



大数据和机器智能对未来社会的影响

吴 军

(丰元创投 美国加利福尼亚州萨拉托加 95070)

摘 要:从 2008 年起,大数据已经被推到了创新、提高竞争力和提高劳动生产率的前沿。过去大部分人对大数据的理解还停留在大量数据和数据的多样性本身,而忽略了它最重要的角色,即帮助机器智能的腾飞。基于此,对大数据和机器智能的发展可能给未来社会带来的全面、颠覆性的变革进行了系统分析。首先介绍了机器智能的发展历程,分析了大数据对解决机器智能问题的重要作用;然后结合实际案例,阐述了大数据的内涵以及大数据和大量数据的异同,重点分析了大数据和机器智能的关键技术;最后探讨了发展机器智能对未来社会的冲击。

关键词:机器智能;大数据;数据驱动

doi: 10.11959/j.issn.1000-0801.2015054

Big Data, Machine Intelligence and Their Impacts to the Future World

Wu Jun

(Zpark Capital, Saratoga CA 95070, United States)

Abstract: Big data has become the next frontier for innovation, competition, and productivity since 2008. However, most people still limited their vision on the quantity and the variety of big data, while ignored the crucial role it played in machine intelligence that was taking off recently. Therefore, comprehensive analysis on global and disruptive impacts to the future world due to big data and machine intelligence development was made. It was started from the introduction on development history of machine intelligence. The important role of big data to solve the problem of machine intelligence was analyzed. Combined with the actual case, the meaning of big data and the similarity and differences of big data and large data were investigated. The key techniques of big data and machine intelligence were described. Finally the challenges for future world due to machine intelligence development were discussed.

Key words: machine intelligence, big data, data driven

1 引言

过去 50 年,人类 GDP 增长根本的动力是摩尔定律。20 世纪 60 年代中期大规模集成电路(后来是超大规模集成电路)的出现,不仅带动了整个 IT 行业的技术革命,而且导致了全球的自动化和信息化,这是在过去的半个世纪里拉动世界经济增长的根本动力(虽然在中国还有房地产

和基本建设,但是在世界范围内,这个市场是在萎缩的)。

图 1 为 1971-2011 年微处理器中晶体管数量的变化,图中的点是不同时期具有代表性的处理器。摩尔定律带来的一个结果就是互联网的兴起以及产业的数字化,而这也带来了一个没有预想到的结果,就是各种数据量的急剧增长(如图 2 所示),最终导致了大数据的应用。摩尔定律和大数据共同带来的另一个结果就是“机器智

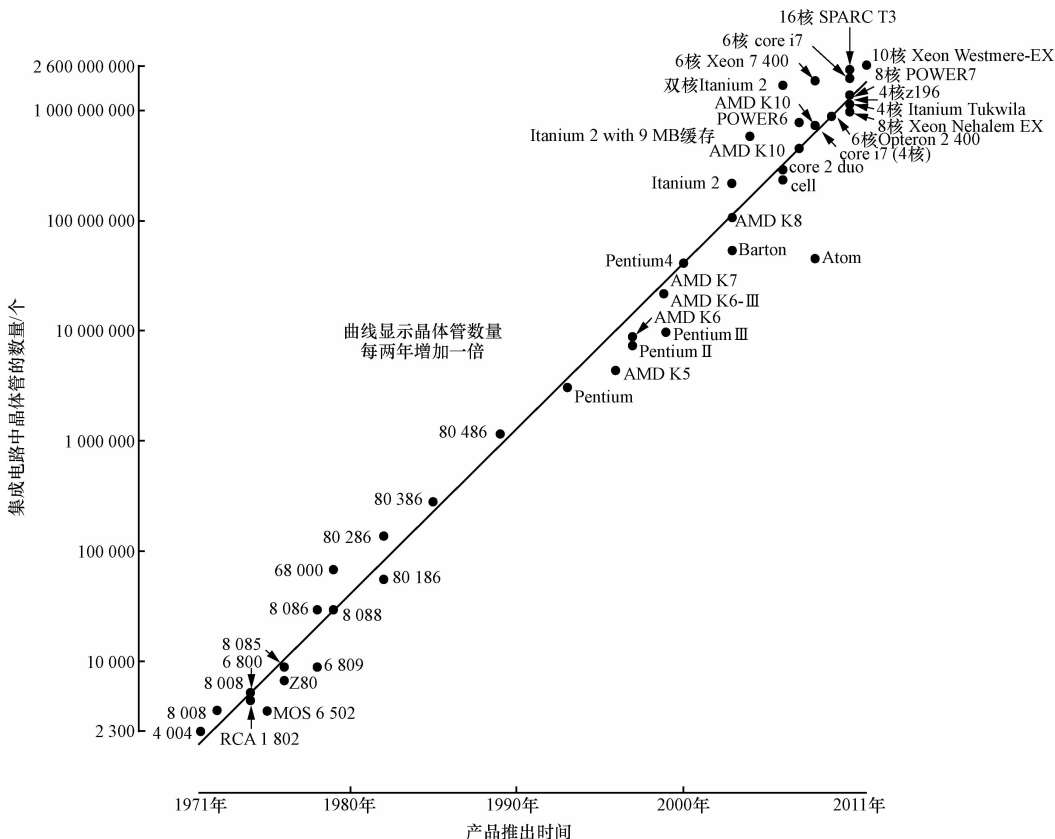


图1 微处理器中晶体管数量变化与摩尔定律

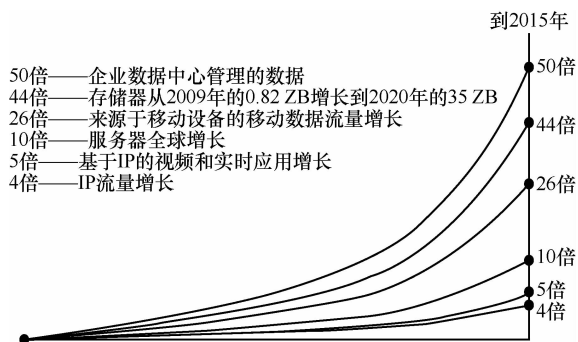


图2 各种数据量的增长趋势(数据来源于美国 Gartner 公司)

