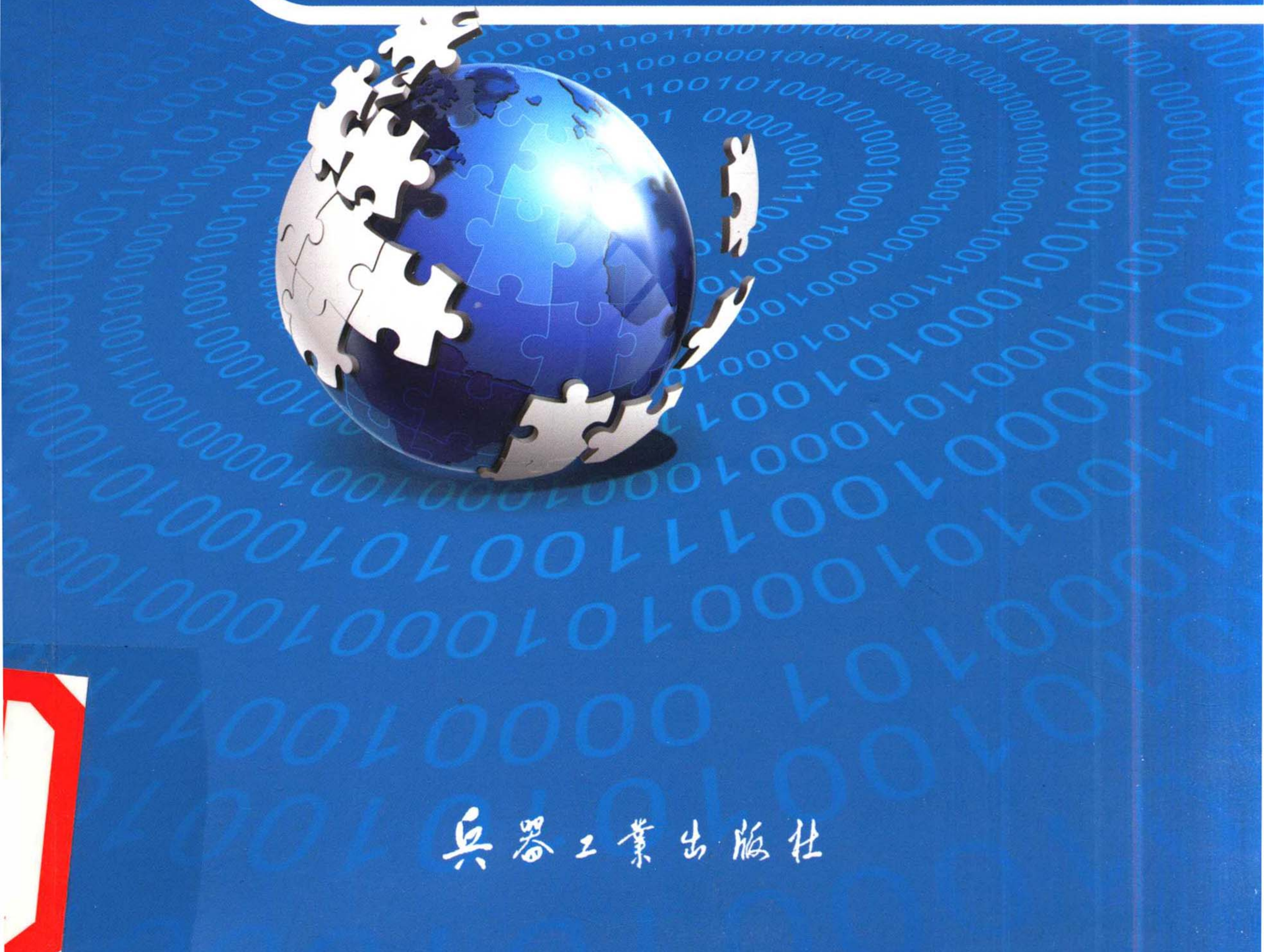
主　编 刘　如
副主编 吴晨生 刘彦君 徐　成

大数据时代科技情报可视化应用研究



兵器工业出版社

大数据时代科技情报可视化应用研究

主　编　刘　如
副主编　吴晨生　刘彦君　徐　成

兵器工业出版社

内容简介

本书较为全面系统地总结了信息可视化的发展脉络和发展趋势、国内外可视化技术及方法的研究现状、情报信息可视化的理论研究、大数据信息向情报信息的转化、国内外经典可视化工具及案例的分析等。在此基础上，构建了科技情报可视化系统，设计了情报辅助系统中的可视化模块方案。本书按照理念形成、技术分析、解决方案和典型案例的思路介绍了科技信息可视化方案的建设。科技信息可视化是大数据时代的必然产物，是一种整体的数据分析解决方案，是一个高效、生动的展示体系。科技信息可视化技术打破了传统的科技信息管理和分析模式，实现了决策高效、信息准确、展示形象等诸多特征，朝最终实现“智慧情报”迈出了一大步。

