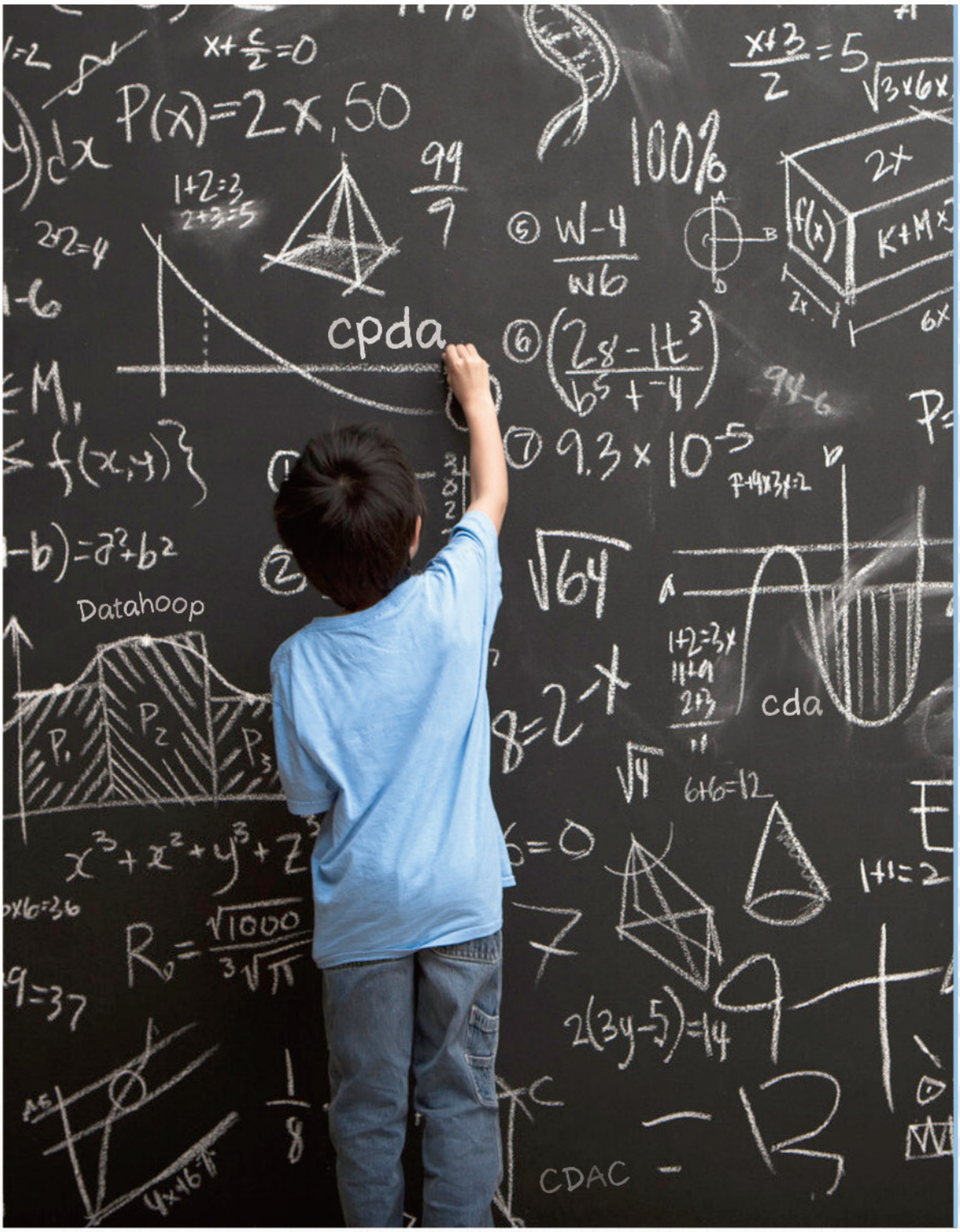
数据分析

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
2016年第02期 总第26期（季刊）
咨询热线：010-59000991 / 59000339
http://www.chinacpda.org/
投稿请至 xiehui@chinacpda.org



China Data Analysis

/ 追溯数据源头,挖掘大数据价值！ /

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 2016年第二期

本期目录 CONTENTS

卷首语
03 追溯数据源头，挖掘大数据价值!

协会动态
05 北京 / 山东 / 云南 数据分析师SALON
06 中国数据分析新远程平台上线

行业资讯
07 专注机器学习:Google在苏黎世成立欧洲研发中心
美发布《联邦大数据研究与开发战略计划》
08 海尔大数据探索成果登上《哈佛商业评论》
阿里云发布《数据安全白皮书》

会议热点
09 引领数据浪潮,创新商业价值 — 大数据应用高峰论坛
10 【数博会演讲】引领数据浪潮,创新商业价值!
13 【数博会演讲】数据分析 — 企业的必修课
15 第四届中国数据分析行业峰会预告

政策导向
17 关于促进和规范健康医疗大数据应用发展的指导意见

会客厅
18 数据的"增值"在于强大的分析能力

数业专攻
20 基于Spark的文本情感分析
24 深入浅出之推荐系统原理应用介绍

运数有道
27 大数据为你预测2016美国大选
29 航空大数据:为你的旅途带来无限便利
31 零售行业的数据挖掘七步走

事务所风采
33 湖南翰林数据分析师事务所
34 云南鼎臻数据分析师事务所
封4 珠海横琴安得信数据分析师事务所

主办
中国商业联合会数据分析专业委员会

编委
佳伊 / 冯伟 / 欧阳琦 / 杜艳丽

出版时间
2016年第二期 07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: +86-10-59000991 / 59000339
传真: +86-10-59000991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!

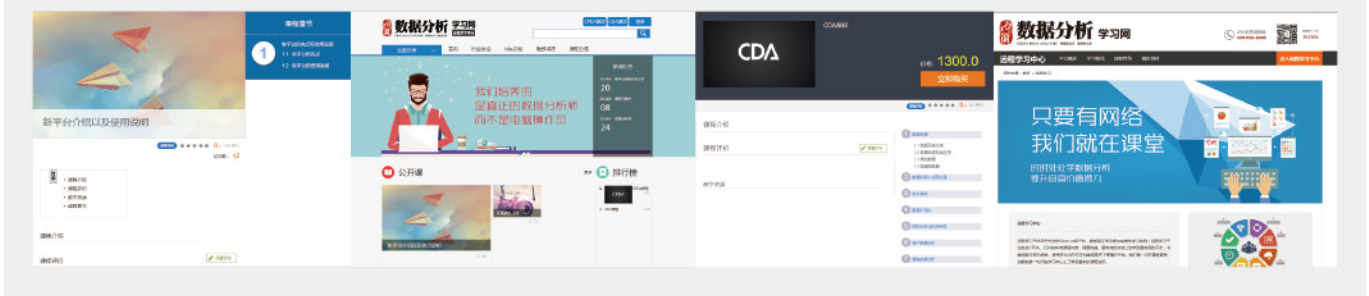
投稿邮箱：xiehui@chinacpda.org

更正说明：本刊2016年第1期"行业热点"栏目中《2016数据分析行业发展战略会议精彩回顾》一文中，"河南翰林数据分析师事务所" 更正为"湖南翰林数据分析师事务所"。
特此声明!

中国数据分析
远程学习平台
重磅升级

2016.07

结构清晰　方便学习　沟通快捷　课件优化　页面整洁



/ 追溯数据源头，挖掘大数据价值! /

“大数据”一词无疑是时下最热门的行业词汇之一。自2015年10月，国家十三五规划正式提出实施国家大数据战略以来，大数据就被各行各业视作战略性资源加以高度重视。

2016年1月7日，国家发改委发布《组织实施促进大数据发展重大工程的通知》，强调大数据应用在各行业落地。然而，大数据应用落地的实现，是以大数据应用价值的实现为目标的，那么如何挖掘大数据价值无疑将是整个行业探讨的重中之重。在我看来，要想挖掘大数据价值，就必须了解数据从哪里来，什么才是真正的大数据?

历经数据分析行业多年的探索与实践，我认为数据的来源有三个：

第一，网络数据，通过网络，我们可以获得很多大体量数据，这个数据量是十分可观的。像BAT三大商业巨头就拥有电商、搜索、社交网络等领域相当大的数据存储量，同时亦具备非常强的数据处理和分析能力。在该源头领域中，尽管已经有大数据应用价值挖掘方面的尝试，但这种尝试仍然停留在传统的互联网模式的基础之上；

第二，政府数据，李克强总理去年9月份提出，要逐步有条件地开放中国政府大数据平台。与网络数据相比，政府数据开放性较弱，保密性高，数据源相对分散，导致政府数据的整合难度高，价值开发难度大。

第三，企业微观运营数据，在我看来，这个源头的数据价值远远超过网络和政府数据。众所周知，近几年阿里也在探索企业微观的运营数据，它虽然拥有庞大的电商交易数据量，价值非常可观，但这巨大的数据价值却并不存在于数据或企业自身，而在于细分市场数据的融合，融合后的数据价值不可估量！可以说，谁能够真正利用所采集的庞大数据衍生数据化产品，将数据增值，谁才可能将大数据的商业价值予以最大程度地开发。但是，要想挖掘企业微观运营层面的价值，我们必须参加企业的构建与运维，全面掌握企业微观运营数据，进而推动企业的数据化建设。这样，我们才能够真正在企业运营端上拥有对微观数据的开发、操盘，甚至深度挖掘的权利，才能够真正实现企业微观运营数据所创造的经济效益。

在DT时代，企业数据才是真正能够创造商业价值的数据来源，而数据价值的创造并不在于数据本身，而在于数据的融合，数据融合才是真正的大数据。只有实现数据融合，并加以分析研究，才能够最大限度地创造大数据的商业价值，即在帮助企业节约运营成本，提高盈利能力的同时，助力企业预见未来趋势，把握机会，使企业在千变万化的市场环境中，制定远见性战略，提前规划布局，有条不紊！

中国商业联合会数据分析专业委员会

- 智能化，一键匹配
- 常用成熟的算法集成
- 操作简单，界面简洁
- 结果输出丰富，可视化度高
- 权威专家实时指导

数据分析 de 魔术师

Datahoop数据分析智能平台
助力您的事业腾飞

CREATIVE

登录 www.datahoop.cn了解更多信息

Datahoop
想你所想 做你所需

以最方便最快捷的方式解决问题
让您集中精力去关注更重要的业务

时间 Time
利用您的闲暇时间，实现化整为零，高效管理。

效率 Efficiency
最高效，最快捷的通过平台解决您所面对的问题，让您解放精力！

精准 Accuracy
海量的数据分析模型，与高效的算法，带给您无比精准的结论。

Datahoop
为数据而生，化整为零
随时随地，随你所用

Datahoop为无处不在的数据而生

- 一键智能匹配
- 操作简单 界面简约
- 成速算法集成
- 输出丰富 可视化度高

/ 北京2016年第三期数据分析公益沙龙圆满落幕 /

文 / 协会市场处 欧阳琦　编辑 / 协会会员处 冯伟　日期 / 2016-06

2016年6月25日，中国商业联合会数据分析专业委员会与猎聘网合作的第3期公益沙龙在朝阳区猎聘网同道小馆成功举办。

猎聘大数据研究院副总监 石晶

炎炎烈日的下午，同学们纷纷到现场签到就坐，主持人简单介绍后，我们的主讲老师开始分享关于大数据的主题报告。此次沙龙分别进行了两个主题的分享，首先由猎聘大数据研究院（DIG）副总监石晶先生带来主题为《大数据分析师职业发展洞察》的报告，他为大家详细解析了数据分析师职业的未来发展前景和薪酬变化。

第二位分享嘉宾是国内某知名互联网公司调研经理谭振华先生，分享主题为数据分析的从业前景和应用实践。他以自己多年的行业经验为大家分析行业前景和实际应用，并教大家如何运用数据做出高大上的数据分析报告。

知名互联网公司调研经理 谭振华

同学们听得十分认真，时不时拿出手机拍摄老师的演讲PPT干货。茶歇时光，同学们边吃着水果，边和到场的同学互相交流讨论。

本次沙龙现场的微信墙互动热烈，大家纷纷把自己对沙龙的期待、感受发送到微信墙大屏幕，在老师演讲过程中把疑问通过微信发送到现场微信墙大屏幕，老师讲解完毕后回答同学们的疑问，同学们表示这种互动形式不仅简单有趣，还能跟老师提问，收获颇丰。

最后是抽奖环节，主持人对现场积极发微信墙提问、问题被老师选中回答的同学们给予了奖品奖励。至此，本次沙龙活动完美落幕，大家反响热烈，纷纷表示期待下期活动的开展。

CPDA数据分析公益沙龙活动是中国最早、最专业、最具影响力的交流平台。在这里你可以免费学习到数据分析软件的技巧应用、实际案例的详细剖析、数据分析应用于公司决策的方式方法，我们真诚的期待更多的数据分析爱好者们加入到沙龙活动中进行分享交流，让我们共同推动数据分析行业在中国的蓬勃发展！

/ 山东数据分析沙龙(第二期)完美落幕 /

文 / 协会市场处 杜艳丽　编辑 / 协会会员处 冯伟　日期 / 2016-06

雨后的青岛格外闷热～但是也比不过数据分析行业朋友们的学习热情，由山东CPDA授权中心组织的－2016年6月26日山东数据分析行业第二届公益沙龙完美落幕！

此次沙龙活动为大家呈现了三个精彩的主题分享。首先，由青岛起啥管理咨询有限公司的特聘讲师马连君老师为大家带来《数据分析行业最新资讯》以及实力干货《用户画像》的主题分享，第二位分享嘉宾是来自K&M Accessories的中国IT主管孙云龙先生，分享主题为《企业信息化管理》。随后大家分别提出了自己的观点和看法，与嘉宾一起进行了探讨和互动。在愉快热烈的氛围中，大家合影留念。

fin

/ 云南省第二届数据分析沙龙活动成功举办 /

文 / 协会市场处 欧阳琦　编辑 / 协会会员处 冯伟　日期 / 2016-06

2016年6月26日《彰显数据价值，智谋企业未来—云南省第二届数据分析沙龙活动》在昆明白云宾馆成功举办。本届沙龙活动由中数委主办，昆明万象教育信息有限公司承办，云南鼎臻数据分析师事务所协办。

活动特邀演讲嘉宾有：中数委邹东生会长，云南鼎臻事务所苟国爱总经理，云南省玉溪师范学院邹国忠老师。

活动现场吸引了当地数据分析师学员，从业机构负责人以及企业高管的热情参与。本届沙龙活动贯穿行业监管，人才培养，从业发展等环节，旨在普及大数据知识，分享行业信息，为广大数据分析师及爱好者提供一个学习交流和资源整合的专业平台，从而促进云南地区数据分析业务发展和培训报告工作的顺利开展。

fin

/ 中国数据分析新远程学习平台上线 /

文 / 协会市场处 欧阳琦　编辑 / 协会会员处 冯伟　日期 / 2016-06

经过半年多的完善和升级，中国数据分析学习网（www.cdachina.cn）新版远程学习平台于2016年7月1日正式上线！新平台增加了在线答疑、在线考试、公开课、预习功能、评测系统、BBS互动等众多模块功能，为学员提供了更加多元化的学习服务和更高效优质的学习环境。

