政府治理大数据的共享、集成与融合

金澈清[1]，陈晋川[2]，刘威[3]，张召[1]
1. 华东师范大学数据科学与工程学院，上海 200062；2. 中国人民大学信息学院，北京 100872；
3. 中山大学数据科学与计算机学院，广东 广州 510006

摘要
为支持政府治理方法科学化、过程智能化、结果精细化，政府治理大数据共享、集成与融合不能局限于提供数据访问接口，而是要从语义层面发现实体、找出关联关系以及演化过程。然而，政府治理大数据的多源、异构、动态、海量、孤岛化特性却使之面临严峻挑战。系统性回顾了大规模分布式异构数据共享、集成、融合的基础理论和方法，并指出了构建面向政府治理大数据的高可信共享、高精准集成、高效率融合技术的迫切性。

关键词
政府治理；数据共享；数据融合；数据集成

中图分类号：TP315 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020012

Sharing, integration and fusion of government-governance big data

JIN Cheqing[1], CHEN Jinchuan[2], LIU Wei[3], ZHANG Zhao[1]
1. School of Data Science and Engineering, East China Normal University, Shanghai 200062, China
2. School of Information, Renmin University of China, Beijing 100872, China
3. School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China

Abstract
To make governance measure scientific, governance progress intelligent, and governance result refined, sharing, integration, and fusion of government-governance big data cannot be limited to data accessing interface, but novel techniques to resolve entities according to the semantics, find out the relationship among entities, and track entities' evolution process. However, this task is challenging due to some characteristics of the government-governance big data, such as multi-source, heterogeneous, dynamic, massive and isolate. The basic theories and methods for large-scale distributed heterogeneous data sharing, integration and fusion were studied, and several important topics to construct high-trustworthy data sharing, high-precise data integration, high-efficient data fusion for government-governance big data in future were pointed out.

Key words
government-governance, data sharing, data fusion, data integration

1 引言

政府治理是指政府行政系统与其他相关主体一道对社会公共事务的治理。政府治理是在坚持中国特色社会主义制度的前提下，破除不适应生产力发展的制度，释放生产力和社会活力[1]。传统的政府管理模式强调政府基于科层制体系而形成的垂直型结构，政府作为单部门封闭式行政；而政府治理模式则强调政府与其他主体（包括企事业单位、行业协会等）之间紧密配合，协作式行政。参与治理的主体之间保持信息沟通顺畅，能真实、客观、全面地描述事态现状，预测发展方向，从而使治理方法科学化、过程智能化、结果精细化。

政府治理大数据泛指支持政府治理行为的所有数据，而非单指政务大数据。政府治理大数据改变人们的思维方式和决策过程，为政府治理能力和治理体系现代化提供强大驱动力。在过去相当长的时间内，政府部门在进行决策时能够采用的数据相对有限，而且准确度较低，因而在决策过程中会融入较多主观因素，不够精准。进入21世纪以来，数据采集方式更加丰富，采集范围更加广阔，经由各个渠道汇集而成的海量行为信息深刻而生动地刻画了治理对象。例如，城市的手机信令数据能够反映出城市的整体交通状况，特别是拥堵路段的位置，交通管理部门可据此优化交通路线；再例如，我国不同地区的能源消耗总量以及工业用电、居民用电的占比能够反映出该地区的工业和经济发展活力，为宏观调控提供依据。这些行为数据规模宏大、到达速度快、类型多样，基于这些数据的分析结果有助于决策者从多层面、多角度洞察和理解社会现象，以进行科学决策。政府治理大数据的出现改变了以往认为人类行为难以预测的旧观点，可以通过电子踪迹监测和预测人类的行为习惯，使政府能够提前进行科学决策，并为用户提供便利、快捷、无缝集成的一体化服务。

有效汇集不同来源的数据能帮助政府从不同维度审视治理对象。例如，当人们衡量一个商圈的活力时，需要了解该商圈的人流量（电信信令数据）、消费水平（支付宝、微信支付、银联刷卡数据）、口碑（互联网、社交媒体）等，而这些数据由不同机构采集和维护，并不隶属于单一机构。因此，政府治理过程是一个多治理主体共同参与的协作式治理。尽管众多企事业单位已经在过去几十年的信息化建设中累积了大量数据，但是由于行政管理和信息技术等方面的障碍，存在严重的信息孤岛现象，大量数据无法被共享使用以支持政府治理。2016年5月，李克强总理在全国推进简政放权放管结合优化服务改革电视电话会议上指出："目前我国信息数据资源80%以上掌握在各级政府部门手里，'深藏闺中'是极大浪费。"《广东省"数字政府"建设总体规划（2018—2020年）》指出：省直单位现有政务信息系统1 068个，其中省级垂直系统475个，建设20个以上系统的单位21个，其中存在37个网络孤岛、44个机房孤岛和超过4 000类数据孤岛。

数据孤岛意味着数据没有被充分共享、难以有效集成、有待深度融合。**表1**总结了数据孤岛现象带来的3个问题以及拟达成的目标。

（1）政府治理大数据没有充分共享

数据共享机制描述数据发布者、使用者（有些场景下还包括监管者）之间的交互规则，使信息能够顺利流转。常用的文件共享机制支持在不同实体之间以文件形式共享信息，但是忽视了各参与实体的其他诉求。例如，数据发布者想对数据进行确权，充分了解数据的传播过程，并可在必要时限制

数据传播；数据使用者想确保所获取的数据是真实、完整、一致的；而数据监管者则期望能确保相关数据共享规章制度被严格贯彻、认真遵循。当前，由于数据未充分共享而造成的治理疏漏并不少见。例如，由于各省间的婚姻系统不联网，2019年1月江苏男子张某被曝分别与3位女士登记结婚。

