卫生大数据的分析框架与技术

李岳峰

国家卫生和计划生育委员会统计信息中心, 北京 100044

E-mail: liyf@nhfpc.gov.cn

2014-12-17 收稿, 2015-03-23 接受, 2015-07-23 网络版发表

摘要 探讨了卫生大数据的应用现状、发展前景及其价值实现, 提出了卫生大数据分析的基本框架, 分析了卫生大数据的收集、传输、维护、集成、分析和模式解释等步骤中的关键技术, 以及卫生大数据发展中的几个认识问题, 提出大数据应用的相关政策建议。

关键词

卫生大数据
分析
方法

1 卫生大数据的应用价值

随着电子计算机和网络通讯技术的广泛应用, 人类社会正处于一个全新的信息时代. 2012年, 我国提出促进工业化、信息化、城镇化、农业现代化同步发展的战略, 明确了信息化在“四化”同步发展中的重要地位, 信息化作为“四化”目标之一, 吸纳了更新的科技水平, 代表了更先进的生产力, 信息化已成为我国经济社会发展的新方向和新动力, 贯穿于“四化”同步发展的始终, 为实现“四化”提供坚强的支撑和保障^[1]. 大数据、云计算、移动互联网和物联网共同构成了人类社会从传统“物质世界”向未来“数字世界”转变的关键技术群. 2015年57%的新建数据中心可能都是云数据中心^[2]. 2008年, *Nature*出版的大数据专刊“Big data”, 从互联网技术、网络经济学、超级计算、环境科学和生物医药等多个方面介绍了海量数据带来的挑战^[3]. 2011年, *Science*推出关于数据处理的专刊“Dealing with data”, 讨论了数据洪流(Data Deluge)所带来的机遇^[4], 认为更有效地组织和利用这些数据, 将对经济社会的发展产生巨大的推动作用^[5]. 2011年, 麦肯锡(MGI)发布了《大数据: 创新、竞争和生产力的下一个新领域》的报告, 认为大数据是无法在一定时间内用传统数据库软件工具对其内

容进行抓取、管理和处理的数据集合, 有效处理这些数据能够释放出巨大的潜在价值^[6]. 2012年初, 在瑞士达沃斯举行的世界经济论坛上, 大数据作为框定的主题之一, 该论坛发布的*Big Data Era*报告, 认为大数据作为一种新型的经济资产, 就像货币或黄金一样珍贵^[7]. 2012年, 奥巴马政府发布*Big Data Research and Development Initiative*的报告, 提出将通过收集庞大而复杂的数据资料, 获得知识, 提升能力, 加速推动其在科学、工程上应用的步伐^[8]. 无论是学术界还是产业领域, 大数据的研究都已成为一个热门话题. 大数据不仅是人们获得新认知, 创造新价值的源泉, 而且还是改变市场、组织机构以及政府与公民关系的方法. 它强调3个转变, 即从强调随机样本到更重视全体数据, 从强调精确性到更重视混杂性, 从强调因果关系到更重视相关关系^[9].

近些年, 国内外的卫生信息化(health informatization)发展都很快. 美国前总统克林顿在2013年医疗卫生信息与管理信息系统协会(healthcare information and management systems society, HIMSS)大会主旨演讲中指出, 卫生信息化不仅是降低成本的工具, 更是实现患者参与的重要手段, 通过数据管理, 才能更好地服务于决策、流程再造和居民健康^[1]. 医疗卫生领域是一个数据高度密集型的行业, 在各级医疗机构中逐

引用格式: 李岳峰. 卫生大数据的分析框架与技术. 科学通报, 2015, 60: 2396-2403

Li Y F. Analytic framework and related technologies for big health data (in Chinese). *Chin Sci Bull*, 2015, 60: 2396-2403, doi: 10.1360/N972014-00721

