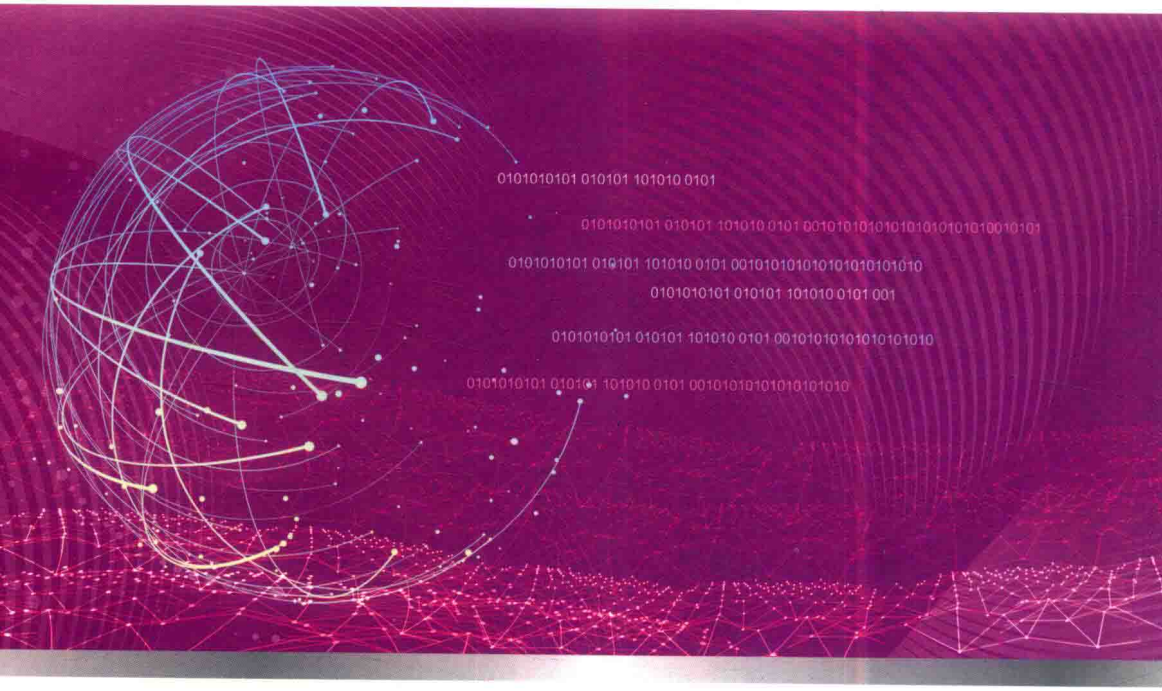
基于科技大数据的情报分析方法与技术研究

曾　文◎著



科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

本书的出版得到中国国家社会科学基金项目“基于事实型科技大数据的情报分析方法及集成分析平台研究”（项目编号：14BTQ038）特别资助

基于科技大数据的情报分析方法与技术研究

曾　文　著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS
·北京·

图书在版编目（CIP）数据

基于科技大数据的情报分析方法与技术研究 / 曾文著. —北京：科学技术文献出版社，2018.7
ISBN 978-7-5189-4630-3

Ⅰ.①基… Ⅱ.①曾… Ⅲ.①数据处理—应用—情报分析—研究 Ⅳ.①G252.8

中国版本图书馆 CIP 数据核字（2018）第 148211 号

基于科技大数据的情报分析方法与技术研究

策划编辑：周国臻　　责任编辑：李　鑫　　责任校对：张吲哚　　责任出版：张志平

出 版 者　科学技术文献出版社
地　　址　北京市复兴路15号　邮编　100038
编 务 部　（010）58882938，58882087（传真）
发 行 部　（010）58882868，58882870（传真）
邮 购 部　（010）58882873
官方网址　www.stdp.com.cn
发 行 者　科学技术文献出版社发行　全国各地新华书店经销
印 刷 者　北京教图印刷有限公司
版　　次　2018 年 7 月第 1 版　2018 年 7 月第 1 次印刷
开　　本　710×1000　1/16
字　　数　226千
印　　张　14
书　　号　ISBN 978-7-5189-4630-3
定　　价　68.00元

