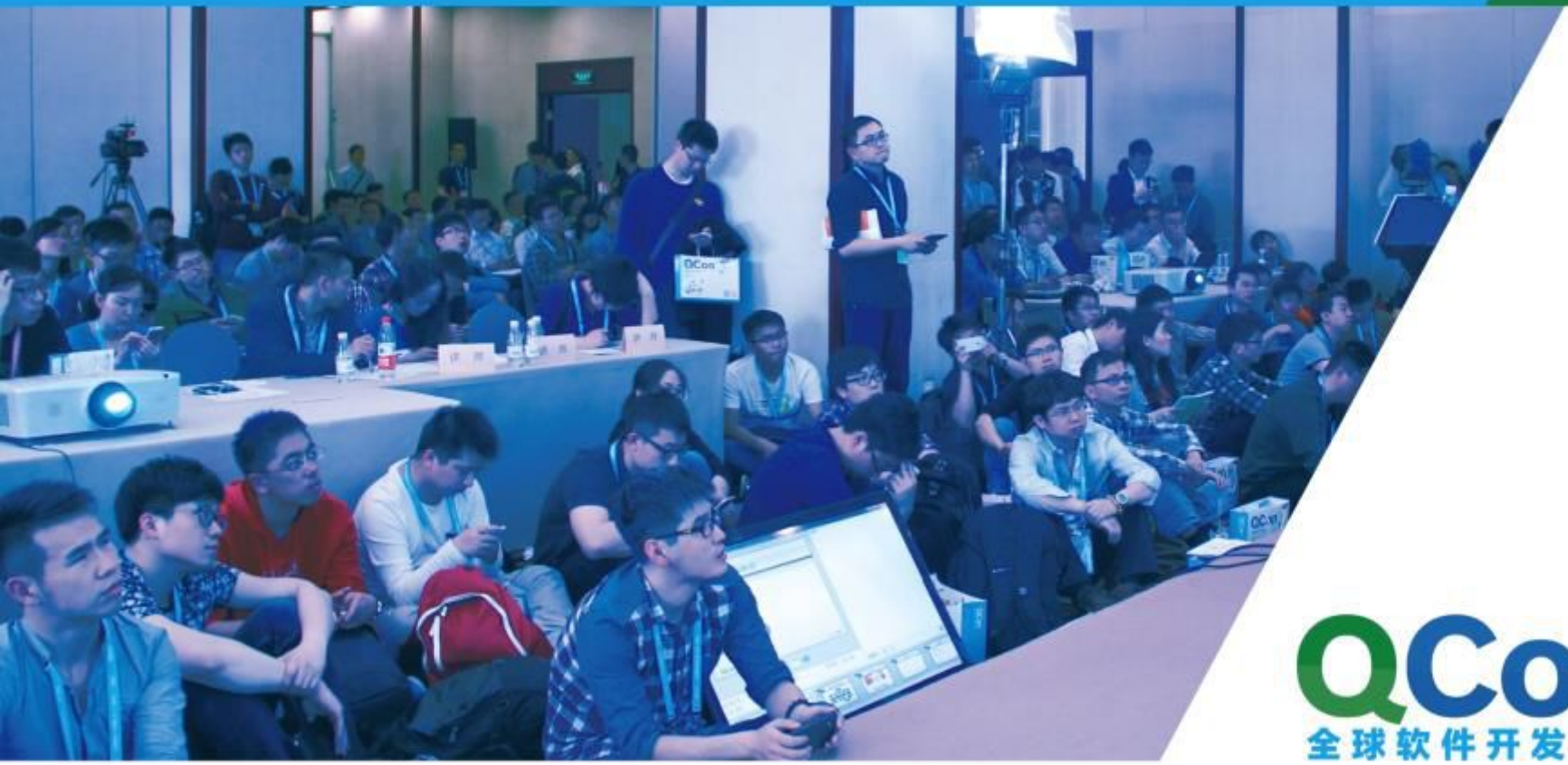
1号店通用精准化平台架构

陈敏敏
微信：24180823
邮箱：chenminmin@yhd.com
2015-10-17

QCon全球软件开发大会

International Software Development Conference



QCon
全球软件开发大会

主要内容

- 1号店精准化平台介绍
- 情景推荐和促销排期中的关联规则
- 精准化平台架构
- 平台中画像的技术实现原理
- 大数据营销案例

推荐产品架构

转化率 ↑
交叉销售 ↑

个性化推荐

推什么？

全面深挖用户购物兴趣

=

短期兴趣 转化率 ↑

+

长期累积兴趣 转化率 ↑ 交叉销售 ↑

+

潜在兴趣 交叉销售 ↑

用户意图引擎	用户画像引擎	千人千面引擎	CF推荐引擎
高转化率 精准定位用户短期兴趣 用于首页栏位、站外广告 EDM及APP营销投放	**长期累积的兴趣图谱** 1亿userID, 5亿GUID 每个用户的类目、品牌、导购属性兴趣偏好	**用户群体兴趣** 覆盖8千万左右用户 购买力level、同学(985,文科) 男/女、地域、同事、邻居、同行、好友群、辣妈、孕妇、新/老客	**相似用户兴趣** 协同过滤算法挖掘相似用户兴趣

什么时间推？

精准定位购物兴趣发生的时机

=

合适的场景 交叉销售 ↑

+

购物周期 转换率 ↑

+

最近热点 交叉销售 ↑

情境推荐引擎	反向推荐引擎	主题推荐
定位特定类目购物情境 公历农历的季度、月和周、中西节日、当地气温变化、当地天气、外出、旅游	**定位复购时间点** 定位复购时间点 覆盖74个高复购率类目 将用户划分成新客、成长期、衰退期、流失期多个阶段，投放不同的类目及负毛利营销品	**用户行为兴趣** 根据评论、标题给商品和用户打标签形成场景词 形成主题场景聚合SKU

