

北京邮电大学
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

Management in Big Data Era

马宝君 博士 讲师

经济管理学院
电子商务中心
2015年1月5日



1

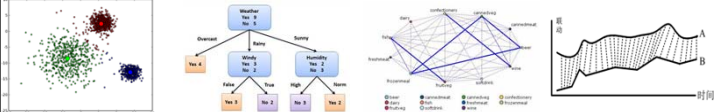
课程回顾1：深度业务分析——原方法

- 聚类 (Clustering)
- 分类 (Classification)
- 关联 (Association)
- 模式 (Pattern)
-

类别

联系

轨迹



课程回顾2：深度业务分析——组合方法及应用

- 信息检索及信息搜索服务 (文本内容、链接)
- 推荐系统及产品推荐
- 情感分析及舆情监测
- 社交网络分析及关系营销
- 用户生成内容 (口碑/评论/社交) 分析
-




以推荐
挖掘并满足用户的潜在需求

amazon.com

PANDORA

NETFLIX

last.fm

StumbleUpon

Discover your web

3

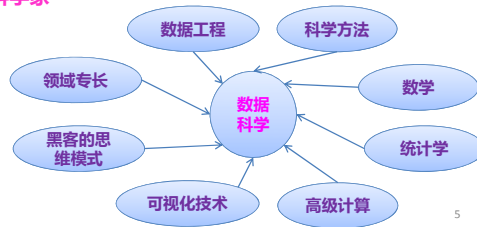
大数据的学科：数据科学 ——影响及应用




4

数据科学简介

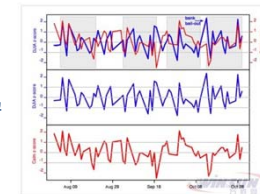
- Data Science
- 1960年由彼得·诺尔提出，当时是作为**计算机科学**的一个分支
- 涉及多方面的内容，涵盖数学、统计学、数据工程、模式识别、机器学习、高性能计算、可视化、数据仓库以及数据建模等多个领域的技术和理论
- 最终目标为从数据中挖掘出有用的信息，让数据增值
- 培养的是：**数据科学家**



5

大数据在商业金融领域的应用

- “啤酒与尿布”
 - 商品交叉销售：关联分析、购物篮分析（美式、日式）
- 比价网站的成功：Forecast, Decide.com
 - 价格趋势及预测：相关关系分析、预测模型
- 基于大数据的个性化推荐系统
 - 亚马逊云：自动推荐、人工推荐、内部竞争机制
 - 基于基因的推荐系统：潘多拉（网络电台）
- Target的大数据营销
 - 孕期用品推荐及销售：女性购买行为在怀孕期间产生变化的模型
- 社交网络数据之于对冲基金
 - Twitter的“平静指数”预测股市



6

大数据在生物医学领域的应用

- 流行病预测
 - 谷歌的流感预测：比官方提前1-2周
 - 利用微博来预测流感、流行疾病
- 大数据与智慧医疗
 - 临床操作、付款定价、研发、新的商业模式、公众健康
- 疾病监控
 - 服务心脏病患者、“魔毯”病人的监控、监测脑外伤病人恢复
- 可穿戴技术、大数据与智慧医疗
 - 生命体征测量T恤、Health Tech产品、Google Glass, iWatch



7

大数据在智慧城市领域的应用

- 智慧城市（Smart City）
 - 新一代信息技术支撑、知识社会下一代创新（创新2.0）环境下的城市形态
- 基本特征
 - 全面透彻的感知、宽带泛在的互联、智能融合的应用、以人为本的可持续创新
 - 智慧的经济、智慧的运输业、智慧的环境、智慧的居民、智慧的管理
- 国际实践
 - 韩国的松岛新城、美国的哥伦布市、英国的智能屋试点、智慧日本、爱沙尼亚的塔林市、荷兰阿姆斯特丹的智慧城市建设计划、巴西里约热内卢的智慧城市建设.....



