

生活中的大数据

车品觉

阿里巴巴集团, 浙江 杭州 311121

摘要

随着大数据时代的来临,大数据逐渐渗入人们的日常生活中。选取了几个方向,对已渗入人们生活的典型大数据应用进行了分析和阐述,希望能给予读者启示。

关键词

大数据;气象;销售;医疗;信用;数据分析;数据产业

中图分类号:TP399

文献标识码:A

doi: 10.11959/j.issn.2096-0271.2016011

Big data in our life

CHE Pinjue

Alibaba Group, Hangzhou 311121, China

Abstract

With the arrival of the era of big data, big data has gradually come into people's lives. Some typical big data applications that have influenced into people's lives were analyzed and described as inspiration.

Key words

big data, meteorology, sales, medical treatment, credit, data analysis, data industry

1 引言

如今,一个大规模生产、分享和应用数据的时代正在开启。美国互联网数据中心指出,互联网上的数据每年将增长50%,每两年便将翻一番,而目前世界上90%以上的数据是最近几年才产生的。此外,数据又并非单纯指人们在互联网上发布的信息,全世界的工业设备、汽车、电表上有无数的数码传感器,随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化,也产生了海量的数据信息。这些以“PB”为单位的数据将现代社会带入了一个“大数据”时代。

“大数据”引起了全球的高度重视。2000年以来,搜索、广告和电商利用大数据获得丰厚回报。2012年3月,美国奥巴马政府宣布投资2亿美元启动“大数据研究和计划”,将“大数据研究”上升为国家意志,将对未来的科技与经济发展带来深远的影响。

“大数据”其实离我们很近,甚至已逐步渗入人们的生活中,并给人们生活带来了极大的便利。

2 你的冗余我的宝贝

人们每天都会收到天气预报,但这对大家来说只不过是个信息,它只能说明人们决定如何适应今天的天气,比如出门是否带伞、是否需要多穿衣服等。但美国有一家公司却把天气数据用于商业中。这家公司叫Planalytics,是一家全球领先的专注天气分析的公司,服务来自美国、欧洲及南美的200多家客户。该公司在天气分析上积累了大量经验,特别是在细节上,在美

国多家大型零售和制造商都有成功案例。目前为止,这家公司已分析了上万亿的交易数据及8 500个类目,基于这些分析再将天气数据转化为有助于企业决策的指针(可理解为数据中间层),进而优化业务结果。

数据的重大价值在于能让人们把决策前置。如果通过数据能够比别人哪怕早一秒对事情做出准确的预判,就比竞争对手更有优势。准确的预判加上快速的行动,就是现在人们所讲的行动智能(actionable intelligence)。但想达到这个目标,需要足够的资源,最关键的原材料就是数据。然而,一些重要数据之间的关系往往不是那么直观,人们常会忽略这些数据之间的关联性。发生这种情况有两种可能:一是这些数据之间的关系较隐性,不容易直接被人发现;二是过去不容易拿到这个数据,或者能拿到但数据太粗或不够准确等。

可以想象,春天的到来对零售商意味着春装的采购、户外用品销售的回升等,然而今年的春天是否会和以往相同?数据表明,重复上一年天气的概率仅有15%。而商户的规划往往基于往年的销售并假设同样的天气再现。

天气对商业的影响远大于人们所意识到的,商户对天气的反应也往往是滞后和临时性的。在美国,天气对服装行业和百货商店的影响都高达30亿美元,对园艺五金工具的影响则达到40亿美元。对于天气敏感的类目,利用天气数据将其业务weatherized(天气化)带来的效益是非常可观的。天气数据看似简单,却价值连城,但其价值并非来自本身,而是通过与具体商业问题的结合,发掘出其对商业的影响,从而创造价值。30°C的天气对业务意味着什么?这取决于所在的地区、所处的类目、所在的季节、该地区的历史天气以及历史销售记录等。因此,天气看似是小数据,但

要利用天气数据驱动业务，却需要融入来自多个来源的大数据。

一直以来，笔者也观察了天气跟服装销售的关系，发现一个很有意思的现象：原来，每年一到冬天刚结束，就有人开始寻找夏装，而当广州的天气第一次变热时，夏装的需求就爆发了。反过来，在冬天，北方城市的首次降温也会带来冬装搜索的热潮，而最早开始搜索的这些人往往都是服装达人。

