuapu.com/). 目前, 华谱系统中已有 1 290 万的家谱人物和

收稿日期 2020-07-06 录用日期 2020-09-14
Manuscript received July 6, 2020; accepted September 14, 2020
国家重点研发计划 (2016YFB1000901), 国家自然科学基金重点项目 (91746209), 教育部创新团队项目 (IRT17R3) 资助
Supported by National Key Research and Development Program of China (2016YFB1000901), The National Natural Science Foundation of China (91746209), The Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education (IRT17R3)
本文责任编委 杨涛
Recommended by Associate Editor YANG Tao

1. 大数据知识工程教育部重点实验室 (合肥工业大学) 合肥 230009 2. 合肥工业大学 计算机与信息学院 合肥 230009 3. 合肥工业大学 大知识研究院 合肥 230009 4. 明略科技集团 北京 100084 5. 北京大学软件与微电子学院 北京 102600
1. Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Hefei 230009 2. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009 3. Research Institute of Big Knowledge, Hefei University of Technology, Hefei 230009 4. Mininglamp Technology, Beijing 100084 5. School of Software and Microelectronics, Beijing University, Beijing 102600

721 个姓氏. 我们应用知识图谱构建和推理技术, 从海量家谱数据采集、汇聚开始, 在数据治理、数据开发、数据资产管理等模块建设过程中辅以 “HI” (人类智能)、“AI” (人工智能) 和 “OI” (组织智能) 三者的交互和协同, 创建标准、可复用的家谱数据中台架构.

随着国富民强, 盛世修谱在全国各地广为流行, 我们可收集、存储和处理的家谱数据正在以指数级的方式增长. 海势原理 (HACE theorem) 指出: 大数据具备海量、异构、分布和分散式控制的自治源、数据间复杂和演化的关联等典型特征[9]. 为解决家谱大数据中数据孤岛、烟囱式开发、数据开发速度慢和数据重复开发等问题, 华谱系统将所有华人姓氏家谱数据当成一项企业, 以 “为华夏写谱, 助百姓寻根”为使命, 其数据中台架构旨在提升家谱数据应用开发的敏捷性和数据服务的质量和速度, 进而将家谱数据转化为数据资产, 形成华谱的核心竞争力.

数据中台是近年来源于国内的一个技术概念, 旨在利用数据技术对海量数据进行采集、存储、计算、加工、统一表示, 形成标准后的数据 API, 进而提高数据的共享能力. 阿里巴巴、明略科技、百度、网易云等大数据和人工智能公司针对不同行业数据近几年建设了多个具有共享服务能力的平台. 但总体上来说, 数据中台的建设尚处于起步阶段, 还没有统一的数据中台建设标准、规范以及评价指标, 如何建设数据中台正在成为学术界和工业界的一个研究前沿.

本文详细介绍华谱系统作为一个家谱数据中台的定义、功能模块、关键技术、挑战问题以及相应的解决方案. 其数据中台框架具有以下特点:

1) 基于数据的逻辑管理和物理管理, 建设了家谱领域数据管理体系, 提高数据服务效率, 加速数据价值变现过程, 打破了信息之间的屏障.

2) 基于 HAO (Human intelligence, artificial intelligence and organizational intelligence) 智能体系[10], 采用数据规范、数据清洗等数据治理技术形成了包括家谱人物数据、人物关联数据、社交数据、日志数据等的数据体系, 更好地赋能于华谱系统前台业务.

3) 融合知识图谱技术, 以家谱知识图谱作为数据中台架构的数据底座, 从业务的角度组织数据. 完成传统数据模式无法支持的节点关联分析、社区发现、用户推荐等复杂计算和挖掘任务.

本文第 1 节对 “数据中台”的定义和功能模块进行阐述; 第 2 节对华谱系统数据管理架构建设和关键技术进行详细描述; 第 3 节对家谱数据体系建设仍存在的问题进行阐述并对家谱系统建设前景进行展望; 第 4 节对全文总结.

1 数据中台: 定义和 7 项核心功能

1.1 数据中台

2017 年, 阿里巴巴出版了构建企业级共享服务体系的中台战略[11], 随之引起了数据中台的研究热潮. 目前, 数据中台尚没有统一且规范的定义, 以下为两个有代表性的定义.

定义 1[12]**.** 数据中台是一套可持续 “让企业数据用起来”的机制, 是一种战略选择和组织形式, 是依据企业特有的业务模式和组织架构, 通过有形的产品和实施方法论支撑, 构建一套持续不断把数据变成资产并服务于业务的机制.

定义 2[13]**.** 数据中台是一个用技术连接大数据计算存储能力, 用业务连接数据应用场景能力的平台.

由这些定义可知, 数据中台的基础是数据, 其主要目标为提供快速共享和服务的数据应用, 实现数据价值最大化. 下面是本文对数据中台的正式定义.

定义 3. 数据中台将一个机构 (企业、事业、或政府部门) 的数据作为战略资产进行管理, 是从数据收集到处理应用的一套管理机制, 以期提高数据质量, 实现广泛的数据共享, 最终实现数据价值最大化. 数据中台建设覆盖数据的逻辑管理和物理管理, 逻辑管理包括数据结构的设计和数据之间相关性的分析, 如数据仓库; 物理管理包括数据的存储和检索.

中台和数据中台是源自国内科技公司的技术名词, 国内外 (包括 Gartner) 的英文翻译还没有统一口径. 将 “中台”翻译成 Middleware、Middle Platform、Middle Office 和 Platform 的都有, 但都没有体现上述定义中数据共享和数据价值最大化的实质. 作者们也曾思考过用中文拼音 Zhongtai 来作为英文名称, 但经过多轮比较, 还是选择了 Central-Platform, 既有 “中”的体现, 但实质是区别于 “前台”和 “后台”的中央位置和核心重要性.

1.2 数据中台的核心功能

数据中台的核心是一个机构的数据资产管理, 其核心功能包括以下 7 个部分:

1) 数据的物理管理: 包括多源数据的采集、汇聚、存储、索引和检索.

2) 数据的逻辑管理: 包括: a) 数据治理[14] (规范、清洗、交换和集成); b) 数据之间的层次建模和相关性分析.

3) 数据服务: 用数据实现多样化的用户服务, 让数据快速应用到业务场景中以发挥数据价值, 如预测、推理推荐和人机交互. 数据服务的一个共性需求是数据的可视化.

4) 知识图谱建设: 融入机构的知识体系和组织智能, 用以界定数据的来源和数据的服务范围.

5) 数据资产管理: 对数据对象和数据服务进行价值定义、保护、组织和管理, 实现数据价值的最大化.

6) 客户关系管理: 采集和分析用户对数据和数据服务的使用行为, 理解和进一步服务用户的需求. 用户信息采集包括线上行为采集 (用户注册和/或通过 “埋点”的方式, 记录重要数据资产的被使用情况)、线下行为采集 (通过传感器、摄像头等智能硬件设备进行采集) 和网络爬虫.

7) 信息安全: 信息安全贯穿在数据管理、数据服务、数据资产管理和客户关系管理的各个核心功能之中, 需要保证中台上的数据和服务在物理层和逻辑层都是安全的.

数据的逻辑管理 (第 2 项功能) 中的数据治理[14]是将一个机构的数据作为战略资产来管理, 需要从数据收集到处理应用的一套管理机制, 以期提高数据质量, 实现广泛的数据共享, 最终实现数据价值的最大化. 数据中台中数据资产管理与数据治理的对象都是数据资产, 目的都是提高数据质量、实现广泛的数据共享, 但是, 我们认为它们仍有以下的不同之处:

1) 驱动不同: 内部风险管理和外部监管合规驱动数据治理技术的发展, 而数据资产管理是以业务应用为驱动, 从而实现资产价值最大化.

2) 目标不同: 数据治理的目标是正确的信息以正确的形式, 在正确的时候交付给正确的人, 数据资产管理的目标是对数据的合理管理和有效应用, 充分发挥数据的价值.

3) 面向数据不同: 数据治理内容包含数据标准管理、元数据管理、数据质量管理、数据安全管理等内容, 面向的是所有原始基本数据, 是对数据的基本处理; 数据资产管理包含数据价值管理和共享管理等内容, 面向的是能够产生价值的数据资源.

数据资产管理 (第 5 项功能) 和客户关系管理 (第 6 项功能) 需要根据中台的使用情况和业务场景进行不断地迭代和演化, 需要人机交互, 需要有企业知识的人类智能 (Human intelligence) 进行协同. 知识图谱建设 (第 4 项功能) 需要应用人工智能 (Artificial intelligence, AI) 技术将组织智能 (Organizational intelligence, OI) 融合到数据中台的建设之中. 所以, 一个数据中台的建设一定是一个数据应用领域 HAO 智能[10] (Human intelligence + Artificial intelligence + Organizational intelligence) 的成功落地.

2 数据中台

本节首先介绍华谱系统, 然后阐述家谱中台系统的主要架构, 以华谱数据中台建设为例, 详细介绍家谱中台系统中核心模块的主要技术和要解决的问题, 最终以家谱数据应用与服务为例, 验证本文提出家谱数据中台架构的合理性、有效性.

