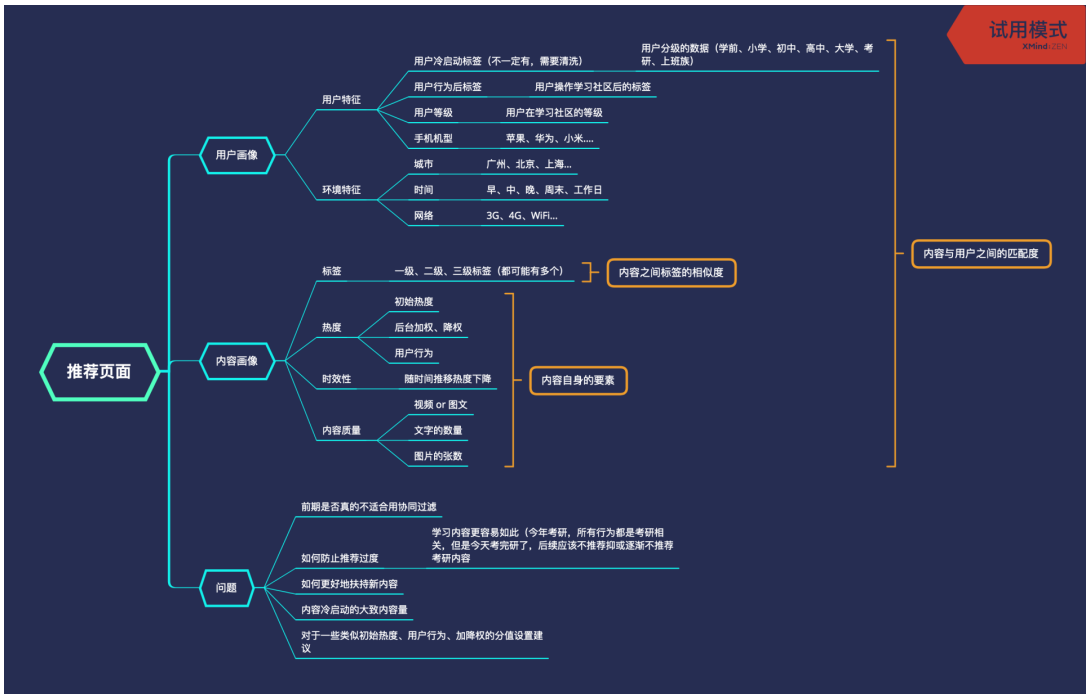
【调研】推荐算法



个性化推荐应用的两个条件：

（1）存在信息过载。如果用户可以很容易就从所有物品中找到喜欢的物品，就不需要个性化推荐了。

- 词典学习社区的内容面向UGC和PGC，内容存在信息过载的情况，但前期内容信息较少的时候，无法深度个性化推荐，更多的需要可能是基于内容属性的相似性推荐。

（2）用户大部分时候没有特别明确的需求。因为如果用户有明确的需求，可以直接通过搜索引擎找到感兴趣的物品。

- 大部分用户进到词典学习社区都没有特别明确的需求，就是简单的刷刷刷（这个得验证验证）

所以词典学习社区前期并不推荐直接用比较深度的个性化推荐：
前期障碍：用户量大，但用户数据少，内容量级处于初始状态
前期计划：以收集用户行为数据为核心，简单试行内容相关推荐
前期推荐机制：热度算法、基于内容属性推荐

一、热度算法基本原理

》〉》初始热度、时间热度
开始：Score = S0 - S(time)
S0：初始热度

- 需要对某一话题、某一用户内容卡片进行提权和降权
- 热词匹配：通过当期时事热度，进行关键词匹配，提升卡片的初始热度。思考点：

1. 初始热度的合理值（基础分50，提权：70、90、100，降权：-10...）

2. 后台直接对标签进行热度打分/改分（后台给基础分，对需要提权的标签进行手动提分）
3. 前期运营驱动，后期如何做到机器驱动（每天把标签直接都拉到百度指数跑一遍，然后rank？）
4. 对内容卡片进行提权和降权

S(time)：热度衰减分

- 需要根据内容发布的卡片时长来降权，发布时间越长的，受到降权的力度更大
- 公式：S（time）= e**[k * (T1 - T0)]

备注：推荐前提是用户登录对应账号，信息和userID绑定，在没有登录的时候也进行热度推荐，去掉用户行为分

即：Score = S0 - S(time)　　///////　　Score = S0 / S(time)

