Computer Science and Application 计算机科学与应用, 2019, 9(9), 1724-1731
Published Online September 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.99193

Hans汉斯

Analysis of Big Data Method and Its Application

Pengfei Guo[1], Gang Li[2*]

[1]College of Computational Science, Zhongkai University of Agriculture and Engineering, Guangzhou Guangdong
[2]Guangzhou E-Government Center, Guangzhou Guangdong
Email: *457616871@qq.com

Received: Aug. 28th, 2019; accepted: Sep. 12th, 2019; published: Sep. 19th, 2019

Abstract

With the advent of the age of information explosion, the property of information data which is Volume, Velocity, Variety, has brought a natural application scenario for data science and big data technology. By introducing the concept, theory and technology of big data, this paper analyzes the application and development status of big data practice, gives a brief panoramic view of big data method and its application, and provides theoretical guidance for big data development in modern information society.

Keywords

Big Data, Machine Learning, Cloud Computing

大数据方法及其应用

郭鹏飞[1]，李　刚[2*]

[1]仲恺农业工程学院，计算科学学院，广东 广州
[2]广州市电子政务服务中心，广东 广州
Email: *457616871@qq.com

收稿日期：2019年8月28日；录用日期：2019年9月12日；发布日期：2019年9月19日

摘　要

随着信息爆炸时代的到来，信息数据本身所具有的大量性、高速性、多样性为数据科学和大数据技术带来了天然的应用场景。本文通过研究大数据的特点、相关理论和技术，剖析了大数据实践应用及发展现状，给出了大数据相关的几个重要方法及其应用的简要全景图，为大数据的推广和应用提供参考。

*通讯作者。

文章引用: 郭鹏飞, 李刚. 大数据方法及其应用[J]. 计算机科学与应用, 2019, 9(9): 1724-1731.
DOI: 10.12677/csa.2019.99193

关键词

大数据，机器学习，云计算

1. 引言

“大数据”的概念已经提出了很多年。从一开始的神秘到如今的快速发展，对人们的生活产生了巨大的影响。从当初在巴士站、的士站人们的久久等待到如今随手滴滴打车的迅速接驾，从起初淘宝购物大海捞针似地寻找心仪的商品到如今淘宝主页精准的商品推荐，从当初为了美食一家一家店的尝试到如今根据美团推荐随心所欲挑选心仪饭店，可以说“大数据”已经渗入到生活的方方面面，为人们的生活提供了巨大的便利。

从二十世纪八十年代进入信息时代以来，信息爆炸和互联网的普及使得数据量呈指数级增长，数据量及类型的增长速度也早已超越了摩尔定律的限制。只 2015 年每天就有 4.7 个万亿字节的数据产生。据 IBM 研究指出目前大概 90%的数据是近两年产生的。百度产生数 PB 的用户搜索数据，八亿八千万在线交易产生的 20 TB 的数据发生在淘宝平台。马云曾在第五届阿里巴巴技术论坛说：“我们正在从 IT 时代走向 DT 时代(数据时代)。IT 和 DT 之间，不仅仅是技术的变革，更是思想意识的变革，IT 主要是为自我服务，用来更好地自我控制和管理，DT 则是激活生产力，让别人活得比你好。”这一切都说明，大数据时代已经到来。

2. 大数据的特点

“大数据”的概念起始于二十世纪九十年代，由 John Mashey 提出并将其发扬光大[1]。该概念通常是指那些传统统计分析软件无法在可容许的时间内抓取、处理的数据集。“大数据”的数据规模不仅仅是一个静态的数值，它可能是一个不断积累的量，所以对其处理所使用的技术手段提出了新的要求。传统的数据分析更多的是采用随机抽样调查的方法，而大数据则是运用所有相关数据来进行挖掘分析。

2001 年，META 集团(Gartner)分析师 Doug Laney 在其报告中从三个维度定义了数据增长：Volume (大量)，Velocity (高速)，Variety (多样)。随后工业界逐渐沿用“3V”模型来描述大数据[2]。2012 年，META 集团对该理论提出了改进指出：大数据是通过新数据科学处理模式才能具有更强决策力、洞察发现力及流程优化能力的海量和多样化的信息资产。海量且多样化的数据为具体问题提供了更多的信息，充分利用海量数据中的信息来解决问题，这就是大数据的价值。2015 年以来，随着深度学习的快速发展，卷积神经网络处理高维空间数据、循环神经网络处理高维序列数据以及深度置信网络实现数据生成模型为大数据产业的升级转化提供了新的技术温床；同时一定程度上实现了真正意义上的大数据高维性($>10^8$)突破[3] [4] [5] [6]。

