中国数据分析

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
2015年第04期 总第24期
咨询热线：010-59000991
www.chinacpda.org

关注行业微博微信 了解数据分析行业前沿知识

《中国数据分析》会员特刊
2015年第04期 总第24期

主办
中国商业联合会数据分析专业委员会

主编
冯伟

编委
崔旭　崔茜茹　周子赫

出版时间
2015年12月

美工
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 100020
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

目 录 CONTENTS

2015——大数据破局之年

匆匆一年，2015年即将成为往事，大数据时代的到来仿佛加快了时间的速度，数据分析行业在不知不觉间在中国已经经历了十二年的发展。回顾过往，感想颇多。如果需要找一个词来形容2015年的大数据行业，那么“变革”是我对大数据在2015年最主要的印象。

在2015年，国务院印发了《促进大数据发展行动纲要》，政府数据资源将逐步公开，一批面向全球的大数据龙头企业和中小型企业将获得重点培育。与此同时工信部正在制定《大数据产业“十三五”发展规划》。结合物联网和车联网的开发应用，国家对大数据的关注可谓空前绝后，这些变革都说明一个核心信息：中国大数据产业已经升级为国家战略， 2015年是中国大数据发展的破局之年。

多年来，协会伴随着中国数据分析行业的成长，经历了从无人问津到成为关注热点的过程，在大家都在炒作大数据的时候，协会一直在关注大数据的应用与落地。2015年协会举办了行业峰会，论坛，沙龙等不同规模的近百场活动，为数据分析行业的普及与推广进行大量的宣传。随着智能大数据分析平台的完善，基于平台使用推出的百家企业扶持计划等大型公益活动，旨在为中国的中小型企业提供优质的数据化服务，打破大数据的应用瓶颈，让更多的企业有机会与大数据进行快速对接。同时，事务所作为数据分析服务专业机构，近年来一直以量化咨询作为主体业务，“技术+咨询”的经营模式得到了大家的一致认可，平台的引入无疑使事务所的整体实力得到显著提升，弥补了业务上的短板。无论是协会自身还是从业机构，破局已经完成，行业升级的变革即将开始。

作为中国数据分析行业协会，我们对2015年所发生的转变欣喜不已的同时，也发现了一些问题，例如：部分事务所在巨大的发展机遇面前准备不足，虚假量化投资类业务屡禁不止，假协会、假培训机构炒作大数据概念、误导企业及分析师等等。为了更好响应国家、政府在大数据领域的政策导向、做好十三五行业发展规划，协会在战略布局、行业监督和工作风格等方面都要顺应大数据时代发展的要求，不遗余力的促进行业健康、稳定、快速的发展。

长风破浪会有时，直挂云帆济沧海。相信2016年的大数据征程不会一帆风顺，但我们坚持科学发展道路，相信在业界同仁们齐心合力下，必将能够到达理想的彼岸。值此岁序更迭之际，我代表协会全体工作人员祝愿业界同仁、合作伙伴以及所有数据分析师们新年快乐！感谢大家为数据分析行业做出的贡献，期盼能够在新的一年里与大家共赴征程，共创未来！

中国商业联合会数据分析专业委员会

大数据校园行(面试文化节活动)

文 / 中商联数据分析专业委员会 市场处　崔旭　　图 / 崔峻珩

由中国商业联合会数据分析专业委员会（以下简称“协会”）推出的主题为“学习数据思维、成就职场精英”之大数据校园行活动，已于2015年10月正式启动！大数据的概念已不再陌生，但不得不承认的是，中国的大数据仍处于发展阶段，许多高校尚未开设大数据相关专业，数据知识的普及也相对缺少。随着国务院《关于促进大数据发展的行动纲要》中“培养专业大数据人才”的提出，全民数据思维的建立将成为大数据时代的发展趋势，因此，本次活动的启动从帮助学生在大数据时代掌握更具有竞争力的数据思维及技能开始。

据悉，本次活动最先从北京六大高校启动，协会联合北京师范大学、北京理工大学、北京农业大学、北京航空航天大学、北京外国语大学、中央财经大学共同举办大数据沙龙、校园公开课、面试节等形式多样的活动，并与花旗银行、雀巢、渣打银行、施耐德等世界著名500强企业共同为学生提供名企就业、学习交流的宝贵机会。同期，贵州中医药大学也将联合协会开设医疗大数据讲堂，通过对医疗案例的技术分析、平台共享和资讯解读，为医学院的学生解析大数据在医疗当中起到的作用，为学生日后接触常规医疗数据工作夯实基础。

目前中国数据分析人才教育体系仍相对缺失，而随着大数据近年来的蓬勃发展，无论是企业、学校还是社会都对数据人才培养保持高度关注。因此，将数据分析知识融入高校教学体系是当下最紧急的事情。据中国商业联合会数据分析专业委员会人才统计报告，未来五年内，大数据人才缺口将会达到1300万人之多。为了弥补人才的巨大缺口，从学生中培养数据人才，“大数据校园行”势在必行！协会随后将陆续在上海、深圳、厦门、西安等地各大高校开设大数据公开课、公益沙龙、交流会等活动，将大数据理念带入全国校园，培养高校学生的数据思维，使其在大数据的浪潮中激流勇进，成为国家急需的数据人才。

旨在大数据时代下，为学生全面提供大数据知识，并带来良好的工作机遇。整个面试节包括大数据知识普及与面试大赛两个方面，协会在六所高校举办大数据讲堂，并与高校联合创办面试大赛，提供赢得500强Offer的宝贵机会！本次校园行活动汇集大数据时代最热门的讯息，从解析大数据到大数据时代下的职业发展无所不包。现场的数据分析专家解答大家对于大数据的一切问题！

1、会议开篇，期待选手的表现

2015年11月7日晚7：00，面试大赛决赛在北京师范大学

敬文讲堂举行。会场座无虚席、秩序井然，一段点燃全场的开场舞后，比赛正式开始。参每位选手在展示个人VCR后，以“大数据”为切入点简单阐述，这是既六强选手走访协会后，首次以自己独特的方式，来诠释他们对大数据的理解和展示，下面就让我们一起来重温一下现场的精彩。

2、选手演绎“大数据”

来自中国农业大学金融系的罗书弈，中国农业大学农林经济管理专业的湛梓瑛，北师大人力资源管理系的女生王梦雨，北师大英语系选手周敏，北师大英语专业的学生冯倩文，北京航空航天大学能源与动力工程学院的工科男周弘毅均对大数据的课题进行了精彩阐述。

3、企业HR激烈抢人

现场京东，联想，甲骨文，宝马的HR看到优秀的选手，简直是抢人“无底限”啊——给出了各种优越条件，这说明优秀的大学生无论何时都会受到企业的欢迎。

选手们的展示结束后，北京中外企业人力资源协会秘书长牛先生进行简短点评，“选手们的优秀从何而来？”牛先生谈道，“必然是长期的学习与实习经历。”经过评委们的综合打分，三号选手王梦雨被评为“最有价值雇员”，成为本届“面霸”得主。

国务院《关于促进大数据发展的行动纲要》中明确提出要“培养专业数据人才”，全民数据思维的建立已成为大数据时代发展的趋势。

这次活动的成功举办，对大数据知识在高校的推展有很大的促进作用。中国商业联合会数据分析专业委员会作为数据行业的全国性协会，与北京各大高校合办的这场大数据校园行活动，全面的揭开大数据的奥秘，让学生们迈入大数据时代的门槛。FIN

河南省大数据沙龙成功举办

文 / 中商联数据分析专业委员会 市场处 崔茜茹 图 / 崔峻珩

2015年河南大数据沙龙成功召开，11月7日下午，由中国商业联合会数据分析专业委员会主办、项目数据分析师（CPDA）河南授权中心承办，河南MBA&EMBA同学会，郑州市一帆教育培训学校，河南智宸项目数据分析师事务所，项目数据分析师（CPDA）河南学友会协办的“如何使用高级算法分析客户行为”沙龙活动在郑大宾馆成功举办。河南CPDA学友会和各行各业数据分析从业者、爱好者共40余人参加了此次沙龙活动，本次沙龙特邀王庆生老师给大家做分享，王庆生老师是国内知名数据分析与数据挖掘领域专家。有10年数据挖掘及数据分析项目经验，擅长使用各种算法模型，有丰富的数据建模经验。精通金融，工业生产，零售行业数据分析与挖掘。从事金融数据挖掘，家电行业数据分析与零售管理数据分析等工作，有丰富的项目经验，给大家带来全新的视角与数据盛宴。

下午13:30沙龙正式开始，首先由中国商业联合会数据分析专业委员会肖坤学主任给大家介绍了数据分析行业的发展历程及现状，数据分析行业作为新兴行业，其发展前景是十分广阔的，肖老师表示，希望越来越多的人加入数据分析行业中来，越来越多的人成为专业的数据分析人才。同时欢迎大家登

陆协会官网，多多提出宝贵意见，为共同发展数据分析行业贡献一份力量。

稍后，王庆生老师就本次沙龙主题开展了生动地演讲。首先，王老师立足行业，介绍了在大数据时代背景下，企业应该如何建立体系化的数据分析环境，以应对市场变化；然后，王老师针对企业产生的客户数据，展示了系统化数据分析技术和算法，并讲解了算法和技术的应用构造，以贝叶斯算法为例，详细阐述了贝叶斯算法在分析客户行为上的应用原理和应用环境；最后，王老师介绍了贝叶斯算法的应用案例，以汽车

公司销售数据为例，结合多种算法，最终形成详细的客户细分模型，并利用这些模型，提出针对性的营销方案，整合资源，做到精准营销。王老师提示，数据分析并不是仅仅局限在企业内部的数据，同时还需要来自外部的行业数据，从整体把控，数据分析才更精准，更有意义。大家目不转睛，认真仔细的听王老师的讲解，同时做笔记。

王庆生老师演讲结束后，沙龙进入白热化阶段，大家争先恐后的向王老师提出自己在数据分析方面的疑问，自己在处理数据工作中遇到的问题以及对算法模型的疑问等等，王老师耐心地一一作答。

茶歇期间，现场进行了抽奖活动，抽出了本次活动的一二三等奖，极大的活跃了沙龙现场的气氛。同时感谢河南智宸项目数据分析师事务所提供的奖品支持。

接着，有河南CPDA授权中心主持，联合郑州市一帆教育，河南智宸项目数据分析师事务所，海马轿车有限公司，太平洋保险四家企业进行现场招聘，四位公司代表逐一阐述了本公司的发展理念，以及数据分析人才稀缺情况，希望广大数据分析人才能加入到自己团队中去，用数据分析技术为企业发展贡献力量。

最后沙龙活动在大家热烈的掌声中圆满结束，许多参与活动的数据分析爱好者对王老师的精彩分享意犹未尽，收获颇多，极大地提升了对数据分析技术的应用兴趣，他们说：“大数据是未来的发展趋势，感谢CPDA河南授权中心给大家带来这样好的机会接触大数据，期望能有机会参与下次活动”。通过王老师的分享，参会人员对大数据的应用有了进一步的了解，认识到数据分析技术对企业发展的重要性，同时表示有兴趣加入数据分析协会，成为数据分析协会大家庭中的一员。

FIN

同道精英汇之CPDA数据分析师职业专场

文 / 猎聘网　　图 / 猎聘网　　编辑 / 赵秋红

2015年11月14日，由中国商业联合会数据分析专业委员会与猎聘网携手共办的同道精英汇之CPDA数据分析师职业专场活动，在猎聘网二层同道小馆胜利举行，本次会议邀请了保险，金融，互联网，广告业等多家知名企业参加。现场干货满满，更有精彩小礼品。

这是我会首次联合知名企业以及第三方招聘平台为数据分析师搭建的职场交流活动。随着企业对大数据价值的认知，数据分析人才在企业中的重要性得以突显，2015年9月，政府也将专业数据人才培养列入大数据发展行动纲要，这正是大数据人才井喷时代即将来临的信号！

精彩瞬间：你不该错过的那些职场机遇

会议中，猎聘网大数据部商业中心数据分析负责人石晶老师发表了《分析师人才竞争力报告》。他用猎聘网自身作为案例引导，通过三点来讲述了现今大数据行业的发展前景，并解析了分析师行业求职者最关心的

猎聘网与职业新机遇

问题——如何在泥沙俱下的“大数据”风口中抓住机会的同时保持清醒头脑。可以说本次活动中猎聘网公告的这份报告，帮助求职者及时了解到职业需求形势、收入水平的变化趋势，并为求职者的职业规划提供了准确指引。

本次会议还邀请了中国商业联合会数据分析委员会市场处张楠主任进行了分享，张老师以“数据分析师你的职业竞争力在哪里”，张主任从大数据的发展历程讲起，在不断发展进步的大数据时代，我们要为什么要提高自身的职场竞争力，我们要怎么提高自身的职场竞争力。具有良好的竞争力以后，我们要通过什么途径来实现自身价值，以及个人的职场定位。

会议最后，是由品友互动资深数据分析师韩静波为大家分享“数据工程师的职业进阶之路” 韩老师主要从两方面聊起首先是数据分析师怎么挑选一个比较靠谱的企业，或者说挑选一个比较靠谱的行业。其次是聊企业是怎么来挑选一个靠谱的数据分析师的。

名企开招，热闹纷呈

今天活动爆场，来晚的只能站着了，猎聘网，品友互动，钰诚集团，阳光保险，数据分析协会，五大靠谱企业，现场人才开抢，大数据人才井喷的时代真的来临了！

现场之声 | 数据分析师，你的竞争力在哪里

演讲人：张楠

引言：在传媒、快消品行业有近十年品牌运营经验，转行到大数据领域后，见识到数据分析在市场营销中的精准应用，深知在大数据时代，企业要想发展，需要转变经营理念、转化经营模式，而要做到这一点，必须学会用数据说话，才能做理性决策！

