

# 京东推荐系统实践

打造千人千面的个性化推荐引擎

推荐搜索部

刘思喆

2015 年 4 月 18 日



# 目录

## 推荐系统



- 1 京东推荐产品及架构
- 2 通用模型的应用
- 3 离线 CTR 预测实例
- 4 实验与监控

# 目录

## 推荐系统



- 1 京东推荐产品及架构
- 2 通用模型的应用
- 3 离线 CTR 预测实例
- 4 实验与监控

# 京东推荐产品

- 80+ 推荐产品，包括移动端和 Web 端
- 20+ 推荐服务，支撑 EDM、广告、微信端等
- 遍布用户网购的各个环节

## 推荐系统的价值


- 挖掘用户潜在购买需求
- 缩短用户到商品的距离
- 用户需求不明确时提供参考
- 满足用户的好奇心

# 推荐产品截图示例

根据浏览猜你喜欢



达尔优 (dare-u) G60 牧马 人游戏鼠标 四色呼吸灯变换



购买了该商品的用户还购买了



康纳(CONNAL) ZTD100K-SD1 家用空气净化器  
直降 ¥1299.00  
加入购物车



迪士尼 (Disney) 可调闪光片 装轮滑鞋 溜冰鞋旱冰鞋 DCY3 D 27-30  
¥239.00

540水槽双槽不锈钢

浏览了该商品的用户最终购买



康纳(CONNAL) CXW-200-TD08A 欧式 吸油烟机  
¥1799.00

猜你喜欢 根据你的喜好精心为你推荐



【京东自营】孔雀 (Peacock) 不锈钢真空保温壶  
¥219.00



沃尔玛GIFT卡吉祥如意卡 购物卡超市卡礼品卡  
¥985.00



欧莱雅 (LOREAL) 男士劲能保湿护肤霜 50ml  
¥85.00



奥乐儿童帐篷游戏屋宝宝帐篷 户外玩具 单个帐篷不含海洋球  
¥39.00

¥15.20

关注此商品的人还关注



快易典EH1 电子词典 翻译机 小学初中  
¥298.00  
加关注



快易典 (Koridy) A990全能词典王  
¥339.00  
加关注



快易典EH2 电子词典 学生英语辞典  
¥418.00  
加关注

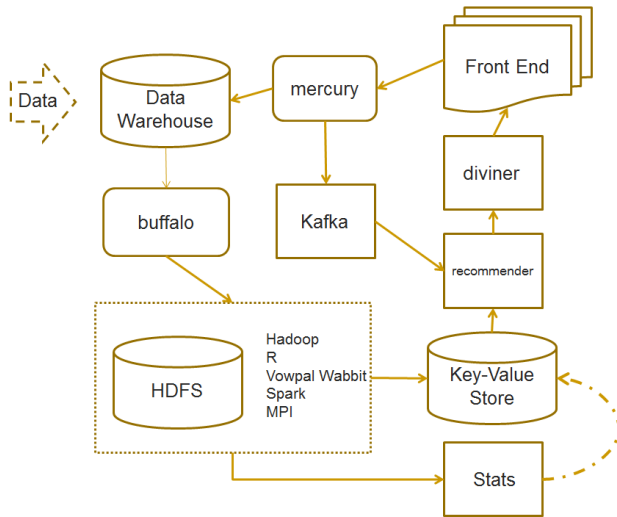


好易通 (besta) 无线 V4半津高防+剑  
¥669.00  
加关注

## 不同位置的推荐产品定位不同

- 单品页：购买意图
- 过渡页：提高客单价
- 购物车页：购物决策
- 无结果页：减少跳出率
- 订单完成页：交叉销售
- 关注推荐：提高转化
- 我的京东推荐：提高忠诚度
- 首页猜你喜欢：吸引用户

