R 语言在电子商务领域的应用

刘思喆

数据部
推荐系统

2014 年 5 月 25 日

目录

推荐系统

使用 R 语言的背景

- 京东业务涉及用户、商品、商家、促销、反作弊、风险控制、精准营销、运营优化等
- 2012 年正式启动了大数据平台的搭建，平台底层数据存储和离线运算由 hadoop 完成

建模环境简述

数据情况：

- 客户维度：亿级 -> 千万级
- 商品维度：千万级 -> 百万级
- 数据量：M-G 级

分析场景：

1. 探索分析：均值、方差、分位数、列联表
2. 基础分析：如假设检验、相关分析、主成分（因子）分析
3. 挖掘模型：回归、kmeans 聚类、决策树、关联规则、时序等
4. 可视化图形：条图、直方图、概率密度图、定制化图形
5. 重复性分析：

从 R 的角度来看

- 数据挖掘领域应用最广泛的软件和语言（KDnuggets 2012,2013）
- 完整且丰富的统计、机器学习、可视化平台
- 数据编程的完美实现
- 便捷的、可扩展的并行方案（如同 hadoop）

目录

推荐系统

大数据的解决方案

Solution 1: R 同一些特定领域的工具的结合 (e.g MapReduce style tools, Hadoop, Streaming, Hive, Pig, Cascading...)

- R + Hadoop: rhbase, rmr, rhdfs, RHIPE
- R + MongoDB: RMongo, rmongodb
- R + MPI: Rmpi, pbdMPI, snow, snowfall
- R + GPU: gputools, HiPLARM

Solution 2: 通过扩展包，增强读取和处理大数据的能力 (e.g bigmemory, ff, biglm...)

- ff: offers file-based access to data sets that are too large to be loaded into memory
- biglars: can use the ff to support large-than-memory datasets for least-angle regression, lasso and stepwise regression.
- bigrf: a Random Forests implementation with support for parellel execution and large memory.

实际应用方案

包或支持	运行环境	优势	劣势
BLAS	单机	直接并行化	只针对于数学计算有效
parallel	单机	轻量级	fork 方式
snow，snowfall	集群	易部署	socket
Rmpi	集群	较成熟	(未采用)
Rhadoop	集群	同现有环境匹配	依然有一定开发量