（2）政府治理大数据难以有效集成

数据规模、来源和质量均深刻影响着数据集成的难度。政府治理场景面向的治理对象涉及面广，与之相关的数据规模宏大，来源广泛。为了使场景描述更加准确，部分政府治理场景使用互联网上的开放数据，这使得信息来源更加复杂，数据源的挑选愈加困难；由于数据平台构建的历史因素、数据采集设备的精度因素、人工录入因素、不同业务领域导致数据标准存在差异，政府治理大数据的质量不高、规格不一，有效集成的难度很大。

（3）政府治理大数据有待深度融合

将低价值密度的大数据通过数据融合转换为高价值密度的知识是政府治理大数据管理的宗旨，而精准发现大数据中的实体及其语义关联是提升大数据价值特征的核心。例如，中国人民银行为国内的法人单位建立资信评级时需要融合多源信息，并挖掘深层的语义关系。行为数据会随着时间增加而动态变化，在某些场景下甚至会急剧变化。例如，信用评级机构基于日常行为数据对法人（或自然人）评级；但在极端情况下（例如经营不善等），法人（或自然人）可能会表现出与其当前等级明显不符的行为。例如，2019年出现多起网贷平台跑路事件，如果能预先将数据进行深度融合，将能有效对网贷平台的信用度进行预警，从而防范社会风险。

近几年来，我国在加快数据开放与共享、推进政府治理创新方面已经前进了一大步。一方面，各地方政府积极推出便民平台，改进工作流程，让数据多跑路，让群众少跑腿，使得用户只需要访问一个平台就能够办理多项业务，例如广东省的“粤省事”、上海市的“一网通办”、浙江省的“浙里办”等App平台。另一方面，各地积极基于大数据技术构建智慧城市，提升城市治理的智能化水平。例如，浙江省“城市大脑”已经形成了一批成熟应用，整合多源信息，在交通等领域进行了创新。可以看出，尽管政府治理大数据的共享与融合能够显著提升政府的治理水平，并且已经在部分地区和领域中有了良好的示范效应，但是还需要努力克服存在的挑战，以深化政府治理体系和治理能力现代化建设。部分学者也已经意识到大数据融合方面的问题与挑战[2]，本文聚焦政府治理领域的数据共享与融合。

表1 数据孤岛引发的问题与目标

问题	目标
数据没有充分共享	防止数据被篡改，数据共享流程合规化，施政措施可溯源
数据难以有效集成	在开放数据中选择合适的数据源，生成全局模式
数据有待深度融合	高效生成实体，动态维护实体之间的关联关系

2 数据共享

数据共享旨在破除不同治理实体之间的数字藩篱，搭建数据流通渠道，在共享过程中需要综合考虑架构、隐私、合规和溯源等因素。首先，不同数据共享架构能够支持的功能差异显著，使用方需要结合应用场景理性选取；其次，隐私保护是数据共享的基础诉求之一，为了鼓励用户共享数据以推进协作，必须要确保用户隐私安全；再次，整个共享过程的合规化操作可避免其他主观因素的影响，增强整体可信度；最后，溯源机制在多方参与的机制中起到事中监管、事后追责的作用，维护整个

过程正常推进。

2.1 数据共享架构

按照数据发布者和使用者构成的网络拓扑不同，可将数据共享架构划分成3种。

第一种也是最常用的数据共享架构是集中式架构。参与政府治理的所有主体之间预先约定好一个公共服务器，继而主动将数据传送到该服务器。服务器设定数据访问规则，允许参与治理的主体以不同权限访问服务器上的数据，例如Web服务器或者文件传输协议（file transfer protocol，FTP）服务器。尽管这种架构的结构简单，但是仍然存在明显的不足之处。首先，在网络部署上可能引发争议。若治理实体之间存在上下级关系，则上级实体可以通过行政手段决定网络部署方式；而若治理实体之间是平级关系（无隶属关系），则服务器由哪个单位进行管理会成为焦点议题。其次，这种集中式架构还存在单点故障和性能缺陷，一旦由于黑客攻击、软硬件故障等原因导致服务器宕机，则所有数据访问服务均会被迫中止，且整个系统的数据访问能力受限于服务器的性能，当大量数据访问请求同时到达时，系统性能会急剧降低。最后，这种架构无法确保数据的可信性，具有管理员权限的治理实体成为强势的一方，具备数据修改的能力，而不具备管理员权限的治理实体则处于相对弱势的一方（通常不将管理员权限赋予所有实体，以保障系统安全性）。

第二种架构基于对等网络，将数据分散部署在整个网络中，该网络中没有特定的服务器节点，所有节点既可提供数据，又可消费数据。由于（多副本）数据分散在不同网络节点，而非单一节点上，因此可避免单点故障缺陷，且可扩展性更强。对等网络的共享方式包括非结构化对等网络和结构化对等网络2种。非结构化对等网络较为简单，对节点之间的拓扑结构并无特别约定，只需要记录邻接节点信息，但是无法保证以低时间复杂度来处理数据查询请求，典型系统如Gnutella。结构化对等网络则对网络节点进行精心部署，使用分布式哈希表（distributed Hash table，DHT）来提升数据访问效率，典型的结构化对等网络包括Chord[3]。与第一种架构相比，这种架构的最大优势是能够克服单点故障，然而这种架构仍然无法确保数据的可信性，不排除数据在共享过程中被篡改的可能。

