04｜特征工程：推荐系统有哪些可供利用的特征？

2020-10-09 王喆

深度学习推荐系统实战 进入课程 >

讲述：王喆
时长 15:33 大小 14.25M

你好，我是王喆。基础架构篇我们已经讲完了，你掌握得怎么样？希望你已经对深度学习推荐系统有了一个初步的认识。

从这节课开始，我们将会开启一个新的模块，特征工程篇。

如果说整个推荐系统是一个饭馆，那么特征工程就是负责配料和食材的厨师，推荐模型是一个大厨做的菜好不好吃，大厨的厨艺肯定很重要，但配料和食材作为美食的基础也同样重要。而且只有充分了解配料和食材的特点，我们才能把它们的作用发挥到极致。

今天，我们就先来讲讲特征工程，说说到底**什么是特征工程，构建特征工程的基本原则是什么，以及推荐系统中常用的特征有哪些。**相信通过这节课的学习，能让你更好地利用起推荐系统相关的数据提升推荐的效果。

什么是特征工程

在第一节课中我们学习过，推荐系统就是利用“用户信息”“物品信息”“场景信息”这三大部分有价值数据，通过构建推荐模型得出推荐列表的工程系统。

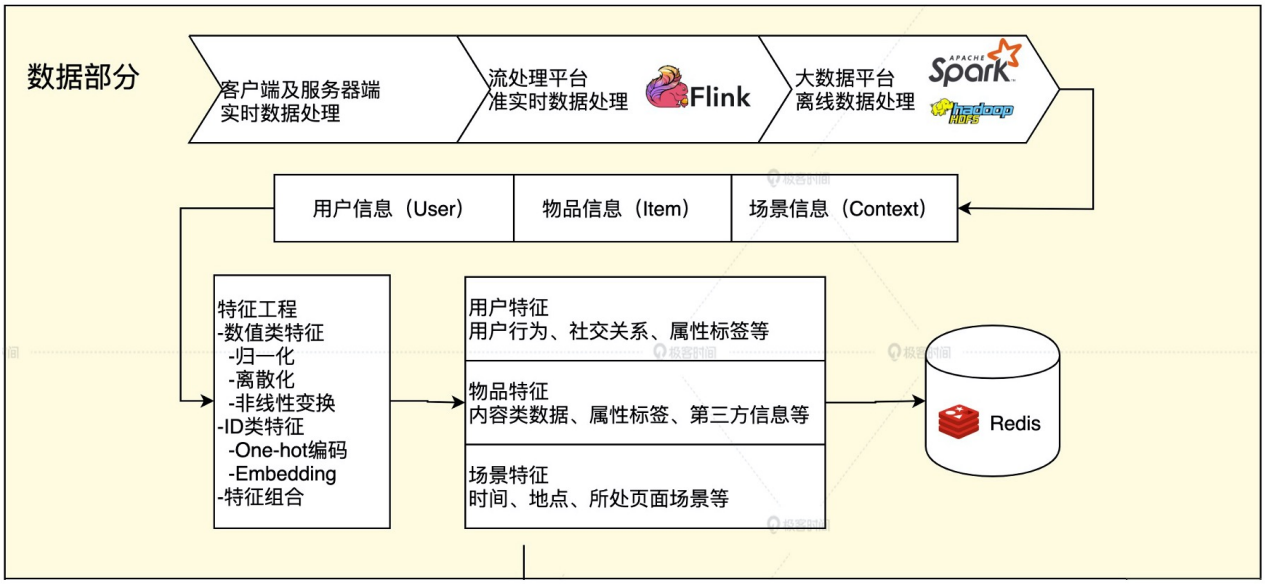

图1 特征工程部分在推荐系统中的位置

在这个系统之中，**特征工程就是利用工程手段从“用户信息”“物品信息”“场景信息”中提取特征的过程。**这个过程说起来容易，但实际做起来其实困难重重。

比如说，一个网站或者 APP 每天收集起来的用户日志，采集来的站外信息，自己公司员工编辑添加的结构化数据那么多，那么庞杂，怎么才能挑出那些对推荐有用的特征呢？

再比如从“推荐模型”的角度来说，一个机器学习模型的输入，往往是一个数值型的向量。那用户性别，用户行为历史这些根本不是数字的信息怎么处理成一个模型可用的数值向量呢？