2.1 华谱系统目标

华谱是一个面向全体华人姓氏的家谱系统, 目标是构建一个完整的家谱数据大知识平台, 解决用户的寻根、溯源、传承和跨姓氏分析等需求, 系统提供修谱、社区分享、人物关联分析等服务, 其发展主要经历了三个阶段:

1) 数据数字化: 家谱起源于先秦时代, 具有重要的历史、文化和经济研究的重要价值. 目前, 家谱数据大多为纸质形式保存, 华谱系统将海量的家谱数据电子化后, 能够快速、便捷地实现家谱数据的分析与研究.

2) 数据标准化: 家谱数据具有海量、多源、异构、自治等特点, 没有标准的家谱数据保存规范. 例如在家谱中, 家谱人物的出生信息可能以 “生辰”、“诞辰”、“出生日期”或 “出生年月日”等形式保存, 其表示的含义相同, 但给家谱数据可视化、关联分析带来困难. 因而华谱系统数据标准化主要目的是提供一个统一的数据标准或行业规范, 以一种更好理解、更直观的方式展示数据资产.

3) 数据服务化: 华谱系统将家谱数据变为一种服务能力, 提供便捷、快速的家谱保存、家谱分卷打印、人物关联分析和家谱数据可视化等具体的前台应用, 实现数据的价值变现.

2.2 华谱数据中台架构

我们设计了一个家谱中台建设架构 Huapu-CP, 结合 HAO 智能、知识图谱等技术, 通过对家谱数据的物理管理和逻辑管理, 实现从数据收集、存储、处理到数据应用的一套完整的数据资产管理机制, 最终实现一个家谱数据共享平台. Huapu-CP框架图如图 1 所示, 我们将第 1.2 节中所述的数据中台 7 个核心功能做了以下实现, 并在接下来的章节中进行详细介绍.

1) HAO 智能. HAO 智能包含人类智能 (Human intelligence)、人工智能 (Artificial intelligence)

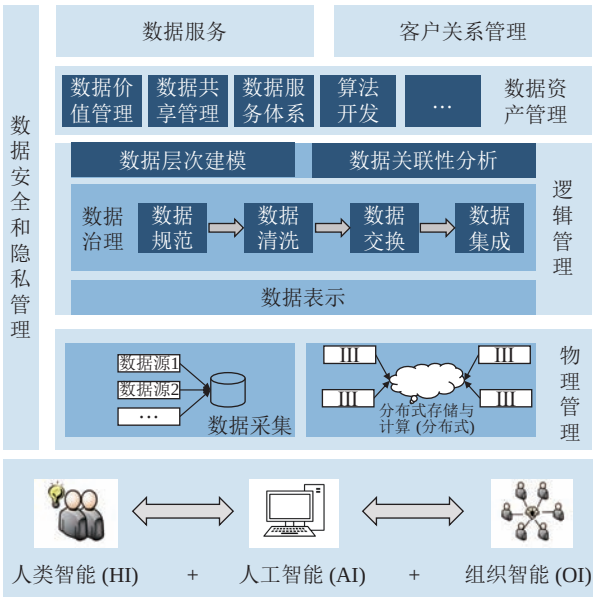


图 1 Huapu-CP 框架图

Fig.1 Overall framework of Huapu-CP

和组织智能 (Organizational intelligence), HAO 智能贯穿整个家谱数据中台的建设过程, 包括家谱知识图谱建设, 提供强有力技术支撑.

2) 数据物理管理. 物理管理模块主要涉及数据采集、索引、大数据分布式存储和计算等工具支撑, 为上层逻辑管理提供底层的工具支撑.

3) 数据逻辑管理. 逻辑管理包含数据融合、数据开发、为家谱数据服务提供数据使用接口, 支撑敏捷数据应用开发.

4) 数据服务. 包括家谱树、谱系图、家谱的分卷打印, 跨姓数据分析.

5) 数据资产管理. 这个模块包括数据和数据服务模块的价值管理、数据共享管理和数据服务体系等子模块.

6) 客户关系管理. 包括用户权限, 在线客户行为采集和分析.

7) 数据安全和隐私保护. 通过基于 HAO 模型的用户权限管理和应用权限管理, 实现数据安全与隐私保护.

2.3 物理管理

2.3.1 物理工具支撑和物理管理框架

家谱数据中台建设主要依托机房、基础网络设施、云资源和云服务等物理工具, 其中云服务是指通过网络获取的数据计算、存储等相关服务, 实现华谱数据中台中计算和存储能力的扩展.

针对数据存储方面的高性能、弹性扩展、容灾与备份等问题, Huapu-CP 选择图数据库集群的方式 (如图 2), 将数据分布存储到多个机器上, 并进行实时同步, 以保障数据的安全性、一致性及性能的可扩展性. 图数据库是以点的形式存储人、物等实体, 边存储实体之间的关联关系, 并且点和边中都可存储丰富的属性信息 (如人物节点中可存储姓名、出生年月等属性值). 图数据库具有数据表示更加直观、数据操作更加便捷等优点, 且数据库操作性能不会因数据量增长而降低.

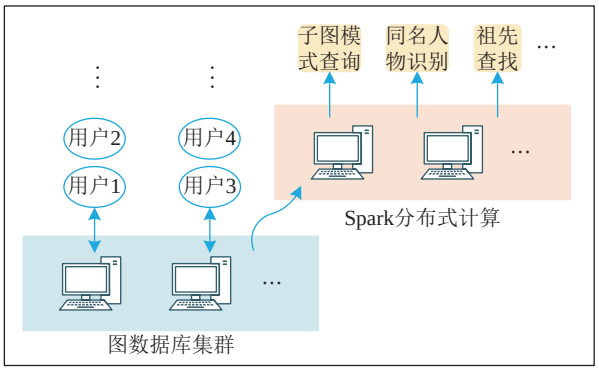


图 2 物理管理框架图

Fig.2 Physical management framework

针对图数据规模较大、图数据分析与挖掘耗时较长的问题, 通过对大数据计算算法和框架的对比研究[15], 华谱系统采用 Spark 分布式计算框架, 即在上层应用 (如子图模式查询、同名人物识别) 中, 利用图划分算法[16] 将原始的图数据拆分为多个子图, 基于分布式计算并行处理, 以节省整体计算时间、提高计算效率.

2.3.2 数据采集

华谱系统中家谱数据包含内部和外部的非结构化、半结构化和结构化的数据, 根据数据来源和数据特点, 实现多样化的数据采集模式和方法. 数据采集架构如图 3 所示, 主要包含 4 种方式实现数据采集:

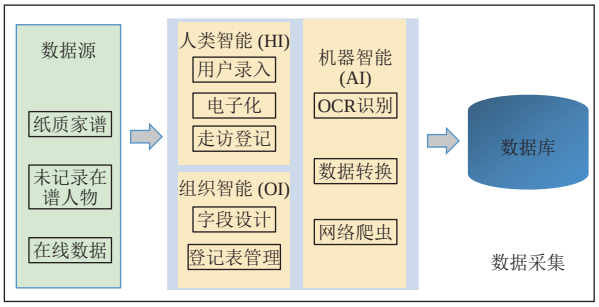


图 3 数据采集架构图

Fig.3 Data gathering framework

1) 线下数据采集包括对纸质族谱采集、新增人物登记. 针对纸质家谱转换为电子谱文本录入困难的问题, 采用光学字符识别 (Optical character recognition, OCR) 技术, 首先, 对预先收集的图片进行噪声去除、图像倾斜矫正等预处理; 然后, 调用文字识别接口将家谱转换成计算机可接受的形式. 另外针对新增人口, 采取入户登记, 责任到人的方式.

2) 互联网数据采集指在遵守相应协议和法规, 同时不给目标网站造成过大请求压力的前提下, 实现互联网新增数据与系统内部数据的有效融合. 目标网站包括百度百科、上海图书馆、中国历代人物传记资料库 (CBDB) 等. 采集工具包括 WebCollector[17]、WebMagic 等.

3) 线上行为采集包括系统自动记录 PC 端和移动端的用户的所有操作请求及内容, 并汇聚到数据库中.

4) 内部数据汇聚中家谱数据汇聚包括根据系统预定义语言模式, 完成 “世系图” “人物描述”数据的抽取存储, 并根据用户反馈调整语言模式; 用户行为数据汇聚是将用户在华谱系统中的操作进行日志记录.

网上家谱作为华谱系统主要数据来源之一, 具有海量、多源、异构、碎片化等特点, 为提升互联网数据采集的精度和准确性, 采用本团队开发的 Web 信息爬取工具 WebCollector[17]; WebCollector 是一个无需配置, 便于二次开发的 JAVA 爬虫框架并提供精简的 API, 在网络中抽取整个每位人物的海量 Web 信息.

