**GSB人工评估方法检测分析**

**评测方法**

1. 对整页的结果进行满意度评判
2. 以用户体验为基本出发点
3. 判断前需充分理解Query意义与用户检索意图
4. 不要求所有的结果都点开看
5. 由于两家对象展现形式有比较大的差异，主要以需求满足度为评价标准，对UI形式不做比较关注。

**评分等级**

【很满意】本次搜索我感到非常的满意，**信息很丰富，结果体验很好**。完全不需要更换搜索词条并再次检索。

【满意】本次搜索我感到的满意，我找到了我要的信息，**有个别瑕疵，但不影响我的体验**，不需要更换搜索词条并再次检索。

【一般】有能满足/部分满足我的需求的结果，且有比较多的毛病，有些影响体验。

【不满意】这次搜索不好，结果比较差，结果几乎不能满足需求，或者有能满足/部分满足需求的结果，但是其他结果太差，可以逆转满足需求的结果给我带来的好感。

【很不满意】这次搜索太糟糕了，完全没有满足我的需求的结果。

**评测细则-精品结果**

基本原则：

* 首条满足需求，最少打分【一般】
* Top5只有一条完全不相关的结果，最多打分【一般】
* 完全不相关的结果出现在首位，最多打分【不满意】
* 未完结的电视剧，更新慢1-2集，打分【满意】

**建模**

可以看到相关性上面是需要加限制的。

样本的选择：

点击的记为正样本，在用户点击的最后一条样本之前（skip\_above）的是负样本，最后一条点击之后的样本随机选取负样本。

GBDT模型中，只能对样本进行加权。

1. 对于每一个session 下相关性最高默认成全命中执行如下的规则。





LR中，需要对T的权重加入正则限制。

利用非负线性回归，进行迭代是最基本的模型。相关性的权重占比不高。那么我们给相关性加正则。

记（navboost, instant\_navboost, pctr, quality, quality, t）的权重分别为

记约束向量为：

建立如下的求解模型

求解得到：

**评测细则-精品结果**

很满意标准：

* 全部都为相关结果，片源能播放，且排序合理，无重复。
* 综艺节目在要求单集时，出现全集且在显著位置标记需求级数时，可以很满意
* 未上映的电影/电视剧，质量高的预告，片花专辑
* 专辑/栏目类query,结果多而全（横向比较）。
* 多意图query,先分清主辅需求，主辅需求明确，出主需求的结果就行，不明确时，出现知名的不同类型的结果，且经典的热门的结果排前。
* 未完结的电视剧，同步更新最新的几集。

**评测细则-综合结果**

|  |  |
| --- | --- |
| **等级** | **规则** |
| **好** | **精品/意图query：**出周边的相关视频，如以query为主体的新闻、花絮、预告、节目片段等；且时新性高，片长够。 |
| **中** | **精品/意图query：**视频内容是与query部分相关的新闻、花絮、预告、节目片段等。或时新性不高，片长短 |
| **差** | 与query 完全不相关。或时新性不高，片长短 |
| **注释** | 1：片长长短与query 内容有关，需点击观视频后做结论，当然10秒一下的视频一般也不会有多少内容呈现  2：好、中、差评级可以根据主观体验进行微调 |

建模：时新性要高，相关性要高，时间长度要够应该排在前面。

**评测细则-综合结果**

**整体打分原则：**

Top5结果全好，6-10里也没有差的结果，打分【很满意】

Top3结果全好，4-10里面出现中的结果，但是没有差的结果，打分【满意】

Top5中有差的结果出现，最多打分【一般】（便于发现badcase）

Top3都是中结果，最多打分【不满意】

另外，整体排序不好时，降1-2级别；整体样本多样性不好时，降1级；精品结果有明显重复的，体验降低1级。