今天很高兴和猎聘网共同举办这次活动，很多老朋友都知道，我们往期的沙龙更多都是关注技术应用、案例分享，那像这种联合招聘方、人才培训方、企业方三方力量为分析师搭建职场平台的活动还是第一次，那为什么会有这样一次活动呢，是因为有一次在猎聘网上去搜企业对数据分析人才的招聘需求时，然后就看到有一些企业在招聘启示里明确标出“持有CPDA数据分析师证书者优先”的字样，然后我就想到，我们协会培养了这么多的分析师，那企业又在找我们培养的人员，那为何不能直接来一个线下的交流会呢，于是，我把想法和猎聘网一说，一拍即合，就有了这样的一场沙龙。

今天有四家企业来到现场，第一个猎聘网，刚刚大家看到了，随后还有品友互动、钰诚集团、阳光保险集团，这四家企业也都为大家带来了数据分析师的岗位，随后四家企业都有展示的机会，最后我们会有一个互动环节，希望分析师和企业牵手成功！

现在我想说，企业有需求，企业也准备好了，作为分析师的你们，准备好了吗？在这样一个迭代更新快速的时代，你们的职场竞争力在哪里？

我会从三个层次与大家交互，第一个就是理解我们现在所处的背景，都说时势造英雄，所以对时代发展阶段的理解可以帮助你把握前进的方向；第二个是对于数据分析师来说，你的职场竞争力是什么，又体现在哪些方面；最后呢，会告诉大家你要怎样拥有这种竞争力。

我们先来看我们处的时代背景，大数据概念在21世纪初就传入了中国，但那时候没有多少人知道，到2010年前后，大数据概念开始被社会各界认知，这个阶段是大数据的认知阶段。到2013年的时候呢，大数据的概念一下子就火了，所以有媒体把这一年称为“大数据元年”，也就是从这个时候起，政府、企业开始主动认知到大数据的价值，在于通过数据分析帮助企业进行精准决策，从而产生商业价值。所以说这个阶段也是进入数据分析师行列的最佳时机，协会作为人才培养机构，我们自己是能切身体会到这点的，我们现在每天接到的咨询电话比去年同期高出35%。再往后，就会进入到大数据2.0应用阶段，我相信，更多的相关项目将陆续落地，大数据也将真正进入应用阶段，数据分析师也将在这个阶段大展宏图。

我想跟大家说了这么多关于大数据发展阶段性的东西，是希望大家能够明晰自己的未来发展方向，但光有信念是不够的，还要有落地的方式。12年来，我们见证了一些优秀分析师的成长，看到有很多优秀分析师在像IBM、京东这样的大公司里担任要职，同时我们也看到有些分析师没有坚持自己的道路最后放弃了。但是终归我们看到优秀分析师为什么能够成长，我在与他们交流后得出一条经验，就是：别人做不到的你可以做，别人做得到的你能够做的更好。这个会取决于三个方面：思维意识、学习能力和实践应用。

有句话说意识决定行为，那在大数据时代，我也希望大家建立一种“用数据看世界”的思维方式，《大数据时代》的作者维克托是这样描述的：每天早上起来想一下，这么多数据我能用来干什么，这些价值在哪里可以找到，能不能找到一个

别人以前都没有做过的事情。你的想法和思路，是最重要的资产。这种思维方式的建立，会帮助你比其他人有更多对数据的敏锐观察和思考，会帮助你服务的企业挖掘到更多的数据价值。那你也会成为企业中不可或缺的人才。所以说要成为一名优秀的数据分析师，意识要先行。

你有了一个好的意识后，还要去学习，因为目前咱们国内的教育体系中还没数据分析专业，数学也好、统计也好、计算机也好，都只是这个学科的一部分，大家从图中可以看到数据分析师的知识体系是很全面的，而且这个时代也是一个“快鱼吃慢鱼”的时代，大数据的技术、知识迭代更新的非常快，协会从03年第一批数据分析人才培养开始起，到现在分析师课程都已经改版五次了，目前已经在着手第六次改版了，这样做就是为了让分析师的知识体系更适应时代和企业的需求。

所以说，即使你参加了8天面授和1年远程，考取了一个证书，那也不代表可以一劳永逸，这只是说你拥有了一个行业的敲门砖，拥有了一套标准的行业知识框架，在这样一种快速更新的时代，不要不断地补充新知识，不断地在实践中磨砺成长。同时把你学到的东西学以致用。

经常会有学员问我这样的问题：我学完了，但是我们企业到现在为止好像和大数据关系还很远。还有的会问，我现在学了这么多，可是实际工作中只用到一小部分，没有更多发挥空间怎么办？我其实想跟各位说的，你们是不是真的把你们所学的主动应用于你们现在的岗位，你们是不是有帮助你们企业构建数据化进程，帮助你的企业真正实现你们企业内部数据打通工作。我想大家都知道，现在企业发展模式是数据即业务，业务即数据。

未来几年，企业运营数据实际上就是运营他的业务。你现在做的可能只是数据收集和清洗的工作，或者一些简单的数据分析，但我希望大家看到企业未来发展方向，所以你们需要学习这种全面知识的体系，珍惜自己每一个实践的机会。同时，协会也定期都会举办公益沙龙，请一些企业人员、政府科研人员、院校专家等来与大家做交流，明年开始，我们还会参加更多互动环节，拿出一些协会正在做的案例、研究项目让学员做实操，帮助你们提升专业性、扩展你们的业务思路。我们也希望你们在实际工作中遇到的难题也可以拿出来与我们的老师、研发人员进行商讨。终归一句话，你自己要有这种学以致用的意识。

具体到做法，有两个方面：

首先，是你要有清晰的职业定位；

其次，是要有正确的成长计划。

大致上来说，取得我们CPDA证书的数据分析师会有三个走向：

一是在专职岗位从事数据分析工作，大家可以看到我这里实际上是分两个层次的，一类是基础性人员，数据分析员，一类是精深岗位的数据分析师，我们在与一些企业去聊他们需求的时候，会发现企业分不清自己的需要，往往有时明明招的就是个分析员的岗位，却叫了分析师。实际上在员和师之间在数据处理的层级上、经验上、技能匹配上还是有差异的。目前协会也与政府相关机构共同制定数据分析的行业标准。希望这个标准到时会帮助企业还有求职者明晰自己的方向。

二是在企业内部，还有大量的非专职岗位的存在，从上一张图中大家也都能看到数据分析是可以应用于企业的各个环节的。比如我，我本身是从传统行业转型到大数据行业的，但我一直做的是市场，之前不懂大数据，到了协会以后才算真正接触，我看到我们数据中心的人研究算法，讨论模型，看到课程部的人每编写一个教学案例的时候，都会拿出来先给员工进行讲解和分享，会看到有传统咨询公司在做到战略分析的时候找到我们去做数据分析，这一切都让我看到了数据发挥的价值，慢慢地，我自己也开始学习分析师的课程，尤其是学到数据分析在市场营销中的应用时，我觉得特别惊讶，原来一直困惑我们的市场预测可以用回归解决。所以说，只要你拥有了数据思维，你想把你所学的引入你的岗位，不管你是不是专职，你都可以找到数据分析发挥作用的地方。

最后的话呢，获得我们CPDA数据分析师证书的学员，如果想创业，协会是支持成立数据分析师事务所的。我们也会为组建人员提供创业指导、执业指导，包括技术平台的支持。

我相信大家已经在思考你自己的发展方向了，那关于成长计划这块，我之前其实已经都有提到，比如你要学会用“数据思维”看世界，建立一套完整的知识体系，还要随时关注大数据前沿动态和变化，不断地学习，不断地应用，与业内人士更多地交流等等。

我想，随着未来更多的企业认识到大数据的价值，认识到数据分析人才对企业发展的重要性，也就会有越来越多的企业将大数据人才当成自己的标配人才，因为不论是营销还是企业管理岗位，无论是基层，还是中高层，都需要数据分析能力。那时会是大数据人才需求和供给的繁荣阶段，大数据人才也会变为最有价值的人才。 所以说，世界是数据的，机会就在眼前！ FIN

全国项目数据分析师事务所更名正式启动

文 / 中商联数据分析专业委员会 会员处　周子赫　　图 / 崔峻珩

未来的十年将是一个“大数据”引领的智慧科技时代，大数据时代已经悄然来临，为了顺应行业发展，结合事务所及行业专家的建议，经协会研究决定：消除项目数据分析师事务所名称的局限性，帮助事务所能够更好的进行行务开展。

从2015年11月13日起对项目数据分析师事务所名称进行统一调整，不在使用“项目”二字，事务所名称统一为“数据分析师事务所有限（责任）公司”。本次更名希望可以帮助事务所与企业能够更好的进行业务对接，拓宽事务所经营，为事务所未来发展提供更大帮助。FIN

数据分析行业专家招募活动

文 / 中商联数据分析专业委员会 会员处　冯伟　　图 / 崔峻珩

为了更好的实施国家大数据战略，促进数据分析研究水平的提高，推动行业快速发展。经研究决定，协会面向社会公开征集行业专家，完善了协会专家库。

自2003年成立以来，我们伴随着中国大数据行业的发展，见证了从无到有，从崭露头角到方兴未艾的过程，在10余年的历程中，我们培养了上万名数据分析师，创建了过百家数据分析师事务所，与数十家大数据行业知名企业和科研院所达成合作，每年举办上千人参与的行业峰会及上百场公益沙龙，论坛，研讨会等活动，并且拥有专业的研发团队，建立数据库，自主研发datahoop智能分析平台……

我们招募的专家不仅热爱数据分析，长期从事数据分析相关工作，对数据分析行业具有独到的见解和突出的贡献，拥有丰富的数据分析理论知识及实战经验，积极参与协会组织的行业活动，科研背景，发表过学术文章，并获得过职称，考取CPDA职业证书的特点。

我们的专家可获得非凡的机遇与挑战，可以拓展行业人脉：加入协会专家库，参加行业专家座谈会等活动，有更多的机会与业界同仁交流沟通；获得行业荣誉，颁发行业专家荣誉证书，优秀专家成为我们的名誉会员;参与科研项目：协助协会进行课题研究，行业白皮书撰写等项目；行业标准制定，进入协会标准化制定工作组，参与数据分析行业标准的制定;更多商业机会，可优先被数据分析师事务所聘请为企业顾问;汲取培训经验，有机会成为数据分析师培训的特邀导师;提升行业知名度，可优先作为嘉宾出席大型行业峰会，专业研讨会，论坛，沙龙等活动;免费个人推广，专家库成员名录将在中国数据分析行业网进行备案，并可进行查询与推荐;智能平台使用，免费或优惠使用Datahoop智能分析平台;行业素养的综合提升的机会。FIN

第六届全国优秀事务所评选结果揭晓

文 / 中商联数据分析专业委员会 会员处　　图 / 崔峻珩

从2015年10月开始，历时一个月多的资料上报和评审工作，2016年数据分析行业优秀事务所评选活动已经圆满结束。优秀事务所评选从2010年开始至今已连续举办六届，本次活动旨在加强行业内事务所间交流与合作，增强数据分析行业建设，树立良好的行业形象，发挥优秀事务所的带头作用，推广技术经验。同时协会通过本次活动可以掌握各地事务所的市场环境和经营现状，有针对性的进行行业规范和指导，并了解事务所对行业发展的意见与建议。

本次评选除了往年的评选方式外，同时结合了事务所在大数据背景下的业务开发能力进行全面综合评估。对于事务所遇到的问题也进行了详尽的统计和分析。从汇总上来的资料和报告中，我们看到：

事务所已经开始认识到品牌推广的重要性，通过各个渠道不留余力的对自身进行宣传。他们的业务量处于长期稳定的增长，并拥有了固定的客户群体。

目前企业已经开始认识到数据分析技术的引入对制定企业战略是至关重要的。企业的主要目的是为了提升决策速度和准确性，从而提高运营效率，或是优化运营过程，改善现有产品、服务和特点，从中挖掘新的收益来源，以及预测消费者行为等，以上这些正是事务所能够提供的服务与支持。

而随着数据分析行业的持续升温，来自于不同行业的人才加入事务所，而事务所也利用他们的丰富的原行业经验，涉足了新的领域，扩展了企业的生存空间。

但是在进行业务开展时，事务所面临着很多的问题，例如：数据获取来源不足，数据体量偏小，研究技术水平偏低等，这些都阻碍对数据的分析效果。对于上述问题的解决，协会将做为2016年重点工作之一。在此，协会希望我们的事务所响应国家大数据发展战略，在良好的发展趋势下，积极迎接

挑战，不断提高专业水平，与行业共同发展进步。

经过评审专家组的综合评审，有4家优秀事务所从众多参选事务所中脱颖而出，这4家事务所不但保持着稳定的发展，同时在规范化经营等方面进行了严格的把控，并且对行业的发展具有突出贡献。

他们分别是：

中颢润（北京）数据分析师事务所

中颢润拥有包括专业数据分析师、数据架构师、资产评估师、注册会计师和专业市场调查人员构成的强大团队，致力于专业大数据分析深度服务，坚持经营数据分析报告、投资数据分析报告、决策数据分析报告的权威性、客观性、中立性和实用性，为众多企业和政府机构的决策者提供了准确的数据分析成果和决策建议。此外，中颢润为了更好的服务于国内外客户在国内开展业务对数据信息的需求，先后通过了美国、英国、法国、德国等大使馆认证及涉外调查等多项资质。

作为北京地区的数据分析师事务所，在面对激烈的市场竞争时，中颢润（北京）数据分析师事务所领导人提出要以互联网思维经营事务所，采取微信、微博、线上线下结合等多种新的营销方式，获得了优异的成果，为全国事务所的经营提供了宝贵的经验，且中颢润事务所在大数据分析领域已有一定的储备，专业实力较强。中颢润的数据分析报告思路清晰，结构严谨，分析细致，数据资料详实，方法得当，论证深入，整体质量较高，为企业做出经营决策提供科学依据。

