CHINA-R

第六屆中國R語言會議

癸巳年五月十八日

高朋滿座勝友如云

統
CAPITAL OF STATISTICS
PROFESSION, HUMANITY & INTEGRITY

目录

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的统计及相关院系。除了这些作者之外，R 还拥有一大批贡献者，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超过 4000 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 2012 年做的“使用何种编程或统计类语言进行分析和数据挖掘”的调查中，R 以 52.5% 的得票率荣登榜首，力压 Python、SQL、JAVA 和 SAS（`http://t.cn/zTexTiL`）。此外 R 还击败了 Excel 和 Rapidminer（2010 和 2011 年排名第一），在“过去十二个月中你在实际项目中使用的数据挖掘或分析工具”的调查中排名第一。

Rexer Analytics 5th 数据挖掘者调查报告指出：R 语言一直保持上升的势头，牢牢占据工具类的第一名。几乎有一半的调查对象（47%）声称使用 R 语言作为数据挖掘工具。（`http://t.cn/zWKsvEZ`）。

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。

中国人民大学应用统计研究中心简介

中国人民大学应用统计科学研究中心前身是成立于 1988 年的统计科学研究所。十几年来，中心积极培育中青年学术骨干，逐渐发展并形成了经济与社会统计、统计调查与数据分析和风险管理与精算三个各具特色的研究方向。几年来，中心建设的重点研究平台是：1) 统计理论与建模方法和应用研究；2) 满意度统计理论、方法和应用研究；3) 国际竞争力理论方法及其应用研究；4) 数据挖掘技术中的统计理论、方法与应用研究；5) 改进我国政府统计数据质量及其抽样调查制度的理论方法研究；6) 统计在社会科学中的应用研究；7) 风险管理与保险精算应用研究；8) 六西格玛管理应用研究；9) 环境经济核算理论方法和应用研究。此外，中心本着创建和发展面向实际应用的研究中心的宗旨，创建了：竞争力与评价研究中心；数据挖掘中心；六西格玛质量管理研究中心；保险精算中心；统计资讯研究中心等子机构，在突出应用主题的研究中心下，本着联系实际和服务实际的思想，创建了面向实际应用的网站，建立新型的学术交流、知识普及和与用户零距离连接的模式。随着我国经济体制的进一步改革，中心积极适应市场经济的需要，面向全国开放，加强国际学术交流与合作，推动重大应用统计项目的研究。

中心现有专兼职研究人员 29 人，学术委员会委员 19 人，其中既有统计科学领域国内外著名的学术带头人，如中科院院士严加安教授、陈木法教授、彭实戈教授；又有一批全国知名学者和业务骨干，如袁卫教授、吴喜之教授、耿直教授、赵彦云教授和原国家统计局局长谢伏瞻研究员等。中心研究队伍强大的教育背景、研究成果和学术声誉将使本中心成为全国一流并具有国际声誉和影响的开放式应用统计研究机构。

中国人民大学统计学院简介

中国人民大学统计学科始建于 1950 年，两年后成立统计学系，是新中国经济学科中最早设立的统计学系， 2003 年 7 月，成立中国人民大学统计学院。多年来，本学科一直强调统计理论和统计应用的结合，不断拓宽统计教学和研究领域，成为统计学全国重点学科。教育部人文社会科学重点研究基地“应用统计科学研究中心”也设在统计学院。学院拥有统计学和风险管理与精算学两个博士点，统计学、概率论与数理统计、风险管理与精算学、流行病与卫生统计学四个硕士点，应用经济学下设统计学博士后流动站。

统计学院现有教师 33 人，其中教授 14 人，副教授 11 人，博士生导师 13 人。国内兼职教授 11 名，海外客座教授 10 人。 50 多年来，共培养不同层次人才 5000 多人。2008 年 9 月，在校学生总人数为 523 人，其中本科生 305 人，硕士生 142 人，博士生 76 人，大多数毕业生在金融、保险、证券、基金、信息等领域从事数据采集和分析工作。