新平台功能亮点：

1、面向所有人群：任何对CPDA或CDA感兴趣的人群都可以免费注册，收看和查阅平台里的资料和视频；

2、学习进度评估：在新平台中，教师可掌握学员学习进展，评估学员学习成果；学员也可通过每个章节的测试评估自己的学习进展；

3、实现资源共享：教师和学员通过平台可实现远程学习、论坛交互、客服系统进行交流，平台更提供强大的资料图书库支持，供教师将相关的延展学习资料分享给学生进行拓展练习，真正实现资源共享；

4、在线答疑：学员听课中遇到的疑问可通过BBS答疑专区随时留言，老师将统一解答，方便师生沟通，达到最佳的学习效果；

5、学习内容更新：平台课程每周都会根据学员的学习进度和面授课时间进行更新，同时设置相应的自测题，让学员真正将面授和远程进行结合，扎实地完成学习并掌握相应技能；

6、丰富、专业的微课呈现：每节微课时长短，内容精炼，重点讲解数据分析在企业中的实际应用，让学员更加直观感受数据分析行业带来变化的同时掌握数据分析技能。

本次远程学习平台的升级历时半年，我们从学生的学习体验与实效出发，更新了多种辅助学习的功能，让学生在远程课程中学习得更扎实有效，掌握更多的数据分析技能。未来我们将持续更新升级学习平台，为学员提供更好的服务，也期待学员在新平台中能学有所获、学以致用！

fin

/ 专注机器学习:Google在苏黎世成立欧洲研发中心 /

编辑 / 协会会员处 冯伟　图 / 崔峻珩　日期 / 2016-06

尽管 Google 在欧州尤其是欧盟委员会那里颇不受待见，但是这依然阻挡不了 Google 在欧洲继续前进的脚步。

Google 将在苏黎世办公室的基础之上，成立一个专注于机器学习的欧洲研发中心。

此前瑞士苏黎世已经成为Google在美国本土之外最大的工程师办公中心，主要负责 Knowledge Graph （中文名为：知识图谱）引擎的开发。除此之外，Google 在今年 I/O 大会上发布的 Allo 聊天应用中内置了机器人会话功能，该功能也是由 Google 苏黎世办公室开发的。

新成立的 Google 欧洲研发中心将专注于以下三个领域：

机器智能；

自然语言处理和理解；

机器感知。

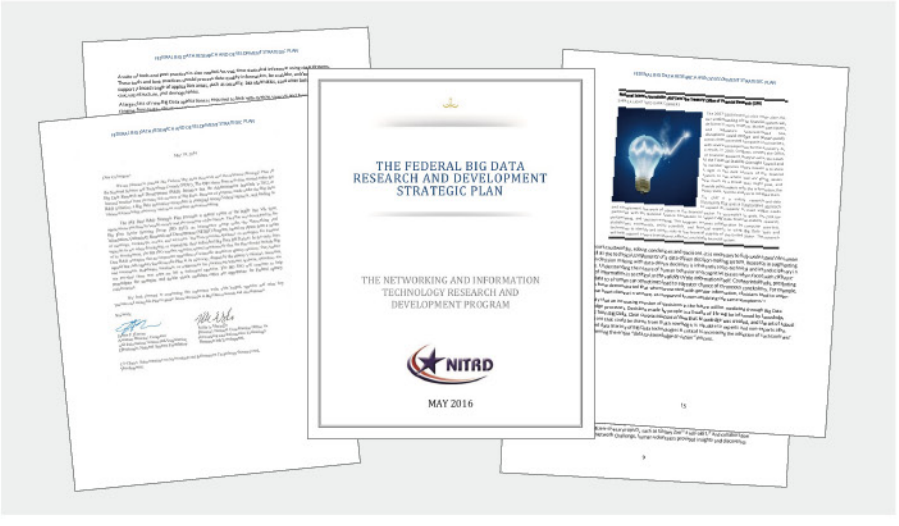
Google 在很早之前就已经发力机器学习领域。Google 旗下的 Translate、Photo 搜索、Inbox 智能回复等功能都是建立在机器学习的基础之上，每一天都有数百万人在使用这些功能。当然，这些功能是 Google 来自全世界的研发者共同努力的成果。至于为何要在欧洲成立研发中心，Google 的官方说法是看中了欧洲一批优质的理工类高校资源，但显然不会是这么简单。 占据统治地位的搜索市场份额让 Google 无法不重视这个庞大的市场；而且 Google 在欧洲还有其他一些前沿业务。

/ 美发布《联邦大数据研究与开发战略计划》 /

编辑 / 协会市场处 佳伊　图 / 崔峻珩　日期 / 2016-05

大数据有可能从根本上改善所有美国人的生活，为了从资源丰富的大数据中获得最大的效益，奥巴马政府于2012年3月就推出了“大数据研究与开发计划”。在此基础上，美国又于2016年5月发布了《联邦大数据研究与开发战略计划》其目标是对联邦机构的大数据相关项目和投资进行指导。

该“计划”主要围绕代表大数据研发关键领域的七个战略进行，包括促进人类对科学、医学和安全所有分支的认识；确保美国在研发领域继续发挥领导作用；通过研发来提高美国和世界解决紧迫社会和环境问题的能力。



/ 海尔大数据探索成果登上《哈佛商业评论》 /

编辑 / 协会市场处 杜艳丽　图 / 崔峻珩　日期 / 2016-05

提到大数据时代，就不得不提及IBM、惠普、Teradata、甲骨文等这些推动我们进入大数据时代的企业，他们利用大数据分析平台的优势资源率先开始掘金大数据市场，成为前互联网时代名副其实的大数据引领者。但是这一次在《哈佛商业评论》探讨的企业大数据战略中，主角不是IBM也不是谷歌，而是变成了引领工业4.0趋势的智能制造企业——海尔。

近日，美国贝恩公司陆原、申文燮在《哈佛商业评论》5月刊上发表的题为《建立大数据能力的6大要素》的文章在业界引发广泛关注，文章在讲述“数据→分析→洞察→决策支撑的产品化、常态化”这一关键因素时，着重以海尔SCRM为例进行了详细阐述，并将海尔SCRM大数据探索定义为企业建立大数据战略与能力成功的典范。

/ 阿里云发布《数据安全白皮书》 /

编辑 / 协会市场处 杜艳丽　图 / 赵金元　日期 / 2016-07

6月29日，阿里云在云栖大会·成都峰会发布《数据安全白皮书》（以下简称白皮书），首次公开了阿里云在保障用户数据安全方面建立的流程、机制以及具体实践办法。

借此机会，针对用户最关心的云上数据传输、使用和存储等问题，阿里云也给出了详实的解读和承诺。并通过权威的认证和审计报告，充分证明其在数据安全方面的合法合规性。

作为行业内主动接受市场和监管检验的云服务商，阿里云开创性地以透明、公开的方式搭建信任关系，再次成为云计算行业的代表，也为云计算市场的健康发展起到了积极的示范作用。

“1+3”的强力安全运营管控理念。从成立第一天起，阿里云就将“用户安全”列为最重要的事情。

白皮书透露，阿里云在架构设计之初就同步考虑了安全架构，不仅将安全的基因植入到整个云平台和各个云产品中，更将数据安全要求嵌入产品开发生命周期的各个环节。

数据是客户资产，阿里云不会移作它用。对于数据的交换、转移与分享，阿里云都提供了标准的加密传输协议，以满足云平台与外界以及系统间传输敏感数据的需求。

全球领先的云计算安全技术。白皮书介绍，所有开发、维护、客服以及其他可能接触到阿里云内部系统的人员，他们的每次登陆都有严格的身份识别，确保帐号与生产设备“不会误用”、“不被盗用”、“不能乱用”的三不原则！

国际多家标准组织肯定阿里云数据安全机制

云计算行业安全认证和行业合规是任何一家云服务提供商正式运营的必备条件，也是提供云服务的资质保障。

/ 引领数据浪潮,创新商业价值 —— 大数据应用高峰论坛 /

文 / 协会市场处 欧阳琦　编辑 / 协会会员处 冯伟　图 / 崔峻珩　日期 / 2016-05

5月27日，中国第二届大数据产业峰会暨大数据应用高峰论坛在贵阳生态会议中心举行，本次大数据应用高峰论坛由数博会组委会主办，中国商业联合会数据分析专业委员会承办，贵州云海数据教育有限公司协办。

CEO巅峰对话大数据时代最有价值的“商业模式”

（从左到右依次是：中国教育电视台著名主持人严楷晨先生，中国商业联合会数据分析专业委员会会长邹东生先生，联想副总裁田日辉先生，阿里巴巴数字化客户运营平台首席架构师周芳雷先生，猎聘网首席数据官单艺先生，中颢润(北京)数据分析师事务所有限公司总经理王芳女士。）

本届“大数据应用高峰论坛”主题是“引领数据浪潮，创新商业价值”，是一场集专业性、实践性、高层次于一体的大数据应用顶级盛会。论坛旨在探寻大数据如何实现商业落地、分享大数据在行业应用中的经典案例以及解决企业大数据应用难题。

中国商业联合会数据分析专业委员会会长邹东生、联想集团副总裁田日辉、阿里巴巴数字化客户运营平台首席架构师周芳雷、猎聘网首席数据官单艺、中颢润(北京)数据分析师事务所有限公司总经理王芳等数据分析行业的领军人士参与论坛并发表演讲。

中国商业联合会数据分析专业委员会会长邹东生发表《引领数据浪潮，创新商业价值》的主题演讲

“大数据,用起来”中国商业联合会数据分析专业委员会会长邹东生在论坛上开门见山亮出自己的观点——“数据是一个工具,只有用起来才能发挥其神奇功效。”邹东生认为,这一工具的核心魅力所在便是数据分析。

美国百货商店之父,约翰·华纳梅克曾经无奈地感叹:“我在广告上的投资有一半是无用的,但是问题是我不知道是哪一半”。困惑约翰·华纳梅克的问题，在大数据时代觅得解决之法。邹东生认为，确切的答案就是数据分析。

“利用数据分析，你将不再为一个决策担忧，你将避免不必要的风险，你可以找到更好的方法，可以预知未来。数据分析让科学决策取代主观判断，一切变得简单可靠。”邹东生对数据分析推崇至极。

让邹东生如此推崇的原因源于一组数据：“运用数据分析的公司盈利能力高出行业平均水平2倍，决策效率高出行业平均水平5倍。”

无独有偶，北京犀数科技有限公司首席数据官孙雪将数据分析视为企业的必修课。“数据分析做好了可以运筹帷幄，决胜千里之外，会引领企业的变革,会引领时代的变革。”

“应用”二字，成为论坛上各个嘉宾演讲的核心。

当求职过程从线下转为线上，大数据驱动了人才资源管理的创新，猎聘网已经将大数据运用于人力资源工作中，成效初显。

中颢润(北京)数据分析师事务所有限公司开发的大数据助力政府精准扶贫方案，运用大数据分析，瞄准精准扶贫“六要素”，即将为全国扶贫工作注入数据力量。

联想集团将大数据作为核心的驱动力，正在运用大数据探索推动制造业企业弯道超车……

一堂盛宴，全国各地大数据运用的先进案例，让大数据不再是虚无的概念，而是人们握在手中的工具，其利刃有劈开一条发展新路之力。

/ 【数博会演讲】引领数据浪潮，创新商业价值！ /

中国商业联合会数据分析专业委员会 邹东生

文 / 协会会员处 冯伟　编辑 / 协会市场处 欧阳琦　图 / 崔峻珩　日期 / 2016-05

大数据，用起来！大数据其实没那么神秘，大数据真正的价值在于帮助企业、帮助消费者，让大家能够看到切身的利益、看到真实的价值，真正把它用起来，让它发挥价值，大数据才能帮我们看到神奇所在。

我今天要跟大家简单地交互几个小问题，到今年我们协会已经成立了8年，实际上应该是最早在中国践行大数据，同时也是最早关注大数据和支持数据分析事务所从业的全国性的行业协会。大家能看到，我们说大数据在改变这个世界，像25年前的互联网，25年前互联网在开始进入的时候，没有人说它很神奇，很多人觉得不过就是个网站，不过就是我需要触网，但是25年以后，互联网创造了太多的神奇，现在能冠以时代名称的，我觉得，除了互联网，就是大数据。而且大数据对于企业、对于个人的影响，我相信它所创造的变化，给我们带来的深远的变化和影响有可能会超越互联网。

那么当然，从实际上，从政府这块，迅速地拉动，或者是密集地出具政策，应该是从2015年起。2015年是做大数据的人应当关注或者记忆的一年，也就是这一年，国家密密麻麻地出台相应的政策，从9月份到12月份，大数据正式作为国家战略。实际上这个名字在很多会议活动上，包括政府议政都一直在提，要尽早地把大数据作为国家战略，但是很高兴我们终于等到了这一天，而且这个时间并不算太晚。

从另外两点聊一聊大数据给我们带来的变化。使用大数据的企业的盈利能力会比行业内的平均企业高出将近1倍，决策效率是行业平均水平的5倍。这意味着什么呢？比如传统行业想在促销或零售活动增加1个点或0.5个点的业绩，都是非常非常不容易的。而它们比平均的行业业态要增长一倍，这意味着一个巨大的竞争力和竞争优势。当然，大数据能带来的变化还有非常非常多，使用大数据帮企业产生的盈利和效果是非常明显的。

第二我们看到了一个新的业态。全球最大的注册服务商IBNB在中国有了分公司，它是全球最大的注册服务商，但是它没有房产。大家可能会知道的优步，现在在中国做专车服务，它是全球最大的出租车公司，但它没有一辆自己的出租车。包括Facebook，没有自己的内容制作人。但是他们有什么？往往进入一个行业、进入一个行业强劲的竞争者往往跟以前不一样。以前是什么呢？是这个行业中产生的竞争对手。现在的业态是什么呢？是别的行业中，跨行业领域，进入到这个行业而且迅速地在这个行业中确认自己的地位。也许它在这个业态并没有比在传统业态更惊艳，但是刚才我提到的3个例子无一例外是运用大数据帮它产生的核心竞争力。这就说明什么呢？在这样一个大数据时代，如果你不数据化，不仅你业内的竞争对手会把你淘汰，而且业外的竞争对手的进入也可能迅速把你淘汰。

回过头来我们再来说说大数据到底是什么。我们这几天看了大数据的各式各样的展览，金融的，各个行业的，技术

的，等等一系列。但是有一点我想和大家分享，我觉得大数据的核心无外乎就是这三点，第一个就是帮助你的用户，或者帮助你自身，提高你的盈利能力，让你赚更多的钱；第二个就是帮助你节约更多的成本，实际上还是跟第一个有关，让你能赚更多的钱；第三个是挖掘机会，这个时候其实反而还不直接，简单地说就是能够真正地预见未来，在未来的千变万化的竞争环境当中，找到自己的先见性的战略，提前做好所有的布局，做好这件事情。大数据无外乎就是带来这三个作用。所以我经常会跟我们的事务所，跟我们的战略合作伙伴说一句话就是：你不用跟我说你使用了多么先进的技术、多么炫的分析方法，或者你把你的经验不停地向别人介绍，你不用说那么多，你就告诉他一件事情，你是不是能帮你的客户赚更多的钱，你能不能做到这一点？如果你能做到，你的客户一定会为你买单。否则的话，大数据就是一种非常光鲜的包装。