研究大数据，需要从理论、技术和实践三个方面来展开。具体框架如图 1 所示。

3. 大数据相关的理论

理论是认知的必经途径，是任何技术方法被广泛认同和传播的基线。通过大数据的特征定义理解各行各业对大数据的整体描绘和定性；通过基于机器学习算法的大数据科学建模，实现大数据信息精准导

向在各行各业中的应用，进而实现大数据价值的探讨来深入解析大数据的珍贵所在；因为精准，所以更需要考虑大数据安全隐私问题，从这一重要的视角审视人和数据之间的长久博弈。

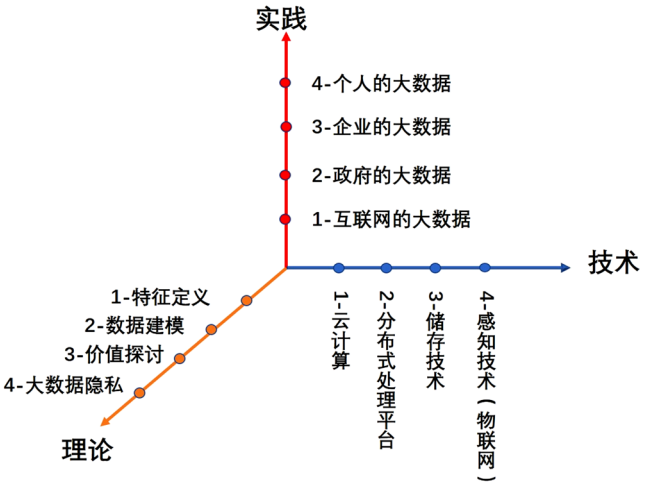

Figure 1. Research dimension
图 1. 研究维度

大数据的理论方法，更多的是从机器学习发展来的。谈到机器学习，首先要理解的是机器学习、数据挖掘和大数据之间的关系。大数据以海量多样数据为研究对象来提出全方位精确解决方案的过程，其中包括数据处理、数据分析和数据挖掘等步骤。然而，数据挖掘却受到很多学科领域的影响，其中数据库、机器学习、统计学无疑影响最大。对数据挖掘而言，数据库提供数据管理技术，机器学习和统计学提供数据分析技术。统计学对大数据的理论的研究发展至关重要，通过统计学界提供的坚实理论基础转化为实用机器学习算法之后进入数据挖掘领域，从而使得基于机器学习算法的大数据研究具有牢固的理论基石。

机器学习是人工智能领域的基础理论方法，主要研究计算机如何模拟或实现人类学习行为，以获取新知识或新技能，重新组织已有的知识结构进而不断改善自身性能。由于部分机器学习并不能处理海量数据，因此需要对相应的算法进行适度改进以适应海量数据，使得算法性能和空间占用达到最优实用的地步[7] [8]。

机器学习算法可以分为监督学习算法和无监督学习算法两类。监督学习算法是对具有标记的训练样本进行学习，以尽可能对训练样本集以外的数据进行标记预测。回归问题就是典型的监督学习问题。无监督学习算法是对没有标记的训练样本进行学习，以发现训练样本集中的结构性知识。聚类就是典型的无监督学习问题。具体的算法分类如图 2 所示。

比较主流、成熟和应用广泛的大数据算法有聚类算法、广义主成分分析、支持向量机(SVM)、决策树算法、逻辑回归、随机森林算法、贝叶斯网、神经网络算法、因子分析和关联分析。正如图中所示，针对不同的问题我们有不同的解决办法。例如：支持向量机、卷积神经网络及贝叶斯网算法是从机器学习算法改进而来的，更多的是针对预测、分类问题；而聚类、广义主成分分析和因子分析等算法是基于统计理论分析不同数据变量之间的相关关系的算法[9]。

探究数据的内在规律，除了高大上的大数据建模算法和数据处理之外，统计学中的基础统计分析方法也尤为重要。在追求更细致、更完美结果的过程中，那些基础的、经典的推理分析所得出的结果或许更能够一针见血的说明问题。