我们这节课先聚焦第一个问题，“怎么挑出有用特征”，下节课我们再解决第二个问题。都说“理论指导实践”，在展开讲有哪些有用的特征之前，我们先看一看构建特征工程有哪些原则或者规律可以遵循。

构建推荐系统特征工程的原则

我给推荐系统中的特征下了一个比较抽象的定义，**特征其实是对某个行为过程相关信息的抽象表达**。为什么这么说呢？因为一个行为过程必须转换成某种数学形式才能被机器学习模型所学习，为了完成这种转换，我们就必须将这些行为过程中的信息以特征的形式抽取出来。

我们来举个最简单的例子，用户的性别有三个，男、女、未知。但推荐模型没办法直接认识这三个类别，它是一个只认识数字的“严重偏科理工男”，所以我们就需要把它转换成1、2、3（为了方便你理解，这里我用的是一个最简单的方法，不一定是最合适的）这样的数字代号它才能处理。

但是，这种从具体行为信息转化成抽象特征的过程，往往会造成信息的损失。为什么这么说呢？

一是因为具体的推荐行为和场景中包含大量原始的场景、图片和状态信息，保存所有信息的存储空间过大，我们根本无法实现。

二是因为具体的推荐场景中包含大量冗余的、无用的信息，把它们都考虑进来甚至会损害模型的泛化能力。比如说，电影推荐中包含了大量的影片内容信息，我们有没有必要把影片的所有情节都当作特征放进推荐模型中去学习呢？其实没有必要，或者说收效甚微。

这其实也是我们构建推荐系统特征工程的原则：**尽可能地让特征工程抽取出的一组特征，能够保留推荐环境及用户行为过程中的所有“有用“信息，并且尽量摒弃冗余信息。**

接下来，我们就结合一个实际的例子，说一说在电影推荐这个场景下，我们该怎么贯彻特征工程原则来挑选特征。

现在，你就可以先把自己当成是一个用户，假设你正在选择看哪部电影。想一想在这个选择过程中，你都会受什么因素影响呢？如果是我的话，可能影响我的因素有 6 个，把它们

按照重要性由高到低排序就是，**电影类型我是否感兴趣、电影是不是大片、导演和演员我是否喜欢、电影海报是否吸引人、我是否已经观看过该影片以及我当时的心情。**

那站在一个工程师的角度，我们能不能用某些特征把这些要素表达出来呢？我尝试用表格的形式把它们特征化的方法列举了出来：

要　素	能利用的数据	特　征
电影类型我是否感兴趣	当前电影的类型，和我的历史观看影片序列	我的兴趣和电影类型的相似度
电影是不是大片	影片的流行热度分数影片	影片流行度特征
导演和演员我是否喜欢	影片的导演、演员等元数据（相关信息、metadata）和我的历史观影记录	元数据标签类特征 和我感兴趣标签的相似度
电影海报是否吸引人	影片海报的图像	图像内容类特征
我是否已经观看过该影片	用户观看历史	是否观看过的Bool型特征
我当时的心情	无法抽取	无

图2 电影推荐的要素和特征化方式

我们详细来讲一个要素，比如，如何知道一个用户是否对这个电影的类型（动作、喜剧、爱情等）感兴趣。一般来说，我们会利用这个用户的历史观看记录来分析他已有的兴趣偏好，这个兴趣偏好可能是每个电影类型的概率分布，比如动作 45%、喜剧 30%、爱情 25%。也可能是一个通过 Embedding 技术学出来的用户兴趣向量。