拉升销量和用户转化率 ↑ 拉升交叉销售GMV ↑ 提升精准化营销效率 ↑

精准化通用平台后台系统

显示栏位配置

页面ID	页面名称	栏位ID	栏位名称	栏位推荐品数量	栏位状态	操作
4	H5产品详情页	53	测试栏位二排1	156	可用	
1	一号店产品详情页	54	验证栏位和算法修改	20	可用	

显示流程配置

流程ID	流程名称	推荐方式	是否团购	是否闪购	是否过滤库存	类目打散方式	品牌打散方式	推出商品数量	流程可用状态	操作
58	测试用户画像修改	用户画像	否	否	是	无	无	12	可用	
59	测试选品池	带故事情景的纯选品池	是	是	是	交替排序	随机打散	11	可用	
60	测试看了还看	相似相关算法库	否	否	是	随机打散	交替排序	10	可用	

效果预览

当前栏位 验证栏位和算法修改

当前算法名称 测试用户画像修改　　当前算法优先级 2

当前算法流量控制 0 至 0　　当前算法推出商品数 12

当前算法名称 测试选品池　　当前算法优先级 2

后台系统生成猜你喜欢

天气维度

换季、气温、雨雪、雾霾

【覆盖全国2954个市、县、区】

节日维度

农历节日：春节、端午、中秋、节气等

西历节日：元旦、国庆、父亲节、母亲节

大促：双11，双12，12.21，店庆

【覆盖全年共50种各类节日】

地域维度

大区：东北、华中、华东、华南等

旅游地、城市级别、小区，公司，

小区档次，学校类型，公司类型

【覆盖全国378个地级市或区】

时间维度

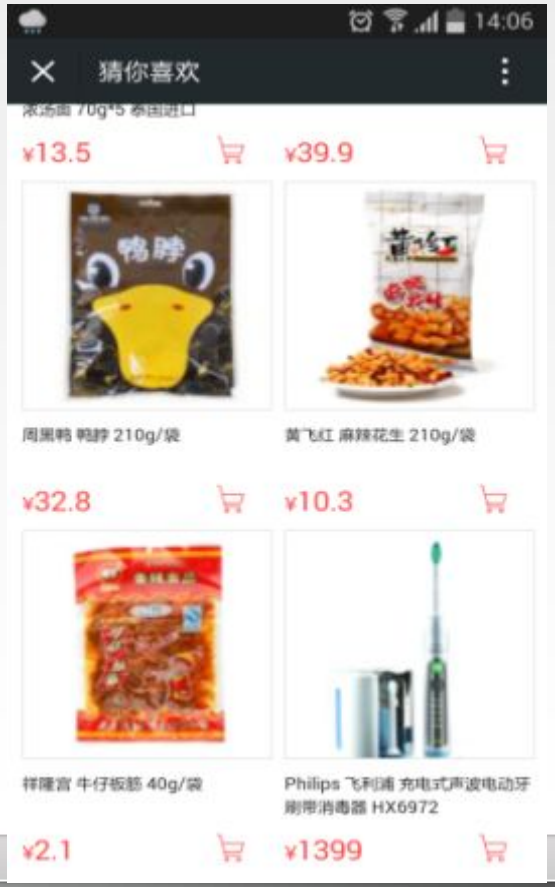
月份、季节、星期

画像和产品维度

性别，促销敏感，校园、公司、一贵就赔等

推荐数据

相似相关产品，类目等



情境推荐引擎

从5.8亿订单相关数据中
共挖掘210万条有效规则
覆盖88.3%1号店用户，含新客

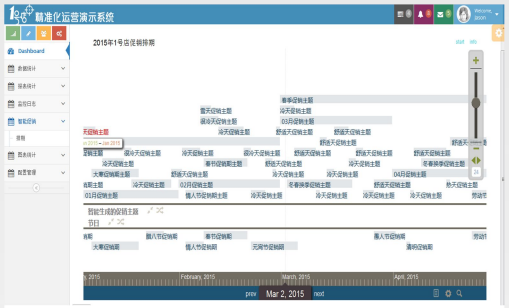
情境维度

天气维度
换季、气温、雨雪、雾霾
【覆盖全国2954个市、县、区】

节日维度
农历节日：春节、端午、中秋、节气等
西历节日：元旦、国庆、父亲节、母亲节
大促：双11，双12，12.21，店庆
【覆盖全年共50种各类节日】

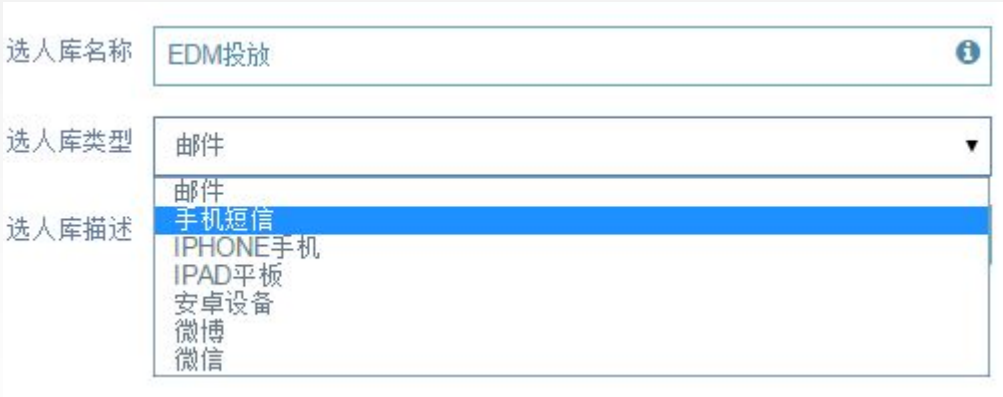
地域维度
大区：东北、华中、华东、华南等
旅游地、经济区、行政级别、城市级别、人口规模
【覆盖全国378个地级市或区】