考是否有一天计算机能够在智力上超过人类。早在 1950 年, 计算机科学的先驱图灵博士就提出了一种衡量机器(包括计算机)是否有类似于人类智能的方法: 让一台机器和一个人坐在幕后, 与一个人展开对话(回答人的问题), 而这个人无法辨别和他讲话的是另一个人还是一台机器, 那么称这台机器具有和人等同的智能。这种方法被称为图灵测试^[1], 如图 3 所示。

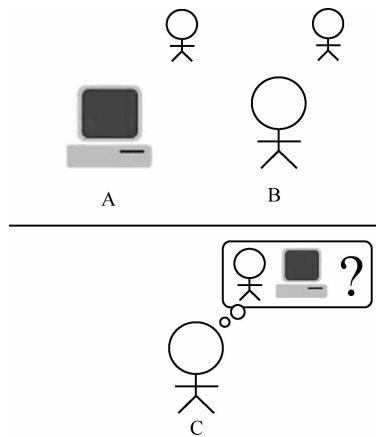


图3 图灵测试

能”将成为可能。在未来的 20 年(甚至更长时间)里, 大数据和机器智能的结合不仅将引领 IT 发展, 而且将彻底改变人类的生活方式以及社会结构和社会形态, 如今正处在这样一个变革的关键时期。为了说明这一点, 本文先从机器智能谈起, 然后谈到大数据的作用, 最后分析它们对今后社会的影响。

2 机器智能及其早期发展

自从 1946 年第一台电子计算机诞生, 人类就开始思

图灵当时只是给出了一个测试机器是否有智能的方法,但是并没有指出机器是否有可能获得类似于人类的智能,如果有这种可能,应该怎么做才能实现?计算机科学家们在20世纪50年代和60年代一直在考虑这个问题,他们最初的思路是让计算机仿造人的做法,并且提出了人工智能的概念。简单地讲,就是让计算机能够学习人的思维方式,掌握逻辑推理,以期望最终能像人一样根据已知条件,应用推理规则推出新的知识。这种设想符合人的直觉,因此在机器智能的发展初期它成为了主流的方法。后来在学术界,把这样的方法论称作“鸟飞派”,意思是看鸟是怎样飞的就能模仿鸟造出飞机,而不需要了解空气动力学。事实上,怀特兄弟发明飞机靠的是空气动力学,而不是仿生学。虽然当时计算机科学和认知科学对机器智能的认识并不深刻,但是整个学术界还是充满了乐观情绪,在后来的10年里,美国在经费上对人工智能的研究也是非常支持的。

从20世纪50年代到60年代,机器智能按照传统人工智能的路线走得非常不顺利,几乎没有拿得出手的成果。而与此同时,计算机科学的其他分支都发展得非常迅速。因此,美国计算机学界开始反思,机器智能是否走错了路?1968年,著名计算机科学家明斯基在Semantic Information Process一书^[2]中分析了所谓人工智能的局限性,他引用了Bar-Hillel使用过的一个非常简单的例子:The pen was in the box,这句话很好理解,如果让计算机理解它,做一个简单的语法分析即可。但是,另一句语法相同的话:The box was in the pen,就让人颇为费解。在英语中,pen还有另外一个不太常用的意思,即小孩玩耍的围栏,在这里理解成这个意思整个句子就通顺了。但是,如果用同样的语法分析,这两句话会得到相同的语法分析树,而仅仅根据这两句话本身是无法判定pen在哪一句话中应该作为围栏,哪一句话应该是钢笔的意思。事实上,人对这两句话的理解并非来源于语法分析和语意本身,而来自于他们的常识或者说关于世界的知识(world knowledge),这个问题是传统的人工智能方法解决不了的。因此,明斯基给出了结论:“目前”(即当时)的方法无法让计算机真正有类似人的智能。由于明斯基在计算机科学界具有崇高的声望,他的这篇论文导致美国政府削减了几乎全部人工智能研究的经费。

3 从人工智能到数据驱动的时代

在机器智能的发展史上,贾里尼克是一个划时代的人

物。1972年,康奈尔大学教授贾里尼克来到IBM沃森实验室进行学术休假,并且承担IBM研制智能计算机的工作。当时,计算机专家们认为如果计算机实现了下面几件事情中的一件,就可以认为它有图灵所说的“智能”:语音识别;机器翻译;自动回答问题。贾里尼克挑选了一个他认为最有可能突破的课题,即语音识别。

贾里尼克不是真正的计算机科学家,他的专长是信息论和通信,因此他看待语音识别问题完全不同于人工智能的专家们,在他看来这是一个通信问题。人的大脑是一个信息源,从思考到找到合适的语句,再通过发音说出来,是一个编码的过程;经过媒介(空气或者电话线)传播到听众耳朵里,是经过了一个长长的信道的信息传播问题;听话人把它听懂,是一个解码的过程。既然是一个典型的通信问题,就可以用解决通信问题的方法来解决,为此贾里尼克用两个马尔可夫模型分别描述信源和信道。为了训练和使用这两个马尔可夫模型,需要使用大量的数据。采用马尔可夫模型,IBM将当时的语音识别率从70%左右提高到90%以上,同时语音识别的规模从几百词上升到两万多词^[3],这样,语音识别就能够从实验室走向实际应用。贾里尼克和他的同事在无意中开创了一种采用统计的方法解决智能问题的途径,因为这种方法需要使用大量的数据,因此又被称为数据驱动的方法。

