普通高等教育“十三五”规划教材

医学大数据应用概论

娄　岩　主编

张志常　马　瑾　副主编

科学出版社

北　京

内 容 简 介

本书是继 2015 年中国医科大学计算机教研室编写的《医学大数据挖掘与应用》之后的又一本面向大数据在医学领域应用的教材。本书遵循定义、特征、技术流程和医学应用典型案例分析的逻辑，抽丝剥茧，由易到难，有助于读者理解和掌握大数据技术。本书应用案例围绕医学大数据及其相关应用这一主线，递进展开，内容具体，过程详尽，并且具有一定的操作性，既方便教师教学，又能引起读者自主学习的兴趣，加深对知识的理解，以及对学习效果的检验。

本书可作为医学院校本科生、研究生的教学用书，也可供医学从业人员，尤其是致力于医学数据处理的人员自学和参考。

图书在版编目（CIP）数据

医学大数据应用概论/娄岩主编．—北京：科学出版社，2017

（普通高等教育“十三五”规划教材）

ISBN 978-7-03-055999-9

Ⅰ．①医… Ⅱ．①娄… Ⅲ．①医学-数据处理-高等学校-教材

Ⅳ．①R319

中国版本图书馆 CIP 数据核字（2017）第 310786 号

责任编辑：宋 丽 陈将浪 / 责任校对：马英菊

责任印制：吕春珉 / 封面设计：东方人华平面设计部

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

印刷

科学出版社发行 各地新华书店经销

*

2017 年 11 月第 一 版 开本：787×1092 1/16

2017 年 11 月第一次印刷 印张：10 1/2

字数：250 000

定价：30.00 元

（如有印装质量问题，我社负责调换〈 〉）

销售部电话 010-62136230 编辑部电话 010-62135927-2014

本书编写人员

主　编　娄　岩

副主编　张志常　马　瑾

参　编　郑琳琳　刘尚辉　李　静　丁　林

徐东雨　曹　阳　庞东兴　霍　妍

前　言

我们正处在一个新技术和传统行业相融合的智能时代，大数据、AR、VR 和人工智能等信息技术必将撬动传统行业的各个板块，为社会发展和时代进步注入新的血液。习近平总书记在十九大报告中提出要“推动互联网、大数据、人工智能和实体经济深度融合”，强调“贯彻新发展理念，建设现代化经济体系”。

国务院在 2015 年印发的《促进大数据发展行动纲要》中明确指出，大数据成为推动经济转型发展的新动力、重塑国家竞争优势的新机遇、提升政府治理能力的新途径。坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

在智能医学与健康服务的大潮中，构建电子健康档案、电子病历数据库，建设覆盖公共卫生、医疗服务、医疗保障、药品供应、计划生育和综合管理业务的医疗健康管理和服务大数据应用体系势在必行；探索预约挂号、分级诊疗、远程医疗、检查检验结果共享、防治结合、医养结合、健康咨询等服务，优化形成包括规范、共享、互信的诊疗流程在内的医学大数据应用也摆到了我们面前。作为医学院校的教育工作者，应该为创新医学院校人才培养模式，建立健全多层次、多形态的应用人才培养体系，培养具有统计分析、计算机技术、医学知识等多学科知识的跨界复合型人才做出贡献。我们希望能够在医学学生中开展大数据知识普及和教育培训，培育具备大数据技术的应用创新型人才，提高医学学生对大数据的整体认知和应用水平。

为此，本书围绕医学大数据应用，从理论、相关技术和实际应用 3 个层面进行了简明扼要的阐述，目的是让广大师生对大数据在医学领域的应用方法和相关知识有所了解，更好地把握科学发展的方向。

我校已连续 4 年将大数据技术及相关课程纳入大学计算机基础教育中，为国家培养了一批又一批掌握最新科学发展动态和技能的数字化医学人才，同时积累了一定的教学经验。本书针对医学学生的特点和大数据在医学领域的应用策略编写，理论联系实际，书中全部案例和解决问题方法均采用与数字医学密切相关的内容。

本书的一大亮点是每章中将大数据在医学领域中的应用落地，注重方法运用、案例解析及可操作性。另外，本书注重启发式的学习策略，便于读者理解和掌握。全书在每章均附有实际应用案例与关键词注释，方便读者查阅和自学。

本书由娄岩担任主编，张志常、马瑾担任副主编。全书包括 11 章，具体编写分工如下：第 1 章大数据概论由娄岩编写，第 2 章医学大数据采集由郑琳琳编写，第 3 章大数据分析由刘尚辉编写，第 4 章大数据可视化由李静编写，第 5 章 Hadoop 由马瑾编写，第 6 章 HDFS 和 Common 由丁林编写，第 7 章 MapReduce 由徐东雨编写，第 8 章 NoSQL 由曹阳编写，第 9 章 Spark 由庞东兴编写，第 10 章云计算与大数据由张志常编写，第 11 章大数据在医疗领域的应用由霍妍编写。

科学出版社对本书的出版做了精心策划和充分论证，在此向所有参加编写的同事们、帮助和指导过我们工作的朋友们及参考文献中的作者们表示衷心的感谢！

由于编者水平有限，加之时间仓促，书中难免存在疏漏之处，恳请广大读者批评斧正！

娄　岩

2017 年 11 月

目　录

第 1 章
大数据概论

导学

本章主要介绍大数据的技术架构、大数据分析的 4 种典型工具及大数据未来的发展趋势。通过对本章的学习，读者可以更好地了解大数据技术。

了解：大数据未来的发展趋势、大数据隐私和安全问题。

掌握：大数据的核心，大数据的数据格式、基本特征，大数据的技术架构，大数据分析的 4 种典型工具，大数据在医学领域的应用。

随着现代科技的发展，人们对数据的认识及处理能力不断提高，开始挖掘、利用、存储、开发、分析大数据（Big Data），从而造福于社会。大数据即目前人们认识的数据全部，其来源广泛，数据格式多元化，用传统的数据挖掘和处理技术已无法对其进行处理，如非结构化，时间敏感或信息量巨大，无法通过关系数据库引擎进行处理的数据。这些类型的数据，需要采用不同的方法和实时且具有分布式处理能力的并行硬件设备来处理。

大数据究竟是什么？有哪些相关技术？医学大数据如何应用？大数据未来的发展趋势如何？本章将一一介绍这些问题。

1.1 大数据技术概述

从技术层面上看，大数据无法用单台计算机进行处理，必须采用分布式计算架构，其特色在于对海量数据的挖掘、分析和处理。同时，大数据必须依托一些现有的数据处理方法，如云式处理、分布式数据库、硬件设备的并行处理等。

大数据在改变人类生活与思考方式的同时，也在推动人类信息管理准则的重新定位。大数据正以不可阻拦的磅礴气势，与当代同样具有革命意义的最新科技（如虚拟现实技术、增强现实技术、人工智能和移动平台应用等）一起，揭开人类新世纪的序幕。

大数据时代已悄然来到我们身边，并渗透到我们每个人的日常生活之中，谁都无法回避。它提供了光怪陆离的全媒体、难以琢磨的云计算、无法抵御的虚拟仿真环境和随处可在的网络服务。随着互联网技术的蓬勃发展，我们一定会迎来大数据的智能时代，即大数据技术和生活紧密相连，它不再只是人们津津乐道的一种时尚，而会成为人们生活中的向导和助手。中国大数据市场的应用如图 1-1 所示。

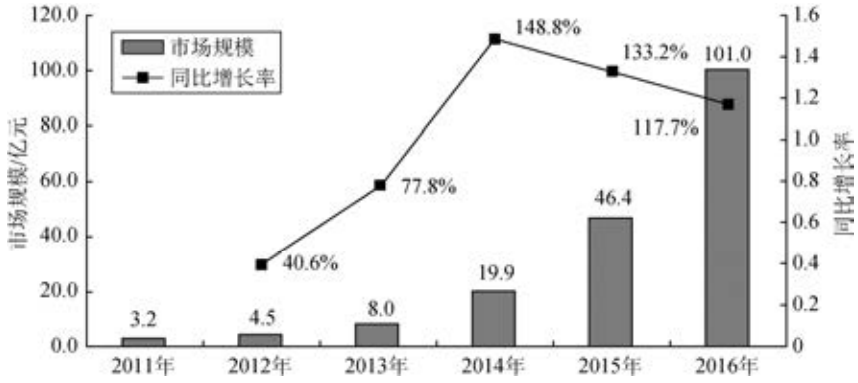

图 1-1 中国大数据市场的应用

1.1.1 大数据的主要来源

大数据的来源非常广泛，如信息管理系统、网络信息系统、物联网系统、科学实验系统等，其数据类型包括结构化数据、半结构化数据和非结构化数据。

1）信息管理系统：企业内部使用的信息系统，包括办公自动化系统、业务管理系统等。信息管理系统主要通过用户输入和系统二次加工的方式产生数据，其产生的大数据大多数为结构化数据，通常存储在数据库中。

2）网络信息系统：基于网络运行的信息系统。网络信息系统是产生大数据的重要方式，如电子商务系统、社交网络、社会媒体、搜索引擎等是常见的网络信息系统。网络信息系统产生的大数据多为半结构化数据或非结构化数据。

3）物联网系统：物联网是新一代信息技术，其核心和基础仍然是互联网，是在互联网基础上延伸和扩展的网络，其用户端延伸和扩展到了物品与物品之间进行信息交换和通信，通过传感技术获取外界的物理、化学、生物等数据信息。

4）科学实验系统：主要用于科学技术研究的补充。可以由真实的实验产生数据，也可以通过模拟方式获取仿真数据。

1.1.2 大数据的核心

大数据的核心就是预测，它使得我们分析信息时需要从不同于以往的角度看待问题。

1. 全新的数据处理理念

1）不只是随机样本，而是全体数据。过去由于受制于技术只能收集与分析随机样本，但是在大数据时代，收集与分析全体数据已成为可能。在大数据时代，我们可以分析更多的数据，甚至可以处理和某个特别现象相关的所有数据，而不再依赖于随机采样，即样本就是总体。

2）不再追求精确性，而是混杂性。大数据时代追求大量数据，而非精确数据，但由于传统处理的信息量较少，因此传统处理方式对数据精确性要求很严格。随着数据量的增加，数据错误率也可能增加，格式也不再单一，只有 5%的数据是结构化数据且适

用于传统统计方法，95%的数据是非结构化数据。因此，只有接受不精确性才能利用大量的数据。由此我们可以断言，大数据时代利用数据快速找出事物的规律更重要。

3）重视的不再是因果关系，而是相关关系，即大数据时代不再热衷于寻找因果关系。

2. 预测

大数据的核心是建立在相关关系分析基础上的预测。相关关系是 A 与 B 经常一起发生，即只要注意到 B 发生，就能预测 A 的发生。

3. 数据价值的获取方式

数据的价值来源于万物数据化和数据交叉复用，大数据的重要价值在于数据深挖掘。

1）数据化。一切事物都可量化，变为数据。数据化不是数字化，数字化即模拟数据转换成用“0”和“1”表示的二进制码，如书页的扫描，无法检索内容；而数据化就是把一种现象转换为可制表分析的量化形式的过程，如书变成数据化文本，可检索。数据化的重点是由 T（Technology，技术）转变为 I（Information，信息）。

