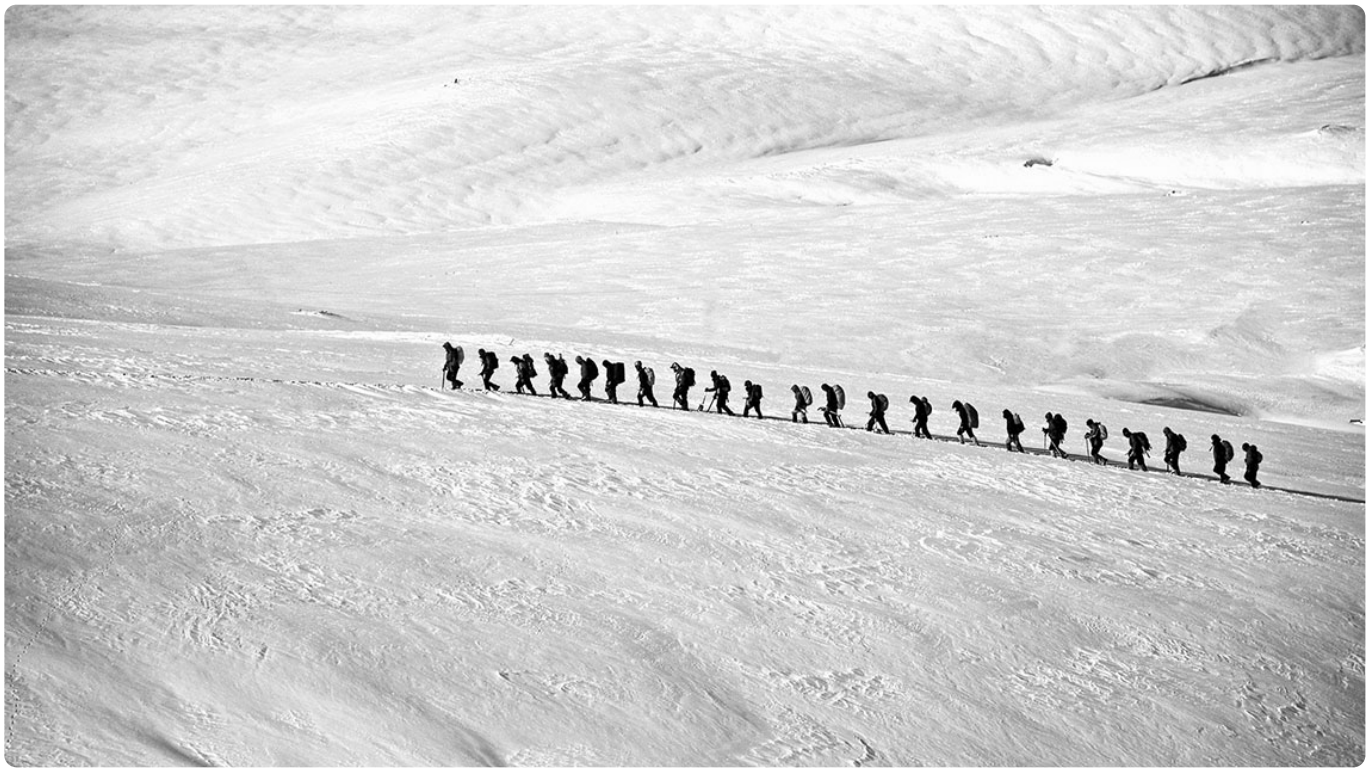
065 | 简单推荐模型之三：基于内容信息的推荐模型

2018-03-02 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 07:33 大小 3.46M

周一的文章中，我们聊了一个最基本的基于流行度的推荐模型。周三我们讨论了基于相似信息的推荐模型。基于相似信息的推荐模型，其核心就是协同过滤的思想，希望能够通过相似的用户或者相似的物品来对当前的场景进行推荐。

然而，不管是基于流行度的推荐，还是协同过滤，这些方法都有一些根本的问题。比如，对于基于流行度预测的推荐来说，推荐结果不是个性化的。因为流行度预测是一种全局的预测，每个人得到的推荐结果是一样的。而协同过滤的问题是强烈依赖相似用户以及相似物品的定义，而且对于新用户或者新物品来说有数据稀缺的问题。因此，在实际应用中，往往不能在整个系统中单独使用协同过滤。

今天，我们来分享一个更加普遍的方法，那就是**基于内容信息的推荐系统**。这种系统在实践中往往更能适应各种不同的推荐场景。

什么是基于内容信息的推荐系统

所谓基于内容信息的推荐系统，其实就是**用特征（Feature）来表示用户、物品以及用户和物品的交互，从而能够把推荐问题转换成为监督学习任务。**

把推荐系统完全定义为监督学习任务，需要有这么几个步骤。

第一，就是我们已经提到的，需要把所有用户、物品的各种信号用特征来表示。这里面往往牵涉非常复杂和繁琐的**特征工程**，也就是看如何能够把不同的信息通过特征表达出来。

第二，就是每一个监督任务都需要面临的问题，如何构造一个**目标函数**，来描述当前的场景。可以说，这是最难的一个部分，也是和基于流行度和基于相似度的推荐系统的最大区别。

内容信息的各类特性

那么，对于物品特性来说，有哪些是比较重要的呢？这里我们肯定没法提供一个完备的列表，那我就来谈一些主要的特性，起到一个抛砖引玉的作用。

第一，**物品的文本信息**。比如商品的名字和描述。这些文字信息可以使用很多文本挖掘（Text Mining）的方式来组成有效的特征。

我们在讲搜索模块的时候，其实就已经提到了一些，比如用 TF-IDF 的方法来形成文本向量。当然，因为文本信息的噪声相对比较大，并且数据维度也比较大（维度等于文本所对应语言的词汇量），很多时候我们都寻求降低这部分数据的维度，降低到一个固定的维度。这种时候，很多所谓“降维”的工具就很有必要了。

