数据挖掘大战给用户画像、贴标签

为何更多的互联却让我们“作茧自缚”

一度被热捧的大数据挖掘，近日站在了舆论的风口浪尖：一些商家利用大数据挖掘技术“杀熟”被网友亲测证实；脸书公司创始人马克·扎克伯格4月11日在美国国会就用户数据泄露丑闻听证；今日头条也以一种它并不情愿的姿态上了“头条”。

实际上，自从以算法主导的媒体生态、商业形态诞生以来，关于其为用户“画像”，通过兴趣推荐制造“信息茧房”的争论一直存在。“信息茧房”究竟是什么？它的危害果真如此巨大吗？“信息茧房”能否打破呢？

本报记者　王昱　整理

大数据挖掘是怎样的技术

据《科技日报》报道，“通过贴标签的方式建立用户画像，是数据挖掘常用的一种技术。”北京大学计算机科学技术研究所多媒体信息处理研究室主任彭宇新教授说，建立用户画像就是利用社交网络的信息，根据用户社会属性、生活习惯和消费行为等信息，抽象出一个标签化的用户模型，目标是使机器实现类似于人的“见信如面”、对用户了如指掌的能力。社交网络数据是实现这一目标的基础，机器对人的“初相见”多是源自于对社交网络数据的挖掘。

“有了标签，计算机就能够自动处理与人相关的信息，能够通过算法、模型逐步‘理解’人。”彭宇新介绍，多个标签共同完成画像，整个过程可分三步走：一是采集数据，即基于文本的信息抓取，口语称为“爬数据”；二是用户行为建模，通过机器学习技术，形成算法模型，判断用户可能的一些行为；三是可视化展现，把机器运算出来的结果，通过能让人类理解的方式展现出来。这三步是多轮调整的，在实际应用中，根据结果的反馈以及业务需求，可能进行二次建模等调整。

整个过程的影响参数是相对多元的，不同的行为类型，对于标签信息的权重影响也不同。以应用最广的商品营销为例，比如网售红酒，如果“购买”权重计为5，仅“浏览”计为1，加上浏览间隔、驻留时长、生活习惯等，通过复杂的算法最终呈现出一个标签的权重，再形成画像。

彭宇新说，基于用户画像技术，大数据挖掘进行分类和关联规则计算等分析：例如喜欢红酒的用户有多少，喜欢红酒的人群中，男、女比例是多少，喜欢红酒的人通常喜欢什么运动品牌等等。

另外，以前文本信息占主流，现在图像、视频等多媒体数据铺天盖地而来。彭宇新说，后者目前占据大数据的80%以上。数据类型发生的巨大变化，使得智能识别的任务更加艰巨。“管不住”和“用不好”的问题日益凸显。

一个20年前的“神预言”

“信息茧房”(Information Cocoons)是美国哈佛大学法学教授凯斯·桑斯坦在2001年的《网络共和国》一书中最早提出的。桑斯坦在该书对“信息茧房”只是粗略地说了个概念，没有进行更深刻的论述。但后来在他的另外一本著作《信息乌托邦》中，桑斯坦进行了更深入的分析和讨论。自此，“信息茧房”的观念才逐渐为知识界承认。

在这两本书中，通过对互联网的考察，桑斯坦指出，在信息传播中，因公众自身的信息需求并非全方位的，公众只注意自己选择的东西和使自己愉悦的领域，久而久之，会将自身桎梏于像蚕茧一般的“茧房”中。而加深化的个体观念差异，最终将带来社会因意见分歧造成的崩解。

桑斯坦的观点，显然是为了反驳当时在美国热议的、被他称为“信息乌托邦”的一种理论。该理论可以追溯到一本经典著作：1995年出版、1997年译为中文的尼葛洛庞帝所著《数字化生存》，这本书一度也被中国的互联网创业者推为“创业圣经”。比如，尼葛洛庞帝在书中提到：互联网可以基于算法，为我们挑选自己想看到的新闻，摒弃那些不想看到、不感兴趣的新闻，从而大幅提高人类的阅读效率。尼葛洛庞帝为这种当时尚在理论中的新闻媒体起了一个名字：Daily Me——我的日报。