2）更有价值。数据价值不会随使用次数的增多而减少，可以重复挖掘。其潜在价值可通过下述 6 种方式释放：数据再利用、重组数据、可扩展数据、数据的折旧值、数据废气、开放数据。

3）角色定位。大数据早期价值来自思维和技术，大数据中后期价值必须从数据本身中挖掘。大数据价值链中主要存在 3 种公司：基于数据本身的公司、基于技能的公司和基于思维的公司。

4. 大数据的安全问题

大数据时代，危险不再是隐私的泄露，而是被预知的可能性，因此需要新的规章制度应对大数据时代的各种隐忧。应用得当，大数据是合理决策的有力武器；应用不当，大数据会变成损害民众利益的工具。大数据时代，告知与许可、模糊化和匿名化三大隐私保护策略都失效。挣脱大数据的困境，是大数据时代人类共同的战争。

1.1.3　大数据的处理流程

大数据的处理流程可以定义为在适合工具的辅助下，对不同结构的数据源进行汲取和集成，并将结果按照一定的标准统一存储，再利用合适的数据分析技术对其进行分析，最后从中提取有益的知识并利用恰当的方式将结果展示给终端前的用户。大数据的处理流程如图 1-2 所示。

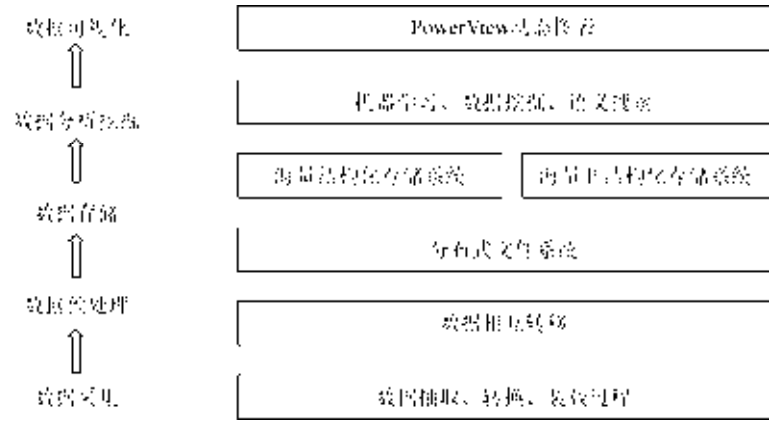

图 1-2　大数据的处理流程

例如，分布式并行处理运算如图 1-3 所示。

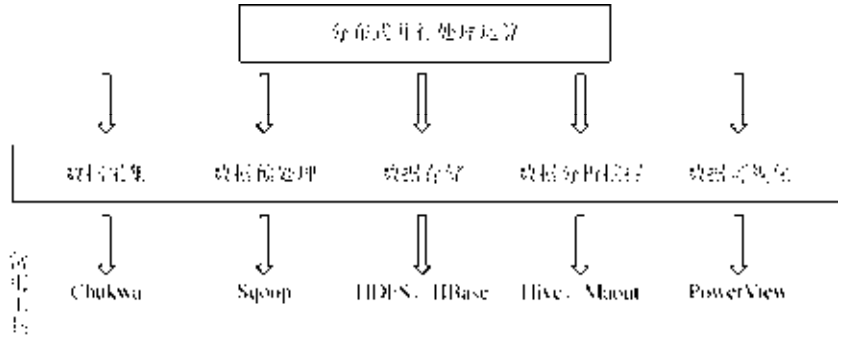

图 1-3　分布式并行处理运算

1. 数据采集

由于大数据处理的数据来源类型广泛，因此其第一步是对数据进行抽取和集成，从中找出关系和实体，经过关联、聚合等操作，再按照统一的格式存储数据。现有的数据抽取和集成引擎有 3 种：基于物化或 ETL（Extraction Transformation Loading，数据抽取、转换和装载）方法的引擎、基于中间件的引擎、基于数据流方法的引擎。

2. 数据预处理

数据预处理的目的是提高数据质量，以便进行数据分析。数据预处理有多种方法，如数据清理、数据集成、数据变换和数据归约等。在数据分析之前使用这些数据处理技术，大大提高了数据分析结果的质量，减少了数据分析所需的时间。

3. 数据存储

大数据存储与管理指用存储器把采集到的数据存储起来，建立相应的数据库，并进

行管理和调用。大数据存储与管理重点解决复杂结构化、半结构化和非结构化大数据管理与处理技术，主要解决大数据的可存储、可表示、可处理、可靠性及有效传输等几个关键问题。开发可靠的分布式文件系统（Distributed File System，DFS）、能效优化的存储、计算融入存储、大数据的去冗余及高效率低成本的大数据存储技术；突破分布式非关系型大数据管理与处理技术、异构数据的数据融合技术、数据组织技术；研究大数据建模技术；突破大数据索引技术；突破大数据移动、备份、复制等技术；开发大数据可视化技术。开发新型数据库技术，主要指的是 NoSQL（Not Only SQL）数据库，分为键值数据库、面向列存储数据库、面向图形存储数据库及面向文档存储数据库等类型。开发大数据安全技术，是指改进数据销毁、透明加解密、分布式访问控制、数据审计等技术，突破隐私保护和推理控制、数据真伪识别和取证、数据持有完整性验证等技术。

4. *数据分析挖掘*

数据分析技术包括改进已有数据挖掘和机器学习技术，开发数据网络挖掘、特异群组挖掘、图挖掘等新型数据挖掘技术，突破基于对象的数据连接、相似性连接等大数据融合技术，突破用户兴趣分析、网络行为分析、情感语义分析等面向领域的大数据挖掘技术。数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘涉及的技术方法很多，有多种分类方法。根据挖掘任务，数据挖掘技术可分为分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等。

从挖掘任务和挖掘方法的角度来看，数据挖掘技术包括以下 5 个方面：

1）可视化分析。数据可视化无论对于普通用户还是数据分析专家，都是最基本的功能。数据图像化可以让数据“自己说话”，让用户直观地感受到结果。

2）数据挖掘算法。数据图像化是将机器语言翻译给人看，而数据挖掘就是机器的母语。我们可以通过分割、集群、孤立点分析等各种算法精炼数据，挖掘价值。这些算法一定要能够应付大数据的量，同时应具有很高的处理速度。

3）预测性分析。预测性分析可以让分析师根据图像化分析和数据挖掘的结果作出一些前瞻性判断。

4）语义引擎。语义引擎需要达到较高的人工智能水平，以满足从数据中主动地提取信息的要求。语言处理技术包括机器翻译、情感分析、舆情分析、智能输入、问答系统等。

5）数据质量和数据管理。数据质量和数据管理是管理的优秀实践，透过标准化流程和机器对数据进行处理，可以确保获得一个预设质量的分析结果。

5. *数据可视化*

数据可视化主要是指借助图形化手段，清晰有效地传达与沟通信息。数据可视化技术的基本思想是将数据库中的每一个数据项作为单个图元元素表示，大量的数据集合构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察

数据，从而对数据进行更深入的观察和分析。使用可视化技术可以将处理结果通过图形方式直观地呈现给用户，如标签云、历史流、空间信息等。

1.1.4 大数据的结构类型

从信息技术（Information Technology，IT）角度来看，大数据的结构类型大致经历了 3 个阶段，即结构化信息阶段、半结构化信息阶段和非结构化信息阶段。必须注意的是，旧的阶段仍在不断发展，如关系数据库的使用。因此 3 种数据结构类型一直存在，只是在不同阶段，其中一种结构类型主导其他结构类型。

1）结构化信息：可以在关系数据库中找到，多年来一直主导着 IT 应用，是关键任务 OLTP（On-Line Transaction Processing，联机事务处理）系统业务所依赖的信息。另外，结构化信息还可对结构数据库信息进行排序和查询。

2）半结构化信息：包括电子邮件、文字处理文件及大量保存和发布在网络上的信息。半结构化信息以内容为基础，可以用于搜索，这也是 Google（谷歌）等搜索引擎存在的理由。

3）非结构化信息：在本质形式上可认为主要是位映射数据，数据必须处于一种可感知（如可在音频、视频和多媒体文件中被听到或看到）的形式中。许多大数据是非结构化的，其庞大的规模和复杂性需要高级分析工具来创建或利用的一种更易于人们感知和交互的结构。

1.1.5 大数据的基本特征

从多种类型的数据中快速获得有价值的信息的能力，就是大数据技术。

大数据呈现出“4V1O”的特征，具体如下：

1）数据量大（Volume）：大数据的首要特征，包括采集、存储和计算的数据量非常大。大数据的起始计量单位至少是 100TB。通过各种设备产生的海量数据，其数据规模极为庞大，远大于目前互联网上的信息流量，PB 级别将是常态。

2）多样化（Variety）：表示大数据种类和来源多样化，具体表现为网络日志、音频、视频、图片、地理位置信息等多类型的数据。多样化对数据的处理能力提出了更高的要求，其编码方式、数据格式、应用特征等多个方面存在差异性，多信息源并发形成大量的异构数据。

3）数据价值密度低（Value）：表示大数据价值密度相对较低，需要很多的过程才能挖掘出来。随着互联网和物联网的广泛应用，信息感知无处不在，信息量大，但价值密度较低。如何结合业务逻辑并通过强大的机器算法挖掘数据价值，是大数据时代最需要解决的问题。

4）速度快，时效高（Velocity）：随着互联网的发展，数据的增长速度非常快，处理速度也较快，时效性要求也更高。例如，搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法要求实时完成推荐，这些都是大数据区别于传统数据挖掘的显著特征。

5）数据是在线的（On-Line）：表示数据必须随时能调用和计算。这是大数据区别于传统数据的最大特征。大数据不仅大，更重要的是数据是在线的，这是互联网高速发

展的特点和趋势。例如，医疗信息和医患互动平台，患者的数据和医生的数据都是实时在线的，这样的数据才有意义。如果把它们放在磁盘中或者离线，显然这些数据远远不及在线数据的商业价值大。

总之，无所遁形的大数据时代已经到来，并快速渗透到了每个职能领域，借助大数据持续创新发展，使企业成功转型，具有非凡的意义。

1.2　大数据的技术架构

各种各样的大数据应用迫切需要新的工具和技术来存储、管理和实现商业价值。新的工具、流程和方法支撑起了新的技术架构，使企业能够建立、操作和管理这些超大规模的数据集和数据存储环境。