时间维度<
月份、季节、星期
【覆盖全年】

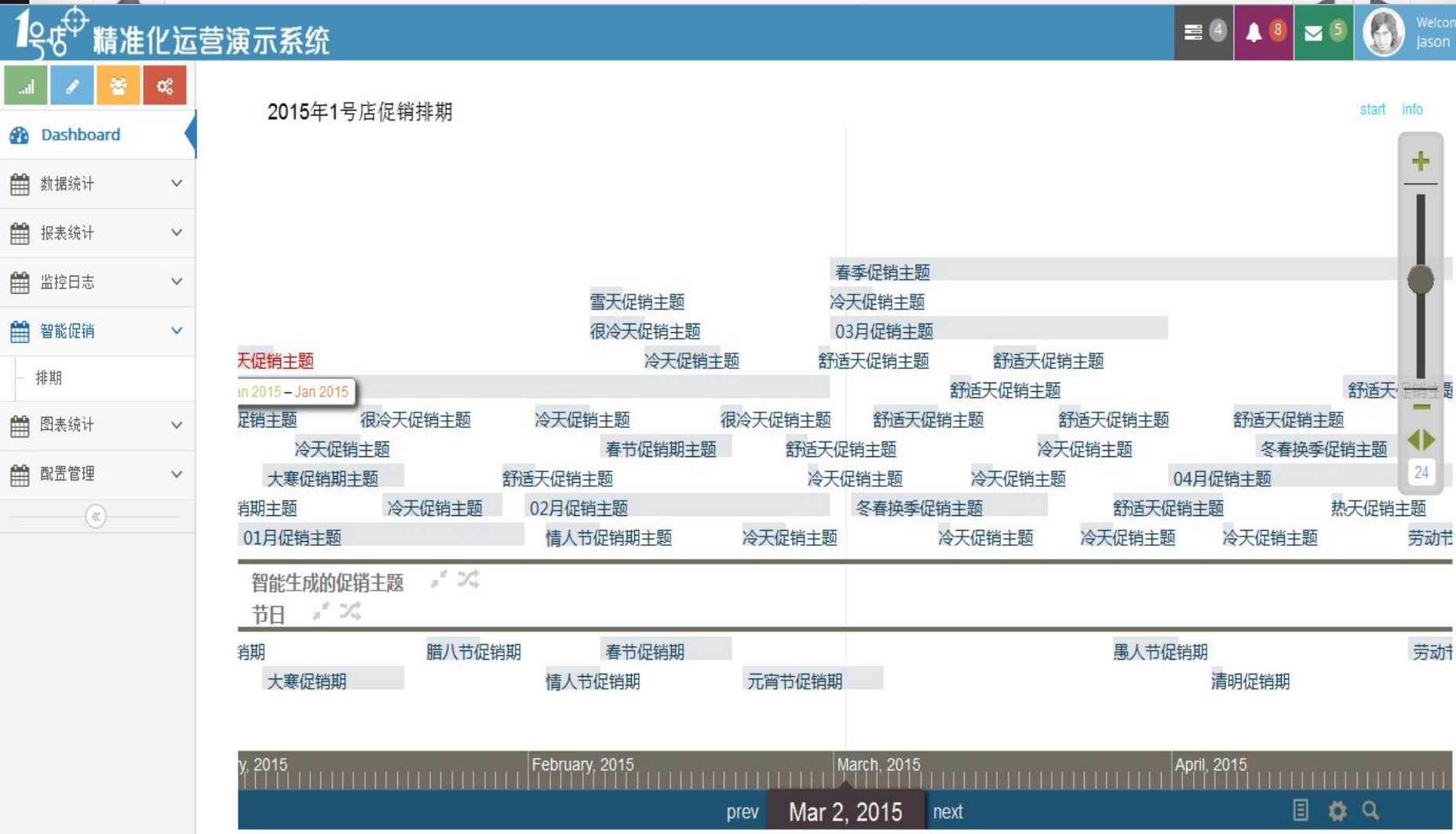


基于情境的投放排期PoC

小区雷购每周选品
按节日、天气、地域、季节、月份对小雷8个城市的每周选品提供3份SKU list给采购以参考，最终采纳的SKU选择在60%以上



基于情景的促销排期系统



端午节自动排期促销类目

Jun 12, 2015 – Jun 20, 2015 端午节

1 选择类目 — 2 选择SKU — 3 智能生成促销主题 — 4 完成

每页显示 10 条　　搜索

类目搜索　批量添加

编号	类目id	类目名称	类目层级	销量	销售额	区段
1	8704	保健滋补/成人计生	c1	35030	779836.76	12
2	968904	破壁灵芝孢子粉	c4	1097	24975.98	9
3	968998	芦荟制剂	c4	1084	10552.12	9
4	968664	植物精华/提取物	c3	5720	139508.06	14
5	957306	大家电	c2	4752	14311960.13	19
6	968876	灵芝2	c3	1299	26466.5	11
7	969013	叶黄素	c4	819	2180.68	8
8	954872	男装	c2	16277	1108216.74	26