8

大数据在影视娱乐领域的应用

● 大数据捧红《纸牌屋》

- Netflix海量的用户数据积累和分析
- 从受众洞察、受众定位、受众接触到受众转化，每一步都由精准细致、高效经济的数据引导，从而实现大众创造的C2B，即由用户需求决定生产



● 谷歌预测电影票房

- Google网页+Youtube搜索量
- 预测好莱坞新电影首映第一个周末的票房（94%准确率）



● 利用数据预测奥斯卡奖项

- 微软亚洲研究院



大数据在其他领域的应用

● 大数据帮助奥巴马赢得大选

- “微观智能”、“微竞争”
- 数据由人创造，反映人的行为和心理



● 棱镜门

- 美国“棱镜计划”：2007年开始
- 加拿大的“棱镜门”：2005年开始
- 隐私安全、伦理忧思



● 大数据帮助寻根问祖

- 家谱网：Ancestry.com
- 庞大的基于家庭关系的资料数据库(4PB)



● 大数据与社会治安

大数据分析的常用工具

● WEKA

● R与RStudio

● Gephi

● NLPIR / ICTCLASS

● 其他

- SPSS
- SAS
- MATLAB
- CLUTO

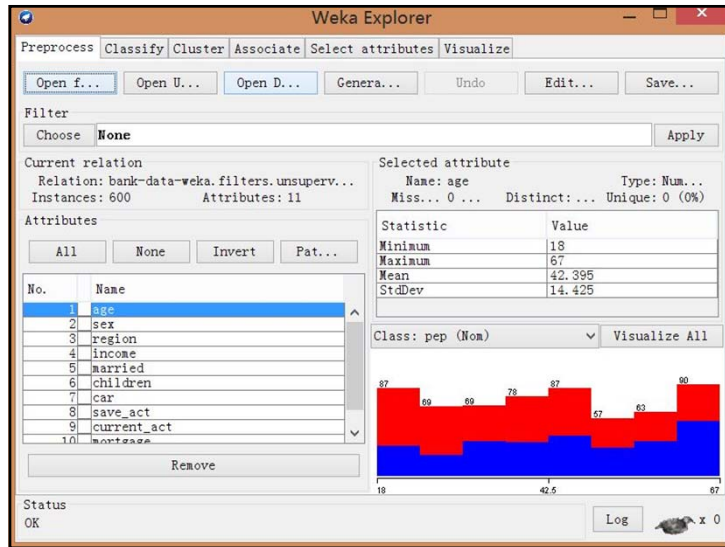


11

大数据分析的常用工具简介 ——Weka



12



Weka实际演示

● 参见课堂演示

- 数据格式转换 (csv => arff)
- 数据预处理 (删除属性、改变属性类型、离散化等)
- 关联规则 (Apriori)
- 分类 (C4.5 : J48)
- 聚类 (K-means : SimpleKMeans)

● 参考资料 :

- weka.ppt
- WEKA使用教程: <http://blog.csdn.net/yangliuy/article/details/7589306>

18

大数据分析的常用工具简介 ——R与RStudio



19

R与RStudio简介

- R是一套完整的**数据处理、计算和制图**软件系统。其功能包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言：可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能
- R语言是**统计领域**广泛使用的，诞生于1980年左右的S语言的一个分支
- R是一个免费的自由软件，它有UNIX、LINUX、MacOS和WINDOWS版本，都是可以免费下载和使用的
- RStudio是R语言的一种集成开发环境，它是免费自由软件

● 软件下载

- R : <http://www.r-project.org/>
- Rstudio : <http://www.rstudio.com/products/RStudio/>

20

R与RStudio简介

● 网络课程

- MOOC : <http://mooc.guokr.com/course/831/R-Programming/>

● 参考资料

● Book

- Software for Data Analysis: Programming with R (Statistics and Computing) by John M. Chambers (Springer), 2008 (数据分析软件: R语言编程)
- R语言统计入门(第2版), Peter, Dalgaard 著; 郝智恒, 何通, 邓一硕, 刘旭华 译, 人民邮电出版社, 2014