今天我们可以看到，以今日头条为代表的算法推荐机制，无疑就是这个预言的现实版。通过获取用户信息和算法推荐，今日头条实现了尼葛洛庞帝梦想中的“基于算法，为我们挑选自己想看到的新闻，摒弃那些不想看到、不感兴趣的新闻”。而这种体验确实会让人上瘾：今日头条以其算法推荐的特点，迅速成为众多网络用户获取资讯的渠道。

看见没？原来今日头条的模式早已被学者所预见并写在书中了，然而，尼葛洛庞帝梦想的实现，真的就意味着人类阅读效率的大幅提高吗？并非如此。正如桑斯坦所忧虑的那样——我们非但没有来到“信息乌托邦”，反而掉入了“信息部落化”的陷阱。

互联网时代的“部落民”

在一些被高度阻隔的小区域中，人们可以发现一些审美、观点都高度特化的小部落。曾几何时，全球化的一大功绩，就是将这种“信息部落化”彻底打破，无论你身在何地，只要能联上网，就可以和全世界的信息沟通互联，基于地理区隔导致的“信息部落化”不复存在。

然而，当互联网进入2.0时代，尤其是“我的日报”真正实现后。人们惊奇地发现，这种趋势居然又反了过来——互联网不再打破“部落化”，而是反过来推动它。在海量的信息中，用户通常会选择自己需要的，而在算法主导的信息分发模式下，很容易过滤掉不感兴趣、不认同的信息，实现“看我想看，听我想听”。如同吸食精神鸦片后所获得的心理上的舒适感。久而久之，信息接受维度变窄，知识获取单一，行为习惯被自己的兴趣引导，在单调的信息中形成了特定思维习惯。而这些有着特定思维习惯的人群又开始在虚拟社区中聚集，彼此激发，最终形成了一个个小的“互联网部落”，生活在这些虚拟部落中的人们不再关心“部落外”发生了什么，而只接受自己喜欢的信息，与有共同嗜好的人交流。

是的，这就是当下互联网的尴尬处境，从某种意义上说，那些沉迷于快手或抖音上某些猎奇视频的人，与泰缅边境迷恋长脖子的部落民并没有多少区别，他们是互联网时代被特定信息“喂养”出来的、思维高度特化的“部落民”。阻隔这些“部落”的不再是崇山峻岭，而是互联网内容供应商的算法、推送机制和用户在不知不觉间构筑的“信息之茧”。

“信息茧房”如何打破

在尼葛洛庞帝的预言实现之后，桑斯坦的预言也在实现。令人遗憾的是，桑斯坦虽然成功预见到了“信息乌托邦”的危害，但并没有为其开出药方。相反，他指出在信息时代人们依照自己的喜好构筑“信息之茧”将是一种本能。毫无疑问，意识到这种前景危害的我们，眼下正在跟这种人类的本能作战。我们希望这场作战能取得胜利——毕竟从本心而言，任谁都不想在互联网链接全球的今天，却活得像个部落民。

诚然，在当今的互联网上，想要打破算法为你构筑的信息茧房是很难的，但并非毫无办法。眼下类似今日头条的新闻推送媒体大都采用给用户“数据画像”的手法，即通过用户的浏览记录，通过某种加权算法得出用户的偏好，而后进行新闻推送。按照这个原理，原先阅读面、思维面越狭窄的人，越容易被画出一张片面的“画像”，从而掉入信息茧房中。相反更广泛的阅读，会帮助你尽量避免受到算法的束缚——这也许才是信息技术升级为人们提出的最大挑战。

