【内容推荐】超越标签的内容推荐系统

我曾在不同公司里都听到过，他们的产品经理或者大佬问过这样的问题：我们的推荐系统标签够不够?

相信你也遇到过类似的问题。这其实是一个很大的误区：基于内容的推荐系统，标签只是很小一部分。

而且就算是标签，衡量质量的方式也不是数目够不够；所以，今天我要讲的内容，就是说一说脱离标签定式思维的内容推荐。

为什么要做好内容推荐

所谓的基于内容推荐，通俗一点来讲，就是一个包装成推荐系统的信息检索系统。这听上去有点残酷，但通常一个复杂的推荐系统很可能是从基于内容推荐成长起来的。

可以说，基于内容的推荐系统是一个推荐系统的孩童时代，所以，我们不能让自己的推荐系统输在起跑线上，得富养才行。那么，首先我就来讲一讲如何养成一个基于内容的推荐系统。

为什么基于内容的推荐系统这么重要呢？因为内容数据非常易得，哪怕是在一个产品刚刚上线，用心找的话总能找到一些可以使用的内容，不需要有用户行为数据就能够做出推荐系统的第一版。

内容数据尤其是文本，只要深入挖掘，就可以挖掘出一些很有用的信息供推荐系统使用。

另外，著名的流媒体音乐网站 Pandora，其音乐推荐系统背后的“音乐基因工程”，实质上就是人工为音乐标注了各种维度的属性，这样，即便使用基于内容推荐的方式，也做出了很好的推荐效果。

听上去，上面这段话特别像是在安慰还处在冷启动阶段的你，事实上呢，其实并不全是，内容推荐的方式还有它的必要性。推荐系统总是需要接入新的物品，这些新的物品在一开始没有任何展示机会，显然就没有用户反馈，这时候只有内容能帮它。

