第一章 好的推荐系统

1.1 什么是推荐系统

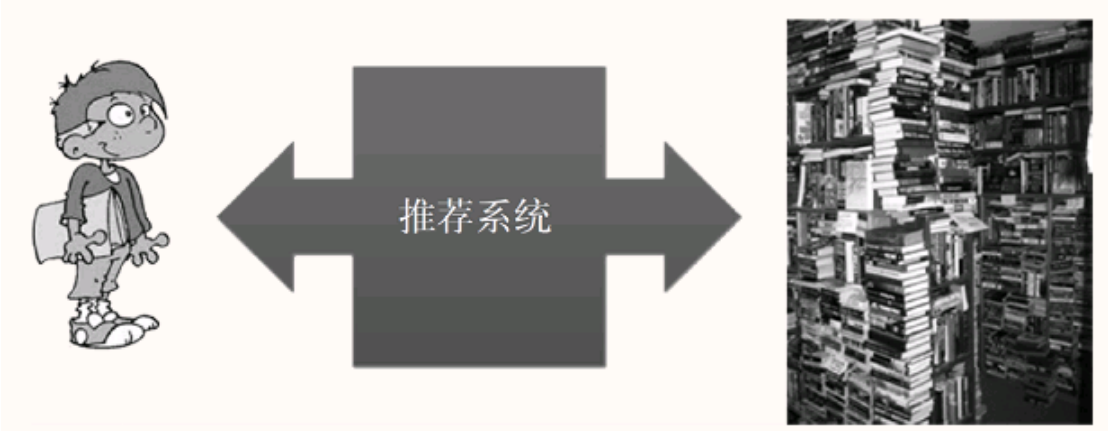

图 1.1 推荐系统的任务就是联系用户和物品，解决信息过载的问题

推荐系统：通过一定的方式自动将用户与物品联系起来的一种工具，它能够在信息过载的环境中帮助用户发现令他们感兴趣的信息，也可以将信息发送给对它们感兴趣的用户。

1.2 个性化推荐系统的应用

应用场景：电子商务、电影视频网站、个性化阅读、个性化音乐电台、个性化邮件、个性化广告、社交网络、基于位置的服务。

1.3 个性化推荐系统的评测

一个完整的推荐系统一般存在 3 个参与方：用户、物品提供者和推荐系统的网站。在评测一个推荐算法时，需要同时考虑三方的利益，一个好的推荐系统就是能够实现三方共赢的系统。

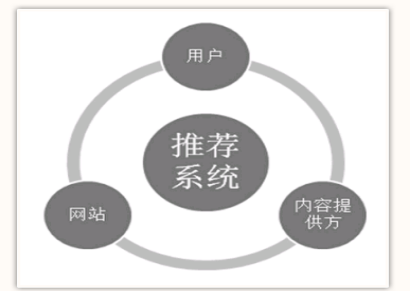

图 1.3.1 推荐系统的参与者

1.3.1 推荐系统实验方法

1. 离线实验（offline experiment）

离线实验的方法一般由如下几个步骤构成：

（1） 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；

（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。

从上面的步骤可以看到，推荐系统的离线实验都是在数据集上完成的，也就是说它不需要一个实际的系统来供它实验，而只要有一个从实际系统日志中提取的数据集即可。

表 1-2 离线实验的优缺点

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	

2. 用户调查（user study）

用户调查是推荐系统评测的一个重要工具，很多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。

用户调查的优缺点也很明显。它的优点是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。缺点是招募测试用户代价较大，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。

3. 在线实验（online experiment）

在完成离线实验和必要的用户调查后，可以将推荐系统上线做 AB 测试，将它和旧的算法进行比较。

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，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。对 AB 测试感兴趣的读者可以浏览一下网站 http://www.abtests.com/ 。

一个大型网站的架构分前端和后端，从前端展示给用户的界面到最后端的算法，中间往往经过了很多层，这些层往往由不同的团队控制，而且都有可能做 AB 测试。如果为不同的层分别设计 AB 测试系统，那么不同的 AB 测试之间往往会互相干扰。比如，当我们进行一个后台推荐算法的 AB 测试，同时网页团队在做推荐页面的界面 AB 测试，最终的结果就是你不知道测试结果是自己算法的改变，还是推荐界面的改变造成的。因此，切分流量是 AB 测试中的关键，不同的层以及控制这些层的团队需要从一个统一的地方获得自己 AB 测试的流量，而**不同层之间的流量应该是正交的。**

