中国科学技术大学

专业硕士学位论文

（专业学位类型）

**基于Spark协同过滤推荐系统研究与实现**

作者姓名： 张宁

专业领域： 软件工程

校内导师： 戴礼荣 教授

企业导师： 程宁 博士

完成时间： 2017年9月13日

University of Science and Technology of China

A dissertation for master’s degree

（Professional degree type）



**Research and implementation of collaborative filtering recommendation system based on Spark**

Author： Ning Zhang

Speciality： Software Engineering

Supervisor：Lirong Dai

Advisor： Ning Cheng

Finished time:

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文,是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：\_\_\_\_\_\_\_\_\_\_\_ 签字日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

□公开 □保密（\_\_\_\_年）

作者签名：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 签字日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 摘 要

大数据时代的背景下，信息的挖掘过滤对于用户来说是十分重要的。大多数应用场景，用户往往对需求不明确或难以表达，以及由于资源信息的过于丰富，很难从中找到自己需要的内容，这就是如今的信息过载的现象。推荐系统作为缓解信息过载问题的有效方式之一，在社交媒体中的作用也日趋重要。系统通过计算提取用户或商品的特征信息。根据一定的算法规则，从而实现更加主动、智能的为用户推送可能感兴趣的产品信息，从而帮助用户从海量数据中找到自己感兴趣的内容。

然而随着信息量的不断增加，以及推荐系统的应用场景的多样性，推荐系统的实现遇到了各种各样的挑战。其一、是推荐信息的时效性，由于用户某一时刻产生的行为日志信息随着时间推移所产生的推荐结果可能是无效的，因此如何根据用户当前行为，为用户提供实时的推荐成为了关键问题之一。其二，对于新用户或商品，缺少必要的特征信息，如何进行推荐(冷启动问题)。

本文针对上述问题，为了解决推荐系统的实时性和冷启动问题，设计实现了基于Spark实时计算的推荐系统。其的主要内容如下：

1）基于Spark分布式计算框架，根据需求，设计推荐系统基础架构，整个系统结构分为三层：一、实时流处理层，基于Spark Streaming对消息队列的数据进行实时流计算。二、离线计算层，用户处理耗时任务，为实时推荐提供数据基础。三、业务逻辑层，整合实时计算和离线计算资源。为上层结构提供推荐服务。

2）为了弥补传统的协同过滤算法的不足，提出了基于矩阵分解的协同过滤的异步分布随机梯度算法。

3）通过用户时间权重的大小，利用三分图的形式描述各个属性间的关系，进一步获取个性化的预测评分公式，从而解决推荐系统的冷启动问题。

**关键词**：信息过载 spark 协同过滤 冷启动

# ABSTRACT

As the Internet develops at fast speed, people’s requirements for personal digital printing also reflect a tendency of linear growth, and more and more people are looking forwards to completing the production of the personalized prints on the mobile ends and transforming the personalized prints into the printing forms with more quality sense and higher collection values. Since traditional printings and Web printing cannot satisfy such requirement from the mobile users any longer, there arises the mobile printing consumption and personalized printing in response to the new trends.

In this thesis, the author designed and implemented the personalized printing system based on iOS (iPhone Operation System). It provides the users with the main functions including login and registration, classified selection, online printing products, order management, shopping cart and personal information management etc. Mobile users could choose on-demand diversified templates like the photo books, desk calendars, common calendar and business cards etc. to custom-make the personalized prints, breaking down the limitations of traditional and Web printing on the PC (Personal Computer) end. This thesis applies C/S (Client/Server) architecture and MVC model to conduct development, and completes the server-end database designs by utilizing MySQL technology. The client end adopts UIKit framework respectively to realize the graphical interface display for the system, GPUImage technology to conduct the image filter processing, XML (Extensive Markup Language) technology to finish the template layout designs, JSON (JavaScript Object Notation) technology to complete the data exchanges between front and back end. Meanwhile, communication with the back stage will be fulfilled in conformity with HTTP (HyperText Transfer Protocol) Protocol, ultimately accomplishing the basic functions delivered by the system.

Based on the above designs, the author applied the MacOS system and Xcode tool to realize the iOS-based personalized printing system that may satisfy the user requirements for mobile-end printing. With the help of this APP (Application), one could complete all the personalized printing designs and operation on the client end for custom-making personalized prints at their own wills with ease and convenience.