上海天元项目数据分析师事务所

上海天元项目数据分析师事务所业务网络遍布海内外与全国30多个省市自治区，已与多家国内外金融机构、大型财团、银行、上市公司、商会等建立了业务合作意向及业务往来，业务涉及农业、林业、建材、石油、石化、化工、通信、电子、电力、电网、水利、铁路、民航、煤炭、冶金、轻工、纺织、医药、机械、市政、交通、房地产、卫生、教育、文化娱乐、环保、旅游等诸多行业领域，在为客户提供服务的过程中，积累了许多宝贵经验，已形成一支执业经验丰富、人员结构合理、高素质的专业队伍，能承担各种类型项目数据分析及相关业务。上海天元撰写的报告内容结构严谨，分析到位。其专业的方法、谨慎的作风、客观的态度、公正的原则以及热情的服务，帮助企业做出切实有效的决策。为推动中国数据分析事业的发展贡献着自己的力量。

湖南翰林数据分析师事务所

湖南翰林数据分析师事务所有限公司具有数据分析师、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是湖南省较早成立规模较大且资质较齐的数据分析师事务所，是一家专业从事各类数据分析及管理咨询的专业机构。

湖南翰林数据分析师事务所一直注重自身行业品牌及事务所形象的推广宣传，事务所拥有多位企业管理方面的复合型项目数据分析师，在提供科学、专业的数据分析服务时有了最基础的保障。事务所的数据分析报告思路清晰，内容丰富详实，专业度较高，数据资料丰富，市场需求部分研究方法得当，为企业科学决策提供了依据。

重庆传晟项目数据分析师事务所

重庆传晟项目数据分析师事务所主要致力于房地产政府管理部门全局解决方案、软件开发、咨询服务和档案整理加工，以及房产数据分析统计处理。公司技术实力雄厚，拥有高素质的数据分析师团队、软件开发队伍和强大的技术支持力量。主要核心人员从事房地产政府管理部门行业软件开发和咨询工作多年，成功推出了拥有自主知识产权的档案管理系统、数据清理系统、搜索引擎、档案生命周期管理平台、档案整理加工系统、房地产管理部门便捷化服务平台、权证防伪、短信平台等系列产品。公司包括成都和重庆两个研发中心，公司已扎根西南，并不断扩展中。

重庆传晟项目数据分析师事务所一直充分发挥着带动当地数据分析行业快速发展的榜样作用，积极拓展业务范围，将企业经营决策数据分析作为事务所发展的长远目标，不断提高自身专业水平和业务能力，获得了当地政府和房地产行业的一致认可。

重庆传晟走的是一条“技术+咨询”的道路，其具备较强的信息化技术能力，结合数据分析师事务所得天独厚的数据分析专业实力，这是其独特的核心竞争力。所内从业的数据分析师均具备多年的项目数据分析、软件技术工作经验，秉承以诚为本、客观公正的职业操守，为重庆地区的房地产政府管理部门提供提供专业、全面、详实、精准的数据分析报告。

重庆传晟撰写的经营类分析报告宏观分析到位，推论严谨，思路较为清晰、公正、客观，可以为需求单位提供最佳决策建议依据。FIN

大数据1.1 时代: 从认知到应用

文 / 中商联数据分析专业委员会 市场处　崔旭　　图 / 崔峻珩

在刚刚过去的十一黄金周里，国内各大景点游人如织，不少景点甚至出现了因为参观游览人数超过其最大承载量而实施限流的现象。相比于国内游，海外游的热度也毫不示弱，中国游客“买爆”日本、韩国等国的新闻也比比皆是。在浏览黄金周游客出行相关报道时，我们不难发现，“大数据”的曝光率非常高，借助大数据分析，我们可以了解黄金周游客出游的距离、高峰、人均费用和消费行为等信息，基于大数据分析而进行的黄金周出游情况分析，已成为人们了解和认知中国黄金周旅游特征的重要手段之一。实际上，大数据的应用远远不止于此，各行各业都能见到它的影子。在其起步阶段，互联网和电商等与IT 技术关联度较大的领域最先了解和使用了大数据，随着大数据逐渐被社会各界认知，越来越多的行业包括大量传统行业在内，都开始了解和应用大数据。大数据对人们的生活究竟产生了怎样的影响？为此，《经济》记者专访了中国商业联合会数据分析专业委员会会长邹东生，与他就大数据大数据1.1 时代: 从认知到应用的相关问题进行了深入探讨。

从1.0 向2.0 时代过渡

邹东生认为，当前中国正处于大数据1.1 时代，是以认知为特征的大数据1.0 时代向以应用为特征的大数据2.0 时代的过渡阶段。“尽管大数据的概念在21 世纪初就传入中国，但是在2010 年前后才开始被中国的社会各界广泛认知。”邹东生向记者解释，“过去的四五年是大数据在中国的普及阶段，政府、企业，特别是一些原本做数据建设和信息技术的公司，开始主动认为大数据是一个趋势。

国家前后出台了很多大数据领域的相关规划和推动大数据发展的政策，一些大数据交易平台和交易所陆续成立，越来越多的大数据相关会议也在中国举办”。总结和展望大数据概念在我国的普及和发展过程，邹东生认为主要可以分为3个阶段，“第一阶段是2010年前后大数据的1.0认知阶段，这是社会各界对大数据概念产生初步认知的时期，政府部门和企业界对大数据的理解千差万别，各不相同；第二阶段就是当前，是大数据1.1阶段，是认知大数据向应用大数据的过渡时期；第三阶段是不远的将来大数据2.0应用阶段，未来几年，大数据的相关项目将陆续落地，产生实用性价值，大数据将进入应用和产生价值的阶段。”

在划分大数据时代的不同阶段和阐释相应特征的基础

上，邹东生向《经济》记者概括了当前我国大数据发展的主要特征，他认为主要可以从大数据的认知、应用和人才培养角度出发，分为3个方面。第一，从认知大数据的角度出发，企业开始探索大数据能为自己带来什么。“最初，人们对大数据的求知欲仅仅停留在‘我希望了解什么是大数据’，而现在，越来越多的企业希望了解大数据能给自己带来什么和大数据项目如何落地。”第二，从应用大数据的角度出发，企业对大数据的关注不仅仅停留在技术层面，更多向分析数据和研究数据发展。“前几年大数据相关会议和活动的主角是技术公司，只要一说大数据，他们就必提一些技术的新名词，从而导致很多人误认为大数据就是一种技术。这其实是大数据1.0时代的一种特殊认知。因为在那一阶段，很多企业还没有构建自己的数据平台，缺乏数据基础，企业开始接触和应用大数据分析的第一步就是数据化，离不开对于技术的接触。但是技术仅仅是大数据的底层，不是大数据的核心。

未来，随着技术的逐渐开源，越来越多先进技术门槛的降低，越来越多的人会明白大数据最具价值的部分不是通过数据搭建的技术平台，不是数据本身的储存，而是它的分析过程，是通过对数据进行的深层次应用，帮助企业提高决策决心，降低运营风险。”第三，从人才培养的角度，企业越来越重视大数据人才的价值。大数据人才即将迎来井喷阶段，“以各企业对数据分析人才的需求为例，前两年相关的人才需求比较零散，而现在的需求十分旺盛，甚至一天之内就可能出现几万个岗位空缺，有些职位也对应着高额的薪水，目前大数据人才仍处于井喷的前期，并未迎来真正的繁荣。未来，越来越多的企业会认识到自己的核心竞争力应该是人才储备，而不是技术储备，越来越多的企业会将数据分析人才当成自己的标配人才，因为不论是营销还是企业管理岗位，都需要数据分析能力。那时才是大数据人才需求和供给的大繁荣阶段，大数据人才才能变为最有价值的人才。”

核心价值

进入21 世纪以来，互联网对人们的生产生活产生了巨大影响，在极大程度上推动了各行各业的创新和进步，互联网自身也在不断的发展中改变和获得突破。时至今日，传统互联网技术在融合了云计算和云存储等新概念的基础上，已进入了大数据时代。对于传统互联网和大数据间的差异，邹东生认为：“传统的互联网是一种服务于各行各业的工具，致力于为人们带来直接的便捷，企业在前期需要进行充分的互联网相关技术搭建，但这一过程并不一定能带来快速和直接的盈利。所以对于互联网企业而言，‘走得早，不一定走得好’。而大数据则不同，在大数据领域，越早使用大数据和相关技术、理念分析处理问题，就能越早获益。”为什么大数据企业“走得早”，就能“走得好”？邹东生告诉记者，这是由大数据的核心价值决定的。“大数据能够帮助企业进行精准分析，从而提高企业的决策效率，帮助企业获取更多的收益，降低成本。换言之，大数据能够给企业带来看得见、摸得着的收益。”除了帮助企业精准决策，在解决“数据孤岛”问题上，大数据也将发挥巨大作用。他认为大数据的出现是解决“数据破碎化问题”的钥匙，“大数据能够通过构建数据沟通平台将‘数据孤岛’打通，最大程度地使得原本不关联的东西相互关联”。

邹东生同时强调如果使用不当，大数据技术非但不能解决，甚至可能加剧数据的破碎化，他说：“如果缺乏对大数据分析过程的深度理解和正确认知，搭建平台前没有充分的研究，大数据的使用将无法打通数据间的‘孤岛’，甚至可能会产生越来越多的数据碎片和信息流失”。那么，如何挖掘大数据的核心价值？邹东生为有意向应用大数据的企业提出了“三步走”的建议。第一步，从企业自身角度出发，“企业越早信息化，未来数据分析的‘弹药’就越充足。企业首先需要将自己的数据保存起来，在此基础上，有目的、有计划地收集外部数据，通过外部数据和内部数据的整合，形成足够大的数据体量”。第二步，从企业和大数据平台间的合作角度出发，“利用数据平台帮忙整合数据”。第三步，从对大数据分析的角度出发，“在结合行业和数据特征的基础上，设计算法，提供科学的量化、引导和分析，使大数据成为帮助企业更好提高决策效率和降低风险的分析工具”。在大数据概念如火如荼的背景下，应用大数据开展精准决策的例子不断涌现。电商可以借助大数据向潜在客户进行产品推荐，医院可以借助大数据分析患者治疗情况以提升自身的运维效率，传统百货商场和超市可以借助大数据进行动态定价等。从浅层次的数据收集和汇总，到深层次的分析和研究，对于大数据的应用遍布了各行各业。邹东生向《经济》记者举例：“假设某家企业目前记录到了大量的消费者数据，就可以研究其目标人群的特征，与数据库中的数据进行对比。如果能准确地‘画出’该企业的客‘画像’，即找到精准的客户人群，那么就可以不花费分文广告费进行精准营销。”从电商到医疗行业，从美国硅谷到中国上海，大数据正在逐步改变着各行各业，在世界上越来越多的角落产生影响。我们身处的这个时代，正在被冠以“大数据”的名称，未来，随着数据化的深入发展，越来越多的领域将借助数据进行记录和表征。FIN

2016年数据分析行业发展战略会议——引爆2016，大数据淘金时代

文 / 中商联数据分析专业委员会 会员处　　图 / 崔峻珩

“2016年数据分析行业发展战略会议”将于2016年1月9日——1月10日在北京前门建国饭店举行。引爆数据增值时代，开启数据收益新战略。这次会议会深度剖析大数据行业发展趋势，讲述当前的机遇与挑战，对Datahoop平台进行展示，展现超凡平台，各领域典型行业数据专家实战应用进行精彩案例的分享，协会数据分析专家会带你领略企业数据化分析定制化，尽情展现专家实力。

本次会议旨在探讨大数据从业者面对如此大好的行业机遇，我们应该如何把控商机？面临的业务问题应如何解决？指引大数据企业发展道路，剖析当前的商业机遇，协助数据企业对接商业平台，增加大数据企业创收。该次会议亮点众多，比如超凡平台展示，各领域典型行业数据专家实战应用精彩案例分享，更有协会数据分析专家带你领略企业数据应用定制化。从企业需求，事务所需求的供求两面手把手教你成功，深度挖掘企业数据化中企业技术分析的深度，让协会成为你前进的智力后盾，帮你解决个性化问题，感受数据化管理的强大优势；此次会议行业专家阵容强大，为你成功之路指点迷津。

会议通过调整数据分析行业发展方向，转变从业机构经营理念，学习行业新战略等一系列布局，进而顺应大数据时代的发展变化，与时俱进，提高行业竞争力，在大数据时代的热潮中脱颖而出，达到推动大数据发展，实现从业者自身价值的目的，帮助你抓住大数据时代下的商业机遇，让数据价值获得质变新生。

会议亮点：

1、针对当前大数据市场风口的现状，剖析大数据时代面临的机遇。

2、通过一系列的真实案例来印证当前我们不仅仅拥有机遇，同时还面临着挑战，结合这些挑战，专家给予专业的指导，帮你排忧解惑。

3、各典型领域的行业数据专家精彩分享大数据的实战应用案例。

4、结合政策导向、市场动态以及行业发展趋势，总结出2016年数据分析行业的发展战略。

5、协会技术研究团队和资深数据分析专家携手带你领略大数据的真正魅力所在。

6、最新技术抢先看，商业合作模式带你畅游数据蓝海。

会议地址：

北京市西城区永安路175号前门建国饭店

FIN

以大数据促进国家治理现代化

编辑 / 光明日报　　文 / 周文彰　　图 / 崔峻珩

当前正是利用大数据推进国家治理现代化的宝贵时机。对这一轮大数据革命，我国作出了非常及时的战略响应。7月1日，国务院办公厅发布了《关于运用大数据加强对市场主体服务和监管的若干意见》；7月4日，国务院发布了《关于积极推进“互联网+”行动的指导意见》；9月5日，国务院发布了《促进大数据发展行动纲要》。这几份重磅文件密集出台，标志着我国大数据战略部署和顶层设计正式确立。