那么我们目前在中国的大数据市场里，有什么样的问题在迎接我们呢？我这里总结了简单的几点，可以跟大家在座的交互。第一个现象是我觉得我们现在不仅是政府，也包括我们行业组织，也包括现在的从业机构，也包括我们的客户，其实都面临一个对大数据认知不够的普遍现象。我们协会也接触了很多希望跟协会合作做大数据的一些企业，我们跟他聊的第一句话，他会告诉你：我对大数据感兴趣，我一定要触碰数据，否则将来我会被淘汰。第二句话他就会告诉你：我不懂大数据，你能告诉我它到底是干什么用的吗？

我们有很多做大数据技术的企业对外称：我们对大数据很懂，我们可以做很多先进的平台、技术等等，但是我们会发现很多这样的公司，它下一步会去找做数据分析的事务所、找专家或者找协会，为什么呢？到底数据分析的算法、模型、应用到底该怎么做，我可能懂技术但我不懂研究，我不懂分析。可是你要知道你不懂分析，技术换不来商业价值的提升。

我们的事务所也面临着很多的问题就是做了很多基于小数据体量的项目，但关于大数据、数据融合方面的项目的认知存在不足。这就意味着虽然很多人认为这个行业已经很热了，但我认为它依然还处于1.0时代，没有真正地被大众认知，或者我们的需求还不真实，这个需求是需要完善和认知的过程的。

第二个就是刚才提到的，我们的很多专业的人才，现在是严重不足，什么叫严重不足呢？大数据是什么？大数据是一个园区？一堆服务器？一堆企业内部的数据，这些都不是。我觉得大数据的真正价值是让客户看到，刚才我提到的，能赚到更多的钱，能够节约更多的成本，能够预见未来的变化。它靠什么来使数据增值？靠的是人的分析能力，靠的是人和IP技术的结合，如果没有专业的人，那么很多平台、园区和厂房是没有商业价值的，所以说要关注人才的培养，关注人才的研究能力的问题，只有这样，我们才能让大众充分认知它离我们并不远。

什么叫真正的大数据公司？能够真正地掌控数据，能够在纷繁复杂的数据里找到真正的决策规律。你有这个能力，你就是大数据公司，不管你是叫事务所，还是叫什么名称。你没这个能力，你即使把自己标榜成大数据公司，你仍然不是真正帮企业创造价值的公司。包括一些应用化的平台，实际上现在重复性开发越来越严重，现在的底层的重复性开发让客户看不到真正商业价值的体现。

如何让企业内部，部门与部门的数据，上层与下层的数据，行业内部的数据和行业外部的数据进行有机融合，做到这件事情，你即使没有在研究深度上做得很深，你照样能够迅速地张起来很多架子。但是如果没有数据融合的体现，你做不到真正的大数据。数据的整合不仅仅在于企业内部的整合，更重要的是企业内部和外部的整合。

我们跟美国、欧洲的很多数据协会也有合作，我最感兴趣的是把他们的研究方法、他们的模型、他们成型的一些东西，比如说，客户行为分析的一系列东西拿到我们国内来，我们一起来帮助中国的企业创造商业价值，但往往在这件事情上，跟和国外的企业对接的时候，难度是比较大的。因为他们害怕把他们核心的东西拿到国内来泄密。很多国际化的融合项目产生的效果并不明显。

说了这么多，我想告诉你，其实所有的问题，你看到的是困境，实际上转念一想，它带来的就是机遇，带来的就是，如果你能克服这些问题，如果你能正视这些问题，你就会成为未来的大数据之王，关键看我们有没有能力去应对这件事。

我们应该怎么应对呢？我觉得不仅仅是想接触大数据的企业，也包括想从事大数据，把大数据作为盈利的数据公司、数据分析事务所，我觉得都应该沉下心来想一想，体会真实的需求到底是什么，客户想解决什么问题，这些问题的数据从哪里来，怎么用，他的数据是什么，怎么样整合数据。可以被记录下来的东西都是数据，企业的所有行为都可以转换为数据。你敲击一次按键，我可以把你的行为转换为5个或10个数据字段来存储到我们的分析系统里，来判断它可能产生的变化。

所以，你记录你所有的行为，你就逐渐能具备大数据的体量，同时就能产生相应的变化。没有数据不可怕，可怕的是从今天以后你不去关注数据、采集数据，那么你将来也不会在未来的大数据时代取得竞争地位。

后面我想跟大家说的，大数据是什么，其实在欧美，大数据做了很多年，或者已经有了发展的下一阶段的企业，大家有一个共识：大数据不是技术，大数据是真正的利用现在先进互联网IT的技术去通过深度的研究和分析帮助企业提供决策的精准依据的过程，所以我们要想明白，不要陷入IT的陷阱。什么叫IT陷阱？不停地升级，不停地优化。现在互联网经常用一个词叫迭代。传统企业在拥抱大数据时应该学会使用互联网的迭代行为，就是能够满足我明年或下一阶段的使用就可以了，更多的是要通过数据的滚动，更多地让他们关注到数据和分析的结合和对决策的支持和帮助上。

我们经常会告诉我们的分析师，去分析一个项目的时候，要从需求、数据、方案来看，要更多地关注企业的决策点分析。什么叫决策点分析？企业的定价、企

业的产品、企业的广告测评、包括现在的一些大的战略制定，现在更多地通过量化战略来引导企业的高端决策。所有的大数据都是跟企业的决策直接相关的。所以关注客户的决策需求是大数据破题的关键。

最后总结一下，我们要关注几个关键点，这样能够帮助我们少走弯路，尽可能准确地找到将来的着力点。

做数据研究是很精深的，做不到层层面面，所有的细节都关注，所以有的时候我们的分析师研究一两个关键性的模型，可能需要很长时间不断地试用，把你的数据收集起来，用各种方法去调整，调整后不理想还要再去优化它。而国外往往一个优化的模型可以卖到几千万美金，因为这样的模型它帮企业创造了精准的决策依据，是值这个钱的。

要做这个事情第一要有人，有人至关重要，甚至你想去组建大数据平台也好，或者在IT上帮你收集存储也好，你要先有人，如果你内部有人你才能整合需求，如果你没人你要整外部优秀的人去帮你梳理需求，帮你做整体的规划，否则直接上技术的话，你可能会走很多弯路。

第二你要有分析能力，有人不代表有分析能力。在北京一位优秀的数据分析师年薪在几十万是很正常的一件事，但不是所有的学过数据分析的都能成为数据科学家，他需要在实践中不断深化自己的数据分析能力，而这个过程没有捷径可言，做数据分析不仅仅是实践经验，还要不断探索不断优化，它需要的工作更多，难度也更大。有了人才企业在大数据领域就不会觉得很迷茫，就会找到非常精准的方向。

我们的分析师遍布全国各行各业，事务所也遍布全国几十个省份。刚刚我们提到企业的微场景搭建，什么叫微场景？就是应用大数据，帮助企业构建自己的行业场景，构建自己的细分市场场景。

大数据给企业带来的是你看得懂的东西，我告诉你这个产品的定价应该是多少，你按这个定价可以使你的收益增加多少；你花了二十万的广告，值不值得花这个钱，它会给你在市场占有率上增加多少；它会使你产生什么样的蜕变。所有的这些都是客户能够解读的，客户能不能很舒服地解读，取决于这个微场景的搭建是否成功。我们现在开放的大数据分析平台，这个平台主要是给分析师使用各种模型和算法，而且它有很庞大的数据处理能力，但是它不是微场景，微场景的搭建一定是给企业定制化搭建的，只有定制化搭建的场景，企业才能解读，才有定制化的价值。

第三个关键点是要学会做外部数据的引入，中国政府已经提到了2018年中国政府会开放一些政府大数据，让公众可以获取一定体量的政府数据。从企业而言，企业也应该关注很多从外部引入的数据，也可以找第三方的数据公司帮我们采集。外部数据的引入取决于你能否真正把大数据融合并真正产生效果。所以我希望大家关注外部数据源的采集和收集。

最后我想简单说一说做大数据，我们的认知和理解是大数据的咨询重于技术，大数据没有那么神奇，也没那么昂贵，它完全可以让很多企业享用到大数据带来的支持和帮助。贵有贵的做法，便宜有便宜的做法，但不管是便宜还是贵，要做得惊喜、做得准确，关注客户的需求，关注客户能力的提升，不管你现在付出大还是小，都可以带来深刻的影响和很多的变化。

所以大家关注大数据，让你所在的企业单位尽早地去拥抱数据，这应该是我们所有的企业决策者应该关注的事情。我们希望跟大家更多地交互，让我们的大数据分析能力帮助大家创造更多的价值。

【数博会演讲】数据分析——企业的必修课

北京犀数科技有限公司　孙雪

文 / 协会会员处　冯伟　　编辑 / 协会市场处 欧阳琦　　图 / 崔峻珩　　日期 / 2016-05

都说大数据分析非常重要，但是其实我们很多人并不理解大数据分析和传统意义上的数据分析到底有什么区别，我想在这分享下我对大数据分析的理解。首先我认为大数据分析不是一个词，它的每个字都有它独特的含义。大数据分析有个公式，即“大数据分析=数+据+分+析”，每个字都有什么样的解释呢?

我先说一下“数”，数指的就是数据，但是这个数据并不是我们想象当中的0—9当中的任何一个，而咱们在传统的数据分析当中由于数据来源比较有限比较单一，往往是企业的一些报表数据或者我们调研的数据等等，而这个数据主要以结构化为准，而在大数据时代，数据的源头是多样化的，大数据都是自发形成的，在大体量自发性的数据当中，多半以半结构化和非结构化为准，其中就包括了我们经常在网上看到的声音、视频、文字等等，所以我认为大数据分析的数据其实是指计算机能够处理的所有数据类型。

“数”说完了后我们再看一下“据”，什么叫“据”呢，其实我对“据”的理解是依据和根据，即在大数据时代，由于体量庞大，所以我们必须依据这种分布式的架构体系把你的离线数据和流数据进行分化出来，我是这么理解的“据”，“分”就是分析，我们企业要根据不同的业务类型来考虑要搭建什么应用场景、选择什么模型、采取什么样的策略、用什么样的算法，把这些数据有效地分析出来，把你的价值发挥到最大化，最后是“析”，这个“析”指的是解析和展示，也就是说数据分析的结果通过什么样的手段或方式把它展示出来，让大家能够看得懂、看得清、看得明白，这就可以做到解析了。

那其实我们说数据分析的最高境界是什么，我的理解是“运筹帷幄之中，决策千里之外”，如果要想实现这点，必须把行业数据和企业内部数据结合在一起综合去分析，才可以实现效果。

其实我们以前在没有数据分析做决策的时候，大部分的决策都是凭感觉，凭主观臆断来做决策的，那我看一下依靠这种做出来的决策到底靠不靠谱。

在这为大家找一篇报道，它是《每日邮报》在4月份的一个报道，它记录了一架航班在等待起飞的时候突然有个人提出她身体不舒服不能继续飞行了，乘务员刚把她送下飞机时，这个乘客又突然改口说，她其实并没有生病。那为什么要下飞机呢，主要是她看到她邻座的大叔正在书写一些奇怪的符号，她据此判断这位大叔可能是中东的恐怖分子，可能要执行他的劫机计划，所以她才假装装病，下飞机报警。这个犯罪嫌疑人头发卷卷的，有点络腮胡，有点像中东的恐怖分子。事情的真相到底是什么？这位大叔其实是宾夕法尼亚大学的经济学教授，当时他在飞机上写的这些，其实并不是劫机方案，只是几道数学题而已，暂且不说刚刚举报的大姐智商是不是短路了，但她的警惕性还是值得学习的。但是基于大姐的主观经验的判断，这么一位热爱科学、热爱工作的一位教授，瞬间就被化身成为恐怖分子了，所以基于主观的判断是个误会。所以作为教授的这位经济学家当时的心理阴影面积，估计他自己都计算不出来。

其实通过这个故事告诉我们一个深刻的道理：有时候人的主观判断是不准确的，而且人眼对人脸的识别效果也是有偏差的。还可以看出，如果你对数据分析进行深挖的话，可以挖掘出它背后不为人知的秘密，所以数据分析师非常靠谱的。既然人眼有时候对人脸识别有偏差，那么机器对人脸识别准确度有多高？在Facebook里有个face的项目是专门对人脸技术进行识别，它在14年发布的时候，当时准确性就已经达到了97.25%，在去年它也开发出一款新的识别算法，在看不清面部的情况下，只凭借发型、身材、脸型等状况去判断一个人，准确性也已达到83%，在今年4月份Facebook举办的发布会上，他们宣布已经把身份识别推向市场领域，以个人的身份可以进行标注识别。

人眼有时对人脸识别有偏差，机器对人脸识别的准确性现在已经有所提高了。机器除了拥有一双智慧的眼睛之外，它还拥有一个最强的大脑。比如在前阵子

引发热议的阿尔法狗与李世石的世纪大战，谷歌无人驾驶的汽车，还有微软研究院针对人工智能技术对世界杯和奥斯卡奖项进行成功预测，预测准确性非常高。其实，微软也就即将召开的里约奥运会进行了奖牌归属情况的预测，为了预测准确，微软专门研究收集了从1896年开始的各类奥运奖牌的分布情况、50多个国家、30000多名运动员历年比赛的成绩，除了这些数据，它还收集了参赛国的GDP、人均收入、当地体育项目的规模等等，同时也会采集一些社交网络平台上的数据，看大家的舆论导向是什么。所以这个结果是建立在对多维数据的综合分析的结果上，所以预测结果已经达到很准确的地步了，当然这个在以前是完全做不到的。

现在我简单地介绍一下优步，它没有一辆出租车，为什么能够做到现在的地步？优步是09年成立的，但是发展到现在已经成为全球估值最高的非上市企业，它的业务范围已经遍布全国60多个国家和地区，最新的估值已经达到700多亿美元，这个其实比当时Facebook达到这个数据的时候要将近快了两年。

一款打车软件为什么能获得这么大的成功呢？在我看来，它的主要成功主要得益于它的大数据平台，通过大数据算法所做的出租车数据分析，它目前做的是最好的，比如说它的基于路径的优化，还有基于定价的策略，包括它的业务、客户的满意度评价等等都是基于大量数据的分析基础上来建立的。国际巨头Facebook、微软，他们的一些先进的技术，正是因为他们掌握了这些技术才是他们不断地突破自己。

数据分析除了可以引领企业变革之外，他们也可以引领一个行业的变革，比如随着车联网技术的普及，我们可以看到更多维度的车辆的数据，基于车辆数据的比如车险的定价，或者是风险的分析和计价系统，为这些系统提供了强大的数据支持，同时也引申了UBI，基于使用量的保险类型，它可以基于驾驶员的驾驶信息来评判驾驶风险，从而确定车险价格。

大数据分析不光为保险业带来创新，它给各个行业都带来了巨大价值。以美国为例，它利用大数据可以让零售商增加60%的利润，帮助制造业减少50%的装配成本，基于大数据的智慧医疗平台产值也很高，而且在大数据预测犯罪方面，准确率也达到了90%多。

