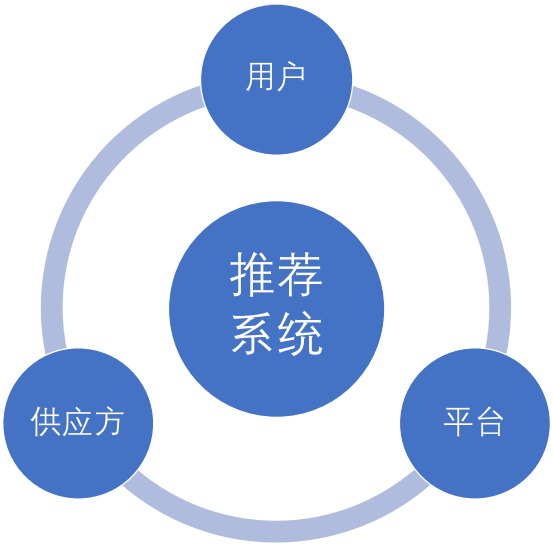
推荐系统的商业价值：什么是好的推荐系统？

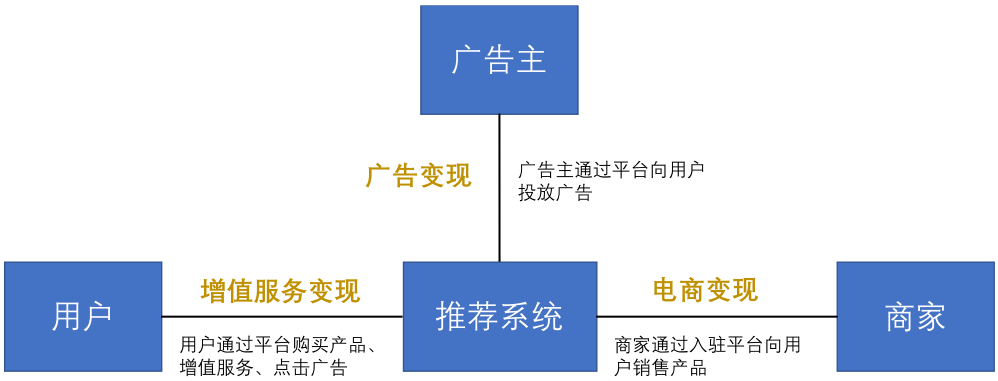
- 一个好的推荐系统应该是能让用户、供应方（商家/广告主）、平台**三方共赢**的系统
  - 准确的预测并不代表好的推荐[i]

    比如，给用户推荐一本他本来就准备购买的书，无论是否给予推荐他都会购买。推荐系统并未使用户购买更多的书，而仅仅是方便用户购买一本他本来就准备买的书。对于用户来说，他会觉得这个推荐结果很不新颖，不能令他惊喜。同时，对于出版社来说，这个推荐也没能增加这本书的潜在购买人数。

