推荐系统离线算法应用介绍

打造千人千面的个性化推荐引擎

推荐搜索部

刘思喆

2014 年 12 月 17 日

目录

推荐系统

JD.COM 京东

多 · 快 · 好 · 省

目录

推荐系统

京东推荐产品

- 90+ 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

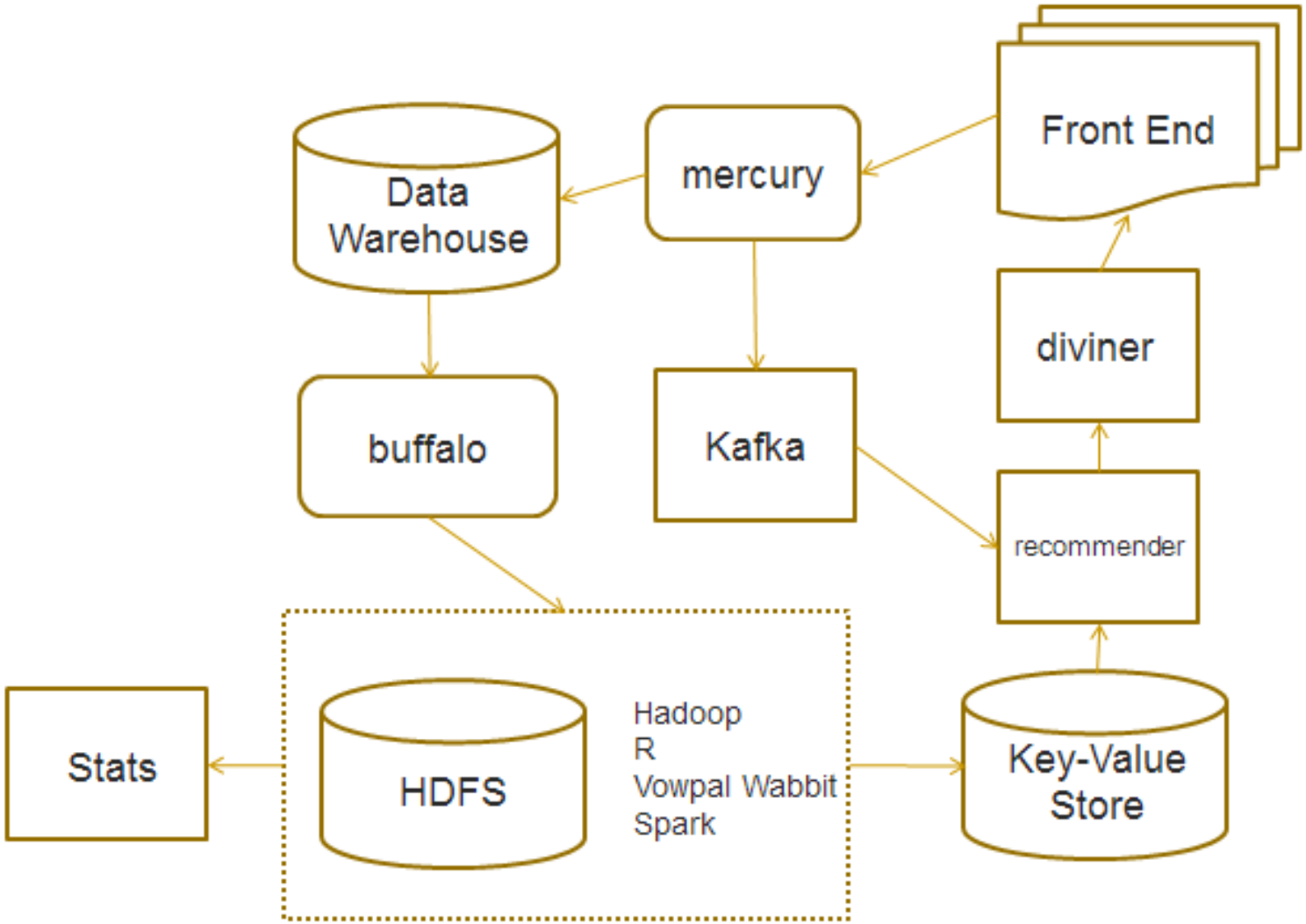

根据浏览猜你喜欢
金士顿（Kingston）8G Class4 TF（micro SD）存储
(已有466365人
¥26.90
购买该商品的用户还购买了
康纳(CONNAL) ZTD100K-SD1 嵌
直降 ¥1299.00
加入购物车
快易典H5 平板学生电脑 小学初中高中同步学习机 英
您可能还需要以下商品
赛尔贝尔（Syllable） G03-002 柯莉
¥59.00
(已有9111人评价)
加入购物车
购买了该商品的用户还购买了
海尔(haier) ES50H-Q1(ZE) 50
热水器
999.0
LG WD-N12435D 滚筒洗衣机
色）
2499.0
浏览了该商品的用户最终购买
康纳(CONNAL) CXW-200-TD08A
欧式 吸油烟机
¥1799.00
华帝(VATTI) JZT-i10008C 台嵌两用
式 燃气灶(天然气)
1299.0
关注此商品的人还关注
快易典EH1 电子词典 翻译机 小学初
¥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) 无敌V4牛津高阶+剑
¥669.00
加关注

不同位置的推荐产品定位不同

- 单品页：购买意图识别
- 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，意图发现
- 我的京东推荐：提高忠诚度

京东推荐系统架构


Front End
mercury
Data Warehouse
diviner
buffalo
Kafka
recommender
Stats
HDFS
Hadoop
R
Vowpal Wabbit
Spark
Key-Value Store