<https://engine.scichina.com/doi/10.1360/N972014-00721>

步得到广泛使用,如电子病历(electronic medical record, EMR)、电子健康档案(electronic health records, EHR)、计算机化医嘱录入(computerized physician order entry, CPOE)、图像存档通信系统(picture archiving and communication system, PACS)、临床决策支持系统(clinical decision support system, CDSS)和实验室信息系统等,将产生大量的医疗卫生服务和健康相关数据,称为“卫生大数据”(health big data)。随着以居民电子健康档案为核心的区域信息平台 and 以电子病历为核心的医院信息系统的建设,以及居民健康卡的推广应用,卫生大数据正在以几何数量级增长。以卫生统计为例,90余万家医疗卫生机构通过统计直报系统上报实时、月报、年报的医疗卫生机构、卫生人力、卫生资源与卫生服务利用、疾病报告与健康监测等数据。以每个电子计算机断层扫描(computed tomography, CT)图像含有约150 MB的数据,每个基因组序列文件约为750 MB,每个标准病理图则接近5 GB计算,乘以我国人口数量和平均寿命,那么每个社区医院或中等规模制药企业均可以生成和累积达到TB甚至PB级的结构化和非结构化数据,数据挖掘分析的经济价值无可估量。据估计,2012年,全球数字医疗数据有500 PB,预计2020年将达到25000 PB^[10]。卫生大数据是卫生系统最重要的信息资源(health information resources, HIR)之一,在整个卫生信息化体系中处于核心地位^[11]。许多卫生管理专家认为,利用大数据发现有价值的知识,能够服务卫生决策、增加病人安全、消除不必要的成本等^[12,13]。联合国已经推出了名为“全球脉动”(global pulse)的新项目,希望利用大数据预测某些地区的失业率、支出削减或是疾病爆发等现象,其目标在于利用数字化的早期预警信号来提前指导援助项目^[14]。本文通过梳理学界有关卫生大数据分析框架的研究成果,从技术层面探讨卫生大数据在卫生领域的应用以及面临的挑战,为卫生管理和相关领域研究提供参考。

2 卫生大数据的分析框架与技术

大数据具有数据体量(volumes)巨大、数据类别(variety)繁多、价值(value)密度低和处理速度(velocity)快等特点,通过大数据技术,分析数据和产生数据的环境及其影响因素,提取有用信息(information),找出信息变化的规律性,总结出知识

(knowledge),利用知识的逻辑框架及其知识库建立,形成智慧(wisdom)这样一个螺旋式学习提升和价值发现过程^[15]。然而,数据操作问题、复杂性和昂贵或不可用的计算方案等,都是医疗保健、生物医学研究数据管理和分析的难点^[16]。按照统计学将统计分析分为统计设计、数据收集、数据整理和数据分析等4个步骤的思路,本文将卫生大数据分析划分为卫生大数据的收集、传输、集成、维护、分析和模式解释等6个步骤(或基本框架),对每个步骤涉及的技术问题与方法进行了较为系统的分析。

2.1 卫生大数据的收集

通过医疗卫生服务活动等运营式系统收集的数据,是卫生大数据的最基本数据,如医院病人的电子病历记录,公共卫生监测数据和卫生统计等数据。传统意义上的卫生数据收集通过普查、抽样调查、统计报表、重点调查和典型调查等形式。从数据类型看,按照数据特征可以有多种分类方式,如定性数据和定量数据,定类数据、定序数据、定距数据和定比数据,时间数据和空间数据,结构化数据、半结构化数据和非结构化数据等,不同的数据类型决定所采用的数据收集方式,也直接关联收集数据的信息系统构建。近年来,随着信息技术的发展,卫生大数据产生的方式发生重大变化,用户原创内容阶段和感知式系统阶段产生的数据成为卫生大数据的主要来源。数据模型作为信息系统的核心,它定义了数据的类型、关系、操作以及保证数据库完整性的规则,建立科学的数据模型,有利于数据的收集和分析^[17]。

用户原创内容阶段数据包括以博客、微博和微信等为代表的新型社交网络平台,和以智能手机、平板电脑等为代表的新型移动互联平台,医疗卫生机构及医务人员、居民、企业和政府相互之间,搜索、交换和共享健康信息,自生成和互交换产生大量的卫生大数据。

感知式系统阶段大数据主要依靠物联网(internet of things, IOT)技术,将各种信息传感设备嵌入物体、设施和环境,与互联网结合形成一个巨大智能网络,将物联网技术运用到健康监测领域,可以实现个体化、连续性、跨区域的远程健康监测服务,实现健康管理的全程可控以及个性化服务。例如,可穿戴技术(wearable technology)是将生理信息检测技术融入人们日常穿戴的衣物或饰品,实现人体基本生理参

数(心电、呼吸、体动、体温等)的长期动态获取和实时诊断,并进行生理信号的无线传输,还可在异常情况下定位和报警.无线传感器网络(wireless sensor network, WSN)是由部署在监测区域内大量微型计算机,通过无线通信方式形成的自组织网络系统,感知、采集和处理网络覆盖区域中被感知对象的信息,并发送给观察者.躯域传感网络(body sensor network, BSN)是以人体为传输媒介进行生物信号、电信号的传输,是实现人体生理信号获取、分析与处理的基本平台^[18].感知式系统收集的数据具有实时动态性、指标多样性和区域灵活性等特点^[19].当然,用户原创内容阶段和感知式系统阶段产生的大数据也存在安全性、数据噪声、医学数据的检测精度要求、数据采集密度与功耗、费用问题和法律法规等需要进一步研究的问题.