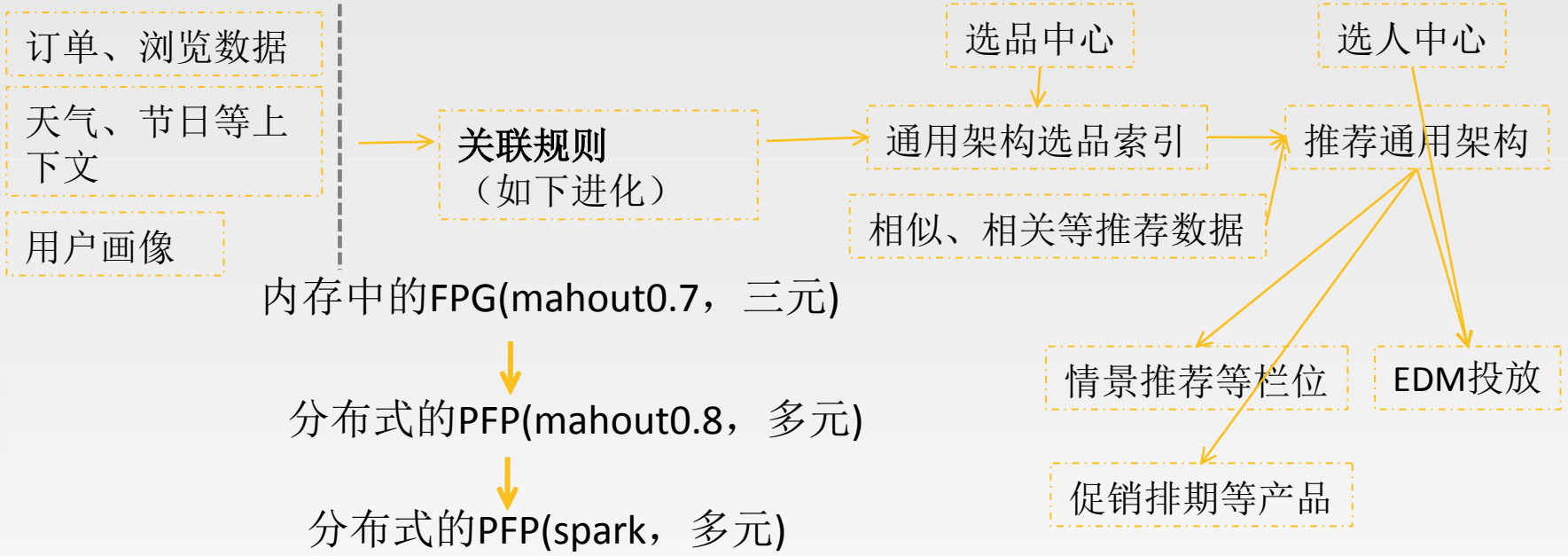
第 1 到 8 条，共 8 条　　上一页 1 下一页

商品类目

- 保健滋补/成人计生
  - 传统滋补营养品
    - 灵芝2
      - 破壁灵芝孢子粉
  - 保健品/膳食补充剂
    - 植物精华/提取物
      - 芦荟制剂

根据类目筛选相关商品做促销

关联规则在精准化通用架构中的位置



显示栏位配置

页面名称	栏位ID	栏位名称	栏位配置ID	栏位推荐品数量	栏位状态	操作
一号店相似商品页	7	相似商品页-精选搭配栏位	8	8	可用	
H5产品精准化聚合页	9	H5聚合页精准化栏位	9	8	可用	
H5产品精准化聚合页	10	H5产品精准化聚合页-购物车搭配推荐	10	20	可用	

Spark关联规则经验

基于英特尔的开源项目修改和调优：

- 分组后数据先reduce，避免fpgrowth树循环次数太大，提前减少运算量
- 我们场景partition里面重复key比较多，不用groupByKey，没有本地combine，shuffle消耗太大，用reduceByKey，combineByKey，foldByKey
- num-executor和executor-core的调整

--executor-cores 20 --num-executors 15

--executor-cores 5-num-executors 60　拥有lower gc,更快

- 分组数越大速度越快，但是相应的分组时间增加

data set size :10000 took 5.629 seconds.

data set size :110000 took 151.805 seconds.

data set size :3403606 took 8157.69 seconds

- persist和checkpoint，步骤出错避免全部重新计算

persist(StorageLevel.DISK_ONLY)

Spark关联规则经验

- 数据用int替代string

70G压缩到10G左右

- sparkContext.set("spark.shuffle.consolidateFiles", "true")
- sparkContext.set("spark.default.parallelism", “300")

