大数据治理的全景式框架

印鉴[1,2]，朱怀杰[1,2]，余建兴[1,2]，邱爽[1,2]
1. 中山大学数据科学与计算机学院，广东 广州 510006；
2. 广东省大数据分析与处理重点实验室，广东 广州 510006

摘要
大数据治理是个复杂的工程，涉及技术和管理2个层面的内容。从系统的角度出发，从大数据和治理2个维度刻画大数据治理的过程，提出了一个全景式系统框架，阐述了该二维框架中每个单元的内容及彼此之间的关联关系。针对大数据维度的一个具体方面，介绍了在理论、方法与技术上的一些研究实践。

关键词
大数据；治理；框架

中图分类号：TP315 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020011

A panoramic framework of big data governance

YIN Jian[1,2], ZHU Huaijie[1,2], YU Jianxing[1,2], QIU Shuang[1,2]
1. School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China
2. Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, China

Abstract
Big data governance is a complex process, involving two aspects, techniques and management. From the perspective of system, the process of big data governance was described from two dimensions of big data and governance. A panoramic system framework was proposed, the content of each unit of the framework and the relationship between them were described. Some research practices in theory, method and technology for a specific aspect of big data dimension were introduced.

Key words
big data, governance, framework

1 引言

自中国共产党第十八次全国代表大会以来，习近平总书记在不同场合多次就推进国家治理体系和治理能力现代化建设发表重要论述。他指出：要运用大数据提升国家治理现代化水平，要建立健全大数据辅助科学决策和社会治理的机制，推进政府管理和社会治理模式创新，实现政府决策科学化、社会治理精准化、公共服务高效化[1]。习近平总书记的重要论断充分肯定了面向政府治理的大数据治理在推动国家治理变革中的重要作用，为推进国家治理体系和治理能力现代化打开了一条技术赋能的路径。将大数据治理应用到国家治理现代化进程中，用数字技术变革推动治理变革，我国在这方面已取得一系列突出成绩[2]。一方面，为老百姓提供了更加便捷的公共服务。各地在推进“互联网+政务服务”方面不断创新，日益实现政务服务“一网、一门、一次”的目标，“让百姓少跑腿、数据多跑路”，给群众带来实实在在的获得感。另一方面，还可以实现政府决策科学化、社会治理精准化。在经济运行、社会治理、信用建设等方面都可以运用大数据的实时、精准和智能等特点进行赋能，从而不断提高政府部门的治理水平。

但是，大数据治理是一个长期的过程，在发展中还存在许多亟待解决的问题。其主要原因在于大数据治理的复杂性。正如梅宏院士[3]指出：大数据治理体系建设涉及国家、行业和组织3个层次，需要从制度法规、标准规范、应用实践和支撑技术方面多管齐下，提供支撑。

笔者认为大数据治理既是技术问题，也是管理问题。大数据治理是一个系统工程，包含大数据和治理2个维度，单独的技术方法和管理方案难以很好地解决大数据治理问题。单纯的大数据存储、分析技术难以被直接应用到复杂的实际问题中，而在实际管理过程中，由于不了解技术现状，很容易产生研究人员好高骛远、研究成果难以实现的问题。因此，研究大数据治理问题的关键在于打通数据科学技术与管理实践问题之间的壁垒，跨学科整合两者之间的关系，梳理出一个行之有效的大数据治理框架。

本文将从上述2个维度展开讨论，分别梳理2个维度的脉络，提出包含9个具体内容的大数据治理全景式框架（PIE框架）。

2 治理流程

从管理学的角度来说，治理通常包括方案、实施和评估3个方面。

（1）方案

方案是针对现实生活中的问题制定的方针、决策或者政策。传统上，一般由企业或政府高层搭建方案的基本框架，然后授权基层制定具体的实践细节，通过在试点执行获得反馈并逐步调整，直至最终出台较完善的方案[4]。这种自上而下的方案制定方式，由于难以全面分析实际问题，需要进行多次试点和改进，存在决策周期长、绩效水平低等缺陷。在大数据时代，研究者可以采集公众多维度的数据，并对这些数据进行融合和分析，进而量化，并预测方案的效益。通过最大化效益，研究者能依靠“众智”制定出最优的民主方案。

（2）实施

实施是指决策方案的具体实践过程。传统的实施由于缺乏对信息的统一管理，各级、各部门较为分散，决策各阶段涉及的审批事项通常需要提供多类办理材料并联合多个部门进行办理，大大降低了工作效率。在共建共享共用的数据时代，构建了统

