R 语言数据挖掘应用 @ 京东商城

刘思喆

商业智能和搜索部
数据挖掘组

2013 年 03 月 29 日

目录

高速成长的京东商城

- 中国最大的网络零售商，增速超过业界平均速度 3 倍以上
- 8000 万注册用户，上万家供应商
- 日均 PV 超过 2 亿，日均 UV 超过 1500 万, 日订单量超过 100 万

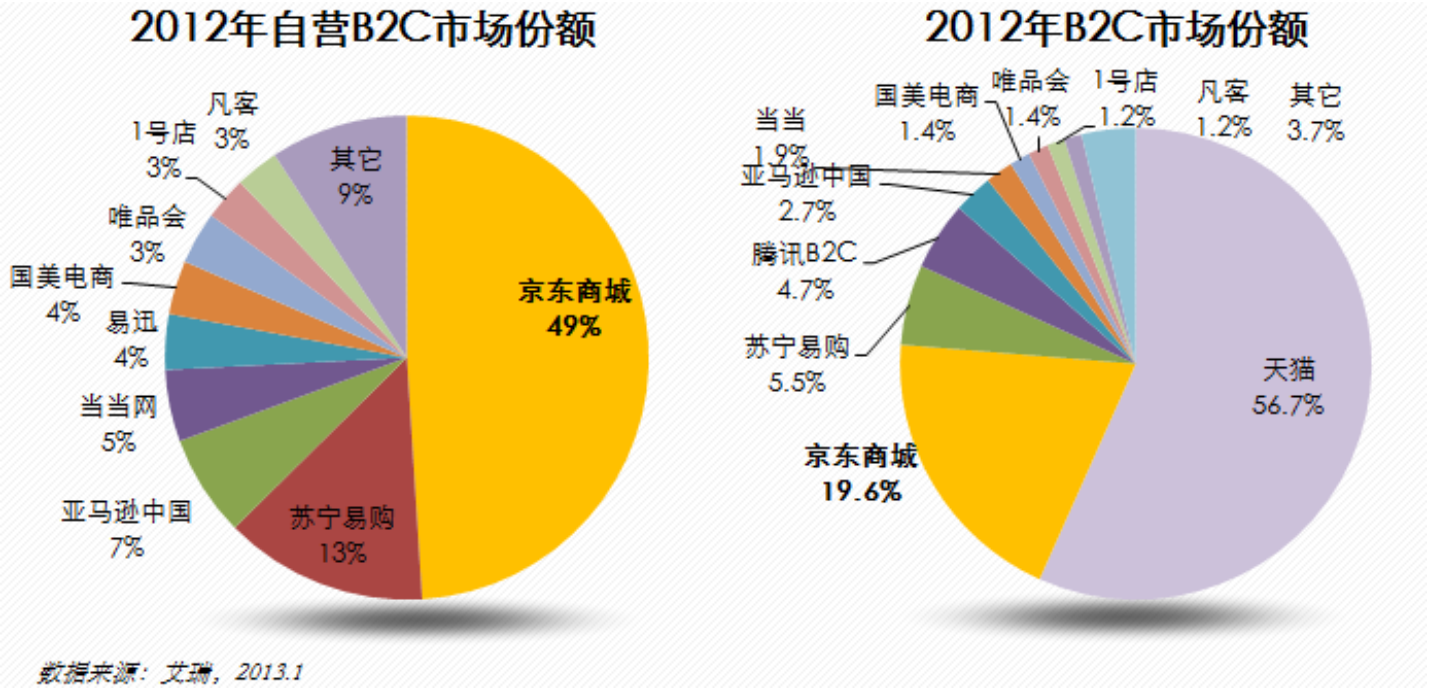


Figure: 京东商城在 2012 年继续强劲增长，在中国自营式 B2C 市场中占据 49.0% 的份额，在中国 B2C 市场中占据 19.6% 的份额。

目录

为什么我们选用 R 作为主要的数据挖掘工具

京东商城线上有千万种商品同时售卖，纯粹的人工补货是不现实的，必须依赖于自动补货系统。而商品的未来需求则是自动补货系统的重中之重，如何准确的预测每件商品未来需求（销量）是数据挖掘团队的其中一项重要的工作。

在 2011 年京东商品销量预测项目直接引出了挖掘工具的选型问题：由于团队成员背景不同，各有偏重，数据挖掘团队选择了 R、SPSS、Java 以及一家国内数据挖掘软件作为候选工具评估：

	R	PASW	Java	AA
准确性	高	高	低	高
扩展性	高	中	高	低
灵活性	高	低	高	低
易用性	高	高	低	中
集成性	高	低	高	低

面对大数据的解决方案

Solution 1: Use R in Conjunction with other specialized tools(e.g MapReduce style tools, Hadoop, Streaming, Hive, Pig, Cascading...)

