软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(9):2830−2856 [doi: 10.13328/j.cnki.jos.005854]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

数据治理技术*

吴信东[1,2,3,4], 董丙冰[2,3,4], 堵新政[1], 杨 威[1]

[1](明略科技集团,北京 100084)

[2](合肥工业大学 大知识科学研究院,安徽 合肥 230009)

[3](大数据知识工程教育部重点实验室(合肥工业大学),安徽 合肥 230009)

[4](合肥工业大学 计算机与信息学院,安徽 合肥 230601)

通讯作者: 吴信东, E-mail: wuxindong@mininglamp.com

摘 要: 随着信息技术的普及,人类产生的数据量正在以指数级的速度增长,如此海量的数据就要求利用新的方法来管理.数据治理是将一个机构(企业或政府部门)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,最终实现数据价值最大化.目前,各行各业对大数据的研究比较火热,但对于大数据治理的研究还处于起步阶段,一个组织的正确决策离不开良好的数据治理.首先介绍数据治理和大数据治理的概念、发展以及应用的必要性;其次,对已有的数据治理技术——数据规范、数据清洗、数据交换和数据集成进行具体的分析,并介绍了数据治理成熟度和数据治理框架设计;在此基础上,提出了大数据 HAO 治理模型.该模型以支持人类智能(HI)、人工智能(AI)和组织智能(OI)的三者协同为目标,再以公安的数据治理为例介绍 HAO 治理的应用;最后是对数据治理的总结和展望.

关键词: 数据治理;数据规范;数据清洗;数据交换;数据集成

中图法分类号: TP311

中文引用格式: 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. http://www.jos.org.cn/1000-9825/5854.htm

英文引用格式: Wu XD, Dong BB, Du XZ, Yang W. Data governance technology. Ruan Jian Xue Bao/Journal of Software, 2019, 30(9):2830−2856 (in Chinese). http://www.jos.org.cn/1000-9825/5854.htm

Data Governance Technology

WU Xin-Dong[1,2,3,4], DONG Bing-Bing[2,3,4], DU Xin-Zheng[1], YANG Wei[1]

[1](Mininglamp Technology, Beijing 100084, China)

[2](Research Institute of Big Knowledge, Hefei University of Technology, Hefei 230009, China)

[3](Key Laboratory of Knowledge Engineering with Big Data(Hefei University of Technology), Hefei 230009, China)

[4](School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China)

Abstract: Along with the pervasiveness of information technology, the amount of data generated by human beings is growing at an exponential rate. Such massive data requires management with new methodologies. Data governance is the management of data for an organization (enterprise or government) as a strategic asset, from the collection of data to a set of management mechanisms for processing and applications, aiming to improve data quality, achieve a wide range of data sharing, and ultimately maximize the data value. Research

* 基金项目: 国家重点研发计划(2016YFB1000901); 国家自然科学基金(91746209); 教育部创新团队项目(IRT17R3)

Foundation item: National Key Researh and Development Program of China (2016YFB1000901); National Natural Science Foundation of China (91746209); Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education (IRT17R3)

收稿时间: 2018-12-25; 修改时间: 2019-03-11; 采用时间: 2019-04-22; jos 在线出版时间: 2019-05-22

CNKI 网络优先出版: 2019-05-22 15:26:23, http://kns.cnki.net/kcms/detail/11.2560.TP.20190522.1525.012.html

and development on big data is nowadays popular in various domains, but big data governance is still in its infancy, and the decision-making of an organization cannot be separated from excellent data governance. This paper first introduces the concepts, developments, and necessity of data governance and big data governance, then analyzes existing data governance technologies—data specification, data cleaning, data exchange, and data integration, and also discusses the maturity measurement and framework design of data governance. Based on these introductions, analyses and reviews, the paper puts forward a "HAO governance" model for big data governance, which aims to facilitate HAO Intelligence with human intelligence (HI), artificial intelligence (AI), and organizational intelligence (OI), and then instantiates the "HAO governance" model with public security data governance as an example. Finally, the paper summarizes data governance with its challenges and opportunities.

Key words: data governance; data specification; data cleaning; data exchange; data integration

随着信息技术的迅速发展,数据规模逐渐扩大.与此同时,劣质数据也随之而来,极大地降低了数据挖掘的质量,对信息社会造成了严重的困扰[1].劣质数据大量存在于很多领域和机构,国外权威机构的统计表明:美国的企业信息系统中,1%~30%的数据具有各种错误和误差[2];13.6%~81%的关键数据不完整或陈旧情况存在于美国的医疗信息系统中[3].根据 Gartner 的调查结果:在全球财富 1 000 强的企业中,超过 25%的企业信息系统中存在错误数据[4].

大多数组织不考虑数据质量对大数据平台建设、分析应用等方面的重要影响而盲目投入,缺乏对大数据资源的整体规划和综合治理,最终导致一些项目实施的终止和失败.项目的失败和数据量的激增,使得数据治理的重要性逐步得到工业界和学术界的共识.随着国家政策支持以及产业实际需求的增长,如何通过数据治理提升组织数据管理能力、消除数据孤岛、挖掘数据潜在的价值,将成为重点发展领域.

数据治理的重要前提是建设统一共享的数据平台,信息系统的建设发展到一定阶段,数据资源将成为战略资产,而有效的数据治理才是数据资产形成的必要条件.同时,在数据共享的时代,享受大数据带来便利的同时,也带来如个人隐私泄露的问题[5].个人隐私信息泄露事件频繁发生,使得人们更加注重保护个人的隐私信息,通常采取一些措施,如在进行网站注册时故意填写虚假信息,这将会严重的影响数据的质量和完整性,低质量的数据将导致低质量的挖掘结果.数据治理不仅要规范数据,实现数据的价值和管控风险,还要做到隐私保护.

本文首先对数据治理做全面的分析,然后介绍我们自己设计的大数据治理模型.本文第 1 节介绍数据治理的定义和发展趋势.第 2 节介绍数据规范技术的内涵以及应用方法.第 3 节介绍数据清洗的背景以及清洗的基本方法.第 4 节对数据交换的基本概念及其实现模式进行阐述.第 5 节介绍数据集成技术的基本概念和数据集成的方法,并说明这些方法的应用场景.第 6 节从数据治理的成熟度模型开始,引出数据治理框架.第 7 节对我们提出的HAO治理模型进行详细说明.第8节以公安数据治理为例,具体介绍治理模型的具体应用.最后是对数据治理技术的总结与展望.

1 数据治理的研究现状

1.1 数据治理的定义

至今为止,数据治理还没有统一标准的定义.IBM 对于数据治理的定义是,数据治理是一种质量控制规程,用于在管理、使用、改进和保护组织信息的过程中添加新的严谨性和纪律性[6].DGI 则认为,数据治理是指在企业数据管理中分配决策权和相关职责[6].

数据治理的目标,总体来说就是提高数据质量,在降低企业风险的同时,实现数据资产价值的最大化,包括:

- 构筑适配灵活、标准化、模块化的多源异构数据资源接入体系;
- 建设规范化、流程化、智能化的数据处理体系;
- 打造数据精细化治理体系、组织的数据资源融合分类体系;
- 构建统一调度、精准服务、安全可用的信息共享服务体系.

其次,我们还需理解数据治理的职能——数据治理提供了将数据作为资产进行管理所需的指导.最后,我们要把握数据治理的核心——数据资产管理的决策权分配和指责分工[7].

由此,数据治理从本质上看就是对一个机构(企业或政府部门)的数据从收集融合到分析管理和利用进行评估、指导和监督(EDM)的过程,通过提供不断创新的数据服务,为企业创造价值[6].

数据治理与数据管理是两个十分容易混淆的概念,治理和管理从本质上看是两个完全不同的活动,但是存在一定的联系,下面我们对这两个概念进行详细的解读.

COBIT5(control objectives for information and related technology)对管理的定义:管理是按照治理机构设定的方向开展计划、建设、运营和监控活动来实现企业目标[6].所以,治理过程是对管理活动的评估、指导和监督,而管理过程是对治理决策的计划、建设和运营.具体分析:首先,数据治理与数据管理包含不同的活动即职能,数据治理包括评估指导和监督,数据管理包括计划建设和运营;其次,数据治理是回答企业决策的相关问题并制定数据规范,而数据管理是实现数据治理提出的决策并给予反馈;最后,数据治理和数据管理的责任主体也是不同的,前者是董事会,后者是管理层.

1.2 大数据治理——数据治理新趋势

近年来,大数据已成为国内外专家学者研究的热点话题,目前基本上采用 IBM 的 5V 模型描述大数据的特征:第 1 个 V(volume)是数据量大,包括采集、存储和计算的量都非常大;第 2 个 V(velocity)是数据增长速度快,处理速度也快,时效性要求高;第 3 个 V(variety)是种类和来源多样化,包括结构化、半结构化和非结构化数据;第 4 个 V(value)是数据价值密度相对较低,可以说是浪里淘沙却又弥足珍贵;第五个 V(veracity)是各个数据源的质量良莠不齐,需要精心甄别[8].随着数据量的激增,可以用“5V+I/O”——体量、速度、多样性、数据价值和质量以及数据在线来概括其特征.这里的“I/O”是指数据永远在线,可以随时调用和计算,这个特征是大数据与传统数据最大的区别.

2014 年,吴信东等人基于大数据具有异构、自治的数据源以及复杂和演变的数据关联等本质特征,提出了 HACE 定理[9].该定理从大数据的数据处理、领域应用及数据挖掘这 3 个层次(如图 1 所示)来刻画大数据处理框架[8].

框架的第 1 层是大数据计算平台,该层面临的挑战集中在数据存取和算法计算过程上;第 2 层是面向大数据应用的语义和领域知识,该层的挑战主要包括信息共享和数据隐私、领域和应用知识这两个方面;架构的第 3 层集中在数据挖掘和机器学习算法设计上:稀疏不确定和不完整的数据挖掘、挖掘复杂动态的数据以及局部学习和模型融合[9].第 3 层的 3 类算法对应 3 个阶段:首先,通过数据融合技术对稀疏、异构、不确定、不完整和多源数据进行预处理;其次,在预处理之后,挖掘复杂和动态的数据;最后,通过局部学习和模型融合获得的全局知识进行测试,并将相关信息反馈到预处理阶段,预处理阶段根据反馈调整模型和参数[9].

Fig.1 A big data processing framework[9]

图 1 大数据处理框架[9]

面对大数据兴起带来的挑战,为了促进大数据治理的发展和变革,目前业界比较权威的大数据治理定义是:大数据治理是广义信息治理计划的一部分,它通过协调多个职能部门的目标,来制定与大数据优化、隐私与货

币化相关的策略[10].此定义指出:大数据的优化、隐私保护以及商业价值是大数据治理的重点关注领域,大数据治理是数据治理发展的一个新阶段,与数据治理相比,各种需求的解决在大数据治理中变得更加重要和富有挑战性[6].

- 海量数据存储:根据本地实际数据量级和存储处理能力,结合集中式或分布式等数据资源的存储方式进行构建,为大数据平台提供 PB 级数据的存储及备份能力支撑.云计算[11,12]作为一种新型的商业模式,它所提供的存储服务具有专业、经济和按需分配的特点,可以满足大数据的存储需求;
- 处理效率:大数据治理提供多样化的海量数据接入及处理能力,包括对各类批量、实时、准实时及流式的结构化、非结构化数据提供快速的计算能力和搜索能力,比如数据加载能力≥130MB/s、亿级数据秒级检索、百亿数据实时分析≤10s、千亿数据离线分析≤30m 等等.对于大数据的搜索能力方面,为了保证数据安全,大数据在云计算平台上的存储方式一般为密文存储,因此,研究人员设计了很多保护隐私的密文搜索算法[13−22],基于存储在云平台上大数据的计算安全问题的解决方法一般采用比较成熟的完全同态加密算法[23−29];
- 数据可靠性:围绕行业数据元相关标准规定,基于行业元数据体系打造大数据平台采集汇聚、加工整合、共享服务等全过程的、端到端的数据质量稽核管控体系,确保数据准确可靠;
- 数据安全性:数据价值是大数据平台的核心价值,所以数据的安全是保证平台运行的基础.数据安全包括数据存储的安全、数据传输过程中的安全,数据的一致性、数据访问安全等,如图 2 所示.数据安全的总体目标是保证数据的存储、传输、访问、展示和导出安全.数据安全措施主要有数据脱敏控制[30]、数据加密控制、防拷贝管理、防泄漏管理、数据权限管理、数据安全等级管理等.

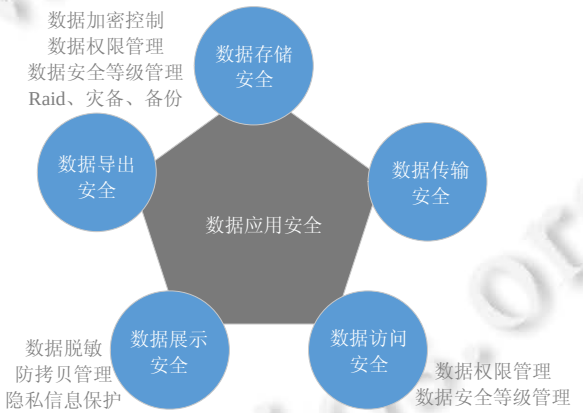

Fig.2 Data application security schematic

图 2 数据应用安全示意图

而数据治理技术就是在数据治理的过程中所用到的技术工具,其中主要包括数据规范、数据清洗、数据交换和数据集成这 4 种技术,下面具体介绍这 4 种技术.