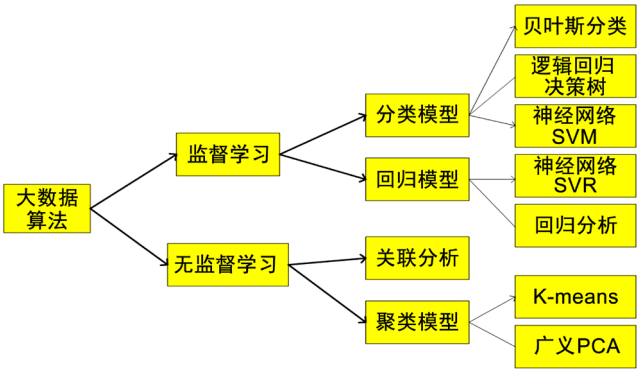

Figure 2. Algorithm category
图 2. 算法分类

4. 大数据的技术支撑

技术是大数据价值的体现手段和前进基石。本部分从云计算、感知技术、分布式处理技术和存储技术的发展来说明大数据从采集、处理、存储到形成结果的整个过程。目前适用于大数据的技术，包括大规模并行处理(MPP)数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

云计算是一种与服务供应商很少交互的可配置的计算资源共享池(资源包括网络，服务器，存储，应用软件，服务)。大数据与云计算的关系就如同一枚硬币正反面一样密不可分。大数据的本质在于对海量数据的挖掘，但这一过程的实现必须依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术。由于实时的大型非结构化数据和半结构化数据集分析需要像 MapReduce 一样的框架来向数十、数百或甚至数千的电脑分配工作，并且这些数据在下载到关系型数据库用于分析时会花费过多时间和金钱，因此需要借助云计算的相关技术以有效地处理大量的容忍经过时间内的数据。大数据与云计算的关系如图 3 所示。

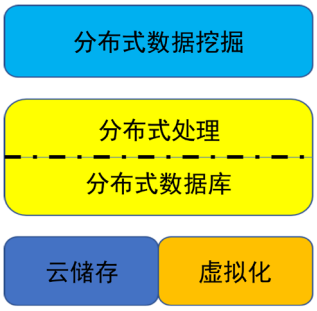

Figure 3. Platform hierarchy
图 3. 平台层次

如图所示，大数据平台的构建，云存储及云计算运算方式是基础。目前，市场上比较流行的 oracle，mysql 数据平台是侧重于数据存储的关系型数据库。随着数据量增大，关系型数据库开始暴露出一些难以克服的缺点，具有扩展性和并发行特点的非关系型数据库得到了快速发展，例如 NoSQL 数据库。

另一类是倾向于数据计算分布式数据库，如 hadoop。分布式数据库是通过高速计算机网络将物理上分散的多个数据存储单元连接起来组成一个逻辑统一的数据库。分布式数据库基本思想是将原来集中式数据库中数据分散存储到多个通过网络连接的数据存储节点上，以获取更大存储容量和更高并发访问量。近年来，随着数据量快速增长，分布式数据库技术也得到了飞速发展。传统的关系型数据库开始从集中

式模型向分布式架构发展，基于关系型的分布式数据库在保留了传统数据库的数据模型和基本特征下，从集中式存储走向分布式存储，从集中式计算走向分布式计算。

由于传统的集成式数据处理方法对于海量数据的限制，分布式处理和并行处理方法应运而生。分布式和并行处理方法是为了提高处理数据速度采用的两种不同体系架构。并行处理是利用多个功能部件或多个处理机同时工作来提高系统性能或可靠性的计算机系统，这种系统至少包含指令级别或指令级别以上的并行。分布式处理是将不同地点、具有不同功能或拥有不同数据的多台计算机通过通信网络连接起来，在控制系统统一管理下，协调地完成大规模信息处理的计算机系统。

