**满意度离线（非人工测评）评估**

**一、目的**

解决离线情况下什么情况下算是好的排序。

**二、IQIYI搜索排序思想**

**1) 精准匹配**，用户搜什么词，展现什么内容，而且需要 top 结果排序；

**2) 内容生态，**爱奇艺的视频搜索不仅仅是站内搜索引擎，而是全网的视频搜索引擎，所以我们会囊括所有中文的视频资源，包括我们没有版权的视频，我们希望打造的是，帮助用户链接到想要的资源，同时我们还支持文学、漫画资源的搜索；

**3) 智能分发**，搜索结果有不同的版权方，我们需要对原创结果进行激励，防止略币驱逐良币，载流量上给优质资源进行扶持；

**4) 冷启动问题**，新视频相比于老视频在特征上相对弱势，我们需要给予冷启动空间，在此做一些探索和利用；

**5) 搜索多样性**，防止靠前结果都是一样的。

**三、排序相关性指标检验**

**1)相关性指标：**

1：全命中（文本相关性最高的是否在第一位）

2：相关性，top10中最低的相关性是全局的最高的前10的相关性（综合单字召回这种情况，无法评测好坏，低俗内容的打压不在这个里面）。但是这个和多样性矛盾

3：相邻的样本内容重复

4：前10条结果内容上的多样性

**2)内容质量上：**

由于quality 已经整合了播放源，清晰度等指标。直接以quality为标准进行评估。

1：如何选择规定前10的quality 必须满足什么样的限制难以界定。

**四、基于点击日志的搜索引擎用户满意度评价研究**

**1：研究出发点**

目前，传统的搜索引擎评价主要采用人工 评价的方式[2-3]，需要巨大的时间和人力成本。随着搜索 引擎数据处理规模的日益庞大，传统的搜索引擎评价方 法已无法满足搜索引擎全面快速的评价要求[4]。搜索引 擎的用户点击日志所记录的用户行为信息，隐含了用户 对搜索引擎结果的判断。

**2：用户满意度评价指标**

用户满意度指用户在搜索引擎中检索查询词时，对搜索引擎结果中的单个网页（文档）的满意程度。基于对搜索 引擎的用户行为和用户点击日志的分析，本文选择网页 搜索结果排名、网页点击率、网页平均浏览时间3个评价指标来衡量用户满意度。表1列出了搜索引擎用户满意度评价及其描述。

表1：用户满意度评价指标

|  |  |
| --- | --- |
| **变量名** | **变量描述** |
| Rank | 文档搜索结果排名 |
| Click | 视频文档点击率 |
| Time | 视频文档平均浏览时间 |

各评价指标变量的定义及计算方法如下：

1）网页搜索结果排名 Rank：用户检索查询词时， 网页在搜索引擎查询结果中的排名。搜索引擎的查询 结果是按照网页与查询词之间的相关性排序的，因此可 以用网页排名代表网页与查询词之间的相关程度。网 页搜索结果排名越小，网页在搜索引擎查询结果中的位 置越靠前，网页与查询词之间的相关性越高，用户满意 度越高。

2）网页点击率 Click：在单个用户会话过程内，网 页被用户点击的次数与用户的总点击次数之比，为该会 话内网页的点击率。对所有会话内网页的点击率取平 均值，即网页的点击率。用户对网页的点击次数越多， 点击率越高，说明用户对网页的满意度越高。 （

3）网页平均浏览时间Time：用户点击该网页与点 击下一网页的时间差，即用户浏览该页面的时间。浏览 时间越长，说明用户对该网页的满意度越高。对所有用 户对该网页的浏览时间取平均值，即该网页的平均浏览 时间。

**3：有监督学习数据收集**

数据集中的一个样本记作。

其中 代表查询词， 代表文档， 表示用户对该查询词下的该文档的满意度。

三个指标值根据用户点击日志中的所有用户信息进行统计计算得出，用户满意度由标注人员打分。标注人员根据 经验，对每一个样本记作给出相应的用户满意度打分。为了尽量减轻打分的主观性，请两名 具有网页用户满意度打分经验的标注人员独立打分，最 终结果根据两人打分共同决定。为了保证人工标注结 果的可靠性，使用公式计算Kappa统计 值来衡量两名标注人员标注的一致性程度。其中 是两名标注人员标注结果一致的比例，是标志结果不一致的比例。如果最终得到的 K>0.8 认为 标注过程非常可靠。

**4：有监督学习数据收集**

得到样本集合,可以采用多种模型预测满意度函数

**5：解释说明**

Tip1: 我们最早做过一版满意度模型，后来下掉了，因为当时的KPI是长点击和点击。现在换成满意度说不定之前的模型可以重新考虑。

Tip2: 记最早一版满意度模型为，看看和 那个能够更好的拟合人工评估集合。

若是好，可以一步到位满意度，然后在考虑长点击点击等其他指标

若是更好，则考虑将模型上线后，统计后根据得到满意度，作为人工评估满意度之前的模型自检方案。