社会化媒体

影视传播

广告

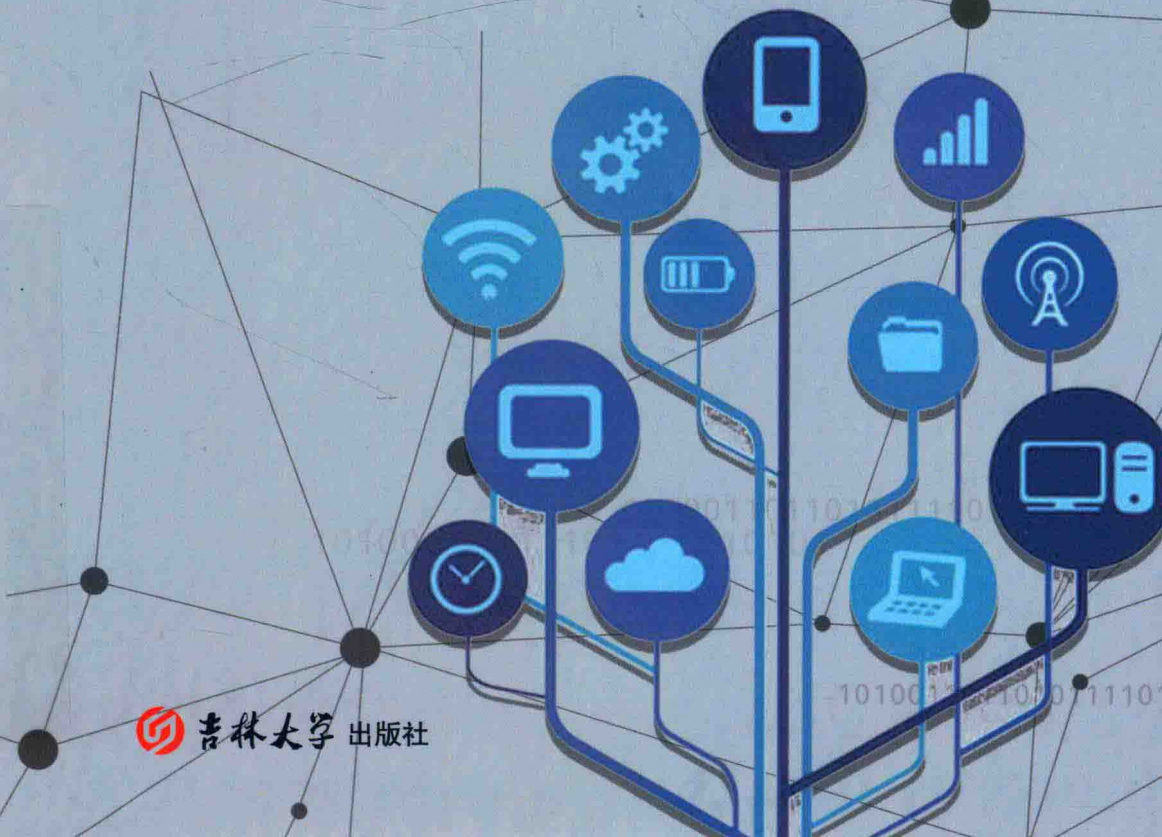
出版业

社会舆情

大数据

时代新闻业态研究

宋 宇 著



吉林大学出版社

大数据

时代新闻业态研究

宋 宇 著

吉林大学出版社

图书在版编目(CIP)数据

大数据时代新闻业态研究 / 宋宇著. — 长春 : 吉林大学出版社, 2018.12
ISBN 978-7-5692-4061-0

Ⅰ. ①大… Ⅱ. ①宋… Ⅲ. ①新闻工作—研究 Ⅳ. ① G21

中国版本图书馆 CIP 数据核字 (2019) 第 008286 号

书　　名：大数据时代新闻业态研究
DASHUJU SHIDAI XINWEN YETAI YANJIU

作　　者：宋　宇　著
策划编辑：邵宇彤
责任编辑：邵宇彤
责任校对：韩　松
装帧设计：优盛文化
出版发行：吉林大学出版社
社　　址：长春市人民大街 4059 号
邮政编码：130021
发行电话：0431-89580028/29/21
网　　址：http://www.jlup.com.cn
电子邮箱：jdcbs@jlu.edu.cn
印　　刷：三河市华晨印务有限公司
开　　本：170mm × 240mm　1/16
印　　张：13.75
字　　数：250 千字
版　　次：2019 年 3 月第 1 版
印　　次：2019 年 3 月第 1 次
书　　号：ISBN 978-7-5692-4061-0
定　　价：49.00 元