》〉》内容初始分

- 不同的内容类型，不同的内容质量，基础权重不一样
- 内容的发布类型、内容的数量（图片、文字）不同的权重
- S(sw)=a*type + b*number

选项	内容发布类型	文案字数	图片数	发布账号的级别
分值	视频=2 图片=1	>100字，+0.1 >200字，+0.3	>1张，每一张+0.1	+0.1

》〉》用户行为分

- 用户对内容的行为，影响到其对应的权重值
- 标签加权：对各个标签进行加权，加权：

1. 用户对内容的行为，影响到对应的权重
2. 后台权重可调配（随时更改数值）
3. 不同行为不同的权重

S(users)：用户行为分

- 公式：S(users) = [a*click + b*like + c*comment + d*share + e*tip + f*focus + g*favor] / DAU*N(固定值）
- 行为事件：关注、点赞、评论、分享、点击、添加内容（到计划）

- **结合热度算法得：Score=S0+S(sw)+S(user)-S(time)**
- 思考点：

1. 各类行为对应的权重
2. K、N的值（前期内容量不够多，K值可以相对比较低，保证展示过的内容对该用户不再展示即可）
3. 用户行为分对热度算法的占比

内容用户行为	分值		计划用户行为	分值
点击	1		点击	1
点赞	2		打卡	3
评论	3		加入计划	5
循环播放			动态数	
视频播放时长				
页面停留时长				
添加内容	4			
分享	5			
举报	–5			

