摘要：随着互联网的发展和普及，网络平台和论坛成为了股票投资者交流与讨论的重要平台，然后一些网络的评论信息对股票市场的影响日益显著。中国的股票市场处于刚刚发展的阶段，理性投资者较少，投资的决策很容易收到网络上的消息以及传闻的影响。在网络发展的现在，用户的评论数据量也越来越大，传统数据处理技术遇到了瓶颈，大数据相关技术为数据处理这方面提供了良好的支持。本文针对股票网站系统中，由于互联网“信息过载”造成的难以精确定位用户兴趣并提供产品推荐的问题，利用大数据平台，通过深入挖掘股票社中丰富的用户评论信息，开发产品特征提取算法，建立用户兴趣偏好模型，结合用户评论打分来改善传统协同过滤推荐的推荐准确性；通过一系列的处理，准确的挖掘出有用的信息，并向用户进行推荐。

关键字：

绪论

1.背景和意义

2.现状分析

3.目的和主要内容

4.中国股市简介

5.影响股市因素

5.1宏观因素

5.2微观因素

5.3市场因素

1.大数据相关技术概述

1.1 hdfs文件系统

HDFS（Hadoop Distributed File System）是一个分布式文件系统，提供高吞吐量和一定高度的容错性的应用程序数据访问，适用于大数据集上的应用，对外部客户端而言，HDFS 类似于一个传统的文件系统。可以创建、删除、移动或重命名文件，等等。但是 HDFS 的架构是基于一组特定的节点构建的，这是由它自身的特点决定的。这些节点包括 NameNode，它在 HDFS 内部管理元数据；DataNode是HDFS里的存储块，所有数据存放在其中。由于仅存在一个 NameNode，因此很容易造成单点故障，但是可以搭建一个高可靠的HDFS集群，拥有两个NameNode，一个NameNode为Active状态，一个NameNode为StandBy状态。

存储在HDFS中的文件被分成块，然后将这些块复制到多个计算机中（DataNode）。这与传统的 RAID 架构大不相同。块的大小默认为128M，低版本为64M，当然块的大小和复制的块数量在创建文件时由客户机决定。

1.1.1 概述

1.1.2

1.1.3 特性

1.HDFS中的文件在物理上是分块存储（block），块的大小可以通过配置参数( dfs.blocksize)来规定，默认大小在hadoop2.x版本中是128M，老版本中是64M

2.HDFS文件系统会给客户端提供一个统一的抽象目录树，客户端通过路径来访问文件，形如：hdfs://namenode:port/dir-a/dir-b/dir-c/file.data

3.目录结构及文件分块信息(元数据)的管理由namenode节点承担——namenode是HDFS集群主节点，负责维护整个hdfs文件系统的目录树，以及每一个路径（文件）所对应的block块信息（block的id，及所在的datanode服务器）

4.文件的各个block的存储管理由datanode节点承担--- datanode是HDFS集群从节点，每一个block都可以在多个datanode上存储多个副本（副本数量也可以通过参数设置dfs.replication）

5.HDFS是设计成适应一次写入，多次读出的场景，且不支持文件的修改

1.2 zookeeper协调服务

1.2.1 概述

1.2.2 选举机制

1.2.3 运行机制

1.3 Kafka消息系统

1.3.1 概述

1.3.2 架构

1.3.3 原理

1.3.4 执行流程

1.3.5 特性

1.3.6 使用场景

1.4 SparkStreaming流处理

1.4.1 概述

1.4.2 特性

1.4.3 执行流程

2.股评文本数据采集概述

2.1 数据爬虫概述

2.2 爬虫技术概述

2.3 Java爬虫Jsoup技术

2.3.1

2.3.2

3.文本数据分析相关技术与算法概述

3.1 分词

3.2 筛选

4.系统设计概述

4.1 系统模块概述

4.2 数据库设计概述

5.基于大数据平台的股评信息文本挖掘实现

5.1 股评数据爬取

2.2 消息系统中间键的搭建

5.2 大数据平台实时分词

5.3 词汇筛选

5.4 股票打分

5.5 结果处理以及股票推荐

结 论

附 录

参考文献

致 谢