前　言

随着社会经济的快速发展，世界正向大数据时代加速推进。大数据是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，需要用新处理模式，唯此才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。大数据的来临，给人们的思维方式带来了巨大的变化，深刻地影响了人们的工作和生活，它已成为社会各界及国家政府持续关注的热点。

在大数据时代，新闻业态也产生了极大的变化。大数据时代的到来，颠覆了传统新闻的表现形式，通过手机、电脑、网络等载体对新闻事件进行直播或动态发布等各种新形式，给人们更多的视觉和听觉感受。大数据时代，传播实现了可视化、便捷化，新闻内容也变得动态化、具体化，新闻传播也加快了数字化建设，电子化阅读越来越受到人们的喜爱，这对新闻行业来说是一个全新的挑战和应用。

笔者一直关注新媒体下新闻的发展及教学方面的一些问题，发现新闻业态要想达到预期的效果，必须重视与现实科技发展的结合。落实到现实中便是与大数据的结合，而目前国内对大数据与新闻学的研究，大多集中在网络新闻中。本书结合前期笔者相关的研究论文，以及现存多种新闻业态的发展现状，切实讨论了大数据时代下主要新闻业态的发展问题，是这方面鲜有的系统性论著。

本书立足大数据的基本理论，结合当前我国新闻事业的发展概况，分析了大数据时代我国新闻业态的发展现实。同时，从社会化媒体、影视传播、广告、出版业和社会舆情等角度研究了大数据时代新闻业态的发展，希望能为未来我国新闻事业的进步提供新的研究思路。

本书在写作过程中得到了大量专家教授的帮助，在此表示感谢。由于时间及作者水平所限，不足之处在所难免，恳请读者批评指正。

目 录

第一章　大数据概述

第一节　大数据的内涵

一、大数据的定义

基因学、天文学，这两个最先经历信息爆炸的学科，是大数据概念最初的发源地，这个概念如今已经广泛应用到人类社会的发展之中。

大数据这个概念并未被准确地定义，它刚被提出时，由于信息量过大，一般电脑的内存不足以支撑这些信息数据的处理，所以工程师们对此工具进行了改造。

大数据包括数据的两个方面：一是数据的信息量，二是对数据处理的速度。大数据在数据分析这一领域，一般处在前端位置上。IT 行业中的新兴词汇数量较多，数据分析、数据挖掘在其中比较明显，数据仓库与数据安全亦较为流行。其商业价值已被各大商业人士互相争抢。大数据概括了以下种类：交通工具、工业器材以及生产设备上安装的传感、互联网信息等，这些都可以实现数据随时随地的测量。新的处理模式使大数据的洞察力和决策力大大加强，流程的优化更加容易实现，增长率也逐渐提高，并能够处理更大的数据信息量。最终，不同种类中有价值的信息都可以从大数据中进行迅速提取。

大数据技术随着网络、服务器和传感器等设施发展，使各大企业根据需求实现自身的经济效益、商业价值和社会价值等。各行业在使用大数据的基础上，获得极大效益，展现出极强的社会能力，而绝非仅由数据所得。因此，使用者可以在合理的时间内使用大数据采取大量信息，从而有效解决社会问题，这就是大数据的定义。

至今，在人们的观念里，大数据处理只能发生在大规模的数据基础上，通过大数据，人们可以获得新的知识，改变一些关系，从而创造出新的价值。

二、大数据的特征

（一）种类繁多，体量超大

互联网及微博、电子商务平台等社交网站的发展，衍生了大能量的数据信息。网络、传感、存储等计算机科学领域在全面发展，网络数据可以实现同时收集大量的数据信息，使人们用不同方法收集的数据量达到从未有过的水平，其包含医疗领域的临床数据、科学研究、电子商务、智能手机、传感器等，如基因组钻研，可以将 GB 级以及 TB 级数据传送到数据库。由于非结构数据的增加速度，总数据的增长速度甚至比结构数据的速度快几十倍，非结构化数据占总数据量 85 % 以上。对于发展网络企业和存储的商业人士，这样的预测可以增强信心。