贾里尼克的同事彼得·布朗在20世纪80年代,将这种数据驱动的方法应用于机器翻译^[4]。由于缺乏数据,最初的翻译结果并不令人满意,虽然一些学者认可这种方法,但是其他学者尤其是早期从事这项工作的学者认为,解决机器翻译这样智能的问题只靠基于数据的统计是不够的。因此,当时SYSTRAN等公司依然在组织大量的人力,写机器翻译使用的语法规则。

如果说在20世纪80年代还看不清楚布朗的方法和传统人工智能的方法哪一个更适合计算机解决机器智能问题,那么在20世纪90年代以后,数据的优势就凸显出来了。从20世纪90年代中期之后的10年里,语音识别的错误率减少了一半,而机器翻译的准确性提高了一倍,其中20%左右的贡献来自于方法的改进,而80%则来自于数据量的提升,这背后的一个原因是互联网的普及使可使用的数据量呈指数增长。

最能够说明数据对解决机器翻译等智能问题有帮助的是2005年NIST对全世界各家机器翻译系统评测的结果^[5]。之前没有做过机器翻译的Google不仅一举夺得了各



项评比的第一名,而且将其他单位的系统远远抛在了后面。例如在阿拉伯语到英语翻译的封闭集测试中,Google 系统的 BLEU 评分为 51.31%,领先第二名将近 5%,而提高这 5 个百分点在过去需要研究 7~10 年;在开放集的测试中,Google 以 51.37% 的得分比第二名领先了 17%,可以说整整领先了一代人的水平。大家能想到的原因是它请到了世界著名的机器翻译专家弗朗兹·奥科 (Franz Och),但是参加评测的南加州大学系统和德国亚琛工学院系统也是奥科写的姊妹系统。从奥科在 Google 开始工作到提交评比结果,中间只有半年多的时间,奥科在方法上没有做任何改进。Google 系统和之前的两个系统唯一的不同之处在于,前者使用了后者近万倍的数据量。

表 1 是 2005 年 NIST 评比的结果。值得一提的是,SYSTRAN 公司的系统是唯一采用传统语法规则进行机器翻译的。它和那些采用数据驱动的系统相比,差距之大已经不在一个时代了。

表 1 2005 年 NIST 从阿拉伯语到
英语的翻译(封闭集)评比结果

公司	BLEU-4 分数
Google	51.31%
南加州大学	46.57%
IBM 沃森实验室	46.46%
马里兰大学	44.97%
约翰·霍普金斯大学	43.48%
SYSTRAN 公司	10.79%

2000 年之后,虽然还有一些旧式的学者死守着传统人工智能的方法不放,但是无论是学术界还是工业界,机器学习的主流方法是基于统计或者说数据驱动的方法。与此同时,机器学习和数据挖掘开始热门起来。

4 从大量的数据到大数据

数据驱动的方法需要大量的数据。从理论上讲,切比雪夫不等式保证了在具有大量代表性的数据后,统计模型的准确性。从应用上讲,Google 等公司的成功也验证了这一点。既然数据是非常有用的,如果具有更多、更完备、全方位的数据,就可能从中挖掘出很多预想不到的惊喜。大数据这个概念就在这样的背景下应运而生。什么是大数据?现在很多人都爱讲这个时髦的词,但是大多数人对它的理解不仅有很大的局限性,有些甚至完全是错误的。

人们对大数据的第一个误解就是将大数据和大量的数据混为一谈,大数据的数据量确实非常大,但是只是量

大还不是大数据。一个人的 DNA 的原始数据,单纯从量上来讲也是非常大的,但是做不了什么事情,因为不具备统计意义。类似地,欧洲核子研究中心(CERN)每一个对撞实验的数据量都很大(一年产生的数据量和整个互联网的数据量在同一个数量级),但是这些数据都被限制在特定的维度,虽然能够验证很多理论(如证实了希格斯玻色子的存在),但是它们只是围绕特定的问题,而不会对其他领域产生影响。

大数据不同于大量数据的第一个关键点是数据的多维度性质。2013 年 9 月,百度发布了一个颇有意思的统计结果——中国十大“吃货”省市排行榜。百度没有做任何民意调查和各地饮食习惯的研究,只是从“百度知道”的 7 700 万条和吃有关的问题里“挖掘”出来一些结论,而这些结论看上去比任何学术研究的结论更能反映中国不同地区的饮食习惯。百度给出的结论:在关于“什么能吃吗”的问题中,福建、浙江、广东、四川等地的网友最经常问的是“什么虫能吃”;江苏、上海、北京等地的网友最经常问“什么的皮能不能吃”;内蒙古、新疆、西藏的网友则是最关心“蘑菇能吃吗”;而宁夏网友最关心的竟然是“螃蟹能吃吗”。宁夏网友关心的事情一定让福建网友大跌眼镜,反过来也一样,宁夏网友会惊讶于有人居然要吃虫子。

百度做的这件小事就是大数据的一个典型应用,它有几个特点:首先,数据本身非常“大”,7 700 万个问题和回答不是一个小数字;第二,数据维度非常多,涉及食物的做法、吃法、成份、营养价值、价格、问题来源的地域和时间等,而且这些维度也不是明确给出的(这与传统的数据库不一样)。在外人看来,这些原始的数据“相当杂乱”,但是恰恰是这些看上去杂乱无章的数据将原来看似无关的维度(时间、地域、食品、做法和成份等)联系起来。经过对这些信息的挖掘、加工和整理,得到了有意义的统计规律,例如不同地域的人的饮食习惯。

