好的推荐系统

首先介绍什么是推荐系统、推荐系统的主要任务、推荐系统和分类目录以及搜索引擎的区别等；然后按照不同领域分门别类地介绍目前业界常见的个性化推荐应用；最后介绍推荐系统的评测，通过介绍评测指标给出“好”的定义，从而最终解答“什么是好的推荐系统”这个问题。

什么是推荐系统

推荐系统的任务就是联系用户和信息，一方面帮助用户发现对自己有价值的信息，另一方面让信息能够展现在对它感兴趣的用户面前，从而实现信息消费者和信息生产者的双赢。

分类目录和搜索引擎

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。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他们兴趣和需求的信息。因此，从某种意义上说，推荐系统和搜索引擎对于用户来说是两个互补的工具。搜索引擎满足了用户有明确目的时的主动查找需求，而推荐系统能够在用户没有明确目的的时候帮助他们发现感兴趣的新内容。

发掘物品的长尾效应：推荐系统通过发掘用户的行为，找到用户的个性化需求，从而将长尾商品准确地推荐给需要它的用户，帮助用户发现那些他们感兴趣但很难发现的商品。

个性化推荐系统的应用

前端的展示页面、后端的日志系统以及推荐算法系统。

电子商务

电影和视频网站

个性化音乐平台

两个条件：第一是存在信息过载，第二是大部分时候没有特别明确的求。

社交网络

个性化阅读

基于位置的服务

个性化邮件

个性化广告

计算广告

- **上下文广告** 通过分析用户正在浏览的网页内容，投放和网页内容相关的广告。
- **搜索广告** 通过分析用户在当前会话中的搜索记录，判断用户的搜索目的，投放和用户目的相关的广告。

- **个性化展示广告** 经常在很多网站看到大量展示广告，它们是根据用户的兴趣，对不同用户投放不同的展示广告。

推荐系统评测

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，推荐系统还要能够帮助商家将那些被埋没在长尾中的好商品介绍给可能会对它们感兴趣的用户。

推荐系统实验方法

评测推荐效果的实验方法：离线实验（offline experiment）、用户调查（user study）和在线实验（online experiment）。

离线实验

- 通过日志系统获得用户行为数据，并按照一定格式生成一个标准的数据集；
- 将数据集按照一定的规则分成训练集和测试集；
- 在训练集上训练用户兴趣模型，在测试集上进行预测；
- 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。

优　　点	缺　　点
不需要有对实际系统的控制权	无法计算商业上关心的指标
不需要用户参与实验	离线实验的指标和商业指标存在差距
速度快，可以测试大量算法	

用户调查

离线时没有办法评测的与用户主观感受有关的指标都可以通过用户调查获得，尽量保证测试用户的分布和真实用户的分布相同。

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。但是测试用户代价较大，很难组织大规模的测试用户，因此会使测试结果的统计意义不足。此外，在很多时候设计双盲实验非常困难，而且用户在测试环境下的行为和真实环境下的行为可能有所不同，因而在测试环境下收集的测试指标可能在真实环境下无法重现。

在线实验

AB 测试是一种很常用的在线评测算法的实验方法。它通过一定的规则将用户随机分成几组，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。

优点是可以公平获得不同算法实际在线时的性能指标，缺点主要是周期比较长，必须进行长期的实验才能得到可靠的结果。因此一般不会用 AB 测试测试所有的算法，而只是用它测试那些在离线实验和用户调查中表现很好的算法。

- 通过离线实验证明它在很多离线指标上优于现有的算法；
- 通过用户调查确定它的用户满意度不低于现有的算法；
- 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。

评测指标

用户满意度

满意度是评测推荐系统的最重要指标。但是，用户满意度没有办法离线计算，只能通过用户调查或者在线实验获得。

预测准确度

预测准确度度量一个推荐系统或者推荐算法预测用户行为的能力，这个指标是最重要的推荐系统离线评测指标。

在计算该指标时需要有一个离线的数据集，该数据集包含用户的历史行为记录。然后将该数据集通过时间分成训练集和测试集。最后，通过在训练集上建立用户的行为和兴趣模型预测用户在测试集上的行为，并计算预测行为和测试集上实际行为的重合度作为预测准确度。

- 评分预测

预测准确度一般通过均方根误差（RMSE）和平均绝对误差（MAE）计算。

RMSE 加大了对预测不准的用户物品评分的惩罚（平方项的惩罚），因而对系统的评测更加苛刻。如果评分系统是基于整数建立的（即用户给的评分都是整数），那么对预测结果取整会降低 MAE 的误差。

- TopN 推荐

预测准确率一般通过准确率（precision）/召回率（recall）度量。

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。那么，推荐结果的召回率定义为：

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

推荐结果的准确率定义为：

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

- 关于评分预测和 TopN 推荐的讨论

覆盖率

覆盖率（coverage）描述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，一个内容提供商会关心的指标。

假设系统的用户集合为 U，推荐系统给每个用户推荐一个长度为 N 的物品列表 R(u)，覆盖率为：

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

通过研究物品在推荐列表中出现次数的分布描述推荐系统挖掘长尾的能力。如果这个分布比较平，那么说明推荐系统的覆盖率较高，而如果这个分布较陡峭，说明推荐系统的覆盖率较低。

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

$$G = \frac{1}{n-1} \sum_{i=1}^{n} (2i - n - 1) p(j_i)$$

评测推荐系统是否具有马太效应的简单办法就是使用基尼系数。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，就说明推荐算法具有马太效应（所谓强者更强，弱者更弱的效应）。

多样性

多样性描述了推荐列表中物品两两之间的不相似性，多样性和相似性是对应的。

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：

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

新颖性

最简单方法是利用推荐结果的平均流行度。

惊喜度

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。

信任度

度量推荐系统的信任度只能通过问卷调查的方式，询问用户是否信任推荐系统的推荐结果。

增加推荐系统的透明度（transparency），而增加推荐系统透明度的主要办法是提供推荐解释。考虑用户的社交网络信息，利用用户的好友信息给用户做推荐，并且用好友进行推荐解释。

实时性

需要实时地更新推荐列表来满足用户新的行为变化；能够将新加入系统的物品推荐给用户。

健壮性

模拟攻击：首先，给定一个数据集和一个算法，可以用这个算法给这个数据集中的用户生成推荐列表。然后，用常用的攻击方法向数据集中注入噪声数据，然后利用算法在注入噪声后的数据集上再次给用户生成推荐列表。最后，通过比较攻击前后推荐列表相似度评测算法的健壮性。如果攻击后的推荐列表相对于攻击前没有发生大的变化，就说明算法比较健壮。

商业目标

	离线实验	问卷调查	在线实验
用户满意度	×	✓	○
预测准确度	✓	✓	×
覆盖率	✓	✓	✓
多样性	○	✓	○
新颖性	○	✓	○
惊喜度	×	✓	×

评测维度

获得一个算法在什么情况下性能最好，可以为融合不同推荐算法取得最好的整体性能带来参考。

用户维度、物品维度和时间维度