中 国 数 据 分 析 行 业 核 心 刊 物

数据分析

CHINA DATA ANALYSIS 用数据说话 · 做理性决策

++ 中国商业联合会数据分析专业委员会 主办 ++

《中国数据分析》会员特刊
2017年第02期 总第30期（季刊）
咨询热线：010-59000991 / 59000339
http://www.chinacpda.org/
投稿请至 xiehui@chinacpda.org

/ 大数据，走下神坛 /

/ 大数据，走下神坛 /

随着数据的可获得性和流动性日益增强，大数据不仅成为与资本、土地、技术、人力资源等共同推动经济发展的要素，而且也成为提升全要素生产率和决定未来现代化水平的关键要素，也正在催生出以互联网和实体经济深度融合发展为内涵的新经济形态。

数据全球战略布局全面升级，大数据发展已处于从概念推广到全面落地的重要转折期，各国发布各种战略措施，积极推动大数据的发展。同时大数据的发展也面临着部分领域建设过热、数据开放进展滞后、制度建设尚不完善、安全管理存在漏洞、人才资源储备不足等突出问题。

目前市面上常见的大数据产品商业模式都是企业对企业的B2B模式。B2C的产品是非常少的。B2B产品是谁在使用？B2B产品是企业客户在买单，但是到最终使用环节，终究还是个人用户在使用。既然是个人用户在使用，那么产品形态和应用场景以及用户体验，还是要围绕着个人用户来做。

用户是不会去关注你背后用了什么技术？搭了多少服务器？用了什么架构？是Hadoop还是Spark又或者是Hive？用户的关注点在于你的产品是否方便快捷？是否有用？是否有趣？是否实用？是否能够节约时间？

就像是我们常用到的微信支付和支付宝支付一样，用户只关注的是我的钱是否安全？我使用的时候是否方面？我是否能够随时随地使用？我用移动支付能做些什么事？诸如此类。用户更不会去关注腾讯和阿里为移动支付架了多少台服务器，使用什么技术架构，用的是什么数据库等等。

任何脱离了用户需求和应用场景的技术，都毫无价值。你需要找到用户的痛点在哪里，你需要找到明朗的商业模式。无论是B2B也好，B2C也好，企业 / 个人只会为“价值”买单。

简单来说，大数据技术再怎么强大，但是对于企业 / 个人没有实际的价值，他们是不会买单的。最终到用户使用的时候，用户关注的还是产品形态和用户体验。

大数据的闭环是从业务中来，到业务中去。在这过程中，大数据起到的是连接 / 串联 / 支持 / 展示 / 结果 / 支撑 / 辅助的作用。大数据更多需要的是接地气，以人为本，以业务为导向。

所以，科技还是要以人为本，大数据更多的是通过深度分析接地气，数据分析师们，到大家发力的时机了！

中国商业联合会数据分析专业委员会

中国数据分析 CHINA DATA ANALYSIS 用数据说话 · 做理性决策

《中国数据分析》 会员特刊 2017年第二期 总第30期

本期目录 CONTENTS

卷首语
01 大数据，走下神坛

协会动态
03 “新起点，大不同”CPDA新课程体系品鉴会圆满落幕
04 “用数据改变人生
——CPDA2017新课程体系分享沙龙完美落幕

数博专题
06 用大数据，享新价值
07 BAT齐聚数博会，大佬们都说了这些
08 “工业大数据与智能制造”高峰对话举行
09 挖掘大数据应用价值，把大机遇变成大红利
11 2017中国电子商务创新发展峰会《贵阳共识》
13 2017中国国际大数据产业博览会：群英荟萃，成果丰硕

政策导向
14 三部委印发《智慧健康养老产业发展行动计划》
17 挖掘数据价值会引发工业革命？

人才培养
19 数据分析躲坑秘籍
21 分分钟搞掂矩阵“特征值”要表示什么“特征”
22 如何用Python玩转TF-IDF之寻找相似文章并自动生成摘要

数业专攻
24 大数据浪潮下，前端工程师眼中的完整数据链图
28 零基础初识：搭建Hadoop大数据处理

运数有道
30 大数据解密:《人民的名义》是怎么火起来的?
32 男子利用人脸识别技术，被拐27年后与亲生父亲相认

事务所风采
34 陕西智诚数据分析师事务所

主办单位
中国商业联合会数据分析专业委员会
编委成员
袁硕 、李缘
出版时间
2017年第2期 7月出版
美工设计
崔峻珩

联系我们
中国商业联合会数据分析专业委员会
地址: 北京市朝阳区朝外soho C座9层
电话: 010-59000991 / 010-59000339
传真: 010-59000991转 607

投稿
欢迎广大读者踊跃投稿，内容包括学术观点、教学体验、教学活动、学习感悟、实战经验、随笔文章等。 稿件附图格式为JPG或TIFF格式，大于1M，分辨率在300dpi以上。

感谢您对《中国数据分析》的支持！

投稿邮箱：xiehui@chinacpda.org

/ “新起点，大不同”CPDA新课程体系品鉴会圆满落幕 /

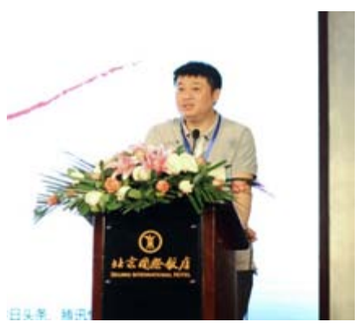
文 / 协会市场处 冯雪 编辑 / 协会会员处 袁硕 日期 / 2017-04

4月12日，中国商业联合会数据分析专业委员会主办的“新起点，大不同”CPDA新课程体系品鉴会于北京建国国际会议中心圆满落幕。与会人员来自全国二十多个省市的专家、学者及大数据行业从业人员。

此次品鉴会采用非常规的分享模式，是用视频串起整场活动，多维立体展示了产品研发的初衷、新课程体系的强大落地性、大数据行业唯一垂直社群以及全面导入Python的Datahoop专业大数据分析平台。

自从大数据战略上升为国家战略，各行业争先恐后和“大数据”扯上关系。中国商业联合会数据分析专业委员会作为中国数据分析行业唯一的全国性行业组织，致力于帮助大数据行业健康有序发展。经过多方筹备，多次验证，提出建立大数据人才全新培育体系。

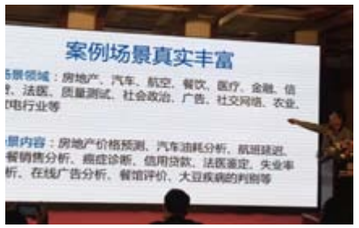
会长邹东生先生为品鉴会致辞，他和大家分享了大数据行业的现状、数据分析人才的需求缺口及从业要求。对行业的发展提出了宝贵的建议，邹东生会长倡导用分析引导大数据落地，这正是此次大数据课程体系的核心理念。



CPDA数据分析师培训体系倡导用8天面授+远程学习+直播互动学习+垂直社群讨论+免费开放的大数据分析平台相结合的方式，帮助数据分析师持续学习，真正实现价值提升。

想要成为一名优秀的数据分析师，甚至实现百万年薪，比取得CPDA数据分析师证书更重要的是要在数据分析行业中不断历练，用专业的数据分析能力为企业带来实际的价值。

主创团队用蓼虫忘辛的情怀并肩前行。（蓼虫忘辛：吃惯了蓼这种有辣味的草的虫子已经不感到蓼是辣的了。比喻人为了做自己喜好的事就会不辞辛苦。）在一次次的探讨中凭着对这个行业的热爱和最初的梦想，用泪水和汗水凝结成了CPDA2017新课程体系。



专家学者们结合行业用人需求，全面导入落地型实操案例。每个案例都极具代表性，实际的应用场景包括房地产、汽车、航空、餐饮、医疗、金融等诸多领域。旨在为大数据行业输送专业能力强、实操水平高的优秀数据分析师。做到让数据分析师学以致用，活学活用！下图为中国科学院理学博士祝捷老师，介绍Datahoop大数据分析平台：

协会自主研发的Datahoop大数据分析平台免费对数据分析师开放，为数据分析师们提供了丰富的应用大数据场景的实战机会，大数据分析理论与大数据计算能力的完善结合，使数据分析师们在未来的职场上可以直接叱咤风云。

自动绘制词云、统计词频道，能够使模型结果信息显示更清晰、模型结果图展示更美观、模型参数交互更简洁而且还支持全数据处理流程，如此强大的处理平台还是免费的，参会的小伙伴们都跃跃欲试。

大数据行业唯一垂直社区，拥有在线直播、在线问答、数据资料下载等强大功能。

让我们共同开启CPDA2017华美篇章吧！

FIN

/ “用数据改变人生 ——CPDA2017新课程体系分享沙龙完美落幕 /

文 / 协会市场处 冯雪 编辑 / 协会会员处 李媛 日期 / 2017-05

我们和数据有多少种关系，你的密码、他（她）的电话、母亲的生日、父亲的车牌、这个月的对账单、下一年的KPI……既然选择了这行，就多结交些同行。既然一直坚持这行，就该让自己持续领先。

中国商业联合会数据分析专业委员会倡导用分析引导大数据落地，结合行业发展和企业用人需求，耗时半年，完成了第7次课程改革……

新课程体系强大的实战性和落地性，将会使CPDA学员在毕业那一天，就具备专业技能和实战经验，2017年5月21日，CPDA新老学员欢聚在北京东郎影视文化产业园娱客众筹咖啡厅，听嘉宾分享此次新课改的心路历程和全新亮点。

新老学员畅所欲言，新学员虚心求教，老学员坦诚分享经验心得。沙龙活动比原计划延时了一小时大家还意犹未

尽，活动当天大家约好，这个初夏，我们再相聚！

2017年5月21日，CPDA新老学员欢聚在北京东郎影视文化产业园娱客众筹咖啡厅。

原定9:30分开始签到，很多学员8点多就来了，说是沙龙活动其实真的像老友聚会。

沙龙活动开始全场座无虚席，很多学员是站着听完全程的。

主持人刘萍老师和大家深度分享CPDA全新课程体系，继续延续8天金牌讲师面授和覆盖全国的远程学习平台，同时增加在线直播课堂，学员随时随地进入直播间与大咖、学者、师兄互动学习；搭建大数据行业唯一垂直社群，集聚万余名数据分析师探讨新资讯、分享新方法、结交国内外同行；同时免费开放大数据专业分析平台，提供大体量数据、各种算法，供学员课后练习。

新的培训体系，旨在引导所有CPDA学员时刻保持学习能力，从毕业到就业，我们一路保驾护航。助力各位学员不断提升专业能力、实操水平、职场竞争力！

参与新课程设计的孙爽老师用网络游戏举例子，生动有趣地呈现出了课改的特点和强大的落地性。新课程全面导入实际案例场景，涉及房地产、汽车、航空、餐饮、医疗、金融、电商、通信、能源、零售等诸多领域，让学员一毕业就有实战经验。

搭建Datahoop大数据分析专业平台的领军者祝捷博士和大家阐释了平台新增的强大功能，丰富的实战场景，大体量数据练习，结合所学知识点，让学员掌握理论知识更有实战经验。

基于Python支持全数据处理流程，数据采集、预处理、综合算法分析、可视化等分析工作轻松实现一键分析。自动绘制词云、统计词频，结果显示更清晰、模型展示更美观、模型参数更简洁。

学员老张，分享多年从业经验，主动解答新学员的各种疑问，这个师兄当得没话说！

新老学员互动频繁，活动延迟一小时大家还舍不得离开！

在热烈的掌声中，此次沙龙活动圆满结束，新老学员意犹未尽。这个初夏我们继续约起来！

FIN

/ 用大数据，享新价值 /

文 / 协会会员处 袁硕　编辑 / 协会会员处 李缘　日期 / 2017-05

分析是大数据的灵魂。今天，大数据行业到处充斥的寒冬说、无用论、皇帝的新装等说法，让大家有种不寒而栗的感觉，我们到底处在一个什么样的大数据时代？

5月28日由数博组委会指导，中国商业联合会数据分析专业委员会主办，贵州云博大数据有限公司承办的“用大数据享新价值”大数据应用高峰论坛在贵阳市举行。

中国商业联合会数据分析专业委员会自成立以来，一直积极推动数据分析技术的普及和应用。

自从十八届五中全会把大数据上升成为国家战略后，各行各业都掀起了探寻大数据应用的热潮。

邹东生会长向大家介绍了大数据发展面临的种种问题，同时从数据为商业带来价值的角度与大家做了更深入的阐述，让大家意识到大数据离我们不遥远，而且大数据不昂贵，只要你想用就用得起，因才适用，可以解决很多问题；数据分析，可以少花钱多办事；真正在做大数据的人就是那些从数据背后寻找、分析规律、并赋予其商业价值的数据分析人才。

大数据走下神坛，培养大数据分析人才才是我们破题的关键。

同时犀数科技、海云数据、泰一指尚、新华三集团、亚信数据等机构的企业数据专家们针对在教育、政府、公安、电信等不同行业的应用，为大家做了精彩的演讲。

2017年中国国际大数据产业博览会在贵阳市举行。本届数博会以“数字经济引领新增长”为年度主题，由国家发改委、工信部、国家互联网信息办公室、贵州省人民政府共同主办。会议主要围绕国家大数据试验区交流、数字经济、区块链开启价值互联网时代等七个板块开展。

FIN

/ BAT齐聚数博会，大佬们都说了这些 /

文 / 财经110　编辑 / 协会会员处 袁硕　日期 / 2017-06

今年的中国国际大数据产业博览会（以下简称：数博会）期间，BAT三家第一次在同一届数博会到齐，这是继乌镇世界互联网大会、深圳IT峰会后，又一次看到BAT三家再一次聚首。然而，关于大数据时代什么最重要的问题，马云、李彦宏和马化腾却给出了不同的答案。

“由于大数据，让计划和预判成为了可能。”马云的这句话意味着其更关注数据的作用。他提出，以前渔民下海捞鱼时，由于对气象的把握不够，只能依靠老船长的经验，但是气象台出现以后，就能够准确的预判什么时候是否暴风雨，使得捕鱼的计划性就有可能出现，其中的气象就代表数据。