第三种架构通过区块链来实现数据共享。区块链技术利用共识机制在不可信网络中为各参与方构建信任关系，确保数据不易被篡改。区块链系统通常可以被划分为公有链和许可链。公有链面向全网公开，无用户授权机制，如比特币、以太坊（Ethereum）等；许可链有用户授权机制，仅允许授权的用户和节点加入，如超级账本（Fabric）等。由于现有区块链系统的数据管理能力较弱，一些学者尝试将区块链与数据库技术结合，提升数据管理性能，华东师范大学提出的师大链数据库（semantics empowered blockchain database，SEBDB）就是构建于许可链之上的区块链数据库系统[4]。典型的共识协议包括工作量证明（proof of work，POW）、权益证明（proof of stake，POS）和实用拜占庭协议（practical Byzantine fault tolerance，PBFT）及其变种。工作量证明机制根据各节点的计算资源进行投票，并要求可信节点控制的计算资源多于一半[5]；权益证明机制根据各用户拥有的权益比重进行投票；实用拜占庭协议能够在$n \geqslant 3f+1$（n是网络节点数，f是不可信节点数）的条件下解决拜占庭将军问题[6]。

表2列举了3种数据共享架构及其特点。

表 2 数据共享架构

架构	代表性系统	是否存在单点故障	是否支持溯源	是否去信任	数据是否易篡改	政府治理应用情况
集中式架构	Web、FTP	是	否	否	是	普遍使用
对等网络架构	非结构化（例如Napster）、结构化（例如Chord）	否	否	否	是	较少使用
区块链系统	公有链（例如以太坊）、许可链（例如超级账本、师大链数据库）	否	是	是	否	较少使用，前景广阔

2.2 数据隐私保护

政府治理大数据共享必须重视隐私保护。我国早已立法明确政府信息公开中“保护个人隐私”的原则。《中华人民共和国政府信息公开条例》中第十四条规定：行政机关不得公开涉及国家秘密、商业秘密、个人隐私的政府信息。但是，经权利人同意公开或者行政机关认为不公开可能对公共利益造成重大影响的涉及商业秘密、个人隐私的政府信息，可以予以公开。由于用户隐私泄露而造成负面社会效应的案例屡见不鲜。在大数据背景下，当来自不同数据源的数据经过整合之后，数据相互关联就会揭示更多知识。例如，2006年8月，美国在线（American Online，AOL）公布了大量旧的搜索查询数据（数据已经经过脱敏处理，包括用户名称和地址等个人信息），《纽约时报》在几天内综合分析“60岁的单身男性”“有益健康的茶叶”“利尔本的园丁”等搜索记录之后，发现第4417749号代表是佐治亚州利尔本的一位62岁的寡妇塞尔玛·阿诺德[7]。典型的隐私保护技术包括匿名化[8]、加密处理[9]和多方隐私技术等。匿名化技术将数据的关键部分模糊化处理，从而保护用户隐私，例如，k-匿名技术就是将当前数据项与其他至少k-1个数据项进行模糊化处理，使得这k个数据项之间不可区分。加密处理将明文转化为密文，以保护私密信息。多方隐私保护下的数据集成技术（或称多方PPRL）还处于起步阶段，主要支持精确匹配，例如将各个数据源的记录编码，然后传入另一方进行对比[10]。参考文献[11]提出了一种基于安全多方计算的精确匹配方法，参考文献[12]提出一种基于k-匿名的支持多约束条件的隐私保护方法。

除了上述以软件和算法的方式来保护用户隐私之外，还可以通过构建细粒度的访问控制以及基于可信执行环境（trusted execution environment，TEE）来保障数据隐私[13]。鉴于政府治理大数据分别属于不同治理实体，且不同治理实体的访问权限不同，可以借鉴面向对象设计（object-oriented design，OOD）的思想，设定多层级访问权限，包括开放可访问、敏感不可访问、部分用户可访问等。通过分级权限来限制对数据的访问。TEE可保护敏感而又无法脱敏的数据。软件防护扩展（software guard extensions，SGX）是典型的TEE，它将敏感数据和操作转移至Enclave（即SGX的可信内存）中进行处理，而数据和操作在其他地方以密文的方式存在。借助于可信硬件的数据保护方式比同态加密、零知识证明等传统密码学方法更灵活和高效。

2.3 共享流程合规化

数据共享流程由多个治理主体共同参

与，并遵循特定管理制度。程序透明增强了共享流程的公平性。为确保整个流程自动化执行，避免人为干预，可将相关规章制度预先编制成可自动运行的程序。当外部条件满足时，该程序自动被触发运行，整体上流程不需要人工介入。智能合约就是一段自动运行、可验证的程序，以数字化方式让各参与方履行特定承诺。在基于智能合约的数据共享流程自动化机制中，行政部门将数据共享的管理制度转化为智能合约代码，采用形式化方式严格定义各参与主体的义务，明确每条义务的实施主体、前提条件、具体内容以及完成期限；同时，定义一项义务的各种状态，如激活、就绪、满足、过期以及违约等，并分析各状态之间的转换条件。当某个参与主体未及时履行预先约定的义务时，管理部门作为实施主体对该参与主体进行处罚。管理制度的运行实例可等价为一个有穷状态机，其运行机制由组成此制度的所有义务共同决定。管理部门可使用图形化建模工具来制定制度，将规章制度自动生成对应的状态机，并展现制度的运行过程，自动分析并显示异常的运行状态，为管理部门对制度改进提供决策支持。

2.4 数据溯源

施政效果评估和责任追究是政府治理的重要内容。基于政府治理行为大数据开展溯源分析，能够评估施政效果和责任认定。数据溯源是指数据产生并随时间推移而演变的过程。2017年，国家食品药品监督管理总局发布了《关于食品生产经营企业建立食品安全追溯体系的若干规定》，推动食品生产经营企业建立食品安全追溯体系。基于关系数据库的溯源系统有DBNotes[14]、Perm[15]、Trio[16]等。DBNotes系统基于关系数据库对溯源标注信息进行管理。Perm系统利用查询重写规则改写SQL查询，以追踪数据溯源信息。Trio系统是一个不确定数据库上的数据世系管理系统，将数据不确定性和溯源信息紧密整合在一起。区块链系统将所有操作按照时间顺序进行存储，难以篡改，且新数据只能以添加的方式加入区块链系统之中，能有效提供数据溯源功能。参考文献[17]研究了如何基于区块链设计食品安全溯源体系。

