第六届中国 R 语言会议（上海会场）

自由的统计语言

主办：
华东师范大学金融与统计学院

协办：
统计之都

赞助：
亿贝软件工程（上海）有限公司
联系新加坡（Contact Singapore）
檬果咨询（Mango Solutions）

2013 年 11 月 2 日 - 3 日

R 语言简介

R 是一个有着统计分析功能及强大作图功能的语言环境和软件系统，由新西兰奥克兰大学统计系的Ross Ihaka 和Robert Gentleman 共同创立。R 语言可以看作是由AT&T 贝尔实验室所创的S 语言发展出的一种方言。

R 是在 GNU 协议 General Public Licence 下免费发行的，它的开发及维护现在则由 R 开发核心小组 *R Development Core Team*具体负责，这个团队的成员大部分来自大学机构（统计及相关院系），包括牛津大学、华盛顿大学、威斯康星大学、爱荷华大学、奥克兰大学等。除了这些作者之外，R 还拥有一大批贡献者（来自哈佛大学、加州大学洛杉矶分校、麻省理工大学等），他们为R 编写代码、修正程序缺陷和撰写文档。

R 的功能很大程度上是通过程序包（Package）来实现的，迄今为止，R 语言官网上的程序包数目已经达到5000个，广泛地覆盖了数据分析应用到的各类行业和领域。各种统计前沿理论方法的相应计算机程序都会在短时间内以软件包的形式得以实现，这种速度是其它统计软件无法比拟的。

在 KDNuggets 于2013年做的“使用何种编程或统计类语言进行分析、数据挖掘及数据科学的工作”的调查中，R 以60.9%的得票率荣登榜首，力压Python、SQL和SAS（`http://www.kdnuggets.com/polls/2013/languages-analytics-data-mining-data-science.html`）。而2012年这个比例是52.5%。

目前，几乎所有的西方大学与研究机构、以及越来越多的金融机构、制药公司、高科技企业都使用 R。R 的灵活性、开放性以及业界最广泛的支持是其不断完善和发展的根本原因，随着 R 越来越被学术界及业界认可，它也将在数据分析和统计建模中发挥越来越大的作用。

华东师范大学金融与统计学院简介

华东师范大学金融学科、贸易学科、概率论与数理统计学科是国内最早培养国际金融、国际贸易、保险精算、概率论与数理统计高级人才的重要基地之一。在我国金融改革、开放、发展力度不断加大，国家和上海面临新一轮重大发展机遇的大背景下，根据国家和上海中长期发展规划及上海建设国际金融、贸易等四大中心的发展战略，为适应现代市场经济对于高素质复合型金融、贸易、保险精算和统计人才的迫切需求，华东师范大学在原有金融学科、贸易学科、概率论与数理统计学科的基础上，于2007年在国内高校中率先组建金融与统计学院，以培养既具有扎实的经济学与数理基础、又擅长专业应用的复合型人才，并希望通过结构创新和资源整合，来奠定华东师范大学在金融、贸易、概率论与数理统计和保险精算领域走向国内领军地位的基础。

学院下设五个系、一个研究院、一个研究中心和一个研究所：金融系、金融工程系、国际贸易系、风险管理与保险学系、统计与精算学系；华东师范大学应用统计科学研究院；华东师范大学人文社会科学重点研究基地—-国际金融与风险管理研究中心和华东师范大学国际金融研究所。学院现设有六个本科专业：统计学（1983）、金融学（1984）、保险（1994）、国际经济与贸易（1999）、金融工程（2009）和经济统计（2012）；五个专业学位硕士点：工商管理硕士（MBA，金融管理方向，2006）、金融硕士（2010）、国际商务硕士（2010）、保险硕士（2010）和应用统计硕士（2010）；两个一级学科博士点：应用经济学（2010）和统计学（2010）（含八个二级学科博士点和八个二级学科硕士点）；两个博士后科研流动站：应用经济学（2012）、统计学（2012）。

目前学院教职员工85人，其中教授30人，博士生导师22人，“千人计划”国家特聘教授3人，长江讲座教授1人，紫江讲座教授5人，副教授30人，有博士学位的教师63人。

学院配备专业资料室、金融实验室、统计与精算实验室。金融实验室和统计与精算实验室拥有先进的金融软件、统计软件和精算软件供教学和科研使用。中国数学类核心期刊、中国数学会概率统计学会《应用概率统计》杂志编辑部设在本学院。

在学校的支持下、在国内外同仁关心和帮助下，在学院全体师生员工的共同参与下，我们将努力把金融与统计学院建设成为文理结合、学科交融、多元发展、国内一流、国际知名、开放与互动的研究型与应用型并重的学院。