**Key Words**: Printing Spark, Production,iOS,XML

**目 录**

[摘 要 I](#_Toc488932004)

[ABSTRACT 2](#_Toc488932005)

[第一章 绪论 5](#_Toc488932006)

[1.1 课题背景 5](#_Toc488932007)

[1.2 研究现状 6](#_Toc488932008)

[1.3 本文的研究内容 7](#_Toc488932009)

[1.4 论文结构 7](#_Toc488932010)

[第二章 相关工作 7](#_Toc488932011)

[2.1 Spark 分布式框架 7](#_Toc488932012)

[2.1.1 Spark简介 7](#_Toc488932013)

[2.1.2 Spark Core技术 8](#_Toc488932014)

[2.1.2 Spark SQL 9](#_Toc488932015)

[2.1.2 Spark Streaming 10](#_Toc488932016)

[2.1.3 Spark MLlib 技术 10](#_Toc488932017)

[2.1.4 Spark与Hadoop区别 10](#_Toc488932018)

[2.2 HDFS 分布式存储 11](#_Toc488932019)

[2.3 推荐算法 11](#_Toc488932020)

[2.3.1 推荐系统介绍 11](#_Toc488932021)

[2.3.2 协同过滤概述 12](#_Toc488932022)

[2.3.3 基于SVD的推荐 12](#_Toc488932023)

[2.3.4 基于人口统计学的推荐 13](#_Toc488932024)

[2.3.5 基于内容的推荐 14](#_Toc488932025)

[第三章 基于Spark 的推荐系统设计 15](#_Toc488932026)

[3.1 推荐系统实时流计算 15](#_Toc488932027)

[3.1.1 业务需求分析 15](#_Toc488932028)

[3.1.2 功能需求和系统整体结构 15](#_Toc488932029)

[3.2 推荐系统离线计算 16](#_Toc488932030)

[3.1.1 业务需求分析 16](#_Toc488932031)

[3.1.2 功能需求和系统整体结构 17](#_Toc488932032)

[3.3 相关详细介绍 18](#_Toc488932033)

[3.3.1 分布式日志框架 18](#_Toc488932034)

[3.3.2 推荐引擎组 19](#_Toc488932035)

[3.3.3 推荐系统整体架构 20](#_Toc488932036)

[第四章 关于推荐算法的相关问题研究 21](#_Toc488932037)

[4.1 基于item相似度计算 21](#_Toc488932038)

[4.1.1 余弦相似度计算 21](#_Toc488932039)

[4.1.2 基于相关的相似性计算 21](#_Toc488932040)

[4.1.3 调整余弦相似度 22](#_Toc488932041)

[4.2 预测计算 22](#_Toc488932042)

[4.2.1 加权求和 23](#_Toc488932043)

[4.2.2 回归 23](#_Toc488932044)

[4.2.3 性能影响及意义 23](#_Toc488932045)

[4.3 基于矩阵分解的协同过滤的异步分布随机梯度算法 24](#_Toc488932046)

[4.3.1 算法概述及引入原因 24](#_Toc488932047)

[4.3.2 基于Item的矩阵计算 25](#_Toc488932048)

[4.3.3 随机梯度优化 26](#_Toc488932049)

[4.4 关于算法冷启动问题的研究 28](#_Toc488932050)

[4.4.1 冷启动问题概述 28](#_Toc488932051)

[4.4.2 解决方案 28](#_Toc488932052)

[第五章 实验与测试 31](#_Toc488932053)

[5.1 实验环境 31](#_Toc488932054)

[5.2 数据集 31](#_Toc488932055)

[5.3 评价指标 31](#_Toc488932056)

[5.4 实验结果 32](#_Toc488932057)

[参考文献 82](#_Toc488932058)

[致 谢 84](#_Toc488932059)

# 第一章 绪论

## 课题背景

当前我国已经进入信息化时代，信息内容也更加丰富。然而信息技术也带来了一些负面问题，例如“信息过载”现象的出现严重干扰了人们的正常生活。大量的信息也给人们的生活带来了一定的困扰。无论是用户要想从海量数据中摆脱出来，还是商家想要引领消费者进行消费，都离不开个性化推荐系统。为了更好的解决“信息过载”问题，推荐系统的概念应运而生，并且已经得到了越来越多的关注，相关的专家学者们对其研究也更加深入。因此，如何解决大数据时代的实时准确推荐是一个非常重要的课题。

推荐系统的出现则在一定程度上弥补捜索引擎的缺陷。推荐引擎用于处理信息过载的情况，推荐系统通过分析日志数据、用户历史行为、用户属性从而创建用户兴趣偏好模型，依靠推荐算法预测用户对商品权重偏好并进行排序。运样在用户没法准确定位需求描述的同时，推荐系统向用户推送满足需求与兴趣的推荐列表。因此二者为互补的工具，当用户有明确需求时采用捜索引擎满足要求，而在用户没有明确自身需求或者以新成员身份进入到一个新的环境中时采用推荐系统帮助他们获取到感兴趣的内容。因此推荐系统在电子商务平台上获取广泛的成功以及巨大的市场空间。然而如何填充用户商品稀疏矩阵、如何推送高质量推荐列表、快速地响应推荐请求、实时生成推荐列表、海量数据分析都成为必须面对和解决的难题。

