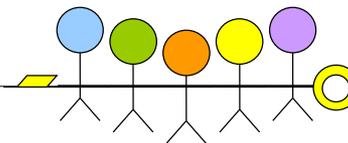


火龙果讲堂：

- 一线专家
- 案例回顾
- 经验分享



大数据中的算法介绍



随时听讲座

每天看新文

追随技术信仰

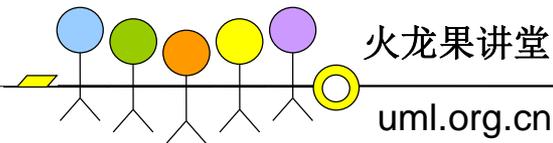
孙超

■ 孙超

- 2012年于浙江大学获得计算机科学博士学位，专业方向：人工智能，机器学习
- 2012年就职于百度（上海）网盟广告算法和策略部门，算法工程师
- 2014年加入阿里巴巴个性化推荐算法部分，任算法工程师
- 主要方向：计算广告学，推荐系统，自然语言处理，机器学习等

- What ——大数据中的算法是什么
- How ——大数据中的算法怎么用
- Can and Cannot ——优势与缺陷
- Will ——未来会怎样

大数据是什么



大数据是什么

网页 新闻 图片 地图 视频 更多 ▾ 搜索工具

找到约 3,030,000 条结果 (用时 0.27 秒)

大数据

网络定义

大数据，或称巨量数据、海量数据，指的是所涉及的数据量规模巨大到无法通过目前主流软件工具，在合理时间内达到截取、管理、处理、并整理成为帮助企业经营决策更积极目的的信息。网络上每一笔搜索，网站上每一笔交易，敲打键盘，点击鼠标的每一个输入都是数据，整理起来分析排行。它的功能可不仅仅止于事后被动了解市场，搜集起来的数据还可以被规画，引导开发更大的消费力量。大数据的常见特点是3V：Volume、Velocity、Variety。“大数据”是由数量巨大、结构复杂、类型众多数据构成的数据集合，是基于云计算的数据处理与应用模式，通过数据的集成共享，交叉复用形成的智力资源和知识服务能力。

<http://zh.wikipedia.org/zh-cn/大数据>

大数据（巨量资料（IT行业术语））_百度百科

baike.baidu.com/view/6954399.htm ▾

第二层而是技术，技术是大数据价值体现的手段和前进的基石。……的决策，了解他们而揣着什么，在充分利用的情况下，大数据可以赋予人们近乎超感官知觉的能力。

什么是「大数据」？ | 问答 | 问答 | 果壳网科技有意思

www.guokr.com/question/457983/ ▾

经常看到有人说大数据，也看到果壳有人在问相关问题，但是到底什么是大数据？... 所谓大数据，狭义上可以定义为难以用现有的一般技术管理的大量数据的集合。

大数据是什么

大数据

搜索 高级搜索 设置 帮助



大数据 北京 <http://weibo.com/bigdata>
大数据世界论坛官方微博
关注 183 | 粉丝 3万 | 微博 1583
最新微博：通过获取与分析海量数据,我们能够获得用来分析人们行为习惯的有效信息,从而使...

精选

央视新闻 央视新闻

【《三体》电影开拍！你期待吗？】国产科幻电影《三体》日前开拍，片中主演均由网友**大数据**选出，其中张静初饰演女主角“统帅叶文洁”。片方特邀美国好莱坞特效团队加盟，小说原作者刘慈欣担任监制提供科幻创意。不过科幻迷们有人期待，有人担忧中国导演无法还原小说精髓...你怎么看？<http://t.cn/RASfUA0>

相关用户

- 数据挖掘与数据分析**
数据挖掘与数据分析自媒体微博
- 数据分析精选**
数据分析精选www.afenxi.com官方微博
- 大数据皮东**
简介：关注大数据在互联网金融的安全和风险管理，致力于大数...
- 微软大数据**
微软 SQL Server 官方微博

LinkedIn 领英 测试版

大数据 精确搜索

所在地区

- 全部
- 中国 (2304)
- 中国 北京市区 (768)
- 中国 上海市区 (268)
- 中国 广东 深圳 (165)
- 中国 上海郊区 (114)

+ 添加

与“大数据”相关的职位

-  **集团总部-大数据中心-数据挖掘算法工程师**
Sohu.com
-  **集团总部-大数据中心-受众定向研发工程师**
Sohu.com
-  **集团总部-大数据中心-大数据平台开发工程师(Oracle方向)**
Sohu.com

卢效鹏 2度
先生

建立联系

■ 大数据：（个人理解）

■ 在海量数据中挖掘有价值的知识，将其转化为生产力的技术

■ 关键字

■ 海量数据：TB, PB, EB, ZB

■ 知识：购买意图（for 淘宝），舆论走向（for 微博），股民信心（for 股票）...