美国咨询企业麦肯锡从单层数据集合定义大数据，如不容易通过数据库软件不容易分析、采集、储存的巨大数据集。大数据最有价值之处在于随着数据的类型日益增多，如文字、图片、符号、视频等，可以发现这些不同数据流间的相关信息。例如，可以将交通状况和供水系统数据作比较，会发现早高峰时间和清晨洗浴紧紧相连，交通事故与司机休息质量、电网运行数据、堵车时间有着地点相关性。

（二）开放公开，容易获得

运用大数据不是为了储存，而是为了方便分析。特定的企业组织和政府机构不仅存在于大数据中，也存在于生活中，如各大运营商积存顾客的通话记录，各大商务网站整理消费者的各类信息。发掘大量的数据，可以使各大企业改善运营服务，提高自身能力，提供决策，商业能力的实现为各大企业带来高经济回报，进而发现了各个大企业发展的不一样的规律。今天，在有规则的开放下，越来越多的政府机构和商业组织爬虫采集技术和应用接口技术开始向研究所和社会各大企业提供自己储存采集的各类大量数据源，特别是美国政府排在前面，积极提供权威的数据源，例如“数据·政府”等其他的开放数据。国内外组织从微博上采集大量信息，从而分析每个人的标签属性和特点，预测商机、企业票房和社会情况。大数据的基本特征，开放公开是易得的数据源，并且产生了超大的社会影响。

（三）重视社会预测

预测是大数据的本质特征。在大数据时期，在将来的发展中预见能力变成了

企业追逐的方向。Netflix 企业发布《纸牌屋》，就是采集 3 000 万客户的播放动作，包括快进、暂停、倒退、打开等，分析百万注册用户的搜索和评级，评论用户对各类不一样电影电视节目给出的不同观念，通过发掘大量的数据，从类型、导演、演员、题材、情节等不同方面了解观众欣赏的节目习惯。这个公司通过微小细节的采集和客户数据的分析，变换了视频行业的制作方法，用计算和逻辑分析方法替代了旧的制造方法，制作剧情要先分析大数据的客户，以此来判断是否可以获得观众。更有趣的例子是，商家通过大数据中客户购物方式，分析出女性怀孕的可能性，一个未成年女儿的怀孕信息，商场比父亲更早知道。人们更为关注大数据预测社会问题的应用能力，大数据会在科学社会区域凸显出更大的作用。

（四）注重发现而非实证

实证研究强调建立理论假设，在范围时间内随机抽取采集数据，收集相关数据信息来验证逻辑、理论假设。大数据重视数据，依靠自上而下的数据采集处理，在不依赖理论假设的情况下发现知识、洞察趋势、预知未来、创造知识、找到规律。大量的交易数据还可以通过大数据的技术进行分析。如周六、日在沃尔玛超市买婴儿纸尿裤的男人会捎带购买啤酒。大数据探究不做所有假设，富有不可知的能力，结果最有用且实用。大数据忽视抽样，重视群体。大数据是存在有信息技术自主采集储存的大量数据，能迅速分析处理得到结果，从而使存储设备成本不断下滑，计算机工具效能越来越发达，处理巨大数据的能力迅速提高，数据探究算法快速改进，尤其是机器学习的神经网络建模技术使抽样查询不再是独有的方法，总体数据在大数据理论上进行把握，更关注整体的数据。

（五）非结构化数据的涌现

有效信息和未知的实用知识被数据探究关注，最多的是非结构化数据，这成为大数据的突出特点。迄今为止，超过 90 % 的数据都是非结构化数据。社会媒体尤其是微博每一刻产生的无限数据文本信息，导致有价值的数据躲藏在巨大的信息中，大数据的分辨能力从大量文本中探究人们的态度和行为，发现了社会需求和重要商机。非结构化的大量数据的被采集处理，使技术发生了新的变化，社会产生了不一样的需要，各种非关系型数据库如 NoSQL、Hadoop 集群及 MapReduce 等很流行，新的 IT 技术也不断涌现。出现了网络挖掘、数据挖掘、文本挖掘、NLP 自然语言处理、机器学习等商业智能技术信息、策划支持和 IT 及社会科学区域的使用。