百度只公布了一些大家感兴趣的结果,只要它愿意,还可以从这些数据中得到更多有价值的统计结果。例如,它很容易得到不同年龄人、性别和文化背景的饮食习惯(假如百度知道用户的注册信息是可靠的,即使不可靠也可以通过其他方式获取可靠的年龄信息),不同生活习惯的人(例如正常作息的人、夜猫子们、经常出差的人或者不爱运动的人等)的饮食习惯。如果百度的数据收集时间跨度足够长,还可以看出不同地区人的饮食习惯的变化,尤其是在不同经济发展阶段饮食习惯的改变。而这些看似很

简单的问题(如饮食习惯的变化)没有“百度知道”的大数据将很难得到。

上面这些统计似乎并不复杂,按照传统的统计方法应该也可以获得,但是采用传统方法的难度比一般人想象得大。如果在过去想要获得这些统计结果必须做的事情:首先,需要先设计一个非常好的问卷,然后要从不同地区寻找具有代表性的人群进行调查(这就是盖洛普一直在做的事情),最后要半人工地处理和整理数据。这样不仅成本高,而且如同盖洛普民调一样,很难在采样时对各种因素考虑周全。如果在后来统计时发现调查问卷中还应该再增加一项,而补上这一项的成本几乎要翻倍。

传统方法难度大的另一个原因是填写的问卷未必反映被调查人真实的想法。大家在百度知道上提问和回答是没有压力也没有功利的,有什么问题就提什么问题,知道什么答案就回答什么,但是在填写调查问卷时就不同了。大部分人都不愿让自己表现得“非常怪”,因此不会在答卷上写下自己有“爱吃臭豆腐”的习惯或者“喜欢吃虫子”的嗜好。据著名媒体人罗振宇介绍,中央电视台过去在调查收视率时就遇到这样的情况,他们发现通过用户填写的收视卡片调查得出的收视率和自动收视统计盒子得到结果完全不同。从收视卡得到的统计结果中,那些大牌主持人和所谓高品位的节目收视率明显地被夸大了,因为用户本能地要填一些让自己显得有面子的节目。笔者也做过类似的实验,从社交网络的数据得到的对奥巴马医疗改革的支持率(大约只有24%)比盖洛普的结果(41%)要低得多。

大数据的好处远不止是成本和准确性的问题,还有它的多维度(或者说全方位)。过去计算机能够存储和处理的数据有限,因此只收集与问题相关的数据,这些数据只有很少的几个维度,而看似无关的维度都被省略掉了。这种限制也决定了特定的数据使用方式,即通常是先有假设或者结论,然后再用数据来验证。如今云计算的出现可以存储和处理大量的、关系很复杂的甚至是原本看上去没有用途的数据,工作的方法因此而改变。除了使用数据验证已有的结论之外,还可以从这些数据本身出发,不带任何固有的想法,看看数据本身能够给出什么新的结论,这样就发现了很多新规律。例如,百度百科中的数据乍一看是杂乱无章的,其实数据之间有很多内在联系。在对这些大数据进行分析之前,产品经理们的头脑里并没有预先的假设,也不知道能得出什么样的结论。但是,通过对这些数据的分析发现了很多新的规律。

大数据和大量数据的另一个不同之处在于它的完备性。在1936年的美国总统大选前夕,当时著名的民意调查机构“文学文摘”(the literary digest)预测共和党候选人兰登会赢。此前,文学文摘已经连续4次成功地预测了总统大选的结果,这一次它收回来240万份问卷,比前几次多得多,统计量应该是足够了,因此民众们相信他们的预测结果。而当时一位名不见经传的新闻学教授(也是统计学家)盖洛普却对大选结果提出了相反的看法,他通过对5万人意见的统计,得到了民主党候选人罗斯福会连任的结论。大选的结果出来后,采用了少量样本的盖洛普预测对了,面对迷惑的民众盖洛普解释了其中的原因:文学文摘统计的样本数虽然多,但是却不具有代表性,它的调查员们是根据电话本上的地址发送问卷的,而当年美国只有一半的家庭安装了电话,这些家庭的收入相对偏高,他们大多支持共和党。而盖洛普自己在设计统计样本时,考虑到了美国选民种族、性别、年龄和收入等各种因素,因此虽然只有5万个样本,却更有代表性。这个例子说明统计样本代表性的重要性。1936年的大选预测还催生出一个现在最权威的民调公司——盖洛普公司,在这之后该公司又成功地预测了1940年和1944年两次大选。在1948年底美国大选的前夕,盖洛普公布了一个自认为颇为准确的结论:共和党候选人杜威将在大选中以比较大的优势击败当时的总统,即民主党候选人杜鲁门。由于盖洛普公司前3次的成功,在大选前很多人都相信这个结论。但是,大选的结果是杜鲁门以比较大的优势获胜,这让大家对盖洛普公司的民调方法产生了质疑。虽然盖洛普公司考虑了选民的收入、性别、种族和年龄的因素,但是还有非常多的其他因素以及上述因素的组合他们没有考虑。

在随后的几十年,盖洛普公司不断地改进采样方法,力求统计准确,但是它对美国大选结果的预测是大局(全国)尚准确,但是细节(每一个州)常常出错,因为再好的采样方法也有考虑不周全之处。

2012年总统选举时,这种“永远预测不准”的情况得到了改变。一位统计学家Nerd Silver通过对互联网上能够取得的各种数据(包括社交网络上用户发表的信息、新闻信息和其他网络信息)进行大数据分析,准确地预测了全部50个州的选举结果(如图4所示),而盖洛普从来没有做到这一点。这个结果是否是蒙的?这个可能性或许存在,但是概率只有 10^{-15} ,因此可以认为这是大数据分析的结果。Silver并没有好的采样方法,只是收集的数据很完备。大数



