知乎 首发于 推荐系统那些事儿

推荐系统之用户画像

龚旭东

汽车之家 人工智能高级产品经理

24 人赞同了该文章

标签体系到用户画像

上一篇文章介绍了标签体系，实际上标签体系是用户画像的基础，本质上用户画像是一系列与用户相关的标签的结构化表示，也是成体系的，同时用户画像各个细分维度的构建也是同样遵守标签体系构建的原则——以业务为导向。从之家的用户画像实践看，大体是分为人口属性、网络属性、兴趣属性、商业属性四个维度。

画像词云

用户画像的构建

时间变化很快，有较强的时效性，之家的实践是分为长期、短期。

人口属性

大部分主流的人口属性标签都和这个体系比较类似

序号	一级分类	二级分类	标签名称
1	人口属性	自然类	性别
2	人口属性	自然类	年龄
3	人口属性	社会类	职业
4	人口属性	社会类	婚姻状况
5	人口属性	社会类	教育水平
6	人口属性	社会类	人生阶段
7	人口属性	自然类	省份
8	人口属性	自然类	城市
9	人口属性	自然类	城市等级
10	人口属性	自然类	商圈
11	人口属性	自然类	常用位置
12	人口属性	自然类	用户LBS信息
13	人口属性	自然类	用户工作地位置信息
14	人口属性	自然类	用户居住地位置信息
15	人口属性	自然类	用户常用行政区县

人口属性维度

很多产品在注册时就会引导用户填写基本信息，这些信息就包括年龄、性别、收入等大多数的人口属性，但完整填写个人信息的用户只占很少一部分。而对于无社交属性的产品（如输入法、团购APP、视频网站等）用户信息的填充率非常低，有的甚至不足5%。

在这种情况下，我们一般会用填写了信息的这部分用户作为样本，把用户的行为数据作为特征训练模型，对无标签的用户进行人口属性的预测。这种模型把用户的标签传给和他行为相似的用户，可以认为是对人群进行了标签扩散，因此常被称为标签扩散模型。