马云认为，数据还将使计划经济有迹可循。“大数据时代，人类获得数据的能力远远超过想象，人类取得对数据进行重新处理以及处理的速度的能力也远远超过想象，我们对世界的认识将会提升到一个新的高度，大数据会让市场变得更加聪明。”

为强调数据的重要性，马云还以能源与其进行类比，“第一次工业革命的主要能源是煤，诞生的商业模式是工厂；第二次工业革命主要能源是石油，诞生的行为是公司；这一切皆是创新。“马云认为，数据将成为未来主要的能源，如果离开了数据，任何组织的创新都基本上是空壳。”

相比于马云的“数据论”，李彦宏则在数博会期间的“人工智能”高峰论坛上提出了“数据和技术，或者和算法到底是什么关系？”的问题，而他给出的答案是“数据确实重要，没有数据训练的话，人工智能走不到今天，但是数据不是根本，推动时代进步的是技术和创新。”

事实上，这一看重技术多于数据的观点也与李彦宏在几天前的百度联盟峰会上的表态类似，当时，李彦宏谈到，足够多的数据，即使算法稍微差一点，得出的结果虽然也不会错，但是，真正推动社会进步的是算法，而不是数据。

李彦宏还以煤与蒸汽机的关系类比数据与技术的关系，“工业革命时期最宝贵、最具标志性的东西不是煤，而是蒸汽机这样的技术革命、革新；所以，人工智能时代最宝贵的也不是数据，而是数据所带来的技术和创新。”

然而，李彦宏也提出了他的担心，即现阶段，技术在很多时候呈现指数式增长，但是人的思维方式在大多数时候却停留在线性增长，如果按照现有的思维方式，很难跟上未来技术的变化。

“最近我在讲要培养AI思维，这样的思维方式不是我们习惯的思维方式，技术革命带来不断的可能性，我们需要极早为未来的这种可能性做准备。”

李彦宏认为，未来数据资源需要变成创新能力，这种创新来源于数据及场景，需要把数据组织起来、把场景吃透，遇到问题解决问题，就产生了创新，“这样的创新又会不断地培养我们的思维方式，跟得上未来科技的进步。”

然而，同在数博会，BAT三家并没有同时出现在一个舞台上，对于李彦宏和马云的观点，马化腾则进行了“隔空”点评，“我相信李彦宏谈的是从0到1，需要由创新技术驱动；马云讲的是从1到N，这个过程需要持续不断的数据驱动；所以，他们谈的是不同阶段。”

不同于马云和李彦宏的观点，马化腾认为，未来互联网发展，更重要的一个要素是“场景”，或者可以称之为“战场”，再通俗一点就是“市场”。“我觉得这是最关键的，有了应用场景，有了市场，数据自然会产生，也会驱动技术发展，人才也会随之而来。”

马化腾表示，从不可复制性的角度来说，计算能力和大数据都是可复制的，但是市场和人才是不可复制的。“我觉得这是一个核心点，就好像今天BAT三家分别在社交、电商和搜索有各自的主战场和场景；滴滴、摩拜有交通出行的场景；微信、支付宝有支付场景。”

马化腾认为，有了各种场景，未来人工智能时代，就可以借助新技术，从而把握先机，所以最关键的还是场景，否则空有技术、空有数据是远远不够的。

值得注意的是，马化腾在演讲中还强调了云在数字经济中的作用。他认为，一个行业的云化程度，是该行业数字经济发展程度的重要指标。“工业革命之后，电力为经济及社会带来深刻改变，人们通常把用电量作为衡量经济好坏的重要指标；数字经济时代，‘用云量’将会成为一个重要的经济指标。”

虽然BAT三家对大数据理解的侧重点各不相同，也很难定论孰是孰非；但是，可以确定的是，大数据时代已经到来，这种趋势只能顺应而不能“逆行”，正如李彦宏所说，“需要极早为未来做准备。”

/ “工业大数据与智能制造”高峰对话举行 /

文 / 贵州人民政府 编辑 / 协会会员处 李缘　日期 / 2017-06

5月26日，“工业大数据与智能制造”高峰对话举行,省委副书记、省长孙志刚出席。

富士康科技集团创始人、总裁郭台铭,全国工商联副主席、传化集团有限公司董事长徐冠巨,浪潮集团党委书记、董事长兼首席执行官孙丕恕,腾讯公司高级执行副总裁汤道生,SAP企业管理解决方案公司副总裁、首席科学家Teodoro Dennis Pristley,深圳市怡亚通供应链股份有限公司董事长兼CEO周国辉,省政协副主席李汉宇,省政府党组成员任湘生等出席。

郭台铭发表演讲,解读了富士康的大数据智能制造新时代。随后,嘉宾们围绕“大数据与智能制造引领行业变革”主题进行了对话,畅谈了变革带来的商机、工厂的变化、生产流程的变化以及各自关心的话题。

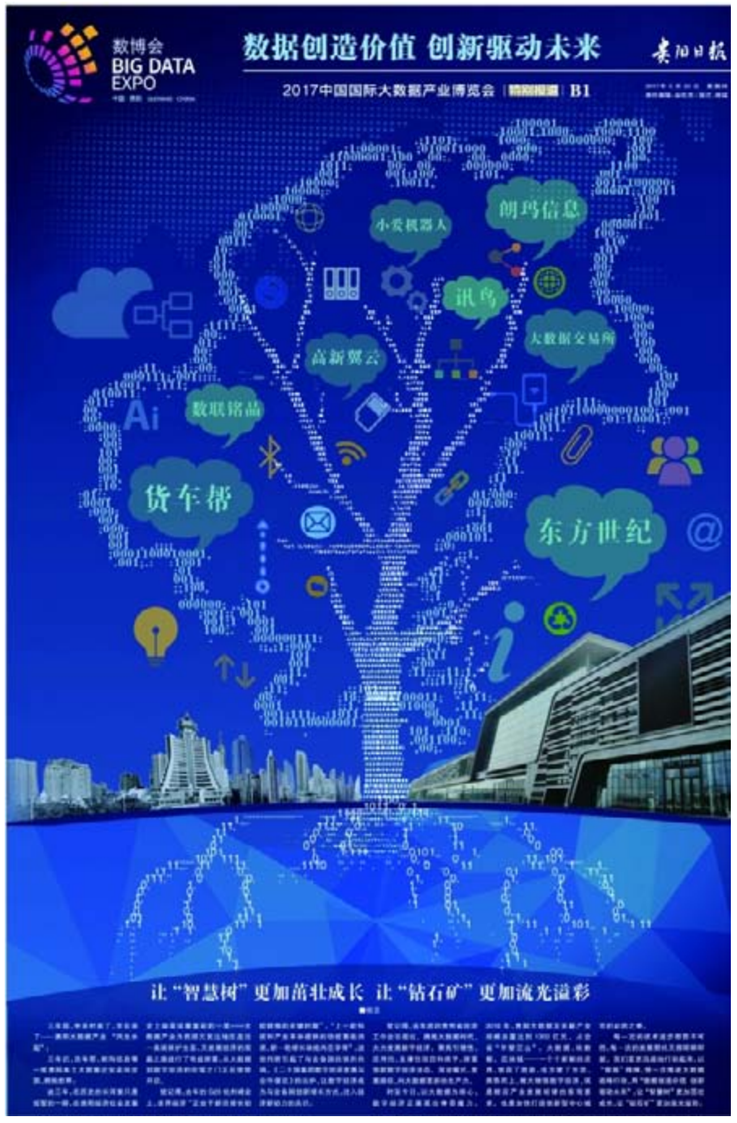
本场活动由国家发改委、工信部、国家互联网信息办公室、贵州省政府主办,来自国内外的70余名行业嘉宾等参加了活动。

中央电视台财经频道《对话》栏目组对对话现场进行了录制。

FIN

/ 挖掘大数据应用价值，把大机遇变成大红利 /

文 / 贵州人民政府 编辑 / 协会会员处 袁硕 日期 / 2017-05



5月24日，提升政府治理能力大数据应用技术国家工程实验室在贵阳国家高新区揭牌成立，这是全国首个大数据领域国家工程实验室，也是贵州第一个国家工程实验室。实验室将致力于为推动政府治理大数据应用的技术进步、产业转型和产业发展提供技术支撑。

实验室的成立，意味着贵阳在大数据应用领域又迈出坚实一步，在贵州大数据发展史上具有里程碑意义。

大数据既是大机遇、大变革，又是大产业、大红利。作为贵州省会，贵阳是“国家大数据(贵州)综合试验区”建设的主战场。在大数据领域，贵阳并不只是探路者、理论者，更是耕耘者，让大数据落地生根，把大机遇变成大红利。

从数据铁笼、大数据综合治税等政务应用，到货车帮、贵阳大数据交易所等商业应用，再到“筑民生”、智慧医疗等民生应用……“接地气”的应用，让大数据真真切切为这座城市带来活力和动力。

去年12月底发布的《贵阳区块链发展和应用》白皮书，明确贵阳发展区块链的顶层设计，提出通过5年努力，建成主权区块链应用示范区的目标。白皮书明确，贵阳市将从经济社会发展的痛点出发，提出区块链的政用、民用和商用应用场景。

随着一批批云应用示范工程的落地生根、开花结果，在贵阳，大数据正在变革政务管理模式、传统产业模式，深刻改变市民的生活方式。

大数据提升政府治理能力

大数据的发展，为提升政府治理能力提供了全新契机，贵阳将大数据思维和手段运用作为政府治理能力提升的最重要、最有力推手。

“数据铁笼”是其中最具亮点的应用之一。

近日，贵阳市运管局信息中心主任陆宇的手机收到一条短信，内容为：窗口办理的一项从业资格证事项，经“数据铁笼”研判为证件异常发放。这条信息显示，工作人员想给熟人开个“人情证”，“数据铁笼”平台核查培训系统后发现“查无此人”，自动研判为“异常数据”。在市运管局内部，这个推送信息就是执法者的“红色警报”。工作人员如果不能在3天内合理说明，信息就将被推送给上级领导、分管局长，直至7天内，最终由局纪检监察室启动问责。

让权力运行全程电子化，处处留痕，置于公众的监督之下，如今，这个大数据“管家”，正在贵阳市运管局内发挥着巨大的“管事能力”：无论是“人”的行为、“事”的过程，还是“权”的运行，所有轨迹都在这个“管家”的掌握之中，构建了“不敢腐、不能腐、不想腐”的监督机制。

在大数据思维引领下，贵阳市以数据政用创新推动管理流程再造，探索运用大数据提升政府治理能力，实施“数据铁笼”行动计划，围绕数据在哪里、问题在哪里、办法在哪里三个问题，以部门自身业务流程和权责规范为出发点，利用大数据云计算等先进技术编织制约权力的笼子，形成用数据说话、用数据决策、用数据管理、用数据创新的大数据工作机制。

从“数据铁笼”到“数据铁笼+电子政务”平台，贵阳市不断完善“网上办事大厅”等公共服务大数据应用，“让数据多跑路，群众少跑腿”，提升城市公共服务水平；

从生态大数据，到精准扶贫区块链应用，贵阳市从经济社会发展的痛点出发，不断提出区块链的应用场景，破解痛点；

从“块数据指挥中心”、“智慧门牌”数据管理平台，到智慧城管、智慧交通等云应用示范工程……贵阳市以大数据创新运用为支撑，破解社会治理难题，已初显成效。

利用大数据，贵阳市不断助推政府决策科学化，助推从经验决策转向数据决策，从全面治理转向数字治理，不断提升政府治理能力。用数据说话、用数据决策、用数据管理、用数据创新，释放出城市发展的更大活力。

大数据引领经济转型升级

大数据真正的价值在于应用。如今，随着贵阳大数据产业的快速发展，越来越多的大数据应用落地生根，见证着贵阳这片大数据沃土的日新月异。

跑了十多年货车的司机蒲杰，以前常常为找货难发愁，吃睡在车上，生怕错过货主。2013年，这样的情况有了改变。这一年，货车帮APP上线。第一次使用，蒲杰就通过APP联系到了重庆回贵阳的货源，没有跑空车，多赚了3000多元。

在贵阳崛起的货车帮，成立的初衷就是为了解决物流信息不对称这一痛点。通过搭建开放、透明、诚信的货运信息平台和社会“公共运力池”，货车帮有效降低了公路物流运输成本，已成为全国最大的公路货运信息平台。

借助贵阳大数据发展的优势，货车帮、朗玛信息、东方祥云等一批本土企业迅猛成长，其中货车帮已成为中国物流行业的“独角兽”企业，目前拥有35万家物流企业会员和230万货车会员，日均平台结算金额80亿元左右；朗玛信息打造了全国首家线上线下相结合的互联网医院，入围“2016年中国互联网企业100强”。

作为大数据领域的探索者，贵阳还成立全国首家大数据交易所，截至今年一季度，交易金额突破1.2亿元，交易会员达到500家，可交易的数据总量超过60PB，接入30个行业领域的数据。

甲骨文、谷歌、英特尔、微软、IBM、惠普、戴尔、富士康、思爱普等世界500强企业先后落户贵阳，阿里巴巴、腾讯、奇虎360等国内互联网领军企业纷纷牵手贵阳。这势必将衍生出更多的大数据应用业态，助推贵阳大数据产业快速发展。

经过几年发展，贵阳大数据各类业态加快聚集，一批新技术、新产品、新模式不断涌现，初步构建了从数据存储、清洗加工、数据安全等核心业态到电子信息制造、软件和信息技术服务等关联业态，再到服务外包与呼叫中心、电子商务、精准营销、大数据金融等衍生业态的产业链条……

在贵阳，大数据融入全社会各个单元、各个细胞中，带动产业转型升级，绘就一幅加速发展、加快转型的美丽“云图”。

大数据服务民生惠及市民

在贵阳，只需要一款手机APP，就可以足不出户在掌上办水电燃气生活缴费、公积金查询、违章查询等与群众生活息息相关的事情。

“查询社保、公积金，再也不需要专门跑一趟办事大厅了，拿起手机一点，全部搞定，真是太方便了。”如今，越来越多的市民受益于“筑民生”便民服务平台，不用再多跑冤枉路，办事越来越方便。