像这样能了解到天气与服装销售之间关系的人，可能就发小财了，这就是数据联结的价值，这就是大数据的奇妙之处。别人的冗余，居然被收集起来“废物利用”，变成了我的宝贝，点石成金。这种经济效益绝对奇妙无穷，也是笔者从事大数据的动力。试想，在地广人多的中国，如果能够把更多的环境数据与人们衣食住行的习惯进行关联，将有可能挖掘出非常有价值的商业机会。

3 用生态数据看本质

如何判断一家公司的文化好坏？找员工填写调查表，还是找管理层深入访谈？可能去这家公司的洗手间看看，五秒就能得到答案。

其实，现在甚少有企业会去量化公司的“生态环境”，更少有城市会去量化城市的“生态健康”。大家常常说“智慧城市”，但有谁为“智慧城市”定过指标？“智慧”如何量化呢？是指城市竞争力强，还是居住舒适度高？

人们常常忽视数据的价值。很多KPI（关键绩效指标）的制定和考核都带有业务数据的痕迹，但是很少听到企业说要用数据量化公司部门及员工的合作关系，更少看到企业用数据去预测“生态规则”的

有效性，比如通过数据去了解一个新政策出台后导致的实际影响和人心变迁。实际上，数据之所以拥有巨大价值，关键在于可以通过数据将“生态”展现出来，并预测随之而来的影响及趋势。相信一切都可度量，这才是大数据的根本价值所在。

近期一个鲜明案例在美国发生，美国总统奥巴马宣布了一个“警察数据倡议”，在全美21个重要的管辖范围执行，目标是打造一个可以分析和分享知识、共同解决问题及建立最佳实践档案的社区。

2015年3月，在特遣队给总统的60条建议中，明确提出了警察数据化的强烈需求。这一建议指出，尽管美国采用了广泛的社交和经济指标来衡量国家的健康度，然而却缺乏衡量警察和执法机关的指标。警方数据的收集是保障公民权利及安全体系中非常重要的一个环节。目前这个数据开放社区主要有两项工作：通过开放数据增加透明度，建立社区信任并支持革新；更好地利用技术手段，例如早期预警，借之识别安全问题，提升责任感及防止滥用警力。

目前，已经有12个警察局承诺加入这个数据开放社区，他们和数据科学家一起深度分析警察/公民之间对抗的数据，从而让警察能够更加有效地在早期介入案件。美国首席数据科学家Patil D J亦在社交网站上，就警方数据和民众进行了交流。

奥巴马当局的聪明之处在于，他们围绕一个非常小的点——警方数据，构建一个数据生态社区。假以时日，这个社区可清晰地勾勒出美国社会安全的隐患，完善美国警力结构，进而降低犯罪率。

建立一个“生态数据”的体系，表面上跟传统的商业智能方案没什么不一样。其中的要点也不外乎数据收集、判断、整合、分析等。但从笔者的经验发现，两者不同的地方在于生态着重的是平衡，数据所显示的是各种角色之间互为关系的结果，即

从外往内看连接点，从垂直细分领域延伸至全面分析。同时，也清晰地认识到，大数据虽大，但目前的数据对于生态环境或系统而言却远远不够。因此，生态数据的构建将是人们长远的课题。

4 医疗业酝酿大数据突破

说起看病，很多人都会觉得痛苦不堪。排队长、看病慢、费用高，如果换一家医院，所有病历统统作废，检查记录需要重新来一遍。所以，普遍中国人都有个愿望，就是有医师、律师、会计师成为自己最贴心的顾问，帮助自己管理身体、司法事务及财富。这是痴人说梦话吗？现在，美梦或许逐渐成真。