基于内容的推荐能把这些新物品找机会推荐出去，从而获得一些展示机会，积累用户反馈、走上巅峰、占据热门排行榜。

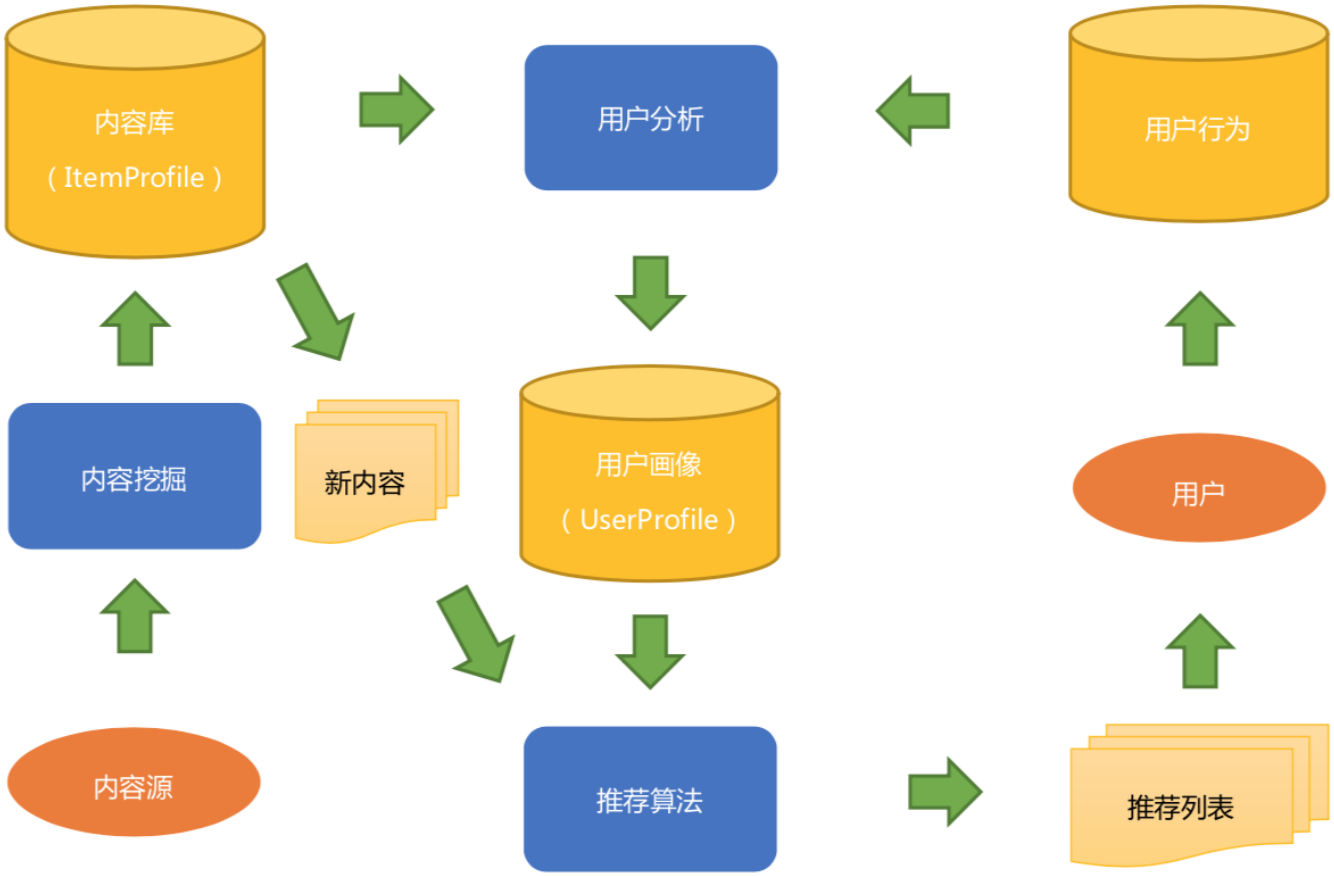
要把基于内容的推荐做好，需要做好“抓、洗、挖、算”四门功课。它们分别是对应了下面的内容。

1. 抓：大厂们从来不公开说的一件事是，他们一直在持续抓数据丰富自己的内容，所以做好一个基于内容的推荐，抓取数据补充内容源，增加分析的维度，两者必不可少。
2. 洗：抓来的数据，相当于捡别人掉地上的东西吃，我们也得注意卫生，洗洗更健康，数据也一样，冗余的内容、垃圾内容、政治色情等敏感内容等等都需要被洗出去。
3. 挖：不管是抓来的数据，还是自家的数据，如果不深入挖掘，那就和捧着金饭碗去要饭一样，浪费了大好资源。可以说，很多推荐系统提升效果并不是用了更复杂的推荐算法，而是对内容的挖掘做得更加深入。
4. 算：匹配用户的兴趣和物品的属性，计算出更合理的相关性，这是推荐系统本身的使命，不仅仅是基于内容的推荐才要做的。

那么，这四门功课到底如何分布在基于内容的推荐系统中呢?

下面我和你一起来看看，基于内容推荐的框架

在文稿中，我放了一张图，一个典型基于内容推荐的框架图是下面这样的：



简要介绍一下这张图的流程和基本元素。

内容这一端：内容源经过内容分析，得到结构化的内容库和内容模型，也就是物品画像。用户这一端：用户看过推荐列表后，会产生用户行为数据，结合物品画像，经过用户分析得到用户画像。

以后对于那些没有给用户推荐过的新内容，经过相同的内容分析过程后就可以经过推荐算法匹配，计算得到新的推荐列表给用户。如此周而复始，永不停息。

内容源

在互联网中，抓数据是一件可做不可说的事情，哪怕是市值几千亿的大厂，也有专门的小分队抓数据，补充推荐系统每天的内

容消耗。因为，只有当内容有多样性了，一个推荐系统才有存在的合法性，所以大厂职工们抓数据也是为了保住自己的饭碗。

爬虫技术本身非常复杂、非常有学问，比推荐算法难多了，这里就不展开讲了。

不论是抓来的数据还是自家用户产生的数据，都离不开清洗数据。由于各家都在相互借鉴来借鉴去，所以抓到重复的内容也是很有可能的，去重与识别垃圾内容、色情内容、政治敏感内容等都是必修课。

