R&S[29] | 浅谈用户理解

原创 机智的叉烧 CS的陋室 2020-06-27

搭配主题，效果更佳

Mission Impossible

Einmannkapelle - TV und Filmmusik

往期回顾：

- R&S[27] | 用户画像初探
- R&S[26] | 搜索领域算法需要掌握的知识
- R&S[25] | 搜索中的意图识别
- R&S[24] | 浅谈Query理解和分析
- R&S[23] | 搜索系统中的纠错问题

最近是做了不少和用户理解有关的工作，前段时间也总结了一些有关用户画像的东西：R&S[27] | 用户画像初探，但用户理解往往不局限在用户画像本身，而还有更多的操作，这些操作又不见得是围绕用户本身进行的，而还要为整个推荐或者搜索系统服务。调出用户理解本身，在搜索系统也好，推荐系统也好，都有至关重要的作用，所谓的“千人前面”和“个性化”，就是来源于用户理解，用户理解的程度直接决定了系统个性化的能力，下面来聊一下近期我对用户理解的思考吧。

其实本文的核心目的，就是想让大家不要简单的把用户理解看成是embedding。

什么是用户理解

用户理解建立在个性化需求基础上搭建，可以说是一个子系统，目标就是理解用户，为用户提供定制化、个性化的服务，那么核心的用户理解就是要解决这么几个问题：

- 如何定位一个用户。游客or账户，个人or群体。
- 用户有哪些特点。性别年龄之类的基础性质，武侠片爱情片之类的偏好，抽象成向量化的模糊偏好。
- 如何描述用户的特点。文本类、向量化，甚至还有一些等级之类的信息，描述用户特点的方式还是很多。
- 用户理解如何命中物料。用户理解的结果最终要被拿来使用，用户偏好是需要和物料进行对比的，那么物料和偏好的对齐就非常重要。

无论是哪个问题，最终的落脚点都是要让我们更好地理解用户，这种理解面向的是最终的搜索和推荐结果，要让我们的理解有用，这个其实和处对象类似，我们要去探求的是用户在我们产品中的期望和意图。

怎么理解用户

理解用户的途径在目前的互联网环境中，其实非常简单，但是又由于场景不同，信息获取的形式和内容又会很不一样。

用户主动输入

用户主动输入是最传统但是又最直接的方法，很多内容都是可以直接拿来用的。这种获取的数据有如下特点：

- 用户的基本信息一般会比较易得。
- 主动填写意味着用户对产品的期待还是不低的。
- 不见得是真的，无论是基础信息类，还是一些偏好信息。基础信息往往会有人隐瞒，而偏好信息，说真的用户自己喜欢什么真不见的知道，或者可能因为什么原因（例如害羞）而不填上去。
- 信息相对简单，不用指望用户填非常复杂的信息。
- 相信我，只有行为才不会骗人。

基本行为挖掘

基本行为推理应该是最基本的用户画像信息，其实逻辑也非常简单，从用户近期点击多的内容抽取一些信息，例如用户最近经常点外设，可能用户最近想了解外设的内容。这种推理非常简单，只需要抽取用户行为的一些“规律”即可。

这种“基本行为”往往是可解释的，简单的，而这种简单的东西往往会产生更多的玩法，例如不同角度维度的偏好衡量，例如对电影偏好，可以有对演员的、导演的、电影类型的、年代的、出品国的等等，从点击行为角度，可以看曝光点击，可以看点击占比，可以看主动点赞等等，时间维度有长期、短期、超短期等等。

标签推理

有些标签，其实我们并不能直接获取，但是我们可以通过一些推理获取，我这里之所以用“推理”，是因为这个推理的方法是有很多的，当然模型是一个非常常用的方法，但是别一天天的表现出自己只会模型的样子，有时候规则就是一个非常好的信息。