大数据的分析能以新视角挖掘企业传统数据，并带来传统上未曾分析过的数据洞察力。大数据一般采用 4 层堆栈技术架构，如图 1-4 所示。

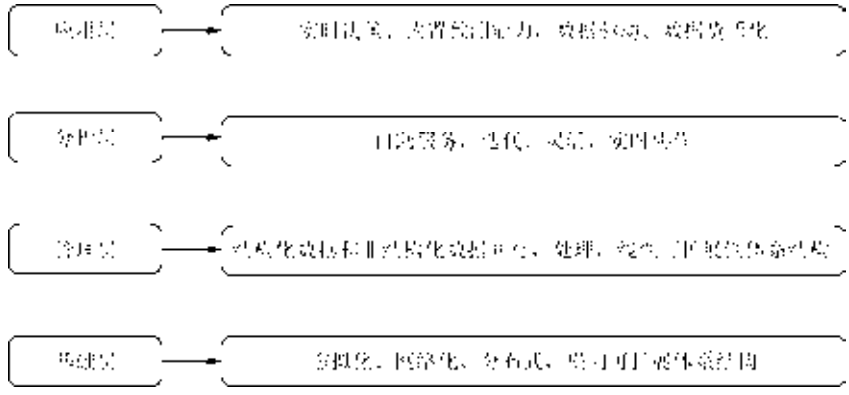

图 1-4　4 层堆栈技术架构

1. 基础层

基础层是整个大数据技术架构基础的最底层。要实现大数据规模的应用，企业需要一个高度自动化的、可横向扩展的存储和计算平台。该存储和计算平台需要从以前的存储孤岛发展为具有共享能力的高容量存储池。容量、性能和吞吐量必须可以线性扩展。

2. 管理层

大数据要支持在多源数据上进行深层次分析，因此在技术架构中需要一个管理平台，即管理层。通过管理层将结构化数据和非结构化数据管理为一体，具备实时传送和查询、计算功能。管理层既包括数据的存储和管理，也涉及数据的计算。并行化和分布式是大数据管理平台必须考虑的要素。

3. 分析层

大数据应用需要大数据分析。分析层提供基于统计学的数据挖掘和机器学习算法，用于分析和解释数据集，帮助企业获得深入的数据价值领悟。可扩展性强、使用灵活的大数据分析平台可成为数据科学家的有力工具，在分析数据时起到事半功倍的效果。

4. 应用层

大数据的价值体现在可以帮助企业进行决策和为终端用户提供服务。不同的新型商业需求驱动了大数据的应用；反之，大数据的应用为企业提供的竞争优势使企业更加重视大数据的价值。新型大数据应用不断对大数据技术提出新的要求，大数据技术也因此在不断的发展变化中日趋成熟。

1.3 大数据分析的 4 种典型工具

大数据分析是在研究大量的数据的过程中寻找模式、相关性和其他有用的信息，以帮助各行政机构与企业更好地适应变化，并作出更明智的决策。

1. Hadoop

Hadoop 是一个能对大量数据进行分布式处理的软件框架，是一个能让用户轻松架构和使用的分布式计算平台。用户可以轻松地在 Hadoop 上开发和运行处理海量数据的应用程序。

Hadoop 带有用 Java 语言编写的框架，因此运行在 Linux 操作系统上是非常理想的。Hadoop 上的应用程序也可以使用其他语言编写，如 C++。

2. Spark

Spark 是一个基于内存计算的开源集群计算系统，目的是更快速地进行数据分析。Spark 由加利福尼亚大学伯克利分校 AMP 实验室以 Matei 为主的小团队使用 Scala 开发，其核心部分的代码只有 63 个 Scala 文件，非常轻量级。 Spark 提供了与 Hadoop 相似的开源集群计算环境，但基于内存和迭代优化的设计，Spark 在某些工作负载方面表现更优秀。

3. Storm

Storm 是一种开源软件，是一个分布式、容错的实时计算系统。Storm 可以非常可靠地处理庞大的数据流，用于处理 Hadoop 的批量数据。Storm 很简单，支持许多种编程语言，使用起来非常有趣。Storm 由 Twitter 开源而来，其他知名的应用企业有 Groupon、淘宝、支付宝、阿里巴巴、乐元素、Admaster 等。

4. Apache Drill

为了帮助企业用户寻找更为有效、可加快 Hadoop 数据查询的方法，Apache 软件基

金会发起了一项名为 Drill 的开源项目。

Apache Drill 开源项目其实是从 Google 公司的 Dremel 项目中获得灵感的，该项目帮助 Google 公司实现了海量数据集的分析处理，包括分析抓取 Web 文档、跟踪安装在 Android Market 上的应用程序数据、分析垃圾电子邮件、分析 Google 分布式构建系统上的测试结果等。

通过开发 Apache Drill 开源项目，组织机构将有望建立 Apache Drill 所属的应用程序编辑接口（Application Program Interface，API）和灵活强大的体系架构，从而帮助支持广泛的数据源、数据格式和查询语言。

1.4　大数据未来的发展趋势

大数据逐渐成为我们生活的一部分，它既是一种资源，又是一种工具，让我们更好地探索世界和认识世界。大数据提供的并不是“最终答案”，只是“参考答案”，它为我们提供的是暂时帮助，以便等待更好的方法和答案出现。

1.4.1　数据资源化

资源化指大数据已成为企业和社会关注的重要战略资源，并成为大家争抢的新焦点，数据将逐渐成为最有价值的资产。

随着大数据应用的发展，大数据资源成为重要的战略资源，数据成为新的战略制高点，大数据也已经演变成不可或缺的资源。华尔街日报在题为《大数据，大影响》的报告中提到，数据就像货币或者黄金一样，已经成为一种新的资源类别。

大数据作为一种新的资源，具有其他资源所不具备的优点，如数据的再利用性、开放性、可扩展性和潜在价值。数据的价值不会随着它的使用次数增多而减少，而是可以不断地被处理和利用。

1.4.2　数据科学和数据共享

1. 催生新的学科和行业

数据科学将成为一种专门的学科，被越来越多的人所认知。越来越多的高校开设了与大数据相关的学科课程，为市场和企业培养人才。

一个新行业的出现，必将会增加工作职位的需求，因此大数据催生了一批与之相关的新的就业岗位，如大数据分析师、大数据算法工程师、数据产品经理、数据管理专家等。因此，具有丰富经验的大数据相关人才将成为稀缺资源。

2. 数据共享

大数据相关技术的发展将会产生一些新的细分市场，针对不同的行业将会出现不同的分析技术。但是对于大数据来说，数据的多少虽然不意味着价值更高，但是数据越多对一个行业的分析价值越有利。

以医疗行业为例，如果每个医院想要获得更多病情特征库和药效信息，就需要对数据进行分析，以便从数据中获得相应的价值。如果想获得更多的价值，就需要对全国甚至全世界的医疗信息进行共享。只有这样，才能对整个医疗平台的数据进行分析，获取更准确、更有利的价值。因此，数据可能发展成为一种共享的趋势。

1.4.3 大数据的隐私和安全问题

1. 大数据引发个人隐私、企业和国家安全问题

大数据时代将引发个人隐私安全问题。大数据时代，用户的个人隐私数据可能在不经意间就被泄露，如网站密码泄露，系统漏洞导致用户资料被盗，手机里的 APP 暴露用户的个人信息等。在大数据领域，一些用户认为根本不重要的信息很有可能暴露用户的近期状况，带来安全隐患。

大数据时代，企业将面临信息安全的挑战。企业不仅要学习如何挖掘数据价值，还要考虑如何应对网络攻击、数据泄露等安全风险，并且建立相关的预案。在企业用数据挖掘和数据分析获取商业价值的同时，黑客也利用这些数据技术向企业发起攻击。因此，企业必须制定相应的策略来应对大数据带来的信息安全挑战。

大数据时代，大数据安全应该上升为国家安全。数据安全的威胁无处不在，国家的基础设施和重要机构所保存的大数据信息，如与石油、天然气管道、水电、交通、军事等相关的数据信息，都有可能成为黑客攻击的目标。

2. 正确合理地利用大数据，促进大数据产业的健康发展

大数据时代，必须对数据安全和隐私进行有效的保护，具体方法如下：

1）从用户角度来看，应积极探索，加大个人隐私保护力度。数据来源于互联网上无数用户产生的数据信息，因此，建议用户在使用互联网或者 APP 时保持高度警惕。

2）从法律角度来看，应提高安全意识，及时出台相关政策，制定相关法规，完善立法。国家需要有专门的法规来为大数据的发展扫除障碍，因此必须健全大数据隐私和安全方面的法律法规。

3）从数据使用者角度来看，数据使用者要以负责的态度使用数据，我们需要把进行隐私保护的责任从个人转移到数据使用者身上。政府和企业的信息化建设必须拥有统一的规划和标准，只有这样才能有效地保护公民和企业隐私。

4）从技术角度来看，应加快数据安全技术研发，尤其应加强云计算安全研究，保障云安全。

1.4.4 开源软件

大数据获得动力的关键在于开放源代码，开放源代码可以帮助分解和分析数据。开源软件的盛行不会抑制商业软件的发展；相反，开源软件将会为基础架构硬件、应用程序开发工具、应用服务等各个方面相关领域带来更多的机会。

从技术的潮流来看，无论是大数据还是云计算，其实推动技术发展的主要力量都来

源于开源软件。使用开源软件有很多优势，这是因为开源的代码有很多人在看、在维护、在检查。了解开源软件和开源模式，将成为一个重要的趋势。

1.4.5　大数据对生活的影响

大数据作为一种重要的战略资产，已经不同程度地渗透到了每个行业领域和部门。现在，用户希望通过大数据掌握真正的便捷信息，从而让生活更有趣。

大数据也将促进智慧城市的发展，是智慧城市的核心引擎。智慧医疗、智慧交通、智慧安防等，都是以大数据为基础的智慧城市的应用领域。大数据将多方位改善我们的生活。

1.5　大数据在医学领域的应用

大数据在社会生活的各个领域得到了广泛应用，如科学计算、金融、社交网络、移动数据、物联网、医疗、网页数据、多媒体、网络日志、RFID（Radio Frequency Identification，射频识别）传感器、社会数据、互联网文本和文件、互联网搜索索引、呼叫详细记录、天文学、大气科学、基因组学、生物和其他复杂或跨学科的科研、军事侦察、医疗记录、摄影档案馆视频档案、大规模的电子商务等。不同领域的大数据应用具有不同的特点，其时间性、稳定性、精确性的要求各不相同，解决方案也层出不穷。本书的重点是大数据在医学领域的应用，包括案例和具体实现流程。

我们正处在一个医学信息爆炸的时代。据统计，医学信息资源占据 30%以上的互联网信息资源，医学文献的数量正以惊人的速度增长。全球医药类期刊近 3 万种，每年发表论文 200 多万篇，并以每年 7%的速度递增。临床医生平均每天必须阅读大量的专业文献，才可能跟上现代医学发展的速度。2012 年，美国政府发布了《大数据研究和发展倡议》，旨在利用大量复杂数据集合获取知识和提升预见能力，投入金额高达 2 亿美元。与此同时，医学科技的发展也离不开大数据。在科研过程中，大数据的利用、开发和整理可能颠覆以往很多研究结果，为我们带来意想不到的效益。

医疗行业很早就遇到了海量数据和非结构化数据的挑战，而近年来很多国家在积极推进医疗信息化发展，这使得很多医疗机构有资金能够进行大数据分析。因此，医疗行业将和银行、电信、保险等行业一起首先迈入大数据时代。

