R&S[27] | 用户画像初探

原创 机智的叉烧 CS的陋室 2020-05-25

Girls Like You

Maroon 5 - Girls Like You (WondaGurl Remix)

往期回顾：

- 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
- R&S[22] | 搜索系统中的召回

R&S | 用户画像初探

无论是推荐系统还是搜索系统，但凡要做所谓的“个性化”，都离不开对用户进行分析，无论是人群的，还是个人的，这一步终究是难以避免的，现在很多的画像工作可能都回归于产品或者数据分析，而在算法领域，这部分工作却被很多人忽视。今天，来给大家介绍用户画像。

为了让大家对用户画像有一个更加深入的了解，我从是什么、为什么、怎么做加上常用的技术点来和大家聊聊。额，文章写完才来补充这句话，可能很多人看完这篇文章会觉得大家也能想到额，见仁见智吧。

懒人目录：

- 是什么
- 为什么
- 怎么做
- 涉及的技术点
- 推荐阅读

是什么

所谓用户画像，就是根据用户的基本信息、行为信息等数据，对用户进行刻画，从而抽象出有利于后续推荐、搜索、商业化等功能的用户信息模型，和一些书籍和资料的描述可能会不一样，关键在于后面的“有利于”，用户画像的构建是要为了后续任务服务的，例如推荐中的召回需要依赖用户偏好进行召回，如果画像内容无法命中物料中的内容，那画像的内容无法产生作用。

为什么

说个例子吧，我们现在要给一个用户进行个性化推荐。那么，怎么给这个用户做个性化推荐，首先就要定义这个个性化，就是针对这个用户的特性进行推荐，什么叫做用户的特性，要描述这个用户，那就需要构建一些特征，然后挖掘一些这个用户的特征，然后根据这个特征再来进行推荐，例如根据用户平时的点击和搜索行为来分析，他喜欢打篮球，喜欢球鞋，那就可以给他推相关的东西了，而这里我们的关心点就在于，要挖掘出这个人喜欢打篮球、喜欢球鞋，这是要做个性化推荐非常重要的先决条件。

怎么做

那么，怎么去挖掘用户画像的内容呢，这其实是一个困难而且是长期的过程。来看看几种常见的画像构建方法。

基础画像

也有叫做用户属性维度的，还是先举几个例子：性别年龄职业位置，这些都是比较基础的画像，这些大都可以通过用户填写个人资料来获取，另外有一些其实可以大概的挖掘推断出来，例如通过位置，可以大概推断职业，如学生，大学生基本就是宿舍教室图书馆，如上班族，工作日基本两点一线。

用户行为

用户行为应该是挖掘空间最大，油水最多的一个，但是需要花费的经历其实也是最多的，可以这么理解，用户的行为是不会骗人的。

然而，用户行为其实又一定程度依赖对物料性质的挖掘，换言之，。例如一篇新闻，是讲新冠肺炎的，那么，我是怎么知道这讲的是和新冠肺炎有关呢，这里面涉及了关键词抽取、主题模型等内容，再例如一双鞋，用户点击了红色的，那么这个是鞋子的性质，就需要被挖掘到。然后根据用户的点击行为，把用户对应“新冠肺炎”的兴趣点、“鞋子：红色”均提取出来。

OK，回到用户画像，用户有了行为，怎么衡量用户是否喜欢？这里提供一个简单的方法：TFIDF。这应该是NLP领域里面非常基础的文本表示方法，那么在这里，我们也要这么去操作。

有标签T，和用户行为（如点击）U，有$\omega(T,U)$表示用户U触发了一次T，则TF可以表示为：

$$TF(P,T)=\frac{\omega(P,T)}{\sum_{T_i=tags}\omega(P,T_i)}$$

这里其实考虑的是两者的关联，用户点击了很多东西，这个TF实质上就是看用户点得多的是哪些。

那么问题就来了，用户点的多真的就表示用户喜欢了吗，不见得，很可能因为我们给他曝光的太多了，此时我们就需要IDF来进行调整了。

$$IDF(P,T) = log\frac{\sum_{P_j=users}\sum_{T_i=tags}\omega(P_j,T_i)}{\sum_{P_j=users}\omega(P_j,T)}$$

没错，这个IDF实质上就是用于分析某个tag的总体曝光程度，他出现的频次低，而会被用户点击，就说明这个用户对它是真爱了，于是，用户偏好就有了。

$$rel(P,T) = TF(P,T) \times IDF(P,T)$$

首先，这只是一种衡量方法，第二，这只是一个简单的衡量方法，还要考虑到时间衰减、归一化等多个内容的汇总。

涉及的技术点

好了，该说说涉及的技术点了。由于用户画像的使用会有多个领域，这里基本上只会从算法技术方向来看吧。

存储

存储应该是最简单直接的需求了，我也建议大家多学学各种数据存储的方式，至少要知道怎么存和怎么取吧，首先一般的存储，会用的是mysql，数据规模较大的会存在hive，另外备份或者是稳定的数据则可以存在hdfs上，但是由于mysql的读取性能其实并不高，因此我们需要一些读取数据的工具，简单的有k-v存储的redis，还有能够见多种索引的搜索引擎elasticsearch，当然还有一些nosql的数据库。

标签抽取

重头戏。

标签的抽取，技术上当然就有mapreduce、spark等大数据工具，进行联表、批量计算等。

算法上，标签挖掘实质上是一个数据挖掘的工作，那么很多数据挖掘的东西我们就可以用起来了，关联分析、聚类分析等，另外一些NLP的方法是非常重要的，关键词提取、主题模型等，在者本身就有一些推荐系统的技术就可以用来进行挖掘，矩阵分解、向量召回、协同过滤。

推荐阅读

下面这些内容其实在我的candyhub上已经提到过了，这里再放一次在这里，方便大家进一步深度学习。

- 推荐系统——用户画像：https://blog.csdn.net/sin_geek/article/details/83064127
- 美团机器学习实践，第五章
- 用户画像-方法论与工程化解决方案。新书，刚到手，后面好好看看。

- 推荐系统之用户画像：https://zhuanlan.zhihu.com/p/103754069，文章来自汽车之家的产品，对技术和业务的理解都很深刻。
- 一文读懂推荐系统用户画像：https://baijiahao.baidu.com/s?id=1665215120349875742&wfr=spider&for=pc
- 用户画像之标签聚类：https://blog.csdn.net/u014156013/article/details/82657290
- 用户画像—打用户行为标签：https://blog.csdn.net/u014156013/article/details/82657080
- 用户画像—计算用户偏好标签及数据指标与表结构设计：https://blog.csdn.net/u014156013/article/details/82656883