shuffle过程调成300个task，默认8个无法handle那么多的数据，值不是越大越好，每一个core分配2到3个任务比较好。

- 修改JAVA_OPTS
- 修改序列化为Kryoserializer

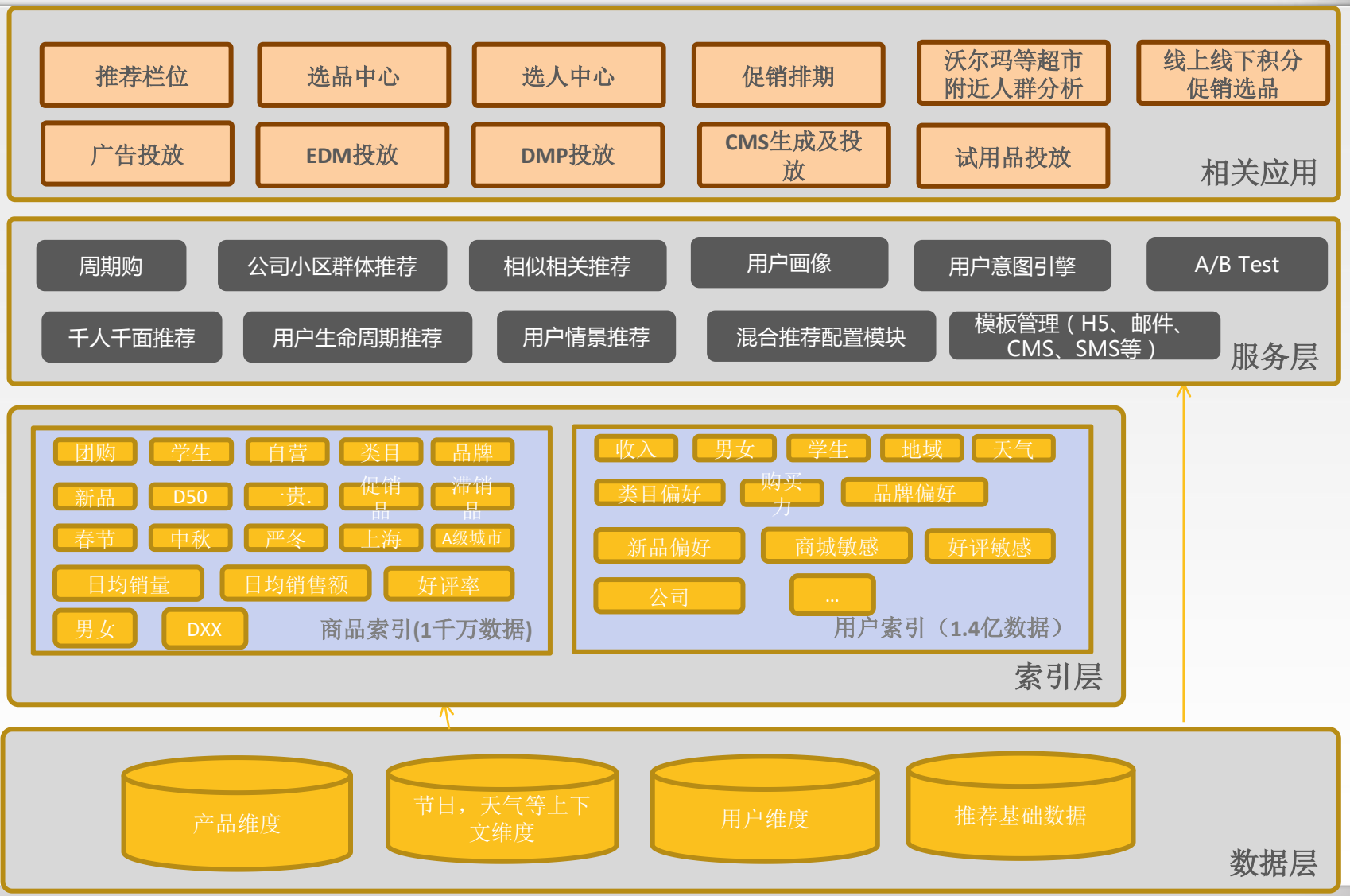
加快序列化时间，使得结果更为紧凑

测试结果
数据集：5亿条交易记录
测试参数：--num-executors 15 --executor-cores 8 --executor-memory 20G
支持度2000，耗时96分钟，频繁项3653447
支持度200，耗时70分钟，频繁项15603604

Spark和Hadoop关联规则比较

算法	测试集和结果
内存中的FPG	5亿条购买数据（70G），分5000组，取频繁模式的top K（目前取80），得到二元,三元频繁项集100多万条，丢失了部分挖掘到的频繁项集，耗时1.5小时左右
分布式的PFP	2.7亿数据，分5000组，top K取20w，支持度设置10个维度，耗时1小时
Spark	5亿购买数据（70G），从18个维度中选择了10个维度进行关联分析，分1000组，支持度设置为2000，挖掘到所有满足支持度的多元频繁项集3,653,447条，耗时96分钟。

精准化通用架构


推荐栏位
选品中心
选人中心
促销排期
沃尔玛等超市附近人群分析
线上线下积分促销选品
广告投放
EDM投放
DMP投放
CMS生成及投放
试用品投放
相关应用
周期购
公司小区群体推荐
相似相关推荐
用户画像
用户意图引擎
A/B Test
千人千面推荐
用户生命周期推荐
用户情景推荐
混合推荐配置模块
模板管理（H5、邮件、CMS、SMS等）
服务层
团购
学生
自营
类目
品牌
新品
D50
一贵.
促销品
滞销品
春节
中秋
严冬
上海
A级城市
日均销量
日均销售额
好评率
男女
DXX
商品索引(1千万数据)
收入
男女
学生
地域
天气
类目偏好
购买力
品牌偏好
新品偏好
商城敏感
好评敏感
公司
...
用户索引（1.4亿数据）
索引层
产品维度
节日，天气等上下文维度
用户维度
推荐基础数据
数据层

商品索引切分与路由

切分

- 按照顶级类目切分
- 关联类目分为一组

路由

- 类目按组路由
- 搜词按照“词->类目”先解析

类目	类目	类目页 pv	PV 占比	产品数	产品占比
食品饮料、酒水、水果	5135	6, 157, 973	0. 243	135121	0. 046
进口食品、进口牛奶	8644	2, 105, 485	0. 083	38762	0. 013
保健滋补、器械计生	8704	528, 165	0. 021	42851	0. 015
统计-0			0. 347	216734	0. 074
美容化妆、个人护理	5009	1, 751, 630	0. 069	102340	0. 035
母婴用品、玩具	5117	1, 317, 868	0. 052	78536	0. 027
厨卫清洁、纸、清洁剂	5134	1, 109, 023	0. 044	28114	0. 010
家居家纺、锅具餐具	950340	1, 084, 479	0. 043	238088	0. 082
统计-1			0. 208	447078	0. 153
生活电器、汽车生活	21266	1, 568, 945	0. 062	69952	0. 024
手机通讯、数码电器	21306	2, 216, 454	0. 088	63669	0. 022
电脑、软件、办公用品	21392	797, 851	0. 032	61882	0. 021
统计-2			0. 181	195503	0. 067
服饰内衣、鞋靴运动	123	4, 850, 002	0. 192	295776	0. 101
运动户外	34140	487, 273	0. 019	71515	0. 024
箱包皮具	22906	462, 207	0. 018	38490	0. 013
珠宝、饰品、手表、眼镜	32258	496, 913	0. 020	94836	0. 032
统计-3			0. 249	500617	0. 171
图书杂志	25228	2, 309	0. 000	1342491	0. 460
统计-4			0. 000		0. 460
生活服务	33078	728	0. 000	24356	0. 008
创意礼品中心、礼品卡	5342	365, 359	0. 014	194212	0. 066
统计-5			0. 014	218568	0. 075
统计-all		25, 302, 664		4280923	

