一文读懂推荐系统用户画像

人人都是产品经理
发布时间：20-04-28 19:21 | 深圳聚力创想信息科技有限公司

本系列文章将从最简单的概念开始，逐步讲解推荐系统的发展历程和最新实践。以产品经理的视角，阐述推荐系统涉及的算法，技术和架构。本文将介绍推荐系统如何给现实世界中的用户打数字化的标签：用户画像。

用户画像，简单来讲，就是我们给用户打上的一系列的标签。它的应用非常广泛，在互联网产品的任何一个领域，任何一种实现用户个性化的功能，都需要用到用户画像。本文只涉及推荐系统的用户画像体系。

一、推荐系统用户画像长什么样

用户画像这个词具有广泛性。它被应用于推荐，广告，搜索，个性化营销等各个领域。任何时候，不管出于什么目的，我们想描述我们的用户是谁的时候，大家都会用到用户画像这个词。

比如：

（1）产品经理定性用户分析

设计产品功能时，会对用户是谁进行描摹。如：目标用户群体的人口属性，社会背景，使用习惯等信息。这种用户画像主要描述用户是谁，以便做好功能定位。

如下图中的定性用户画像分群：

作者最新文章

信息容器的归类与应用，这些知识点get到了吗?

如何快准狠找到数字化转型的“目标客户”?