数据分析会引领我们变革，引领时代的变革，最终引领整个时代发生巨大变革。那么支撑大数据有这种巨大威力的根源是什么？大数据分析跟刚刚所说的人工智能、机器学习、数据挖掘、计算机视觉，以及模式识别等等这些概念到底有什么关系？在我看来，所有都是基于大数据来分析的，所以大数据分析是目的，而其他所有的研究内容和方向都是手段，而最核心的机器学习也是大数据分析的技术实现。

刚刚我把数据分析的用途简单给大家介绍了下，那么是什么支撑着数据分析的强大决策能力？在这我把算法简单地做了归纳。针对每种算法的功能，可以分成分类算法、回归算法、聚类算法、关联分析、异常值检测、特征提取和选择以及我们用这些算法所衍生出来的预测的模型，比如客户价值分析模型、还有一些竞价模型等等。

看了这么多算法可能大家不太理解，我举一个简单的应用实例，比如我现在想把你坐的桌子和椅子有效地分离出来，你要机器去分别，你要怎样去给它数据呢？首先你要给它一些支持这种决策的数据，包括长、宽、高、体积、面积、重量、颜色、材质，你会罗列出来好多信息点，然后你再采集好多种类型的椅子和桌子的数据，然后告诉机器哪个是桌子哪个是椅子，然后就让它去训练模型了，这个模型训练好之后，你再给它相似的数据，它就可以帮你预测出哪个是桌子哪个是椅子，这个是一个简单的应用场景。

如果你有标识信息，也就是你告诉机器哪个是桌子哪个是椅子，那这些算法你会采用分类的算法，如果你的标识信息是分类型的，你要用分类算法，如果是连续性数值，你要用回归来对它进行预测。而如果我只给你一堆数，不告诉你这里面哪些是桌子哪些是椅子，那你就要用物以类聚的聚类算法，把它们按照相似度聚在一起，相似度低和相似度高就可以分离出来了，我们就可以通过数据把桌子和椅子匹配出来了，这是咱们的聚类算法。分类算法和回归算法都属于监督学习，聚类就属于非监督学习。

现在随着标记数据获取越来越难，又衍生出一种新的算法，像半监督学习，我简单跟大家介绍一下。如果你想找桌子椅子的横向信息、纵向信息有什么关联，那你就会用关联信息。而异常值检测，就是如果你发现里面有人为录入的错误，有异常的数据或孤立点存在的话，就可以用异常检测的算法，异常检测算法的实际应用领域像咱们的基于客户的个性化检测，可以通过这个算法把它识别出来。刚才我们说在有效的决策里，哪个是桌子哪个是椅子的时候我们找了好多维度的数据，但这些数据可能会存在冗余，这就需要用特征提取和选择算法对它进行降维的处理，这几个类型跟大家简单介绍一下。

刚才讲了那么多复杂的算法，这些算法也会成为企业在大数据时代如何来挖掘企业核心资产价值的主要性的障碍。这些往往会制约企业在数据化转型时如何能够正确地决策出你的商业价值，我们设计开发出了一款大数据开发平台Datahoop，它整合了刚才所说的大部分算法，所以那些算法你是不用去自己编的，我在这个平台里都往里整合了，你不需要去理解它的内部原理是什么，所以它可以为企业和数据分析师减少大量工作。

这个平台是通过R语言来对接的，R语言作为常用的数据分析语言，它的算法包、算法库是很强大的，这个平台非常好用的一点是它采用了智能化的设计理念，它可以使算法有自动分析的功能，它可以基于你的数据类型自动为你匹配算法，所以就不用讨论使用什么算法，它是可以自动匹配出来的。这对于对算法一窍不通的人也可以顺畅使用。这个平台的使用人次每天都可达到几千人，这些人大部分是分析师或企业的决策者，他们在使用过程中也给了我们很多建议，我们每天也在对算法平台进行优化，相信在这么多人全身心的投入下，这个平台会越做越好。

中国国际大数据及云计算展览会 暨

第四届中国数据分析行业峰会

THE 4TH CHINA DATA ANALYSIS INDUSTRY SUMMIT. 2016

赢在股权 共享未来

2016年 8月6日

北京 · 中国国际展览中心

北三环东路6号中国国际展览中心

一、关于峰会

“2016中国国际大数据及云计算展览会”由中华人民共和国工业和信息化部、中华人民共和国商务部、国家互联网信息办公室指导，经中国国际贸易促进委员会批准，中国国际展览中心集团公司主办，中国商业联合会数据分析专业委员会协办。

第四届中国数据分析行业峰会作为本次展会的一大亮点，由中国商业联合会主办，中国商业联合会数据分析专业委员会承办。

二、峰会亮点

150+专业及大众媒体进行立体全方位多层次报道

政府牵头，市场化运作，响应“双创”号召，政企面对面深度对话

业界专家、行业大咖，共同打造全生态链产业

权威协会联合支持，知名企业落座帝都，各展风采

十余城市主题展区，数十位市长高峰对话，共同搭建产学研一站式推广平台

三、精彩议题

行业发展：探讨“大数据+”的发展之路

技术前沿：大数据产业应用 大数据与智慧城市的创新发展

人才培养：深度剖析数据分析人才培养的重要性

应用创新：大数据时代信息安全、

高峰对话：数据分析人才与技术在企业发展中如何实现平衡

深度探讨：媒体、行业专家、数据分析师事务所、企业畅谈大数据的精准应用

四、与会嘉宾

汪 洋
国务院副总理

苗 圩
工业与信息化部部长

高虎城
国家商务部部长

姜增伟
中国国际贸易促进委员会会长

邹东生
中国商业联合会
数据分析专业委员会会长

雷 军
小米CEO

贾跃亭
乐视CEO

刘强东
京东CEO

吴湘东
中国电信云计算公司总经理

焦 刚
中国联通云数据公司总经理

郁智华
腾讯数据中心总监

李崇辉
中国工商银行数据中心高级经理

五、报名参会

会场地址：北京·中国国际展览中心(北京北三环东路6号)

联系方式：中国商业联合会数据分析专业委员会

协会市场处 010-59000067

/ 关于促进和规范健康医疗大数据应用发展的指导意见 /

编辑 / 协会市场处 欧阳琦　图 / 崔峻珩　日期 / 2016-07

新华社北京6月24日电， 经李克强总理签批，国务院办公厅日前印发《关于促进和规范健康医疗大数据应用发展的指导意见》（以下简称《意见》），部署通过“互联网+健康医疗”探索服务新模式、培育发展新业态，努力建设人民满意的医疗卫生事业，为打造健康中国提供有力支撑。

《意见》指出，要坚持以人为本、创新驱动，规范有序、安全可控，开放融合、共建共享的原则，以保障全体人民健康为出发点，大力推动政府健康医疗信息系统和公众健康医疗数据互联融合、开放共享，积极营造促进健康医疗大数据安全规范、创新应用的发展环境。

到2017年底，实现国家和省级人口健康信息平台以及全国各级药品招标采购业务应用平台互联互通，基本形成跨部门健康医疗数据资源共享共用格局。

到2020年，建成国家医疗卫生信息分级开放应用平台，依托现有资源建成100个区域临床医学数据示范中心，基本实现城乡居民拥有规范化的电子健康档案和功能完备的健康卡，适应国情的健康医疗大数据应用发展模式基本建立，健康医疗大数据产业体系初步形成，人民群众得到更多实惠。

《意见》从夯实应用基础、全面深化应用、规范和推动“互联网+健康医疗”服务、加强保障体系建设等四个方面部署了14项重点任务和重大工程。

主要包括：建设统一权威、互联互通的人口健康信息平台；推动健康医疗大数据资源共享开放；推进健康医疗行业治理、临床和科研以及公共卫生的大数据应用；培育健康医疗大数据应用新业态；研究推广数字化健康医疗智能设备；发展智慧健康医疗便民惠民服务；全面建立远程医疗应用体系；推动健康医疗教育培训应用；推进网络可信体系建设；加强健康医疗数据安全保障；加强法规和标准体系以及健康医疗信息化复合型人才队伍建设等。

《意见》强调，要建立党委政府领导、多方参与、资源共享、协同推进的工作格局。

要从人民群众迫切需求的领域入手，重点推进网上预约分诊、远程医疗和检查检验结果共享互认等便民惠民应用；支持发展医疗智能设备、智能可穿戴设备，加强疑难疾病等重点方面的研究；加快推进基本医保全国联网和异地就医结算；选择一批基础条件好、工作积极性高、隐私安全防范有保障的地区和领域开展健康医疗大数据应用试点。

要研究从财税、投资、创新等方面制定政府支持政策，鼓励和引导社会资本参与健康医疗大数据的基础工程、应用开发和运营服务。

要加快健康医疗数据安全体系建设，加强对涉及国家利益、公共安全、患者隐私、商业秘密等重要信息的保护。

/ 数据的"增值"在于强大的分析能力 /

—— 专访中国商业联合会数据分析专业委员会会长 邹东生

文 / 协会会员处 冯伟　编辑 / 协会市场处 欧阳琦　图 / 崔峻珩　日期 / 2016-07

邹东生先生是中国数据分析行业发起人、奠基人、资深数据分析专家。中商联数据分析专业委员会成立于2008年，是我国数据分析行业唯一的行业协会，目前拥有百家数据分析师事务所及万名数据分析师，积极推动着我国数据分析技术的普及和应用，培养专业人才，促进着行业的健康发展。

在“中国大数据产业峰会暨中国电子商务创新发展大会”即将在贵阳召开前夕，记者通过电话对话邹东生会长，就数据分析的意义、行业的现状、最新的数据观点，以及我省我市致力于发展大数据产业，给予了不少的好建议。

"专业化分析" 才是大数据技术的战略意义！

记者：大数据技术的战略意义是什么？

邹东生：2012年以来，“大数据”已经成为最火热的行业词汇。哈佛大学社会学教授加里·金说：“这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。”这样，围绕大数据商业价值的利用，逐渐成为争相追捧的利润焦点。

但是，大数据技术的战略意义不在于庞大的数据信息，数据本身不能产生价值，只有对数据进行科学有效的分析，才能彰显数据价值。只有通过“专业化分析”，提高对数据的“分析能力”，通过“分析”实现数据的“增值”。

“专业化分析”数据，需要专业人士和专业机构，为民众进行专业化的服务。

记者：看来“大数据分析”已经成为一种产业，请您简单介绍一下这个行业的世界发展状况。

邹东生：当今和未来，政府、企业包括个人决策，基于数据和分析而作出，而非基于经验和直觉。由此数据分析成为行为的决定性因素，同时发展为一个全新的、迅速发展的行业。

数据分析在国外早已广泛应用于各个领域，很多国家成立了相应的数据分析专业服务机构，拥有专业的数据分析人员。

在发达国家，差不多大中型企业里都有专业的数据分析师，从事相关的数据分析工作。

日本有15万多名，瑞典也有10多万名数据分析专业技术人员；美国有近万家专门从事数据分析的服务公司，年营业额达到几千亿美元，英国有3000多家，日本

有1000多家，瑞典有500多家。

蒸蒸日上的中国“大数据分析”产业

记者：正处转型、升级、创新的中国，"大数据分析" 产业的情况是怎样的?

邹东生：我举一个例子，在北京一位专业的数据分析师的年薪可达50万、60万，甚至上百万，可见产业的发展在国内，蒸蒸日上。

在我国，数据分析行业刚刚走过13个年头。从无到有，直至今天不断发展壮大，主要经历了几个阶段，2003年底，原信息产业部职鉴中心正式设立了“数据分析师”认证培训项目，并制定出管理办法；2004年我国首批“数据分析师”诞生；2008年12月，全国数据分析专业委员会成立，直至今日全国已有百家数据分析师事务所、上万名数据分析师。

目前，在贵州有贵州经纬，贵州新阳光，贵州北辰星，贵州华鑫成4家数据分析师事务所，最早的成立于2008年，在金融业，制造业等领域都有所涉及。相信随着大数据事业在贵州的不断发展，事务所也在不断壮大中。

去年9月国务院印发《促进大数据发展行动纲要》，系统部署大数据发展工作，而工信部正在制定《大数据产业“十三五”发展规划》，这些将为数据分析行业带来前所未有的机遇。

记者：这些为“大数据分析”产业带来机遇，同时也会带来挑战。请您谈谈相关的情况。

邹东生：我先说说机遇。数据分析是一个跨学科的边缘学科，横向看其业务涉足各行各业，不再受行业框框的约束；纵向看其服务能力，从大决策到驾驭微管理，可以深入行业内部。同时有数据的、有能力的公司都可以进军这一行业，为服务对象提供实实在在的数据分析效益，使这一行业占得时代先机，前途无量,面对的挑战也不容小觑。

在我国首先是传统观念的排斥，数据分析领域没有很多成熟、成功的研究方法及案例进行推广，还处于探索阶段；还有数据分析师的水平良莠不齐，数据分析师事务所的发展参差不齐，要避免不深化分析水平、一味追求“短平快”的分析，失去了真正的研究能力；同时数据分析师的培育机制还有待完善，普及培训网点分布，培训内容要有实战性。只有以扎实的数据分析能力，为企业决策提供行之有效的观点和决策支撑，才能使数据分析让更多人认知和认可。

数据分析行业发展的五大趋势 不会被“淘汰”！

记者；您是数据分析的专家，对于已经从事或正要跨入这一行业的企业和个人，请分析一下其未来发展的趋势。

邹东生：首先数据及数据业务将成为企业重要的资产。未来企业的竞争，将是对数据分析和运用的竞争，数据已成为民族和国家乃至于企业的新核心资产。数据业务必成为一种具有变革意义的商业模式，成就更多智慧企业。

其次，数据分析能力将成为企业的核心竞争力。互联网时代，企业行为更加依赖数据的分析，数据分析将成为企业IT工作的重心，数据处理将实现平台化，处理大规模数据的应用、统计和定量分析，推动着企业的决策和行动。

数据价值的核心将是数据分析解读和应用。数据分析揭示着各个变量之间可能的关联，具象到行业实践中，执行人能够解读大数据分析的结论，制定出解决问题的方案，可见深谙数据分析的人成为制胜关键。由此数据分析人才更加专业化，数据科学家将成为高端、高薪、受人尊敬的职业。

《哈佛商业评论》宣布，“数据科学家”是二十一世纪最性感的职业。数据科学家就是数据洞察的工程师，具备统计分析、对数据的提取与综合、数据的可视化表示三种能力;知晓计算机科学、数理统计学、图形设计学和人机交互学。

最后，数据分析从业机构的服务领域开始向纵深发展。比如商业公司、软件公司等，依靠在本领域积累的经验，有了可供深度分析和挖掘的数据库的积累，成为业内短期无法撼动的公司。

总的来看，全球数据分析行业正处在蓬勃发展的时期，未来数据产业的规模和产值将不断扩大，将是一个不会被“淘汰”的行业。

贵阳应发展大数据产业,培养自己的“数据精英”。

记者：您关注贵阳数博会吗?