(a) Nerd Silver 对 2012 年美国大选的预测结果



(b) 2012 年美国大选的实际结果

图4 Nerd Silver 对 2012 年美国大选的预测和实际结果的对比^[6]
(深灰色表示共和党获胜,浅灰色表示民主党获胜)

据的完备性不仅有用,甚至有点可怕。

大数据的第三个特征体现在它的英文提法 big data 上,这里使用的是 big data,而不是 large data。big 更强调抽象意义上的大,而 large 是强调数量(或者尺寸)大。big data 的提法不仅表示大的数据量,更重要地是强调思维方式的不同。这种以数据为主的新做法,在某种程度上颠覆了长期以来在科学和工程上的方法论。过去强调做一件事情的因果关系,通过前提和假设推导出结果。本文以 Google 和微软在网页搜索上的工作为例来说明这个方法论的变化。

大多数人认为 Google 的搜索比微软的 Bing(在质量上)做得略好一点的原因是 Google 的算法好,这种看法 2010 年以前是对的,因为那时 Bing 搜索在技术和工程方面确实明显落后于 Google。但是如今这两家公司在技术上已经相差无几,Google 还能稍稍占优,除了产品设计略微好一些之外,主要是数据的力量。在 2006 年以前,Google 改进产品的思路是先找到产品的不足再研究改进的方法,同时进行大量的实验,最后在工程上实现。但是,从 2004 年开始 Google 发现,在历史上无意间收集到的大量关于用户使用互联网习惯的数据对产品质量的提升有意想不到的好处。由于大数据的完备性,通过分析数据可以先得到正确的结论,即使它背后的原因一时还考虑不清楚。这样,公司就面临一个选择——是否相信这些无法解释的结论。起初,很多人还试图在找到合理的解释后才使用通过数据得到的结论,但是当商业上的竞争不断加剧时,2006 年之后大家的工作方式不自觉地转到了先相信结论,再搞清楚原因这样一种新思路上。久而久之,Google 在工作中对数据产生了严重的依赖。如今它的产品比微软和雅虎等竞争对手做得稍好一些,主要的原因是它不仅拥有更多的数据,而且比其他公司更早地具有了大数据的思维方式,即不再刻意

追求假设条件、推理过程和结论之间的因果关系,而是直接得到问题的解答。

综上所述,大数据不仅仅是大量的数据,也不仅仅是一些公司为了推销产品而强调的异构数据或者非结构化数据,而是具有了以往数据中难以具备的多维度和完备性等特点,更重要的是它是一种新的思维方式、一种新的方法论。

5 大数据和机器学习

现在大数据炙手可热的原因不仅是各行各业都可以通过对数据的分析极大地提升自身的业务,更重要的是它将带来机器智能的全面革命,并且最终改变世界的产业格局和社会生活。

从数据驱动的时代开始,计算机领域的科学家和工程师已经意识到,让机器具有所谓的智能需要走一条和人的认知方法完全不同的道路,这就是发挥计算机在计算和存储方面的特长,利用大数据的完备性,发现人难以发现的规律,得到传统方式无法得到的结果,从而在某些方面超越人的智力。概括起来,如今从学术界到工业界都意识到,实现机器智能需要 3 个支柱:摩尔定律、大数据、数学模型。

摩尔定律的作用是保证计算机的计算能力和存储能力能够适应解决复杂智能问题的需求;大数据的多维度和完备性特点是保证智能性问题能够找到答案的关键;数学模型则是将现实生活中的问题转化成计算的桥梁。本文的重点是阐述大数据扮演的角色,下面通过 Google 的两个例子来说明。

第一个例子是关于计算机自动回答人的问题。根据图灵的观点,当机器可以回答人的问题而提问者无法判定回答问题的是机器还是人时,机器就有了和人同等的智能。

这要求计算机不仅能够识别人的语音,还能回答人的提问,而后面一件事在半个多世纪里都做得不是很好。虽然可以通过模式匹配找到一些简单问题的答案,但是直到 2012 年,机器自动回答复杂问题(如“为什么”和“怎么办”)在学术界依然被认为是一个尚未解决的难题。

2012 年,在 Google 内笔者领导的研究小组开始着手解决这个问题,直到 2014 年年初,Google 的搜索引擎已经能够回答 30% 的复杂问题了,这远远超过了学术界迄今为止同类研究的水平。究其原因,除了 Google 在自然语言处理等基础算法上做到了世界领先之外,更重要的是 Google 将这个过去认为是自然语言理解的问题变成了一个大数据的问题。首先,Google 发现对于用户在互联网上问的各种复杂问题,有 70%~80% 可以在前 10 条自然搜索结果(去掉广告、图片和视频等结果)中找到答案,而只有 20% 左右的复杂问题,答案存在于搜索结果的摘要里。因此,Google 将机器自动问答这样一个难题转换成了在大数据中寻找答案的摘要问题。这里有 3 个前提:首先,答案存在,这就是大数据的完备性;其次,计算能力足够,Google 回答这样一个问题的时间小于 10 ms,但是需要上万台服务器同时工作;最后,要用到非常多的自然语言处理算法,包括对全部的搜索内容进行语法分析和语义分析,能够从文字的片段合成符合语法而且读起来通顺的自然语言等。第一个前提只有 Google 等少数大公司具备,而学术界不具备,因此决定了 Google 而非学术界最早解决图灵留下的这个难题。图 5 为 Google 自动问答的实例。

第二个例子是关于 Google 无人驾驶汽车。无人驾驶汽车可以算是一个机器人,因为它需要像人一样对各种随机突发性事件快速地做出判断。2004 年,经济学家们^[7]依然