这个时候，我们就可以根据这个电影本身的特征，计算出用户对电影的感兴趣程度了。对于其他的特征，我们也都可以通过类似的分析，利用日志、元数据等信息计算得出。

不过，并不是所有的要素都能特征化。比如，“自己当时的心情”这个要素就被我们无奈地舍弃了，这是因为我们很难找到可用的信息，更别说抽取出特征了；再比如，“电影海报是否吸引人“这个要素，我们可以利用一些图像处理的方法提取出海报中的某些要点（比如海报中有哪些演员？是什么风格？），但想面面俱到地提取出海报中所有的图像要素，几乎是不可能的。

因此，**在已有的、可获得的数据基础上，“尽量”保留有用信息是现实中构建特征工程的原则。**

推荐系统中的常用特征

前面我以电影推荐为例，讲解了特征工程的基本原则，互联网中的推荐系统当然不仅限于电影推荐，短视频、新闻、音乐等等都是经典的推荐场景，那么它们常用的特征之间有没有共性呢？确实是有的，推荐系统中常用的特征有五大类，下面我一一来说。

1. 用户行为数据

用户行为数据是推荐系统最常用，也是最关键的数据。用户的潜在兴趣、用户对物品的真实评价都包含在用户的行为历史中。用户行为在推荐系统中一般分为显性反馈行为（Explicit Feedback）和隐性反馈行为（Implicit Feedback）两种，在不同的业务场景中，它们会以不同的形式体现。具体是怎么表现的呢？你可以看我下面给出的几个例子。

业务场景	显性反馈行为	隐性反馈行为
电子商务网站	对商品的评分	点击、加入购物车、购买等
视频网站	对视频的评分、点赞等	点击、播放、播放时长等
新闻类网站	赞、踩等行为	点击、评论等
音乐网站	对歌曲、歌手、专辑的评分	点击、播放、收藏等

图3 不同业务场景下用户行为数据的例子

对用户行为数据的使用往往涉及对业务的理解，不同的行为在抽取特征时的权重不同，而且一些跟业务特点强相关的用户行为需要推荐工程师通过自己的观察才能发现。

在当前的推荐系统特征工程中，隐性反馈行为越来越重要，主要原因是显性反馈行为的收集难度过大，数据量小。在深度学习模型对数据量的要求越来越大的背景下，仅用显性反馈的数据不足以支持推荐系统训练过程的最终收敛。所以，能够反映用户行为特点的隐性反馈是目前特征挖掘的重点。

2. 用户关系数据

互联网本质上就是人与人、人与信息之间的连接。如果说用户行为数据是人与物之间的“连接”日志，那么用户关系数据就是人与人之间连接的记录。就像我们常说的那句话“物以类聚，人以群分”，用户关系数据毫无疑问是非常值得推荐系统利用的有价值信息。

用户关系数据也可以分为“显性”和“隐性”两种，或者称为“强关系”和“弱关系”。如图 4 所示，用户与用户之间可以通过“关注”“好友关系”等连接建立“强关系”，也可以通过“互相点赞”“同处一个社区”，甚至“同看一部电影”建立“弱关系”。

图4 社交网络关系的多样性

在推荐系统中，利用用户关系数据的方式也是多种多样的，比如可以将用户关系作为召回层的一种物品召回方式；也可以通过用户关系建立关系图，使用 Graph Embedding 的方法生成用户和物品的 Embedding；还可以直接利用关系数据，通过“好友”的特征为用户添加新的属性特征；甚至可以利用用户关系数据直接建立社会化推荐系统。

3. 属性、标签类数据

推荐系统中另外一大类特征来源是属性、标签类数据，这里我把属性类和标签类数据归为一组进行讨论，是因为它们本质上都是直接描述用户或者物品的特征。属性和标签的主体可以是用户，也可以是物品。它们的来源非常多样，大体上包含图 5 中的几类。

主　体	类　别	来　源
用户	人口属性数据（性别、年龄、住址等）	用户注册信息、第三方DMP（Data Management Platform，数据管理平台）
	用户兴趣标签	用户选择
物品	物品标签	用户或者系统管理员添加
	物品属性（例如，商品的类别、价格；电影的分类、年代、演员、导演等信息）	后台录入、第三方数据库