北京大学商务智能研究中心简介

北京大学商务智能研究中心是一个面向互联网大数据的科研平台。中心尤其关注具备以下三种特征的互联网大数据：（1）中文文本数据；（2）网络结构数据；（3）地理位置数据。为此，中心依托北京大学光华管理，联合众多互联网企业，以及科研机构。中心现有合作学者十余人，横跨统计学、营销、管理科学、计算机等众多学科。合作学者来自海内外知名大学。例如：北京大学、人民大学、中央财经、四川大学、西安欧亚学院、俄亥俄州立大学等。中心现有合作企业多个，覆盖互联网众多细分行业。例如：博雅立方，百度，新浪等。中心在研项目包含但不局限于：（1）基于微博数据理解企业竞争态势；（2）基于搜索 Cookie 数据了解消费者特征；（3）基于 URL 首页文本数据萃取企业行业特征；（4）基于社交媒体评论热度，理解股市走向等。中心诚挚邀请有共同兴趣的学术机构、企业伙伴、以及个人相互切磋，共同学习，互相提高！

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”的建设需要您的帮助，如果您从统计之都网站中获取了有价值的帮助，可以登录我们的捐赠页面 (`http://cos.name/donate/`) 对我们的工作予以支持。

统计之都活动回顾

统计之都（下文简称 COS）虽以网站和论坛起家繁荣，但是随着越来越多喜爱统计的朋友们加入，大家对于线下活动和书稿撰写翻译等等的需求也越来越旺。目前，COS 的线下活动从一年两次的 R 会议（春季北京、秋季上海），逐渐发展到沙龙、交流会、竞赛、讲座、培训等等。我们希望更多的新鲜血液可以就近加入 COS 的线下活动中。COS 线下活动总结：

1. 中国 R 语言会议：目前已开展到第五届，分别在中国人民大学（北京）和华东师范大学、上海财经大学（上海）举行，一般为期两天。历届会议纪要和幻灯片共享都可以在 COS 主站上找到：
   - 第五届中国 R 语言会议：http://cos.name/chinar/chinar-2012 /
   - 第四届中国 R 语言会议：http://cos.name/chinar/chinar-2011/
   - 第三届中国 R 语言会议：http://cos.name/chinar/chinar-2010/
   - 第二届中国 R 语言会议：http://cos.name/chinar/chinar-2009/
   - 第一届中国 R 语言会议：http://cos.name/chinar/chinar-2008/
2. 北上广（深）三地沙龙：目前我们定期在北京、上海和广州深圳开展线下沙龙活动。不同于规模庞大的 R 语言会议，沙龙形式更为轻巧，注重讨论交流。北京的沙龙已经进行了 7 期，上海的沙龙进行了 7 期，而广深沙龙只有 1 期。
   - 参与方式：每期沙龙开始前将在微博和论坛上进行通知，同时我们将建立参与过沙龙的朋友们的邮件列表。
   - 新浪微博：@ 统计之都　或者 `http://weibo.com/cosname`
   - COS 论坛：`http://cos.name/cn/`
   - 沙龙邮箱：`salon@cos.name`
3. 比赛、交流会、讲座：除了定期的活动之外，我们还将不定期举行或者和其他组织共同举办一些交流会和讲座。北京的交流会主要以中国人民大学为中心，而上海的则更偏重于沟通合作。目前，我们主办或协办过的活动包括：
   - 数据挖掘竞赛
   - 科普讲座：与科学松鼠会合作，分别在北京、上海、杭州联合举办过统计科普讲座——别让数字吓到你

- 校内交流会：不定期在学校内举办统计建模、统计实际应用、统计学出国等经验交流会，包括中国人民大学和南开大学等。
- 培训：我们亦致力于 R 语言的普及，已有培训包括：(1). 七周七语言：与 Top Geek 合作开展七周七语言系列的 R 语言培训，由资深 useR 李舰主讲。(2). Supstat 夏令营：2012 年 9 月，我们第一次尝试开展了为时两天的 Supstat 夏令营活动，由二十多名 useR 相互交流讨论，系统而深入的讨论了 R 语言的方方面面。