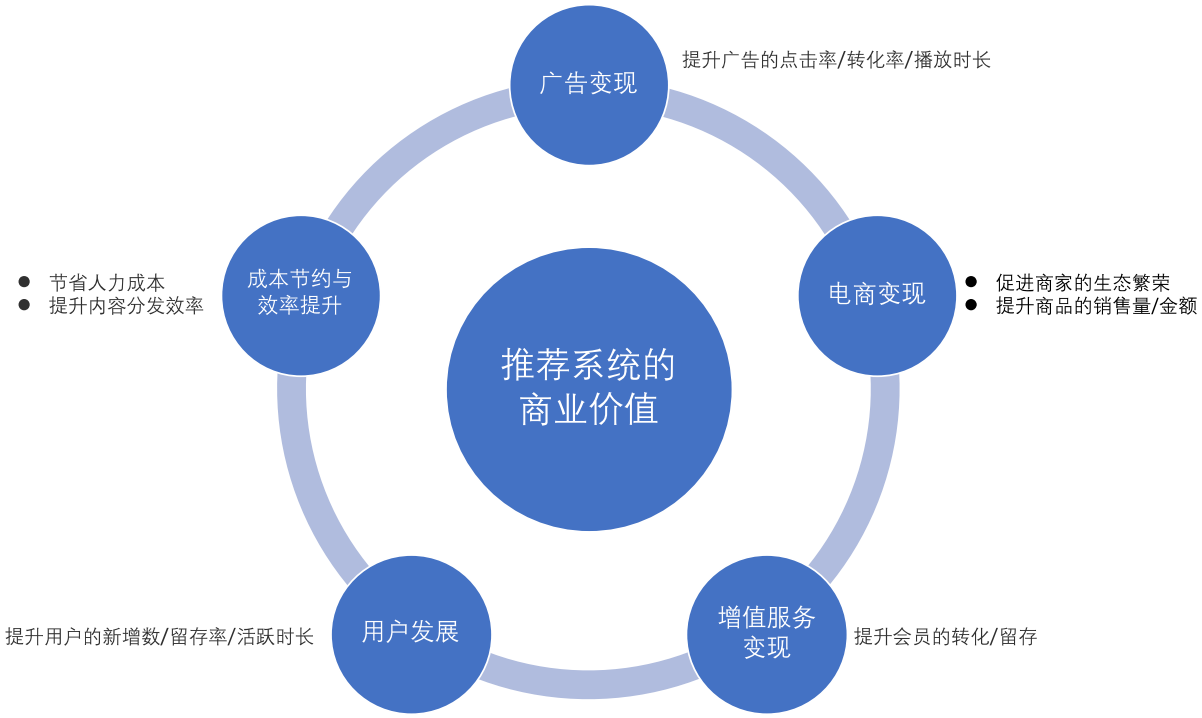
  - 好的推荐系统不仅仅能够准确预测用户的行为，而且能够扩展用户的视野，帮助用户发现那些他们可能会感兴趣，但却不那么容易发现的东西
  - 同时，推荐系统还要能够帮助商家将那些被埋没在长尾中的好商品介绍给可能会对它们感兴趣的用户



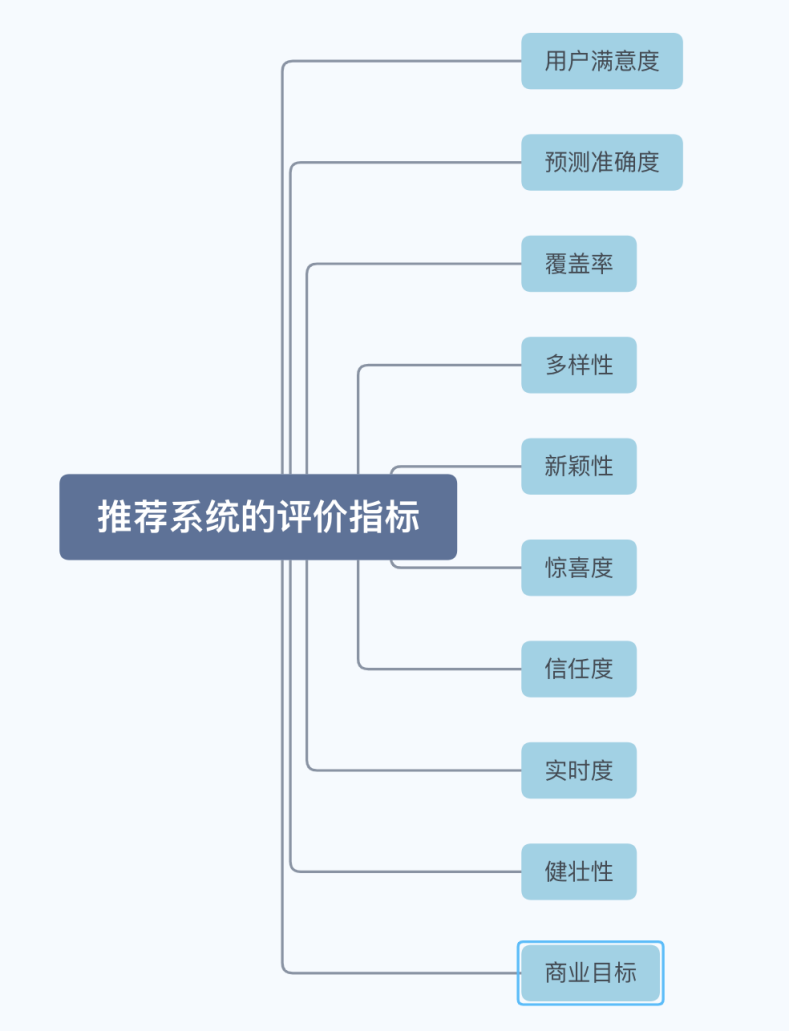
- 推荐系统相关的互联网产品盈利模式：广告、电商、增值服务



■ 推荐系统商业价值的体现维度[ii]



