推荐系统产品需求文档

——V4.0 版本 潘一鸣

1. 背景

1.1 外部环境

目前京东和天猫、唯品会等大型电商公司都引入千人千面测策略，同时取得了很好的效果，在电商公司进入精细化运营的大背景下，我们也需要建立自己的推荐系统。

1.2 目标位置

首页推荐模块：位置会在爆款之下召回部分有限数量的商品（暂定 10 个）。作为为你推荐模块。底部依旧使用首页单品团专场混排的形式。

1.3 首页现状

因为品类扩充已经比较多，但是依旧使用排期的方式，首页目前面面俱到，面面俱到的结果是对任意一个用户而言，首页都充斥的大量的不感兴趣的内容，导致用户对于首页的浏览越来越少。几个典型的 case 如下：

- 一般的年轻在校生，首页前两屏出现纸尿裤。
- 南方不需要购买毛裤的用户，首页前两屏出现毛裤。
- 只购买高端兰蔻雅诗兰黛 YSL 的用户，首页充斥着悦诗风吟等大众品牌。

2. 项目目标

提高首页 CTR，提高用户阅读深度。提高用户访问时长和访问频次。核心衡量制表位，提高首页 CTR 和用户的阅读深度。

3. 具体方案

3.1 主要思路：

- 通过算法在首页召回一定数量推荐商品，Top N 依靠推荐因素和业务数据综合确定。
- 商品标签为商品的品牌和分类，用户标签为根据用户发生行为的商品标签计算出的用户的标签，已经用户对于每个标签的偏好程度。
- 根据标签命中情况推荐商品和提权进行排序。
- 使用数据除了商品数据外，还需要引入冷启动引导的数据。

3.2标签的分类

- 普通标签【召回，排序】

  商品三级分类；品牌；商品产地，品牌调性。（如：兰蔻。一个标签关联有限商品，一个标签标签关联部分用户）
- 特殊指标【排序】（一个标签关联全部用户和商品，但是分数不同）

  消费指数：用户消费指数匹配商品消费指数
- 业务数据：商品数据销量、流量数据

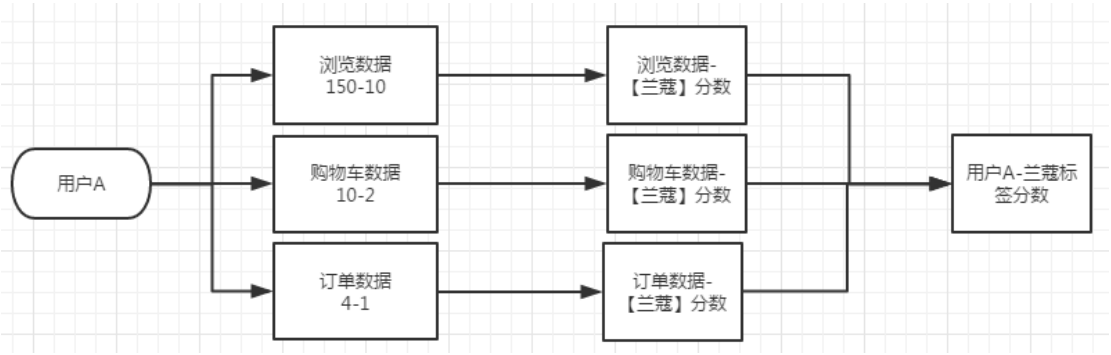
3.3构建商品标签：

基本按照现在的商品信息使用脚本，给出标签规则。比如：轻奢，食品等。

商品的标签暂定：

- 品类标签：根据商品的分类进行归类，选取三级进行分析
- 品牌标签：根据商品的品牌进行归类
- 商品产地标签：根据商品的产地进行归类
- 品牌调性标签：根据品牌调性进行分类

3.4构建用户标签



用户标签分类和商品标签分类相同，根据用户行为数据中的商品数据，作为用户喜好。

3.5用户标签数据收集的数据类型（注意数据优先级）

1. 【P0】浏览（view）：对最近 90 天浏览数据，分析商品作为数据源。
2. 【P0】订单（order）：对最近三年的订单，分析订单中的商品标签作为数据源。
3. 【P0】引导（cold start）：用户手动选择，作为数据源。
4. 【P1】购物车（shopping cart）：对最近 90 天加入购物车的商品，分析商品标签作为数据源。
5. 【P2】收藏（enjoy）：对于用户最近 90 天收藏商品，分析收藏商品中的标签作为数据源。
6. 【P2】订阅（subcribe）：对于用户最近 90 天订阅商品，分析订阅商品中的标签作为数据源
7. 【P2】心愿单（wish）：对于用户最近 90 天心愿单商品，分析心愿单商品中的标签作为数据源。

