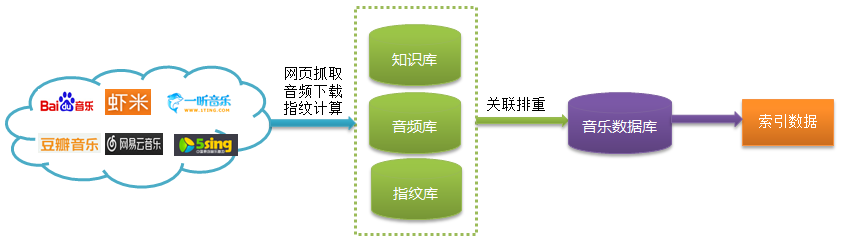
**1   概述**

音乐搜索的数据来自业务运营的库内数据和外网下载的库外数据。库内数据以正规发行的音乐为主，由于版权等原因，与竞品相比会存在数据缺失，需要从外网补充库外数据。本文主要介绍由搜索团队运营的库外数据，以及离线数据处理。

**1.1  库外数据补充**

库外数据主要来自百度，虾米，网易等竞品站点。抓取到页面经过结构化抽取，歌曲信息，专辑信息等元数据会保存到知识库，知识库对数据进行去噪和聚合，生成规整完备的音乐信息集合。如果能够抽取出音频链接，就会下载音频并上传微云，相关的信息保存到音频库。因为音频存储在微云，不存在死链问题。下载的音频也会计算音频指纹，播放时长，码率等信息，经过音频指纹匹配得到指纹聚合关系，这些信息会保存到指纹库。



知识库，音频库，指纹库分别由三个团队维护，数据存储方式不一致，音乐搜索的索引又同时需要三者的数据。为了解决这个问题，把知识库，音频库，指纹库的信息关联排重之后写入音乐数据库，从这个中间数据库出库建索引。

目前知识库包含数千万的歌曲信息和上百万的专辑信息；音频库存储了上千万的音频文件；指纹库计算了数百万的音频指纹。

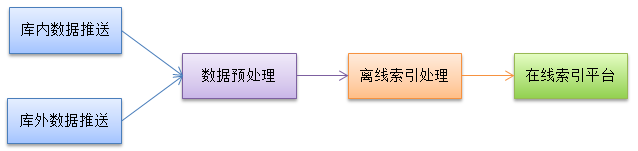
**1.2  离线索引流程**

1)  全量数据索引流程

库内和库外数据每天凌晨推送一次全量数据。数据预处理模块把两份数据合并，进行排重，预排序，索引字段填充，数据格式转换。预处理后的数据推送到离线索引平台[3]，生成顺排，倒排，摘要，属性等索引数据。最后在线索引平台灌入索引数据，提供线上服务。

2)  增量数据索引流程

如果库内或库外数据有变化，会实时推送更新或删除请求。数据预处理模块只进行索引字段填充，数据格式转换，不需要排重和预排序。后续流程同上。



**2   数据补充与优化**

**2.1  数据补充**

库外数据补充基于数据抓取平台[1]和结构化抽取平台[2]，主要通过以下三种方式。

1)  竞品站点定向下载

抓取百度，虾米，豆瓣等竞品站点的数据，能够做到抓全抓透，并监控数据更新。

2)  特殊音频下载

库内数据主要是正规发行的音乐，缺少伴奏，DJ，铃声等特殊音频，通过抓取一些专业网站补充这几类数据。目前已下载了上百万首伴奏，数十万首DJ和铃声。

3)  基于业务反馈的数据补充

根据用户的搜索行为，可以有针对性的发现数据缺失问题。目前已经建立了零结果补充流程，高QV低CTR补充流程，用户不满意反馈补充流程。以零结果补充为例，每个小时收集一次零结果query，下载竞品的搜索结果。

**2.2  数据去噪**

外网抓取的数据存在大量噪音，例如歌曲名中包含歌手名、版本等不相关信息。这些噪音会影响在线检索的文本相关性和知识库聚合。知识库使用多字段关联去噪的方法[4]，去除数据噪音并填入对应的字段，准确率为99%，召回率为78%。

常见的噪音以及去噪后的结果：

|  |  |  |
| --- | --- | --- |
| 去噪前的歌曲名 | 去噪后的歌曲名 | 其他字段修改 |
| 约定 - 光良 | 约定 | 歌手字段修改为光良 |
| Boy Friends (国) | Boy Friends | 语言字段修改为国语 |
| 驼铃(DJ版) | 驼铃 | 版本修改为DJ |
| 病人（电影《百份百感觉II》主题曲） | 病人 | 影视剧字段修改为电影《百份百感觉II》主题曲 |
| 我愿意 (feat. 丁于) | 我愿意 | Feat歌手字段修改为丁于 |

      由于数据去噪不能覆盖到所有噪音，并且有些文档存在字段缺失的问题，这部分数据质量不高，不适合在搜索结果中展现。知识库计算文本规整分来表示文本规整程度，在线排序时根据文本规整分打压低质量的库外数据。

**2.3  数据聚合**

库外数据和库内数据存在大量的重复，把重复数据聚合起来，可以用来索引排重，避免音频重复下载和存储。知识库基于文本相似度进行聚合，指纹库基于音频指纹进行聚合。知识库使用文本聚合和指纹聚合关系，把相同的歌曲聚合为一个实体，整体的聚合占比约为23%。

1)  基于文本相似度的聚合

选择一些关键字段，比如歌曲名，歌手名，专辑，版本等；对每个字段进行文本预处理，计算文本相似度得分；每个字段的得分线性加权，如果大于阈值认为是相同的歌曲，聚合为一个实体。不同字段的文本相似度算法不尽相同，目前使用的算法有余弦相似度，编辑距离，jaccard相似度，simhash相似度等。

