

预习 02 | 大数据应用发展史：从搜索引擎到人工智能

2018-11-01 李智慧

从0开始学大数据

[进入课程 >](#)



讲述：李智慧

时长 11:29 大小 5.26M



上一期我们聊了大数据技术的发展历程，事实上，我们对大数据技术的使用同样也经历了一个发展过程。从最开始的 Google 在搜索引擎中开始使用大数据技术，到现在无处不在的各种人工智能应用，伴随着大数据技术的发展，大数据应用也从曲高和寡走到了今天的遍地开花。

Google 从最开始发表大数据划时代论文的时候，也许自己也没有想到，自己开启了一个大数据的新时代。今天大数据和人工智能的种种成就，离不开全球数百万大数据从业者的努力，这其中也包括你和我。历史也许由天才开启，但终究还是由人民创造，作为大数据时代的参与者，我们正在创造历史。

大数据应用的搜索引擎时代

作为全球最大的搜索引擎公司，Google 也是我们公认的大数据鼻祖，它存储着全世界几乎所有可访问的网页，数目可能超过万亿规模，全部存储起来大约需要数万块磁盘。为了将这些文件存储起来，Google 开发了 GFS（Google 文件系统），**将数千台服务器上的数万块磁盘统一管理起来，然后当作一个文件系统，统一存储所有这些网页文件。**

你可能会觉得，如果只是简单地将所有网页存储起来，好像也没什么了不起的。没错，但是 Google 得到这些网页文件是要构建搜索引擎，需要对所有文件中的单词进行词频统计，然后根据 PageRank 算法计算网页排名。这中间，Google 需要对这数万块磁盘上的文件进行计算处理，这听上去就很了不起了吧。当然，也正是基于这些需求，Google 又开发了 MapReduce 大数据计算框架。

其实在 Google 之前，世界上最知名的搜索引擎是 Yahoo。但是 Google 凭借自己的大数据技术和 PageRank 算法，使搜索引擎的搜索体验得到了质的飞跃，人们纷纷弃 Yahoo 而转投 Google。所以当 Google 发表了自己的 GFS 和 MapReduce 论文后，Yahoo 应该是最早关注这些论文的公司。

Doug Cutting 率先根据 Google 论文做了 Hadoop，于是 Yahoo 就把 Doug Cutting 挖了过去，专职开发 Hadoop。可是 Yahoo 和 Doug Cutting 的蜜月也没有持续多久，Doug Cutting 不堪 Yahoo 的内部斗争，跳槽到专职做 Hadoop 商业化的公司 Cloudera，而 Yahoo 则投资了 Cloudera 的竞争对手 HortonWorks。

顶尖的公司和顶尖的高手一样，做事有一种**优雅的美感**。你可以看 Google 一路走来，从搜索引擎、Gmail、地图、Android、无人驾驶，每一步都将人类的技术边界推向更高的高度。而差一点的公司即使也曾经获得过显赫的地位，但是一旦失去做事的美感和节奏感，在这个快速变革的时代，陨落得比流星还快。

大数据应用的数据仓库时代

Google 的论文刚发表的时候，吸引的是 Yahoo 这样的搜索引擎公司和 Doug Cutting 这样的开源搜索引擎开发者，其他公司还只是“吃瓜群众”。但是当 Facebook 推出 Hive 的时候，嗅觉敏感的科技公司都不淡定了，他们开始意识到，大数据的时代真正开启了。

曾经我们在进行数据分析与统计时，仅仅局限于数据库，在数据库的计算环境中对数据库中的数据表进行统计分析。并且受数据量和计算能力的限制，我们只能对最重要的数据进行统计和分析。这里所谓最重要的数据，通常指的都是给老板看的数据和财务相关的数据。

而 Hive 可以在 Hadoop 上进行 SQL 操作，实现数据统计与分析。也就是说，**我们可以用更低廉的价格获得比以往多得多的数据存储与计算能力**。我们可以把运行日志、应用采集数据、数据库数据放到一起进行计算分析，获得以前无法得到的数据结果，企业的数据仓库也随之呈指数级膨胀。