随着购物网站规模的不断发展，用户和项目的数量不断增多，这也就导致了用户项目评分阵的稀疏性问题越来越无法避免，协同过滤算法中的数据稀疏性过于严重，推荐的结果更加不准确，从而影响用户对购物网站的信任，造成用户的流失。并且在协同过滤的推荐算法中，往往不能很好的展现用户的爱好迁移的问题，因此也会对推荐系统的性能造成一定的影响。所以，缓解数据稀疏性问题和适应用户喜好化问题已经成为了网络购物系统发展的严峻考验之一。

在新一代推荐系统中为了解决如上难题以及保证系统的可扩展性，需要存储不断增长的数据集以及快速并行化处理、加速建模阶段。常用的技术主要包括：大数据存储技术以及大数据处理技术。推荐算法主要特点在于多个并行操作之间重用数据工作集，属于非循环应用模型的应用，因此引入SPARK计算框架。SPARK是由UC Berkeley推出的大数据处理框架，它提出分布式内存抽象，称为弹性分布式数据集RDD，它允许在计算过程中将中间工作集缓存在内存中，而在后续操作中重用该工作集，运样可以极大加速处理过程。由于大部分数据挖掘算法设及到数据集迭代计算，因此SPARK框架更符合推荐算法的需求，本课题利用改进后的协同过滤算法和用SPARK大数据处理技术实更快、更准确的推荐平台。

## 1.2 研究现状

推荐系统的研究至今已有约20多年的历史，其最早期的研究源自于美国明尼苏大学GroupLens[18]研究小组对一个叫MovieLens的电影推荐系统的研究。研究内容主要是对网站用户打分情况进行分析，从而得到用户可能感兴趣的电影，从而进行推荐。早期的算法主要是基于内容的推荐，通过对推荐物品本身进行特征提取。该方式会导致如果用户没有对某类产品评价，该类产品就永远不会被推荐。随着时间推移，旧的推荐系统问题逐渐暴露，新的推荐算法应运而生。

协同过滤推荐（Collaborative Filtering，简称CF）起源于1992年。从算法理论方面来说，最早期的个性化推荐算法是GroupLens小组提出的协同过滤算法(即基于用户的协同过滤推荐算法)，至于为止该算法在推荐领域仍占有一席之地。基于该算法，人们陆续提出了基于物品的协同过滤推荐算法，基于模型的协同过滤推荐算法等等。经过近20多年的发展，当前推荐系统衍生出多种多样的推荐方式。一、基于上下文的推荐，即推荐系统会根据用户的访问时间、所在城市、天气、季节等因素研究给用户生成TopN推荐，提高推荐效果。二、基于用户画像的推荐，所谓用户画像，就是想用户的所有特征信息标签化，是对用户相关联数据可视化的一种展现形式。多角度抽象用户信息，从而实现个性化推荐系统等。三、基于协同过滤的推荐系统，其中有分为两大类，其一是基于内存的协同过滤推荐算法，包括基于用户的协同过滤推荐算法和基于项目的协同过滤推荐算法等等，其二是基于模型的协同过滤推荐算法，即利用数学建模的手段，评估你的意图，包括基于贝叶斯的协同过滤推荐算法、基于隐语义的协同过滤推荐算法和基于线性回归的协同过滤推荐算法等等。四、基于内容的推荐、即对数据内容进行分析并得到物品的结构化描述信息，分析用户在过去时间中评分活评论的商品，利用用户画像构建预测模型，从而生成推荐策略。五、基于机器学习的推荐系统、根据数据的多样性选取合适的算法策略，自动的归纳逻辑和规则，并根据归纳总结的结果(模型)来对未来数据进行推荐。除此之外，还有基于关联规则的推荐、基于效用的推荐、基于知识的推荐等等。但是，上述每种单一方式的推荐系统都有各自的缺点，比如基于内容的推荐有严重的矩阵稀疏问题，基于关联规则的推荐较为耗时只能做离线处理等等。

由于各种推荐方法都有优缺点，因此目前主流推荐系统都是以上各个策略的混合，研究和应用最为广泛的是内容推荐和协同过滤推荐的组合。尽管从理论上有多种组合推荐的方式，但并非对于某一具体问题一定有效，组合推荐中最重要的原则是通过各种策略的组合，使得各自的推荐技术能够相互弥补，从而实现精准高效的推荐。