由于大数据的“3V”特征，使得传统的数据分析工具对此无能为力，因而需要更高层次的数据平台布局以及存储。一个完整的商用大数据平台架构如图 4 所示。

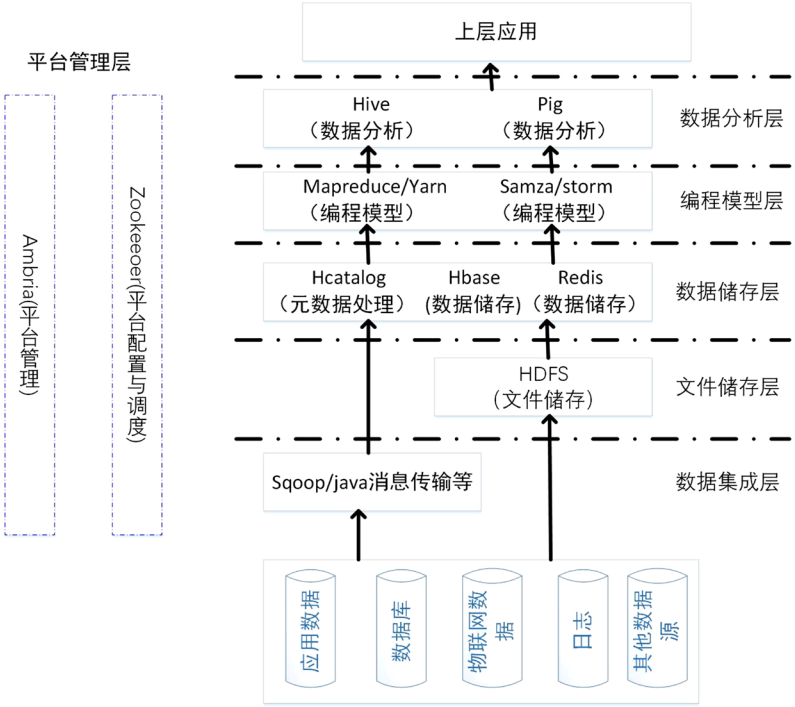

Figure 4. Typical commercial big data architecture
图 4. 典型的大数据商用架构

作为大数据开发中的重要一环，数据平台搭建的合理性往往决定着企业的发展程度。如果没有完善的数据平台和合适的大数据技术，就没有坚实的根基，后期的精准决策更无从谈起。由于大数据平台建设及云计算的发展非常依赖于计算机技术的发展，因此随着计算机后摩尔时代的到来，大数据技术的更新变革将更加迅速。

5. 大数据实践应用及其发展现状

实践是大数据的价值体现。随着大数据浪潮的到来，各行各业都在将大数据应用到实际生产中来。无论是政府、企业还是个人都对大数据的概念和应用充满热情。从图 5 的中国大数据 IT 应用行业投资结构来看，互联网大数据作为技术革新的桥头堡，获得投资份额最高。由于大数据初涉的实用性，新型行业如金融、电信乃至政府也对大数据寄予厚望。下面分别从企业的大数据，政府的大数据和个人的大数据三个方面来描绘大数据已经展现的美好景象及即将实现的蓝图。

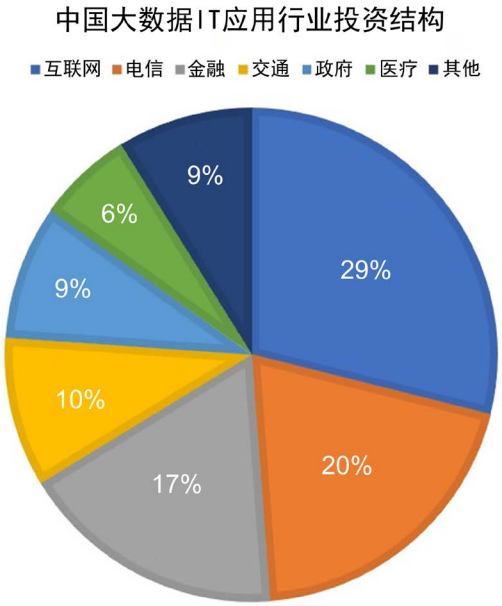

Figure 5. China's IT industry investment structure
图 5. 中国大数据 IT 应用行业投资结构

1) 企业大数据。随着互联网数据大爆发，大数据在新型产业得到迅速发展。目前热门的无人驾驶系统开发、语音识别、自然语言处理、智能推荐等技术都得益于大数据的发展。这些技术的本质是机器学习和模式识别，通过巨量的数据来训练精确模型，使得模型能通过样本数据量的增加不断改进、完善，更精确地指导生产和生活。