4. 书籍出版，包括撰写和翻译：

- **撰写：** 吴喜之《复杂数据统计方法 —基于 R 的应用》(已经由中国人民大学出版社出版)，谢益辉《现代统计图形》，谢益辉、肖楠等《R 忍者秘籍》，陈丽云《Play Econometrics with R》，李舰、肖凯《数据科学中的 R 语言》，魏太云、肖楠《knitr 与动态报告》；
- **翻译：** R in Action(《R 语言实战》，已经由人民邮电出版社出版)、ggplot 2: Element Graphics for Data Analysis (《ggplot2：数据分析与图形艺术》已经由西安交大出版社出版)、The Art of R programming(即将由机械工业出版社出版)，R in a Nutshell(即将由电子工业出版社出版)，R Graphics Cookbook(即将由人民邮电出版社出版)。

第六届中国 R 语言会议北京会场日程

5 月 18 日 人大国学馆 114・113 不见不散		
演讲及活动		**时间**
陈昱、吴喜之	会议主席致开幕辞	08:55-09:10
赵彦云	统计学院院长致欢迎辞	09:10-09:20
谢益辉	我的第 8 个 R 包	09:20-09:50
Graham Williams	Data Mining with Rattle and R	09:50-10:20
合影・茶歇		
谢邦昌、刘思喆	DataMining 云端决策平台 CDMS Smart Score II——以 R 为基础	11:00-11:30
John Maindonald	Rethinking Data Analysis and Data Analysis Tools	11:30-12:00
午休		
李舰、周扬	禽流感分析中的 R——MSToolkit, Rweibo, html5vis 的介绍	14:00-14:30
张晓华	displayHTS: a R package for displaying data and results from high-throughput screening experiments	14:30-15:00
Lightning Talk：杏树林、Mango、泰山投资、Merck、万达信息、Careerfocus、中信银行		15:00-15:35
抽奖		
张常有	Julia 语言与并行计算	16:00-16:30
肖楠	Web Scraping with R	16:30-17:00

5 月 19 日 人大国学馆 114・113 不见不散		
演讲及活动		**时间**
庄宝童	机器学习在互联网广告中的应用	09:00-09:30
李忠、潘佳鸣	R 在 eBay 大数据分析中的应用	09:30-10:00
Lightning Talk：阿里巴巴、Amazon、eBay、Spinger、京东、百度、豆瓣		10:00-10:35
茶歇		
稳国柱	R 的工程实践和 Data Scientist	11:00-11:30
王浩	用户产生内容的质量评价与智能排序	11:30-12:00
午休		
王汉生	On the ultrahigh dimensional linear discriminant analysis problem with a diverging number of classes	14:00-14:30
周庭锐	移动应用里的线上行为：一个 R 的尝试	14:30-15:00
李欣海	用 R 和 WinBUGS 实现贝叶斯分级模型	15:00-15:30
抽奖		
王贺	网络舆情监测——基于 R 语言的网络文本挖掘与数据可视化	16:00-16:30
关菁菁	Data cloning: easy maximum likelihood estimation for complex models: an application to zero-inflated responses of Internet ads	16:30-17:00

注：Lightning Talk 演讲次序随机，顺序由现场抽签决定。

中国人民大学地图

注：图上标记的餐厅都可以现金消费。

我的第 8 个 R 包

谢益辉*
Iowa State University

摘要

2007 年我向 CRAN 提交了第一个 R 包 `animation`，此后陆续提交了 `formatR`，`Rd2roxygen`，`R2SWF`，`MSG`，`iBUGS`，`fun`和 `knitr`。本次演讲我想谈谈这五年开发 R 包过程中的一些观察和感想，主要包括：（1）“好玩”是最强的生产力（2）“不推销”是最好的推销策略（3）最好的搭档是会用 C 的人，但要做最炫的自己还得学 JavaScript（4）命名是最难的任务（5）“好看”本身就是最好的“功能”（6）需求可以源于小处（参见朱重八发家史）（7）编程的社交化是必然趋势。