美国初创公司Kyruus正在探挖医疗领域数据，高效地配对医生和病人。如果看看航运、旅游业的做法，就很容易理解Kyruus公司在做什么事情。

人们预订机票、酒店时，能很方便地根据指定时间、地点、星级等找到合适目标，那为何不能用这种数据驱动方式寻找医生呢？瓶颈原来在数据的可获取性。

Kyruus公司花费了大量精力探索多个数据系统，把所有与医生相关的数据整合在一起，包括预约系统、评价系统、索赔和账单数据。医生信息系统包括专科、语言、地点、电子邮箱、最早可预约时间等。当近距离看这些数据时会发现，大家一方面感叹医疗资源不足，另一方面竟然有30%~40%的医生每天都有空档。通过Kyruus这个平台，将大大减少医生的空档，提高医疗效率。

波士顿和洛杉矶的一些医疗保险公司已经开始使用这一项服务，Kyruus公司采用对医生收费的盈利模式，包括首次数据采集发布费用及后续的月费。目前Kyruus平台只向医疗保险公司开放，但很快将直

接面向病人。

在上述例子中看到，Kyruus公司整合了医生的数据，但在医疗领域上，大部分数据仍未打通。例如，哪位医生开了什么药给哪种病人，这个病人吃药后效果如何，这个病人以前的病历是否对药效有影响，甚至无数不同的病人对不同药物的反应是怎样，人的基因与药物测试结果的关系等。不同医生的诊断数据能否相互关联，形成一个庞大数据库，进而让机器从中学习，并找出最好、最有效的治疗方法呢？

试想还有多少专业领域正等待大数据去革新？越是看起来牢不可破的行业，就蕴藏着越大商机，当大家发现数据资源丰富（已数据化），但基于技术或组织原因而分散的行业时，该行业就值得立即革新。

“关联”就是大数据的创新精神。

笔者相信人类正在加速这方面的发展，希望借助大数据的力量，打破和革新更多传统行业的服务壁垒，让人们过上更轻松、更幸福的生活。

5 我的信用我做决定

大家试过信用卡逾期未还款吗？如果大家告诉大家逾期还款3次以上，会影响日后贷款的利息，大家可能就不会轻易忘记还款了。信用分数一直是个黑盒子，人们不知道自己为何会得到这样的评分，得不到适时反馈的结果是我行我素，甚至自暴自弃。每一次人们悔不当初之时，心里总会想：早知道就好了。

其实，眼下大部分的现实反馈都是滞后的：等到体检后才知道“三高”，等到失眠后才知道喝太多咖啡，而这时候再弥补和改正可能已经晚了。未来，大数据可以帮助人们“早知道”一点吗？再进一步看，大数据可以帮助人们迅速纠正错误的判断和

行为吗? 答案必然是肯定的。

现实中已有很好的例子, 比如健康手环, 实时地收集走路、休息、睡眠、心率等生活活动数据, 让人们更好地自律, 更健康地生活。根据这些数据的反馈, 可以马上采取修正行动。

下面为大家介绍的一家创新金融公司 Credit Karma (CK), 是一个打破“讯息黑盒子”的真实案例。该公司成立于2007年, 业务完全基于信用数据, 继而发展到繁多的个人金融服务, 目前已有4 000万用户, 占美国人口的12%, 非常可观。在美国, 信用分数 (credit score) 至关重要, 大至房屋贷款, 小至租房租车, 对一个人的信用判断都强烈依赖信用分数, 几乎所有人对自己的信用分数都非常重视。

CK是一个免费的个人信用分数管理平台, 把曾经是信用机构和金融机构才能取得的信用数据透明化。在得到用户授权的情况下, 该公司从美国三大征信局提取信用评分数据并提供给用户, 又在用户的分数发生突变时及时通知。同时, 帮助用户理解影响其信用分数的关键因素, 例如申请信用卡太多造成负面影响。这家公司还提供信用工具, 例如信用分数仿真器, 仿真发生某种个人金融行为对信用分数的影响。

在提供个人信用管理服务的基础上, CK公司基于对个人信用数据的了解, 进而发展为个人金融服务平台。该公司相信他们拥有足够的数据来为用户提供金融建议, 当用户的信用分数上升时, 理应获得更低的利率, CK公司会在第一时间为用户提供最佳的产品推荐, 现已包括信用卡、贷款及保险等。

有趣的是, 梵文 karma 可简单翻译为因果, 这个因果可能是有延迟效应的 (佛家叫“业”), 其中可分为共业与自业, 呼应于数据收集, 便可解释为个体反馈与集体反馈。信用评价与大环境的经济气候向来息