2.4　逻辑管理

逻辑管理是家谱数据中台中数据资产建设的主战场, 是数据价值产生的核心环节. 通过数据采集得到的数据, 一般以最原始的状态堆积而成, 相当于商品的原材料. 逻辑管理相当于商品加工的流水线, 将数据转换为资产. 华谱系统中逻辑管理框架如图 4 所示, 将知识图谱技术、专家智能、组织智能等技术融入数据表示、数据治理等子模块中, 协调逻辑管理整个流程.

2.4.1　数据表示

数据表示是基于 HAO 智能和本体粒度划分技术, 构建家谱知识图谱.

1) 首先, Huapu-CP 基于 HAO 智能构建亲属关系模型. 模块接入分析加工后的数据, 请求领域专家解决特殊问题, 如生僻属性的定义与发展. 专家们查询数据库并提出问题分析需求, 然后讨论和掌握相应的处置原则, 即专家知识. 并在组织的协助下, 经过反复磋商, 总结得出一套问题陈述和定义. 获取的知识必须进行验证和修改, 领域专家们会反复讨论所收集的知识, 直至问题得到满意解释为止. 同时, 获取的知识也需要经过加工处理和管理发现才可用于解决实际问题.

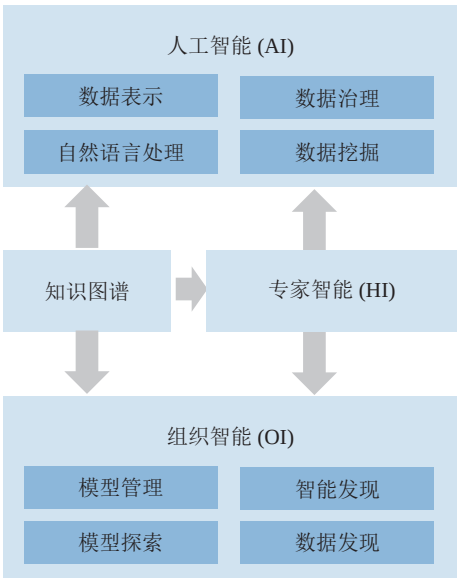


图 4　逻辑管理框架图

Fig.4　Logical management framework

2) 其次, 采用本体粒度划分技术, 分别以 “家谱”、“人物”为基本单元构建知识图谱. 具体地, 若根据 “家谱”划分, 则关于这个家谱的堂号、家训、家谱画像等被划分为静态信息. 家谱中的人物则被认为是动态信息. 若按照 “人物”划分, 则性别等与生俱来的属性被认为是静态信息. 婚姻状态、官职经历等被认为是动态信息. “人物”粒度的范围小于 “家谱”粒度. 并且 “人物”既可以作为 “家谱”粒度的属性, 又可以作为单独的粒度单元, 关联二者. “家谱”和 “人物”基本单元如图 5、图 6 所示.

2.4.2　数据治理

为提高家谱数据价值, 实现有效的数据治理, Huapu-CP 利用 HAO 数据治理构架 (图 7) 进行数据规范、数据清洗、数据交换和数据集成等数据治理工作 (流程图如图 8 所示).

1) 数据规范

华谱系统数据规范使用的技术有启发式规则、数据字典. 启发式规则指领域专家们通过对接入的家谱、人物数据集的特点进行分析总结后, 制定出一套计算机可理解的规则库, 然后根据规则对数据完成解析. 具体地, 系统现有的规则包括 NULL 值替换、正则处理、化 “繁”为 “简”原则、化 “简写”为 “全称”原则等. 数据字典指为不同的描述与标准数

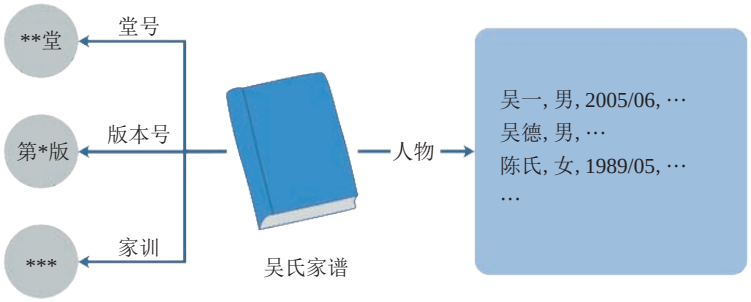


图 5 “家谱”基本单元

Fig.5 Basic unit of a genealogy

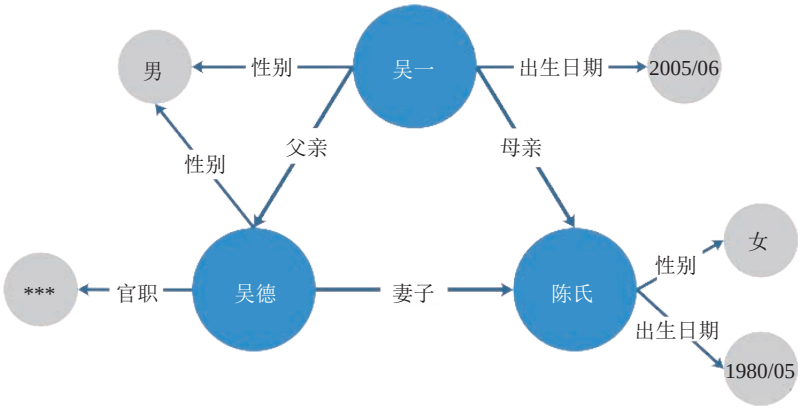


图 6 “人物”基本单元

Fig.6 Part of personal knowledge unit

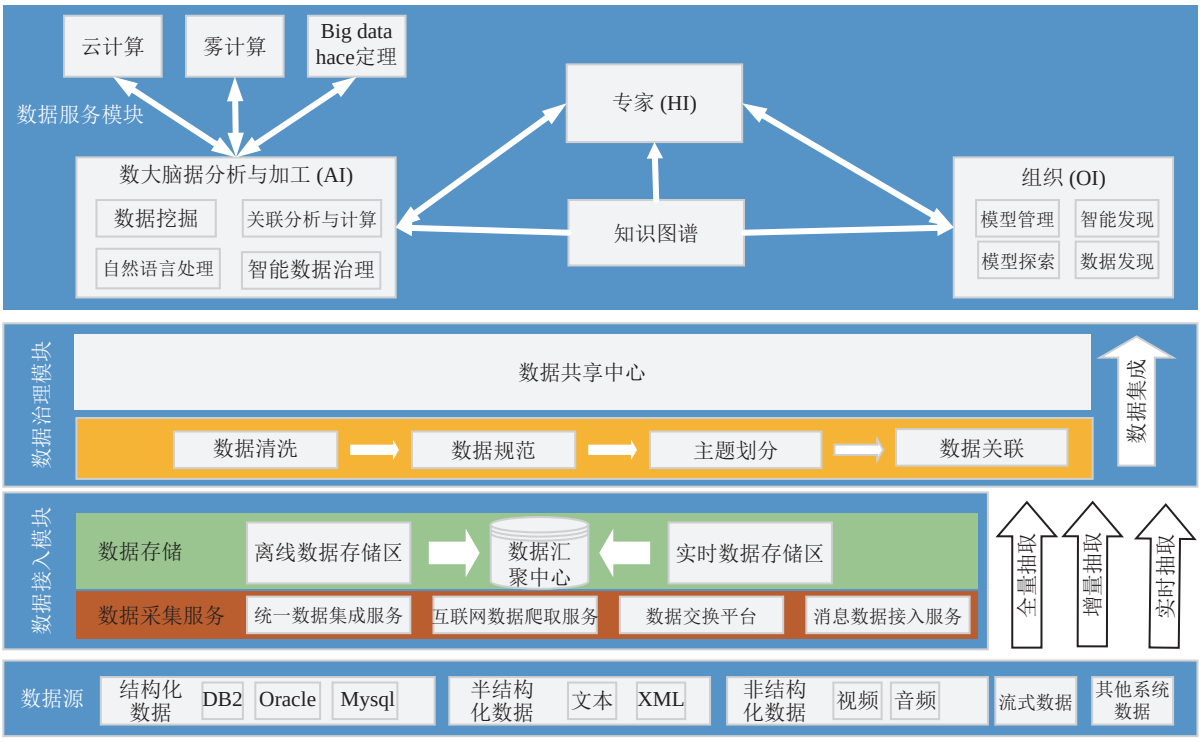


图 7 数据治理架构图[14]

Fig.7 Overall framework of data governance[14]

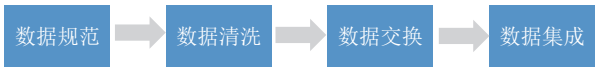


图 8　家谱数据治理流程图

Fig.8　Basic processes of data governance

据之间建立键值对, 一般可通过现有的知识库加工得到, 如根据中国古代纪年表、中国古代皇帝年号时间对照表可建立起中国古代纪年、中国古代皇帝年号与标准公元纪年之间的映射.