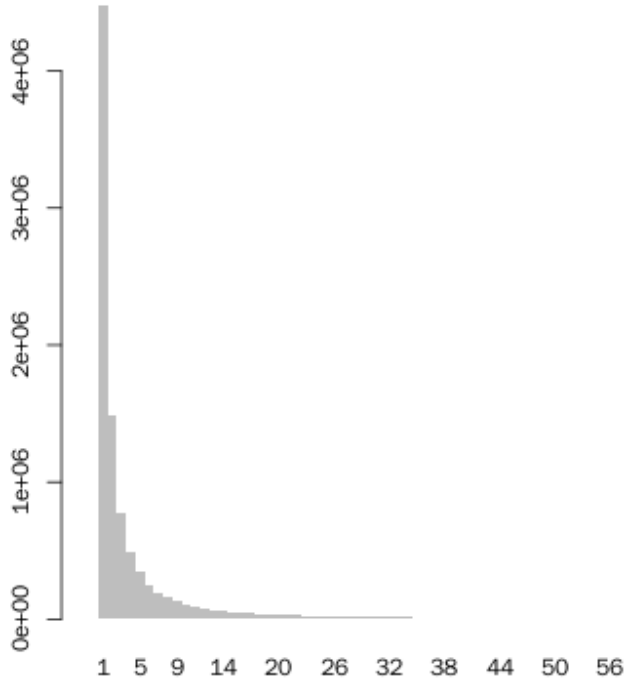
目录

推荐系统

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



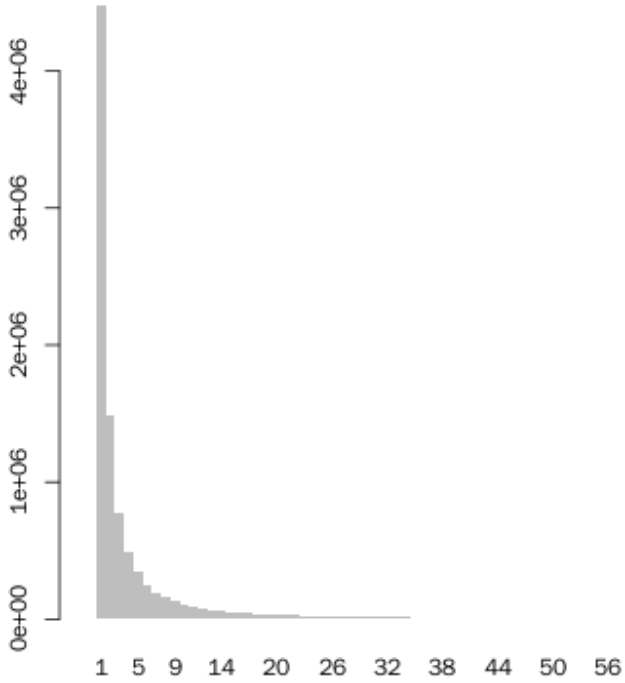
基于内容

- 标题
- 扩展属性
- 评论
- 描述
- ...

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

关联规则（Association Rules）

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Figure 1：关联规则三个重要指标的示例

频繁项集（以 FP Growth 算法为例）

transaction database	lexicographically sorted
a, d, e	a, c, d
b, c, d	a, c, d, e
a, c, e	a, c, d, e
a, c, d, e	a, c, e
a, e	a, d, e
a, c, d	a, d, e
b, c	a, e
a, c, d, e	b, c
b, c, e	b, c, d
a, d, e	b, c, e

prefix tree representation

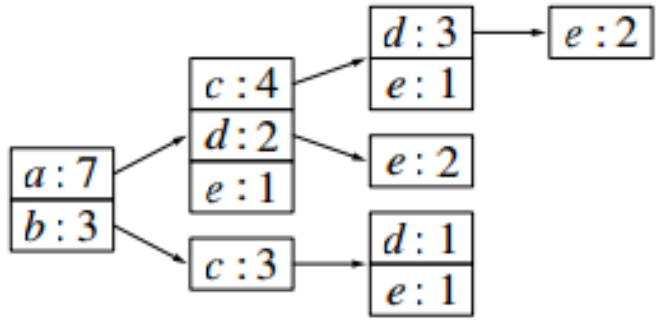


Figure 2：频繁项集是推荐系统中基础算法之一，很多推荐位都有体现

协同过滤

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) 向量大约有 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
- RA(AA)
- 欧 (闵) 式距离
- ...

序列挖掘

用户购买行为的序列 $(B \to AC)$ 是另一序列 $(AB \to E \to ACD)$ 的子序列，因为 $B \subseteq AB$ 且 $AC \subseteq ACD$，且序列的顺序没有改变

