我国大数据发展指数构建及实践应用——从政务数据与社会数据融合的视角

郭明军[1,2]，陈沁[3]，安小米[2,4]，王建冬[1]，易成岐[1]
1. 国家信息中心，北京 100045；2. 中国人民大学信息资源管理学院，北京 100872；
3. 成都数联铭品科技有限公司，四川 成都 610041；4. 中国人民大学智慧城市研究中心，北京 100872

摘要
针对大数据发展指数研究数据源较单一、无法覆盖到各城市的不足，从政务数据、社会数据的全量数据融合视角，在充分融合政务数据、企业数据、互联网数据的基础上，从基础能力、创新应用、综合保障3个维度，构建了政务数据与社会数据相融合、全景式展示各城市大数据画像的大数据发展指数，客观评估我国大数据的发展水平，为政府治理、产业发展及民生服务能力提升提供客观数据参考。

关键词
政务数据；社会数据；数据融合；大数据发展指数

中图分类号：C37 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022023

Construction and practical application of big data development index in China-from the government data and social data fusion perspective

GUO Mingjun[1,2], CHEN Qin[3], AN Xiaomi[2,4], WANG Jiandong[1], YI Chengqi[1]
1. State Information Center, Beijing 100045, China
2. School of Information Resource Management, Renmin University of China, Beijing 100872, China
3. Chengdu Shulian Mingpin Technology Co., Ltd., Chengdu 610041, China
4. Smart City Research Centre (Renmin University of China), Beijing 100872, China

Abstract
In view of the two deficiencies of the current big data index research, data sources are relatively limited and unable to cover each city, based on the perspective of full data fusion of government data and social data, a big data development index was constructed, which integrates government data with social data and presents the "portrait" of big data of various cities in a panoramic manner from the three dimensions of basic capability, innovative application and comprehensive guarantee.

The development level of big data was evaluated objectively, which providing objective data reference for government governance, industrial development and improvement of people's livelihood service capability.

Key words
government data, social data, data fusion, big data development index

0 引言

指数是客观反映事物发展水平的有效工具之一。随着大数据的快速发展，大数据指数研究成为学术界的重点议题，成为客观反映中国大数据发展现状的重要手段，有助于政府、企业、社会了解大数据领域的发展状况，为相关决策的制定提供参考。

近年来，随着我国大数据战略的持续推进，如何对大数据发展水平进行量化评估，为政府科学决策、精准施策提供客观数据支撑，成为众多研究机构的一项重要任务。许多学者围绕大数据指数进行了深入研究，形成了丰富的指数类研究成果。但受限于数据的可获得性，目前大数据指数领域研究利用的数据或者仅涉及经济或民生等某一具体领域，或者仅使用政府统计类小数据或互联网大数据，数据种类及数量不足，难以真正实现“统筹规划政务数据资源和社会数据资源”的目标。这成为我国大数据指数研究的重要瓶颈，亟待突破。本文以2019年贵阳数博会发布的《中国大数据发展指数报告（2018年）》提出的大数据发展指数（2018）为例，充分融合政务数据和社会数据资源，为我国大数据指数研究探索新的路径。

1 已有工作

在国外，大数据指数研究首先在预测流行病的发生概率方面取得了显著成效[1]，随后，大数据方法逐步被拓展到经济学领域[2]，许多学者围绕个人消费行为[3]、失业率预测[4-5]、劳动供需分析[6]、就业歧视[7]等进行了大数据指数监测预测研究。近年来，采用综合集成方法，融合多视角、多层面、多利益方，把各方面构成要素有机融合，成为大数据治理体系框架构建的基本思路[8]。在国内，大数据指数研究的力度逐渐加强，形成了两种主要类型，第一类是以高校为代表的学术类研究，第二类是权威信息化机构发布的评估类研究报告。

1.1 学术类大数据指数研究

国内学术类大数据指数研究主要由高校牵头，往往聚焦某个特定领域。比如，一些学者围绕消费[9]、交通[10]、就业[11]、融资[12]、生活质量[13]等民生领域，编制了大数据指数进行预测分析研究；也有学者聚焦经济领域，通过形成电力大数据指数[14]、金融风险指数[15]、制造业指数[16]、实体经济指数[17]等，对经济发展趋势进行监测预测研究；还有一些学者关注政府治理[18]、互联网舆情[19]、居民情绪[20]、社会信用[21]、城市发展[22]等社会治理领域，开发了相关的大数据指数。

1.2 评估类大数据指数研究

评估类大数据指数研究主要由国家信息中心、中国信息通信研究院（以下简称中国信通院）、中国电子信息产业发展研究院

（以下简称赛迪研究院）等国家级信息化机构编制，从国家层面、区域层面、省级层面、产业层面等对我国大数据的发展情况进行评估排名，阿里巴巴集团、南方基金管理股份有限公司、新浪财经等企业也基于自身数据资源优势，针对某个领域开展大数据指数评估。

