EasyChair Preprint

№ 2828

Where Will Artificial Intelligence Go in Big Data'S Environment?

Yang Jinghua

March 3, 2020

在大数据的环境下人工智能将何去何从

摘要：随着大数据时代的到来，数据已经不能简单地用二维形态来表述了，并且工业上对数据价值的要求也越来越高，这就促进了许多大数据衍生行业的产生，人工智能就属于这个新型时代产物之一。经过科学家们不断地努力，人工智能已经发展到了前所未有的局面，但图灵奖获得者朱迪·玻尔却认为目前的人工智能只是弱人工智能，要想达到通用的智能机器必须学会因果，只有攀登了因果论的第二层，才有可能达到强人工智能的巅峰。本文对大数据进行了简单的概述，对因果发展历程进行了回顾，介绍了因果涉及到的部分方法，最后结合当今人工智能发展的障碍进行了分析，得出了只有发展因果才能帮助人工智能取得新的成就的结论。

关键词：人工智能；大数据；因果论

Where will artificial intelligence go in big data's environment?

Abstract: With the arrival of the era of big data, data can no longer be expressed simply in two-dimensional form, and the demand for data value in industry is getting higher and higher, which promotes the emergence of many industries derived from big data. Artificial intelligence is one of the products of this new era. Through the continuous efforts of scientists, artificial intelligence has developed to an unprecedented situation, but Turing Award winner Judy Bohr believes that the current artificial intelligence is only weak artificial intelligence, in order to achieve universal intelligent machines must learn causality, only by climbing the second layer of causality theory can we reach the peak of strong artificial intelligence. This paper gives a brief overview of big data, reviews the course of causal development, introduces some methods involved in causality, and finally analyzes the obstacles to the development of

artificial intelligence today. it is concluded that only the development of causality can help artificial intelligence to make new achievements.

Key words: Artificial intelligence; big data; Causality theory

1 大数据相关内容概述

随着新一代信息技术的到来，数据正在以前所未有的速度增长，大数据已经成为了社会发展的一种潮流。大数据相比于传统数据，主要体现在种类多、规模大、真实性高、增长速度快等特点。环绕在我们周围的数据已经不再是简单的结构化数据，不能简单地以二维形态描述图片、声音、视频这些非结构化数据；被遗忘在数据库中的数据也会被重新利用，对其进行分析和挖掘，得到的信息能够帮助我们对未来的事情做出合理的预判和抉择。

大数据最早起源于谷歌公司在 2003 年到 2006 年发表的三篇论文，它们分别是分布式文件系统 GFS、大数据分布式计算框架 MapReduce 和 NoSQL 数据库系统 BigTable。这三篇论文研究了海量数据的存储及计算问题，他们将分布式、大数据之类的研究带入了大众视野。接着，Lucene 开源项目的创始人 Doug Cutting 根据论文原理初步实现了类似 GFS 和 MapReduce 的功能；在 2006 年，Doug Cutting 启动了一个专门用于开发和维护大数据技术的独立项目，即 Hadoop(包含分布式系统 HDFS 和大数据计算引擎 MapReduce)；2008 年，Hadoop 正式成为 Apache 的顶级项目，随即出现了专门运营 Hadoop 的商业公司 Cloudera，这对 Hadoop 进一步发展提供了商业支持；随后，Yahoo 因 Map Reduce 编程太麻烦，开发了脚本语言 Pig，Facebook 开发了使用 SQL 语法的 Hive，随着众多 Hadoop 周边产品的出现，大数据生态体系逐渐形成。

基于大数据技术的广泛适用性，许多国家已经将其与人工智能技术一起提升到国家战略层面。人们最初接触人工智能可能是 2016 年 3 月在韩国首尔举行的“人机大战”，棋手李世石同人工智能围棋程序 AlphaGo 进行了五番比赛，结果李世石以 1 比 4 的成绩输给了 AlphaGo。这场人机比赛不仅仅让人类见识到了人工智能强大的一面，更掀起了一场人工智能的发展热潮。