下面我们从医疗服务业的 5 大领域（临床操作、付款/定价、研发、新的商业模式、公众健康）介绍大数据分析和应用是如何提高医疗效率和医疗效果的。

1.5.1　临床操作

在临床操作方面，有 5 个主要场景的大数据应用。麦肯锡估计，如果这些应用被充分采用，仅是美国，国家医疗健康开支一年就将减少 165 亿美元。

1. 比较效果研究

通过全面分析患者的特征数据和疗效数据，以及比较多种干预措施的有效性，可以找到针对特定患者的治疗途径。

2. 临床决策支持系统

临床决策支持系统可以提高工作效率和诊疗质量。大数据分析技术将使临床决策支持系统更智能，这得益于对非结构化数据的分析能力的日益加强。例如，可以使用图像分析和识别技术识别医疗影像［如 X 光、CT（Computed Tomography，电子计算机断层扫描）、MRI（Magnetic Resonance Imaging，磁共振成像）等］数据，或者挖掘医疗文献数据建立医疗专家数据库（如 IBM Watson 等），从而为医生提供诊疗建议。此外，临床决策支持系统还可以使医疗流程中的大部分工作流向护理人员和助理医生，使医生从耗时过长的简单咨询工作中解脱出来，从而提高治疗效率。

3. 医疗数据透明度

提高医疗过程数据的透明度，可以使医疗从业者、医疗机构的绩效更透明，间接促进医疗服务质量的提高。例如，根据医疗服务提供方设置的操作和绩效数据集，可以进行数据分析并创建可视化的流程图和仪表盘，促进信息透明。

4. 远程患者监控

从对慢性患者的远程监控系统收集数据，并将分析结果反馈给监控设备（查看患者是否正在遵从医嘱），从而确定以后的用药和治疗方案。

5. 对患者档案的先进分析

在患者档案方面，应用高级分析可以确定哪些人是某类疾病的易感人群。

1.5.2 付款/定价

对医疗支付方来说，通过大数据分析可以更好地对医疗服务进行定价。以美国为例，这将有潜力创造每年 500 亿美元的价值，其中一半来源于国家医疗开支的降低。

1. 自动化系统

自动化系统（如机器学习技术）可以检测欺诈行为。

2. 基于卫生经济学和疗效研究的定价计划

在药品定价方面，医药公司可以参与分担治疗风险，如基于治疗效果制定定价策略。这对医疗支付方的好处显而易见，有利于控制医疗保健成本支出。对患者来说，好处更加直接，如他们能够以合理的价格获得创新的药物，并且这些药物经过了基于疗效的研究。而对医药公司来说，它们可以获得更高的市场准入可能性，也可以通过创新的定价

方案，推出更有针对性的疗效药品，获得更高的收入。

1.5.3　研发

医疗产品公司可以利用大数据提高研发效率。以美国医疗产品公司为例，其可以创造每年超过 1000 亿美元的价值。

1. 预测建模

医药公司在新药物的研发阶段，可以通过数据建模和分析，确定最有效率的投入产出比，从而配备最佳资源组合。模型基于药物临床试验阶段之前的数据集及早期临床阶段的数据集，尽可能及时地预测临床结果。评价因素包括产品的安全性、有效性、潜在的副作用和整体的试验结果。通过预测建模可以降低医药公司的研发成本，在通过数据建模和分析预测药物临床结果后，可以暂缓研究次优的药物，或者停止在次优药物上的昂贵的临床试验。

2. 分析临床试验数据

分析临床试验数据和患者记录可以确定药品更多的适应症并发现副作用。在对临床试验数据和患者记录进行分析后，可以对药物重新定位，或者实现针对其他适应症的营销。实时或者近乎实时地收集不良反应报告可以促进药物警戒（药物警戒是上市药品的安全保障体系，对药物不良反应进行监测、评价和预防）。或者在一些情况下，临床试验暗示出了一些情况但没有足够的统计数据去证明，现在基于临床试验大数据的分析可以给出证据。

这些分析项目是非常重要的，因为药品撤市可能给医药公司带来毁灭性的打击。例如，2004 年从市场上撤下的止痛药 Vioxx 给默克公司造成 70 亿美元的损失，短短几天内股东价值损失 33%。

3. 个性化治疗

另一种在研发领域有前途的大数据创新，是通过对大型数据集（如基因组数据）的分析发展个性化治疗。这一应用考察遗传变异对特定疾病的易感性和对特殊药物的反应的关系，然后在药物研发和用药过程中考虑个人的遗传变异因素。

个性化治疗可以改善医疗保健效果，如在患者发生疾病症状前就提供早期的检测和诊断。很多情况下，患者用同样的诊疗方案但是疗效却不一样，部分原因就是遗传变异。针对不同的患者采取不同的诊疗方案，或者根据患者的实际情况调整药物剂量，可以减少副作用。

4. 分析疾病模式

通过分析疾病模式和趋势，可以帮助医疗产品企业制定战略性的研发投资决策，帮助其优化研发重点，优化配备资源。

1.5.4 新的商业模式

大数据分析可以为医疗服务行业带来新的商业模式。

1. 汇总患者的临床记录和医疗保险数据集

汇总患者的临床记录和医疗保险数据集并进行高级分析，将提高医疗支付方、医疗服务提供方和医药公司的决策能力。例如，对医药公司来说，他们不仅可以生产出具有更佳疗效的药品，而且能保证药品适销对路。临床记录和医疗保险数据集的市场刚刚开始发展，扩张的速度将取决于医疗保健行业完成电子病历（Electronic Medical Record，EMR）和循证医学发展的速度。

2. 网络平台和社区

另一个潜在的大数据启动的商业模型是网络平台和社区，这些平台已经产生了大量有价值的数据。例如，patientslikeme.com 网站，患者可以在该网站上分享治疗经验；sermo.com 网站，医生可以在该网站上分享医疗见解；participatorymedicine.org 网站，该非营利性组织运营的网站鼓励患者积极进行治疗。这些平台可以成为宝贵的数据来源，如 sermo.com 网站向医药公司收费，允许它们访问会员信息和网上互动信息。

1.5.5 公众健康

大数据的使用可以改善公众健康监控。公共卫生部门可以通过覆盖全国的患者电子病历数据库，快速检测传染病，进行全面的疫情监测，并通过集成疾病监测和响应程序，快速进行响应。这将带来很多好处，如减少医疗索赔支出、降低传染病感染率，卫生部门可以更快地检测出新的传染病和疫情。通过提供准确和及时的公众健康咨询，将会大幅提高公众健康风险意识，同时也将降低传染病感染风险。所有的这些都将帮助人们创造更美好的生活。

本 章 小 结

近年来大数据应用取得了令人瞩目的成绩，作为新的重要资源，世界各国都在加快大数据的战略布局，制定战略规划。

目前我国大数据产业还处于发展初期，市场规模仍然比较小，2012 年仅为 4.5 亿元，而且主导厂商仍以外国企业居多。据估计，我国到 2020 年，技术先进、应用繁荣、保障有力的大数据产业体系将基本形成，大数据相关产品和服务业务收入将突破 1 万亿元。

关键词注释

1．联机事务处理（On-Line Transaction Processing，OLTP）：也称面向交易的处理，其基本特征是顾客的原始数据可以立即传送到计算中心进行处理，并在很短的时间内给出处理结果。

2．ETL（Extraction Transformation Loading）：数据抽取（Extract）、转换（Transform）和装载（Load）的过程，即将业务系统的数据经过抽取、转换之后加载到数据仓库的过程，是构建数据仓库的重要环节，其目的是将企业中分散、零乱、标准不统一的数据整合到一起，为企业决策提供分析依据。

习　题　1

一、填空题

1．大数据的首要特征是数据量大，起始计量单位至少是_____，_____级别将是常态。

2．大数据处理的基本流程可概括为 4 个阶段，即_____、_____、_____、_____。

3．大数据呈现出的“4V1O”特征是_____、_____、_____、_____、_____。

4．大数据的 4 层堆栈技术架构是_____、_____、_____、_____。

5．大数据处理分析的 4 种典型工具是_____、_____、_____、_____。

二、简答题

1．简述大数据的特点。

2．简述大数据未来的发展趋势。

3．简述大数据在医学应用中的某一种场景。

第 2 章
医学大数据采集

导学

本章主要介绍大数据采集、医学大数据采集的基本概念、数据来源与方法。通过对本章的学习，读者可以对医学大数据采集有一个概括性的了解和掌握。

了解：医学大数据采集的方法，Chukwa 数据采集的基本过程，网络爬虫的相关概念、工作流程和抓取策略等。

掌握：大数据采集的基本概念、大数据采集与传统数据采集的区别、医学大数据采集的数据来源。

在大数据环境下，医学数据的来源、种类繁多，数据结构也非常复杂。大数据时代，对数据存储和处理的需求量大，数据表达的要求高，因此数据处理的高效性与可用性非常重要。因此，必须在数据的源头，即数据采集上把好关，数据源的选择和原始数据的采集方法是大数据采集的关键。本章将着重介绍医学大数据的采集。

2.1 大数据采集概述

2.1.1 大数据的采集

大数据出现之前，计算机所能够处理的数据都需要前期进行相应的结构化处理，并存储在相应的数据库中。但大数据技术对数据结构的要求大大降低，人们在互联网上留下的社交信息、地理位置信息、行为习惯信息、偏好信息等各种维度的信息都被进行实时处理。

1. 大数据的数据采集

大数据的数据采集是在确定用户目标的基础上，针对该范围内所有结构化数据、半结构化数据和非结构化数据使用某种技术或手段，将数据收集起来并存储在某种设备上。采集后对这些数据进行处理，从中分析出有价值的信息。与传统的数据采集相比，大数据的数据采集在数据收集和存储技术上是不同的，如表 2-1 所示。

表 2-1　传统的数据采集与大数据的数据采集对比

项目	传统的数据采集	大数据的数据采集
数据来源	来源单一，数据量相对大数据较小	来源广泛，数据量巨大
数据类型	结构单一	数据类型丰富，包括结构化数据、半结构化数据和非结构化数据
存储技术	关系型数据库和并行数据库	分布式数据库

（1）大数据的收集过程

在收集阶段，大数据采集的数据与传统的数据采集相比，在时间和空间两个方面都有显著的不同。在时间维度上，为了获取更多的数据，大数据收集的时间频度更大；在空间维度上，为了获取更准确的数据，大数据采集点设置得会更密集。

以收集一个面积为 $100m^2$ 细菌培养室的平均温度为例。传统数据采集时，由于成本所限，研究员只能在细菌培养室的中央设置一个温度计来计算温度，而且每隔 1h 观测一次，这样一天只有 24 个数据。而在大数据采集时，在空间维度上可以设置 100 个温度计，即每隔 $1m^2$ 设置一个温度计；在时间维度上，每隔 1min 观测一次，这样一天就有 144000 个数据，是原来的 6000 倍。有了大量的数据，我们就可以更准确地知道细菌培养室的平均温度，如果再加上时间维度，还可以得出一个时间序列曲线。

