基于Elastic Search的推荐系统“召回”策略

当我们打开一个资讯APP刷新闻时，有没有想过，系统是如何迅速推送给我们想看的内容？资讯APP背后有一个巨大的内容池，系统是如何判断要不要将某条资讯推送给我们的呢？这就是今天想跟大家探讨的问题——[**推荐系统**](https://www.tuijianxitong.cn/#home)中的“召回”策略。  
  


1. 推荐系统中的召回

召回策略主要满足的是能够从海量的数据中召回一部分相关的候选集，在实现方法上也有很多种，主要可以分为离线召回和在线召回。离线召回的方法主要有协同过滤、聚类算法。协同过滤可以分为基于用户的协同过滤和基于物品的协同过滤。在线召回主要是基于搜索的方式，从该用户的历史记录中利用不同的算法抽取相关信息，这里的相关信息可以是物品的基本信息，也可以是离线标注的标签，然后将这些抽取到的关键信息从索引中快速的查询出相关的结果集。

1. Elastic Search与TF-IDF

Elastic Search是基于Apache Lucene(TM)的一个开源搜索引擎，是一个分布式且具有高扩展性的全文检索的搜索引擎，而且还提供了近乎实时的索引、分析、搜索功能。 Lucene是现今搜索领域被认为速度最快、性能最稳定、功能最全的搜索引擎库，而Elastic Search是以Lucene为核心进行二次开发的搜索引擎，主要完成索引和搜索的功能，它可以通过简单的接口隐藏Lucene的复杂性，从而让整个搜索引擎变得更方便。  
  
在基于内容推荐的算法中，关键词的提取是至关重要的一环，关键词提取直接影响了物品相似度计算的效果。假如没有关键词提取的话，物品全部信息作为特征维度会造成维度灾难，使得维度巨大，且构造的矩阵会相当稀疏，不利于计算。在关键词提取的过程中，TF-IDF算法是很有效率的一种算法。关于TF-IDF算法我们不再这里展开详述。

1. Elastic Search与推荐系统的结合

Elastic Search的搭建过程可以参考官网。安装配置完成Elastic Search之后，当数据进入Elastic Search并完成分词和索引以后，现在只能够根据指定输入的词语进行搜索，与推荐系统并没有任何联系。因此，我们可以根据用户的历史记录进行TF-IDF构造关键词并且输入Elastic Search，采用倒排索引的方式进行存储，以便全文搜索。  
  
简单来说，用户的浏览、收藏等行为会被记录在数据库中，推荐系统会收集用户的各种行为记录，利用TF-IDF关键词提取算法，实现一个概率模型，这个概率模型可以计算出近期用户记录中用户偏好的关键词，由于这些关键词由用户的记录产生，因此就具有了个性化的特性，再将关键词输入到Elastic Search，就可以召回与用户强相关的内容集合。  
  
利用这种方法搭建的推荐系统，其优势也很明显：一是能够对内容池中的所有物料做可计算的精细整合，实现对所有物料的召回，特别是对长尾物料的召回，从而极大地提高物料利用率；二是自带物品冷启动，在没有大量用户数据的情况下也能实现个性化推荐，并且让用户对推荐结果满意。

**先荐是一款赋能媒体的AI产品，是集内容上传、内容管理、内容分发、推荐干预、前端渲染于一体的一站式推荐服务可视化平台，支持PC、WAP、APP全平台接入，帮助媒体从0到1搭建推荐系统，显著提升用户活跃、留存、观看时长等重要业务指标，在减少技术成本投入的同时，大幅提高媒体运营效率，从而实现业务智能化转型。目前已服务人民日报、环球网、花瓣网、果壳网、段友、36氪、简书等三百余家内容平台，其中在环球网web端的兴趣推荐项目上，实现了点击率58%的提升，同时访问量和营收分别增长了69%、20%。**