关键字： 开发 R 包；感想；knitr

*邮箱：xie@iastate.edu

DATA MINING 雲端決策平台 CDMS Smart Score II ——以 R 为基础

谢邦昌 刘思喆*
台湾辅仁大学 京东商城

摘要

随着 Google, Amazon, IBM, Microsoft, Yahoo 等知名业者跟进云端服务后，云端运算几乎成了网络服务的代名词，它将是未来几年 IT 的必然趋势。云端运算是将庞大运算操作拆成千百个较小的操作，再交给远端的多台服务器同时运算。CDMS Smart Score II 将云端运算整合至 data mining 的范畴之中，进而形成云端智慧并提供云端服务。它克服了其他数据挖掘软件的缺点，降低成本并提高了企业的获利。

关键字： 关键词：云端运算；CDMS Smart Score II；R；iSmartScore

*谢邦昌邮箱：025674@mail.fju.edu.tw；刘思喆邮箱：sunbjt@gmail.com

Data Mining with Rattle and R

Graham Williams[*]
Australian Taxation Office

Abstract

Data Mining and Analytics have become fundamental tools for any organisation today. The amount of data being collected, and importantly analysed, is growing rapidly, and businesses gain competitive advantage through the knowledge they gain from data. Google, Amazon, eBay/Paypal, and GroupOn are all prime examples of the importance of the analysis of data. R directly supports or provides access to all of the tools and technology required for data mining. In this presentation we will discover how to quickly get started with data mining in R using the Rattle graphical user interface. The basic algorithms for data mining will be introduced, particularly decision trees and random forests, and then demonstrated through Rattle. We will then illustrate how to migrate from using the graphical user interface into using R directly, through the scripts captured from Rattle.

Keywords: data mining; Rattle; decision tree; random forest

About Lecturer

Professor Graham Williams is a Chinese Academy of Science Senior International Expert and Visiting Professor, hosted at the Shenzhen Institutes of Advanced Computing, Chinese Academy of Sciences. He is also the Director and Senior Data Miner for the Australian Taxation Office, and holds adjunct professorships with the Australian National University and the University of Canberra. He is an active machine learning researcher and regularly teaches data mining courses. Graham is author of the Rattle software for data mining and of the Rattle book published in 2011: "Data Mining with Rattle and R: The Art of Excavating Knowledge from Data". Graham has been involved in data mining projects for clients from government and industry over 25 years. His research contributions include the development of ensemble learning and hot spots discovery. He is involved in numerous international artificial intelligence and data mining research activities and

[*]Email: Graham.Williams@togaware.com

conferences and has edited a number of books and has authored many academic and industry papers. He is chair of the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD), and the Australasian Conference on Data Mining (AusDM).

Rethinking Data Analysis and Data Analysis Tools

John Maindonald[*]
Australian National University

Abstract

The abilities of R and other such software have greatly enlarged the kitbag of tools that are available for handling of practical statistical analysis. The changes that are needed to take advantage of these new tools, affecting research and teaching as well as statistical practice, have proceeded slowly.

Additionally, experience with the new tools feeds back into a demand for even better tools. Thus some aspects of R and of R packages still reflect, to too large an extent, approaches from the pre-computer era. This is even more true for the textbooks, even for the book that I wrote with John Braun!

One large set of changes arises from the ability to handle large simulation and/or bootstrap calculations. These calculations can take us places where statistical theory is, on its own, unable to go. The view should be that simulation is the basic tool for determining sampling distributions. In certain fortunate circumstances, theoretical mathematical results make the recourse to simulation unnecessary.

I will comment on implications for handling regression. Simulation abilities that supplement plot.lm() were recently added to the DAAG package. Simulation can provide an important reality check, also, when there has been some modest amount of model or variable selection. A further set of issues has to do with the extent to which statistical learning approaches that make an automatic choice of smoothing parameters may give misleading results when used with data where there is, for example, spatial dependence. Here, resampling methods can be a useful recourse. These examples, while important, are in no way exhaustive.