三、大数据的分类

根据信息来源可将大数据分为以下几大类：科研数据、互联网数据、感知数据、企业数据、政府的大数据、企业的大数据、个人的大数据。

（一）科研数据

科研数据在大数据时代前就存在，来自于生物工程、天文望远镜或粒子对撞机等不同方面。这些数据存于封存系统内，使用者最初全部是高性能计算（HPc）企业，许多大数据源自 HPc。

科研机构拥有最大计算速度且性能优越的机器，包含生物工程和天文学望远镜或粒子对撞机。如欧洲的国际核子研究所装备的大型强子对撞机，在满能量的工作下每分每秒的速度就能产生 PB 级的数据。

（二）互联网数据

互联网数据在大数据时代是主流。近年来，社交媒体的发展是大数据的主要来源。比如，以搜索位于互联网企业前列的百度和谷歌的数据模式都已经达到千 PB 的级别，影响最大、应用广泛的阿里巴巴、亚马逊、雅虎、脸谱的数据都突破上百 PB。互联网数据的上升动力有两个：一是扎克伯格来回引用的信息分享理论：个人分享的信息每一两年翻番；二是梅特卡夫定律（互联网企业和用户数的平方成正比）。

大型互联网企业的大数据生态系统很突出，其一是维护自己的生态系统，其二是各方面不同程度上参与开源，甚至连硬件都依赖自己。大型互联网企业不单自己产生大量的数据，另外还有平台联动的影响，如 Facebook 到 Zynga，阿里巴巴起头做的数据交换平台。中型互联网企业也能维持大数据技术团体，但是与大型互联网企业核心开发能力和贡献能力相比，中型企业更注重在外围开发、优化和运维。但是它们也有自己的好方法，如豆瓣的推崇、暴风 Hadoop 管理。小型互联网企业有数据但是没有大数据的能力，所以产生了一些大数据技术和服务的机会，如百分点为电商做有个性的推荐和营销手段，各类营销服务、广告联盟、移动应用平台及贡献统计分等。

（三）感知数据

互联网时代有感知数据加入后，移动平台的 LBS 的普及和移动平台的感知功能，互联网的数据和感知数据渐渐增加，感知数据的惊人能力仅次于社交媒体。

Teradata 曾预测，感知数据的总量会超过社交媒体，并达到后者的 10 ~20 倍。

（四）企业数据

企业数据类别繁杂，感知数据和企业数据在理论上也没有不重复、不遗漏的区分，公司也同样可以依据互联网收集海量的感知数据实现快速增长。因为在传统的理解上，工商所数据是人产生的，所以划分为两类；传感器、标识物等是感知数据所产生的。公司外部数据每天吸取社会媒体数据，内部数据不仅有结构化数据，最多的是非结构化数据，如早期的电子邮件、文本等发展到感知数据和社会媒体阶段，包括各式各样的模拟信号、音频、图片、视频等。

传统产业发展中，企业数据和感知数据都涉及，能将它们放在一起讨论，在经济总量上比互联网产业大很多，因为传统产业自己的大数据能力有限，所以是大数据技术和服务行业的重要目标市场。但对单个企业来讲，拥有大数据需求的不是很多。比如，麦肯锡的报告中显示出制造业为大数据数量最大的行业，但制造业很少开展大数据项目。即使有也是在市场营销上增加了互联网的招数，取得来自终端的需求，供应链和生产方面相比大数据之前没有太大变化，如 Zara。工业互联网通过数据采集和分析来提升制造业的工作效率，在这方面有最大的市场，但并不是大数据。

互联网的大数据不便于分类。阿里巴巴根据商业价值区分为信用数据、交易数据、社交数据和移动数据；百度把数据区分为用户搜索产生的需要数据和通过公共网络取得的数据；腾讯喜欢探究用户关系数据，并且在这一用户关系数据基础上生出社交数据。人们的很多行为和想法可以通过数据进行分析，从而发现社会行为、文化活动、政治治理、身体健康、商业发展等区域的各类信息，进而分析未来。互联网的大数据能分为互联网金融数据、用户消费产生的行为、地理位置、社交等。从各个社会角度根据使用的重要成分来分为三类：企业的大数据、个人的大数据、政府的大数据。