关于这个环节的边角算法，我们在后面的文章中会专门花一些篇幅来讲。

内容分析和用户分析

基于内容的推荐，最重要的不是推荐算法，而是内容挖掘和分析。内容挖掘越深入，哪怕早期推荐算法仅仅是非常硬的规则，也能取得不俗的效果。举个例子，如果推荐物品是短视频，我们分几种情况看：

1. 如果短视频本身没有任何结构化信息，如果不挖掘内容，那么除了强推或者随机小流量，没有别的合理曝光逻辑了；
2. 如果对视频的文本描述，比如标题等能够有内容分类，比如是娱乐类，那么对于喜欢娱乐的用户来说就很合理；
3. 如果能够进一步分析文本的主题，那么对于类似主题感兴趣的用户就可能得到展示；
4. 如果还能识别出内容中主角是吴亦凡，那就更精准锁定一部分用户了；
5. 如果再对内容本身做到嵌入分析，那么潜藏的语义信息也全部抓住，更能表达内容了。

举这个例子是为了说明：随着内容分析的深入，能抓住的用户群体就越细致，推荐的转化率就越高，用户对产品的好感度也就增加了。上一篇中我列举了文本数据——这也是内容数据最常见形式的分析方法。

内容分析的产出有两个：

1. 结构化内容库；
2. 内容分析模型。

结构化的内容库，最重要的用途是结合用户反馈行为去学习用户画像，具体的方法在上一篇中已经介绍了。容易被忽略的是第二个用途，在内容分析过程中得到的模型，比如说：

1. 分类器模型；
2. 主题模型；
3. 实体识别模型；
4. 嵌入模型。

这些模型主要用在：当新的物品刚刚进入时，需要实时地被推荐出去，这时候对内容的实时分析，提取结构化内容，再于用户画像匹配。

内容推荐算法

对于基于内容的推荐系统，最简单的推荐算法当然是计算相似性即可，用户的画像内容就表示为稀疏的向量，同时内容端也有对应的稀疏向量，两者之间计算余弦相似度，根据相似度对推荐物品排序。

你别嫌糙，如果你内容分析做得深入的话，通常效果还不错，而且基于内容的推荐天然有一个优点：可解释性非常强。

如果再进一步，要更好地利用内容中的结构化信息，因为一个直观的认识是：不同字段的重要性不同。

比如说，一篇新闻，正文和标题中分析出一个人物名，评论中也分析出其他用户讨论提及的一些人物名，都可以用于推荐。直观上新闻的正文和标题中更重要。

那么，我们可以借鉴信息检索中的相关性计算方法来做推荐匹配计算：BM25F算法。常用的开源搜索引擎如Lucene中已经实

现了经典的BM25F算法，直接拿来使用即可。

前面提到的两种办法虽然可以做到快速实现、快速上线，但实际上都不属于机器学习方法，因为没有考虑推荐的目标，而我们在之前的专栏中就专门强调了目标思维，那么，按照机器学习思路该怎么做呢?