还有一种方式：使用 R 生成规则，在 hadoop 平台做并行

目录

推荐系统

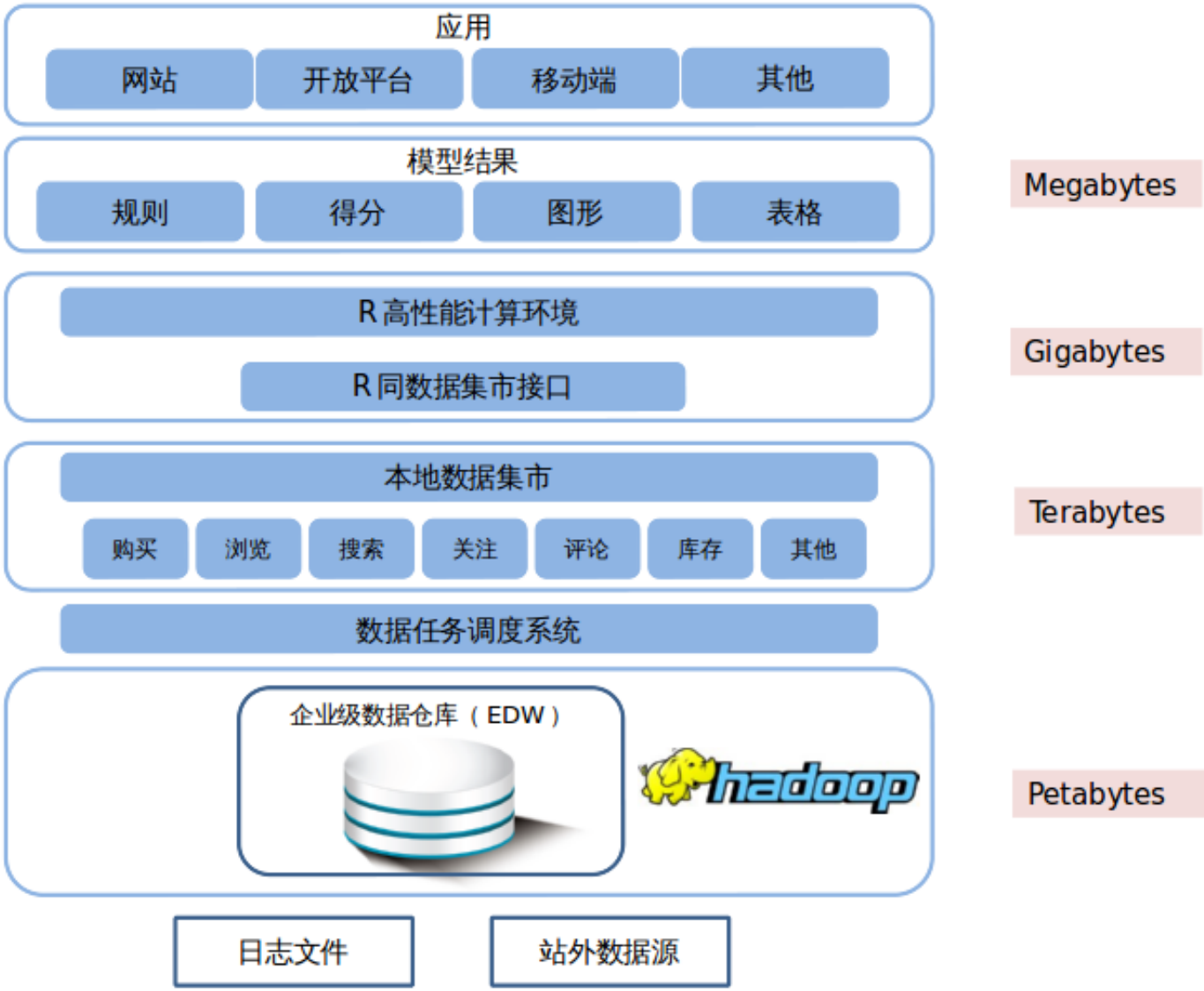
JD.COM 京东

多 · 快 · 好 · 省

一般工作流程

1. 通过 Hive 集群获取目标数据
2. 在 R 环境下进行数据探索、清洗、转换工作
3. R 环境下分析建模 (Feature Selection, Benchmark)
4. 评估 (离线评估和分流量测试)
5. 线上集成 (R, Hive QL, Java, C++, Python...)

数据的流动


应用
网站
开放平台
移动端
其他
模型结果
规则
得分
图形
表格
R 高性能计算环境
R 同数据集市接口
本地数据集市
购买
浏览
搜索
关注
评论
库存
其他
数据任务调度系统
企业级数据仓库（EDW）
hadoop
日志文件
站外数据源
Megabytes
Gigabytes
Terabytes
Petabytes

涉及数据挖掘、分析技术的相关 R 包

- 数据传递及服务（RHive、RServe、rJava、RJDBC）
- 清洗及预处理（sqldf、stringr、XML、data.table）
- 抽样、预测、分类、关联规则、特征选择、稀疏矩阵运算、矩阵分解、社交网络、分词、模型评估等
- 高性能计算（rhdfs、rmr2、Rcpp、snow）
- 自动化报告（knitr、slidify）
- 其他

目录

推荐系统

如何评价一款商品的好坏？

1. 普通消费者在购买商品后，会对商品有一个全面的感受，比如销售者满意度、客户体验等。但如何收集消费者对商品的感受数据则是难点。
2. 消费者对一件（一组）商品满意度非常差，则消费者会选择不再回来购物。这种流失的源头是出现了“不良”商品，需要商家进行优化（比如配送、商品质量、商品价格等）