AB 测试的优点是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。AB 测试的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。其次，一个大型网站的 AB 测试系统的设计也是一项复杂的工程。

一般来说，一个新的推荐算法最终上线，需要完成上面所说的 3 个实验。

☐首先，需要通过离线实验证明它在很多离线指标上优于现有的算法。

☐然后，需要通过用户调查确定它的用户满意度不低于现有的算法。

☐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。

1.3.2 评测指标

在介绍各种指标之前，说一说各种参数表示的意义，以便读者更好的理解。

u: 一个用户；

U: 用户集合；

i: 一件商品；

I：商品集合；

||：取模操作，一般指一个集合包含元素的多少；

R(u): 暂先理解为用户 u 喜欢的物品集合；

p (i)：暂先理解为喜欢 i 物品的用户数/总用户数，也称物品流行度。

1.用户满意度

用户作为推荐系统的重要参与者，其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。

一般的情况下，我们可以用**点击率、用户停留时间和转化率**等指标度量用户的满意度。

2.预测准确度

这个指标是最重要的推荐系统离线评测指标，从推荐系统诞生的那一天起，几乎 99%与推荐相关的论文都在讨论这个指标。

由于离线的推荐算法有不同的研究方向，因此下面将针对不同的研究方向介绍它们的预测准确度指标。

☐评分预测

评分预测的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。对于测试集 T 中的一个用户 u 和物品 i，令 rui 是用户 u 对物品 i 的实际评分，而 rui(^)是推荐算法给出的预测评分，那么 RMSE 的定义为：

$$RMSE = \sqrt{\frac{\sum_{u,i \in T} (r_{ui} - \hat{r}_{ui})^2}{|T|}}$$

MAE 采用绝对值计算预测误差，它的定义为：

$$MAE = \frac{\sum_{u,i \in T} |r_{ui} - \hat{r}_{ui}|}{|T|}$$

关于 RMSE 和 MAE 这两个指标的优缺点，Netflix 认为 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。研究表明，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 预测**

网站在提供推荐服务时，一般是给用户一个个性化的推荐列表，这种推荐叫做 TopN 推荐。TopN 推荐的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：

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

推荐结果的准确率定义为：

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

评分预测一直是推荐系统研究的热点，绝大多数推荐系统的研究都是基于用户评分数据的评分预测。很多研究人员都将研究精力集中在优化评分预测的 RMSE 上。对此，亚马逊前科学家 Greg Linden 有不同的看法。2009 年，他在 Communications of the ACM 网站发表了一篇文章，指出**电影推荐的目的是找到用户最有可能感兴趣的电影，而不是预测用户看了电影后会给电影什么样的评分。因此，TopN 推荐更符合实际的应用需求**。也许有一部电影用户看了之后会给很高的分数，但用户看的可能性非常小。因此，预测用户是否会看一部电影，应该比预测用户看了电影后会给它什么评分更加重要。

3.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。覆盖率有不同的定义方法，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。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)。那么推荐系统的覆盖率可以通过下面的公式计算：

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

从上面的定义可以看到，**覆盖率是一个内容提供商会关心的指标**。

上面的定义过于粗略。在信息论和经济学中有两个著名的指标可以用来定义覆盖率。第一个是信息熵：

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

这里 p(i)是物品 i 的流行度除以所有物品流行度之和。

第二个指标是基尼系数（Gini Index）：

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

这里，ij 是按照物品流行度 p()从小到大排序的物品列表中第 j 个物品。

社会学领域有一个著名的马太效应，即所谓强者更强，弱者更弱的效应。那么，推荐系统是否有马太效应呢？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对它们感兴趣的某一类人群。但是，很多研究表明现在主流的推荐算法（比如协同过滤算法）是具有马太效应的。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。

4.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。因此，多样性和相似性是对应的。假设 s(i,j)∈[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：

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

而推荐系统的整体多样性可以定义为所有用户推荐列表多样性的平均值：

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

关于推荐系统多样性最好达到什么程度，可以通过一个简单的例子说明。假设用户喜欢动作片和动画片，且用户 80%的时间在看动作片，20%的时间在看动画片。那么，可以提供 4 种不同的推荐列表：