今年4月，“筑民生”平台上线之

初，便整合了市政务服务中心、市人社局、市交管局、市教育局、市公积金中心等十余个部门的资源，融合推出生活账单、医疗保健、就业服务、户籍证照、安居乐业、出行服务六大类便民服务，市民只需通过手机APP便可“一站式”享受近百项服务。经过几个月的持续努力，平台上线的服务已增至152项。

发展经济的最终目的，是保障和改善民生。在探索大数据蓝海的同时，贵阳决定走出一条“接地气”的道路，让大数据走近群众，走进生活，让市民共享大数据发展成果。大数据、块数据、数字经济……这些词语放在几年前，公众闻所未闻，但在今天的贵阳，却实实在在地与老百姓日常生活紧密相连。

在贵阳城区，市民用智能手机登录全域免费WIFI“D-Guiyang”，即可上网冲浪。每天产生的商业、社会、政府、人文等各个领域的大量行为数据，通过这张WIFI网，实现“块”上集聚。

守护学生“舌尖上的安全”，“阳光厨房”APP目前已在贵阳市属16所学校试点运行。在这些试点学校，后厨成为“透明厨房”，食品加工流程随时在“天眼”的监督之下，食堂存在的食品安全问题可随时被发现和整改。

通过“智慧门牌”，民警扫描二维码就能知道房主信息，实现警务信息采集专业化和社会化。

“社会和云”通过资源整合，将涉及社会治理和群众工作的相关数据汇集到管理平台，实现社会治理“管理扁平化、工作服务化、网格具体化、业务综合化”。

如今，在贵阳，大数据发展正从风生水起转入落地生根，给民众思维模式和生活方式带来显著改变。大数据广泛应用在民生服务中，让越来越多市民享受到大数据产业发展带来的红利。

FIN

/ 2017中国电子商务创新发展峰会《贵阳共识》 /

文 / 贵州人民政府 编辑 / 协会会员处 李缘 日期 / 2017-06

2017年5月25日至28日，2017中国电子商务创新发展峰会在贵阳举行。

峰会由国家发展和改革委员会、贵州省人民政府主办，中央网信办、司法部、农业部、商务部、中国人民银行、海关总署、国家税务总局、国家工商总局、国家质检总局、国家统计局、国家林业局、国家邮政局、国家旅游局、国家标准委等14个部委支持，中国网络电视台、贵阳市人民政府承办。

来自全国31个省市自治区70多个城市100多家企业、12家研究机构和科研院所，以及俄罗斯、澳大利亚、美国、韩国等国内外嘉宾，共计3000多人相聚贵阳，围绕国家重大战略部署，探讨电子商务创新发展之路。

本次峰会以“聚合创新要素、赋能实体经济”为主题，由开幕式主论坛和国家电子商务示范城市工作交流会、产业电商、电商创新要素、跨境电商行业新布局、农业电商、反侵权假冒、品质电商和电商物流等八个主题分论坛组成。

与会者一致认为，电子商务是数字经济的重要组成部分，电子商务已经成为我国经济增长的新引擎。我国已经成为全球规模最大、发展速度最快的电子商务市场，电子商务促进传统产业转型升级，跨境电商成为外贸增长新希望，农村电商探索出电商扶贫新模式，电商“双创”催生出规模化就业新领域，民生电子商务推动公共服务创新。同时，电子商务快速发展也带来区域发展不平衡、新旧市场主体竞争不公平、线上线下市场秩序不同步等一系列问题，这些新现象和新问题需要政府、企业和社会共同努力、实施有效的治理。

2017年是电子商务提质升级和赋能实体经济之年。与会者一致认识到，中国电子商务创新发展需要政府、企业和社会共商、共建、共治和共享，达成以下六点共识：

第一，要发挥电子商务在政府投资结构调整方面的引导作用，加快建设电子商务信息基础设施。

健全电子商务发展支撑体系，重点开展电子商务综合管理平台、电子商务公共信息服务平台和电子商务商业信息服务平台的建设和发展工作，加强公共服务，降低电子商务运行成本。围绕电子商务共性过程和关键环节基础信息，建立采集、管理、维护、共享的机制与设施，逐步形成电子商务市场完整、准确、实时、动态的经济信息资源，建设电子商务经济运行监测平台，提升电子商务市场运行效率，辅助政府相关部门宏观调控决策。

第二，要发挥电子商务在供给侧结构性改革方面的促进作用，推动电子商务进入新一轮高速发展阶段。

电子商务是实体经济活动的网络化表现形式，在农业、工业和服务业领域，电子商务应进一步向消费、流通、生产过程深度渗透和融合，利用电子商务提高供给的效率与质量、适应性和灵活性，助推供给侧结构性改革。建立适应电子商务创新创业需求的新载体，重点培育线上线下融合发展、跨境电子商务、社交电子商务、电子商务促进县域经济、“电商扶贫”等新模式、新业态，加快电子商务商业模式、科技水平及市场组织方式创新，推动电子商务发展提质升级。

第三，要发挥电子商务在打造数字经济产业链和生态链方面的协同作用，加快发展电子商务要素市场。

培育和健全电子商务要素市场，大力发展电子商务人才和信息服务业、技术服务业、产业载体及物流服务业、金融及支付服务业，发挥电子商务对金融、物流、科技等要素产业的升级创新作用，带动电子商务上下游产业互动、协同、融合发展。政府相关部门应把建立与完善电子商务要素市场作为职责和使命，引导电子商务要素产业延伸发展，但要发挥市场在资源配置中的决定性作用，完善电子商务要素产业网络化运行体系。

第四，要发挥电子商务在“精准扶贫、就地脱贫”方面的支撑作用，促进区域经济协调发展。

完善农村电子商务双向流通服务体系，通过电商带动农村产业聚集，多方位、全渠道利用电子商务的新型产业链带动贫困人口实现脱贫增收，创造大量劳动力返乡创业就业的机会实现就地脱贫，依托电子商务为贫困地区提供便捷的信息资讯、远程教育和网上办事，缩小城乡社会生活差距。加速农林产品的商品化、品牌化和电商化进程，创新农产品上行的渠道和模式，发挥电子商务的辐射作用，形成纵横交错、布局合理、优势互补的县域电子商务创新发展新局面，以电子商务助推区域协调发展。

第五，要发挥电子商务在实施“一带一路”建设中的先导作用，提升电子商务对外开放水平。

围绕落实“一带一路”建设，推动与“一带一路”沿线国家重要节点城市的对点合作，推动跨境电子商务合作通道建设，带动对外贸易和产业合作；支持电子商务企业建设国际合作平台，促进国际化发展；深入推进跨境电子商务综合试验区建设，促进国内关检税汇业务协同，便利跨境电子商务发展。建设以产品、服务和资本自由流动为目标的“一带一路”网络经济带和数字丝绸之路，致力于实现沿线国家和地区间电子商务相关政策、标准及管理制度的协调、一致或互认。

第六，要发挥电子商务在数字经济发展新秩序中的引领作用，推进电子商务治理体系建设。

重视新型产业形态与传统产业形态的关系，允许新模式挑战旧业态，清除阻碍电子商务健康快速发展的体制机制、法规政策障碍。鼓励和保护市场竞争，遏制滥用市场支配地位的垄断行为。运用大数据加强和改进市场监管，推进产品追溯体系和电子商务信用体系建设，积极推动电子商务规制创新、市场治理、网络交易安全保障、绿色电子商务等工作，建立跨部门、跨地域、跨行业的政府、企业和社会协同治理机制。重视电子商务平台数据开放、信息共享引发的风险问题，维护国家经济安全。

总之，与会者一致认识到，“十三五”期间，电子商务从经济增长“新动力”进一步壮大成为“关键动力”的新趋势日益凸显，正在孕育全球经济合作新机遇。与会者一致同意，电子商务发展要贯彻落实“创新、协调、绿色、开放、共享”五大发展理念，推动电商经济进入创新发展的新阶段。与会者一致赞同，推助电子商务成为三次产业转型升级、提质增效的新动能，努力形成电子商务支撑体系与要素市场协同发展的新态势，打造全体人民共享电子商务快速发展成果的新景象，营造有利于电子商务市场规范发展的新秩序。

FIN

/ 2017中国国际大数据产业博览会：群英荟萃，成果丰硕 /

文 / 贵州人民政府 编辑 / 协会会员处 李缘 日期 / 2017-06

5月28日,2017中国国际大数据产业博览会成果新闻发布会在贵阳召开。为期4天的2017国际大数据产业博览会无论是在规模上、还是影响力都创出新高,取得了丰硕成果。

本届数博会举办了开幕式、电商峰会和5场高峰对话,以及论坛77场,展馆发布31场、新闻发布12场、系列活动15场,来自全球各地的大数据业界高管精英、专家学者、科研机构、咨询机构、中小企业负责人和企业创新者累计超过5万人参加了此次活动。

本届数博会开幕盛况引人瞩目。开幕式以突出国家大数据战略,注重科技前沿探索,运用大数据手段促进传统产业转型和实体经济融合发展为内容。据不完全统计,仅数博会开幕当天,不同媒体各渠道转载发布信息点击浏览量达到16.978亿人次。

来自20多个国家和地区的参会嘉宾,专业公众21000人参加活动；23位省部级领导出席参加数博会相关活动,马云、马化腾、李彦宏等百余名国内知名企业负责人,苹果、微软、谷歌、亚马逊、英特尔、甲骨文、IBM、戴尔、思科、高通、以太坊、新思科技、通用电气、通用汽车等世界500强企业、互联网企业和大数据企业等146位全球高管,以及白春礼、邬贺全、倪光南等18位两院院士,北京大学、清华大学、复旦大学、香港中文大学等66所国内知名高校的负责人及专家学者,哈佛大学、麻省理工大学、斯坦福大学等国外著名学府30名专家以及联合国开发计划署驻华代表处、美国全国移动通信系统协会、印度软件服务业企业协会、日本贸易振兴机构、美国CSA研究院、世界经济论坛、印度国家信息学院等知名行业协会及研究机构的负责人参加了活动。

高峰对话,引领潮流。本届数博会针对最新行业发展趋势,围绕核心嘉宾举办了“智能制造”、“机器智能”、“区块链”、“工业大数据与人工智能”、“数字经济”5场高峰对话。

专业探讨,精彩纷呈。期间,围绕国家大数据试验区交流、数字经济、区块链开启价值互联网时代、数字安全与风险控制、数据共享与开放、人工智能、智能制造7大板块,共举办了77场论坛。主委会统一主办和支持承办论坛有55场,电商峰会论坛有9场,贵州省各市州策划举办外场论坛13场,涉及承办机构364家。

前沿展示,增强体验,是本届数博会的特色之一。设专业展馆6个,参展企业共316家,其国际企业51家,超过5万人次前来观展。此外,举行了2017十大黑科技、大数据蓝皮书暨大数据发展指数、贵阳主权区块链联盟暨区块链技术蓝皮书等30余场发布活动。

国家部委、外国驻华使领馆、知名企业和贵州省主办共举办了15项活动,30多位国家部委领导和来自30个省、自治区、直辖市以及100多个城市的企业、行业、媒体等3700余名嘉宾、8000余名专业观众参加。

在数博会上,“数聚华夏 创响未来”中国数据创新行活动正式启动,活动由国家发改委、工信部、国家网信办共同指导,行业协会和8个国家大数据综合试验区共同参与组织。

今年数博会参会媒体共210家,1268人。传统媒体高度关注,央视新闻联播、新华社、人民日报、经济日报、光明日报等及时报道。网络媒体全覆盖,人民网、新华网、新浪网、网易等10家网络进行了全网直播,共计1.49亿人次在网上进行了观看。

据组委会初步统计,本届数博会共对接企业1479家,其中500强企业112家,达成签约意向项目235个,意向金额256.1亿元,签约项目119个,签约金额167.33亿元。

FIN

/ 三部委印发《智慧健康养老产业发展行动计划》 /

文 / 中国智慧城市 编辑 / 协会会员处 袁硕 日期 / 2017-05

我国正处于工业化、城镇化、人口老龄化快速发展阶段，生态环境和生活方式不断变化，健康、养老资源供给不足，信息技术应用水平较低，难以满足人民群众对健康、养老日益增长的需求。

智慧健康养老利用物联网、云计算、大数据、智能硬件等新一代信息技术产品，能够实现个人、家庭、社区、机构与健康养老资源的有效对接和优化配置，推动健康养老服务智慧化升级，提升健康养老服务质量效率水平。

为加快智慧健康养老产业发展，培育新产业、新业态、新模式，促进信息消费增长，推动信息技术产业转型升级，特制定本行动计划。

一、总体要求

（一）总体思路

牢固树立和贯彻落实创新、协调、绿色、开放、共享的发展理念，着力推进供给侧结构性改革，深入实施创新驱动发展战略，充分发挥信息技术对智慧健康养老产业的提质增效支撑作用，丰富产品供给，创新服务模式，坚持政企联动、开放融合，促进现有医疗、健康、养老资源优化配置和使用效率提升，满足家庭和个人多层次、多样化的健康养老服务需求。通过发挥新消费引领作用，促进产业转型升级。

（二）发展目标

到2020年，基本形成覆盖全生命周期的智慧健康养老产业体系，建立100个以上智慧健康养老应用示范基地，培育100家以上具有示范引领作用的行业领军企业，打造一批智慧健康养老服务品牌。健康管理、居家养老等智慧健康养老服务基本普及，智慧健康养老服务质量效率显著提升。智慧健康养老产业发展环境不断完善，制定50项智慧健康养老产品和服务标准，信息安全保障能力大幅提升。

二、重点任务

（三）推动关键技术产品研发

突破核心关键技术。发展适用于智能健康养老终端的低功耗、微型化智能

传感技术，室内外高精度定位技术,大容量、微型化供能技术，低功耗、高性能微处理器和轻量操作系统。加强健康养老终端设备的适老化设计与开发。突破适用于健康管理终端的健康生理检测、监测技术。支持大容量、多接口、多交互的健康管理平台集成设计。推进健康状态实时分析、健康大数据趋势分析等智能分析技术的发展。

丰富智能健康养老服务产品供给。针对家庭、社区、机构等不同应用环境，发展健康管理类可穿戴设备、便携式健康监测设备、自助式健康检测设备、智能养老监护设备、家庭服务机器人等，满足多样化、个性化健康养老需求。