推荐系统的评价指标：如何评估推荐系统的效果？[iii]



评估指标	具体含义
用户满意度	用户满意度无法进行离线计算，只能通过用户调查或在线实验获得： ● 用户调查获得用户满意度主要通过调查问卷的形式，设计调查问卷不是简单地询问用户对推荐结果是否满意，而是要考虑到用户各方面的感受 ● 通过在系统中设计一些用户反馈页面也可用于收集用户满意度 ● 更一般的情况下，可以通过点击率、用户停留时间和转化率等指标来度量用户满意度

<table>
<tr><td rowspan="2">预测准确度</td><td>评分预测</td><td>对于离线测试集中的某个用户u和物品i，令r_{ui}为用户u对物品i的实际评分，$\hat{r}_{ui}$为推荐算法给出度预测评分
● 均方根误差（RMSE）
$$\mathrm{RMSE}=\frac{\sqrt{\sum_{u,i\in T}(r_{ui}-\hat{r}_{ui})^2}}{|T|}$$
● 平均绝对误差（MAE）
$$\mathrm{MAE}=\frac{\sum_{u,i\in T}|r_{ui}-\hat{r}_{ui}|}{|T|}$$</td></tr>
<tr><td>Top-N 推荐</td><td>● Precision/Recall/F1
假设$R(u)$表示根据用户在训练集中的行为给用户做出的推荐列表，$T(u)$表示用户在测试集上的行为列表
$$\mathrm{Precision}=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|}$$
$$\mathrm{Recall}=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|}$$
$$\mathrm{F1}=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$$
● ROC AUC
AUC 表示 ROC 曲线下的面积，衡量的是推荐系统将用户喜欢/不喜欢的商品区分出来的能力。如果所有的预测都是随机产生的，那么 AUC=0.5，因此 AUC 大于 0.5 的程度衡量了算法比随机推荐方法的精确程度
AUC 指标仅用一个数值评估推荐系统的整体表现，而且涵盖了所有不同长度推荐列表的表现。但 AUC 指标没有考虑具体排序位置的影响，导致在 ROC AUC 相同的情况下难以区分算法的好坏，其适用范围也受到一些限制
● Hit Rate (HR)
在 Top-N 推荐中，HR 是常用的衡量召回率的指标。分母是所有测试集的总数，分子是每个用户 Top-N 推荐列表中属于测试集的个数总和
$$\mathrm{HR}=\frac{\#\mathrm{hits}}{\#\mathrm{tests}}$$
例如，三个用户在测试集中的商品个数分别是 10、12、8，模型得到的 Top-10 推荐列表中在测试集中的个数分别是 6、5、4，</td></tr>
</table>

		那么 HR 的值是(6+5+4)/(10+12+8)=0.5 ● Average Reciprocal Hit Rate (ARHR) $$\text{ARHR} = \frac{1}{\#\text{tests}} \sum_{i=1}^{\#\text{hits}} \frac{1}{\text{rank}_i}$$ ARHR 是一种加权版本的 HR，度量的是一个物品被推荐的强度，其中rank_i是推荐列表中该物品排列位置，ARHR 可以用来衡量排序对该召回路径是否友好
	排序	当推荐项目的数量很大时，用户会更加重视推荐列表中排在前面的物品，这些物品中发生的错误比列表中排在后面的物品中的错误更严重，因此需要按排名列表对推荐效果进行加权评估 ● Half-life Utility (HL) ● Discounted Cumulative Gain (DCG) DCG 的主要思想是用户喜欢的商品被排在推荐列表前面比排在后面会更大程度上提升用户体验，定义为 $$\text{DCG}(b, L) = \sum_{i=1}^{b} r_i + \sum_{i=b+1}^{L} \frac{r_i}{\log_b i}$$ 其中r_i表示排在第i位的商品是否被用户喜欢，b是自由参数一般设为 2，L为推荐列表长度 ● Rank-biased Precision (RBP) RBP 和 DCG 指标的唯一不同点在于 RBP 把推荐列表中商品的浏览概率按等比数列递减，而 DCG 则是按照 log 调和级数形式 ● Normalized Discounted Cumulative Gain (NDCG) 由于用户之间 DCG 没有直接的可比性，需要进行归一化处理 ● Mean Reciprocal Rank (MRR) ● Mean Average Precision (MAP)
覆盖率		● 物品覆盖率 即推荐系统能够推荐出来的物品占总物品的比例，覆盖率越高表明模型能够针对更多的物品产生推荐，从而发掘长尾的能力就越强