4月11日，脸书首席执行官马克·扎克伯格说，脸书将在其应用首页添加工具，引导用户进行隐私设置。

据《科技日报》报道，北京邮电大学教授杨义先认为，打破信息控制权几乎不可能，但隐私保护有个很便捷的方法。他比喻说，如果数据在网上“裸奔”，为了不被溯源，最便捷的安全手段是“把脸捂住”。这就是所谓的“匿名化处理机制”。

不难想象，随着技术的不断创新，会有更多用于信息安全的技术突破，不是一门心思用于大数据挖掘，也能用于制衡“信息控制权”。

看食品营养标签 选择低盐食品

统计显示，慢性病死亡在我国总死亡人数中的占比超过80%，且慢性病患者呈年轻化的趋势。不健康的生活方式和行为，如高盐、高油、不当膳食和缺少锻炼等，是当前慢性病发生发展的主要因素。

在日常生活中，该如何避免高盐高油的饮食呢？这就要用到营养标签了。食品营养标签标示了一个食品的基本营养特性和营养信息，是消费者了解食品的营养成分和特征的来源，也是保证消费者知情权、引导和促进健康消费的重要措施。

根据相关规定，食品标签必须标示能量和蛋白质、脂肪、碳水化合物、钠这四种核心营养素的含量值及占营养素参考值(NRV)的百分比。对着营养成分表，消费者可以拿起计算器来算一算。首先看看一个包装或一份食品的净含量，然后和营养成分表中的含量值相乘，就可以知道吃了整份食品后到底摄入了多少能量和多少营养成分。

为了避免买到高盐食品，消费者需要从多个细节入手。首先要注意查看营养标签，了解食品中钠含量的高低，是避免摄入过量钠的第一步。尤其对于那些有特别饮食需要的人群(如高血压患者)，钠含量的标注可以帮助他们根据自身健康状况作出有益的选择。

低钠食品在标签上会有相关的标注。低钠盐的钠含量要比普通食用盐低三分之一左右，低钠盐不但可以降低钠的摄入量，还可以帮助膳食中的钠、钾和镁元素达到更好的平衡。但同时也要注意：不能因为低钠食品就放松了数量的控制，即使是低钠食品也应适量。

此外，对于一些低盐产品的名称，消费者也要注意辨别，如“无钠”指每份钠含量低于5mg，“低钠”指钠含量低于140mg，“限盐”指钠含量降低25%，“无盐”或“未添加盐”指食品加工过程中未添加常规钠盐，但食品本身仍含有盐。

需要注意的是，有些食品成分里虽然没有明确写着“钠”，但仍可能含有钠，要注意不同形式的钠成分或不同的表述方法，如：海藻酸钠、碳酸氢钠(小苏打)、苯甲酸钠、谷氨酸钠(味精)等。

(志方)

美拟再建 百亿亿次级超级计算机

美国能源部4月9日称，计划投资18亿美元用于开发至少两台新的百亿亿次级(Exascale)超级计算机，以寻求美国在高性能计算领域的领导地位。

美国能源部列出了三台拟开发的百亿亿次级超级计算机，但第三台属于“潜在系统”，是否开发需视情况而定。而拟定开发的两台超级计算机，需满足能源部科学办公室先进科学计算研究(ASCR)项目和国家核安全管理局高级仿真和计算(ASC)项目的任务需求，它们将分别部署在橡树岭国家实验室和劳伦斯·利弗莫尔国家实验室，先后于2022年和2023年投入使用。

“这些新系统代表了下一代超级计算机，对美国科学家和美国工业来说是至关重要的工具，”能源部部长里克·佩里称，“它们将有助于确保美国在高性能计算这一重要领域的持续领导地位。”

能源部称，拟定开发的超算系统的运算速度将比目前美国最快的超级计算机高出50倍到100倍，将极大推动美国的科学研究和产业发展，有望在新材料研发、核安全评估、密码破译、癌症研究等领域大显身手。

(据《科技日报》)