（五）政府的大数据

各个不同级别的政府部门都拥有海量的原始数据，促成社会发展和运行的基础，包括各式各样的环保、电力、气象等生活数据，道路住房交通、自来水等公共数据，海关、旅游、安全等管理类数据，信用、医疗、教育及金融等服务数据。在完整的政府单个部门里面，无数数据整体化并没有产生任何价值，如果将这些数据连接在一起作整体分析并有条理地管理，将会产生超大的经济效益和社会价值。

现代城市依据网络智能向智慧化发展，不论智能电网、智慧医疗，还是智能交通、智慧环保，都不可以离开大数据的支撑。智慧城市的主要核心就是大数据。在建设智慧城市的过程中，大数据能在各方面提供决策和智力支撑。政府可以将数据渐渐开放，可以让更有能力的机构组织或个人来分析利用，可以更快地造福人类。奥巴马任职期内的一个最重要举措，就是使美国的政府机构创建了 data.gov 网站，且要求政府公开，主要就是政府机构的开放。至今已开放了上万个数据库。

（六）企业的大数据

公司离不开大数据的有效决策和支持。企业为了对快速消费者群体提供不一样的产品和服务，需要依据大数据的帮助实现准确营销。网络企业根据大数据达到服务升级和方向转型。传统公司面对到处都有的互联网压力，一定要实现融合、寻求发展，不停前进。

随着信息技术的走向，大数据已然是各行各业的主要资产和基础要素，数据成为产业并成长为供应链方式，互联网时代，互相自由连接的外部数据的重要性渐渐超越单一的内部网络数据，公司个体的内部数据更不容易和全部互联网数据并列。综合提供数据、整理数据加工、推动数据应用的新型企业明显有竞争趋势。

（七）个人的大数据

不同人都能使用互联网创造属于个人的信息中心，储存、记录、积累、采集个人的全部大数据信息。依据有关的法律规定，凡是通过本人亲自授予，全部的个人信息就会转换成有价值的数据，会被第三方采集，经过处理，获得有个性的数据服务。经过信息技术能够获得不一样的可穿戴设备，包含带进的各种芯片信息都可以经过感知技术获得单人的大数据，还可以包含视力、心率、体温等身体数据及地理位置、购物获得、社会关系等数据。能将个人的身体数据授权给医疗服务，以便检查身体情况，制定私密的健康计划；还能把个人的理财数据授权给最专业的金融理财专家，便于定制适合自己的理财计划并检测收益。同时国家相关部门会在法律允许范围内经过严谨的程序流程进行防御监控，每时每刻监控公共安全，防止犯罪。

每个人的大数据通过法律保护，剩余的第三方机构要遵照法律制度授权使用，数据必须要接受全方面监督、公开透明。采集个人数据应该明确依据国家法定规程，由用户自己决策采集内容和范围。数据只能由本人确定使用后才可以严肃处理。

第二节　大数据的发展历程

一、数据的开始

人类在劳动中发明了图形、文字和语言，但只用这些还不能准确地描述世界，数字作为一项重要的改造世界的工具而产生了。它把抽象的概念具体表达，如“很多”人、“非常”多人可以理解为不同的程度，但如果说 1 000 人、10 000 人就清清楚楚了。物质的交换、人们的生产等活动全部是以数据为基础发展的，如货币、度量衡等的出现和发明，极大地促进了人类文明的发展。

数据的测算产生了最初“有依据的数字”，就是数据是对客观世界测算结果的记录，不是随便发生的。测算从一开始就是为科学服务的，从以前到现在，测量都是科学的重要手段，其重要性可以解释为：没有测量，就不会有科学。测量来的数据可以由计算再衍生出新数据，这样看来，一切数据都是人为的产物。但这时的数据还只是传统意义上的数据，它和信息、知识是有严格区别的：信息是数据的背景条件，数据是信息的原体，知识是通过处理后展现出来的有原则的信息。

