< 推薦系統實驗設計 > July 27, 2019

希望提供更好的使用者導向推薦系統 提升女性購買的選擇多樣性

推薦算法本質上就是一個機器學習問題，

需要構建推薦算法模型 選擇認為合適的效果好算法模型

將算法模型部署到線上推薦業務中

利用算法模型來預測用戶對標的物的偏好

通過用戶的真實反饋（是否點擊 是否購誤 是否收藏等）來評估算法效果

問題點：

1. 如何收集用戶日誌，來證明男女的購買行為不同

2. 經過個性化分類器＋推薦系統後的成效透過“離線測試”跟“在線測試”

3. 測量標準是用戶滿意系統推薦的商品 給出十個商品的平均分數

三種推薦系統實驗方法

1、離線實驗（offline experiment）

實施步驟：

（1）將從日誌系統收集來的用戶行為數據，生成為一個標準的數據集；

（2）按照一定的規則，將數據集分成訓練集和測試集兩個部分；

（3）在訓練集上訓練用戶興趣模型，在測試集上進行測試；

（4）利用定義好的離線指標評測算法，在測試集上進行結果預測。

優點：所有實驗在數據集（從系統日誌中提取而來）上完成，對實際系統和用戶參與依賴度低，方便快捷；

缺點：對商業上關注的指標獲取能力弱；

2、用戶調查（user study）

用戶調查，是一種通過分析被調查用戶（真實的用戶），在被測推薦系統上完成任務時的行為和回答問題的情況，來了解測試系統性能的一種實驗方法。它旨在為上線測試提供準備工作，以防範上線測試所潛在的降低用戶滿意度的問題。

優點：在離線測試解決不了的「用戶主觀感受相關指標」的獲取上，有著優越的性能；風險易控。

缺點：實驗成本高，難組織大規模測試；雙盲實驗設計困難，影響測評結果。

3、在線實驗（online experiment）

這裡的在線實驗方法，偏指AB測試方法。

實施步驟：

（1）通過一定規則將用戶隨機分組；

（2）對不同組的用戶採用不同算法；

（3）統計不同組用戶的不同評測指標，以比較不同的算法。

優點：公平獲得不同算法實際在線的性能指標，包括商業上關注的指標。

缺點：試驗周期較長；設計AB測試系統的工程量大，且流量切分設計一般必不可少。

參考資料



**十個推薦系統測評指標**

1、**用戶滿意度**

「用戶滿意度」測評指標，適用的實驗方法為用戶調查法與在線測試法。

用戶調查法設計要點：將用戶對被測推薦系統的滿意程度分為多個層次，並照顧到用戶各方面感受，為用戶給出準確答案創造環境。

在線測試法設計要點：根據不同情境，區別「購買率」「用戶反饋按鈕統計」「用點擊率」「停留時間」「轉化率」等統計指標的設計，以度量用戶滿意度。

原文網址：<https://kknews.cc/tech/89g2kkq.html>

實驗測試：

在線對比測試

那就是AB test 分流量做真實的皮評測

這需要一個支持流量證交切分的AB test框架

AB test 在樣本充分的前提下

基本上可以定性新的推薦系統是否比老的推薦系統更加優秀

用戶訪談

系統有多好 一類是深度類 廣度類

1. 评分准确度。通常就是均方根误差 RMSE，或者其他误差指标，反映预测评分效果的好坏

2. 排序。检测推荐系统排序能力非常重要, 因为把用户偏爱的物品放在前面是推荐系统的天职。

由于推荐系统输出结果是非常个人化的，除了用户本人，其他人都很难替他回答哪个好哪个不好，所以通常评价推荐系统排序

推薦系統需要測試 只是它不同於傳統的功能測試

傳統軟件的功能測試 功能的響應式有預期的

點擊一個加關注按鈕 應該有什麼響應

推薦系統的測試方法有四種

業務規則掃描

離線模型測試

再現對比測試

用戶訪談

推荐系统中，主要有三种评测推荐效果的实验方法：

1）离线实验。 往往是从日志系统中取得用户的行为数据，然后将数据集分成训练数据和测试数据，比如80%的训练数据和20%的测试数据（还可以交叉验证），然后在训练数据集上训练用户的兴趣模型，在测试集上进行测试。 优点：只需要一个数据集即可，不需要实际的推荐系统（实际的也不可能直接拿来测试），离线计算，不需要人为干预，能方便快捷的测试大量不同的算法。缺点是无法获得很多实际推荐系统的指标，比如点击率，比如转化率（谁让没有人为干预呢。。）

2）用户调查。 离线实验往往测的最多的就是准确率，但是准确率不等于满意度，所以在算法上线之前，需要用户调查一下，测试一下用户满意度。

3）AB测试，通过一定的规则把用户随机分成几组，并对不同组的用户采用不同的推荐算法，这样的话能够比较公平的获得不同算法在实际在线时的一些性能指标。但是缺点是周期比较长，需要长期的实验才能得到可靠的结果。

如何判断一个推荐系统好不好，主要的测量指标如下：

1）用户满意度。这是最最关键的指标，推荐系统推荐物品干嘛，就是希望推荐出来的物品能让用户满意。可以有两种方法，一是用户问卷调查，二是在线评测满意度，比如豆瓣的推荐物品旁边都有满意和不满意的按钮，亚马逊这种可以计算推荐的物品有没有被用户购买等等，一般用点击率，用户停留时间，转化率等指标来度量。