Solution 2: Packages that enable new functionality for reading and processing very large data sets. (e.g bigmemory, ff, Enhance function, but no enhancements to the core language)

面对大数据的解决方案

Solution 1: Use R in Conjunction with other specialized tools(e.g MapReduce style tools, Hadoop, Streaming, Hive, Pig, Cascading...)

Solution 2: Packages that enable new functionality for reading and processing very large data sets. (e.g bigmemory, ff, Enhance function, but no enhancements to the core language)

主要针对如下领域

除了销量预测系统以外，R 语言还应用在

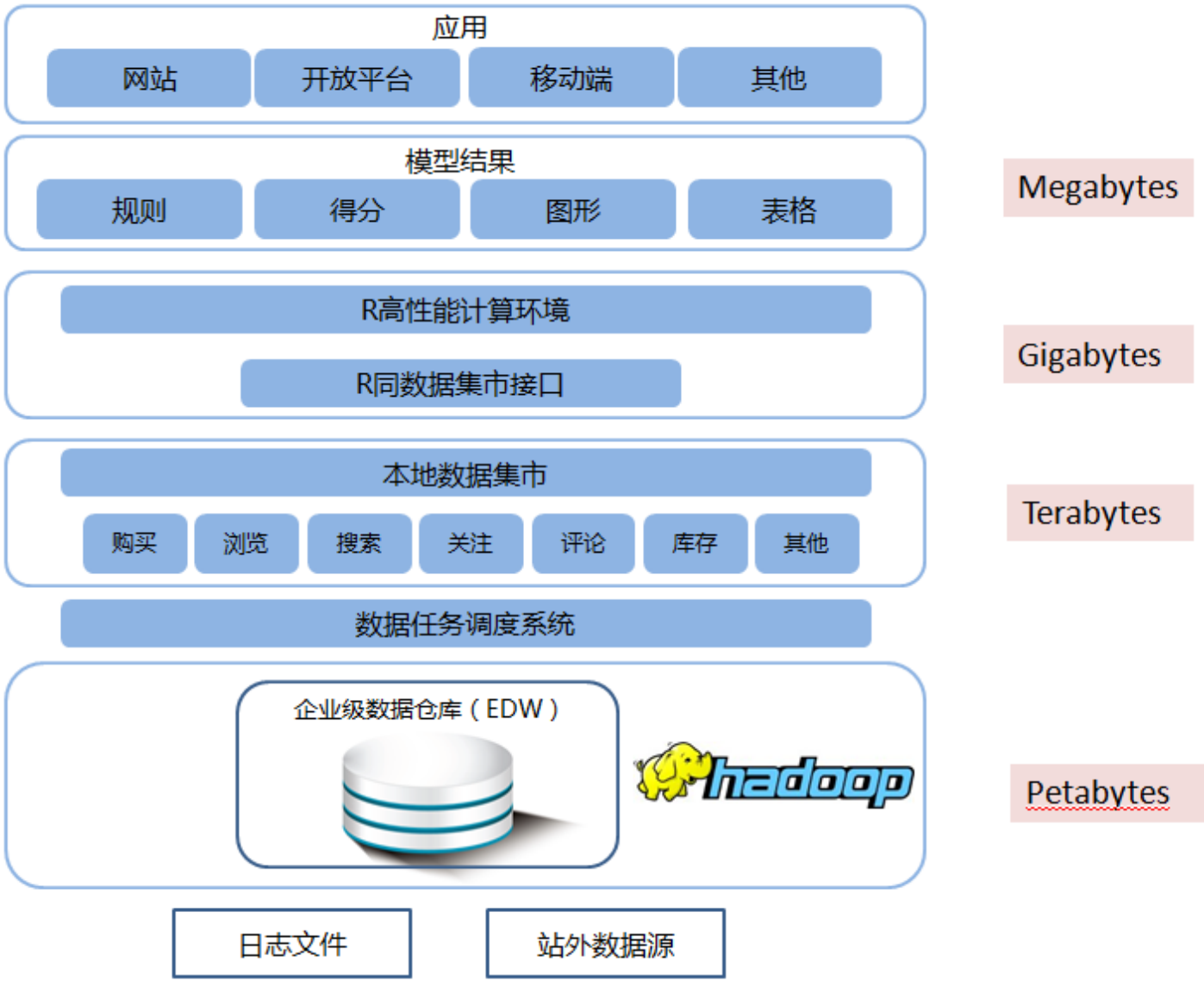
- 集群数据的调度清洗
- 建模过程中的数据预处理
- 统计分析和建模
- 数据可视化
- 算法的原型实现