大数据是一场管理革命，“用数据说话、用数据决策、用数据管理、用数据创新”，会给国家治理方式带来根本性变革。

“四个结合”助力国家大数据战略

实施国家大数据战略部署和顶层设计，需要我们做到“四个结合”：把政府数据开放和市场基于数据的创新结合起来。政府拥有80%的数据资源，如果不开放，大数据战略就会成为无源之水，市场主体如果不积极利用数据资源进行商业创新，数据开放的价值就无从释放；把大数据与国家治理创新结合起来。国务院的部署明确提出，“将大数据作为提升政府

治理能力的重要手段”“提高社会治理的精准性和有效性”，用大数据“助力简政放权，支持从事前审批向事中事后监管转变”“借助大数据实现政府负面清单、权力清单和责任清单的透明化管理，完善大数据监督和技术反腐体系”，并具体部署了四大重大工程：政府数据资源共享开放工程、国家大数据资源统筹发展工程、政府治理大数据工程、公共服务大数据工程；把大数据与现代产业体系结合起来。这里涉及农业大数据、工业大数据、新兴产业大数据等，我国的产业结构优化升级迎来难得的历史机遇；把大数据与大众创业、万众创新结合起来。国务院专门安排了“万众创新大数据工程”，数据将成为大众创业、万众创新的肥沃土壤，数据密集型产业将成为发展最快的产业，拥有数据优势的公司将迅速崛起。

此外，我国作为世界制造业第一大国，需要高度关注一个现实——大数据重新定义了制造业创新升级的目标和路径。无论是德国提出的工业4.0战略，还是美国通用公司提出的工业互联网理念，本质正是先进制造业和大数据技术的统一体。大数据革命骤然改变了制造业演进的轨道，加速了传统制造体系的产品、设备、流程贬值淘汰的进程。数字工厂或称智能工厂，是未来制造业转型升级的必然方向。我国面临着从“制造大国”走向“制造强国”的历史重任，在新的技术条件下如何适应变化、如何生存发展、如何参与竞争，是非常现实的挑战。

推动大数据在国家治理上的应用

在大数据条件下，数据驱动的“精准治理体系”“智慧决策体系”“阳光权力平台”将逐渐成为现实。大数据已成为全球治理的新工具，联合国“全球脉动计划”就是用大数据对全球范围内的推特（Twitter）和脸谱（Facebook）数据和文本信息进行实时分析监测和“情绪分析”，可以对疾病、动乱、种族冲突提供早期预警。在国家治理现代化进程中推动大数据应用，是我们繁重而紧迫的任务。

在政府治理方面，政府可以借助大数据实现智慧治理、数据决策、风险预警、智慧城市、智慧公安、舆情监测等。大数据将通过全息的数据呈现，使政府从“主观主义”“经验主义”的模糊治理方式，迈向“实事求是”“数据驱动”的精准治理方式。

经济治理领域也是大数据创新应用的沃土，大数据是提高经济治理质量的有效手段。互联网系统记录着每一位生产者、消费者所产生的数据，可以为每个市场主体进行“精确画像”，从而为经济治理模式带来突破。判断经济形势好坏不再仅仅依赖统计样本得来的数据，而是可以通过把海量微观主体的行为加总，推导出宏观大趋势；银行发放贷款不再受制于信息不对称，通过贷款对象的大数据特征可以很好地预测其违约的可能性；打击假冒伪劣、建设“信用中国”也不再需要消耗大量人力、物力，大数据将使危害市场秩序的行为无处遁形。

在公共服务领域，基于大数据的智能服务系统，将会极大地提升人们的生活体验，智慧医疗、智慧教育、智慧出行、智慧物流、智慧社区、智慧家居等等，人们享受的一切公共服务将在数字空间中以新的模式重新构建。

加强大数据动态的跟踪研究

我国要从“数据大国”成为“数据强国”，借助大数据革命促进国家治理现代化，还有几个关键问题需要深入研究。

切实建设数据政策体系、数据立法体系、数据标准体系。以数据立法体系为例，一定要在数据开放和隐私保护之间权衡利弊，找到平衡点。

重视对“数据主权”问题的研究。

借助大数据技术，美国政府和互联网、大数据领军公司紧密结合，形成“数据情报联合体”，对全球数据空间进行掌控，形成新的“数据霸权”。思科、IBM、谷歌、英特尔、苹果、甲骨文、微软、高通等公司产品几乎渗透到世界各国的政府、海关、邮政、金融、铁路、民航系统。在这种情况下，我国数据主权极易遭到侵蚀。对于我国来说，在服务器、软件、芯片、操作系统、移动终端、搜索引擎等关键领域实现本土产品替代进口产品，具有极高的战略意义，也是维护数据主权的必要条件。

“数据驱动发展”或将成为对冲当前经济下行压力的新动力。大数据是促进生产力变革的基础性力量，这包括数据成为生产要素，数据重构生产过程，数据驱动发展等。数据作为生产要素其边际成本为零，不仅不会越消耗越少，反而保持“摩尔定律”所说的指数型增长速度。这就可能给我国经济转型升级带来新动力，对冲经济下行压力。

需要建设一个高质量的“大数据与国家治理实践案例库”。国家行政学院一直重视案例库的建设，在中央的重视和支持下，就大数据促进国家治理这一主题，各部门、各地方涌现出大量创新性的实践案例，亟须进行系统梳理和总结，形成一个权威的“大数据与国家治理实践案例库”，以方便全国领导干部进行借鉴和推广。

在大数据时代，个人如何生存、企业如何竞争、政府如何提供服务、国家如何创新治理体系，都需要重新进行审视和考量。我们不能墨守成规，抱残守缺，而是要善于学习，勇于创新，按照党中央、国务院的战略部署，政府和市场两个轮子一起转，把我国建设成“数据强国”。FIN

北京电视台 —— 财经锋汇 未来无处不在的大数据

文 / 中商联数据分析专业委员会　会员处　　图 / 崔峻珩

大数据时代已经来临，它将在众多领域掀起变革的巨浪。但我们要冷静地看到，大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式研究将是大数据产业健康发展的关键。我们相信，在国家的统筹规划与支持下，通过各地方政府因地制宜制定大数据产业发展策略，通过国内外IT龙头企业以及众多创新企业的积极参与，大数据产业未来发展前景十分广阔。

本期会客厅，我们就针对邹会长的考察情况，对他进行一次深入的访谈。旨在通过本次采访，了解未来互联网加大数据将会是怎样的一幅图景。

主持人：十八届五中全会公报把大数据上升成为国家战略，那么未来互联网加大数据将会是怎样的一幅图景，今天我们来为您进行解读。我们一直说大数据，其实很多人还是对这个大数据只是一个很模糊的概念，可能认为是一堆数字，或者是一些电脑上的一些数据。那么邹会长先给我们解读一下这个大数据的内涵到底还包括哪些内容。

邹会长：我们一直在聊大数据，但是其实我们被大数据经常的一个概念所迷惑，就是认为大数据只是数据很大，但实际上大数据真正的起因是大体量的数据产生的决策，使企业更加精准地判断。实际上大数据的应用、大数据的研究使它加以落地，给企业带来实际的决策价值，才是大数据真正的内涵所在。

主持人：有很多领域、行多行业，是能够通过我们大数据进行改造，甚至进行颠覆的，给我们列出来医疗、教育、金融、农业、政府调控、生物科学、电商等等。能不能再给我们举一些具体的例子，这些到底在实际当中大数据在这些行业当中是怎么样来应用的?

邬玉良：实际上我们九三中央在对大数据进行调研的过程中，我们走访了上海，然后我们在上海走访了上海的盛康医院医疗中心。在上海盛康医疗中心然后把上海地区的38个医院统一联网连起来，然后实现了我们现在希望做到的我拿一张医疗卡，可以到任何一个医院去看病。完整记录以后就形成了一整套治疗的方法，还形成了一整套有效地预防的方法。

主持人：还有哪些大数据在其他领域应用的例子呢?

邹会长：其实大数据在交通领域的应用是非常广泛的，以美国的一个例子为例，就是GE他在美国的一个南部城市做了一个整个铁路交通网的大数据的分析。那么1年的时间里，这个铁路交通运输网1小时负责能提速1公里，那么在这片地区1年能节约整个的铁路交通的营运成本超过2个亿。那么这种分析它是包括整个的运维，包括所有的环节，它要做非常缜密的模型。

那我们国内你看我们现在每个人手机上全有各种APP，那比如说高德地图、百度地图等等这些，那么我原来以为仅仅我们是做大数据的，所以我原来以为经常会外出看它的交通流量状况。

它是通过大量的搜索数据，比如说我想去北京旅游，我想来看红叶节，我就要提前搜这个红叶节，搜这样的信息数据。那么这样的信息汇总多了，那么就能找到了一些每个人的趋势和偏好，那么在这种情况下，我们就可以公布一些这个节假日可能某个地区、某个旅游网点大家集中会去的地方，那么你就可以体现地调配自己的时间和所要去旅游的目的地。那么这些都是通过各种各样的大数据的汇总，帮我们解决一些实际问题的例子。

那么最后一个的话就是我们说的农业，其实现在中国农业整个来说，那么农业的话其实在没有大数据支持的情况下，我们经常说靠天吃饭。但是现在有了大数据之后，我们就可以把全中国的气象数据给分析出来，那我们分析这些气象数据的话我们就可以知道，2016年大概是什么地区可能会有这种狂风暴雨闪电雷鸣，有这种洪涝灾害，有这种泥石流。我们就可以判断2016年什么地区的农业有可能有减产的这种风险等等。

主持人：这是我们在政府层面的应用，那十三五同时也说企业化，提出企业化数据建设，邹会长在这方面有没有什么例子呢?

邹会长：我觉得就是从我们角度来讲，我们一直是从事数据研究的，那么我们协会2008年就开始成立。就像刚才两位似的，我们第一次在这次五中全会中明确提出来，国家提出

来大数据战略，这个也是我们兴奋的，为什么？因为从国家角度开始推进这件事了，而且明确提出来企业要构建自己的这种数据化的一些服务，或者不管是技术也好，不管是分析也好，这个意义重大。拿一句心里大家经常会聊到的话，就是目前来讲，现在来讲业务即数据，在未来几年以后，数据即业务。你现在是把企业的所有的行为、所有的业务，把它数据和进行存储和分析。那么到了几年以后，企业去运维自己的数据，实际上就是在经营自己的业务。

主持人：在这个大数据时代，就是说这个大数据时代需要什么样的人才呢?

邹会长：在国外把大数据人才分成两个领域，一个是偏技术型，就是偏技术型人才，虽然他也需要懂后面的东西，但是他对于技术构建这方面要求是比较高的。还有一方面是偏后面的基于研究和分析这一块的人才。实际在欧美国家他也有一个大数据发展的轨迹，在刚开始的时候他们更多地偏技术型引入的人才。

而在这个方面领域的人员，他对不仅仅是需要技术，他更多地需要经济学决策型人才的知识储备去做这件事情，把他两个结合起来当然是最完美的。

邹会长：我觉得很关键的一点就是什么呢？它能够帮助中国的企业，包括中国的经济进行合理地转型，让大数据真正能够助推企业发展，让大数据真正能够帮中国的企业产生越来越好的、越来越棒的竞争力。

总结:此次访谈，我们了解了大数据在各个领域的应用。政府对大数据的展望和未来这个行业在各个领域的发展有了更多的认知,大数据上升到国家战略上的非凡意义。

FIN

大数据工具比较：R语言和Spark谁更胜一筹？

文 / 36大数据　Vivek Murugesan　图 / 崔峻珩

由于R语言本身是单线程的，所以可能从性能方面对比Spark和R并不是很明智的做法。即使这种比较不是很理想，但是对于那些曾经遇到过这些问题的人，下文中的一些数字一定会让你很感兴趣。

你是否曾把一个机器学习的问题丢到R里运行，然后等上好几个小时？而仅仅是因为没有可行的替代方式，你只能耐心地等。所以是时候去看看Spark的机器学习了，它包含R语言大部分的功能，并且在数据转换和性能上优于R语言。

曾经我尝试过利用不同的机器学习技术——R语言和Spark的机器学习，去解决同一个特定的问题。为了增加可比性，我甚至让它们运行在同样的硬件环境和操作系统上。并且，在Spark中运行单机模式，不带任何集群的配置。

在我们讨论具体细节之前，关于Revolution R 有个简单的说明。作为R语言的企业版，Revolution R试图弥补R语言单线程的缺陷。但它只能运行在像Revolution Analytics这样的专有软件上，所以可能不是理想的长期方案。如果想获得微软Revolution Analytics软件的扩展，又可能会让事情变得更为复杂，比方说牵扯到许可证的问题。

因此，社区支持的开源工具，像是Spark，可能成为比R语言企业版更好的选择。

数据集和问题

分析采用的是Kaggle网站 上的数字识别器的数据集，其中包含灰度的手写数字的图片，从0到9。

每张图片高28px，宽28px，大小为784px。每个像素都包含关于像素点明暗的值，值越高代表像素点越暗。像素值是0到255之间的整数，包括0和255。整张图片包含第一列在内共有785列数据，称为“标记”，即用户手写的数字。

分析的目标是得到一个可以从像素数值中识别数字是几的模型。

选择这个数据集的论据是，从数据量上来看，实质上这算不上是一个大数据的问题。

对比情况

针对这个问题，机器学习的步骤如下，以得出预测模型结束：

在数据集上进行主成分分析和线性判别式分析，得到主要的特征。对所有双位数字进行二元逻辑回归，并且根据它们的像素信息和主成分分析以及线性判别式分析得到的特征变量进行分类。

在全量数据上运行多元逻辑回归模型来进行多类分类。根据它们的像素信息和主成分分析以及线性判别式分析的特征变量，利用朴素贝叶斯分类模型进行分类。利用决策树分类模型来分类数字。