人工智能概念最早是在 1956 年的达特茅斯会议上被提出来的，维基百科将

人工智能定义为由人制造出来的机器所表现出来的智能，通常人工智能是指通过普通计算机程序来呈现人类智能的技术；该词也指出研究这样的智能系统是否能够实现，以及如何实现。那么什么是智能呢，1950 年英国数学家艾伦·麦席·图灵将“智能”的概念用“机器能思考吗？”代替，进而提出了图灵测试来判断机器是否具有智能。“图灵测试”很简单，测试要求一个人和一台拥有智能的机器设备在互不相知的情况下，进行随机的提问交流，如果超过 30%的测试者没有发现对方是机器设备，那就代表了这台设备拥有“人类智能”。既然智能，自然有强弱之分，强人工智能就是机器具有自我意识，要求机器有知觉有意识；弱人工智能是指没有知觉意识的智能，机器按照事先写好的程序进行工作，并不拥有智能，从定义来看，我们现在接触的人工智能都是弱人工智能。人工智能经过孕育、形成、知识运用及综合集成四个阶段的发展，已经实现了虚拟人工智能助理、自动驾驶技术、智慧医疗等现代科技。

2 因果发展历史及模型概述

因果和相关是两个不同的概念，如果两个事物有因果他们一定相关，但是相关的两个事物不一定有因果，比如闪电过后一定有雷声，雷声和闪电是相关的，但闪电并不是雷声的因。由此我们可以看出，相关性是指在数据分布中，我们可以观测到 X 与 Y 相关，如果仅仅观测 X 的分布，也能推测出 Y 的分布；而因果性是指在改变 X 后，Y 也会随着 X 的改变而改变，则 X 是 Y 的因。

因果概念最早出现在两千多年前，亚里士多德等西方哲学家思考事件之间的“导致”关系并提出了因果的概念，亚里士多德把因分成了四类：目的因、动力因、质料因及形式因。除了哲学家外，佛教也特别强调因果，《因果经》中记载：“欲知前世因，今生受者是；欲知来世果，今生作者是”。其意思相当明了：想知道前世的事，今生这番就是最好的体现；想知道下辈子如何，取决于今生的作者。17 世纪末、18 世纪初德国数学家、哲学家戈特弗里德·莱布尼茨(1646-1716)在其著作《单子论》中明确提出了“充足理由律”；18 世纪苏格兰哲学家、经济学家和历史学家大卫·休谟给出了因果关系的现代定义；德国哲学家亚瑟·叔本华

(1788-1860)在《充足理由律的四重根》中，则把充足理由律分为四种表现形式：因果关系、逻辑推论、数学证明、行为动机。充足理由律衍生出很多产物，比如牛顿的经典力学、爱因斯坦的相对论、量子理论等，这也从侧面反映出因果关系在科学研究中的重要地位。

因果推断常用的模型有两种：一种是潜在结果模型，另一种是因果网络模型。潜在模型主要用在作因果判断的两个事物是已知的，进而评价原因对结果的影响。因果网络模型表示了多个变量间的因果关系，图中的节点代表变量，而节点间的箭头代表事件的因到事件的果或是变量中数据的变化过程。因果网络框架主要研究因果作用的可识别性及对其中的因果关系进行学习，通过因果网络可以很容易地判断变量中是否存在混杂因素以及哪些因素是混杂因素，也是因果网络的建立解决了几个世纪未曾解决的悖论问题，如："运动-胆固醇水平"的辛普森悖论、有关"饮食-体重"的罗德悖论以及有关止血带作用和好坏胆固醇的悖论等。潜在结果模型和因果网络模型的不同点在于，前者不需要知道完整的因果网络，但是需要可忽略处理分配假定或者工具变量假定；后者则需要事先知道一个完整的因果图。

