# 03 | 数据处理：我们应该如何获取和处理数据？

2023-04-14 黄鸿波 来自北京

《手把手带你搭建推荐系统》




你好，我是黄鸿波。

在前面的课程中，我们对推荐系统在企业中的作用有了比较深刻的认识，也借助一个具体的 Netflix 案例对推荐系统的整体架构和数据流有了充分的了解。从本章开始，我们就要真正地进入到推荐系统的开发当中，从头来搭建一个企业级的推荐系统了。

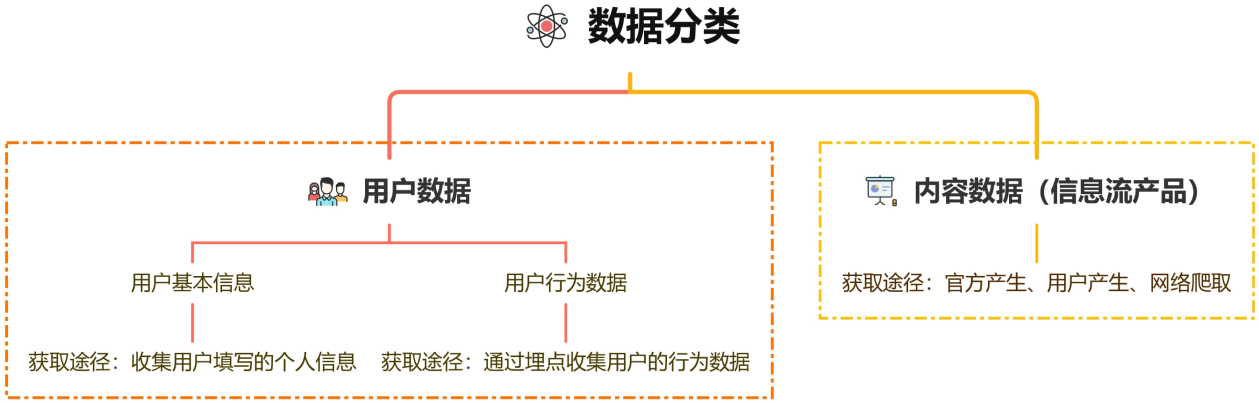
这一章的核心内容是数据，我们的话题都是围绕着数据来展开的。一个好的推荐系统，我们首先要知道怎样获取和处理数据，然后将数据存入适当的数据库中。紧接着，我们还要将原始数据处理成我们推荐系统所需要的画像，作为后续推荐算法需要的特征数据。

这节课我们就从数据的来源开始，讲解应该从哪里获得数据，要怎么处理它们。

## 数据获取方式

想要处理数据，第一步是要先获取数据。我们要获取的数据有哪些呢？根据上一讲的内容我们知道，推荐系统主要由离线层、在线层和近似在线层三个部分构成，我们需要两种类型的数据：用户数据和内容数据。








一般来讲，用户数据又可以分为两部分，一个是用户本身的属性数据，比如说用户的基本信息，经常看的内容等，这些我们可以统称为用户基本画像。另外一种是用户的行为数据。也就是一个用户从进入到我们的 App 开始到离开为止都留下了哪些行为，这些数据有助于我们分析用户的喜好，这些数据和用户基本画像组合起来，就成为了我们的用户画像。

获取用户数据的渠道一般可以分成两种。第一个渠道就是直接获取用户在注册时填写的个人信息，使用个人信息能获取到很多用户的基础数据，一般来讲，这些基础数据就可以刻画一个人或一类人。第二个渠道是在我们的产品中埋点，然后收集用户的行为数据，借此，我们可以知道用户经常看哪一类的文章，后续将其处理成用户画像，方便在算法中使用。

除了用户数据，在推荐系统中，还有一类内容数据也尤为重要，因为只有有了内容数据才会有用户可以阅读的内容。

在推荐系统中，根据产品的不同，内容数据的获取方式也会有所不同。作为信息流产品，内容可以由官方产生，也可以由用户产生。

所谓的官方产生，就是运营 App 的公司自己来写稿子，创作相关的视频放到内容池中，供用户阅读和观看。这类数据一般是由专业人员写出来的，内容的质量相对比较高。另外一种内容数据由用户产生（用户通过投稿或者在 App 中自由发帖产生），质量往往参差不齐。所以在推荐系统进行推荐计算时，一般也会通过算法把质量相对较低的数据给筛掉，留下质量好的数据进行推荐。



除了官方和用户产生的内容，我们还可以通过爬虫爬取内容。比如说在一些新闻资讯类的 App 中，运营者没有办法写出很即时的稿子（或者没有相关的记者），为了及时地发布最新资讯，就可以使用爬虫爬取互联网上的新闻，再通过筛选和过滤，形成推荐的内容。

