

数据挖掘的 “行情”

□ 文 / 魏瑾瑞

一、引言

人类历史上许多最重要的发明，从语言到电脑，无一不是用来生成信息、捕捉信息和消费信息的。但是不是有点儿过头了？就像旧石器时代的猎人们从一次猎杀一头长毛象到两头，再到把整群猎物逐下山崖，这就有点儿进步过头了。这世界被信息污染的还不够多吗？Gordimer（1991年诺贝尔文学奖得主）在《Loot》里有这么一句：A woman from the village cooks and cleans and doesn't bother him with any other communication. 这才是好woman！

平安集团就只销售一份保险给你吗？很可能不会，利用数据挖掘技术，它可以推测你潜在的理财需求，然后推送其他产品，比如由其控股的深圳发展银行的某款理财类或小额贷款产品……某人没有购买保险，这是否说明他没有丝毫购买保险的意愿？不，很可能是因为保险产品太单一，价格太高，他没有找到适合他的险种。100块钱的保险他买不起，为什么不能为他量身定做做一个80块钱的险种从而发展一个新的客户？注意，“他”可能并不是一个人，而是一类人的代表，一类人就是一个很大的市场；第二，一个新的客户可能会带

来长久的收益。所以，数据挖掘可以帮助企业找到“丢失的客户”（潜在的顾客）。一项对财富500强企业的调查显示，65%的企业正在使用数据挖掘，它为这些企业平均每年创造2000万至2400万美元的净利润。

数据挖掘的对象——大规模数据集——多集中在金融保险、零售、电信、生物信息、气象等领域。从服务企业的角度，Gartner（美国咨询公司）的分析师Doug Laney在2000年提出了刻画big data的三个基本特征：多样性、时效性与海量性。IBM公司也有类似的论述。当然，何为“大”，是一个变动的范畴。就像我们买电脑的时候对硬盘容量的关注一样，原先80G、120G就是大了，而现在500G也不算大。大规模数据集难在哪儿呢？最基本的，存储、分析和可视化等都是问题，特别是非结构化数据（雅虎首席产品官Blake Irving指出，世界上只有5%的数据是结构化的，而非结构化数据一直保持极大的增长）以及数据社会化所带来的挑战与机遇。

二、回归到繁荣时期

在美国网络搜索中，数据挖掘与算法(algorithm)、java代码(Java code)

这些关键词的网络活跃程度以及相关性(0.9779;0.9807)。大势已去？

可是麦肯锡2011年5月的研究报告《Bigdata: The next frontier for innovation, competition, and productivity》从商业视角就大规模数据集进行了讨论，涉及的领域有美国卫生保健和零售行业、欧洲公共部门、全球制造业和个人定位信息。以美国为例，该报告预测，按照现在的发展趋势，美国在2018年将有140万到190万能做大规模数据深度分析的人才缺口，而且这并不是一朝一夕可以弥补的，因为这种人才的培养需要时间。这实际上也是个全球性的问题，因为人才会流动（比如A国有人才缺口，它可能会高薪招聘，于是B国的人才就会奔赴A国就业）。事实上早在2010年8月，麦肯锡就出过类似主题的季节。

另一方面，从科学研究的视角，2008年9月4日刊出的《自然》(Nature)以big data作为专题(封面)进行了广泛的研讨。2011年2月11日，《科学》(Science)携其子刊《科学-信号传导》(Science Signaling)、《科学-转译医学》(Science Translational Medicine)、《科学-职业》(Science Careers)专门就日益增长的研究数据展开了一场大讨论。“数据”俨然为科学研究领域一个重要的核心关键词。根据这次专题讨论的词频（以大小表示）制作的，数据(data)自然是提到最多的词，其次是信息(information)、研究(research)、知识(knowledge)、分析(analysis)、可视化(visualization)……但具有讽刺意味的

是,我没有找到统计(statistics)、挖掘(mining)这样的词。这与前文麦肯锡报告提到的数据分析人才缺口相印证?

这是否说明,科学研究滞后于流行?数据挖掘在公众眼中逐渐冷却的时候,恰是科学研究开始重视的时候?还是发展到一定程度,大家开始关注数据挖掘的细节而不是数据挖掘这个名词了?热词从追捧到冷却,是网络时代的一贯特征?那既然如此,上面的图也就没什么意义了。

大规模数据集不仅是挑战,更是机遇。2010年《科学》对他们的同行评议员(peer reviewers)做过一个调查(有效样本1700),其中对“在你的研究中,曾经使用过或产生过的数据集有多大?”该问题的统计结果是,大约20%的研究者接触到的数据量在100G以上(其中超过1T的有7.6%);1-100G之间的有32%,小于1G的有48.3%。

云计算(cloud computing)之所以日益流行,可能就是因为它改变了我们处理数据(特别是大规模数据)的方式:原先,工具(比如软件)捆绑于本地,我们只能下载或拷贝数据到本地,让二者结合,然后才能进行分析;但是当数据量庞大到无法下载或移动时,为什么不能让工具移动到数据那边去?进一步,为什么不能让更多闲着的工具参与进来去加速处理呢?这便是我所理解的云计算(超级计算)。云就是通过网络。云特别依赖于畅通的网络环境。