认为开车这件事是计算机难以取代人的几件事之一,他们在得出这个结论时,除了分析技术上和心理上的难度外,还参考了当年 DARPA 组织的自动驾驶汽车拉力赛的结果:当时排名第一的汽车花了几个小时才开出 8 英里,然后就抛锚了。但是仅仅过了 6 年,2010 年 Google 的自动驾驶汽车不仅被研制出来,而且已经在高速公路和繁华的市区行驶了 14 万英里,没有出一次事故^[8]。

为什么 Google 能在不到 6 年的时间里研制出自动驾驶汽车?最根本的原因是它把这个机器人问题变成了一个大数据问题。首先,自动驾驶汽车项目是 Google 街景项目的延伸,Google 自动驾驶汽车能去的地方都是它“扫过街”的地方,这个汽车在行驶到任何地方时,对周围的环境是非常了解的,不像过去那些研究所里研制的自动驾驶汽车每到一处要临时识别目标。其次,Google 的自动驾驶汽车上面装了十几个传感器,每秒钟几十次的各种扫描,这不仅超过了人所谓的“眼观六路、耳听八方”,而且积攒下来的大量数据对各地的路况以及不同交通状况下车辆行驶模式有准确的了解,计算机学习这些“经验”的速度远远比人快得多。这两点是过去学术界所不具备的条件,因此 Google 才能在非常短的时间里实现汽车的自动驾驶。

无论是计算机自动回答问题还是智能汽车的自动驾驶,都涉及大量的预先完成的离线计算。为了实现计算机的自动问答,需要事先对全部网页的内容进行语法和语义分析;为了实现自动驾驶,要事先识别出各个街道上的所有目标,这些都需要海量的计算。因此,世界上最强大的机器人其实不是那些做成人形、能够活动的机器,而是在数据中心里面几万、几十万甚至上百万的服务器。

最后,回顾一下 Bar-Hillel 和明斯基用的那个关于 pen

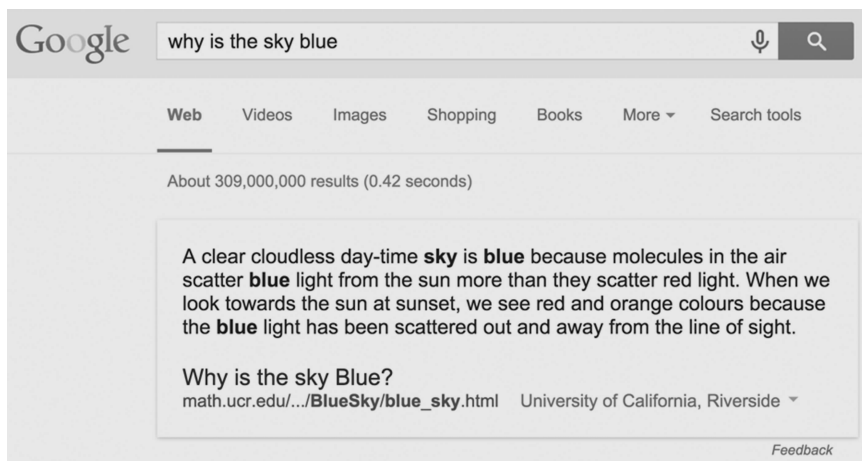


图 5 Google 自动问答(问题为“天为什么是蓝色的”,问题下面是计算机产生的答案)



含义的实例。明斯基指出,传统人工智能的方法无法让机器获得世界的知识,这是机器产生智能的障碍。现在有了大数据的完备性,从某种角度上讲,解决了世界知识的障碍,因此使机器具有智能成为可能。

6 大数据和机器智能的关键技术

实现大数据的应用和机器智能涉及很多关键技术,涵盖了计算机科学、电机工程、通信、应用数学和认知科学等许多方面,本文重点介绍其中最主要的几项。

6.1 数据的收集

大数据离不开数据,而数据的收集非常关键。与传统的利用采样收集数据不同,大数据需要全面地、在无意间收集各种可能有用的数据。强调“无意间”是因为有时可以收集的数据会变形,不具有统计意义,关于央视收视率调查的例子就说明了这个道理。

虽然收集大量的数据并不难,但是收集全面却不容易。在 Silver 所做的美国大选预测的例子中,准确性来自于数据的全面性。下面用 Google 产品中另外一个例子来说明收集完备的数据和大量的数据在难度上的区别。

统计语言模型^[9]是语音识别、机器翻译甚至输入法的基础,而训练一个好的统计语言模型需要大量的数据。过去为了保证训练数据具有代表性,通常是从各种来源的文本数据中抽样,凑足即可。这样得到的模型是针对所有人的,当然对于用词非常怪异的人(小概率事件)来讲,通用的模型就未必合适。在大数据时代,有可能为每一个用户量身定做一个语言模型,这就要收集每一个用户完备的输入数据,要做到这一点而不涉及用户的隐私就有相当大的难度了。在大数据时代,拥有数据就等于拥有了财富。2014年初,Google 以 32 亿美元的天价收购几乎没有什么收入的智能空调控制器公司 Nest,这不是为了给家庭省一点电,而是通过 Nest 的产品收集人们在家活动的数据,而这些数据采用传统的数据采集方法是得不到的。

6.2 数据的存储

摩尔定律使存储成本成倍下降,但是当大数据出现后,数据量增长的速度可能超过摩尔定律增长的速度(如图 2 所示)。例如,Google 眼镜可能将人一辈子看到的事情全部记录下来,如果这件事做成了,会彻底改变人们对世界,甚至对自己人生的了解。但是,将这些视频(包括音频)数据都存下来不是一件容易的事情。

对于拥有和使用大数据的公司,存储数据的数据中心

就成了未来业务的关键点。如果这个点出了故障,如何保证业务的正常运行就是一个挑战。另外,在很多时候数据的存储不仅仅是成本问题。处于保护隐私的考虑,很多与用户相关的数据还不能集中存储,而是需要存在每一个人的智能终端上,这也给数据的存储和使用带来了挑战。