二、基于内容属性的相似性推荐

- 标签相似性越高的内容更应该被推荐
- 标签相互排斥性更高的，更应该不被推荐
- 同个用户不同标签也得有三六九等

通过内容标签的相似性来进行排序。

一、召回筛选规则：

1、在用户画像有三级标签的情况下，除了三级标签本标签（考研心理学-考研心理学），另外相似的三级标签中，权重D降权，不加入排序

2、该用户看过的内容，不用进入筛选（暂定）

二、排序规则：

1、标签重合数量多的权重》标签重合数量少的权重

2、同等标签重合度，二级标签权重》一级标签权重

3、三级标签含权重，A1》A2》...》B》C》D（D为两个标签相互独立的表示，有他无我，有我无他）

4、四级、五级...标签逻辑同三级标签即可

5、根据用户的画像标签的权重进行排序推荐

三、用户画像规则

1、前期（用户没有在词典圈有任何互动），画像只拿“用户分级”的身份标签（研究生、高中、大学...）

2、用户的行为会提升/降低其各个标签的权重，扩展、更新其用户画像

四、内容画像规则

1、尽可能多的标签，有的内容可以有2-3个二级标签（比如考研心理学，可以是考研，也可以是心理学）

2、一级和二级标签为显式标签，即用户可见

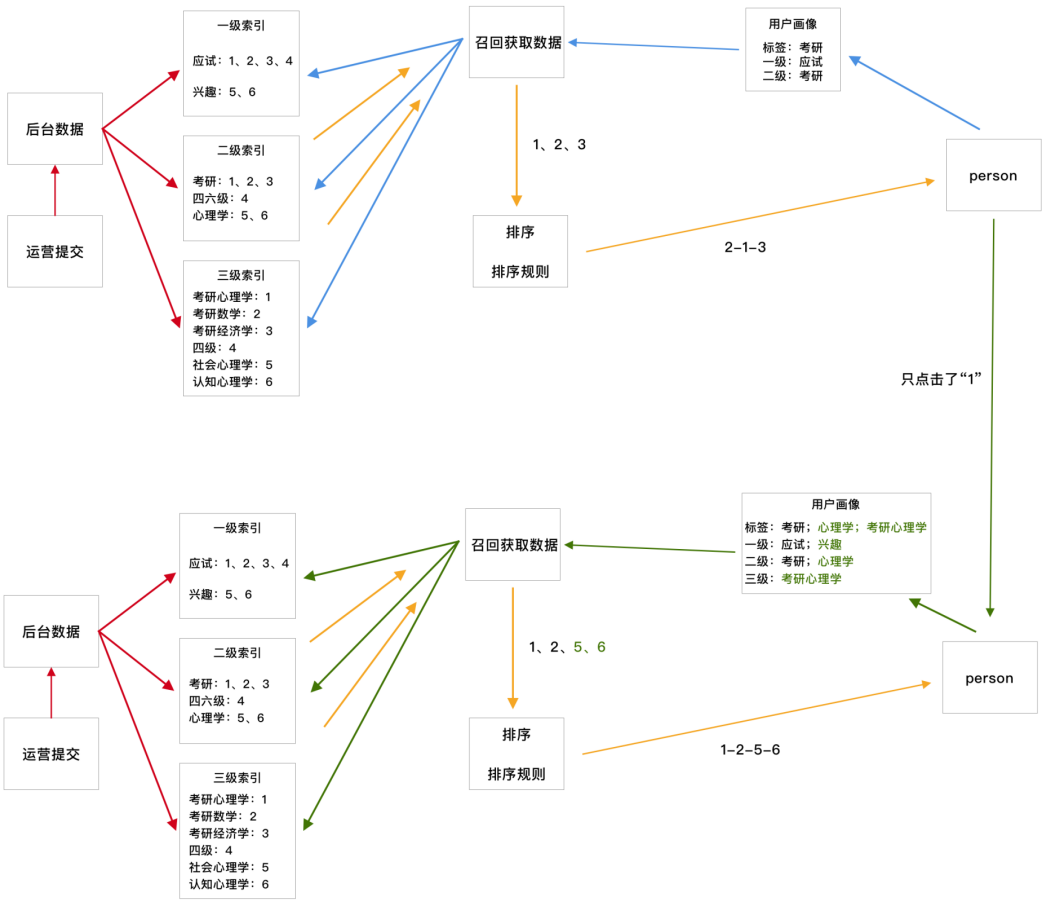
3、三级标签为隐式标签，运营进行标记

4、三级标签有权重，A1、A2的区分可以先跑市场数据来定义，后续根据词典显示数据调。这里主要是为了排序以及让一些不该出现的内容不出现

（比如：考研心理学标签下，考研数学是哪个专业都得考，他的权重要大于考研经济学，而考研经济学是不应该出现的，所以应该再降权）

举例1:

内容	一级标签（频道）	二级标签（分类）	三级分类（tag）
内容	一级标签（频道）	二级标签（分类）	三级分类（tag）
内容1	应试	考研、（心理学）	考研心理学，权重D3
内容2	应试	考研、（数学）	考研数学，权重A2
内容3	应试	考研、（经济学）	考研经济学，权重D1
内容4	应试	四六级	四级
内容5	兴趣	心理学	社会心理学，权重C1
内容6	兴趣	心理学	认知心理学，权重C2



三、推荐算法的评测方法

循序渐进：离线实验法——用户调查法——线上ABtest

1、离线实验法

在测试阶段，输入参与测试的用户的词典id，就能看到用户的行为以及各行为的分值，同时能够看到给这个用户推荐的最终结果（注有推荐原因，比如分值）

- 在客户端demo上，通过用户操作（内部用户即可），服务端获得行为日志
- 在demo获得推荐结果
- 在后台输入操作的某个用户的词典id，可以看到对其推荐的内容，以及推荐的原因

优点：快速验证算法准确率

缺点：推测准确度高不代表用户满意度高

2、用户调查

- 在调研小组里，给用户操作demo，同时给予新的推荐结果
- 通过侧面的提问来获取用户的满意度

优点：可以在一定程度获取用户的意见

缺点：样本少、成本高

3、线上实验法

- 线上abtest
- 两个或以上算法不确定时的较好验证方法

四、推荐算法的测评关键指标

1、推荐准确度

- 用户的点击率，以及点击外的各种互动数据、停留时长

2、用户满意度

- 通过用户调查、nps等来测验

3、内容覆盖率

- 整个内容分发体系的基尼系数

将每个标签下的内容展示数从低到高进行排列，然后计算每个标签内容展示数占总展示数的比例。比如有20个标签，总展示数为210，占比分别为1/210、2/210...20/210.

这时候获得所有洛伦兹曲线的点。在基尼系数模型里，下方的面积B=这20个标签所在的梯形面积总和，而A+B=1/2，所以Gini=1-2B

基尼系数的值越低，表示越平均。0.2-0.3为比较平均，0.3-0.4为相对合理，所以争取内容覆盖的基尼系数在0.2-0.4的区间波动

4、及时性

- 用户在内容进行互动后，推荐系统的反应速度
- 新内容产生后，对其的处理能力