

机器学习算法系列（13）：推荐系统（1） 一简介

一、定义

1.1 从搜索引擎说起

人们在寻找信息时，常常需要借助搜索引擎，主动地提供准确的关键词来进行相应的搜索，但是当用户无法找到准确描述自己需求的关键字时，搜索引擎就无能为力了。

和搜索引擎一样，推荐系统也是一种帮助用户快速发现有用信息的工具。但是和搜索引擎不同的是，推荐系统不需要用户提供明确的需求，而是通过分析用户的历史行为给用户的兴趣建模，从而主动给用户推荐能够满足它们兴趣与需求的信息。

1.2 推荐系统的具体应用

近十年来，推荐引擎对互联网用户来说随处可见。以亚马逊为例，简单说说实际的应用。Amazon有个性化商品推荐列表和相关商品的推荐列表，以下为一种基于物品的推荐算法（item-based method），它会给我推荐那些和之前我喜欢的物品相似的物品，因为我之前查找过关于算法数据结构相关的书籍，所以就给我推荐了以下这些计算机领域的经典图书

您最近查看的商品和相关推荐

猜您喜欢



算法导论(原书第3版)
Thomas H. ...
★★★★☆ 350
平装
¥102.40 Prime



普通高等教育“十一五”国家
级规划教材•高等 ...
沈备军
★★★★☆ 4
平装
¥42.00 Prime



图灵程序设计丛书:算法(第
4版)
塞奇威克 (Rob ...
★★★★☆ 185
平装
¥78.20 Prime



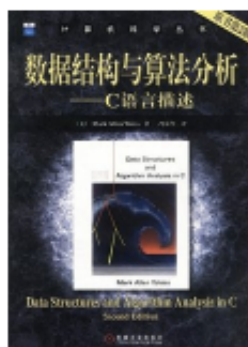
现代操作系统(原书第3版)
塔嫩鲍姆 (Tan ...
★★★★☆ 142
平装
¥59.20 Prime



C++ Primer(中文版)(第5
版)
斯坦利·李普曼 (...
★★★★☆ 650
平装
¥94.70 Prime

除了个性化推荐列表，亚马逊另一个重要的推荐应用就是相关推荐列表。一种是包含购买此商品的顾客也同时购买，另一种是包含浏览过这个商品的顾客购买的其他商品

购买此商品的顾客也同时购买



数据结构与算法分析:C语言描述(原书第2版)

维斯

★★★★☆ 220

平装

¥27.60 Prime



编程珠玑(第2版·修订版)

乔恩·本特利 (J ...)

★★★★☆ 36

平装

¥30.80 Prime



算法导论(原书第3版)

Thomas H. ...

★★★★☆ 350

平装

¥102.40 Prime

看过此商品后顾客买的其它商品？



深入理解计算机系统(原书第3版) 平装

兰德尔·E.布莱恩特(Randal E. Bryant)

★★★★☆ 23

¥107.10 Prime



算法(英文版)(第4版) 平装

塞奇威克 (Robert Sedgewick);韦恩 (Kevin Wayne)

★★★★☆ 10

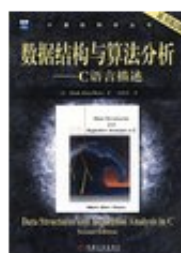
¥103.40 Prime

此外，还有一个应用就是打包销售，同时购买这些商品往往会给你一定的折扣。

经常一起购买的商品



+



+



总价: ¥208.20

全部加入购物车

- ✓ 本商品: 图灵程序设计丛书:算法(第4版) - 塞奇威克 (Robert Sedgewick) 平装 ¥78.20
- ✓ 数据结构与算法分析:C语言描述(原书第2版) - 维斯 平装 ¥27.60
- ✓ 算法导论(原书第3版) - Thomas H.Cormen 平装 ¥102.40

除了亚马逊等电子商务领域的推荐系统应用，此外还有像Netflix会向用户推荐电影，豆瓣电台给用户推荐喜欢的歌曲，Facebook给用户推荐个性化物品和好友，Flipboard给特定的用户推荐特定领域的阅读资讯，雅虎的个性化广告投放等等。

二、推荐系统评测

一个好的推荐系统不仅仅能够准确预测用户的行为，而且能够扩展用户的视野，帮助用户发现那些他们可能会感兴趣，但却不那么容易发现的东西。同时还要能帮助商家将那些被埋在长尾的好商品介绍给可能对他们感兴趣的用户。