■ 生产力：促成交易，指定政策，走势预测...

■ 关键技术

■ 操作数据：NoSQL, Hadoop, Spark, Storm...

■ 数据挖掘：关联分析，分类算法，聚类算法...

■ 数据采集

- 技术难点：高并发数据采集响应，硬件采集等
- 典型场景：12306，双11淘宝秒杀，智能手环

■ 数据存储

- 技术难点：超大规模数据预处理和保存
- 典型场景：Twitter中消息数据，Google中网页数据

■ 数据分析和挖掘

- 技术难点：海量数据中的有效数据挖掘
- 典型场景：“猜你喜欢”

■ 大数据中最受关注的技术: Hadoop? NoSQL?

- 大数据分析 (12.91%)
- 云数据库 (11.82%)
- Hadoop (11.73%)
- 内存数据库 (11.64%)
- 数据安全 (9.21%)

根据ESM国际电子商情针大数据应用现状和趋势的调查显示

■ 数据分析中最受关注Top3

- 实时分析 (21.32%)
- 挖掘模型(17.97%)
- 可视化界面(15.91%)

■ 大数据中的工程技术与算法

■ 工程技术：

- 如何操作挖掘机
- 如何保存宝藏

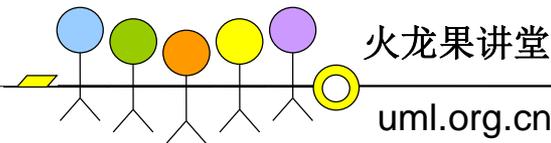
■ 挖掘算法：

- 如何定位宝藏位置
- 如何从泥土中识别宝藏

■ 工程与算法都是“器”，关键问题是

■ 算法能够做什么

案例0:



■ “啤酒”与“尿布”

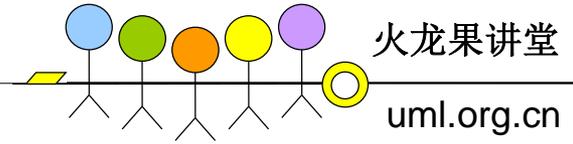
- 数据：超市的购物结算清单
- 知识：啤酒与尿布同订单的比例很高
- 价值：搭配销售

- 挖掘技术：关联规则算法
- 数学理论：贝叶斯理论

■ X易游戏的潜在客户挖掘

- 场景：X易公司上线一款网游，对用户免费，但是很多场景和条件下需要付费（比如升级，买装备，打副本...）。现在上线四个礼拜后，具有付费用户**100万**，非付费用户**500万**，希望能够从刚刚上周注册的**10万名**免费账户中，识别出哪些可能在未来成为客户，即订购付费项目，对其有针对性的做推广
- 怎么做？

案例1:



■ X易游戏的潜在客户挖掘

■ 数据源: ?

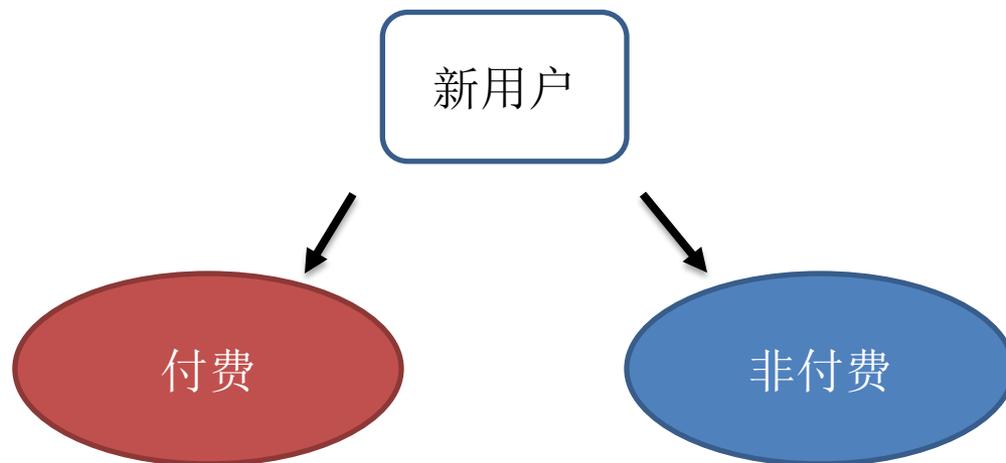
■ 知识: ?