图书在版编目（CIP）数据

大数据时代科技情报可视化应用研究 / 刘如主编.
-- 北京 : 兵器工业出版社, 2015.12
ISBN 978-7-5181-0153-5

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①科技情报－可视化软件
Ⅳ. ①G350

中国版本图书馆CIP数据核字(2015)第257987号

出版发行：兵器工业出版社
发行电话：010－68962596，68962591
邮　　编：100089
社　　址：北京市海淀区车道沟10号
经　　销：各地新华书店
印　　刷：北京集特印刷有限公司
版　　次：2015年12月第1版第1次印刷

责任编辑：朱　婧
封面设计：理想设计
责任校对：郭　芳
责任印制：王京华
开　　本：787×1092　1/16
印　　张：8.5
字　　数：150千字
定　　价：30.00元

本书编委会

主　编： 刘　如

副主编： 吴晨生　刘彦君　徐　成

编　委（按姓氏拼音排序）：

董晓晴　贾明慧　李　荣　李梦辉

李楠欣　王　康　吴　琼　吴玉辉

张鲁冀　赵俊超

前　言

21世纪以来，数字化、数据化与网络化为特征的信息技术迅速发展，成为促进经济发展和社会进步的主导技术，使全球经济增长方式发生了根本性的变化。在信息化的进程中，信息和信息技术的影响已经渗透到社会各个方面，极大地改变了人们的生活和工作方式。大数据时代下经济、社会的飞速发展所产生的海量科技情报，在数量上呈现指数增长的态势，在内容上不断细化，在学科上不断交叉渗透和汇聚融合，并随着互联网等信息传输工具的大量普及而得以更加快速地在世界范围内传播。数据越庞大，所能透露的信息量也就越多，对科技情报工作来说，是大数据时代新一轮的机遇与挑战。但同时，如何让这些庞大的数据显示或者展示出科技信息为科技情报服务提出了严峻的挑战。信息的可视化作为一种统计学工具，用于创建一条快速认识数据集的捷径，并成为一种令人信服的沟通手段，传达存在于数据中的基本信息。而最原始的统计图表只能呈现基本的信息，发现数据之中的结构，可视化定量的数据结果。面对大数据时代复杂或大规模异型数据集，比如产业分析、专利分布、科研人员研究行为数据等，信息可视化面临处理的状况会复杂得多。面对这些新的挑战，科技情报服务不再满足于提供未加筛选的情报信息，而是需要更加“智慧”的情报，即提供经过鉴别的有价值的情报信息、有事实依据的分析性情报信息、有预见未来发展的预测性情报信息。这种高价值的精准情报服务是需要大量数据支撑的，并使其可视化展示。

本书首次提出“情报信息可视化”的新理念，通过大数据驱动信息设计，为科技情报服务带来“大智慧”。在分析研究了大数据时代下可视化研究背景、研究现状的基础上，分析了适应大数据环境下情报形成的路径，研究了情报信息可视化的表述分类，分析了经典的可视化平台与开发工具，最终设计了嵌入在科技情报辅助系统中的各阶段的模块，改变了传统的情报分析模式，改善了当前情报分析工具或软件的可视化效果，增强了大数据环境下情报分析的智能化能力，使得情报在辅助决策的功能上更直观、有效。本书对科技信息可视化模块的研究设计，旨在聚合科技相关数据，逐步形成北京市科学技术情报可视化方案，最终为政府决策、行业引领、城市规划与公共服务提供全方位数字化解决方案，从而推动科

学进步及社会发展。

《2011~2012年世界经济论坛报告》指出，“信息可视化”是聚合世界这一复杂系统的有效手段，它能有效处理和消化海量数据，百万级的数据可以被转化成一张图表或者一段3D动画，把几十年的数据量“压缩”于瞬间。而科技情报信息可视化是创造性设计美学和严谨的情报科学的卓越产物。用极美丽的形式呈现可能非常沉闷繁冗的数据，从而有利于从数据中挖掘分析出潜在的情报产物。

科技情报可视化的应用研究可以有效提高情报机构为情报服务的数据资源建设和处理能力，对于情报预测预警服务可以提供智能展示的支撑，进而增强情报所的行业竞争力。

编著者

2015年9月5日

目　录

第1章　大数据时代科技情报可视化的研究背景

随着这些年全球化的经济发展，中国拥有了更加庞大的应用市场，这使得中国成为了拥有最庞大数据的国家之一。大数据时代的来临是网络技术发展及海量数据处理能力提高所带来的必然产物。2013 年 2 月，工业和信息化部发布《关于数据中心建设布局的指导意见》[1]，提出了数据中心建设和布局的基本原则，即市场需求导向原则、资源环境优先原则、区域统筹协调原则、多方要素兼顾原则和发展与安全并重原则。可见政府对大数据时代的应对措施已经落到实处，从政策支持到财力扶持都做了充足的准备工作。国际数据公司(IDC)发布的关于中国大数据技术和服务市场的首份报告《中国大数据技术与服务市场 2012 ~2016 年预测与分析》[2]显示，该市场规模将会从 2011 年的 7760 万美元增长到 2016 年的 6.17 亿美元，未来 5 年的复合增长率达 51.4%，市场规模增长近 7 倍。而全球大数据技术及服务市场 2016 年收入将达 238 亿美元，接近 1500 亿元人民币。大型国际企业，包括 IBM、谷歌、亚马逊和微软，都纷纷涉足大数据掘金的这一市场。