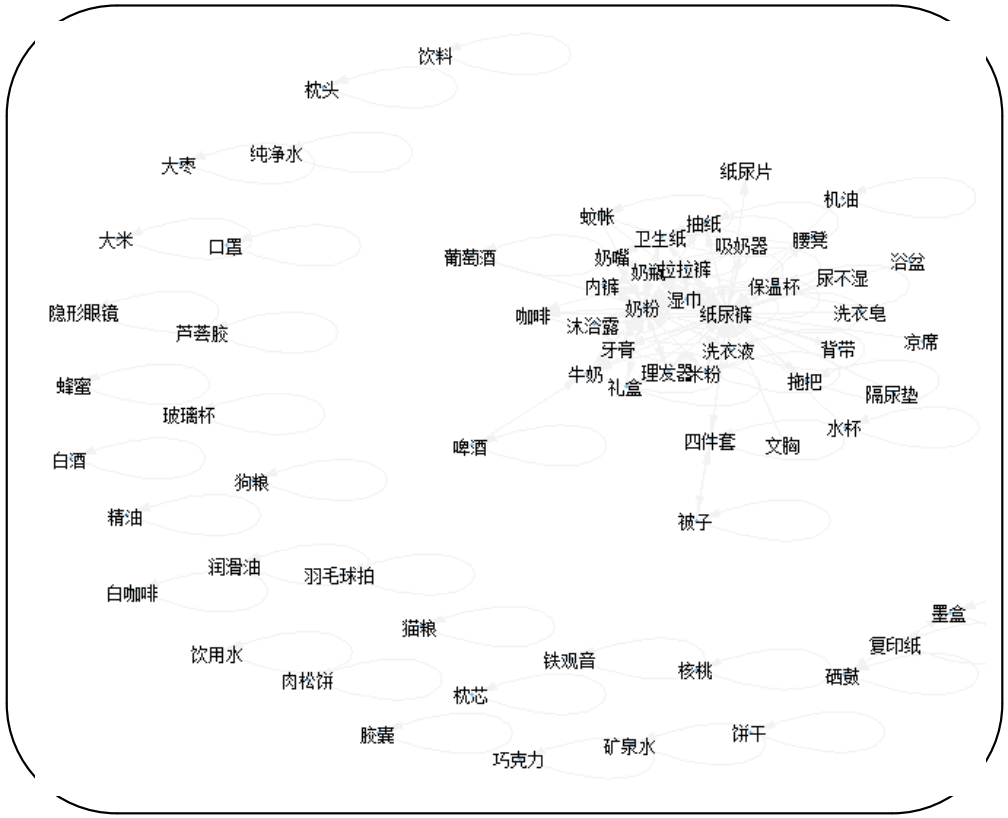


Figure 3：通过挖掘频繁发生的最小的子序列，即可得到被反复购买的商品。上图是根据京东实际订单挖掘的有周期性购买倾向的商品（包括非自身类）

基于内容的相似

- 图书简介（LDA）
- 标题（LSH）
- 扩展属性

目录

推荐系统

推荐的 CTR 预测

什么是推荐商品的 CTR（Click Through Rate）？

- 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
2. 用于多模型的融合
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

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
- 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：

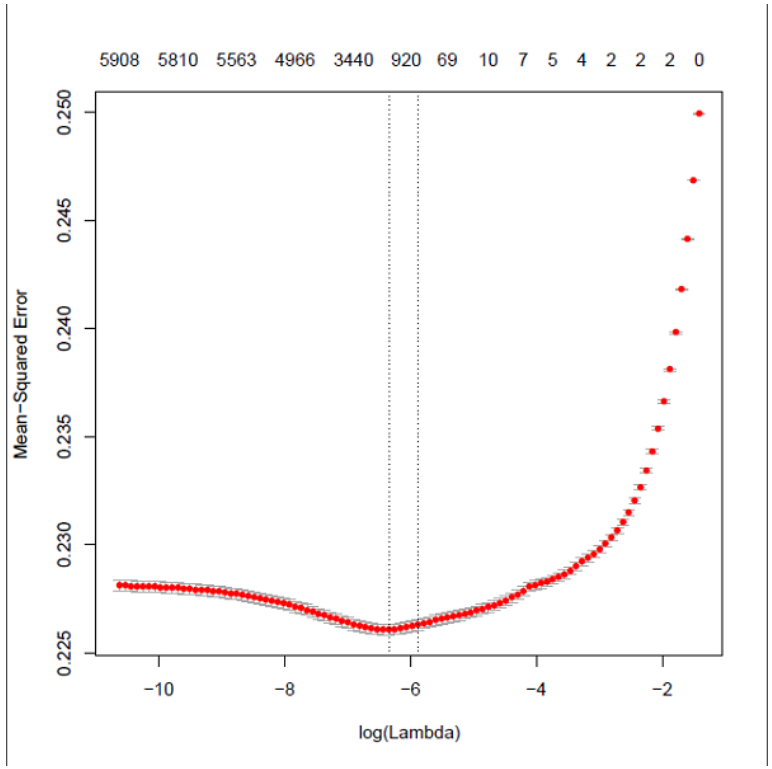
- 相同类属性的组合：如时序上的销量（趋势系数），销量的方差
- 不同类属性的组合：如商品的展示和点击组合（如 CTR）、点击和购买的组合（如 CVR）
- 推荐主商品和推荐品属性的组合。比如品牌词是否一致，价格的比值是否在一定范围内。

