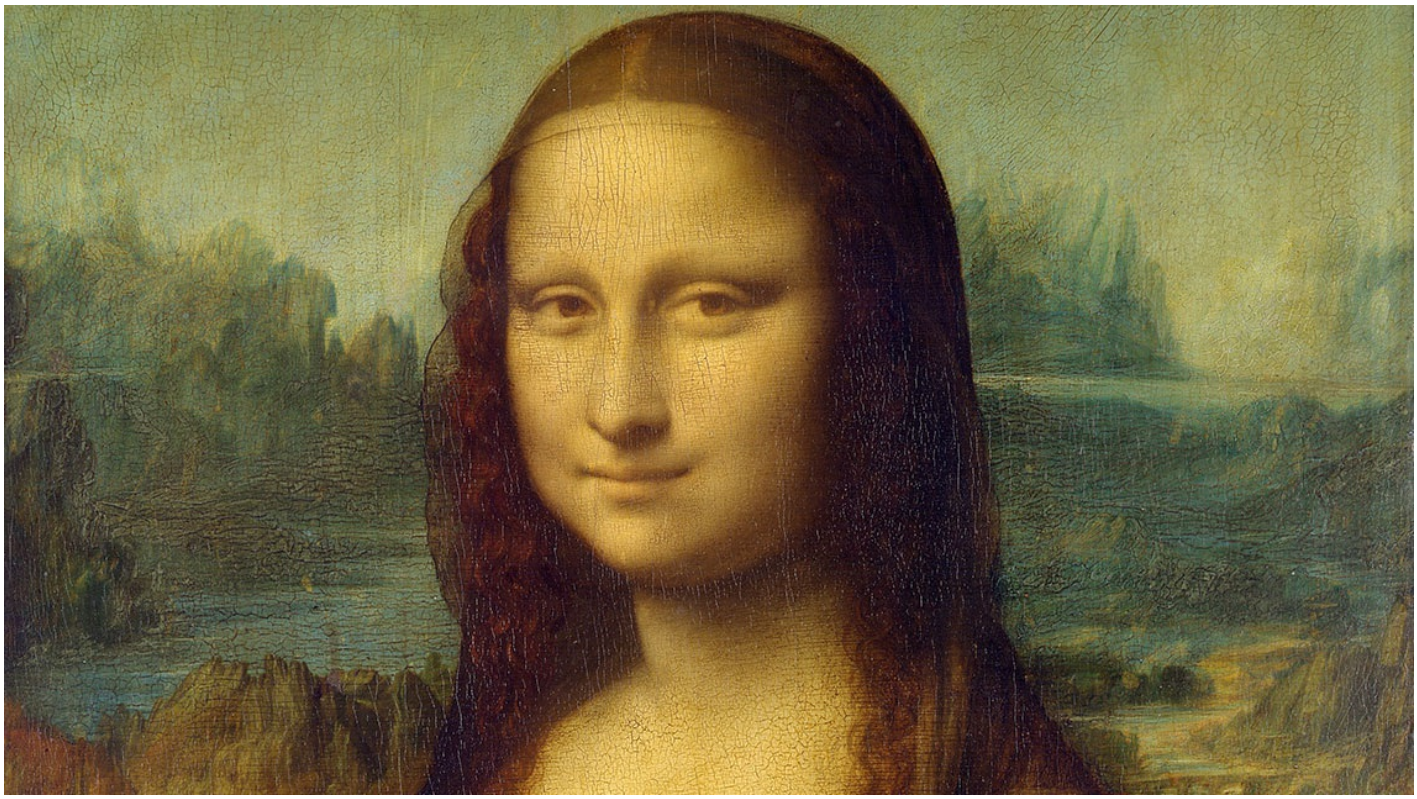
【内容推荐】画鬼容易画人难：用户画像的“能”和“不能”



做好一个推荐系统，总共分三步：

1. 认识每一个用户；
2. 给他推荐他感兴趣的东西；
3. 坐等各项指标上升。

开个玩笑，如果这么简单的话，那么你和我都要失业了；但是话说回来，认识用户是必须的，不过不用担心，认识用户不用请他们吃饭，这就是我们常常听说的“用户画像”这个词。今天，我就来跟你聊一聊：用户画像的那些事儿。

用户画像比较抽象，就像每个人都听说过鬼，但很少有人见过。事实上，它也没有那么神秘，只是大家对它有误解，要么觉得没什么用，要么觉得它是“银弹”，可能相信后者的人略多一些，但实际上这两种看法都不准确。

什么是用户画像

先说说“用户画像”这个词，它对应的英文有两个：Personas 和 User Profile。Personas属于交互设计领域的概念，不在本文讨论范围内，请出门右转去找交互设计师们聊，留下来的人，我们聊聊 User Profile 这种用户画像。