根据数据源及分析方法的不同，评估类研究又可分为互联网大数据分析与统计数据分析两种情况。互联网大数据分析的典型代表是国家信息中心编写的《中国大数据发展报告（2017）》，该报告基于40多亿条互联网数据，应用大数据分析技术，从政策环境、人才状况、投资热度、创新创业、产业发展、网民信心等方面，对全国各省大数据的发展情况进行评估。阿里研究院2017年发布的《品质消费指数报告》[23]从全网采集数据，使用大数据技术分析了我国居民消费的升级趋势。南方新浪大数据指数则在利用上市公司数据的基础上，将互联网大数据引入指数编制中，利用大数据对市场主体的情绪进行刻画和量化。这类评估的好处在于，基于海量的互联网数据、使用大数据技术进行的分析评估，其数据样本很大，有的指数涵盖数十亿甚至数百亿条数据，且很多数据能够实时动态更新，同时使用了数据挖掘、文本分析、语义分析等大数据技术，是真正意义上的大数据分析。不足之处在于，由于数据采自互联网，容易被认为是网络舆情分析或不能直接反映现实情况。统计数据分析的典型案例是中国信通院发布的《中国大数据发展调查报告（2018年）》，其主要通过现场访问、电话采访、在线调研和专家访谈等方式获取数据。赛迪研究院发布的《中国大数据发展指数报告（2018年）》也主要以统计数据为基础，对我国31个省、自治区、直辖市的大数据发展环境、大数据产业、大数据应用、技术研发创新及数据共享开放情况进行评估。这类评估主要基于官方统计数据或调查研究数据，使用传统数据分析方法进行分析评估，好处是数据相对权威，不足之处在于这类评估并非真正意义上的大数据分析，而是基于传统统计方法开展的小样本数据分析。

从以上分析可以看出，无论是学术类大数据指数研究还是评估类大数据指数研究，虽然都可以在一定程度上进行大数据预测或反映大数据的发展水平，但也存在一些不足。学术类大数据指数研究由于缺少海量数据支撑，即便拥有较好的分析技术方法，往往也只能对某一领域进行研究，分析对象无法覆盖各个省市。权威机构发布的评估类指数也存在一定缺陷，从数据源来看，该类评估仅仅基于互联网大数据或统计小数据，不能体现多源数据广泛代表性的优势，同时由于数据获取困难，评估对象只能覆盖到省级地区，无法延伸到地市层面。

2 数据来源及加工处理

为了弥补上述两类方法的不足，国家信息中心联合相关单位，探索将政府部门数据、政府网站数据、统计数据等政务数据与互联网大数据、企业数据等社会数据进行对接融合，构建了大数据发展指数（2018），使用大数据方法对全国及各省市大数据发展情况进行评估，从而既体现政务数据的直接性和权威性，也保证互联网等社会数据的鲜活性和大样本量。

一是数据层面，拓展多源数据。“指标好编、数据难得”是评估工作中普遍存在的问题，是否拥有广泛的数据源和足够的数据量，直接关系到定量评估的效果的好坏。本文通过大数据评估解决了数据源问题：一是通过搭建大数据指标监测系统，

抓取互联网中与大数据评估相关的指标，包括政府大数据规划及政策、企业注册数据、专利数据、招聘数据、经济金融产业中间投入及数量等数据；二是采集国家统计局、工信部、知识产权局发布的相关政务数据；三是参考借鉴现有研究成果，包括财新BBD（成都数联铭品科技有限公司）数字经济指数中的产业指数、新华三数字经济指数和国家行政学院电子政务研究中心提出的政务应用指数等。通过互联网数据与政务数据融合分析，保证多源数据的代表性，同时也兼顾了数据的权威性。相关数据见**表1**。

二是对象层面，覆盖所有城市。此前的评估主要针对国家层面、区域层面及省级层面，缺少全方位对城市层面进行的专门评估。本文将评估对象下沉到城市级，获取全国各个城市的多源异构数据，对各城市的大数据发展进行全景式分析，力求展现各城市的大数据发展水平。

三是技术层面，使用先进技术。传统的统计分析仅能处理小样本数据，而对于数十亿甚至数百亿条的数据，往往束手无策。更重要的是，对于互联网上的大量非结构化数据，只能通过大数据手段，才能进行有效的分析挖掘。大数据发展指数（2018）的数据加工处理综合使用了文本挖掘、语义分析、情感分析、机器学习等先进技术手段，体现了应用大数据技术开展大数据评估的特点。

3 指数构建

3.1 构建原则

为了确保评价指标的有效性，准确、全面地衡量我国大数据的发展水平，本文在设计大数据发展指标时坚持以下4项原则。

一是完备性。指标体系中的指标能够全面地反映评估对象的发展情况，确保不遗漏重要指标项。本评估指标体系包含了基础能力、创新应用、综合保障三大方面，从50个维度评估省市的大数据发展情况。