# 京东推荐系统架构

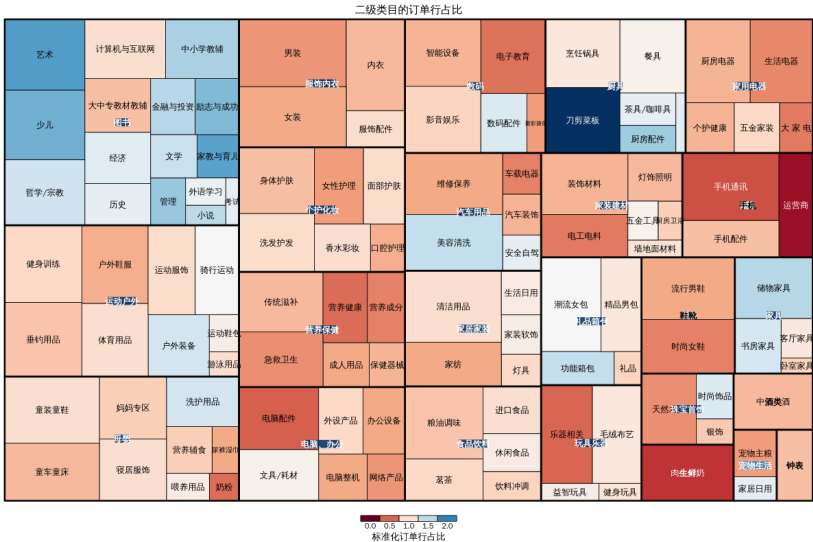


## 京东推荐算法优化方向

- 以数据分析为工具，提升数据的质量和覆盖度，增强对业务的理解（25%）
- 测试不同算法在不同数据源的效果，提高召回模型的质量，增加结果辨识度（50%）
- 以用户反馈为依据，融合不同类型、不同维度据源，对推荐结果重排序（15%）
- 增加数据的更新频率（5%）
- 其他（5%）



# 推荐系统效果全景图



注：出于公司数据发布安全考虑，已对品类订单占比数据做了随机变换，仅为演示所用

# 目录

---

推荐系统



JD.COM 京东

多·快·好·省

- ① 京东推荐产品及架构
  - ② 通用模型的应用
  - ③ 离线 CTR 预测实例
  - ④ 实验与监控
-

# 京东对推荐数据的理解

## 用户行为

- ① 浏览
- ② 点击
  - 普通点击
  - 搜索点击
- ③ 加入购物车（或关注）
- ④ 购买
  - 订单
  - 用户
- ⑤ 评分

## 基于内容

- 标题
- 扩展属性
- 评论
- 描述
- ...

# 典型推荐系统技术

按照数据的分类： 协同过滤、内容过滤、社会化过滤

按照模型的分类： 基于近邻的模型、矩阵分解模型、图模型

# 协同过滤 I

用户和商品的共现阵：

	I
U	1, 0, 0, 0, 0, 1,
	0, 1, 0, 0, 0, 0,
	1, 1, 0, 0, 0, 1,
	0, 0, 0, 0, 1, 0,
	0, 0, 1, 0, 1, 0,
	0, 0, 1, 0, 1, 0,
	0, 0, 0, 1, 0, 0,
	0, 0, 0, 0, 0, 1,
	0, 0, 0, 0, 1, 0,
	0, 0, 1, 0, 0, 1,

对于商品 (item) 向量至少有 10+ 的距离计算公式来计算商品间的距离，一般有：

- Jaccard 距离
- (修正)cosine 距离
- Manhattan 距离
- Chebychev 距离
- 欧 (闵) 式距离
- Pearson 相关系数
- Spearman 相关系数
- Kendall 相关系数
- ...