2）预测准确度。如果是类似电影评分机制，则一般计算均方根误差(误差平方和取均值)和平均绝对误差（误差绝对值和取平均）。如果是topN推荐的话，则主要计算召回率和准确率。准确率就是指我推荐的n个物品中有多少个是对的，其所占的比重。 召回率则是指正确结果中有多少比率的物品出现在了推荐结果中。两者的不同就是前者已推荐结果个数当除数，后者已正确结果个数当除数。

3）覆盖率。 就是指推荐出来的结果能不能很好的覆盖所有的商品，是不是所有的商品都有被推荐的机会。最简单的方法就是计算所有被推荐的商品占物品总数的比重，当然这个比较粗糙，更精确一点的可以信息熵和基尼系数来度量。

4）多样性。推荐结果中要体现多样性，比如我看电影，我既喜欢看格斗类的电影，同时又喜欢爱装文艺，那么给我的推荐列表中就应该这两个类型的电影都有，而且得根据我爱好比例来推荐，比如我平时80%是看格斗类的，20%是看文艺类的，那么推荐结果中最好也是这个比例。可以根据物品间的相似度来计算，一个推荐列表中如果所有物品间的相似度都比较高，那么往往说明都是同一类物品，缺乏多样性。

5）新颖性。不能说系统推荐的物品其实我都知道，那这样推荐系统就完全失去了存在的意义，一般都希望推荐一些用户不知道的商品或者没看过没买过的商品。方法一是取出已经看到过买过的商品，但这还不够，一般会计算推荐商品的平均流行度，因为通常越不热门的物品越会让用户觉得新颖。比如我爱周星驰，那么推荐《临岐》就很有新颖性，因为大家都不知道这是周星驰出演的

6）惊喜度。 这个和新颖度还是有区别的，惊喜度是讲我直觉想不出来为什么会给我推荐这物品，比如电影，但是我看了之后觉得很符合我的胃口，这就是惊喜度。像上面一个例子，只要我知道是周星驰演的，那可能就没什么惊喜度，因为我知道是因为演员才给我推荐的这部电影。 注：新颖性和惊喜度暂时没有什么可以度量的标准

7）信任度。如果用户信任推荐系统，那么往往会增加与推荐系统的互动，从而获得更好的个性化推荐。增加信任的方法往往是提供推荐解释，即为什么推荐这个商品，做到有理有据。也可以通过类似facebook间的好友关系来增加信任度，一般相比于陌生人的推荐，总会选择好友给的推荐。

8）实时性。新闻等一些物品具有很强的实时性，一般得在具有有效性的时候进行推荐，必须考虑推荐系统处理物品冷启动的能力。

9）健壮性。要能防止被攻击，例如有些商家为了提高自己的排名，注册很多假的帐号，给与自己的商品高分这样类似的情况，要能防止。

10)商业目标。 一般推荐系统都是为了更好的盈利。。。。。当然这个比较难测。。。

---------------------

作者：wangyuquan

来源：CSDN

原文：https://blog.csdn.net/wangyuquanliuli/article/details/36684817

版权声明：本文为博主原创文章，转载请附上博文链接！

「好的推薦系統不僅僅能夠準確預測用戶的行為，而且能夠擴展用戶的視野，幫助用戶 發現那些他們可能會感興趣，但卻不那麼容易發現的東西。同時，推薦系統還要能夠幫助商家將 那些被埋沒在長尾中的好商品介紹給可能會對它們感興趣的用戶。」  
  
原文網址：<https://kknews.cc/tech/89g2kkq.html>

**三種推薦系統實驗方法**

1、**離線實驗（offline experiment）**

實施步驟：

（1）將從日誌系統收集來的用戶行為數據，生成為一個標準的數據集；

（2）按照一定的規則，將數據集分成訓練集和測試集兩個部分；

（3）在訓練集上訓練用戶興趣模型，在測試集上進行測試；

（4）利用定義好的離線指標評測算法，在測試集上進行結果預測。

優點：所有實驗在數據集（從系統日誌中提取而來）上完成，對實際系統和用戶參與依賴度低，方便快捷；

缺點：對商業上關注的指標獲取能力弱；

2、**用戶調查（user study）**

用戶調查，是一種通過分析被調查用戶（真實的用戶），在被測推薦系統上完成任務時的行為和回答問題的情況，來了解測試系統性能的一種實驗方法。它旨在為上線測試提供準備工作，以防範上線測試所潛在的降低用戶滿意度的問題。

優點：在離線測試解決不了的「用戶主觀感受相關指標」的獲取上，有著優越的性能；風險易控。

缺點：實驗成本高，難組織大規模測試；雙盲實驗設計困難，影響測評結果。

3、**在線實驗（online experiment）**

這裡的在線實驗方法，偏指AB測試方法。

實施步驟：

（1）通過一定規則將用戶隨機分組；

（2）對不同組的用戶採用不同算法；

（3）統計不同組用戶的不同評測指標，以比較不同的算法。

優點：公平獲得不同算法實際在線的性能指標，包括商業上關注的指標。

缺點：試驗周期較長；設計AB測試系統的工程量大，且流量切分設計一般必不可少。

****

**十個推薦系統測評指標**

1、**用戶滿意度**

「用戶滿意度」測評指標，適用的實驗方法為用戶調查法與在線測試法。

用戶調查法設計要點：將用戶對被測推薦系統的滿意程度分為多個層次，並照顧到用戶各方面感受，為用戶給出準確答案創造環境。

在線測試法設計要點：根據不同情境，區別「購買率」「用戶反饋按鈕統計」「用點擊率」「停留時間」「轉化率」等統計指標的設計，以度量用戶滿意度。

原文網址：<https://kknews.cc/tech/89g2kkq.html>