A 列表中有 10 部动作片，没有动画片；

B 列表中有 10 部动画片，没有动作片；

C 列表中有 8 部动作片和 2 部动画片；

D 列表有 5 部动作片和 5 部动画片。

在这个例子中，一般认为 C 列表是最好的，因为它具有一定的多样性，但又考虑到了用户的主要兴趣。A 满足了用户的主要兴趣，但缺少多样性，D 列表过于多样，没有考虑到用户的主要兴趣。B 列表即没有考虑用户的主要兴趣，也没有多样性，因此是最差的。

5.新颖性

评测新颖度的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。因此，如果推荐结果中物品的平均热门程度较低，那么推荐结果就可能有比较高的新颖性。但是，用推荐结果的平均流行度度量新颖性比较粗略，因为不同用户不知道的东西是不同的。因此，要准确地统计新颖性需要做用户调查。

通过牺牲精度来提高多样性和新颖性是很容易的，而困难的是如何在不牺牲精度的情况下提高多样性和新颖性。

6.惊喜度

惊喜度（serendipity）是最近这几年推荐系统领域最热门的话题。但什么是惊喜度，惊喜度与新颖性有什么区别是首先需要弄清楚的问题。

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。

7.信任度

让用户对推荐结果产生信任是非常重要的。同样的推荐结果，以让用户信任的方式推荐给用户就更能让用户产生购买欲，而以类似广告形式的方法推荐给用户就可能很难让用户产生购买的意愿。

提高推荐系统的信任度主要有两种方法：

首先需要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**。只有让用户了解推荐系统的运行机制，让用户认同推荐系统的运行机制，才会提高用户对推荐系统的信任度。

其次是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**。这是因为用户对他们的好友一般都比较信任，因此如果推荐的商品是好友购买过的，那么他们对推荐结果就会相对比较信任。

8.实时性

在很多网站中，因为物品（新闻、微博等）具有很强的时效性，所以需要在物品还具有时效性时就将它们推荐给用户。比如，给用户推荐昨天的新闻显然不如给用户推荐今天的新闻。因此，在这些网站中，推荐系统的实时性就显得至关重要。

推荐系统的实时性包括两个方面。

首先，推荐系统需要实时地更新推荐列表来满足用户新的行为变化。

其次，推荐系统需要能够将新加入系统的物品推荐给用户。这主要考验了推

荐系统处理物品冷启动的能力。关于如何将新加入系统的物品推荐给用户，本书将在后面的章节进行讨论，而对于新物品推荐能力，我们可以利用用户推荐列表中有多大比例的物品是当天新加的来评测。

9.健壮性

在实际系统中，提高系统的健壮性，除了选择健壮性高的算法，还有以下方法：

□设计推荐系统时尽量使用代价比较高的用户行为。比如，如果有用户购买行为和用户浏览行为，那么**主要应该使用用户购买行为，因为购买需要付费，所以攻击购买行为的代价远远大于攻击浏览行为**。

□在使用数据前，进行攻击检测，从而对数据进行清理。

10.商业目标

不同的网站具有不同的商业目标。比如电子商务网站的目标可能是销售额，基于展示广告盈利的网站其商业目标可能是广告展示总数，基于点击广告盈利的网站其商业目标可能是广告点击总数。因此，设计推荐系统时需要考虑最终的商业目标，而网站使用推荐系统的目的除了满足用户发现内容的需求，也需要利用推荐系统加快实现商业上的指标。

11.总结

表 1-3　获取各种评测指标的途径

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

对于可以离线优化的指标，我个人的看法是应该在给定覆盖率、多样性、新颖性等限制条件下，尽量优化预测准确度。用一个数学公式表达，**离线实验的优化目标是**：

最大化预测准确度

使得：

覆盖率 > A
多样性 > B
新颖性 > C

其中，A、B、C 的取值应该视不同的应用而定。

1.3.2 评测维度

上一节介绍了很多评测指标，但是在评测系统中还需要考虑评测维度，比如一个推荐算法，虽然整体性能不好，但可能在某种情况下性能比较好，而增加评测维度的目的就是知道一个算法在什么情况下性能最好。

一般来说，评测维度分为如下 3 种。

□**用户维度**　主要包括用户的人口统计学信息、活跃度以及是不是新用户等。

□**物品维度**　包括物品的属性信息、流行度、平均分以及是不是新加入的物品等。

□**时间维度**　包括季节，是工作日还是周末，是白天还是晚上等。

如果能够在推荐系统评测报告中包含不同维度下的系统评测指标，就能帮我们全面地了解推荐系统性能，找到一个看上去比较弱的算法的优势，发现一个看上去比较强的算法的缺点。