## 协同过滤 II

以及不太常见的：

- simrank
- Mahalanobis 距离
- 基于条件概率的 interest
- Log likelihood ratio
- Mutual information

# 支持类模型

- 离线推荐 CTR 预测模型
- 用户购买力模型
- 周期购买商品识别模型 (商品识别 + 购买周期)
- “不良”商品识别模型
- 基于图书内容的 LDA 模型
- 用户行为加权组合的 SVD、SVD++

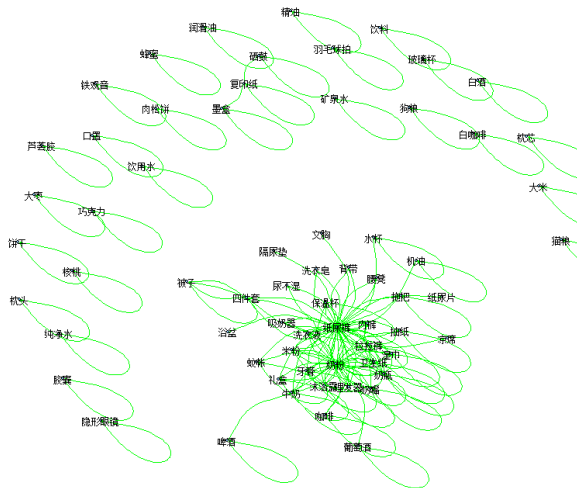
# 关于冷启动

对于“瓜子”我们应该推荐什么

1	1591_瓜子	1590_锅巴	1.000
2	1591_瓜子	1590_薯片	0.596
3	1591_瓜子	1590_花生	0.443
4	1591_瓜子	1591_开心果	0.318
5	1591_瓜子	1591_花生	0.274
6	1591_瓜子	1591_西瓜子	0.265
7	1591_瓜子	1591_腰果	0.235
8	1591_瓜子	1595_饼干	0.230
9	1591_瓜子	1590_豆腐干	0.227
10	1591_瓜子	1592_牛肉干	0.226
11	1591_瓜子	1594_口香糖	0.206
12	1591_瓜子	1591_炒货	0.204
13	1591_瓜子	1590_肉松饼	0.203
14	1591_瓜子	1671_卫生纸	0.172
15	1591_瓜子	1593_大枣	0.165



## 周期类商品 (部分)



# 作弊和反作弊

- 用户行为的复杂
- 过度 SEO
- 直接作弊

策略：

- 异常行为降权
- 异常用户直接过滤
- 点击流规则过滤

# 目录

## 推荐系统



- 1 京东推荐产品及架构
- 2 通用模型的应用
- 3 离线 CTR 预测实例
- 4 实验与监控

## 推荐的 CTR 预测

- 关联推荐的情境下，根据给定主商品推出的推荐商品，在用户浏览后被点击的概率。
- 可以理解为条件概率  $P(Y = 1|X)$

为什么要预测推荐商品的 CTR？

- ① 调整推荐商品的排序，推断潜在模式
- ② 多模型融合的方式
- ③ 发现影响推荐商品点击率的重要因素

# 特征表征方法

用目标问题所在的特定领域知识或者自动化方法来生成、提取、删减或组合变化来得到特征。

## 领域经验法

- 条件关系 ( $=, !=$ )
- 几何运算
- 分段及比例
- 其他

## 自动化技术

- PCA, ICA, NMF
- Linear Discriminant Analysis
- Collaborative Filtering
- AutoEncoder

## 最优子集 (Feature selection) 的优点

- 提高模型的可解释性
- 减少训练和预测的时间
- 有效降低过拟合，提升模型的适应能力

模型选用的是基于  $L1 + L2$  正则的 elastic net

## 最优子集 (Feature selection) 的优点

- 提高模型的可解释性
- 减少训练和预测的时间
- 有效降低过拟合，提升模型的适应能力

模型选用的是基于  $L1 + L2$  正则的 elastic net

# 如何对商品属性进行描述

## 对商品的形容：

品牌词、中心词、修饰词；类目属性、扩展属性；

## 基于用户行为的在商品上的反映：

- 销量、PageRank、评论数、好评度、浏览深度
- 商品的标签（如时间标签、地域标签、性别标签等）

对于商品标签（以时间差异构建的时间 feature 为例）：

假设 9:00 - 19:00 为白天 (D)，19:00 - 9:00 为夜间 (N)，则在这两个时间段内的用户购买则构成了该商品的时间标签，该商品标签的一般性定义为：

$$\frac{\sum_{u \in D} M_{u,i}}{\sum_{u \in D} M_{u,i} + \sum_{u \in N} M_{u,i}} - \frac{\sum_{u \in D} M_u}{\sum_{u \in D} M_u + \sum_{u \in N} M_u}$$