2.2 卫生大数据的传输

卫生大数据通常涉及时间(实时还是非实时、不同时间段)、空间(区域划分、空间地理位置)和人员组织(数据的采集方、数据的利用方)等维度,建立数据仓库是一项非常重要的任务.数据仓库技术涉及管理大量数据、管理多介质存储、索引和监视数据、多种技术的接口、对数据存放位置的控制、数据的并行存储/管理、元数据管理技术、语言接口、数据的高效装入和高效索引技术、数据压缩技术、复合键码技术、变长数据、锁管理切换技术等^[20],数据库处理技术是卫生大数据收集的关键技术.卫生大数据传输采用两种传输技术:(i)高速文件传输技术.由中国科学院北京基因组研究所(Beijing Institute of Genomics, BGI)开发完成的EasyGenomics系统,采用Aspera's fasp™高速文件传输技术,持续以每秒10千兆比特率横跨太平洋转移基因组数据^[21].(ii)数据压缩技术.数据压缩能大大降低网络负担的同时降低存储要求,并便于快捷传输.随着我国深化医改工作,转诊服务、远程医疗、信息共享和纵向医疗资源流动等,推动了医疗机构之间、医疗卫生机构与区域信息平台之间、区域信息平台相互之间的庞大的数据传输.相比而言,医学图像压缩和传输技术更为复杂,也应用更为普遍.医学图像较普通图像纹理更多,分辨率更低,相关性更大,存储空间要更大,其处理技术包括图像变换、图像压缩、图像增强、图像平滑、边缘锐化、图像分割、图像识别等^[22].在医学

图像传输中,吴晶等人^[23]引入可逆整数小波变换、渐进编码和熵编码,提出了一种连续帧医学图像无损渐进传输方案,给出了基于C/S结构的VC实现.黄祥国等人^[24]借鉴了eFilm Workstation和开放源代码软件在局域网中传输医学图像的方法.在医学图像压缩中,陈家新等人^[25]提出了基于各向异性扩散的分水岭分割算法.吕洁和熊春荣^[26]提出一种结合Live Wire算法和轮廓插值算法的三维医学图像分割算法.在医学图像融合上,张少敏等人^[27]提出了一种融合梯度信息的最小生成树医学图像配准算法,增强鲁棒性.夏云等人^[28]提出一种基于拉普拉斯谱的医学图像配准算法,提高了传感器医学图像配准的精度.目前大数据传输中主要困难在于压缩算法往往难以计算,并导致转换变慢.

2.3 卫生大数据的集成

卫生大数据非常大、分散、非结构化和异质性,使得集成和转换面临诸多问题.集成非结构化数据对于大数据分析是一个主要挑战,即使是结构化的电子病历数据,Kuo等人^[29]认为集成方面存在功能、元数据和实例等挑战.

功能集成是2个或2个以上的系统借助读写相同文件格式(file formats)以及采用相同协议(protocols)交换信息的能力,实现互操作(interoperability).目前比较主流的Oracle数据库和MySQL数据库,使用不同的数据结构存储数据,2个数据库的编码系统不一致,会导致功能集成问题.

元数据是描述资源特点的结构化的数据,在关系数据库模型中,列名称作为元数据来描述存储的数据的特点.关于元数据集成有2个主要问题:(i)不同的数据库系统使用不同的元数据描述内容;(ii)简单元数据映射复合元数据存在问题.

数据实例整合的问题是,在卫生术语、计量单位和卫生信息系统之间使用的代码集有很多的变量,不同的诊断信息编码会导致编码映射问题.当前编码系统的交叉映射进展较慢.

在原始数据中有效整合和处理非结构化数据是非常困难的.目前集成结构化数据的解决方案主要有2种:(i)用户干预方法.自动模式(多个元数据)或数据实例映射算法总是产生错误,因为大数据包含了太多的元数据,不能手动检查错误,一些研究认为使用人群的反馈来提升集成水平.计算机系统在

加载目标架构后,采用匹配规则找出模式中最适当的匹配表和相关领域,当应用程序生成的映射定义完成,生产映射规范,其中包含所有的数据库表和与HL7 RIM类和属性相关的列.该模型的主要好处是,用户的干预显著提高模式定位精度.存在问题,确定用户反馈的可靠性非常困难,且人为干预大量元数据映射是不现实的.(ii) 概率方法.概率集成方法利用概率选择模式中对象之间的关系,得出概率后,设定选择匹配和非匹配对象的阈值,消除整合过程中的不确定性.概率方法还能够从一组数据源中自动创建一个中间模式,进行数据源和模式之间的语义映射^[30].