步入信息时代后，巨大的变化产生了。20 世纪 60 年代，数据库的发明，软件科学的发展，电脑的数据库用来存储一切文本、图片、数字。这时，数据开始不单指“有根据的数字”，其内涵发展到所有存在于电脑中的文件，包含视频、文本、图片等。数据也发展成信息的名词，因为这些信息只是一种对世界的记录，数据因此多了一个来源：记录。数据库出现以后，信息总量与日俱增，增速也越来越快。20 世纪 90 年代初，就有美国人提出了“大数据”概念，这个时候还不是真正的大数据时代，数据的重要性在上升，在价值上的重要性已经被预见。21 世纪开始，特别是 2004 年新社交媒体产生以后，数据开始爆炸，大数据的提法又一次出现，这时的大数据既指容量大，又指价值大。争议开始了：到底什么算大？多大才是真正的大？

二、大数据的开始

有史以来，处理各种不断增长的数据成了人类社会的难关。

前期大数据的现代发展历史能追溯到美国的统计学家赫尔曼·霍尔瑞斯。赫

尔曼·霍尔瑞斯，被后世称为“数据自动处理之父”。他创造了一台“电动打孔卡片制表机”，对卡片指定位置上的孔洞进行识别，且可以另外进行自动统计。这一创造被用在1890年的人口调查数据统计中，这个机器用两年多的时间就完成了预算耗时13年的工作，这个骇人的速度就是全球数据自动处理的新起点。

1943年“二战”期间，在英国，为粉碎纳粹的扰乱，其领导人决定借用第一台能够实现编程的电子计算机对大规模数据进行处理。“巨人”也就应运而生。为寻出一种拦截信息中的潜在模式，其读取纸卡的速度上升至每秒钟5000个字符，将原计划的数周耗时缩短至几个小时完成。

美国国家安全局是一个刚刚成立9年，并且拥有12 000个密码学家的情报机构，在1961年这个处处盛行间谍的冷战时期，在超量信息压力下，该机构借助计算机收集信号并对各种情报进行处理，同时处理掉了堆积于仓库内的各种模拟磁盘信息，仅在当年的7月份一个月就收获了近2万卷磁带。

自1940年以来，人们就梦想能够建立一个世界范围的数据信息库。从而使用户能在此信息库中修改或者读取各种重要信息。20世纪60年代，英国计算机科学家蒂姆·伯纳斯·李发明了一个全球网络资源唯一认证的系统——统一资源标识符，在其设计的超文本系统中，任一有用的事物均被命名为“资源”，且由“统一资源标识符”进行标识，从而把超文本嫁接到因特网上，命名为万维网，并且经由超文本传输协议将资源传输给用户，用户在点击相应链接时得到所需的资源，从而让人们经由互联网在全世界实现信息共享。

15年后，戈登·摩尔，英特尔创始人，在研究计算机硬件的发展过程中发现摩尔定律。在此定律的说明中，面积相同的芯片每1至2年便能在数量上多容纳1倍的晶体管，微处理器的性能亦能够提升2倍，价格也能便宜近半。在近半个世纪以来，硬件的发展基本符合这一定律。到今天，一根头发尖大小的地方就能放上万个晶体管。后来，英特尔公司又发明了22纳米的3D晶体管，比以前的晶体管小了大约1/3，摩尔定律的生命进一步得到延续，从而使信息产品的功能日渐完善，而设备的体积也逐渐变小，甚至存储器的成本持续节约了1亿倍还多，实现了低成本存储海量数据的梦想。摩尔定律已经成为描述一切呈指数级增长的事物代名词，这为大数据时代的到来铺平了硬件道路，打下了物质基础。

除了便宜、功能强大，摩尔定律使计算设备也变得越来越小。马克·伟泽是一名美国科学家，在1988年提出了普适计算，即各类计算机可以在任何时候任何场所获得数据并对其进行计算处理。在此理论中指出了计算机发明所需经历的三个阶段：一是主机型阶段，一台占据大半个房间的大型机器被很多人共享；二是个人电脑阶段，每个人拥有一台变小了的个人电脑；三是普适计算阶段，在计算

机逐渐从人们视线中消失的过程中，各种微小计算设备被广泛配置于日常的生活环境中，并随时随地获取处理数据，最后再融入周围环境中。现今的各种流行设备或工具，如各种传感器、可穿戴式设备、RFID（射频识别）标签、智能手机等都达到了随时随地自动采集数据的能力，大大加强了人类采集数据的能力，这就是大数据时代来临前的物理基础准备。