（2）大数据的存储技术

通过增加数据采集的深度和广度，数据量会越来越大，数据存储问题就显现出来了。原来 1TB 的数据使用一块硬盘就可以实现数据存储；而现在变成了 6000TB，即需要 6000 块硬盘来存储数据，而且该数据每天都会增加。此时，计算机技术中的分布式存储开始发挥优势，它可以将 6000 台甚至更多的计算机组合在一起，让它们的硬盘组合成一块巨大的硬盘。

2. 医学大数据采集

生命科学领域所涉及的大数据与经济、社交媒体、环境科学等领域的大数据存在明显不同。医学大数据泛指所有与医疗和生命健康相关的大数据，其与人的健康、疾病和生命息息相关，而且具有更复杂的多样性，以及更多需要研究探讨的未知事件。因此，医学大数据在医学临床研究和医疗健康等领域具有重要的意义，医学大数据的采集更具有价值。

医学大数据采集的结构化数据包括电子病历、临床数据、患者数据、健康管理和医保报销等；非结构化数据包括影像记录、CT 图片和诊断视频数据等。

2.1.2　医学大数据的数据来源

根据数据来源，医学大数据分为医疗大数据、服务平台医疗健康大数据、医学研究或疾病监测大数据、自我量化大数据、网络大数据和生物大数据六大类。这些不同种类的数据具有不同的性质、医学价值及问题。

1. 医疗大数据

医疗大数据来源于医院常规临床诊治、科研和管理过程，包括各种门/急诊记录、住院记录、影像记录、实验室记录、用药记录、手术记录、随访记录和医疗保险数据等，具有数据量庞大、产生速度快、数据结构复杂和价值密度低等典型大数据的特征。医疗大数据大多数是用医学专业方式记录下来的，是最原始的临床记录。从临床管理或研究角度看，这些数据是关于患者就医过程的真实记录，或者也可以说是临床医疗行为留存的痕迹，每一个数据都是有价值的，包括记录不完善或错误的数据，都可能隐藏了有待发掘和利用的重要医学信息。

2. 服务平台医疗健康大数据

依托于服务平台的大数据是未来医疗健康大数据的发展方向，一方面，服务平台汇集整合了区域内很多家医院和相关医疗机构的医疗健康数据，致使数据量大幅度增加；另一方面，服务平台数据的收集事先都经过了充分论证和规划，比原始的医院数据更为规范。

3. 医学研究或疾病监测大数据

除了上述原生态医学大数据以外，还有一些医学大数据来自专门设计的基于大量人群的医学研究或疾病监测。

例如，中国卫生部开展的脑卒中筛查与防治项目，计划在全国各地筛检 100 万脑卒中高危人群，随后对筛检出的高危人群的疾病及其治疗效果进行长期追踪；中国环境与遗传因素及其交互作用对冠心病和缺血性脑卒中影响的超大型队列研究，评估 50 余万自然人群遗传和环境危险因素及其复杂的交互作用等。

因为这些医学研究或疾病监测都经过了仔细的专业设计，所以数据内容较多，数据质量也较高，能够产生较为理想的研究结果。这些专项大数据与医疗过程数据相互融合后，可在疾病治疗和预防中发挥更大的作用。

4. 自我量化大数据

基于移动物联网的个人身体体征和活动的自我量化大数据是一种新型的医学大数据。自我量化大数据包含血压、心跳、血糖、呼吸、睡眠、体育锻炼等信息，除了有利于帮助了解自身健康状况以外，经过一定时期的累积，其在医学上会变得很有价值，不仅有助于识别疾病病因或防控疾病，而且有助于个性化临床诊疗，塑造全新的医疗或健康管理模式。

5. 网络大数据

网络大数据指互联网上与医学相关的各种数据。这类网络大数据经常与其他各类医学大数据混为一谈，造成了对大数据效用的误解。网络大数据产生于社交互联网关于疾

病、健康或寻医的话题，互联网上的购药行为，健康网站的访问行为等。网络大数据杂乱无章，同一主题的数据既可来自同一网站众多不同的网络用户，也可来自大量不同的网站，而且其中包含着大量的音频、视频、图片、文本等异构性数据。与自我量化大数据相比，网络大数据是被动性存在的，随机性很大，数据中蕴含的信息缺乏稳定性。

6. 生物大数据

生物大数据具有很强的生物专业性，主要是关于生物标本和基因测序信息的数据，其中组学大数据是重要内容。与过去的分子生物学研究相比，组学研究使基础研究由碎片连接为整体，数据容量大、动态性强、复杂性高、异质性明显。生物大数据直接关系到临床的个性化诊疗及精准医疗。生物信息数量巨大，据估计，人类基因测序一次，产生的数据量可高达 100～600GB。生物大数据目前面临的最大难题是，如何能使标本及数据标准化、测定结果实用化，以及测定结果与患者临床数据无缝连接等。

2.2　医学大数据采集的实现

2.2.1　医学大数据采集的方法

由于医学大数据的复杂性、敏感性和不易共享等特点，医学大数据的采集仍然面临比较困难的局面，能够采集到的数据远远小于理论上可以采集的数据。因此，解决医学大数据的隐私性问题、数据孤岛和数据标准化问题等是实现数据采集的重要目标。现阶段的医疗机构数据更多来源于内部，外部数据并没有得到很好的应用。

1. 医疗机构内部数据

对于医疗机构内部数据，可使用 ETL 技术对医院信息管理系统（Hospital Information System，HIS）、影像归档与通信系统（Picture Archiving and Communication System，PACS）、实验室信息系统（Laboratory Information System，LIS）等的数据以全量采集或增量采集的方式进行采集。目前主流的 ETL 产品有 Datastage、Powercenter、Automation 和 Kettle 等。

例如，某医院使用 ETL 技术采集医院信息系统数据，系统对接方式如图 2-1 所示。

方式 1：ETL 可以直接访问医院信息系统，直接读取需要抽取的数据。

方式 2：在医院信息系统和 ETL 之间建立一个用于数据交互的中间库，医院信息系统将数据按照原始格式写入中间库，ETL 通过读取中间库的数据来抽取数据。

方式 3：将医院信息系统中的数据按照双方约定的方式导出数据包文件，再将该文件导入中间库，ETL 通过读取中间库的数据来抽取数据。

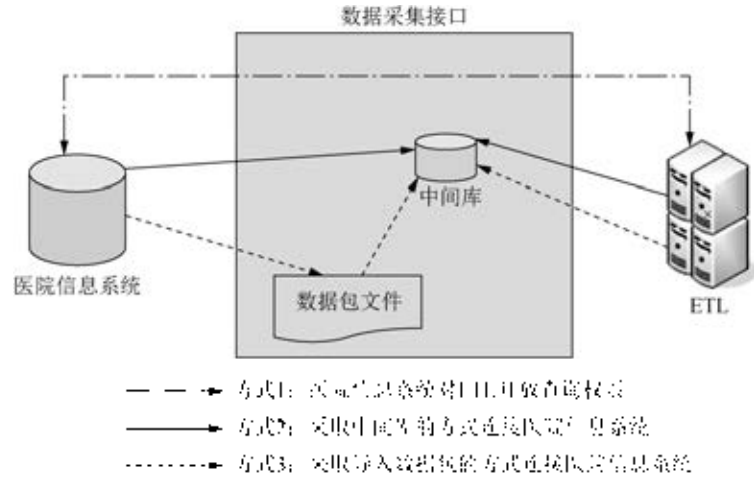

图 2-1　系统对接方式

2. 医疗机构外部数据

对于外部数据，医疗机构可以借助如百度公司、阿里巴巴公司、腾讯公司等第三方数据平台解决数据采集难题，很多互联网企业有自己的海量数据采集工具，多用于系统日志采集，如 Hadoop 的 Chukwa 等。

Chukwa 构建在 Hadoop 的 HDFS 和 MapReduce 框架之上，包含一个强大和灵活的工具集，可以将各种类型的数据收集成适合 Hadoop 处理的文件，保存在 HDFS 中供 Hadoop 进行 MapReduce 操作，可用于展示、监控和分析已收集的数据。

Chukwa 基本架构如图 2-2 所示。

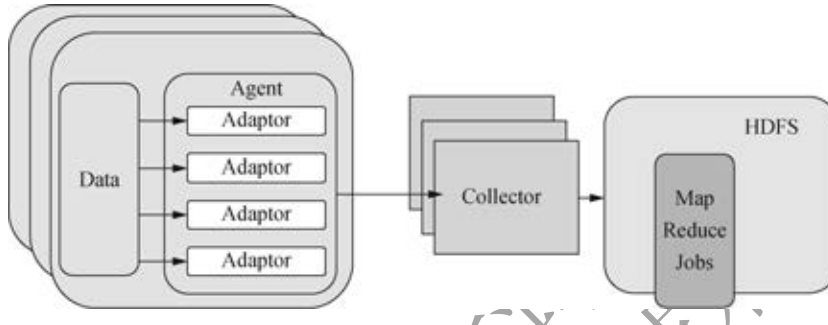

图 2-2　Chukwa 基本架构

Chukwa 基本架构中的主要部件功能如下：

1）Agent：负责采集最原始的数据，并发送给 Collector。

2）Adaptor：直接采集数据的接口和工具，一个 Agent 可以管理多个 Adaptor 的数据采集。

3）Collector：负责收集 Agent 收送来的数据，并定时写入 HDFS 中。

4）MapReduce Jobs：定时启动，负责把 HDFS 中的数据分类、排序、去重和合并。

3. 保密性要求较高的数据

政府部门数据和学科研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。

4. 互联网医学大数据

互联网医学大数据主要通过网络爬虫采集数据。

1）网络爬虫是一种按照一定的规则自动采集与整理互联网信息的程序或脚本。网络爬虫可用 Java、Python、PHP 和 C++等计算机语言实现。

简单的网络爬虫架构如图 2-3 所示。

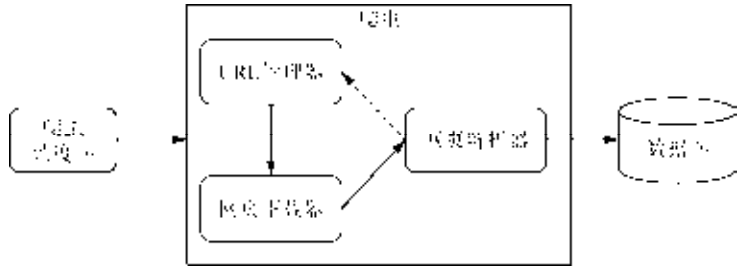

图 2-3　简单的网络爬虫架构

网络爬虫架构中的主要部件功能如下：

① 爬虫调度端：启动、执行、停止爬虫，或者监视爬虫中的运行情况。

② URL（Uniform Resource Locator，统一资源定位符）管理器：对将要爬取的 URL 和已经爬取过的 URL 这两个数据进行管理。

③ 网页下载器：下载 URL 管理器里提供的 URL 对应的网页并将其存储为字符串，将字符串传送给网页解析器进行解析。

④ 网页解析器：解析出有价值的数据，当页面有很多指向其他页面的网页时，将这些 URL 解析出来并补充进 URL 管理器（图 2-3 中虚线箭头）。

⑤ 数据端：存储解析出来的数据。

2）通用的网络爬虫的基本工作流程如图 2-4 所示。

① 获取初始 URL。

② 将 URL 放入待抓取 URL 队列。

③ 从待抓取 URL 队列中读取 URL，解析 DNS，得到主机 IP，并将 URL 对应的网页下载下来，存储到已下载的网页库中。此外，将这些 URL 放进已抓取 URL 队列。