Keywords: practical statistical analysis; statistical tools; large simulation; regression; resampling methods

About Lecturer

[*]Email: jhmaindonald@gmail.com

Dr John Maindonald is a Visiting Fellow in the Centre for Mathematics and Its Applications, at the Australian National University. His research interests include Statistical Computation, Statistical Perspectives on Data Mining, Use of the R system for Practical Data Analysis, and Research Planning. He has had wide experience in practical data analysis, in areas ranging from historical wage rate statistics to molecular biology. He is the lead author of ‘Data Analysis and Graphics Using R: An Example-Based Approach’, now in its third edition

禽流感数据分析中 R 的应用

李舰 周扬*
Mango Solutions

摘要

本次演讲介绍了演讲者及 Mango Solutions 公司原创的 `MSToolkit`、`Rweibo`、`Rwordseg`、`tmcn`、`googleVis.Mango`等 R 包，以近期出现的禽流感的传播为例，使用 R 进行了一个完整的分析流程。在这个例子中，使用 MSToolkit 来进行临床数据的模拟，使用 Rweibo 抓取微博中的文本信息，使用 `Rwordseg`进行中文分词，使用 `tmcn`进行中文相关操作，使用 `googleVis.Mango`进行动态的可视化展现。同时也会介绍这几个包的开发经验和相关资源。

关键字： MSToolkit; Rweibo; Rwordseg; tmcn; googleVis.Mango; 禽流感案例

*李舰邮箱：`jian.li@188.com`；周扬邮箱：`zhouyanga9@gmail.com`

displayHTS: a R package for displaying data and results from high-throughput screening experiments

Xiaohua Zhang[*]
Merck

Abstract

The R package `displayHTS` implements recently developed methods and figures for displaying data and hit selection results in high-throughput screening (HTS) experiments. It generates not only certain useful distinctive graphics such as the plate-well series plot, plate image and dual-flashlight plot but also other commonly used figures such as volcano plot and plate correlation plot. These figures are critical for visualizing the data and displaying important features of HTS data and hit selection results.

Keywords: displayHTS; high-throughput; screening

About Lecturer

Dr. Zhang is the head of Early Development Statistics – Asian Pacific in Merck Research Laboratories. He got his Ph.D. in statistics from Carnegie-Mellon University, PA, USA. He has worked on data analysis for genome-wide RNAi research and microarrays in drug discovery and development for various diseases for many years. He and his colleagues have continuously developed and published novel analytic methods and experimental designs for quality control and hit selection in genome-scale RNAi research in peer-reviewed journals such as Bioinformatics, Nucleic Acids Research, Cell Host&Microbe, et al. His book titled "Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-scale RNAi Research" was published by Cambridge University Press, Cambridge, UK in 2011. He has membership in professional competition committee, serves in the Editorial Board in multiple peer-reviewed journals, has served as a referee in many peer-reviewed journals, and has been invited to be an expert evaluator for grant proposals in both the European Union 7th Framework Programme and USA NIH Study Section. He has also been invited to give presentations in many international conferences and seminars in many research universities in both USA and China.

[*]Email: `xiaohua_zhang@merck.com`.

Julia 语言与并行计算

张常有*
中国科学院软件研究所

摘要

Julia 是一个新的高性能动态高级编程语言，提供了精度和分布式并行运行方式，高效支持外部函数的调用。第一部分主要介绍 Julia 语言的基本语法，以理解 Julia 语言的基本程序结构。第二部分介绍制作和调用 C 语言函数库的方法，方便构建和利用现有的高性能函数库适应更广泛的计算平台。第三部分介绍 Julia 语言对并行计算的支持能力，通过程序案例体验其问题求解的提速效果。最后，介绍面向 Julia 的协作云服务平台和 OpenBlas 算法库方面的工作进展。

关键字： Julia 语言；外部函数；并行计算；云服务

*邮箱：changyou@iscas.ac.cn

Web Scraping with R

肖楠*
Central South University

摘要

The web itself is the world's largest, public-accessible data source. Knowing how to scrape data from the web has become one must-have skill, particularly for data hackers. In this report, you will learn the basic coding strategies and neat tricks for web scraping with R. While introducing how to retrieve data from the web and parse a variety of data formats, we will summarize the usage and application scenarios of several useful R packages. At last but not least, this report emphasizes the suitable exception handling and parallelization methods, which is crucial for the construction of a robust and high performance web scraper with R.