$$\text{Coverage}_{\text{Item}} = \frac{|\cup_{u \in U} R_u|}{|I|}$$

其中U是所有用户的集合，I是所有物品的集合，R_u是给用户u推荐的所有物品集合。

从以上定义可以看出，热门排行榜的推荐覆盖率是很低的，它只会推荐那些热门的物品，这些物品在总物品中占的比例很小。但是该定义过于粗略，覆盖率 100%的系统可以有无数种可能的物品流行度分布。

为了更细致地描述推荐系统发掘长尾的能力，需要统计推荐列表中不同物品出现次数的分布。如果所有的物品都出现在推荐列表中，且出现的次数差不多，那么推荐系统发掘长尾的能力就很好。信息论中的信息熵和基尼系数可以用来定义覆盖率

- 信息熵

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \log p_i$$

其中p_i是物品i的流行度除以所有物品的流行度之和

- 基尼系数（Gini Index）

$$\text{Gini} = \frac{1}{n-1}\sum_{j=1}^{n}(2j - n - 1)p_{i_j}$$

i_j是按物品流行度p_i $(i = 1, \dots, n)$从小到大排列的物品列表中的第j个物品

- UV 覆盖率

首先定义有效推荐为推荐结果列表长度大于 C 的列表，独立访问的用户去重得到 UV，有效推荐覆盖的独立去重用户数除以独立用户数即 UV 覆盖率

$$\text{Coverage}_{\text{UV}} = \frac{\#(l_{\text{uv}} > C)}{\#\text{uv}}$$

- PV 覆盖率

PV 覆盖率计算方法类似，唯一区别就是计算时分子分母不去重

$$\text{Coverage}_{\text{PV}} = \frac{\#(l_{\text{pv}} > C)}{\#\text{pv}}$$

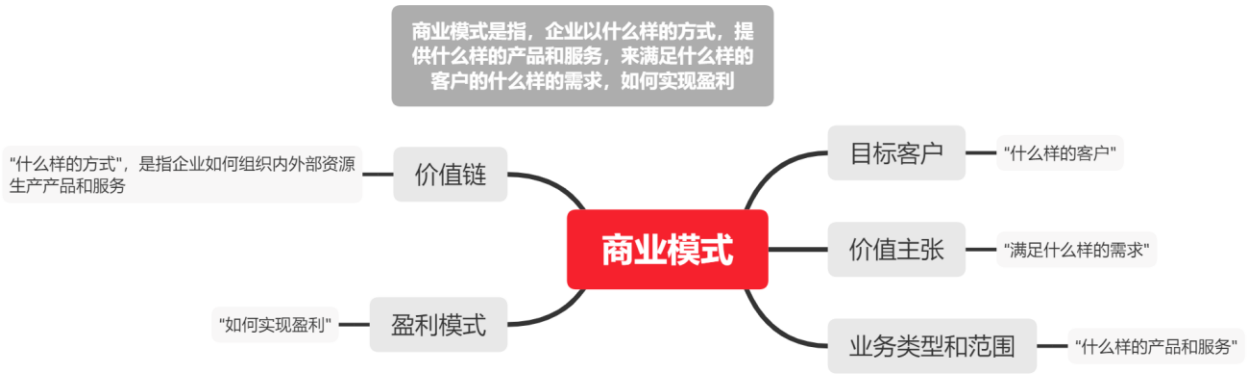
多样性	多样性描述了推荐列表中物品两两之间的不相似性。假设$s(i.j) \in [0,1]$定义了物品i和j之间的相似度，那么用户u的推荐列表$R(u)$的多样性定义如下 $$\text{Diversity} = 1 - \frac{\sum_{i,j \in R(u), i \neq j} s(i,j)}{\frac{1}{2}\lvert R(u)\rvert(\lvert R(u)\rvert - 1)}$$ 而推荐系统的整体多样性定义为所有用户推荐列表多样性的平均值 $$\text{Diversity} = \frac{1}{\lvert U\rvert}\sum_{u \in U}\text{Diversity}\big(R(u)\big)$$ 推荐系统列表多样性与用户实际兴趣多样性应尽可能接近
新颖性	新颖的推荐是指给用户推荐那些他们以前没有听说过的物品。在一个推荐系统中实现新颖性的最简单办法是，把那些用户之前对其有过行为的物品从推荐列表中过滤掉。 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。
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。 目前并没有什么公认的惊喜度指标定义方式，这里只给出一种定性的度量方式。用户满意度只能通过问卷调查或者在线实验获得，而推荐结果和用户历史上喜欢的物品相似度一般可以用内容相似度定义。也就是说，如果获得了一个用户观看电影的历史，得到这些电影的演员和导演集合 A，然后给用户推荐一个不属于集合 A 的导演和演员创作的电影，而用户表示非常满意，这样就实现了一个惊喜度很高的推荐。因此提高推荐惊喜度需要提高推荐结果的用户满意度，同时降低推荐结果和用户历史兴趣的相似度。

