“大数据”时代如何做新闻？

文 | 陈昌凤

一年前“大数据”还是少数专业人士使用的概念，华尔街日报在2012年1月曾刊出物理学家和工学院院长合作的文章《科技变革即将引领新的繁荣》，声称：2012年1月，人类正处于三场宏大技术变革的开端，即“大数据”、智能制造和无线网络革命[①]。2月13日纽约时报网站即刊文《Age of Big Data》称，“大数据时代”已经来临。而就在2012年，“大数据”概念在中国已经普及至电子商务、经济战略、政治建设等各个领域。在美国，2012年3月29日奥巴马政府宣布投资2亿美元启动《“大数据”研究和发展计划》，希望增强收集海量数据、集中提取知识和观点的能力，加快在科学与工程中的步伐，加强国家安全，并改变教学研究。美国的大学开始培养新一代的“数据科学家”，数据分析也成为美国最热门的职业领域之一。

“大数据”与数据挖掘

“大数据”(Big Data, Massive Datasets)一词几年前开始出现，首先被世界IT大企业重视。“大数据”是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合，其主要特点是海量、非结构化和半结构化、实时处理，业界将其归纳为4个“V”：Volume(数据量大)，Variety(数据类型多样)，Velocity(处理速度快)，Value(价值密度低)。“大数据”首先是数据量大：过去常用的千字节(KB)，已经升级为兆(MB)和吉(GB)，甚至是太(TB)，乃至拍(PB)。这不是简单的数据增多，而是全新的问题，比如全球范围内的工业设备、汽车、电子仪表和装运箱中，都有着无数的数字传感器，这些传感器能测量和交流位置、运动、震动、温度和湿度等数据，甚至还能测量空气中的化学变化。[②]数据容量增长的速度大大超过了硬件技术的发展速度，引发了数据存储和处理的危机。

“大数据”浪潮成了全球政治、经济、文化、社会的变革之引，它成了加速企业创新、引领社会变革的利器。2012年1月在瑞士达沃斯世界经济论坛上，“大数据”是讨论的主题之一，论坛上发布的一份题为《“大数据”，大影响》(Big Data, Big Impact)的报告宣称，数据已经成为一种新的经济资产类别，就像货币或黄金一样。联合国推出了名为“全球脉动”(Global Pulse)的新项目，进行所谓的“情绪分析”，使用自然语言解密软件来对社交网站和文本消息中的信息作出分析，用来帮助预测某个指定地区的失业率、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，以阻止某个地区重新陷入贫困等困境，促进全球经济发展。

数据挖掘(Data Mining)，也称为网络挖掘(Web Mining)，斯坦福大学数年前就开设了一门课程“Web Mining”并出版了讲义《数据挖掘》(Mining of Massive Datasets)[③]。数据挖掘是“通过仔细分析大量数据来揭示有意义的新的关系、趋势和模式的过程。”[④]新闻界是数据的重要应用者，在互联网时代媒体经营、新闻实务等几乎一切都离不开“大数据”、数据挖掘。“大数据”时代大部分数据都是在自然环境下产生的，比如说网络言论、图片和视频等网民自发上传的内容，以及来自于传感器的数据等，即所谓的“非结构化数据”，通常不能为传统的数据库所用。因此从互联网时代非结构化数据的庞大宝库中获得知识和洞察力的计算机工具正在迅速发展，目前已经具备人工智能(AI)技术，比如

[大数据时代来临] 进入2012年，大数据（Big Data）一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。它已经上过纽约时报、华尔街日报的专栏封面，进入美国白宫官网的新闻，甚至被嗅觉灵敏的国金证券、国泰君安、银河证券等写进了投资推荐报告。

自然语言处理、模式识别和机器学习。

传媒运用数据挖掘：彭博案例

西方媒体对数据的运用越来越重视，出现了不少专门与数据打交道的记者，通过数据挖掘的方式进行新闻报道。他们在繁杂琐碎的非结构化数据之后，发现常规新闻中不能体现的逻辑，帮助读者对新闻事件进行深度解读。数据挖掘的新闻往往比传统新闻报道更有力度，也对记者提出了更高的要求。这里以彭博社一个数据挖掘类的报道栏目“今日图表”（Chart of the Day）为例，解读数据挖掘在新闻报道中的应用。

彭博新闻社依托其全球终端建立起来的海量的数据库，使得记者进行数据挖掘非常得心应手。彭博的“今日图表”“这个栏目将彭博新闻、彭博数据与彭博分析整合起来”[5]，其深度、速度和灵活性都非常高，工作难度也很大。彭博主编Matthew Winkler声称这几乎是竞争对手无法复制的栏目，至多能滞后些做出来。它通过图表和简单的事实而非说教来阐明道理，是彭博新闻“show, don't tell”理念的体现，是一种“简单而优雅的呈现观点以及点燃想象力的”[6]方式。