2) 数据清洗

华谱系统数据噪音主要来源于: a) 由于传承时间过于久远或保管不善等原因造成的原始数据缺失; b) 自动采集录入数据导致的数据噪音. 目前针对数据噪音有以下三种处理方式:

a) 若缺失的数据不会对主体对象造成过大的影响, 则直接忽略. 如人物除 ID 以外的属性信息, 官职、葬地等.

b) 有些数据的缺失虽然不会对主体对象造成影响, 但影响上层应用的性能; 如若人物的 “姓名” 属性缺失, 则无法显示完整的家谱树, 原因是家谱树的每一个结点以人物姓名作为标识. 对于这样的数据, 系统采用自动填充或人工补全的方式, 如填充吴氏男性人物的陈氏配偶姓名为 “吴陈氏”.

c) 对于人物关系或属性错误等噪音数据, 如个人父子关系自环, 即人物与自身建立了父子关系; 翁媳关系保存为父女关系; 两男性数据间建立了配偶关系等噪音数据. 这类数据错误严重影响家谱树展示、家谱打印等服务. 因而, 华谱系统采用了基于子图匹配的方法实现数据去噪: 即 “HI”归纳亲属关系知识图谱中存在的错误, 将其转化为错误数据模式图, 基于错误模式图通过子图匹配算法准确定位噪音, 并将检测的噪音数据反馈给用户修正实现数据降噪.

3) 数据交换与数据集成

华谱系统数据交换机制主要是协议式交换. 即在其他数据源与本系统之间定义了一个交互协议, 将其他数据源数据接入移植到本系统数据库中进行储存时, 必须遵循本系统的数据准则. 我们设计了碎片化家谱数据融合框架 FDF-HAO (Fragmented data fusion framework with human intelligence, artificial intelligence and organizational intelligence)[18] 对家谱数据进行集成, 实现碎片化家谱知识的融合, 利用实体对齐、冲突消解和属性融合等技术, 力求构建一个全面、准确的家谱数据库.

2.5　数据资产管理

家谱数据和家谱服务需要一系列资产管理和用户关系管理算法支撑, 华谱系统已形成较为完善的算法开发体系, 算法之间能够相互调用实现个性化用户需求, 且能够根据前端的应用需求变化及时调整算法开发的内容和策略; 本节中首先详细介绍华谱系统中包含的主要算法模块, 然后介绍数据价值管理数据共享管理内容.

1) 用户权限算法开发. 现有的系统主要采用基于角色的访问控制模型 (Role-based access control, RBAC), 该模型中包括用户、角色和权限三类实体, 其中角色指一系列不同权限的集合, 如果用户被指定为某个角色, 就能够拥有该角色所拥有的所有权限. 然而由于华谱系统的应用中存在: a) 同一个用户针对同一份家谱中不同的数据可能拥有不同的角色; b) 同一份家谱中的某些数据可能只能由同一角色中的部分用户修改的情况, Huapu-CP 采用了粗细粒度结合的权限模型 (见第 2.6 节). 并且, 我们设计了基于 HAO 模型的闭环权限管理机制 (见第 2.6 节). 因此, 权限算法设计主要包含:

a) 粗粒度权限判别算法: 类似于 RBAC 模型, 本文中的权限模型也设定了不同的 “角色”, 每个角色作为权限的载体, 是一定数量的权限的集合. 如表 1 中列出华谱系统中部分角色类型. 针对某用户对某一数据的操作请求, 用户权限控制中心首先获取该用户的角色集合 (粗粒度权限), 判断是否可根据该粗粒度权限集合直接判断是否有该数据的请求权限, 如果有权限则直接返回; 否则, 应基于细粒度权限判别算法处理.

b) 细粒度权限判别算法: 在部分应用中 (如家谱编修应用中多人协作修改同一份家谱), 存在同一个用户有多个角色、以及某些数据可能只能由同一角色的部分用户修改的情况. 针对该类情况, 无法完全基于角色判断用户的权限, 因此, 需设计细粒度 (面向具体数据) 的权限判别算法.

c) 日志分析与挖掘算法: 为了实现闭环权限控制, 系统实时地采集、存储用户行为数据和用户操作数据, 并设计智能的日志分析算法. 该模块包括检测用户行为是否异常、检测数据是否存在修改异常以及检测优质用户等.

2) 家谱人物导入/导出算法开发. 根据用户需求不同, 目前在家谱数据中, 华谱系统中开发了如下算法:

a) 单个人物导入/导出算法: 目前主要分为用户家谱登记表录入和系统新增单个人物两个部分, 均提供定制化人物录入功能, 可根据录入人物信息选择录入方法.

b) 批量导入/导出算法: 用户可将整份家谱数据

表 1 角色表
Table 1 Role table

角色类型	角色诞生及身份转变方式
普通用户	注册华谱系统的普通用户, 可进行创建家谱、查看公开数据等.
普通家谱成员	普通用户向某一共建家谱申请成为家谱成员, 只拥有针对该家谱的最基本权限, 例如查看该家谱中的基本信息.
家谱共建者	家谱成员向家谱创建者或核心修谱成员申请成为家谱共建者, 拥有上传数据、对本人上传的数据以及其他用户分享的数据拥有基本修改权限. 若涉及家谱主树结构变化, 需经过审核.
数据合作拥有者	包含部分、全部合作拥有者, 家谱成员向数据拥有者申请数据增加、修改等权限, 成为数据合作拥有者.
数据拥有者	家谱成员保存个人家谱数据, 成为数据拥有者, 拥有数据的全部权限, 且在共建家谱中其权限可转让或共享.
核心修谱成员	家谱成员向家谱创建者申请成为核心修谱成员. 拥有对该家谱数据的大部分权限, 包括查看、编辑、审核本家谱所有家谱人物. 不具备指定核心修谱成员、修改家谱名称等少数信息的权限.
家谱创建者	拥有该家谱的所有权限, 可指定每位家谱成员的权限级别.

保存至 Excel 或 Word 文件中, 通过解析文件中人物信息及人物关联信息实现家谱数据地批量导入.

c) 家谱打印算法: 将家谱数据按照指定格式进行处理: 按照统一、格式化的文本格式提供给用户.

d) 家谱分卷算法: 将家谱数据过多的家谱按照逻辑清晰、人物划分均衡的要求拆分为多个子谱, 便于家谱校对、保存工作的开展.

3) 家谱人物分析算法开发. 针对海量家谱数据, 为了实现对家谱数据的有效分析和利用, 我们实现了如下数据分析算法:

a) 家谱数据可视化: 系统为用户提供家谱树、家谱人物展示算法和工具开发.

b) 人物关联分析算法: 对任意两人间分析其关联路径分析, 利用数据可视化算法将分析结果直接展示给用户.

c) 同名人物检测: 相同/不同家谱中, 通过分析人物属性信息如姓名、出生日志等, 判断并提醒用户系统中可能是相同人物的列表, 实现避免重复人物录入, 建立不同家谱中的关联.

d) 人物合并: 系统检测两人物为同一人物时, 提供人物自动合并功能, 其主要涉及属性合并、家庭关系合并和社会关系合并, 合并完成后, 提供自动删除冗余人物的功能.

e) 家谱中孤立子树检测: 一份规范的家谱应是以一位家谱先祖为根节点的完整树, 用户在录入家谱数据时可能会录入未与家谱先祖建立关系的人物, 为家谱数据分析、家谱打印带来困难, 因而, 孤立子树检测算法为检测未连接至家谱树的子树根节点算法.

f) 人物辈分推算: 系统提供家谱辈分录入功能, 当保存人物辈分属性时, 自动推算人物的所处 “世”, 保证录入数据的一致性、准确性.

4) 用户行为分析算法开发 (用户关系管理): 根据家谱系统中用户行为数据提出日志分析算法[19], 实现异常用户检测、用户画像以及针对游客的标签

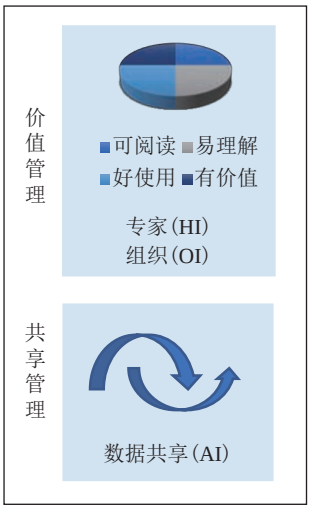


图 9 数据价值管理和数据共享管理
Fig.9 Data value management and data sharing management

记录, 进而实现高效地系统管理, 用户画像能够有效地优化系统服务、推荐并吸引用户.

5) 数据价值管理和数据共享管理: 图 9 描述了华谱系统数据价值管理架构. 数据价值管理的目标是面向业务人员提供可阅读、易理解、好使用、有价值的数据资产. 可阅读包括业务能直接通过查找、分类检索、自动推荐的方式直接读取数据; 易理解包括为数据新建名称、描述、取值类型、约束等标签; 好使用包括通过参数设置的方法完成业务人员的数据请求与变更. 有价值包括通过访问量、好评率、意见反馈等形式评估数据资产的价值. 针对数据共享管理, 系统采用流量控制、访问次数控制、权限认证、异常警告、安全协议 5 个监控措施管理数据的输入输出, 共享使用情况.