● Manual

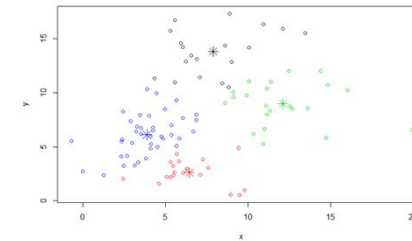
- An Introduction to R (pdf), R Data Import/Export (pdf)
- <http://www.ibm.com/developerworks/cn/linux/l-r1/>
- <http://www.ibm.com/developerworks/cn/linux/l-r2/index.html>
- <http://www.ibm.com/developerworks/cn/linux/l-r3.html>



21

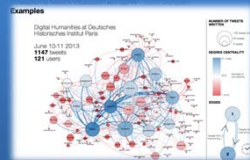
R软件实际演示

- `x <- rbind(matrix(rnorm(100, mean = 5, sd = 2), ncol = 2),`
- `matrix(rnorm(100, mean = 10, sd = 4), ncol = 2))`
- `colnames(x) <- c("x", "y")`
- `cl <- kmeans(x, 4)`
- `plot(x, col = cl$cluster)`
- `points(cl$centers, col = 1:4, pch = 8, cex = 2)`



22

大数据分析的常用工具简介 ——Gephi



23

Gephi简介



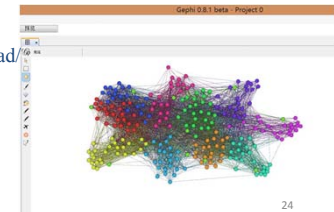
- Gephi是一款开源免费跨平台基于JVM的**复杂网络分析软件**。其主要用于各种网络和复杂系统，动态和分层图的交互可视化与探测开源工具。可用作：探索性数据分析，链接分析，社交网络分析，生物网络分析等
- Gephi不大能够处理大规模数据集并生成漂亮的可视化图形，还能对数据进行**清洗和分类**，其使用也相对比较复杂

● 软件下载

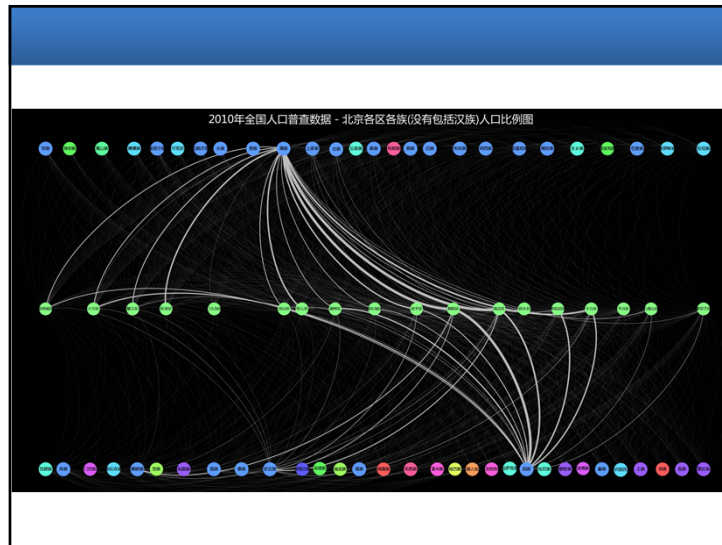
- <https://gephi.github.io/users/download/>