从图 5 的大数据行业结构来看，由于互联网企业对数据掌控较多导致其对大数据相关项目投入最多。Google、Facebook、IBM、微软、Uber 等互联网公司在搜索、无人驾驶、社交、移动互联网及网约车等方面都做出了重要的贡献。谷歌通过其支柱业务搜索引擎布局大数据架构，通过大数据聚类和关联算法将信息更精确的分类，提供贴近用户习惯的搜索结果。随之而来的是庞大的信息量，进而使得谷歌在大数据平台架构发展上也走在世界前列，导致目前常用的数据平台(GFS、MapReduce 和 BigTable)均与其有着千丝万缕的联系。此外，Google 通过 Tensor flow 开源平台推动人工智能、深度学习的发展。依托 Google 丰富的数据库，谷歌的子公司 DeepMind 开发的围棋机器人 AlphaGo 通过神经网络、深度学习、价值网络、蒙特卡洛树搜索法等机器学习算法打败了围棋九段高手李世石。微软的 cortana，苹果的 siri 等语音助手通过网页搜索技术、知识搜索技术、知识库技术和语言模型技术实现智能语音控制，在生活中担任了管家的角色；特斯拉的无人驾驶通过卷积神经网络的深度学习算法实现视觉识别计算达到数据的积累自主学习的目的，进而实现智能控制的无人驾驶。

2) 政府大数据。互联网企业通过大数据技术崛起的势头，让政府看到了提高行政能力、提高政策透明度的希望，图像视频识别和数据聚类技术可以增加预防犯罪和改善国家安全的效率；全生活链的大数据分析使得通过精确分类让公民获得更好的教育和医疗福利的可能性得到极大的提高。政府在行政和管理的过程中，拥有五类独特类型数据。分别为：a) 政府行政管理采集的数据：资源类、税收类、财政类等；b) 政府行政管理汇总的数据：如建设、农业生产总值、工业生产总值等；c) 政府公共资源产生的数据：如城市基建、交通基建、医院、教育师资等；d) 政府监管职责产生的数据；如人口普查、食品药品管理等；e) 政府公共服务的客户级消费和档案数据：如社保、水电、教育信息、医疗信息、交通路况、公安等。由数据属性分类，政府数据有自然信息类(地理、气象、环境、资源、水利等)、城市建设类(旅

游景点、住宅建设、交通设施等)、城市健康管理统计监察类(人口、机构、企业、工商、税收、商品等)和服务与民生消费类(水、电、通信、医疗、燃气、出行等)等。基于上述数据，下面从具体的交通管理、医疗卫生、公共安全和教育方面展开讨论。

交通管理——交通管理方面，通过对道路交通信息实时数据挖掘，有效缓解交通拥堵，并快速响应突发状况，为城市交通的良性运转提供科学的决策依据[10]。通过整合道路交通、公共交通、对外交通大数据，汇聚气象、环境、人口等行业数据构建交通大数据平台，提供道路交通状况判别及预测，辅助交通决策管理，支撑智慧出行服务，加快了交通大数据服务模式的创新(图 6)。

Figure 6. Traffic big data
图 6. 交通大数据

医疗卫生——通过整合医疗、药品、气象和社交网络等相关医疗信息数据，构建医疗大数据平台，形成智能临床诊断模式和自主就医模式的创新，为市民、医生、政府合理优化医疗资源配置。同时提供流行病跟踪与分析、临床诊疗精细决策、疫情监测及处置、疾病就医导航、健康自我检查等服务[11]。

公共安全——在公共安全领域，通过大数据的挖掘，可以及时发现人为或自然灾害、恐怖事件，提高应急处理能力和安全防范能力。针对公共安全领域治安防控、反恐维稳、情报研判、案情侦破等实战需求，建设基于大数据的公共安全管理和应用平台。汇聚融合涉及公共安全的人口、警情、网吧、宾馆、火车、民航、视频、人脸、指纹等海量业务数据，建设公共安全领域的大数据资源库，全面提升公共安全突发事件监测预警、快速响应和高效打击犯罪等能力。

教育——针对全民学习、终身教育的需求，建设教育大数据服务平台。积累数字教育资源，收集教育服务平台学习者行为数据和学习爱好数据，能够为千万级学习者提供个性化的终身在线学习服务，提高教育资源的共享和利用率，实现因材施教，优化教学过程，提高教学质量，为教育政策调整提供决策支持。