可选的数据解决方案

- 客户投诉数据
- 商品评论的文本数据
- 基于用户购买行为

前两种方案优点是非常明显的，是了解商品满意度的直接途径。但也存在部分缺点：

- 投诉数据量有限，有些用户并不倾向通过投诉来表示不满，而是直接“用脚投票”
- 使用评论数据的前提是：所有的不满意用户都会在网站上留言，但事实并非如此
- 用户对于商品挑剔程度不同：有的用户所有的评论都呈现攻击性态度，但这些差评并不影响未来的购物

基本原理

在京东第一次购物的用户体验非常重要，体验好则成为存量用户，反之则失去用户。已知这些新用户第一次购买的商品清单和用户未来的状态（未来是否再发生购物行为），则可以生成如下矩阵：

status		p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
1	u1	0	0	0	1	1	0	0	0	0	0
1	u2	0	0	1	1	0	0	0	1	0	0
1	u3	1	1	0	0	0	0	0	0	1	0
1	u4	1	0	0	1	0	0	0	0	0	0
1	u5	0	1	0	0	0	1	0	0	0	0
0	u6	0	0	1	0	0	0	0	0	1	0
0	u7	0	1	0	0	0	0	1	0	0	1
0	u8	1	0	0	0	1	0	0	0	0	0
0	u9	0	0	0	1	0	0	0	0	1	0
0	u10	0	0	1	0	0	0	0	1	0	1

- 如果用户未来 X 个月又购买了某种商品，则在这个矩阵 status 对应的位置标记为 1，反之为 0。
- 矩阵的行代表了用户，列代表了商品：对于第一行来说，u1 用户购买了 p4、p5 两件商品。

基本原理

在京东第一次购物的用户体验非常重要，体验好则成为存量用户，反之则失去用户。已知这些新用户第一次购买的商品清单和用户未来的状态（未来是否再发生购物行为），则可以生成如下矩阵：

status		p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
1	u1	0	0	0	1	1	0	0	0	0	0
1	u2	0	0	1	1	0	0	0	1	0	0
1	u3	1	1	0	0	0	0	0	0	1	0
1	u4	1	0	0	1	0	0	0	0	0	0
1	u5	0	1	0	0	0	1	0	0	0	0
0	u6	0	0	1	0	0	0	0	0	1	0
0	u7	0	1	0	0	0	0	1	0	0	1
0	u8	1	0	0	0	1	0	0	0	0	0
0	u9	0	0	0	1	0	0	0	0	1	0
0	u10	0	0	1	0	0	0	0	1	0	1

- 如果用户未来 X 个月又购买了某种商品，则在这个矩阵 status 对应的位置标记为 1，反之为 0。
- 矩阵的行代表了用户，列代表了商品：对于第一行来说，u1 用户购买了 p4、p5 两件商品。

实际的清单结果及解释

Table：部分“不良”商品的清单

sku_id	score	product_name
10xxxxxxxx	-0.52	carslan 卡姿兰 xxxxxxx
10xxxxxxxx	-0.41	xxxx xxxx 秋冬新款男士亮面休闲 xxx xxx xxx
10xxxxxxxx	-0.55	xxxx 秋冬新款男装修身中长款 xxxxx xxxxx xxx 黑色 XL
xx41xx	-0.37	xxxx xxxx 18 升电烤箱 xxx xxxx xx
xxxx76	-0.40	诺基亚（NOKIA）xxxx GSM 手机（x）非定制机
xxxx81	-0.38	诺基亚（NOKIA）xxxx GSM 手机（x）非定制机
xxxx23	-0.33	联想 xxxx 3G 手机 xxxxxxxx 双卡双待单通
xxxx25	-0.52	xxxxx xxxx 4G 录音笔
xxxx32	-0.41	HTC xxxxx 3G 手机（xxxx）TD-SCDMA/GSM