3 数据集成

政府治理大数据来源丰富、领域多样、发展历程迥异，因而不同来源的数据格式不一，且存在质量问题。数据集成旨在以统一模式访问不同数据，包括数据源选择和数据模式匹配2个方面。

3.1 数据源选择

精准选择数据源是实现数据集成结果准确的前提。当数据源数量较少时，使用人工方式就能够较为有效地筛选出合适的数据源。而当数据来源较多时，难以借助人工方式有效地挑选出合适的数据源。特别地，如果尝试结合互联网数据进行治理，则数据源的数量就急剧增多，需要设计算法来高效、精准地选择数据源，以解决应用需求。由于政府治理大数据包含大量行为数据，在选择数据源时不仅需要考虑更加广泛的质量维度，以解决面向实体和行为数据的集成，还要根据目标模式自动构建候选模式集成处理路径。数据源选择方法可分为按需驱动的选择方法和基于多质量维度的选择方法2种。

（1）按需驱动的数据源选择方法

这种方法在目标模式和数据源模式之间匹配关联信息，反向构建出包含多个模式集成处理路径的候选集合，并最终找出

满足集成需求的数据源模式结构与集成方式。目标模式通常是一个以实体为核心的关联数据整体，其结构可以映射到共享数据的模式关联图上。首先，基于共享数据生成模式关联图，采用基于图结构的查询方法寻找与目标模式匹配的候选模式集合。然后，基于候选数据模式间的匹配关系，利用数据集成算子创建由集成操作构成的有向无环图集合。最后，进一步提出约简策略，以减少不必要的操作，降低数据集成的运算代价。

（2）基于多质量维度的数据源选择方法

这种方法通过面向数据质量的数据源选择策略管理参与数据集成的数据源，从而保证集成结果在完整性、精确性和时效性等维度上的质量需求。首先，从数据源的多质量维度（同一性、完备性、精确性、时效性以及综合质量）构建数据源质量评价模型，用于独立评价数据源的各维度质量；其次，定义多维度的综合评价模型和数据源集成代价评估模型；最后，构建利益代价模型，并以此选择集成数据源。参考文献[18]意识到数据准确性的重要性，提出了面向数据融合的数据源选择方法，从数据质量和集成代价的平衡上选择数据源。参考文献[19]进一步提出了融合覆盖率、新鲜性和准确性质量等多个维度的数据源选择方法，并在此基础上实现了数据源选择系统SourceSight[20]。

3.2 数据模式匹配

数据模式匹配内容丰富，包括基于实例的匹配、基于模式信息的匹配、混合匹配等[21]。在20世纪90年代，数据模式匹配主要以描述逻辑（description logic）为主，包括使用视图处理查询。参考文献[22]在样本数据上运用机器学习方法发现属性之间的映射关系。参考文献[23]基于目标数据样本获取匹配路径，而不需要用户指定连接路径。当数据源比较多时，一些学者也采用一些半自动化映射技术，例如learn to match等。为了应对大数据挑战，有不少工作采用即付即用（pay as you go）方式，即采用简单的模式匹配方法搭建一个初步可用的数据服务系统，然后在系统服务过程中根据用户的反馈逐渐更新模式匹配[24-25]。近期有部分工作采用机器学习特别是深度学习来提高模式匹配的效果，包括采用概率推理方法从所有候选模式中找出最优结果[26]。

数据模式匹配的一个难点在于部分数据源质量低下、缺乏表头信息、规模庞大且增长迅速，无法精确匹配模式。在此情况下，可以采用概率模式匹配方法筛选出潜在的匹配模式，并评估其可信度。当数据源的数目较多时，简单罗列出所有潜在的模式匹配组合及其发生概率的计算开销太大，需要灵活运用剪枝策略缩小搜索空间，构造出一个包含少量模式匹配组合的候选集合，并最终生成概率模式。参考文献[27]提出了一种基于概率模型的全局数据模式生成方法。另外，为解决开放数据规模庞大的问题，还可以划分原始数据，将任务分摊到不同节点之中，采用分布式架构提升效率。例如，以Spark为代表的通用并行处理框架具备良好的水平扩展能力，可支持海量开放数据的模式匹配。

4 数据融合

数据融合指将来自政府治理中不同数据源的同一实体（如企业、个人）的不同表象融合成单一表象，消除潜在的数据冲突[28]。数据融合包括实体匹配、实体链接与关联、动态数据的语义关联3个方面。首先，通过实体匹配在多个数据源中找出指向同一实体的

记录；其次，需要明确实体之间的链接与关联关系；最后，实体本身以及实体之间的关联关系都会随着时间推移而不断演化。

4.1 实体匹配

实体匹配也被称为记录连接、重复数据删除，旨在找出存在于多个数据源中但指向同一实体的记录集合。例如，同一企业对应的地址信息在政府的不同数据源中，往往存在多种表述方式。通过实体匹配不仅可以减少数据的冗余，而且拼接碎片化数据可以提高数据质量。当前基于实体局部结构特性（实体属性或实体间关系）进行匹配的方法具有复杂性较高的缺点[29]。Chaudhuri S等人[30]在索引和磁盘访问方面进行了优化，从而提高了运行效率；Firmani D等人[31]研究了如何在线进行实体匹配；Konda P等人[32]设计了Magellan系统，使用数据科学栈完成实体匹配任务，从而提升了效率。具有演化属性的实体匹配研究正在展开，但精准性和效率都有待改善[33-34]。