3.6推荐分数的计算

对于每个数据源有相应的权重，对于每个命中的标签需要进行打分。对于标签 i，和数据类型 j，用户某类型商品出现标签次数为u_{ij}：

每种数据来源类型下的标签分数计算：

【暂定，冷启动交互待确认】对于 j 属于冷启动类型的数据来源，其中 U 为用户选择的标签数量。t_k为数据 k 发生的日期。t_0新一轮推荐计算开始时间。ω为系数，暂时取 1。

$$s_{ij} = \frac{1}{\sqrt{U}} \cdot e^{-\omega|t_0 - t_k|}$$

对于 j 属于商品类型的数据来源，n 代表用户在数据来源类型 j 中的数据样本量，$lg(u_j + 1)$为数据可靠性提权，U_{ij}为在数据来源类型 j 中、用户u命中标签 i 的商品集合，U_j为数据来源类型 j 中用户u全部数据的结合，t_k为数据 k 发生的日期。t_0新一轮推荐计算开始时间。ω为系数，暂时取 1。

$$s_{uij} = \sqrt{\frac{\sum_{k \in U_{ij}} e^{-\omega|t_0 - t_k|}}{\sum_{k \in U_j} e^{-\omega|t_0 - t_k|}}} \cdot lg(u_j + 1)$$

所有数据来源类型下的标签分数综合计算：

γ_j代表数据类型 j 的权重，X代表月活跃用户数，X_i代表某个标签命中的用户数，Y代表月活跃用户数，Y_i代表某个标签的平均销售额，$\sqrt{\sum {\gamma_j}^2}$为归一化参数，$lg(X_i + 1)$表示对流行度高的标签进行降权，α_i代表对于标签 i 给出的业务权重。对于标签 i，用户的三级分类标签得分如下：

$$s_{ui} = \frac{\alpha_i}{\sqrt{\sum {\gamma_j}^2}} \cdot \frac{100}{lg(Y_i + 50)} \cdot \frac{1}{lg\left(\frac{X_i}{10000} + 4\right)} \sum_j \gamma_j \cdot s_{uij}$$

不包含流行度降权，算例如下：

数据类型	类型权重(γ_j)	数据量(n_j)	兰蔻商品数(u_{ij})	特定数据标签得分（Sij）	分数
浏览数据	1	150	10	0.562	
购物车数据	5	10	2	0.447	
订单数据	2	4	1	0.301	
兰蔻标签用户数为10000 业务权重为5				求和分数($\sum_j \gamma_j \cdot s_{ij}$)	1.757
				综合分数	0.401

3.7根据标签召回商品的策略

A：设置召回销售量阈值：月销售额大于 5000。

B：每个候选集召回商品数量最大为，暂按照销售排序取 TOP。M_c为分类的销售额，N_c召回商品数上限。

$$N_c = \left\lfloor \frac{\sqrt[3]{M_c}}{1.5} \right\rfloor$$

逻辑为：A 并 B。

召回标签只包括分类标签和品牌标签，分类标签最多召回 25 个分类的商品。品牌标签最多召回 15 个品牌的商品。

3.8消费能力指数计算

使用数据：用户的订单数据

分类商品队列确定原则：

三级分类下月销量大于 0 的商品大于 10 个，则分类商品队列为月销售额大于 0 的商品。

如果三级分类下月销售额大于 0 的商品小于 10 个，则此三级分类不计入消费指数的计算。（用户的此三级分类直接忽略，不进入a_u的）

商品 i 的三级分类为 C，p_i代表商品 i 的价格。L_{ci}表示商品 i 在分类 C 中的价格排名等级，$Max(p_C)$代表分类 C 中商品最高价，$\overline{pc}_j$代表分类 C 中平均价。U 代表所有分类的全集。对于商品 i 而言，消费能力指数为：

$$a_i = L_{ci} \cdot \sqrt{\frac{lg(\max(pc_i)+1)}{Max_{j\in U}\{lg(\max(pc_j)+1)\}}}$$

$$L_{ci} = \frac{\left\lfloor 5 \cdot lg\left(\frac{p_i}{20}+1\right)\right\rfloor}{\left\lfloor 5 \cdot lg\left(\frac{Max(p_C)}{20}+1\right)\right\rfloor}$$

对于用户而言，$N(u)$表示用户订单内所有商品的集合。N_u代表用户订单内的商品数， 用户 u 消费能力指数（订单内商品消费指数的平方平均数）为：

$$a_u = \sqrt{\frac{\sum_{i\subseteq N(u)} {a_i}^2 \cdot lg(\max(pc_i)+1)^2}{\sum_{i\in N(u)} lg(\max(pc_i)+1)^2}}$$