一种最典型的场景：提高某种行为的转化率，如点击、收藏、转发等。那么标准的做法是：收集这类行为的日志数据，转换成训练样本，训练预估模型。

每一条样本由两部分构成：一部分是特征，包含用户端的画像内容，物品端的结构化内容，可选的还有日志记录时一些上下文场景信息，如时间、地理位置、设备等等，另一部分就是用户行为，作为标注信息，包含“有反馈”和“无反馈”两类。

用这样的样本训练一个二分类器，常用模型是逻辑回归（Logistic Regression）和梯度提升树（GBDT）或者两者的结合。在推荐匹配时，预估用户行为发生的概率，按照概率排序。这样更合理更科学，而且这一条路可以一直迭代优化下去。

总结

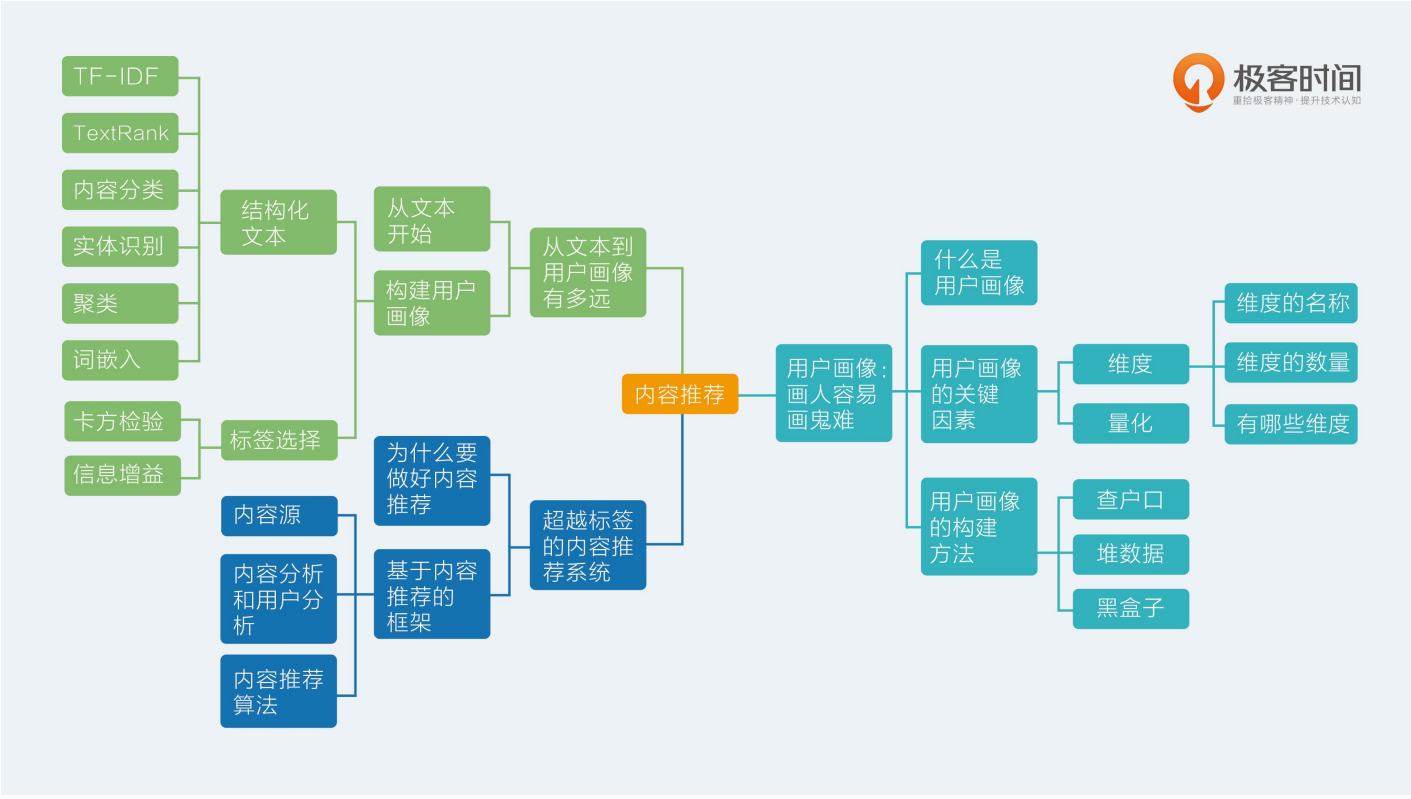
基于内容的推荐一般是推荐系统的起步阶段，而且会持续存在，它的重要性不可取代。因为：