此外，还可以充分利用数据间丰富的关联关系从以下3个方面提升实体匹配的准确性和效率。其一，利用图能够有效表示数据对象间拓扑关系的能力，可以将共享集成的结构化数据集构建为数据对象关系图，再基于图迭代进行实体匹配；对象之间的相似度可以综合属性相似度、结构相似度、语义路径相似度来计算；针对复杂数据记录匹配，可以依据数据之间的关联关系构建有向依赖图，按依赖关系确定匹配顺序，减少匹配次数。其二，可以综合采用哈希方法和位计算提高匹配准确性和效率。针对快速到来的时序数据，采用哈希方法对数据记录进行快速分块，不仅具有高效率和高准确性，且不需要进行全局数据排序。可优先选择识别度高的属性进行哈希处理，提高分块中候选匹配对的数量，对于块可匹配估计方法，可以结合哈希计算和位计算提高块中可匹配候选对的准确性和效率。优先选择块匹配冗余度高的分块进行实体匹配，从而在最短时间内获得更多的匹配对。其三，可以通过分布式架构提高实体匹配的效率。在利用分布式并行处理平台的同时，尽量减少通信代价，可以采用多属性哈希实现更精准的分块；均衡分布节点上的处理任务，降低总匹配时间，可以通过构建分层的分块模型和优化组合来均衡不同处理节点上的匹配任务。

4.2 实体链接与关联

政府治理中的同一实体通常并不仅仅在一个系统中出现，而是存在于多个系统中，且互相链接与关联。例如，同一企业法人的信息既有来自工商管理部门系统的基本信息，又有蕴含于开放的互联网中的大量交互行为信息。为了更全面地刻画企业的诚信特征，需要将互联网中的多个记录与工商管理部门知识库中的该实体链接起来。实体链接技术通过基于属性的模型和基于关系的模型在不同系统中找出针对同一实体的描述记录，从而形成更加全面的实体信息，其中，涉及实体链接、消除实体歧义和复杂数据之间实体关联。实体链接与关联通过建立知识库中的知识条目与待消歧实体的对应关系实现消歧，它包含2个步骤：候选集生成、候选实体消歧[35]。候选集生成的方法主要有基于信息检索的方法[36]、基于查询表述上下文的方法[37]等。参考文献[38]提出了一种减少候选集规模的方法。候选实体消歧方法大致有2类：基于相似度计算的实体链接方法、基于有监督学习的实体链接方法。其中，基于有监督学习的实体链接方法在性能上有进一步改进[39-41]。由于实体语义模糊和异构网络知识有限，

Shen W等人[42]考虑了实体的流行度，提出了基于概率链接模型的知识流行度算法，将链接模型以高可靠性映射到上下文信息，迭代丰富网络实体，从而提高链接性能。

为了提升实体链接和关联的效率，可以从以下3个方面进行改进。其一，考虑政府领域、跨系统语料变化和社交媒体短文本等特点，基于用户行为特征进行实体关联，即将用户行为特征抽象为时间、地点和主题三维模型，通过学习训练用户行为数据的多维度特征，聚类用户的三维行为特征，完善用户的行为模式；再构建基于用户行为聚类特征的相似度度量模型，改善基于用户行为特征的用户匹配准确性。其二，为克服复杂文本、噪声数据和半结构化数据的挑战，可以通过深度学习方法研究跨系统结构化和非结构化数据之间实体关联技术，提高实体关联模型的鲁棒性和扩展性。其三，利用政府治理领域知识和机器学习方法、结构化数据相似性判别技术，聚类同一实体的所有记录，保证高内聚、低歧义。在跨系统实体链接和关联过程中存在数据冲突，可基于各系统的数据源质量解决冲突问题。

4.3 动态数据的语义关联

在政府治理场景中，实体会随着时间推移而变化，需要准确关联用户行为，以捕获序列事件的演化规律。例如，一个法人（用户）的信用会随着时间的推移而发生改变，尽快检测到语义变化有助于及时制定应对措施。实体的属性值会随时间变化，同一实体对应的多条记录会出现不一致的情况，为了发掘动态数据中的语义关联，需要细粒度地分析变化。文本词语会随着时间发生语义变化，参考文献[43]提出了动态统计模型以学习时间感知的词语表示，获取动态数据中语义关联。尤其是随着移动社交网络的发展，同一实体在空间和时间上会有多样记录，参考文献[44]提出了基于K-L散度的关联模型链接两类数据源中的时空记录，并通过时间和空间过滤机制降低匹配的搜索空间。针对高动态性及实效敏感的数据源，参考文献[45]提出了扩散随机梯度下降算法，对不同样本分配实效感知权重，增强模型对动态数据的处理能力。在非结构化数据中，传统词嵌入方法无法表征语料信息的变化历史，参考文献[46]提出了时态词向量法，可以有效分析实体的演化过程。

为提升动态数据的语义关联效率，可以从以下3个方面进行改进。首先，可以面向演化数据对实体进行关联，为精准关联具有演化特性的同一实体，可定义精准的时间模型和相应的相似度计算算法，并通过基于深度学习的动态分布表示法刻画语义迁移和涌现，提高关联演化实体的准确性。其次，针对实体关联关系的实时演化技术，为结合行为数据准确关联用户或事件的演化规律，克服由于实体名称改变或隐匿造成的实体重复副本，可定义结合实体语义相关性、实体关联性和实体的时序特征的事件演化模型，为每个实体构建时间活动路径，通过路径相似度判别潜在相同实体。最后，为解决现有实体关联预测技术大多针对静态数据的问题，可以考虑增量式的动态语义关联维护技术，通过结合已有匹配结果实现快速计算，从而捕获用户的演化特性。

