# 浅析领域知识对大数据发展的影响

冷莉华[1] 廖伊婕[2] 廖鸿志[1]
(云南大学软件学院 昆明 650500)[1] (云南开放大学资产管理与评价处 昆明 650032)[2]

**摘 要** 大数据正在成为当今社会的热门话题,除了 IT 领域对它的不断探索外,大数据也在持续影响着经济、社会的进程。各行各业对大数据的炒作愈演愈烈,应该认真思考大数据研究与应用过程中面临的诸多问题。知识领域对于一项新技术来说尤为重要,若不清楚各行各业的知识领域,则新技术在不同领域中的发展及应用将会面临很多阻碍[1]。分析了大数据研究与应用的几个环节,指出了大数据处理涉及到的数据采集、数据管理、数据分析、数据建模和数据应用等过程,其中领域知识是沟通大数据处理的核心环节。由于大数据应用的关键是数据分析,数据分析的基础是领域知识,因此大数据处理必须通过领域知识才能前后贯通。

**关键词** 大数据,领域知识,系统分析
**中图法分类号** TP311 **文献标识码** A

## Analysis of Influence of Domain Knowledge on Development of Big Data

LENG Li-hua[1] LIAO Yi-jie[2] LIAO Hong-zhi[1]
(College of Software, Yunnan University, Kunming 650500, China)[1]
(Office of Asset Management and Evaluation, Yunnan Open University, Kunming 650032, China)[2]

**Abstract** Big data is becoming a hot topic in today's society. Except of the IT field for its continuous exploration, it also continues to affect the economic and social progress. In the face of all walks of life to the big data hype, we should think seriously about the problems faced in the process of big data research and application. Domain knowledge is very important for a new technology, if we don't know the field of domain knowledge in all walks of life, the development and application of them in different fields will face a lot of obstacles. This paper analyzed several aspects of the research and application of big data, and pointed out that big data processing involves data collection, data management, data analysis, data modeling and data application. Domain knowledge is the core of communication of data processing. As the key to the application of big data is data analysis, and data analysis is based on domain knowledge, the processing of big data must be through the field of domain knowledge to breakthrough.

**Keywords** Big data, Domain knowledge, System analysis

## 1 引言

"The prospect of a fresh start is certainly alluring"(新事物的前景无疑是诱人的),就如同当前公众对于大数据、云计算这样的新技术、新趋势趋之若鹜,各界既有理性的思考,也有跟风的哄抬,使得有些新技术、新概念被扭曲、异化。因此,除了对新技术保持极大的原始渴望外,还应该认真思索新技术在发展中存在的一些问题。

当今,社会信息化和网络化的发展导致数据爆炸式增长,据 IDC《数字宇宙》[2]的研究报告表明,2020 年全球新建和复制的信息量已经超过 40ZB,是 2012 年的 12 倍;而中国的数据量在 2020 年将超过 8ZB,比 2012 年增长 22 倍。大数据的爆炸式增长引发学术界和商业界对它的密切关注。在中国,北京、武汉、黑龙江、贵阳等多地建立了大数据中心和大数据交易服务平台[3];《Nature》[4]、《Science》[5]等杂志相继推出了关于大数据的专刊;各行各业都争先恐后地涌进大数据的潮流中,希望能够从中获得最新、最有价值的信息。中国拥有庞大的消费群体和大量互联网用户(2016 年,互联网用户和移动互联网用户预计分别为 6.5 亿和 9.8 亿[6]),这为大数据带来了前景无限的发展机会。

大数据的处理模式主要分为流处理[7](stream processing)和批处理[8](batch processing)。批处理指先存储后处理,而流处理指直接处理。大数据的处理步骤涉及到数据采集、数据管理、数据分析、数据建模和数据应用等过程,其中数据采集和数据管理通常是由信息技术专家完成,数据建模是由数学与统计学专家完成,数据分析涉及到领域知识的问题,但是目前在大数据研究与应用中,熟悉领域知识的领域专家经常被忽略,造成许多领域内大数据的发展鲜有实质性进展。一些大数据企业虽然考虑到了大数据产业链上的所有环节,但也不能解决产业链贯通的问题。目前的大数据企业往往衍生于原来的信息企业,可以很好地解决数据管理的问题,并可以建立大数据处理的模型库,提供了一批大数据处理平台,但