因此，在大数据时代，中国的产业提升、科技发展的一个重要手段就是要探索以大数据处理为基础的动态情报解决方案。传统的科技情报处理手段可能已经无法处理数据飞涨的情报源，信息的瞬间爆炸给情报的时效性带来新的考验，于是研究如何选择建立一个适应大数据时代的科技情报可视化方案是情报工作者首当其冲的研究课题。

1.1　科技情报工作面临的新环境——大数据时代的来临

随着智能手机、移动终端的大量普及，以及 3G/4G/WIFI 网络的大面积覆盖，情报工作自身相关的数据都可能成为了可被记录和分析的数据，这是科技情报所面临的新的机遇与挑战，也是科技情报不可摆脱的新环境。大规模发布、分享和应用数据的时代开启了一次科技情报重大的时代转型。

1.1.1 科技情报新环境——大数据时代特征

目前对大数据的描述最常见的就是4V(见图1-1)：Volume(数据海量化)、Variety(形式多样化)、Velocity(产生快速化)、Value(最大价值化)。

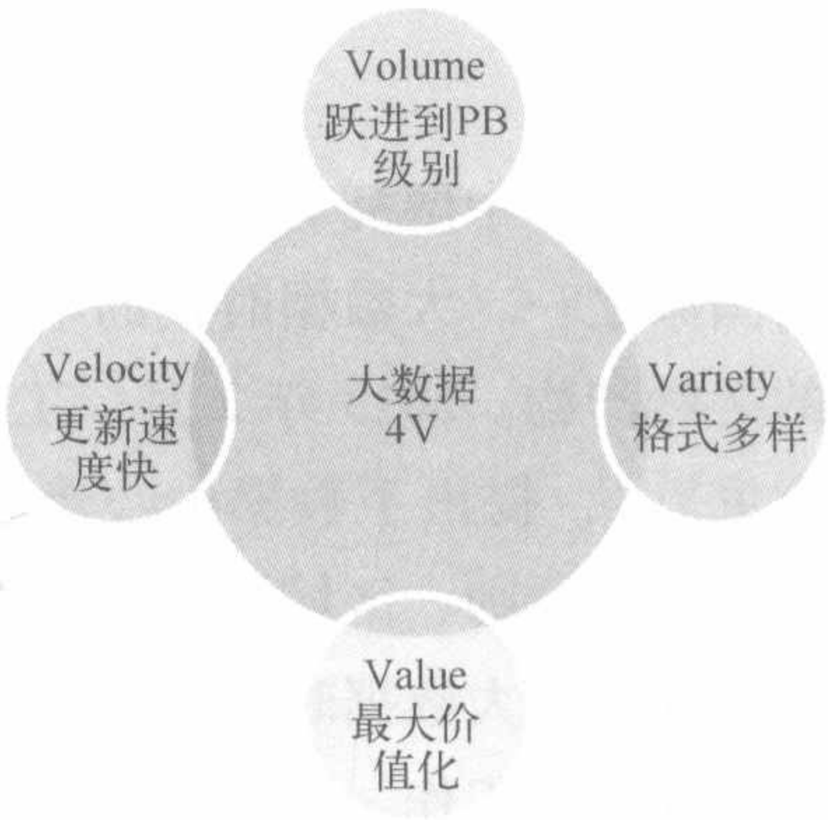


图1-1 大数据的4V特征

(1)数据海量化

随着信息技术的提高，目前数据化(Datafication)发展日趋广泛，其核心动力主要是来源于人类对记录和分析数据信息的渴望，这一点在情报界显得尤为突出。在海量的数据中，绝大部分是非结构化数据，只有少部分结构化数据可适用于传统的数据库。而将这绝大多数非结构化的数据转变成可制表分析的量化形式的过程，就是数据化的过程。数据化使得书籍文字、人们的交流、方位信息等这些信息资源变成人类可阅读、机器可分析的数字资源。截至2013年，世界上存储的数据预计能达到约1.2ZB，如果把这些数据全部记录在书中，这些书可以覆盖整个地球[3]。在这些数据中有多一半是近几年才产生的，在未来，互联网数据每年将会增长近50%，每两年便将翻一番。

(2)形式多样化

大数据环境下结构化和非结构化数据的处理应用是一个新的挑战。相对于以往便于存储的以文本格式的结构化数据，非结构化数据越来越多，比例也越来越大，包括微博日志、图片、音频、视频、地理位置信息等，这些多类型的数据对数据的处理能力提出了更高要求，这也是情报分析技术手段革新的新挑战。