2.6 数据安全和隐私保护

本节从用户权限管理和应用权限管理两个方面介绍 Huapu-CP 架构中保障数据安全与隐私的机制.

1) 用户权限管理

现有的系统主要采用基于角色的访问控制模型 (Role-based access control, RBAC), 该模型中包括用户、角色和权限三类实体, 其中角色是 RBAC 模型的核心要素, 被定义为一系列不同权限的集合. 即如果用户被指定为某个角色, 就能够拥有该角色所拥有的所有权限. 该模型能够简化权限的管理过程.

然而, RBAC 在一些场景下并不适用. 我们以华谱系统中的 “共建家谱”应用为例说明: 在家谱编修过程中, 需要多位系统用户分工协作录入数据并对数据进行校正. 为保障数据的安全性和隐私性, 防止用户肆意修改、删除家谱数据, 需对用户权限进行统一管理和设置, 以有效地减少家谱编修的总时间. 虽然修谱的用户也可以被划分为不同角色 (见表 1), 然而同一用户针对不同的数据可能拥有不同的角色、家谱中的某个数据可能只能由同一角色中的部分用户修改 (见表 2). 例如, 某位用户可能是同一份家谱中部分数据的 “数据拥有者”角色 (拥有该数据的所有权限)、另一部分数据的 “共同拥有者”角色 (只拥有该数据的修改权限, 没有删除权限) 以及其他剩余数据的 “家谱共建者”角色 (只能查看, 没有修改和删除权限). 由于该用户针对同一家谱数据同时拥有三重角色, 原 RBAC 模型因此失效.

表 2　数据权限表

Table 2　Data authority table

数据类型公开级		相关描述
完全公开10		如公开家谱、百度百科人物; 所有人可进行查看、编辑
完全公开9		如公开家谱、百度百科人物; 所有人可进行查看.
8	针对共建家谱而言	部分用户可查看
7		部分用户可修改人物信息
6		部分用户可修改人物关系
5		部分用户可删除人物
4		仅数据录入者可修改
3		私有家谱
2		普通管理员可修改查看
1		超级管理员可修改、查看

为了解决上述问题, Huapu-CP 架构中提出了基于图数据库的 “粗细粒度结合”的权限管理方法 (见图 10), 并且提出了基于 HAO 模型的权限管理闭环架构 (见图 11). 该权限管理架构的基本思想如下.

a) “粗细粒度结合”的权限管理方法: 粗粒度表示该用户所拥有的角色, 细粒度指针对数据层面的权限管理. 权限控制的基本过程为: 针对某用户对某一数据的操作请求, 用户权限控制中心首先获取该用户的角色集合 (粗粒度权限), 判断是否可根据该粗粒度权限集合直接判断是否有该数据的请求权限, 如果有权限则直接返回; 否则, 在图数据库中查询该用户针对请求数据的角色并以此判断是否拥有所请求的权限 (细粒度权限控制). 由于 Huapu-CP 中采用图数据表示模型 (见第 2.4.1 节), 即以点的形式存储实体 (用户、家谱中的人物等), 以边的形式存储实体之间的关联关系 (用户是否有人物的修改权限等), 因此, 图数据库能够以 O(1) 的复杂度查询某一用户是否拥有所请求数据的权限.

我们以 “共建家谱”应用中两类用户请求删除某个家谱人物为例说明 “粗细粒度结合”的权限管理方法以及在系统中的程序执行流程.

情况 1. 用户 A (“家谱创建者”角色) 请求删除人物数据 X. 该请求会发送给 “用户权限控制中心”; “用户权限控制中心”通过调用统一的数据接口, 判断用户 A 为该家谱的创建者角色 (粗粒度权限)、拥有该家谱所有数据的所有权限, 因此将允许删除的指令传送给 “应用权限控制中心”, 应用权限控制中心判断 “家谱编修应用”的最大权限集合, 判断出该删除请求在权限集合范围之内, 因此将删除请求传递给数据接口, 由数据接口执行指令并将结果返回给用户.

情况 2. 用户 B (部分数据的 “数据拥有者”角色以及部分数据的 “共同拥有者”角色) 请求删除人物数据 Y 和 Z, 其中, 用户 B 实际是数据 Y 的数据拥有者 (有删除权限)、以及数据 Z 的 “共同拥有者” (没有删除权限). 类似于情况一的执行流程, “用户权限控制中心”首先判断出该用户在该家谱中有多重角色, 因此无法直接根据所拥有角色的权限判断是否可以删除人物数据, 需通过进一步调用数据接口获取该用户在所请求数据上的具体权限. 最终返回的结果为用户 B 可删除人物 Y、但无法删除数据 Z. 由于数据在系统中以图的方式表示, 即用户与家谱人物都以节点的方式存储、用户针对家谱人物的操作权限以边的形式存储, 因此读取该人物针对所请求数据的时间复杂度是 O(1).

综上, 粗细粒度结合的权限管理方法首先解决了原 RBAC 模型不能处理同一用户针对同一家谱有多重角色的问题, 当用户拥有角色中权限重复时, 系统能够根据细粒度管理的关联 “边”判断用户拥有的最高权限; 并且, 相较于关系数据库保存家谱人物数据, 采用图数据库查询细粒度的权限时间复杂度更低.