邹东生：从去年贵阳首次举办“数博会”，数据分析专业委员会就给予了高度的关注，不少的数据分析师事务所参加了博览会，现场分享、展示了各自的研究成果。首届我们看到了“数据的引入”，今年委员会同样要参加贵阳“数博会”，让大家看到数据带来的实实在在“益处”。

记者：对于贵州贵阳，近几年正全力发展大数据产业，您有什么好的建议?

邹东生：我建议贵州省从政府、行业组织、企业到个人，要树立强烈的“数据分析意识”，建立良好合作机制，注重数据的储存和应用。政府发挥桥梁作用，让企业了解数据分析带来的收益。去年9月，行业协会举办的“大数据走进校园系列公益讲座”走入了包括贵州在内多所高校，同时开展的“百家企业公益扶持计划”，助推企业数据建设，推动“更多人了解大数据”。

在贵州高校建议设立“数据分析”专业，出版针对高校教学的教材，为今后的发展进行有效的人才储备。

在本地企业内部一定要培养自己的数据分析师，致力于获取数据、数据整理、数据观察到数据挖掘，变成本行业的“数据精英”。

/ 基于Spark的文本情感分析 /

编辑 / 协会会员处 冯伟 图 / 崔峻珩 日期 / 2016-07

文本情感分析是指对具有人为主观情感色彩文本材料进行处理、分析和推理的过程。文本情感分析主要的应用场景是对用户关于某个主题的评论文本进行处理和分析。比如，人们在打算去看一部电影之前，通常会去看豆瓣电影板块上的用户评论，再决定是否去看这部电影。另外一方面，电影制片人会通过对专业论坛上的用户评论进行分析，了解市场对于电影的总体反馈。本文中文本分析的对象为网络短评，为非正式场合的短文本语料，在只考虑正面倾向和负面倾向的情况下，实现文本倾向性的分类。

文本情感分析主要涉及如下四个技术环节。

1.收集数据集：本文中，以分析电影《疯狂动物城》的用户评论为例子，采集豆瓣上《疯狂动物城》的用户短评和短评评分作为样本数据，通过样本数据训练分类模型来判断微博上的一段话对该电影的情感倾向。

2.设计文本的表示模型：让机器“读懂”文字，是文本情感分析的基础，而这首先要解决的问题是文本的表示模型。通常，文本的表示采用向量空间模型，也就是说采用向量表示文本。向量的特征项是模型中最小的单元，可以是一个文档中的字、词或短语，一个文档的内容可以看成是它的特征项组成的集合，而每一个特征项依据一定的原则都被赋予上权重。

3.选择文本的特征：当可以把一个文档映射成向量后，那如何选择特征项和特征值呢？通常的做法是先进行中文分词（本文使用 jieba 分词工具），把用户评论转化成词语后，可以使用 TF-IDF（Term Frequency Inverse Document Frequency，词频-逆文档频率）算法来抽取特征，并计算出特征值。

4.选择分类模型：常用的分类算法有很多，如：决策树、贝叶斯、人工神经网络、K-近邻、支持向量机等等。在文本分类上使用较多的是贝叶斯和支持向量机。本文中，也以这两种方法来进行模型训练。

为什么采用 Spark?

传统的单节点计算已经难以满足用户生成的海量数据的处理和分析的要求。比如，豆瓣网站上《疯狂动物城》电影短评就有 111421 条，如果需要同时处理来自多个大型专业网站上所有电影的影评，单台服务器的计算能力和存储能力都很难满足需求。这个时候需要考虑引入分布式计算的技术，使得计算能力和存储能力能够线性扩展。

Spark 是一个快速的、通用的集群计算平台，也是业内非常流行的开源分布式技术。Spark 围绕着 RDD（Resilient Distributed Dataset）弹性分布式数据集，扩展了广泛使用的 MapReduce计算模型，相比起 Hadoop的 MapReduce 计算框架，Spark 更为高效和灵活。Spark 主要的特点如下：

1.内存计算：能够在内存中进行计算，它会优先考虑使用各计算节点的内存作为存储，当内存不足时才会考虑使用磁盘，这样极大的减少了磁盘 I/O，提高了

效率。

2.惰性求值：RDD 丰富的计算操作可以分为两类，转化操作和行动操作。而当程序调用 RDD 的转化操作（如数据的读取、Map、Filter）的时候，Spark 并不会立刻开始计算，而是记下所需要执行的操作，尽可能的将一些转化操作合并，来减少计算数据的步骤，只有在调用行动操作（如获取数据的行数 Count）的时候才会开始读入数据，进行转化操作、行动操作，得到结果。

3.接口丰富：Spark 提供 Scala，Java，Python，R 四种编程语言接口，可以满足不同技术背景的工程人员的需求。并且还能和其他大数据工具密切配合。例如 Spark 可以运行在 Hadoop 之上，能够访问所有支持 Hadoop 的数据源（如 HDFS、Cassandra、Hbase）。

本文以 Spark 的 Python 接口为例，介绍如何构建一个文本情感分析系统。作者采用 Python 3.5.0，Spark1.6.1 作为开发环境，使用 Jupyter Notebook编写代码。Jupyter Notebook 是由 IPython Notebook 演化而来，是一套基于 Web 的交互环境，允许大家将代码、代码执行、数学函数、富文档、绘图以及其它元素整合为单一文件。在运行 pyspark 的之前，需要指定一下 pyspark 的运行环境，如下所示：

清单 1. 指定 pyspark 的 ipython notebook 运行环境

```
export PYSPARK_PYTHON=ipython3 PYSPARK_DRIVER_PYTHON_OPTS="notebook"
```

接下里就可以在 Jupyter Notebook 里编写代码了。

基于 Spark 如何构建文本情感分析系统。

在文章开头，介绍了文本情感分析主要涉及的四个技术环节。基于 Spark 构建的文本分类系统的技术流程也是这样的。在大规模的文本数据的情况下，有所不同的是文本的特征维度一般都是非常巨大的。试想一下所有的中文字、词有多少，再算上其他的语言和所有能在互联网上找到的文本，那么文本数据按照词的维度就能轻松的超过数十万、数百万维，所以需要寻找一种可以处理极大维度文本数据的方法。

在本文后续章节中，将依次按照基于 Spark 做数据预处理、文本建模、特征提取、训练分类模型、实现待输入文本分类展开讨论。

爬取的数据说明

为了说明文本分类系统的构建过程，作者爬取了豆瓣网络上《疯狂动物城》的短评和评分。

示例数据如下所示：

评分	评论文本
5	做冰棍那机智的不像话！！！全片最爱！！！想吃！！
5	绝对的好片子裂墙推荐。实在是因为另一场满了…随手挑了这个片子。真是 5 分钟一小笑 10 分钟哄堂大笑。看那个又懒又慢树獭简直要锤墙了。旁边法国妹子精辟的吐槽！看！这是我们法国人。我要憋到内伤了。最后散场大家都静坐着等着整首歌放完…五 星好评。2016 年度十佳。
5	不要看任何影评，如果可以预告片都别看，直接买票就好了。你要啥这电影里有啥！
3	最精彩的动画是用想象力拍出真实世界难以实现的故事，而不是用动物化填充一段如果是真人就普通到不能再普通的烂俗故事。笑料有，萌趣有，但更有的是莫名其妙的主旋律和政治正确，恐怕没有评分所体现的那么出色。
4	换了新领导就是不一样。迪士尼暗黑大电影，洛杉矶罪案片风格和内核。还真是动物乌托邦，美国针对有色人种，欧洲针对难民，天朝针对公知和五毛吗？人设精彩，细节丰富，但要说创意超《头脑特工队》显然就不实事求是了。
.....	

表 1. 示例数据

表格中每一行为一条评论数据，按照“评分，评论文本”排放，中间以制表符切分，评分范围从 1 分到 5 分，这样的数据共采集了 116567 条。

数据预处理

这一节本文是要说明用 Spark 是如何做数据清洗和抽取的。在该子系统中输入为爬虫的数据，输出为包含相同数量好评和坏评的 Saprk 弹性分布式数据集。

Spark 数据处理主要是围绕 RDD（Resilient Distributed Datasets）弹性分布式数据集对象展开，本文首先将爬虫数据载入到 Spark 系统，抽象成为一个 RDD。可以用 distinct 方法对数据去重。数据转换主要是用了 map 方法，它接受传入的一个数据转换的方法来按行执行方法，从而达到转换的操作它只需要用一个函数将输入和输出映射好，那么就能完成转换。数据过滤使用 filter 方法，它能够保留判断条件为真的数据。可以用下面这个语句，将每一行文本变成一个 list，并且只保留长度为 2 的数据。

清单 2. Spark 做数据预处理

```
originData=sc.textFile('YOUR_FILE_PATH')
originDistinctData=originData.distinct()
rateDocument=originDistinctData.map(lambda line : line.split('\t')).\
filter(lambda line : len(line)==2)
```

清单 3. 统计数据基本信息

```
fiveRateDocument=rateDocument.filter(lambda line : int(line[0])==5)
fiveRateDocument.count()
```

本文得到，五分的数据有 30447 条，4 分、3 分、2 分、1 分的数据分别有 11711 条，123 条，70 条。打五分的毫无疑问是好评；考虑到不同人对于评分的不同偏好，对于打四分的数据，本文无法得知它是好评还是坏评；对于打三分及三分以下的是坏评。

下面就可以将带有评分数据转化成为好评数据和坏评数据，为了提高计算效率，本文将其重新分区。

清单 4. 合并负样本数据

```
negRateDocument=oneRateDocument.union(twoRateDocument).\
union(threeRateDocument)
negRateDocument.repartition(1)
```

通过计算得到，好评和坏评分别有

30447 条和 2238 条，属于非平衡样本的机器模型训练。本文只取部分好评数据，好评和坏评的数量一样，这样训练的正负样本就是均衡的。最后把正负样本放在一起，并把分类标签和文本分开，形成训练数据集

清单 5. 生成训练数据集

```
posRateDocument=sc.
parallelize(fiveRateDocument.
take(negRateDocument.
count())).repartition(1)
allRateDocument=negRateDocument.
union(posRateDocument)
allRateDocument.repartition(1)
rate=allRateDocument.map(lambda
s : ReduceRate(s[0]))
document=allRateDocument.
map(lambda s: s[1])
```

文本的向量表示和文本特征提取

这一节中，本文主要介绍如何做文本分词，如何用 TF-IDF 算法抽取文本特征。将输入的文本数据转化为向量，让计算能够“读懂”文本。

解决文本分类问题，最重要的就是要让文本可计算，用合适的方式来表示文本，其中的核心就是找到文本的特征和特征值。相比起英文，中文多了一个分词的过程。本文首先用 jieba 分词器将文本分词，这样每个词都可以作为文本的一个特征。jieba 分词器有三种模式的分词：

1.精确模式，试图将句子最精确地切开，适合文本分析；

2.全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；

3.搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

这里本文用的是搜索引擎模式将每一句评论转化为词。

清单 6. 分词

```
words=document.map(lambda
w:"/".\
join(jieba.cut_for_search(w))).\
map(lambda line: line.split("/"))
```

出于对大规模数据计算需求的考虑，spark 的词频计算是用特征哈希（HashingTF）来计算的。特征哈希是一种处理高维数据的技术，经常应用在文本和分类数据集上。普通的 k 分之一特征编码需要在一个向量中维护可能的特征值及其到下标的映射，而每次构建这个映射的过程本身就需要对数据集进行一次遍历。这并不适合上千万甚至更多维度的特征处理。

特征哈希是通过哈希方程对特征赋予向量下标的，所以在不同情况下，同样的特征就是能够得到相同的向量下标，这样就不需要维护一个特征值及其下表的向量。

要使用特征哈希来处理文本，需要先实例化一个 HashingTF 对象，将词转化为词频，为了高效计算，本文将后面会重复使用的词频缓存。

清单 7. 训练词频矩阵

```
hashingTF = HashingTF()
tf = hashingTF.transform(words)
tf.cache()
```

缺省情况下，实例化的 HashingTF 特征维数 numFeatures 取了 220次方维，在 spark 的源码中可以看到，HashingTF 的过程就是对每一个词作了一次哈希并对特征维数取余得到该词的位置，然后按照该词出现的次数计次。所以就不用像传统方法一样每次维护一张词表，运用 HashingTF 就可以方便的得到该词所对应向量元素的位置。当然这样做的代价就是向量维数会非常大，好在 spark 可以支持稀疏向量，所以计算开销并不大。

词频是一种抽取特征的方法，但是它还有很多问题，比如在这句话中“这几天的天气真好，项目组的老师打算组织大家一起去春游。“的”相比于“项目组”更容易出现在人们的语言中，“的”和“项目组”同样只出现一次，但是项目组对于这句话来说更重要。

本文采用 TF-IDF 作为特征提取的方法，它的权重与特征项在文档中出现的评率成正相关，与在整个语料中出现该特征项的文档成反相关。下面依据 tf 来计算逆词频 idf，并计算出 TF-IDF

清单 8. 计算 TF-IDF 矩阵

```
idfModel = IDF().fit(tf)
tfidf = idfModel.transform(tf)
```

至此，本文就抽取出了文本的特征，并用向量去表示了文本。

训练分类模型

在这一小节中，本文介绍如何用 Spark 训练朴素贝叶斯分类模型，这一流程的输入是文本的特征向量及已经标记好的分类标签。在这里本文得到的是分类模型及文本分类的正确率。

现在，有了文本的特征项及特征值，也有了分类标签，需要用 RDD 的 zip 算子将这两部分数据连接起来，并将其转化为分类模型里的 LabeledPoint 类型。并随机将数据分为训练集和测试集，60%作为训练集，40%作为测试集。

清单 9. 生成训练集和测试集

```
zipped=rate.zip(tfidf)
data=zipped.map(lambda line:Labele
dPoint(line[0],line[1]))
training, test = data.randomSplit([0.6,
0.4], seed = 0)
```

本文用训练数据来训练贝叶斯模型，得到 NBmodel 模型来预测测试集的文本特征向量，并且计算出各个模型的正确率，这个模型的正确率为 74.83%。

清单 10. 训练贝叶斯分类模型

```
NBmodel = NaiveBayes.
train(training, 1.0)
predictionAndLabel = test.
map(lambda p : (NBmodel.predict(p.
features), p.label))
accuracy = 1.0 *
predictionAndLabel.filter(lambda x: 1.0 \
if x[0] == x[1] else 0.0).count() / test.
count()
```

可以看出贝叶斯模型最后的预测模型并不高，但是基于本文采集的数据资源

有限，特征提取过程比较简单直接。所以还有很大的优化空间，在第四章中，本文将介绍提高正确率的方法。

分类未标记文档

现在可以用本文训练好的模型来对未标记文本分类，流程是获取用户输入的评论，然后将输入的评论文本分词并转化成 tf-idf 特征向量，然后用 3.4 节中训练好的分类模型来分类。

清单 11. 分类未分类文本

```
yourDocument=input("输入待分类的评论：")
yourwords="/".join(jieba.cut_for_search(yourDocument)).split("/")
yourtf = hashingTF.transform(yourwords)
yourtfidf=idfModel.transform(yourtf)
print('NaiveBayes Model Predict:',NBmodel.predict(yourtfidf),'
```