息相关, 大数据能给个人带来的就是集体经验的反馈。收集自己的健康数据, 可以调节生活习惯, 促进健康。若能导入类似人群的健康数据进行比较分析, 必然会得到更好的健康指导。

笔者认为, 大数据的趋势是运用集体智慧去优化个体的意愿, 这种互动是一种新的生活方式。以CK公司为例, 人类与大数据之间的互动, 将会加速变革很多传统行业, 使之产生翻天覆地的变化。

6 大数据走向平民化

笔者曾经跟一位美国零售集团的高管交流, 得知数年前当传统零售业意识到电子商务所带来的改变时, 除了大量招聘数据科学家之外, 第一件事就是去收集在线对手大量的实时商品数据。经历了几年后, 如今美国的大型零售公司都已具备选择商品、动态定价及多终端多渠道的管理能力。为了支持每一秒钟与对手的竞争, 可以想象收集更广、更实时数据的重要性, 即使是一个简单的天气变化, 都有可能令当前商品的竞争力落后于对手。

兵法有云“知己知彼, 百战不殆”, 对于企业应该如何使用数据, 这一句话最为贴切。当很多公司问该怎么使用数据时, 笔者必然会告诉他, 用数据了解自己 (知己) 是第一步。第二步就是用数据来了解竞争对手 (知彼), 并根据对手的动态变化做出攻守判断。大数据的关键, 就在于如何量化数据大小与决策好坏的关系。

大家可能会问: 这么高难度的事情, 小企业如何跟得上呢? 好消息是在这几年间, 美国的大数据生态已逐步走向平民化。

最近, 美国一家专注数据收集爬取的公司 Import.io, 就让大家眼前一亮。互联网是世界上最大的开放数据来源, 通过

Import.io公司的工具,用户可以瞬间将众多不同网站转化为数据表格或API(应用程序编程接口)。Import.io公司每天从互联网上搜集千万条纪录,至今已从30万个网站上收集了上千亿条数据纪录。Import.io公司在同类公司中技术领先,曾多次被评选为最佳创业公司。目前该工具对个人用户免费开放,对企业用户则源源不断地提供大规模、定制化及具有质量保证的数据。

令人感到兴奋的是,他们把采集半结构化数据这项本来只有技术工程师才能做的事,变成了人人都可使用的平民化服务。而健康的大数据生态,应该是可以让数据从收集、加工到应用的各个环节变得愈来愈精细,整个数据处理的过程形成一个互惠互利的产业链,大家都在大数据的海洋中共同分工协作。

不得不赞叹美国大数据产业链的发达:有人做数据的中间层,比如把天气跟零售的数据关联起来给需要的人使用;

有人把机器学习的门坎降低,变成MLS(machine learning as a service),让不懂数据的人也具备数据学习的能力;也有人把数据可视化变成简单易用的工具。Import.io这样的公司就把数据收集变成如水电煤一样每个人都能使用的服务。

虽然大数据从原始数据到加工再到使用是很长的链条,但如果有的数据生态和产业链,每个人专注其中的一个环节,各司其职,分工协作,就能把数据的价值发挥到最大。希望在不远的将来,中国也能跟美国一样,早日建立丰富完整的数据产业链。

7 结束语

大数据已在人们的生活中开花结果,在不经意间,给人们的生活带来了极大的便利,使科技发展日新月异。

作者简介



车品觉(1965-),男,现任阿里巴巴集团副总裁,浙江大学管理学院客座教授,中国计算机学会大数据专家委员会副主任,全国信息技术标准化技术委员会大数据标准工作组副组长,华人大数据学会执行会长,中国计算数学学会第九届理事,清华大学教育指导委员(大数据项目)等职。畅销书《决战大数据》的作者。多年来致力于研究大数据和互联网的发展趋势。曾在美、英、澳等地接受西方教育,毕业于新南威尔士大学,后于清华大学及英士国际商学院(INSEAD)获得双硕士学位,先后在HSBC、PCCW、微软、eBay等多家跨国公司任高管职务,未加入阿里巴巴集团之前,在敦煌网任CPO。2010年发起桑珠助学公益计划,致力于改善藏区儿童教育。

收稿日期:2015-11-03