达观数据 原创
2020/04/02 10:06

陈运文
作者

达观数据：怎样评价推荐系统的结果质量？

推荐系统是互联网发展至今最常见也重要的技术之一。如今各类APP、网站、小程序等所有提供内容的地方，背后都有推荐系统在发挥作用。

开发好一套真正优秀的推荐系统非常有价值，但也非常艰巨。达观数据是国内推荐系统主要第三方供应商，一直在摸索中前进。**在想办法开发出强大的推荐系统服务好客户时，也一直在思考推荐系统的评估方法。**

众所周知业界有一句俗话："没有评价就没有进步"，其意思是如果没有一套科学的评价推荐系统效果的方法，那就找不到优化改进的方向，打造优秀的推荐系统就无从谈起。

笔者在几年前写过《**怎样量化评价搜索引擎的结果质量**》一文并首发于InfoQ（也可见知乎 https://zhuanlan.zhihu.com/p/30910760）。和搜索引擎相比，移动互联网时代的推荐系统应用面更广阔，评价指标也更复杂。

评价指标像一把尺子，指引着我们产品优化的方向。到底怎样才能科学合理的评价推荐系统的结果质量？从各类文献资料和网上文章里能看到数十种评估公式，让人眼花缭乱。这些指标各自有什么优缺点，应该怎样取舍？本文从我们的实践经验出发，对此进行一些深入的分析，期望对大家有所裨益（达观数据 陈运文）。

针对不同的推荐场景，一定要因地制宜的选择合适的评估方法

推荐场景是制定评价指标时最为关键的，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，而且这种关系体现的有时还很微妙。**

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，谁都不愿意傻傻的在一堆电影名称+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"、1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。

用这个简单例子我们达观想为各位读者们解释的是，**一定要从产品的场景来深刻理解推荐的作用，才能更好的选择评估方法，才能让那些茫茫多的推荐评估公式找到合适的用武之地。**

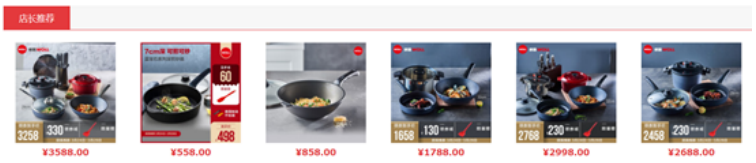
影响推荐系统评估方式的几类因素

场景的细微差异，决定着评估方法应该有所不同。俗话说"什么样的场合穿什么样的衣服"，西装也好运动服也好，都有适配的场合。根据我们对场合细微差异的理解，有以下几个因素在发挥作用：

因素一：推荐展示槽位是固定数量，还是不断延展的信息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条结果的点击率CTR（Click-Through-Rate）是常见指标（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 Cumulatie Gain)，MRR(Mean Reciprocal Rank)，或MAP（meanaverage precision）都融入了位置因素）



常见的展示推荐/广告

还有一类是展示型的推荐，和经典的效果广告非常类似，区别只在于收费方式，如上图。这种情况下推荐系统可以借用广告系统的常见评价方式，例如AUC，ROC等指标。

而如果是在移动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，此时位置先后因素并没有特别重要，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。**此时首屏首条结果并不像Top-N推荐那么重要，因此评估指标也不同。**

因素二：推荐背后的商业模式是以电商交易型、还是广告收益型的

很多推荐系统用于电商平台上，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。推荐最核心目的是促成交易（例如用户完成商品购买，或者用户点播观看某部电影，或用户开始阅读某本小说）；**此时推荐带来的交易笔数占总交易的比例、或者交易总金额与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。

而有一些平台是以广告点击、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，那么期望推荐系统能更好的扩大用户在APP上停留的时间，提高用户点击数等，这些意味着平台能获得更多的广告收入，因为无论是CPM或CPC计费的广告形式，用户越活跃，翻阅次数越多，平均收益就越高。

这种情况下，推荐系统争取满足的用户需求是消磨时间、或"闲逛"的场景，**此时用户平均停留时长、推荐引导下的成功阅读次数等，则更符合需求。**

因素三：推荐评估是离线进行，还是在线实时完成

离线评估和在线评估因为数据准备的条件不同，适合采取的手段也不同。离线数据采集通常很难做到完全细致全面的情况下（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），离线评估方法会有所不同。

典型的离线评估例如有著名的Netflix Prize竞赛、以及KDD Cup、Kaggle上的一些大数据算法竞赛，这些比赛数据集固定，采用静态的评估方法，MSE（Mean Absolute Error）平均绝对误差、RMSE（Root Mean Squared Error）均方根误差，或者R-Squared（R方）来计算：