2.1 推荐系统实验方法

首先，人们需要通过实验的办法来计算和获取一些指标，从而全面地测评推荐系统的好坏。主要的实验方法有离线实验、用户调查、在线AB测试。

2.1.1 离线实验

一般离线试验的步骤如下：

- 1) 通过日志系统获取用户行为数据，并按照一定格式生成一个标准的数据集；
- 2) 将数据集按照一定的规则分成训练集和测试集；
- 3) 在训练集上训练用户兴趣模型，在测试集上进行预测；
- 4) 通过事先定义的离线指标评测算法在测试集上的预测结果。离线试验不需要有对实际系统的控制权、不需要用户参与实验、速度快、可以测试大量算法。但是它无法计算商业上关心的指标（点击率、转化率）、离线试验的指标和商业指标存在差距。

2.1.2 用户调查

因为离线实验的指标与实际的商业指标存在差距，所以需要将算法直接上线测试，但在这之前必须进行用户调查，否则直接进行在线实验会有较高的风险，因为对算法会不会降低用户满意度谁都没有把握。

用户调查可以获得很多体现用户主观感受的指标，相对在线实验风险很低，出现错误后很容易弥补。但是招募测试用户代价较大，很难组织大规模的测试一款能过户，因此会使得测试结果的统计意义不足。

2.1.3 在线实验

AB测试时最常用的在线测评算法的实验方法。它通过一定的规则将用户随机分成机组，并对不同组的用户采用不同的算法，然后通过统计不同组用户的各种不同的评测指标比较不同算法，比

如可以统计不同组用户的点击率，通过点击率比较不同算法的性能。

AB测试可以公平获得不同算法实际在线时的性能指标，包括商业上关注的指标。但是周期较长，必须进行长期的实验才能得到可靠的结果，因此常常只会用它测试那些在离线实验和用户调查中表现很好的算法。

以上就是一个新的推荐算法最终上线所需要做的事。总结一下：

- 1) 首先，需要通过离线试验证明它在很多离线指标上优于现有算法；
- 2) 然后，需要通过用户调查确定它的用户满意度不低于现有算法。
- 3) 最后，通过在线的AB测试确定它在我们关心的指标上优于现有的算法。

2.2 离线评测指标

令 $R(u)$ 是根据用户在训练集上的行为给用户作出的推荐列表，而 $T(u)$ 是用户在测试集上的行为列表。

- 准确率 (Precision) :

$$Precision = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|}$$

- 召回率 (Recall) :

$$Recall = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|}$$

- 覆盖率(Coverage)，描述了一个推荐系统对物品长尾的发掘能力，

- 最简单的定义为推荐系统能够推荐出来的物品占总物品集合的比例:

$$Coverage = \frac{|\cup_{u \in U} R(u)|}{|I|}$$

- 信息熵：

$$H = - \sum_{i=1}^n p(i) \log p(i)$$

- 基尼系数:

$$Gini = \sum_{i=1}^n p_i (1 - p_i)$$

- 马太效应：即强者更强，弱者更弱。若一个系统会增大热门物品和非热门物品的流行度

差异，让热门的物品更加热门，不认的物品更加不热门，那么这个系统就有马太效应。推荐系统的初衷是希望消除马太效应，使得各种物品都能被展示给它们感兴趣的某一类人群。但是现在主流的推荐算法都具有马太效应。可以使用基尼系数来评测推荐系统是否具有马太效应。如果G1是从初始用户行为中计算出的物品流行度的基尼系数，G2是从推荐列表中计算出的物品流行度的基尼系数，那么如果G2>G1，就说明推荐系统具有马太效应。

- 多样性：如果推荐列表比较多样，覆盖了用户绝大多数的兴趣点，那么就会增加用户找到感兴趣物品的概率。因此给用户的推荐列表也需要满足用户广泛的兴趣，即多样性。
 - 多样性描述了推荐列表物品两两之间的不相似性。假设 $s(i, j) \in [0, 1]$ 定义了物品 i 和 j 之间的相似度，那么用户 u 的推荐列表 $R(u)$ 的多样性定义如下：

$$Diversity(R(u)) = 1 - \frac{\sum_{i, j \in R(u), i \neq j} s(i, j)}{\frac{1}{2}|R(u)|(|R(u)| - 1)}$$

- 而推荐系统整体的多样性可以定义为所有哟京沪推荐列表多样性的平均值：

$$Diversity = \frac{1}{|U|} \sum_{u \in U} Diversity(R(u))$$

- 新颖性：给用户推荐那些它们以前没有听说过的物品。最简单方法是利用推荐结果的平均流行度，因为越不热门的物品越可能让用户觉得新颖。但是这个指标比较粗略，因为不同用户不知道的东西是不同的。需要用户调查来准确统计新颖性。
- 惊喜度：惊喜度和新颖性两者之间是有区别的，如果推荐结果和用户的历史兴趣不相似，但却让用户觉得满意，那么就可以说推荐结果的惊喜度很高，而推荐的新颖性仅仅取决于用户是否听说过这个推荐结果。