在上述步骤之前，我已经将标记的数据分成了训练组和测试组，用于训练模型和在精度上验证模型的性能。

大部分的步骤都在R语言和Spark上都运行了。详细的对比情况如下，主要是对比了主成分分析、二元逻辑模型和朴素贝叶斯分类模型的部分。

主成分分析

主成分分析的主要计算复杂度在对成分的打分上，逻辑步骤如下：

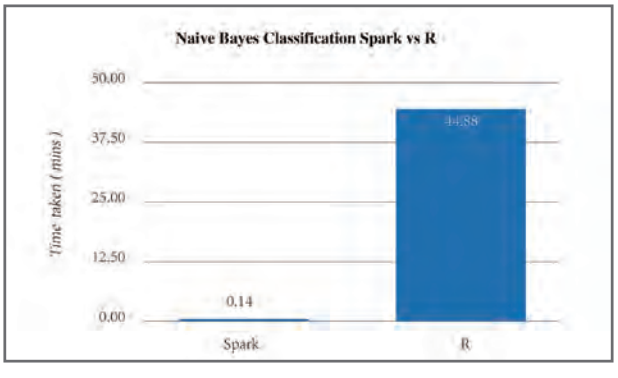
通过遍历数据以及计算各列的协方差表，得到KxM的权重值。（K代表主成分的个数，M代表数据集的特征变量个数）。

当我们对N条数据进行打分，就是矩阵乘法运算。

通过NxM个维度数据和MxK个权重数据，最后得到的是NxK个主成分。N条数据中的每一条都有K个主成分。

在我们这个例子中，打分的结果是42000 x 784的维度矩阵与784 x 9的矩阵相乘。坦白说，这个计算过程在R中运行了超过4个小时，而同样的运算Spark只用了10秒多矩阵相乘差不多是3亿次运算或者指令，还有相当多的检索和查找操作，所以Spark的并行计算引擎可以在10秒钟完成还是非常令人惊讶的。

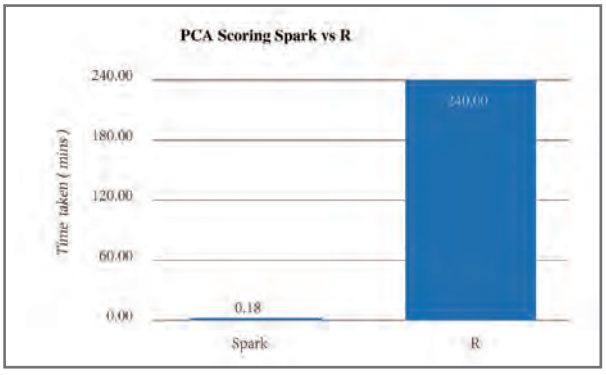
我通过查看前9个主成分的方差，来验证了所产生的主成分的精度。方差和通过R产生的前9个主成分的方差吻合。这一点确保了Spark并没有牺牲精度来换取性能和数据转换上的优势。



逻辑回归模型

与主成分分析不同的是，在逻辑回归模型中，训练和打分的操作都是需要计算的，而且都是极其密集的运算。在这种模型的通用的数据训练方案中包含一些对于整个数据集矩阵的转置和逆运算。

由于计算的复杂性，R在训练和打分都需要过好一会儿才能完成，准确的说是7个小时，而Spark只用了大概5分钟。



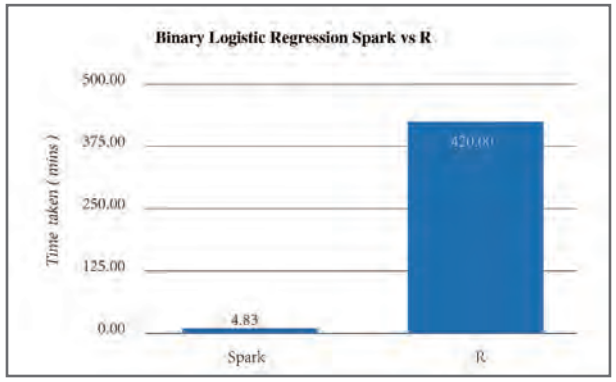
这里我在45个从0到9的双位数字上运行了二元逻辑回归模型，打分/验证也是在这45个测试数据上进行的。

我们也并行执行了多元逻辑回归模型，作为多类分类器，大概3分钟就完成了。而这在R上运行不起来，所以我也没办法在数据上进行对比。

对于主成分分析，我们采用AUC值来衡量预测模型在45对数据上的表现，而Spark和R两者运行的模型结果的AUC值差不多。

朴素贝叶斯分类器

与主成分分析和逻辑回归不一样的是，朴素贝叶斯分类器不是密集计算型的。其中需要计算类的先验概率，然后基于可用的附加数据得到后验概率。



如上图所示，R大概花了45余秒完成，而Spark只用了9秒钟。像之前一样，两者的精确度旗鼓相当。

同时我也试着用Spark机器学习运行了决策树模型，大概花了20秒，而这个在R上完全运行不起来。

Spark机器学习入门指南

对比已经足够，而这也成就了Spark的机器学习。 最好是从编程指南开始学习它。不过，如果你想早点尝试并从实践中学习的话，你可能要痛苦一阵子才能将它运行起来吧。

为搞清楚示例代码并且在数据集上进行试验，你需要先去弄懂Spark的RDD 支持的基本框架和运算。然后也要弄明白Spark中不同的机器学习程序，并且在上面进行编程。当你的第一个Spark机器学习的程序跑起来的时候，你可能就会意兴阑珊了。

以下两份资料可以帮你避免这些问题，同时理顺学习的思路：

Spark机器学习所有的源代码，可提供任何人拿来与R语言作对比：

Docker容器的源代码，Spark和上述项目的包已预置在内，以供快速实施：

容器中已事先安装Apache Hadoop，并且在伪分布式环境下运行。这可以将大容量文件放进分布式文件系统来测试Spark。通过从分布式文件系统加载记录，可以很轻松地来创建RDD实例。

产能和精度

人们会使用不同的指标来衡量这些工具的好坏。对我来说，精准度和产能是决定性的因素。

大家总是喜欢R多过于Spark机器学习，是因为经验学习曲线。他们最终只能选择在R上采用少量的样本数据，是因为R在大数据量的样本上花了太多时间，而这也影响了整个系统的性能。

对我们来说，用少量的样本数据是解决不了问题的，因为少量样本根本代表不了整体。所以说，如果你使用了少量样本，就是在精度上选择了妥协。

一旦你抛弃了少量样本，就归结到了生产性能的问题。机器学习的问题本质上就是迭代的问题。如果每次迭代都花费很久的话，那么完工时间就会延长。可是，如果每次迭代只用一点时间的话，那么留给你敲代码的时间就会多一些了。

结论

R语言包含了统计计算的库和像ggplot2这样可视化分析的库，所以它不可能被完全废弃，而且它所带来的挖掘数据和统计汇总的能力是毋庸置疑的。

但是，当遇到在大数据集上构建模型的问题时，我们应该去挖掘一些像Spark ML的工具。Spark也提供R的包，SparkR可以在分布式数据集上应用R。

最好在你的“数据军营”中多放点工具，因为你不知道在“打仗”的时候会遇到什么。因此，是时候从过去的R时代迈入Spark ML的新时代了。 FIN

基于灰色模型的重庆市快递行业业务量及收入预测分析

文 / 杜长海 博士　　图 / 周子赫

近年来，随着电子商务的爆发式增长，我国快递行业发展突飞猛进。作为新兴的基础性产业，快递服务联系各行各业，贴近人民生活，服务生产消费，对于促进国民经济和社会发展具有重要作用。因此，提前对快递业务量及收入进行预测，不仅能为快递企业提供预警，使之制定应对策略，提高顾客满意度，而且能为快递行业的持续发展提供科学决策服务。

灰色系统理论研究的对象是“部分现象已知，部分现象未知”的小样本不确定系统。在任何一个系统中，有已知的信息和未知的信息，信息完全明白的为白色系统. 完全不明白的为黑色系统，信息部分明确，部分不明确的为灰色系统。对这样的系统建立数学模型进行预测时，总是力图使那些完全不明确的即灰色信息由“灰”变“白”，使预测达到一定精度。灰色预测是灰色系统理论的一个重要组成部分，是通过将无规律的原始数据生成、建立模型、拟合、探求系统内在规律，预测未来的数学预测模型系统。

利用灰色模型处理数据对数据没有很强的限制，尤其是数据较少时，预测比较准确。本文运用灰色GM(1,1)模型对重庆市2008–2014年快递行业的年度业务量及收入进行建模，检验其精度，并对未来2年的重庆市快递业务量及收入进行年度预测，以期望为重庆市制定快递行业发展战略提供参考依据。具体步骤如下：

一、数据来源

本文数据来源于重庆市统计局公布的《重庆统计年鉴》，具体为2008–2014年快递行业的年度业务量及收入数据，见表1。

表1 重庆市2008–2014年度快递业务量及收入

数据种类	2008	2009	2010	2011	2012	2013	2014
快递业务量（万件）	1608.40	2240.00	2829.40	4068.00	5497.90	10614.80	13886.30
快递业务收入（万元）	35346.10	48879.20	60268.10	76828.80	103426.50	136957.50	201060.00

二、灰色GM(1,1)模型的建立

GM(1,1)模型是基于随机的原始时间序列，经按时间累加后所形成的新的时间序列呈现的规律可用一阶线性微分方程的解来逼近。

设非负原始时间序列$X^{(0)}$有n个观测值$X^{(0)}=\left(x^{(0)}(1),x^{(0)}(2),\mathrm{L}\ ,x^{(0)}(n)\right)$，其中$x^{(0)}(k)>0$，$k=1,2,\mathrm{L}\ ,n$；

对其做一次累加，得$X^{(1)}=\left(x^{(1)}(1),x^{(1)}(2),\mathrm{L}\ ,x^{(1)}(n)\right)$，其中$x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i)$，$k=1,2,\mathrm{L}\ ,n$

GM(1,1)模型的微分方程为：$\dfrac{dX^{(1)}}{dt}+aX^{(1)}=b$ （1）

其中，a为发展系数，b为内生控制灰数。

设$\hat{a}$为待估参数向量，$\hat{a}=[a,b]^T$，利用最小二乘法可求得a和b的值。

$$\hat{a}=\left(B^TB\right)^{-1}B^TY_N \quad （2）$$

其中：$B=\begin{pmatrix}-\frac{1}{2}\left(x^{(1)}(1)+x^{(1)}(2)\right) & 1\\ -\frac{1}{2}\left(x^{(1)}(2)+x^{(1)}(3)\right) & 1\\ \mathrm{M} & \mathrm{M}\\ -\frac{1}{2}\left(x^{(1)}(n-1)+x^{(1)}(n)\right) & 1\end{pmatrix}$，$Y_N=\left(x^{(0)}(2),x^{(0)}(3),\mathrm{L}\ ,x^{(0)}(n)\right)^T$ （3）

求微分方程，可得预测模型为：$\hat{x}^{(1)}(k+1)=\left(x^{(0)}(1)-\dfrac{b}{a}\right)e^{-ak}+\dfrac{b}{a}$；$k=1,2,\mathrm{L}\ ,n$ （4）

通过对$\hat{x}^{(1)}$做一次累减还原，则：$\hat{x}^{(0)}(k+1)=x^{(1)}(k+1)-x^{(1)}(k)$ （5）

通过上述公式，就能计算未来的预测值，但是一定要经过检验才能判定这个模型是不是合理，只有通过精度检验合理的模型才能用来做预测。

同时，由于灰色GM(1，1) 模型是根据最小二乘法原理建立的类指数增长模型，如果对原始数据序列进行一些预处理，提高数据序列光滑度，就可以在很大程度上改善灰色预测的精确度。对原始数据序列的预处理如下：

首先，需要对原始数据序列做对数变换，令$Y^{(0)}=\left(y^{(0)}(1),y^{(0)}(2),\mathrm{L}\ ,y^{(1)}(n)\right)$，其中，$y^{(0)}(k)=\ln x^{(0)}(k)$，$k=1,2,\mathrm{L}\ ,n$；然后，利用平滑度更高的序列进行灰色预测；最后$Y^{(0)}$将预测数据进行逆变换即可得到原始数据的预测值。

三、模型精度检验

一般最常用的是残差检验和后验差检验，具体为：

1、残差检验

计算原始数列$x^{(0)}(k)$与模型拟合值$\hat{x}^{(0)}(k)$的残差$\mathrm{d}^{(0)}(k)$和相对误差$M^{(0)}(k)$，残差$\mathrm{d}^{(0)}(k)=x^{(0)}(k)-\hat{x}^{(0)}(k)$，

相对误差$M^{(0)}(k)=\dfrac{\mathrm{d}^{(0)}(k)}{x^{(0)}(k)}$，平均相对误差$\bar{M}^{(0)}=\dfrac{1}{n}\sum_{k=1}^{n}\left|M^{(0)}(k)\right|$。

根据经验，一般认为$\bar{M}^{(0)}<0.2$时，模型残差检验是合格的。

2、后验差检验

先计算原始数列的平均值$\bar{x}$，残差平均值$\bar{\mathrm{d}}$，$\bar{x}=\dfrac{1}{n}\sum_{k=1}^{n}x^{(0)}(k)$，$\bar{\mathrm{d}}=\dfrac{1}{n}\sum_{k=1}^{n}\mathrm{d}^{(0)}(k)$，再计算原始数列的方差$s_1^2$，

残差方差s_2^2，$s_1^2=\dfrac{1}{n}\sum_{k=1}^{n}\left(x^{(0)}(k)-\bar{x}\right)^2$，$s_2^2=\dfrac{1}{n}\sum_{k=1}^{n}\left(\mathrm{d}^{(0)}(k)-\bar{\mathrm{d}}\right)^2$，就可以计算出方差比$c$和小误差概率$p$，