序　言

大数据影响了情报研究的任务对象和工作环境，如何在大数据环境下进行情报的感知、刻画和响应是情报工作者不可回避的一个重要问题。本书作者以科技情报工作者特有的敏锐，观察解读与科技大数据相关的情报分析方法和技术，为新时期的情报研究理论与实践做出了突出贡献。

大数据作为一个非学术性的操作概念，映射于管理和科研领域的方方面面，背景相异的人们出于不同目的给予“大数据”多种解读。基于科研管理关切对科技大数据进行术语解读和实操分析，则是情报学者应尽的本分。作者基于科技情报的业务本分阐释了与科技大数据相关的术语，形成了一套对科技大数据的认知理念，在此基础上依照数据处理的流程，渐次讨论了对科技大数据的采集处理、建模与分析问题，总结梳理出适用于科技大数据管理问题的情报分析基本方法。

大数据的分析使用不是孤立事件，对科技大数据的应用更要厘清相关的环境和操作条件。作者解析了科技大数据所对应的资源建设问题，抓住其中关键的术语，自动获取技术进行展示分析，对能够运用科技大数据的情报分析方法逐一进行评介，给读者呈现出一幅完整的科技大数据情报分析方法图景。

科技大数据的应用在数据准备和方法处理上均有特殊之处，需

要通过特定平台予以实现。作者阐述了科技大数据情报分析平台的构建问题，提取和剖析了科技大数据应用中的关键算法与技术，为科技大数据应用提供了平台技术和知识保障条件。

本书是科技大数据应用领域为数不多的专门著作，从内容组织到表达方式都反映出科技情报的业务特色，值得与关注科技管理、科技大数据的专业读者分享，也值得非专业读者借鉴。

北京大学 信息管理系

2018 年 7 月 10 日

前　言

我们生活在一个充满“数据”的时代，“数据”已经渗透到人类的工作和生活中。人类不仅是“数据”的使用者，同时也是“数据”的生产者，大数据已经与我们的工作和生活息息相关、须臾难离。中国工程院院士高文说：“不管你是否认同，大数据时代已经来临，并将深刻地改变着我们的工作和生活。”2015 年 5 月，习近平总书记在给国际教育信息化大会的贺信中指出：“当今世界，科技进步日新月异，互联网、云计算、大数据等现代信息技术深刻改变着人类的思维、生产、生活、学习方式，深刻展示了世界发展的前景。”以习近平同志为核心的党中央，站在时代最前沿，带领全国人民迈入大数据时代。十八届五中全会通过的“十三五”规划建议提出：“实施国家大数据战略，推进数据资源开放共享。”

计算机技术的发展给情报学，特别是情报分析的方法带来了强烈的冲击和影响，传统的情报分析方法面对时代的发展和技术进步的现实，同样需求新方法和新技术充实到科技情报分析的事业中，情报分析的技术和工具需要更具实用性和应用性。本书内容以科技大数据为视角和分析对象，注重情报分析的理论和实践内容相结合，所介绍的分析方法可用于情报分析的实际工作中。本书较为完整地论述了科技大数据的概念、内涵及技术架构；详细介绍了科技大数据的采集和预处理、科技大数据的建模和分析方法等；既阐述传统的科技情报分析方法，又论述科技大数据资源建设及科技情报分析的新方法和新思路。例如，本书将重点介绍如何把深度学习这一新技术，应用于科技大数据的情报分析过程中。本书不仅注重情

报分析方法的研究内容介绍，而且对情报分析工具的研发和使用也做了比较详尽的阐述。

在编写原则上，本书既保持大数据技术本身应有的系统性和理论性，又着重体现其在科技情报分析的针对性与应用性。全书在内容上共分成9章。第一章、第二章和第三章分别为科技大数据的相关术语解读、科技大数据的采集处理及科技大数据所对应的建模与分析问题。第四章、第五章、第六章和第七章分别从科技大数据所使用的基本分析方法、科技大数据所对应的资源建设问题、科技大数据中的术语自动获取技术、科技大数据所支持的情报分析方法论述科技大数据分析的基本理论和方法与技术。第八章为科技大数据情报分析平台的建设。第九章从云计算的角度介绍科技大数据应用中的关键算法与技术。