图5 属性、标签类数据的分类和来源

用户、物品的属性、标签类数据是最重要的描述型特征。成熟的公司往往会建立一套用户和物品的标签体系，由专门的团队负责维护，典型的例子就是电商公司的商品分类体系；也可以有一些社交化的方法由用户添加。图 6 就是豆瓣的“添加收藏”页面，在添加收藏的过程中，用户需要为收藏对象打上对应的标签，这是一种常见的社交化标签添加方法。

图6 豆瓣的“添加收藏”页面

在推荐系统中使用属性、标签类数据，一般是通过 Multi-hot 编码的方式将其转换成特征向量，一些重要的属性标签类特征也可以先转换成 Embedding，比如业界最新的做法是将标签属性类数据与其描述主体一起构建成知识图谱（Knowledge Graph），在其上施以

Graph Embedding 或者 GNN（Graph Neural Network，图神经网络）生成各节点的 Embedding，再输入推荐模型。这里提到的不同的特征处理方法我们都会在之后的课程中详细来讲。

4. 内容类数据

内容类数据可以看作属性标签型特征的延伸，同样是描述物品或用户的数据，但相比标签类特征，内容类数据往往是大段的描述型文字、图片，甚至视频。

一般来说，内容类数据无法直接转换成推荐系统可以“消化”的特征，需要通过自然语言处理、计算机视觉等技术手段提取关键内容特征，再输入推荐系统。例如，在图片类、视频类或是带有图片的信息流推荐场景中，我们往往会利用计算机视觉模型进行目标检测，抽取图片特征，再把这些特征（要素）转换成标签类数据供推荐系统使用。

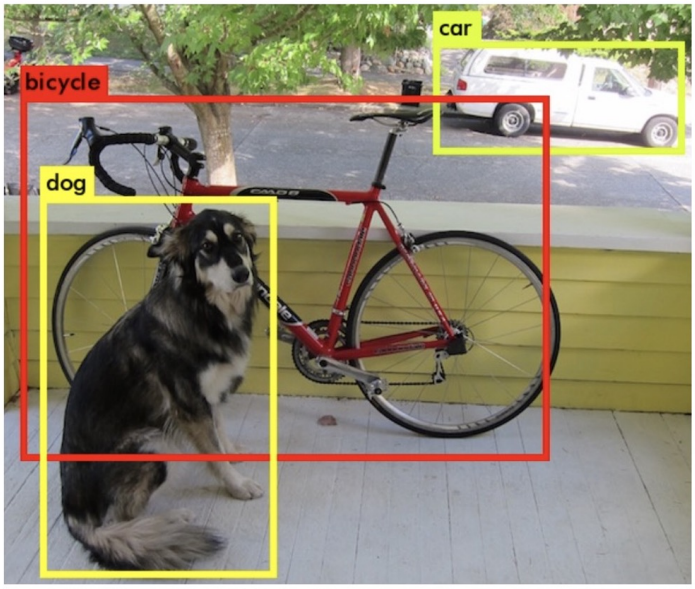

图7 利用计算机视觉模型进行目标检测，抽取图片特征

而文字信息则更多是通过自然语言处理的方法提取关键词、主题、分类等信息，一旦这些特征被提取出来，就跟处理属性、标签类特征的方法一样，通过 Multi-hot 编码，Embedding 等方式输入推荐系统进行训练。

5. 场景信息（上下文信息）

最后一大类是场景信息，或称为上下文信息（Context），它是描述推荐行为产生的场景的信息。最常用的上下文信息是“时间”和通过 GPS、IP 地址获得的“地点”信息。根据推荐场景的不同，上下文信息的范围极广，除了我们上面提到的时间和地点，还包括“当前所处推荐页面”“季节”“月份”“是否节假日”“天气”“空气质量”“社会大事件”等等。

场景特征描述的是用户所处的客观的推荐环境，广义上来讲，任何影响用户决定的因素都可以当作是场景特征的一部分。但在实际的推荐系统应用中，由于一些特殊场景特征的获取极其困难，我们更多还是利用时间、地点、推荐页面这些易获取的场景特征。