当程序输入待分类的评论：“这部电影没有意思，剧情老套，真没劲，后悔来看了”。

程序输出为“NaiveBayes Model Predict: 0.0”。

当程序输入待分类的评论：“太精彩了讲了一个关于梦想的故事剧情很反转制作也很精良”。

程序输出为“NaiveBayes Model Predict: 1.0”。

至此，最为简单的文本情感分类系统就构建完整了。

提高正确率的方法

在第三章中，本文介绍了构建文本分类系统的方法，但是正确率只有 74.83%，在这一章中，本文将讲述文本分类正确率低的原因及改进方法。

文本分类正确率低的原因主要有：

1.文本预处理比较粗糙，可以进一步处理，比如去掉停用词，去掉低频词；

2.特征词抽取信息太少，搜索引擎模式的分词模式不如全分词模式提供的特征项多；

3.朴素贝叶斯模型比较简单，可以用其他更为先进的模型算法，如 SVM；

4.数据资源太少，本文只能利用了好评、坏评论各 2238 条。数据量太少，由于爬虫爬取的数据，没有进行人工的进一步的筛选，数据质量也无法得到 100%的保证。

下面分别就这四个方面，本文进一步深入的进行处理，对模型进行优化。

数据预处理中去掉停用词

停用词是指出现在所有文档中很多次的常用词，比如“的”、“了”、“是”等，可以在提取特征的时候将这些噪声去掉。

首先需要统计一下词频，看哪些词是使用最多的，然后定义一个停用词表，在构建向量前，将这些词去掉。本文先进行词频统计，查看最常用的词是哪些。

清单 12. 统计词频

```
text=words.flatMap(lambda w:w)
wordCounts = text.map(lambda word: (word, 1))\
.reduceByKey(lambda a, b: a+b).\
sortBy(lambda x: x[1],ascending=False)
wordCounts.take(10)
```

通过观察，选择出现次数比较多，但是对于文本情感表达没有意义的词，作为停用词，构建停用词表。然后定义一个过滤函数，如果该词在停用词表中那么需要将这个词过滤掉。

清单 13. 去掉停用词

```
stopwords = set([" ","的","了","是","就","吧",……])
def filterStopWords(line):
 for i in line:
 if i in stopwords:
 line.remove(i)
 return line
words=words.map(lambda w : filterStopWords(w))
```

尝试不用的分词模式

本文在分词的时候使用的搜索引擎分词模式，在这种模式下只抽取了重要的关键字，可能忽略了一些可能的特征词。可以把分词模式切换到全分词模式，尽可能的不漏掉特征词，同样的模型训练，正确率会有 1%～2%的提升。

清单 14. 全分词模式分词

```
words=document.map(lambda w:"/".join(jieba.\
cut(w, cut_all=True))).\
map(lambda line: line.split("/"))
```

更换训练模型方法

在不进行深入优化的情况下，SVM 往往有着比其他分类模型更好的分类效果。下面在相同的条件下，运用 SVM 模型训练，最后得到的正确率有 78.59%。

清单 15. 用支持向量机训练分类模型

```
SVMmodel = SVMWithSGD.train(training, iterations=100)
predictionAndLabel = test.map(lambda p : (SVMmodel.predict(p.features), p.label))
accuracy = 1.0 * predictionAndLabel.filter(lambda x: 1.0 if x[0] == x[1] else 0.0).count() / test.count()
```

训练数据的问题

本文只是为了演示如何构建这套系统，所以爬取的数据量并不多，获取的文本数据也没有人工的进一步核对其正确性。如果本文能够有更丰富且权威的数据源，那么模型的正确率将会有较大的提高。

作者对中国科学院大学的谭松波教授发布的酒店产品评论文本做了分类系统测试，该数据集是多数学者公认并且使用的。用 SVM 训练的模型正确率有 87.59%。

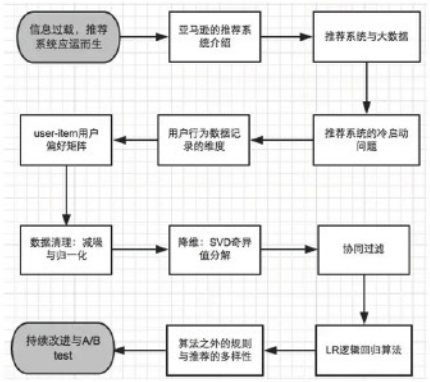
总结

本文向读者详细的介绍了利用 Spark 构建文本情感分类系统的过程，从数据的清洗、转换，Spark 的 RDD 有 Filter、Map 方法可以轻松胜任；对于抽取文本特征，Spark 针对大规模数据的处理不仅在计算模型上有优化，还做了算法的优化，它利用哈希特征算法来实现 TF-IDF，从而能够支持上千万维的模型训练；对于选择分类模型，Spark 也实现好了常用的分类模型，调用起来非常方便。最后希望这篇文章可以对大家学习 spark 和文本分类有帮助。

/ 深入浅出之推荐系统原理应用介绍 /

文 / neil 编辑 / 协会市场处 杜艳丽 插图 / 协会会员处 冯伟 日期 / 2016-06

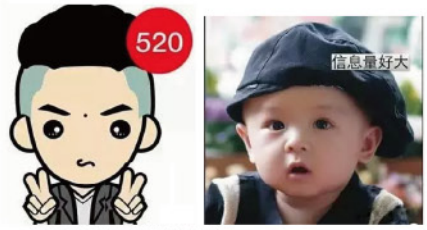
A First Glance



为什么需要推荐系统——信息过载

随着互联网行业的井喷式发展，获取信息的方式越来越多，人们从主动获取信息逐渐变成了被动接受信息，信息量也在以几何倍数式爆发增长。举一个例子，PC时代用google reader，常常有上千条未读博客更新；如今的微信公众号，也有大量的红点未阅读。垃圾信息越来越多，导致用户获取有价值信息的成本大大增加。为了解决这个问题，我个人就采取了比较极端的做法：直接忽略所有推送消息的入口。但在很多时候，有效信息的获取速度极其重要。

由于信息的爆炸式增长，对信息获取的有效性，针对性的需求也就自然出现了。推荐系统应运而生。



亚马逊的推荐系统

最早的推荐系统应该是亚马逊为了提升长尾货物的用户抵达率而发明的。已经有数据证明，长尾商品的销售额以及利润总和与热门商品是基本持平的。

亚马逊网站上在线销售的商品何止百万，但首页能够展示的商品数量又极其有限，给用户推荐他们可能喜欢的商品就成了一件非常重要的事情。当然，商品搜索也是一块大蛋糕，亚马逊的商品搜索早已经开始侵蚀谷歌的核心业务了。

在亚马逊的商品展示页面，经常能够看见：浏览此商品的顾客也同时浏览。

这就是非常典型的推荐系统。八卦

一下：”剁手族”的兴起，与推荐系统应该有一定关系吧，哈哈。

推荐系统与大数据

大数据与云计算，在当下非常热门。不管是业内同事还是其他行业的朋友，大数据都是一个常谈的话题。就像青少年时期热门的话题：“性”,大家都不太懂，但大家都想说上几句。业内对于大数据的使用其实还处于一个比较原始的探索阶段，前段时间听一家基因公司的CEO说，现在可以将人类的基因完全导出为数据，但这些数据毫无规律，能拿到这些数据，但根本不知道可以干什么。推荐系统也是利用用户数据来发现规律，相对来说开始得更早，运用上也比较成熟。

冷启动问题

推荐系统需要数据作为支撑。但亚马逊在刚刚开始做推荐的时候，是没有大量且有效的用户行为数据的。这时候就会面临着“冷启动”的问题。没有用户行为数据，就利用商品本身的内容数据。这就是推荐系统早期的做法。

基于内容的推荐：

1.tag 给商品打上各种tag：运动商品类，快速消费品类，等等。粒度划分越细，推荐结果就越精确

2.商品名称，描述的关键字 通过从商品的文本描述信息中提取关键字，从而利用关键字的相似来作推荐

3.同商家的不同商品 用户购买了商店的一件商品，就推荐这个商店的其他热销商品

4.利用经验，人为地做一些关联 一个经典的例子就是商店在啤酒架旁边摆上纸尿布。那么，在网上购买啤酒的人，也可以推荐纸尿布?

由于内容的极度复杂性，这一块儿的规则可以无限拓展。基于内容的推荐与用户行为数据没有关系，在亚马逊早期是比较靠谱的策略。但正是由于内容的复杂性，也会出现很多错误的推荐。

比如：小明在网上搜索过保时捷汽车模型。然后推荐系统根据关键字，给小明推荐了价值200万的保时捷911.....

用户行为数据—到底在记录什么

在游戏里面，我们的人物角色是一堆复杂的数据，这叫做数据存储；这些数据以一定的结构组合起来，这叫做数据结构。同样地，在亚马逊眼里，我们就是一张张表格中一大堆纷繁复杂的数字。举一个栗子：

小明早上9点打开了亚马逊，先是浏览了首页，点击了几个热销的西装链接，然后在搜索栏输入了nike篮球鞋，在浏览了8双球鞋后，看了一些购买者的评价，最终选定了air jordan的最新款。

这就是一条典型的用户行为数据。亚马逊会将这条行为拆分成设定好的数据块，再以一定的数据结构，存储到亚马逊的用户行为数据仓库中。每天都有大量的用户在产生这样的行为数据，数据量越多，可以做的事情也就越强大。

user-item 用户偏好矩阵

收集数据是为了分析用户的偏好，形成用户偏好矩阵。比如在网购过程中，用户发生了查看，购买，分享商品的行为。这些行为是多样的，所以需要一定的加权算法来计算出用户对某一商品的偏好程度，形成user-item用户偏好矩阵。

喜好程度	iphone	macbook	冰淇淋
小明	0.9	0.8	0.99
小芳	0.8	0.9	0.98

数据清理

当我们开始有意识地记录用户行为数据后，得到的用户数据会逐渐地爆发式增长。就像录音时存在的噪音一样，获取的用户数据同样存在着大量的垃圾信息。因此，拿到数据的第一步，就是对数据做清理。其中最核心的工作，就是减噪和归一化：

减噪：用户行为数据是在用户的使用过程中产生的，其中包含了大量的噪音和用户误操作。比如因为网络中断，用户在短时间内产生了大量点击的操作。通过一些策略以及数据挖掘算法，来去除数据中的噪音。

归一化：清理数据的目的是为了通过对不同行为进行加权，形成合理的用户偏好矩阵。用户会产生多种行为，不同行为的取值范围差距可能会非常大。比如：点击次数可能远远大于购买次数，直接套用加权算法，可能会使得点击次数对结果的影响程度过大。于是就需要归一算法来保证不同行为的取值范围大概一致。

最简单的归一算法就是将各类数据来除以此类数据中的最大值，以此来保证所有数据的取值范围都在[0,1]区间内。

降维算法——SVD奇异值分解

通过记录用户行为数据，我们得到了一个巨大的用户偏好矩阵。随着物品数量的增多，这个矩阵的列数在不断增长，但对单个用户来说，有过行为数据的物品数量是相当有限的，这就造成了这个巨大的用户偏好矩阵实际上相当稀疏，有效的数据其实很少。SVD算法就是为了解决这个问题发明的。

偏好程度	白菜	菠菜	生菜	西瓜	苹果	牛仔裤	针织衫
小明	0.9	0.9	0.9		0.9	0.9	
小芳	0.9		0.9	0.9	0.9		0.9

偏好程度	蔬菜	水果	休闲服
小明	0.9	0.9	0.9
小芳	0.9	0.9	0.9

将大量的物品提取特征，抽象成了3大类：蔬菜，水果，休闲服。这样就将稀疏的矩阵缩小，极大的减少了计算量。但这个例子仅仅是为了说明SVD奇异值分解的原理。

真正的计算实施中，不会有人为的提取特征的过程，而是完全通过数学方法进行抽象降维的。通过对矩阵相乘不断的拟合，参数调整，将原来巨大的稀疏的矩阵，分解为不同的矩阵，使其相乘可以得到原来的矩阵。这样既可以减少计算量，又可以填充上述矩阵中空值的部分。协同过滤算法

我一直在强调用户行为数据，目的就是为介绍协同过滤算法做铺垫。协同过滤，Collaborative Filtering，简称CF，广泛应用于如今的推荐系统中。通过协同过滤算法，可以算出两个相似度：user-user相似度矩阵； item-item相似度矩阵。

user-user 相似度

相似度	用户C	用户D	用户E
用户A	0.9	0.9	0.9
用户B	0.9	0.9	0.9

item-item 相似度

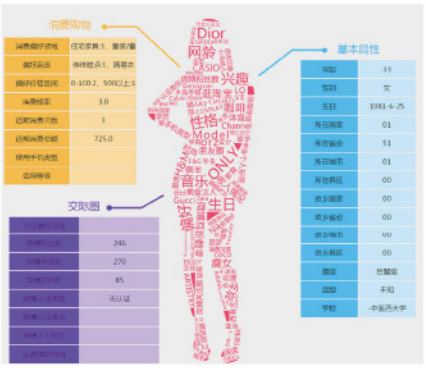
相似度	奔驰	凯迪拉克	沃尔沃
奥迪	0.9	0.9	0.9
宝马	0.9	0.9	0.9

为什么叫做协同过滤？是因为这两个相似度矩阵是通过对方来计算出来的。举个栗子：100个用户同时购买了两种物品A和B，得出在item-item相似度矩阵中A和B的相似度为0.8; 1000个物品同时被用户C和用户D购买，得出在user-user相似度矩阵中C和D的相似度是0.9. user-user, item-item的相似度都是通过用户行为数据来计算出来的。

计算相似度的具体算法，大概有几种：欧几里得距离，皮尔逊相关系数，Cosine相似度，Tanimoto系数。具体的算法，有兴趣的同学可以google.