5 案例分析

本文成稿之时，正逢新型冠状病毒引发的肺炎疫情在我国肆虐，疫情凶猛。截至2020年3月1日24时，据31个省（自治区、直辖市）和新疆生产建设兵团报告，累积报告确诊病例80 026例，确诊病例远超17年前的非典疫情。全国上下众志成城、万

众一心，以极大的努力和决心投入抗击疫情的工作之中。作为数据科学研究人员，笔者也在深入反思这次抗击疫情过程中暴露出来的问题是否能够以更高效的方式解决。以下是政府治理大数据的共享、集成与融合方面面临的一些实际挑战。

（1）信息孤岛现象依然存在

科学应对疫情的前提是能够准确了解与疫情相关的关键性数据。但是在对抗疫情的过程中，一些关键性的数字掌握得不够及时、准确，例如当地医疗物资的储备和消耗情况、区域内的医疗物资的生产能力和调拨能力等。相关信息的互联互通有助于统一决策、统一规划，以充分利用有限的资源抗击疫情。

（2）确保共享数据的真实性

疫情暴发之后，网上谣言满天飞，并且通过社交工具迅速传播。造谣一张嘴，辟谣跑断腿。数据的真实性非常重要。如何通过技术手段识别信息的真伪，如何及时发现并切断虚假的甚至是恶意的信息传播，如何分析谣言传播的路径等，都非常值得进一步探讨。

（3）确保共享数据可追溯，提升可信性

由于疫情暴发具有突然性，这使得医疗物资（例如口罩）成为紧俏物资，不少厂商纷纷加大生产力度，支援抗疫一线。但是在这种紧急情况下，仍然有不法商家生产假冒伪劣产品，借以牟利，造成了恶劣的社会影响。在此，如果能够构建基于区块链技术的物资数据可溯源平台，则能够排除伪劣产品，保障物资安全。另外，在本次疫情中，世界各地的爱心人士捐款捐物，非常踊跃。捐赠系统中数据的透明性和可信性能够极大地影响捐赠热忱。

（4）综合多个数据来源的数据集成

将不同来源的数据集成起来能够增加对整体事件的透视性。在抗击疫情过程中，数据来源众多，及时集成相关数据才可客观评判事态发展。在2020年1月29日中央指导组派出督查组赶赴黄冈市进行督查核查时，黄冈市卫生健康委员会主任对黄冈市定点医院收治能力、核酸检测能力的明确数据等均不了解。推而广之，在政府治理过程中实时汇聚多源数据，可以辅助领导层快速应对突发事件。

（5）实体关联与融合提升服务民众

疫情暴发以来，各地政府和机构通过不同渠道发布疫情通报，不仅有病例数据、密切接触者寻找通知，也有关于公共交通车次的调整信息。这些信息来源杂、数量大、增长快。如果能够从实体层级汇聚多源信息，并且找出不同实体之间的关联关系，则能够更加清晰地表明疫情发展情况。

（6）动态数据的实时演化

疫情的发展随时间变化而不断演变，从疫情暴发以来，腾讯、新浪等门户网站每日实时发布疫情地图，显示不同地域确诊案例、疑似案例、重症案例等关键信息的变化轨迹。分析动态数据的实时演化过程能够让人们更加清晰地了解疫情发展的整个过程以及各项措施所取得的成效，从而不断调整应对方案。

6 结束语

综上所述，政府治理大数据的共享、集成与融合需要从理论、机制、实践等方面进行深入的研究。现有的方法都存在一些不足。为了构建面向政府治理大数据的高可信共享模型、高精准集成机制、高效率融合机理，还需要从以下3个方面进行努力。首先，研究政府治理大数据高可靠共享技术，包括可确保所共享数据可信、可验证的数据证明机制，可复现数据演化过程的数据溯源技术，可确保数据管理制度自动实施的流程合约化机制等。其次，研

究政府治理大数据高精度集成技术，包括在数据抽取过程中的持续闭环迭代能力、在数据源选择过程中基于目标约束的自动优选能力、在模式匹配过程中的劣质数据容忍能力等。最后，研究政府治理大数据高效率融合技术，包括在实体识别阶段采用分布式计算机系统提升可扩展性、在跨系统实体链接与关联阶段充分结合用户行为数据提升效率、在实体演化分析方面采用增量式策略提升处理效率等。

参考文献:

[1] 王浦劬. 国家治理、政府治理和社会治理的基本含义及其相互关系辨析[J]. 社会学评论, 2014, 2(3): 12-20.
WANG P Q. The inherent meaning and interrelationship of state governance, government administration and social governance[J]. Sociological Review of China, 2014, 2(3): 12-20.

[2] 孟小峰, 杜治娟. 大数据融合研究: 问题和挑战[J]. 计算机研究与发展, 2016, 53(2): 231-246.
MENG X F, DU Z J. Research on the big data fusion: issues and challenges[J]. Journal of Computer Research and Development, 2016, 53(2): 231-246.

[3] STOICA I, MORRIS R, LIBEN-NOWELL D, et al. Chord: a scalable peer-to-peer lookup protocol for internet applications[J]. IEEE/ACM Transactions on Networking, 2003, 11(1): 17-32.

[4] ZHU Y C, ZHANG Z, JIN C Q, et al. SEBDB: semantics empowered blockchain database[C]//The 35th IEEE International Conference on Data Engineering, April 8-11, 2019, Macao, China. Piscataway: IEEE Press, 2019: 1820-1831.