在这名科学家提出普适计算的后一年，第一届数据挖掘学术年会由英国计算机协会下属的数据挖掘及知识发现专委会成功举办，期间伴随着相关期刊的出版等，这是大数据时代的一个最重要的里程碑。而后，数据挖掘的发展如火如荼。所谓数据挖掘，是指为找寻数据中暗含的趋势或规律，对大量数据使用特定算法进行自动分析计算，从而供决策者参考。而数据挖掘进步的根本原因在于人类模式识别算法的逐步强大，这也是大数据时代技术基础最集中的体现。现代网络信息中，各种数据可以记录某件产品的流向及消费者的各种情况，运用数据挖掘，为客户量身打造，使服务和消费完美结合，实现质的飞跃。这种数据的挖掘借助于网络，成本较小，同时不必逐一进行调查，也不需要制作问卷，最重要的是其分析的实时性。因此，这些特性使数据挖掘取代抽样技术，成为分析预测中地位显著的工具。总而言之，其优越性主要体现在“实时、量大、多源”这三个特点上。

在大数据中，机器学习是其热点和前沿，相较于数据挖掘，这种算法并不固定，它含有自调适参数，能使其在逐步挖掘计算过程中自动调整算法参数，使结果更加精确。“机器学习”，顾名思义，即提供给机器大量数据，使其在逐步学习的过程中进行自我修改和完善。不仅机器学习、数据挖掘，数据的分析和使用技术也已经基本成熟，且自成谱系，如多维联机分析处理、数据仓库、内存分析、数据可视化等都是其体系的重要组成部分。

大卫·埃尔斯沃斯和迈克尔·考克斯是美国的研究员，在1997年就曾使用“大数据”这个新兴词汇，对超级计算机产生的超出主存储器的海量信息进行描述，使数据远远超出远程磁盘的承载能力。

2004 年之前，传播分享信息是互联网的主要作用，而静态网站的建立则是其主要的组织形式。自此年起，脸书网（Facebook）、推特（Twitter）等相关社交媒体争相出现，开启了互联网的时代。人们借此进行互动交流，拉开了新时代的序幕。这速度有多快呢？一个鲜明的实例可以表达这种时速：美国弗吉尼亚州在 2011 年 8 月 23 日发生 5.9 级地震，纽约市的居民首先看到的是推特上的消息发布，经过几秒后，地震波的震感才从地震中心传播而来。这是一个比地震波传播还快的社交媒体传播信息时代！

同时，此类社交媒体给全世界的网民提供了一个平台，使他们随时随地都可以记录自己的行为想法，这种记录其实就是贡献数据。由于全世界网民对数据数量上的巨大贡献，使人类历史上出现了最为庞大的数据爆炸。李塔鲁是乔治敦大学的一名教授，其在2012年考察了推特上产生的数据量后估算出一个惊人的数字：在过去的50年间，《纽约时报》所产生的单词信息量大约30亿个，而现在仅一天的单词信息量足以达到过去的1.3倍。换种说法即是：现今的数据总量大约等于《纽约时报》过去一个世纪所产生的数据总量。

从第一台计算机问世并踏入信息时代起，直至社交媒体产生之前，产生和收集数据的工具一般都是信息系统或者传感器，但社交媒体改变了这一现状。人们通过互联网，使用微博、微信或者推特记录并表达自己的活动与心情，以此产生“行为数据”。这种“行为数据”大都是非结构化的，缺乏严谨的结构，因此十分难于处理。到目前为止，大约75％都是非结构化数据。现在的大数据＝结构化数据＋非结构化数据，推动人类进入大数据时代的真正理由是人们对于数据的使用能力提升和突破，主要体现在数据挖掘上。近年来，数据挖掘的应用还在不断推陈出新，达到了前所未有的高度。例如，凭借长久积累的用户资金记录，阿里巴巴等一些互联网公司踏入金融领域，并且借用这些记录在几分钟内决定用户的信用资质和发放贷款与否。又如，奈飞公司能够精准地进行营销，主要以其对用户网上点击率的统计判断并做出预测。自此之后，“预判发货”这项专利顺理成章地在2014年1月被美国电子零售巨头亚马逊提出。此专利规定了在客户想下订单网购之前包裹已被寄出。这项专利的核心技术依然依赖于数据挖掘，这些数据源于客户日常的搜索、消费记录或者心愿单，更甚者能够从客户鼠标在某项商品页面的停留时间预测判断客户的需求爱好。专利的本质概括来说就是发货“外包”的智能化实现过程，利用预测判断将发货的过程交给数据算法，从而实现自动发货。回顾以往，这个大数据现象的物理基础基本来源于1966年的摩尔定律，从此晶体管体积逐渐减小，成本逐渐降低，使人类有了可以承载海量数据的容器制造力。1989年兴起的数据挖掘技术是让大数据产生“大价值”的关键，因为大数据之大并不只是容量之大，更在于价值之大。而产生这“大容量”的主要根源则在于20世纪初社交媒体的逐渐流行，使全世界的美国人人无时无刻不在制造着数据。