2.4 卫生大数据的维护

卫生大数据维护有多种解决方案,包括云计算、网格计算、NoSQL/NewSQL和其他存储系统等,每种方案各有其优势和劣势^[31~33].云计算作为一种性能、存储和网络资源共享和多租户环境,为卫生大数据提供强大的存储和计算能力,来自大数据的业务需求,能为云计算的落地找到更多更好的实际应用,云计算是最符合成本效益和有发展前景的大数据维护方式,其架构包括:云计算环境层(包括分布式文件系统、并行编程环境、分布式系统管理);数据采集层(从各个数据源并行采集数据);数据清洗层(包括数据约束检查、冗余处理、抽取转换及加载等);并行分析层(包括维度定义、关联规则定义、并行算法及报表中心等)^[34].从存储系统看,分布式缓存、分布式数据库和非关系型数据库系统技术有效提升了存储效能.如Memcached是一个高性能的分布式内存对象缓存系统,用于动态Web应用以减轻数据库负载.Greenplum 数据引擎软件基于大规模并行处理和完全无共享架构.Hive是一个基于Hadoop的数据仓库平台,将转化为相应的MapReduce程序,基于Hadoop执行.Big Table是建立在谷歌文件系统上的专用分布式数据库系统.Cassandra是一个开源数据库管理系统,处理分布式系统上的大量数据.HBase是一个开源分布式非关系型数据库,可在PC Server上搭建大规模结构化存储集群.Dynamo是由亚马逊开发的专用分布式数据存储系统等.当然集中存储和共享租赁的空间意味着敏感的数据可能会受到恶意攻击,关于数据安全,相比本地存储数据,云计算在某种程度上更安全.基于云计算技术的有关数据

隐私、使用与知识产权等法律问题需要探讨^[35].

2.5 卫生大数据的分析

卫生大数据的分析是卫生大数据应用的落脚点,也是其价值实现的过程.关键之处在于选择或建立预测模型,处理多类型和大规模的数据集,分析和预测复杂的卫生问题.选择或建立预测模型的难度在于:(i) 分析问题的复杂性.如果研究的问题是NP难度,即多项式复杂程度的非确定性问题,计算时间可能是超级指数级的^[36].例如,贝叶斯网络是计算生物学和生物信息学领域常用建模算法,当数据量增加时,寻找最好网络的计算时间呈指数增加.(ii) 数据级的影响.对于一些复杂的分析,当数据集不是很大时,SQL可以快速查询反馈结果.当列表包含亿行级无索引的数据时,许多传统的统计分析方法和数据挖掘算法很难并行化,并行处理(massively parallel processing, MPP)工具难以有效分析,计算时间会很长.(iii) 计算资源的可及性.卫生大数据所使用超级计算资源,获取和维护成本是非常高昂.

针对建模分析存在的难题,有以下可行的解决措施:

(i) 按数据及时性要求、数据量大小和网络带宽资源及时间段流量约束,选择流处理(stream-processing)或批处理(batch-processing)方式.流处理是实时获取动态数据,进行及时监测和预警.由于响应时间的要求,流处理的过程基本在内存中完成,通常不会对所有的数据进行永久化存储,其处理方式更多地依赖于在内存中设计巧妙的概要数据结构(synopsis data structure),内存容量是限制流处理模型的一个主要瓶颈,以相变存储器(phase change memory, PCM)为代表的储存级内存(storage class memory, SCM)设备的出现或许可以使内存未来不再成为流处理模型的制约.批处理是先存储后处理(store-then-process),如Google公司在2004年提出的MapReduce^[37],编程模型是最具代表性的批处理模式.

(ii) 按照数据分析的关系和结构要求,研究选择合适的计算平台(硬件).用于大型数据集的计算平台主要分为超级计算、网格计算、云计算和异构计算等4种模式^[36].超级计算机是指大型电脑(超级计算机)或连接在一起的廉价集群计算机(计算机集群),通过一个快速的局域网(local area network, LAN)作为一个超级计算机运行.超级计算机提供更快处

理速度、更大的存储容量、更好的数据完整性和可靠性,可以利用现有的计算机资源,其实施和维护费用比较昂贵. 网格计算是一个结合了计算机资源分布位置的处理器架构. 网格的分布式性质计算对用户是透明的,具有松耦合、异质性和地理分布性,面对一个共同的计算任务时,可以更好地利用大规模的计算资源. 执行好这样的计算任务需要集体学习、标准共享和良好的协作. 云计算指的是一种按需、自助服务的互联网基础设施,用户可以随时随地访问计算资源. 但是隐私权保护、数据管辖权、数据损失问题以及网络带宽瓶颈等,都是云计算要逐步解决的问题. 异构计算是指使用不同类型的计算处理器如通用处理器(general-purpose processor, GPP)、专用处理器和自定义加速逻辑的计算机架构,能够为开发人员提供更具灵活性、更高的速度和精度的计算性能. 但是由于空前的并行规模、异质性计算、互连和存储单元,使得编写该架构高性能的程序及其困难.