二是客观性。在指标选取及整个指标体系的确立中，每个环节均选取客观、可量化的指标，除了能够在网上直接采集的客观数据，也采用了政府部门及相关机构发布的客观统计数据，并借鉴了已有研究成果，尽可能规避主观因素带来的干扰。

三是导向性。整个指标设计突出了应用导向，将“创新应用”指标作为最重要的指标，并赋予最大的权重，重点评估各个省市在政务、经济及民生三大领域的应用成效。

四是易操作性。所有指标的数据均能获取，有些数据甚至可以做到实时获取。

表1 大数据发展指数的政务数据与社会数据使用情况

数据类别	数据级别	具体来源
政务数据	国家级	工信部、商务部、国家市场监督管理总局、国家统计局、知识产权局、国家行政学院、国家互联网应急中心、中国互联网络信息中心、国家信息中心
	地区级	省市政府门户网站、各地统计年鉴
社会数据	企业数据	BBD大数据、全国中小企业股权交易中心、私募通数据库、新华三集团
	第三方机构数据	中国电子信息产业统计年鉴
	互联网数据	互联网各大招聘网站

评价方法采用专家赋权法，该方法具有很强的操作性。同时，专业分析人员及大数据分析平台保证了整个评估工作的易操作性。

3.2 指标体系

大数据发展指数（2018）以“应用”为核心，围绕“能力-成效-保障”3个方面，构建了由基础能力、创新应用、综合保障构成的一级指标。按照每个一级指标包含的核心要素，设置了9个二级指标，同时，根据数据的可获得性以及为了充分体现政务数据与社会数据的融合，构建了50个三级指标，形成了我国大数据发展指数。大数据发展指数属于大数据指数范畴，两者在数据采集汇聚、加工处理以及编制思路等方面基本相似，只是大数据发展指数更加强调发展应用，“能力-成效-保障”3个维度更加突出发展的导向。

（1）基础能力指标

基础能力指标主要衡量地区的大数据基础，包括数据、算力、算法3个二级指标和20个三级指标，见**表2**。

（2）创新应用指标

创新应用指标包括政务应用、经济应用和民生应用3个二级指标和24个三级指标，体现大数据在政府治理、产业发展和便民服务领域的应用，见**表3**。

（3）综合保障指标

综合保障指标包含政策保障、合作保障和安全保障3个二级指标及6个三级指标，着重分析大数据应用的稳定性及可持续性，见**表4**。

指标使用了所有可获得的能够度量大

表 2 基础能力指标

一级指标	二级指标	三级指标	数据来源
基础能力	数据	政务数据共享交换平台	国家信息中心
		开放数据平台	国家信息中心
		市政大数据平台	国家信息中心
		政府数据开放条数	国家信息中心
	算力	局用交换机容量	工信部
		移动电话普及率	工信部
		光缆线路长度	工信部
		互联网宽带接入用户数	工信部
		互联网宽带接入端口	工信部
		信息传输计算机服务和软件业全社会固定资产投资完成额	国家统计局
		固网宽带应用渗透率	各地统计年鉴
		移动网络应用渗透率	各地统计年鉴
		网民普及率	各地统计年鉴
	算法	大数据行业专利占总专利比例	知识产权局
		软件产业研发经费	工信部
		大数据行业专利数量	知识产权局、工信部
		大数据行业专利转移数量	知识产权局
		大数据行业总薪酬	互联网各大招聘网站
		大数据行业就业岗位数量	互联网各大招聘网站
		软件产业从业人员数	工信部

表 3 创新应用指标

一级指标	二级指标	三级指标	数据来源
创新应用	政务应用	政务服务数字化水平	国家信息中心
		在线服务成效指数	国家行政学院
		服务事项覆盖度指数	国家行政学院
		服务方式完备度指数	国家行政学院
		在线办理成熟度指数	国家行政学院
		办事指南准确度指数	国家行政学院
	经济应用	互联网上网营业收入	工信部
		大数据行业新三板上市注册资本总额	全国中小企业股权交易中心
		大数据行业新企业注册资本总额	国家市场监督管理总局
		大数据行业风险投资总额	私募通数据库
		软件产业信息安全软件收入	中国电子信息产业统计年鉴
		企业数电子商务交易	国家统计局
		电子商务采购额	国家统计局
		电子商务销售额	国家统计局
		经济金融产业中间投入	BBD大数据
		经济金融产业中间投入数量	BBD大数据
	民生应用	教育服务数字化	中国互联网络信息中心
		医疗服务数字化	新华三集团
		社会产业中间投入	BBD大数据
		社会产业中间投入数量	BBD大数据
		共享经济服务业中间投入	BBD大数据
		共享经济服务业中间投入数量	BBD大数据
		文化产业中间投入	BBD大数据
		文化产业中间投入数量	BBD大数据