信任度	如果用户信任推荐系统，就会增加用户和推荐系统的交互。特别是在电商推荐系统中，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。 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。 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，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
实时性	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。比如，当一个用户购买了 iPhone，如果推荐系统能够立即给他推荐相关配件，那么肯定比第二天再给用户推荐相关配件更有价值。与用户行为相应的实时性，可以通过推荐列表的变化速率来评测。如果推荐列表在用户有行为后变化不大， 或者没有变化，说明推荐系统的实时性不高 ● 实时性的第二个方面是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。而对于新物品推荐能力，我们可以利用用户推荐列表中当天新加入物品的比例来评测。
健壮性	任何一个能带来利益的算法系统都会被人攻击，这方面最典型的例子就是搜索引擎。搜索引擎的作弊和反作弊斗争异常激烈，这是因为如果能让自己的商品成为热门搜索词的第一个搜索结果，会带来极大的商业利益。 推荐系统目前也遇到了同样的作弊问题，而健壮性指标衡量了一个推荐系统抗击作弊的能力。算法健壮性的评测主要利用模拟攻击。 ● 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，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的相似度来评测算法的健壮性。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 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
商业目标	商业目标和平台的盈利模式是息息相关的。一般来说，最本质的商业目标就是平均一个用户给公司带来的盈利。不过这种指标不是很难计算，只是计算一次需要比较大的代价。因此，很多公司会根据自己的盈利模式设计不同的商业目标。 不同的平台具有不同的商业目标。比如电商平台的目标可能是销售额，基于展示广告盈利的平台其商业目标可能是广告展示总数，基于点击广告盈利的平台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，而平台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。

	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	×	✓	×

评估指标的影响因素：如何根据场景设计推荐系统的评估方式？[iv]



推荐场景是制定评价指标时最为关键的影响因素，脱离了推荐场景来谈评测指标就像无水之鱼。所谓"推荐场景"，与所推荐的**内容类型、展现方式、推荐所满足的用户需求**都有莫大的关系。一定要从产品的场景来理解推荐的作用，才能更好地选择评估方法，从众多的评估公式中找到适合当前产品场景及商业目标的指标。

> 例如同样都是推荐视频，但在推荐电影（典型的长视频）、和推荐短视频（一般只有几秒钟长度），其背后所面对的用户需求完全不同。前者展示的是电影海报、名称、评分、主演和故事梗概，用户查看这些内容的目的是尽快挑选出一部适合观赏的电影，因此推荐系统强调的是如何更快更准的给出优质结果。而后者的短视频推荐（例如常见的抖音快手等）用户在浏览过程中目的性不强，而且因为时长短，决策成本低，用户浏览目的是为消磨时间，推荐系统的目的是让用户在这个 app 上停留的时间足够长，粘性足够大。
>
> 对前面这个场景来说，用户在推荐页（注意不是在播放页）停留的时间越长，满意度一定是越低的，谁都不愿意傻傻地在一堆电影名称+海报的挑选页面花费太多的时间，如果挑了十几分钟还没能找出一部接下来值得观看的电影，用户一定会对推荐系统的印象大打折扣。但对后者来说，推荐的过程本身就在不断观赏短视频，为了满足用户 kill time 的需求，多样性、新颖性等更重要。
>
> 如果从评估方法的角度来看，推荐电影等长视频时更多要看在足够短的时间里推出了满足用户持续观看的电影，而且用户看后认为是"高分好片"、一个多小时的观影时间花的值得，是最理想的指标。而对后者来说，黏住用户，增加浏览时长，同时照顾到平台上短视频制作方的曝光和健康生态，则对推荐系统来说是关键考核因素。