b) 基于 HAO 模型的权限管理闭环架构: 现有的主流架构没有考虑到权限管理的闭环问题, 有以

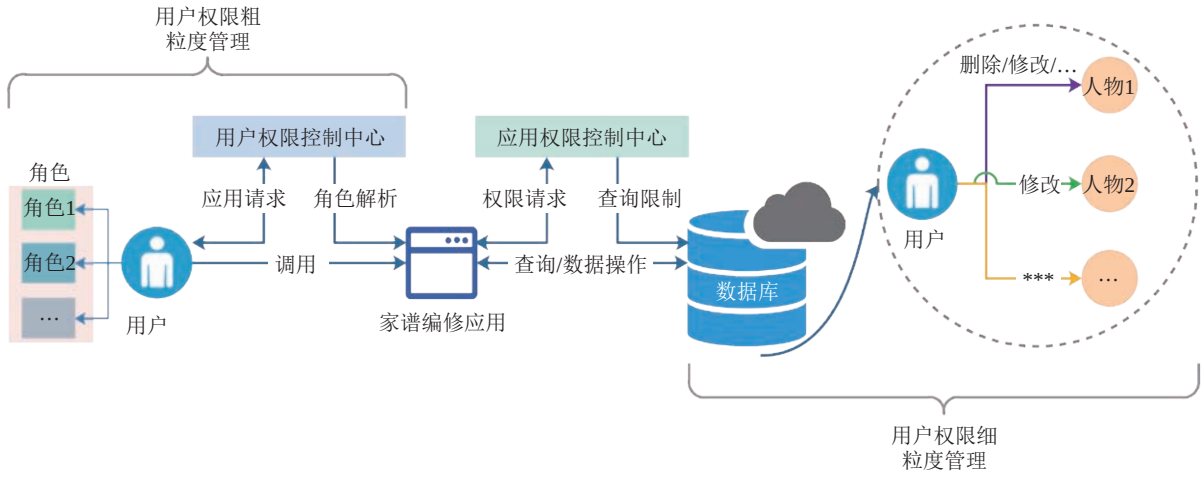


图 10 “粗细粒度结合”的权限管理方法

Fig.10 Multi-granularity based authority management method

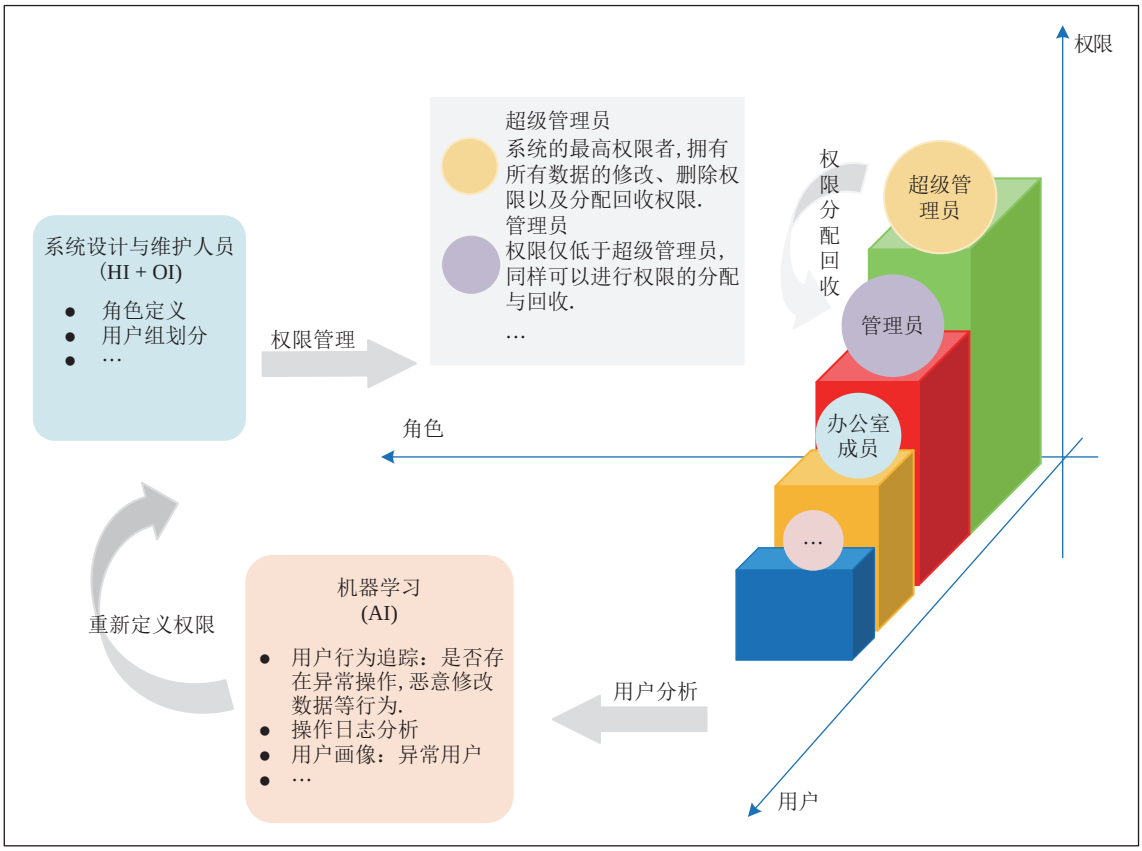


图 11 基于 HAO 模型的用户权限管理架构

Fig.11 HAO-Model based user authority management architecture

下两方面因素导致对用户的权限以及权限设计需要进行调整.

i) 权限体系的优化: 虽然基于领域专家 (HI) 和组织智能 (OI) 设计的权限体系具有一定程度上的合理性, 然而很难全面地考虑到用户在各方面的实际需求. 而能够填写产品反馈意见的用户可能只是少数, 公司可能很难全面收集与掌握用户对产品的实际评价. 因此, 有必要通过分析与挖掘用户的操作日志 (AI), 调整与优化权限设计.

ii) 异常用户或异常数据的检测与权限限制、优

质用户的权限升级: 考虑到系统中可能存在异常用户 (如恶意修改和删除数据的用户)、或者异常数据 (如被多位用户重复修改的数据), 系统也需要针对用户日志进行分析 (AI), 检测出异常用户和数据并进行暂时性权限限制, 并通知有更高权限级别的用户组进行权限审核. 另外, 为了提升应用的服务效果, 系统也可基于操作日志分析可能的优质用户, 将检测结果告知具有更高权限级别的用户组, 由更高权限级别的用户组可决定是否提高该优质用户的权限级别. 我们以 “共建家谱”应用中多位用户协作校对同一份家谱为例进行说明. 若某用户可能由于不熟悉系统操作或者恶意操作等原因, 错误地修改了多位家谱人物数据, 而这些数据被其他多位用户 (特别是被权限级别更高的用户) 修改回原数据, 系统可根据修改日志判断该用户存在异常, 应短暂性的限制该用户的修改权限并通知该家谱的创建者或者核心修谱成员审核该用户的权限. 再比如, 若一个家谱数据在相近的时间段内被多位用户修改, 但修改的内容差距较大、且存在重复修改的情况, 则可能原因是这几位用户对该数据的理解存在差异, 权限控制中心可锁定该数据、通知管理员用户判断数据正确版本并释放数据的修改权限.

基于上述分析, Huapu-CP 中提出了图 11 所示的基于 HAO 模型的用户权限管理闭环架构, 该权限设计架构的主要流程为: 由领域专家 (HI) 和组织智能 (OI) 设计用户的权限体系, 由人类专家 (HI) 审核后在权限控制中心提供接口, 最后基于日志的智能分析 (AI) 提供权限调整方面的反馈, 并再由 HI 或 OI 审核、优化. 其中, 为了实现闭环权限控制, 系统应实时地采集、存储用户行为数据和用户操作数据, 并设计智能的日志分析算法.

2) 应用权限管理

应用权限管理是为辅助敏捷应用开发而设置的权限管理中心. 通过设置应用读取数据的权限, 避免恶意修改读取数据程序而导致的数据泄露问题; 同时, 加入 HAO 模型实现应用权限管理的闭环, 便于应用的改进和升级. 开发一个新应用所需的权限管理流程如下: 首先, 由于每个应用只需部分数据的读取权限, 为了系统中数据资产的安全性与隐私性, 应限制数据访问权限为该应用所需的最小集合. 因此, “HI”与 “OI”可基于该应用的实际需求设计该应用所具备的最大权限集合. 其次, 由于应用的需求可能存在变化, 该应用交付使用后可针对用户行为数据分析或用户反馈数据分析 (AI) 得出该应用改进方向和内容, 反馈至专家 (HI) 和组织 (OI)重新设计应用权限, 形成权限管理的闭环.

2.7 数据服务

Huapu-CP 的建设是以数据应用为驱动的, 通过开发数据服务 API, 实现敏捷的应用开发, 最终实现广泛的家谱数据共享. 数据中台中规范统一的数据接口为敏捷的应用/服务研发奠定基石, 因而本节中将首先介绍华谱系统中所使用的数据接口, 其次介绍华谱系统中核心服务与应用.

1) 数据接口

数据中台中数据接口是应用/服务获取数据的唯一途径 (如图 12), 相较于没有采用数据中台的应用/服务开发方式, 主要区别和优点是:

a) 应用/服务开发者仅需关注核心算法的开发, 不需要对数据模型、数据获取权限进行设计和开发, 也无需考虑数据维护、数据模型更新、权限维护等问题, 因此可极大地减少开发者的工作量、节省应用的开发时间, 解决了应用/服务开发速度受限于数据开发的问题;

b) 由于所有数据的获取、存储直接调用统一的数据接口, 可极大地保障系统数据的安全性和隐私性, 并有助于对系统数据的统一管理.

Huapu-CP 中权限管理架构与数据接口分别如图 10 和图 11 所示, 与现有的数据中台相比, 主要创新在于如下两点:

a) 权限控制中心分为用户权限控制中心和应用权限控制中心两个部分 (如图 11 所示). 其中用户权限控制中心用于管理用户在具体应用中所具有的不同权限; 应用权限控制中心用于限制系统中每个应用所具有的最大权限, 以保证数据资产的隐私和安全.

b) 采用了 HAO 智能模型的权限管理机制, 改进了现有数据中台的权限设计 (详见第 2.6 节).

数据接口的主要功能包括: i) 数据模型的研发, 以及对数据结构、数据操作和数据约束等进行设计、维护; ii) 解析数据调用语句, 转换为数据库查询语句获取目标数据; iii) 对从数据库中获取的数据进行预处理, 转化为应用/服务可直接进行处理或应用的数据; iv) 对数据库中的数据进行维护.

以新开发家谱编修应用为例, 若未采用 Huapu-CP 中的数据接口, 系统界面中每处数据读写操作都需编写数据库读写以及权限管理的代码, 代码量较大、且容易导致权限管理混乱、数据泄露等数据安全和隐私问题; 并且, 当出现数据模型更新、数据库更改、权限重新设计等情况时, 需要修改每处的数据库读写代码以及权限管理代码, 系统维护代价极大、开发周期较长.

Huapu-CP 中数据接口可解决上述问题实现应

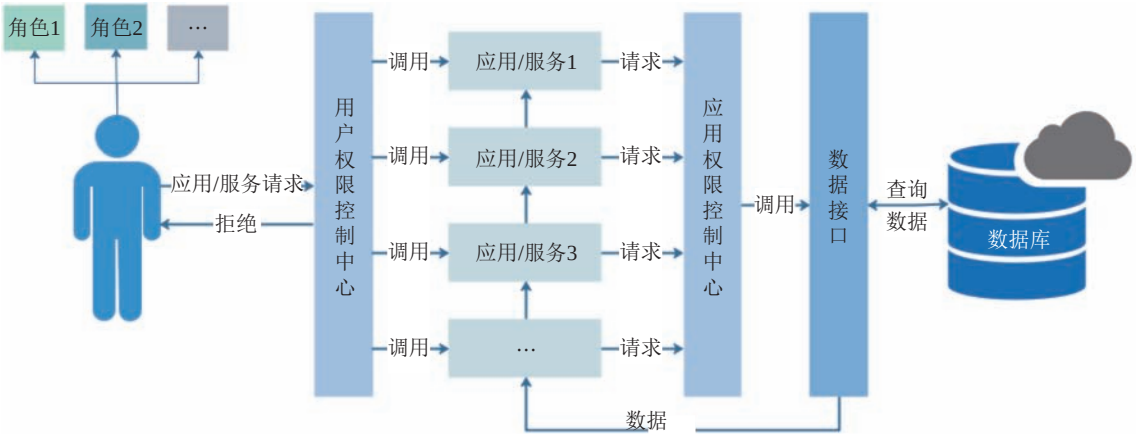


图 12 数据接口示意图

Fig.12 Date interface diagram

用的敏捷开发, 系统中新应用的具体开发流程为: i) 权限设计环节: 由组织智能 (OI) 设计该应用读取/修改数据的最大权限集合、并基于第 2.6 节的权限管理架构设计粗细粒度结合的用户权限管理机制, 并由人类智能 (HI) 审核后分别在应用权限控制中心和用户权限控制中心给予相应接口; ii) 应用开发过程: 该应用的所有数据读取操作直接调用数据接口, 无需编写数据库读写和权限限制相关的代码; iii) 权限闭环反馈过程: 该应用开发完成交于使用后, 基于针对用户行为数据和用户应用使用反馈数据的分析 (AI), 反馈给 OI 和 AI, 修改用户权限控制中心和应用权限控制中心的数据接口, 以升级和维护该应用.