关键字： R; web scraping; web crawling

*邮箱：road2stat@gmail.com

机器学习在互联网广告中的应用

庄宝童*
一淘

摘要

自从互联网广告诞生以来，机器学习对广告收益的增长贡献越来越大。无论是google/百度为代表的关键词广告，还是 yahoo 为代表的品牌广告而言，都有 2-3 倍以上的效果提升。机器学习在帮助这些公司取得可观收益的同时，也给广告主带来相对传统广告更高的 ROI，给用户带来更好的体验。在实现这些目标的过程中，涌现出一些有趣的机器学习问题，有相对比较成熟的点击率/转化率/跳出率预估问题，也有尚在探索中的实时出价预估问题，用户满意度/用户疲劳问题，广告创意质量审核等。涉及机器学习的方方面面，包括分类、回归、增强学习等，非常值得研究，这里将会对这些问题进行引介和讨论。

关键字： 机器学习；互联网广告

*邮箱：zhuangbao@gmail.com

R 在 eBay 大数据分析中的应用

潘佳鸣 李忠*
eBay

摘要

随着大数据时代的来临，如何有效地利用和分析大数据成为各行各业关注的焦点，本文将演示两个如何利用 R 来分析 eBay 大数据的应用案例。第一个案例演示了如何对 eBay 的移动用户购买行为进行深度分析，从三个不同的视角 (地理分布，性别，年龄段) 分析了用户数量，订单量，购买频率，购买金额，购买类别和购后评价，最后演示了如何分析和展现用户的保持率。第二个案例演示了如何对 eBay 的系统错误日志进行分类，来更好的帮助后台支持人员解决问题，并且让公司管理层对整个运营系统的错误类别有全面的掌控。

关键字： eBay；大数据；移动用户购买行为；错误日志分类

*潘佳鸣邮箱：jipan@ebay.com；李忠邮箱：zholi@ebay.com.

R 的工程实践和 Data Scientist

稳国柱*
豆瓣

摘要

我们的工程环境中，使用 R 的场景通常有两类：大规模数据的处理，精细化的数据分析/挖掘工作。对于前者，我会介绍这些年我们用 R 对大数据量处理的实践心得，以及正在搭建的并行计算的框架性工作。对于后者，我会介绍 Data Scientist 这种豆瓣的新员工角色，以及他们每天会面临的挑战。

关键字： 并行计算；大数据；Data Scientist

个人介绍

阿稳是豆瓣的算法工程师，在近五年的互联网数据挖掘生涯里，做过日志分析、推荐系统、用户行为分析、BI 系统搭建等脏活累活，其中 R 也作为一个工具参与其中，所以有那么一点实践的心得。

*邮箱：GuozhuWen@douban.com

基于机器学习的互联网内容质量评价与智能排序

王浩*
宝宝树信息技术有限公司

摘要

线上社交网络中的海量用户产生的内容 UGC，其信息量与价值参差不齐，如何对 UGC 的质量进行评价并智能排序，是非常有意思的工作。已有的内容质量评价方法，主要包括利用众包的投票机制（如 Digg 和 Reddit），以及基于用户间互动的社交关系（如 Facebook 的 Newsfeed 和新浪微博智能排序）。这些方法，主要借助于用户消费内容后的自然反馈来觉得内容质量。在我们的工作中，我们借助于机器学习技术，基于 UGC 的文字内容本身，对其进行质量评估，并针对具体的 UGC 展示场景进行智能排序。

关键字： UGC；评价；智能排序；机器学习

*邮箱：lexwhu@163.com

On the ultrahigh dimensional linear discriminant analysis problem with a diverging number of classes

王汉生*
Peking University

摘要

This paper is concerned with the problem of variable screening for Fisher's linear discriminant analysis with a diverging number of classes and an ultrahigh dimensional predictor. In the presence of a diverging number of classes, the total number of relevant features may go to infinity at a rate faster than usual. This makes the related statistical problem much more challenging than the conventional one with a fixed number of classes. To solve the problem, we propose here a novel pairwise sure independence screening method for the linear discriminant analysis with an ultrahigh dimensional predictor. The proposed procedure is directly applicable for the situation with a finite number of classes and with a diverging number of classes. We further prove that the proposed method enjoys the strong screening consistency property. Simulation studies are conducted to assess?the finite sample performance of the proposed procedure. We also demonstrate the proposed methodology via an empirical analysis of a real-life example on hand-written Chinese character recognition.