一的信息平台，原来分散的各级、各部门、各类审批事项得到了梳理整合，办理所需材料的时间大大减少，服务领域的二次开发成为常态，决策的实施走向精细化、人性化、个性化和高效化[5]。

（3）评估

评估是对实施结果和预期目标进行比对和分析的过程。传统的评估通常只是主管部门的自我检查、自我评价。这往往导致“自拉自唱”，评估结果和群众的实际感受存在较大差距。依托于大数据，各类信息数据都可以上传至网络进行记录，实现了信息实时对比共享、全程监控的智能化评估体系。评估主体已从“人控”转为“技控”，评估结果具有独立性、科学性、权威性、公正性和客观性。这实现了对决策实施和管理效果更准确的评估，能够及时捕获管理漏洞，迅速、精确地制定和实施管理方案，大大提高管理效能。

3 大数据赋能机制

大数据具有容量大、增速快、种类多、价值高的特征。作为一种新型的治理技术，大数据通过挖掘、预测、诊断和应用，能够实现科学决策、精细管理、精准服务、精确监管和高效协同，是提升国家现代化和发展国民经济的重要手段。从技术的角度来说，大数据治理是一个大数据赋能的机制和过程，主要包括数据基础、数据服务以及数据生态3个方面。

（1）数据基础

数据基础是指对原始数据进行采集、加工以及存储的硬件设施。数据基础是大数据的来源，是数据服务、数据生态的基石。如何形成优质、完整的数据基础是当今大数据时代的一大难点。

（2）数据服务

数据服务是指提供数据传输、数据处理（包括计算、分析、可视化等）、数据交换等功能的一种由信息技术驱动的服务。数据服务是数据价值的体现方式，能让大数据“活起来”。如何提供好的数据服务，也是目前研究的热点，尤其是在数据处理方面，目前虽然出现了很多与深度学习相关的算法，但是大部分算法仍然缺少可解释性。

（3）数据生态

数据生态是指数据的开放与共享、互动与协同，数据生态使得数据变得透明可信，从而形成大数据感知、管理、分析与应用服务的新一代信息技术体系。随着大数据技术与物联网、云计算、人工智能等新技术的相互融合渗透，数据生态变得越来越重要，对数据进行开放与安全共享变得尤为重要，需要各类型数据之间的互动与协调。

虽然目前大数据产业生态体系将迈入成熟阶段，但是大数据体系还需要不断完善（如开放的方式、安全的共享机制等），还有许多问题没有解决（如数据的透明与可信问题）。

4 大数据治理全景式框架

基于大数据赋能机制和治理流程2个维度，本文提出大数据治理全景式框架，如**图1**所示。

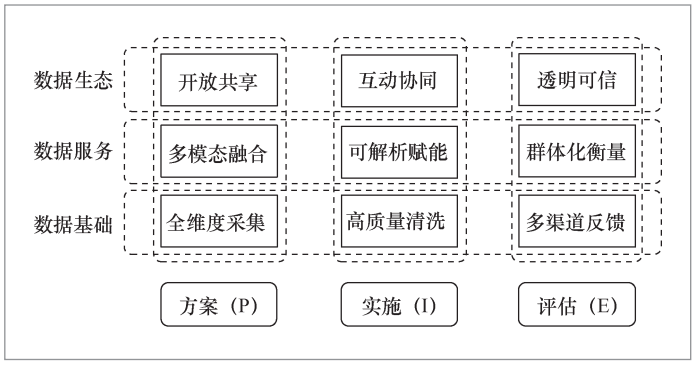

图1 大数据治理全景式框架

4.1 数据基础

（1）方案的数据基础

方案的制定依赖于全面且广泛的数据基础，数据的维度越丰富，制定的方案就越客观可靠。为了能全维度地采集数据，企业或者政府应始终坚持大力发展大数据相关技术和信息智能化建设工作。

基层应树立起对大数据概念、应用价值的全面认知理念，积极响应企业或者政府就促进大数据发展提出的各项指示。着力推进数据开放步伐，打破数据碎片、数据孤岛和数据壁垒，对散落在各个委、办、局、企业、部门的数据进行整合，形成最广泛的数据联合平台。重点采集与经济和民生相关的多维度数据，并设立数据决策机构，针对具体业务和实践状况的数据特性和区域特色，制定更加适宜的发展方案。