● MOOC中文课程

- <http://www.udemy.com/gephi>



24



大数据分析的常用工具简介 ——NLPIR / ICTCLASS



NLPIR 汉语分词系统
又名: ICTCLAS 2015



26

NLPIR汉语分词系统 / ICTCLASS简介

- NLPIR汉语分词系统(又名ICTCLAS2015), 主要功能包括**中文分词**; **词性标注**; **命名实体识别**; **用户词典功能**; 支持GBK编码、UTF8编码、BIG5编码。新增**微博分词**、**新词发现与关键词提取**; 北京理工大学张华平博士先后倾力打造十余年, 内核升级10次。
- 全球用户突破20万, 先后获得了2010年钱伟长中文信息处理科学技术奖一等奖, 2003年国际SIGHAN分词大赛综合第一名, 2002年国内973评测综合第一名。
- 软件下载:
 - <http://ictclas.nlpir.org/>



27

NLPIR大数据搜索与挖掘共享开发平台

- NLPIR文本搜索与挖掘开发平台针对互联网内容处理的需要, 融合了自然语言理解、网络搜索和文本挖掘的技术, 提供了用于技术二次开发的基础工具集

12项主要功能



软件下载

- <http://www.lingjoin.com/cn/download/downloads.html>

28

其他一些常用数据挖掘、数据分析工具

- SPSS
 - “统计产品与服务解决方案”
- SAS
 - “统计分析系统”
 - SAS系统在国际上已被誉为统计分析的标准软件，在各个领域得到广泛应用
- MATLAB
 - 美国MathWorks公司出品的商业数学软件，用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境
- CLUTO
 - 一款专门用于聚类的工具包，有多种不同情景下的聚类方法实现
 - <http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download>

29

《大数据时代的管理》

课程总结



30

《大数据时代的管理》课程教学内容

- 本课程围绕大数据（Big Data）时代**新兴IT应用特征和经济管理挑战**，了解和分析数字化社会和网络经济活动中的**新商务模式**，以及企业和管理者驾驭数据、把握竞争优势的新课题等**管理问题**。



- 本课程的主要内容包括：

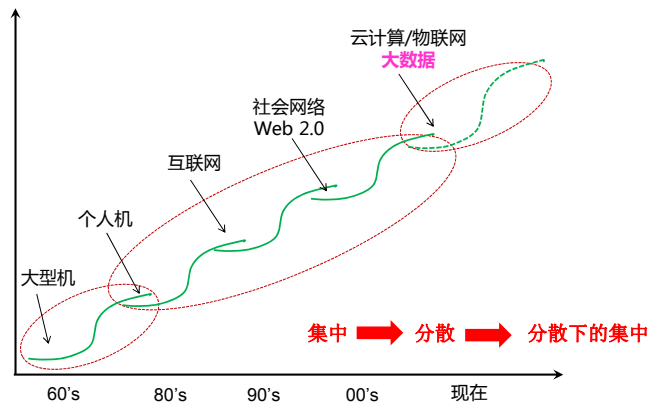
- （1）新兴技术融合带来的**新的社会经济变革**，如：第三次工业革命、技术时代沿革、数字化生活等；（C1-4）
- （2）大数据时代的**管理喻意**，如：新模式-新人群-新业态等；（C5-6）
- （3）企业大数据的管理与分析，如：数据商务、**深度商务分析**（Business Analytics）等；（C7-16）
- （4）大数据时代的若干实践与管理场景。

教学进度安排

上课周次	课程内容主题	备注
1 (09/22)	课程简介与引言	
2 (09/28)	引言与大数据概念	国庆假期调课（周日）
3 (09/29)	大数据特征与变革	
4 (10/13)	大数据时代的思维变革	
5 (10/20)	大数据时代的管理挑战之“三个融合”	
6 (10/27)	大数据时代的管理挑战之“三个新”	
7 (11/03)	大数据分析之聚类1	
8 (11/10)		APEC 开会放假停课
9 (11/17)	大数据分析之聚类2	
10 (11/24)	大数据分析之关联分析	
11 (12/01)	大数据分析之分类分析1	
12 (12/08)	大数据分析之分类分析2	
13 (12/19)	大数据分析之信息检索与链接分析	12/15 出差调课一次
14 (12/22)	大数据分析之推荐系统与产品推荐	
15 (12/29)	大数据分析之情感分析与舆情监测	
16 (01/05)	大数据分析常用工具介绍及课程总结	