(iii) 按照数据分析的关系和结构要求,研究选择适宜的开发工具(软件). 大数据分析通常涉及到大量的数据维护和密集计算. 传统的统计软件包如Excel, SPSS, MATLAB等处理任务的能力有限,运用商业智能(business intelligence, BI),经过抽取(extraction)、转换(transformation)和装载(load),即ETL过程的数据挖掘过程受到信息系统的限制. MapReduce作为编程模型,用于处理大数据集的并行运算,或者靠近存储资产减少传输数据和处理. 原则上可以使用任何编程语言,如Java, C++或R等,实现一种基于MapReduce的解决方案, Hadoop是较为成功的开源的MapReduce实现软件平台,支持数据密集型分布式计算. R是一个免费软件的编程语言和软件环境,可以用于开发统计软件和数据分析,它提供了各种各样的统计和图形技术,包括线性和非线性建模、经典统计检验、分类、聚类、时间序列分析等. 在大数据分析和虚拟可视化中, R和Hadoop是可以相互弥补的. 几个R可以用来提高Hadoop的功能(如RHadoop, Ricardo, Bluesnp和PBDR)^[37,38]. 除R包外,统计分析系统(statistics analysis system, SAS)也被使用来提高Hadoop的性能. 黄山等人^[39]分析了MapReduce列存储、索引、连接、迭代计算、科学计算及调度算法方面的优化技术,指出了MapReduce技术研究的挑战性问题.

(iv) 按照数据分析的关系和结构要求,研究选

择科学的算法. 大数据分析的最重要的方面是选择一个合适的分析算法. 并行随机存取模型(parallel random access machine, PRAM)是依据并行算法(parallel algorithm)建立的模型,是目前使用很广泛的一种抽象的并行计算模型. 并行算法的性能分析比串行算法考虑的因素要多,不仅要考虑算法的时间和空间复杂性,而且还要考虑算法的其他性能参数(加速比、效率、可扩展性等)^[40]. 目前算法研究比较多,有学者概括为集成学习、自然语言处理、情绪分析、遗传算法、神经分析、模式识别和监督式(或无监督式)学习等形式,也有基于机器学习的模式分类技术(或数据挖掘技术),如决策树、Bayes分类、k-近邻、神经网络、遗传算法、支持向量机等经典算法^[32].

2.6 卫生大数据的模式解释

通过卫生大数据分析,可能揭示医疗卫生服务和居民健康现状、趋势的可能模式或结果(pattern/outcome),为政府卫生循证决策、医疗服务干预和居民自我保健等提供数据支持. 在运用描述性统计分析、Cronbacha信度系数分析、探索性因素分析(exploratory factor analysis)和验证性因素分析(confirmatory factor analysis)、结构方程模型分析(structural equations modeling)等基本方法的基础上,可以在终端设备展示实时动态的数据及其关联项变化. 对卫生大数据的模式解释,由于分析的复杂性,需要卫生、信息等相关领域专家的合作. 此外,要考虑大数据也存在并不能避免扭曲、误差和错误的假设,因此,通过信息透明,共享数据仓库、数据建模、数据模式或结果,使发现的模式/结果成为医疗卫生领域有用的知识. 验证模式的方法有3种:(i)数据或模型的统计有效性检验;(ii)将数据分为训练集和测试集来验证模型精度;(iii)相关领域专家论证.

3 有关问题的思考

3.1 防止将卫生大数据价值泛化

2007年计算机图灵奖得主Jim Gray认为基于协同化、网络化与数据驱动为其主要特征的大数据改变了传统的研究范式,在学界引起了巨大反响,数据价值提到一个前所未有的高度^[41]. 目前全球范围内广泛兴起的数据关联运动、政府数据开放运动,增加了

数据资源的可获得性与可用性^[42]。但是也要看到,在大数据时代,如个人隐私的保护能力、数据所有权与知识产权保护、数据使用的溯源、数据再利用价值的评价等问题也备受关注。在2013年美国HIMSS大会上一个重要的议题是“患者要求将我的数据还给我”,需要深思的是,患者的数据所有权、医疗机构的患者数据采集和保存权如何界定,研究机构和企业如何利用这些卫生大数据等。因此,研究人员一方面要重视卫生大数据的价值,另一方面也需要利用自己的专业知识思考有关大数据产生的方法、分类和语境,防止可能的概念炒作^[43]。

