11 | 基于规则的召回：如何使用规则找到用户的兴趣点?

2023-05-10 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。

当我们学习完推荐系统的定义以及推荐系统中的数据处理后，紧接着就要进入到推荐系统算法的实质性部分，我会从召回开始带你进入推荐系统算法的大门。

首先我们来讲一讲，召回到底是什么，以及它在推荐系统中的意义。

召回对于推荐系统的意义

在推荐系统中用户能接收到的任何推荐内容都来自推荐池。我们可以把推荐词理解为能够给用户推荐的所有数据集合，也就是总的内容。

一般来讲，所有用户目前能看到的内容都可以在推荐池中找到，但是推荐池中的内容并不一定就是所有的数据源内容，这点是非常需要注意的。在推荐系统中我们可以将所有的内容大体分

成三个类别：可以给用户推荐的内容、已经过期的内容以及还没有对外发布的内容。

可以给用户推荐的内容。如果我们把内容加上状态这个标签的话，那么这些内容的状态就是对用户可见。也就是说，用户可以通过推荐系统或者其他的形式能够找到这篇内容，这类数据会出现在推荐池和推荐列表中。

已经过期的内容。一般指的是这篇内容之前可以推荐，但是目前不会出现在用户的推荐列表中。比如说几个月前的公告，或者去年的新闻等，这些数据对于用户来说实际上已经是无用的数据了，这些就不会再出现在推荐池或推荐列表中。

还没有对外发布的数据内容。指的是内容已经写好了，但是还没有对外公布，这类数据一般也不会出现在推荐池或推荐列表中。

实际上，推荐系统的过程就是一个对数据集精简的过程。我们从最初的原始数据集中，通过一步步精简、筛选和过滤，逐渐地找到用户感兴趣的话题，而召回集就是这个筛选过程中的第一步。召回集是推荐系统的根，召回集的质量往往决定着整个推荐系统的推荐质量。

基于规则的召回

从大的类别来看，我们可以把召回集分成基于规则的召回算法、经典的召回算法和基于机器学习的召回算法。

类别	概述
基于规则的召回算法	通过一些规则筛选就能够得到召回集结果的算法，比如时间、热度、关键词等
经典的召回算法	常见的召回算法，比如协同过滤算法、LFM算法、NMF算法等
基于机器学习的召回算法	和机器学习或者深度学习相挂钩的一些算法，如YoutubeDNN，embedding等

在这一章，我们主要聚焦于基于规则的召回算法。

我认为，之所以能称之为规则，说明这种召回算法能够有比较强的可解释性，能够通过一系列的规则和特点来进行归纳和总结，从而得到想要的召回结果。时间规则、热度规则、搜索规则等都属于常见的规则。相应的，常见的基于规则的召回算法会有下面三种。

基于时间的召回算法。该算法主要根据内容产生的时间来做一定排序，从而进行召回。我认为从严格意义上来讲，基于时间的召回算法其实也不能完全算是一种算法，它更像是一种对内容的展示顺序。将最新的内容排在最前面让用户最先看到，这个实际上就是最简单的基于时间的召回算法。

基于热度的召回算法。在这种算法中，一般每个内容都会被设置一个初始的热度值，每篇文章都会有对应的阅读、评论、点赞、收藏、转发等指标。我们可以将这些指标的每一项按照重要性不同来赋予不同的权重，然后根据权重和对应的数量计算出热度值。在计算热度值时还会涉及热度的更新和衰减，也就是说热度既会随着点击的增加而上升，也会随着时间的流逝而下降，最后我们再算出一个总的热度值，作为当前的热度。

基于关键词搜索的召回算法。和前面两种算法不同，该算法不仅用到了规则相关的技术，同时还会融入一些简单的数据统计和 NLP 相关的内容，算是一种简单的结合体。在目前大部分的产品中，都会用到搜索技术，无论是搜索商品还是搜索文章标题，实际上都是在对其中的关键词进行检索。更确切来说，是将搜索的关键词和文章中的关键词进行匹配，然后再按照某一种排序规则将其展示出来。

基于规则的召回算法	概述
基于时间的召回算法	该算法主要根据内容产生的时间来做一定排序，从而进行召回。基于时间的召回算法其实也不能完全算是一种算法，更像是一种对内容的展示顺序
基于热度的召回算法	在这种算法中，一般每个内容都会被设置一个初始的热度值，每篇文章都会有对应的阅读、评论、点赞、收藏、转发等指标，可以将这些指标的每一项按照重要性不同来赋予不同的权重，然后根据权重和对应的数量计算出热度值
基于关键词搜索的召回算法	该算法不仅用到了规则相关的技术，同时还会融入一些简单的数据统计和NLP相关的内容，算是一种简单的结合体