本文受国家自然科学基金项目(61263022,61303234),云南省科技创新人才计划项目(2015HB038)资助。
冷莉华(1992—),女,硕士生,主要研究方向为大数据发展问题和心电信号的可视化,E-mail:970194088@qq.com;廖伊捷(1976—),女,博士,主要研究方向为经济管理学、经济历史学等;廖鸿志(1946—),男,博士,教授,博士生导师,主要研究方向为计算数学、系统工程、信息安全等。

这些平台不能根据领域知识解决数据与处理贯通的问题。对于一直接触纯技术层面的技术人员来说，他们迫切需要了解已经获取的这些数据所涉及到的相关领域知识，这对下一步的系统建模和系统应用尤为重要。本文将着重讨论领域知识与大数据发展之间的关系，以及领域知识对大数据发展所产生的深远影响。

## 2　大数据概述

### 2.1　大数据的基本概念和特点

单从字面上来看，大数据表示数据的规模很庞大，这通常也是绝大多数人对大数据概念的理解。但是仅仅从数量上理解大数据肯定是不够的。Gartner 公司给出了大数据的定义：大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程化能力的海量、高增长率和多样化的信息资产。从最具代表性的 3V[9] 定义（即：规模性、高速性和多样性），到 IBM 在 3V 的基础上提出的 4V[10] 的定义（即：规模性、高速性、多样性和价值性），目前对大数据的定义尚未有一个标准，我们暂且将大数据理解为体量巨大、类型繁多、处理速度较快、数据价值密度较低的数据集合[11]。

大数据的特点也是从定义的 4 个层面来理解：1）数据体量巨大，只有数据体量达到 PB 级别以上的才能被称为大数据，其中 1PB＝1024TB，1TB＝1024GB，它超过传统数据库存储容量的百倍；2）数据类型繁多，例如：网络数据、通信数据、交通数据、视频、图片、音乐等；3）处理速度快，一秒定律，通过算法对数据的逻辑处理速度非常快，因此能从各类信息中快速获得高价值信息；4）价值密度低，数据本身是没有价值的，只有被充分利用起来才能体现数据的潜在价值。

### 2.2　大数据的处理步骤

大数据处理涉及到数据采集、数据管理、数据分析、数据建模和应用等过程。

大数据采集是指利用多个数据库接收来自客户端的数据，由于数据的类型多种多样，因此对数据的采集方式也各不相同，主要包括：系统日志采集法，比如 Facebook 的 Scribe 和 Hadhoop 的 Chukwa；网络数据采集法，包括网络爬虫、URL 队列和数据处理等模块。

大数据管理主要是数据的存储，主流的技术是云存储技术，它将数据存储在云主机上，保证数据的安全、高效。云存储的出现从功能上弥补了传统存储的不足，通过虚拟化大容量存储、分布式存储和自动化运维等功能，实现了存储空间无限扩充，其自动化和智能化功能提高了存储效率[12]。

大数据分析包括数据统计和数据分析，是大数据处理的核心[13]。众所周知，大数据已经不是简单地指数据体量大，只有通过分析才能获得更多深入、有价值的信息。不同的行业对应不同的领域知识，如果领域知识理解不到位很容易误导用户。因此，理解不同领域的领域知识对于大数据的发展尤为重要。

大数据建模也叫数据挖掘，其主要是基于各种算法的计算来预测可以将该批数据应用到的领域，从而发挥大数据的价值。

## 3　大数据应用的关键

大数据应用的关键是数据分析，数据分析的基础是领域知识。

### 3.1　拥有领域知识才能知道大数据的用途

阿里巴巴是中国最大的电商，通过电子商务获取的数据可以分析商品的市场需求及变化趋势，甚至可以分析不同区域对商品的嗜好，但由于缺少网上购物者的真实身份信息以及商品直接使用者的信息，因此除了有直接销售对象的商品外，不能分析其他商品对不同年龄段人群的吸引力[14]；熟悉经济活动的规律才能分析金融数据反映的经济变化等。

数据需求是由领域知识决定的，领域知识能告诉我们已有数据能做什么，为了某种分析需要什么数据。缺少领域知识就可能白白浪费已有的数据资源，也可能去做已有数据不能支撑的分析。

### 3.2　拥有领域知识才能建立正确的分析模型

影响区域内人口变化的因素包括不同年龄段育龄妇女人数、生育模式、全社会育龄妇女平均总和生育率等，若了解人口发展的领域知识，就可利用人口发展方程来进行人口变化预测；若不懂领域知识，则只会用时间序列模型进行趋势预测。