关键字： diverging number of classes; novel pairwise; sure independence screening method; Simulation studies;

*邮箱：hansheng@gsm.pku.edu.cn

移动应用里的线上行为：一个 R 的尝试

周庭锐*
中国人民大学商学院

摘要

国内某卫视集团自今年元月份起，率先开始实验电视节目与移动终端双屏结合的商业模式，由元月一日起截止至三月底，已经拥有超过 300 万注册用户，同时上线并发的最高历史峰值曾经达到每分钟 100 万人次以上，为了面对这样的高并发数据接入，数据库采用 MongoDB，因此必须先进行文本数据结构的转换，然后才能进行统计运算。本报告针对这样一个数据集，利用 R 进行一些初步探索，和大家分享一部分探索结果，并讨论分析过程所遭遇的问题。

由于数据结构同时包含结构的定量数据与非结构的文本数据，且导出的数据是.json 与.bson 的格式，必须转换为 R 可以读取的文本格式，然后导入 R 进行运算。R 里面有 RMongo 包，不过这次还来不及进行实验。我们使用正则表达式对 MongoDB 的数据进行解析，转换为.csv 格式，使用 `read.table()` 或 `scan()`读入 R。所取得数据主要包括：访问记录、注册用户数据、用户行为数据、社会网络里的文本数据。为了满足数据之间的可比性，我们仅提取截止至元宵节为止的数据进行分析，总数据量大约 30G。因为数据量比较巨大，使用了一台自己组装的 16 核（模拟为 32 核）、64G 内存的 Linux 机器进行计算，多线程脚本采用 snowfall 包编写，全部分析时间约一个星期，所遭遇最主要困难为内存不足。

分析主要包括下列几个部分：（一）利用访问记录，观察了这个移动应用的用户来源，所采用的技术是解析国外某些提供 IP 地址坐标转换网站里的 java 反馈信息，直接将用户访问 IP 转换为地理经纬度坐标，很惊讶地发现，这个应用才刚放入苹果商店与安卓市场，用户的地理分布居然已经遍及全球。我们同时使用 `maptools` 包读入中国地图 `bou2_4p.shp`，观察不同类型用户在时间轴上的地理分布；（二）解析用户的行为记录，进行时间轴的行为串联。首先计算相同特定用户时间轴上发生行为的时间差，利用聚类分析，计算出最适当的切割时间，用以切割用户同一次使用手机应用的浏览向量，然后利用这个数据计算用户的行为特征；（三）利用前述数据计算用户流失率与转化率；（四）通过前述过程发现不活跃用户与僵尸用户之间的关联，并通过适当计算，确认人为灌水的僵尸来源；（五）通过文本资料进行用户之间的互动内容的语义分析；（六）研究用户之间社交网络的结构。

*邮箱：tingjui.chou@gmail.com

本报告除了介绍所观察到的移动应用线上行为外，将讨论在分析过程所遭遇的两个最主要的运算问题：(1) 内存不足，以及 (2) 多线程脚本返回列表的汇总处理方法，期待抛砖引玉，向与会朋友学习更好的解决方案。

关键字： 高并发数据；移动应用线上行为；snowfall；maptools；聚类分析

用 R 和 WinBUGS 实现贝叶斯分级模型

李欣海*
中国科学院动物研究所

摘要

分级模型（Hierarchical modeling）是把不同的描述随机过程的模型整合在一起的模型体系。例如一个物种在地点 i 的数量与地点 i 的植被、海拔和温度等变量有关，可以用波松回归描述物种数量同环境变量的关系。同时，该物种每个个体被人发现的概率与调查时间、调查人的经验等变量有关，可以用逻辑斯蒂回归描述发现率和其解释变量的关系。在地点 i 调查到的该物种的数量是波松回归和逻辑斯蒂回归的集合。分级模型可以是很多模型（多于两个）的集合。分级模型的思想起源于上世纪 90 年代，在近十年中有较大的发展，逐渐成为描述物种分布的主流方法。贝叶斯方法通过 MCMC 估计每个模型的参数，是当前分级模型参数估计的主要方法。研究者一般用 R 整理数据，然后通过 R2WinBUGS 包调用 WinBUGS 进行参数估计和模型选择。最后利用 MCMC 算出的参数在 R 中进行模型验证。