表 4 综合保障指标

一级指标	二级指标	三级指标	数据来源
综合保障	政策保障	大数据相关专题政策	省市政府门户网站
		大数据进入政府长期规划	省市政府门户网站
	合作保障	外商投资企业投资总额	商务部
		外商投资企业数	商务部
		大数据技术的国际合作	商务部与EconLit数据库
	安全保障	感染计算机恶意程序的主机数量占本地区活跃IP地址数量的比例	国家互联网应急中心

数据发展的年度城市数据，并创新地纳入了全网大数据。例如，“数据”二级指标主要来自其他指数的评估结果，包括复旦大学发布的中国开放数林指数、新华三的中国城市数字经济指数；“算力”二级指标主要来自各项统计指标；而“算法”部分则主要来自大数据，从人才总量和专利申请、专利流转3个方面度量大数据的技术水平，选用的数据总量达20亿条。各种类型数据的使用保证了在更细、更高频维度上度量城市大数据的发展情况，这也是目前其他机构发布的大数据指标中未能体现的特点。

3.3 权重设计

大数据发展指数的测算采用主观赋权法对指标体系中的各指标进行赋权，通过专家打分再求平均值获得各指标的权重，在保证指标体系科学性、全面性的同时，力求指标权重的稳定性。

使用定基标准化方法将测算结果进行标准化处理，既可以较方便地进行横向和纵向比较，也能与2018年北京的数值进行对标，有利于各个地方找到大数据发展过程中自身存在的短板。

使用定基标准化方法的原理如下：

- 设原始值为X_{ijt}，其中i表示指标项，j表示城市，t表示时间；
- 选择2018年北京各项数值作为基期，标准化数值为100，记录其缩放比例，即当j=北京、t=2018时，记录$\Phi_i = 100 / X_i$；
- 将所有指标数值乘以对应的Φ_i，得到$\bar{X}_{ijt} = X_{ijt}\Phi_i$；
- 将$\bar{X}_{ijt}$作为新的指标数值进行计算。

按照此设计方法，本文对9个二级指标权重进行了赋值，见**表5**。从**表5**可知，创新应用指标的权重最高，达到45%，体现了创新应用的重要性，突出了发展导向的指标制定思路。

表 5 大数据发展指数的指标权重赋值

一级指标	权重	二级指标	权重
基础能力指标	40%	数据	14%
		算力	18%
		算法	8%
创新应用指标	45%	政务应用	10%
		经济应用	25%
		民生应用	10%
综合保障指标	15%	政策保障	9%
		合作保障	4%
		安全保障	2%

4 实践应用

大数据发展指数（2018）编制形成之后，从横向、纵向两个维度验证了指标体系在评估国家及城市大数据发展水平方面的有效性。通过对2015—2018年全国大数据发展水平进行纵向对比评估，比较了历年大数据发展的变化情况。通过对351个城市的大数据发展水平进行横向对比评估，分析了各城市大数据发展特点及存在的短板。

4.1 全国大数据发展水平评估

基于构建的大数据发展指标，笔者对2015—2018年全国大数据发展指数进行了测算，如**图1**所示。结果显示，4年来，我国大数据发展总体呈现上升趋势，但2018年略有下降。从具体指标来看，2015—2018年我国大数据发展的基础能力始终保持上升态势，但由于创新应用水平在2018年有所下降，总指数下降了0.3。

通过进一步的分析发现，2018年指数下降主要有两方面原因。一是我国经济增速放缓，经济下行压力增大，“资本寒冬”持续存在，给大数据行业的发展造成了影响，数据显示，2018年大数据行业的风险投资额度只有2017年的20%；二是前几年由政府主导推动的大数据应用因补贴高速扩张而在2018年资金短缺时无法持续，一些大数据公司2018年支付高级人才的工资也不如2017年高，招聘数量也出现明显萎缩。

4.2 各城市发展水平评估

笔者对2018年全国351个城市大数据的发

展水平进行了评估，2018年大数据发展指数排名前40名的城市如**图2**所示。结果显示，2018年城市大数据发展指数排名前十的分别为深圳、北京、上海、杭州、成都、广州、天津、南京、东莞、武汉，呈现出一线城市引领，成都、天津等地快速跟进的特点。

● 分项指标方面，一线城市在经济及民生应用方面遥遥领先，而江苏、浙江、广东等地的政务应用水平位居前列，算法优势集中于北京、上海、深圳等地，而算力资源在东北地区较为丰富。

● 数据指标排名前五的分别为成都、深圳、广州、青岛和福州，它们均为数据资源及经济基础较好的城市。

● 算力指标排名前五的为大庆、牡丹江、哈尔滨、长春和北京。东北之所以排名靠前，主要是因为算力指标主要的计算来源是与大数据相关的信息系统基础，东北三省在人均光纤长度、人均互联网端口数量等方面均处全国领先地位。

