**修订记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本号 | 修订日期 | 修改内容 | 备注 | 修订人 |
| USS\_video \_v1.0 | 2021-8-10 | 1：初稿 |  | 孙玲 |

# 超编编辑部需求

超级编辑部系统搜索功能接入统一搜索。支持5家子公司内容的搜索及筛选。8月版本支持视频数据；9月版本支持音乐数据。

## 统一搜索提供热搜榜单查询能力

统一搜索目前有视频、音乐2个热搜榜单接口，这两个接口复用，超级编辑部直接调用这两个接口，并过滤掉直接跳转方式的数据；

## 统一搜索热词文件生成

【需求背景】

超级编辑部搜索框下的热词原始数据是来自搜索统计文件：



运营管理台增加一个定时任务，解析getResult\_WordRate\_shipin\_new\_\*\*\*\*\*\*\*.csv文件内容去掉第一行，取第一列、第三列数据（排序顺序不变），生成新的csv文件，文件名格式为video\_hot\_words\_\*\*\*\*\*\*\*\*.csv，如video\_hot\_words\_20210804.csv 并上传到。

该定时任务每6小时执行一次，优先检查是否有该日期下的文件了，如果有就不再执行后续操作，如果没有就按照上面要求生成文件，并上传。

## 统一搜索提供全内容的筛选能力

【需求背景】

超级编辑部需要调用统一筛选能力，在超编系统上为用户提供筛选能力，包括筛选视频、音乐、数媒、动漫、游戏。前端共5级筛选标签。

无搜索词筛选的场景（5级标签）：





此种场景第3级标签单选；第4级标签不会作为筛选项传递到搜索；第5级标签可多选；即：同维度并集，不同维度是交集。

此种场景中筛选匹配规则：

优先按照传递上来参数进行全匹配筛选，如果筛选到数据返回给超编；

如果没有全匹配筛选到数据，超编会将搜索条件作为关键字，调用统一搜索的搜索接口进行检索。搜索范围按照向上一级。

举例：比如“长视频-电影-动作-经典”无筛选结果，超编调用搜索接口，把“经典”放到作为搜索关键字，在长视频电影动作下面去搜，如果还是无结果，

超编再次调用搜索接口，把“动作经典”放到搜索关键字，在长视频电影下面去搜索。如果还是无结果，

超编再次调用搜索接口，把“动作经典”放到搜索关键字，在长视频下面去搜索，如果还是无结果，

超编会再次调用搜索接口，把“动作经典”放到搜索关键字中，在整个媒资内容下去搜索，有内容返回结果，无内容就不返回。

注：此处已经给超编建议前端做控制，如果当前级别已经是搜索出的结果，再选择其下级的筛选条件，就不要调筛选接口，直接调用搜索接口。把当前级别及选择的下级标签统一作为搜索词搜索，尽量避免超编要反复调用接口，前端展示结果页面时间过长问题。