(3)产生快速化

大数据时代，海量数据不断攀升，情报分析有了越来越丰富的信息资源。美

国互联网数据中心指出，目前世界上 90% 以上的数据是最近几年才产生的[3]。这些庞大的快速增长的数据形成了可供情报分析的宝贵资源。通过对这些庞大的数据进行分析，原本不可捉摸的事物规律、人类习惯等变得可被解析、描述和量化，甚至能够对其进行预测和预警。但同时，数据的快速增长也为实时的情报监测分析带来了新的挑战。

（4）最大价值化

据美国权威研究机构——透明度市场研究最新发布的报告《大数据市场：2012 ~ 2018 年全球形势、发展趋势、产业分析、规模、份额和预测》显示，2012 年全球大数据市场产值为 63 亿美元，预计到 2018 年将会达到 483 亿美元，将增长近 7 倍。而美国权威调查咨询机构麦肯锡 2012 年的大数据报告中显示，大数据产业每年为美国医疗系统带来 3000 亿美元的收益；为欧洲公共管理部门带来 2500 亿欧元的收益；为零售业增加 60% 的净利润；为制造业减少了 50% 的产品研发等成本；个人地理位置信息的利用，为服务商带来超过 1000 亿美元的收益，为用户带来超过 7000 亿美元的价值。这些大数据带来的价值其本质就是各行业、各领域通过对大数据的分析，挖掘大数据潜在的价值情报，从而获得最大化的经济效益。

大数据所产生的这些潜在的具有价值的情报信息可以促进相关产业的发展与革新、推动科技创新、催生数据服务性企业；此外，大数据在社会管理、智慧城市、金融服务、医疗卫生、生产制造、商业零售、个人数字生活等方面都具有巨大的价值。

1.1.2　针对科技情报的“大数据”概念

需要特别指出的是，大数据的概念近几年被疯狂炒作，大数据究竟是什么？很多人至今仍然有些模糊。首先值得肯定的是，业内专家普遍认同“大数据不只是更多的数据”。最常见到以下两种概念：

概念 1：2001 年 Doug Laney① 最先提出大数据的“3V”特征：数量（Volume）、速度（Velocity）、种类（Variety）。之后，又被扩展到了 11V，包括有效性（Validity）、真实性（Veridicality）、价值（Value）和可见性（Visibility）等。

① Doug Laney 是 Gartner 副总裁，研究领域涉及商业分析、大数据应用等相关领域。

概念2：大数据是一种技术，早在2001年，大数据就被提出，现在大数据的普及并不是简单地比十几年前有更多的数量、速度和种类，而是因为新技术的推动，数据处理能力的提升、数据处理方式的创新等。也就是说，目前的大数据是之前因为技术限制而被忽略的数据。