2 数据规范

2.1 数据规范的含义

数据治理的处理对象是海量分布在各个系统中的数据,这些不同系统的数据往往存在一定的差异:数据代码标准、数据格式、数据标识都不一样,甚至可能存在错误的数据.这就需要建立一套标准化的体系,对这些存在差异的数据统一标准,符合行业的规范,使得在同样的指标下进行分析,保证数据分析结果的可靠性.例如,对于数据库的属性值而言,可以建立唯一性规则、连续性规则以及空值规则等来对数据进行检验和约束:唯一性规则一般是指为主键或其他属性填写 unique 约束,使得给定属性的每个值与该属性的其他值不同;连续性规则

是指属性的最大值和最小值之间没有缺失值并且每个值也是唯一的,一般用于检验数;空值规则是指使用其他特殊符号来代替空值,以及对于这样的值应该如何处理.

数据的规范化能够提高数据的通用性、共享性、可移植性及数据分析的可靠性.所以,在建立数据规范时要具有通用性,遵循行业的或者国家的标准.

2.2 数据规范方法

数据治理过程中可使用的数据规范方法有:规则处理引擎、标准代码库映射.

(1) 规则处理引擎

数据治理为每个数据项制定相关联的数据元标准,并为每个标准数据元定义一定的处理规则,这些处理逻辑包括数据转换、数据校验、数据拼接赋值等.基于机器学习等技术,对数据字段进行认知和识别,通过数据自动对标技术,解决在数据处理过程中遇到的数据不规范的问题.

- 根据数据项标准定义规则模板,图 3 中“出生日期”的规则如下所示。
  - 值域稽核规则:YYYY:MM:DD 或 YYYY-MM-DD;
  - 取值范围规则:1900<YYYY<=2018,1<=MM<=12,1<=DD<=31.
- 将数据项与标准库数据项对应。

借助机器学习推荐来简化人工操作,根据语义相似度和采样值域测试,推荐相似度最高的数据项关联数据表字段,并根据数据特点选择适合的转换规则进行自动标准化测试.根据数据项的规则模板自动生成字段的稽核任务.

规则体系中包含很多数据处理的逻辑:将不同数据来源中各种时间格式的数据项,转化成统一的时间戳(timestamp)格式;对数据项做加密或者哈希转换;对身份证号做校验,检验是否为合法的 18 位身份证号,如果是 15 位的,则将其统一转换成 18 位;将多个数据项通过指定拼接符号,连接成一个数据项;将某个常量或者变量值赋给某个数据项等.

规则库中的规则可以多层级迭代,形成数据处理的一条规则链.规则链上,上一条规则的输出作为下一条规则的输入,通过规则的组合,能够灵活地支持各种数据处理逻辑.例如:对身份证号先使用全角转半角的规则,对输出的半角值使用身份证校验转换规则,统一成 18 位的身份证号;再对 18 位身份证号使用数据脱敏规则,将身份证号转成脱敏后的字符串.

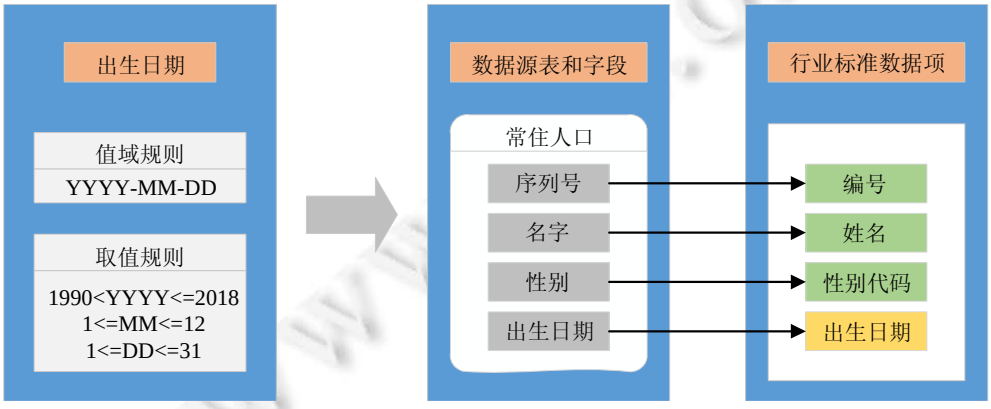

Fig.3 Rule processing schematic

图 3 规则处理示意图

(2) 标准代码库映射

标准代码库是基于国标或者通用的规范建立的 key-value 字典库,字典库遵循国标值域、公安装备资产分类与代码等标准进行构建.当数据项的命名为 XXXDM(XXX 代码)时,根据字典库的国标或部标代码,通过字典

规则关联出与代码数据项对应的代码名称数据项 XXXDMMC(XXX 代码名称).

例如,我们想要将所有表示性别“男”的字段都转换成“男”这种同一的表示方式,可以先建立一个数据字典,其中的键的取值范围是所有不同表示方式的集合,值为最终我们想要归一化表示的“男”.

{
“男” => “男”,
“男性” => “男”,
“male” => “男”,
“man” => “男”,
“1” => ”男”
…
}

使用数据转换规则时查找数据字典,将所有不同的表示方式统一成一种表示方式.

3 数据清洗

3.1 数据清洗背景

数据质量一般由准确性、完整性、一致性、时效性、可信性以及可解释性等特征来描述,根据 Rahm 等人在 2000 年对数据质量基于单数据源还是多数据源以及问题出在模式层还是实例层的标准进行分类,将数据质量问题分为单数据源模式层问题、单数据源实例层问题、多数据源模式层问题和多数据源实例层问题这 4 大类[31].现实生活中的数据极易受到噪声、缺失值和不一致数据的侵扰,数据集成可能也会产生数据不一致的情况,数据清洗就是识别并且(可能)修复这些“脏数据”的过程[32].如果一个数据库数据规范工作做得好,会给数据清洗工作减少许多麻烦.对于数据清洗工作的研究基本上是基于相似重复记录的识别与剔除方法展开的,并且以召回率和准确率作为算法的评价指标[33,34].现有的清洗技术大都是孤立使用的,不同的清洗算法作为黑盒子以顺序执行或以交错方式执行,而这种方法没有考虑不同清洗类型规则之间的交互简化了问题的复杂性,但这种简化可能会影响最终修复的质量,因此需要把数据清洗放在上下文中结合端到端质量执行机制进行整体清洗[35].随着大数据时代的到来,现在已经有不少有关大数据清洗系统的研究[36,37],不仅有对于数据一致性[38−40]以及实体匹配[41]的研究,也有基于 MapReduce 的数据清洗系统的优化[42]研究.下面对数据清洗具体应用技术以及相关算法进行分析.

3.2 数据清洗基本方法

从微观层面来看,数据清洗的对象分为模式层数据清洗和实例层数据清洗[43].数据清洗识别并修复的“脏数据”主要有错误数据、不完整的数据以及相似重复的数据,根据“脏数据”分类,数据清洗也可以分为 3 类:属性错误清洗、不完整数据清洗以及相似重复记录的清洗,下面分别对每种情况进行具体分析.

3.2.1 属性错误清洗

数据库中很多数据违反最初定义的完整性约束,存在大量不一致的、有冲突的数据和噪声数据,我们应该识别出这些错误数据,然后进行错误清洗.

(1) 属性错误检测

属性错误检测有基于定量的方法和基于定性的方法.

- 定量的误差检测一般在离群点检测的基础上采用统计方法来识别异常行为和误差,离群点检测是找出与其他观察结果偏离太多的点,Aggarwal 将关于离群点检测方法又分为 6 种类型:极值分析、聚类模型、基于距离的模型、基于密度的模型、概率模型、信息理论模型[44],并对这几种模型进行了详尽的介绍;
- 定性的误差检测一般依赖于描述性方法指定一个合法的数据实例的模式或约束,因此确定违反这些

模式或者约束的就是错误数据.

图 4 描述了定性误差检测技术在 3 个不同方面的不同分类,下面我们对图中提出的 3 个问题进行分析.

- 首先,错误类型是指要检测什么.定性误差检测技术可以根据捕捉到的错误类型来进行分类,目前,大量的工作都是使用完整性约束来捕获数据库应该遵守的数据质量规则,虽然重复值也违反了完整性约束,但是重复值的识别与清洗是数据清洗的一个核心(在后续小节将会单独介绍);
- 其次,自动化检测.根据人类的参与与否以及参与步骤来对定性误差检测技术进行分类,大部分的检测过程都是全自动化的,个别技术涉及到人类参与;
- 最后,商业智能层是指在哪里检测.错误可以发生在数据治理的任何阶段,大部分的检测都是针对原始数据库的,但是有些错误只能在数据治理后获得更多的语义和业务逻辑才能检测出来.

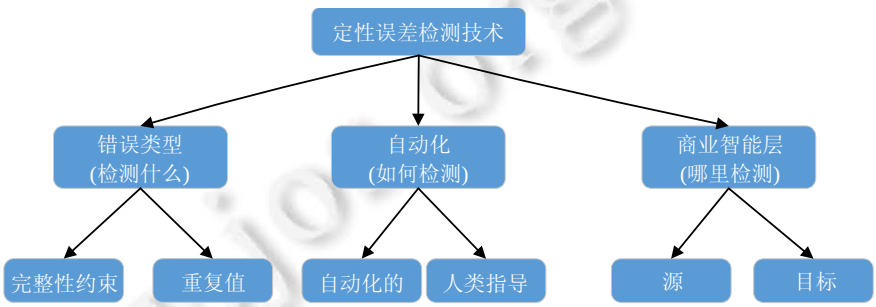

Fig.4 Classification of qualitative error detection techniques[45]

图 4 定性误差检测技术分类[45]

不仅可以使用统计方法来对属性错误进行检测,使用一些商业工具也可以进行异常检测,如数据清洗工具以及数据审计工具等.Potter's Wheel[46]是一种公开的数据清洗工具,不仅支持异常检测,还支持后面数据不一致清洗所用到的数据变换功能.

(2) 属性错误清洗

属性错误清洗包括噪声数据以及不一致的数据清洗.

- 噪声数据的清洗也叫光滑噪声技术,主要方法有分箱以及回归等方法:分箱方法是通过周围邻近的值来光滑有序的数据值但是只是局部光滑,回归方法是使用回归函数拟合数据来光滑噪声;
- 不一致数据的清洗在某些情况下可以参照其他材料使用人工进行修改,可以借助知识工程工具来找到违反限制的数据,例如:如果知道数据的函数依赖关系,通过函数关系修改属性值.但是大部分的不一致情况都需要进行数据变换,即定义一系列的变换纠正数据,也有很多商业工具提供数据变换的功能,例如数据迁移工具和 ETL 工具等,但是这些功能都是有限的.

3.2.2 不完整数据清洗

在实际应用中,数据缺失是一种不可避免的现象[42].有很多情况下会造成数据值的缺失,例如填写某些表格时需要填写配偶信息,那没有结婚的人就无法填写此字段,或者在业务处理的稍后步骤提供值,字段也可能缺失.处理缺失值目前有以下几种方法.

- 忽略元组:一般情况下,当此元组缺少多个属性值时常采用此方法,否则该方法不是很有效.当忽略了此条元组之后,元组内剩下的有值的属性也不能被采用,这些数据可能是有用的;
- 人工填写缺失值:这种方法最大的缺点就是需要大量的时间和人力,数据清理技术需要做到最少的人工干预,并且在数据集很大、缺失很多属性值时,这种方法行不通;
- 全局变量填充缺失值:使用同一个常量来填充属性的缺失值.这种方法虽然使用起来较为简单,但是有时不可靠.例如,用统一的常量"NULL"来填写缺失值,在后续的数据挖掘中,可能会认为它们形成了一个有趣的概念;

- 中心度量填充缺失值:使用属性的中心度量来填充缺失值.中心度量是指数据分布的“中间”值,例如均值或者中位数,数据对称分布使用均值、倾斜分布使用中位数;
- 使用最可能的值填充:相当于数值预测的概念.回归分析是数值预测最常用的统计学方法,此外也可以使用贝叶斯形式化方法的基于推理的工具或决策树归纳确定缺失值.

鉴于现在很多人为了保护自己的隐私或者为了方便,随意地选择窗口中给定的值,Hua 等人于 2007 年提出了一种识别伪装缺失数据的启发式方法,当用户不愿意泄露个人信息时故意错误地选择窗口上的默认值(如生日字段),这时数据就会被捕获[47].