统计之都简介*

“统计之都”（Capital of Statistics，简称 COS）网站成立于 2006 年 5 月，其主旨为传播统计学知识并将其应用于实际领域。纵观现今国内统计学理论和应用的发展，一方面我们不难发现统计学在应用领域的巨大潜力——现代管理、咨询、商业、经济、金融、医药、生物等等，无不需要数据的力量，而另一方面我们也不得不承认，国内统计学的应用很大程度上受理论的制约——无论是应用界的人们对统计学基础理论知识的欠缺，还是学术界所研究的理论对应用领域问题的轻视。

“统计之都”网站便是基于这样的认识而创建的。我们希望，统计理论研究者能充分关注应用问题，而统计应用者也能正确把握统计学基本知识，将统计学这门应用学科真正的潜力开发出来。

“统计之都”为非赢利性质网站，但大力欢迎所有商界和研究领域的朋友与我们在实际应用问题上合作。我们的口号是：

中国统计学门户网站，免费统计学服务平台

我们怀着“十年磨一剑”的决心，要将“统计之都”创建成中国的统计学门户网站；我们抱着“己欲立而立人、己欲达而达人”的信条，要将“统计之都”以免费统计学服务平台的形式坚持办下去。我们希望“统计之都”在专业知识体系上有真正的王者风范，在面对用户需求时却又以谦恭的态度为大家服务。

*统计之都网址：http://cos.name/

eBay 简介

eBay 集团（纳斯达克上市公司代码：EBAY）1995 年 9 月成立于美国加州圣荷西，是全球商务与支付行业的领先者，为不同规模的商家提供公平竞争与发展的机会。eBay 集团旗下的主要业务包含全球领先的在线交易平台 eBay、在线支付工具 PayPal，以及为全球企业提供零售渠道以及数字营销便利的 eBay Enterprise。同时，eBay 集团还有其它专门的交易平台来服务数以百万的用户，其中包括全球最大的票务市场 StubHub 和 eBay classifieds 社区分类广告网站。

eBay 在中国致力于推动中国跨境电子商务的发展，为中国卖家开辟直接面向海外的销售渠道。通过 eBay 在线交易平台和 PayPal 支付解决方案，数以千计的中国企业和个人用户在 eBay 全球平台上将年销售规模达数十亿美元的产品和服务销售给世界各地的消费者。

数据分析平台与交付部 (Data and Data Infrastructure)

eBay 的分析平台与交互部门致力于开发一个企业级的数据平台，该平台以公司战略为依归，具备高度的可扩展性和可靠性。只有具备这样的数据平台，eBay 整个网上集市的各个部门才有可能运用高端的数据分析解决方案做出及时有效的业务决策。他们通过 PB 级别的数据仓库平台支持成千上万的 ETL 处理，为遍布全球的业务用户提供数据分析和商业智能解决方案。

搜索科学和流量部 (Search Science and Traffic)

搜索科学和流量部门在专注于保持一个健康的市场的基础上给用户提供他们最想购买的商品。一方面，这个团队拥有着业界的顶尖人才和先进科技，致力于通过信息检索、大数据挖掘和迭代实验来建立世界级的搜索引擎。另一方面，这个团队又致力于通过和外部搜索引擎的合作，给 eBay 带来高质量的互联网流量。

搜索科学部门在上海打造着一个世界级的团队，他们利用网络日志数据和人工智能来分析用户行为和意图，从而提高 eBay 的搜索性能和用户体验，并基于数据和用户反馈为其它产品的改进提供专业意见和解决方案。流量部门在上海专注于建立实验平台，利用数据和人工智能，帮助 eBay 更高效地获取流量。

中国分析中心 (CAC)

中国分析中心成立于 2007 年，现已发展成为 eBay 全球营运提供分析支持的核心部门。该中心的使命是为 eBay 全球市场营销及核心业务提供快速、准确且高效的分析支持。该中心具有提供全方位分析的能力，其支持范围涵盖 eBay 全球互联网在线营销分析、综合营销分析、 客户分析、投资金融分析、以及网站运营分析、物流配送优化以及智能搜索分析等领域。该中心一向致力于利用最先进、完善的分析方法和技术，通过对数据的量化解析来驱动 eBay 全球营销策略和市场运营的流程优化，从而创造价值并为 eBay 的上亿用户提供最优质的在线购物体验。

关注亿贝中国研发中心

Company Twitter : weibo.com/ebay twitter.com/ebay

Company LinkedIn profile URL : http://www.linkedin.com/company/ebay

Company website address URL: www.ebay.com www.ebay.cn

联系新加坡简介

联系新加坡是由新加坡经济发展局和人力部共同成立的联盟。我们与国际人才建立联系，并且协助他们到新加坡工作和生活。联系新加坡积极为新加坡本地雇主和专业人士牵线搭桥，为本国支柱产业的发展提供支持。