纵观目前各行各业对大数据概念的描述，不难看出，大数据的本质是基于数据量的剧增和技术的成熟而产生的(见图1-2)。

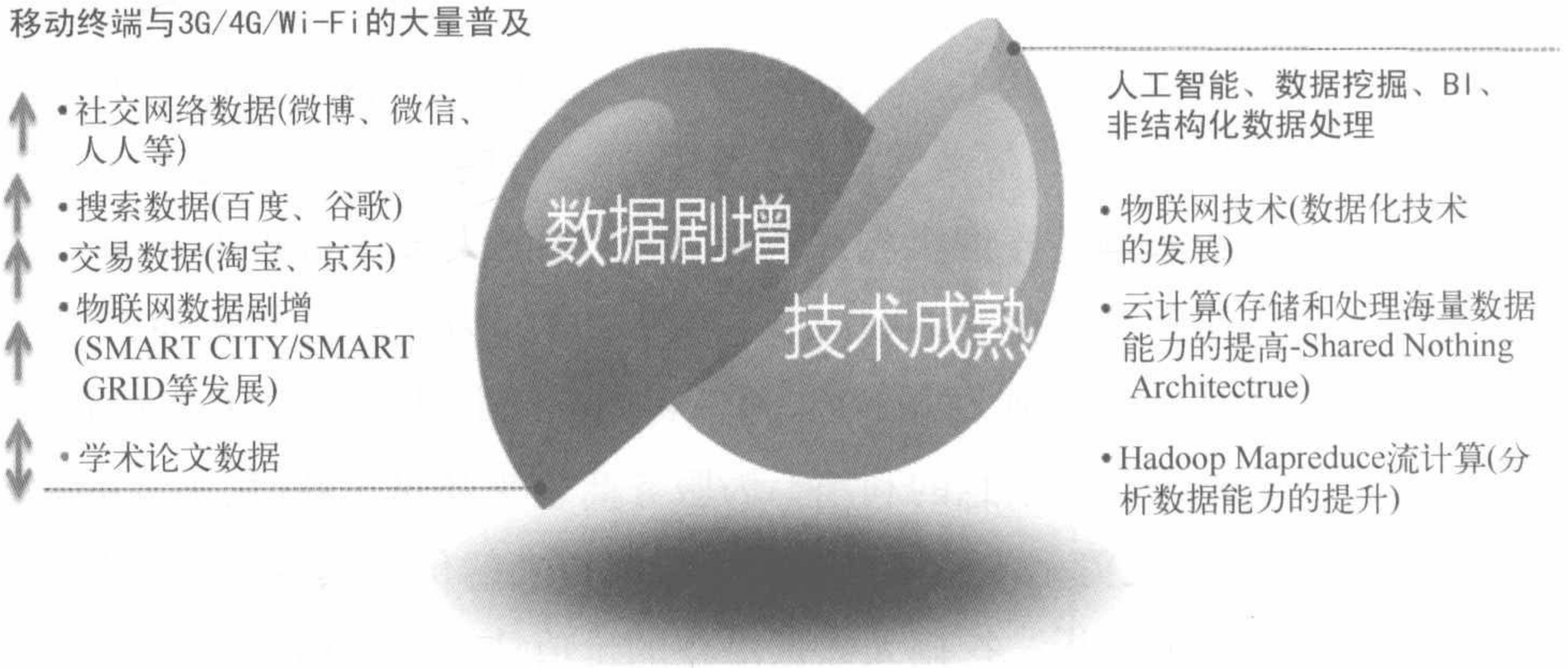


图1-2　大数据形成的本质

首先是数据的剧增。随着移动终端的普及与网络大面积覆盖，社交网络数据、搜索数据、网上交易数据、物联网数据飞速增长，使得大数据的分析有了最基本的条件；其次是大数据相关的各方面技术日趋成熟，比如数据化技术、存储和处理海量数据的能力、分析海量数据和实时数据的能力等，这些技术在近几年突飞猛进，为大数据时代的到来起到了决定性的作用。

虽然“大数据”这个概念没有统一的界定，但从科技情报的角度思考，大数据的目的其实就是为了在海量的结构化、非结构化数据中挖掘出有价值的科技情报，为各方所用，最终推动科技的进步、国家的发展。这些庞大的数据只不过是我们需要的情报信息的一层表面，真正隐藏在表面之下的情报信息是我们目前所无法估量的。正如钱学森同志将科技情报定义为激活的知识，那么大数据就是激活这个知识的催化剂。

因此，以科技情报工作的视角去审视大数据，可将大数据界定为5V特征：Volume(数据海量化)、Variety(形式多样化)、Velocity(产生快速化)、Visualization(可视化)、Value(最大价值化)，如图1-3所示。