【具体方案】

统一搜索提供筛选接口：

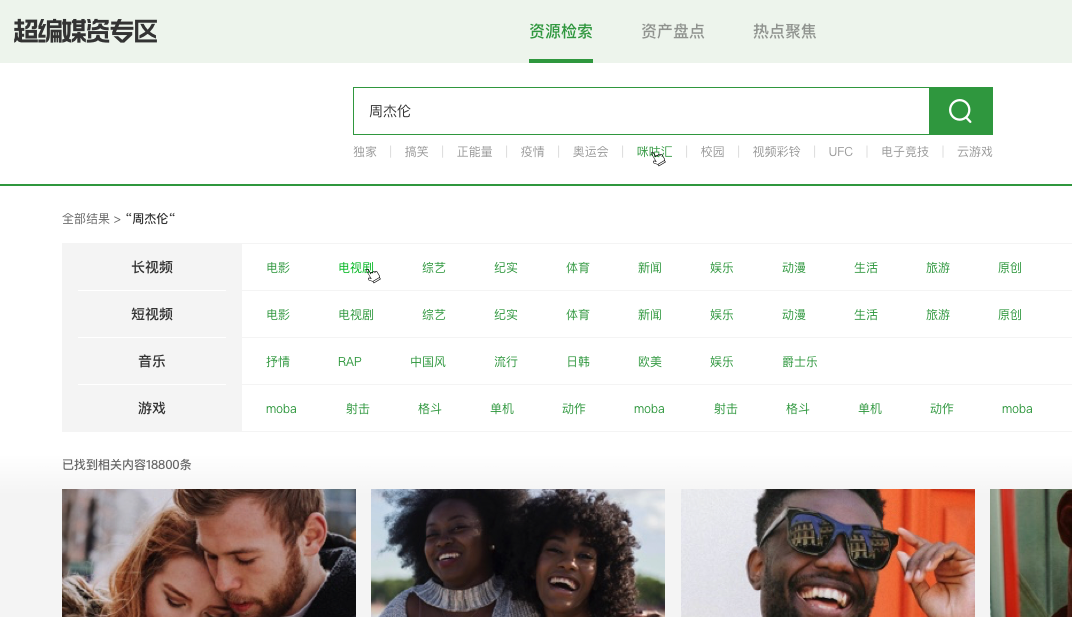
筛选接口匹配标签字段、一级分类字段及长短视频字段，按照如下规则排序：score > op\_hot\_tune > heatvalue > createTime > programid 降序排列；筛选接口支持分页，实际最多返回1600条数据，但筛选内容的总数还是按照ES筛选结果总数返回，搜索控制页数最多返回100页。

## 统一搜索提供全内容的搜索能力

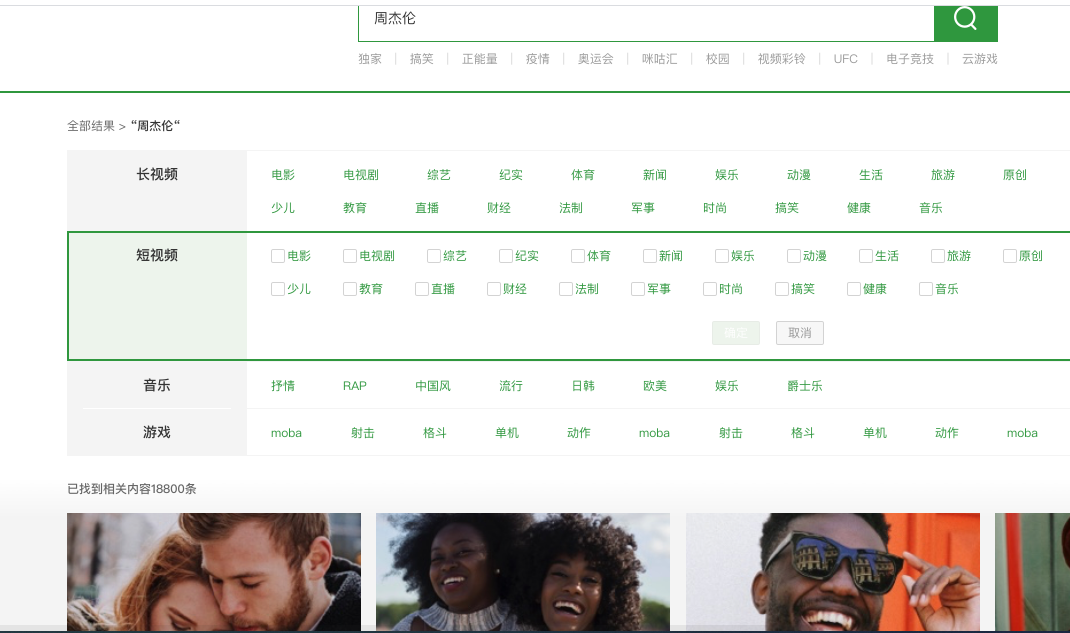
【需求背景】

超级编辑部需要调用统一搜索能力，在超编系统上为用户提供搜索能力，包括搜索视频、音乐、数媒、动漫、游戏。前端有两种调用统一搜索能力的模式，一种为单独搜索词搜索，另外一种为带筛选条件及搜索关键字的搜索。