注：图为14年初的老数据

用户索引切分

- 按用户id取模24

Shard命中记录数列表 | 查询结果

	Shard命中数记录
1	10.4.37.112:8080 s20, hit 4933631, group 0, facet false
2	10.4.7.81:8080 s1, hit 4933278, group 0, facet false
3	10.4.7.81:8080 s13, hit 4933572, group 0, facet false
4	10.4.21.37:8080 s16, hit 4933347, group 0, facet false
5	10.4.7.81:8080 s15, hit 4933022, group 0, facet false
6	10.4.37.112:8080 s21, hit 4933361, group 0, facet false
7	10.4.7.81:8080 s3, hit 4933407, group 0, facet false
8	10.4.21.37:8080 s19, hit 4933181, group 0, facet false
9	10.4.37.112:8080 s9, hit 4933712, group 0, facet false
10	10.4.37.112:8080 s10, hit 4933073, group 0, facet false
11	10.4.37.112:8080 s8, hit 4933428, group 0, facet false

- 利用HBase的特性，按用户id数均分

```
hbase(main):001:0> count "user_profile_base",INTERVAL=>5000000
Current count: 5000000, row: 116383103
Current count: 10000000, row: 126797118
Current count: 15000000, row: 137147958
Current count: 20000000, row: 146769164
Current count: 25000000, row: 152306173
```