请关注微博：@联系新加坡 http://e.weibo.com/contactsingapore

同时请关注我们的官方网站：www.contactsingapore.sg

job portal：www.contactsingapore.sg/jobs

在今年的11月22-23日，新加坡总部会举办全球范围的网络视频招聘会，此次招聘会针对data analytics（数据分析）的人才，届时会有新加坡科学技术研究所，新加坡国立大学，南洋理工大学等多家知名企业和院校参与其中。欢迎大家踊跃报名，给自己开创一个崭新的职业生涯。

报名地址：www.contactsingapore.sg/Research/VCF2013

报名截止日期：2013.11.08

Mango Solutions简介

Mango Business Solutions Ltd.（简称Mango Solutions或Mango）2002年成立于英国，主要为客户提供定制化的数据分析解决方案、培训及定制的软件产品。公司总部位于英国，业务遍及美国及欧洲。10年来为各行各业的公司、政府部门与学术机构成功地提供了咨询服务，客户中既包含辉瑞制药、诺华、谷歌、索尼这样家喻户晓的跨国企业，也有许多本地的中小型企业与机构。

Mango的团队由建模师，统计分析师与软件工程师组成，既可以为客户解决数学、统计、建模方面的特定问题，如提供数据分析方案、设计与执行数据分析IT平台的搭建，也可以为客户定制开发与数据分析相关的软件工具，也可以提供相关工具的培训，如R, S-PLUS等。

Mango Solutions的中文网站: www.mango-solutions.com.cn

英文网站：www.mango-solutions.com

新浪微博：weibo.com/mangosolutions

第六届中国 R 语言会议（上海会场）

会议指南

1. 日程安排

11 月 2 日	注册和报告	华东师大中山北路校区，科学会堂
11 月 3 日	报告和讨论	华东师大中山北路校区，科学会堂

2. 会议议程

演讲	时间
11 月 2 日	
上午（地点：科学会堂二楼报告厅）	
会议开幕	09:00 ~ 09:10
Chih-Jen Lin：Experiences and Lessons in Developing Machine Learning and Data Mining Software	09:10 ~ 09:50
林祯舜：R 与大数据对统计教育的影响	09:50 ~ 10:15
拍照、茶歇	
李忠：EBAY Multiscreen Insight	10:35 ~ 11:00
赵扬：Large Data Analysis using Rhipe/Rhadoop	11:00 ~ 11:25
朱筠珺：How R helps personalization analysis in marketing campaign	11:25 ~ 11:50
午餐时间	
下午（地点：科学会堂二楼报告厅）	
Lightning Talks eBay、携程、Mango Solutions、SupStat、点融网、北京数衡科技有限公司、北京华章图文信息有限公司、浙江大学软件学院、统计之都、中国统计网、数盟社区等公司和单位	13:30 ~ 14:30
甘华来：R 语言在旅游行业中的应用	14:30 ~ 14:50
严紫丹：x13季节调整方法的 R 实现及应用	14:50 ~ 15:10
茶歇	
魏太云，周扬：通向高富帅图表的 R 包——recharts	15:30 ~ 16:00
朱雪宁：微博那些事儿	16:00 ~ 16:25
何通：豆瓣网标签的整理和分析	16:25 ~ 16:50
第一天会议结束	

演讲	时间
11 月 3 日	
上午（地点：科学会堂二楼报告厅）	
刘思喆：R 语言企业级应用	09:00 ~ 09:25
孙哲：小分队撬动大估值——关于零售金融服务行业的大数据应用模式探讨	09:25 ~ 09:50
许亮：复杂交易网络中的白富美挖掘	09:50 ~ 10:15
茶歇	
李洪成：用 R 进行高频金融数据分析简介	10:35 ~ 10:55
邓一硕：quantstrat 包与 R 中的量化投资之路	10:55 ~ 11:15
专题讨论：R 语言与企业应用 （主持人 林祯舜）	11:15 ~ 12:00
午餐时间	
下午（地点：科学会堂二楼报告厅）	
罗立辉：R 语言在陆面建模系统上的应用研究	13:30 ~ 14:00
李舰：中文文本挖掘和 tmcn 包	14:00 ~ 14:30
陈逸波：kaggle 数据比赛的一些经验分享——以 Amazon Employee Access Challenge 为例	14:30 ~ 15:00
刘辰昂：use R for fun	15:00 ~ 15:30
会议结束	

3. 会议机构

主办单位：

华东师范大学金融与统计学院

协办单位：

统计之都（http://cos.name/）

赞助单位：

亿贝软件工程（上海）有限公司
联系新加坡 （Contact Singapore）
檬果咨询 （Mango Solutions）

会议主席：

汤银才 李舰 沈羽

会议详情请参见：http://cos.name/2013/10/2013-china-r-conference-shanghai-notice/

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 9 页

Experiences and Lessons in Developing Machine Learning and Data Mining Software

Chih-Jen Lin[1,*]

[1]

Abstract

Traditionally academic machine learning and data mining researchers focus on proposing new algorithms. The task of implementing these algorithms is often left to companies that are developing software packages. Recently, the rise of open-source software development has lead many researchers to start creating software packages.