因素一：推荐展示槽位固定数量 VS 不断延展的信息 Feed 流

- 固定槽位数量的推荐，更接近搜索引擎或者定向广告的结果。因为展示数量有限，且可能还有先后次序（类似搜索结果从上到下排列），对推荐结果的准确率要求高，这类场景称为 **Top-N 推荐**
  - 此时推荐结果前 N 条的点击率（Click-Through-Rate, CTR）是常见指标，即前 N 条结果的曝光点击率（点击/曝光）
  - 如果推荐结果有明显的先后顺序（如 APP 上从上到下展示结果），那么往往还可以把位置衰减

因素予以考虑，例如 NDCG (Normalized Discounted Cumulative Gain)，MRR (Mean Reciprocal Rank)，或 MAP (Mean Average Precision）都融入了位置因素

- 如果是在移动 APP 上常见的 Feed 流推荐，因为推荐展示槽位数量很多（甚至可视为无限多），用户滑屏又可轻易实现，此时位置因素并不是特别重要，常用曝光点击率（点击量/曝光次数）来衡量推荐质量，此外 PV 点击率（点击量/总 PV）、UV 点击率（点击量/总 UV）也是 Feed 流中常用方法

因素二：电商交易型 VS 广告收益型

- 电商平台推荐系统的目的是更好的促成买卖双方交易（例如各大电商网站、外卖生活类 APP 等）
  - 此时推荐带来的交易笔数占总交易数目的比例、或者交易总金额占 GMV 的比例，就是最直接的评价指标
  - 因为从推荐激发购物者兴趣，到用户完成订单，有漫长的操作链条，所以还可以分解动作以更好的衡量每个环节的效果，例如加购物车率（通过推荐引导的加购物车数量/推荐曝光总数），商品详情页阅读率（通过推荐引导进入商品详情页数量/推荐曝光总数）等
- 广告平台是以广告点击、曝光等作为主要收入来源的（例如常见的各类新闻资讯类 APP，或者短视频类、免费阅读（漫画、小说）类 APP）
  - 广告作为主要收入来源，那么就期望推荐系统能更提升用户在 APP 上停留的时间、点击数等，因为无论是 CPM（cost per “mille” or 1,000 impressions）或 CPC（cost per click）计费的广告形式，用户停留时间越长，越活跃，翻阅次数越多，平均收益就越高

因素三：离线评估 VS 在线实时评估

- 离线数据采集很难做到完全细致全面（例如大量用户的隐式反馈数据很难完整记录，因为性能代价太大），离线评估通常采用静态的评估方法，例如 RMSE、MAE 等
- 在线评估的好处时可以随时进行 AB Test 分流测试，效果好坏一目了然，但其难点有以下两个：
  - 线上环境极为复杂，会受到很多其他因素的干扰，未必真正能反映推荐算法效果的好坏

    例如一些指标很容易受攻击和作弊，一些运营活动也会干扰效果，尤其当抽取比对的流量占比过小时，数据抖动很大，AB Test 的结果未必真能体现实际效果
  - 第二个难点是评估数据往往体现的是最终结果，而不是中间某个模块的直接好坏。如果想用 AB 测试传导到内部更深层次的算法模块，往往需要在工程架构上做大量开发，把内部参数传递出来才行

    例如通过在线评估虽然可以很容易的计算推荐排序策略（Ranking Strategy）孰优孰劣，但如果需要分析之前的召回策略（Recall Strategy）哪个更有效，通过在线评估就困难的多，向前的参数传导需要在大数据工程架构上下功夫（？）