千人千面引擎

用户群体千人千面

校园标签【244万校园用户

按校园收货地址、IP、GPS聚合，形成同学群
含985/211、综合/师范/理工/文科/医药/语言

亲朋好友群

按用户手机通讯录聚合客户的亲朋好友及同事，
形成社交网络图

公司标签/同事群

按单位收货地址的工作单位聚合，形成同事群
行业、公司规模、收入档次

小区标签/邻里群

按小区收货地址及坐标的聚合，形成邻居群
小区名、楼盘档次

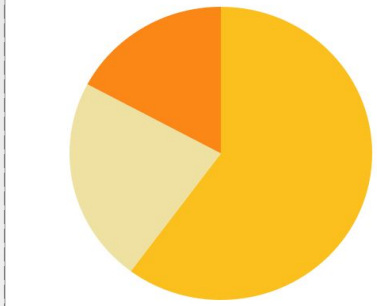
活动地点标签

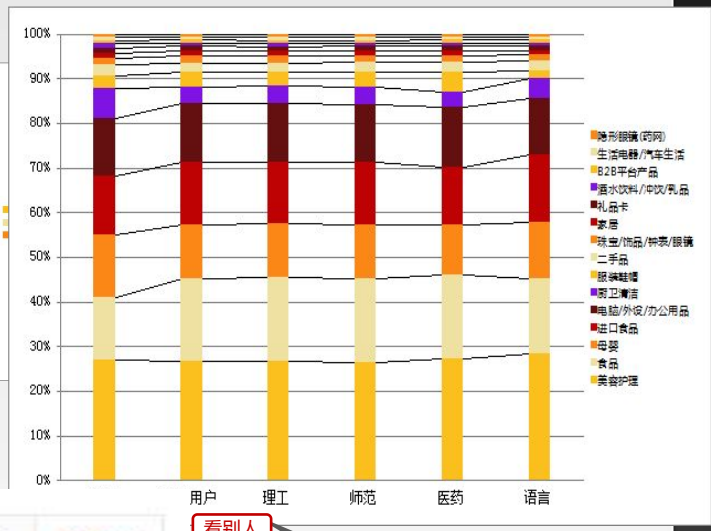
关注重点购物中心、旅游地，推荐土特产，
聚恒隆女，沃尔玛超市附近

辣妈群、性别、购买力等标签

用户画像基础标签

校园、小区、公司用户画像标签





可以换成其它千人千面，如同城、同事、小区、同城孕妇

1号店详情页

用户标签画像

公司、小区、校园标签：

用户群体	数量
公司	覆盖3558家公司，591个行业
小区	覆盖293个城市的4.26万个小区
校园	覆盖全国1334所高校

校园、小区千人千面引擎优化上线

完整的地址处理系统包含三部分：

- **地址结构化**
- **命名实体识别**

公司名识别模型的F1值（提高到80.6%）

- **地址匹配**

地址结构化

- **地址结构化**

第一版：后缀激活二元的Bi-Gram + 自动机识别地名

第二版：

1.抓取高德地图全国三千万条POI信息形成地址词库：

2.没有识别的利用第一版中后缀去激活相应的BiGram模型，如果通过了BiGram的校验，则进行召回。

- **命名实体识别**
- **地址匹配**

命名实体识别步骤

- **命名实体识别**

1.通过赶集网等构建小区名、公司名地址库（省、市、区、小区名称、房价、经纬度）

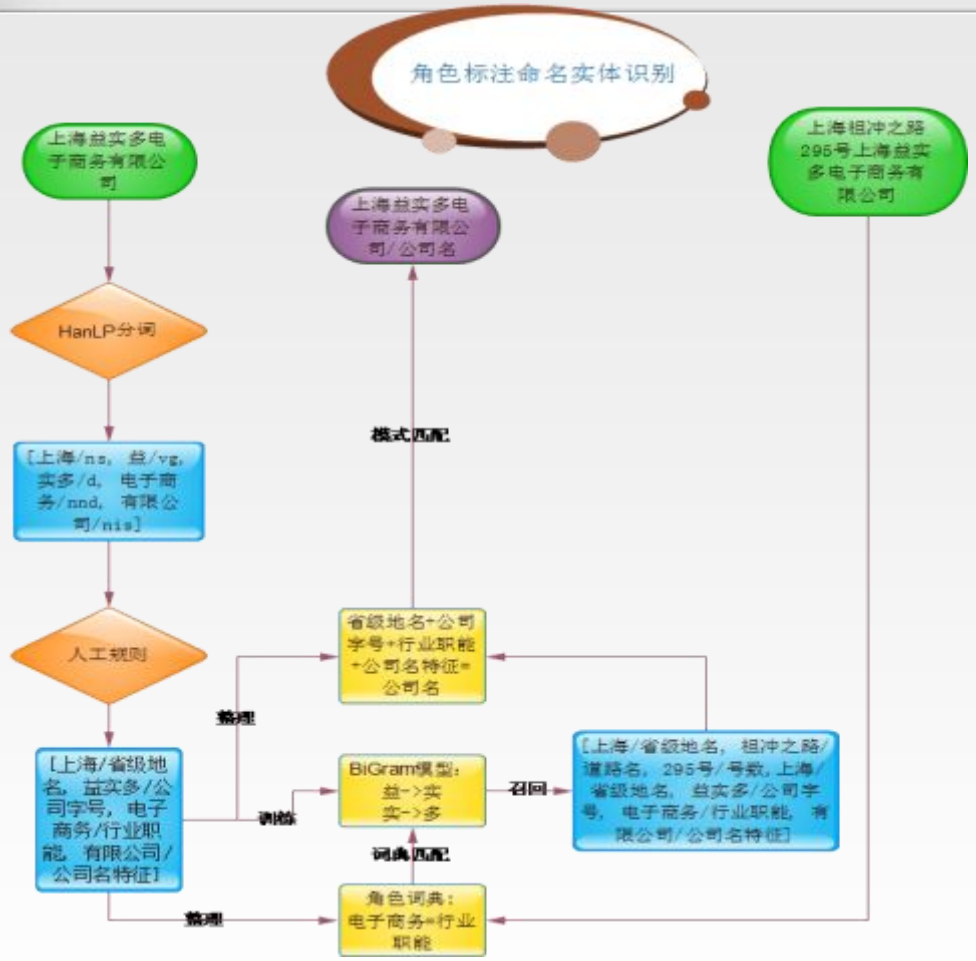
```
黑龙江 哈尔滨市   南岗区    科大小区   6570   126.63725  45.721832
黑龙江   哈尔滨市   null   科技公寓   6818   126.704511 45.748821
黑龙江   哈尔滨市   道里区    穆斯林小区  6455   126.56034  45.713382
```

2.角色标注法（ CRF、HMM等模型需要人工标注大量语料，少量语料没法命中几千万的命名实体）

编码	意义	例子
1	乡镇级地名	轸溪乡、轵城镇
2	乡镇级地名后缀	乡、镇
32	道路名后缀	路、环路
33	隧道	复兴东路地道
34	高速路入口	杨高南路立交桥入口
35	高速路出口	松卫北路出口

整理“地名——角色词典”，将大多数地名标注为准确的角色

命名实体识别之角色标注



1.从百度地图POI库中抓取城市的所有公司名和学校名，对其分词，人工编写规则对分词结果进行处理，转化为一个角色标注序列

2.收集所有的词语和角色，得到一个角色词典，并对角色去除后缀训练对应的NGram模型

3.收集所有的角色标注序列中的角色序列，制定为一个规则集

4.对于输入的一个订单地址，先使用角色词典标注，然后使用相应后缀的NGram进行召回，得到一个角色标注序列。对一个角色标注序列，满足上述规则集的就是一个命名实体。

命名实体识别之试验

输入："上海祖冲之路295号3楼上海益实多电子商务有限公司“:

角色标注： [上海/省级地名, 祖冲之路/道路名, 295号/号数, 3楼/层数, 上海/省级地名, 益实多/公司字号, 电子商务/行业职能, 有限公司/公司名特征]
命中模式串：省级地名+公司字号+行业职能+公司名特征=公司名[上海/省级地名, 益实多/公司字号, 电子商务/行业职能, 有限公司/公司名特征]
地址抽取： [上海/省级地名, 祖冲之路/道路名, 295号3楼/号数, 上海益实多电子商务有限公司/公司名,]
==========识别结果==========
上海益实多电子商务有限公司/公司名 [上海/省级地名, 益实多/公司字号, 电子商务/行业职能, 有限公司/公司名特征]

地址匹配

- **地址匹配**

需要知道一条订单地址对应地址库中的哪一个实体，采用的是基于树的匹配方法：

1.通用字典树
2.面向地址的字典树
有如下特点：
2.1 有些用户喜欢跳过部分地址段，比如省市区跳过市直接填区，这会导致city为null，地址树可以自动跳过这些null，将区这个节点直接挂在省节点上。
2.2 有些地址对应多个实体，比如一栋楼里面有很多公司，地址树被改造为支持多value。

地址匹配之地址树

面向地址的字典树：

```
AddressTrie<String> trie = new AddressTrie<String>();
trie.put(“1号店", "祖冲之路", "295号");
trie.put("纽海信息技术(上海)有限公司", "祖冲之路", "295号");
trie.put("某家公司", "祖冲之路", "300号");
System.out.println(trie.get("祖冲之路", "295号"));
```