3.2.3 相似重复记录清洗

- 相似重复记录识别

消除相似重复记录,首先应该识别出相同或不同数据集中的两个实体是否指向同一实体,这个过程也叫实体对齐或实体匹配.文本相似度度量是实体对齐的最基础方法,大致分为 4 种:基于字符的(例如编辑距离、仿射间隙距离、Smith-Waterman 距离、Jaro 距离度量、Q-gram 距离[48])、基于单词的(例如 Jaccard 系数)、混合型(例如 softTF-IDF)和基于语义的(例如 WordNet).随着知识表示学习在各个领域的发展,一些研究人员提出了基于表示学习的实体匹配算法,但均是以 TransE 系列模型为基础构建的.TransE[49]首次提出基于翻译的方法,将关系解释为实体的低维向量之间的翻译操作,随之涌现出一些扩展的典型算法,下面对这些算法进行简单介绍.

a) MTransE 算法[50]:基于转移的方法解决多语言知识图谱中的实体对齐.首先,使用 TransE 对单个的知识图谱进行表示学习;接着,学习不同空间的线性变换来进行实体对齐.转移方法有基于距离的轴校准、翻译向量、线性变换这 3 种.该知识模型简单复用 TransE,对于提高实体对齐的精度仍存在很大局限;
b) JAPE 算法[51]是针对跨语言实体对齐的联合属性保护模型,利用属性及文字描述信息来增强实体表示学习,分为结构表示、属性表示.IPTransE 算法[52]使用联合表示的迭代对齐,即使用迭代的方式不断更新实体匹配.该方法分为 3 部分:知识表示、联合表示、迭代对齐.但这两种算法都是基于先验实体匹配,将不同知识图谱中的实体和关系嵌入到统一的向量空间,然后将匹配过程转换成向量表示间距离的过程;
c) SEEA 算法[53]分为两部分:属性三元组学习、关系三元组学习.该模型能够自学习,不需要对齐种子的输入.每次迭代,根据前面迭代过程所得到的表示模型,计算实体向量间的余弦相似度.并选取前β对添加到关系三元组中更新本次表示模型,直到收敛.收敛条件:无法选取前β对实体对.

实体对齐方法不仅应用于数据清洗过程中,对后续的数据集成以及数据挖掘也起到重要的作用.除此之外,也有很多重复检测的工具可以使用,如 Febrl 系统、TAILOR 工具、WHIRL 系统、BigMatch 等,但是很多匹配算法只适用于英文不适合中文,所以中文数据清洗工具的开发还需要进一步的研究.

- 相似重复记录清洗

相似重复记录的清洗一般都采用先排序再合并的思想,代表算法有优先队列算法、近邻排序算法、多趟近邻排序算法.优先队列算法比较复杂,先将表中所有记录进行排序后,排好的记录被优先队列进行顺序扫描并动态地将它们聚类,减少记录比较的次数,匹配效率得以提高,该算法还可以很好地适应数据规模的变化.近邻排序算法是相似重复记录清洗的经典算法,近邻排序算法是采用滑动窗口机制进行相似重复记录的匹配,每次只对进入窗口的 w 条记录进行比较,只需要比较 $w\times N$ 次,提高了匹配的效率.但是它有两个很大的缺点:首先是该算法的优劣对排序关键字的依赖性很大,如果排序关键字选择得不好,相似的两条记录一直没有出现在滑动窗口上就无法识别相似重复记录,导致很多条相似重复记录得不到清洗;其次是滑动窗口的值 w 也很难把控,w 值太大可能会产生没必要的比较次数,w 值太小又可能会遗漏重复记录的匹配.多趟近邻排序算法是针对近邻排序算法进行改进的算法,它是进行多次近邻排序算法每次选取的滑动窗口值可以不同,且每次匹配的相似记录采用传递闭包,虽然可以减少很多遗漏记录,但也会产生误识别的情况.这两个算法的滑动窗口值和属性值的权重都是固定的,所以也有一些学者提出基于可变的滑动窗口值和不同权重的属性值来进行相似重复记录的清

洗.以上算法都有一些缺陷,如都要进行排序,多次的外部排序会引起输入/输出代价过大;其次,由于字符位置敏感性,排序时相似重复的记录不一定排在邻近的位置,对算法的准确性有影响.

4 数据交换

4.1 数据交换的基本概念

数据交换是将符合一个源模式的数据转换为符合目标模式数据的问题,该目标模式尽可能准确并且以与各种依赖性一致的方式反映源数据[54,55].

早期数据交换的一个主要方向是在关系模式之间从数据交换的上下文中寻求一阶查询的语义和复杂性.2008 年,Afrati 等人开始系统地研究数据交换中聚合查询的语义和复杂性,给出一些概念并做出了技术贡献[56].在一篇具有里程碑意义的论文中,Fagin 等人提出了一种纯粹逻辑的方法来完成这项任务[55].从这时起,在数据库研究界已经对数据交换进行了深入研究.近年,Xiao 等人指出,跨越不同实体的数据交换是实现智能城市的重要手段,设计了一种新颖的后端计算架构——数据隐私保护自动化架构(DPA),促进在线隐私保护处理自动化,以无中断的方式与公司的主要应用系统无缝集成,允许适应灵活的模型和交叉的服务质量保证实体数据交换[57].随着云计算和 Web 服务的快速发展,Wu 等人将基于特征的数据交换应用于基于云的设计与制造的协作产品开发上,并提出了一种面向服务的基于云的设计和制造数据交换架构[58].

完善合理的数据交换服务建设,关系到大数据平台是否具有高效、稳定的处理数据能力.

4.2 数据交换的实现模式

数据整合是平台建设的基础,涉及到多种数据的整合手段,其中,数据交换、消息推送、通过服务总线实现应用对接等都需要定义一套通用的数据交换标准,基于此标准实现各个系统间数据的共享和交换,并支持未来更多系统与平台的对接.平台数据交换标准的设计,充分借鉴国内外现有的各类共享交换系统的建设经验,采用基于可扩展标记语言(XML)的信息交换框架.XML 定义了一组规则,用于以人类可读和机器可读的格式编码文档,它由国际万维网联盟设计.XML 文档格式良好且结构化,因此它们更易于解析和编写.由于它具有简化、跨平台、可扩展性和自我描述等特征,XML 成为通过 Internet 进行数据传输的通用语言[59].XML 关心的重点是数据,而其他的因素如数据结构和数据类型、表现以及操作,都是有其他的以 XML 为核心的相关技术完成.基于基本的 XML 语言,通过定义一套数据元模型(语义字典)和一套基于 XML Schema 的描述规范来实现对信息的共同理解,基于此套交换标准完成数据的交换.数据交换概括地说有以下两种实现模式.

(1) 协议式交换

协议式数据交换是源系统和目标系统之间定义一个数据交换交互协议,遵循制定的协议,通过将一个系统数据库的数据移植到另一个系统的数据库来完成数据交换.Tyagi 等人于 2017 年提出一种通用的交互式通信协议,称为递归数据交换协议(RDE),它可以获得各方观察到的任何数据序列,并提供单独的性能序列保证[60];并于 2018 年提出了一种新的数据交换交互协议,它可以逐步增加通信大小,直到任务完成,还导出了基于将数据交换问题与秘密密钥协议问题相关联的最小位数的下限[61].这种交换模式的优点在于:它无需对底层数据库的应用逻辑和数据结构做任何改变,可以直接用于开发在数据访问层.但是编程人员基于底层数据库进行直接修改也是这种模式的缺点之一,编程人员首先要对双方数据库的底层设计有清楚的了解,需要承担较高的安全风险;其次,编程人员在修改原有的数据访问层时需要保证数据的完整性和一致性.此外,这种模式的另一个缺点在于系统的可重用性很低,每次对于不同应用的数据交换都需要做不同的设计.下面我们举一个通俗易懂的例子:安徽人和新疆人有生意上的往来,但由于彼此说的都是家乡话,交易很难进行,于是双方就约定每次见面都使用安徽话或者新疆话.假如他们规定一个协议,每次见面都以安徽话来交谈,那么新疆人每句话的语法结构和发音标准都按照安徽话来修改,同时要保证每句话的完整性和准确性,保证双方顺利的交谈.然而在下次的生意中,新疆人可能面对的是一位广东人,那么交流依旧出现了困难,此时新疆人又需要把自己的新疆话转换为广东话.

(2) 标准化交换

标准化数据交换是指在网络环境中建立一个可供多方共享的方法作为统一的标准,使得跨平台应用程序之间实现数据共享和交换.下面我们依旧以安徽人与新疆人作交易为例来解释这种交换模式.为了解决双方无法沟通的困境,双方约定每次见面交易都使用普通话这种标准来交流,当下次即使遇到全国各地的人,也可以使用普通话来交流,而且大家只需要熟悉普通话的语法规则即可,不需要精通各地的语言.这种交换模式的优点显而易见,系统对于不同的应用只需要提供一个多方共享的标准即可,具有很高的可重用性.

实现基于XML的数据交换平台确实需要一系列的努力和资源来创建/管理交换,但它不是对现有系统的大规模改变而是有限的改变,所以使用基于 XML 数据交换的关键优势是信息共享的组织不需要更改其现有的数据存储或标准,使得异构系统之间可以实现最大限度的协同,并能在现有数据交换应用的基础上扩展更多新的应用,从而对不同企业间发展应用集成起到促进作用.

5 数据集成

5.1 数据集成的基本概念

在信息化建设初期,由于缺乏有效合理的规划和协作,信息孤岛的现象普遍存在,大量的冗余数据和垃圾数据存在于信息系统中,数据质量得不到保证,信息的利用效率明显低下.为了解决这个问题,数据集成技术[62]应运而生.数据集成技术是协调数据源之间不匹配问题[63−67],将异构、分布、自治的数据集成在一起,为用户提供单一视图,使得可以透明地访问数据源.系统数据集成主要指异构数据集成,重点是数据标准化和元数据中心的建立.

- 数据标准化:数据标准化的作用在于提高系统的可移植性、互操作性、可伸缩性、通用性和共享性.数据集成依据的数据标准包括属性数据标准、网络应用标准和系统元数据标准.名词术语词典、数据文件属性字典、菜单词典及各类代码表等为系统公共数据,在此基础上促成系统间的术语、名称、代码的统一,促成属性数据统一的维护管理;
- 元数据中心的建立:在建立元数据标准的基础上,统一进行数据抽取、格式转换、重组、储存,实现对各业务系统数据的整合.经处理的数据保存在工作数据库中,库中所有属性数据文件代码及各数据文件中的属性项代码均按标准化要求编制,在整个系统中保持唯一性,可以迅速、准确定位.各属性项的文字值及代码,也都通过词库建设进行标准化处理,实现一词一义.建立元数据中心的基本流程如图 5 所示.

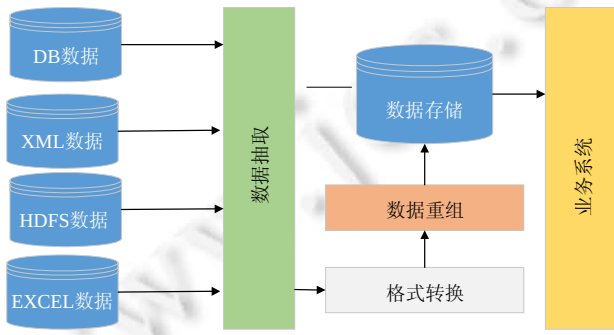

Fig.5 Metadata center

图 5 元数据中心

5.2 数据集成方法

数据规范和数据交换的完成,对数据集成的有效进行提供了很大的帮助,但在数据集成时仍然需要解决以下难题.

首先是异构性.数据异构分为两个方面:其一,不同数据源数据的结构不同,此为结构性异构;其二,不同数据

源的数据项在含义上有差别,此为语义性异构;其次是数据源的异地分布性;最后是数据源的自治性.数据源可以改变自身的结构和数据,这就要求数据集成系统应具有鲁棒性.

为了解决这些难题,现在有模式集成方法、数据复制方法和基于本体的方法这几种典型的数据集成方法:

(1) 模式集成方法

模式集成方法为用户提供统一的查询接口,通过中介模式访问实时数据,该模式直接从原始数据库检索信息(如图 6 所示).该方法的实现共分为 4 个主要步骤:源数据库的发现、查询接口模式的抽取、领域源数据库的分类和全局查询接口集成[68−73].

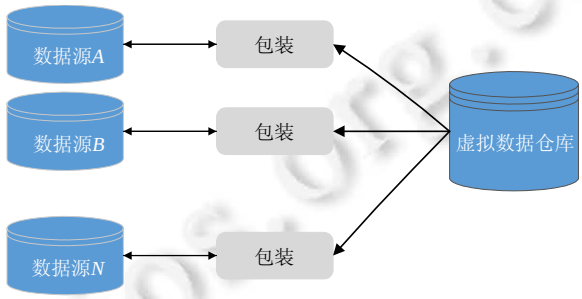

Fig.6 Schematic diagram of a pattern integration approach

图 6 模式集成方法示意图

模式集成方法依赖于中介模式与原始源模式之间的映射[74],并将查询转换为专用查询,以匹配原始数据库的模式.这种映射可以用两种方式指定:作为从中介模式中的实体到原始数据源中的实体的映射——全局视图(GAV)方法[75],或者作为从原始源中的实体到中介模式——本地视图(LAV)方法的映射[76].后一种方法需要更复杂的推理来解析对中介模式的查询[67,77,78],但是可以更容易地将新数据源添加到稳定中介模式中.

模式集成方法的优点是为用户提供了统一的访问接口和全局数据视图;缺点是用户使用该方法时经常需要访问多个数据源,存在很大的网络延迟,数据源之间没有进行交互.如果被集成的数据源规模比较大且数据实时性比较高更新频繁,则一般采用模式集成方法.

(2) 数据复制方法

数据复制方法是将用户可能用到的其他数据源的数据预先复制到统一的数据源中,用户使用时,仅需访问单一的数据源或少量的数据源.数据复制方法提供了紧密耦合的体系结构,数据已经在单个可查询的存储库中进行物理协调,因此解析查询通常需要很少的时间[79],系统处理用户请求的效率显著提升;但在使用该方法时,数据复制需要一定的时间,所以数据的实时一致性不好保证.数据仓库方法是数据复制方法的一种常见方式[80],第一个数据集成系统便是使用该方法于 1991 年在明尼苏达大学设计的.该方法的过程是:先提取各个异构数据源中的数据,然后转换、加载到数据仓库中,用户在访问数据仓库查找数据时,类似访问普通数据库.