2) “大”家谱知识服务

家谱数据服务是华谱系统数据中台的核心模块, 目前, 我们提供如下数据服务与应用: 华夏修谱, 家谱交互式数据可视化、寻根问祖、家族变迁等, 由于篇幅限制, 本节仅以华夏修谱为例, 介绍从数据录入、校验、打印等详细修谱过程:

a) 家谱编修

华谱系统提供 “我的家谱”和 “共建家谱”功能供用户编修家谱. “我的家谱”面向个人家谱修建功能, 适用于人数较少、数据隐私较高的家谱建设. “共建家谱”提供多人家谱修建功能: 当家谱数据过多时, 数据校验工作量较大、较复杂时, 共建家谱能够供给多人修建同一份家谱, 并通过权限设置限制其他家谱共建人对此数据的修改、删除等操作. 而且, 能够将 “我的家谱”中数据合并至 “共建家谱”中, 为同宗共源家谱修建提供便利.

数据录入时, 提供多样、定制化数据录入方式如单个或批量数据导入; 单个数据录入时提供文本解析功能实现自动解析文字中包含的人物信息, 批量导入时用户可利用 Excel 表、家谱登记表等实现批量数据保存.

而且, 华谱系统中能够实现特殊人物关系保存: 如隔代相连人物关系和待考人物关系, 隔代相连人物关系是指两位人物间已知某一人物是另一人物的直系后辈, 且相隔辈分已知, 提供隔代相连关系, 避免无价值信息保存, 先祖称为 “隔代相连先祖”, 后辈称为 “隔代相连后辈”, 并作 “隔代”标记. 相应的, 待考关系是指某一人物姓名未知, 其他信息已知, 且与另一已知人物存在未知关系, 用户能够确认是另一人物后辈, 系统提供待考关系, 先祖称为 “待考先祖”, 后辈称为 “待考后辈”, 并作 “待考”标记.

家谱编修过程中需不断对家谱人物信息和关系不断校对和完善, 特别是多人编修的 “共建家谱”中存在数据重复录入、人物关系未建立或建立错误等, 华谱系统提供孤立家谱子树检测、同名人物识别、人物合并、人物关联路径分析等服务实现家谱编修的完整性和正确性.

b) 家文化传承

家文化包含姓氏源流、世系表、家训、家传和家谱图像等重要信息, 为寻根问祖、跨姓分析等服务提供史料支撑, 是家谱编修过程中应重点记录的内容; 华谱系统中 “家文化”模块将家谱文化全部保存下来.

c) 家谱印刷

在共建家谱中, 华谱系统提供交互式家谱分卷功能, 首先根据家谱数据特点提供自动家谱分卷, 然后用户根据个性化需求调整家谱分卷逻辑, 分卷结果便于数据校对或保存.

家谱印刷时, 华谱系统实现个性化数据导出服务, 如指定打印先祖、家谱先祖、打印截止辈分、至指定后代等, 目前, 系统中主要提供家谱印刷格式包含:

i) 谱系图: 将家谱树拆分成多棵三层的家谱子树, 便于理清前后辈关系, 特别是当数据较多时, 此方法展示的家谱树信息更加清晰.

ii) 家谱: “段落式/表格式”家谱人物信息处理展示.

3) 用户中心

用户中心是将家谱人物查询、家谱人物分析、家谱行为分析和系统推荐服务集于一体的用户个性化中心, 其中: 家谱人物查询查询提供了多个查询标识 (keys), 包括人物编码、人物姓名、年龄、字号等; 家谱人物分析服务通过大数据分析组件, 对人物数据进行分析, 形成同名人物关联、跨姓氏人物打通等 API 供上层调用; 用户行为分析通过对系统采集的用户数据进行分析, 为用户提供更加优质的家谱数据服务, 如系统界面优化、功能优化等; 此外用户行为数据还可为用户自画像、用户地图、用户积分等功能提供更为准确的数据支撑, 为系统优化、家谱数据价值提升提供更广泛的扩展空间.

推荐服务基于用户的基本信息、操作习惯, 家谱人物的基本信息、从属关系、价值属性和查询、修改、分享、聊天等内容形成群组推荐、家谱推荐等 API 服务.

4) “智能”社交数据服务

华谱系统提供社交服务模块, 包含您可能认识的人、群组推荐、话题推荐和网络祭祀等.

5) 其他数据服务

华谱系统是面向全体华人的家谱数据服务系统, 目前已积累较为广泛的数据来源, 为海内外为华人寻根问祖等提供数据和工具支撑.

3 数据中台的挑战和前景

面向海量、多源、异构、碎片化的家谱数据, 我们提出并实现了一个结合 HAO 智能的家谱数据中台架构 Huapu-CP, 对数据中台的开发和实践做了一个成功的案例. 数据中台的研究和开发目前还面临以下挑战.

挑战 1. 数据安全和隐私保护问题

数据中台的目标之一是实现广泛的数据共享, 盲目地打通数据之间的关联将不可避免地带来数据安全与隐私保护的问题. 目前数据安全和隐私保护主要是从数据的层面去管理, 可能会导致数据安全或者有隐私漏洞, 同时, 严格的数据管理也会阻碍数据价值的产生. 因而, 如何从业务场景出发建设数据的安全和隐私保护体系是未来的研究的重点之一.

挑战 2. 无统一的数据中台模板

为实现企业的数字化转型, 企业构建数据中台系统满足迅速变化的应用需求, 解决数据开发和应用开发速度不匹配的问题. 但是, 数据需求千人千面, 企业数据应用也是不断更新迭代, 企业的中台系统也需要不断变化, 因而, 无法创建统一、规范的数据中台模板供其他企业借鉴与使用, 企业需根据自身业务的需要, 构建适合于本企业发展的数据中台.

挑战 3. 数据中台是不断建设不断完善的过程

数据中台建设不是一朝一夕能够完成的, 需要业务人员、数据分析人员和管理人员等具备高效的沟通和更广泛的业务视角, 紧随企业核心业务的变化迅速调整中台建设规划, 通过数据应用的实践, 评估数据中台的建设效果进而不断反馈修正.

挑战 4. 缺乏丰富经验的技术团队和成熟的检验产品或工具

数据中台建设团队涉及数据管理、数据开发、数据服务等团队之间的协作, 单一的开发团队无法完成数据中台的所有任务, 且对尚未有成熟的对数据中台建设效果进行检验的产品或工具, 中台建设的优劣短期内可能仍旧停留在数据应用的效果这个单一评价指标上. 本文第 1.2 节提出的 7 个数据中台核心功能为数据中台发展提供了广泛的前景, 主要包含以下几个方面:

前景 1. 实现机构数据资产的高效管理和数据价值最大化: 在爆炸式数据增长的时代, 海量数据的存储、管理和价值的实现是企业面对的一个主要问题, 数据中台可望能够有效地盘活机构数据资源, 将其转化为数据资产, 通过更贴近业务数据服务 API 实现数据价值的最大化.

前景 2. 能够迅速根据时代变化调整机构的发展方向和快速创新相应用户需求: 数据中台是一个完整的数据服务体系, 为机构带来了数据平台化的运营机制, 可望解决应用开发与数据开发速度不匹配的问题, 因而, 数据中台为一个机构根据时代发展要求调整机构的战略提供了契机.

前景 3. 提升机构内团队协作能力: 原始机构的业务各自发展, 可能导致出现烟囱式应用开发和数据孤岛等问题, 数据中台的出现, 可以将机构的核心技术或团队凝聚在一起, 建设机构内强大的数据开发、运营等团队, 提升机构的团队的硬实力和软实力.

4 总结

本文以 HAO 智能为技术支撑, 结合知识图谱技术等提出并实现了一个家谱数据中台 Huapu-CP. 本文首先并给出了数据中台的定义和相应的 7 项核心功能, 然后以华谱数据中台构建的架构为例, 详细

地介绍了家谱数据中台中包含的模块和使用的相关技术.