输出：

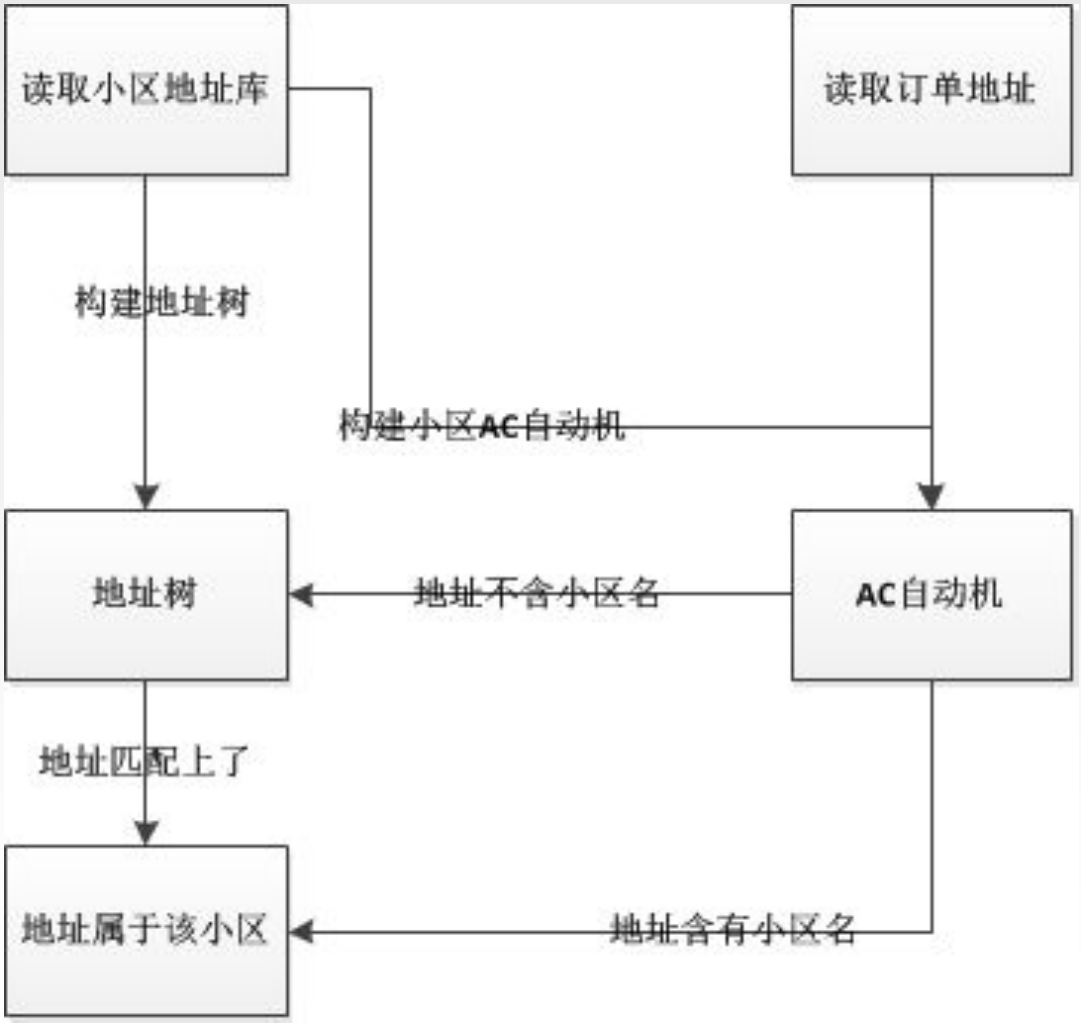
[1号店,纽海信息技术(上海)有限公司]

基于编辑距离相似度的地址树（允许每个地址段有少量的不同）

```
AddressTrie<String> trie = new FuzzyAddressTrie<String>();
trie.put(“1号店”, “祖冲之路”, “295号”);
trie.put(“纽海信息技术(上海)有限公司”, “祖冲之”, “295号”);
```

默认的相似度阀值是60%以上，可以调用setFuzzy来调节。

匹配小区的地址库形成画像标签



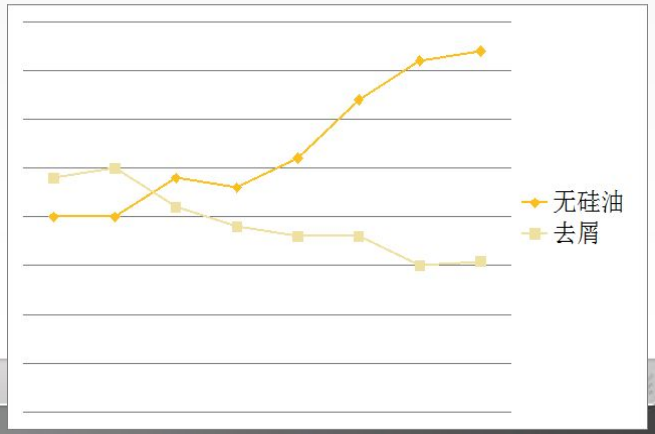
输入小区地址库和结构化订单地址，输出这些地址与小区的对应关系

用户画像在大数据营销中的应用

根据画像的校园和偏好标签做营销：
男生买女性用品销量=》暖男排行
零食销量=》吃货排行
化妆品销量=》颜值排行
单反等销量=》潮人排行
安全套销量=》性福排行；
等等。

捕捉到用户画像标签属性的变迁调整新品

用户偏好画像的标签是通过用户的搜索、浏览、购买等所有的站内行为计算而来，针对标签的监控，可以体现用户的喜好和关注度的迁移变化。



THANKS