● 算法指标排名前五的为北京、深圳、东莞、上海和西安，这主要得益于城市科教资源丰富、对大数据领域人才吸引力较强。

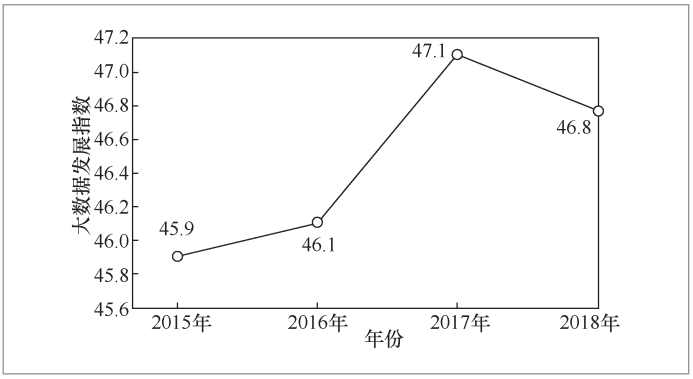

图 1　2015—2018 年全国大数据发展指数

● 政务应用指标排名前五的分别为东莞、深圳、无锡、南通和金华，主要集中于江苏、浙江、广东，其政府部门的大数据整体应用水平相对较高。

● 经济应用和民生应用方面，上海、北京、深圳和杭州位居前四，这表明这4个城市在大数据推动产业发展及提升便民服务水平方面位居全国前列。

● 综合保障指标方面排名前五的分别是天津、杭州、宁波、武汉和成都，表明这些城市在政策、合作和安全方面的保障力度较大，可持续发展能力较强。

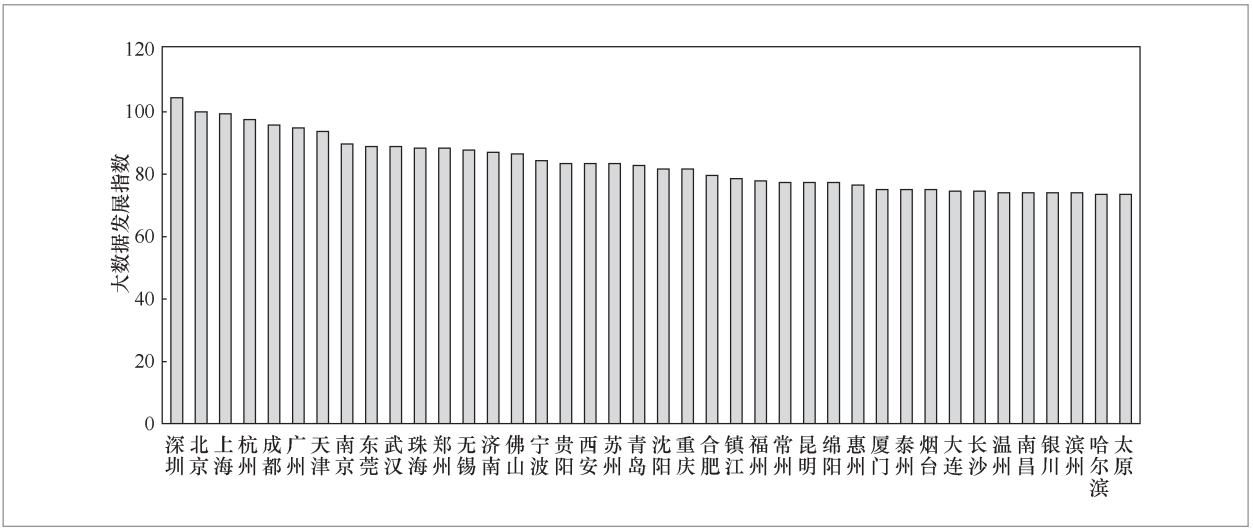

图 2　2018 年大数据发展指数排名前 40 名的城市

5 结束语

本文将政务数据与社会数据进行融合，构建了国内第一个将政务数据与社会数据融合、全景式展示城市大数据“画像”的大数据发展指数，弥补了大数据指标存在的数据源单一、无法下沉到城市的不足。总体来看，该指标体系具有“专、新、多”三大特点：“专”，评估对象聚焦到市，覆盖全国351个城市；“新”，充分利用大数据手段，包括数据挖掘、语义分析、社会网络分析等技术；“多”，将政务数据、统计数据、企业数据、互联网数据融合，应用数据量达20亿条。未来研究将基于多类数据融合的思路，进一步拓展数据资源范围，充分利用大数据分析技术，构建更为合理的指标体系，客观评估我国国家、区域、领域及省市县各级大数据发展水平，为政府治理、产业发展及民生服务能力提升提供客观参考。

参考文献:

[1] GINSBERG J, MOHEBBI M H, PATEL R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009, 457(7232): 1012–1014.