不仅是老板，公司中每个普通员工比如产品经理、运营人员、工程师，只要有数据访问权限，都可以提出分析需求，从大数据仓库中获得自己想要了解的数据分析结果。

你看，在数据仓库时代，只要有数据，几乎就一定要进行统计分析，如果数据规模比较大，我们会想到要用 Hadoop 大数据技术，这也是 Hadoop 在这个时期发展特别快的一个原因。技术的发展同时又促进了技术应用，这也为接下来大数据应用走进数据挖掘时代埋下伏笔。

大数据应用的数据挖掘时代

大数据一旦进入更多的企业，我们就会对大数据提出更多期望，除了数据统计，我们还希望发掘出更多数据的价值，大数据随之进入数据挖掘时代。

讲个真实的案例，很早以前商家就通过数据发现，买尿不湿的人通常也会买啤酒，于是精明的商家就把这两样商品放在一起，以促进销售。啤酒和尿不湿的关系，你可以有各种解读，但是如果不是通过数据挖掘，可能打破脑袋也想不出它们之间会有关系。在商业环境中，如何解读这种关系并不重要，重要的是它们之间只要存在关联，就可以进行**关联分析**，最终目的是让用户尽可能看到想购买的商品。

除了商品和商品有关系，还可以利用人和人之间的关系推荐商品。如果两个人购买的商品有很多都是类似甚至相同的，不管这两个人天南海北相隔多远，他们一定有某种关系，比如可能有差不多的教育背景、经济收入、兴趣爱好。根据这种关系，可以进行关联推荐，让他们看到自己感兴趣的商品。

更进一步，大数据还可以将每个人身上的不同特性挖掘出来，打上各种各样的标签：90 后、生活在一线城市、月收入 1~2 万、宅.....这些标签组成了用户画像，并且只要这样的标签足够多，就可以完整描绘出一个人，甚至比你最亲近的人对你的描述还要完整、准确。

除了商品销售，数据挖掘还可以用于人际关系挖掘。你听过“六度分隔理论”吗，它认为世界上两个互不认识的人，只需要很少的中间人就能把他们联系起来。这个理论在美国的实验结果是，通过六步就能联系上两个不认识的美国人。也是基于这个理论，Facebook 研究了

十几亿用户的数据，试图找到关联两个陌生人之间的数字，答案是惊人的 3.57。你可以看到，各种各样的社交软件记录着我们的好友关系，通过关系图谱挖掘，几乎可以把世界上所有的人际关系网都描绘出来。

现代生活几乎离不开互联网，各种各样的应用无时无刻不在收集数据，这些数据在后台的大数据集群中一刻不停地在被进行各种分析与挖掘。这些分析和挖掘带给我们的是美好还是恐惧，依赖大数据从业人员的努力。但是可以肯定，不管最后结果如何，这个进程只会加速不会停止，你我只能投入其中。

大数据应用的机器学习时代

我们很早就发现，数据中蕴藏着规律，这个规律是所有数据都遵循的，过去发生的事情遵循这个规律，将来要发生的事情也遵循这个规律。一旦找到了这个规律，对于正在发生的事情，就可以按照这个规律进行预测。

在过去，我们受数据采集、存储、计算能力的限制，只能通过抽样的方式获取小部分数据，无法得到完整的、全局的、细节的规律。**而现在有了大数据，可以把全部的历史数据都收集起来，统计其规律，进而预测正在发生的事情。**

这就是机器学习。

把历史上人类围棋对弈的棋谱数据都存储起来，针对每一种盘面记录如何落子可以得到更高的赢面。得到这个统计规律以后，就可以利用这个规律用机器和人下棋，每一步都计算落在何处将得到更大的赢面，于是我们就得到了一个会下棋的机器人，这就是前两年轰动一时的 AlphaGo，以压倒性优势下赢了人类的顶尖棋手。

再举个和我们生活更近的例子。把人聊天的对话数据都收集起来，记录每一次对话的上下文，如果上一句是问今天过得怎么样，那么下一句该如何应对，通过机器学习可以统计出来。将来有人再问今天过得怎么样，就可以自动回复下一句话，于是我们就得到一个会聊天的机器人。Siri、天猫精灵、小爱同学，这样的语音聊天机器人在机器学习时代已经满大街都是了。

将人类活动产生的数据，通过机器学习得到统计规律，进而可以模拟人的行为，使机器表现出人类特有的智能，这就是人工智能 AI。

现在我们对待人工智能还有些不理智的态度，有的人认为人工智能会越来越强大，将来会统治人类。实际上，稍微了解一点人工智能的原理就会发现，这只是大数据计算出来的统计规律而已，表现的再智能，也不可能理解这样做的意义，而有意义才是人类智能的源泉。按目前人工智能的发展思路，永远不可能出现超越人类的智能，更不可能统治人类。