β代表业务参数，用于调整权重，暂定为 0.2。用户和商品消费指数相似度计算如下：

$$\theta = \frac{lg(N_u+1)}{\left(\left\lceil\frac{|a_u - a_i|}{\beta}\right\rceil + 1\right)}$$

3.9 品牌调性策略

品牌调性打分两位 1~4 分。分为四个标签。

品牌调性的计算策略和普通标签相同，但是最终只保留得分最高的品牌调性标签。

3.10 标签协同过滤策略

$N(i)$为购买过标签 i 的用户数，$N(j)$为购买过标签 j 的用户数，$N(i)\cap N(j)$为同时购买 i 和 j 的用户数，标签相似度定义：

$$w_{ij} = \frac{|N(i)\cap N(j)|}{\sqrt{|N(i)||N(j)|}} \cdot \frac{1}{1+\lg\left(1+\left(\frac{N(j)}{100000}\right)^2\right)}$$

比如w_{ij}标识手链和面膜的相关度，则$\lg\left(1+\frac{N(j)}{50000}\right)$中$N(j)$为购买过面膜的用户数。

s_{uj}表示针对用户 u 计算出来的标签 j 的分数。p_{uj}表示针对用户 u 考虑标签 CF 参数后的标签分数。标签 i 为用户得分最多的 C 个标签（只有这些标签作为协同召回的根标签），C 暂定为 5。s_{ui}为和用户相关的标签的分数。$N(u)$表示用户购买的商品的集合，$S(j,K)$是和标签 j 最相似的 K 个同类标的集合。K 暂定为 5。λ为调节参数，暂定为 3。

$$p_{uj} = s_{uj} + \lambda \sum_{i \in N(u) \cap S(j,K)} w_{ij} s_{ui}$$

备注：系统需要同时存储s_{uj}和p_{uj}

3.11 最终排序策略

M暂定为召回商品的 30 天内销售额，用户召回的商品的推荐效用分数为：

$$v = \Delta\left(\varepsilon \cdot lg(M+1) + \rho \cdot Min\left\{\frac{\sqrt{\sum_{j \subseteq N(j)} {p_{uj}}^2}}{Max\{15,\ Min\left(2 \cdot \overline{U_{(S'_{50})}},\ Max\left(U_{(S')}\right)\right)\}},\ 1\right\} + \zeta\theta\right)$$

$$U_{(S')} = \sqrt{\sum_{i \subseteq N(j)} {s_j}^2}$$

$$U_{(S'_{50})} = \sqrt{\sum_{i \subseteq N(j)} {s_j}^2}\ \left(Top\ 50\ of\ U_{(S')}\right)$$

其中$N(j)$ 表示商品命中的用户标签集合。ε为业务权重，ζ为消费指数权重，暂时取 1。ρ暂时取 8。

得到分数之后根据候选集分类 j，i 表示推荐效用分数在分类 j 中的排序。v_{ij}表示分类 j 排序为 i 个商品的得分。μ代表离散率。

$$v_{i,j}' = \left(\sum_{k=1}^{i} v_{kj} \right)^{\mu}$$

$$v_{i,j}'' = v_{i,j}' - v_{i-1,j}' \left(v_{0,j}' = 0 \right)$$

3.12 最终排序过滤

最终排序需要进行业务层过滤：

不推荐用户 30 天内购买过的商品（product ID 级别过滤）

部分三级分类用户 30 天内购买过不进行推荐（分类 ID 级别过滤），具体三级分类名单随后提供。

3.13 业务层逻辑：

根据 product ID 调用最优 deal 或者商城商品，只推荐可售单品。

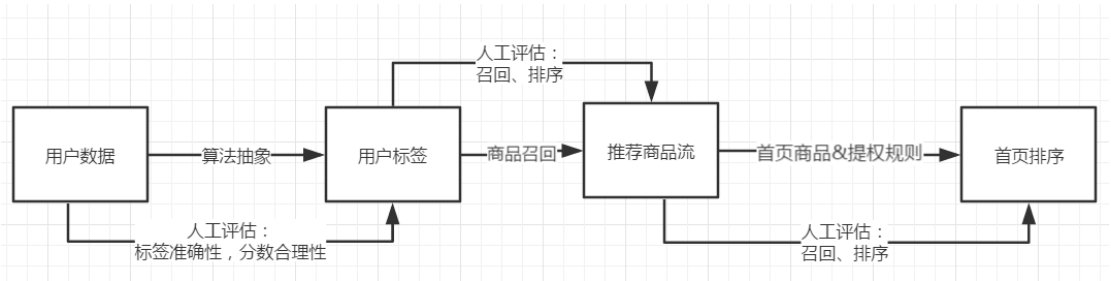
每天推荐更新一次。过滤可售商品后形成商品列表。对于位置为 K 的商品。排序的依据是：