User Profile 原本用于营销领域。营销人员需要对营销的客户有更精准的认识，从而能够更有针对性地对客户和市场制定营销方案。

这个理念本身没有错，但是有一个问题：传统营销领域，是以市场销售人员为第一人称视角去看待客户的，也就是用户画像为营销人员服务。

在这种用途下谈论的用户画像，和我们即将在推荐系统领域谈论的相差有点大；但是很遗憾，今天在媒体上看到的大多数“用户画像”案例分享，都停留在这个意思上。

比如最常见的用户画像出现在高大上的PPT上：用标签云的方式绘制一个人的形状，或者在一个人物形象旁边列出若干人口统计学属性，以此来表达“用户画像”这个概念。

看上去非常酷炫，但是我得悄悄告诉你一个赤裸裸的真相：越酷炫的用户画像越没什么用。

为什么会这样？根本原因是：用户画像应该给机器看，而不是给人看。

既然是给机器看的，那么画像是不是酷炫、是不是像、维度是不是人类可读，都不重要。那它到底是个什么样子呢？先别急，听我慢慢讲。

一个推荐系统来到这个世界上，它只有一个使命，就是要在用户（User）和物品（Item）之间建立连接。

一般方式就是，对用户和物品之间的匹配评分，也就是预测用户评分或者偏好。推荐系统在对匹配评分前，则首先就要将用户和物品都向量化，这样才能进行计算。

而根据推荐算法不同，向量化的方式也不同，最终对匹配评分的做法也不同，在后面讲到具体推荐算法时你会看到这一点。

用户向量化后的结果，就是User Profile，俗称“用户画像”。**所以，用户画像不是推荐系统的目的，而是在构建推荐系统的过程中产生的一个关键环节的副产品。**

另外，通常大型推荐系统一般都分为召回和排序两个阶段，这个在后面我会专门讲到。

因为全量物品通常数量非常大，无法为一个用户（User）逐一计算每一个物品（Item）的评分，这时候就需要一个召回阶段，其实就是预先筛选一部分物品（Item），从而降低计算量，用户画像除了用于最终匹配评分，还要用在召回。所以，构建用户画像就要以这两个阶段为目的。

用户画像的关键因素

举个例子，我想去吃点夜宵，楼下有五家大排档，那么从推荐系统的思路来看，我怎么选择呢？

首先就是将五家大排档向量化，我暂定向量的维度有：

1. 价格，1~5分，最贵的1分，最便宜的5分；
2. 种类，1~5分，只烤馒头片的是1分，天上飞的、海里游的、地上跑的、地里种的都有就是5分；
3. 味道，1~5分，根据以前吃的，最难吃的是1分，最好吃的是5分。

现在每一个大排档都有一个向量，我自己也要有一个对应的向量，就是你有多看中这三个元素：

1. 价格： 1~5分，土豪不差钱就是1分，囊中羞涩就是5分。
2. 种类： 1~5分，早就想好吃什么了不在乎选择多不多1分，看看再说就是5分
3. 味道： 1~5分，只是果腹就是1分，资深吃货就是5分

这样一来就可以对五家大排档做匹配打分了，你很容易得出哪家大排档最适合。

假如我的向量是：

价格： 3
种类： 5
味道： 5

这就是一个大排档推荐系统的简单用户画像了，是不是很简单！

这里可以简单计算一下：每一个因素相乘后再相加，就得到每一个大排档的评分了。

接下来我来围绕这个大排档推荐系统的用户画像，看看建立用户画像的关键因素：**第一个是维度，第二个是量化。**

首先我先来说说“维度”。

看前面这个例子，我定下来的几个维度：价格、种类、味道。这几个维度有三个特点：

1 每个维度的名称都是可理解的。

当我们去给每一个大排档计算评分时，想象你是一台计算机，你读取了用户画像的“价格”取值为3，再去取出一个大排档的“价格”评分，两者相乘，用户画像的维度“价格”和大排档的“价格”天然匹配上了。

因为是同一个名字；但是计算机很傻，你把大排档的这个维度换成“价钱”，它就不知道该如何是好了。

另一方面，对这三个维度，把两边同时换成1、 2、3或者a、b、c都是可以的，也不影响计算结果，计算机依然能够匹配上；所以用户画像的维度不一定需要人类能够理解，只要计算机能把两边对应上就可以了。