1、单独搜索词搜索场景：



2、带筛选条件及搜索关键字的搜索：



当搜索结果页返回的筛选条件，同类型筛选条件支持多个取并集；不支持不同维度选择；

当首页带筛选条件的搜索时，同类型筛选条件支持多个取并集，不同类型筛选条件取交集；

【具体方案】

统一搜索新增全内容搜索接口：

1）该接口支持搜索视频、音乐、数媒、动漫、游戏的能力。本版本仅先支持搜索视频。

2）该接口基于视频搜索进行改造，支持根据关键字查询及带条件筛选搜索。分如下场景：

**视频名称搜索：**

召回排序逻辑：去掉关联搜、标签聚类、节目单、挂件搜索、模块化样式、专题搜、大图搜、图集搜、相关视频搜、演员表、角色聚类效果，体育标签，菜单专题、组合专题；增加标签召回权重，并且标签搜索命中级别更高；score得分计算方法：长短视频、热度加分、连载都不需要；目前仅按照标题、标签等、匹配字段的匹配度计算。（视频名完整匹配的,包含去清洗词的完全匹配>标签命中）

语义解析是一级、二级视频分类，优先召回分类下的长视频，然后短视频

**人物搜索：**当系统识别搜索词是人物搜索时，优先展示该人物的长视频数据，再展示根据标题等所有匹配字段匹配到的其他视频；当有多个同名人物时，多个人物的长视频都需要展示，按人物热度排序人物间的长视频；多个人物的短视频就按照正常排序规则排序。

**体育搜索：**体育搜索支持体育搜索倾向。

**语义搜索：**语义搜索在当前视讯搜索的类型基础上做减法，去掉搜索结果不是视频类型的语义，比如“\*\*\*\*\*的演员表”、“\*\*\*\*\*的角色” “\*\*\*\*\*\*的扮演者”。

一级、二级分类搜索及视频人工标签搜索词命中的搜索，搜索结果优先查询了视频知识库数据，优先排序内容是视频标签聚类的内容，排序和视频标签内排序大致相同。

3）该接口有筛选条件带上来时，支持根据筛选条件进行搜索；

4）搜索范围： 内容命中字段除视频当前命中字段外，再增加标签字段；

5）搜索支持纠错、屏蔽词、语义能力使用已有能力；

6）排序：score > op\_hot\_tune > heatvalue > createTime > programid 降序排列；

7）搜索结果支持分页，每页内容数由前端传递；单页内容数不超过100个，搜索最多返回1600条数据，最大支持返回100页；

8）搜索结果返回数据总数（ES实际检索的数据总数）、每条内容的节目id；媒资类型

9）搜索结果返回结果一级分类列表；

当前：精确匹配标题的长视频 》精确匹配标题的短视频 》

标题前缀匹配+标签匹配的长视频 》标题前最匹配+标签匹配的短视频 》书名号内标题匹配+标签匹配视频 》

标题中后缀+标签匹配视频》标签匹配长视频》标签匹配短视频》keyword搜索 》分词匹配

数据处理：

1）从当前视频表中根据packageId获取数据到超编的索引表。PackageId列表为：1002581,1002601,1003861,1003862,1003863,1003864,1003865,1003866,1004041,1004121,

1004261,1004262,1004281,1004321,1004262,1004281,1004322,1004261,1004421,1004422,

1002781,1004301,1004641,1004761,1005061,1005261,1005301,1005321

该packageId支持可配置。

2）热度数据：为各媒资类型的数据热度可做比较，需将heatvalue字段数据做归一化，取热度的最大值，新热度 = 实际热度/最大热度值

3）视讯的运营管理的视频标签增加到新的通用搜索的内容表中，作为一种可筛选的属性进行存储。

特殊： 日志打印增加新模块名及appid；

## 统一搜索提供联想提示能力

通用搜索的联想提示使用的方式是和视频联想提示类似。采用联想提示词库方式进行检索。开发时基于视频联想提示进行改造。具体改造点：

1. 去掉专题查询
2. 去掉节目查询
3. 去掉直播挂件查询
4. 去掉关联提示词

5）保留从纠错获取联想提示能力；其余逻辑和视频联想提示一致；

数据处理：