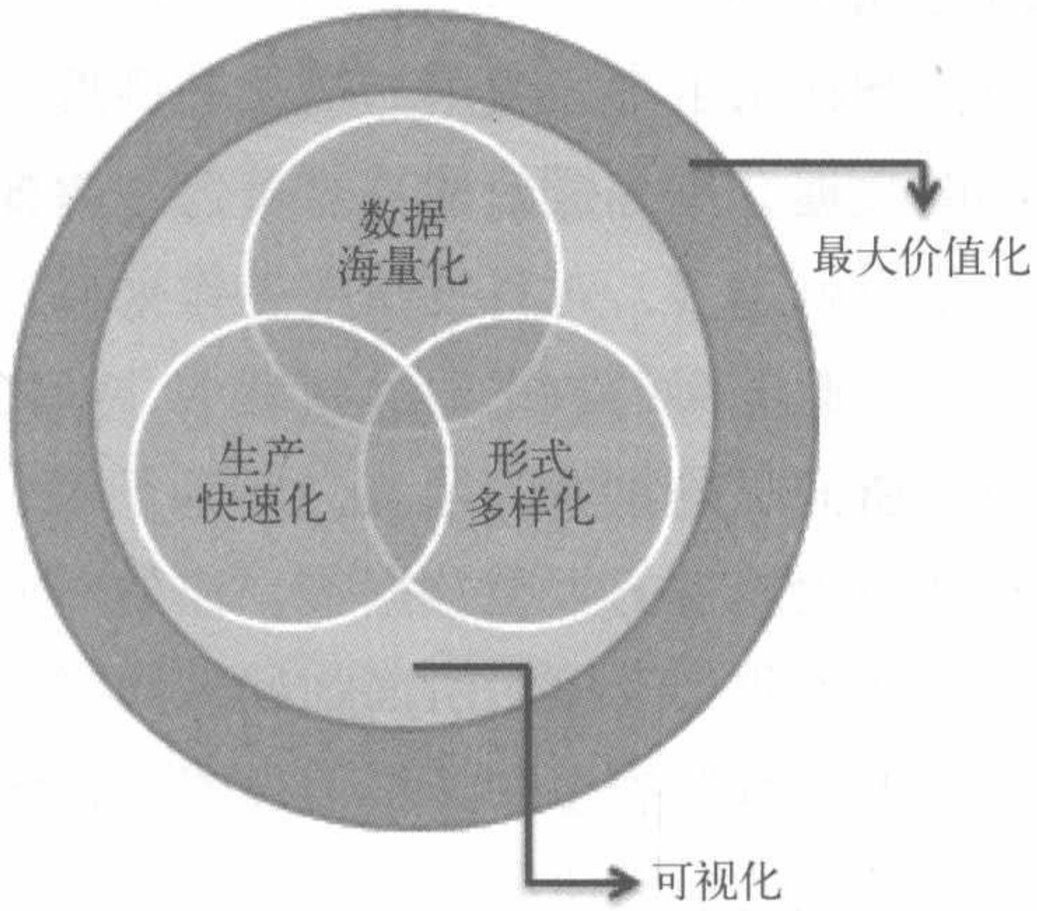


图 1 -3　情报研究领域的大数据 5V 特征

大数据时代之前的情报工作可以简单地概括为资源的获取、技术手段的应用、最终提供情报判读的服务。在大数据时代，这条情报工作的思路并没有彻底改变，而是将资源替换成了具有 3V(Volume、Variety、Velocity)特征的大数据资源，通过分析海量大数据，并进行数据描述的可视化(Visualization)，最终判读情报，提供服务，使大数据的价值最大化(Value)。换言之，如果把大数据作为一种产业，那么这种产业实现盈利的关键，在于提高对大数据的“可视化加工能力”，通过“可视化加工”实现数据的“增值”。

如何利用这些大数据产生价值，是大数据的主题；如何利用这些大数据及时准确地产生情报价值，是大数据时代科技情报的主题。大数据时代，信息的飞速产生，使得科技情报在科技发展中占据主要位置，科技情报是针对特定目标进行跟踪、监测、分析和展示的一系列活动，其信息源就是大数据。对于科技情报来说，大数据是一个机遇、一次革新，是一种获取情报、解读情报、预测预警的新方式。探索大数据所隐含的情报将成为中国产业提升、科技发展的一个重要手段。

1.2　大数据为科技情报可视化带来的机遇

大数据时代下，经济、社会的飞速发展所产生的海量情报数据，在数量上呈现指数增长的态势，在内容上不断细化，在学科上不断交叉渗透和汇聚融合，并随着互联网等信息传输工具的大量普及而得以更加快速地在世界范围内传播。

“大数据”理念悄然兴起，正催生社会、技术、科学和经济的变革，同时对情报服务的理念、模式、方法和技术带来了崭新的机遇和挑战。当今世界上最大的数据库就是存在于互联网的数据，也是最新的甚至可以称为即时性的数据。数据越庞大，所能透露的信息量也就越多，对科技情报工作来说，是大数据时代新一轮的机遇与挑战。但同时，在此庞大的数据库的基础上，如何对有效情报进行收集并应用，是目前情报机构所面临的重大问题。

面对大数据时代复杂或大规模的异型数据集，比如产业分析、专利分布、科研人员研究行为数据等，数据收集及应用所面临处理的状况变得更加复杂。而现代科学的发展不仅要科研人员掌握静态的情报信息资源，而且要掌握多维的情报信息资源，及时了解学科领域的最新研究成果和研究动态，以及本学科的发展方向、前景、今后的预测，等等。科技情报的可视化在大数据环境下通过更合理的组织和展示，把繁杂的数据转化为人们更容易且更能迅速探索的形式。这就是情报可视化的核心任务。

情报可视化由两个重要的部分构成。首先是实现数据描述必需的支撑性技术基础，其次就是情报理解。如果对情报搜集任务和情报可视化用户目标没有深刻的认识，很难开发出实际可用的可视化系统。要想开发出有效的科技情报可视化系统，就必须对情报感知和认知有着深刻的理解。

大数据时代下情报可视化所面临的机遇主要有：

(1)大数据时代，科技相关数据由数字化向数据化转变，为科技检索可视化提供了基础

随着目前电子图书馆和各种知识数据库的发展，越来越多的期刊和书籍，通过 PDF 等格式，转变成数字形态存入这些数据库中。但是对这些数字化的资料进行查询分析并不是十分方便。首先要知道所需资料在哪本书中，其次还要仔细翻阅这些数字化的资源，以便找到所需要的信息。这种方法和在书本中找没有本质的区别。

若将这些文本相关信息数据化，文本中的字、词、段落、作者信息、关键词和其他相关信息都能够一一被识别，利用竞争情报辅助搜索引擎加以检索就会方便很多。所谓数据化就是将一种现象转化为可以制表分析的可量化的过程。量化，是数据化的核心，也是表述情报的可视化前提。信息只有被数据化，其巨大的潜在情报价值才有可能被释放出来。

另外，以往的科技情报工作在定量分析的时候，最常应用的是分析结构化数据库的文献计量学方法；而在大数据时代，除了通过对万方数据、NSTL 科技文

献数据库、CNKI 数据库、EI、Derwen 专利数据库、维普科技期刊数据库等结构化数据的文献计量分析之外，非结构化数据的情报分析开始进入我们的研究视野。大数据提供了大量实时信息的基础，通过技术手段可视化后，就可得到有用的情报信息。