专栏一：智能健康养老服务产品供给工程

健康管理类可穿戴设备。重点发展健康手环、健康腕表、可穿戴监护设备等，对血压、血糖、血氧、心电等生理参数和健康状态信息进行实时、连续监测，实现在线即时管理和预警。

便携式健康监测设备。重点发展用于家庭、家庭医生、社区医疗机构的集成式、分立式智能健康监测应用工具包，便于个人、医护人员和机构在家庭和移动场景中实时监测各项生理指标，并能借助在线管理系统实现远程健康管理等功能。

自助式健康检测设备。重点发展用于社区机构、公共场所的自助式智能健康检测设备，便于用户在不同社区、机构中随时、随地、自助地完成基础健康状态检测，提升用户自我健康管理的能力水平。

智能养老监护设备。重点发展用于家庭养老及机构养老的智能轮椅、监护床等智能监测、康复、看护设备，开发预防老年痴呆症患者走失的高精度室内外定位终端，实现自主自助的养老功能，提高用户自主养老、自主管理的能力，提升社会和家庭养老资源的使用效率。

家庭服务机器人。重点发展满足个人和家庭家居作业、情感陪护、娱乐休闲、残障辅助、安防监控等需求的智能服务型机器人，提供轻松愉快、舒适便利、健康安全的现代家庭生活，提高老年人生活质量。

发展健康养老数据管理与服务系统。运用互联网、物联网、大数据等信息技术手段，推进智慧健康养老应用系统集成，对接各级医疗机构及养老服务资源，建立老年健康动态监测机制，整合信息资源，为老年人提供智慧健康养老服务。发展健康养老数据管理和智能分析系统，实现健康养老大数据的智能判读、分析和处理，提供便捷、精准、高效的健康养老服务。

三部委关于印发《智慧健康养老产业发展行动计划（2017-2020年）》的通知

工业和信息化部 民政部 国家卫生计生委关于
印发《智慧健康养老产业发展行动计划
（2017-2020年）》的通知

工信部联电子〔2017〕25号

各省、自治区、直辖市工业和信息化主管部门、民政厅（局）、卫生计生委：

为贯彻落实《国务院关于积极推进“互联网+”行动的指导意见》（国发〔2015〕40号）、《国务院办公厅转发卫生计生委等部门关于推进医疗卫生与养老服务相结合指导意见的通知》（国办发〔2015〕84号）、《国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见》（国办发〔2016〕47号）、《国务院办公厅关于全面放开养老服务市场提升养老服务质量的若干意见》（国办发〔2016〕91号）等文件要求，加快智慧健康养老产业发展，工业和信息化部、民政部、国家卫生计生委制定了《智慧健康养老产业发展行动计划（2017-2020年）》。现印发你们，请认真贯彻落实。

工业和信息化部
民政部
国家卫生计生委
2017年2月6日

（四）推广智慧健康养老服务

培育智慧健康养老服务新业态。推动企业和健康养老机构充分运用智慧健康养老产品，创新发展慢性病管理、居家健康养老、个性化健康管理、互联网健康咨询、生活照护、养老机构信息化服务等健康养老服务模式。

专栏二：智慧健康养老服务推广工程

慢性病管理。重点发展病情监测、档案管理、个性化评估、趋势分析、诊疗建议、异常预警、紧急救助、康复服务等。

居家健康养老。重点发展健康体检、居家环境监测、远程看护、亲情关怀、健康干预、健康评估反馈等。

个性化健康管理。重点发展信息采集、健康计划、健康教育、健康跟踪、病情诊断、风险筛查、健康信息查询等。

互联网健康咨询。依托互联网平台，发展在线咨询、预约挂号、诊前指导、诊后跟踪等。

生活照护。基于互联网平台，为老年人提供家政配餐代买等智慧便民服务和关怀照料等养老互助服务。

养老机构信息化服务。重点发展机构内老年人的无线定位求助、跌倒监测、夜间监测、老人行为智能分析、老年痴呆症患者防走失、视频智能联动、门禁系统联动、移动定位、消费娱乐等。

推进智慧健康养老商业模式创新。充分发挥市场主体作用，探索民办公助、企业自建自营、公建民营等多种运营模式，鼓励社会资本投入，推进基本、保障性服务由政府保底购买，高端、个性化需求由市场调配的运作机制，推动用户、终端企业、系统集成平台、健康养老机构、第三方服务商等实现共赢，形成可持续、可复制的成熟商业模式。

（五）加强公共服务平台建设

建设技术服务平台。建设智慧健康养老创新中心，解决行业共性技术供给不足问题，不断创新产业生态体系。集聚产学研医等各方面资源，推动关键技术、核心器件、重点产品研发，完善产品检测认证、知识产权保护等服务，提升智慧健康养老产业的协同创新能力和产业化能力。

建设信息共享服务平台。充分利用现有健康信息、养老信息等信息平台，基于区域人口健康信息平台，建设统一规范、互联互通的健康养老信息共享系统，积极推动各类健康养老机构和服务商之间的信息共享、深度开发和合理利用，开展健康养老大数据的深度挖掘与应用。

建设创新孵化平台。支持智慧健康养老领域众创、众包、众扶、众筹等创业支撑平台建设，鼓励创客空间、创业咖啡、创新工场等新型众创空间发展，推动建立一批智慧健康养老产业生态孵化器、加速器，为初创企业提供资金、技术、市场应用及推广等方面的扶持。

（六）建立智慧健康养老标准体系

制定智慧健康养老设备产品标准，建立统一的设备接口、数据格式、传输协议、检测计量等标准，实现不同设备间的数据信息开放共享。优先制定适用于个人、家庭和社区的血压、血糖、血氧、心律和心电五大类常用生理健康指标智能检测设备产品及数据服务标准。完善智慧健康养老服务流程规范和评价指标体系，推动智慧健康养老服务的规范化和标准化。制定智慧健康养老信息安全标准以及隐私数据管理和使用规范。

（七）加强智慧健康养老服务网络建设和网络安全保障

加强宽带网络基础设施建设，到2020年实现城市家庭宽带接入能力达到100Mbps，打造覆盖家庭、社区和机构的智慧健康养老服务网络。落实智慧健康养老服务平台网络安全防护要求，提高防攻击、防病毒、防窃密能力。加强智慧健康养老个人信息保护，严格规范用户个人信息的收集、存储、使用和销毁等行为。落实数据安全和用户个人信息保护安全标准要求，加强智慧健康养老服务平台的数据管理和安全管控。

三、组织实施

（八）建立部际协同工作机制

工业和信息化部、民政部、国家卫生和计划生育委员会建立部际联席会议制度，加强统筹协调，密切协作配合，形成工作合力，探索体制机制创新，共同研究解决行动计划落实过程中遇到的重大问题，推动行动计划的顺利实施。制定年度落实计划和分工方案，确保行动计划各项任务措施落实到位。

（九）强化组织落实

各地区工业和信息化、民政、卫生计生等主管部门要高度重视智慧健康养老产业发展，建立省级联席会议制度，结合本地实际制定实施方案，明确各部门资源投入，形成合力，联合开展试点示范，科学组织实施。工业和信息化部、民政部、国家卫生和计划生育委员会适时开展联合督导，对各地实施进展和效果进行评估，

总结先进经验并向全国推广。

（十）完善多元化资金投入机制

充分发挥工业转型升级资金、专项资金、地方财政资金等财政资金扶持作用，推动各部门资金集约化整合和精准投放，加大对智慧健康养老的扶持力度。探索与国有资本投资公司合作，充分发挥国有资本的引领和放大作用，通过发起设立智慧健康养老产业投资基金等方式，引导社会资本参与智慧健康养老产业发展，与政府资金形成支持合力。积极推进政府购买智慧健康养老服务，逐步扩大购买服务范围，完善服务内容。探索政府和社会资本合作（PPP）模式，积极引导社会资本参与智慧健康养老服务推广。

（十一）培育和规范消费市场

制定智慧健康养老产品及服务推广目录，推动在养老机构、医疗机构等有关政府采购项目建设中优先支持目录内产品。加强对消费者的使用培训，鼓励有条件的地方通过补贴等形式支持家庭和个人购买使用智慧健康养老产品和服务。

（十二）开展应用试点示范建设

按照企业主体、政府扶持、市场化运作的方式，开展覆盖多级区域、多种类型的应用试点，培育100个智慧健康养老示范企业，建设500个智慧健康养老示范社区，创建100个具有区域特色、产业联动的智慧健康养老示范基地。引导医院、养老机构、社区服务中心和相关企业机构参与支持试点项目建设，支持企业探索可推广、可复制的智慧健康养老服务模式，为智慧健康养老服务提供优质的医疗、养老资源保障。

FIN

/ 挖掘数据价值会引发工业革命？ /

文 / 大数据产业观察　编辑 / 协会会员处 袁硕　日期 / 2017-06

当阿尔法狗战胜了世界诸多一流的围棋手时，人工智能再一次霸屏网络。有人说，谁掌握了数据，谁就拥有未来的世界。而在信息技术迅猛发展的今天，大数据时代会引发下一场工业革命吗？

在贵阳中国国际大数据产业博览会期间，记者采访了英国国际贸易部和多家英国公司的人士，多数人都认为“数据是钻石矿”，如果深度挖掘数据的价值，会有深度的工业革命。大数据技术是英国八大领先科技之一，英国拥有欧洲最大的数据中心市场，占整个欧洲的26%多；2016年英国数据科技的收入超过1700亿英镑，英国数字科技投资超过68亿英镑；阿尔法狗的背后是一家英国公司DeepMind，2014年被谷歌收购。

找出“钻石”

大数据的定义是什么？是关于数据的捕捉、存储、管理和分析，它可能不仅是各种数据集，还可能侧重于从数据中所产生的延伸价值，更多的分析和可视化，更重要的是开发。

“很多年前都还没怎么谈大数据时，我就曾对朋友说谁拥有数据就拥有将来。”全球地球信息科学专家、英国伦敦大学学院时空实验室（Space Time Lab）创始人程涛教授告诉记者，数据本身是钻石矿，怎样把钻石矿里面的钻石找出来，这才是真正的核心。只有把数据转化成为有价值的东西，才会是真正的工业革命。

英国国际贸易部技术与智慧城市贸易主管安德鲁·柯本对记者表示，“我们讲工业革命，其实数据一直都存在着，也有人一直在收集数据，但怎样应用这些数据，决定了是否能开启新一轮的工业革命。有很多公司已经意识到了这一点。”

“大数据是‘新原油’，确实是这样。”Utterberry的首席执行官Heba Bevan告诉记者。这是一家伦敦的传感器创业团队，采用智能技术为各行业提供环境监测解决方案。“现在有AI，有机器学习，还有很多的计算法，来帮助人们理解这些数据是什么。数据分析的结果最后产生作用，才能够智慧地利用大数据。”Heba说。

而Speechmatics首席执行官Benedikt非常肯定地说：“我们现在正在进行第四次工业革命，因为利用大数据才建立了一个快速的机器学习新语言的框架或是系统，而我们的设备学习一门新的语言仅仅只需要两到三周的时间。”

英国从数据中心产业中受益颇多，因为拥有强劲的市场驱动，包括云计算、创意内容、大数据、物联网、智慧城市等，还有金融、电商、医疗健康、智能制造、文化创意等，这都是数据增长的驱动力。而英国在欧洲科技创业领域极具优势，几乎1/3（29.57%）的金融科技（Fintech）投资发生在英国。

开放数据

英国政府还利用大数据为政府作出决策。比如利用大数据检索技术，清查出的逃税与诈骗数额达200亿英镑。而英国媒体对伊拉克伤员的数据报道最终令政府决定撤军。英国大数据的“黑科技”与创新成果刷新着人类的认知，其蓬勃发展得益于政府的战略引导和政策支持。

早在2009年，英国政府打造在线数据公开网站，公布英国财政、交通、医疗、教育等部门的政府信息。此后，英国政府修订《自由保护法案》等，对公共数据开放的形式、格式、许可使用范围等作出规定，为大数据产业提供保障。

2013年，英国与美国、俄罗斯、巴西等国共同发布《开放数据宪章》，通过分享经验和工具支持开放数据的创新者，加强国际合作。

英国政府投入上亿英镑把数据全部开放，以提升政府的治理水平。大数据伦敦计划就将所有的基层和政府管理工作在这个平台上完成高效的数据分析。安德鲁·柯本称：“英国政府是开放数据的政府，最重要的不止是搜集数据或只是去开发数据，而是利用数据提高产业的效率。

英国的数字化战略可以把数字化发展的想法转化成实际应用，来刺激中小企业、大企业及初创企业的合作。英国政府支持产学研之间的合作，数据可以广泛应用到医学和教育方面。我们邀请中国企业、学术界一起在大数据领域加强合作，共同产生价值。”

英国目前有世界上最大的开放数据研究所，还有阿兰·图灵研究所，是英国的国家数据科学研究所。程涛教授表示，英国政府对大学的资助是大数据战略重要的部分，英国所有重要的基金会都有大数据研究策略性的资助，从工程物理学会，到社会经济学、自然环境、减灾保护，这几个委员会都有大量的资金投入。最近政府觉得用数据来做决策很重要，所以正在组织一个DataPolicy会议，这是很高层次的会议，有很多学校和政府组织，今年9月将在英国国会大厦内召开。

FIN

/ 数据分析躲坑秘籍 /

文 / 西南财经大学 吕雪芬 编辑 / 协会会员处 李媛 日期 / 2017-06

有人说数据分析师这条路，是艰辛的、孤独的、漫长的。很多时候兜兜转转竟又回到原点，跑一个算法只能自己和自己死磕，提升段位感觉像在魔兽世界练到110级。

为此，我们特邀科班出身美女硕士——吕雪芬（西南财经大学硕士，主修数据挖掘、多元统计分析、高级数据库应用等，擅长数据建模理论及数据建模工具。在银行、证券、期货领域有丰富的实战经验。目前实操中国证监会资本风险监测项目。），从四个方面，总结自己曾经的误区，并跟大家讲解自己是怎么掉进这个“坑”，又是怎么爬出这个“坑”，最后怎么避免这个“坑”。