2 维度的数量是我拍脑袋定的。

假如是根据用户的阅读历史挖掘阅读兴趣标签，那么我们无法提前知道用户有哪些标签，也就不能确定用户画像有哪些维度，所以第二点也不是必须的。

3 有哪些维度也是我拍脑袋确定的。

因为这一点也不是必须的，用户画像的维度个数可以不用确定。理论上来说维度越多，画像越精细，但带来的计算代价也是很大的，需要权衡。

虽然这里以标签作为例子，但是你要注意，用户画像是向量化结果，而不是标签化。标签化只是向量化的一种，因为向量的维度不一定需要人理解。

其次，我来说说量化。

我们这里的量化都是主观的，而在实际生产系统上，用户画像每个维度的量化，应该交给机器，而且以目标为导向，以推荐效果好坏来反向优化出用户画像才有意义，像这里这个简单的例子，没有去管推荐效果而先行主观量化了每一个维度，是大忌。

所以用户画像的量化是和第三个关键元素“效果”息息相关的。前面已经说过，不要为了用户画像而用户画像，它只是推荐系统的一个副产品，所以要根据使用效果（排序好坏、召回覆盖等指标）来指导用户画像的量化。

用户画像构建的方法

再来整体说说怎么构建用户画像，按照对用户向量化的手段来分，用户画像构建方法分成三类：

1. 第一类就是查户口。

直接使用原始数据作为用户画像的内容，如注册资料等人口统计学信息，或者购买历史，阅读历史等，除了数据清洗等工作，数据本身并没有做任何抽象和归纳。这就跟查户口一样，没什么技术含量，但通常对于用户冷启动等场景非常有用。

2. 第二类就是堆数据。方法就是堆积历史数据，做统计工作，这是最常见的用户画像数据，常见的兴趣标签，就是这一类，就是从历史行为数据中去挖掘出标签，然后在标签维度上做数据统计，用统计结果作为量化结果。这一类数据贡献了常见的酷炫用户画像。

3. 第三类就是黑盒子。就是用机器学习方法，学习出人类无法直观理解的稠密向量，也最不被非技术人员重视，但实际上在推荐系统中承担的作用非常大。

比如使用潜语义模型构建用户阅读兴趣，或者使用矩阵分解得到的隐因子，或者使用深度学习模型学习用户的Embedding 向量。这一类用户画像数据因为通常是不可解释，不能直接被人看懂。

我会在后面专门讲解这些技术手段，以及它们在推荐系统中的实际使用。

总结

现在总结一下今天的内容：

1. 用户画像到底是什么？它是对用户信息的向量化表示，为什么不向量化表示不行呢？因为没办法交给计算机计算，而且，用户画像是给机器看的，而不是给人看的。
2. 用户画像的关键元素有哪些？维度、量化。用户画像是跟着使用效果走的，用户画像本身并不是目的。
3. 通常构建用户画像的手段有哪几类？有三类，第一类只会查户口做记录，第二类就是堆数据做统计，第三类就是黑盒子看不懂。

你可以分享一下你现在正在经历的用户画像是什么样的，它有哪些优点和哪些问题，是不是为了展示给人看而构建的酷炫用户画像呢？欢迎留言和我一起讨论。感谢你的收听，我们下次再见。

精选留言

你这个狗东东

写的真好。我们有时候在做用户画像就跟做BP一样，纯粹是为了彰显我们的用户净值多高，我们的市场定位多么明确。其实推荐系统里的用户画像，是根据目标来定，比如你需要提高内容的点击率，需要了解的维度可能是用户以往阅读内容的维度，用户订阅的分类维度……而其他无关紧要的标签其实不重要

2018-03-12 16:06

作者回复

是这样的。

2018-03-13 07:43

yxj

刑老师，我理解的第三类方法用户行为画像机器学习也要通过前两种方法堆积数据或者人为去选择定义输入的特征（feature），然后再通过机器学习或者深度学习的算法黑箱学习匹配以求得一个准确率很高的算法，但我们学到的是算法，要对用户画像还是要这些数据做为这个学得的算法输入该用户的feature向量才能准确画出。

2018-03-12 09:10

作者回复

比如用户点击物品组成的时序数据，是你说的feature ，经过word2vec(学习算法)之后得到物品的向量，进一步变成用户的向量

。就是用户画像的一部分了。

2018-03-14 23:23

谢烟客

1. 查户口做记录一个无奈之处是：总要给公司领导一个交代。

2. 先不抱怨领导不理解推荐系统，反转位置，非技术体系领导如何考核管理推荐系统团队？

2018-03-12 18:39

作者回复

是个大话题，三两句无法表达我那些心里话。

2018-03-14 23:16

XzAmrzs

您好，讲的很专业，但是这里“每个因素相乘后再相加，就得到每一个大排挡的评分了”有点没看懂，比如(3,5,4)是大排档，“我”的向量是(2,3,5)，那么相乘后的匹配评分是2×3+5×3+5×4=41吗？那么是怎么个匹配标准？这里是分数越高越匹配还是越低越匹配？照我的想法应该是当两向量平行的时候最匹配才对