[2] EDELMAN B. Using Internet data for economic research[J]. Journal of Economic Perspectives, 2012, 26(2): 189–206.

[3] VOSEN S, SCHMIDT T. Forecasting private consumption: survey-based indicators vs. Google trends[J]. Journal of Forecasting, 2011, 30(6): 565–578.

[4] ASKITAS N, ZIMMERMANN K F. Google econometrics and unemployment forecasting[J]. Applied Economics Quarterly, 2009, 55(2): 107–120.

[5] ZHI S. Chinese online unemployment-related searches and macroeconomic indicators[J]. Frontiers of Economics in China, 2014(4): 573–605.

[6] CAPILUPPI A, BARAVALLE A. Matching demand and offer in on-line provision: a longitudinal study of monster.com[C]//Proceedings of 2010 12th IEEE International Symposium on Web Systems Evolution. Piscataway: IEEE Press, 2010: 13–21.

[7] KUHN P, SHEN K L. Gender discrimination in job ads: evidence from China[J]. The Quarterly Journal of Economics, 2012, 128(1): 287–336.

[8] 安小米, 郭明军, 洪学海, 等. 政府大数据治理体系的框架及其实现的有效路径[J]. 大数据, 2019, 5(3): 3–12.
AN X M, GUO M J, HONG X H, et al. Framework of government big data governance system and effective way of implementation[J]. Big Data Research, 2019, 5(3): 3–12.

[9] 郭洪伟. 基于网络大数据的消费者信心指数编制[J]. 统计与信息论坛, 2015, 30(6): 111–112.
GUO H W. Consumer confidence index is compiled based on online big data[J]. Statistics & Information Forum, 2015, 30(6): 111–112.

[10] 邹伟. 基于道路交通指数大数据的上海市主城区交通拥堵特征研究[J]. 上海城市规划, 2017(2): 76–81.
ZOU W. Analysis on traffic congestion of Shanghai central city based on road transportation index big data[J]. Shanghai Urban Planning Review, 2017(2): 76–81.

[11] 耿林, 毛宇飞. 中国就业景气指数的构建、预测及就业形势判断: 基于网络招聘大数据的研究[J]. 中国人民大学学报, 2017, 31(6): 24–35.
GENG L, MAO Y F. Construction of China employment market prosperity index(CIER) and employment situation analysis based on CIER—research based

on large data on network recruitment[J]. Journal of Renmin University of China, 2017, 31(6): 24–35.

[12] 张晓芳. 基于民间融资大数据的地区民间融资风险指数模型研究: 以温州为例[J]. 时代金融, 2017(6): 41–43, 47.
ZHANG X F. Regional private financing risk index model based on private financing big data: a case study of Wenzhou[J]. Times Finance, 2017(6): 41–43, 47.

[13] 魏颖, 刘厉兵. 居民生活质量大数据指标体系的构建与运用[J]. 中国经贸导刊, 2019(16): 15–17.
WEI Y, LIU L B. Construction and application of big data index system of resident's quality of life[J]. China Economic & Trade Herald, 2019(16): 15–17.

[14] 王凯军, 龙厚印, 吴良良, 等. 基于电力大数据的产业结构调整及经济指标研究[J]. 经济研究导刊, 2017(25): 38–39.
WANG K J, LONG H Y, WU L L, et al. Research on industrial structure adjustment and economic indicators based on power big data[J]. Economic Research Guide, 2017(25): 38–39.

[15] 李崇纲, 许会泉. 冒烟指数: 大数据监测互联网金融风险[J]. 大数据, 2018, 4(4): 76–84.
LI C G, XU H Q. Smoke index: big data technologies monitor Internet financial risks[J]. Big Data Research, 2018, 4(4): 76–84.

[16] 钱斌华. 基于大数据的长三角制造业指数建构与预测研究[J]. 现代管理科学, 2019(5): 41–43.
QIAN B H. Research on construction and prediction of manufacturing index in Yangtze River Delta based on big data[J]. Modern Management Science, 2019(5): 41–43.

[17] 高欣东, 马冬妍, 师丽娟. 大数据与实体经济深度融合指数构建及评估实证研究: 以贵州省的实践为例[J]. 经营与管理, 2019(12): 100–104.
GAO X D, MA D Y, SHI L J. An empirical study on the construction and evaluation of the index of deep integration of big data and real economy: a case study of Guizhou Province[J]. Management and Administration, 2019(12): 100–104.

[18] 张宇杰, 安小米, 张国庆. 政府大数据治理的成熟度评测指标体系构建[J]. 情报资料工作, 2018(1): 28–32.
ZHANG Y J, AN X M, ZHANG G Q. Construction of maturity evaluation index system for government big data governance[J]. Information and Documentation Services, 2018(1): 28–32.