However, without a careful design and implementation, the resulting package may not be widely used. In this talk, we discuss the experiences in developing two machine learning packages LIBSVM and LIBLINEAR, which have been popular in both academia and industry. We demonstrate that the interaction with users leads us to identify some important research problems. For example, the decision to study and then support multi-class SVM was essential in the early stage of developing LIBSVM. The birth of LIBLINEAR was driven by the need to classify large-scale documents in Internet companies. For fast training of large-scale problems, we had to create new algorithms other than those used in LIBSVM for kernel SVM. We present some practical use of LIBLINEAR for Internet applications. Finally, we give lessons learned and future perspectives for developing industry-strength machine learning and data mining software.

[*]Chih-Jen Lin is currently a distinguished professor at the Department of Computer Science, National Taiwan University. He obtained his B.S. degree from National Taiwan University in 1993 and Ph.D. degree from University of Michigan in 1998. His major research areas include machine learning, data mining, and numerical optimization. He is best known for his work on support vector machines (SVM) for data classification. His software LIBSVM is one of the most widely used and cited SVM packages. For his research work he has received many awards, including the ACM KDD 2010 best paper award. He is an IEEE fellow and an ACM distinguished scientist for his contribution to machine learning algorithms and software design. More information about him can be found at `http://www.csie.ntu.edu.tw/~cjlin`.

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 10 页

R与大数据对统计教育的影响

林祯舜[1,*]

[1]

摘要

在十多年前，“数据挖掘”这个时髦的名词出现之后，许多统计学家开始思考统计科学的课程体系（请参考2000年中华数据挖掘协会成立大会的研讨会各个专家的发言及台湾辅仁大学统计系更名为统计信息系的缘由），同时有更多的统计学家开始担忧由计算机科学家开发的数据挖掘算法及工具对传统统计数据分析方法及理论的冲击。然而十多年过去了。现今，大数据这个名词又非常的火，数据科学已经成为一个交叉学科的领域，而近十年 R 的快速发展对传统的统计教育会产生哪些冲击?统计学家要如何在这个浪潮下调整思路，培养有领导力的未来数据科学家?这个报告从顶层设计和自下而上的两个方向，阐明大数据和 R 语言对统计教育的影响，希望能给高校及企业提供一个清晰的大数据人才培养思路。

*林祯舜博士是数据科学及营销科学方面的专家，毕业于人民大学统计学院并获得博士学位，在企业界，目前担任信息技术咨询公司的总经理，在学术界，目前是兰州商学院及中山大学数学学院的兼职教授。林博士学术领域的研究方向包括数据挖掘，机器学习，统计计算，网站效果测量与点击流数据分析。

EBAY Multiscreen Insight

Zhong Li

[1] eBay Data and Data Infrastructure

Abstract

From smartphones and tablets to laptops and television, 90% of all media interactions today are screen-based.So Multiscreen is becoming more and more important for e-commerce, therefore in this presentation, I willintroduce multiscreen basic facts firstly, and then share the interesting EBAY multiscreen case study whichintegrated R and machine learning technology to you.

Big data analysis using Rhipe/RHadoop

Yang Zhao

[2] eBay Data and Data Infrastructure

Abstract

1. Introduce the basic structure of R+Hadoop platform.
2. Major advantages and drawbacks about using Rhipe(R and Hadoop Integrated Programming Environment), when dealing with large data set.
3. Two specific user cases about data analysis using Rhipe? One is a simple map reduce job; the other one is a complex project about DNS traffic analysis.
4. Talk about how to learn R+Hadoop more efficiently(Might be the best way to learn R+Hadoop).

How R helps personalization analysis in marketing campaign

Junjun Zhu

[3] eBay CAC

Abstract

eBay isn't really much of an auction site anymore, with the "Buy It Now" option available on site, more and more big brands tend to sell new merchandise products with discount on eBay sites, which means marketing campaign are critical for these sale events to be successfully turning a visitor into a real buyer.

So this presentation will mainly focus on the personalization analysis we made for better targeting customer in a certain marketing campaign, as well as how we use R to help this analysis accomplished.