32

技术时代沿革



33

数字化生活（新兴IT）特征

- 移动泛在性
- 虚拟时空/体验
- 个性化（推荐）
- 社会性
- 富媒体（交叉服务）



大数据的“4V”特征

体量Volume	海量数据：比传统数据仓库增长速度快10倍到50倍
多样性Variety	多源异构性：不同形式（文本、图像、视频数据）、无模式或者模式不明显、不连贯语法或句义
价值密度Value	低价值密度：大量的不相关信息、需深度分析
速度Velocity	实时分析：流信息、即时需求、连续商务

35

大数据时代的思维变革

- 更多
 - 不是随机样本，而是**全体数据**
- 更杂
 - 不是精确性，而是**混杂性**
- 更好
 - 不是因果关系，而是**相关关系**



36

大数据时代的管理喻意

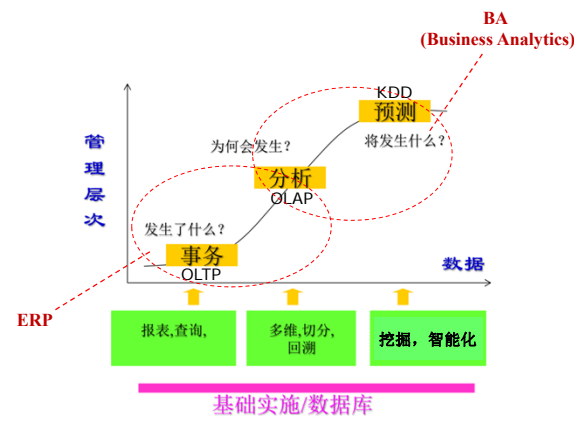
- 三个“融合”
 - IT融合
 - 内外融合
 - 价值融合
- 三个“新”
 - 新模式
 - 新业态
 - 新人群






37

企业数据分析的管理层次



38

深度业务分析 (Business Analytics-BA)

Methods & Tools

- 原方法
- 组合方法

39

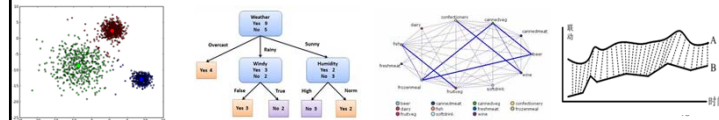
深度业务分析——原方法

- 聚类 (Clustering)
- 分类 (Classification)
- 关联 (Association)
- 模式 (Pattern)
-

类别

联系

轨迹



40

聚类分析方法的种类

- 划分法 Partitioning approach
 - 构建分区: **K-means**, k-medoids, CLARANS
- 层次法 **Hierarchical approach**
 - 分层分解: Diana, Agnes, BIRCH, ROCK, CAMELEON
- 基于密度的方法 Density-based approach
 - 基于连接性和密度函数: DBSCAN, OPTICS, DenClue
- 基于模型的方法 Model-based approach
 - 根据假设为每个类构建一个模型: SOM, EM, COBWEB
- 基于频繁模式法 Frequent pattern-based approach
 - 基于频繁模式的分析: pCluster
 - 多层次粒度结构: STING, WaveCluster, CLIQUE
-

41

分类分析典型方法

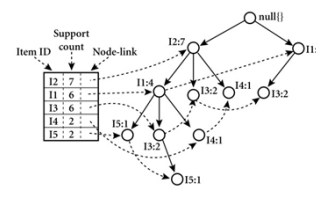
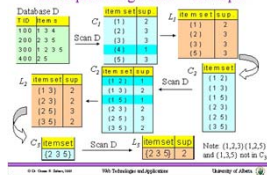
- K近邻方法 (K nearest Neighbors)
- 决策树方法 (Decision Tree)
- 朴素贝叶斯分类方法 (Naïve Bayes)
- 关联分类方法 (Associative Classification)
- 神经网络方法 (Neural Network)
- 支持向量机方法 (Support Vector Machines)