[19] 徐映梅, 高一铭. 基于互联网大数据的CPI舆情指数构建与应用: 以百度指数为例[J]. 数量经济技术经济研究, 2017, 34(1): 94–112.
XU Y M, GAO Y M. Construction of the public opinion index of CPI based on the Internet big data[J]. The Journal of Quantitative & Technical Economics, 2017, 34(1): 94–112.

[20] 黄燕芬, 张超. 大数据情绪指数与经济学研究: 现状、问题与展望[J]. 教学与研究, 2018(5): 40–50.
HUANG Y F, ZHANG C. Big data sentiment index and economic research–current situation, problems and prospects[J]. Teaching and Research, 2018(5): 40–50.

[21] 毛通, 谢朝德. 基于百度大数据的信用舆情指数构建与实证研究[J]. 征信, 2020, 38(1): 11–20.
MAO T, XIE C D. Construction and empirical study of public opinion index on credit based on the big-data of Baidu[J]. Credit Reference, 2020, 38(1): 11–20.

[22] 胡晓珂, 张岩. 大数据视角下新城评价指标体系的构建[J]. 住宅产业, 2016(9): 34–38.
HU X K, ZHANG Y. Construction of new city evaluation index system from the perspective of big data[J]. Housing Industry, 2016(9): 34–38.

[23] 阿里研究院. 去年居民品质消费占比超三成[N]. 北京日报, 2017–04–17.
The Ali Institute. Quality consumption accounted for more than 30% in the last year[N]. Beijing Daily, 2017–04–17.

作者简介

郭明军（1978– ），男，博士，国家信息中心大数据发展部副处长、经济师，主要研究方向为政府信息资源管理与大数据治理协同创新。

陈沁（1986– ），男，博士，成都数联铭品科技有限公司首席经济学家，主要研究方向为数据分析处理。

安小米（1965– ），女，中国人民大学信息资源管理学院、中国人民大学数据工程与知识工程教育部重点实验室、中国人民大学智慧城市研究中心教授、博士生导师，主要研究方向为政府信息资源管理与知识管理、智慧城市及其数据治理、大数据与人工智能应用场景下的数据治理。

王建冬（1982– ），男，博士，国家信息中心大数据发展部处长、研究员，主要研究方向为政府大数据分析、数字经济等。

易成岐（1988– ），男，博士，国家信息中心大数据发展部副研究员，主要研究方向为大数据支撑政府决策、社会网络大数据分析等。

收稿日期： 2021-05-07
基金项目： 国家社会科学基金资助项目（No.20&ZD161）
Foundation Item: The National Social Science Fund of China (No.20&ZD161)

·学术科普·

漫威电影中的机器学习

王元卓　中国科学院计算技术研究所
沈英汉　中国科学院计算技术研究所
陆　源　北京科技大学

在《美国队长3：内战》中，钢铁侠与美国队长因观念冲突导致了内战，钢铁侠先不断挨揍以积累大量美国队长的战斗数据，再将这些数据进行计算，分析美国队长的出拳规律，从而扭转局势。这一针对对战数据进行的分析归功于钢铁侠强大的机器学习能力，如图1所示。

那么，什么是机器学习呢？

机器学习是通过对已有数据或者经验的学习来自动改进算法性能的人工智能的重要研究方向。机器学习的目的就是把人类思考归纳经验的过程转化为计算机通过对数据的处理得出模型的过程。得出的模型能够以近似于人的方式解决很多复杂的问题。机器学习的核心是使用算法解析数据，通过一系列运算从数据或经验中学习知识，对某个任务做出决策或者预测，并对效果进行评估。

机器学习中的“训练”与“预测”的过程可以对应到人类的“归纳”和“推测”的过程。通过这样的对应，我们可以发现，机器学习的思想并不复杂，它仅仅是对人类在生活中学习成长的一个模拟。机器学习不是基于编程形成的结果，它的处理过程不是因果的逻辑，而是通过归纳思想得出的相关性结论。

图1《美国队长3：内战》中的机器学习片段

机器学习的发展如图2所示。从20世纪50年代开始研究机器学习以来，不同时期的研究途径和目标并不相同，机器学习的发展可以划分为3个阶段。

第一阶段是从20世纪50年代中叶到70年代中叶，这个时期主要研究“有无知识的学习”。这个时期，由于缺乏丰富的知识，远不能实现真正的智能。最具有代表性的成果就是于1952年创建的第一个真正的机器学习程序——一个简单的棋盘游戏。此后，主要研究工作是将各领域的知识植入系统里，目的是通过机器模拟人类学习的过程。在这一研究阶段，主要用各种符号表示机器语言，科研人员将专家学者的知识加入系统里，并取得了一定的成效。