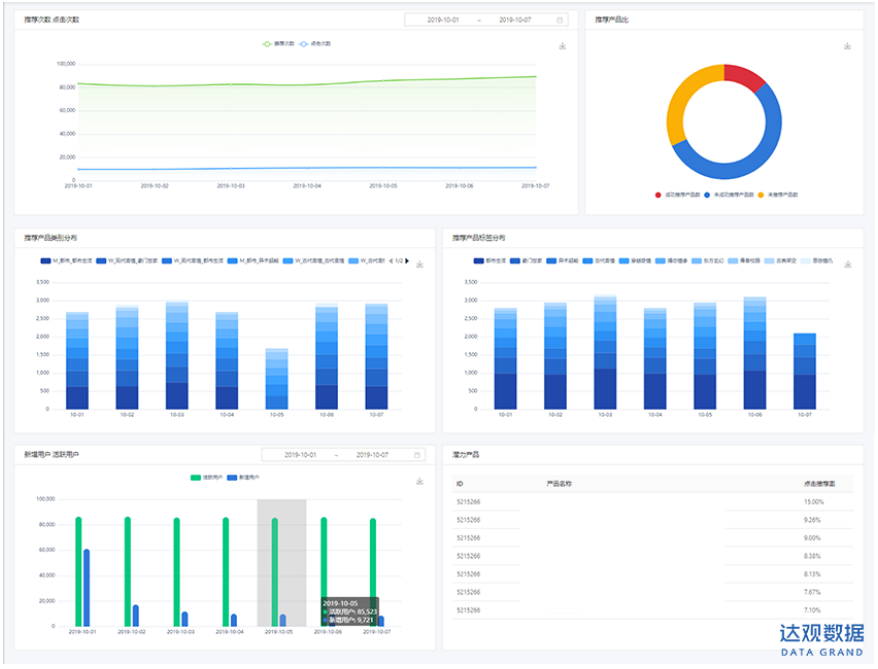
$$MAE = \frac{1}{N}\sum|r_i - \hat{r}_i|$$

$$RMSE = \sqrt{\frac{1}{N}\sum(r_i - \hat{r}_i)^2}$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\hat{y}_i - y_i)^2}$$

例如在电影、电视剧的推荐中，用户-物品评分矩阵（User-Item-Rating）就是常用于离线评估，在学术界尤其常见。因为高校、学术界很难接触真实线上环境，用离线评估是比较方便来评估算法好坏的，也算是学术界的无奈吧。

但我们都知道用户真正给产品评分的显式数据（Explicit Feedbacks）是非常稀缺的，有时我们不得不拍出一些评分映射关系，例如分享映射为几分、点赞映射为几分等，来近似的生成评估矩阵并计算上述这些静态指标。



在线实时计算各类推荐效果指标

而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，工程师们很喜欢。**其难点有以下两个：**

1. **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。另外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。
2. **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

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，但如想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，这也是达观智能推荐一直致力于的。

还有个恐怕是一线算法工程师常常会遇到的难题，就是离线评估的结果和在线测试的结果南辕北辙。离线测下来效果顶呱呱的算法，上线后可能石沉大海一点浪花也看不到。这也恰好证明了正确选择评估方法是多么重要。

因素四：推荐系统当前的目标是最大化运营指标，还是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。

但还有一类复杂的情况，一些平台的待推荐内容来自各个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。毕竟大树之下寸草不死一定不是平台乐意看到的现象。**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将内容来源覆盖率（Source Coverage）、多样性（Novelty）等指标。**

经济学中的基尼系数(Ginicoefficient)，也可以作为辅助的指标用来评价生态的健康程度。推荐系统的初衷就是消除马太效应，使各种物品都能被展示给某类人群。但研究表明主流的推荐算法（比如协同过滤）都是具有马太效应的。基尼系数就是用来评测推荐系统马太效应强弱的。如果Gini1 是从初始用户行为中计算出的物品流行度的基尼系数，Gini2 是从推荐列表中计算出的物品流行度的基尼系数，如果 Gini2 > Gini1则说明推荐算法具有马太效应。

因素五：推荐结果要迎合人性，还是引导人性

推荐系统本质上是让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。

推荐系统一味地迎合人性，会显得"媚俗"，最终也会被用户唾弃。例如人性都有猎奇、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。

但想要引导人性，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。（达观数据 陈运文）

实战中推荐评估指标设置的常用方法

方法一：为不同的细分人群来设置不同的评价指标

基于用户的整体式评估，会让推荐算法导向满足"大多数人口味"的推荐结果，但这背离了千人千面的个性化的初衷。我们期望社群里不同的人都能通过推荐来形成满意的体验。小众的人群偏好往往会淹没在整体数据中，我们一线的算法工程师经常有体会，就是某个新的推荐算法上线后，看整体指标明显好很多了，但是你的领导/客户可能来投诉，说感觉推给他的东西感觉没以前好了。个体和群体经常存在类似的矛盾。某个推荐算法可能对整体有利，但对其中另一类人未必如此。