“今日图表”的构成有两部分，一部分是由彭博制作的图表，另一部分是一个4至6段的文字报道。首先，记者或编辑从纷繁复杂的数据、报道中寻找灵感的过程。“今日图表”灵感一般都来自最近发生的新闻。记者或编辑的“想象力，对数据的深入分析，每天的新闻标题，市场价格的异常变化，或者与分析师、投资者、经济学家的谈话也能提供灵感”[7]。

哪些数据值得挖掘？正在或刚发生的、读者关注的重大新闻事件，通过用数据挖掘得出的不同视角，可以丰富读者对事件的认知。彭博社很重视相关深度信息的呈现，如下面几例：

2012年11月2日，“今日图表”对过去10年在上海证券交易所交易的股票进行分析，自2002年11月至2012年10月，贵州茅台酒业股票上涨高达3451%，

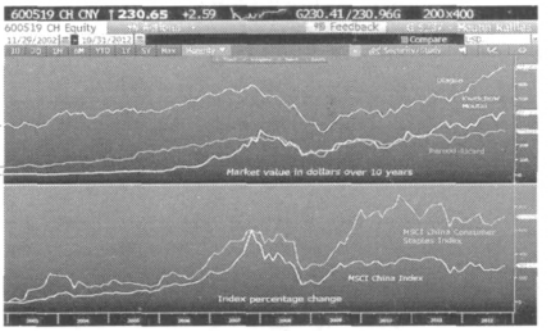

图1 世界三大造酒公司过去10年股价变化

A 帝亚吉欧(Diageo)酒业公司
B 贵州茅台(Kweichow Moutai)
C 保乐力加集团(Pernod-Ricard)

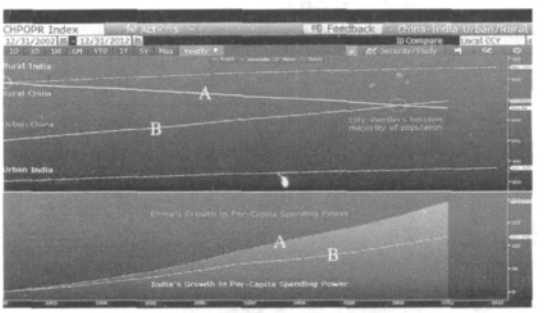

图2 中印两国过去10年城乡人口变化图

A 中国城乡人口变化
B 印度城乡人口变化

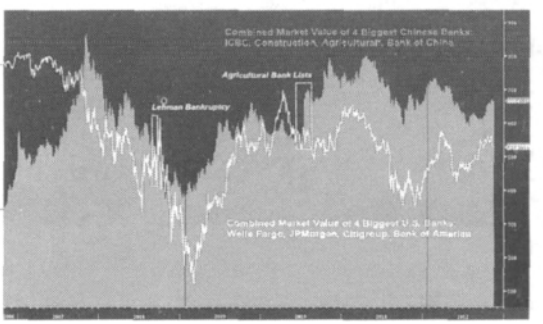

图3 中美两国4家最大银行市值对比

A 美国4家最大银行股值
B 中国4家最大银行股值

市值从不到10亿美元到今天的410亿美元，成为世界第二大造酒公司（图1以世界三大酒厂中的另外两家作为对照）。由此引申，中国自2002年以来的经济发展，已经造就了一个富裕阶层，他们对奢侈品类的需求，刺激了相关消费，因此出现茅台酒的大幅增长。而3451%的股价飙涨，对呈现中国经济的变化，非常有说服力。

2012年11月8日，彭博“今日图表”对比了过去10年中印两国城乡人口的变化：中印两国农村人口几乎都是7.8亿，但经过10年发展，有超过1亿农民进城，中国城市人口已超过农村（见图2）。同样作为新兴大国，中国的发展在城市化方面至少是远超印度的，而城市化被作为衡量现代化的重要指标。

2012年11月13日“今日图表”对比了中国4家最大银行与美国4家最大银行的市值变化：自2006年10月，4家中国银行的总市值超过美国同行，之后绝大多数时期，它们的市值一直都保持优势（见图3）。彭博社通过分析，认为中国几大银行未来几年还会保持优势。政府与

银行之间的关系的不同，就导致了两国银行获利方式的区别；中国银行的优势，是在过去10年的经济结构调整和发展中逐渐形成的。

数据挖掘也用于日常报道，从而对现实世界做出更深入的解释。2012年10月，科技新闻提到联想超越惠普，成为全球第一大个人电脑厂商。在移动设备风生水起，个人电脑销量下滑的今天，联想成为第一有多大意义？彭博社以2004年底联想和IBM签合同为起点，对比了世界五大个人电脑生产商的股价变化，发现联想股价在8年中上涨130%多，IBM的股价也提升了超过100%，而其他几家电脑厂商却有不同程度的下跌。这说明8年前联想并购Thinkpad的决定，至少从资本市场来说，对双方都是一个双赢决定；并对其他电脑厂商造成了一定打击。这样的数字就很说明问题，也是对质疑者的有力回应。(见图4)