对于经常更新的数据集,数据仓库方法不太可行,需要连续重新执行提取、转换、加载(ETL)过程以进行同步.根据数据复制方法的优缺点可以看出:数据源相对稳定或者用户查询模式已知或有限的时候,适合采用数据复制方法.数据仓库方法示意图如图 7 所示.

下面举例说明这两种集成方法具体应用的区别:目前我们想要设计一个应用程序,该应用程序的功能为用户可以利用该程序查询到自己所在城市的任何信息,包括天气信息、人口统计信息等.传统的思想是,把所有这些信息保存在一个后台数据库中,但是这种广度的信息收集起来难度大且成本高,即使收集到这些资源,它们也可能会复制已有数据库中的数据,不具备实时性.

此时,我们选择模式集成方法解决该应用程序面临的问题,让开发人员构建虚拟模式——全局模式,然后对各个单独的数据源进行“包装”,这些“包装”只是将本地查询结果(实际上是由相对应的网站或数据库返回的结果)转换为易于处理的表单,当使用该应用程序的用户查询数据时,看似是本地查询,实则数据集成系统会将此查询转换为相应数据源上的相应查询.最后,虚拟数据库将这些查询的结果反馈给用户.

如果我们选择使用数据复制方法来解决此问题的话,首先,我们需要把所有的数据信息复制到数据仓库中,每当数据(如天气情况)有所更新时,我们也要手动集成到系统中.所以,两种数据集成方法的使用需根据具体的情形来选择.

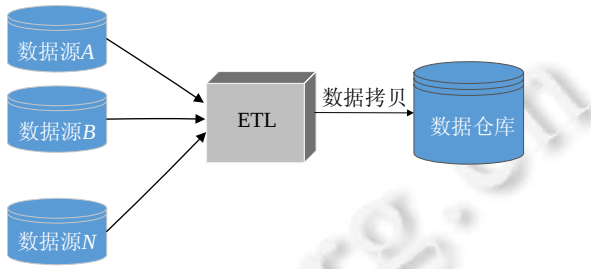

Fig.7 Schematic diagram of data warehouse method

图 7 数据仓库方法示意图

(3) 基于本体的数据集成

根据上述介绍,数据异构有两个方面:前两种方法都是针对解决结构异构而提出的解决方案;而本体技术致力于解决语义性异构问题.语义集成过程中,一般通过冲突检测、真值发现等技术来解决冲突,常见的冲突解决策略有如下 3 类:冲突忽略、冲突避免和冲突消解.冲突忽略是人工干预把冲突留给用户解决;冲突避免是对所有的情形使用统一的约束规则;冲突消解又分为 3 类:一是基于投票的方法采用简单的少数服从多数策略;二是基于质量的方法,此方法在第 1 种方法的基础上考虑数据来源的可信度;三是基于关系的方法,此方法在第 2 种方法的基础上考虑不同数据来源之间的关系.

本体是对某一领域中的概念及其之间关系的显式描述,基于本体的数据集成系统允许用户通过对本体描述的全局模式的查询来有效地访问位于多个数据源中的数据[81].陶春等人针对基于本体的 XML 数据集成的查询处理提出了优化算法[82].目前,基于本体技术的数据集成方法有 3 种,分别为:单本体方法、多本体方法和混合本体方法.

由于单本体方法所有的数据源都要与共享词汇库全局本体关联,应用范围很小,且数据源的改变会影响全局本体的改变.为了解决单本体方法的缺陷,多本体方法应运而生.多本体方法的每个数据源都由各自的本体进行描述,它的优点是数据源的改变对本体的影响小,但是由于缺少共享的词汇库,不同的数据源之间难以比较,数据源之间的共享性和交互性相对较差.混合本体方法的提出,解决了单本体和多本体方法的不足:混合本体的每个数据源的语义都由它们各自的本体进行描述,解决了单本体方法的缺点.混合本体还建立了一个全局共享词汇库以解决多本体方法的缺点,如图 8 所示.混合本体方法有效地解决了数据源间的语义异构问题.

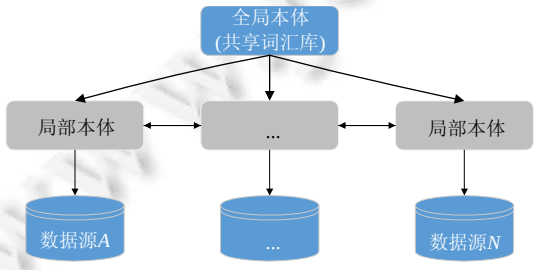

Fig.8 Hybrid ontology approach

图 8 混合本体方法

6 数据治理框架

6.1 数据治理成熟度模型

一个机构的数据治理能力越高,所享受到数据治理带来的价值也会越多,如增加收入、减少成本、降低风险等.于是,很多机构想要准确地评估本公司的数据治理能力,可以利用数据治理成熟度模型方法,包括 DQM, Dataflux 和 IBM 在内的一些组织都开发了相类似的数据治理成熟度模型.

我们先介绍一下 DQM 集团的数据治理成熟度模型[83],此数据治理成熟度模型共分为 5 个阶段.

(1) 意识阶段:当公司数据不统一的情况随处可见,数据质量很差却难以提高,数据模型的梳理难以进行时,公司会意识到数据治理对于数据平台的建设发挥着至关重要的作用,但并没有定义数据规则和策略,基本不采取行动;

(2) 被动的反应阶段:公司在出现数据上的问题时,会去采取措施解决问题,但并不会寻其根源解决根本问题,也就是说,公司的行动通常是由危机驱动的.该类反应性组织的数据仍然是“孤立”存在的,很少进行数据共享,只是努力达到监管的要求;

(3) 主动的应对阶段:处在这个阶段的组织最终可以识别和解决根本原因,并可以在问题出现之前将其化解.这个阶段的组织将数据视为整个企业的战略资产,而不是像第 1 阶段将数据作为一种成本开销;

(4) 成熟的管理阶段:这个阶段的组织拥有一组成熟的数据流程,可以识别出现的问题,并以专注于数据开发的方式定义策略;

(5) 最佳阶段:一个组织把数据和数据开发作为人员、流程和技术的核心竞争力.

IBM 的数据治理成熟度模型也分为 5 个阶段[84],分别是初始阶段、基本管理、定义阶段(主动管理)、量化管理、最佳(持续优化)阶段(影响数据治理成熟度的关键因素有以下 3 个:严格性、全面性以及一致性).

(1) IBM 的初始阶段是指企业缺乏数据治理流程,没有跟踪管理,也没有一个稳定的数据治理的环境,仅仅只能体现个人的努力和成果,工作尚未开展;

(2) 基本管理阶段是指该阶段有了初始的流程定义,开展了基本的数据治理工作,但仍然存在很多问题;

(3) 定义阶段是指企业在相关成功案例的基础上积累了相关的经验,形成了部分标准但仍不完善的流程;

(4) 量化管理阶段的企业能够运用先进的工具对数据治理的效果进行量化,数据治理已经能取得持续的效果,并且能根据既定的目标进行一致的绩效评估;

(5) 最佳阶段是持续地关注流程的优化,达到了此阶段的企业已经具有创新能力,成为行业的领导者.

从这些企业的数据治理模型可以看出:数据治理从来都不是一次性的程序,而是一个持续的过程,这个过程必须是渐进式迭代型的,每个组织必须采取许多小的、可实现的、可衡量的步骤来实现长期目标.

6.2 数据治理框架

Khatri 等人使用 Weill 和 Ross 框架进行 IT 治理,作为设计数据治理框架的起点[85],IBM 的数据治理委员会以支撑域、核心域、促成因素和成果这 4 个层次来构建数据治理框架[84],如图 9 所示.

图 9 的数据治理框架所包含的 11 个域并不是相互独立运行的而是相关联的,例如,数据的质量和安全/隐私要求需要在整个信息生命周期中进行评估和管理.IBM 的数据治理框架注重数据治理的方法以及过程,IBM 数据治理委员会最关键的命题是数据治理的成果,在下面 3 层的支撑作用下,组织最终实现数据治理的目标提升数据价值.

在 IBM 数据治理框架的基础上加以扩充,文献[6]设计了一个大数据背景下的数据治理框架,如图 10 所示.

结合 IBM 公司的数据治理框架,我们对文献[6]给出的大数据治理框架进行了几处修改得到图 10.为了与图 9 保持一致,将文献[6]中大数据治理框架图的“范围”修改为“核心域”,文献[6]的大数据治理框架图的“大数据质量”修改为“数据质量管理”,文献[6]的大数据治理框架图的“大数据生命周期”修改为“数据生命周期管理”.图 10 从原则、核心域、实施与评估这 3 个方面来对大数据治理全面地进行描述,企业数据治理应该遵循战略一致、风险管理、运营合规以及价值创造这 4 个基本的指导性原则,治理的核心域或者说叫决策域包括战略、组织、

数据生命周期管理、数据质量管理、大数据服务创新、大数据安全以及大数据架构这 7 个部分,实施与评估维度指出大数据治理在实施评估时重点需要关注促成因素、实施过程、成熟度评估以及审计这 4 个方面.一个大数据治理组织要在 4 个基本原则下对 7 个核心域进行数据治理,不断地推进大数据治理的工作.

Fig.9 IBM data governance framework[84]

图 9 IBM 数据治理框架[84]

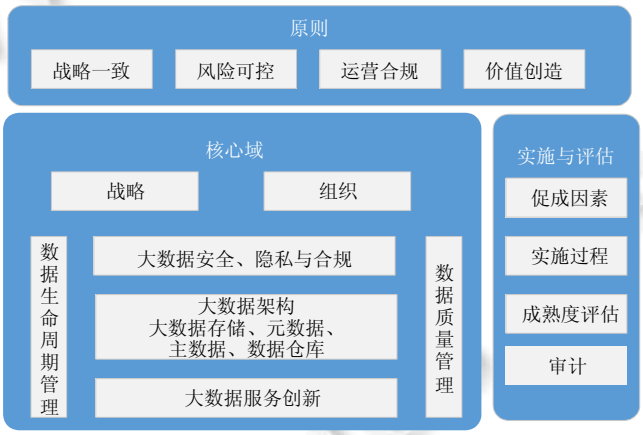

Fig.10 Big data governance framework[6]

图 10 大数据治理框架[6]

框架顶部的 4 个原则是数据治理自上而下的顶层设计,对大数据治理的实施具有指导作用,它为所有其他的管理决策确定方向.战略一致是指数据治理的战略要和企业的整体战略保持一致,在制定数据治理战略时要融合企业的整体战略、企业的文化制度以及业务需要,来绘制数据治理实现蓝图;大数据的到来不仅伴随着价值同时也会带来风险,企业要保持风险可控有计划地对风险进行不定期的评估工作;运营合规是指企业在数据治理过程中要遵守法律法规和行业规范;企业的数据治理要不断地为企业提供创新服务创造价值.

框架的核心域也可以叫做决策域,指出数据治理需要治理的核心对象,下面对数据治理的 7 个核心域进行一一介绍,其中:战略制定要根据大数据治理目标来制定,根据战略的制定,企业应该设置对应的组织架构把战略实施落到实处,明确各个部门相关职责;数据生命周期管理是从数据的采集、存储、集成、分析、归档、销毁的全过程进行监督和管理,根据出现的问题及时优化的过程;数据质量管理不仅要保障数据的完整性、准确

性、及时性以及一致性,而且还包括问题追踪和合规性监控.

2014 年 10 月,美国摩根大通公司电脑系统发生数据泄露,被窃取的信息包括客户姓名、地址、电话号码和电子邮箱地址,将对 7 600 万家庭和 700 万小企业造成影响.2018 年 1 月,有一家数据分析公司对 Facebook 超过 8 700 万用户进行非法的数据挖掘,接下来的 3 月、9 月以及 12 月,Facebook 又多次发生用户数据泄露事件.大数据背景下的信息开放和共享,使得隐私和信息安全问题被显著放大,IBM 数据治理专家 Soares 在其著作《Big Data Governance an Emerging Imperative》中以清晰的案例介绍电信行业利用地理位置数据来侵犯个人隐私[10],因此在大数据治理过程中,采取一定的措施和策略保证信息安全和隐私保护尤为重要.下面从大数据安全防护和隐私保护两个方面来介绍它们的关键技术.

(1) 首先,大数据安全防护主要包括以下关键技术.

- 大数据加密技术:对平台中的核心敏感数据进行加密保护,结合访问控制技术,利用用户权限和数据权限的比较来防止非授权用户访问数据;
- 大数据安全漏洞检测:该技术可以采用白/黑/灰盒测试或者动态跟踪分析等方法,对大数据平台和程序进行安全漏洞检测,减少由于设计缺陷或人为因素留下的问题;
- 威胁预测技术:利用大数据分析技术,对平台的各类信息资产进行安全威胁检测,在攻击发生前进行识别预测并实施预防措施;
- 大数据认证技术:利用大数据技术收集用户行为和设备行为数据,根据这些数据的特征对使用者进行身份判断;

(2) 其次,对于隐私保护,现有的关键技术分析如下.

- 匿名保护技术:针对结构化数据,一般采用数据发布匿名保护技术;而对于类似图的非结构化数据,则一般采用社交网络匿名保护技术;
- 数据水印技术:水印技术一般用于多媒体数据的版权保护,但多用于静态数据的保护,在大数据动态性的特点下需要改进;
- 数据溯源技术:由于数据的来源不同,对数据的来源和传播进行标记,为使用者判断信息真伪提供便利;
- 数据审计技术:对数据存储前后的完整性和系统日志信息进行审计.