R语言在旅游行业中的应用

甘华来

[1] 携程商业智能部

摘要

携程是国内最大的在线旅行商（OTA），拥有丰富的用户行为数据和订单数据，将简单介绍携程的大数据挖掘，同时以携程Noshow订单预测项目为例，介绍如何利用R使用GBM模型对订单进行预测。

x13季节调整方法的R实现及应用

严紫丹

[1] 携程商业智能部

摘要

国家统计局自11年起开始发布GDP季度环比数据和固定资产投资、社会消费品零售总额等月度环比数据，计算环比数据的关键技术是对数据进行季节调整，国家统计局使用的季节调整方法主要是在美国普查局x12方法的基础上，增加了中国的节假日因素。

美国普查局去年7月正式发布了x-13-arima-seat方法的源程序，这一方法是对x12的优化升级。目前这一方法很少能在普通软件中实现，本文将介绍如何用R调用美国普查局的源程序进行季节调整和大批量数据处理的流程，并介绍针对中国节假日因素的参数设定方法。

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 13 页

通向高富帅图表的R包——recharts

魏太云[1,*]　　周扬[2,†]

[1]凯普质量研究院
[2]Mango Solutions

摘要

大数据时代，重新定义数据图表的时候到了：ECharts基于Canvas，纯Javascript图表库，提供直观，生动，可交互，可个性化定制的数据可视化图表。创新的拖拽重计算、数据视图、值域漫游等特性大大增强了用户体验，赋予了用户对数据进行挖掘、整合的能力。GitHub主页：`http://ecomfe.github.io/echarts/`

周扬和魏太云一起编写了R下的接口包recharts，以方便广大R用户使用。

*电子邮件：weitaiyun@gmail.com；微博：http://weibo.com/taiyun/
†微博http://weibo.com/zhouyummy

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 14 页

微博那些事儿

朱雪宁[1,*]

[1] 北京大学

摘要

微博，这一新生代大规模杀伤性社交武器近年来迅速在国内走红，其来势之汹，范围之广，威力之猛当不可小觑。该演讲者将主要介绍用R分析微博数据的方法及建模点滴。

*北大光华商务统计2013级硕士

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 15 页

豆瓣网标签的整理和分析

何通[1,*]

摘要

豆瓣网有众多的书、影、音条目，更有众多用户为它们打上了个性化的标签。这些标签内容丰富，但同时也有着噪声大、文本短小等特点。演讲者以在豆瓣实习时的工作与大家分享对标签信息的整理与分析结果，希望能够引出更多有益的探索。

*豆瓣算法组实习生

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 16 页

R语言企业级应用

刘思喆[1,*]

京东商城

摘要

1. R语言企业级应用的架构
2. 常用的技术方案介绍
3. 分享案例

*京东商城个性化推荐组负责人

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 17 页

小分队撬动大估值——关于零售金融服务行业的大数据应用模式探讨

孙哲

摘要

1 大数据不是新概念但出现了新趋势

（1）关于上世纪80年、90年代美国金融服务行业的大数据应用的趋势回顾。

（2）从数据垄断到优势扩散，关于2008年前后数据采集及数据应用格局出现的转变。

2 国内某银行大数据应用典型案例介绍

（1）基于数据的资产证券化技术；

（2）以最大化价值为基础的大数据经营体系；

（3）基于互联网技术的新客户获取探索。

3 金融服务+大数据=巨大的估值想象空间

（1）关于Capital One等信息公司成长逻辑的回顾；

（2）关于金融服务行业中大数据本质的理解

（3）关于大数据与当前国内金融服务行业的机遇；

（4）关于互联网公司进入金融服务行业现象的探讨。

4 关于以小型团队撬动企业估值的一种可行性探讨

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 18 页

复杂交易网络中的白富美挖掘

许亮[1,*]

[1] 天猫

摘要

复杂网络理论

交易网络中的生态群体

交易网络中的白富美买家发掘

交易网络中的高品质卖家挖掘

*天猫数据挖掘专家，曾就职于微软亚洲研究院，腾讯，阿里金融等公司，主要从事电商交易平台上的复杂人际网络研究和优质买卖家挖掘。

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 19 页

用R进行高频金融数据分析简介

李洪成[1,*]

[1] 上海金融学院

摘要

R的高频添加包给出了大量的分析高频金融数据的工具，包括管理、清理和匹配高频交易和报价数据的许多函数。应用该包提供的工具函数，可以计算各种流动性指标、波动率等，同时也可以探测噪声的微观结构。本文简单介绍了该添加包，并给出应用示例。

*上海金融学院教师。主要研究兴趣为统计学和数据挖掘。著作主要有《SPSS数据分析教程》等；译著有《数据挖掘与R语言》、《R经典实例》和《金融数据分析导论:基于R语言》等R相关书籍。

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 20 页

quantstrat包与R中的量化投资之路

邓一硕[1,*]

[1]SupStat

摘要

quantstrat 是一个由Peter Carl、Brian 等联合开发的用于量化投资的R包。该包为用户提供了一个测试和模拟基于信号的量化策略模型的泛型框架。基于它我们不仅可以构建交易系统，还可以对构建的交易系统进行仿真测试，它能够支持对多资产类别和多币种组合进行回测（backtesting）等。目前，这个包还在开发中，不过已经被很多业内人士所使用。由于该包尚没有完备的帮助文档可供查询，因此，本演讲着重以实例来介绍quantstrat 包的使用方法。