图 1 -4 所示为发生在 2013 年年初的 H7N9 公共卫生事件的可视化分析图，这些数据的来源主要是新闻媒体报道，官方微博发布的非结构化数据。像这样将多个来源、不同格式的数据进行统一格式的数据化后，整合分析，并可视化，使得用户可以非常明确地看到 H7N9 发生的重灾区，再通过饼图颜色的不同，可了解各个地方 H7N9 的具体发展事态，从而可以很快作出相关决策，大大提高决策的效率与准确度。

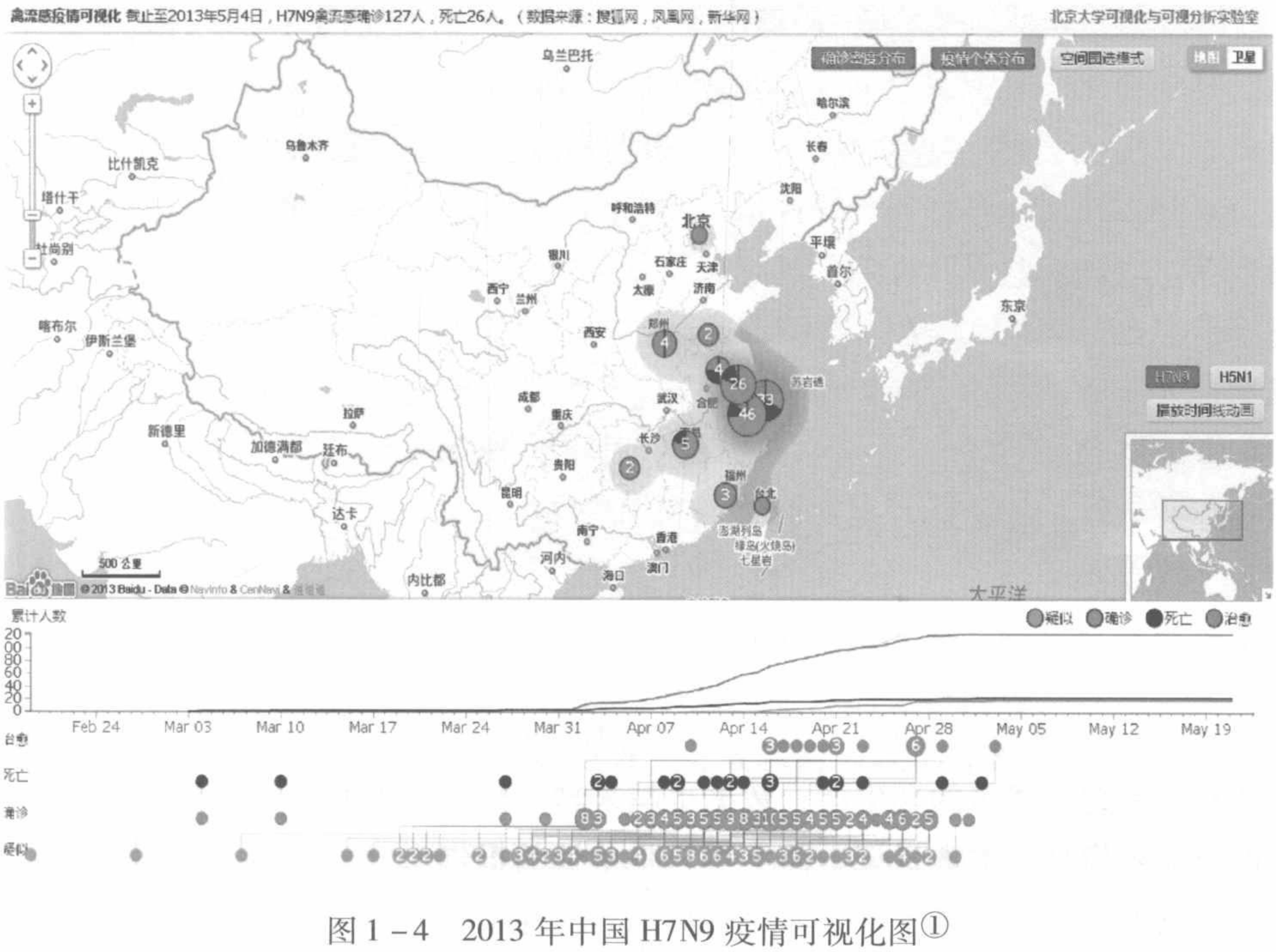


图 1 -4　2013 年中国 H7N9 疫情可视化图①

(2)大数据时代使人的行为量化变为可能，为发现和把握社会主体行为的可视化提供了基础

如今的人们已经离不开手机、电脑、智能电视等智能终端设备，其日常生活

① 北京大学机器感知与智能教育部重点实验；http：//vis. pku. edu. cn/birdflu/

基本上都可以数字化地表示。人们的一举一动都产生了大量的数据。任何行为，皆有前兆。互联网恰恰保留了大量前兆性的数据，通过对这些数据的收集和分析，可以预判物理世界中人类的未来行为。