3 因果才是人工智能的出路

早期的人工智能可以分为两类：一类是基于符号逻辑的演绎推理；一类是基于概率的归纳推理。而因果推断则是将演绎推理和归纳推理两个角度进行了结合产生的新的算法。

图灵奖获得者朱迪·玻尔将因果论分为三层：第一层为"关联"，第二层为"干预"，第三层为"反事实推理"。关联层主要是被动地观察到什么，比如百货公司的经理常常思考买纸尿布的顾客同时购买啤酒的概率是多少；干预层主要是主动地改变现状，比如百货公司的经理思考假如我把纸尿裤的价格提高一倍，那么买啤酒的人会增多吗；反事实层是因果论的最高层，它将现实存在的情况进行完全地否定，并作出精准地判断，比如，百货公司的总经理思考假如我把纸尿布的价格提高两倍，那么之前买纸尿裤的顾客依然买纸尿裤的概率是多少。玻尔认为目前的人工智能只是处于因果论的的最底层，智能也只是相对于特定的环境而言的，即使我们把人工智能做得很优秀，也只是处于底层的弱人工智能。如果专

家想做到强人工智能，攀登上因果的第二层，令机器能够像人类一样对复杂的环境做出相应的反应，就必须让机器人拥有因果推理的能力。

目前的人工智能虽然取得了巨大的成就，但它对复杂的环境做相应判断的能力却不如一个小孩子。与机器获得的数据相比，小孩子根据从外界获得的少量数据，结合孩子自身的因果判断能力，对周围的环境做出了正确地反应。故强人工智能的目标不再是大数据智能，而是通过小数据做出正确抉择，灵活地完成复杂的任务。如今影响人工智能发展的三大障碍分别是：鲁棒性、可解释性及对因果掌握的程度。鲁棒性主要是机器对于陌生的或复杂的环境适应能力差，不能在未经过训练的环境下正常工作，比如现在的手机语音助手可以很轻松地回答“今天天气怎么样？”、“给我讲个笑话”等，但是你问它我是吃冰淇淋还是吃汉堡，它可能就不知道怎么回答了，这就显示了人工智能的适应能力不强，只能在特定环境下回答特定的问题，而不能根据所处的环境随机应变；可解释性主要是指对系统产生的结果没有办法进行合理的解释。系统处理的过程对于用户来说，像一个黑盒子，是看不见摸不到的，而处理过程中出现任何的微小的扰动都会对结果产生很大的影响，产生的噪音却是不可溯源的，这对进一步研究系统的模型、提升系统的性能来说都是一种障碍。最后一点掌握因果的能力，形式化人类自身具有的因果能力，尝试让机器对周围的环境进行简单的记录，对接触的概念进行抽象升华并尝试回答如果我不这么做结果会怎么样，这也是让机器登上因果论巅峰的一个尝试。

如此重要的因果论在最近才崭露头角，一个重要的原因就是它的基础学科统计学在 20 世纪后期才得到发展，而现在火爆的大数据不过是对多个变量进行统计分析，深度学习不过是多了一层神经网络，AI 火爆的原因不在于理论的创新，而是算力的提高，应用层面的创新远远超过了理论的创新点。相比于人工智能，强化学习可通过与环境交互进行学习，允许对结果进行干预，不从已知的知识获取经验为反事实成立提供了条件。因此，我们应该像玻尔教授交给我们那样，不再停留在底层，努力攀登“干预”层，最终拿下“反事实”的旗帜，顺着因果的指引，走向强人工智能的阳光大路上。

参考文献

[1] 苗旺, 刘春辰, 耿直. 因果推断的统计方法[J]. 中国科学:数学, 2018, 48(12):3-28.

[2] 梅剑华. 人工智能与因果推断——兼论奇点问题[J]. 哲学研究, 2019.

[3] 姜奇平. 因果推断与大数据[J]. 互联网周刊, 2014(18):70-71.

[4] 江生. 人工智能不能缺少因果推断[N]. 中国证券报,2019-08-03(009).

[5] 王杉.大数据背景下的人工智能范式综述[J].科技风,2019(34):1.

[6] 叶磊. 大数据综述[J]. 2015.

[7] 李学龙, 龚海刚. 大数据系统综述[J]. 中国科学:信息科学, 2015, 45(1):1-44.

[8] 刘瑞玲. 浅谈大数据应用现状及发展趋势[J]. 科技展望(15):16.

[9] 潘文. 我国大数据发展现状与趋势[J]. 领导科学论坛(4).

[10] 李芬, 朱志祥, 刘盛辉. 大数据发展现状及面临的问题[J]. 西安邮电大学学报, 2013, 18(5):100-103.