④ 分析已抓取 URL 队列中的 URL，当页面有很多指向其他页面的网页时，分析其中的 URL，并且将 URL 放入待抓取 URL 队列，然后进入下一个循环。

3）抓取策略即决定待抓取 URL 排列顺序的方法，在爬虫系统中，待抓取 URL 队列是很重要的一部分。待抓取 URL 队列中的 URL 以什么样的顺序排列也是一个很重要

的问题，因为这涉及先抓取哪个页面，后抓取哪个页面。

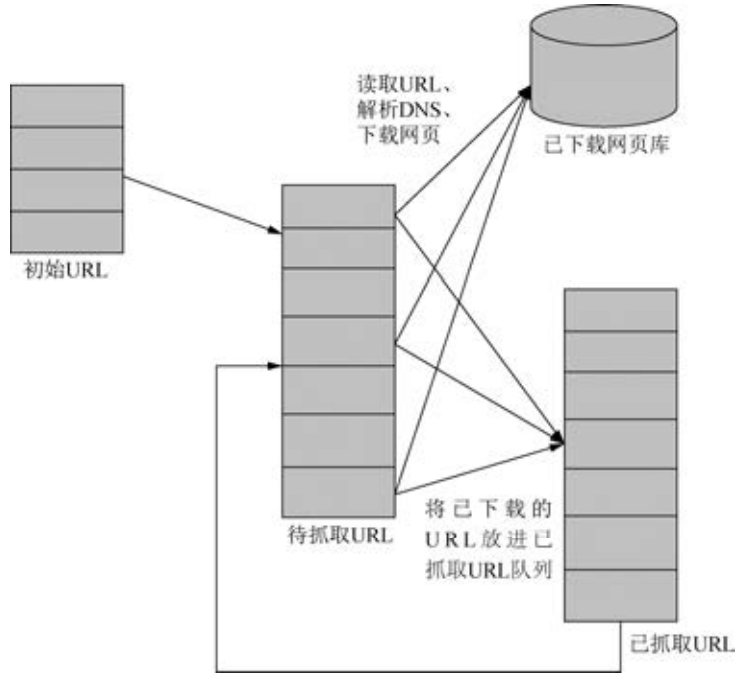

图 2-4　通用的网络爬虫的基本工作流程

抓取策略主要有深度优先遍历策略和宽度优先遍历策略等。如图 2-5 所示，假设有一个网站，A、B、C、D、E、F、G、H、I 分别为站点下的网页，图中箭头表示网页的层次结构。

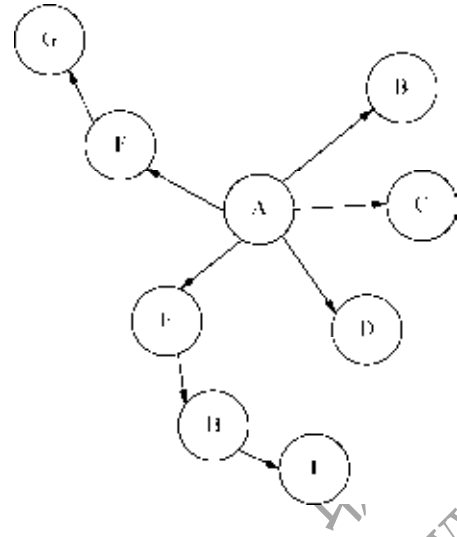

图 2-5　某假设网站的层次结构

① 深度优先遍历策略：网络爬虫会从起始网页开始，一个链接接着下一个链接地跟踪下去，处理完这条线路之后再转入下一个起始网页，继续跟踪链接。

遍历图 2-5 的路径：A—F—G、E—H—I、B、C、D。

② 宽度优先遍历策略：将新下载网页中发现的链接直接插入待抓取 URL 队列的末尾，即网络爬虫会先抓取起始网页中链接的所有网页，然后选择其中的一个链接网页，继续抓取在此网页中链接的所有网页。

遍历图 2-5 的路径：A—B—C—D—E—F、G、H—I。

一般来说，抓取系统需要面对的是整个互联网上数以亿计的网页，单个抓取程序不可能完成这样的任务，往往需要多个抓取程序一起处理。抓取系统往往是一个分布式的 3 层结构，如图 2-6 所示。

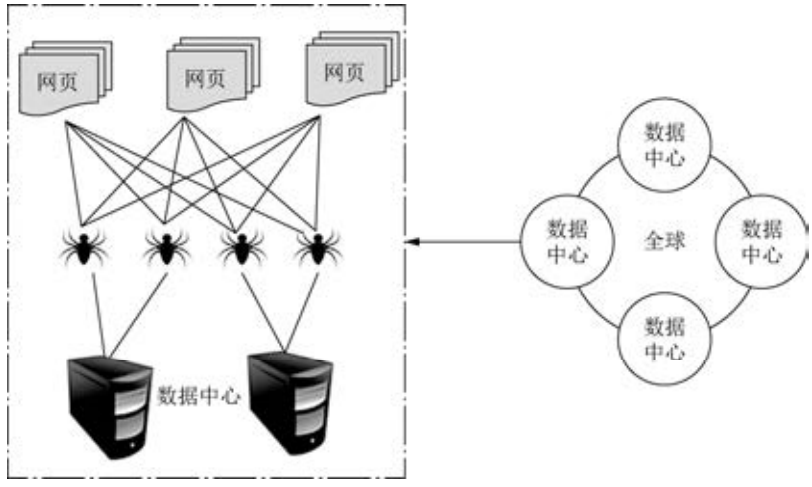

图 2-6　分布式抓取系统结构

分布式抓取系统中，最下一层是分布在不同地理位置的数据中心，每个数据中心中有若干台抓取服务器，而每台抓取服务器上可能部署了若干套爬虫程序，这就构成了一个基本的分布式抓取系统。

2.2.2　网络爬虫采集的实现

下面以网络矿工数据采集软件为例，采用网络爬虫的方式采集百度学术网站的医学数据。网络矿工数据采集软件是一款集互联网数据采集、清洗、存储、发布为一体的工具软件，具有高效的采集性能，从网络获取数据，从中提取需要的内容，存储最终的数据。网络矿工官方网页（http://www.minerspider.com）如图 2-7 所示。

该软件操作步骤如下：

1）进入网络矿工官方网站，下载免费版（通常免费版有试用期限，一般为 30 天），本例下载的是网络矿工数据采集软件免费版 V5.33。运行网络矿工数据采集软件需要.Net Framework 2.0 环境，建议使用 Firefox 浏览器。

图 2-7　网络矿工官方网页

2）下载的压缩文件内包含多个可执行程序，其中 SoukeyNetget.exe 为网络矿工数据采集软件，运行此文件即可打开网络矿工数据采集软件，操作界面如图 2-8 所示。

图 2-8　网络矿工数据采集软件操作界面

3）单击“新建采集任务分类”超链接，在弹出的“新建任务类别”对话框中输入类别名称，并设置存储路径，如图 2-9 所示，单击“确定”按钮。

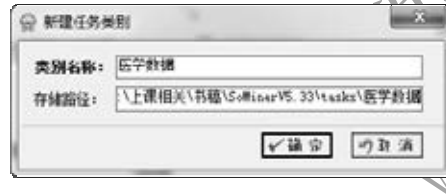

图 2-9　“新建任务类别”对话框

4）在树形菜单中右击“医学数据”，在弹出的快捷菜单中选择“新建采集任务”命令，如图 2-10 所示。弹出“新建采集任务”对话框，如图 2-11 所示，输入任务名称。

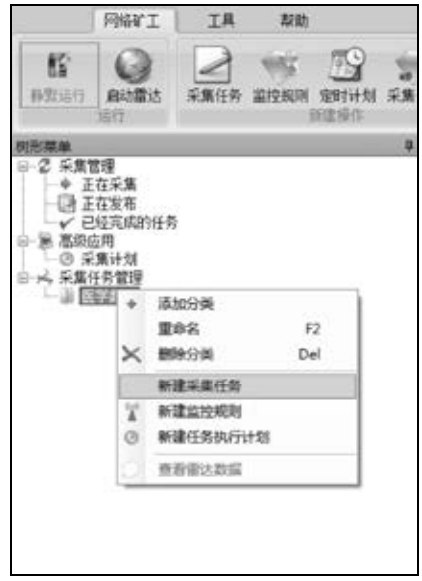

图 2-10　选择“新建采集任务”命令

图 2-11　“新建采集任务”对话框

5）在“新建采集任务”对话框中单击“增加采集网址”按钮，弹出“增加采集网址”对话框，如图 2-12 所示，输入采集网址，如 http://xueshu.baidu.com/s?wd=%E5%BF%83%E8%A1%80%E7%AE%A1%E7%96%BE%E7%97%85&rsv_bp=0&tn=SE_baiduxueshu_c1gjeupa&rsv_spt=3&ie=utf-8&f=3&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsp=0（此网址是在百度学术中搜索“心血管”后得出的）。选中“导航采集（多级采集）”复选框，单击“增加”按钮，增加导航规则。