主修数据挖掘、多元统计分析、高级数据库应用等，擅长数据建模理论及数据建模工具。在银行、证券、期货领域有丰富的实战经验。目前实操中国证监会资本风险监测项目。

一、轻理论、重工具的学习

现在，各种分析软件层出不穷，到底应该学哪个？如何学好？是理论重要还是工具重要？怎么驾驭？

Excel是所有数据分析师都应该掌握的工具，在你对接业务或跟领导汇报项目的时候，你会选择最直接的用Excel进行数据展现，加以解释，和他们沟通，作用是很强大的，能链接Visio，“竖仓”，还可以进行基础的数据分析回归操作等。SAS前几年用的比较多，内容涵盖的也比较广，包含一些上层、中游、底层的内容，但它是一个收费软件，收费之前一直是政府在用，现在淡化了很多；SPSS，适用于数据分析基础人群，都是模块化，如果想依据自己的想法，再加之和他人的交流经验对其模型或者算法进行完善，可能还达不到那么完美。

网上的招聘简章有的写明要求熟悉很多软件，是不是就要学很多工具？我本人也做过很多面试，在面试应聘者的时候，通常会问一些理论知识，并且针对面试官自己熟悉的点以及应聘者简历的交集着重进行细致的讨论。

至于技术层面的知识，一般面试官会针对应聘者相对熟悉的工具，对一些关键的处理技术进行询问，来评判应聘者对于技术的掌握程度，以此检验简历的真实成分。需要面试者对自己熟悉的工具要有足够的功底，如果能再多掌握一些工具，更是锦上添花。但千万不要贪多，否则等真正用的时候发现无所适从。

总的来说，对于一名分析师，分析是首要的，因此需要对理论知识和业务知识有更加深入和本质的了解，当然对于技术处理也不能疏忽，但绝不是厚此薄彼，舍本逐末。

现在是一个快鱼吃慢鱼的社会，也是一个趋于精细化分工的社会，如果你在前期自己什么都没学过，像概率论、统计学等都不太清楚的情况下，我建议从视频上学习或去上面授课，它能快速的引领你到这条轨道上，然后你再自己慢慢拓宽知识面。我奉行的理念：专业的事情还是交给专业的人处理，能达到效用最大化。如果你本身是学统计学或是数学出身，对于理论知识应该还比较熟悉，具体对哪些方面感兴趣，则可以着重培养。

二、无章法，走了过多弯路

现在做数据服务或者咨询服务的公司在开发工具和制度流程方面，设计的非常规范，非常专业，而对于个体或者不成熟公司，规则制度不健全，总想抄近道，最终越走越远，不得已返工，得不偿失。无规矩不成方圆，做事还是要有规范，有流程。

我走过一个很典型的弯路，以前上学的时候第一次做项目不懂流程，少了很多需求确认和沟通的环节，最终到出结果时，才发现跟甲方的结果千差万别，开始反思讨论，发现取数口径不一致，最后加班加点，修改文档，编码，测试，调度，前期做了很多无用功。相比现在，有很多人还在这个弯路上，奋力前行，只是最终发现南辕北辙，无功而返。这就是缺乏沟通，很多事情都是自己想当然。因此沟通，确认，框架，流程都很重要，缺一不可。

三、过度依赖数据，忽略业务内涵

俗话说：“兼听则明，偏听则暗”。现在比较盛行的量化交易，一般分三个方面分析：宏观的基本变量分析、技术分析、行为经济学。以前做量化的人，直接依据模型，依据数据，然后再做他们自己的东西，之后得出结论，这样很不科学，会造成很大的误区。现在技术人员已经不单单看数据本身了，而是宏观研究员，基本面分析师来定未来的走势、方向。技术分析师运用具体策略给出建仓或者清仓的点位。

所以，工具只有被我们驾驭才能发挥威力，结果只有被我们利用才能彰显他的价值。不能过度依赖数据的结果，要综合考虑业务内涵，定性定量均能提供一定的指导，切勿独断专行。

四、橄榄球式错误

特点是“两头小中间大”，前期后期都想略过，造成中间过程无端增大。小学时，老师让写作业，把一篇课文抄十遍，不管老师为什么让这么做，也不管抄完之后，从中学到了什么。这就是一个为完成任务式学习，这不是一个正确的思维方式。

我们怎么可以避免这样的错误呢？举个例子，如果一个案例分析需要将数据存储到excel中，前期你需要思考：Excel表的数据承载量最大是多少，能不能存储，如果后期领导让你把做好的东西放到他的平台上做展示，要考虑它的可移植性是否很强，全样数据里有没有特殊情况，比如文字、乱码，遇到这些情况，你要怎么处理。这些问题你都需要在excel表中罗列清楚，一项一项去核对。

当然现在有专门的测试部门来做测试，他们按照测试用例逐条检验出你的错误，把问题尽量提前避免，即使前期你把工作做得特别完美，到后期可能也会出现问题。前期不重视，后期必定让你手忙脚乱。

“中间小，两头大”还有一个最直观的体现，本来前期的数据处理工作以及完成模型后的展示汇报工作要占很大比重，但很多情况下，大家容易犯的错误就是拿来的数据直接扔到模型开始计算，根本不去理会数据的量级是否正确，是否有重复值，是否有缺失值、异常值等，这样做会严重影响到模型结果的准确性和正确性，同时如果做完数据模型部分，不去认真研究分析结果，展现工作成果时也会让别人忽视你的付出，同时怀疑你的专业水准。

所以，最后提醒大家一点，一定要重视总结，不断积累。希望我的这些小建议能够帮助大家少走些弯路，多长些见识。

FIN

/ 分分钟搞掂矩阵“特征值”要表示什么“特征” /

文 / statist3927，暨南大学金融系 经济学硕士 编辑 / 协会会员处 袁硕 日期 / 2017-04

从很多年前接触到“特征值”这个词开始，我就一直有个疑问没搞明白，为啥矩阵 “特征值”和“特征向量”中的“特征”，与我们日常理解的、一般口语中的“特征”差异怎么就那么大呢?

比方说张飞的“特征”是高大，黑，大胡子……，但矩阵的“特征值”却是：设A是n阶方阵，如果数λ和n维非零列向量x使关系式Ax=λx 成立，那么这样的数λ称为矩阵A的“特征值”。张飞的特征例子和特征值的定义放在一块，真的非常的风马牛不相及！矩阵的“特征值”想要表现矩

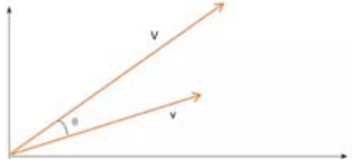
阵的“特征”是什么?

直到最近我开始留意国外教材中，关于“特征值”的英文单词的“eigenvalue”的词根“eigen”后，最后形成了一把解开我心中这个结的关键钥匙。

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sigma_{ij}^{2}$$

Eigen 翻译过来叫做“本征”、“固有的”、“自己的”。在这里，我认为“固有的”、“本征”解释会比“特征”更加贴近原意。在线性代数中，如果将一个向量在空间中

展示的话，会有2个“固有”或“本征”的特性，一个是其方向，一个是其长度。一个方阵A左乘一个非零列向量x，往往是表示“将这个向量x进行线性转换”的意思。而转化的过程，就是通过A。举个例子来说，下图，要将向量v转化成向量V，那么就需要改变v的方向，然后拉长其长度。



那怎么样在矩阵代数中实现呢？过程如下：

设v的坐标向量为$(x_1, x_2)^T$，向量的长度为1，
V的坐标向量为$(x'_1, x'_2)^T$，向量的长度为a。
向量v如果要变换成V，则需要旋转θ角，
并拉长其长度至a，
这个过程可以通过左乘一个转换矩阵实现。

这个转换矩阵就是$\begin{bmatrix} a\cos\theta & -a\sin\theta \\ a\sin\theta & a\cos\theta \end{bmatrix}$，

因为$\begin{bmatrix} a\cos\theta & -a\sin\theta \\ a\sin\theta & a\cos\theta \end{bmatrix}(x_1, x_2)^T = (x'_1, x'_2)^T$

这个时候回过头来再看看“eigenvalue”和“eigenvector”的定义，我们就能够借鉴上面的转换例子来了解到：这个矩阵A，对向量x进行了转换后，并没有改变其“固有”的方向，却只改变了向量的长度λ倍。所以说，这个时候和向量的“固有”特性、“本征”从语言和逻辑上就扯上关系了，而且还不是牵强附会那种。

为了进一步直观的理解上面的话，我们还是举二维向量的例子吧：

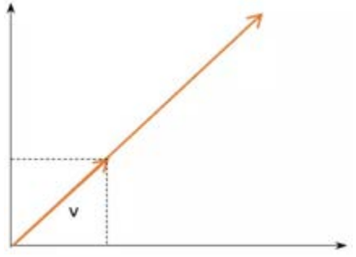
设v的坐标向量为$(1,1)^T$，

v左乘一个转换矩阵$\begin{bmatrix} 1 & 2 \\ 1.5 & 1.5 \end{bmatrix}$后，

v的角度或方向没有变，但长度延长了3倍。

即$\begin{bmatrix} 1 & 2 \\ 1.5 & 1.5 \end{bmatrix}(1,1)^T = (3,3)^T = 3(1,1)^T$

用图表示就是：



综合以上的分析过程看来，“eigenvalue”应该是这样理解：矩阵A在不改变某些向量“固有”方向的基础上，对向量只进行长度λ倍的变换，因此λ就是矩阵A能将这些向量进行保留“本征”变换的倍数值。而“eigenvector”就是矩阵A能够进行这种保留“本征”变换λ倍的那些向量。因此，有些港澳台或者国外翻译的教材，并不把“eigenvalue”翻译成特征值，而是“本征值”，我觉得这么翻译反而更有道理，更贴近原意。

当然了，说到最后，我还是觉得这次解惑的探索过程，应验了某句名言“问题讲清楚了，就解决了一半”。确实，很多求知过程中遇到的问题，理解不了，主要是因为没搞懂“概念”或者没能搞懂确切的含义。当搞清楚或澄清了“是什么”的问题后，问题反而就解决了！

FIN

/ 如何用Python玩转TF-IDF之寻找相似文章并自动生成摘要 /

编辑 / 协会会员处 李缘 日期 / 2017-04

“优雅”、“简单”、“明确”的小特点，让Python成为越来越多数据分析师的新宠。而我更钟爱他的是：简单容易上手，庞大的数据处理能力，和轻松的可移植性……

很荣幸请到中国社科院理学博士祝捷先生，分享有关自然语言处理相关的知识，主要包含两个方面，如何寻找相似的文章，以及如何自动生成摘要，单独与大家进行探讨。

1、了解自然语言处理及应用

从研究内容来看，自然语言处理包括语法分析、语义分析、篇章理解等。从应用角度来看，自然语言处理应用广泛，特别是在信息时代，自然语言处理的应用包罗万象，例如：机器翻译、手写体和印刷体字符识别、语音识别及文语转换、信息检索、信息抽取与过滤、文本分类与聚类、舆情分析和观点挖掘等，我们接触过的怎么通过爬取相关的文本来抽取里面的实体词，最终形成完整的知识图谱等，它涉及与语言处理相关的数据挖掘、机器学习、知识获取、知识工程、人工智能研究和与语言计算相关的语言学研究等。

2、需要知道的背景知识

作为一门语言，它的基本承载信息量的单元，我们认为是一个句子，只要想表达一个意思，通过话语的形式描述出来。但作为一句话，在自然语言处理里，是不太适合直接拿来用的，因为，

它的颗粒度还是太大，我们希望把它拆的再小一点，这样就没有其他选择了。把一句话分成相关的词，那我们先要做的就是分词，来看下面的例子：

1）分词

“一直以来，我都希望有机会学习更多的数据分析知识。” 这句话，拆分成词，用斜杠分隔后就是：“一直以来/我/都/希望/有/机会/学习/更多/的/数据分析/知识。”这就是一句完整的话，按照词语给分隔开来。

但对于我们，每一个词并不是都有意义，上面例句中的“的”字，就是一个虚词。如果把不加选择、无意义的词放到要用的模型中（或其它任何一个模型），就会给整个分析带来干扰，会把整个分析结果拉向一个不准确的方向。

2）去除停用词

把有意义的词保留，无意义的词去掉后：“我/希望/有/机会/学习/数据分析/知识”，就是“去除停用词”步骤。这些词都是有意义的，删掉掉一个，对这句话的信息含量都会造成一定损失。

上面两步是所有自然语言处理之前，必须要做的，这是自然语言预处理的一个过程。

通过举例，引申出怎样去找两篇文章相似的重点是去寻找关键词。得到关键词后，看能否把文章摘要找出来。其实，找出关键词，就可以确定关键句（包含了关键词的句子），而文章摘要就是把关键句组合到一起。

3、如何得到文章的关键词

文章关键词的特点：

1）一篇文章的关键词在文中出现的频率（F1）较高；

$$F1 = \frac{\text{某个词在文章中出现的次数}}{\text{文章的总词数}}$$

2）一篇文章的关键词在其他文章中出现的频率（F2）较低。

$$F2 = \frac{\text{包含该词的文章个数}}{\text{语料库中的文章总数}}$$

假如有个语料库，语料库中计算一篇文章的某个词在语料库中包含的文章个数，就要去除以这个语料库中文章的总数。

为了方便应用，把F1、F2合成一个数学表达式，需要用到TF-IDF：TF-IDF是一种统计方法，用来评估语料库中一个词语对于其所在文档的重要程度。词语的重要性正比于它在文档中出现的次数，反比于它在语料库其他文档中出现的频率。TF-IDF得分高的词语对文章的代表性更强，可以被认为是文章的关键词。通过对TF-IDF得分排序，来发现文章关键词。

$$\text{词频(TF)} = \frac{\text{某个词在文章中出现的次数}}{\text{文章的总词数}} = F1$$

$$\text{逆文档频率}(IDF) = log\left(\frac{\text{语料库中的文章总数}}{\text{包含该词的文章个数}+1}\right) = log\left(\frac{1}{F2}\right)$$

$$\text{TF-IDF} = \text{词频(TF)} \times \text{逆文档频率（IDF）}$$

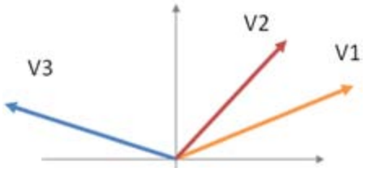
一个有代表性的关键词，F1比较大，这个词出现的频率高，F2要小，为了区分，这个词不能在其它文章里出现的频次过高，TF-IDF值就比较大。