大数据架构是从系统架构层面进行描述,不仅关心大数据的存储,还关心大数据的管理和分析.我们首先要明确元数据和主数据的含义:元数据是对数据的描述信息,而主数据就是业务的实体信息.所以对于元数据和主数据的管理是对基础数据的管理.数据治理不仅要降低企业成本,还要应用数据创新服务为企业增加价值,大数据服务创新也是大数据治理的核心价值.

大数据治理的实施与评估主要包括促成因素、实施过程、成熟度评估和审计:促成因素包括企业的内外部环境和数据治理过程中采用的技术工具;大数据治理是一个长期的、闭环的、循序渐进的过程,在每一个阶段需要解决不同的问题,有不同的侧重点,所以应该对数据生命周期的每个阶段有一个很好的规划,这就是实施过程的内涵所在;数据治理成熟度模型我们已经在本节的上半部分介绍了它的内容,但成熟度评估主要是对数据的安全性、一致性、准确性、可获取性、可共享性以及大数据的存储和监管进行评估;审计是第三方对企业数据治理进行评价和给出审计意见,促进有关数据治理工作内容的改进,对于企业的持续发展意义重大.

在企业的数据治理过程中,治理主体对数据治理的需求进行评估来设定数据治理的目标和发展方向,为数据治理战略准备与实施提供指导,并全程监督数据治理的实施过程.通过对实施成果的评估,全面了解本公司数据治理的水平和状态,更好地改进和优化数据治理过程,以致达到组织的预期目标.

7 HAO 治理模型

下面介绍我们自己设计的 HAO 治理模型.该模型从大数据开始,为 HI(人类智能)、AI(人工智能)和 OI(组织智能)三者协同的 HAO 智能[86]提供数据治理支持.

HAO 治理模型旨在实现以下需求.

(1) 建立全面、动态、可配置的数据接入机制,满足数据采集、数据汇聚、任务配置、任务调度、数据加密、断点续传等需求;

(2) 建立标准化的数据处理流程,形成面向数据内容的数据规范、清洗、关联、比对、标识等转换处理规范模式,为一个组织的数据融合建库提供支撑;

(3) 统筹建设多元集成、融合建库的数据组织模式,按照业务类型、敏感程度、隐私内容等关键要素分级分类推进云建库和存储管理,采用特征标签、归一集成等多种手段实现不同来源的数据资源关联融合;

(4) 构建知识图谱分类,建设多渠道、多维度的数据服务模式,面向使用者提供查询检索、比对排序等基础数据服务,面向专业人员提供挖掘分析、专家建模等智能数据服务;

(5) HI 和 AI 通过知识图谱和 OI 实现交互和协同,存取和共享治理过的集成数据,并利用大数据处理模型(以 HACE 定理开始的三级结构,如图 1 所示)、云计算和雾计算机制来实现数据服务和隐私保护.

HAO 治理模型如图 11 所示.

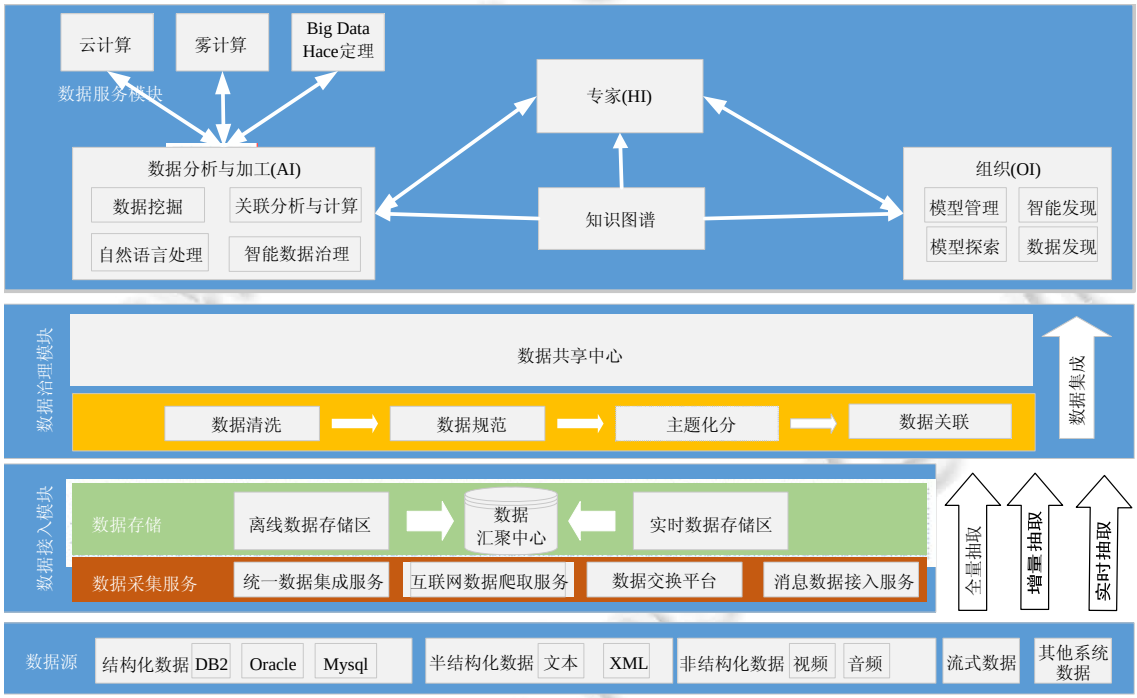

Fig.11 Architecture diagram of HAO governance model

图 11 HAO 治理模型架构图

该模型具备以下功能.

- 支持不同种类、不同数据源、不同目标库的数据抽取传输.常用数据源、目标库类型包括 Oracle,SqlServer,MySql,Hbase,Hive,GreenPlum,Gbase,PostgreSQL,SOLR,Redis,ODPS,OTS,GDS 等主流数据库,常用文件类型包括 FTP,XML,CSV,JSON,EXCEL 等,常见消息处理类型包括 Kafka 和 Webservice;
- 支持不同类型的抽取汇聚任务配置,主要包括异构数据库之间数据传输汇聚,不同类型、跨服务器的文件型数据传输,数据库和文件类、服务接口间相互传输等;
- 支持数据清洗和数据规范的规则自定义,主要包括 NULL 值替换、字符串操作、数据类型转换、函数依赖、正则处理、组合字段、数据比对、自定义 SQL 脚本执行、JSON 输出等数据转换规则,以及对

相似重复记录和属性值异常等问题数据清洗规则,以及 MD5 加密规则;

- 实现基于数据元的异构数据自动解析,并能按照业务场景进行自定义配置,实现智能化、可视化、组件式数据汇聚整合任务构建;
- 通过构建知识图谱实现作业流程的可视化设计,各组件、连接线等以图形控件形式提供,并按不同功能分组,支持复制、粘贴、剪切、撤销等功能,数据整合任务在流程设计器中可直观显示;
- 支持插件二次开发:提供第三方开发平台,方便根据现场实际业务需求,定制项目插件.

HAO 治理模型的设计准则包括:(1) 数据源和治理功能的模块化;(2) 模型的可分解性;(3) 快速原型系统构建;(4) 数据更新和融合能力;(5) 交互的灵活性和(6) 实时反应.

下面对 HAO 治理模型包括的 3 个核心模块——数据接入模块、数据治理模块、数据服务模块分别进行介绍.

7.1 数据接入模块

大数据工程的数据来源包含企业内部数据和企业外部数据,其中:企业内部数据由资源服务平台、综合资源库、各业务系统生产库中的结构化数据和文件服务器上的文本、图片等非结构化数据组成,其中包括人财物记录、财物报表、原材料、顾客信息、气测数据以及企业的文化和规章制度等;企业外部数据由社会数据、互联网数据和设备采集数据组成,外部数据一般包括地理环境、人口数据、经济市场、金融数据、社会关系、社交数据等等.

在数据接入之前,首先需要进行数据采集,如图 12 所示.数据采集基于云计算和分布存储之上的采集工具,采用标准化、规范化的抽取模式,实现结构化、半结构化、非结构化资源的统一抽取、整合、加工、转换和装载.数据采集工具主要包括了数据层、接入层、交互层和监控层.其中,工具的数据层即涉及整个采集平台中总体架构的数据层即数据支撑层,工具背后的接入层是采集逻辑处理部分,交互层即对应总体架构的采集门户.

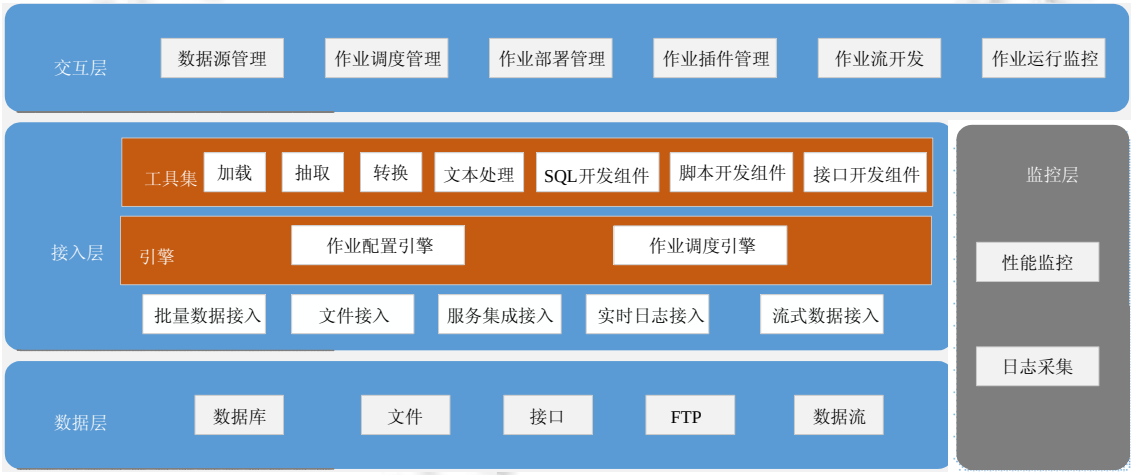

Fig.12 Data acquisition tool architecture diagram

图 12 数据采集工具架构图

数据层指出企业内部和企业外部数据的主要数据来源方式,数据库可以是指业务系统的 Oracle;文件方式是各种文件或 FTP 接入的文件包;接口主要是用来企业对接外部系统使用的;数据流是指可以使用 Kafka 平台处理的实时数据流式方式这种来源.接入层主要提供丰富的工具集,针对不同的数据接入方式提供相应的工具组件,依赖作业配置引擎和作业调度引擎实现数据抽取.监控层可监控作业执行情况,采集作业日志,对问题作业及时告警,方便后期用户排除故障、维护作业.交互层提供可视化页面便捷地实现数据接入与作业管理.

对采集后各种类型的源数据进行数据抽取,该模型的数据抽取支持 3 种方式:全量抽取、增量抽取、实时

抽取,将经过数据抽取后的数据汇入到汇聚库中;对于其他的数据库系统,可以直接通过数据交换平台,把数据汇入到汇聚库中.

7.2 数据治理模块

数据治理模块主要包括对汇聚库中的数据进行数据清洗和数据规范,必要时进行主题划分和数据关联,然后进行数据集成,治理完成后的数据汇聚到数据共享中心中.

数据清洗是对数据进行审查和校验,过滤不合规数据、删除重复数据、纠正错误数据、完成格式转换,并进行清洗前后的数据一致性检查,保证清洗结果集的质量.数据清洗的方法除了以上介绍的几种基本方法以外,该模型还支持自定义清洗规则,数据清洗规则是由业务需求人员与开发人员配合制定数据处理逻辑,经过这些规则进行数据清洗后,保证数据的一致性、准确性和规范性更能满足业务上的需求.

数据治理技术及基本方法在前面几节进行了详细介绍.

7.3 数据服务模块

数据服务模块以数据共享中心构建知识图谱为起点,早在 2006 年,Web 创始人 Berners-Lee 就提出数据链接的思想,随后掀起了语义网络的狂潮[87],知识图谱在此基础上形成.但是直到 2012 年,知识图谱的概念才被谷歌正式提出[88].知识图谱是由节点和边组成的巨型知识网络,节点代表实体,边代表实体之间的关系,每个实体还由(key-value)键值对来描述实体的内在特性.新的知识图谱中还增加了实体与实体之间的事件,即边表示关系或事件.杨玉基等人提出用四步法来构建知识图谱,即领域本体构建、众包半自动语义标注、外源数据补全、信息抽取[89].

数据服务模块基于知识图谱面向不同用户提供多渠道、多维度的数据服务,面向使用者提供模型管理、智能发现、模型探索、数据探索、数据订阅等数据服务,面向专业人员提供挖掘分析、专家建模等智能数据服务.模型管理主要是对实体、关系进行编辑和处理;智能发现是根据日志等元信息,将配置到系统的数据源反向推导出物理模型关系,将多个异构物理模型归一到同一实体后自动生成语义层的业务视图;模型探索是支持关键词搜索实体、关系等,将搜索结果拖拽到画布探索实体之间以及关系之间的核对关系,用户在了解业务模型的同时,也可以了解到业务模型背后对应的物理模型,以及物理数据表的生产血缘关系;数据探索是对业务模型视图可以进行知识问答式的搜索,在路径的任意节点上设置标签的条件,再在另外的节点上设定对应标签的答案,使得用户对数据的业务关系充分地了解;数据订阅满足外部其他平台对本平台各类数据的需求,通过对不同用户下放的不同权限,再结合数据资源目录服务的开放数据内容,为外部用户提供数据订阅/退订流程,并通过资源总线服务完成最终的数据投递.