理想的做法是将其中的人群进行细分，例如电商网站中既有价格敏感型的大众用户，也有追求品质的高端用户。在计算指标时如果划分不同人群来计算，更能体现推荐效果作用后的具体差异。例如我们期望新用户能迅速完成交易并沉淀下来，那么针对这群人的推荐指标，下单率和次日或7日留存就非常重要。而针对高端人群的则有所不同。个体的差异性和小众品味要得到更大程度的重视。

方法二：按不同的推荐位置来制定不同的指标

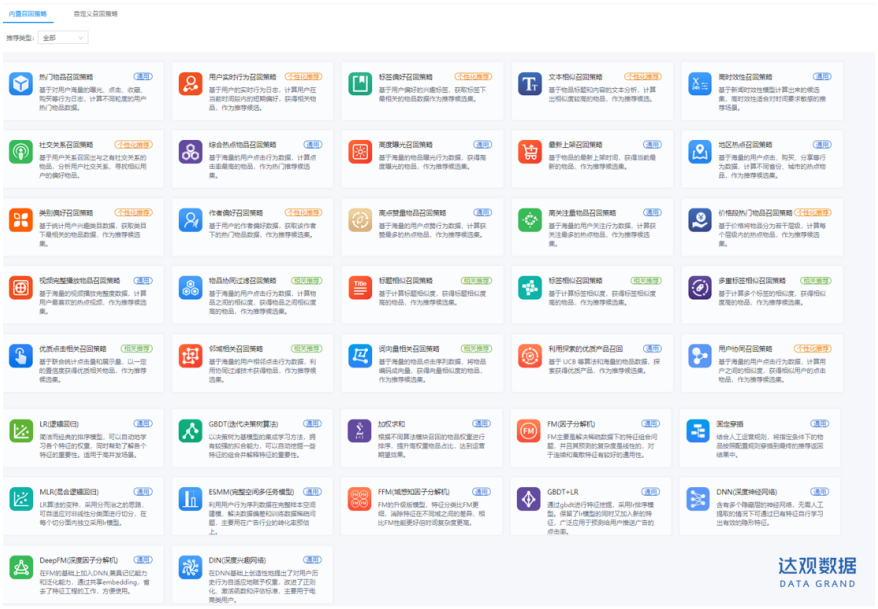
在同一个推荐APP或产品里，不同位置的推荐需要针对性的设置推荐评价指标。前文中提到的不同位置、不同场景，推荐指标制定规则可以有所不同。例如首页首屏的banner推荐（Top-N推荐），信息流Feed推荐，内容详情页下面的相关推荐（常用precision-recall或者F1-score）来计算。还有在搜索无结果页、购物车页面，退出确认页等等，不同的位置一定需要因地制宜的选择适合的评价指标。

方法三：综合几种不同的评估指标来获得整体数据指标

每个指标都有局限性，推荐系统需要平衡很多因素（商业、用户体验、技术实现、资金、人力等），怎么做好平衡是一种哲学。通常可以把几个因素加权求和来作为整体指标。

指标的选择和产品主打定位有关系。例如一个特别强调内容快速新鲜的APP，那么结果的时效性就应该占更大的权重。而一个强调格调品味的APP，单篇阅读时长就显得更可贵。而强调社群活跃度的平台，用户对内容的分享率，互动率等，在整体指标中要更突出。（达观数据 陈运文）

在产品运营的不同阶段，倾向性不同指标的选择也应该有所不同。产品上线前期可能要照顾用户体验，大力拓展新用户。当用户量足够多后，可能会侧重商业变现（推荐的付费视频，在列表中插入较多广告等），想办法通过推荐让产品尽快盈利。如果是电商类推荐，要细致的考虑用户购买前和购买后的差异，以及标品和非标品的差异。购买前往往可以多推荐同类产品以更好的让用户进行比选。当购买动作完成后，尤其是耐用消费品，再继续推荐就适得其反了。



各类推荐算法和指标的灵活选择

推荐指标小结

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

在同一个推荐场景下，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。**

关于作者

> 陈运文：达观数据创始人&CEO，复旦大学计算机博士，国家"万人计划"专家，第九届上海青年科技英才，任复旦大学、上海财经大学校外研究生导师。在人工智能领域拥有丰富研究成果，在IEEE Transactions、SIGKDD等国际顶级学术期刊和会议上发表数十篇高水平科研成果论文，译有人工智能经典著作《智能Web 算法》（第2 版），曾多次摘取ACM KDD CUP、CIKM、EMI Hackathon等世

达观数据

达观数据是一家专注于文本智能处理技术的国家高新技术企业，获得2018年度中国人工智能领域最高奖项“吴文俊人工智能科技奖”，也是本年度上海市唯一获奖企业。达观数据利用先进的...

理论　达观数据　推荐系统

1

相关数据

复旦大学 • 机构

达观数据 • 机构

陈运文 • 人物

权重 • 技术

机器学习 • 技术

展开全部数据

登录 后评论

暂无评论~