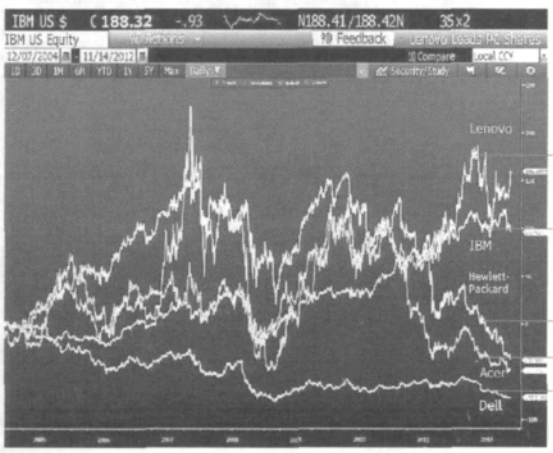

图4 全球5大电脑厂商过去8年股价变化

如今男女平等的概念已深入人心，但是现实情况如何？2012年10月22日，“今日图表”就做了一个有趣的数据挖掘，分析了全球最大的50家跨国企业董事会成员性别，结果发现，事实与人们嘴上说的大相径庭：在宝洁公司，女性占了45.5%的董事会席位，也是50家跨国巨头中唯一女性董事超过40%的公司；而在三星、本田等公司，董事会竟全是男性。西方国家董事会女性比例高些，这可以解释文化以及政策等多方面的情形。

如何用有效的数据支撑数据挖掘？10年前，记者如想获知许多国家的关键经济数据，还只能通过打电话到相关统计部门，经过繁琐的过程后才能拿到。今天，网络已经使得世界各国的数据触手可及，在类似彭博、路透及道琼斯这样的专业金融数据机构，这些数据更易获取，归类方式也更为合理。除了这些专业的金融数据机构外，有很多途径可以获得相关数据。例如股票市场的数据，在互联网上几乎都可以得到，因为每个上市公司都需要将随时发生的重大调整上报相关股票交易所，也需要每季度对外公开财报，这些数据都随时可查证。从彭博报道提及的信源看，有相当的数据都是外部机构发布的，记者只是根据报道需求去寻找数据。极少数难以直接获取的数据，可以请数据专家帮忙。最重要的，依然是对数据的解读方式和切入点。

媒体经营如何使用“大数据”？英国卫报网站2012年9月发表法国数字集团ePresse总经理Frédéric Filloux文章《数字新闻读者的“大数据”蕴藏巨大价值》称，其他行业得以有效利用的“大数据”，同样适用于数字媒体行业，读者的“大数据”蕴藏着尚未被挖掘的巨大价值，行为数据可用于使得新闻服务更能吸引读者，并为内容发行商带来更大的收益。数字发行的价值被严重低估，很多数字内容发行商都无法留住读者，读者的个人资料和其浏览的内容，以及它们与互联网“噪音”之间有怎么样的对比？这些是尚未被挖掘的最大价值来源。

如何使用无法被传统数据库管理工具吸收和分析的庞“大数据”集？其他行业的实践也许可以给予借鉴。通过强大的数据挖掘技术，美国超市连锁店塔吉特(Target)能够查出哪些顾客到了怀孕的第三个月，那是他们消费习惯中的一个重要时段。“谷歌流感趋势”(Google Flu Trends)对流感爆发的追踪比任何政府机构做得都要好。Google的搜索和广告业务及其实验中的机器人汽车，便利用了大量人工智能技术，它们对数量庞大的数据进行分析，并作出即时的决策。苹果公司在2010年收购的Siri网站，就在变成一种日益成熟的“个人助理——它能向用户提供提醒服务、天气预报、餐饮建议和对用户提出的大量问题作出解答等。

数据挖掘技术能够用于从数字新闻中提取更多的价值。互联网目前已经能够提供了解各类数据的必需工具，如谁在访问网站、他们喜欢什么等等，从中可以更加准确地了解用户和预测他们的需求，增进对读者的了解，从而推送甚至定制更契合需要的新闻、服务信息、精准的广告。

(作者系清华大学新闻与传播学院教授、副院长；本文的合作者刘少华，为清华大学新闻与传播学院硕士研究生)

注释：

①Mark P. Mills and Julio M. Ottino, The Coming Tech -led Boom, The Wall Street Journal, Jan.30,2012.

②本段落及以下3段，参引：The Age of Big Data, The New York Times, Feb.12, 2012,http://www.nytimes.com/2012/02/12/sunday -review/big -datas -impact -in -the -world.html?pagewanted=all

③Anand Rajaraman and Jeffrey D. Ullman, Mining of Massive Datasets, Copyright ? c 2010, 2011 Anand Rajaraman and Jeffrey D. Ullman.

④王光宏、蒋平：《数据挖掘综述》[J].同济大学学报自然科学版，2004(2):P246。

⑤ ⑥ ⑦ Matthew Winkler. The Bloomberg Way[M]. U.S.: Wiley, 109.