（2）实施的数据基础

由于采集的数据一般是从多个业务系统中抽取而来的，难免会出现各类数据问题，例如数据目录不标准、多源数据未能融合、数据质量无法保证、数据安全机制不全等。这些混乱、重复、错误、缺失甚至不一致的数据会严重影响决策系统的有效运行。为了解决这个问题，需要在微观上对这些“脏数据”进行清洗处理，包括对数据进行重新审查和校验删除重复信息、纠正存在的错误，并提供数据一致性。在清洗的过程中，需要根据各个部门提供的资料（如数据字典、样本数据等）分析可能存在的数据质量问题，并制定数据质量检查规则、数据清洗的内容，然后构建清洗的任务，解决可纠正的问题，标记不可纠正的问题，并汇报数据中存在的质量问题。

（3）评估的数据基础

方案执行的效果可以通过广泛地收集公众意见等方式反馈数据感知。这些数据能够及时、有效地对国家治理现代化进程进行监管。将海量碎片化且无序排列的信息变成有用、有序的数据，可以使监管更有针对性。对方案执行的每个阶段进行事中评估，及时了解进展情况，发现问题，纠正偏差，避免出现方向性错误。这需要扩展公众意见反馈渠道，使公众能够通过互联网平台等多种方式向机构、窗口部门建议和提问，使得工作透明化，调动公众参与国民经济建设的积极性。另外，这些多种渠道的数据收集并不是简单的堆积，而是需要根据数据的质量、类型、用途、领域等实施科学、有效的整理和归纳，并对公众反馈的数据严格把关，制定恰当的管理和分析体系，确保对各领域的有效评估，进一步降低治理成本，提高治理效率，从而进一步提升治理的整体效能。

4.2 数据服务

（1）方案的数据服务

方案制定过程是一种宏观规划，其充分依赖于以往决策的评价以及现有条件。方案制定过程受多方面因素影响，可能需要整合许多不同类型、不同来源甚至不同模态的数据。因此，在方案的数据服务中，多模态数据的融合与处理是一项关键技术。方案制定过程中各类影响因素之间的关系较为复杂，多种实体之间存在的推理关系可构成知识图谱。因此，方案制定过程也涉及基于知识图谱的推理问题。大数据技术为方案制定过程提供的数据服务主要包括对已有方案的效果预判、根据现有复杂影响因素自动构建知识图谱以及推理生成方案。其中涉及的关键科学问题主要包括多模态数据融合和知识图谱的构建与推理。

（2）实施的数据服务

实施是治理流程中的主体环节，是第

一阶段制定的方案的执行过程，也是方案中大数据技术具体赋能治理过程的关键步骤。实施是根据方案执行的一系列步骤，是一个过程。在每个步骤中，如何对方案进行应用，并根据效果实时地对执行方案进行微调，是实施过程中遇到的主要问题。为此，需研究可解释的大数据分析技术。只有对方案中的逻辑思路有深入的了解，才能找到实施过程中出现问题的原因，更好地指导实施过程的执行。例如，大数据、人工智能技术赋能产业实施的过程中，可通过用户画像、兴趣推荐以及行为预测等方式促进各类具体金融、商务应用的实施。只有当大数据技术可解析时，才能使推荐、画像更具可信性，更能应对实施过程中出现的特殊情况。

(3) 评估的数据服务

在治理流程中，评估是必不可少的一环。合理的评估机制可以形成良好的反馈环，促进产业良性增长。在大数据技术的支撑下，可构建群体化评估体系，整合用户评价大数据。在大数据赋能的数据服务层，对评估的支持具体体现在对海量评价数据的自动处理与分析方面，从而得到详实可靠的评估打分、正/负面情感倾向。此过程中包含2项关键技术：一是直接针对评估文本的自然语言处理技术，如情感分析技术等；二是针对体现评估效果的数据（如点击率、打分分值）的数据挖掘技术。数据服务层提供自动评估处理服务接口，用户可以接入并对众包的评估数据进行自动处理，快速生成业务、服务的智能评估。

4.3 数据生态

(1) 方案的数据生态

方案是指从目的、要求、方式、方法、进度等方面部署具体、周密并且有很强的可操作性的计划。对于方案的数据生态来说，一个开放和共享的环境能让方案变得更加透明和可信，能够为方案的良性循环提供一个好的生态环境。在大数据的背景下，借助众包、众智等技术有利于形成透明可信的方案。另外，在方案的设计过程中会产生更多的数据，这样方案和数据生态构成了一个互动与协同的体系。