下面我们用之家性别年龄画像的例子来说明标签扩散模型是如何构建的：

对于购车用户我们也希望尽可能了解其性别，不同性别的用户在购车的意向车系上还是有较大差别的。

假设我们有40%的用户填写了个人信息，我们将这40%的用户作为训练集，来构建全量用户的性别画像，数据如下所示：

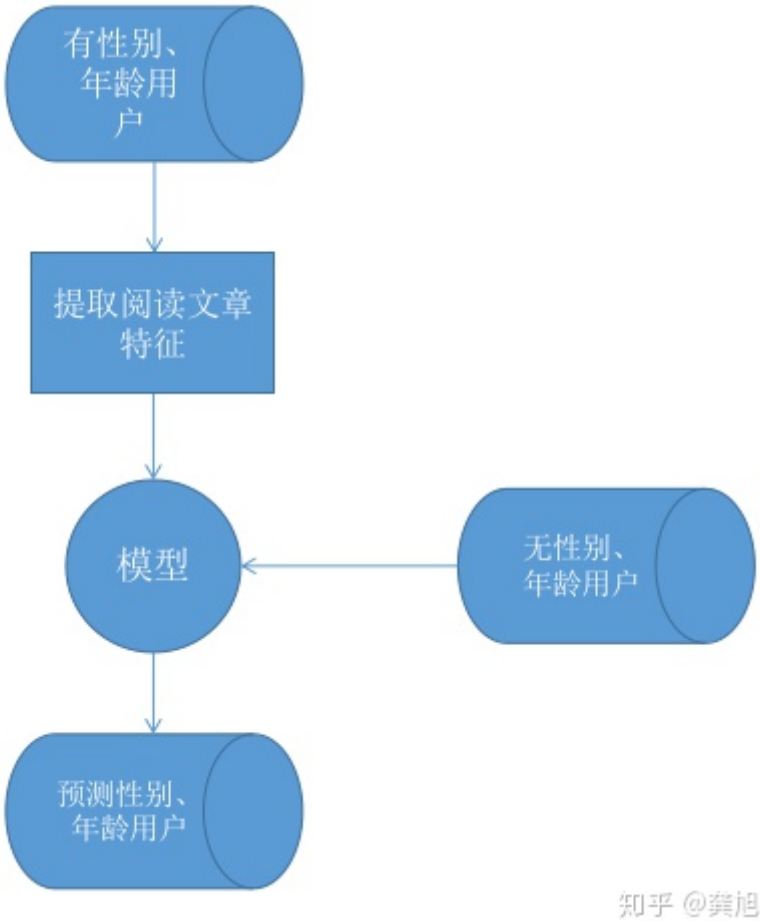
知乎 首发于 推荐系统那些事儿

528232		mini cuper
563232	Female	沃尔沃XC40、宝马X1

性别预测

下面我们来构建特征。通过分析，我们发现男性和女性，对于意向车系的偏好是有差别的，因此使用阅读车系相关的文章列表来预测用户性别有一定的可行性。此外我们还可以考虑用户阅读文章的点击率、完读率、评论、转发、点赞等行为，为了简化，这里我们只使用用户浏览文章特征。

由于阅读文章的特征是稀疏特征，我们可以使用调用MLlib，使用LR、线性SVM等模型进行训练。考虑到注册用户填写的用户信息本身的准确率不高，我们可以从40%的样本集中提取准确率较高的部分（如用户信息填写较完备的）用于训练，因此我们整体的训练流程如下所示。



性别预测流程

对于人口属性标签，只要有一定的样本标签数据，并找到能够区分标签分类的用户行为特征，就可以构建标签扩散模型。其中使用的技术方法主要是机器学习中的分类技术，常用的模型有LR、FM、SVM、GBDT等

兴趣属性

序号	一级分类	二级分类	标签名称
28	兴趣属性	车相关	车系偏好
29	兴趣属性	车相关	品牌偏好
30	兴趣属性	车相关	车型偏好
31	兴趣属性	车相关	厂商偏好
32	兴趣属性	车相关	产地偏好
33	兴趣属性	车相关	级别偏好
34	兴趣属性	车相关	派系偏好
35	兴趣属性	车相关	车系偏好（短期）
36	兴趣属性	车相关	品牌偏好（短期）
37	兴趣属性	车相关	车型偏好（短期）
38	兴趣属性	车相关	厂商偏好（短期）

车相关兴趣属性

兴趣属性中的各个标签在个性化推荐、互联网广告、精准营销中显得尤为重要。兴趣属性主要是从用户海量行为日志中进行核心信息的抽取、标签化和统计，因此在构建用户画像的兴趣属性之前需要先对用户有行为的内容进行标签体系（分类-主题-关键词）构建。关于内容标签体系构建上一篇文章中已经讲过，不再赘述。对于兴趣属性，我们需要认识到用户的兴趣点是受各种各样因素影响的，所以我们要考虑到兴趣和时间这个维度的关系。

兴趣衰减

我们可以根据用户点击，计算用户对分类、主题、关键词的兴趣，得到用户兴趣标签的权重。最简单的计数方法是用户点击一篇文章，就把用户对该篇文章的所有标签在用户兴趣上加一，用户对每个词的兴趣计算就使用如下的公式：

$$score_{i+1} = score_i + C \times weight$$

其中：关键词在这次浏览的新闻中出现，则 C=1，否则C=0，weight表示词在这篇新闻中的权重。这样做有两个问题：一个是用户的兴趣累加是线性的，数值会非常大，老的兴趣权重会特别高；另一个是用户的兴趣有很强的时效性，昨天的点击要比一个月之前的点击重要的多，线性叠加无法突出近期兴趣。

为了解决这个问题，需要要对用户兴趣得分进行衰减，我们使用如下的方法对兴趣得分进行次数衰减和时间衰减。

$$score_{i+1} = \alpha \times score_i + C \times weight \quad (0 < \alpha < 1)$$

其中，α是衰减因子，每次都对上一次的分数做衰减，最终得分会收敛到一个稳定值 ，α取0.9时，得分会无限接近1。

$$score_{day+1} = score_{day} \times \beta \quad (0 < \beta < 1)$$

它表示根据时间对兴趣进行衰减，**这样做可以保证时间较早的兴趣会在一段时间以后变的非常弱，同时近期的兴趣会有更大的权重。**根据用户兴趣变化的速度、用户活跃度等因素，也可以对兴趣进行周级别、月级别或小时级别的衰减。

关于网络属性和商业属性就不再细讲，其实所有画像的维度都差不多。关键是在于定义某每个标签时，标签计算的方式和标签最终的取值，要和业务真实的情况相匹配，能够有效的支撑业务的需求。