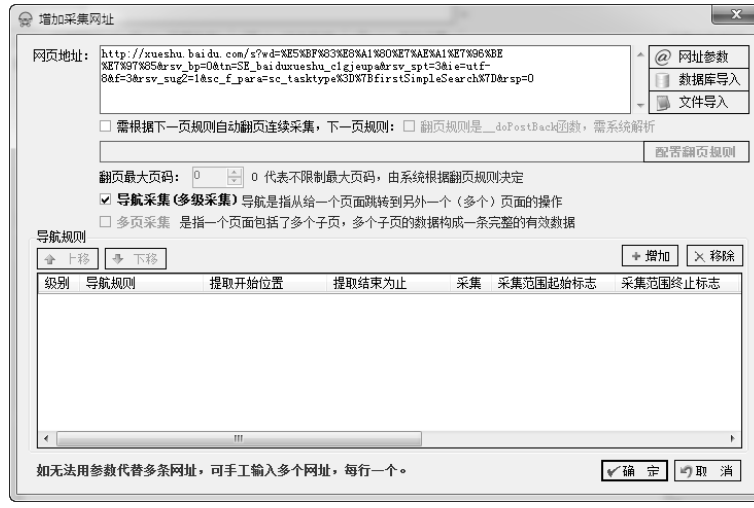

图 2-12 “增加采集网址”对话框

6）弹出“导航页规则配置”对话框，如图 2-13 所示，导航规则可设置为“前后标记配置”“可视化配置”“自我导航”等。

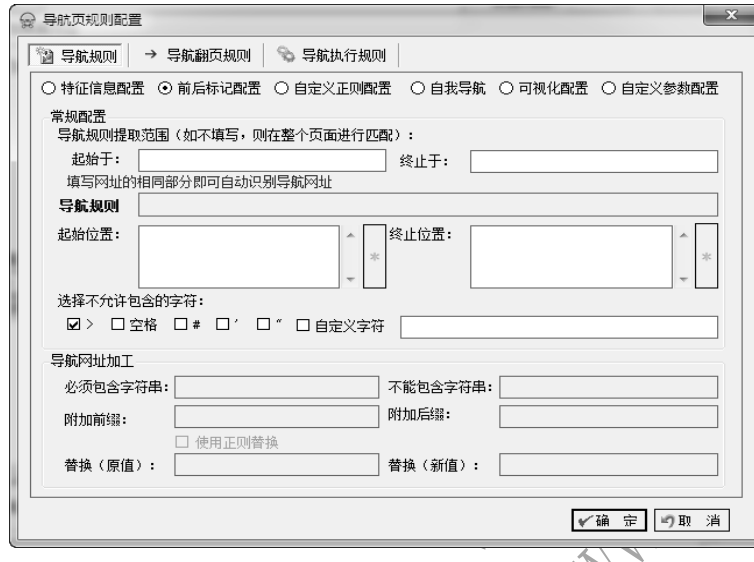

图 2-13 “导航页规则配置”对话框

7）选中“可视化配置”单选按钮，打开“可视化配置导航规则”工作页面，如图 2-14 所示。

图 2-14　“可视化配置导航规则”工作页面

8）单击“可视化提取”按钮，打开“可视化采集配置器”窗口，单击“转到”按钮，出现百度学术搜索结果。单击“开始捕获”按钮，鼠标指针在页面滑动时，会出现一个蓝色边框，用蓝色边框选第一篇要采集的心血管相关文章，单击，再选最后一篇文章，单击，系统会自动捕获导航规则，如图 2-15 所示。

图 2-15　“可视化采集配置器”窗口

9）单击“确定退出”按钮，配置完成。选中刚才配置的网址，单击“测试网址解析”按钮，如图 2-16 所示，可以看到系统已经将需要采集的新闻网址解析出来，这表示配置成功，如图 2-17 所示。

图 2-16　单击“测试网址解析”按钮

图 2-17　网址解析配置成功

10）配置采集数据的规则。应采集文章的正文、标题和发布时间，可以用 3 种方式完成：智能采集、可视化采集和规则配置。以智能采集为例，在“新建采集任务”对话框中单击“2、采集数据”按钮，再单击“配置助手”按钮，如图 2-18 所示。

图 2-18　分别单击“2、采集数据”和“配置助手”按钮

11）打开“采集规则自动化配置”窗口，在地址栏输入采集地址，单击“生成文章采集规则”按钮，系统即将文章的智能规则输入系统中。单击“测试”按钮，可以检查采集结果是否正确，如图 2-19 所示。单击“确定退出”按钮，配置完成。

图 2-19　“采集规则自动化配置”窗口

12）在返回的“新建采集任务：心血管”对话框中单击“采集任务测试”按钮，在打开的工作页面中单击“启动测试”按钮，如图 2-20 所示。

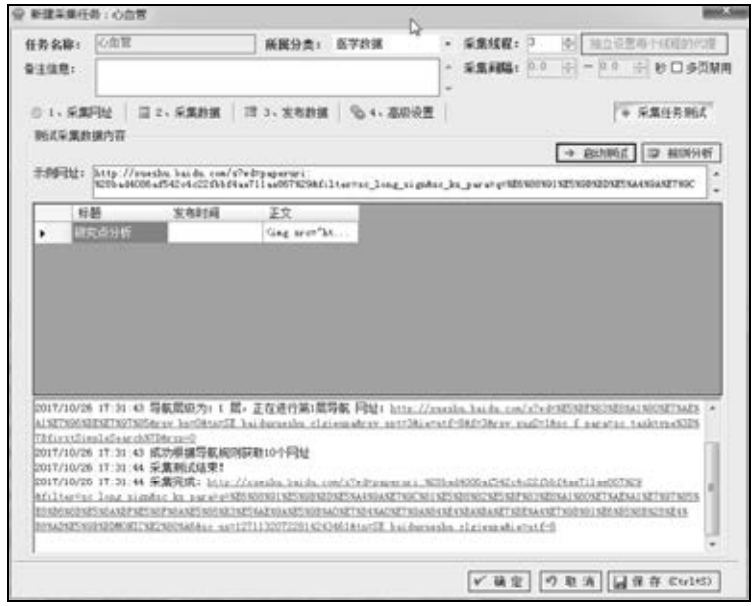

图 2-20　单击“启动测试”按钮

13）设置完成后，返回最初操作页面。右击任务，在弹出的快捷菜单中选择“启动任务”命令（图 2-21），可看到下方屏幕滚动，停止后则采集完成。

图 2-21　启动采集任务

14）采集任务完成后，任务将以.smt 文件格式保存在安装路径的 tasks 文件夹内。

右击采集任务的名称，在弹出的快捷菜单中选择数据导出格式（包括文本、Excel 和 Word 等），如图 2-22 所示。例如，选择“导出 Excel”命令，导出结果如图 2-23 所示。

图 2-22　选择数据导出格式

	A	B	C
1	标题	发布时间	正文
2	研究点分析		<img src="http
3	《中国心血管病报告2012》概要		 我们已与文
4	中国心血管病报告2013概要		 我们已与文
5	研究点分析		<img src="http
6	研究点分析		 我们已与文
7	中国心血管病预防指南简介		作者
8	心血管病预防的现状和展望		作者
9	我国中年人群向心性肥胖和心血管病危险因素及其聚		<i class="ico
10	现代心血管病学		<i class="c-i
11	心理行为因素与心血管疾病的发生发展		作者

图 2-23　导出 Excel 结果

上述采集任务完成后，即可实现在互联网上采集数据。可在“已经完成的任务”列表中查看已经下载的数据，选中任务后右击也可以查看、编辑和发布数据等。

本 章 小 结

本章主要介绍大数据采集、医学大数据采集的基本概念、数据来源与方法。从医学大数据中能获取大量的医学经验和知识，也能更为可靠地获得解决各种医学问题的新途径，造福于患者并保障人民健康。然而，不同种类的医学数据的性质是不同的，并且它们的价值和问题也是不同的，医学大数据的发展目前仍面临一系列障碍，包括技术限制、成本高昂、处理及分析数据对于多学科知识的要求等。因此，构建统一的数据标准、解决数据敏感性问题、进行数据共享是进一步增大数据量的重要手段，也是医学大数据采集实现的重要目标。

关键词注释

1．关系型数据库：建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据。

2．并行数据库：通过并行实现各种数据操作，如数据载入、索引建立、数据查询等，可以提高系统的性能。

3．分布式数据库：利用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上统一的数据库。数据分布存储于若干场地，并且每个场地由独立于其他场地的数据库管理系统（Database Management System，DBMS）进行数据管理。

4．组学大数据：生物大数据中的组学包括基因组、转录组、蛋白质组和表观基因组学等。

5．HIS（Hospital Information System，医院信息管理系统）：医院管理和医疗活动中进行信息管理和联机操作的计算机应用系统。

6．PACS（Picture Archiving and Communication System，影像归档与通信系统）：应用在医院影像科室的系统，主要任务是把日常产生的各种医学影像通过各种接口以数字化的方式海量保存起来，当需要时在一定的授权下能够很快地调回使用，同时增加一些辅助诊断管理功能。

7．LIS（Laboratory Information System，实验室信息系统）：自动接收检验数据，打印检验报告，系统保存检验信息，可根据实验室的需要实现智能辅助功能。

8．URL（Uniform Resource Locator，统一资源定位符）：对可以从互联网上得到的资源的位置和访问方法的一种简洁的表示，是互联网上标准资源的地址。互联网上的每个文件都有一个唯一的 URL，其包含的信息包括文件的位置及浏览器应该怎么处理它。

习 题 2

一、填空题

1．大数据的数据采集是在确定用户目标的基础上，针对该范围内所有结构化数据、半结构化数据和_____数据的采集。

2．医学大数据采集的结构化数据包括电子病历、_____、患者数据、健康管理和医疗报销等。

3．医学大数据采集的非结构化数据包括影像记录、_____和诊断视频数据等。

4．根据数据来源，医学大数据分为医疗大数据、服务平台医疗健康大数据、医学研究或疾病监测大数据、自我量化大数据、网络大数据和_____六大类。

5．医疗大数据来源于_____、科研和管理过程，包括各种门/急诊记录、住院记录、影像记录、实验室记录、用药记录、手术记录、随访记录和医疗保险数据等。

6．医疗大数据具有数据量庞大、产生速度快、_____和价值密度低等典型大数据的特征。

7．_____数据的收集事先都经过了充分地科学论证和规划，比原始的医院数据更为规范。

8．_____大数据有助于个性化临床诊疗，塑造全新的医疗或健康管理模式。

9．生物大数据具有很强的生物专业性，主要是关于生物标本和基因测序信息的数据，_____是其中的重要内容。

10．Chukwa 基本架构中，数据被_____收集，并传送到 Collector，由 Collector 写入 HDFS，然后由 MapReduce Jobs 进行数据的预处理。

二、简答题

1．简述大数据采集的概念。

2．简述传统的数据采集与大数据的数据采集的异同。

3. 简述医学大数据的数据来源。
4. 简述网络大数据的概念。
5. 简述网络爬虫的基本工作流程。

第 3 章
大数据分析

导学

本章主要介绍大数据分析的基础知识、大数据分析的主要技术及大数据分析处理系统，以及医学大数据分析实证应用案例。通过对本章的学习，读者可以对大数据分析有一个概括性的了解和掌握。

了解：大数据分析的基本思想、目前国内外大数据分析的主要状况、大数据分析的应用案例。

掌握：大数据分析的基本概念，大数据分析过程，大数据分析的研究方向，大数据分析的基本技术，大数据分析处理系统的类型、特点和作用。

大数据分析就是研究包含各种数据类型的大型数据集的过程。大数据技术可以发现隐藏的数据模式、未知数据的相关性、发展趋势和其他有用的商业信息。就医学大数据分析而言，其分析结果可以带来更有效的医疗诊治、更好的医疗服务，提高医疗效率，获得竞争优势及其他医疗和商业利益。