2018-03-12 10:10

作者回复

一般是按照分数排序后，矮子中选高的。

2018-03-14 06:38

小鑫

能讲讲如何从历史行为数据中挖掘（产生）标签么？标签是事先拍脑袋定的，还是通过算法自动产生的？有监督学习不都是事先定义好标签么？

2018-03-26 23:56

叶晓锋

我们现在就有一个用户画像系统，而且还有一个后台进行管理和维护，画像系统其实是一个标签系统，分成九大类数百个标签，提供给产品，运营，管理人员查看。我们的模型有时也会部分标签数据，但基本还会在此基础上做演算，例如归一化，独热编码等。

2018-03-12 12:44

作者回复

给人看和给机器看的用户画像不一样。

2018-03-13 07:44

Mars

目前我们画像的策略是将用户画像分为四类：基础类，行为类，衍生类和模型类，首末分别对应到老师讲的查户口和黑箱，中间两个一起是统计类。这里行为类的直接统计的结果，衍生类的通过行为类二次计算的行为结果。主要目的是支撑运营和推荐，看了这节，确实要多想想，不能为了用户画像而画像，要用推荐的结果去反推优化画像。

顺便请教下老师，画像除了我上面提到的两个目的，还有其他目的或者还是其他什么的副产品嘛?

2018-05-26 18:00

惜心（伟祺）

用户画像是对用户特征的向量化 这个定义很好

把用户画像的定义泛化了 不一定非得人懂的标签才是用户画像 具体的维度跟目的走 在这个目的下描述用户的特征都是用户的画像标签

2018-03-24 16:46

auroroa

用户画像两个目的：

1、向量化后让机器读懂每个用户

2、召回阶段减少计算量

可以这么理解吗？？？

2018-03-23 17:44

作者回复

减少计算量不是目的。应该是这样理解：

1. 计算机只能处理结构化信息，所以必须结构化才能让计算机去计算。

2. 用结构化的方式让用户表示得越细腻越好。

2018-03-23 19:30

jt120

之前看过推荐系统实战

对这里的大排档例子有疑问，如果是UC，那么是找相似用户，如果是IC，是找相似物品，这里的例子，为什么是直接找用户和物品的相似，这两个实体，是不是没可比性

2018-03-14 08:50

作者回复

两个向量的维度一致：数量一致，每个维度意义一致，就可以去衡量他们的距离远近。

2018-03-14 23:08

holysky

已买,赶紧更新啊，我要从0开始推荐体统了

2018-03-13 22:10

作者回复

骚年，不要急，慢慢来。

2018-03-14 23:10

技术猿

邢老师，用户画像向量做好啦，如何去跟物品做匹配，对应的应该是一个区间还是一个固定的物品？

还有就是用户画像维度的定义主要依据是什么

2018-08-13 19:00

吴峰

这里用户画像的范畴，除了user的向量化，是否也包含item的向量化。毕竟item有哪些维度、如何量化，也存在拍脑袋的空间。

2018-07-19 18:29

作者回复

用户画像就是user profile

item profile 则被叫做物品画像。

用户画像被计算出来的部分，通常和item的维度一致，而自然属性部分则各有各的。

2018-07-23 13:35

黄鸿强

邢老师，我们产品是款唱歌类app，现在在做画像时年龄这个维度非常难收集到，从用户演唱歌曲来看没有很明显的界限，有没有什么好的建议？

2018-03-20 19:01

作者回复

还是得先收集标注数据。

2018-03-21 09:59

栈

隐语义模型和矩阵分解有什么区别？我的理解是，它们是同一个东西的两个不同名称而已。

2018-03-17 15:14

作者回复

方法不同。

2018-03-21 09:50

185

酷炫画像也很好，能让土豪老板或小白领导作出马上签合同的决定，现在我想到了echart

2018-03-15 06:49

作者回复

哪怕是一张卫生纸，一条底裤，都有它的作用。

2018-03-15 20:47

孤帆

老师会写根据用户实时行为数据推荐物品的文章吗？

2018-03-13 23:51

作者回复

内容都在开篇词的目录中的。后续会有图书出版，会比专栏更丰富和深入。

2018-03-14 23:09

禾子先生

听了老师的课，才知道自己原来构建的推荐系统是多么的low，期待后续章节

2018-03-12 08:39

作者回复

我加油。欢迎多分享传阅。

2018-03-13 07:45

Drxan

不错

2018-03-12 08:36

作者回复

欢迎再来。

2018-03-14 23:25

钱姚

第三种embedding是类似词向量那样的结果吗？具体生成步骤能讲详细点吗？

2018-12-17 23:49