[*]SupStat 合伙人、统计之都沙龙理事

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 21 页

R语言在陆面建模系统上的应用研究

罗立辉[1]

[1] 中国科学院寒区旱区环境与工程研究所

摘要

陆面过程是指发生在地表, 控制地气之间水分、热量和动量交换的作用过程，它涉及了地球系统五大圈(大气圈、岩石圈、生物圈、水圈和冰冻圈)中几乎所有的圈层。陆面作为全球气候系统的重要组成部分，在气候变化中的作用显著。考虑到大多数陆面过程模型的强迫数据集来源于气象站和涡度相关仪器，其输入和输出数据格式为NetCDF，所以在陆面过程模型的前处理（pre-processing）和后处理（post-processing）采用R语言来进行数据制备、质量控制、统计分析与可视化。在模型模拟的前处理，R用于制备数据、质量控制和空值填补，数据制备加载了RNetCDF包，采用相邻7个值的方法对数据进行了质量控制，而且对每个变量都设置了旗标（flag）来限制其值域的范围。模型模拟的后处理中加载了R语言的ncdf、gdata、gplots、plotrix、Hmisc、lattice和泰勒R软件包，然后在此基础上构建了陆面过程模型模拟后处理的R语言功能函数，其中集成了15种不同数学统计分析方法来评估陆面过程模型，然后将可视化以pdf、eps、emf、ps等格式文件输出。陆面过程模型由于对计算和存储需求高，一般在高性能计算环境中进行模拟计算，而具有大型计算能力的机构和单位一般要求使用SSH等进行登录，然后采用作业调度系统（如Platform LSF）提交计算作业，且针对NetCDF处理分析的软件目前较为成熟，为了填补R语言在这方面的不足，又采用了NCO、CDO、GrADS和NCL脚本语言，结合R语言来实现陆面过程模型的从其前处理、模拟运行到后处理的自动化过程。

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 22 页

中文文本挖掘和tmcn包

李舰[1,*]

[1]Mango Solutions

摘要

R 在进行文本分析和挖掘方面有很多很好的包，比如 tm，但是这些包对中文的支持不是很好，而且基于一些复杂面向对象的程序开发，不是很适合于对 R 不熟练的用户。

演讲者开发了一个 tmcn 包，专门用来进行中文的文本挖掘，包含了一些中文编码处理和字符处理的函数，与作者开发的另一个做中文分词的 R 包 Rwordseg 配合使用，可以很方便地进行常用的中文文本分析比如文本相似度、文本分类、主题模型等。此外还引入了一些新的分析方法的库，比如条件随机场 CRF++ 以及 Google 的 word2vec。

*电子邮件：lijian.pku@gmail.com

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 23 页

kaggle数据比赛的一些经验分享——以Amazon Employee Access Challenge为例

陈逸波

[1] 统计之都

摘要

以一场kaggle比赛为例，介绍数据挖掘/有监督学习过程中的数据处理、模型训练、模型集成及效果评估等内容。

第六届中国 R 语言会议手册，上海， 2013 年 11 月 2 日 – 3 日
主办：华东师范大学金融与统计学院；第 24 页

use R for fun

刘辰昂[1,*]

[1]浙江大学

摘要

R除了在学术上助我们一臂之力外，同样也可以在生活中带给我们很多快乐。神奇的魔术不再是魔术师的专利，自己动手也能开发出好玩的小游戏，绚丽的图案美妙的音乐也都可以信手拈来……不怕做不到就怕想不到，在寻找快乐的同时这也不失为正经之余修炼内功享受生活的一个好途径。

*电子邮件：liuchenang@gmail.com；个人主页：chenangliu.info/cn/；微博:http://weibo.com/liuchenang

Job Vacancies

Apply in www.ebaycareers.com

CAC-Business Analyst/Sr. Business Analyst-88945BR/88790BR/89666BR

- 2+ years experience in analyzing multi-dimensional datasets, using SQL, SAS, Teradata, etc.
- Good communication skills both in English & Chinese
- Sound business judgment and quantitative analytic ability

CAC-Manager, Business Analytics-88239BR

- 7+ years experience in analyzing large, multi-dimensional datasets (SQL) and extracting insights from diversified data/results
- Sound business judgment and quantitative analytic ability

CAC-Sr. CRM Manager, Business Analytics-88390BR

- Lead ad-hoc projects, collaborate with other analytics functions and effectively communicate results to analytics partners and cross functional groups
- Capability and experience in leading a team of analysts to achieve their potential is also the key.
- Partner with global team to scope out action items to ensure the team is on track to deliver towards the company's CRM goal

ebaycareers.com

Paypal-Business System Analyst-89199BR

- 7+ years of experience with Bachelor degree
- For data specialist background candidate: must have data-warehouse skills in one of followings : data modeling, ETL, or BI; For business analyst background candidate: must have strong skills in large dataset manipulations for financial analysis, data mining or statistical modeling;
- Must have good understandings of Dimensional Modeling
- Hands-on experiences with SAS or Hadoop are strong plus
- Teradata SQL and BTEQ – Plus point