$$p = Random(0,1)^{\frac{1}{4 \cdot lg(3+k)}}$$

4. 评估

4.1 离线评估

能够生成离线数据并且调整参数并观察排序，前期可以用 Excel 进行评估，后期需要考虑作为 APP 功能。



离线评估主要分为三个部分：

数据准确性：

数据正确：包括数据源，销量，价格等，是否与真实数据一致

计算正确：计算过程和得分是否和既定公式一致

人工评估用户标签：

标签准确性：是否用户被打上了应该打上的标签。

分数合理性：标签的分数分布是否和用户行为数据是否一致

人工评估商品流排序：

召回合理性：首页展示结果是否符合用户特征

排序合理性：首页排序是否符合用户特征

排序多样性：结果排序是否多维度符合用户标签

业务数据得分：销量好的商品需要在相对靠前位置

首屏差异性：针对典型用户首屏商品要体现足够的推荐效果

4.2数据评估指标：

用户选择年销售额大于 0 用户。

$T_i(U_i)$表示针对每个用户筛选出的 TOP100 的商品数据，$T(U)$表示所有在召回硬指标范围内的商品集合，覆盖率定义如下：

$$\alpha = \frac{\cup_1^n T_i(U_i)}{T(U)}$$

$R_i(U_i)$表示用户 i 真实购买的商品数据，$R_i(U_i) \cap T_i(U_i)$表示用户 i 真实购买的商品数据且被筛选出来的 TOP100 的商品数据，用户召回率定义如下：

$$\beta = \frac{1}{n}\sum_1^n \frac{R_i(U_i) \cap T_i(U_i)}{R_i(U_i)}$$

$T_i(U_i)$表示针对用户 i 筛选出来的 TOP100 的商品，$R_i(U_i) \cap T_i(U_i)$表示用户 i 真实购买的商品数据且被推荐的商品数据，用户准确率定义如下：

$$\gamma = \frac{1}{n}\sum_1^n \frac{R_i(U_i) \cap T_i(U_i)}{T_i(U_i)}$$

综合 F 值：

$$F = \frac{2\gamma \cdot \beta}{\beta + \gamma}$$

销售命中率，Mij表示用户 i 和商品 j 的销售额，$[R_i(U_i) \cap T_i(U_i)]$表示用户 i 真实购买的商品数据且被推荐的商品数据的逻辑 bool 值，命中则为 1，不命中则为 0 ：

$$\delta = \sum_i \sum_j \text{Mij} \cdot [R_i(U_i) \cap T_i(U_i)]$$

4.3 AB test

前端 AB test 机制：

需要能够根据业务需要控制一定比例用户在首页应用不同的算法策略。根据同时在线多套算法策略，并在统计参数中带上算法策略版本号。

4.4 报表机制：

cube 和神策报表中，现有的首页 CTR 以及浏览数据报表，首页销售报表，需要增加策略版本号筛选功能，默认展示总数据。但同时可以导出不同推荐策略的报表。

5. 及时响应策略补充

用户冷启动及时获得了大量的数据，对于推荐系统而言，无法及时计算，需要及时响应策略的补充。补充策略如下：

根据用户选择数据确定输入数据作为搜索调用数据条件：

- 品牌调性：用户选择品牌，最多的品牌调性作为输入品牌调性。
- 品牌：用户选择品牌作为输入品牌。
- 分类：用户选择分类作为输入分类

搜索调用逻辑为：

品牌调性命中 且｛品牌或分类命中｝，每个 3 级分类只召回最多 3 个销量最高的商品

排序逻辑为：

$$score = lg(M) + \pi A + \varpi B$$

其中M为商品销售额，π，ϖ为参数权重，A 代表分类的是否命中的布尔值，B 代表品牌的是否命中的布尔值。

6. 涉及团队

前三页：前端展示和冷启动数据收集

搜索推荐系统：数据存储，计算

大数据：基础数据提供

BI：基础数据提供，埋点规范确定，神策和 Cube 报表输出

专场：专场标签获取

用户系统：用户标签存储

7. 后续待补充策略

及时响应性推荐策略：加入购物车首页立即推荐商品

时间衰减性策略：最近的用户行为应该占有更多权重

商品流行度降权策略：商品流行度越高，点击量不满足则被降权

系统过滤策略混合：购买了 A 商品的人还购买了的商品

标签之间协同过滤：命中了 A 标签的人，可以推荐其他的标签内容

更多标签挖掘策略：数据扩充，地理信息，年龄信息

搜索系统增加个性化策略：搜索系统召回策略和排序策略根据用户特征调整