单纯的TF或IDF都不能分离出对文档重要程度高的词语。词语在某一文档内的高出现率，同时在整个语料库中的低出现率，会产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语。

TF-IDF的优点是计算简单，便于理解，性价比高。缺点是忽略了其他一些关键信息，如位置（段首以及句首位置的一般重要程度更高）；其次，不排除这种情况：能代表文章内容的核心关键词只出现1-2次，但其余内容都在对其进行阐述和解释，所以此时单纯靠TF-IDF仍然不能得到理想的结果。

在全部关键词所张成的矢量空间中，文章可以用矢量来表示，文章矢量的距离(余弦距离)近（文章关键词重合度高，文章内容重叠度高），可以判定文章相似。

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{A \cdot B}{|A| \times |B|}$$



寻找相似文章的步骤：

第一步：分词&去除停用词；

第二步：计算TF-IDF得到前n个关键词；

第三步：生成词频向量（原理：如果两篇文章相似，则他们使用的词语也会比较相似。从词频统计的角度，建立文章的词频向量，再来计算他们的相似程度。）

余弦相似度：用两个向量夹角的大小来表示他们之间的相似度，夹角越小越相似。

最后，祝博士给大家现场演示了如何用Python去寻找相似文档，实现生成摘要。

FIN

/ 大数据浪潮下，前端工程师眼中的完整数据链图 /

文 / 推酷网 编辑 / 协会会员处 袁硕 日期 / 2017-05

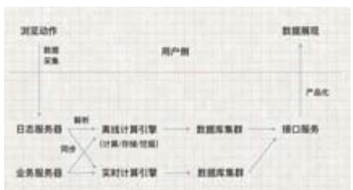
作者：陈屹，前端架构师，现就职于阿里巴巴数据技术与产品部，负责数据产品前端团队，花名流形。《深入React 技术栈》作者，知乎专栏《pure render》创办人。深入在前端架构，可视化等领域多年。坚信 DT 时代的未来，是人工智能与人机交互的双重革命。

马云曾经说过『人类正从IT时代走向DT时代』。正如他说言，今天几乎所有的互联网公司背后都有一支规模庞大的数据团队和一整套数据解决方案作决策，这个时代已经不是只有硅谷巨头才玩数据的时代，是人人都在依赖着数据生存，可以说如今社会数据价值已经被推到前所未有的高度。

我作为一名前端工程师在阿里巴巴数据团队工作多年，深入了解数据生产加工链路与产品化。我们这群前端是与界面最近的工程师们，似乎与数据离得很远，对于我们来说与数据有些怎样连接呢。

完整数据链路

首先，我用直观的一张图绘制出数据采集到产出的流程，中间省略技术细节。



业界常提到的hadoop，Spark，Druid 都在用户侧的下方，也就是数据研发与数据挖掘职能的工作。相对于前端职能而言，一定是与输出终端相关，包括本职工作数据类产品的研发，如阿里指数或百度指数这样的数据展示型产品，还有较为复杂的 BI 工具等，细分起来，最特别的工作应该是数据采集和数据可视化的工作。

但到今天而言，数据研发工程师已经很难说只精通其中一种技术。任何一环深入下去都涉及到整条链路的打通，我就从数据采集，数据可视化，数据产品研发到人工智能几个板块来写写我的体会与经验。

数据采集

过去还是流量为王的年代，流量就意味着钱，互联网都用着简单粗爆的方式导流。在过去做过站长的对数据采集已然不陌生，包括著名的第三方平台CNZZ(现友盟+) 和 google analytics 两个平台几乎都使用过。

基本原理

Web 端的数据采集的链路从客户端或后端开始一直到存储结束。因此，数据采集这个动作涉及到了前端，客户端，数据研发，产品经理等职位的参与。在这个过程中，前端工程师的工作集中在不同客户端上(PC、iOS、Android)的信息收集及埋点上。

再说到采集应用信息这件事，系统可以采集到的信息越丰富，那么数据可分析的内容就越多。数据采集在 Web 端可以分为几个步骤，包括前端 JS 加载，触发事件时收集各种浏览器端的信息，根据用户行为上报日志服务器。

Web 端的数据采集可以收集哪些信息呢?这是一件同时带有『用户体验』和『业务反馈』的工作。

为什么这么说呢。

UX 和前端工程师都非常关注用户体验反馈，而用户体验反应在界面上可以表现为习惯性的浏览轨迹，点击热区等。一般可以采样一部分用户的行为来分析，比如发起一次纯交互改版后，我们需要做一次 AB Test，对于等比例用户分布的点击热力图，同时收集一部分用户的调查反馈，来验证改版后的效果。如果是交互与功能同时改版的话，相对较难判断影响面。

业务反馈就比较自然了。对于网站应用来说，一定会带有用户登录的信息，或关联业务的一次活动，那么自然会把业务信息都带到一起。此外，业务形态不同，会设计业务独有的打点采集方式。如区块明显的应用， 我们需要精确的确定站点，页面，区块和链接，可以达到流量的精准定位，就是典型的为业务形态定制采集方案。对于区块化不明显的后台类应用就不完全适用。精细化采集更需要运营或产品经理对于产品功能有细节上的思考。

此外，业务反馈还可以从前端本身看，前端需要的稳定性指标也是从界面上采集到的，比如加载性能、JS 报错等。大规模应用背后开发一般都有自己的监控平台，而前端的监控就从用户界面开始。

再说说埋点，这是需要大量前端开发的工作。其实近年来为了减少开发的投入，业界有好一些方案产生，其中比较流行的是可视化埋点和无埋点。

可视化埋点对于 PD 或运营来说就十分友好了，仅通过界面就可以配置打点的位置。但有一个极大的问题，我

们的界面经常会改变，导致了元素的变化，埋点就需要经常作更新。

另外就是无埋点，这是一些新兴数据公司如 GrowingIO 推崇的概念。其实无埋点的理念十分简单就是全采集，不论是有效还是无效，所有的点击行为都会被采集下来。这种方案最大的优点就是零工作量，采集的数据非常全，缺点就有大量的无效信息需要到计算或存储过滤。这是计算资源、硬件存储与人力成本之间的考量。

新采集

前端只是客户端采集的源头。比如音频，图像，视频等新媒体都需要一种可以量化成数据的方法，沉淀成指标;在 IoT 领域，更是涉及到硬件设备上采集技术。这些采集可以说与前端都没有联系，但都是采集的一部分。

数据多样的同时，我们需要什么样的数据才是有价值的问题。同样的信息有这样收集，那样收集的，但最好采集的方式是要考虑的。今天非常多的数据是被动的收集的，用户没有发现价值还被强迫采集。数据其实是业务的一部分，业务应该会自生产我们需要的数据，通过智能化手段改善业务，这是主动的方式。

数据可视化

经过清洗，计算与存储后达到数据展现的阶段。无论是面向哪个群体的数据产品都绕不开对数据的可视化，可以说产品端除了考虑分析链路或操作链路外，最重要的工作就是如何更好的反应它们，可视化在其中至关重要。

数据可视化包括科学可视化、信息可视化和可视分析学。数据可视化绝不是单纯的视觉，也不是单纯的图表，它是帮助人类从原始信息中做到对信息有一定程度的认知，任何可视化手段都为了这个过程，而非结果。如果计算机已经可以得出结论，那么可视化是没有意义的。

数据可视化对于我们而言其实是一个跨界的领域，交互视觉知识远远不够，还涉及硬件、客户端编程、数据分析、机器学习等领域。

什么是优秀的可视化图表

数据可视化常指将数据用统计图表方式呈现，从大的分类上看可以分为统计数据可视化、关系数据可视化、地理空间数据可视化，当然还有时间序列数据可视化、文本数据可视化等。

在这里，我就从数据可视化中的图表这个宏观概念来讲讲，如何来构建可视化图表，什么是优秀的可视化图表。

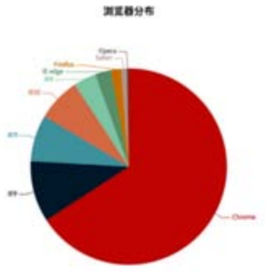
我们看过形形色色的图表，可视化图表是从数据 –> 清洗 –> 交互 –> 视觉 –> 开发的整个过程下创造的。那么优秀的可视化作品具备哪些特征呢?正如上图所说，优秀的可视化作品 = 信息 + 故事 + 目标 + 视觉形式。

定义合适的可视化图形，可以说是最为关键的。在统计领域的运用从 18 世纪已经有经典的案例。我相信每个人都熟知大部分的统计图表，如线柱饼等形式，但统计图表的使用方式往往被经验化，其实它也是一门学问而不是经验。一般情况来看，线柱饼的确可以完成我们大部分的需求，但对于大数据场景或具体业务场景下就需要更多理论在背后作基础。

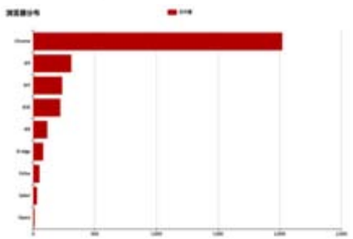
我给出对于定义图表的原则是即直观又丰富，即相关又离散，看上去是两对矛盾的词。对于统计图表与表格最大的不同在于它经过了一次对于数据的加工，反映在我们脑中的信息也许更容易触达所想，同样也可能会形成误导。因此，我们需要展现出多方面的信息。

举个例子，比如我们要研究不同浏览器在应用中的比例，主流的有 IE 系列，chrome，firefox 以及 Safari 等。

在统计图表中，饼图是我们最常用的图形。但大部分用户并不清楚饼图仅适合少分类的情况，而在实际场景下大多是多分类的数据，这时饼图是非常糟糕的图表选择。



这种情况下横向柱形图更容易看到直观的对比。



此外，饼图还有一个致命的问题就是不能反应时间维度的变化。加入时间维度的分析，等于加入了变化的趋势，不再是定量分析，这时候用堆积面积图(Stacked Area Chart)是最合适的。

回到我说的原则了，即直观又丰富表示了我们尽可能的展示我们所能提供的数据，用一种最直观的形式。即相关又离散表示了我们需要把数据之间的关系表现出来，是趋势还是对比，而优秀的可视化图表在设计的时候还会考虑留下一些可分析的内容，让阅读者思考。

我在前年负责浏览器升级的事项，首先对于这件事有一个明确的目标，即 IE8 的比例占总数的 5% 以下就可以放弃，接着在网站上放一些公告提示，以及客服导向，近一个月的效果非常显著，IE 8 从 10% 下降到 4% 左右。在汇报时我就用了堆积面积图来说明效果。

利用可视化图表就可以完成初级的一些运用，比如做信息图，产品中统计数据的展示等。

分析领域

理解可视化图表是进军数据可视化的第一步。数据可视化这个术语被工业界最常运用到的场景是 BI(商业智能)工具上，我们最熟悉的 BI 工具就是Excel，它是以表格为主体的工具。

BI 工具的目标定位是数据分析师，最知名产品的有 tableau、chartio，还有开源的 metabase。它们都拥有完整的数据链路，包括数据源的接入(Data Source)，到选择图表(Chart)，最后可以制作仪表盘(Dashboard)。

我们可以查阅 [Gallery | Tableau Public](Gallery) Tableau Gallery 中一些优秀的仪表盘制作，有一个感官上的认识，看得到很多报表在交互与视觉上均做了深度的定制。

在这个领域中，有几个关键的支持:

多数据源的支持，一定能够快速建立 OLAP 模型，这是基础，大规模系统会提前部署好相应的集群。

可视化图表的丰富度与可交互性上是决定性的。在分析过程中，需要很多辅助手段，过去在表格当中是很常见的，但在图表上一样需要这样的支持，如去噪点(如峰值)，不同图表同维度的联动，值与比例的转换等等。

非常便捷的报表服务。比如现在 BI 工具都会提供仪表盘的功能，对于分析人员提高工作效率非常重要。

到今天，大规模的数据集上的数据分析已是常态，那我们不得不引入更多的分析方法，像 tableau 在去年版本更新中已经支持了聚类的可视化展现。从丰富的可视化手段来看，不可否认，BI 工具不再局限在表格上的操作，新兴产品更多地利用了可视化手段。

当然，BI 工具方面我也是浅尝辄止，它本身涉及复杂的操作与数据分析思路，需要很多专业背景才能了解和掌握。

算法领域

再说到算法领域，在分析领域我们已经看到会引入像聚类的可视化手段。而在更底层的算法领域其实早就在利用可视化做工作了。我们最熟知的是 R 语言中的 ggpolt 库功能十分强大，它用来作模型评估在该领域中是必备工具了 。

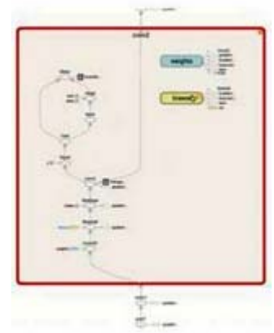
这里就提到了可视化在算法领域的主要工作之一——模型评估。对于一个场景而言，比如定性分析用户的类别，我们可能会同时跑逻辑回归或决策树多个算法，怎么知道我们的算法欠拟合或过拟合呢，当然可以直接看结果。更好的方式就是通过可视化的方式直观的对比。此外，以下还会提到深度学习中的应用。

另外，算法过程可视化近年来慢慢流行起来。这个页面就展示了决策树的可视化过程 A visual introduction to machine learning。此外，著名的深度学习工具 TensorFlow 官方也推出了一个开源的可视化工具 Playground，用于简单神经网络演示与实验。