用户画像

用户画像关联阅读：经典：系统性阐述用户画像数据建模方法。



提到大数据，不能不说用户画像。经常看到有公司这样宣传：“掌握了千万用户的行为数据，描绘出了极其有价值的用户画像，可以为每个app提供精准的用户数据，助力app推广。” 这样的营销广告经不起半点推敲。

用户对每个种类的app的行为都不同，得到的行为数据彼此之间差别很大，比如用户在电商网站上的行为数据，对音乐类app基本没有什么价值。推荐系统的难点，其中很大一部分就在于用户画像的积累过程极其艰难。简言之，就是用户画像与业务本身密切相关。

LR逻辑回归

基于用户偏好矩阵，发展出了很多机器学习算法，在这里再介绍一下LR的思想。具体的逻辑回归，又分为线性和非线性的。其他的机器学习算法还有：K均值聚类算法，Canopy聚类算法，等等。有兴趣的同学可以看看July的文章。链接在最后的阅读原文。

LR逻辑回归分为三个步骤：

1.提取特征值

2.通过用户偏好矩阵，不断拟合计算，得到每个特征值的权重

3.预测新用户对物品的喜好程度

举个栗子：

小明相亲了上千次，我们收集了大量的行为数据，以下数据仅仅是冰山一角。

女生姓名	个性开朗程度	颜值	喜爱程度
小红	1	9	45%
小绿	2	8	40%
小黄	9	5	30%
……			

通过大量的拟合计算得出，特征值“个性开朗程度”的权重为30%，“颜值”的权重为70%。哎，对这个看脸的世界已经绝望了，写完这篇文章，就去订前往韩国的机票吧。

然后，通过拟合出的权重，来预测小明对第一千零一次相亲对象的喜爱程度。

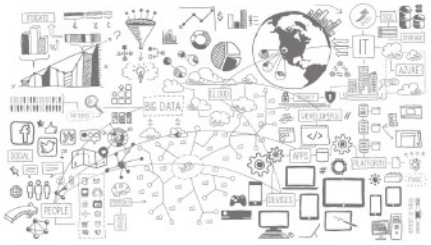
姓名	个性开朗程度	颜值	预测的喜爱程度
翠花	3	8	42%

这就是LR逻辑回归的原理。具体的数学算法，有兴趣的同学可以google之。

如何利用推荐系统赚钱

还是以亚马逊为例。小明是个篮球迷，每个月都会买好几双篮球鞋。通过几个月的购买记录，亚马逊已经知道小明的偏好，准备给小明推荐篮球鞋。但篮球鞋品牌这么多，推荐哪一个呢？笑着说：哪个品牌给我钱多，就推荐哪个品牌。这就是最简单的流量生意了。这些都叫做：商业规则。

但在加入商业规则之前，需要让用户感知到推荐的准确率。如果一开始就强推某些置顶的VIP资源，会极大地损害用户体验，让用户觉得推荐完全没有准确性。这样的后果对于推荐系统的持续性发



展是毁灭性的。

过滤规则

协同过滤只是单纯地依赖用户行为数据，在真正的推荐系统中，还需要考虑到很多业务方面的因素。

以音乐类app为例。周杰伦出了一张新专辑A，大部分年轻人都会去点击收听，这样会导致其他每一张专辑相似专辑中都会出现专辑A。这个时候，再给用户推荐这样的热门专辑就没有意义了。所以，过滤掉热门的物品，是推荐系统的常见做法之一。这样的规则还有很多，视不同的业务场景而定。

推荐的多样性

与推荐的准确性有些相悖的，是推荐的多样性。比如说推荐音乐，如果完全按照用户行为数据进行推荐，就会使得推荐结果的候选集永远只在一个比较小的范围内：听小清新音乐的人，永远也不会被推荐摇滚乐。这是一个很复杂的问题。在保证推荐结果准确的前提下，按照一定的策略，去逐渐拓宽推荐结果的范围，给予推荐结果一定的多样性，这样才不会腻嘛。

持续改进

推荐系统具有高度复杂性，需要持续地进行改进。可能在同一时间内，需要上线不同的推荐算法，做A/B test。根据用户对推荐结果的行为数据，不断对算法进行优化，改进。要走的路还很长：路漫漫其修远兮，吾将上下而求索。

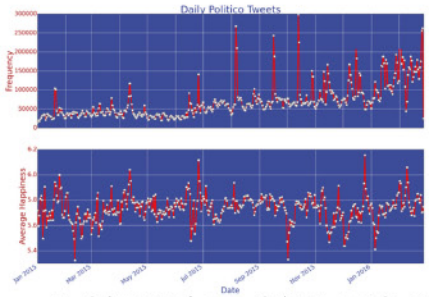
/ 大数据为你预测2016美国大选 /

编辑 / 协会市场处 佳伊　插图 / 崔峻珩　日期 / 2016-04

近年来，社交媒体逐渐成为民众在大选时发表观点和对候选人意见的渠道。推特，做为一个公共和广泛使用的渠道，提供了一个衡量和预测竞选动态的平台。现在，“超级星期二”已经过去，让我们来分析一下党内初选的情况。

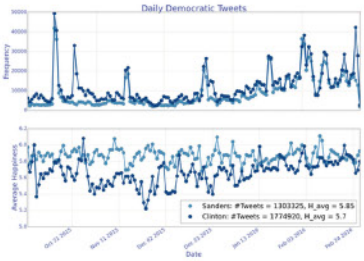
我们以模式匹配的方式从推特的API 10%的样本中收集了关于每个候选人的政治性推文。以下，我们展示了2015年1月1日至2016年2月25日的推文日频率以及相应的平均感情指数（由我们的LaBMT幸福感数据库得出）。

2015年期间的峰值发生在电视辩论期间。至于平均感情指数则需要进行进一步的分析，我们将分析集中在几个主要候选人上。

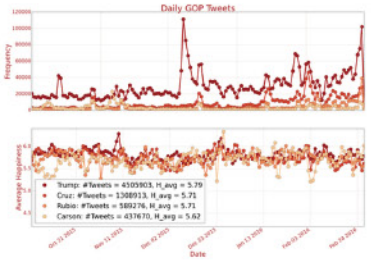


让我们从两个民主党候选人开始分析：希拉里·克林顿和伯纳德·桑德斯。对于每个候选人，分析都包括各自支持者和反对者的推文，以及包括多位候选人的推文。在幸福感时间序列中，我们可以看到桑德斯的平均指数比克林顿的略高（5.85比5.7）。

在两党领先的候选者中，关于桑德斯的推文的关键词有最高的平均幸福指数。在本文后面部分，将会分析是哪些具体的词语造成了这个差异。



至于共和党，在同样的时间序列中，特朗普在推文中的提及率占有明显的领先优势。其他候选人在同一个坐标轴中几乎不可见。而关于每位共和党候选人的推文的幸福感指数则可进行比较，特朗普相对于克鲁兹和卢比奥有微弱的优势，对于卡森有明显的优势。并且，特朗普的平均幸福指数比克林顿略高（5.79比5.70），但仍比桑德斯低（5.79比5.85）。



知道了相对幸福指数还只是一个开始，我们要知道对于每位候选人而言是哪些词推动了分值。在下面的词语图中，词语的颜色根据感情状态表示（越蓝越快乐，越紫越悲伤），词语的大小由加权平均tf-idf值决定，一种原始频率和相对“惊奇”因素的结合。

桑德斯

在桑德斯的词图中，词语的大小差不多，表明关于桑德斯的竞选有比较多的可述范围，他的推文反映了一系列的讨论主题，其中，偏正面的词汇包括了“支持”，“真相”，“赞同”，“胜利”。偏负面的词汇有着政治的天然属性，反映了他对华尔街的贪婪的抵制以及一般民众所面对的财务问题，而这也正是他的竞选运动的焦点。

克林顿

克林顿的词图中，正面词汇与负面词汇相交织，其中比较重要的正面词汇有“经验”，“才能”，“女性”，“世界”；负面词汇有“犯罪”，“调查”和“谎言”，也许是与电子邮件服务器丑闻相关。

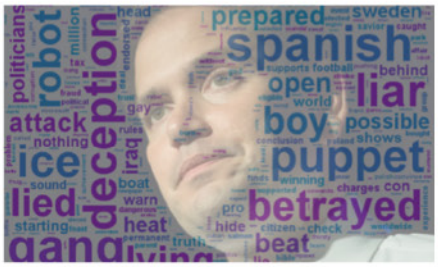
共和党的词图有着相似的主题，特朗普的权重最大的词汇包括支持者形容的“前行”，以及现在的共和党初选中的“胜利”；负面词汇，或许来自于他的反对者，包括“羞辱”，“攻击”，“种族主义”，“骗子”和“危险”。

参议院克鲁兹有着相似的词图，反映了他的支持者的政治理想，而负面词汇则多是政治行话及“谎言”。卢比奥和卡森则是负面词汇偏多。

特朗普

克鲁兹

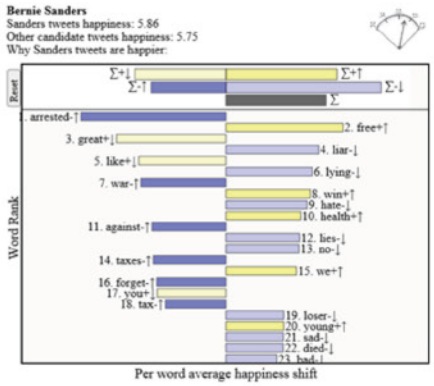
卢比奥



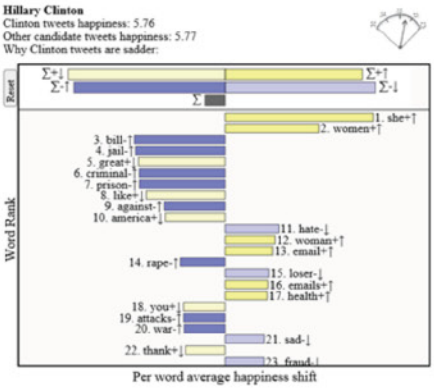
卡森

接下来，我们用词汇转移图对平均幸福指数进行了量化分析。词汇转移图展示了最重要的词汇在比较集和参照集之间的相对位置。比较集包括了我们所收集的所有政治推文，而不包括竞选者，这些推文都产生于2016年 1月至2月25日之间。

第一个词汇转移图比较了与桑德斯相关的推文和其他政治推文。无论是对比参照分布（5.86比5.75）还是其他参选者，与桑德斯相关的推文相对来说更积极一些。这种转移是由于对诸如“自由”，“胜利”，“健康”，“我们”，“年轻”，“民主”，“大学”等正面词汇的越来越多的提及，以及对诸如“骗子”，“谎言”，“憎恨”，“不”，“失败者”，“坏”等负面词汇的越来越少的提及。使用率最高的负面词汇是“逮捕”，“抗议”，也许是由于参议院桑德斯民主权利运动中因抗议隔离政策而被捕。

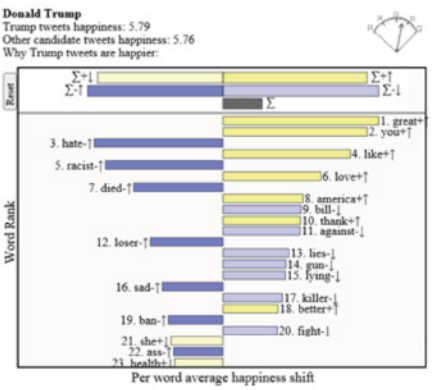


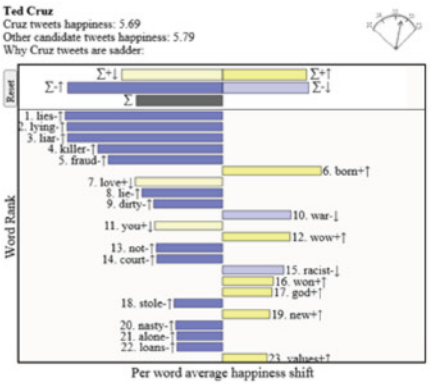
克林顿的词汇转移图与参照分布比较相似（5.76比5.77）。负面词汇包括电子邮件调查及“监狱”，“犯罪”，“囚犯”，“丑闻”等。此外，“票据”是以负面形式呈现的词汇（被理解为支付票据），但是在克林顿这里则是指比尔·克林顿。正面词汇主要有“她”，“女性”，“感谢”，“健康”，而负面词汇“憎恨”，“悲伤”，“失败者”，“诈骗”，“种族主义”较少被提及。

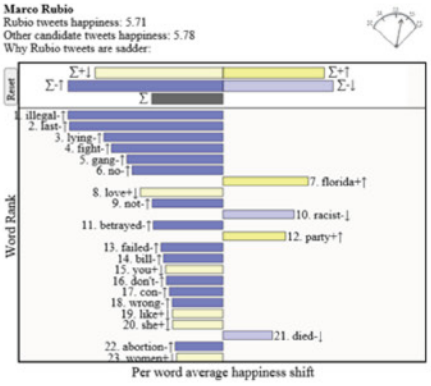


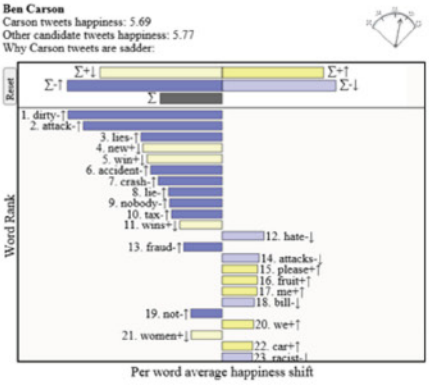
在共和党候选人中，特朗普有最高的幸福水平（5.79），其正面词汇有“伟大”，“爱”，“美国”，“更好”等，显然，这与他的宣传口号相关：使美国再次伟大。有趣的是，他的负面词汇包括了更多：“憎恨”，“种族主义”，“死亡”，“失败者”，“悲伤”，“禁止”，以及以亵渎的方式，反映了他的反对者的观点。有两个需要注意的地方：我们没有特别分析特朗普和他的竞选团队的推文；而“悲伤”这个词语几乎完全来自于莱克斯康的广播。

最后三个共和党候选人有非常多的负面词汇，这也反映了他们在辩论时所关注的焦点和他们团队的关注中心。









我们要再次强调这次初步分析没有调查推文是来自每位候选人的支持者还是反对者。为此，我们现在的工作内容就在于解决这个问题。

社交媒体正在使一般民众有机会去表达他们的政治观点，而这有可能会显著地影响大选的结果。从我们目前的结果来看，如果正面指数是唯一的预测指数（当然有可能并不是），那么我们可以期待一场桑德斯VS特朗普的美国大选。

/ 航空大数据: 为你的旅途带来无限便利 /

编辑 / 协会市场处 欧阳琦　插图 / 崔峻珩　日期 / 2016-05

作为目前世界上最快速同时也是最安全的民用交通工具，飞机成为了越来越多的人们特别是商旅人士出行时的第一选择。但是，飞机晚点、行李遗失等情况真的让人很不爽，大数据能让航班服务变得更优质，我们来看看东方航空是如何做的。

当航空公司提供的服务变得日益及时、精准、周到时，您是否能够想象得到，这背后其实也有着大数据的功劳呢?

飞机晚点行李遗失真的无法避免?

作为目前世界上最快速同时也是最安全的民用交通工具，飞机成为了越来越多的人们特别是商旅人士出行时的第一选择。那么在乘坐航班出行的时候，大家有没有遇到过飞机晚点、行李遗失等情况呢?

看到这个问题，想必不少读者朋友都会感同身受，包括笔者自己也曾经有过这样的经历。然而非常让人气馁的是，每当出现这些问题的时候，大部分人除了在朋友圈抱怨吐槽之外，似乎并没有更加有效的方法。

那么问题就来了：飞机晚点、行李遗失等现象产生的原因究竟是什么?难道这些问题真的无法根治么?

数据分析处理滞后导致巨额损失

在对东方航空高级项目经理刘建国的访谈中，笔者对航空行业的诸多痛点有了初步的了解与认识。

一趟出行从开始到结束，这个过程中航空公司与旅客的接触，会涉及到旅客的行前、行中和行后各个方面。在这些分散的信息数据中，对于航空公司来说最关键的就是离港定位，因为这直接关系到飞机晚点、旅客行程等问题。

然而在中国，离港数据并不是航空公司自己拥有，而是要通过中国航信才能拿到，并且获取这种数据还是分批次的。因此在数据的获取和分析方面，往往都会产生很多的延迟。举例来说，许多旅客往往是在到达机场开始排队之后，才接到航班延迟的信息通知，导致旅客的时间被大量浪费。而当遇到飞机晚点、天气原因、航空管制或者其他问题时，旅客又经常会因为不知道实际情况和最终原因，而将责任全部推给航空公司一方，并将其作为发泄怒气的对象。这也导致许多无辜的一线地服人员承受了本不该有的委屈和巨大的压力。

这种数据分析处理的滞后，究竟会带来多大的损失?刘建国透露，过去因为航空公司系统与机场系统是两个独立的系统，数据处理、分析的不够及时，导致旅客的行程延误、行李丢失等问题，给航空公司带来了巨额的经济损失。根据来自国外的一项数据统计报告，仅在2015年，因为行李丢失问题而给全球航空公司造成的损失就高达15亿美金。

Streams分析技术让企业更懂顾客

面对数据分析处理滞后的种种弊端，航空公司又应该如何升级转型呢?