小结

这节课我们一起进入推荐系统中一个非常重要的模块，特征工程模块的学习。推荐系统中可用的特征非常多，但它们基本上可被划分到“用户行为”“用户关系”“属性标签”“内容数据”“场景信息”这五个类别，而且挑选特征的方法也遵循着“保留有用信息，摒弃冗余信息”的原则。

就像本节开头说的一样，特征工程是准备食材的过程，准备食材的好坏直接影响到能不能做出好菜。同时，要准备的食材也和我们要做什么菜紧密相连。所以针对不同的推荐系统，我们也要针对它们的业务特点，因地制宜地挑选合适的特征，抓住业务场景中的关键信息。这才是特征工程中不变的准则，以及我们应该在工作中不断积累的业务经验。

从工程的角度来说，除了特征的挑选，特征工程还包括大量的数据预处理、特征转换、特征筛选等工作，下节课我们就一起学习一下特征处理的主要方法，提升一下我们“处理食材”的技巧！

课后思考

如果你是一名音乐 APP 的用户，你觉得在选歌的时候，有哪些信息是影响你做决定的关键信息？那如果再站在音乐 APP 工程师的角度，你觉得有哪些关键信息是可以被用来提取特征的，哪些是很难被工程化的？

欢迎在留言区畅所欲言，留下你的思考和疑惑。如果今天的内容你都学会了，那不妨也把这节课转发出去。今天的内容就到这里了，我们下节课见！

提建议

上一篇 国庆策划 | 深度学习推荐系统基础，你掌握了多少？

下一篇 05 | 特征处理：如何利用Spark解决特征处理问题？

精选留言 (12)

写留言

朱月俊
2020-10-10

选择音乐，哪些因素是我关注的？
我经常会听五类歌曲：
1.听网络流行歌曲（听大家听的）；
2.听一些我喜欢的风格的歌曲（励志类，空灵类，感伤类）；
3.听一些我喜欢的歌手唱的歌，比如汪峰等；...

展开

作者回复: 非常好和全面，赞一个。

10

张弛 Conor
2020-10-12

我在选歌的时候，信息重要性从高到低依次是：
1.听歌的目的。比如是为了放松，冥想，学习还是运动。目的决定了歌曲是安静还是激昂，舒缓还是节奏感强烈。
2.歌曲或歌单是否受欢迎。定下基调后，我一般会选择收藏或播放量较多的歌曲。这样一般不容易采坑。...
展开

作者回复: 非常好的答案，比我之前看到的答案还细致一些。能够完全从用户角度考虑，然后反映到工程实践上。

5

金鹏
2020-10-09

音乐产品更加依赖场景性和心情，在工作、学习、跑步、睡眠、开车、高兴、优伤等等，希望听到的音乐是不同的。所以市面上的音乐目前主要以歌单的形式来推，可以更好的让用户快速找到符合自己当下场景的音乐，感觉更是个强搜索型的产品，音乐的推荐策略更像是一种补充。
...
展开

作者回复: 非常好了。如果是歌单的形式的话，确实歌单更像一个搜索词，更接近一个搜索问题。

2

马龙流
2020-10-12

像多模态或者是通过其它预训练方法得到的向量，直接加到推荐排序模型作为特征的话，感觉都没有效果。不知道你这边有没有碰到类似问题呢。我理解是预训练学习的目标和排序学习目标并不一致，不知道大佬怎么看这个问题
展开

作者回复: 确实存在多模态特征效果不强的问题。我觉得还是目前多模态的技术本质上还处于比较初期的阶段。

比如用一些CV的技术去处理视频图像，识别出一些物品，比如视频里有汽车，有甜品之类。但你要说这些物品对于推荐效果到底有没有影响，我觉得还是过于微弱了。远不及知名演员一个要素的影响大。