对算法过程作可视化对于非专业人员去理解算法来说很有必要。一方面可以作为算法在学校或工作中的教学辅助，另一方面可以给非专业人员讲解算法的运算过程。

在算法与可视化结合作得非常惊艳的不得不再说到 TensorFlow。它提供了 tensorboard 这样的内置功能作为可视化的工具。

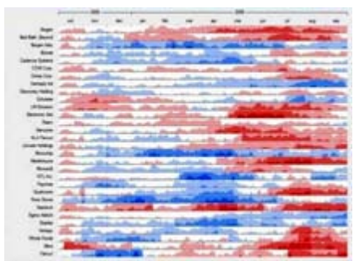
我们很多时候做好神经网络但没有一个图像的概念。用了 tensorboard 就可以看到整体神经网络的框架结构，如第一张是 GRAPH 网络结构，第二张图是整个训练过程中，各个参数的变换情况:



可以说 Tensoflow 的运用大部分都在这样的可视化界面中操作并得到相应的反馈。这一点是不是联想到我们日常工作是否可以在可视化界面上作，事实上，前端也有很多可视化编辑器，概念上等同于把代码与流程进行了封装，用更简单的方式去工作。

数据可视化发展

可视化基础图表的发展一直在推陈出新。像 Datawatch 这家公司在 2008 年设计了一种图表名叫 Horizon Graphs，这个图表还在 2016 评为重要的可视化发展之一。它可以帮助你在狭小的空间下看到极多的类目基于时间序列的变化。这个图表在视觉上很特别，非常有创意。对一名数据从业者，如果知道更多更多展现方式一定会事半功倍。



从工程领域来说，数据可视化会向更多基础的工程领域扩展。在未来，智能化是一个持续的热点，怎么理解

算法，怎么评估模型的有效，怎么做平台化，都需要可视化参与。另外，像 jupyter 或 zeppelin 提供了直观从代码到图形化展现的功能，这一点非常吸引数据全栈开发者去玩数据。

从技术实现方面，今天的 web 开发者们也不再满足于传统的开发方式，会思考用一些专业的 3D 引擎或游戏引擎去做可视化效果，甚至会利用增强或虚拟设备(AR/VR)。在可视化上的探索是永远不会局限的，人的想象力有多大就可以做得多不可思议。

数据产品研发

再说到数据产品研发。产品研发一直是前端工程师的主旋律，我们的工作除了基础架构，稳定性保障，大部分都是在产品研发中。

我们在基础架构上做了不少的探索，从过去的 Backbone 到今天的 React + Redux。在这个篇章，我想说的并不是怎么选框架的问题，而是针对不同类型的产品使用不同的开发与思考方式。

从业务形态映射开发模式

一般来说我们构建应用的顺序一定是从组件 –> 页面 –> 应用。从这三个层面都有不同的协作对象的侧重。数据产品与业务产品最大的区别在于展现的形式上极其区块化，区块内与数据内容的展现强关联。从这一点也映射出数据前端在开发产品上的一些思考，用一句话说就是组件，页面与应用切分开。

在组件层面我们与视觉交互会非常紧密，近而团队一般会有带有产品风格的组件库。我们在这个层面考虑的问题往往是如何与 UED 形成一套高效的沟通机制。在数据产品这个层面，团队与团队之间一定会从理解上的一致到共同沉淀认可一致的规范。

到页面层面，我们会关注两点：

第一，业务模块间的逻辑。这时候会定义一种输入输出固定的通信协议，形如使用 React Component 的方式调用。保证了页面上模块的加载的通信性，此外，用同样协议的模块是无关框架的。

第二，整体数据流。这时候与后端工程师的交流会多起来。大部分数据产品极少会出现在客户端抽象的实体，大部分都是指标类的约定。因此，我们在设计上极难利用像 GraphQL 这样的设计。而是转为约定一些固有的返回形式，重转换的过程，如格式化数据。

到应用层面，我们会关注产品本身的设计。比如导航，涉及到路由的配置。

当然，这三者是有一些交汇，只是当团队做得较好时，会有比较清晰的流程控制。对于任何前端开发的 web 产品而言都不会逃离这个步骤，那么我们如果不仅仅停留在开发阶段，想通过工具化的方式来提升这三者的效率，会怎么做呢。

自然地，我们会想到一种思路构建平台，把固有逻辑抽象出来成为一种标准化描述语言，然后通过引擎进行渲染，非常像可视化编辑器的逻辑。但事实上我们团队并没有走这条路，因为当业务中加入很多个性化的元素之后，引擎的维护难度也是相当的工作量。这里也留下一个伏笔，以后来讲讲在这里我们的具体思考。

总之，在产品研发上，总体思路都很相似，轻数据流的抽象界面，重数据流的抽象数据。但世界上哪有这么方便归类的而用同一种思路的，用一句经典的台词，看起来代数问题，却是几何问题。

数据精细化运营

产品上线对于研发而言，就已经结束了，除了后续的问题反馈修复 bug。但我还是提下数据精细化运营。

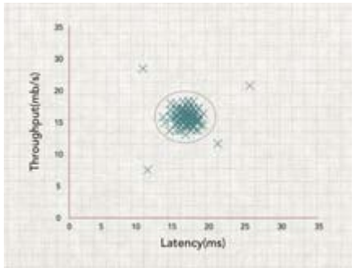
我几乎每天在看产品的数据报表，换个角色研究产品的增长。今天在阿里对于涉及到做营销的角色都会希望有数据支持，比如内容IP，需要看它发表的文章或视频的效果，用户停留的时间多少，他们的粉丝喜欢读哪一类的文章。那我会去看我们产品的功能是否得到满足，哪些功能 MAU 很低。希望从用户的角度去思考怎么帮助他们更好的活着。

工程师具备一定的数据化运营的能力，可以思考产品的功能与研发做连接，问一下作为前端工程师能做什么事。此外，自己在做一些技术推广的时候，也会有一样的思路，我写的文章为什么没有人看，是太简单，太难了?我做的专栏定位是所有前端，还是资深前端?很多问题，你就可以用数据采集的思路去构建更精细的指标。

前端人工智能

最后，讲讲前端在 AI (人工智能) 时代的位置。目前，前端涉及到 AI 的主要是算法数据可视化，这一点在上述也讲到了。

很有意思的是，去年我们在做一款前端监控平台也涉及到了机器学习。我们都知道常规异常报警思路是一旦发生错误就发生通过。传统异常检测是机器学习算法的一个常见应用，利用多维度的值的分布符合某个参数的正态分布来判断。



但前端错误本身，我们无法判断是否会造成影响，有时只是一个报错而已，需要前端工程师自己去排查，这一点与传统异常检测的思路就不一样。我们就利用出现的规模，时长，影响人数等因素利用统计学中的3σ原则，当然，进一步我们利用特征工程的方法实时来检测错误的影响程度。

除了在稳定性方面，只要是生产力工具都可以去思考是否让 AI 改变我们的开发现状。这个地方留给所有的工程师思考。

总结

不论讲到采集还是可视化，还是做数据产品，我都想讲两点：

第一，数据的完整链路。没有『好』的数据，没有看到其中的意义，没有这条链路中清洗计算部分，都是没有意义的，像离线计算，实时流计算都是背后非常关键的技术。这也告述前端工程师专注在一个领域，不等于只看到冰山一角。

第二，不同的思考方式。就说可视化与机器学习这一对。从某种意义上来说思路完全相反，可视化需要人类从感知数据到认知数据，而机器学习是通过大量样本学习得到结论。现在的科技由机器学习的技术还无法做到的事，都还会通过类似于可视化的方式传递给人类。如果某一天机器也可以做到能理解世界，那么真正的人工智能就来到了。

因此，人工智能今天还是技术，也是思路，我们可以用在任何环节，不论是哪个岗位的工程师都应该掌握。在过去，前端的工作只与界面相关，而今天前端在一定程度上已经具备了全栈开发的能力，前端工具化平台化已经很常见，可以利用机器学习完善工具。

还有另一种说法，人工智能的起点是人机交互的革命。要让机器变得更智能，我们会用更多增强体验的方式去改变今天的人机交互。因此，前端技术是有很大延展空间的。今天立足在 Web 领域我们是有优势的，那么在其它领域呢，我们今天的技能是否做到了编程语言与平台不受限。由此也看到前端工程师在大数据时代涉及的一些工作非常需要有综合能力。前端工程师的基础能力从过去纵深到现在更趋向于 T 字型发展。我相信这是未来工程师们的基本形态。

FIN

/ 零基础初识：搭建Hadoop大数据处理 /

文 / 36大数据 编辑 / 协会会员处 李缘 日期 / 2017-05

在互联网的世界中数据都是以TB、PB的数量级来增加的，特别是像BAT光每天的日志文件一个盘都不够，更何况是还要基于这些数据进行分析挖掘，更甚者还要实时进行数据分析，学习，如双十一淘宝的交易量的实时展示。

大数据什么叫大？4个特征

体量化 Volume，就是量大。多样化 Variety，可能是结构型的数据，也可能是非结构行的文本，图片，视频，语音，日志，邮件等快速化 Velocity，产生快，处理也需要快。价值密度低 Value，数据量大，但单个数据没什么意义，需要宏观的统计体现其隐藏的价值。

可以看出想只要一台强大的服务器来实时处理这种体量的数据那是不可能的，而且成本昂贵，代价相当大，普通的关系型数据库也随着数据量的增大其处理时间也随之增加，那客户是不可能忍受的，所以我们需要Hadoop来解决此问题。

优点：

Hadoop是一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在Hadoop上开发和运行处理海量数据的应用程序。它主要有以下几个优点：

高可靠性。Hadoop按位存储和处理数据的能力值得人们信赖。

高扩展性。Hadoop是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

高效性。Hadoop能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。

高容错性。Hadoop能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

低成本。与一体机、商用数据仓库以及QlikView、Yonghong Z-Suite等数据集市相比，hadoop是开源的，项目的软件成本因此会大大降低。

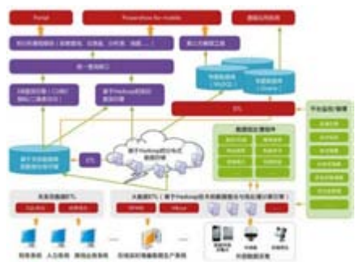
Hadoop得以在大数据处理应用中广泛应用得益于其自身在数据提取、变形和加载(ETL)方面上的天然优势。Hadoop的分布式架构，将大数据处理引擎尽可能的靠近存储，对例如像ETL这样的批处理操作相对合适，因为类似这样操作的批处理结果可以直接走向存储。Hadoop的MapReduce功能实现了将单个任务打碎，并将碎片任务(Map)发送到多个节点上，之后再以单个数据集的形式加载(Reduce)到数据仓库里。

Hadoop在各应用中是最底层，最基础的组件，所以其重要性不言而喻。

框架结构

Hadoop主要由HDFS (分布式文件系统)和 MapReduce (并行计算框架)组成。

Hadoop 由许多元素构成。其最底部是 Hadoop Distributed File System(HDFS)，它存储 Hadoop 集群中所有存储节点上的文件。 HDFS(对于本文)的上一层是MapReduce引擎，该引擎由JobTrackers和TaskTrackers组成。



通过对Hadoop分布式计算平台最核心的分布式文件系统HDFS、

MapReduce处理过程，以及数据仓库工具Hive和分布式数据库Hbase的介绍，基本涵盖了Hadoop分布式平台的所有技术核心。

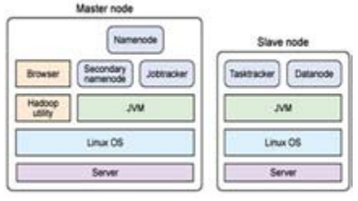
HDFS

对外部客户机而言，HDFS就像一个传统的分级文件系统。可以创建、删除、移动或重命名文件，等等。但是HDFS的架构是基于一组特定的节点构建的，这是由它自身的特点决定的。这些节点包括 NameNode(仅一个)，它在 HDFS 内部提供元数据服务；DataNode，它为 HDFS 提供存储块。由于仅存在一个 NameNode，因此这是HDFS 的一个缺点(单点失败)。

存储在HDFS中的文件被分成块，然后将这些块复制到多个计算机中(DataNode)。这与传统的 RAID 架构大不相同。块的大小(通常为 64MB)和复制的块数量在创建文件时由客户机决定。NameNode可以控制所有文件操作。HDFS 内部的所有通信都基于标准的TCP/IP 协议。

单节点物理结构

主从结构



主节点，只有一个: namenode；

从节点，有很多个: datanodes；

namenode负责：接收用户操作请求 、维护文件系统的目录结构、管理文件与block之间关系，block与datanode之间关系。

namenode是一个通常在 HDFS 实例中的单独机器上运行的软件。它负责管理文件系统名称空间和控制外部客户机的访问。

datanode负责：存储文件文件被分成block存储在磁盘上、为保证数据安全，文件会有多个副本。

MapReduce

MapReduce是处理大量半结构化数据集合的编程模型。编程模型是一种处理并结构化特定问题的方式。

例如，在一个关系数据库中，使用一种集合语言执行查询，如SQL。告诉语言想要的结果，并将它提交给系统来计算出如何产生计算。 还可以用更传统的语言(C++，Java)，一步步地来解决问题。这是两种不同的编程模型，MapReduce就是另外一种。

MapReduce和Hadoop是相互独立的，实际上又能相互配合工作得很好。

主从结构

主节点，只有一个: JobTracker

从节点，有很多个: TaskTrackers

JobTracker负责：接收客户提交的计算任务、把计算任务分给TaskTrackers执行、监控TaskTracker的执行情况

TaskTrackers负责：执行JobTracker分配的计算任务

Hadoop能做什么?