3.2 防止实用主义主导的寻因分析不足

医学研究是关乎居民生命健康的领域,卫生大数据提供了可能的预警模式,但是这种变量之间的关系是基于相关性,缺乏因果分析证据。因果分析需要严格的抽样框或实验样本设计、群组和时间序列研究、混杂因素的控制等传统统计学方法,即便如此,健康的影响因素还存在交叉混杂、时间继起性、归因系数难以准确测量等问题。因此,研究者要重视信息化支撑下的大数据价值,也需要详实的因果研究,其实两者并不矛盾,是相辅相成的关系,通过卫生大数据分析,可以揭示因果关系的可能概率,为因果分析提供证据,因果分析是对大数据分析结论的验证。

3.3 防止过分关注技术忽视数据质量,导致数据检错成本增加

数据质量直接影响大数据分析模式或结果。曹建军归纳数据质量定义有:(i)数据质量是指模式和数据实例的一致性、正确性、完整性和最小性等4个指标在信息系统中满足的程度;(ii)数据质量是数据适合使用的程度;(iii)数据质量是数据满足特定用户期望的程度;(iv)数据质量是数据的一组固有特性满足要求的程度^[44]。大数据技术立足庞大的

样本量,放宽了对数据质量的要求。但是倘若出现系统性误差,大数据纠偏的成本比小样本要大得多,甚至涉及系统的改变。通过设定进入程序分析的数据区间、逻辑检错和不同系统收集数据的比对分析可以有效提高大数据质量。物联网、传感网及互联网等网络环境的开放性,使得大数据知识服务体系中的数据质量面临考验,数据正在向着异质、异构、半结构化、非结构化发展。正如美国Gartner咨询公司所指出的那样,全球1000强公司中有超过25%的关键数据不正确或不精确^[45]。王元卓等人^[5]提出从数据计算伦理学、数据密码学、分布式编程框架中的安全计算、远程数据计算的可信度、数据存储和日志管理的安全性、基于隐私和商业利益保护的数据挖掘与分析、强制的访问控制和安全通信、多粒度访问控制以及数据来源和数据通道的可信等方面加强对大数据的研究。

4 总结与展望

卫生大数据具有不可替代的巨大的应用价值。目前卫生大数据的研究应用还处于探索和发展阶段,一些技术尚不成熟,即使是较为成熟的技术,也有学习和知识扩散的过程。可以预期是,随着社会和学界的日益重视,卫生大数据分析技术将更加成熟,其发展前景将更加广阔。本文认为,针对卫生大数据要有开放的心态,主动的学习探索意识,更加重视卫生大数据的研究和应用推进工作。要立足国内卫生大数据的发展现状和需求,借鉴国外大数据技术发展经验,开展联合协作和攻关研究,探索适合我国卫生大数据发展的推进模式。要看到技术的渐进性与超越发展的有机关系,目前卫生大数据处于探索和积累经验阶段,随着卫生大数据所依赖的区域卫生信息平台、云数据中心、隐私保护、网络与数据安全、标准、法律法规和人才队伍建设的快速推进,卫生大数据发展必将步入快车道。

参考文献

- 1 Li Y F, Meng Q. The enlightenment to China's health informationization from new ideas and technology of HIMSS Conference (in Chinese). Chin J Health Inform Manage, 2013, 3: 233-238 [李岳峰, 孟群. HIMSS 大会新理念和技术对我国卫生信息化建设的启示. 中国卫生信息管理杂志, 2013, 3: 233-238]
- 2 Wu H Q. Evolution and security challenges of the internet (in Chinese). Speech on China Internet Security Conference 2013. Beijing: 360 Center for Internet Security, 2013 [郭贺铨. 互联网的演进和安全挑战. 2013 中国互联网安全大会讲话. 北京: 360 互联网安全中心, 2013]