方差比$c=\frac{s_2}{s_1}$，小误差概率$p=P\left(\left|\mathrm{d}^{(0)}(k)-\bar{\mathrm{d}}\right|<0.6745s_1\right)$。

模型的精度等级一般由p和c共同刻画（见表2）。

表2 精度等级参照表

精度等级	c	p
一级（好）	<0.35	>0.95
二级（合格）	<0.5	>0.8
三级（勉强）	<0.65	>0.7
四级（不合格）	≥0.65	≤0.7

四、应用实例与预测分析

将第一节中的快递业务量及收入经对数变换后按照第二节GM(1,1)模型进行计算求解并根据第三节进行模型精度检验，为简化起见，中间步骤与中间结果不再赘述，计算结果见表3、表4与表5。

表3

年度	业务量实际值（万件）	对数值	对数预测值	残差	相对误差(%)	业务量预测值（万件）
2008	1608.40	7.3830	7.3830	0.0000	0.0000	1608.4000
2009	2240.00	7.7142	7.6324	0.0818	1.0610	2063.9661
2010	2829.40	7.9478	7.9827	−0.0349	−0.4394	2929.9512
2011	4068.00	8.3109	8.3492	−0.0383	−0.4605	4226.7139
2012	5497.90	8.6121	8.7324	−0.1203	−1.3971	6200.8406
2013	10614.80	9.2700	9.1333	0.1367	1.4748	9258.4630
2014	13886.30	9.5387	9.5525	−0.0139	−0.1456	14080.5156
2015	–	–	9.9910	–	–	21830.1332
2016	–	–	10.4497	–	–	34533.1391

表4

年度	业务收入实际值（万元）	对数值	对数预测值	残差	相对误差(%)	业务收入预测值（万元）
2008	35346.10	10.4729	10.4729	0.0000	0.0000	35346.1000
2009	48879.20	10.7971	10.7464	0.0507	0.4694	46463.8455
2010	60268.10	11.0066	11.0144	−0.0078	−0.0713	60742.6028
2011	76828.80	11.2493	11.2891	−0.0397	−0.3531	79941.7459
2012	103426.50	11.5466	11.5706	−0.0239	−0.2073	105932.2528
2013	136957.50	11.8274	11.8591	−0.0317	−0.2676	141361.5516
2014	201060.00	12.2114	12.1548	0.0566	0.4632	190002.3319
2015	–	–	12.4579	–	–	257269.9160
2016	–	–	12.7685	–	–	350995.3631

表5

数据种类	a	b	$\bar{M}^{(0)}$	S_2	S_1	c	p
快递业务量	−0.0449	7.1310	0.8297%	0.0781	0.7383	0.1058	1
快递业务收入	−0.0246	10.3567	0.3053%	0.0358	0.5600	0.0640	1

从而，可以建立重庆市快递业务量灰色预测模型为：

$$\begin{cases}\hat{x}^{(1)}(k+1)=166.2687e^{0.0449k}-158.8857\\ \hat{x}^{(0)}(k+1)=x^{(1)}(k+1)-x^{(1)}(k)\end{cases} \quad (6)$$

建立重庆市快递业务收入灰色预测模型为：

$$\begin{cases}\hat{x}^{(1)}(k+1)=430.9639e^{0.0246k}-420.4910\\ \hat{x}^{(0)}(k+1)=x^{(1)}(k+1)-x^{(1)}(k)\end{cases} \quad (7)$$

从表5来看，2种模型的平均相对误差分别为0.8297%、0.3053%，均明显小于5%；方差比 $c=0.1058<0.35$，$c=0.0640<0.35$，$p=1>0.95$，从表2可知，2种模型的精度等级均为一级。可见，所建模型达到比较高的预测精度，模型评价等级为好，可以用于重庆市快递业务量及收入预测。

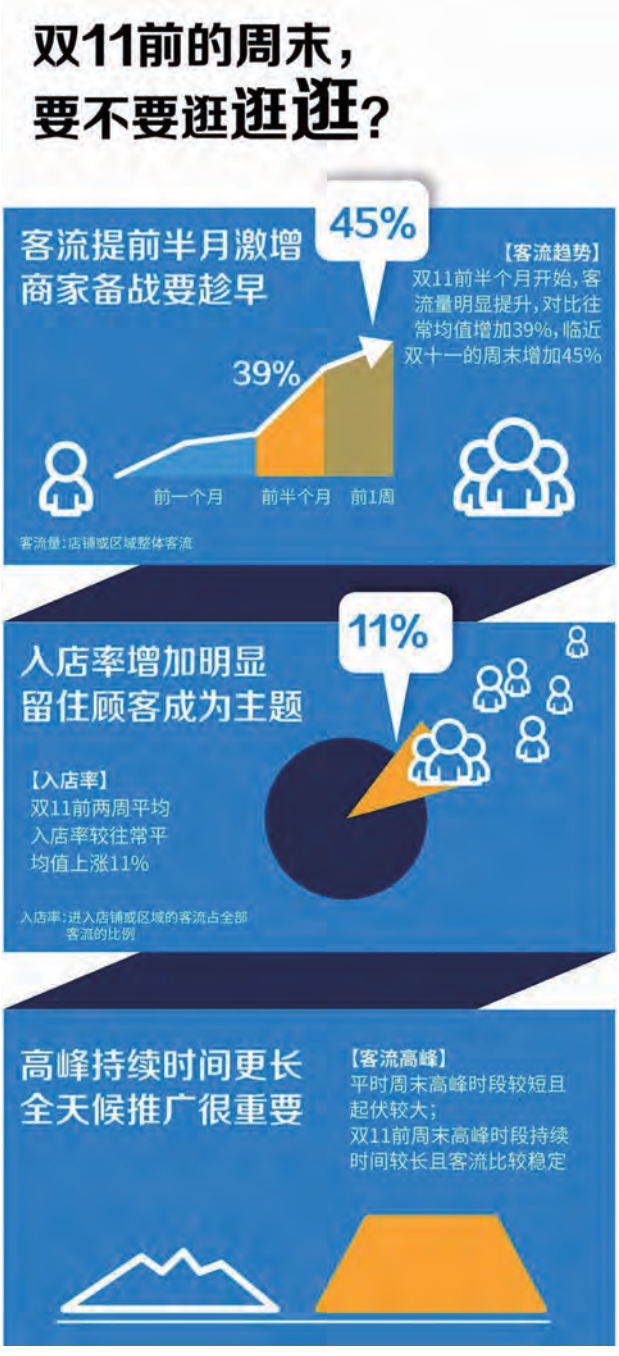
因此，使用模型（6）与（7）对2015年、2016年重庆市快递行业的业务量与收入进行预测，经过对数逆变换后的得到的预测值，见表3、表4的最后两行，快递业务量预测值分别为21830.1332万件、34533.1391万件，快递业务收入预测值分别为257269.9160万元、350995.3631万元，均呈现出显著持续增长趋势，说明重庆市快递行业的发展具有广阔的市场和巨大的发展潜力。需要指出的是，若精度达不到预期要求，可通过残差模型进行修正或对原始数据序列进行适当取舍，以提高模型进度。

综上所述，随着我国将物流产业并入我国十大发展产业以及电子商务的迅速普及，快递行业管理将更加规范化，快递企业业务量将会得到进一步的提升，其收入也会迅速增加。为此，快递企业应不断进行科技创新，提高其科技水平，引进先进的分拣设备，合理配置资源，规范运营机制，这对于规划和建立重庆市快递物流网络体系具有十分重要的意义。 FIN

数据告诉你：面对双11，线下商机何在

文 / 大数据文摘　　编辑 / 周子赫　　图 / 崔峻珩

每年的双11都呈愈演愈烈之势，今年双11，天猫成交额更是达到了前所未有的912亿。线上的购物狂欢对实体商业究竟有什么影响？实体商业在双11的大背景下，还有没有可以挖掘的商机？



双11为实体商业带来了大量的销售机会。芝麻科技联合阿里巴巴大数据平台、意略明市场营销咨询带来了实体商业（以服装与化妆品为代表）的线下客流分析和消费者大数据画像报告。研究数据涉及北京、武汉、深圳重点商圈的男装、女装、化妆品店在“双11”前一个周末（11月7日、11月8日）的客流及客群画像与“双11”前三周的对比。

【商机何在？】

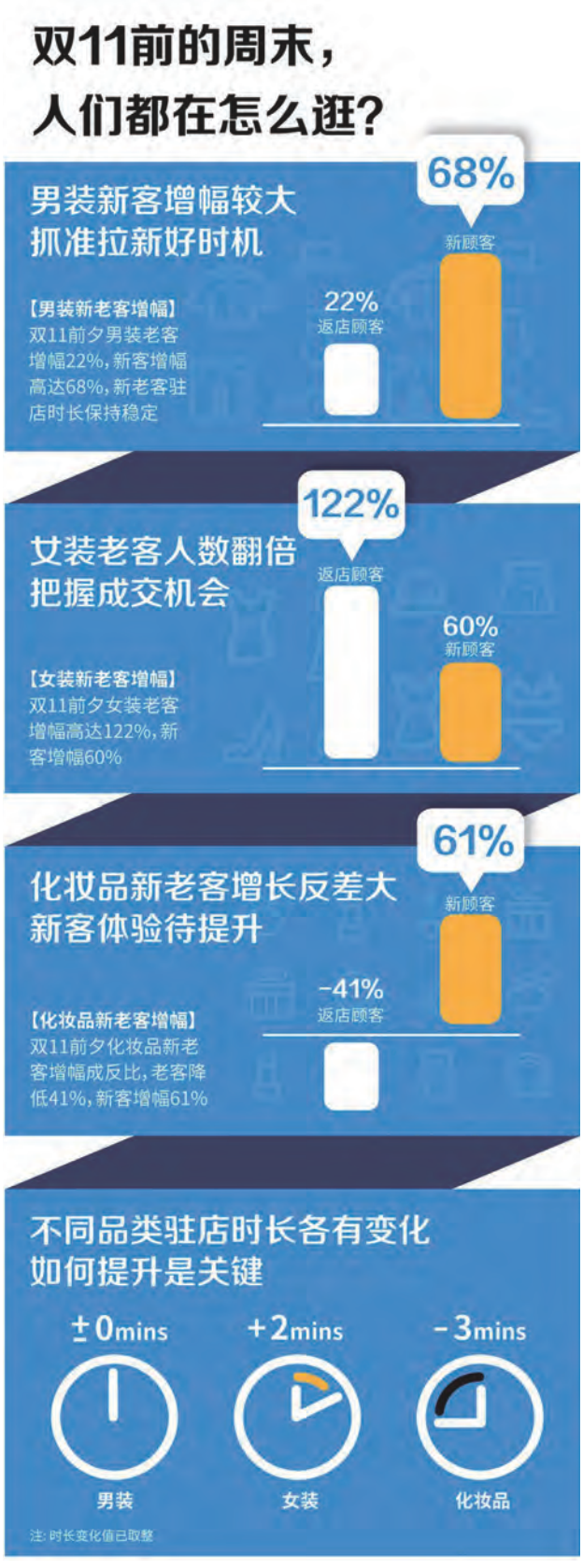
根据对客流量、入店量的统计，双11给实体商业带来了大量的客流和潜在的销售机会，如果品牌门店能够做出有针对性的营销活动，将有很大机会抢夺线上流量，将客流量转化为销售量。

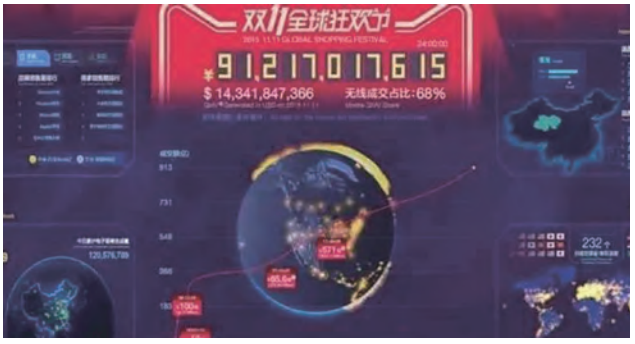
双11的周末客流高峰持续时间更久、峰值更高，这给品牌门店带来了更大挑战，同时也提示门店进行全天候的营销活动。

【商机何在？】

热力图直观呈现了品牌受众在双11前周末的分布情况，如果品牌能够充分利用客群的聚集效应，可以高效地将人气转化为销量。芝麻科技的大数据消费者画像除了有近500项标签刻画消费者画像外，还能够提供品牌受众地理位置分布热力图，了解品牌受众的逛街习惯、集中地理位置等，可作为品牌新店选址、基于地理位置广告投放策略参考等。

【商机何在？】





统计结果显示，双11激励了男装消费者进入实体店接触新品牌，且男装消费者的新老顾客停留时长都没有显著变化，品牌在营销与导购策略上更加注重给新顾客留下良好的品牌印象，鼓励消费者体验试穿。

双11前，女装店吸引了大量老顾客，她们对品牌熟悉，更愿意在门店里长久停留选购（从平均5.3分钟提升至8.5分钟），或是为即将到来的线上抢购做准备，品牌如果能满足老客们已有的购买意愿，将有可能促成她们的线下购买。

化妆品店的老客对于品牌与产品已经非常熟悉，不再需要进店体验，而大量新客会进入门店试用产品。如果导购或BA能为这部分新客提供满意的服务体验，将有可能引发他们的购买意愿，并形成最终转化。值得注意的是，与平时相比，无论新客还是老客，在店内都不会停留太长时间，如果能开展更多体验性营销活动，增加消费者的停留时间，也会带来更多销售机会。

双11不是实体商业的黑色周，相反，无论是客流数据，还是客群画像，都证明了旺盛的购物意愿会为实体商业带来大量销售机会。与其自怨自艾，实体商业不如赶紧修炼内功，好好统计、分析品牌与门店的各项数据，让数据说话，从数据中寻找商机。FIN