用户画像的评估

人口属性画像的相关指标比较容易评估，而兴趣属性的标签比较模糊，兴趣属性的人为评估比较困难，我们常用评估方法是设计小流量的A/B-test进行验证。

我们可以筛选一部分标签用户，给这部分用户进行和标签相关的推送，看标签用户对相关内容是否有更好的反馈。

例如，在内容推荐中，我们给用户构建了兴趣画像，我们从改装车类兴趣用户中选取一小批用户，给他们推送改装车类新闻，**如果这批用户的点击率和阅读时长明显高于平均水平，就说明标签是有效的。**

效果评估

用户画像效果最直接的评估方法就是看其对实际业务的提升，如互联网广告投放中画像效果主要看使用画像以后点击率和收入的提升，精准营销过程中主要看使用画像后销量的提升等。但如果把一个没有经过效果评估的模型直接用到线上，风险是很大的，因此我们需要一些上线前可计算的指标来衡量用户画像的质量。

用户画像的评估指标主要是指**准确率、覆盖率、时效性**等指标

准确率

标签的准确率指的是被打上正确标签的用户比例，准确率是用户画像最核心的指标，一个准确率非常低的标签是没有应用价值的。准确率的计算公式如下：

$$precision = \frac{|U_{tag}|}{|U_{tag}|}$$

其中| Utag |表示被打上标签的用户数，| Utag=true |表示有标签用户中被打对标签的用户数。准确率的评估一般有两种方法：一种是在标注数据集里留一部分测试数据用于计算模型的准确率；另一种是在全量用户中抽一批用户，进行人工标注，评估准确率。

由于初始的标注数据集的分布和全量用户分布相比可能有一定偏差，故后一种方法的数据更可信。准确率一般是对每个标签分别评估，多个标签放在一起评估准确率是没有意义的。

覆盖率

标签的覆盖率指的是被打上标签的用户占全量用户的比例，我们希望标签的覆盖率尽可能的高。但覆盖率和准确率是一对矛盾的指标，需要对二者进行权衡，一般的做法是在准确率符合一定标准的情况下，尽可能的提升覆盖率。

我们希望覆盖尽可能多的用户，同时给每个用户打上尽可能多的标签，因此标签整体的覆盖率一般拆解为两个指标来评估。一个是标签覆盖的用户比例，另一个是覆盖用户的人均标签数，前一个指标是覆盖的广度，后一个指标表示覆盖的密度。

用户覆盖比例的计算方法是：

$$coverage = \frac{|U_{tag}|}{|U|}$$

其中| U |表示用户的总数，| Utag |表示被打上标签的用户数。人均标签数的计算方法是：

$$average = \frac{\sum_{i=1}^{n} tag_i}{|U_{tag}|}$$

其中| tagi |表示每个用户的标签数，| Utag |表示被打上标签的用户数。覆盖率既可以对单一标签计算，也可以对某一类标签计算，还可以对全量标签计算，这些都是有统计意义的。

时效性

新机制，以保证标签时间上的有效性

用户画像应用实例——召回

在推荐系统中，每一篇用户能够看到的文章，大体都经历了，召回、排序、干预（量控、打散、调权、过滤、强插）几个过程。其中召回阶段是推荐产品和算法同学下功夫应该是比较多都一个过程，在《内容算法》这本书中，作者作为今日头条的资深推荐产品提到，头条召回的路数有一万多条，可见其算法粒度做到多么的精细、业务规则是多么复杂。本篇文章我们还是重点介绍画像，就说用户画像中的兴趣属性在召回中的使用。

如某用户兴趣属性中有体育的兴趣偏好，兴趣关键词有：足球、篮球、排球...

索引就会按用户画像兴趣偏好中关键词取回包含这些关键词的物料，并且是按值排好序。这里的排序是根据物料的综合得分的排序，关于物料得分的计算模型，我们会在后面的篇章中介绍。这一路召回可以由其它一些实际情况来决定最后召回多少条。

试想实际上每个用户的兴趣偏好的关键词是各不相同的，那么在用户画像兴趣偏好召回阶段，就实际上是千人前面了。

对于之前提到的用户画像标签AB实验，实际上也是在某一路召回上对同一标签，不同的定义、计算方法来验证是否标签的更新对于实际的业务指标，如CTR是否有提升。

下一篇文章将详细介绍召回，这里只简单介绍。

编辑于 2020-02-21

推荐系统 用户画像 召回