在编撰本书的过程中借鉴使用了多种媒介上的成果，亦有辗转摘录的资料无法一一注明出处，在此向同行学者表示敬意和谢意。

感谢中国科学技术信息研究所领导和同事的支持。感谢徐红姣、桂婕、李智杰、翟娟华和刘旭等同志对本科研项目研究工作及本书撰写过程中给予的支持和帮助。感谢科学技术文献出版社周国臻老师为本书的顺利出版付出的艰苦劳动。

衷心欢迎读者朋友提出宝贵意见。

曾　文

2018年6月

目　录

第一章　科技大数据的相关术语解读

由于互联网技术的发展，科技数据处理、商业智能数据分析等具有海量需求的应用变得越来越普遍，面对日益巨大的数据量，无论从形式上还是内容上，均已无法用传统的方式进行采集、存储、操作、管理和分析。当人们认识到数据的价值，那么分析大数据就成为工作的扩展和延伸。而云计算的兴起使原本很难收集和使用的数据开始变得容易被利用起来，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

用于科技情报分析的科技大数据是一种非数值型的数据，如科技文献数据和科技政策数据等。无论是从事情报分析的专家学者，还是从事科学技术研究的科技人员，面对日益增长和积累的庞大数据集都会期待运用某种手段或方法以发现有价值的情报信息或技术趋势。由于科技大数据多是呈现结构化、半结构化或非结构化的数据结构和状态，处理起来烦琐且需要时间，传统的数据管理和处理方法已难以满足这种需求，亟须解决。因此，无论是从科学研究角度来看还是从应用的角度看，科技大数据的应用已经成为科技信息发展的自然延伸。

1.1　概述

1.1.1　科技大数据的基本概念

早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地称为“第三次浪潮的华彩乐章”。2012 年，美国高德纳咨询公司认为，大数据是大量、高速、多变、真实的信息资产，它需要用新型的处理方式去促成更强的决策能力、洞察力与优化处理。2012 年 3 月，美国政府公布“大数据研发计划”，其目标是使人们具有从现有海量和复杂的数据中获取知识的能力，从而加速在科学与工程领域发明的步伐，增强国家安全，转变人们现有的工作、学习和生活方式。2012 年 3 月 22 日，奥巴马政府宣布投资 2 亿

美元拉动大数据相关产业发展，将“大数据战略”上升为国家战略。大数据将成为信息社会未来的“新能源”。

科技大数据是一种特殊类型的大数据，是指与科技信息相关的非数值型数据，也称为事实型科技大数据。具体来说，科技大数据是指长期积累形成的与科技创新全过程相关的各类非数值型科技信息，它涵盖了客观描述科技创新决策和具体的科技创新活动全过程的各类科技信息。最常见的一种类型科技大数据即科技文献数据，以国家科技数字图书馆为例，科技文献数据资源的规模截至 2017 年 4 月，已拥有西文期刊数据 27 765 728 条、中文期刊数据 64 089 206 条、俄文期刊数据 910 264 条、日文期刊数据 2 303 297 条，外文会议数据 8 240 295 条、中文会议数据 2 238 344 条，外文学位论文数据 482 517 条、中文学位论文数据 3 283 673 条，国外科技报告数据 1 325 109 条，中国大陆专利数据 15 121 727 条、中国台湾专利数据 1 352 167 条、美国专利数据 5 097 875 条、英国专利数据 1 926 386 条、法国数据专利 990 869 条、德国专利数据 3 446 017 条、瑞士专利数据 717 265 条、日本专利数据 23 615 455 条、欧洲专利数据 3 223 554 条、韩国专利数据 4 798 727 条、印度专利数据 430 835 条、以色列专利数据 271 030 条、俄罗斯专利数据 1 020 362 条、加拿大专利数据 198 027 条、世界知识产权组织专利数据 2 985 051 条，科技丛书数据 299 154 条。面对如此庞大的数据规模，科技情报分析工作尤为重要。

科技大数据是科技情报分析研究和工作的重要数据源，情报分析人员利用情报分析方法，实现从科技大数据的数据内容中分析出有“价值”的情报信息，科技情报分析的基本工作流程是根据特定需要进行的情报搜集和信息整序工作，透过现象，揭示具体领域科技大数据所蕴含的数据特征、规律和关联等信息，实现“源于科技大数据，高于科技大数据”的分析结果。