3.1 大数据分析概述

在方兴未艾的大数据时代，人们要掌握大数据分析的基本方法和分析过程，从而探索出大数据中蕴含的规律与关系，解决实际业务问题。

3.1.1 大数据分析简介

大数据分析指对规模巨大的数据进行分析，是一组能够高效存储和处理海量数据，并有效达成多种分析目标的工具及技术的集合。

以下案例为美国利用大数据分析实现精准推送健康诊疗知识宣传，我们通过此案例来初步认识大数据分析。其分析过程如图 3-1 所示。

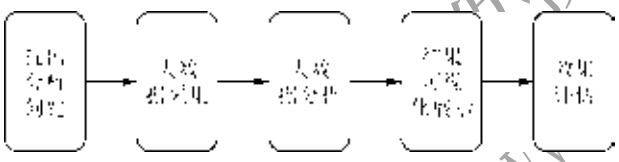

图 3-1 大数据分析过程

第 1 步：提出分析问题，精准定向投放健康诊疗知识材料。将健康诊疗知识精准地送到需要的人手中，提升公共卫生宣传效果是有社会意义的问题。一般医疗宣传的做法是大量投放广告，这需要大量的人力物力，而且很难分清广告的作用。大数据技术可以对某个地区某些疾病的相关数据进行收集和分析，从而找到需要健康诊疗知识的人群。

第 2 步：大数据采集，获得居民的医院诊疗及医学网站上咨询的数据。分析团队搜索采集数据，如该地区居民的诊疗数据、相关的医学网站上的问诊数据，形成数据集，为数据分析做准备。

第 3 步：大数据分析，给出具体的健康诊疗知识投放方案。对采集的数据进行分析挖掘，为需要帮助的患者提供精准可靠的健康诊疗知识，哪个地区的患者对某种健康诊疗知识有需求，相应的健康诊疗知识就送到其电子邮箱和地区的报纸上，非常精准，节省人力物力。

第 4 步：结果可视化展示，将健康诊疗知识投放方案图形化。根据数据分析结果，用图表等生动、易理解的方式将解决方案展示出来。

第 5 步：效果评估，提升健康宣传工作效率。与传统的健康诊疗知识宣传相比，通过大数据分析的创新方案，相关公共卫生宣传部门可提高工作效率，大幅度提高了健康宣传对象的精准度。

3.1.2　大数据分析的研究方向

大数据分析包括预测性分析、可视化分析、大数据挖掘分析 、语义引擎分析、数据质量和数据管理分析 5 个主要方向。

1. 预测性分析

大数据分析最普遍的应用就是预测性分析，即从大数据中挖掘有价值的知识和规则，通过科学建模呈现出结果，然后可以将新的数据代入模型，从而预测未来的情况。

例如，麻省理工学院的研究者创建了一个计算机预测模型来分析心脏病患者丢弃的心电图数据。他们利用数据挖掘和机器学习在海量的数据中筛选，发现心电图中出现 3 类异常者 1 年内死于第 2 次心脏病发作的概率比未出现者高 1～2 倍。这种新方法能够预测出更多的、无法通过现有的风险筛查被探查出的高危患者。

2. 可视化分析

不管是数据分析专家还是普通用户，他们对大数据分析最基本的要求就是可视化分析，因为可视化分析能够直观地呈现大数据特点，同时能够非常容易地被用户所接受。可视化分析可以直观地展示数据，让数据自己说话，让用户看到结果。数据可视化是数据分析工具最基本的要求。图 3-2 是某社区卫生服务站分布情况的地理位置可视化分析。

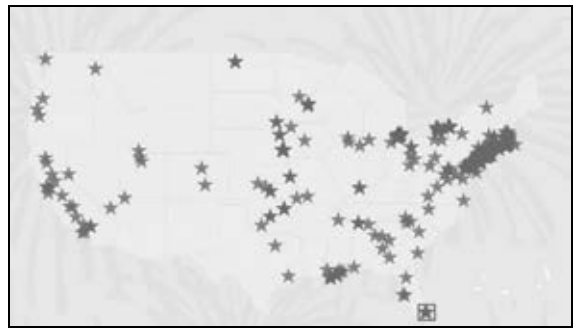

图 3-2　某社区卫生服务站分布情况的地理位置可视化分析

3. 大数据挖掘分析

可视化分析结果是给用户看的，而数据挖掘算法是给计算机看的，通过让机器学习算法，按人的指令工作，从而呈现给用户隐藏在数据之中的有价值的结果。大数据分析的理论核心是数据挖掘算法，算法不仅要考虑数据的量，而且要考虑数据处理的速度。目前许多领域的研究是在分布式计算框架上对现有的数据挖掘理论加以改进，进行并行化、分布式处理。

常用的数据挖掘方法有分类、预测、关联规则、聚类、决策树、描述和可视化、复杂数据类型挖掘（Text、Web 、图形图像、视频、音频）等。很多学者对大数据挖掘算法进行了研究并发表了相关文献。例如，有文献提出了对适合慢性病分类的 C4.5 决策树算法进行改进，对基于 MapReduce 编程框架进行算法的并行化改造；有文献提出了对数据挖掘技术中的关联规则算法进行研究，并通过引入兴趣度对经典 Apriori 算法进行改进，提出了一种基于 MapReduce 的改进的 Apriori 医疗数据挖掘算法。

4. 语义引擎分析

数据的含义就是语义。语义技术指从词语所表达的语义层次上来认识和处理用户的检索请求。

语义引擎通过对网络中的资源对象进行语义标注，以及对用户的查询表达进行语义处理，使自然语言具备语义上的逻辑关系，能够在网络环境下进行广泛有效的语义推理，从而更加准确、全面地实现用户的检索。大数据分析广泛应用于网络数据挖掘，可从用户的搜索关键词来分析和判断用户的需求，从而实现更好的用户体验。

例如，一个语义搜索引擎试图通过上下文来解读搜索结果，它可以自动识别文本的概念结构。例如，搜索“血型”，语义搜索引擎可能会获取包含“A 型血”“B 型血”“O 型血”的文本信息，即语义搜索可以对关键词的相关词和类似词进行解读，从而扩大搜索信息的准确性和相关性。

5. 数据质量和数据管理分析

数据质量和数据管理指为了满足信息利用的需要，对信息系统的各个信息采集点进

行规范，包括建立模式化的操作规程、原始信息的校验、错误信息的反馈、矫正等一系列过程。大数据分析离不开数据质量和数据管理，高质量的数据和有效的数据管理，无论是在学术研究还是在商业应用领域，都能够保证分析结果的真实和有价值。

3.2　大数据分析的主要技术

大数据分析的一个核心问题是如何对数据进行有效表达、解释和学习，无论是对图像、声音还是文本数据，要挖掘大数据的价值必然要对大数据进行内容上的分析与计算。深度学习和知识计算是大数据分析的基础，而可视化在数据分析和结果呈现的过程中均起作用（关于可视化的具体处理方法见本书第 4 章）。本节主要介绍深度学习和知识计算这两个大数据分析的关键技术。

3.2.1　深度学习

1. 深度学习的概念

深度学习是机器学习研究中的一个新领域，其动机在于建立、模拟人脑进行分析学习的神经网络，其模仿人脑机制来解释数据，如图像、声音和文本。

2016 年年初，AlphaGo 击败了前世界第一的围棋选手李世石，使“深度学习”这个名词吸引了全球的关注目光。深度学习的概念源于人工神经网络的研究，即让计算机具有人一样的智慧。深度学习利用层次化的架构学习，使研究对象在不同层次上得到表达，这种层次化的表达可以帮助解决更加复杂抽象的问题。在层次化中，高层的概念通常是通过低层的概念来定义的，深度学习可以对人类难以理解的低层数据特征进行层层抽象，从而提高数据学习的精度。让计算机模仿人脑机制来分析数据，建立类似人脑的神经网络进行机器学习，从而实现对数据的有效表达、解释和学习，这种技术在人工智能上无疑是前景无限的。

2. 深度学习的应用

近几年，深度学习在健康医疗领域、自然语言处理、语音、图像等领域取得了一系列重大进展。

（1）深度学习在健康医疗领域的应用

深度学习在健康医疗领域的应用主要有七大方向：①提供临床诊断辅助系统等医疗服务，应用于早期筛查、诊断、康复、手术风险评估场景；②医疗机构的信息化，通过数据分析，帮助医疗机构提升运营效率；③进行医学影像识别，帮助医生更快、更准确地读取患者的影像；④利用医疗大数据，助力医疗机构大数据可视化及数据价值提升；⑤在药企研发领域，解决药品研发周期长、成本高的问题；⑥在健康管理服务领域，通过包括可穿戴设备在内的手段，监测用户个人健康数据，预测和管控疾病风险；⑦在基因测序领域，将深度学习用于分析基因数据，推进精准医疗。

目前比较常见的是自然语言理解类辅助诊断系统和医学影像识别类辅助诊断系统

两个领域。

在自然语言理解类辅助诊断系统领域，著名的 IBM Watson 机器人（图 3-3）经过 4 年多的训练，学习了 200 本肿瘤领域的教科书、290 种医学期刊和超过 1500 万份的文献后，开始被应用在临床上，在肺癌、乳腺癌、直肠癌、结肠癌、胃癌和宫颈癌等领域向人类医生提出建议。2015 年，Watson 用 10min 左右为一名 60 岁的女性患者诊断出白血病，并向东京大学医科学研究所提出了适当的治疗方案。

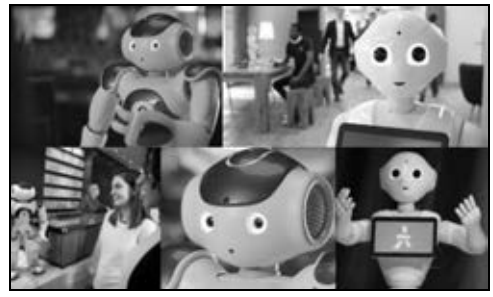

图 3-3　自然语言理解类辅助诊断机器人 Watson

在医学影像识别类辅助诊断系统领域，中国的人工智能企业 Airdoc 目前已经掌握了世界领先的图像识别技术，在心血管、肿瘤、神内、五官等领域建立了多个精准深度学习医学辅助诊断模型。例如，Airdoc DR 系统可帮助医生识别筛查糖尿病视网膜病变。

（2）深度学习在自然语言处理领域的应用

深度学习在自然语言处理等领域主要应用于机器翻译及语义挖掘等方面，国外的 IBM（International Business Machines Corporation，国际商用机器公司）、Google 等公司快速进行了语音识别的研究；国内的阿里巴巴公司、科大讯飞公司、百度公司、中科院自动化所等公司或研究单位也在进行深度学习在语音识别上的研究。

（3）深度学习在图像领域的应用

深度学习在图像领域也取得了一系列进展，如 Microsoft 公司推出的网站 how-old.net，用户可以在此网站上传自己的照片进行年龄评估。系统根据照片会对瞳孔、眼角、鼻子等 27 个“面部地标点”展开分析，并判断照片上人物的年龄，如图 3-4 所示。

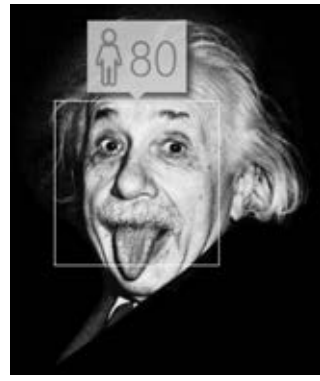

图 3-4　人脸识别判断年龄