用户画像之标签权重算法

赵宏田 人工智能爱好者社区 2018-08-08

点击上方蓝字 轻松关注我们

作者：超人赵，人工智能爱好者社区专栏作者
知乎：
https://www.zhihu.com/people/chao-ji-sai-ya-ren/posts

感谢大家长期以来对文章的关注，最近工作比较忙，好久没更新了。接下来的几篇文章想和大家分享下关于用户画像的一些东西。今天我们先从用户画像的标签权重开始聊起吧。

用户画像：即用户信息标签化，通过收集用户社会属性、消费习惯、偏好特征等各个维度数据，进而对用户或者产品特征属性的刻画，并对这些特征分析统计挖掘潜在价值信息，从而抽象出一个用户的信息全貌，可看做是企业应用大数据的根基，是定向广告投放与个性化推荐的前置条件。
先举个场景，程序员小Z在某电商平台上注册了账号，经过一段时间在该电商平台的web端/app端进行浏览、所搜、收藏商品、下单购物等系列行为，该电商平台数据库已全程记录该用户在平台上的行为，通过系列建模算法，给程序员小Z打上了符合其特征的标签（如下图所示）。此后程序员小Z在该电商平台的相关推荐版块上总能发现自己想买的商品，总能在下单前犹豫不决时收到优惠券的推送，总是在平台上越逛越喜欢....

上面的例子是用户画像一些应用场景。而本文主要分享的是打在用户身上标签的权重是如何确定的。

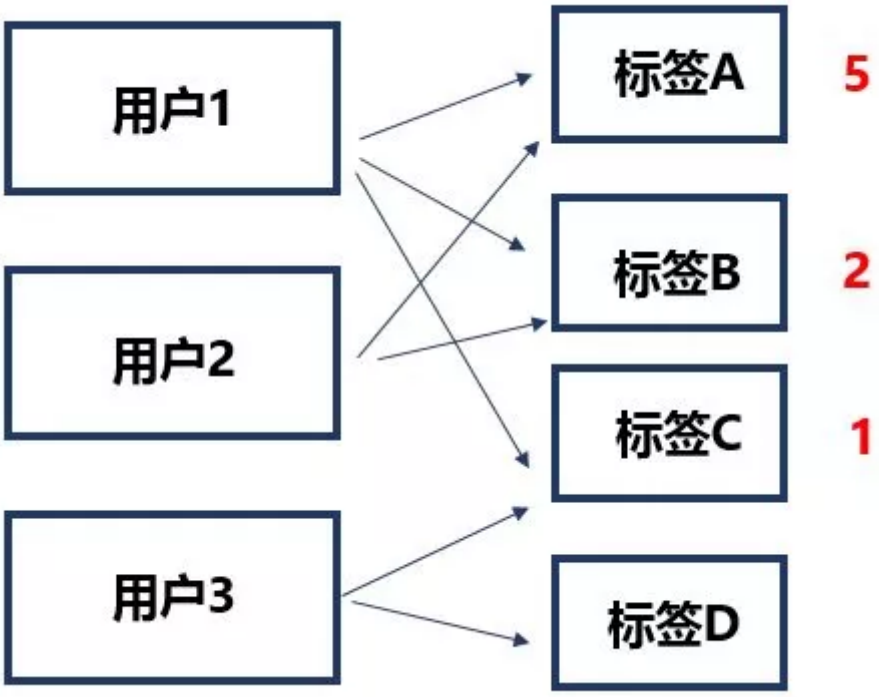
用户id	用户姓名	标签id	标签名称	行为次数	行为类型id	行为类型	行为时间	标签权重
28188144	刘备	ccc90fadsgdsdf	大数据	2	1	搜索	2017/5/1	2.5
79766135	关云长	dsgsd9009	小说	2	2	浏览	2017/6/10	5.66
54619827	张飞	sdfafsadf890	mysql从入门到精通	1	3	购买	2017/7/10	3.2

今天要讲述的重点

如上图所示，一个用户标签表里面包括常见的字段如：用户id、用户姓名、标签id、标签名称、用户与该标签发生行为的次数（如搜索了两次“大数据”这个关键词）、行为类型（不同的行为类型对应用户对商品不同的意愿强度，如购买某商品>收藏某商品>浏览某商品>搜索某商品），行为时间（越久远的时间对用户当前的影响越小，如5年前你会搜索一本高考的书，而现在你会搜索一本考研的书）。**最后非常重要的一个字段是标签权重，该权重影响着对用户属性的归类，属性归类不准确，接下来基于画像对用户进行推荐、营销的准确性也就无从谈起了。**下面我们来讲两种权重的划分方法：

1、基于TF-IDF算法的权重归类

TF-IDF算法是什么思想，这里不做详细展开，简而言之：一个词语的重要性随着它在该文章出现的次数成正比，随它在整个文档集中出现的次数成反比。



比如说我们这里有3个用户和4个标签，**标签和用户之间的关系将会在一定程度上反应出标签之间的关系。**
这里我们用w（P，T）表示一个标签T被用于标记用户P的次数。TF（P，T）表示这个标记次数在用户P所有标签中所占的比重，公式如下图：