(2) 实施的数据生态

实施是体现大数据价值的重要环节，实施的最终状态是一个互动协同的过程。在大数据、人工智能赋能产业的实施过程中，会不断产生各种各样的数据，如实施的结果数据、用户反馈的数据、商家/政府的数据等。这些数据能够对进一步实施产生影响。因此，实施在数据生态这一层面是互动协同的，而这一过程中得到的数据能够促进大数据产业的进一步发展和实施，实施和数据生态相互作用，共同协调发展。因此，需要循环有效地利用数据，从而促进实施的进行，大数据的有效和合理实施也能促进产生更多有价值的数据。

(3) 评估的数据生态

评估是数据治理的最终环节，在数据生态层面，其应该是一个透明可信的过程。透明和可信的评估机制能够形成良好的数据生态。评估首先需要的是当前实施后的结果数据，通过对背景、规则、政策、法制等数据进行有效的整理、归纳以及分析，才能得出透明和可信的评估结果。对这些数据的整理和归纳不仅涉及多模态数据融合、数据拼图、数据对齐等技术，还涉及更复杂的数据，需要利用众包、众智等技术得出有效而公正的评估。另外，评估的结果可以对方案的设计和修正起到有效的作用，进一步对实施构成影响，从而形成一个健全的评估体系。

5 数据服务的研究实践

笔者所在课题组目前正负责国家自然科学基金委员会-广东省人民政府大数据科学研究中心项目：深度学习支持的政府治理大数据分析与预测关键技术研究。基于前文提出的PIE框架，该研究在数据服务层进行了一些探索与实践[6-9]。

在多模态融合方面，课题组提出数据拼图的理念。数据拼图是指将所有相关数据整合起来，呈现一张完整的信息图，是对数据的存储、共享、语义等方面进行综合考虑的一个全局视图。从互联网、社交网络、政府日常运行中产生的数据出发，以多视图学习、增量学习等技术为入口，从不同的角度构造多模态政府大数据的整体画像，完成政府大数据的各主体数据拼图。

在多模态数据检索[6]方面，度量不同模态数据的相似性是多模态数据面临的基本挑战。而不同模态数据的相似性度量的最基本任务是抽取有效的不同模态数据的特征表示。为了抽取多模态数据的特征表示以及度量多模态数据的相似性，研究了基于注意力机制的深度对抗网络，对数据进行特征学习以及相似性的挖掘。自动学习背景数据并将其除去，从而学习有用的特征，更加准确地表示多模态数据。

在语义解析和描述方面，课题组也进行了相关的研究。语义解析的目的是把自然语言自动转化为一种机器可以理解并执行的表达形式，如将用户查询转换为可以在结构化知识库上执行的SPARQL语句。人们在对样例做决策时，往往不是从头开始，而是先从已有的知识库中找到相似的样例，然后进行改写。传统的检索-编辑（retrieve-and-edit）方法通常只考虑一个样例，而一个结构化规范语义表示可能来自多个相关的样例。以此为出发点，课题组提出一种结合检索与元学习的语义解析方法[7]。

在可解析赋能方面，课题组提出了可解释性的深度学习算法。政府在治理大数据过程中遇到的一个科学问题是如何实现对预测结果、决策过程的可解释，使得每一个决策步骤都有充分的依据，实现决策的科学性。通过对可解释性深度学习的理论研究，建立面向具体政府大数据的新型深度学习分析框架与技术，实现对具体问题的全方面辨识，提供可解释的辅助决策数据分析，为科学决策奠定理论基础。可解释性在机器学习系统中非常重要，传统的线性模型或各种浅层的机器学习算法均具有很强的可解释性，其中比较著名的当属贝叶斯网络（概率图模型），这是一种能够对复杂不确定系统进行推理和建模的有效工具，在过去几十年被广泛运用于处理涉及智能推理、诊断、决策风险及可靠性分析等方面的问题，如电网故障诊断、核能运行的可靠性分析和航空交通管理等。然而传统贝叶斯模型的一个较大的问题在于其推理方法通常较慢，特别是在大数据背景下很难适应新的模型的要求。深度学习能够处理政府治理大数据，但其学习网络结构大多是靠手工设计的，这是一个费时费力的工作，需要有丰富的工程经验才能设计出一个合理的网络，而且后验权重概率分布高度复杂，这导致学习又不具有可解释性。为了解决以上问题，可以利用贝叶斯网络做可解释的深度学习模型，先选择出网络结构，再去学习网络权重，进而推理学习最佳网络结构。进一步地，课题组研究了可解析的深度学习模型[8]，该模型将神经网络作为自编码器模拟环境，利用近似推理的方法拟