2)   基于音频指纹的聚合

使用音频指纹技术，生成音频指纹的聚合关系，为知识库的数据聚合，搜索结果混排，搜索结果聚合等功能提供参考。

**3    离线索引流程**

**3.1  库外索引选取**

库外索引数据选取有三个条件，分别是有音频播放链接，因为搜索结果需要支持播放；经过知识库的去噪，保证数据质量；知识库聚合的多个数据只选取其中一个进入索引。

知识库，音频库，指纹库关联之后的数据保存在音乐数据库，全量索引的库外数据按照上述条件从音乐数据库导出。目前有接近一半的库外数据会进入索引。

**3.2  数据预处理**

数据预处理的处理流程包括以下四个步骤。

1)  数据合并，排重

把库内全量数据和库外全量数据合并，按照docid排重，并校验数据量和字段正确性。

2)  索引字段填充

离线计算或补充一些对业务需求有用的字段。比如语义搜索的标签，混排需要的各种文本指纹，数据统计需要的信息等。

3)  数据预排序

因为倒排索引求交时可能出现超时截断或求交截断，数据预排序可以保证优质数据被优先召回。预排序根据数据类型的优先级，其中**库内正版>库内盗版>库内临时库>库外**，每种类型的数据按照播放量排序。

4)  数据格式转换

生成离线索引平台需要的数据格式。

**3.3  数据校验**

数据是搜索引擎的基础，如果数据出问题必然会导致搜索结果异常。另外离线索引流程较长，涉及多个步骤和团队，也容易出错，通过以下数据校验方法来保证数据正确性。

1)  数据量校验

数据预处理时会先校验数据量，如果数据量与前一天的差距超过阈值，会终止流程、发出告警。

2)  字段校验

数据预处理时会校验重要字段是否存在，格式是否正确。有问题的数据会丢弃，达到一定比例会发出告警。

      在线索引会进行在线的字段校验，从原始数据中随机选取测试集，通过在线检索获得摘要，对比摘要中的重要字段是否与原始数据一致，如果有不一致的文档会发出告警。

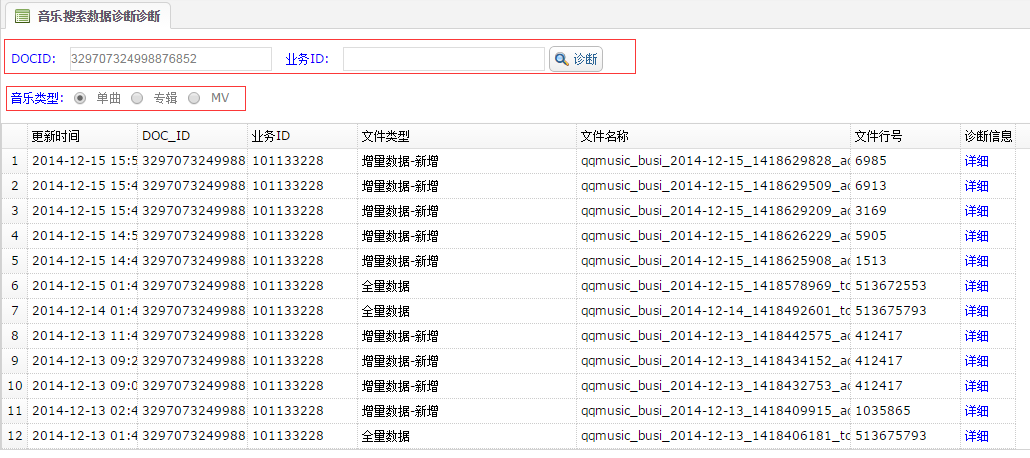
3)  在线覆盖率校验

在线索引平台在数据滚动上线前，使用固定测试集，对比离线新数据和在线老数据的搜索结果，覆盖率低于阈值时停止新数据上线，并发出告警。

**3.4  数据诊断**

音乐经常会有首发的歌曲，或者因为版权问题需要马上下架的歌曲，如果这个时间增量数据的更新流程出问题，会有很大影响。为此我们开发了数据诊断系统，帮助快速定位数据更新问题。

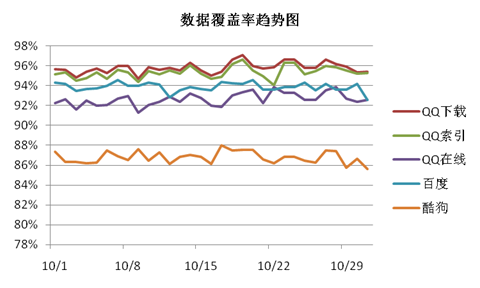
      数据诊断系统支持查询单曲，专辑，MV三类数据，包括最近3天的增量和全量数据推送情况。如下图所示，输入docid或业务id，输出该文档在哪些增量数据和全量数据中出现过，以及相关信息。这样可以快速定位是业务的增量数据推送问题，还是搜索引擎的数据更新问题。后续还会增加离线索引流程中其他环节的诊断。



**3.5  数据覆盖率**

为了评估数据质量，每天选取一个测试集进行覆盖率测试，和竞品对比。覆盖率结果如下图所示，下载和索引的覆盖率高于百度，在线覆盖率略低于百度，高于酷狗。

测试集的选取方法是从音乐的搜索日志里随机选取一批query，分别到QQ音乐，百度，酷狗去搜索，选取前三条搜索结果的歌曲名+歌手名作为测试集。这种测试集的选取方法对三家竞品相对公平。



**4   总结**

本文介绍了音乐搜索团队在数据方面的工作，后续的主要工作是优化数据更新流程，完善监控和容灾，提高可靠性；提高数据质量，进一步优化数据去噪，字段规整；协助补充优质的库外数据进入音乐库内。