1.1.2 科技大数据的来源和特征

科技大数据主要包括两类数据，一类是客观的科研产出和技术产出数据，如科技期刊文献数据、专利数据、学位论文数据和科技报告、技术标准数据等，这类数据相对较为集中，数据格式较为规范，呈结构化或半结构化的特征。另一类是各级组织、科研机构、企业发布的科技政策、新闻等网页信息、科研个体的个人学术网站、微博，以及科研论坛等产生的动态、实时和交互式网络事实型数据，这类数据较为离散，数据格式规范性差，主要呈非结构化的特征。

在大数据背景下，科技大数据的采集、分析、处理较传统方式有了明显的

变化，两者间的比较如表 1-1 所示。

表 1-1　传统数据与科技大数据的特征比较

	传统数据	科技大数据
数据产生方式	被动采集数据	主动生成数据
数据采集密度	采样密度较低，采样数据有限	利用大数据平台，针对科技文献或事件的数据进行密度采样，精确获取科技文献或事件的全局数据
数据源	数据源获取较为孤立，不同数据之间的数据整合难度较大	利用大数据技术，通过分布式技术、分布式文件系统、分布式数据库等技术对多个数据源获取的数据进行整合处理
数据处理方式	大多采用离线处理方式，对生成的数据集中分析处理，不对实时产生的数据进行分析	较大的数据源、响应时间要求低的应用可以采取批处理方式集中计算；响应时间要求高的实时数据处理采用流处理的方式进行实时计算，并通过对历史数据的分析进行预测分析

与大数据一样，科技大数据同样呈现出“4V1O”的特征，具体如下。

①数据量大（Volume）是科技大数据的首要特征，包括采集、存储和计算的数据量都非常大。大数据的起始计量单位至少是 100 TB。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）是指科技大数据种类和来源多样化，以科技文献数据为例，有科技期刊文献、专利文献、学位论文文献、科技图书文献等，这些数据的类型和来源各不相同。编码方式、数据格式、应用特征等多个方面都存在差异性，多信息源并发形成大量的异构数据。

③数据价值密度化（Value）是指科技大数据价值密度相对较低，需要经过很多处理过程才能挖掘出来。随着互联网和物联网的广泛应用，科技信息无处不在，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，是科技大数据时代最需要解决的问题。

④速度快、时效高（Velocity）是指随着互联网的发展，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科技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，这些都是科

技大数据区别于传统数据挖掘的显著特征。

⑤数据在线(On-Line)是指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。数据不仅规模大,而且相当一部分数据是在线的,这是互联网高速发展的特点和趋势。

科技大数据呈现的主要特点除了数据量大且增长速度快,数据来源和数据结构类型多,有价值的数据相对比例小之外,科技大数据具有敏感性和积累性,会涉及国家安全和利益。因此,科技大数据的处理和数据分析与其他类型大数据相比,更具有一定的复杂性。

1.1.3 科技大数据处理的基本流程

科技大数据的处理流程可以定义为在适合工具的辅助下,对异构数据源进行抽取和集成,结果按照一定的标准统一存储,利用合适的数据分析技术对存储的数据进行分析,从中提取有益的科技情报或知识,并利用恰当的方式将结果展示给终端用户。科技大数据处理的基本流程如图 1-1 所示。