本研究以朱鹮在陕西汉中地区 95 个流域的营巢数为因变量，分析每个流域的环境变量和野外调查对营巢数的影响。结果显示流域内的稻田面积、水体（河流和池塘）的面积以及人类活动对营巢数有显著影响。其中稻田面积和水体面积的交互作用是贡献最大的因素，说明朱鹮同时需要一定面积的稻田（繁殖期的觅食地）和水体（非繁殖期的觅食地）。

关键字： 分级模型；贝叶斯方法；MCMC；R2WinBUGS 包

*邮箱：lixh@ioz.ac.cn

网络舆情监测——基于 R 语言的网络文本挖掘与数据可视化

王贺*
中国人民大学

摘要

作为当今信息传播的重要载体，互联网为各行业的分析者提供了丰富的素材。其中海量的数据、丰富的媒介、多样的平台、复杂的联系与迅速的更新也向传统的数据分析者提出了新的要求。笔者借助 R 语言，尝试批量获取网络新闻、微博内容与网上商城的用户评价等具有代表性的网络文本数据，在此基础上进行分析与可视化输出。结构如下：

1. 文本挖掘在网络数据中有广阔的使用前景
2. R 语言中相关程序包简介
   - `Rurl`：下载数据
   - `XML`：解析网页结构，生成 XML 格式的文件供后续分析
   - `Rwordseg`：中文分词
   - `tm`：构建词频 -文档矩阵
   - `Rweibo`：接入新浪微博，获取所需数据
   - `lda`：LDA（潜在狄利克雷分配模型），一个非监督的浅层语义分析模型
3. 实际应用
   - 单个网页的例子：使用 Rurl，tm 包
   - 网络新闻：借助 Google 高级搜索，结合 R 语言，批量获取网页新闻；初步的描述统计分析
   - 微博：使用 Rweibo 接入新浪微博，搜索并提取包含关键词的微博；建立词与词之间的关系，生成 XML 格式文件；使用社会网络分析软件 Gephi 绘制关系图
   - 产品评论：获取金士顿某款 U 盘在亚马逊网站上的 5290 条产品评论；建立词与词之间的关系并绘制关系图；使用 LDA 模型对评论进行浅层语义分析，得到评论主题

关键字： 网络文本挖掘；电商；微博；可视化；R 语言；XML 结构；LDA 主题模型

*邮箱：wang6uem@163.com

Data cloning: easy maximum likelihood estimation for complex models: an application to zero-inflated responses of Internet ads

关菁菁*
香港城市大学

摘要

Objective: Predicting click-through behaviors of customers is crucial to resource/information allocation of online marketing websites. Zero responses frequently generate due to limited attention from customers and numerous information from suppliers, such as advertisers. Traditional statistical models such as logistic regression fail when zero responses are excess. I propose to adapt a Zero-Inflated Poisson (ZIP) regression for customer click-through behavior prediction. Complex statistical models which captures time tendencies, individual differences and multilevel data structure is proposed.

Methodology: Data Cloning, first proposed in 2007, is a simple computational Bayesian method for complex statistical models, especially for generalized linear mixed models. Data Cloning leads to priordistribution invariant maximum likelihood estimators (MLE) and a simple estimator of asymptotic variance of MLE. R is flexible for implementing Data Cloning algorithm. The dclone R package contains low level functions for implementing Data Cloning algorithm with support for JAGS, WinBUGS and OpenBUGS. It is also convenient and practical to code Data Cloning algorithm in R.

关键字： Data Cloning; dclone R Package; Zero-Inflated Poisson Regression; Multilevel Model; Maximum Likelihood Estimation

*邮箱：jjguan@cityu.edu.hk

主办：

中国人民大学统计学院
应用统计科学研究中心
北京大学商务智能研究中心
统计之都