目录

一般工作流程

1. 通过 Hive 集群获取目标数据
2. 在 R 环境下进行数据预处理
3. R 环境下分析建模 (Featrue 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

涉及数据挖掘技术和相关的 R 包

- 数据传递及服务（RHive、RServe、rJava、RJDBC）
- 清洗及预处理（sqldf、stringr、XML）
- 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、分词等
- 高性能计算（rhdfs、rmr2、Rcpp）
- 其他

挖掘模型服务对象

- 在线广告优化
- 在线商品推荐
- 搜索词优化
- 邮件营销
- 移动客户端
- 活动及促销推送
- 开放平台的 PoP 商户
- …

目录

典型场景

用户 A： 男性、28 岁、北京、累计购买金额 13428 元、没有投诉记录、最近 2 个月购买过 ipad4 MD513CH，购买过图书三体，搜索过莫言、剃须刀、HDMI 转接线、手机等关键词，关注 Sony KDL-46HX750 3D LED 液晶电视，促销偏好度高……

用户 B： 女性、33 岁、上海、累计购买金额 3420 元、曾有过投诉记录，记录关键词为安装慢、退货等，近 2 个月购买过 ONLY 圆领立体剪裁无袖修身连衣裙 E(黑)，蓝月亮亮白增艳自然清香洗衣液 3000g，关注飞利浦 PT720 三刀头电动剃须刀，搜索过雅培、多美滋，促销偏好度低……

用户 C：

京东商城要做红酒专场活动，请问上述哪个用户更可能是目标客户群。

典型场景

用户 A： 男性、28 岁、北京、累计购买金额 13428 元、没有投诉记录、最近 2 个月购买过 ipad4 MD513CH，购买过图书三体，搜索过莫言、剃须刀、HDMI 转接线、手机等关键词，关注 Sony KDL-46HX750 3D LED 液晶电视，促销偏好度高……

用户 B： 女性、33 岁、上海、累计购买金额 3420 元、曾有过投诉记录，记录关键词为安装慢、退货等，近 2 个月购买过 ONLY 圆领立体剪裁无袖修身连衣裙 E(黑)，蓝月亮亮白增艳自然清香洗衣液 3000g，关注飞利浦 PT720 三刀头电动剃须刀，搜索过雅培、多美滋，促销偏好度低……

用户 C：

京东商城要做红酒专场活动，请问上述哪个用户更可能是目标客户群。

模型的线下测试效果

- 涉及用户数：9832608
- 购买概率大于 0.34 用户数：303641
- 未来 5 天实际购买用户数：14290
- 预测命中用户数：10337

对用户： 最小程度地打扰客户，提高客户体验

对企业： 减低营销成本，提高客户忠诚度

模型的线下测试效果

- 涉及用户数：9832608
- 购买概率大于 0.34 用户数：303641
- 未来 5 天实际购买用户数：14290
- 预测命中用户数：10337

对用户： 最小程度地打扰客户，提高客户体验

对企业： 减低营销成本，提高客户忠诚度

紧接着.....

筛选的客户我们还需要做以下工作

渠道： 网页直接推荐、邮件推送（提醒）、移动客户端推荐、短信告知、站内提醒

时间： 工作日、周末、节日、日间、晚间等

方式： 直减、满减、活动、优惠券、捆绑销售等

部分应用案例

- 基于京东评论的新词识别模型
- 商品的价格弹性模型
- 商品性别色彩模型
- 京东商城“不良”商品识别模型
- PoP 商家分群模型
- 京东商城三级类目购买关系模型
- 某品类评论关键词网络模型
- 商品销量预测模型
- 促销活动兴趣度模型
- 类目偏好模型（用于定向营销）
- 潜在用户识别模型（用于定向营销）
- 搜索桥梁词识别

Q & A

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

Jump to first slide