用户画像必会的行为偏好计算方法

原创 不等雨 浅梦的学习笔记 2019-11-14

收录于话题
#推荐&广告算法技术原理与实践

55个

“本文简要介绍了用户画像中行为偏好计算方法和常用的行为加权方法，以及面对行为稀疏和统计不置信问题时的解决方案。”

作者：不等雨
来源：知乎专栏 1024，已通过作者授权。

用户行为偏好计算

常用的统计方法

1. 点击率（ctr）：

$pre = \frac{click_i + m}{impl_i + C}$,$m$和$C$是平滑系数

2. 正样本频次（fc）：

$pre = click_i$，点了多少次

3. 正样本中分布比例(rate)：

$pre = \frac{click_i}{click}$,在正样本中分布比例

4. 交叉后的相对偏好(cross-rate)：

$pre^* = \frac{pre}{E(pre)} > 1$

比如pre是点击率，$E(pre)$是此用户的一般点击率或者相似用户对此类内容的一般点击率，也可以是大盘用户对此类内容的一般点击率；当然pre也可以是正样本分布比例。

5. 用IDF来体现稀缺度(tfidf)：

$pre^* = pre \times \log(\frac{\sum_{i=1}^{C} user_i}{user_1 + 1})$

小结

那这么多统计维度都需要用吗？至少在画像特征层面，建议都给做了，提供给模型层统一的特征集合。实际使用中，点击率、相对偏好、正样本频次、正样本分布比，都有自己的优缺点，也可以认为是关于偏好的N种表达或者转换或者交叉，能给模型层很多新的信息量。

用户行为加权

用户的行为应该带有权重的原因

1. 分类型
   点击时长、短停、完成率、分享、聊天、长停、关注、送礼、搜索、红包等等，代表的偏好意义不同；

2. 分场景
   分位置分样式分内容；比如首页还是划屏、第一个还是第N个位置、大尺寸还是小标题、特别热的内容（突发热点几乎都会点或者消费）等；

3. 分时段
   分节日分运营活动；比如主推的活动，或者特殊节日，会给用户行为带来短期的偏差；

4. 时间带来兴趣衰变
   3个月前的行为，权重肯定比不上最近的行为；

所以实际中，用户不同的行为会有对应的权重系数来配置，时间因素会通过记忆力公式来降权或者牛顿系数来降权。对特殊行为（比如活动、热点突发）会做特殊的降权处理。**权重系数base表可以线下去换算或者拍公式，基本原则是越稀少的行为权重越高**(比如一次关注可能相当于10次有效点击等)。

举例：

fc偏好特征： $pre = \sum_{i=1}^{n}(w_{click} * w_{time1} + w_{share} * w_{time2} + \dots)$

实践需注意的细节：

1. 正负样本的判定
   一次曝光10条，其中位置靠后的负样本，需要降权；原因是滚屏采样的误差，或者伪曝光；

2. 时间降权
   还得考虑有效时间。比如用户7天没有来，跟7天都来，那么降权速度应该不一样；（对无效时间可以折扣，减慢降权速度；或者取消时间衰变，给用户出90-180天、30-90天、30天、7天、实时等多种窗口特征，特征之间交叉可以提供兴趣衰变信息）

平滑或expand

行为稀疏和置信度低的问题：

1. 用户对此内容的行为少，比如只点击了1次，曝光了2次；或者点击了0次，曝光了1000次等。统计上不置信；
2. 用户对此内容没有发生曝光；

解决方法：

1. 数据处理。
   既然已经稀疏，最好就不要把数据再分开了，比如把长停、送礼、关注、发言等行为都分析下相关性矩阵，然后统一换算成一种特征，而不是好几个；或者把同手机的用户合并、不同手机同一注册账号合并等等；

2. 平滑
   平滑的技巧是在分子和分母上加一个a和b。考虑加一平滑 或者 bayes平滑。

加一平滑：

- $pre = \frac{click_i+a}{impl_i+b}$ a和b是平滑系数
- 假设考虑颜值（特高、高、中、低）,那么特高取：$b=4, a=1$。但是这样明显不符合先验，实际中高颜值的偏好度远大于低颜值，所以可以根据颜值的偏好统计，做简单的换算。换算后： $b=4, a=1.5$，会看起来合理多了。
- 工程实现
  - 离线：1）type1的a和b计算；
  - 在线：1）用户对type1的在线平滑；

bayes平滑：

- 更推荐相似用户的bayes平滑，实际中效果很好。可以用矩估计来简单求解α和β。比如取同手机同地区同行为的用户群， 然后求解出群体里每个用户的点击率，得到总体的均值和方差，然后根据*beta*分布的方差和均值公式，矩估计，推导出最后的α和β，看效果如何；
- 如果求解的是主播或者广告的点击率，可以用相似主播来做bayes平滑；如果求解的是用户对某种type1的偏好点击率，可以用相似用户对type1的训练集做bayes 平滑，或者也可以找type1的相似type集合，用type集来平滑，假设的是相似type都采样自同一个beta分布
- 当然加一平滑、分组折扣、指数平滑（历史累计数据权重衰变后，成为先验）很实用，效果说话；
- 工程实现
  - 离线：1）用户 - 用户group；2）group对各个type的先验α和β，入库；
  - 在线：1）用户 - 用户group；2）根据group对偏好做平滑；

用户画像的统计偏好还应该提供什么信息？

更多交叉统计：

我们算出了用户对某种内容的各类统计偏好，为排序层的特征工程直接服务。画像特征层应该做好集中管理，并尽可能给排序层做好有意义的特征交叉，类似相对偏好率、细分交叉统计点击率、实时行为特征等这种特征。排序层只需要做进一步的特征定制（选择）、转换等处理。这样就可以尽量提供画像特征、排序特征的解耦。

实时偏好：

正样本ID 、短期窗口统计等

偏好的置信度：

用户的偏好应该提供置信度，比如看统计样本个数。策略层也可以对重要的偏好特征，刻意去试探以提升置信度；

推荐阅读

- **绝对干货！NLP预训练模型：从transformer到albert**
- **NLP与推荐系统的比较、联系与未来**
- **【GraphEmbedding】DeepWalk算法原理，实现和应用**
- **【CTR预估】CTR模型如何加入稠密连续型和序列型特征?**
- **快速掌握TensorFlow中张量运算的广播机制**

想了解更多关于用户画像**与推荐系统**的内容，欢迎关注公众号**浅梦的学习笔记**，回复“**加群**”可以一起参与讨论交流！