Table：对于三类商品的标记以及表现

表现	表现	id	名称	购买用户数	再次购买	不再购买
不良：	-	10xxxxxxxx	xxxxxxxxx 纯棉男袜	32	3	29
正常：	0	38xxxxxx	xxx 超薄干爽纸尿裤箱装 xxxx 片	13	6	7
优秀：	+	30xxxxxx	xx 螺旋藻 xxxxxxxxx*1 桶	15	14	1

实际的清单结果及解释

Table：部分“不良”商品的清单

sku_id	score	product_name
10xxxxxxxx	-0.52	carslan 卡姿兰 xxxxxxx
10xxxxxxxx	-0.41	xxxx xxxx 秋冬新款男士亮面休闲 xxx xxx xxx
10xxxxxxxx	-0.55	xxxx 秋冬新款男装修身中长款 xxxxx xxxxx xxx 黑色 XL
xx41xx	-0.37	xxxx xxxx 18 升电烤箱 xxx xxxx xx
xxxx76	-0.40	诺基亚（NOKIA）xxxx GSM 手机（x）非定制机
xxxx81	-0.38	诺基亚（NOKIA）xxxx GSM 手机（x）非定制机
xxxx23	-0.33	联想 xxxx 3G 手机 xxxxxxxx 双卡双待单通
xxxx25	-0.52	xxxxx xxxx 4G 录音笔
xxxx32	-0.41	HTC xxxxx 3G 手机（xxxx）TD-SCDMA/GSM

Table：对于三类商品的标记以及表现

表现	表现	id	名称	购买用户数	再次购买	不再购买
不良：	-	10xxxxxxxx	xxxxxxxxx 纯棉男袜	32	3	29
正常：	0	38xxxxxx	xxx 超薄干爽纸尿裤箱装 xxxx 片	13	6	7
优秀：	+	30xxxxxx	xx 螺旋藻 xxxxxxxxx*1 桶	15	14	1

有益的效果

按品类，造成用户流失的原因分析略 . . .

- 识别过程更加规整化、流程化。为日常运营中干预“不良”商品提供了一个有效、快速、便捷的方式。
- 有效减少不良商品对于客户的负面影响。阻止这些客户流失或流向竞争对手，对其他（潜在）顾客的负面影响降低至最低。
- 对于使用以天为记录单位的不良商品识别方法的应用，每天大约能记录 5-10 种不良商品，平均覆盖 100-150 个客户。保守地，按照每位客户一年再购买一次商品，客单价 250 计算，未来一年累计额外带来 900 万 -1350 万的销售额。

有益的效果

按品类，造成用户流失的原因分析略 . . .

- 识别过程更加规整化、流程化。为日常运营中干预“不良”商品提供了一个有效、快速、便捷的方式。
- 有效减少不良商品对于客户的负面影响。阻止这些客户流失或流向竞争对手，对其他（潜在）顾客的负面影响降低至最低。
- 对于使用以天为记录单位的不良商品识别方法的应用，每天大约能记录 5-10 种不良商品，平均覆盖 100-150 个客户。保守地，按照每位客户一年再购买一次商品，客单价 250 计算，未来一年累计额外带来 900 万 -1350 万的销售额。

总结

- 并不是所有的数据场景都适合使用 R，但 R 可以帮我们对业务迅速做出反应
- 一般情况下，传统的统计模型优于复杂的推断模型

Q & A

- 邮件：liusizhe<at>jd.com
- 博客: `http://www.bjt.name`
- 微博：@ 刘思喆

Jump to first slide

谢谢！
Thank you!

北京市朝阳区北辰西路8号北辰世纪中心A座6层
6F Building A, North-Star Century Center, 8 Beichen West Street,
Chaoyang District, Beijing 100101
T. 010-5895 1234 F. 010-5895 1234
E. xingming@jd.com www.jd.com

JD.京东.COM