传统上，有用“**话题模型**”（Topic Model）对文本进行降维的。也就是说，我们针对每一个文字描述都可以学习到一个话题的分布，这个分布向量可能是 50 维、100 维等等，但是肯定要比原始的词汇向量要小。

近些年，很多人又开始使用各种“**词嵌入向量**”（Word Embedding）的方法来为文字信息降维，从而能够使用一个固定的维度来表达文字信息。

第二，**物品的类别信息（或者物品的知识信息）**。对于新闻文章来说，类别信息是新闻的话题类别，像娱乐新闻、财经新闻或者时政新闻等。而对于商品来说，类别信息是商品的品

类，像电器、床上用品或者生活用品等。这些类别信息往往能够非常有效地抓住物品的整体属性。通常情况下，这样的属性比直接使用文字信息更加直接。

如何能够得到这样的类别信息呢？在有些情况下，这些类别信息是在数据输入的时候获取的。比如通过合作渠道取得新闻文章的时候，类别往往是编辑加上去的。再比如，商品的类别很多时候也是卖家在输入商品的时候加上去的。

当然，也有一些情况，这些类别信息并不是直接获得的；或者是在数据中有很多缺失的情况下，就需要利用机器学习的手段，来构造分类器以获取这些类别信息。我们在这里就不展开讨论这些分类器该如何构建了。

最后需要说明的一点是，除了最基本的类别信息，最近一段时间比较火热的研发领域，就是**利用知识图谱（Knowledge Graph）来对物品的各种信息进行深入挖掘。**很多信息是通过知识图谱推断出来的。

举个例子，某一篇新闻文章是关于美国总统特朗普的，于是这篇文章可能就会自动被打上美国总统、美国政治等其他标签。这种通过一些原始的信息来进一步推断更加丰富的知识信息，也是重要的物品类别特征的处理工作。

最后需要提及的是图像或者其他多媒体的信息。在信息如此丰富的今天，很多物品都有多样的表现形式，比如比较常见的图像、视频等。

那么，如何从这些媒介中提取信息也是非常关键的物品特征工程。和文字信息正好相反，很多多媒体的信息都是稠密（Dense）的向量，因此需要对这些向量进行特殊处理，比如我们首先学习一个分类器，然后再和其他特征的不同分类器组合。

前面我们简单谈了谈物品的特征，下面我们再来看看用户的特征。

对于用户来说，最基础、最首要的肯定是用户的基本特性，包括性别、年龄、地理位置。这三大信息其实可以涵盖用户特性工程中非常大的一块内容。

这里不仅是最基本的这三个特性的值，还有围绕这三个特性发展出来的三大种类的特性。比如，不同性别在文章点击率上的差异，不同年龄层在商品购买上的差异，不同地理位置对不同影视作品的喜好等，这些都是根据这三个特性发展出来的更多的特性。

然后，我们可以为用户进行画像（Profiling）。有显式的用户画像，比如用户自己定义的喜好，或者用户自己认为不愿意看到的物品或者类别。

但是在大多数情况下，用户都不会为我们提供那么精准的回馈信息，甚至完全不会有任何直接的反馈。在这样的情况下，绝大多数的用户画像工作，其实是通过用户的“隐反馈”（Implicit Feedback），来对用户的喜好进行建模。关于如何进行用户画像，我们今天就不在这里展开了。

目标函数

讨论了物品和用户特征的一些基本情况后，我们再来简单聊聊另外一个话题，那就是**目标函数**。我们前面提到，整个基于内容信息的推荐系统就是把所有的信号都当做特征，然后构建一个监督学习任务。

监督学习的一个关键的就是目标函数。对于一个推荐系统来说，都有什么样的目标函数呢？

和纯粹的基于评分（Rating）的协同过滤推荐系统一样，我们可以设置监督学习的目标函数是**拟合评分**。当然，已经有很多学者指出评分并不是推荐系统的真正目标。

那么，在实际系统中比较常见的目标函数有点击率和购买率，也有一些相对比较复杂的目标函数，比如预测用户在某一个物品上的停留时长。

对于究竟在哪种场景中使用什么样的目标函数，这依然是当前的一个主要研究方向。

小结

今天我为你讲了基于内容信息的推荐系统。通俗地说，就是如何把推荐系统当做监督学习任务来看待。

一起来回顾下要点：第一，我们简要介绍了整个基于内容推荐的内涵以及我们这么做的基本假设；第二，我们详细介绍了如何构造一个基于内容的推荐系统，特别是如何构造物品和用户的特征；第三，我们简要地介绍了目标函数的重要性。

最后，给你留一个思考题，如何把我们前面介绍的两种推荐系统模式，也就是基于流行度和协同过滤，也融进基于内容的推荐系统中去呢？

欢迎你给我留言，和我一起讨论。

上一篇 064 | 简单推荐模型之二：基于相似信息的推荐模型

下一篇 066 | 基于隐变量的模型之一：矩阵分解

精选留言 (2)

写留言

微微一笑
2018-03-02

6

能不能再深入一点,增加些工程方面的内容

展开

林彦
2018-03-05

2

1. 基于流行度的算法对于新用户的冷启动问题来说是一个优秀的解决方案。这些算法通过某些流行度的测量标准，比如下载最多的或者购买最多的，来对物品进行排名，并将这些

流行度最高的物品推荐给新用户。当拥有合适的流行度衡量指标时，这个办法虽然基础却很有效，通常可以为其他算法提供很好的基线标准。流行度算法也可以单独作为算法使用，以引导推荐系统在换到其他更切合用户兴趣点的算法（比如协同过滤算法以及基于...

展开