■ 价值: 游戏收入

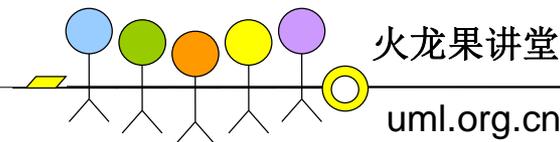
案例1:

■ X易游戏的潜在客户挖掘

- 数据源: 100万付费用户行为, 500万非付费用户行为, 10万新用户行为
- 知识: **付费用户的行为特征**
- 价值: 游戏收入



案例1:



■ 解决方案:

■ 规则法:

- rule 0: 每天上线4小时以上 ->付费用户
- rule 1: 每小时操作行为数大于1000次 -> 付费用户
- rule 2: 与付费用户PK成功率小于20% ->付费用户
-
- rule 256: rule 0 && rule 1 || rule2 && !rule3...

■ 优点: 简单直接

■ 缺点: 耗时费力, 人工成本高, 准确率差

■ 解决方案:

■ 机器学习法:

- 付费用户: 每天上线率, 上线时间.... 标记: 1
 - 非付费用户: 每天上线率, 上线时间.... 标记: 0
 - 新用户: 每天上线率, 上线时间.... 标记: ?
 - 前两个数据扔给机器学习算法, 计算得到模型, 然后把后一个数据作为输入, 自动得到输出
- ### ■ 优点: 成本低, 可扩展, 通用性好
- ### ■ 缺点: **Badcase**修正不直接*
- 百度网盟中投诉的例子

■ 规则与模型

■ 传统认知：

- 运营团队偏规则，算法团队偏模型
- 头部数据偏规则，长尾数据偏模型
- 百度偏规则，Google偏模型

■ 实际应用：两者结合

数据清洗与预处理

特征选择

模型选择与训练

效果评估

■ 数据清洗与预处理

■ 在海量数据中去除异常数据

- 典型案例0：淘宝刷单数据

- 典型案例1：百度网盟中的低俗页面去除

小问题：根据**Case1**，设计一个低俗页面识别的方案

■ 数据预处理：把数据转换成可用的格式

- 数据归一：所有数据统一在同一个范围内

- 典型案例：淘宝商品的价格归一

■ 特征工程

■ 选择可以反映问题的特征

- 算法中的核心部分之一

- 决定了算法最终能够多大程度上解决实际问题

■ 在案例1中：

- 每天上线时间即为一个特征

- 每周上线天数即为一个特征...

■ 特征的数量：

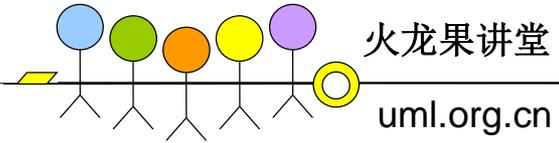
- 很少（5个特征）~ 巨大（2亿个特征）

■ 模型选择与训练

- 模型的选择：常见的机器学习算法（商业中应用的）
 - 分类算法：逻辑回归，随机森林...
 - 聚类分体：K-means，图聚类...
- 模型算法相当于工具库，不需要自己制造工具，只需要选择合适的工具
 - Hadoop中的Mahout
 - Spark中的Mllib
 - Python中的Numpy
 - R语言

- 评估算法多大程度（置信度）上解决了问题
 - 测试数据集
 - 线上验证(A/B test)

Case 1 Review



- 数据源：玩家的线上日志数据，分布在N台服务其中
- 数据清洗和预处理：
 - 编写awk脚本，通过MapReduce任务把需要的日志字段提取出来，并且在awk中设定规则，过滤某些数据
- 特征抽取：
 - 根据定义的特征，编写MapReduce任务，统计分析得到需要的特征
 - 或数据入库（Hive），通过SQL语句统计得到需要的特征

■ 模型选择和训练

- 调用Hadoop中的Mahout中的分类算法（逻辑回归，SVM等），得到预测模型

■ 效果评估

- 在已标记的测试用户中使用该模型，确定其效果
- 模型上线观察预测效果

- 如果现在我们拿到了10万个新用户，在其他游戏上的付费情况，怎么用？
 - 数据维度的增加
- 如果这是一款阿里运营的游戏，因此我们有这个用户在淘宝上的购物数据怎么用？支付宝记录？信用等级？年龄，性别....怎么用？
 - 异构数据的融合

■ 数据框架

- Hadoop(Hive)

- Spark

- Storm

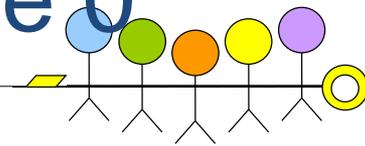
■ 执行工具

- 各种可用的语言：**awk**(对于格式化数据)，**HSQL**(针对数据库)，**Python**，**Java**，**Scala**，**R**

■ 常用算法

- 逻辑回归，**SVM**，随机森林，**KNN**，**Kmeans**，**Deep Learning**....