## 数据具体形态

知道了怎么获取这些数据，那么这些数据的具体形态是怎样的呢？它们由哪些信息组成？

**对于用户数据来说，用户属性数据和用户行为数据的形式有所不同。**用户属性数据指的就是用户最基本的信息，具体可能包含用户的性别、年龄、星座、常住地等，但是根据 App 的不同，我们能获取到的内容也不同，如果再受到登录方式的限制，我们能获取到的内容会更加有限。

用户行为数据往往依赖用户的访问路径，它可以是一个用户经常阅读的文章题材、篇幅长短、内容形式（视频还是看文字）等等。借助用户行为数据，我们可以了解用户的偏好，让推荐的内容更加精准。

**对于内容数据来说，我们需要根据不同的内容类型来刻画不同的数据，最终数据将以内容画像的形式来展现。**例如，对于文章类的内容，我们可能会着重关注这篇文章的字数、用户阅读完成率、这篇文章中的 topN 关键词是什么。而对于视频类的内容，我们更需要关注视频的总时长、用户平均观看时长、观看完成百分比、视频分辨率、视频分类信息等。

但是，无论是视频还是文章类的内容，在制作画像时，我们都可以需要几个共有的属性，即点赞数、阅读 / 观看数、评论数、转发数、收藏数等，因为这些数据是无论哪种类型的信息流产品都有的，也是非常重要的。我们在设计整个产品以及推荐算法的时候，务必要把它们考虑进去。

**获得了用户数据和内容数据之后，我们还需要把它们处理成我们需要的格式。**我比较常用的方式是把数据以 Key-Value 的形式存储在 MongoDB 数据库中，当需要使用的时候，可以直接以 JSON 的形式取出。有时，我们还会直接将内容的列表和用户的列表以 Zset 的形式存在 Redis 数据库中。

## 数据处理方法



好了，假设我们已经获取到了足够的数据，我们该怎么把数据处理成我们想要的形态呢？

一般来讲，根据算法的不同，算法所需要的特征也不同，这会导致我们在处理原始数据时所用的方法也不同。在信息流的推荐系统中，我们主要将处理数据的方法分成下面三个大类。

1. 使用 NLP 提取特征和画像数据。
2. 使用机器学习处理特征和画像数据。
3. 使用统计学提取特征和画像数据。

下面我们来详细说说这几类处理方法和它们所对应的应用场景。

## 使用 NLP 提取特征和画像数据

使用 NLP 来提取特征和画像数据是最常用的方法之一。一般来讲，在信息流推荐中我们离不开对内容的描述和分类，因此，处理数据的方法也不尽相同。

比如说，在纯文字或者图文类的信息流推荐系统中，我们可以提取文章的关键词，取出关键词中的 topN 作为整篇内容的关键词特征。具体的执行上我们可以用 TFIDF、TextRank 等常用的关键词提取方法（也可以几种方式同时使用），然后取交集作为全文的关键词。

有的时候，我们为了更好地刻画人与人、内容与内容之间的关系，还会在推荐系统中提取各种画像信息，然后再使用知识图谱等方式刻画出各种信息之间的关联，方便我们更好地利用其中的相关性信息，让推荐的内容更加合理，提高整体用户的点击率和黏度。

还有的时候，我们会使用命名实体识别和关系抽取等方法，抽取原始数据中非结构类的数据，然后再将需要的数据提取出来，以便后面做成相关的特征和画像。

除此之外，有些系统中还会存在无法分类的老数据。针对于这类数据，我们可能不能很好地给它们打标签，这个时候我们也可以用聚类的方式得到一个粗分类，然后再使用文本分类模型进行精细分类，以此作为内容的标签等等。



总之，NLP 在信息流推荐系统中发挥着非常重要的作用，我们会在后面的课程中详细展开。

### 使用机器学习和统计学处理特征数据

机器学习对于特征处理来说也起着不可替代的作用，甚至可以说，几乎所有的特征处理工作都离不开机器学习。在机器学习常见的方法包括距离计算、特征向量化、聚类操作等。

有些时候，我们需要找到特征与特征之间的相互关系。这时我们就会先将我们需要的原始数据进行向量化处理，然后再利用距离计算函数计算出它们的相对关系。还有的时候，我们会使用机器学习做一些拟合计算，取得某些特征。

我们还可以使用一些统计学的方法来处理特征。比如说，我们可以通过统计字数来确定一篇文章是长文还是短文，也可以通过统计学的方式知道每篇文章或者每个视频的完成情况，这样更有利于我们后面的推荐。

上面这几种方法都是推荐系统中常用的数据处理方式，实际上可以使用的方法会更多，这节课先给你带来一个整体的认识，在后面的课程中，我们会边用边讲。

## 总结

这节课我们就讲完了，我们来回顾一下这节课的重要知识点。

1. 推荐系统中的数据一般分为用户数据和内容数据，其中用户数据包括了用户的行为数据和用户的属性数据。内容数据一般指的是用户要消费的内容，比如文章、视频、帖子等。
2. 我们的数据来源有很多，其中比较常见的是官方产生的信息流文章和用户所产生的内容，当然，我们也可以使用爬虫来爬取数据。

