NEWS PRO

资讯系统

12345 TEAM

目录

概要介绍（S1B）

1. 目标问题

随着信息技术的不断发展，近些年来，海量的数据成为最具价值的财富。在信息传播极其迅速的今天各种数据渗透着我们的生活，它们以指数级的速度增长，数据爆炸将我们带入了大数据时代。

在大数据时代，每个人都有很多的信息接收渠道，却无法很好地从这些信息中找到自己真正感兴趣的内容。为此我们需要设计并开发一个资讯网站平台，通过对用户兴趣的追踪和建模，从而有效地发现用户所关心的内容，并向用户推荐符合其阅读兴趣的内容，且避免向推荐用户强烈不感兴趣的内容。

2. 解决思路

我们希望结合大数据处理技术，通过 Hadoop 提供的 HDFS 文件分布式存储技术，和 Mahout 底层计算框架，并使用组合算法（基于用户的协同过滤和基于物品的协同过滤）针对不同种情况（如游客推荐，用户在线推荐，用户离线推荐），取长补短，搭建一个符合题目需求，满足客户要求的，性能良好的推荐引擎。

3. 具体做法

3.1 对用户兴趣的追踪和建模

我们通过相关技术的学习，我们通过用户的浏览记录和相关浏览文章的时间计算，以及利用用户使用搜索引擎进行搜索的记录，使用 redis-cluster 技术按键值对存储用户的海量数据，并进行 TF-IDF 关键词提取，然后进行用户分类以及可视化建模。

3.2 有效发现用户关心的内容，推荐符合阅读兴趣的内容

我们通过计算用户浏览资讯的有效时间，后台通过一定范围的检测，判断该时间是否有效，去除无效时间，将时间转换为用户评分，然后使用 HDFS 技术，将用户-文章-评分保存在服务器端，实时推荐引擎讲根据基于物品的推荐算法推荐用户评分高的相似物品。

3.3 避免重复推荐

首先用户在线阶段，基于物品的推荐引擎部分会根据用户的点击决定下一步推荐的资讯，当用户喜好有改变之后，新推荐的物品也会不同。

其次，用户离线之后，基于用户的推荐离线引擎会开始运行，将该用户在线阶段的浏览记录与其他邻居用户进行比对，然后结合其他用户的喜好推荐该用户可能喜好的资讯，达到新推荐的物品不同。

4.项目亮点

4.1 强大的技术支持，提供高效的，用户体验良好的资讯网站

通过使用 Hadoop，nginx，solrcloud，redis-cluster，nosql，mahout，scala，kafka，zookeeper 等相关技术，整个系统能处理高并发量下的大数据计算。

4.2 多种推荐算法的组合处理多种情况

基于用户的协同过滤算法，基于用户的无偏好协同过滤算法，基于物品的协同过滤算法，基于物品的无偏好协同过滤算法，多种算法的组合应用。

即使面对推荐系统的冷启动，游客，用户在线，离线推荐等问题依然能以用户的兴趣为主要兴趣源。

4.3 多项式朴素贝叶斯分类算法，对商品进行实时分类

系统采用多项式朴素贝叶斯算法，对资讯进行实时分类，为推荐引擎提供更加精确的推荐。

4.4 提供文章检测错别字功能，和敏感词审核功能

用户写资讯的时候，我们提供资讯的错别字检查功能，用户可以检测自己是否有错别字。针对后台管理审核部分，我们提供文章敏感词审核功能，尽可能让一切自动化。

4.5 多端支持，让整个资讯系统更符合实际需要

我们整个系统包括 PC 端，WAP 端，微信端，安卓端，为用户提供良好的应用体验。

4.6 知识图谱与搜索引擎的结合，提供更全面的搜索信息

solrcloud 集群和 neo4j 图数据库的结合，让横向搜索信息更加准确。