Follow Our Weibo
@eBayTech

Paypal-Product Manager-89197BR

- 8+ years of experience in data warehousing or big data domain
- Must have good understandings of Dimensional Modeling
- Hands-on experiences with SAS or Hadoop are strong plus
- Teradata SQL and BTEQ – Plus point
- Extensive experience in leading and developing teams of analysts and engineers in data technology organization
- Hands on development experience in enterprise data warehousing and big data area

Job Vacancies

Apply in *www.ebaycareers.com*

Search Science-Data Platform Development Manager-79523BR

This team is responsible for build the data platform to support whole traffic team include each channels as well as applied scientist to optimize traffic performance. In short, the team will deal with Big data, complex distribute system and challenges all over the world. We are looking for someone have hands on coding experience with management experience.

Search Science-Data Platform Development Engineer-85708BR

Build traffic data platform which will be on top of various data sources, 5+ years' work experience in software development area with distribute system/big data experience

DDI–Big Data Platform Architect-87690BR

- Experience on building analytics from unstructured data on Hadoop platform is preferred
- Extensive experience of Java development.
- Experience with large data warehouse environments (preferably Teradata and/or Oracle)

DDI–Manager, Business Analytics-86117BR

- Proven records of the data science and data product practical experience is required.
- PHD degree in Computer Science, Statistics, Mathematics or equivalent academic record is required.
- Hands on experience of applying various modeling algorithms, data mining tools, data analysis tools, and statistical packages.

DDI–Software Development Manager (Hadoop)-87689BR

- Excellent understanding of computer science fundamentals, data structures, and algorithms.
- Experience on building analytics from unstructured data on Hadoop platform
- Familiar with search/recommendation/classification applications and domains

MMP-Payment Manager-85355BR

We are the team building the Payments Platform for eBay. Our mission is to create an easy to use, innovative, organized and unified payment experience on eBay, we are providing from front-end UI to backend streamlined money movement services for integrating with other eBay domains. We need candidate have 8+ years' work experience in software development area with 3+ years' managing software development teams

ebaycareers.com

Follow Our Weibo
@eBayTech

Research Virtual Career Fair 2013

If you are looking to make a professional leap into Asia to advance your career, you should join us for our **Research Virtual Career Fair (VCF) on 22 – 23 November**. We are looking for candidates in **Data Analytics and Materials & Manufacturing Technology**. Through the VCF platform, you will have a chance to interact with employers directly via **live video and text chats**, right from the comfort of wherever you are!

Singapore Time (GMT +8)
Date:22-23 November 2013
Time: 9.00am - 9.00am

PARTICIPATING EMPLOYERS

Materials & Manufacturing

- **Advanced Remanufacturing and Technology Center (ARTC)**
- **Institute of Materials Research and Engineering (IMRE)**
- **Nanyang Technological University- Faculty of Materials Science and Engineering (NTU-MSE)**
- **Singapore Institute of Manufacturing**

HOT JOBS

Contact Singapore invites you to embark on a rewarding career in Singapore's Biomedical Sciences. Click on the job titles below to know more about each position.

Alternatively, visit our job portal www.contactsingapore.sg/jobs for other career opportunities in Biomedical Sciences.

Experimental Power Grid Centre

- Research Fellow- Electrical and Electronics Engineering
- Research Scientist- Condition Monitoring
- Research Scientist- Power System Protection

Technology (SIMTech)

Data Analytics

- **Institute of Infocomm Technology (I2R)**

And more!

Register your interest now at:
www.contactsingapore.sg/Research/VCF2013

Read about these employers and get a sneak-preview of the talent they are looking for in the Virtual Career Fair below!

Track 1: Materials and Manufacturing Technology

Advanced Remanufacturing and Technology Centre

The Advanced Remanufacturing and Technology Centre (ARTC) is a new initiative by the Agency for Science, Technology & Research (A*STAR), in collaboration with Nanyang Technological University (NTU). The ARTC initiative is based on a contemporary model of public-private partnership to transform remanufacturing and manufacturing research into ready solutions for cross-sectoral industry members by bringing together best-in-class global OEMs, state-of-the-art equipment makers and R&D minds in Singapore.

The ARTC is now entering an exciting phase of development and we are looking for people with a passion for innovation and the ability to turn ideas into practical solutions for industry members within three core technical areas of Repair & Restoration, Surface Enhancement and Product Verification.