在公共安全领域中，掌握了领域知识就可知道应该从哪些数据中分析犯罪嫌疑人的行为、犯罪手段和行踪。例如侦破震惊全国的甘肃白银连环杀人案的关键是基因分析技术，警方已经多次获得犯罪嫌疑人的体液信息，却不懂得遗传基因的规律，导致二十多年不能正确锁定嫌疑人群，惨案连续发生。

在生产、商业、军事、科研开发和民生等几乎所有的领域，不了解过程的规律就不能选择正确的分析模型[15]，不能充分地利用大数据技术分析问题。

### 3.3　拥有领域知识才能处理分析样本

在已有数据和模型的基础上分析样本的选择也离不开领域知识。科布-道格拉斯生产函数是一个简单的经济活动分析模型：

$$Y=AL^{\alpha}K^{\beta}$$

其中，$Y$ 是劳动产出，$L$ 是劳动投入，$K$ 是资金投入。如果有了若干年的样本数据就直接进行数据拟合，那么往往会得到错误的结果。其原因是在部分年份中，当年的投入并不会在同年就产生产出，例如基础设施的投入以及周期较长的技术改造的投入等。因此必须依靠领域知识对数据进行重新匹配来获得新的分析样本，否则分析的结果会大相径庭。

## 4　大数据技术大有可为

大数据分析对于提升经济发展、社会进步、国家与社会安全水平有重大作用。

我国电子政务第一期工程就规划了经济运行数据库的建设，目前已经积累了海量经济运行数据，并已经进行了部分分析，在国家经济转型、产业升级、对外贸易、金融政策制定、物价控制等领域起到了一定作用，运用大数据技术并关联其他领域的数据还可以大大扩充应用领域，提升应用水平。

在社会管理方面，我国连续启动了教育、科技、卫生、医疗、人口、社会保障、气象、减灾防灾等众多领域的数据建设，对可能影响社会稳定和人民生活的一些问题进行了预控。例如，通过诊疗系统的数据，可以分析持不同种类医保卡的用户所用药物是否为自己所用。2015 年，国家针对未来人口结构可能失调、老龄化问题初现端倪的情况进行大数据分析，适时

（下转第 74 页）

[8] 刘清. Rough 集及 Rough 推理[M]. 北京:科学出版社,2001.
[9] JELONEK J, KRAWIEC K, SLOWINSKI R. Rough set reduction of attributes and their domains for neural networks[J]. Computational Intelligence, 1995, 11(2): 339-347.
[10] WANG J, WANG J. Reduction algorithm based on discernibility matrix the ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489-504.
[11] 王治和,崔晓慧. 改进的差别矩阵启发式属性约减算法[J]. 计算机工程与设计,2016,37(4):1032-1036.
[12] SKOWRONA, RAUSZER C. The discernibility matrices and functions in information systems [M]. Dordrecht: Kluwer Academic Publishers, 1992: 331-362.
[13] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001:30-62.
[14] 杨明,杨萍. 差别矩阵浓缩及其属性约简求解方法[J]. 计算机科学,2006,33(9):181-183.
[15] 殷志伟,张健沛. 基于浓缩布尔矩阵的属性约简算法[J]. 哈尔滨工程大学学报,2009,30(3):307-311.
[16] CHEN D G, YANG Y Y, DONG Z. An incremental algorithm for attribute reduction with variable precision rough sets[J]. Applied Soft Computing, 2016, 45: 129-149.
[17] JAMES N, LIU K, HU Y X, et al. A set covering based approach to find the reduct of variable precision rough set[J]. Information Sciences, 2014, 275(3): 83-100.
[18] 黄卫华,晏林,冯云再. 广义变精度粗糙集模型与其它粗糙集模型的比较[J]. 兰州文理学院学报,2015,29(5):6-8.
[19] 杨勇,朱影. 一种基于 MapReduce 的粗糙集并行属性约简算法[J]. 重庆邮电大学学报,2015,27(1):90-96.