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。

数据预处理及建模过程

- 去掉样本量较小的类，共 25 个一级类需要预测
- 对不均衡样本采取了 undersampling 策略，同时配置 5 次重复抽样预测 (data.table)
- 训练数据量为 500w，在并行 CV 选取 λ 的时间为 15-20 分钟 (glmnet,doMC)
- 预测重排序数据为 6 亿条
- 预测所有数据，16 线程情况约为 1 小时 (snow)

不同 λ 交叉验证的 MSE 曲线



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	婴幼奶粉	水壶/水杯	-0.84

CTR 预测模型实验

1. 过渡页实验效果
   - 实验流量 10%
   - 请求点击率：提升 14%
   - 千次请求订单行数：提升 1%
2. 单品页实验效果
   - 实验流量 25%
   - CTR 提升 30%

目录

推荐系统

冷启动

三级类和三级类关系，产品词和产品词关系

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

对用户的降权

异常行为的用户产生的规则，在推荐中会被降权

商品的价格区间和性别

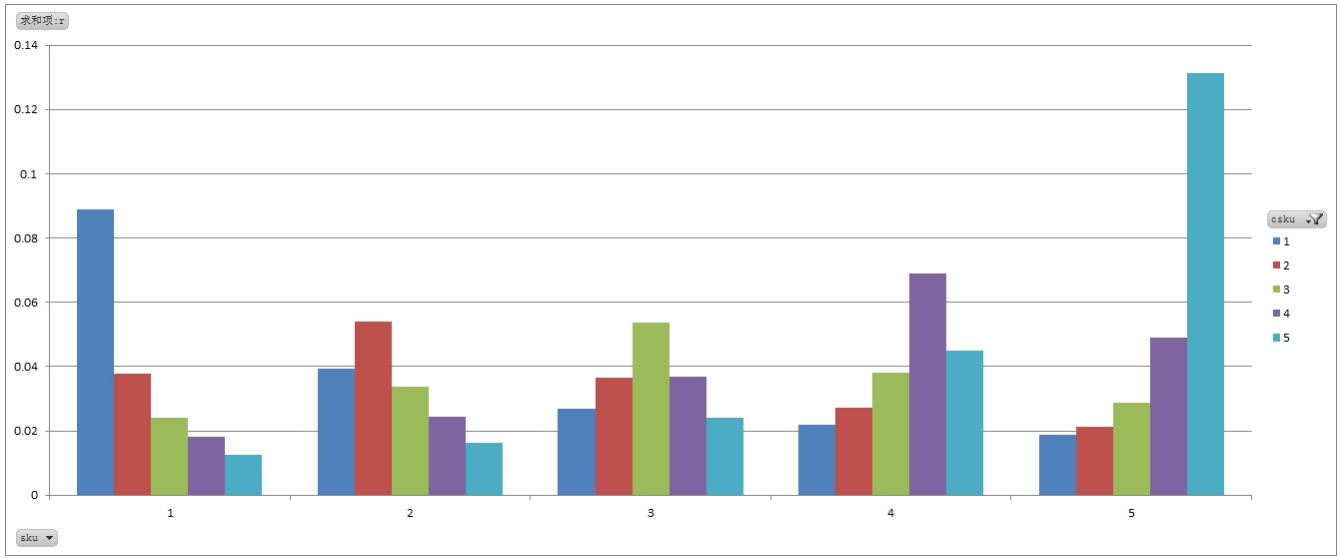


Figure 4：主 SKU 商品价格等级低（高）时，点击或购买的 CSKU 商品价格等级也低（高）。

总结和回顾

1. 推荐的优化是一个迭代过程
2. 巧妇难为无米之炊
3. 数据! 数据!
4. ...

未来算法团队优化方向

User Profile 系统 用户提取用户的所有已知信息，包括 demographic 以及推断信息；

商品属性系统 对商品的全方位描述，不限于人工录入体系；

Recall Model Profile 系统 离线召回模型的集合，适用于模型融合场景；

搜索推荐的技术融合 自然语言处理技术同推荐技术的混搭；

实时兴趣引擎 个性化和千人千面的基础

Q and A