1. 内容数据始终存在并且蕴含丰富的信息量，不好好利用就可惜了；
2. 产品冷启动阶段，没有用户行为，别无选择；
3. 新的物品要被推荐出去，首选内容推荐。

基于内容的整体框架也是很清晰的，其中对内容的分析最为重要，推荐算法这一款可以考虑先使用糙快猛的相似度计算，也可以采用机器学习思路训练预估模型，当然这必须得有大量的用户行为做保证。

好的，今天的内容就到这里，你可以在留言中谈一谈你对整个内容推荐链条各个环节的理解吗？欢迎和我一起讨论，感谢你的收听，我们下期再见。

本周知识要点



精选留言

叶晓锋

这一篇的含金量太高了，我要认真仔细的读，因为信息量太大，需要花点时间消化。

2018-03-16 14:29

作者回复

不要急，慢慢来。

2018-03-16 21:27

江枫

老师好，问个工程实现上的问题。用lda训练出来的k个主题概率分布作为ctr模型的其中k维特征，进行训练和预测，这是个很好的思路。但是训练一般是离线的，预测是在线的，需要一个kv存储特征，那么训练好的模型更新到线上服务器，势必需要确保离线训练特征和在线预测特征的一致性。这样问题来了，lda的抽取主题，是无监督的，没法保证两次抽取的主题是一个顺序的，导致训练和预测特征没法绝对一致，直到kv库和模型都更新完成。老师这边有好的解决方案吗？不知道我又没有表达清楚，哈哈。

2018-04-14 12:03

嘉文

我也有相同的困惑：

老师好，问个工程实现上的问题。用lda训练出来的k个主题概率分布作为ctr模型的其中k维特征，进行训练和预测，这是个很好的思路。但是训练一般是离线的，预测是在线的，需要一个kv存储特征，那么训练好的模型更新到线上服务器，势必需要确保离线训练特征和在线预测特征的一致性。这样问题来了，lda的抽取主题，是无监督的，没法保证两次抽取的主题是一个顺序的，导致训练和预测特征没法绝对一致，直到kv库和模型都更新完成。老师这边有好的解决方案吗？不知道我又没有表达清楚，哈哈。

2018-05-30 19:03

其中说到的抓数据为了就是丰富内容源避免产品单调（有法律风险），还是说抓了用来分析热度来有利于自己内容的推荐

2018-03-19 09:13

作者回复

前者。

2018-03-21 09:54

Drxan

点赞

2018-03-16 09:16

爱谁谁

推荐系统的表格那里，内容源和用户行为分析时间说的是：跟据消费的内容来矫正用户行为

吗

2018-11-14 10:14

gaolinjie

老师你好，请问下您所说的采用机器学习的方法训练预估模型和吴恩达机器学习中说的Content Based Recommendations是一样的吗？谢谢!

2018-09-23 17:35

明华

老师您好! 对于这句话

"每一条样本由两部分构成：一部分是特征，包含用户端的画像内容，物品端的结构化内容，可选的还有日志记录时一些上下文场景信息，如时间、地理位置、设备等等，另一部分就是用户行为，作为标注信息，包含“有反馈”和“无反馈”两类。"

想问:

当训练的时候是选取一个用户的所有行为训练呢，还是选择所有用户的所有行为进行训练呢。如果是所有用户，那逻辑回归训练出来的模型意义又是什么呢?

2018-07-20 17:35

作者回复

你想问的是为每个用户构建一个模型还是为所有用户构建一个模型吗？答案是不冲突，对那些非常活跃和深度的用户，他的数据足够多，有必要给他个人构建一个模型。而更多的用户数据是稀疏的，需要靠全局数据去泛化。

2018-07-23 13:32

我要飞上月球

“每一条样本由两部分构成：一部分是特征，包含用户端的画像内容，物品端的结构化内容，可选的还有日志记录时一些上下文场景信息，如时间、地理位置、设备等等，另一部分就是用户行为，作为标注信息，包含“有反馈”和“无反馈”两类。”我理解这里的用户端画像内容是包含用户行为统计的，这两个分开怎么理解；后面提到的有无反馈是否可以理解为用户有没有响应推荐的物品?