# 商品的组合属性

基于单一属性组合产生的属性，有以下三种：

- 相同类属性的组合：如时序上的销量（趋势系数），销量的方差
- 不同类属性的组合：如商品的展示和点击组合（如 CTR）、点击和购买的组合（如 CVR）
- 推荐主商品和推荐品属性的组合。比如品牌词是否一致，价格的比值是否在一定范围内。

推荐主商品和推荐品三级类目关系需要使用两两配对的 feature 表征形式。

1 VS 0

## 部分三级类组合系数展示

	前项	后项	权重
1	产后塑身	孕妇装	-1.55
2	月子装	孕妇装	-1.32
3	婴儿外出服	羽绒服/棉服	-1.28
4	水壶/水杯	洗衣液/皂	-1.27
5	宝宝洗浴	爬行垫/毯	-1.25
6	待产/新生	湿巾	-1.17
7	待产/新生	宝宝护肤	-1.13
8	婴儿鞋帽袜	防辐射服	-1.12
9	扭扭车	日常护理	-1.04
10	宝宝零食	钙铁锌/维生素	-1.00
11	日常护理	孕妈美容	-0.99
12	奶瓶奶嘴	驱蚊防蚊	-0.97
13	婴儿内衣	防辐射服	-0.97
14	婴儿鞋帽袜	摇铃/床铃	-0.97
15	滑板车	日常护理	-0.87
16	拉拉裤	婴幼儿奶粉	-0.87
17	奶瓶奶嘴	吸奶器	-0.85
18	婴儿尿裤	调味品	-0.84
19	婴幼儿奶粉	水壶/水杯	-0.84

# 目录

## 推荐系统



- 1 京东推荐产品及架构
- 2 通用模型的应用
- 3 离线 CTR 预测实例
- 4 实验与监控

# 实验配置平台

- 配置实时生效
- 任意百分比流量切换
- 可使用 random、partition by user 等策略分流
- 支持版本回溯
- 有权限管理体系



# 监控和报警

## 周期监控

- 按照一周为周期的推荐位指标监控，包括 PV、Click、OrderLine
- 推荐位实验级别的逐日监控
- 分品类的点击率监控（周单位）

## 实时监控

- 重点推荐位覆盖以及准确率监控
- 分钟级别
- 一旦异常邮件预警

## 效果跟踪：模型效率

实验ID	请求	展示	点击
0	21,122,504	18,642,676	1,262,533
100	2,840,850	2,465,516	179,912
101	1,376,364	1,191,754	83,400
102	228,428	200,583	15,352
103	18	0	0

实验ID	模型Tag	请求	展示	点击
0	31	420,212,225	371,504,024	2,794,269
	32	184,240,038	161,916,780	2,049,036
	33	10,998,370	9,701,956	109,560
	34	30,705,936	26,926,951	167,277
	88	116,665,785	101,582,356	235,038
实验ID	模型Tag	请求	展示	点击
100	0	51,677,045	45,608,618	415,691
	88	8,519,445	7,416,507	27,150



## 一些感受

- 推荐系统是完整的工程实现，算法 + 工程，二者缺一不可；
- 用户行为和业务的主要连接是数据，
- 数据的理解高于算法的理解，简单模型配以优质有效数据有更加的效果；
- 算法优化是逐步迭代的过程，更多需要的是灵感；
- ...

# I'm hiring!

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

Jump to first slide