个性化推荐系统（八）--- 机器学习深度学习召回集扩量

杉枫 探索互联网 2017-10-27

Where'd You Go

Fort Minor - The Rising Tied(Deluxe Version)

个性化推荐系统评价有两个重要指标，一个是召回率一个是准确率。召回率就是：召回率=提取正确信息条数/样本中信息条数。准确率就是：准确率=提取出正确信息条数/提取信息条数。召回率大小直接影响准确率，直接影响机器学习模型、深度学习模型线上效果。

模型实时计算第一步是模型上线，将spark、TensorFlow训练模型通过实时加载，使用到线上实时CTR点击量预估。是机器学习模型第一步，第二步是不断扩大线上召回集，增加新特征来提升点击量预估准确率。

今天主要分享下线上实时模型召回素材、特征集扩容，最开始线上召回集数量是100，扩展到200，整个性能下降到70ms，加上线上逻辑性能已不可接受。这时我们想了个方法用多线程进行多核计算提升性能。经过上线测试每个线程计算50个数据，性能优化到计算只消耗3ms，已经线上使用。

进一步线上召回集扩到1000，采用增加线程每个线程100个特征组，线上能能25ms，这种召回集扩量已在线上使用。

下一步在扩量，性能瓶颈已经是IO，而不是多线程计算。将计算服务改成jar包此时召回集可以进行扩量到2000。

在下一步扩召回集，取素材特征与提供接口服务拆分、接口服务通过并发分布式方式进行请求，此时召回集量应为几种方式最大。需要调整接口服务与素材、特征以及计算服务，通过测试得到IO、线程计算结果合并、多核计算的平衡，需排期配合。

最后一步已基本和开源分布式搜索引擎计算方式类似，后续会持续调研新的优化方式，并引入到线上。总结一下，主要思路是先分开并采用多线程，在合并减少IO，最后通过分布式计算实现召回集扩量。

微信搜索：debugme123

微信扫码或长按：