东方航空高级项目经理刘建国

刘建国介绍说，利用IBM InfoSphere Streams分析技术，东航即将实现航班离港数据、订座信息的实时跟进、24小时内航班状态的分析、实时营销决策、航空安全预警等功能。

举例来说，如今每当东方航空获取到航班状态、航空机员机务情况、油空问题、机场问题、天气原因等数据时，会通过IBM InfoSphere Streams分析技术进行及时处理分析，并做出相关的预判，譬如这些情况会导致哪些旅客受到影响，然后及时反馈给对应的旅客，让他们能够及时知晓航班信息并合理安排自己的行程。

除此以外，在航空安全保障方面，通过IBM InfoSphere Streams分析技术，东方航空也大大降低了信息延迟，有效地避免了旅客行李丢失的问题。

另外在与IBM开展合作之后，东方航空还进一步提升了对旅客的精准服务。通过将外部数据和内部数据有机地结合，东方航空可以获悉旅客的关注点或喜好，又或是其最近身体健康状况、工作职位的变化等等，从而在实际提供服务的时候做出相应的调整，让旅客感觉更加贴心舒适。

譬如旅客还在换票的时候，航空公司就能知道旅客是一位VIP，是哪一个级别的VIP，以及有哪些喜好等等;包括在航班出现延误时，航空公司也可以针对旅客的特点，为其给出相应的建议，这对于旅客来说都是非常新颖的个性化体验。

综上所述，如果您在自己的行程中体会到了来自航空公司未卜先知式的周到服务，不需要感到太过惊讶。航空公司其实并没有掌握看穿人心的魔法，只不过IBM的数据分析技术让他们获得了同样的能力。

将数据商业价值转换成服务价值

看到这里，也许有读者会问：为什么IBM的数据分析技术会有如此神奇的效用？

IBM数据与分析事业部大中华区数据管理及大数据产品线销售总监甘佳凌表示，大数据分析技术的价值，就在于可以将海量的、多样性的、结构化以及非结构化的数据进行很快的分析处理，并迅速给出相应的结果反馈。这也成为了大数据时代企业能否成功的关键因素。而海量的实时数据，也让东方航空看到了大数据时代下的商机，并通过与IBM的携手合作全面提升了旅客的飞行感受与体验。

甘佳凌透露，针对数据信息处理滞后的问题，IBM的流计算技术可以做到毫秒甚至是微秒级别，只要一接收到信息数据就可以马上进行分析处理，并在第一时间将反馈结果通知到客户。譬如针对行李的信息，旅客的行李目前是什么状态、具体在哪个位置等等，都可以实时查询并进行相应处理。而流计算在大数据的应用，就在于可以通过数据的实时分析为客户提供差异化的服务。这就将数据的商业价值转换成了对客户服务的价值。

洞经济加速产业转型与升级

伴随着经济与科技的高速发展，IBM于2016年初发布了认知商业时代的战略思想。而认知商业时代的核心，就是通过大数据分

析及认知计算技术的应用，让企业可以更好地拥抱洞察经济，领先竞争对手获得更多的商机，并加速产业的转型与升级。

在海量的数据中，其实隐藏着巨大的经济价值，这一点已经成为业界的共识。因此更好地理解数据、应用数据，将帮助企业释放更多的潜能，甚至改变现有的商业游戏规则。而通过更加深入地了解客户需求及内部运作状况，企业也可以更有自信地采取下一步行动。

最终，企业可以直接或间接地从产品及服务中获得洞察，并将其转化为实际的商业价值。

针对东航与IBM的此次合作，刘建国表示："东航致力于打造世界一流的现代航空服务，为每一位旅客提供绝佳的飞行体验。通过IBM的帮助，我们可以实时洞察航班状态、旅客服务需求及航班全行程情况，最大程度地避免了突发事件的发生，使我们可以更加专注于服务本身。"

由此可见，借助数据分析技术，企业可以将更多的时间、人力、资源投入到服务客户本身，从而不断改进服务品质，提升用户体验。而东航与IBM的成功合作案例，也为其他企业提供了值得参考与借鉴的典范。

/ 零售行业的数据挖掘七步走 /

编辑 / 协会市场处 杜艳丽 插图 / 协会会员处 冯伟 日期 / 2016-05

对于沃尔玛、华润万家、百佳等零售大超市而言，每天都有很多客户通过会员卡进行购买，不断积累了很多销售数据，如何利用这些数据，从数据中挖掘金矿，很值得每个商家去思考。尽管目前零售商有不少的IT系统去支撑企业常规的分析（如销售量、销售额、热销SKU等），但实际上还是未能从数据角度深入挖掘客户的价值，仅仅从经营分析的角度来满足了常规分析工作。

本文从个人的角度去谈一下如何使用数据挖掘帮助零售商提升生意，让数据真正地去指导企业经营，最大限度地发挥数据提供商业决策的作用。

第一、开展会员制能够帮助企业采集更多会员数据，更有利于开展数据挖掘的工作，同时也有利于培养客户忠诚度。

在实施会员制的时候，必须要特别注意两个关键信息的采集：会员卡ID、客户联系号码或者邮箱，因为这两个关键信息对信息采集及后期的精准营销有很大的帮助作用。而微信、微博等社交媒体的横行，若零售商能够通过相关活动让客户关注企业的微信、微博，对培

养客户忠诚度也是有很大的帮助。

会员制有助于为企业培养众多忠实的顾客，建立起一个长期稳定的市场，提高企业的竞争力。通过会员制，可以有效稳定老客户，同时开发新顾客。

因为零售商给会员提供的是优惠的价格，对新顾客吸引力很大，同时大部分会员卡是可以外借的，也给新客户提供了机会，大大增加其成为会员的可能性。

会员制营销能够促进企业与顾客双向交流。顾客成为会员后，通常能定期收到商家有关新商品的信息并了解商品信息和商家动态，有针对性地选购商品。除此之外，企业能够及时了解消费者需求的变化，以及他们对产品、服务等方面的意见，为改进企业的营销模式提供了依据。

第二、开展零售商的数据挖掘项目，必须要重点提供以下几个表的关键信息：

1.销售表：卡号、销售店ID、销售日期、产品名称、产品价格、销售数量、销售金额、折扣等信息。

2.产品表：产品ID、产品名称、建议零售价、实际销售价、一级类别、二级类别、三级类别、四级类别、品牌等信息。

3.客户表：卡号、发卡店ID、城市、号码、邮箱、企业或个人标识、企业名称、所在行业、地址等。

4.零售店表：店ID、店名、所属城市、店等级等。

其中销售表、产品表、客户表比较重要，而产品表梳理对数据分析及数据挖掘团队而言，是做好项目的关键，必须要耗费大量的时间。

第三、与零售商明确数据挖掘目的，能够让分析团队与零售商之间获得更大的信任，同时有利于项目的顺利开展。

成熟的分析团队，比较关注零售商的商业出发点，从客户商业价值出发，抓住客户关注点，一点一点地做好相应的落地分析工作。

客户最常见想让数据帮助其解答的几大问题：

如何让活跃的客户购买更多的产品，最大程度地释放其价值?

如何唤醒沉默客户，让其转化为活跃客户?

哪些客户是我的重点客户群? 其有什么样的特征?

哪些重点客户流失了? 为什么流失? 后期怎样开展挽留手段?

……

第四、通过数据开展客户细分，明确各个群体的特征。

对于零售数据而言，必须要深入零售行业两大客户群：企业及个人。企业客户的特征和个人客户的特征有很大的区别。

企业特征主要表现：采购量比较大，经常进行团购或批发，销售量和销售额都比较大，为零售商的重点客户群。尽管数量不多，但是却贡献了零售商的60%以上的销售额。

而企业的行为经常有：超大型采购、中型采购、一般采购。对企业数据挖掘，需要深入了解企业的所属行业、采购额度、采购规律、采购产品偏好、是否流失、流失的原因调查等信息，有助于帮助零售商开展相应的营销策略。

对于个人，则需要关注哪些是活跃客户、哪些是新增客户、哪些是沉默客户、客户价值是怎样的、哪些节日是重点高峰期、偏好的产品是哪些等等，这些有助于零售商开展销售、备货等工作。

第五、结合5W1H分析法开展零售分析与挖掘。

What：销售情况怎么样? 有多少用户? 来了多少次? 每次消费多少钱? 买了什么东西…….

Where：哪些门店销售最好? 为什么呢? （交通、地区等）…….

When：哪个月份销售得最好? 哪个节日是销售高峰期…….

Who：是哪些客户? 有什么样的特征? 偏好买哪些产品? 产品规格是怎么样的…….

Why：为什么买哪些产品? 为什么买那么多?会不会继续购买…….

How：怎样提高客户重购? 怎样唤

醒客户? 怎么进行交叉销售? 怎样帮助铺货……

第六、协助零售商开展营销活动设计、营销活动执行、营销评估与优化。

因为数据挖掘是一个闭环的流程，不是撰写挖掘报告、输出营销客户名单就是项目成功的，必须协助零售商开展相应的营销设计、营销活动执行、营销评估及优化工作。从而确保数据挖掘有效落地，为客户真实产生商业价值，扩大生意规模。

营销活动设计常有：优惠打折、派发试用装、赠送礼品、多倍积分等，可以通过不同的细分客户群有针对性地开展不同的营销活动，并计算不同群体及不同活动的投入产出比，便于后期不断优化数据挖掘规则。

第七、关键成果固化IT系统，实现数据挖掘成果固化落地。

对于零售商而言，数据挖掘是个不大不小的投入，对于关键的成果输出，总希望能够把成果规则进行IT固化，实现自动代替手工操作，这个时候经常需要搭建一个成果固化模块或系统，让数据挖掘能够最大限度帮助企业。

/ 湖南翰林数据分析师事务所 /

文 / 湖南翰林数据分析师事务所　编辑 / 协会会员处 冯伟　插图 / 崔峻珩　日期 / 2016-06

湖南翰林数据分析师事务所是一家专业从事项目投资融资咨询、各类数据分析及管理咨询的专业咨询机构。事务所2013年、2014年、2015年连续三年荣获中国数据分析行业全国优秀事务所。

事务所由具有数据分析师、资产评估师、房地产估价师、土地估价师、注册税务师、会计师、司法鉴定人等多种执业资格的复合型专业人员申请发起，系中国数据分析行业事务所会员单位，接受中国数据分析行业监管机构——中国商业联合会数据分析专业委员会的监管。事务所位于湖南省湘中城市邵阳，是湖南省较早成立规模较大附和资质较齐的数据分析师事务所。

事务所可为企业和项目人提供各类数据分析服务、编制数据分析报告、编制可行性研究报告及商业计划书等，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及相关咨询报告。业务范围涉及多个省市地区，涉及行业包括房地产、政府融资平台、行政事业单位、教育、农产品加工、煤炭资源、社会福利等行业和部门，绝大部分数据分析报告已经被有关方面采信并使项目获得成功。

湖南翰林数据分析师事务所坚持“三年打基础、三年上台阶、三年大发展”的发展规划，努力探索适合于本所实际情况的发展路径，结合实际情况努力开拓数据分析业务的着陆点，使数据分析业务首先能保证“接地气”，提高项目成功率和回报率，建立适合于经济新常态下的数据分析盈利模式。在此基础上整合资源，加强制度化管理，构建多渠道数据服务平台，以专业服务赢得社会信任和市场认可。

办公地址：湖南省邵阳市大祥区西湖路南端市人民银行对面万基银座小区1单元402室。
联 系 人：卿启伟
联系电话：13187299268　　办公电话：0739-5189006

/ 云南鼎臻数据分析师事务所 /

文 / 云南鼎臻数据分析师事务所　编辑 / 协会会员处 冯伟　插图 / 崔峻珩　日期 / 2016-07

云南鼎臻数据分析师事务所（以下简称“鼎臻事务所”）是云南省成立较早的数据分析师事务所。自2013年成立以来，鼎臻秉承“前瞻精准超越”的理念，不断积累、创新，自2013年到2015年已连续三年获得云南省优秀事务所荣誉称号，成为行业公认大数据研究领军机构，并获得数据分析师云南省唯一人才培训基地的资格。

鼎臻事务所是一家专业从事大数据咨询分析、大数据环境构建与实施的服务机构，专注于大数据智能分析平台研发与实施，大数据环境下模型算法的创新研究与设计。事务所拥有专业从事大数据挖掘分析、模型算法研究、平台软件研发等复合型人才。积累了多年如零售、医疗、房地产、等行业经验，突破行业的传统分析理念，从决策性数理研究入手，为企业和政府提供了各类“咨询+技术”服务。

作为云南省数据分析行业的领跑者，鼎臻事务所以专业的服务精神、先进的大数据分析体系及优质的服务，成为大数据时代下的量化研究服务专家，也肩负起云南大数据人才的培养和大数据人才库的建设，为促进数据分析行业在云南地区的快速健康发展作出努力！

借助云南被纳入国家桥头堡战略规划建设蓝图的契机，云南也是国家与东南亚国家经济交流的窗口，云南产业升级及改扩建项目将会成倍增长，特别省政府投巨资面向一带一路东南亚节点建设的玉溪大数据中心为公司市场开拓带来巨大商机。

鼎臻事务所具有强大的业务推广能力，借助自身团队、外部力量突出专业地位与优势，快速占领并赢得市场，战略定位是立足云南及周边省市，辐射全国，走向世界……以最低的成本为客户创造最大的价值。

地址：昆明市环城南路331号春天映象大厦B1507
手机：18987888667　电话：0871-65118667　传真：0871-65118667　邮编：650101
电子邮箱：gga@cpda.co　公司网址：http://www.cpda.co

安德信数据分析师事务所

珠海横琴安德信数据分析师事务所有限公司是珠海第一家由数据分析师发起设立，具有独立法人资格的专业从事数据分析咨询服务的机构。

公司拥有一批税务、财会、评估、法律、咨询、管理等方面的业内资深人士，以及高级经济师、中国注册税务师、中国注册会计师、中国注册资产评估师等专业人员。公司的业务涵盖数据采集、数据处理、数据分析、数据咨询、战略分析、数据研究等多个领域，为企业仍至政府提供数据支持。

公司以人为本，尊重知识、尊重人才的现代企业管理思想，旨在建立一个高起点、高效率、高质量的中介服务机构；秉承独立、客观、公正的执业理念，恪守执业准则，坚持质量第一、诚信为本的服务宗旨，以专业的技术水准、亲和的沟通方式、诚信求实的工作态度，为客户提供优质的服务。

公司办公地址：珠海银桦路8号2401房　联系人：张进宏
联系电话：13192252898　邮箱：adxzjh@126.com