关键/JIM GUAN

算法研究员

• 手机: 15624901008

• 邮箱: guanwwjian@163.com

关于我

• 2年算法研究员工作经验(个性化推送相关)

• 技术涉及面广

• 创新能力强

教育工作经历

2017.4至今 搜狗公司 算法研究员 2016.6~2016.12 SAP Java开发实习生 2014.9~2017.4 北京理工大学 计算机科学与技术 硕士 2010.9~2014.7 同济大学 计算机科学与技术 本科

职业技能

- 主要编程语言: Python、Shell、SQL
- IDE: vim
- 大数据处理: Spark (个性化推送模型预测流程)、Hadoop Streaming (日志清洗)、Hive (数据统计)
- 机器学习: TensorFlow (Wide&Deep个性化推送模型)、Spark ML
- 数据库: Redis、MySQL、SQL Lite
- Restful接口服务: Sanic (个性化推送服务)
- Restful接口服务部署相关: nginx、PM2、expect
- 其他编程语言: Java(自定义Hadoop InputFormat/OutputFormat)、C++(个性化后端)
- 数据统计及可视化: Kylin、Redash
- 了解国内推送技术现状及部分细节

工作经历

搜狗"今日十大热点"通知栏推送流程技术侧负责人

- 通知栏推送简介:通知栏推送是APP提升用户留存的一项重要机制,对于已经安装APP的用户,有一定概率可以在APP进程被杀死的情况下展示通知栏消息,一旦用户点击通知栏消息,APP就会被激活至前台,从而引导用户使用APP
- 用户规模 DAU: 50万 推送目标用户总量500万
- 制定了客户端的推送Payload接口格式
- 制定了推送相关Pingback回执格式
- 根据pingback回执日志,进行Push相关的统计工作
- 根据pingback回执日志, 收集并更新客户端Pushid集合
- 推送相关的客户端、前端、后端、运营的协调以及问题排查工作
- 参与推送通道调研及推送整体架构设计

"今日十大热点"个性化推送CTR预估模型

- 功能: 计算用户-文章对的CTR预估评分
- 使用编程语言: Python、Shell
- 使用基于Tensorflow构建的Wide&Deep模型
- 使用Spark进行训练数据的分布式预处理,完成特征提取并将特征处理成TFRecord格式
- 使用GPU训练模型,训练过程使用了动态学习率
- 使用Spark加载tensorflow模型实现分布式CTR评分预测流程,当Tensorflow模型结构发生变化时,无需修改预测流程代码

"今日十大热点"个性化推送服务

- 功能: 收到推送指令后,为每一个用户选择目前CTR预估评分最高的文章,根据文章聚合用户,并把每一篇文章的推送用户集合提交至推送平台进行推送
- 使用编程语言: Python3
- 工作流程:为每个用户维护一个优先队列,接收CTR预估模型计算的文章评分,并将该 文章评分插入到对应用户的优先队列中,评分最高的文章在队首。每当收到推送指令 时,取出每个用户的队首文章,按文章聚合提交推送
- 优先队列使用Redis的Sorted Set结构实现
- Restful接口服务使用Python3下的Sanic框架
- 为了保证推送速度,使用Celery框架提供的生产者-消费者模式来分布式获取每个用户的 队首元素,使发送速度达到了100万用户/分钟

"今日十大热点"pingback回执日志通用统计流程

- 功能:根据配置,对"今日十大热点"pingback日志进行分钟级PV和天级UV统计,该流程支持按多个不同字段维度对日志进行筛选和分类
- 使用编程语言: Python、Shell、SQL
- 优势:
 - o 配置简单: 配置文件中每种统计维度组合只需要一行配置项
 - 。配置项无需频繁修改: pingback日志中的已有字段新增字段值无需作出任何修改, 新字段值的相关统计结果将自动列在统计表中
 - 。 维护简单: 当日志中出现新的字段,只需对hive视图作出修改,无需建新的Hive表
- 架构:
 - o Hadoop Streaming对数据进行清洗
 - 。 使用Hive加载清洗好的数据
 - 。 根据配置文件生成SQL语句并提交至Hive
 - o 将Hive的统计结果插入Mysql
 - 。 使用Redash对Mysql中的统计结果进行可视化

"搜狗搜索"文字搜索个性化"猜你想搜"个性化后端流程维护

- 简介:负责维护已有的"猜你想搜"服务,猜你想搜就是对用户的搜索词进行个性化推 荐,以提升用户使用体验
- 使用编程语言: C++、Python、Shell
- 线下流程根据用户搜索pingback日志使用协同过滤算法计算"搜索历史-推荐词"映照表
- 搜索历史-推荐词映照表以及用户搜索历史存储在Redis当中
- 线上Restful接口使用C++开发,负责接收pingback记录用户搜索历史,以及根据搜索历史查询"搜索历史-推荐词"映照表,并把对应的推荐词返回给用户客户端
- 使用AC自动机算法配合运营黑名单进行黑名单模糊匹配