6.3 数据的表示、检索和随机访问

对于互联网的网页数据、公司运营的日志数据、用户使用互联网习惯的数据,虽然其数据量大,但是颗粒度都很小(一个字段一般只有几个字节到几十个字节),因此它们的表示(描述)、检索和随机访问并不是大问题。但是,对于富媒体数据(如视频),要想随机访问其中一个画面就不是一件容易的事情。还有很多比网络富媒体颗粒度大得多的数据,例如很多和医疗相关的数据,一个基本单元就几百兆甚至更多。目前,检索一个词组是件容易的事情,但是检索一段基因就不容易。除了医疗,很多行业(如半导体设计、飞机设计制造)的数据量都很大。

在过去,各个领域甚至各个公司会有自己的数据格式,它们只在自己的领域使用自己的数据。但是,到了大数据时代,希望通过数据之间的相关性尤其是大数据多维度的特性,找到各种事务之间的关联。例如第 4 节的“百度知道”的例子,如果能够将每一个用户的饮食习惯收集起来,通过可穿戴式设备了解他们的生活习惯,然后再与他们的医疗数据甚至是基因数据联系起来,就能研究出不同人、不同生活习惯下各种疾病的发病可能性,并且可以建议他们改进饮食习惯,预防疾病。这个前景看起来很美好,但是要实现它就必须先解决数据的表示、检索和随机访问等问题。显然,对于世界上各种各样的大数据,无法用一个统一的格式来描述,但是需要一些标准的格式,以便于相互交换数据和使用数据。

数据表示的另一个挑战来自于数据保密和对隐私的诉求。很多时候,希望处理/使用大数据的人,既能得到想要的统计规律,又无法看到数据的内容。

6.4 数据的使用和挖掘

使用大数据,相当于在一堆沙子中淘金,不经过处理的原始数据给不出新知识,大数据能产生的效益在很大程度上取决于使用和挖掘数据的水平。在 Google,至少有四成的工程师每天在处理数据。

大数据不同于过去为了某个特定目的获取或者产生的数据,在结构和格式上比较规范,大数据的原始数据常常是杂乱无章的,因此“从沙子里淘金”的本领是使用大数

据的必要条件。在第 5 节的机器自动问答的例子中,虽然问题的答案存在于网页之中,但是答案的内容通常是零碎地分布在不同网页里,对网页的结构、内容进行分析就成为了使用大数据的先决条件。而对于文本大数据来说,自然语言理解技术是使用它们的前提。

从纷杂的数据中获取有用的信息,通常只是使用大数据实现机器智能的第一步,而接下来关键的一步就是机器学习。对于同样的数据、同样的算法,采用不同深度的机器学习方法得到的结果不同。Google 的 Jeff Dean 等人采用大规模并行的人工神经网络,对语音识别的参数重新训练,将识别的错误率降低了 15%(相对值)^[10]。对于机器翻译,效果同样显著。这说明机器学习的重要性。

6.5 安全性和隐私性的维护

大数据由于数据量大而且完备,一旦丢失,损失将是巨大的,而一旦被盗取,后果更是不可想象。因此,大数据的安全性是 IT 领域新的挑战。

大数据的完备性可以带来很多好处,例如很多原本针对一个群体的产品和服务可以做到针对每一个人,但是同时也会带来隐私的忧患。对隐私的保护不是仅在法律层面,而是指如何通过技术手段,使得在使用大数据时既能发挥它的功效,还能够维护个人的隐私。

6.6 数学模型和智能算法

目前,机器智能做得最好的领域都有一个共同的特点,就是找到了把实际问题变成计算的数学模型和智能算法。为了让计算机发挥更大的作用,变得更加智能,需要在数学模型研究上进行更大的投入,而这需要一个漫长的技术积累,很难在一夜之间有突破,因此除了长期坚持研究,别无他法。

7 机器智能和未来社会

机器智能可以帮助改善人类的生活,但是人们在欢呼机器智能到来的同时,是否准备好了它对未来社会带来的冲击? 2011 年德国提出工业 4.0 的概念,即通过数字化和智能化提升制造业的水平。其核心是通过智能机器、大数据分析来帮助工人甚至取代工人,实现制造业的全面智能化。这在提高设计、制造和供应销售效率的同时,也会大大减少产业工人的数量。在中国,全球最大的 OEM 制造商富士康,一直在研制取代生产线上工人的工业机器人。未来将有上百万的机器人取代装配线上的工人,这使得工人们不再需要到生产线上从事繁重而重复的工作,也使工厂里

的工人数量大幅度地减少。

很多人会说,自从大机器出现后工人的数量就在减少,劳动力会被分配到其他行业。但是,如同在 2004 年经济学家低估了机器可以取代驾驶员的可能性一样,如今可能在低估机器智能对未来社会的冲击。这一次由机器智能引发的技术革命不仅替代那些简单的劳动,而且将在各个行业取代原有的从业人员,因为这将是人类历史上第一次机器在智能方面超越人类。

在美国,专科医生(如放射科医生)是社会地位和收入最高的群体,也是需要专业知识最多、智力水平最高的群体,他们需要在大学 and 医院学习和训练 13 年才能获得行医的执照,过去认为这样的工作是不可能被机器取代的。但是,现在智能的模式识别软件通过医学影像的识别和分析,可以比有经验的放射科医生更好地诊断病情,而这个成本只是人工的 1%^[11]。