大数据算法能够实现什么-Case 0

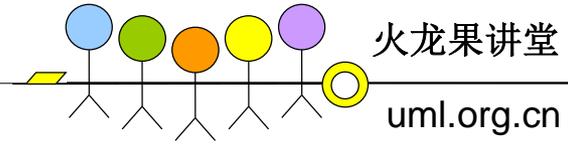


火龙果讲堂

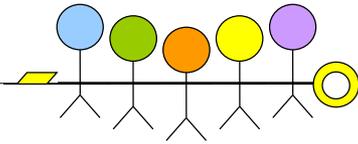
uml.org.cn

- 打开大众点评，查找想吃的餐厅
 - 个性化推荐算法选择最适合你口味的餐厅排在前面
- 打开Uber轿车
 - 算法根据你的地理位置，附近可用车的GPS定位，车辆评价等选择最合适的叫给你
- 吃完用支付宝付款
 - 安全算法根据你的支付记录和支付习惯，确认支付的安全性

大数据算法能够实现什么



- 从最简单的手机定位（也有机器学习算法：为什么不开**GPS**，也能知道位置）到最复杂的天气预报，都是算法在起作用
- 经典的大数据算法场景
 - 搜索引擎：PageRank + **CTR**预估
 - 推荐系统：**协同过滤**，**矩阵分解**
- 非典型大数据算法场景
 - 舆情监控
 - 股票预测



■ “大数据不擅长社会关系分析”？

[北京]世纪佳缘招聘推荐算法|数据挖掘工程师

发布时间：2012-07-31

工作地点：北京

职位类型：全职

来源：北邮人BBS

■ “大数据不擅长上下文情景分析”？

“情景感知”：下一个智能科技新趋势

91.com移动互联网第一平台 时间：2014-07-14 [网站合作] [快速评论](#) [关注](#) [~](#)

■ “大数据不擅长处理真正的巨型问题”？

大数据如何改变2012美国总统大选

大数据只是工具，不能提出问题

析，提炼出关键信息：谁云投票？竞选活动如何接触选民，甚至如何更好地回复一条推文或者电子邮件。

■ 大数据的陷阱

■ “统计陷阱”

- 飞机加固问题
- 医院死亡率最高

■ 相关性不代表因果性

- 草莓味冰淇淋导致熄火问题

■ 牢记成本

- 大数据带来高成本

■ 安全问题

- ebay数据泄露导致股价大跌

■ 更多的数据来源

■ 穿戴式计算

- 智能手环/智能手表
- Google Glass
- 便携式脑电电极帽

■ 普适计算

- 更精准GPS
- 更多的户外监控系统
- 更多的智能家电

■ 更多的数据形式

■ 关于人

- 健康状况（微软的智能医疗愿景）
- 行为记录（各种数据融合）

■ 关于商品

- 材质，产地（百度筷搜）
- 物流数据（全球**GPS**定位）

■ 关于环境

- 气候，天气，雾霾...

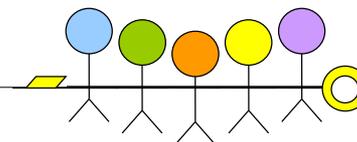
■ 更多的应用

- 金融资金流动预测
- 政府政策模拟执行
- 全球交通状况分析
- 智能家居，智能城市...

■ 更多的数据

■ 更快的计算资源

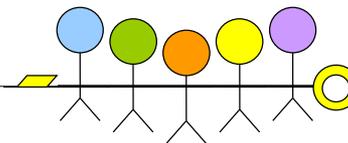
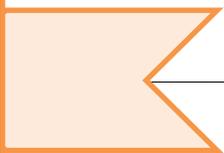
- 大数据需要用到机器学习吗
 - 不一定，但是主流的场景都不同程度的使用了
- 大数据的发展趋势，会一直火下去么
 - 不会，因为大数据本身就只是一个概念而已
- 大数据挖掘在用户画像中的应用
 - 用户模型的建立和利用
- 大数据处理的算法，与传统数据挖掘算法的区别在哪里
 - 本质上没有区别，但是有的算法在大数据下不好用(SVM)，有的非常好（深度学习）



- 大数据在物联网方面的价值及应用举例
 - 浙大已经有物联网的项目
 - 微软的智能医疗项目
- spark是不是未来的大数据处理最有技术
 - 很有可能，Spark的使用范围增长迅速
- 大数据算法需要哪些数学知识
 - 概率，统计，微积分...
 - 机器学习算法
- 如何尝试
 - 阿里巴巴天池大数据竞赛

交流时间





火龙果讲堂

uml.org.cn

讲座	2015年4月25日 大型分布式网站架构初探
讲座	2015年5月16日 Qt 在移动端应用开发实践



随时听讲座

每天看新文

追随技术信仰