[5] ASPNES J, JACKSON C, KRISHNAMURTHY A. Exposing computationally-challenged Byzantine impostors[R]. 2005.

[6] LAMPORT L, SHOSTAK R, PEASE M. The Byzantine generals problem[J]. ACM Transactions on Programming Languages and Systems, 1982, 4(3): 382-401.

[7] 维克托·迈尔-舍恩伯格, 肯尼思·库克耶. 大数据时代: 生活、工作与思维的大变革[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.
MAYER-SCHÖNBERGER V, CUKIER K. Big data: a revolution that will transform how we live, work, and think[M]. Translated by SHENG Y Y, ZHOU T. Hangzhou: Zhejiang People's Publishing House, 2013.

[8] 王智慧, 许俭, 汪卫, 等. 一种基于聚类的数据匿名方法[J]. 软件学报, 2010, 21(4): 680-693.
WANG Z H, XU J, WANG W, et al. Clustering-based approach for data anonymization[J]. Journal of Software, 2010, 21(4): 680-693.

[9] 黄刘生, 田苗苗, 黄河. 大数据隐私保护密码技术研究综述[J]. 软件学报, 2015, 26(4): 945-959.
HUANG L S, TIAN M M, HUANG H. Preserving privacy in big data: a survey from the cryptographic perspective[J]. Journal of Software, 2015, 26(4): 945-959.

[10] QUANTIN C, BOUZELAT H, ALLAERT F, et al. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure[J]. International Journal of Medical Informatics, 1998, 49(1): 117-122.

[11] O'KEEFE C M, YUNG M, GU L, et al. Privacy-preserving data linkage protocols[C]// The 2004 ACM Workshop on Privacy in the Electronic Society, October 28, 2004, Washington, DC, USA. New York: ACM Press, 2004: 94-102.

[12] 杨晓春, 刘向宇, 王斌, 等. 支持多约束的K-匿名化方法[J]. 软件学报, 2006, 17(5): 1222-1231.
YANG X C, LIU X Y, WANG B, et al. K-anonymization approaches for supporting multiple constraints[J]. Journal of Software, 2006, 17(5): 1222-1231.

[13] MCGILLION B, DETTENBORN T,

NYMAN T, et al. Open-TEE: an open virtual trusted execution environment[C]// 2015 IEEE Trustcom/BigDataSE/ISPA, August 20-22, 2015, Helsinki, Finland. Piscataway: IEEE Press, 2015: 400-407

[14] CHITICARIU L, TAN W C, GAURAV V. DBNotes: a post-it system for relational databases based on provenance[C]//The 24th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, USA. New York: ACM Press, 2005: 942-944.

[15] GLAVIC B, ALONSO G. Perm: processing provenance and data on the same data model through query rewriting[C]// 2009 IEEE 25th International Conference on Data Engineering, March 29-April 2, 2009, Shanghai, China. Piscataway: IEEE Press, 2009: 174-785.

[16] JENNIFER W. Trio: a system for integrated management of data, accuracy, and lineage[C]//The 2nd Biennial Conference on Innovative Data System Research, January 4-7, 2005, Pacific Grove, USA. [S.l.:s.n.], 2005: 262-276.

[17] 李明佳, 汪登, 曾小珊, 等. 基于区块链的食品安全溯源体系设计[J]. 食品科学, 2019, 40(3): 279-285.
LI M J, WANG D, ZENG X S, et al. Food safety tracing technology based on block chain[J]. Food Science, 2019, 40(3): 279-285.

[18] DONG X L, SAHA B, SRIVASTAVA D. Less is more: selecting sources wisely for integration[J]. Proceedings of the VLDB Endowment, 2012, 6(2): 37-48.

[19] REKATSINAS T, DONG X L, SRIVASTAVA D. Characterizing and selecting fresh data sources[C]// International Conference on Management of Data, June 22-27, 2014, Snowbird, USA. New York: ACM Press, 2014: 919-930.

[20] REKATSINAS T, DESHPANDE A, DONG X L, et al. SourceSight: enabling effective source selection[C]// International Conference on Management of Data, June 26-July 1, 2016, San Francisco, USA. New York: ACM Press, 2016: 2157-2160.

[21] RAHM E, FALCONER S M, NOY N F, et al. Schema matching and mapping[J]. Data-Centric Systems and Applications, 2011, 30(7): 121-160.

[22] CATE B T, DALMAU V, KOLAITIS P G. Learning schema mappings[J]. ACM Transactions on Database Systems, 2013, 38(4): 28.

[23] QIAN L, CAFARELLA M J, JAGADISH H V. Sample-driven schema mapping[C]// International Conference on Management of Data, May 20-24, Scottsdale, USA. New York: ACM Press, 2012: 73-84.

[24] BELHAJJAME K, PATON N W, EMBURY S M, et al. Incrementally improving data spaces based on user feedback[J]. Information Systems, 2013, 38(5): 656-687.

[25] EL-ROBY A. Utilizing user feedback to improve data integration systems[C]// The 32nd IEEE International Conference on Data Engineering, May 16-20, 2016, Helsinki, Finland. Piscataway: IEEE Press, 2016: 206-210.

[26] VERGA P, BELANGER D, STRUBELL E, et al. Multilingual relation extraction using compositional universal schema[C]// The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 12-17, 2016, San Diego, USA. [S.l.:s.n.], 2016: 886-896.

[27] DONG X L, HALEVY A Y, YU C. Data integration with uncertainty[J]. The VLDB Journal, 2009, 18(2): 469-500.

[28] DONG X L, GABRILOVICH E, HEITZ G, et al. From data fusion to knowledge fusion[J]. Proceedings of the VLDB Endowment, 2014, 7(10): 881-892.