领域专家们(人类智能,HI)可以根据知识图谱中的实体、关系、属性等核心数据进行建模,并进行高层次的数据挖掘分析和加工,可以同知识图谱、数据分析与加工模块(AI)和组织智能(OI)相互交互和协同,实现 HAO 智能的大智慧问题求解[86].吴信东等人于 2008 年所编著的《数据挖掘十大算法》一书详细地介绍了用途最广、影响最大的 10 种数据挖掘算法[90],并于 2018 年,吴信东等人基于分布式计算对大数据分析的两种算法——MapReduce 与 Spark 从背景、原理以及应用场景进行了具体的分析与比较[91].HACE 定理的大数据处理框架中(如图 1 所示),第 1 层架构解决了流数据存储的计算问题,第 2 层架构考虑了隐私保护和模式发现,第 3 层架构主要描述复杂的数据挖掘算法,HACE 定理在数据服务模块如关联分析与计算以及数据挖掘得到了广泛应用[8];自然语言处理的应用更加广泛,例如我们平时使用的私人助手 Siri 以及出行助手等,都能给人们带来更加便利的服务.HAO 治理模型涵盖了数据治理的全过程,从数据的采集、交换、清洗、规范、集成、应用等融为一体,完成了智能数据治理.

HAO 智能的核心是在大数据问题环境下,用人机协同来实现组织智能(HI+AI+OI),所以数据治理功能的模块化和交互的灵活性是上面提到的 HAO 治理模型 6 个设计准则中的两个.

8 数据治理具体应用

下面以公安数据治理为例,具体介绍 HAO 治理模型的大数据治理过程.

8.1 公安数据治理架构

图 13 描述的是公安数据治理框架,平台架构主要包括数据存储、数据计算、数据管理、数据应用这 4 个部分.

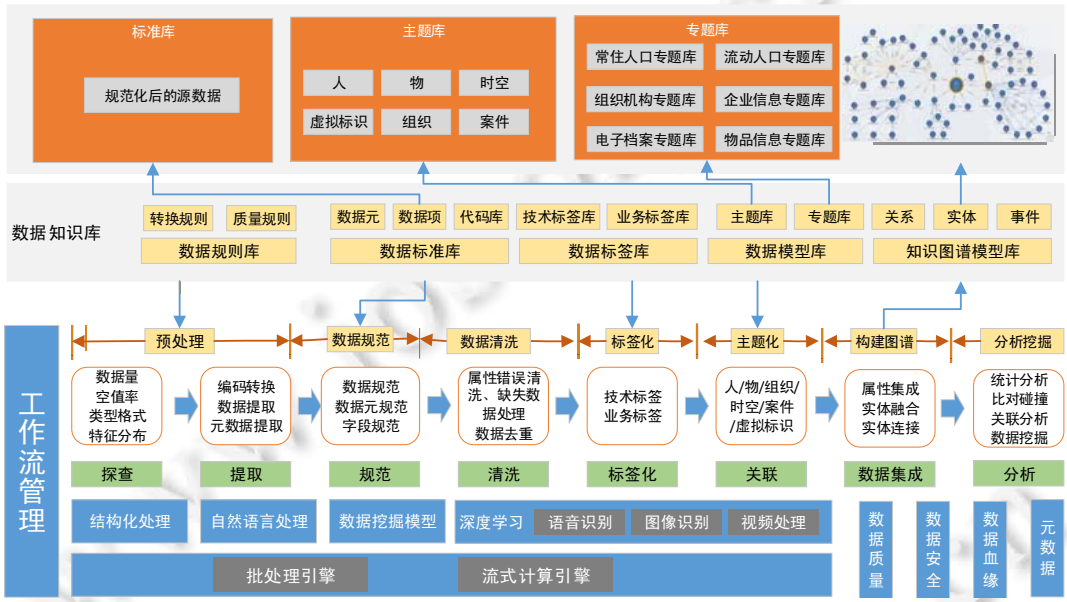

Fig.13 Public security data governance architecture diagram

图 13 公安数据治理架构图

(1) 数据存储:基于分布式的大数据存储平台,具有很强的存储能力和扩张能力;

(2) 数据计算:这是数据治理的最主要部分,包括数据的探查、提取、清洗、转换、集成等.这些计算任务都是基于大数据分布式的计算能力,应用 MapReduce 批处理和 spark streaming 流式处理技术,通过 scheduler 任务调度器,实现对调度任务的执行、管理与监控.

- 数据探查:通过对数据量、数据质量、数据特征等指标的分析来评估后续数据治理任务的工作量;
- 数据提取:抽取分布在各个系统中的各种类型的源数据,提取元数据,基于深度学习的语音识别、图像识别、视频处理技术,实现对非结构化的数据提取;
- 数据清洗:对缺失数据的处理,过滤掉重复相似的记录,清除值错误的数据;
- 数据转换:将不符合规范的数据,按照规范化的处理规则,转化成符合标准的数据,如编码统一、格式统一、元数据统一等;
- 数据集成:将转化后的规范化数据进行整合,按照一定方式重新组织,如数据属性的融合、关系融合、数据的主题化、标签化等;

(3) 数据管理:对集成后的数据统一维护与管理,包括对数据质量的检测、数据安全控制、数据血缘的监控、元素管理等.

- 数据质量检测:从各个维度(唯一性、准确性、完整性、合法性等)检测,并形成数据质量报告;
- 数据安全控制:对数据的使用与访问,进行权限的管理与控制;

- 数据血缘监控:追踪数据的来源与去向的整个过程;
- 元数据管理:数据知识库的建立与维护,包括对代码库、标准库、标签库、模型库、图谱库等的管理;

(4) 数据应用:这是数据价值最直接的体现,基于自然语言处理、数据挖掘算法模型等技术对数据分析挖掘,包括统计分析、比对碰撞、关联分析、数据挖掘等,将分析结果提供给上层应用,如构建专题库、主题库、构建知识图谱等.

8.2 数据处理流程

数据处理流程是对源数据到目标数据整个处理过程的监管,并描述了数据采集、数据处理及数据展现这 3 个方面所用到的技术架构和处理逻辑.本节主要介绍了处理流程中数据接入、数据预处理、数据规范化、数据清洗、数据标签化、数据主题化、构建知识图谱以及数据分析与挖掘 8 个方面的内容.

(1) 数据接入

公安系统中的源数据,包括结构化文本、关系型数据库、非结构化的文本及视频、hadoop 平台中的数据以及流式数据,经过批处理引擎或流式计算引擎,接入到统一的数据源系统中,形成最初的数据集市.

(2) 数据预处理

在对数据集市中的数据做处理前,根据数据规则库定义的规则,首先对数据进行预处理,包括数据质量的评估、空值率的计算、数据特征分析、数据格式的分析等;然后判断数据是否有治理的价值;然后提取需要治理的数据、提取元数据,经过统一的编码转换处理后,过滤掉脏、乱、差的数据;然后进行数据去重等清洗处理.

(3) 数据规范化

数据规范是将预处理后的数据,根据数据标准知识库的标准,将数据统一处理成符合行业标准、省部级标准及国标等标准的规范化数据,提高数据的可移植性、共享性及复用性.数据规范过程(标准化过程)中所依赖的数据规范来源于权威性的行业规范、国标、部标等,对数据、名称、字段及元数据等进行标准化.

(4) 数据清洗

数据清洗是对不完整的数据、不一致的数据以及异常的数据进行清洗,并过滤掉重复相似的记录.

(5) 数据标签化

数据标签根据数据标签库可以分为技术标签和业务标签:技术标签是基于表、字段的技术元数据,例如空间占用、条目数、最新更新时间、更新频率、访问频率、数据格式、字段数据类型、是否压缩等,通过规则引擎进行规则计算,为库、表、字段等打上相应的技术标签,例如最近一天更新的数据、大数据集、小数据集、频繁更新数据集、压缩文件、图片、视频等;业务标签基于库、表、字段的业务定义、描述,值域的具体内容,对于数据进行业务标签生成,例如对于库表来说,数据来源/数据种类(人口、教育、医疗等)标签、数据内容标签(姓名、组织、地址、电话、商品等).

(6) 数据主题化

数据按照一定的主题进行关联来构造一个模型.公安数据治理分别以人、物、时空、组织、虚拟标识、案件等作为主题,分别建立模型,如图 14 所示.

- 以人作为主题时,提取自然人为主体进行描述的数据资源,并按照公安部的数据分类进行主题模型的构建;
- 以物作为主题构建模型时,提取特定的物为主体进行描述的数据资源,针对不同情况涵盖不同的内容,包括物品、物证、微小痕迹、尸体等;
- 以时空作为主题时,提取以时间、地点为主体进行描述的数据资源来构建时空主体模型;
- 以组织作为主题时,提取法人、单位、特定人群组织结构(如:户)为主体进行描述的数据资源来构建组织类主题模型;
- 以虚拟标识作为主题时,以一个物品的标签或者分类信息作为主题进行构建模型;
- 以案件作为主题构建模型时,根据执行主体的不同,案件又分为侦查调查行为和违法犯罪行为:侦查调

查行为是指公安机关行使打击犯罪,维护社会治安进行侦查破案的行为;而违法犯罪行为是指犯罪嫌疑人进行违法犯罪的行为.

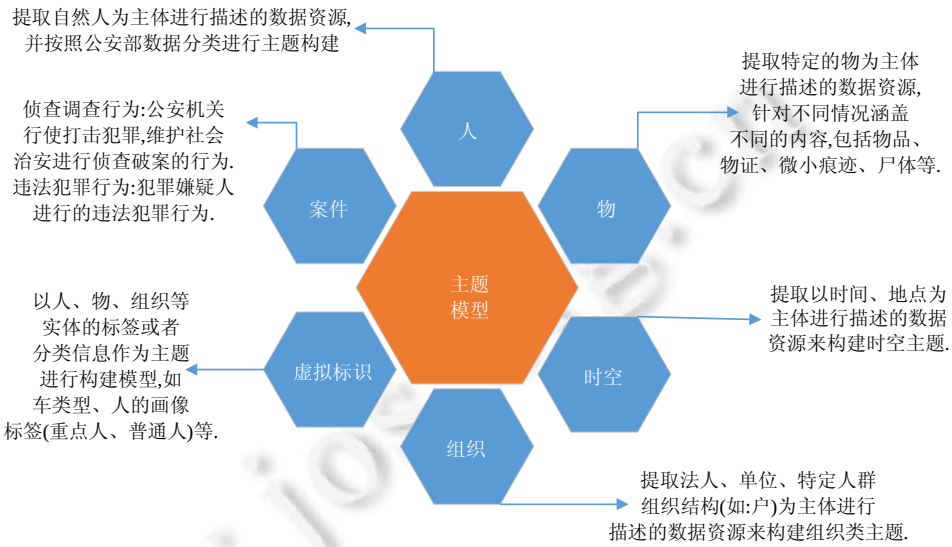

Fig.14 Public security governance theme model diagram

图 14 公安治理主题模型图

(7) 知识图谱构建

知识图谱按照目标数据可以分为实体、事件、关系这 3 种类型来建立数据之间的关联关系,将数据抽象化的内在联系,以可视化的形式有效表现出来.图 15 是以人为中心实体构建的一个简单的知识图谱.以人为中心实体,建立人与电话号码所属关系、人与护照所属关系及人与人的关系,同时建立了人与航班的出行事件、人与旅馆的住宿事件.

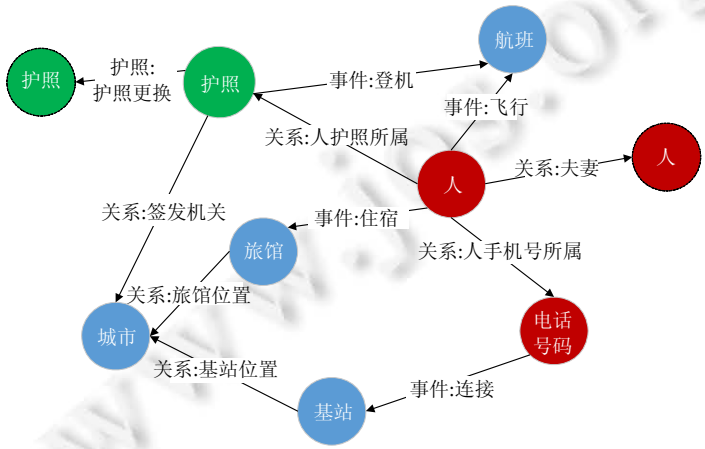

Fig.15 Knowledge gragh of character tracking

图 15 人物追踪知识图谱

(8) 数据分析与挖掘

对治理后的标准化数据,采用一定的数据挖掘算法模型,对数据进行统计分析、关联分析、比对碰撞、数据挖掘等,为上层应用提供数据服务.公安机关作为侦查一线的最实用的技术是数据比对碰撞分析,数据比对碰

撞分析是指运用计算机对数据进行分析,将两组以上同类型的数据集进行梳理,通过关联查询,筛选数据集取交集的一种方法.