$$TF(P,T) = \frac{w(P,T)}{\sum_{T_i \in \text{该用户全部标签}} w(P,T_i)}$$

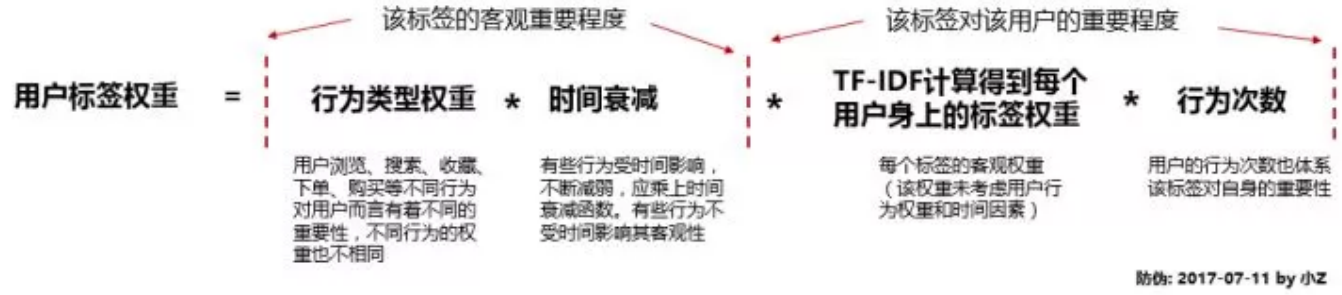
分子：打在某用户身上某个标签的个数
分母：该用户身上全部标签个数

对上面的图来说，用户1身上打了标签A 5个，标签B 2个，标签C 1个，那么用户1身上的A标签TF=5/（5+2+1）。
相应的IDF（P，T）表示标签T在全部标签中的稀缺程度，即这个标签的出现几率。如果一个标签T出现几率很小，并且同时被用于标记某用户，这就使得该用户与该标签T之间的关系更加紧密。

$$IDF(P,T) = \frac{\sum\sum w(P_j,T_i)}{\sum_{P_i \in \text{全部用户}} w(P_j,T)}$$

分子：全部用户的全部标签之和
分母：所有打T标签的用户之和

然后我们根据TF * IDF即可得到该用户该标签的权重值。到这里还没结束，此时的权重是不考虑业务场景，仅考虑用户与标签之间的关系，显然是不够的。还需要考虑到该标签所处的业务场景、发生的时间距今多久、用户产生该标签的行为次数等等因素。我用个图总结下：



关于时间衰减的函数，根据发生时间的先后为用户行为数据分配权重。

时间衰减是指用户的行为会随着时间的过去，历史行为和当前的相关性不断减弱，在建立与时间衰减相关的函数时，我们可套用**牛顿冷却定律数学模型**。牛顿冷却定律描述的场景是：一个较热的物体在一个温度比这个物体低的环境下，这个较热的物体的温度是要降低的，周围的物体温度要上升，最后物体的温度和周围的温度达到平衡，在这个平衡的过程中，较热物体的温度F(t)是随着时间t的增长而呈现指数型衰减，其温度衰减公式为：

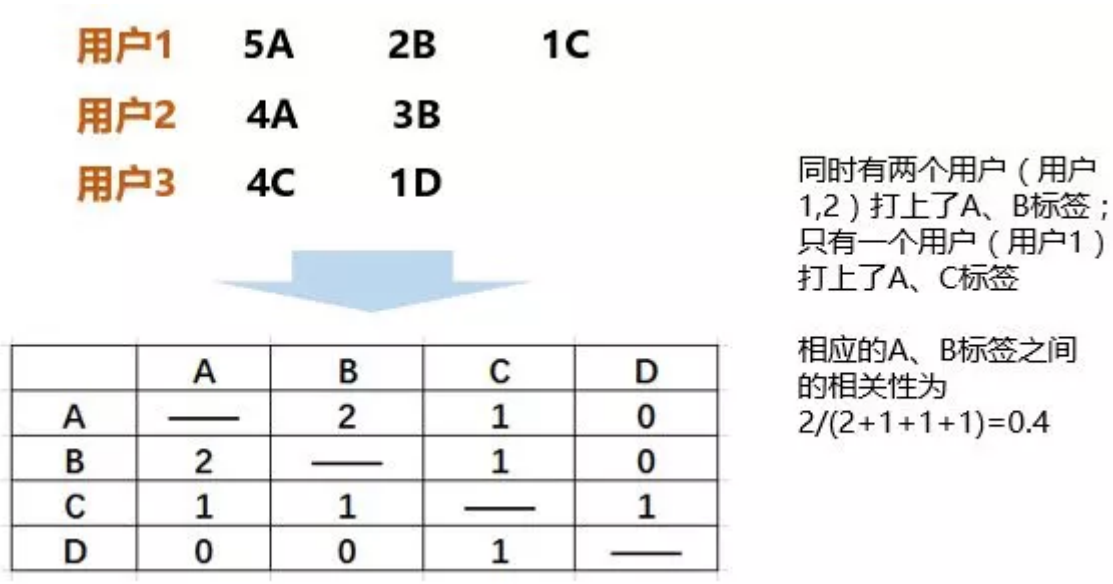
F(t)=初始温度×exp(-冷却系数×间隔的时间)

其中α为衰减常数，通过回归可计算得出。例如：指定45分钟后物体温度为初始温度的0.5，即0.5=1×exp(-a×45)，求得α=0.1556。

2、基于相关系数矩阵的权重归类

这个相关系数矩阵听title挺困难，其实道理十分简单。举个例子：用户1身上打上了5个A标签、2个B标签、1个C标签；用户2身上打上了4个A标签，3个B标签；用户3身上打上了4个C标签、1个D标签。

用个图形象表示一下：



	A	B	C	D
A	——	2	1	0
B	2	——	1	0
C	1	1	——	1
D	0	0	1	——

那么同时打上A、B标签的用户有两个人，这就说明AB之间可能存在某种相关性，当用户量、标签量级越多时，标签两两之间的相关性也越明显。

今天先聊这么多，大家可以留言交流。后面再更新 ...