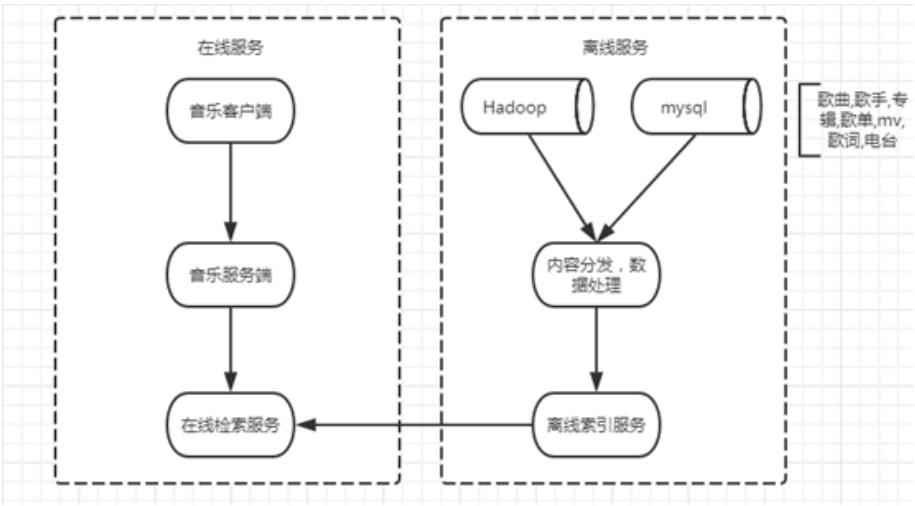
搜索架构 - 搜索提示篇

一、简介

搜索提示服务旨在为用户在输入关键词时给出建议，减少用户的输入量，降低输错的概率，提升搜索体验。

1.1架构方案

整体架构方案：



a) 业务数据推送

原始数据来自于爱听运营的数据库，包括歌手，歌曲，专辑，歌单，电台，用户日志这几类数据，每天推送一次全量数据，每5分钟推送一次增量数据。

b) 数据预处理

数据预处理会做基本的滤重，反垃圾，补充和关联一部分数据。并对文档进行拆分，最后进行文档排序。

c) 离线索引处理

离线索引处理，会生成范围查找的线段树，以及正排的源文档数据。

音乐数据由于有多个种类，可以分多份索引，也可以用同一份索引。因而存在单曲，专辑可能存在同样的词条情况。

搜索提示一般采用前缀分词方案，并支持全拼和简拼。

d) 在线检索

在线检索时，如果能区分query词类别，则根据词类别分别请求索引。或者，请求所有的索引，综合展现结果。

二、索引方案

搜索提示需要展示歌手，歌曲，专辑，歌单，电台等几类数据，最直接的方案是采用多份索引，但我们的索引文档数据量小，这样的运营成本较高，此处采用单索引方案。

2.1前缀分词

1.前缀分词包括原始文本的前缀分词，全拼/简拼的前缀分词，以’刘德华’为例，分词结果如下：

前缀分词：刘；刘德；刘德华

全拼前缀分词：l；li；liu；liud；liude；liudeh；liudehu；liudehua

缩写前缀分词：l；ld；ldh

2.混拼分词

刘德华；

刘德hua；刘dehua；Liudehua

刘德h；刘dh；ldh

2.2特殊索引

普通索引 = 分词

特殊索引 = 分词 + 特殊字符

如 f(x) -> f#x#

s.h.e –> s#h#e#

三、数据预处理

数据预处理是对文档进行有效的扩充，关联歌手-歌曲，歌手-专辑等信息，建立组合索引。

3.1 数据拆分

四类数据都有多种字段需要参与检索，以单曲数据为例，目前有歌曲名，歌手名，歌手别名三个字段。一类数据只会有一个字段建前缀索引，并且前端展示时也要根据命中的字段选择展示方式，因此要把需要建索引的字段进行合并，枚举出所有可能的排列组合方式，把一个原始文档拆分成多个文档。当前的文档分拆方式如下表所示。