9 总结与展望

身处于大数据时代,数据已经成为一个组织最宝贵的财富之一,组织如果想要利用庞大且宝贵的数据资产来挖掘其中的商业价值,在数据挖掘之前就需要使用数据治理技术,提高数据质量,减少实际挖掘所需要的时间.通过第 8 节对公安数据治理流程的具体介绍我们会发现:数据治理技术融入到数据治理的每一个阶段中,而不是孤立使用的,每一个阶段都可能用到多个数据治理技术.

数据治理的核心目标是在降低风险的同时,为企业增加价值.合理的数据治理,能够建立规范的数据应用标准,消除数据的不一致性,提高数据质量,推动数据广泛共享,充分发挥数据对政府及企业的业务、管理以及战略决策的重要作用.大数据治理对于确保大数据的优化、共享和安全是至关重要的,有效的大数据治理计划可通过改进决策、缩减成本、降低风险和提高安全合规等方式,将价值回馈于业务,并最终体现为增加收入和利润[6].根据上述几节的描述,数据治理包括元数据管理、数据质量管理、数据安全合规、数据模型设计以及数据的应用这 5 个基本功能.

一个组织数据治理的好坏是否达到自己预期的目标,可以通过以下几个方面进行评价.

- 从数据的质量方面考虑;
  - a) 数据的准确性:经过数据治理后的数据应该是准确的,而不能在治理过程中给正确的数据带去噪音;
  - b) 数据的完整性和一致性:数据治理之后,数据的完整程度以及数据的一致性;
  - c) 数据的安全性:好的数据治理要充分地保护敏感数据;
- 从数据治理的效率进行考虑:使用每秒处理多少条数据进行直观对比,这直接影响到数据的及时性;
- 数据治理模型的成熟度:数据治理过程中,选择的数据模型的成熟度直接影响数据治理的结果;
- 从是否能追根溯源,找到数据质量问题产生的原因;
- 人工干预程度:发现质量问题以后,是系统自动处理,还是需要人工干预处理.

然而,现在大数据治理也面临一系列的问题和挑战.

- 随着数据产生方式的不断扩展,大数据不仅量大、类型多样,而且数据内容的维度和知识范畴的粒度也以多样性展现,体现的是数据与知识之间的立体关系[92],所以大数据治理技术的复杂性也将加大;
- 数据量的庞大和增长速度之快,就要求数据清洗活动要具有可伸缩性和及时性,虽然已经提出了多种错误检测的方法,但是仍然有很多错误不能被检测到.要设计更具表现力的完整性约束语言,使得数据所有者可以轻松地指定数据的质量规则,并有效地让人类专家参与错误检测[45];
- 数据治理技术面临着更加严峻的隐私安全的挑战.多源数据的集成技术使得数据之间的关联性无形地被公开化,很可能会暴露用户的个人隐私.所以,需要研究主动降低隐私泄露风险的策略和风险评估模型,用来有效地预测隐私泄露的风险程度并提供风险预警[92].Ni 等人于 2010 年提供了一种支持隐私感知访问控制机制的综合框架,即,一种适用于对包含个人身份信息的数据实施访问控制的机制[93];
- 由于数据治理是一个长期的过程,短期投入的人力、技术不一定能够得到实质性的回报,所以数据治理面临着更大的投资回报风险.

本文主要介绍了数据治理技术,数据治理方法不仅需要数据治理技术,还需要企业的制度规范以及生态运营来配合加强数据治理工作.在制度保障方面,一个组织应当定义模型设计规范、数据开发规范、数据变更规范、数据质量管理规范、数据安全规范、元数据规范等;在组织保障方面,组织应当设立数据委员会包括决策小组、安全小组、质量小组以及稳定性小组等来执行管理职责,设立数据资产部门包括部门数据负责人和数据生产团队来执行建设职责.一个组织应该对数据治理进行长期的规划,建立有效的数据治理体系,挖掘数据资产

的潜力,从而发挥数据资产在企业中的核心价值.

References:

[1] Li JZ, Wang HZ, Gao H. State-of-the-Art of research on big data usability. Ruan Jian Xue Bao/Journal of Software, 2016,27(7): 1605−1625 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5038.htm [doi: 10.13328/j.cnki.jos.005038]

[2] Redman TC. The impact of poor data quality on the typical enterprise. Communications of the ACM, 1998,41(2):79–82. [doi: 10.1145/269012.269025]

[3] Miller Jr DW, Yeast JD, Evans RL. Missing prenatal records at a birth center: A communication problem quantified. In: Proc. of the AMIA Annual. Bethesda: American Medical Informatics Association, 2005. 535–539.

[4] Swartz N. Gartner warns firms of ‘dirty data’. Information Management, 2007,41(3):6.

[5] Huang LS, Tian MM, Huang H. Preserving privacy in big data: A survey from the cryptographic perspective. Ruan Jian Xue Bao/ Journal of Software, 2015,26(4):945−959 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4794.htm [doi: 10.13328/j.cnki.jos.004794]

[6] Zhang SH, Pan R, Zong YW. Big Data Technology and Application Series. Shanghai: Shanghai Scientific & Technical Publishers, 2016. 1−224 (in Chinese).

[7] Otto B. Data governance. Business & Information Systems Engineering, 2011,3(4):241−244. [doi: 10.1007/s12599-011-0162-8]

[8] Wu XD, He J, Lu RQ, Zheng NN. From big data to big knowledge: HACE+BigKE. Acta Automatica Sinica, 2016,42(7):965−982 (in Chinese with English abstract). [doi: 10.16383/j.aas.2016.c160239]

[9] Wu X, Zhu X, Wu G, Ding W. Data mining with big data. IEEE Trans. on Knowledge and Data Engineering, 2014,26(1):97−107. [doi: 10.1109/TKDE.2013.109]

[10] Soares S. Big Data Governance: An Emerging Imperative. Boise: MC Press, 2012. 3–286.

[11] Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I. A view of cloud computing. Communications of the ACM, 2010,53(4):50−58. [doi: 10.1145/1721654.1721672]

[12] Feng DG, Min Z, Yan Z, Zhen X. Study on cloud computing security. Ruan Jian Xue Bao/Journal of Software, 2011,22(1):71−83 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3958.htm [doi: 10.3724/SP.J.1001.2011.03958]

[13] Baek J, Safavinaini R, Susilo W. Public key encryption with keyword search revisited. In: Proc. of the Int’l Conf. on Computational Science and ITS Applications. Heidelberg: Springer-Verlag, 2008. 1249−1259. [doi: 10.1007/978-3-540-69839-5_96]

[14] Fang L, Susilo W, Ge C, Wang J. A secure channel free public key encryption with keyword search scheme without random oracle. In: Proc. of the Int’l Conf. on Cryptology and Network Security. Heidelberg: Springer-Verlag, 2009. 248–258. [doi: 10.1007/978-3-642-10433-6_16]

[15] Di Crescenzo G, Saraswat V. Public key encryption with searchable keywords based on Jacobi symbols. In: Proc. of the Int’l Conf. on Cryptology in India. Heidelberg: Springer-Verlag, 2007. 282–296. [doi: 10.1007/978-3-540-77026-8_21]

[16] Bellare M, Boldyreva A, O’Neill A. Deterministic and efficiently searchable encryption. In: Proc. of the Annual Int’l Cryptology Conf. Heidelberg: Springer-Verlag, 2007. 535–552. [doi: 10.1007/978-3-540-74143-5_30]

[17] Bellare M, Fischlin M, O’Neill A, Ristenpart T. Deterministic encryption: Definitional equivalences and constructions without random oracles. In: Proc. of the Annual Int’l Cryptology Conf. Heidelberg: Springer-Verlag, 2008. 360−378. [doi: 10.1007/978-3-540-85174-5_20]

[18] Wee H. Dual projective hashing and its applications—Lossy trapdoor functions and more. In: Proc. of the Annual Int’l Conf. on the Theory and Applications of Cryptographic Techniques. Heidelberg: Springer-Verlag, 2012. 246–262. [doi: 10.1007/978-3-642-29011-4_16]

[19] Xie X, Xue R, Zhang R. Deterministic public key encryption and identity-based encryption from lattices in the auxiliary-input setting. In: Proc. of the Int’l Conf. on Security and Cryptography for Networks. Heidelberg: Springer-Verlag, 2012. 1–18. [doi: 10.1007/978-3-642-32928-9_1]

[20] Boneh D, Waters B. Conjunctive, subset, and range queries on encrypted data. In: Proc. of the Theory of Cryptography Conf. Heidelberg: Springer-Verlag, 2007. 535−554. [doi: 10.1007/978-3-540-70936-7_29]

[21] Hwang YH, Lee PJ. Public key encryption with conjunctive keyword search and its extension to a multi-user system. In: Proc. of the Int’l Conf. on Pairing-based Cryptography. Heidelberg: Springer-Verlag, 2007. 2−22. [doi: 10.1007/978-3-540-73489-5_2]

[22] Katz J, Sahai A, Waters B. Predicate encryption supporting disjunctions, polynomial equations, and inner products. In: Proc. of the Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Heidelberg: Springer-Verlag, 2008. 146−162. [doi: 10.1007/978-3-540-78967-3_9]

[23] Gentry C. Fully homomorphic encryption using ideal lattices. In: Proc. of the STOC, 2009. 169−178. [doi: 10.1007/978-3-642-13013-7_25]

[24] Smart NP, Vercauteren F. Fully homomorphic encryption with relatively small key and ciphertext sizes. In: Proc. of the Int'l Workshop on Public Key Cryptography. Heidelberg: Springer-Verlag, 2010. 420−443. [doi: 10.1007/978-3-642-13013-7_25]

[25] Gentry C, Halevi S, Smart NP. Better bootstrapping in fully homomorphic encryption. In: Proc. of the Int'l Workshop on Public Key Cryptography. Heidelberg: Springer-Verlag, 2012. 1−16. [doi: 10.1007/978-3-642-30057-8_1]

[26] Brakerski Z, Gentry C, Halevi S. Packed ciphertexts in LWE-based homomorphic encryption. In: Proc. of the Public-Key Cryptography (PKC 2013). Heidelberg: Springer-Verlag, 2013. 1−13. [doi: 10.1007/978-3-642-36362-7_1]

[27] Brakerski Z. Fully homomorphic encryption without modulus switching from classical GapSVP. In: Proc. of the Advances in Cryptology (CRYPTO 2012). Heidelberg: Springer-Verlag, 2012. 868−886. [doi: 10.1007/978-3-642-32009-5_50]

[28] Van Dijk M, Gentry C, Halevi S, Vaikuntanathan V. Fully homomorphic encryption over the integers. In: Proc. of the Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Heidelberg: Springer-Verlag, 2010. 24−43. [doi: 10.1007/978-3-642-13190-5_2]

[29] Coron JS, Naccache D, Tibouchi M. Public key compression and modulus switching for fully homomorphic encryption over the integers. In: Proc. of the Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Heidelberg: Springer-Verlag, 2012. 446−464. [doi: 10.1007/978-3-642-29011-4_27]

[30] Luo C, He F, Yan D, Zhang D, Zhou X, Wang BY. PSpec: A formal specification language for fine-grained control on distributed data analytics. In: Proc. of the 39th Int'l Conf. on Software Engineering Companion. Buenos Aires: IEEE Press, 2017. 300−302. [doi: 10.1109/ICSE-C.2017.120]

[31] Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 2000,23(4):3−13.

[32] Tang N. Big data cleaning. In: Chen L, ed. Proc. of the Web Technologies and Applications. Cham: Springer Int'l Publishing, 2014. 13−24. [doi: 10.1007/978-3-319-11116-2_2]

[33] Lee ML, Ling TW, Low WL. IntelliClean: A knowledge-based intelligent data cleaner. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 290−294.

[34] Monge AE. Matching algorithms within a duplicate detection system. IEEE Data Engineering Bulletin, 2000,23(4):14−20.

[35] Chu X, Ilyas IF, Papotti P. Holistic data cleaning: Putting violations into context. In: Proc. of the 2013 IEEE 29th Int'l Conf. on Data Engineering (ICDE). Brisbane: IEEE, 2013. 458−469. [doi: 10.1109/ICDE.2013.6544847]

[36] Dallachiesa M, Ebaid A, Eldawy A, Elmagarmid A, Ilyas IF, Ouzzani M, Tang N. NADEEF: A commodity data cleaning system. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2013. 541−552.

[37] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR), 2009,41(3):16.

[38] Beskales G, Ilyas IF, Golab L, Galiullin A. On the relative trust between inconsistent data and inaccurate constraints. In: Proc. of the 2013 IEEE 29th Int'l Conf. on Data Engineering (ICDE). Brisbane: IEEE, 2013. 541−552. [doi: 10.1109/ICDE.2013.6544854]

[39] Fan W, Ma S, Tang N, Yu W. Interaction between record matching and data repairing. Journal of Data and Information Quality (JDIQ), 2014,4(4):16. [doi: 10.1145/1989323.1989373]

[40] Fan W, Geerts F, Tang N, Yu W. Inferring data currency and consistency for conflict resolution. In: Proc. of the 2013 IEEE 29th Int'l Conf. on Data Engineering (ICDE). Brisbane: IEEE, 2013. 470−481. [doi: 10.1109/ICDE.2013.6544848]

[41] Shen W, DeRose P, Vu L, Doan A, Ramakrishnan R. Source-Aware entity matching: A compositional approach. In: Proc. of the IEEE 23rd Int'l Conf. on Data Engineering (ICDE 2007). Istanbul: IEEE, 2007. 196−205. [doi: 10.1109/ICDE.2007.367865]

[42] Yang DH, Li NN, Wang HZ, Li JZ, Gao H. The optimization of the big data cleaning based on task merging. Chinese Journal of Computers, 2016,39(1):97−108 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2016.00097]

[43] Guo ZM, Zhou AY. Research on data quality and data cleaning: A survey. Ruan Jian Xue Bao/Journal of Software, 2002,13(11): 2076−2107 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/2076.htm [doi: 10.13328/j.cnki.jos.2002.11.003]

[44] Aggarwal CC. Outlier Analysis. Cham: Springer Int'l Publishing, 2015. 237−263. [doi: 10.1007/978-3-319-14142-8_8]

[45] Chu X, Ilyas IF. Qualitative data cleaning. Proceedings of the VLDB Endowment, 2016,9(13):1605−1608.