三、大数据的未来

身处大数据时代，不仅人和社会领域的计算逐渐兴盛，物理环境领域的计算也将逐渐发展起来。这个领域的计算历史已经不短，而人和社会领域的计算才是大数据时代的关键，通过计算，大数据将为我们解决越来越多的社会问题。换言

之，大数据使社会也可以逐步实现计算。而这正是由于人们日常生活的状态将会被转换为数据进行记录存储，这种数据记录的密度和频度都将处于上升阶段，从而为社会领域的计算打下坚实的数据基础。

社会领域的计算也被很多学者称为“社会计算”。1990 年，从“社会软件”这个角度出发，美国的学者就曾经提出过这个概念。而早期的“社会软件”指的是能实现群体交流功能的软件，如 QQ、MSN 等。这种软件能大大减少人们互动交流的成本，从而使大规模合作成为可能。经过 4 年的发展，社交媒体应运而生。社会软件更是物尽其用，通过推特、微博和脸谱网等一些工具，人们可以尽情地记录下日常生活中的心情、思想及活动，甚至有学者直接将基于社交媒体的行为分析定义为“社会计算”。近年来，更多的学者加入大数据的使用和分析行业中，他们认为，通过以计算为特点的定量方法分析处理这些数据，未来的一切社会过程、问题和现象都应该并且能够解决，而且十分科学、精准。因为在他们看来，人和社会本身的数据足够，并且仍在不断增长。虽然这只是初步的定义，但说明其正在演变之中，而且并未达成国际共识，但是相关研究依旧可以继续开展。

借助计算，对社会问题的分析解决正在变成一种流行的趋势。2013 年的 7 月，在华东师范大学，一条温暖的短信被发送至一位女在校大学生手机上，言之“同学你好，发现你上个月餐饮消费较少，不知是否经济困难？”这个温暖的问候正是由数据挖掘而来。通过对学生饭卡的消费数据进行统计分析，发现该生每次餐饮消费较低，从而进行联想，发出温暖的关心。虽然只是这名女同学单纯地节食减肥，但这个美丽的误会则可以归咎于数据量不够充足。大数据不仅是数据，更要突出“量大”“多源”的特点，若加上该学生其他生活方面的数据就足以判断出其目前的减肥现实。同一年，在美国肯塔基大学也有类似大数据平台的使用，在此平台上对学生各种在校状况的数据整理分析计算，类似于各个课程的出勤情况、最终成绩、在线学习的活跃状况、图书馆使用情况等数据进行记录分析，确认可能存在问题的同学并对之进行辅导交流，帮助学生顺利完成毕业。

社会计算可以捕捉到人类历史上那些微妙的、精细的知识和关系，并将其转化为显性知识。麻省理工学院的布林约尔松教授就曾将大数据比作四个世纪之前的人类发明 —— 显微镜，因为它为人类社会带来了质的飞跃，将人类对物理环境的观察和测量精确到了“细胞”层面，也将成为我们进一步研究人类社会行为和自身行为的有力工具。

2008 年末，“计算社区联盟”提出了独特的详细报告 ——《大数据计算：在商务、科学和社会领域创建革命性突破》，报告指出人们要考虑的不仅是机器的数据，更要包括更广泛的领域，寻找更广泛的用途和更加新颖的简介。“社会计算”，