小结

大数据从搜索引擎到机器学习，发展思路其实是一脉相承的，就是想发现数据中的规律并为我们所用。所以很多人把数据称作金矿，大数据应用就是从这座蕴含知识宝藏的金矿中发掘中有商业价值的真金白银出来。

数据中蕴藏着价值已经是众所周知的事情了，那么如何从这些庞大的数据中发掘出我们想要的知识价值，这正是大数据技术目前正在解决的事情，包括大数据存储与计算，也包括大数据分析、挖掘、机器学习等应用。

美国的西部淘金运动带来了美国的大拓荒时代，来自全世界各地的人涌向美国西部，将人口、资源、生产力带到了荒蛮的西部地带，一条条铁路也将美国的东海岸连接起来，整个美国也随之繁荣起来。大数据这座更加庞大的金矿目前也正发挥着同样的作用，全世界无数的政府、企业、个人正在关注着这座金矿，无数的资源正在向这里涌来。

我们不曾生活在美国西部淘金的繁荣时代，错过了那个光荣与梦想、自由与激情的个人英雄主义时代。但是现在，一个更具划时代意义的大数据淘金时代已经到来，而你我正身处其中。

思考题

通过统计历史数据的规律进行机器学习，这样的例子还有很多，比如统计人的驾驶行为进行机器学习，就是无人驾驶；统计股票的历史交易数据进行机器学习，就得到量化交易系统。你还能想到哪些可以进行机器学习的例子？

欢迎你写下自己的思考或疑问，与我和其他同学一起讨论。

从 0 开始学大数据

智能时代你的大数据第一课

李智慧

同程艺龙交通首席架构师
前 Intel 大数据架构师



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 预习 01 | 大数据技术发展史：大数据的前世今生

下一篇 预习 03 | 大数据应用领域：数据驱动一切

精选留言 (67)

写留言



小千 置顶

2018-11-02

26

啤酒与尿布这个经典案例广为流传，但是我专门咨询了沃尔玛超市的人员，他表示这个案例是虚构的，是当年ibm为了卖天价解决方案而编出来的。

实际上，我专门去考察了不下100个超市，没有一家把啤酒喝尿布放在一起。数据挖掘如何在传统企业落地，是非常艰难复杂的过程。

作者回复：刚刚在京东沃尔玛旗舰店搜啤酒，哈尔滨啤酒详情页下方推荐六件商品，两件尿不湿。

传统零售商没有很好的货架策略摆放关联商品，这可能也是传统零售商没落的一个原因。



Shy

2018-11-01

38

统计大家p图的参数进行智能美颜

展开

作者回复: 赞



o°cboy

2018-11-01

27

搜索引擎》数据仓库》数据挖掘》机器学习
总结的很好



虎虎

2018-11-01

21

人工智能超过人类智能不要太简单。。。人类受人工智能/算法支配的例子也比比皆是。比如曾经有新闻表示，有人跟随地图智能导航，把车开进海里的。随着智能的发展，你我可能会因为习惯或者依赖智能的服务，而丧失了某些能力。比如自动驾驶出现可能会导致没人会开车了。这等于把命交给了智能，毕竟算法会替你决定撞别人(一群幼童)还是牺牲自己。对于人工智能是否会统治人类，以现在对人工智能的理解，应该还没有这个可能。...

展开

作者回复: 有道理，非机械方式运作的AI，不能机械地下结论。



想飞就飞

2018-11-15

15

未来的软件开发不再是需求-分析-设计-实现的确定性过程，而是定义问题和目标，收集数据，提供数据，再由神经网络不断探索最优解的非确定性过程。

作者回复: 赞



拿笔小星

2018-11-04

👍 10

第一知道AI，是大学里接触了一款叫“DOTA”的游戏，里面有张人机对战图，地图名字会在结尾被标注AI。当时还不知道AI的意思，后来才知道人工智能啊。现在AI运用在游戏里又火起来了。他也是统计了世界各地顶级高手的数据，完成英雄操作和对战！