[46] Raman V, Hellerstein JM. Potter's wheel: An interactive data cleaning system. In: Proc. of the 27th VLDB Conf. Roma: VLDB, 2001. 381−390.

[47] Hua M, Pei J. Cleaning disguised missing data: A heuristic approach. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2007). New York: ACM Press, 2007. 950−958. [doi: 10.1145/1281192.1281294]

[48] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans. on Knowledge and Data Engineering, 2007,19(1):1−16. [doi: 10.1109/TKDE.2007.250581]

[49] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems (NIPS 2013). Curran Associates Inc., 2013. 2787−2795.

[50] Chen M, Tian Y, Yang M, Zaniolo C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. AAAI Press, 2017. 1511−1517.

[51] Sun Z, Hu W, Li C. Cross-Lingual entity alignment via joint attribute-preserving embedding. In: Proc. of the Int'l Semantic Web Conf. Springer-Verlag, 2017. 628−644. [doi: 10.1007/978-3-319-68288-4_37]

[52] Zhu H, Xie R, Liu Z, Sun M. Iterative entity alignment via joint knowledge embeddings. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. AAAI Press, 2017. 4258−4264.

[53] Guan S, Jin X, Jia Y, Wang Y, Shen H, Cheng X. Self-Learning and embedding based entity alignment. In: Proc. of the 2017 IEEE Int'l Conf. on Big Knowledge (ICBK). Hefei: IEEE, 2017. 33−40. [doi: 10.1109/ICBK.2017.15]

[54] Chirkova R, Libkin L, Reutter JL. Tractable XML data exchange via relations. In: Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2011. 1629−1638. [doi: 10.1145/2063576.2063813]

[55] Fagin R, Kimelfeld B, Kolaitis PG. Probabilistic data exchange. Journal of the ACM (JACM), 2011,58(4):15. [doi: 10.1145/1989727.1989729]

[56] Afrati F, Kolaitis PG. Answering aggregate queries in data exchange. In: Proc. of the 27th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. Vancouver: ACM Press, 2008. 129−138. [doi: 10.1145/1376916.1376936]

[57] Xiao Z, Fu X, Goh RSM. Data privacy-preserving automation architecture for industrial data exchange in smart cities. IEEE Trans. on Industrial Informatics, 2018,14(6):2780−2791. [doi: 10.1109/TII.2017.2772826]

[58] Wu Y, He F, Zhang D, Li X. Service-Oriented feature-based data exchange for cloud-based design and manufacturing. IEEE Trans. on Services Computing, 2018,11(2):341−353. [doi: 10.1109/TSC.2015.2501981]

[59] Wu M, Li Y. Investigations on XML-based data exchange between heterogeneous databases. In: Proc. of the 2012 Ninth Web Information Systems and Applications Conf. Haikou: IEEE, 2012. 21−24. [doi: 10.1109/WISA.2012.44]

[60] Tyagi H, Watanabe S. Universal multiparty data exchange and secret key agreement. IEEE Trans. on Information Theory, 2017, 63(7):4057−4074. [doi: 10.1109/TIT.2017.2694850]

[61] Tyagi H, Viswanath P, Watanabe S. Interactive communication for data exchange. IEEE Trans. on Information Theory, 2018,64(1): 26−37. [doi: 10.1109/TIT.2017.2769124]

[62] Hernández MA, Stolfo SJ. The merge/purge problem for large databases. In: Proc. of the ACM Sigmod Record. San Jose: ACM Press, 1995. 127−138. [doi: 10.1145/223784.223807]

[63] Doan A, Halevy A, Ives Z. Principles of Data Integration. Burlington: Elsevier, 2012. 19−58.

[64] Halevy AY. Answering queries using views: A survey. The VLDB Journal, 2001,10(4):270−294. [doi: 10.1007/s007780100054]

[65] Hull R. Managing semantic heterogeneity in databases: A theoretical prospective. In: Proc. of the 16th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems. New York: ACM Press, 1997. 51−61. [doi: 10.1145/263661.263668]

[66] Lenzerini M. Data integration: A theoretical perspective. In: Proc. of the 21st ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. New York: ACM Press, 2002. 233−246. [doi: 10.1145/543613.543644]

[67] Ullman JD. Information integration using logical views. In: Proc. of the Int'l Conf. on Database Theory. Berlin, Heidelberg: Springer-Verlag, 1997. 19−40. [doi: 10.1007/3-540-62222-5_34]

[68] Ipeirotis PG, Gravano L, Sahami M. Probe, count, and classify: Categorizing hidden Web databases. In: Proc. of the 2001 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2001). Santa Barbara: ACM Press, 2001. 67−78. [doi: 10.1145/376284.375671]

[69] Wu W, Yu C, Doan A, Meng W. An interactive clustering-based approach to integrating source query interfaces on the deep Web. In: Proc. of the 2004 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2004). Paris: ACM Press, 2004. 95−106. [doi: 10.1145/1007568.1007582]

[70] He H, Meng W, Yu C, Wu Z. Automatic integration of Web search interfaces with WISE-Integrator. The VLDB Journal, 2004, 13(3):256−273. [doi: 10.1007/s00778-004-0126-4]

[71] He H, Meng W, Yu C, Wu Z. Constructing interface schemas for search interfaces of web databases. In: Proc. of the Int'l Conf. on Web Information Systems Engineering. New York: Springer-Verlag, 2005. 29−42. [doi: 10.1007/11581062_3]

[72] Wu Z, Raghavan V, Du C, Komanduru SC, Meng W, He H, Yu C. SE-LEGO: Creating metasearch engines on demand. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Informaion Retrieval. Toronto: DBLP, 2003. 464−464. [doi: 10.1145/860435.860555]

[73] Liu W, Meng XF, Meng WY. A survey of deep Web data integration. Chinese Journal of Cumputers, 2007,30(9):1475−1489 (in Chinese with English abstract).

[74] Calì A, Calvanese D, De Giacomo G, Lenzerini M. Accessing data integration systems through conceptual schemas. In: Proc. of the Int'l Conf. on Conceptual Modeling. Berlin Heidelberg: Springer-Verlag, 2001. 270−284. [doi: 10.1007/3-540-45581-7_21]

[75] Goh CH, Bressan S, Madnick S, Siegel M. Context interchange: New features and formalisms for the intelligent integration of information. ACM Trans. on Information Systems (TOIS), 1999,17(3):270−293. [doi: 10.1145/314516.314520]

[76] Duschka OM, Genesereth MR. Answering recursive queries using views. In: Proc. of the 16th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems. Tucson: ACM Press, 1997. 109−116.

[77] Halevy AY. Theory of answering queries using views. ACM SIGMOD Record, 2000,29(4):40−47. [doi: 10.1145/369275.369284]

[78] Abiteboul S, Duschka OM. Complexity of answering queries using materialized views. In: Proc. of the 17th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems. Seattle: ACM Press, 1998. 254−263. [doi: 10.1145/275487.275516]

[79] Widom J. Research problems in data warehousing. In: Proc. of the 4th Int'l Conf. on Information and Knowledge Management. Baltimore: ACM Press, 1995. 25−30.

[80] Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM Sigmod Record, 1997,26(1):65−74. [doi: 10.1145/248603.248616]

[81] Benedikt M, Grau BC, Kostylev EV. Logical foundations of information disclosure in ontology-based data integration. Artificial Intelligence, 2018,262(2018):52−95.

[82] Tao C, Zhang L, Shi BL. Query processing for ontology-based XML data integration. Journal of Computer Research and Development, 2005,42(3):112−121 (in Chinese with English abstract).

[83] Gregory A. Data governance—Protecting and unleashing the value of your customer data assets. Journal of Direct, Data and Digital Marketing Practice, 2011,12(3):230−248. [doi: 10.1057/dddmp.2010.41]

[84] Wróbel A, Komnata K, Rudek K. IBM data governance solutions. In: Proc. of the 2017 Int'l Conf. on Behavioral, Economic, Socio-Cultural Computing (BESC). Krakow: IEEE, 2017. 1−3. [doi: 10.1109/BESC.2017.8256387]

[85] Khatri V, Brown CV. Designing data governance. Communications of the ACM, 2010,53(1):148−152. [doi: 10.1145/1629175.1629210]

[86] Wu M, Wu X. On big wisdom. Knowledge and Information Systems, 2018,58(2019):1. [doi: 10.1007/s10115-018-1282-y]

[87] Bizer C, Berners-Lee T. Linked data-the story so far. Int'l Journal on Semantic Web and Information Systems, 2009,5(3):1−22. [doi: 10.4018/jswis.2009081901]

[88] Liu Q, Li Y, Duan H, Liu Y, Qin ZG. Konwledge gragh construction techniques. Journal of Computer Research and Development, 2016,53(3):582−600 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20148228]

[89] Yang YJ, Xu B, Hu JW, Tong MH, Zhang P, Zheng L. Accurate and efficient method for constructing domain knowledge graph. Ruan Jian Xue Bao/Journal of Software, 2018,29(10):2931−2947 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5552.htm [doi: 10.13328/j.cnki.jos.005552]

[90] Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. Knowledge and Information Systems, 2008,14(1):1−37. [doi: 10.1007/s10115-007-0114-2]

[91] Wu XD, Ji SK. Comparative study on MapReduce and spark for big data analytics. Ruan Jian Xue Bao/Journal of Software, 2018, 29(6):1770−1791 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5557.htm [doi: 10.13328/j.cnki.jos.005557]

[92] Meng XF, Du ZJ. Research on the big data fusion: Issues and challenges. Journal of Computer Research and Development, 2016, 53(2):231−246 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20150874]

[93] Ni Q, Bertino E, Lobo J, Brodie C, Karat CM, Karat J, Trombeta A. Privacy-Aware role-based access control. ACM Trans. on Information and System Security (TISSEC), 2010,13(3):24. [doi: 10.1145/1805974.1805980]

附中文参考文献:

[1] 李建中,王宏志,高宏.大数据可用性的研究进展.软件学报,2016,27(7):1605−1625. http://www.jos.org.cn/1000-9825/5038.htm [doi: 10.13328/j.cnki.jos.005038]

[5] 黄刘生,田苗苗,黄河.大数据隐私保护密码技术研究综述.软件学报,2015,26(4):945−959. http://www.jos.org.cn/1000-9825/4794.htm [doi: 10.13328/j.cnki.jos.004794]

[6] 张绍华,潘蓉,宗宇伟.大数据治理与服务.上海:上海科学技术出版社,2016.1−224.

[8] 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. [doi: 10.16383/j.aas.2016.c160239]

[12] 冯登国,张敏,张妍,徐震.云计算安全研究.软件学报,2011,22(1):71−83. http://www.jos.org.cn/1000-9825/3958.htm [doi: 10.3724/SP.J.1001.2011.03958]

[42] 杨东华,李宁宁,王宏志,李建中,高宏.基于任务合并的并行大数据清洗过程优化.计算机学报,2016,39(1):97−108. [doi: 10.11897/SP.J.1016.2016.00097]

[43] 郭志懋,周傲英.数据质量和数据清洗研究综述.软件学报,2002,13(11):2076−2107. http://www.jos.org.cn/1000-9825/13/2076.htm [doi: 10.13328/j.cnki.jos.2002.11.003]

[73] 刘伟,孟小峰,孟卫一.Deep Web 数据集成研究综述.计算机学报,2007,30(9):1475−1489.

[82] 陶春,张亮,施伯乐.基于本体的 XML 数据集成的查询处理.计算机研究与发展,2005,42(3):112−121.

[88] 刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述.计算机研究与发展,2016,53(3):582-600. [doi: 10.7544/issn1000-1239.2016.20148228]

[89] 杨玉基,许斌,胡家威,仝美涵,张鹏,郑莉.一种准确而高效的领域知识图谱构建方法.软件学报,2018,29(10):2931−2947. http://www.jos.org.cn/1000-9825/5552.htm [doi: 10.13328/j.cnki.jos.005552]

[91] 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):1770−1791. http://www.jos.org.cn/1000-9825/5557.htm [doi: 10.13328/j.cnki.jos.005557]

[92] 孟小峰,杜治娟.大数据融合研究:问题与挑战.计算机研究与发展,2016,53(2):231−246. [doi: 10.7544/issn1000-1239.2016.20150874]

吴信东(1963－),男,安徽枞阳人,博士,教授,博士生导师,主要研究领域为数据挖掘,大数据分析,知识工程.

董丙冰(1996－),女,学士,主要研究领域为数据挖掘,数据治理.

堵新政(1989－),男,学士,软件开发工程师,主要研究领域为计算机辅助几何设计,计算机图形学,科学计算可视化,医学图像处理.

杨威(1982－),男,硕士,主要研究领域为大数据,知识图谱,数据治理.