

拟公示算法机制机理内容

| | |
|--------|---|
| 算法名称 | 喜马拉雅热搜榜排序算法 |
| 算法基本原理 | 喜马拉雅热搜榜排序算法借助线上统计，采用线性模型，再配上文本过滤、去重、打散、运营干预等手段构建热搜、热点和各专项榜单。 |
| 算法运行机制 | <p>喜马拉雅热搜榜排序算法包含 3 个模块：数据输入、算法模型、干预策略。下面分别从这 3 个模块来描述算法运行机制。</p> <p>数据输入：榜单的排序数据来自于上一个小时内所有 APP 用户的搜索点击量。每天的搜索曝光在千万量级，搜索点击在百万量级。数据类型为文本，不涉及生物特征及生物特征信息，包含榜单搜词候选，各搜词历史得分，各搜词最新线上统计量。</p> <p>算法模型：使用的是线性模型，这个模型是一个线性加权公式：$rankScore = w_1 \times onlineCount + w_2 \times lastRankScore$得分模型受最新的线上统计量和上一个状态的排序分影响。参数和之和为 1，可以设计多组方案进行 AB 实验，选择线上效果最佳的一组参数。</p> <p>干预策略：包括敏感内容屏蔽、去重打散和运营调整三种策略。敏感内容屏蔽是由算法得出的榜单会经过一层严格的敏感文本过滤处理，确保榜单内容健康，不会涉黄涉暴涉政，敏感词为人工添加统一维护，更新榜单时自动化接入。去重打散是通过词的实体提取和编辑距离相似性进行打散，保证相同话题内容不会重复出现。运营调整是运营会对算法得出的最终结果进行微调。包括主动配置一些需要扶持的搜词，对一些容易引起争议或引战的内容下榜。</p> |

| | |
|------------------|--|
| 算法应用场景 | 喜马拉雅热搜榜 |
| 算法目的意图 | 为用户提供热门的分类热搜词榜单，帮助用户发现喜马拉雅热点内容，起到发现性推荐的作用。 |
| 算法公示情况 (选填) | |