目前, 关于数据中台建设尚处于起步阶段, 面临着技术不成熟、框架验证标准不一、技术人员缺乏等困难和挑战. 本文的 Huapu-CP 建设经验, 如果用来构建其他领域的数据中台, 需要根据数据特点和应用需求调整各个模块的具体实现. 例如, 在物流领域的数据中台建设中, 面对数据维度高、数据类型复杂、数据量大、实时数据采集困难等问题, 需要有高性能的数据分析和计算平台, 会给现有数据中台建设的技术带来很大的挑战. 此外, 新一代的数据中台技术, 在融合数据的基础上, 更需要关心是否能够很好地沉淀行业知识. 知识图谱技术相对于传统的二维表使用图描述实体与关系. 这种复杂的图结构更有利于探索数据之间的关联关系、获取知识. 我们将继续研究如何优化本文提出 Huapu-CP 数据中台框架, 使其不仅适用于家谱数据领域, 力求能够扩展到其他领域的数据中台建设中, 为数据应用服务提供更好的 API 支撑.

致谢

华谱系统是由合肥工业大学大知识研究院的教师和研究生们集体建设的, 除了本文的作者, 建设系统的主要人员还包括: 朱毅博士、周鹏博士、张启平博士、嵇圣硙、刘古刘、刘啸剑、李姣、董丙冰、洪炎、钟凌锋、赵海霞、邵健轩等.

References

1 Chen Ning-Ning. History and current status of genealogical research. *Library Journal*, 1998, **17**(2): 12−1318
(陈宁宁. 家谱研究历史现状. 图书馆杂志, 1998, **17**(2): 12−1318)

2 Carolina N, Nils G, Hilary C, and Alexander L. Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs. *IEEE Transactions on Visualization and Computer Graphics*, 2019, **25**(1): 544−554

3 Hayden E C. Colossal family tree reveals environment's influence on lifespan. Nature, 2018.

4 Zhan Lu. Literature questions in genealogy. *Journal of Peking University (Philosophy and Social Sciences)*, 2007, (1): 150−151
(湛庐. 家谱中的文献问题. 北京大学学报 (哲学社会科学版), 2007, (1): 150−151)

5 Ou Yang-Kang. The reform and innovation of big data and humanities and social science research. Guangming Daily, 2016-11-10(016).
(欧阳康. 大数据与人文社会科学研究的变革与创新. 光明日报, 2016-11-10(016).)

6 Sun Jian-Jun. How to develop humanities and social sciences in the age of big data. Guangming Daily, 2014-07-07(011).
(孙建军. 大数据时代人文社会科学如何发展. 光明日报, 2014-07-07(011).)

7 Wu X D, Chen H H, Wu G Q, et al. Knowledge engineering with big data. *IEEE Intelligent Systems*, 2015, **30**(5): 46−55

8 Wu X D, Zhu X Q, Wu G Q, Ding W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*. 2014, **36**(1): 97−107.

9 Wu Xin-Dong, He Jin, Lu Ru-Qian, Zheng Nan-Ning. From big data to big knowledge: HACE + BigKE. *Acta Automatica Sinica*, 2016, **42**(7): 965−982
(吴信东, 何进, 陆汝钤, 郑南宁. 从大数据到大知识: HACE+BigKE. 自动化学报, 2016, **42**(7): 965−982)

10 Wu M H, Wu X D. On big wisdom. *Knowledge and Information Systems*, 2018, **58**(1): 1−8

11 Zhong Hua. The Transformation of IT Framework in Enterprises: the Strategic Thinking and Framework of Alibaba. Beijing: China Machine Press, 2017.
(钟华. 企业 IT 架构转型之道: 阿里巴巴中台战略思想与架构实战. 北京: 机械工业出版社, 2017.)

12 Fu Deng-Po, Jiang Min, Ren Yin-Zi, etc. *Data Middle Office: Make Data Valuable*. Beijing: China Machine Press, 2020.
(付登坡, 江敏, 任寅姿等. 数据中台: 让数据用起来. 北京: 机械工业出版社, 2020.)

13 Chen Xin-Yu, Luo Jia-Ying, Deng Tong, Jiang Wei. *Middle-Platform Strategy: Middle-Platform Construction and Digital Commerce*. Beijing: China Machine Press, 2019.
(陈新宇, 罗家鹰, 邓通, 江威. 中台战略: 中台建设与数字商业. 北京: 机械工业出版社, 2019.)

14 Wu Xin-Dong, Dong Bing-Bing, Du Xin-Zheng, Yang Wei. Data governance technology. *Ruan Jian Xue Bao/Journal of Software*, 2019, **30**(9): 2830−2856
(吴信东, 董丙冰, 堵新政, 杨威. 数据治理技术. 软件学报, 2019, **30**(9): 2830−2856)

15 Wu Xin-Dong, Ji Sheng-Wei. Comparative study on MapReduce and Spark for big data analytics. *Ruan Jian Xue Bao/Journal of Software*, 2018, **29**(6): 260−281
(吴信东, 嵇圣硙. MapReduce 与 Spark 用于大数据分析之比较. 软件学报, 2018, **29**(6): 260−281)

16 Ji S W, Bu C Y, Li L, W XD. Local graph edge partitioning with a two-Stage heuristic method. In: Proceedings of the 39th IEEE International Conference on Distributed Computing Systems. Dallas, Texas, USA. IEEE, 2019.

17 Wu Gong-Qing, Hu Jun, Li Li, Xu Zhe-Hao, Liu Peng-Cheng, Hu Xue-Gang, Wu Xin-Dong. Online web news extraction via tag path feature fusion. *Ruan Jian Xue Bao/Journal of Software*, 2016, **27**(3): 714−735
(吴共庆, 胡骏, 李莉, 徐喆昊, 刘鹏程, 胡学钢, 吴信东. 基于标签路径特征融合的在线 Web 新闻内容抽取. 软件学报, 2016, **27**(3): 714−735)

18 Wu Xin-Dong, Li Jiao, Zhou Peng, Bu Chen-Yang. A fusion technique for fragmented genealogy data. *Ruan Jian Xue Bao/Journal of Software*, 2020 http://www.jos.org.cn/1000-9825/6010.htm.
(吴信东, 李姣, 周鹏, 卜晨阳. 碎片化家谱数据的融合技术. 软件学报. http://www.jos.org.cn/1000-9825/6010.htm.)

19 Liu X J, Zhu Y, Ji S W. Web log analysis in genealogy system. In: Proceedings of the 11th IEEE International Conference on Knowledge Graph. Nanjing, China. IEEE, 2020.

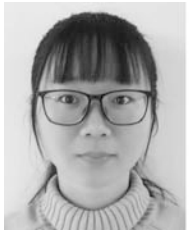
吴信东 合肥工业大学特聘教授, IEEE Fellow, AAAS Fellow. 明略科技集团首席科学家、高级副总裁和明略科学院院长, 营销职能国家新一代人工智能开放创新平台负责人. 主要研究方向为数据挖掘, 大数据分析. 知识工程. 本文通信作者.

E-mail: xwu@hfut.edu.cn

(**WU Xin-Dong** Distinguished professor of the Research

Institute of Big Knowledge, Hefei University of Technology. IEEE Fellow and AAAS Fellow. Hefei University of Technology. He is currently the Chief Scientist of Mininglamp Technology and President of Mininglamp Academy of Sciences. His research interest covers data mining, big data analytics, and knowledge engineering. Corresponding author of this paper.)

盛绍静 合肥工业大学计算机与信息学院博士研究生. 主要研究方向为数据挖掘, 知识图谱.
E-mail: jssheng@mail.hfut.edu.cn
(**SHENG Shao-Jing** Ph. D. candidate at the School of Computer Science and Information Engineering, Hefei University of Technology. Her research interest covers data mining and knowledge graph.)



蒋婷婷 合肥工业大学博士生. 主要研究方向为知识图谱, 知识表示学习, 实体对齐.
E-mail: jiangtt@mail.hfut.edu.cn
(**JIANG Ting-Ting** Ph. D. candidate at the School of Computer Science and Information Engineering, Hefei University of Technology. Her research interest covers knowledge graph, knowledge graph embedding, and entity alignment.)

卜晨阳 合肥工业大学讲师. 2017 年获得中国科学技术大学博士学位. 主要研究方向为演化计算及其在知识图谱、教育数据挖掘、电力系统等领域中的应用.
E-mail: chenyangbu@hfut.edu.cn
(**BU Chen-Yang** Lecturer at the School of Computer Science and Information Engineering, Hefei University of Technology. He received his Ph. D. degree of computer science from the University of Science and Technology of China (USTC) in 2017. His research interest covers evolutionary algorithms and their applications in areas such as knowledge graphs, educational data mining, and power systems.)

吴明辉 明略科技集团创始人兼首席执行官. 北京大学数学系学士学位、计算机软件与理论硕士学位, 中国公安大学特聘教授. 主要研究方向为人工智能和大数据分析.
E-mail: wuminghui@minginglamp.com
(**WU Ming-Hui** Founder and CEO of Mininglamp Technology. He received his bachelor degree in mathematics, master degree in computer software and theory from Peking University. Distinguished professor at the Public Security University of China. His research interest covers artificial intelligence and big data analytics.)