42

关联规则挖掘的有效方法

- **Apriori** (Agrawal & Srikant@VLDB'94)
- Freq. pattern growth (FPgrowth—Han, Pei & Yin @SIGMOD'00)
- Vertical data format approach (Charm—Zaki & Hsiao @SDM'02)
-

The Apriori Algorithm -- Example



深度业务分析——组合方法及应用

- 信息检索及信息搜索服务 (文本内容、链接)
- 推荐系统及产品推荐
- 情感分析及舆情监测
- 社交网络分析及关系营销
- 用户生成内容 (口碑/评论/社交) 分析
-



44

信息检索及信息搜索服务

● 信息检索基础

- 基本概念
- 文档表示、文档权重计算、**向量空间模型**
- 内容相似度计算



● 链接分析基础

- 链接信息的利用
- **PageRank算法**的思路、基础模型、扩展模型
- PageRank的求解
- 搜索结果的综合排序

45

推荐系统基础知识

● 推荐系统出现的背景

● 推荐系统简介

● 推荐系统的模块

- 用户建模模块
- 推荐对象建模模块
- **推荐算法模块**
 - 基于内容的推荐方法
 - 基于用户的协同过滤方法 (UserCF)
 - 基于物品的协同过滤方法 (ItemCF)
 - 混合推荐方法



● 推荐系统的评测指标

46

情感分析与舆情监测

● 情感分析与观点挖掘问题定义

● 情感分析与观点挖掘的任务

- 文档层次情感分类
- 句子层次情感分析
- **基于特征/方面的情感分类**



● 舆情监测的一类分析方法

- 概率主题建模：LDA建模及其结果分析利用
- 案例



47

大数据分析的常用工具

● WEKA

● R与RStudio

● Gephi

● NLPIR / ICTCLASS

● 其他

- SPSS
- SAS
- MATLAB
- CLUTO



48

《大数据时代的管理》

课程论文要求



49

期末课程论文说明

● 主题要求

- 必须与“大数据管理”相关
- 建议围绕所学专业背景下的“大数据管理问题”展开

● 内容要求

- 不少于4000字，版式：word中正文小四字体，1.5倍行距
- 独立完成，不得大段拷贝或直接引用网上、书上及他人已发布内容，需要适当引用时请在引用位置注明参考文献来源（查重）
- 论文内容框架（建议）：
 1. 学习本课程的心得体会、感受，对本课程教学的建议和意见（必有）
 2. 论文背景介绍
 3. 论文涉及的大数据问题及管理需求、策略和意义（可举实例说明）
 4. 本人对该大数据问题的看法、观点及讨论
 5. 总结
 6. 参考文献和资料

50

期末课程论文说明（续）

● 论文提交要求

- 需要以电子版提交，建议提交word版本
- 作业提交邮箱：bigdata_homework@163.com
- 作业提交截止时间：第19周周日（2015.01.11）24时

● 其他说明

- 邮件标题和电子版论文文件请务必按照“学号_班级_姓名.docx”命名，例如“2014211234_2014212103_张三.docx”，也请在邮件中留下姓名、学号及联系方式，以备论文有问题时能够联系到；
- 请在截止时间之前提交论文（不要在截止时间附近，以避免系统原因过期），过期将不再接收论文提交，成绩为0，请务必注意；
- 每次提交论文后，作业邮箱都会有“已收到邮件”的自动回复，如未收到自动回复，表示发送不成功，请在截止时间内重新提交；
- 论文评分的关注重点
 - 有效的课程建议和意见
 - 关注问题的新颖度
 - 个人分析和讨论的深度
 - 论文的整体工作量

51

谢谢大家！

祝大家期末顺利、寒假愉快！

52