我们会为一些资料存档,但真到用的时候往往找不到。这可能是由于乱丢的坏习惯,就像邮箱里的邮件因邮箱容量无限制而再没有去删减过一样;但也可能像老太太藏钱,谨慎备至,到最后却忘了藏哪个柜子里了。我经常怕忘记一件事儿,所以今天找个纸条记下来,这样第二天有个提醒,但是第二天竟然忘记看纸条。数据庞杂,急需一个有效的工具去抽丝剥茧。然后工具发明出来之后,直接纵容了人更加肆无忌惮地制造麻烦。所以,工具不得不升级,技术不得不进步。所以,可能更迫切的是需要从源头上管理好组织好数据,而不是任其随意产生出来之后再挖掘。

我想,对有组织的数据进行挖掘会是高效的。有组织的数据需要挖掘吗?需要!

回顾技术发展的历史,技术进步大都是被迫的。内燃机替代蒸汽机,当时若不是因为爆发口蹄疫需要将饮水槽拆除,我们很难想象还要等多久蒸汽机才能让位于内燃机。新石器时代转变成以农耕为主,可能就是因为当时狩猎过度已经发展到无法可持续的境地而不得不做的改变。所以一个优于旧事物的新事物,并不必然发生替代现象,它至少需要等待时机。

为什么是被迫?因为(1)根据牛顿定律,如果没有干扰,任何物体都有保持它先前姿势的惰性或惯性。(2)旧事物在本能上会排斥新事物,比如发现不是数的人被抛入了大海;再比如罗马时代有人发明了滑轮和杠杆,但维斯帕辛皇帝知道后立刻废除了这个发明,他说“自有我的奴隶帮我做事情,我用不着这个”; (3)当旧事物成为一个标准之后,即便它是落后的,也能成功的限制新事物进入,至少这里面有一个转换成本。

三、回归到原始时期


逻辑、方程……这不是统计的要害,数据才是。而从数据中挖掘出有用的信息和知识,这是数据分析的最终目标。可能是为了使结果易于理解,或索性就是数据处理的一种方式,可视化技术越来越受到追捧,甚至造出了一个新词“信息图形”,既直观也美观,“信息可以是美的”。

注意区分数据与信息(数据仅是信息的载体,信息是数据的表达、含义和解释),所以严格来讲,数据可视化与信息可视化是两码事儿。数据可视化相当于“看图说话”,你得把其中的隐喻表达出来;而信息可视化直接将隐喻这些抽象的东西绘制成图,自然不必多加解释什么。“描述和阐释一部电影几乎是不可能的。如果我能,我就不需要拍摄了。我以为:语言和文字无法表达的,才通过电影来表达。影像的气息是难以言表的。”姜文说。

下面给出了一些“信息图形”。我们看到了世界劳动力价格的惊人差距,

为赚到美国的最低工资,中国人平均要工作8年10个月;而英国只要8个月。儿童玩具广告中的出现频率最高的词竟然是“战争”(battle)!我国专利申请量与授权量稳步上升的事实;图6给出了各个国家移民美国的数量趋势。图7,迈克尔·安德森(Michael Anderson)竟然将简历也做成了信息图形。

但是请注意,信息图形并不是什么新发明,它其实很早就有了。交通路标、校徽、地图、几何等其实说穿了也就是信息图形,如图8。继几何、笛卡尔直角坐标系很久以后,大约在1750-1800年,统计图形才问世,其中William Playfair(1759-1823)对统计图形的发展与改进做出了很大的贡献。我们还能找到更早的,如象形文字。见下图10。所以信息图形与其说是革新,不如说是回归。不使用文字来表达,而是用图,这难道不是回归原始社会(如象形文字)了吗?动态图算进步吗?

图解金刚经、图解经济学、图解黄帝内经、图解易经等类似现象的盛行,在我看来,至少说明了两点,其一,分工细化导致的功能退化,比如自行车解放双脚的同时也滋生了双脚的懒惰、计算器和笔记本解放大脑的同时也弱化了我们的计算力和记忆力;其二,浮躁、缺乏耐心,想千方百计地把将来的东西尽可能多的放到今天享用。比如网络上的电影,再等两天就可以免费观看,可为什么大家还趋之若鹜地花钱去看呢?因为很多人没有耐心。在“我知道我想要什么!我要马上得到!”这样快节奏的社会背景下,坚持不再是重要的品质。人人都寄希望于一夜暴富、不劳或少劳而获。这可能吗?没有投入何来产出?从投入到产出也不是立竿见影的,种庄稼也得等到秋天才能见收成吧。“揠苗助长”的现代版本是,自己跟自己过不去。药吃下去要立刻见效,写了第一篇文章要立刻发表,点击文件、网页要立刻打开……太急了吧。电脑若反应太快你脑子能跟的上吗? 

作者单位:厦门大学计划统计系