Nicholas Felton[①]因其个人年度报告而成为“大数据量化生活”领域的知名人物。这些个人年度报告彰显了 Felton 的设计天赋和在个人数据收集上的严谨性。除了地理位置，他还追踪了每年他相处的人、吃饭的地方、看的电影、读的书籍以及他大量的信息。图 1 -5 是 Felton2012 年的个人年度报告中的其中一页。

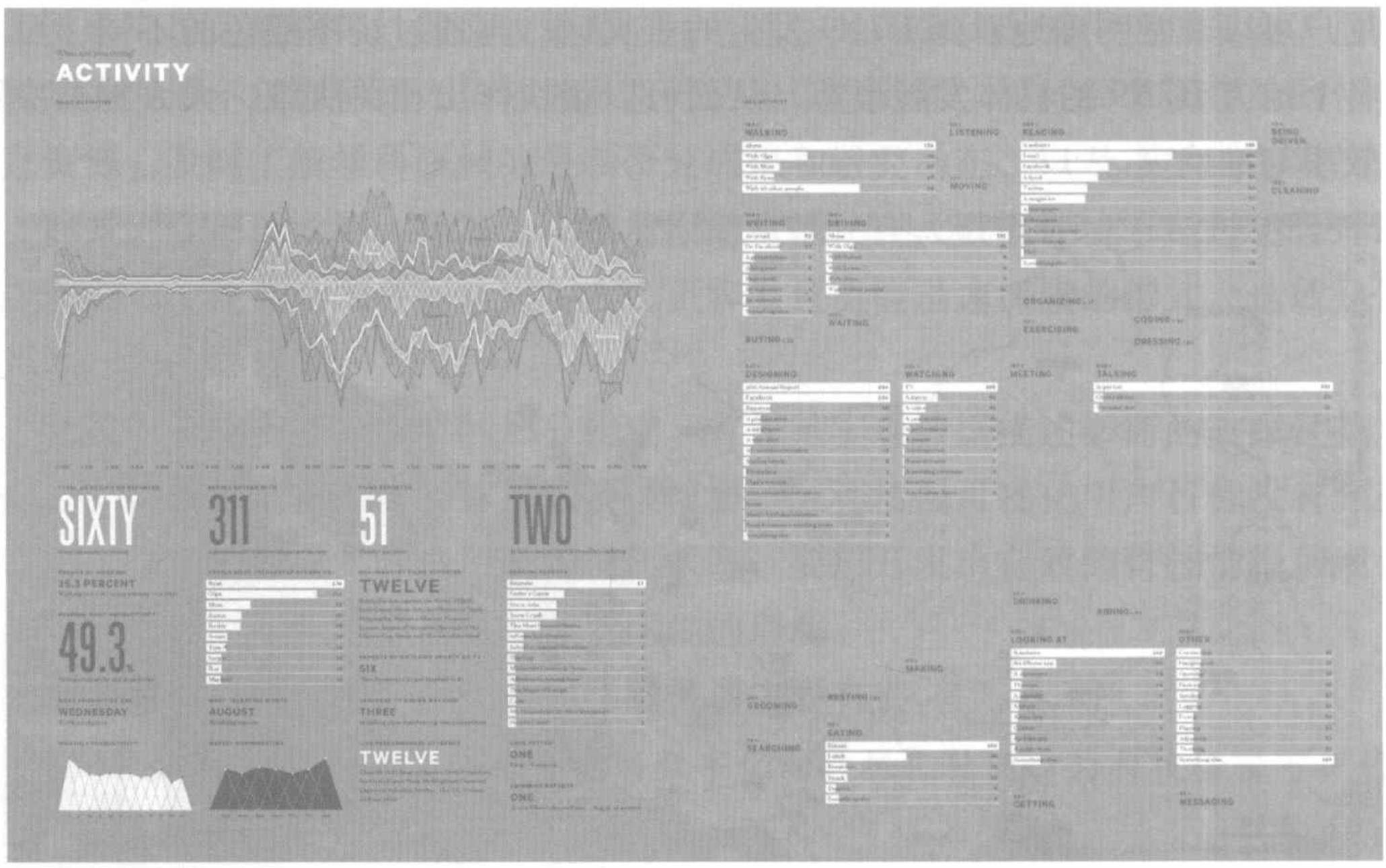


图 1 -5　Felton2012 年的个人年度报告中的其中一页[②]

从 2005 年至今，Felton 已经设计了八本年度报告。值得关注的是，他的报告逐渐暴露了越来越多的私生活，数据也越来越丰富。随着时间的推移，这些数据变得越来越像是一本本日记，而不再只是单纯的报告。

虽然大数据分析基于各人习惯进行预测的准确度确实会因人而异，但总体上其精确度比我们想象的要高。由于人类从来没有像今天这样如此依赖网络和电子设备，因此，大数据时代产生如此多的电子踪迹让研究每个人、每个群体，甚至整个人类的习惯成为了可能。这些可能性和现实性为科技情报事业的发展提供了

① Nicholas Felton 是 Daytum. com 的创始人之一，目前是 Facebook 产品设计团队成员。
② http：//feltron. com