因素四：最大化运营指标 VS 考虑生态平衡和来源多样性

- 推荐的内容如果都来源于平台自身，那么往往只需重点考虑平台关键运营指标最大最优，例如达成更多的交易提升 GMV，或者读者的留存率更高，或者提升整个平台用户的活跃度等就行
- 一些平台的待推荐内容来自 UGC 或 PGC，这些内容提供者依赖平台的推荐来进行内容曝光并获利
  - 在这种情况下，平台要从自身生态平衡、系统长期健康的角度来出发，需要考虑出让一些推荐曝光机会给到长尾 UGC 或 PGC，以避免出现被少量顶部内容渠道绑架导致的“客大欺店”的问题，同时扶植更多的中小内容创作者能让生态更健康繁荣
  - 此时推荐系统作为最重要的指挥棒，其评价指标中一定需要内容来源覆盖率（Source Coverage）、多样性（Novelty）、基尼系数（Gini Index）等指标

因素五：迎合人性 VS 引导人性

推荐系统本质上是利用计算机系统通过大规模数据挖掘来“揣摩”人性。但略微深刻一些来说，人性是最为复杂、矛盾的东西。既有理性的一面，又有感性的一面。

人性都有猎奇、贪婪的一面，而且人性通常是没有耐心的——这也证明了为什么几秒钟的短视频越来越受欢迎，连续剧为什么要有“倍速”功能，以及标题惊悚的短文章总是比内容深刻篇幅长的文章在推荐的时候指标更好看。

人是从众的动物，内心总是关心同类们在看些什么。大量基于协同过滤思想的算法，满足了相关需求。如果充分迎合，会发现大量人群喜欢看的往往是偏低俗、快餐式的内容。如果不加干预，黄赌毒、标题党、危言耸听、猎奇刺激的内容、或者廉价低劣的商品往往会充斥在推荐结果中。

引导人性，倡导更有质量的内容，是推荐系统要肩负的责任，这个时候的评价指标一定不能只单纯看重点击率、转化率等量化指标，因为如果只用这些指标来优化算法，最终结果一定是低劣内容会充斥着版面，降低整个平台的格调。

在推荐系统评估时大家往往语焉不详的“惊喜度”（Serendipity）、“新颖性”（Novelty）等，往往就是在人性揣测的方面进行探索。这些指标计算时最大的难点是评价指标偏主观，很难直接使用在线行为计算。一般只能用事后问卷或者用户对内容的评价评分、转发等行为来间接佐证。或者以 7 日或者 N 日留存率等来判断用户对推荐结果整体的满意度。

总结

推荐系统本质上就是让每个消费者满意，这些指标只是从不同的角度来衡量“满意”这件事情的程度高低。在此小结下常见的指标种类，包括如下几种类型：

- **场景转化类指标**：曝光点击率，PV 点击率，UV 点击率，UV 转化率，加购物车率，分享率，收藏率，购买率，人均点击个数，CTR，AUC 等

- **推荐内容质量指标**：结果多样性（Diversity），结果新颖性（Novelty），结果时效性（Timeliness），结果信任度(Confidence & Trust)等
- **内容消费满意度指标**: 留存率, 停留时长, 播放完成率, 平均阅读时长, 交易量, 沉浸度（Engagement），惊喜度(Serendipity)等

同一个推荐场景下，指标不宜过多，因为太多了不利于最终优化决策，把握准每个场景核心发挥的作用的几个推荐指标就行。**但也不能只有一个指标**，因为过于单一的指标会把推荐算法的优化引入歧途。迷信单一的指标表现好不能说明产品好，而且物极必反，过度优化后的指标虽然上去了，但用户的体验往往会降低。

很多推荐评价指标本身也是脆弱和易受攻击的，一些推荐算法如果严重依赖各类反馈指标来自动优化结果，往往会被恶意利用，所以既要灵活运用推荐评价指标，又不要完全迷信技术指标。因为指标背后体现的是用户的人性。在商业利益和人性之间拿捏到最佳平衡点，是推荐系统开发、以及推荐效果评估的至高境界。

---

[i] 参见 Sean M. McNee、John Riedl、Joseph A. Konstan 的论文“Being accurate is not enough: how accuracy metrics have hurt recommender systems”。

[ii] https://www.infoq.cn/article/pwPEy-LjDtzo86FqR7Hh

[iii] https://zhuanlan.zhihu.com/p/67287992

https://www.infoq.cn/article/Hz_w4DA4RdeBvdBVVfbX

https://juejin.cn/post/6844904065638350861

https://blog.csdn.net/lly1122334/article/details/104221360

[iv] https://www.jiqizhixin.com/articles/2020-04-01-12