这节课，我们就先来讲一讲其中的基于时间的召回算法。

基于时间的召回算法

基于时间的召回看似简单，但是在实际的操作过程中有着很多细节需要注意。

基于时间的召回算法对时间非常敏感，这种召回算法主要是考虑了用户的兴趣随时间的推移而发生变化的情况，也能够反映用户的实时兴趣特征。

该算法并不是完全按照时间的远近来进行排序，一般还会加入时间衰减因子调整内容权重（用户历史行为记录中距离当前时间越远，内容权重越低，反之亦然）。通过计算用户与物品之间的权重得分，选择得分较高的内容作为推荐结果。

基于时间的召回算法逻辑比较简单，它有以下三个主要优点。

1. 时间敏感性。随着时间的变化用户的兴趣爱好也会变化，基于时间的召回算法可以较好地反映出用户当前的兴趣点。
2. 个性化。基于时间的召回算法可以根据用户的历史行为提供个性化的推荐结果，满足用户个性化需求。
3. 实时性。在一定程度上，基于时间的召回算法能够快速响应用户需求，提供实时的推荐结果。

同时，基于时间的召回算法也有下面两个主要的缺点。

1. 基于时间的召回算法需要处理大量的历史数据，同时需要对时间信息进行分析，因此需要大量的计算资源和数据存储资源。
2. 基于时间的召回算法很难考虑到用户行为与时间之间的非线性关系，精度偏低，需要与其他算法相结合，提高性能。

基于时间的召回算法	
优点	可以较好地反映出用户当前的兴趣点
	可以满足用户个性化需求
	能够快速响应用户需求，提供实时的推荐结果
缺点	需要大量的计算资源和数据存储资源
	很难考虑到用户行为与时间之间的非线性关系，精度偏低

在实际的工程项目中，我们用到基于时间的召回一般主要是在前期（也就是冷启动时），这个时候用户的行为数据相对比较少无法获取到更多的特征，那么时间就是我们最好的特征。

对于推荐来说，基于时间的召回只是召回算法中一个非常小的分支，因此它的推荐效果十分有限。我们可以将基于时间的召回和基于热度、关键词、以及深度学习的召回算法相结合。

在基于关键词的召回算法中，可以根据历史数据中某个词语的出现频率等属性，赋予不同词语不同权重；在基于时间的召回算法中，可以根据时间因素给各个内容赋予不同的权重。通过这种方式将时间和关键词两种因素结合起来进行推荐，提高推荐准确度。

我们也可以将基于热度的召回算法和基于时间的召回算法组合使用，得到多维度的召回结果。例如基于时间的推荐可以提供新鲜有趣的内容，而基于热度的推荐可以提供最近最热门的内容。

在基于深度学习的召回算法中，我们可以将时间因素加入到训练数据里以增强模型效果。例如在电商网站中，可以根据用户的浏览和购买记录来收集数据，将每个商品的流行度和产生时间作为特征，然后通过深度学习算法来训练模型，进而预测用户可能感兴趣的商品。

基于时间的召回算法与其他算法相结合	
时间与关键词结合	提高推荐准确度
时间与热度结合	新鲜内容与热门内容结合
时间与深度学习结合	增强模型效果

总结

我们来总结一下这节课的内容，学完这节课，希望你能够记住以下要点。

1. 召回集就是一批用户可能会感兴趣的数据集，召回算法就是用来找到用户兴趣内容的一类算法。
2. 召回算法有很多分类，常见的分类是基于规则的召回、经典召回和基于机器学习的召回。
3. 基于规则的召回一般包括基于时间的召回、基于热度召回和基于关键词的召回。
4. 基于时间召回是最简单的一种召回算法，主要是依照内容产生时间的顺序进行召回，一般需要配合其他召回方式一起使用。

课后题

学完这节课，给你留两道思考题。

1. 想一想，我们如何将规则的召回进行组合。
2. 预习一下什么是基于热度的召回，以及热度的衰减方式。

欢迎你在留言区与我交流讨论，我们下节课再见。

精选留言 (2)

董俊俊

2023-05-27 来自湖北

纯理论吗？没有示例代码？


Geek_ccc0fd

2023-05-10 来自广东

1. 规则组合的话，就是尝试各种单一规则的笛卡尔积吧，时间+热度，关键词+热度，关键词+时间
2. 基于热度的召回，热度其实就说用户对物品的关注程度吧，通过点击，点赞，收藏，转发等来计算加权和，可以加个时间限制，近7/30/90天的热度榜单，各种关键词的热度榜单，物品的热度可以乘一个时间衰减函数来做热度衰减

作者回复: 是的，这么理解是没问题的。