所以问题本质上还是出在你对特征的理解和业务场景本身的理解上。

3 1

夜雨声烦
2020-10-09

影响因素：当时的心情，时间，天气，歌名，歌手，歌曲类型，播放量，是否top，好友中有人听过；
工程师角度：
当时的心情(无法获取)、时间(24小时划分成十个时间段表示)、天气(晴、阴、雨)、歌名(Embedding)、歌手(Embedding)、歌曲类型(onehot)、播放量、是否top(onehot)、好...
展开

作者回复: 非常好，唯一确实的可能是自己的行为历史，就是听过哪些歌，以及这些历史跟当前要推荐歌曲的联系。

1

shenhuaze
2020-10-09

王老师，在线预测的时候，模型所需的特征是直接从数据库读取，还是在线实时组装？我在想如果只是用户或者物品自身的特征的话，可以从数据库读取，但如果是用户和物品的交叉特征的话，是不是必须实时组装？

作者回复: 非常好的点。一般来说如果组合特征可以在线处理，最好能够在线处理，因为组合特征有组合爆炸问题，为了节约宝贵的存储资源，一般不直接存储。

但是对于有些不得不存储的组合特征，比如 用户x物品的曝光、点击记录，如果线上模型需要的话，还是要存储到数据库中，因为这些特征你没办法在线组合。

2 1

pedro
2020-10-09

回答课后问题，按照电影的套路，关联的信息大概有:歌是谁唱的？(比如我的idol)，歌的风格是什么？(比如我超喜欢r&b)，歌的时长(太长的一律跳过)，至少点击过或者单曲循环过？至少听过类似的或者听过该歌手的？
歌的内容即歌词，旋律，副歌？等等
难以被工程化的是歌词内容，毕竟大家都不是专业音乐人，flow和verse这种东西没法去...
展开

作者回复: 歌词通过nlp，旋律通过一些模式识别也许可以提取出一些风格相似性之类的特征。但正如课程中说的，这些内容类信息都需要进一步处理后才可被推荐系统利用。

1

科科科科科科名儿
2020-10-27

老师您在提到了业界最近的做法是将标签类数据和与描述的主体，并结合graph Embedding，再输入推荐模型，现在业界有比较成熟的开源代码么？
展开

作者回复: 业界比较知名的工作是阿里的EGES paper可以参考https://github.com/wzhe06/Reco-papers/blob/master/Embedding/%5BAlibaba%20Embedding%5D%20Billion-scale%20Commodity%20Embedding%20for%20E-commerce%20Recommendation%20in%20Alibaba%20%28Alibaba%202018%29.pdf

相关开源的代码我这没有资源，如果有知道的同学可以回复一下。

2

书豪
2020-10-23

音乐类推荐要考虑，歌手信息，歌曲信息，用户信息。哈手主要是歌手的基本信息，分类标签。歌曲分类，用户特征，用户喜好！
展开

张弛 Conor
2020-10-12

老师，想请问一下为什么在用户行为数据中，将评论数据作为隐性反馈行为呢？因为我的理解，显性反馈行为就是用户对物品的直接评价(评分，赞等)，但是评论也算是用户对物

品的直接评价，所以我很好奇为什么评论会是隐性反馈呢？（我个人的解释是，在某些场景下，用户的评论并不一定是对物品的评价，比如对于新闻来说，评论可能是对于内容本身的讨论，而不是用户是否喜欢该新闻，但是对于电商类网站，对于物品的评论则可以...

展开

作者回复: 评论应该一般是以文本的形式存在的，所以其实很难直接判断出用户的好恶，需要进一步通过nlp等手段处理。这和评分是不一样的，评分能可以直接通过分数的高低判断用户的喜好程度。

波

2020-10-10

影响因素
1.歌曲风格
2.歌曲发布时间
3.歌手标签(流行，摇滚，轻音乐)
4.听歌时间(工作日，节假日，白天，晚上)...

展开

作者回复: 赞

李@君

2020-10-09

我在选歌的时候主要是听前几十秒，如果觉得好听就会继续，如果不好听，就跳过。这个也是很难被量化的吧。

展开

作者回复: 其实是一种正负样本的定义问题。在训练中的作用还是非常大的。

1