妖精的盒子

2018-12-03

👍 9

在老家一直和合伙人运营着一款类似于淘宝客的机器人，可以聊天可以看电影可以购物智能推荐，所以，去培训班学习了大数据，希望在此基础上不断积累，然后学习数据挖掘还有机器学习。谁说女孩子不需要成就感和事业的。。。



江

2018-11-02

👍 5

之前用过一款应用 微软识花，由微软亚洲研究院和中科院植物研究所合作开发，据说研究了植物所提供的几百万张花的照片，对机器识别模型进行训练，才达到拍张照就可以识别花，了解花的详细介绍。

展开 ∨

作者回复: 👍



贾洵

2018-11-02

👍 4

上一代互联网革命是电脑时代，是人找物，即搜索。下一代互联网革命是移动时代，是物找人，即推荐。必然离不开大数据与人工智能相结合。所以任何时候学习都不算晚

展开 ∨

作者回复: 确实如此



Beckwin

2018-11-05

👍 3

听一个清华教授的演讲其中对人工智能的寄语，挺好的，她说“人工智能对人类不是replace，而是be a partner”。

大数据，人工智能确实给我们现代生活带来了很大便利，但很多地方也侵入生活太深，头条的智能推荐阅读，淘宝小红书的推荐购物，有些时候反而有点适得其反让我们陷入到人工智能给我们的范围和圈子中，限制了我们思维，它们拼命推荐给我们喜欢和想要看到...
展开 ▾

作者回复: 是的，这也是各类推荐系统需要克服的缺点，也有各种尝试



会飞的鱼

2018-11-03

👍 3

这样的技术也可以用在医疗领域上，通过对病人病例的统计，得到最理想的特征相关关系，从而得出最可能的病情

作者回复: 是的



yaw

2018-11-02

👍 3

我觉得把正确的合理的算法理论用于便捷生活，即使人的主导地位下降了，但是社会还是在进步中。说到底人只是存在于地球上的一个物种，过分的追求人的主宰地位会导致技术、社会发展遇到瓶颈，世界还是需要一些开创先河的人物。

作者回复: 境界太高👍



庆增

2018-11-02

👍 3

学了专栏之后我可以成为一个大数据从业者吗?

展开 ▾




Heshher

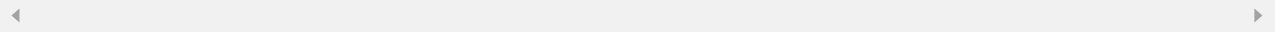
2018-11-02

👍 3

推荐系统、广告系统、估价系统、风控系统都是现在广泛使用了机器学习的

展开 ▾

作者回复: 



技术小工

2018-11-01

 3

现在出现了很多在网上爬数据分析的

展开 ▾



小千

2018-11-05

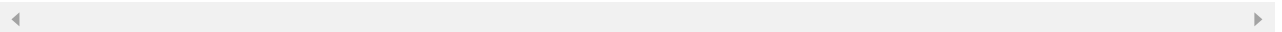
 2

传统零售，货架摆放不可能像互联网页面一样做到对每个消费者实时调整。只能把所有消费者看做是一个整体来处理。对于大多数人来说，把啤酒喝尿布放在一起并不能促进购买欲

还是举啤酒喝尿布的例子，我刚才在京东搜索啤酒，给我推荐是垃圾袋，和给你推荐的尿布不同。这就是互联网做产品展示优于传统零售的地方，传统零售不可能来一个人就把...

展开 ▾

作者回复: 对



wmz

2018-11-01

 2

数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。

展开 ▾



刘胜

2019-01-18

 1

我觉得在制造业领域引入大数据，通过分析海量数据优化公益流程，提高效益。类似于工业4.0。看完老师的文章，我有一种真切的感受，大数据是一个大的潮流。现在是2019年1月，起帆远航。

展开 ▾



 1



2018-12-19

1

输入法收集我的输入信息，根据我输入的信息预测下一个输入

展开 ∨



杰之7

2018-11-22

👍 1

统计过去在网上的商品浏览进行机器学习，就有了亚马逊的智能推荐物品；蚂蚁金服上统计过去的消费能力和信贷进行机器学习，来预测可在借呗上借多少钱，这是一个大数据和机器智能的时代，我们身处其中，需要的是适应并不断学习前行才不会在一段接一段的浪潮中被退去。

作者回复: 是的