(上接第 49 页)

调整了计划生育政策。

近日,接连发生的电信诈骗事件受到广泛关注,引起了政府和公众对社会安全问题的担忧。徐玉玉被骗 9900 元学费死亡、清华教师遭诈骗犯以“公检法”名义诈骗 1760 万元等事件无不揭露了电信诈骗的危害所在。这类事件的发生除了有关部门监管不严之外,大数据应用的滞后是主要原因之一。

电信诈骗源于可疑的短信、电话和邮件等,据统计,在日常生活中 50%的电话为无用电话,其中 40%是推销电话,10%是诈骗电话。从诈骗电话的号码源类型来看,固定电话呼出的诈骗电话数量超过半数,其次是 400 和 800 电话占 27.1%;手机呼出的诈骗电话仅占 15.4%;1.2%的诈骗电话来自境外呼入。运营商知道这些规律以后应该标记来电,提醒用户谨防受骗。此外,诈骗电话有一定规律,如往往针对同一对象以不同电话号码、不同用户身份连番实施,如果电信运营商以保护公众通信安全为己任,就可以通过大数据分析针对可疑通话提醒用户,减少用户损失。

银行业也是阻止电信诈骗的一道有力屏障。针对诈骗方式分析,犯罪嫌疑人通常是将一笔数额较大的款项以最快的速度分散到若干账户并在极短时间内通过不同网点的取款机取走。只要具备银行业的领域知识,不难从银行交易大数据中发现这种规律。以前银行业对可疑交易的分析主要是针对洗钱行为,一般银行的可疑交易系统实际上只是反洗钱系统,对于目前日益猖獗的诈骗犯罪,银行业应该考虑尽快采用大数据技术开发反诈骗系统,最大限度地保障用户的金融资产安全。

结束语 由于对大数据研究与应用的系统分析不够,以及领域专家的知识结构限制,目前的大数据研究基本上处于不同专家各抒己见、不同专家都在渲染自己所熟悉的研究阶段,所以大数据研究与应用还是声势大而实效小,投入大而成果小。解决上述问题的途径包括:1)信息技术专家、应用领域专家、模型与统计专家密切配合,并且各类专家都要增强对其他专家知识的理解;2)要有一批能将几个环节的知识融会贯通的“杂家”。各领域的专家要放下身段,学习一些相关领域的知识,身体力行、扎扎实实地进入一两个应用领域做好一两个项目,才能进一步推进大数据的研究与应用。

## 参 考 文 献

[1] 彭怡,寇纲. 基于领域知识的数据挖掘理论框架研究[C]// 第三届(2008)中国管理学年会--信息管理分会场论文集. 北京:中国管理现代化研究会,2008.
[2] Cloud information. 2020 global information will be more than 40ZB, reaching 12 years 12 times[EB/OL]. [2016-09-05] https://www.aliyun.com/zixun/content/1_1_15290.html.
[3] 冯登国,张敏,李昊. 大数据安全与隐私保护[J]. 计算机学报,2014,37(1):246-258.
[4] Nature. Big Data[EB/OL]. [2016-09-05]. http://www.nature.com/news/specials/bigdata/index.html.
[5] Science. Special online collection: Dealing with data[EB/OL]. [2016-09-05]. http://www.sciencemag.org/site/special/data.
[6] Gartner report: big data will be popular in Chinese[EB/OL]. [2016-09-05]. http://mobile.163.com/16/0824/06/BV7BBSM800118024.html.
[7] KUMAR R. Two computational paradigm for big data [EB/OL]. [2016-09-05]. http://kdd2012.sigkdd.org/sites/images/sum-merschool/Ravi-Kumar.pdf.
[8] Information Week Report. The big data managementchallenge [R/OL]. [2016-09-05]. http://reports.informationweek.com/abstract/81/8766Business-intelligence-and-information-management/research-the-big-Data-management-challenge.html.
[9] GROBELNIK M. Big-data computing: Creating revolutionary breakthroughs in commerce, science, and society [R/OL]. [2016-09-05]. http://videoLectures.net/eswc2012_grobelnik_big_data.
[10] BARWICK H. The “four Vs” of Big Data, Implementing InformationInfrastructure Symposium [EB/OL]. [2016-09-05]. http://www.computerWorld.com.au/article/396198/iiis_four_vs_big_data.
[11] 李广建,化柏林. 大数据分析与情报分析关系辨析[J]. 中国图书馆学报,2014,40(213):14-22.
[12] 陈杰. 大数据场景下的云存储技术与应用[J]. 中兴通讯技术,2012,18(6):47-51.
[13] ROUSSEAU R. A view on big data and its relation to Informetrics[J]. Chinese Journal of Library and Information Science, 2012(3): 12-26.
[14] 庄峻,褚燕. 基于领域知识的企业知识模型[J]. 华东电力,2013,41(5):1101-1104.
[15] 石杰. 基于知识的企业战略管理系统及其模型研究[D]. 西安:西北工业大学,2003.
[16] WANG Y, et al. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations[OL]. http://dx.doi.org/10.1016/j.techfore.