[29] 庄严, 李国良, 冯建华. 知识库实体对齐技术综述[J]. 计算机研究与发展, 2016, 53(1): 165-192.
ZHUANG Y, LI G L, FENG J H. A survey on entity alignment of knowledge base[J]. Journal of Computer Research and Development, 2016, 53(1): 165-192.

[30] CHAUDHURI S, GANTI V, MOTWANI R. Robust identification of fuzzy

duplicates[C]//The 21st International Conference on Data Engineering, April 5-8, 2005, Tokyo, Japan. Piscataway: IEEE Press, 2005: 865-876.

[31] FIRMANI D, SAHA B, SRIVASTAVA D. Online entity resolution using an oracle[J]. Proceedings of the VLDB Endowment, 2016, 9(5): 384-395.

[32] KONDA P, DAS S, PRASAD S, et al. Magellan: toward building entity matching management systems[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 1197-1208.

[33] CHIANG Y H, DOAN A H, NAUGHTON J F. Modeling entity evolution for temporal record matching[C]// International Conference on Management of Data, June 22-27, 2014, Snowbird, USA. New York: ACM Press, 2014: 1175-1186.

[34] LI F R, LEE M L, HSU W, et al. Linking temporal records for profiling entities[C]// International Conference on Management of Data, May 31-June 4, 2015, Melbourne, USA. New York: ACM Press, 2015: 593-605.

[35] HAN X, ZHAO J. Named entity disambiguation by leveraging Wikipedia semantic knowledge[C]//The 2nd ACM Workshop on Social Web Search and Mining, November 2-6, 2009, Hong Kong, China. New York: ACM Press, 2009: 215-224.

[36] MIHALCEA R, CSOMAI A. Wikify! linking documents to encyclopedic knowledge[C]// Conference on Information and Knowledge Management, November 6-10, 2007, Lisbon, Portugal. New York: ACM Press, 2007: 233-242.

[37] CUCERZAN S. Large-scale named entity disambiguation based on WikiPedia data[C]// Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic. [S.l.:s.n.], 2007: 708-716.

[38] ZHANG W, SIM Y C, SU J, et al. Entity linking with effective acronym expansion, instance selection, and topic modeling[C]// The 9th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems, July 16, 2011, Barcelona, Spain. New York: ACM Press, 2011: 1909-1914.

[39] GANEA O E, GANEA M, LUCCHI A, et al. Probabilistic bag-of-hyperlinks model for entity linking[C]//The 25th International Conference on World Wide Web, April 11-15, 2016, Montreal, Canada. New York: ACM Press, 2016: 927-938.

[40] CHENG G, XU D Y, QU Y Z. Summarizing entity descriptions for effective and efficient human-centered entity linking[C]//The 24th International Conference on World Wide Web, May 18-22, 2015, Florence, USA. New York: ACM Press, 2015: 184 - 194.

[41] SIL A, KUNDU G, FLORIAN R, et al. Neural cross-lingual entity linking[C]// The 32nd AAAI Conference on Artificial Intelligence, February 2-7, 2018, New Orleans, USA. Palo Alto: AAAI Press, 2018: 5464-5472.

[42] SHEN W, HAN J, WANG J, et al. SHINE+: a general framework for domain-specific entity linking with heterogeneous information networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(2): 353-366.

[43] YAO Z J, SUN Y F, DING W C, et al. Dynamic word embeddings for evolving semantic discovery[C]// The 11th ACM International Conference on Web Search and Data Mining, February 5-9, 2018, Los Angeles, USA. New York: ACM Press, 2018: 673-681.

[44] BASIK F, GEDIK B, ETEMOGLU C, et al. Spatio-temporal linkage over location-enhanced services[J]. IEEE Transactions on Mobile Computing, 2017, 17(2): 447-460.

[45] CHEN X, CUI P, YI L, et al. Scalable optimization for embedding highly-dynamic and recency-sensitive data[C]// The 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 19-23, 2018, London, UK. New

York: ACM Press, 2018: 130-138.
[46] BARRANCO R C, DOS SANTOS R F, HOSSAIN M S, et al. Tracking the evolution of words with time-reflective text representations[C]//2018 IEEE International Conference on Big Data, December 10-13, 2018, Seattle, USA. Piscataway: IEEE Press, 2018: 2088-2097.

作者简介

金澈清(1977-),男,博士,华东师范大学数据科学与工程学院教授、博士生导师、副院长。中国计算机学会高级会员,数据库专业委员会委员。已发表学术论文100余篇,研究成果曾获得教育部科技进步奖二等奖、上海市科技进步奖一等奖、霍英东教育基金会青年教师奖。担任《计算机研究与发展》编委,主要研究方向为区块链、计算教育学、基于位置的服务等。

陈晋川(1978-),男,博士,中国人民大学信息学院副教授,中国计算机学会会员,区块链专业委员会通信委员,主要研究方向为区块链和分布式数据管理。

刘威(1989-),男,博士,中山大学副研究员,中国计算机学会数据库专业委员会通信委员,主要研究方向为时空大数据分析、推荐系统、个体行为数据分析与挖掘。

张召(1977-),女,博士,华东师范大学数据科学与工程学院副教授,主要研究方向为区块链系统研发、分布式数据管理,多项研究成果发表在VLDB、ICDE和DASFAA等数据管理领域的重要国际会议上。先后主持多项国家自然科学基金项目,作为骨干技术人员,参与开发的“面向大型银行应用的高通量可伸缩分布式数据库系统”获得2017年教育部高等学校科学研究优秀成果科技进步奖一等奖。

收稿日期: 2020-01-31
基金项目: 国家自然科学基金资助项目(No.U1911203, No.U1811264)
Foundation Items: The National Natural Science Foundation of China(No.U1911203, No.U1811264)