2018-06-08 13:29

会飞的书2008

讲解得很好，一口气反复读了三遍，谢谢大牛

2018-05-10 18:28

潘多拉魔盒

你好，有个问题想咨询下，我这是要做电影，电视剧等推荐，用户正向反馈，电影有标签，那么，如何去给用户做标签权重划分，

2018-04-08 08:14

尹士

你好，购买了你的作品，非常好！是我见过的最好的推荐系统整体剖析，不知道作者能否建群，交流沟通?

2018-03-31 22:42

作者回复

请下载“知识星球”(以前叫小密圈)，搜resyschina加群。

2018-04-01 23:30

Vito

非常感谢老师的分享！我们目前在做视频资源的推荐，也是从内容推荐起步，用户数在千万级，媒资数据在90万，这样做余弦相似度，形成的矩阵太大，计算效率太低，老师有什么好的建议吗?

2018-03-30 08:46

travi

有个点没看懂：文章最后你提到2分类器，我理解输入是<item向量，user向量>，输出是点击/不点击。既然是2个类别，后面提到按概率排序，这个概率是怎么由模型得到的?

2018-03-28 10:35

EAsY

能否详细介绍下 物品画像怎么作用到用户画像 比如阅读物品的次数或时间 怎么影响用户画像的更新 感觉这个没做好很影响推荐效果

2018-03-20 21:21

握力王

老师您提到短视频推荐，现在确实非常火，像抖音快手都在做。那么视频方面的内容分析有没有做的比较好的方法呢，例如提取视频特征，提取音频特征方面？谢谢老师

2018-03-19 01:24

作者回复

有。

2018-03-21 09:51

yxj

这两期要反复看看好好琢磨下了

2018-03-17 16:23

bowen

“爬虫技术...难多了”, 能否给个深入的阅读方向?

2018-03-17 00:01

作者回复

去用scrapy抓个网站就知道了。

2018-03-23 02:07

叶晓锋

文中提到了时间，地理位置和设备这些特征，尽管只提了一句话，事实上在实际业务中特别是垂直领域非常的重要，甚至需要把时空这两类特征还要再细化。

2018-03-16 14:40

作者回复

可以分享一下你的经验。

2018-03-16 21:27

林彦

谢谢邢无刀老师的分享。

本文提到“结构化的内容库，最重要的用途是结合用户反馈行为去学习用户画像”。回顾了上一篇“从文本到用户画像有多远”的用户反馈行为部分的内容。发现当时这一块介绍了卡方检验和信息增益2种特征选择方法，但对于把词归入哪一类这个数据处理如何和用户行为关联自己理解得还不深。现在想问下这里的类是用户有操作(点击，阅读，购买，评分等)和无操作的类别，还是用户的各种与用户行为无关的标签，比如(1)单个用户自身的各种属性，以及分类之前传递给单个用户的物品结构化信息或(2)整个系统的的用户类别标签，可能是基于之前学习到的TopK主题词？如果特征选择方法里用到的类不是用户行为，是通过用户行为来判定某个词属于哪一类的二分类问题吗？比如包含某个词的物品被某个类的用户操作了，可以把这个词归入相应的类中。

前面几节还未来得及阅读，可能问的问题有错误。请多包涵。

很期待之后有基于GBDT或GBDT+Logistic Regression来预测用户行为概率的实践分享和开源工具/模块推荐。

2018-03-16 14:00

作者回复

简单理解就是检验特征和目标之间的关联性。预测目标是行为，那么类别就是行为。

2018-03-16 21:29