- 3 Donovan S. Big data. *Nature*, 2008, 455: 1–136
- 4 Steven J B. Dealing with data. *Science*, 2011, 331: 639–806
- 5 Wang Y Z, Jin X L, Chen X Q. Network big data: Present and future (in Chinese). *Chin J Compute*, 2013, 6: 1126–1138 [王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望. *计算机学报*, 2013, 6: 1126–1138]
- 6 Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. MacKinsey Global Institute, 2011-05, http://www.mckinsey.com/mgi/publications/big_data/pdfs/MGI_big_data_full_report.pdf
- 7 Science and Technology of China (in Chinese). Big Data Era [EB/OL], 2012-11-15, <http://www.kejizhongguo.org/> [科技中国. 大数据时代[EB/OL], 2012-11-15, <http://www.kejizhongguo.org/>]
- 8 The White House. Big Data Research and Development Initiative [EB/OL], 2012
- 9 Viktor M S, Kenneth C, Sheng Y Y, et al., Translated. Big Data: A Revolution That Will Transform How We Live, Work, and Think (in Chinese). Hangzhou: Zhejiang People's Publishing House, 2013 [维克托·迈尔-舍恩伯格, 肯尼思·库克耶著. 盛杨燕, 周涛译. 大数据时代: 生活、工作和思维的大变革. 杭州: 浙江人民出版社, 2013]
- 10 Sun J, Reddy C K. Big data analytics for healthcare. In: *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining*. New York: ACM, 2013. 1525–1525
- 11 Lu Z K. Health information resources (in Chinese). *Chin Health Res*, 2001, 6: 282–284 [陆章铿. 卫生信息资源. *中国卫生资源*, 2001, 6: 282–284]
- 12 Shah N H, Tenenbaum J D. The coming age of data-driven medicine: Translational bioinformatics' next frontier. *J Am Med Inform Assoc*, 2012, 19: e2–e4
- 13 Garrison L P Jr. Universal health coverage—big thinking versus big data. *Val Health*, 2013, 16: S1–S3
- 14 Wang Z. Strategic value and inspiration of accelerating big data technology development by US Government (in Chinese). *China Dev Obs*, 2012, 6: 44–45 [王忠. 美国推动大数据技术发展的战略价值及启示. *中国发展观察*, 2012, 6: 44–45]
- 15 Zhou G H, Li Y F. Discussion of applying big data in the healthcare industry (in Chinese). *Chin J Health Inform Manage*, 2013, 8: 296–305 [周光华, 李岳峰. 医疗卫生领域大数据应用探讨. *中国卫生信息管理杂志*, 2013, 8: 296–305]
- 16 Anderson N R, Lee E S, Brockenbrough J S, et al. Issues in biomedical research data management and analysis: Needs and barriers. *J Am Med Inform Assoc*, 2007, 14: 478–488
- 17 Date C J. *An Introduction to Database Systems*. Boston: Addison-Wesley Publishing Company, 1995
- 18 Yan J L, Li N, Du X J, et al. Discussion on the development of wearable mobile health technology application research in telemedicine (in Chinese). *China Digit Med*, 2013, 5: 105–108 [闫军玲, 李楠, 杜小加, 等. 穿戴式移动医疗技术在远程医疗中的应用研究进展综述. *中国数字医学*, 2013, 5: 105–108]
- 19 Meng X F, Ci X. Big data management: Concepts, techniques and challenges (in Chinese). *J Comput Res Dev*, 2013, 1: 146–169 [孟小峰, 慈祥. 大数据管理: 概念、技术与挑战. *计算机研究与发展*, 2013, 1: 146–169]
- 20 Wang Y. Relation of data warehouse and data mining and its safety (in Chinese). *J Comput Res Dev*, 2008, 5: 144–149 [王预. 数据仓库与数据挖掘的关系及其安全性问题. *计算机技术与发展*, 2008, 5: 144–149]
- 21 Dai L, Gao X, Guo Y, et al. Bioinformatics clouds for big data manipulation. *Biol Direct*, 2012, 43: 1–7
- 22 Zhou G H, Li Y F, Meng Q. Fuzzy clustering analysis in medical image processing applications (in Chinese). *Chin J Health Inform Manage*, 2011, 4: 69–73 [周光华, 李岳峰, 孟群. 模糊聚类分析在医学图像处理中的应用. *中国卫生信息管理杂志*, 2011, 4: 69–73]
- 23 Wu J, Chen S Z, Zeng L B, et al. Research and implementation of Lossless progressive transmission of medical image sequence (in Chinese). *Comput Eng*, 2006, 23: 259–261 [吴晶, 陈淑珍, 曾立波, 等. 连续帧医学图像无损渐进传输的研究与实现. *计算机工程*, 2006, 23: 259–261]
- 24 Huang X G, Xu F, Tang T F. The economical and practical picture and text report system of medical images (in Chinese). *Med Device Inform*, 2006, 7: 80–81 [黄祥国, 徐芳, 唐陶富. 经济实用的医学影像图文报告系统. *医疗设备信息*, 2006, 7: 80–81]
- 25 Chen J X, Wu Y, Li W. Watershed segmentation algorithm for medical image based on anisotropic diffusion filtering (in Chinese). *J Comput Appl*, 2008, 6: 1527–1529 [陈家新, 吴颖, 黎蔚. 基于各向异性扩散的医学图像分水岭分割算法. *计算机应用*, 2008, 6: 1527–1529]
- 26 Lü J, Xiong C R. Interactive medical image segmentation algorithm (in Chinese). *Comput Simulat*, 2010, 12: 262–266 [吕洁, 熊春荣. 交互式医学图像分割算法. *计算机仿真*, 2010, 12: 262–266]
- 27 Zhang S M, Zhi L J, Zhao D Z, et al. Minimum spanning tree integrated with gradient information for medical image registration (in Chinese). *J Northeast Univ: Nat Sci*, 2010, 31: 1393–1396 [张少敏, 支力佳, 赵大哲, 等. 融合梯度信息的最小生成树医学图像配准. *东北大学学报(自然科学版)*, 2010, 31: 1393–1396]
- 28 Xia Y, Liang D, Bao W X, et al. Medical image registration algorithm based on Laplacian spectrum (in Chinese). *Comput Eng*, 2011, 14: 231–235 [夏云, 梁栋, 鲍文霞, 等. 于拉普拉斯谱的医学图像配准算法. *计算机工程*, 2011, 14: 231–235]