3. 有了数据，我们就可以使用机器学习、NLP、统计学等方式，把它们处理成我们想要的特征或者画像，供后续推荐算法使用了。

## 课后题

学完这节课的内容，给你留两道思考题。



1. 你想要做一个什么样的产品，需要收集哪些数据来组成用户画像？
2. 请你根据你设想出的产品，设计一套简单的用户画像和内容画像。

欢迎你在留言区留下你的观点，我们一起交流讨论，下节课见。



## 精选留言 (4)

**欢少の不忘初心**

2023-04-30 来自江苏

电商系统，博客系统

电商需要用户的浏览记录，购买记录，性别，商品的相似性信息
博客需要用户的浏览记录，点赞，评论，收藏，关键字，技术栈，喜好

作者回复: 是的，理解的很对，针对不同的系统要学会变通。

1

**阿难**

2023-05-31 来自上海

这节内容是我比较关注的，我浏览完全篇后，对这节课的理解和定位是数据获取-数据“预处理”的一个导引。
数据处理在一个成熟商用的推荐系统按照我的理解应该是有至少：“事前-事中-事后”三部分的。这节内容更多的是在讲事前这个阶段。
那么事中：也就是训练推荐引擎，的阶段数据是如何在不同的组键和模块只用吊起，传输，存

储的呢？这些组键分别支持什么样形式/格式的数据？处理完会存储在哪里？为什么要这么存储？

事后：也就是模型离线训练完成如何做预测？预测的数据形式是什么样子？是key value对吗？训练好的模型参数如何部署？为了线上大规模实时/近实时预测？什么样的数据格式更优？接口传输的数据又是如何定义，存储，流转？



希望这些唯独都能被老师和课程负责人关注到，这些对于一个完整的学习框架搭建才是有价值的。谢谢

**peter**

2023-04-16 来自北京

请教老师几个问题：

Q1：内容的质量怎么确定？

Q2：开发一个新闻APP，用爬虫爬取内容后用在自己的APP上，会有法律或版权纠纷吗？

Q3：推荐系统可以利用chatGPT吗？

作者回复: 答：同学你好。首先，对于一个推荐系统来讲，内容的质量有几种不同的判别方式，比如人工判断，分类，以及通过优质的内容产出者进行辅助判断等。所谓的人工判断，就是由专业的审核人员对内容进行审核，从而判断质量；分类的话实际上就是把优质的内容和普通的内容提前打上标签，然后再造出一批数据集后，训练一个文本分类的模型；还有第三种方式就是很多信息流推荐系统，都会找一些优质的写手或者官方的运营人员进行写作，这部分的内容我们一般也会认为是优质的内容。

关于爬虫爬取是否有法律纠纷主要是看你的爬虫程序是否遵循了网站的robots协议，一般来讲，如果遵循其协议，不会产生纠纷。但是如果你把这些内容进行商用的话，要看对方有没有版权要求，一般来讲是需要征求原作者的同意后才可以。

推荐系统在某些情况下可以和ChatGPT进行结合，但是这样需要在ChatGPT上写很多prompt才可以完成。

**徐曙辉**

2023-04-14 来自湖南

1. 短视频内容，需要用户昵称/性别/唯一标识/ip/地址，用户点赞，收藏，评论，分享，视频时长，完播率，平均播放时长，视频分类，关键词，标签，视频地区。

另外我想强调一下视频封面，封面是用户对视频的第一印象，所以可以生成同一个视频的不同封面图推荐给不同用户，字体/颜色/内容关键字，根据用户点击和观看进行优化，具体还

没实践过。
另外根据同一类用户关注的大V进行推荐，按用户之间粉丝交集关系进行推荐，A跟B都关注了C，A关注D，B没有，是不是可以给B推荐D。
2. 没有头绪，不知道从哪里下手，希望老师给答案

作者回复: 同学你好，正如你所说，一般来讲，短视频内容的画像相比于信息流文章来讲，更注重的是视频的播放完成度、视频时长、平均播放时长等，所以我们在构建内容画像的时候，应该着重去注意这几个点的特征处理，要去想，这些点应该用连续的特征还是离线的特征比较好，另外就是这类特征需要怎么去表达，才能更好的把握住重点。
封面的确是在视频推荐中最重要的一个点，对于封面特征的提取一般会用到图像特征提取的一部分，然后再根据图像的特征来进行封面的提取。另一方面，有一个简单的办法，就是把比较热门的内容的封面和普通内容的封面做一个特征或者图片的分类，也许也能找到一些比较好的特征分类点，这里面其实就不限于文字的颜色、大小和排版等。
关于你说的大V推荐的方法，其实就是协同过滤中的一部分，在后面的章节中你会看到这一部分的内容。