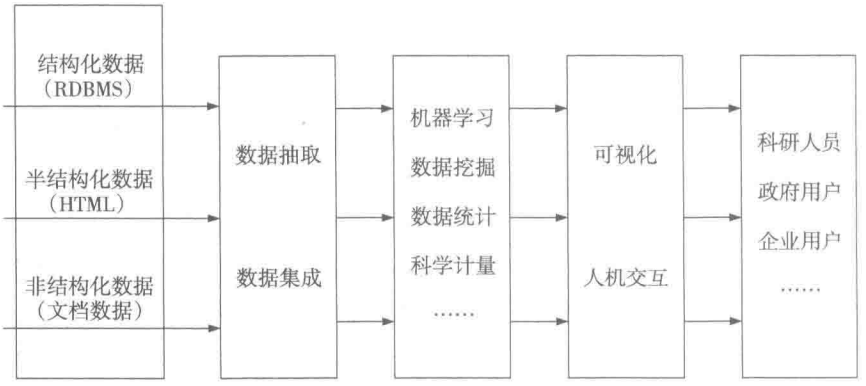


图 1-1 科技大数据处理的基本流程

(1)数据的抽取与集成

由于科技大数据处理的数据来源类型广泛,所以其第一步需要对数据进行抽取和集成,从中找出数据的关系和实体,经过关联、聚合等操作,再按照统一的格式对数据进行存储。

(2)数据分析

数据分析是科技大数据处理流程的核心步骤,通过抽取和集成环节,从异

构的数据源中获得用于大数据处理的原始数据，用户根据需求对数据进行分析处理。从技术角度看，通常采用的数据分析技术主要有数据统计、数据挖掘、机器学习、科学计量等。

（3）数据解释

用户最关心的是数据处理的结果及以何种方式在终端上显示结果，因此采用什么方式展示处理结果非常重要。就目前来看，可视化和人机交互是数据解释的主要技术。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。人机交互技术可以引导用户对数据进行逐步分析，参与并理解数据分析结果。

1.2　科技大数据的技术架构

科技大数据应用需要新的工具和技术来存储、管理和实现价值。新的工具、流程和方法支撑起了新的技术架构，计算机技术是支撑科技大数据应用的基础，其技术架构必须能够以经济的方式存储比以往数量更大、类型更多的数据。此外，还必须适应数据变化的速度，即解决海量数据难以在当今的网络连接条件下快速“移动”的问题，即大数据的基础技术架构必须具有分布计算能力，以便能在接近用户的位置进行数据分析，减少跨越网络所引起的延迟。云计算为科技大数据的处理提供一种灵活的选择，可以实现大数据分析所需的效率、可扩展性、数据便携性和经济性，但其存储和提供数据还不够，必须采用新技术去融合、分析和关联数据，才能提供更多有价值的数据。有部分大数据方法要求处理未经建模的数据。因此，可以用不相关的数据源，不同类型的数据进行模式匹配。从而使科技大数据的分析能以新视角挖掘数据价值。基于上述考虑，一般可以构建出适合科技大数据研究和开发的四层堆栈式技术架构，如图 1-2 所示。

（1）基础层

第一层作为整个科技大数据技术架构基础的最底层，也叫基础层。要实现大数据规模级的应用，需要一个高度自动化的、可横向扩展的存储和计算平台。这个基础设施需要从以前的存储“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。云模型鼓励访问数据并通过提供弹性资源池来应对大规模的数据问题，解决如何存储大量数据及如何积聚所需的计算资源来操作数据的问题。在云中，数据跨多个结点调配和分

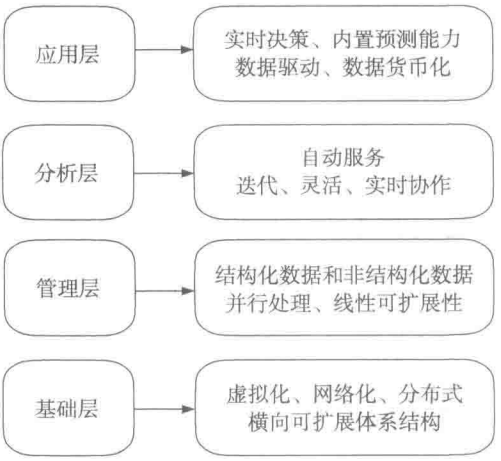


图 1-2　四层堆栈式技术架构

布，使数据更接近需要它的用户，从而缩短响应时间，提高效率。

（2）管理层

科技大数据需要支持多源数据的深层次的分析，因而在技术架构中需要一个管理平台，即管理层，它使结构化和非结构化数据管理为一体，具备实时传送、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问题。

（3）分析层

科技大数据的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，帮助用户获得深入的数据价值分析结果。可扩展性强、使用灵活的大数据分析平台更可成为科技工作者的重要辅助工具。

（4）应用层

不断涌现的大数据应用对大数据技术提出更多的新要求，大数据技术也因此在不断地发展中日趋成熟。科技大数据的价值体现在帮助用户进行决策和为终端用户提供预测服务的应用。不同的需求驱动科技大数据的不同应用。

1.3　科技大数据技术

科技大数据需要特殊的技术，以有效地处理在允许时间范围内的大量数