合复杂的后验分布，提高了数据利用率，降低了对硬件资源的要求，并在基于笔画的图像生成领域取得了较好的效果。

在群体化衡量方面，课题组提出了数据众智的设想，将数据众智与政府大数据治理结合。在大数据治理过程中，为了克服传统决策与评估方法的手段滞后、主观性强等问题，需充分发挥专家、媒体、利益群体和人民大众在决策和评估中的作用，发挥数据决策和评估中大众的智慧，为公共政策制定和评估提供有效的分析工具。为此，课题组开展了一项具体的基于数据众智的决策与实时政策评估实践[9]。为了提升地方经济的竞争力，地方信息资源产业的发展变得越来越重要。如何评估地区信息资源产业的竞争力成为一个非常重要的问题。课题组提出了一个改进优劣解距离法，用于评价我国各省市自治区的信息咨询产业的竞争力。将数据众智、群体化衡量引入大数据治理，可为决策者提供更为客观、详实、及时的评估和决策建议。

6 结束语

以新一代信息技术为代表的科技革命方兴未艾，大数据技术成为新一代产业转型的关键点。研究大数据赋能的治理机制成为中国现代化发展的紧迫任务。本文从大数据赋能机制和治理流程2个维度展开，对大数据治理的全景式框架进行探索和展望。在大数据赋能机制维度上，主要聚焦数据基础、数据服务、数据生态3个层次；在治理流程维度上，依次研究方案、实施、评估这3个要素。在此基础上，本文提出了包含9个重要内容的大数据治理全景式框架，并分别对其技术要点进行了介绍。最后，以框架的数据服务层为例，介绍了笔者团队在具体理论、方法与技术上的一些研究探索与实践。

参考文献:

[1] 欧阳康. 以大数据促进国家治理现代化[N]. 光明日报，2019-10-25.
OUYANG K. Promoting the modernization of national governance with big data[N]. Guangming Daily, 2019-10-25.

[2] 张建峰. 用大数据助力治理现代化[N]. 人民日报，2019-10-17.
ZHANG J F. Modernizing governance with big data[N]. People's Daily, 2019-10-17.

[3] 梅宏. 大数据治理成为产业生态系统新热点[J]. 领导决策信息，2019(5): 26.
MEI H. Big data governance becoming a new hotspot in industrial ecosystem[J]. Information for Deciders Magazine, 2019(5): 26.

[4] 徐晓新. 社会政策过程：新农合中的央地互动[M]. 北京：中国社会科学出版社，2018.
XU X X. Central-local interaction in new rural cooperative medical system[M]. Beijing: China Social Sciences Press, 2018.

[5] 马跃明. 大数据提高政府能力[J]. 今日浙江，2018(5): 20-21.
MA Y M. Promoting government capacity with big data[J]. Zhejiang Today, 2018(5): 20-21.

[6] ZHANG X, LAI H J, FENG J S. Attention-aware deep adversarial hashing for cross-modal retrieval[C]// European Conference on Computer Vision, September 8, 2018, Munich, Germany. Heidelberg: Springer, 2018: 614-629.

[7] GUO D Y, TANG D Y, DUAN N, et al. Coupling retrieval and meta-learning for context-dependent semantic parsing[C]// Annual Meeting of the Association for Computational Linguistics, July 28, 2019, Florence, Italy. Stroudsburg: ACL Press, 2019: 855-866.

[8] ZHENG N Y, JIANG Y F, HUANG D J. StrokeNet: a neural painting environment [C]// The International Conference on Learning Representations, April 30, 2019, New Orleans, America. [S.l.:s.n.], 2019.

[9] QIAN M H, WANG Y X, WEI X, et al. An improved TOPSIS approach for the competitiveness analysis of provincial information resource industries in China[J]. Expert Systems, 2019, 36(4): e12407.

作者简介

印鉴（1968- ），男，中山大学数据科学与计算机学院教授，主要研究方向为大数据分析、机器学习。

朱怀杰（1988- ），男，博士，中山大学数据科学与计算机学院特聘副研究员，主要研究方向为大数据管理、时空数据库、社交网络。

余建兴（1985- ），男，博士，中山大学数据科学与计算机学院特聘副研究员，主要研究方向为自然语言处理领域的智能对话与问答系统。

邱爽（1990- ），女，博士，中山大学数据科学与计算机学院特聘副研究员，主要研究方向为自然语言处理。

收稿日期： 2020-02-01

基金项目： 国家自然科学基金资助项目（No.U1711262，No.U1711261，No.U1911203）

Foundation Items: The National Natural Science Foundation of China（No.U1711262，No.U1711261，No.U1911203）