- 大数据量存储：分布式存储
- 日志处理: Hadoop擅长这个
- 海量计算: 并行计算
- ETL:数据抽取到oracle、mysql、DB2、mongdb及主流数据库
- 使用HBase做数据分析: 用扩展性应对大量的写操作—Facebook构建了基于HBase的实时数据分析系统
- 机器学习：比如Apache Mahout项目
- 搜索引擎:hadoop + lucene实现
- 数据挖掘：目前比较流行的广告推荐
- 大量地从文件中顺序读。HDFS对顺序读进行了优化，代价是对于随机的访问负载较高。
- 数据支持一次写入，多次读取。对于已经形成的数据的更新不支持。
- 数据不进行本地缓存(文件很大，且顺序读没有局部性)
- 任何一台服务器都有可能失效，需要通过大量的数据复制使得性能不会受到大的影响。
- 用户细分特征建模
- 个性化广告推荐
- 智能仪器推荐

扩展

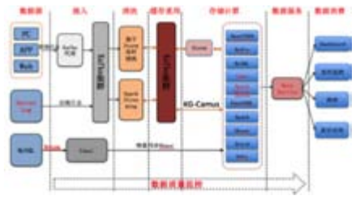
实际应用：

Hadoop+HBase建立NoSQL分布式数据库应用

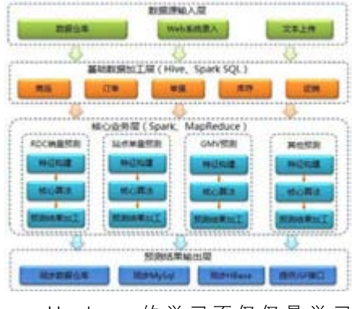
Flume+Hadoop+Hive建立离线日志分析系统

Flume+Logstash+Kafka+Spark Streaming进行实时日志处理分析

酷狗音乐的大数据平台



京东的智能供应链预测系统



Hadoop的学习不仅仅是学习Hadoop，还要学习Linux，网络知识，Java、还有数据结构和算法等等，所以万里长征才开始第一步，希望Hadoop学习不是从了解到放弃。

FIN

/ 大数据解密:《人民的名义》是怎么火起来的？ /

文 / 中国大数据 编辑 / 协会会员处 李缘 日期 / 2017-04

最近这几天，坐地铁上下班的时候，总能看到有人捧着手机，痴迷于《人民的名义》的剧情中。看到此番景象，出于职业本能，不禁想要透过现象看本质，迫切想要知道《人民的名义》怎么这么火?

随即着手于大数据监测，并对结果进行分析，终于发现《人民的名义》是怎么火的，促成它火起来的几个关键点是什么，它走火的路径又是什么。

1）借助大数据工具：新浪微舆情的全网事件分析、微博事件分析、微博传播分析。

2）数据采集起止时间：为了更真实的还原真相，全网事件分析和微博事件分析，将以首映当天(2017年3月28日)，向之前倒退两周时间，也就是2017年3月14日00：00：00为数据采集起始时间，以首映当天(2017年3月28日)，向之后顺推一周时间，也就是2017年4月5日23：59：59为数据采集的截止时间。从这段时间内的数据变化，来发现《人民的名义》走红的路径。

3）分析说明：首先，让我们仔细看看该时间段内，《人民的名义》话题热度的概览图。其次，我们会对《人民的

名义》在该时间段内话题热度的峰值，进行一个一个的分析还原。

1、强势媒体+知名演员联合预热宣传片，引发数万次转发，覆盖人数3000万。

从上图中可以看到，《人民的名义》话题热度在2017年3月17日达到了第一个高点。

原来是在这一天，芒果台“湖南卫视芒果捞”在新浪微博，发布了《人民的名义》概念+阵容宣传片，而与此同时，该戏的主演陆毅转发了该微博，强势媒体的自媒体发声+老牌演员助力+发布阵容宣传片，自然而然推动《人民的名义》蹿红网络，两个微博累计转发达20000余次。

2、强势媒体芒果台播出，引得观众在社交媒体竞相评论，共计十余万条。

话题热度在3月28日达到了第二个高点。这一天是《人民的名义》首映日，在芒果台的黄金时间段播出，再加上一反人们心目中关于芒果台的刻板印象，结果，微博上出现了十几万条关于此剧的讨论。

3、超重量级媒体人民日报微博推荐，外加众大V推荐，再引热议。

眼看，3月30日，《人民的名义》相关信息回落至低点。就在这个时候，重要角色又上场了。

人民日报官微、标榜全职看电视剧的一点茄子酱等发布原创信息，就此，再次激起网络舆论场关于《人民的名义》的信息浪花，转发，原创再次席卷开来。

2017年3月30日，人民日报官微单条微博转发接近20000条。该条微博，也就此成为人民日报官微的超级热门微博。正常情况下，人民日报的热门微博，以千为单位。

4、清明假期，视频网站首发平台PPTV《人民的名义》被转1242次。

事件发展到这个时候，《人民的名义》引得了越来越多媒体在社交媒体上发文，也引得了越来越多影视剧迷的关注观看。至此，《人民的名义》顺利走红，成为清明假期期间，非常多人的观剧必选项。在清明假期首日，《人民的名义》，在视频网站的首发平台——PPTV，被转发1242次，成为热门信息。

从以上热度走势可以看出，《人民的名义》在网络上的热度，可谓一潮未落，一潮又起，又是什么原因促成了这种现象?总的来看，有一下四点：

1、定位于市场空白，在黄金档播出

现如今影视剧题材大同小异，从定位的角度来看，该剧成功抓住了反贪剧登陆黄金档的市场机会，而且是芒果台的黄金档。要知道，2004年以来，广电总局下达指示：因为数量泛滥和编剧粗糙，反腐剧被整顿，从此不能登陆卫视黄金档。

2、最高人民检察院专业支撑，老手编剧周梅森执笔。

该剧是最高人民检察院出品，在专业性上显得专业，比如犯罪动机、犯罪手法、缴获赃款现场等等，与此同时，该剧由反腐剧老手周梅森担任编剧，从而实现了专业和普罗大众两个要素的统一。

3、赢得年轻人关注、观看、转发。

该剧在强势媒体芒果台黄金档播出，借助芒果台所拥有的庞大年轻观众基础，因而快速获得了大量年轻人的关注。之后通过有优质内容、和别样题材，使得大量的年轻粉丝，不仅看剧追剧，还在网络上自发传播《人民的名义》相关话题，引发了该剧网络传播的热潮。

4、众多官方媒体、非官方媒体的开始大力推荐。

该剧在芒果台播出，受到观众热烈讨论，这其中不乏有影响力的媒体人。所以，在这个时间点之后，众多官方媒体、非官方媒体的开始大力推荐，而大量媒体的介入，自然推动该剧的网络热度再次走高，成功走出原本芒果台的粉丝这个圈子，被越来越多的人注意到，正式走红。

FIN

/ 男子利用人脸识别技术，被拐27年后与亲生父亲相认 /

文 / 搜狐新闻　编辑 / 协会会员处 李缘　日期 / 2017-04

通过“宝贝回家”志愿者的不懈努力，33岁的福建男子胡奎终于弄清楚了自己的身世：他其实是重庆人，原本的名字叫“付贵”。4月8日，付贵通过视频与重庆的家人“见面”，被拐27年后，依赖于技术的进步，付贵终于见到了家人。这是跨年龄人脸识别系统成功用于寻找走失儿童的首例。

病床上的付贵与家人在进行视频通话

付贵生活在福建，做过厨师、当过小工头、现在的工作是和花花草草打交道。虽然每天的生活离不开喝茶，但是他特别嗜辣。“我们家里人口味都很轻，只有我口味很重，喜欢吃麻的、辣的，特别是万州烤鱼上面的那层花椒，吃到嘴里又麻又辣，我最喜欢吃。”因为饮食习惯上的巨大差异，付贵一直觉得自己和身边人格格不入。而童年时被拐的记忆，让他寻亲的念头越来越强烈。

“我记得是在上学的时候，或者是在放学的时候，被人拐走的。我有做梦坐过长长的火车，好像是经过了沙漠一样的地方，很大的一片，印象很深刻。然后遇见一间屋子，后面就被拐到这里来了。”付贵回忆，来到福建之后他生了一场大病，所以很多事情都记不清了。

2009年，付贵在宝贝回家网站上登记了自己的信息。在他的登记信息中，失踪地点填的是福建。付贵不敢设想寻亲的结果：“如果没有结果，那我就当自己是一个‘过客’吧。”

付贵等了8年，终于等来了找到亲生父亲的消息。对于付贵的亲生父亲付光发、姑姑付光友而言，他们的等待更为漫长。他们等了27年。

时间回到1990年10月16日。重庆市石柱县大歇镇。这天早上，付光友送6岁的侄子付贵上学，路上还给他买了爆米花。“付贵你去好好读书，下午早点回

来。”这是付光友给付贵讲的最后一句话，她现在都记得自己当天给付贵穿的衣服，涤卡衣服和裤子，背了黄颜色的帆布书包。

幼儿园距离付光友家不到一公里，在大歇镇这样的地方，孩子们都是放学后自己走回家。然而这天下午放学后，付贵并没有回到家中。付光友认为，可能是孩子的外婆把孩子接走了，于是没有在意。直到第二天，他们才发现孩子丢了。付贵的父亲付光发立即报了警，拉上邻里亲戚把周边能找的地方找了一个遍。

那时候，石柱县还没有通火车，于是一家人就坐着船出去找。听到哪儿有找到走失小孩的消息，他们都跑过去拿着照片核实，看是不是付贵。随着时间的推移，希望越来越渺茫，付光发也渐渐不再向别人提起寻找儿子的事情了。

付贵丢失后的第三年，付光发跟着熟人来到辽宁鞍山打工。付光发从没跟工友提过儿子丢失的事情，但晚上下了工休息的时候，他常常会想到儿子，有时候会梦见儿子小时候拽着他，有时想起付贵感冒了上医院打针，其他小孩都哇哇哭，只有付贵不哭的场景。

“孩子也特别聪明，给他起名的时候，我就想富贵人人爱，就给他起了‘付贵’。”付贵走失后，付光发的小儿子出生，从孩子记事起他就给小儿子说付贵丢失的事，让他能时时记得这个事情。

今年1月，付光发兄妹在宝贝回家网站登记了付贵的信息。上面的照片是付贵4岁时继母带着他去拍的，付贵戴着一顶公安帽，看上去乖巧可爱。这张照片也是付贵留下的唯一一张照片。

在双方DNA尚未入库的情况下，登记信息的不一致为宝贝回家的工作人员寻找带来了很大的阻力。但正是这张照片，成为了寻亲成功的关键。

今年3月，将人工智能的跨年龄人脸识别技术应用于寻找走失儿童中，首批超过2万条寻亲图片数据接入跨年龄人脸识别系统进行对比评测，通过照片比对，初步筛选出数十组疑似案例，付贵就在其中。

“我们第一眼看到提供过来的付贵资料，就觉得这个应该是了，除了照片像之外，还有一个是名字，‘付贵’与‘胡奎’的发音很近。”宝贝回家的工作人员在进一步核实了信息后，开始联系双方进行DNA的入库比对。

4月1日，DNA比对成功。胡奎就是付贵！“我寄出血样去做比对还不到一周就找到了，我很惊讶。” 最终确认消息传来的那天，付贵失眠了。身在鞍山的付光发高兴坏了，他第二天就坐火车赶回了重庆，说要等付贵回家。他说，他暂时不准备回工地了，要在家陪儿子。付贵的姑姑付光友也从东莞赶回了重庆，她今年已经50岁了，大儿子是北大高材生，一手养大的另一个侄子也在读研。她最遗憾的，就是没有机会带大付贵。

4月8日，因为付贵意外住院，让这家人的团聚提前了。这一天，病床上的付贵，通过视频与重庆的家人“见面”了。下午3点43分，视频提前接通，姑姑付光友举着手机，手抖得厉害：“付贵啊，你还认识我吗？知道我是谁吗？我是你姑姑啊，我没有一刻不想你啊！”说到这里，父亲付光发也止不住啜泣，一直紧紧握着的手松开了。被拐27年后，依赖于技术的进步，付贵终于见到了家人。

“宝贝回家”网站创始人张宝艳介绍，付贵是“宝贝回家”采用跨年龄人脸识别系统为被拐儿童成功找到亲人的首例。张宝艳说，照片比对是“宝贝回家”的主要工作途径之一，过去都是工作人员用肉眼进行比对。由于寻亲线索越来越多，肉眼比对的方式已经不能满足需求。她对于人工智能技术助力打拐的项目前景表示乐观，并称许多走失家庭将会在这一新技术中受益。“对于众多被拐儿童家庭来说，会让他们在寻亲路上少走了很多的弯路，会让更多的被拐儿童早日与家人团圆。”

FIN

/ 陕西智诚数据分析师事务所 /

文 / 陕西智诚数据分析师事务所　编辑 / 协会会员处 李缘　日期 / 2017-05

陕西智诚数据分析师事务所是中国商业联合会数据分析专业委员会（简称：中数委）事务所会员单位（中数委团证：第039号）。

事务所成立于2009年，是陕西省内从事投资数据分析、市场调研、可行性分析、经营策划、融资咨询的一家管理先进、功能齐全的专业机构，是中国商业联合会数据分析专业委员会推荐事务所之一，以专业的数据分析服务为核心，致力于为中小型企业、国内外银行、投资公司等提供投资数据分析、经济效益评价、数据处理、投资策划、社会经济咨询、投资中介等专业的、系统的服务。为投资决策提供具有经济性、权威性、客观性、公正性、实用性的数据分析报告。

事务所以专业的数据分析服务为核心，以强有力的专业团队为后盾，广泛开展各项咨询服务业务，为促成具有项目、技术、场地、设备、资金优势的各方友好合作出具投资数据分析报告。

依靠数据分析师的良好专业知识背景和聘请高校专家顾问的丰富经验为后盾，以市场为导向，注重数据的精准获取，并视数据的真实获取为公司发展的生命线。在工作中，已承办各类业务，涉及农业、工业、建材、化工、电子、电力、房地产、教育、煤炭、娱乐、环保、旅游等诸多行业领域，在为客户提供报告的过程中，积累了许多的宝贵经验，已形成了一支执业经验丰富，人员结构合理，高素质的专业队伍。

我们坚持社会效益与经济效益并重，以“服务、合作、勤勉、卓越”为宗旨，秉承“以质取信，恪守德操”的职业操守，审时度势，勤勉尽责。我们期待与您真挚的合作！

办公地址：
西安市西稍门十字开元商住楼
一单元1005室
联 系 人：王经理
联系电话：13165719555
029-88623260

FIN

数据决胜未来

Big Data Control The Future

服务号：CPDA大数据圈

CPDA® 数据分析师
CERTIFIED PROJECTS DATA ANALYST