- 29 Kuo M H, Kushniruk A, Borycki E. A comparison of national health data interoperability approaches in Taiwan, Denmark and Canada. *Electron Healthc*, 2011, 2: 14–25
- 30 Fagin R, Kimelfeld B, Kolaitis P G. Probabilistic data exchange. *J ACM*, 2011, 4: 1–55
- 31 Kumar A, Bawa S. Distributed and big data storage management in grid computing. *Int J Grid Comput Appl*, 2012, 2: 19–28
- 32 Moniruzzaman A B M, Hossain S A. NoSQL Database: New era of databases for big data analytics classification, characteristics and comparison. *Int J Database Theory Appl*, 2013, 4: 1–14
- 33 Lith A, Mattson J. Investigating storage solutions for large data. Master Dissertation. Goteborg: Chalmers University of Technology, 2010
- 34 Yan X F, Zhang D X. Big data research (in Chinese). *Comput Technol Dev*, 2013, 4: 168–172 [严霄凤, 张德馨. 大数据研究. 计算机技术与发展, 2013, 4: 168–172]
- 35 Buyya R, Ranjan R. Special section: Federated resource management in grid and cloud computing systems. *Future Gener Comp Sy*, 2010, 26: 1189–1191
- 36 Schadt E E, Linderman M D, Sorenson J, et al. Computational solutions to large-scale data management and analysis. *Nat Rev*, 2010, 11: 647–657
- 37 Das S, Sismanis Y, Beyer K S, Ricardo: Integrating R and Hadoop. In: *Proceedings of the 2010 ACM SIGMOD/PODS International Conference on Management of data*. New York: ACM, 2010. 987–998
- 38 Huang H, Tata S, Prill R J. BlueSNP: R package for highly scalable genome-wide association studies using Hadoop clusters. *Bioinformatics*, 2013, 1: 135–136
- 39 Huang S, Wang B T, Wang G R, et al. A survey on MapReduce optimization technologies (in Chinese). *J Front Comput Sci Technol*, 2013, 10: 865–885 [黄山, 王波涛, 王国仁, 等. MapReduce 优化技术综述. 计算机科学与探索, 2013, 10: 865–885]
- 40 Chen G L, Sun G Z, Xu Y, et al. Methodology of research on parallel algorithms (in Chinese). *Chin J Comput*, 2008, 9: 1493–1498 [陈国良, 孙广中, 徐云, 等. 并行算法研究方法学. 计算机学报, 2008, 9: 1493–1498]
- 41 Jim G. On eScience-A transformed scientific method. In: Tony H, Stewart T, Kirstin T, eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, 2009: 19–33
- 42 Hou J C, Fang J Y. Review on data citation in the context of big data (in Chinese). *J Library Sci China*, 2012, 203: 112–117 [侯经川, 方静怡. 大数据时代的数据引证研究: 进展与展望中国. 图书馆学报, 2012, 203: 112–117]
- 43 Editor. What's the big fuss about 'big data'? *J Health Serv Res Policy*, 2014, 19: 67–69
- 44 Cao J J, Diao X C, Wang T, et al. Basic questions in research on data quality control (in Chinese). *Microcomput Inform*, 2010, 3: 12–14 [曹建军, 刁兴春, 汪挺, 等. 数据质量控制研究中若干基本问题. 微计算机信息, 2010, 3: 12–14]
- 45 Qin X Z, Li C H, Mai F J. An investigation of the connotation, typical characteristics and conceptual model of the big data knowledge service (in Chinese). *Inform Doc Serv*, 2013, 2: 18–21 [秦晓珠, 李晨晖, 麦范金. 大数据知识服务的内涵、典型特征及概念模型. 情报资料工作, 2013, 2: 18–21]

Analytic framework and related technologies for big health data

LI YueFeng

National Health and Family Planning Commission Statistical Information Center, Beijing 100044, China

This paper discusses the development status and value of big health data; proposes the basic framework for big health data; describes the key technologies for the collection, transmission, maintenance, integration, analysis, and interpretation of big health data; discusses several awareness problems related to the development of big health data; and finally presents the proposal of related policy suggestions.

big health data, analysis, method

doi: 10.1360/N972014-00721