始于需求，而终于需求的最终落地

相关文章

产品设计需求分析研究

5000字详解：如何从0到1搭建私域流量

用户画像、用户分层、用户分群：精细化运营的万川之源

万字干货：越过18个让数据变成谎言的陷阱

（2）数据分析用户画像

分析用户行为，用户进行聚类行为分析。如：数据分析师可能会给出，观看电商直播的男女比例，得出女性用户更喜欢看我们的电商直播这样的结论。

（3）推荐系统用户画像

为建立个性化功能，用各种办法给用户大规模打上几万甚至几千万个标签。这种标签不仅仅有偏好，还有偏好程度值。

本文所指的用户画像，仅仅涉及第三种情况。一般地，推荐系统的用户画像长成这个样子：

基础信息

属性	值	概率
性别	男	0.776
年龄	24-30	0.832
城市	广州	0.993

近期喜欢

标签	偏好值
金融战争	78.55
三傻大闹宝莱坞	66.52
郭德纲	33.12

不喜欢

标签	偏好值
风水算命	-125.6
心灵鸡汤	-85.12
樊登	-68.33

长期喜欢

标签	偏好值
海贼王	123.56
王者荣耀	95.12
火影忍者	78.33

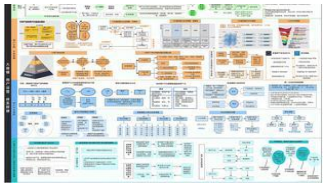
菠萝王子AI分享

推荐系统的用户画像，一般包括用户基础信息和偏好信息。而偏好画像是重点，数量上占了推荐系统用户画像的绝大多数，是我们召回和模型训练的基石。

因为机器跟人不同，一个词“中国”对于人来说是有意义的，对于机器只是一个汉字编码。因为用户画像，为了能让机器计算，需要带上概率值或者偏好值（权重值）等。

我们接下来就聊一聊，在推荐系统中，这种带了一些列数字的用户画像怎么构建出来的。

做好业务项目：产品和运营都不能少



第一章的介绍过，推荐过程分为：召回、初排和精排三个阶段。用户画像主要用在召回和初排两个阶段。

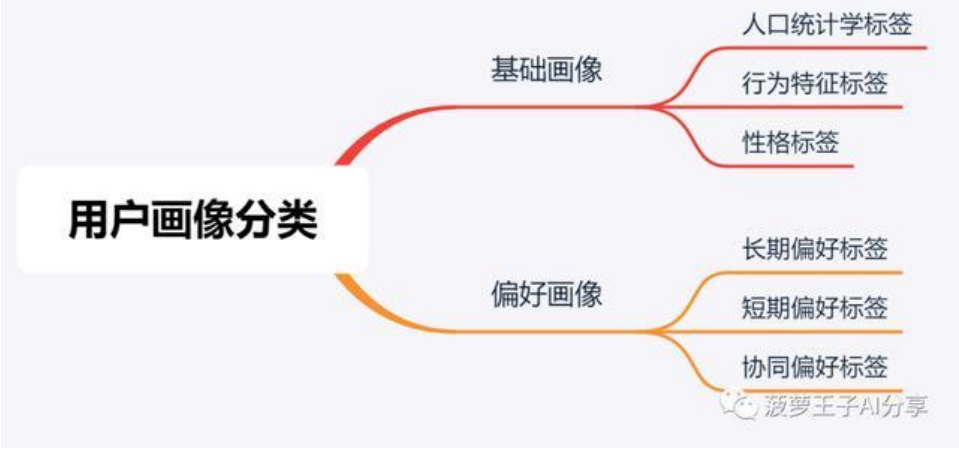
召回阶段使用用户画像，主要是通过用户画像召回相似的物品。比如一个短视频APP上，用户海贼王偏好值比较高，就可以针对海贼王进行内容召回。

初排阶段使用用户画像，是在模型上使用的。模型将用户画像数据作为一部分的特征值，用于模型的离线训练或者实时模型更新。

三、用户画像的分类

用户画像是一个比较大而全的概念，标签是用户画像最基本的单元，用户画像是有成千上万个标签组合而成的。当我们想对用户画像进行分类时，通过对用户标签的分类就可以了。每个平台有自己的用户画像体系。对推荐系统的构建来说，一般从以下维度来做标签分类。

如下图所示：



其中：

（1）基础用户画像

- 人口统计学标签：用户的性别，年龄，地区等信息。
- 行为特征标签：用户在互联网平台的注册，活跃，付费，浏览等方面的行为记录产生的用户标签。
- 性格标签：豪爽大方，精打细算，冲动消费等类型标签

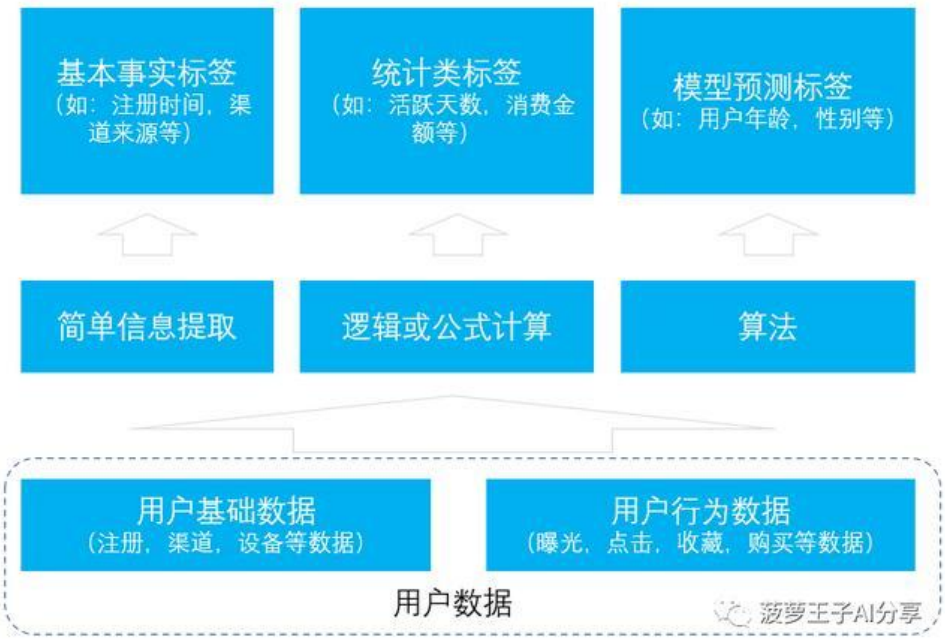
- 长期偏好标签：用户对较长时间内，几个月甚至是几年内，对某类事物的稳定偏好。
- 短期偏好标签：用户最近较短时间内，七天内甚至是几分钟内，对某类事物的偏好。
- 泛化偏好标签：众多的用户偏好中，不同的偏好之间有关联性或者相似性，就像啤酒和尿布那样。用户对啤酒有过直接的行为，但对尿布还没有，那么尿布可能是他的泛化偏好。

以上的五小分类中，前面两类只占了用户标签数量的很小一部分。而推荐系统中，数量最为庞大的要数偏好类的标签了。平台有多少个物品标签，就会产生多少偏好标签。另一方面，偏好类的标签的产生，依赖于物品标签。因为用户对物品的偏好程度，是通过他对平台物品的曝光，点击，购买等行为计算出来的。

四、基础用户画像的怎么来?