首先就是聊规则。举个例子，早上8点定时从一个地方出发，然后去另一个地方，在这个地方到晚上9点离开后回到早上8点出发的位置，一天天如此循环，我们只需要根据用户的位置信息和时间结合，就能够推理出这人大概率是上班族，应该是早上8点出门上班，晚上9点下班，进一步就有上班地和下班地，更为激进

的，记录路线和速度，还能知道用户的上下班交通方式，这些，都不需要用户主动提供，我们根据用户的日志就能获取。

当然很多内容我们是无法直接通过规则推理，这时候通过多特征模型的方式就能推理出来，例如性别，基于用户看的电影、平时经常去的地方（商店 or 篮球场），这些因素综合起来，就能推理出。

这里要强调一个点，我们的目标是推理没错，但是别忘了我们的目标——用户个性化推荐or搜索，而不在乎用户是否真的是这个所谓的标签，如用户性别，我们可能不会去care让是否真的是男女，而是在乎我们给他标记的性别背后所体现的行为特点，男生也有爱逛街的，女生也有爱打球的，男生也有更多是女生喜欢的偏好，女生亦然，我们要做的是，即使我们给一位男生真的推了女装，他也不会反感，因为这位男生最近真的逛了女装店（背后的原因按下不表，手机给女友用了也好，自己突然想穿女装也罢，还是想给女朋友买礼物都行对吧）。

embedding

虽然embedding的方法很多，但是核心目的还是想通过一串数很好地表现一个用户的偏好，这个思路其实就和NLP类似，我不在乎这些数的含义，我在乎的是基于这串数真的能给用户推荐他想要的东西。

我这里想分为3类，方便大家更好地理解。

首先就是仅基于用户行为的，这种一般会比较传统，最基本的就是基于word2vector的序列型向量化抽取，另外一种就是矩阵分解，通过对用户-物料的点击矩阵进行序列化。

第二种是基于相似度模型的，其实这种很多就是用点击率作为目标去做的，用户和item以点击率为目标训练模型，然后抽取embedding层结果作为训练结果。最典型的是DSSM和Youtobe的模型了，另外FM系列之类的还是可能会被用到。这种模型可能会用到用户点击序列，但是还可以加入更多的外部特征，例如性别年龄，物料的上线时间等等。

第三种就是比较新的图模型了。其实像word2vector从某种程度上说也是图模型，单独划分出来第三部分是因为现在的图模型在原有分析一步关系到分析多步关系，这种进步是非常巨大的，玩法也比较多，deepwalk、node2vector之类的都是比较新潮的东西。

用户理解的结果怎么用

要想你的东西有用，就必须让别人知道这个东西怎么用。抛开产品、数据分析层面的应用，我们只谈算法是怎么用的。

拓展用户理解

用用户理解来拓展用户理解，是一种套娃，不过事实确实如此，举个案例吧，基础画像在注册的时候很多用户都会填，聚类的到的群体画像往往可以为新用户打上和它类似的用户所拥有的标签（说白了就是一种userCF）。

召回

召回层对用户理解应该是强依赖，与其说召回是找到和用户相关的东西，不如说是抛弃用户肯定不会喜欢的东西，这就是召回层用户理解的核心作用。

基于用户的标签，如用户喜欢叉烧的文章，那就当然可以给用户继续叉烧的文章，召回层就可以召回很多回来。

基于embedding就有更多的玩法了，向量召回直接召回相似的item（矩阵分解之类的方法往往可以把user和item的embedding对齐），向量召回相似用户再召回相似item等等。

排序

做点击率预估之类的操作，大家都懂，不多废话。

小结

但凡是个性化，无论是搜索还是推荐，都需要有非常精准的用户理解，然而在日常的应用中，我们可能只关注用户特征怎么装入模型，怎么使用，很多时候会忽略用户理解的核心工作，用户理解信息从哪来，怎么做，到哪去之类的，在本文的讨论中，我也是阐述了这块的来龙去脉，也为大家推荐系统知识体系添加一块拼图吧~