服装行业的销量预测

文 / 河南智宸项目数据分析师事务所 周彦锋　　图 / 周子赫

摘要：在服装市场营销竞争日趋激烈的今天，提高企业供应链管理的销率，降低滞销库存压力是企业立足服装市场的最基本保障。服装销售预测是根据服装市场的销售信息和营销计划，用科学的方法进行市场和销售分析，确定在未来某一时间段的销售量(额)。准确的预测可以使企业或公司有计划地安排面辅料采购、服装生产、合理进仓和减少库存、加快资金流动，对服装企业良性营销活动起着重要作用。本文主要结合某服装企业的实际情况，探究在小数据集的情况下，利用灰色模型和三次曲线的方法对不同类型服装产品分别分别进行销量预测。

一. 背景介绍

客户对象：河南某服装二级批发商

企业在行业所处位置：在生产商–代理商–批发商–实体店零售商–消费者的商品流通过程中处于中间位置，从上游服装代理商大批量订货，下游面对实体店多批次小批量出货。

可以从企业内部调用的数据：过去一年仓库中的各个品牌的每日调拨入库数量，盘盈入库数量，批发出库数量以及上期存货数量，本期结存数量。

客户需求：准确的预测将来的需求量以满足合理进仓和避免库存积压，减少库存成本。

二. 商业理解

数据理解：通过与业务人员的交流我们知道这些数据指标之间的关系：本期末结存数量=（本期初存货数量+调拨入库数量+盘盈入库数量–批发出库数量）=下期初存货数量。

业务知识及洞察：（1）服装是一个以年为一个循环周期的产品，做长期预测需要至少四年的数据，而我们只有一年的数据，这迫使了我们必须放弃了去预测企业长期的销量，而去追寻销量短期的规律，并且不能太短，如果预测一天为单位的销量随机因素太多，没有意义，这使得我们想到了以周为单位去考虑这种情况。

（2）通过与具体业务人员的交流，我们想要去发现影响服装出库的主要因素，首先这种影响实体店的进货的原因与影响我们消费者购买的动机有一些异曲同工之处，但是也有着失之毫厘差之千里的差别。然后我们得出了温度，时间这两个主要的影响因素，但是需要注意的两点是：一、温度和时间之间具有关联关系，他们是交互作用影响销量，二、温度对顾客的购买行为的影响是一个提前影响，并且实体店也会基于温度变化提前备货，这一点我们从实际生活出发就可以验证，我们买短袖都不是在夏天最热的时候，而是在温度高峰来临前早就买好了，买羽绒服也是在我们感觉天气要变冷的时候而不是大雪纷飞的时候，同样店铺进货的还要比顾客买衣服的时间早得多，早在顾客要卖货之前就应该备货了。

（3）通过对这40多种服装每周销售量的趋势图对比我们可以发现，服装行业的产品大概可分为三类：需求确定型，需求随机型，需求季节型。对于需求确定型的服装销量不需要我们做预测，但是实际生活中真正满足需求确定的产品很少，所以基本上只需考虑后两种类型的销量预测。

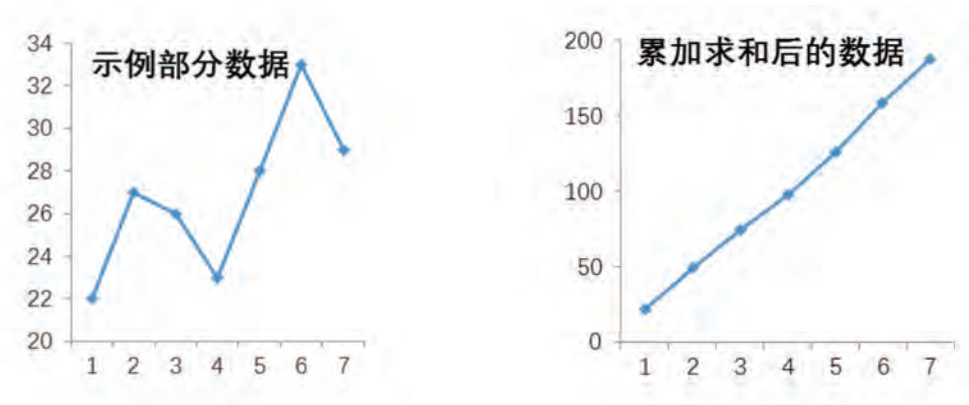
（4）预测模型的选取问题：由于我们面对的企业客户只能提供一年销售数据，这就造成了很难用简单直接的模型方法去预测，比如：我们在CPDA的课程中学过的移动平均、指数平滑、季节型分解等方法，尤其是需求季节型服装理论上用季节型分解的方法做出来的预测效果是最好的方案。但是季节分解要求我们至少有四年的数据才能产生一个比较可信的季节规律。对于不同的情况总有对应适合它的方法，大数据也好，传统数据也好，不论数据大小，只要它有挖掘的价值，它所能为我们带来的信息总是有的，你见或不见。

三. 需求随机型服装销量预测–灰色模型

灰色模型从灰色系统中抽象出来的模型。灰色系统是既含有已知信息，又含有未知信息或非确知信息的系统，这样的系统普遍存在。与之对立的还有黑色系统和白色系统，研究灰色系统的重要内容之一是如何从一个不甚明确的、整体信息不足的系统中抽象并建立起一个模型，该模型能使灰色系统的因素由不明确到明确，由知之甚少发展到知之较多提供研究基础。

灰色模型的基本思想是用原始数据组成原始序列(0)，经累加生成法生成序列(1)，它可以弱化原始数据的随机性，使其呈现出较为明显的特征规律。对生成变换后的序列(1) 建立微分方程型的模型即GM模型。

核心思想如下图所示：



同样GM模型中我们选取的是GM（1,1）模型，其数据建模步骤如下：

给定观测数据：$x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \ldots x^{(0)}(N)\}$

第一步求累加和：

$$x^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \ldots x^{(1)}(N)\}$$

第二步，计算

$$[x^{(1)}(1) + x^{(1)}(2)] / 2$$

$$[x^{(1)}(2) + x^{(1)}(3)] / 2$$

$$\cdots [x^{(1)}(N-1) + x^{(1)}(N)] / 2$$

第三步，令

$$A = \begin{pmatrix} -[x^{(1)}(1) + x^{(1)}(2)] / 2 & 1 \\ M & M \\ -[x^{(1)}(N-1) + x^{(1)}(N)] / 2 & 1 \end{pmatrix} \quad B = \left(x^{(0)}(2) \ldots x^{(0)}(N)\right)$$

设 $\hat{\partial} = \begin{pmatrix} \partial \\ u \end{pmatrix}$，$\hat{\partial} = (A^T A)^{-1} A^T B^T$，解微分方程可得，

预测方程为：$x^{(1)}(k+1) = [x^{(1)}(1) - \frac{u}{a}]e^{-ak} + \frac{u}{a}$

其中–a称之为发展灰数，u称之为内生控制灰数。

当k=1,2,3...N–1时，$x^{(1)}(k+1)$ 为拟合值，当k≥N时，为预测值。

注意：这是相对累加的拟合和预测值，还应该用后减运算还原，即：$x^{(0)}(k+1) = x^{(1)}(k+1) - x^{(1)}(k)$，根据以上理

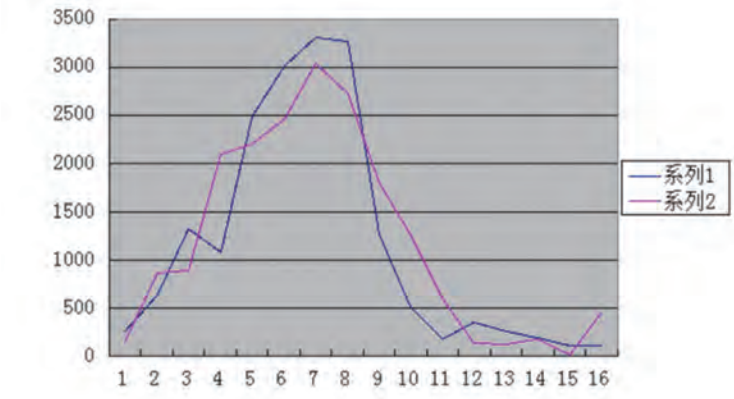
论方法就能得出销量的预测值。真正进行计算时我们使用matlab软件编程来计算最终预测值，由于过程较为复杂本文暂不详细叙述。

四．需求季节型服装销量预测–曲线拟合

曲线拟合也是CPDA课程关于预测方法的一种，通常我们选用的曲线拟合的方法有二次曲线、三次曲线、幂级数、对数以及logstic等，我们最终选取的是三次曲线拟合是一种在众多方法筛选之后（通过比较不同模型预测的结果）最优的一种后验选择。但是有一点必须注意的是：温度及时间都是影响销量的关键因素，并且他们之间有很强的关联性，所以这一点决定了我们在使用spss进行回归的时候一定要选择逐步回归这一项，目的是为了削弱二者之间的关联影响。

最后这些假设模型即每个函数表达式，在通过f检验，t检验的基础上选择调整r具有最大值的的函数表达式，那么这个函数便是我们得到的最优的预测模型。

即Y=b0+x*(b1t+b2t^2+b3t^3)，（x代表温度，t代表时间），调整r方为0.836，f检验与t检验都通过，拟合曲线与原始销量曲线对比如下图所示：



注：系列1为真实销量曲线，系列2为拟合曲线

五．小结

使用模型优点：使用灰色数学处理不确定量使之量化，充分利用了已知信息发现事物的运动规律。其次我们建立的模型，选取的方法完全是从业务需求出发，从拟合结果与实际值得对比检验预测模型的合理性，并能用于指导业务人员的判断。

但是如果我们仔细考虑还是会发现模型构建的一些不足之处;

没有考虑服装自身对顾客吸引力的变化，产品的好坏自然决定着受欢迎的程度，我们的模型其实是在假设顾客近期对同一款产品的需求没有太大波动的情况下考虑的，比如遇到爆款，或者爆冷款的服装，具体业务人员在利用此方法分析的时候应适当调高或者调低自己的预测值。

对于模型选取指标的有一定的偏差，比如在考虑与温度相关的因素时，拿到的是河南地区每天的最高气温与最低气温，我们选取用近似用平均值作为一天的平均气温，在考虑温度对出库的提前影响时，业务经验让我们把它看成是两周来看，等等类似的指标选取都存在一些偏差。

面对这些问题让我们很清楚的认识到此模型预测结果虽然目前看来最为准确，但实际还有许多优化之处，如果想要把做的预测准确的应用于服装企业的销量判断中，我们还有很多工作要做。 FIN

Yelp，如何使用深度学习对商业照片进行分类

文 / InfoQ.com 张天雷　　编辑 / 周子赫

Yelp是美国最大点评网站，拥有世界各地的Yelper上传的成千上万的照片。各种各样的照片给进入当地的商业提供了一个丰富的窗口。通过开发一个照片理解系统使Yelp能够创建有关个人照片的语义数据。跟Yelp第一次在基于内容的照片多样化方面所做的尝试一样，由系统生成的数据正在增强Yelp近期推出的封面照片多样化、标签式照片浏览等服务。

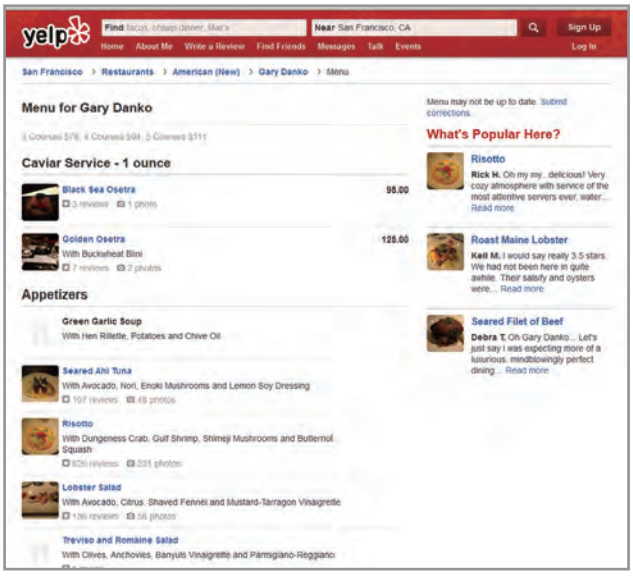
构建一个照片分类器

对于理解照片中的模棱两可的目标，其实有许多不同的方式。一开始，为了帮助简化Yelp的问题，Yelp只专注于将照片分类为几个预定义的类。之后，Yelp又只专注于关于饭店的照片类别。

事实上将照片进行分类，就可以将其当做机器学习中的分类任务，需要开发一个分类器，Yelp首先需要做的就是收集训练数据，在图片分类任务中就是收集很多标签已知的照片。Yelp收集这些信息可以通过几种不同的方式：

照片标题：在很多照片的标题中都包含代表照片自身含义的词汇，例如，很多“菜单”照片的标题中包含单词“菜单”。

为了识别这些关于食物的项目，Yelp依靠自己的菜单结构，它保留了每种食物的商业名单。Yelp发现，将列表中的食物项目与照片的标题进行匹配产生了一个高准确率的数据集。

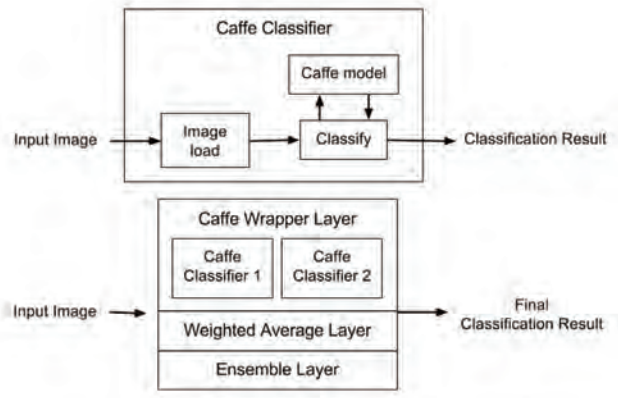


照片属性：当上传照片到Yelp上时，用户允许标记照片的一些属性，虽然它们并不总是准确的，但仍然可以很有效地帮助照片分类。

众包：通过众包可以让大众自动参与照片的标注，并同时纠正一些错误的标注。Yelp已经发现，通过众包Yelp通过合理的成本 （在时间和金钱）获得了质量总体良好的标签。众包体现了一种群体智能。