律师也被认为是最“高大上”的职业,但是他们的工作受到了自然语言处理软件的威胁。如今打一场像苹果和三星这样的官司,要分析和处理上百万份法律文件(因为美国是判例型法律),律师费高得惊人。但是,位于硅谷 Palo Alto 的 Blackstone Discovery 公司发明了一种处理法律文件的自然语言处理软件,使得律师的效率可以提高 500 倍,而打官司的成本可以下降 99%。这意味着未来将有相当多的律师可能失去工作。事实上,这件事情在美国已经发生,新毕业的法学院学生找到正式工作的时间比以前长了很多。

面对势不可挡的机器智能大潮,人类在未来需要重新考虑工作和生活方式,尤其是劳动力的出路问题。从 100 多年前开始的农业革命使得发达国家 2%~5%(根据美国劳工部的统计,美国农业工人与劳动力人口的比例不到 2%)的人提供了全部人口所需的食物,随着机器智能的发展,或许只需要 5% 以下的劳动力就能提供人类所需的所有工业品和大部分的服务工作。当然,会有一小部分人参与智能机器的研发和制造,但是这只会占劳动力的很小一部分。现在必须考虑未来劳动力的出路,这是一个在机器智能发展过程中无法回避的问题。

8 结束语

机器智能曾经是无数代人的梦想。在 1946 年人类制造出第一台电子计算机以后,这个梦想似乎离人们非常近了,但是直到上个世纪末,机器的智能水平还比较低。主要有两个原因:一是很多技术条件还不具备,机器智能的发



展在整体上还处于早期阶段;二是人们习惯于按照人的思维方式去要求机器,并且让机器模拟人,导致走了很多弯路。但是,人类最终认识到计算机实现智能和人采用的方法是完全不同的。经过半个多世纪的摸索,人类找到了采用适合机器特长(计算快、容量大、知识复制容易等)的发展机器智能的方式。由于人类的智能水平在过去的几千年里没有太大的变化,但是机器的智能水平却可以随着摩尔定律呈指数增长,因此在某个时间点,一定会出现机器在智能上超过人类的可能性,如今人们可能就处在这个转折点,而这一切在很大程度上是因为有了大数据。

参考文献

- 1 Turing A M. Computing machinery and intelligence. Mind, 1959: 433~460
- 2 Minsky M. Semantic Information Process. Cambridge MA: MIT Press, 1968
- 3 Jelinek F. Continuous speech recognition by statistical methods. Proceedings of the IEEE, 1976, 64(4): 532~556
- 4 Brown P F, Cocke J, Della Pietra S A, *et al.* A statistical approach to machine translation. Computational Linguistics, 1990, 16(2): 79~85
- 5 NIST. NIST 2005 Machine Translation Evaluation Official Results. http://www.itl.nist.gov/iad/mig/tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html, 2005

- 6 Watercutter A. Fact: Nate Silver Predicts Election Outcome, Becomes Nerdy Chuck Norris. <http://www.wired.com/2012/11/nate-silver-facts-election/>, 2012
- 7 Murnane F L. The New Division of Labor. Princeton: Princeton Press, 2004
- 8 Markoff J. Google cars drive themselves, in traffic. The New York Times, 2010
- 9 吴军. 数学之美. 北京: 人民邮电出版社, 2014
- 10 Wu J. Beauty of Mathematics. Beijing: Posts & Telecom Press, 2014
- 11 Le Q, Ranzato M R, Monga R, *et al.* Building high-level features using large scale unsupervised learning. Proceedings of the 29th International Conference in Machine Learning, Edinburgh, Scotland, UK, 2012
- 12 Brynjolfsson E, McAfee A. Race Against the Machine. Lexington: Digital Frontier Press, 2011

【作者简介】



吴军,男,博士,毕业于清华大学和约翰霍普金斯大学,丰元创投创始合伙人,计算机科学家,前 Google 高级资深研究员,前腾讯副总裁,Google 中日韩搜索算法的作者,Google 自动问答系统和许多创新产品的负责人,工业和信息化部顾问,约翰霍普金斯大学工学院董事,《浪潮之巅》、《数学之美》和《文明之光》的作者。

2015054-10

2014 年 11 月通信业主要指标完成情况(一)

指标名称	单位	本年本月止累计到达	比上年同期累计	本月
电信营业收入	亿元	12 277.3	-0.3%	1 060.9
其中:电信主营业务收入	亿元	10 643.5	4.3%	888.1
电信固定资产投资完成额	亿元	2 915.8	10.3%	420.5
固定本地电话通话时长合计	万分钟	23 983 055.8	-13.6%	2 067 837.3
区间电话通话时长	万分钟	2 407 144.0	-17.4%	207 180.3
区内电话及拨号上网通话时长	万分钟	21 575 911.8	-13.1%	1 860 657.0
固定长途电话通话时长合计	万分钟	4 873 240.2	-9.9%	432 039.0
国内长途电话通话时长	万分钟	4 782 489.2	-9.5%	424 590.0
国际电话通话时长	万分钟	51 516.0	-32.2%	4 038.8
港澳台电话通话时长	万分钟	39 234.9	-19.0%	3 410.2
移动电话通话时长合计(含本地)	万分钟	268 120 906.6	1.2%	23 797 675.0
移动电话国内长途通话时长	万分钟	62 606 598.6	1.9%	5 611 315.2
移动电话国际电话通话时长	万分钟	118 707.0	-1.1%	10 154.3
移动电话港澳台电话通话时长	万分钟	84 515.3	-4.8%	7 425.6
移动短信业务量	万条	69 789 641.4	-16.1%	6 280 366.7
移动互联网接入流量	万吉字节	182 135.5	60.3%	21 653.8

注:(1)电信主营业务收入增长率按可比口径计算。(2)固定长途电话通话时长和移动电话通话时长均包含相应的 IP 电话通话时长。(3)通话时长各项指标均为去话通话时长。

(来源:工业和信息化部运行监测协调局)