3) 个人大数据。对个人来说，时时刻刻都与大数据接触着，其实我们正是企业、政府大数据实践中的受益者。网上购物时，通过浏览记录及购买记录，淘宝或者京东等电商利用这些个人信息建立用户画像，更好的描述个人的偏好，实现精准推荐，使我们能够更快的发现想购买的物品和评价更高的商品；手机打字时，第三方输入法通过我们的日常的词语组合习惯将常用词优先级提前，使我们的打字更加快捷；医院则会通过个人就诊记录对每个人的健康进行评估与预测，防患于未然；交通部门大数据中心通过各观测点交通量的观测，实时调整运营线路，优化线路配置，缓解拥堵，使我们出行更加便利。最终，

我们做一切事产生的所有数据都将服务于我们本身，这就是大数据所带给我们的。正是因为当前大数据的兴起，个人数据不仅对企业、政府重要，同样对不法分子有很大的吸引力。因此，对于个人隐私的保护和信息安全是今后需要努力改进的方向。

6. 总结

随着大数据浪潮的到来，政府通过出台一系列大数据产业发展规划与战略文件，持续推动大数据技术产业与传统领域实现全方位融合，促进我国经济结构转型的升级、提升经济发展质量和国际竞争力。本文从大数据由来、大数据的概念(“3V”)、大数据理论核心算法、实现大数据中心化分布式计算技术及其应用前景等方面给出了经典机器学习统计背景下的大数据简化全景图。尽管近几年大数据理论技术及其在各行各业中的应用迅猛发展，但通过本文层次化分析不难发现：目前数据科学家对于大数据相关底层基础概念还存在一定的分歧；对于大数据核心算法的应用局限于储存和局部的搜索匹配而非整体链式进化驱动应用；并且大数据核心技术的中心化性与目前先进信息安全技术(区块链：去中心化)有一定的冲突。正是因为上述问题的存在，让我们认识到大数据理论技术与产业的发展挑战与机遇同在。紧跟国家大数据产业发展的战略步伐，加强大数据核心理论技术创新意识，助力传统领域产业结构升级是数据科学家关注的核心问题。

参考文献

[1] Mashey, J.R. (1997) Big Data and the Next Wave of Infra-Stress. Computer Science Division Seminar, University of California, Berkeley.

[2] Laney, D. (2001) 3-D Data Management: Controlling Data Volume, Velocity and Variety. Application Delivery Strategies by META Group Inc., Stamford, 949.

[3] Majumder, N., Poria, S., Gelbukh, A. and Cambria, E. (2017) Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intelligent Systems*, **32**, 74-79. https://doi.org/10.1109/MIS.2017.23

[4] Zhang, Q., Yang, L.T. and Chen, Z. (2016) Deep Computation Model for Unsupervised Feature Learning on Big Data. *IEEE Transactions on Services Computing*, **9**, 161-171. https://doi.org/10.1109/TSC.2015.2497705

[5] Grushka-Cockayne, Y., Jose, V.R.R. and Lichtendahl, K.C. (2017) Ensembles of Overfit and Overconfident Forecasts. *Management Science*, **63**, 1110-1130. https://doi.org/10.1287/mnsc.2015.2389

[6] Amin, J., Sharif, M., Yasmin, M. and Fernandes, S.L. (2018) Big Data Analysis for Brain Tumor Detection: Deep Convolutional Neural Networks. *Future Generation Computer Systems*, **87**, 290-297. https://doi.org/10.1016/j.future.2018.04.065

[7] Zhou, L.N., Pan, S.M., Wang, J.W. and Vasilakos, A.V. (2017) Machine Learning on Big Data: Opportunities and Challenges. *Neurocomputing*, **237**, 350-361. https://doi.org/10.1016/j.neucom.2017.01.026

[8] Huck, N. (2019) Large Data Sets and Machine Learning: Applications to Statistical Arbitrage. *European Journal of Operational Research*, **278**, 330-342. https://doi.org/10.1016/j.ejor.2019.04.013

[9] Jeon, S. and Hong, B. (2016) Monte Carlo Simulation-Based Traffic Speed Forecasting Using Historical Big Data. *Future Generation Computer Systems*, **65**, 182-195. https://doi.org/10.1016/j.future.2015.11.022

[10] 赵鹏军, 李铠. 大数据方法对于缓解城市交通拥堵的作用的理论分析[J]. 现代城市研究, 2014(10): 25-30.

[11] 惠榛, 李昊, 张敏, 等. 面向医疗大数据的风险自适应的访问控制模型[J]. 通信学报, 2015, 36(12): 190-199.