王有文

13021008001 · 1131547112@qq.com · johnnwen.github.io

个人信息

- 求职意向：机器学习算法工程师
- 工作经验：2 年
- 学历：计算机硕士(北京邮电大学)

工作经历

- 新浪微博，机器学习部门，算法工程师，2017.04~至今

项目经历

- **视频个性化推荐项目**

  【关键词】Word2Vec、协同过滤、LR+GBDT、Redis

  【目的】对视频流做个性化内容分发

  在视频推荐系统中，采用协同过滤算法、Query、兴趣标签、关注关系等多路召回视频候选集，候选集中都是用户可能会感兴趣的视频集。排序阶段则是在召回的候选集上，使用LR+GBDT算法模型对各个召回通道的候选视频集进行更加精细化的打分排序，达到从候选集中进一步挑选出用户最可能的感兴趣的高质量的视频列表。该推荐列表为top100，以zset存储结构写到Redis库。

- **视频重复检测项目**

  【关键词】CNN、Faiss、ffmpeg、OpenCV、I3D

  【目的】对客户端视频流做内容去重，防止重复内容分发给用户

  该项目主要是对用户新发布的视频进行视频重复检测，主要有两个功能模块，分别是视频特征向量提取、相似视频检索。视频特征向量提取采用I3D网络模型，首先使用ffmpeg音频转换工具包对视频抽样分帧，然后使用OpenCV库将帧图片存储为二进制文本文件，最后将二进制文本文件作为I3D网络模型的输入数据，生成视频特征向量。视频检索阶段，为了加快海量视频的检索速度，使用了Facebook开源的相似性检索类库Faiss，对每一个视频生成一个视频识别码，相似的视频具有相同的视频识别码，在内容分发阶段，相似的视频根据一定的策略规则选择一个相对优质视频分发给用户。

- **用户等级、标签流实时计算项目**

  【关键词】Storm、Flume、Kafka、Hbase、ES

  【目的】 为了提高平台DAU和提高用户粘性

  用户等级项目主要负责对用户经验等级的实时计算。采用flume-kafka-storm-Hbase架构，使用flume实时采集用户行为日志到Kafka，使用storm实时消费Kafka中的行为日志，实时计算每一个用户的经验等级，将计算好的经验实时等级同步到Hbase分布式数据库。标签流项目主要负责对直播、视频、小说流进行实时标签流排序计算，采用flume-kafka-storm-ES架构，对每一个流根据一定的策略实时计算topN个内容并写到ES。上线后，有效的提高了DAU和用户的留存率。

- **直播间舆情分析服务项目**

  【关键词】HLS、ffmpeg、语音转文字、分词、Redis、OSS

  【目的】 提供语音转化为文字的能力，对直播内容和粉丝评论进行审核

  该项目使用了生产者消费者模式和多线程技术。首先根据网页地址抓取内容页直播间url列表缓存到队列，从缓冲队列取出直播间url，拉取每个直播间页面获取M3U8文件，然后解析M3U8文件，下载ts音频片段文件。使用ffmpeg将ts音频文件转换为mp3格式的音频文件。最后使用语音转文字服务，转化语音文件到文本文件并上传到OSS文件，为后期的视频字幕提供服务。对生成的文本文件进行分词后，判断是否有敏感词。上线后，确保了一些色情、政治等敏感词汇不出现在平台。

- **热门微博ABtesting数据系统项目**

  【关键词】Hive、Shell、MR、UDF、Sqoop

  【目的】 为热门微博个性化推荐提供各种指标数据评估

  完成对灰度用户上线后的的各种数据评估指标统计。对hdfs上的有关数据进行hive表的设计、ETL开发，当统计逻辑比较简单时，主要使用hive脚本进行统计，当统计逻辑比较复时，主要使用udf或MR开发。对于定期跑的统计需求，结合python编程，在crontab里启动定时任务。并添加日志和监控。最后使用sqoop从hive仓库同步数据到mysql，从而在UI界面可视化展示。

个人能力

- ★★★ 熟悉java、python语言，熟悉MySQL、Redis数据库
- ★★☆ 掌握贝叶斯、LR、决策树、RF、GBDT等传统机器学习算法
- ★★☆ 掌握CNN、RNN、LSTM等深度学习算法
- ★★☆ 掌握Hadoop、Hbase、Storm、Hive等大数据相关组件
- ★★☆ 掌握Linux 常用命令，Shell脚本书写
- ★★☆ 掌握常用的数据结构与算法
- ★★☆ 掌握simhash原理和应用，掌握相似性检索库faiss原理和应用