那基础的用户画像是怎么产生的呢？一般可分以下几种来源：

- **简单信息提取**：基于实际基本事实而产生标签，如注册时间，渠道来源，用户所在地区等。
- **逻辑或公式计算**：使用简单的逻辑或公式，对用户的行为进行统计而产生标签，如用户活跃天数，用户消费金额等。
- **算法学习**：基于机器学习模型对用户的属性预测产生的标签，如性别，年龄，有车一族等。



五、简单举例：通过模型产生基础用户画像

国内某公司，在Kaggle举行过一个预测用户年龄和性别的比赛。他们公布了一个用户数据集，数据集中包含了手机上安装的APP列表，手机型号和GPS信息等数据用于模型训练。参赛选手通过这些数据建模，预测用户的性别和年龄。准确度高的获胜。

一个用户的手机里安装的APP，跟他的年龄和性别存在着一定的关联。如：女性用户常用美柚，小红书等APP；而男性用户可能会装更多的游戏。

这个是有监督学习，橙色部分数据是特征，蓝色部分数据是label。Label就是我们需要预测的目标。通过大量的数据和算法调优，就可以训练出较为准确的模型。

用训练好的模型，就可以给其他的未知性别和年龄的用户做评分预测。这部分比较简单，就简单举例一下。

六、物品标签

物品画像，则是每个物品的一系列标签。物品画像其中一个作用就是可以作为推荐模型中的物品特征。另外一方面，**在推荐系统中，物品画像是用户画像的基础：物品画像+用户行为=用户画像。**

举个简单的例子，一个用户点击了一系列的阿克苏苹果（物品画像：阿克苏，苹果，阿克苏苹果），这个用户就会被打上阿克苏，苹果和阿克苏苹果的偏好标签。

物品画像的产生，不同的内容形式有不同的做法。但大体可分为两类：

- 人工的方式给物品打标签；
- 机器学习的方式给物品打标签。

如在音乐领域，一些音乐平台是通过一组音乐专家对平台的音乐进行打标签后，再对用户进行推荐。这种人工的方式成本比较高，而且依赖于专家的专业程度。另外，不同专家之间的标准可能不一样，需要统一标准或者拉平差异。但是这也是没有办法的办法，有些场景下，物品标签匮乏，不得不依赖与人工打标的方式。

七、偏好画像的怎么计算得来?

偏好画像如何产生？为了直观简单，直接以图文数据的方式来讲述。**假设一个短视频平台有4个用户使用，有4个视频需要被推荐。**

其中，4个视频分别为：

内容标签：金融战争，做空

内容标签：海贼王，路飞

内容标签：海贼王，甚平，路飞

内容标签：三傻大闹宝莱坞，学霸

整理一下，我们可以得到以上4个视频的物品画像：

视频ID	标签1	标签2	标签3
1	金融战争	做空	
2	海贼王	路飞	
3	海贼王	甚平	路飞
4	三傻大闹宝莱坞	学霸	

另外，为了简单一点，这里只考虑用户的观看行为，看完一次得分为1。4个用户的数据分别如下，数字代表观看次数。如下图中，用户A看了视频1一共2次。

A	2			3
B		2	2	
C		4		5
D	1		3	

先说结论，一般地，用户画像的公式为：**用户偏好程度 = 行为类型权重值 × 次数 × 时间衰减 × TFIDF值。**

- **行为类型权重值**是人为给用户行为的赋值。比如：看完=1，收藏=2，分享=3，购买=4等。我们这里只考虑“看完”这个行为。
- **次数**则是行为发生的次数。
- **时间衰减**则是按一定的衰减系数，随着时间衰减。一般用牛顿热力学公式来取衰减系数。
- **TFIDF值**本来是文本处理领域的算法，用来提取一篇文章中的关键字。这里用来衡量标签的对一个用户的关键程度。