1. Technical Director
2. Principal/Senior/Development Scientists
3. Senior Project Managers

Data Storage Institute

- Scientist- Data Center Technologies
- Scientist- Security and Applied Cryptography

Institute of Chemical and Engineering Sciences

- Scientist- Chemical Engineering
- Scientist- Organic Chemistry

Institute of High Performance Computing

- Scientist – Integrated City Planning

Institute of Microelectronics

- Scientist – Wafer Level Packaging for MEMS device
- Scientist- RF MEMS based on A|N
- Scientist- GaN-on-Si Substrate and Power Electronics Device

Mango Solutions 精于数据分析及数据相关的服务，为客户提供精心定制的培训、咨询、与软件，协助客户挖掘复杂统计分析的能量，来支持关键性的商业决策。我们的总部位于英国，在美国、瑞士、中国设有分支机构。

Mango 掌握数据分析行业的前沿技术，客户涵盖制药、金融、互联网、农业、食品等多种行业，积累了宝贵的技术经验与资源，竭力提供高质量的客户体验。

公司业务

- 数据分析方案咨询
- 统计编程
- R 语言技术支持
- 数据分析外包
- R 语言培训
- 软件开发

产品

Navigator
基于 NONMEM 的数据处理及报告的系统。用户可以上传他们的 NONMEM 模型，在分布式的网格平台上运行这些模型，并对运行的结果生成自动化的图表报告。更方便的是，用户可以使用 Navigator 把这些图表报告自动插入到他们的 Word 文档中，或把它们单独下载为 Word 或 PDF 格式。

ModSpace
Mango 开发的 ModSpace 系统为企业与组织的知识共享与管理提供了新的方案。ModSpace 是一个集知识共享、搜索、与版本控制于一体 的新型知识管理软件。ModSpace 最初被制药公司用于他们的 M&S 团队，用于药物研发中的统计模型的共享与管理，但经过产品化的开发， 它也可适用于多种通用类型的环境，比如公司内部文档的管理与共享。

iNCAS
可以为初次临床研究确定安全并有效的药物剂量提供帮助。通过与 GSK 全球安全委员会的紧密合作，Mango 开发出了一个能够集中储存实验数据，并能够以统一的各式展示的应用程序。通过这个应用程序，研究人员可以使用一个简单、统一的图形界面来对研究数据进行分析，以此来进行预测建模并出安全有效的药物剂量。

客户

Google LLOYD'S NOVARTIS bp Allianz Roche

服务

金融与保险
金融、保险公司与机构极度依赖基于高质量数据做出的快速反应。Mango 帮助金融与保险客户完成各种统计算法及应用程序。我们的服务包括小至为 Excel 的使用提供便利的咨询服务，到大型企业级软件的开发。

制药
新药研发是一项需要大量时间与劳动的工作。一种新药从研发到投放市场可能要经历二十年的时间、耗资几十亿美元。Mango 为制药与生物科技公司提供相关的服务与技术，可以帮助他们减少新药研发所需的时间。我们所提供的帮助包括软件系统、咨询、培训等。

能源
能源公司使用多种不同的统计技术来流程化他们的工作活动，以提高效率，增加投资回报。从管道分析，到能源交易方案，Mango 使统计学的力量为能源公司所用。

医学图像分析
Mango Imaging 为 MRI（Magnetic Resonance Imaging ）与 PET（Positron Emission Tomography）研究提供全面的量化图像分析与统计推理服务。我们为不同的形态及各类的实验建立了不同的数据分析管道，并应用紧密跟踪的 质量监控手段。

食品与感官分析
感官分析由应用统计分析方法与实验设计组成。消费品测试的结果会被逐条记录，随后据此找出有意义的数据，并做出推论。Mango 开发了 R 与 S PLUS 的应用程序，使研究人员能够通过简单的菜单点击就可以实现复杂、可定制的统计分析。

客户关系管理
Mango 可以利用统计分析技术帮助市场推广组织与部门使他们的市场活动更加有效与经济。我们使用复杂的统计算法来识别最有利润的目标客户，以此来节省大量的时间与人力。

其他行业
统计与数据分析可以应用在众多行业。我们在互联网、快速消费品、农业研究等众多领域都有诸多案例。如果您觉得我们可能对您的业务有所帮助，请尽快与我们联系。我们非常乐意与您探讨可能的解决方案。

R 语言培训与咨询
我们是世界上最大、最专业的 R 语言服务提供商。我们为各种公司与机构提供 R 语言培训与技术咨询、支持。为企业定制的培训课程受到众多客户的好评。

英国（总部）
地址：2 Methuen Park, Chippenham, Wiltshire, UK
电话：+44 - (0)1249 - 705 450
电子邮件：info@mango-solutions.com

中国（分公司）
地址：上海徐汇区肇嘉浜路 1065 号飞雕国际大厦 1607C
电话：+86 - (0)21 - 5178 1325
电子邮件：fshao@mango-solutions.com