一旦Yelp有了标签数据，Yelp就开始采用“AlexNet”形式的深度卷积神经网络（CNNs）来识别这些图片（因为这种方法是一种监督学习方法，非监督学习目前仍然是深度学习的难点方向）。CNNs是由多个卷积层组成，ReLU层、pooling层、局部响应正则化层和全连接层。Yelp的CNN被建立在基于Caffe架构的AWS EC2 GPU实例上。Yelp喜欢Caffe，因为它简单易用、高性能、模块化、开源、还一直在不断完善。为了应对Caffe的软件依赖，Yelp使用Docker封装了Yelp的CNN，以便它可以更容易地部署。

Yelp还创建了抽象，以确保Yelp的CNN可以很容易地与其他形式的分类器进行集成，包括CNN的不同实例。如下图所示，Yelp的基线是一个“Caffe分类器”，它通过Caffe的方式运行CNN；它是一个抽象分类器的一种特殊形式，可以采取不同的信号，并执行不同的分类算法。Yelp目前的“facade”分类器，是一个集成分类器，采用了不同分类结果的加权平均。如果Yelp决定进一步集成依赖于其它信号的新的分类器，这将让问题变得更加简单。



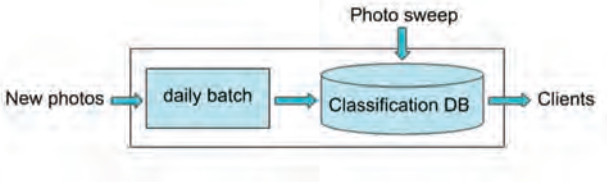
Yelp在一个均匀黄金分割的2500张照片的测试集上进行试验，Yelp目前的“facade”分类器的整体精确度达到了94%，召回率达到了70%。根据Yelp的描述，虽然这些数字绝对可以

再提高，但Yelp发现对于下面描述的应用它们已经足够了。

照片分类服务

Yelp使用面向服务的架构（SOA），Yelp做了一个RESTful照片分类服务，用来支持现有的和即将推出的Yelp的应用程序。由于服务预计拥有不止一个分类器（例如，不同的版本或为不同类型的业务），该服务API使用一个分类器ID，一个行业ID，以及可选的类，然后返回所有属于该行业的照片，其已经通过分类器被归类：

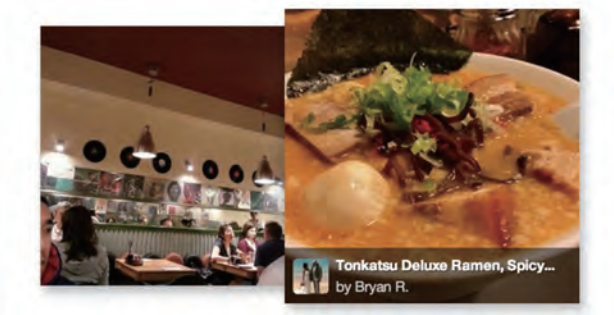
Yelp使用一个标准的MySQL数据库服务器来承载所有的分类结果，所有的服务请求可以通过简单的数据库查询被处理。为了避免更昂贵的实时分类，因为Yelp目前的应用并不取决于最新的照片分类，所以Yelp只执行线下分类。该架构如下图所示：对于每一个新的分类器，Yelp扫描所有的照片，并且将分类结果存储在一个数据库中。扫描在计算上消耗很大，但通过将分类器在任意多的机器上进行并行处理，Yelp可以减轻这一点。扫描结束后，Yelp会每天自动收集新的照片，并将它们发送到一个进行分类和数据库负载的批次中：



应用：封面照片多样化

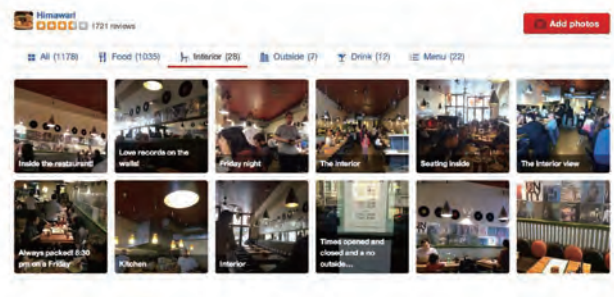
一旦有了照片分类服务，就可以有效地增强Yelp的许多关键功能。Yelp的业务详细信息页面显示了一组“封面照片”，基于用户的反馈和某些照片的属性，它们能够通过照片评分引擎进行推荐。但是，目前Yelp的封面照片存在一个典型问题，即所选的照片缺乏多样性，例如，如下图所示，所有封面照片都是关于食物的（拉面），用户无法看到其他方面的照片，除非他们点击“查看全部”按钮。

通过照片分类服务，现在就可以让封面照片变得多样化，Yelp可以容易地确定最高得分的非食品的照片，然后将其纳入封面照片。通过严格的A / B测试，Yelp已经证实饭店的浏览者更愿意看到一个显示突出的“食品”照片和突出的“非食品”照片，以及两个小“食品”的照片和另外两个“非食品”照片，如下图所示。多样化大大增加了Yelp用户与照片之间的互动。



应用：标签式浏览照片

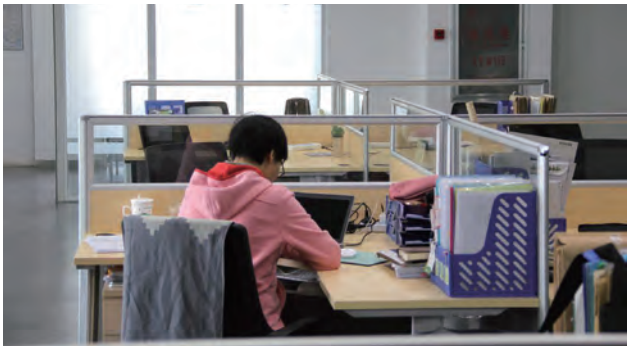
因为任何人浏览Yelp照片都是在有了解之前，大部分来自于饭店的Yelp照片都是食物。但Yelp从用户中得到反馈，他们发现用户关心的可不仅仅是食物。有些人使用Yelp的图片用来检查一个特殊事件的气氛或导航到一个第一次去的地点，而其他人使用Yelp的照片用于一些更严肃的应用，如发现餐厅是否能容纳残疾的顾客。随着标签式照片浏览的推出，所有这些任务现在都变得更容易、更高效。 Yelp表示，标签式照片浏览是他们的照片分类服务现在提供的最显著的应用。照片现在在各自的标签（类）下进行组织；从下图可以看出，跳到你正在寻找的准确信息现在变得更加容易。



下一步是什么？

任何机器学习系统都不可能是完美的。Yelp表示，如果你想帮助提高Yelp照片分类的质量，请随意标注你看到的任何未分类的照片。FIN

云南智财汇项目数据分析师事务所



云南智财汇项目数据分析师事务所（以下简称云南智财汇）是于2015年8月经昆明市西山区工商局注册，并经中国商业联合会数据分析专业委员会正式批准入会（会员资质证书编号：中数委团证第107号），具有独立法人资格的专业从事数据分析咨询服务的机构。

事务所由取得数据分析师资质的人士共同发起成立，事务所吸收融合了一批数据分析师、注册会计师、注册资产评估师、注册税务师、造价师、工程师、律师、企业管理咨询专家等专业人员组成的精英服务团队，致力于将“云南智财汇”打造成云南专业数据分析服务行业的生力军。

事务所经营服务范围：

各行业数据分析：根据客户的需求对数据通过专业的技术手段进行整理、清洗、处理、分析，建立数据分析的数学模型，为客户编制数据分析报告，为客户解读经营管理数据中深层次的信息，提供企业经营决策的战略参考方案。

各行业投资项目的评估、分析、规划、策划：编制投资项目的市场调研报告、可行性研究报告、商业计划书，投融资方案等，为客户提供专业的项目投资风险控制分析服务。

各行业企业经营管理咨询：为客户提供财务管理方案、税收策划方案、企业内控管理方案、企业ERP系统建设方案、企业管理咨询、投资咨询等服务。

经营理念：诚信、客观、公正、高效

服务宗旨：服务第一、至诚信任、精益求精、共同发展

公司地址：云南省昆明市广福路387号
云南泛亚民营金融产业园区五楼

联 系 人：饶先生

联系电话：15398551048

电子邮箱：876055297@qq.com

深圳市星盘项目数据分析师事务所

深圳市星盘项目数据分析师事务所，是深圳市第一家申请入会的专业项目数据分析事务所机构， 事务所自2015年7月开始申请入会，致力于为企业提供一体化数据分析解决方案。依托大数据分析建模平台，既能够为企业提供数据整合和数据展示，又能够进行数据建模和数据分析，同时结合企业的管理需求，梳理企业的数据流和业务流，使数据与业务紧密结合，提高企业的运营管理效率，同时提高企业的经营指标，帮助企业发现市场价值。

事务所由一批在各个领域从事多年数据分析工作的专家组成。熟悉电子商务、零售、电信、投资等领域，在相关行业拥有完整的数据分析解决方案。我们的业务涵盖数据采集、数据处理、经营数据分析、投资价值和收益分析、战略分析、数据分析培训和咨询、行业研究等多个领域。在所的项目数据分析师，全部通过数据分析专业委员会的严格考核，持有项目数据分析师（CPDA）资格证书。

目前已对该事务所进行了实地考察，通过深入沟通和现场考察，已具备事务所成立条件。

公司地址：广东省深圳市南山区高新南7路R2-A509

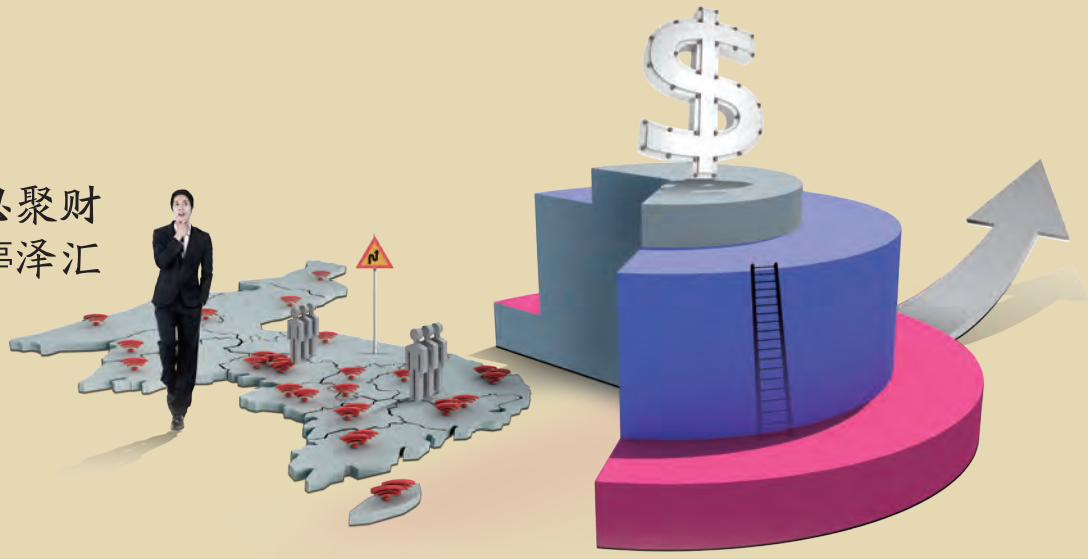
联 系 人：康华

联系电话：13316960649

电子邮箱：conniekh@foxmail.com

知行合一

智思聚财
渊渟泽汇



云南智财汇

项目数据分析师事务所有限公司

云南智财汇项目数据分析师事务所有限公司（以下简称云南智财汇）是于2015年8月经昆明市西山区工商局注册，并经中国商业联合会数据分析专业委员会正式批准入会（会员资质证书编号：中数委团证第107号），具有独立法人资格的专业从事数据分析咨询服务的机构。

公司由取得“国家工业和信息化部教育与考试中心”及“中国商业联合会数据分析专业委员会”颁发的“项目数据分析师”资质的人士共同发起成立，公司吸收融合了一批注册会计师、注册资产评估师、注册税务师、造价师、工程师、律师、企业管理咨询专家等专业人员组成精英服务团队，致力于将“云南智财汇”打造成云南专业数据分析服务行业的生力军。

公司经营服务范围：

※各行业数据分析：根据客户的需求对数据通过专业的技术手段进行整理、清洗、处理、分析，建立数据分析的数学模型，为客户编制数据分析报告，为客户解读经营管理数据中深层次的信息，提供企业经营决策的战略参考方案；

※各行业投资项目的评估、分析、规划、策划：编制投资项目的市场调研报告、可行性研究报告、商业计划书，投融资方案等，为客户提供专业的项目投资风险控制分析服务；

※各行业企业经营管理咨询：为客户提供财务管理方案、税收策划方案、企业内控管理方案、企业ERP系统建设方案、企业管理咨询、投资咨询等服务。

“云南智财汇”的执业经营理念：“诚信、客观、公正、高效”

“云南智财汇”的服务宗旨：“服务第一、至诚信任、精益求精、共同发展”

“云南智财汇”期待为您提供最大价值的服务！

公司地址：云南省昆明市广福路387号云南泛亚民营金融产业园区五楼

联系人：饶先生　　电话：15398551048　　Email：876055297@QQ.COM