原始文档类型	拆分后的文档类型id：索引内容
单曲	100：歌曲名 101：歌曲名+歌手名 102：歌手名+歌曲名 103：歌曲名+歌手别名 104：歌手别名+歌曲名
歌手	200：歌手名 201：歌手别名
专辑	300：专辑名 301：专辑名+歌手名 302：歌手名+专辑名

例如一个单曲文档，歌曲名是十年，歌手名是陈奕迅，歌手别名是Eason，会被拆分成5个文档，文档类型id和索引内容如下表所示。

文档类型id	索引内容
100	十年

101	十年陈奕迅
102	陈奕迅十年
103	十年Eason
104	Eason十年

如果拆分后的多个文档同时召回，会根据原始文档的id进行排重。

前端会根据文档类型选择展示的方式，比如搜“十年”，前端展示“十年 - 陈奕迅”；搜“陈奕迅”，前端展示“陈奕迅–十年”

如果一个歌曲是多个歌手合唱的，跟歌手相关的索引内容会生成多个独立的文本，分别建前缀索引。比如“因为爱情”这首歌，是陈奕迅和王菲唱的，索引内容是“陈奕迅因为爱情；王菲因为爱情”，这样用户只输入一个歌手，也能提示出来。

另外一种是歌曲串烧，一个歌手，有多个单曲进行混合串烧。比如：陈奕迅 – 婚礼的祝福+我是不是该安静的走开+半梦半醒之间。

3.2数据热度

热度值：

单曲的热度是最近7天的播放次数

歌手的热度是最近7天该歌手唱的所有歌曲的播放次数之和

专辑的热度是最近7天专辑里的所有歌曲的播放次数之和

MV是热度是最近7天的播放次数

用户日志热度是最近7天的所有点击次数之和

四、相关性排序

由于索引数据是混合索引，请求时，会根据产品需求，按照不同类别进行展示。

展示顺序为：

歌曲>歌手>专辑>用户日志>电台。

当query有明确分类时：

歌手query：关联该歌手下的所有歌曲，按照热度排序

歌曲query：关联该歌曲下的所有歌手中，热度最高的版本相应的歌手

专辑query：关联该专辑下的所有歌手中，热度最高的版本相应的歌手

a) 热度

一般是根据播放次数，有的情况下也会考虑收藏数，粉丝数

b) 时新

单曲or专辑的发行时间

c) 混排

混排是指在搜索意图不明确时，尽量展示多样化的结果。如用户输入“刘”时，同时会有多个歌手命中，单曲结果中不是展示最热门歌手的4首歌，而是展示4个热门歌手的热门单曲

后来 – 刘若英

我很快乐 – 刘惜君

冰雨 – 刘德华

走在冷风中 – 刘思涵

d) 匹配模式

目前是按照全匹配，精准命中索引内容。但有些长尾词(是中缀，后缀的一部分),无法命中

五、用户运营

目前音乐搜索提示使用文档数为700万+，使用的机器数为4台，峰值qps为2000，支持最大的qps为20000。

搜索提示的日活500万+，点击占比(sug结果点击次数/搜索次数)=54.6%

5.1产品迭代

目前产品还存在一些问题，如：

a) 纠错，点击模型还未接入。

b) 同义词，主题词，复合词，词频度，时新准确度都需要进行更新和维护。

c) 非前缀无法命中，如德华无法找回刘德华，需要接入分词。

d) 无结果query分析

e) 多语言支持

5.2工程迭代

目前工程存在的问题,如：

a) 数据推送中心目前是单机各自推送各自的数据，可能存在不同机器数据不一致的问题。

b) 增量索引问题。

c) Tars特性监控无上报，如query经过纠错，改写，转拼音，关联，原串召回，扩展召回的请求数。

d) 分词词典维护，词权重维护。

e) 反垃圾服务，无效用户查询日志的处理。