下面我们来计算用户A的用户画像和偏好值。

第一步：列一下行为类型权重值，因为我们只考虑观看行为，权重都为1：

第二步：统计用户A的行为次数。用户A看了视频1两次，所以视频1带的标签“金融战争”和“做空”次数都记为2：

用户A	金融战争	做空	海贼王	路飞	甚平	三傻大闹宝莱坞	学霸
行为次数	2	2				3	3

第三步：计算时间衰减，假设用户A看视频1是两天前的行为，看视频4是今天的行为。衰减按照天来计算，衰减系数等于0.1556，热度计算公式为：热度=1×exp(-0.1556×天数)。按照这个衰减系数，45天后热度衰减到0.5。

按照这个计算方式，视频1的热度 = 1×exp(-0.1556×2) = 0.73，今天看的视频4，热度还为1。

第四步：计算TFIDF值。

这步比较复杂。我们先说下TFIDF的公式，TF和IDF是两个不同的值，两两相乘可以得到TFIDF值。

首先说TF。

TF是Term Frequency的缩写，意思是可以理解为词频，计算公式如下：

签重复的），标签的TF值=1÷4=0.25。

用户A	金融战争	做空	海贼王	路飞	甚平	三傻大闹宝莱坞	学霸	总数
标签个数	1	1				1	1	4
TF值	0.25	0.25				0.25	0.25	

而对于用户B，因为有看过两个海贼王的视频。一个视频带标签：海贼王，路飞。另外一个视频带标签：海贼王，路飞，甚平。所以，海贼王和路飞标签个数都是2，甚平的标签个数是1。

这样，计算出用户B的TF值为：

用户B	金融战争	做空	海贼王	路飞	甚平	三傻大闹宝莱坞	学霸	总数
标签个数			2	2	1			5
TF值			0.4	0.4	0.2			

然后说IDF。

IDF是Inverse Document Frequency，意思是逆文档频率。先说怎么计算，公式如下：

这个是为了计算一个标签的稀缺程度。如果一个标签全部的用户都，IDF值就比较小。相反，一个标签只有少部分用户有，则IDF值比较大。公式中，“带该标签的用户数+1”部分加1是为了防止分母为0的情况。

下表的灰色部分是每个用户行为，计算出用户的标签个数统计。如海贼王标签，因为有三个用户带了这个标签，所以“带该标签的用户数”为3。它的IDF值 = 4 ÷ 3 = 1.33，这里4是因为有4个用户。

	金融战争	做空	海贼王	路飞	甚平	三傻大闹宝莱坞	学霸
用户A	1	1				1	1
用户B			2	2	1		
用户C			1	1		1	1
用户D	1	1	1	1	1		
带该标签用户数	2	2	3	3	2	2	2
IDF值	2	2	1.33	1.33	2	2	2

第五步，汇总计算出用户A的每个标签偏好值。

如下图中，用户A对三傻大闹宝莱坞的偏好值为：1×3×1×0.25×2=1.5。

行为权重	1	1	1	1	1	1	1
行为次数	2	2				3	3
衰减系数	0.73	0.73				1	1
TF值	0.25	0.25				0.25	0.25
IDF值	2	2	1.33	1.33	2	2	2
偏好值	0.73	0.73	0	0	0	1.5	1.5

用这种方式，我们就可以为用户打上海量的标签，只用用户行为足够多，我们就能捕捉的用户的偏好数据。

八、总结

推荐系统的用户画像主要有两种：基本画像和偏好画像。基本画像是用户的个人属性，如年龄，性别，居住城市等。用户偏好画像是推荐系统中的重点，它一般用用户偏好程度 = 行为类型权重值 × 次数 × 时间衰减 × TFIDF值计算出来。用户画像在推荐系统中用于召回和模型训练。

本文由 @菠萝王子 原创发布于人人都是产品经理。未经许可，禁止转载

题图来自Unsplash，基于CC0协议

发表评论

发表神评妙论

发表

评论列表（3条）

关访天4h

如果一个人深刻了解并耐心利用用户画像就会产生巨大的价值！非常感谢主编对用户画像的详细介绍！

2020-04-30 回复 3

永生永世爱杰娜

社会不断在进步,系统也在升级

2020-04-28 回复 1

堪投象