第二阶段从20世纪70年代中叶到90年代。机器学习逐渐从学习单个概念扩展到学习多个概念，开始把学习系统与各种应用结合起来，并取得很大的成功。同时，专家系统在知识获取方面的需求也极大地推动了机器学习的发展，自动知识获取成为机器学习应用的研究目标。尤其是从20世纪80年代中叶开始，机器学习已成为新的学科，它综合应用了心理学、生物学、神经生理学、数学、自动化和计算机科学等学科知识，形成了机器学习理论基础；融合各种学习方法，产生形式多样的领域应用系统研究方向。

第三阶段从21世纪初期开始，机器学习进入一个全新的阶段。随着深度学习模型和大数据的出现，机器学习成为人工智能新发展的重要方向之一。

接下来我们看看一次典型的机器学习是如何工作的。

图 2 机器学习讲解图

（选自《科幻电影中的科学：科学家奶爸的AI手绘》）

首先是选择数据。这里的数据可以分为三部分，分别是训练数据、验证数据和测试数据。有了数据后，第二步是对数据进行建模，使用训练数据构建涉及相关特征的模型。得到数据之后，第三步就是要验证模型，用之前准备的验证数据验证建立的模型效果。第四步是调试模型。为了提升模型的性能，使用更多的数据、不同的特征来调整参数，这也是最耗时耗力的一步。模型准备完毕后，第五步是使用模型。部署训练好的模型，对新的数据进行预测。最后需要测试模型，使用测试数据验证模型，并评估模型的性能。

几十年来，机器学习的方法种类很多，常见的也是经典的3类方法包括有监督学习、无监督学习和强化学习。

所谓有监督学习，就是我们常说的分类。也就是通过已有的信息获得一个最优的处理模式，再利用这个模式将所有输入的信息处理成输出信息，计算机通过对输出信息的简单判断，将已有信息分成不同的种类，这样人工智能就有了对未知数据进行分类的能力。比如家长经常教育孩子苹果是能吃的，石头是不能吃的。苹果、石头就是输入信息，而家长给出的判断——能吃和不能吃，就是相应的输出信息。当孩子的认识能力达到一定水平时，就会逐步形成一种模式，遇到类似石头的东西就知道不能吃。有监督学习可以说是通过有标签的数据结合标定结果的直接反馈来预测结果或未来的。

与有监督学习不同，无监督学习并没有放置任何可以参考的样本或者已经分类的参考目标，给定的数据集也没有“正确答案”，计算机需要直接对已有数据建立模型，挖掘出潜在的结构。也许有人会问，在没有样本的情况下，计算机如何自己建立模型呢？其实在人类思维过程中，无监督学习是时常发生的。比如在我们对音乐并不十分了解的情况下，如不知道什么是古典音乐，什么是摇滚乐，能自发地将其进行分类，这就是无监督学习。虽然没有人给我们提供模型将听到的音乐进行分类，但是我们依然能够将不同的音乐区分开。当我们根据某些事物的特性将其归为一类时，使用的就是无监督学习中的聚类分析法。

俗话说“物以类聚”，所谓的类就是具有相似元素的事物的集合。聚类分析的目的是在相似的基础上收集数据进行分类。聚类分析的对象被称为描述数据，通过衡量它与不同数据源之间的相似性，就能把不同的数据归到不同的类别。比如我们找到一种植物，并发现它具有青菜的特征，只是颜色不一样，那么我们就可以将其归类到蔬菜中。

还有一类机器学习方法我们称之为强化学习。AI在自己所属的环境中，一边试错一边寻找最合适行动的过程被叫作强化学习。首先要清楚地表现出自己的行动和状况；其次要认识到在什么样的状况下采取什么样的行动，在该环境下会产生什么样的结果；然后从中学习并采取最优行动。学习的线索是获得回报，回报是相对结果的评价值。比如，在格斗游戏中让人类玩家和AI对战。最开始AI会毫无章法地出招，回报是人类玩家的体力有一定程度的消耗。这最初可能无法对玩家造成伤害，但在反复对战的过程中，偶尔会对玩家造成伤害，那么AI就会记住这些场景。通过不断地对战，AI就会学习在什么样的状况下采取什么招式可以削弱对方的体力。

机器学习有巨大的潜力改变和改善世界，我们正朝着真正的人工智能迈进。比如目前火热研发中的无人驾驶汽车，通过机器学习可以实现自动导航，并保证安全行驶。一个例子是交通标志传感器，它使用监督学习算法识别和解析交通标志，并将它们与一组有标记的标准标志进行比较。这样，汽车就能看到停车标志，并认识到它实际上意味着停车，而不是转弯、单向行驶或人行横道等。

大数据

BIG DATA RESEARCH

ISSN 2096-0271

邮发代号：2-537　国外代号：C9118　定价：48.00元

大数据 第八卷 第二期 2022年3月