## 1.3 本文的研究内容

## 1.4 论文结构

本文余下部分组织如下：

第二章介绍了论文的相关工作：主要是介绍了论文中使用到的关键技术，如Spark 的相关技术栈，以及HDFS分布式存储系统的原理。并对推荐算法以及协同过滤相关算法做了大致的介绍。

第三章主要介绍了基于Spark推荐系统的系统设计，包括离线计算和实时计算部分，以及它们各自使用到的技术以及在整个推荐系统的角色。整个系统以数据收集、数据存储、数据计算、数据合并、数据存储为主线进行描述。

第四章主要介绍了推荐算法的相关问题以及推荐系统的冷启动问题，包括基于Item的推荐算法的优化，以及对性能影响的因素进行分析。对推荐系统冷启动问题进行分析研究，通过结合传统的K-means算法和特征映射法，解决了协同过滤算法的冷启动问题。

第五章对改进的算法使用标准数据集进行测试，并对相关的实验结果进行分析。

第六章总结与展望，对本文的相关工作做出总结并对下一步工作进行展望。

# 第二章 相关工作

## 2.1 Spark 分布式框架

### 2.1.1 Spark简介

Spark发展成立于2012年，以利用当时的五年WordPress经验。 近四年半之后，Spark与广泛的客户合作，从个人和中小企业到跨国公司和政府机构。在2015年底之前，Apache Spark已经成为Hadoop生态系统中最为活跃的项目。作为2009年加州大学伯克利分校AMPLAB研究项目的一部分，Spark与Apache Hadoop并仍在积极发展。 2015年有三个版本，从1.3到1.5，分别包含DataFrames API，SparkR和Project Tungsten等功能。版本1.6在2016年初发布，包括新的Dataset API和数据科学功能的扩展。 Spark 2.0于2016年7月发布，这是一个主要版本，有很多新功能和增强功能，值得拥有自己的一部分。

目前，在Spark核心框架基础上，其主要提供了四个范畴的计算框架其中包括Spark Streaming, Graphx, MLbase, SparkSQL等，其中Spark Streaming是Spark核心API的一个扩展，它对实时流式数据的处理具有可扩展性、高吞吐量、可容错性等特点。Spark Streaming 可以从各种数据源中获取数据并进行复杂计算后推送到文件系统、数据库等存储介质保存。Spark Streaming接收实时的输入数据流，然后将这些数据切分为批数据供Spark引擎处理，Spark引擎将数据生成最终的结果数据。Graphx主要并行图计算，Spark Graphx性能良好，又具有非常丰富的API,并可以基于海量数据进行复杂的图计算从而得到了较为广泛的应用。MLbase 作为Spark生态系统中的一部分，由AMPLab实验室的Mlbase团队开发，是Apache Spark中的核心机器学习库，专门提供机器学习相关的丰富的API。从而让机器学习的门槛更低，从而使那些不了解机器学习的开发人员也能通过MLbase来处理自己的数据。Spark SQL的引入专门用于处理结构化数据比如（DB,JSON），Spark SQL 不仅是SQL,它对各种数据源做了抽象，能够对各种数据源进行ETL操作，并提供丰富的对复杂数据读写的API,使用非常方便，并对多种数据源都提供了很好的支持。其中DataFrame API作为一种具备高伸缩性的数据操作方式，相比RDD API 效率更高。此外Spark SQL还提供了Catalyst Optimizer查询优化器，可以自动优化程序，使得程序更加高效。

Spark引入的弹性分布式数据集RDD，是一种高度受限的共享内存模型，是基于内存的。 这使得其计算效率和速度相比Hadoop 的 MapReduce （基于磁盘，每次执行时都要从磁盘读取数据，计算完毕后都要把数据存放到磁盘上）提升很多。

Apache Spark是围绕速度、易用性和复杂分析构建等方面的大数据处理框架，适用于多种不同的情形的分布式应用场景，包括批处理，流处理，交互式查询等等，并大大减轻了原先需要对各种平台分别管理的负担。统一的RDD抽象和操作，基于统一的技术堆栈，使得Spark称为大数据通用计算平台。在大数据时代逐渐普及的背景下得到了越来越广泛的应用。

### 2.1.2 Spark Core技术

Spark Core是spark整个生态系统的核心，它主要是提供了一套基于RDD的API以及分布式的计算引擎。Spark采用了Master-Slave的分布式计算模型，Master作为主节点，负责任务调度，管理Worker节点，并命令Worker 启动Executor。Worker则是计算节点，负责管理本节点资源已经接收Master的命令并向Master汇报心跳。此外Spark还引入了弹性分布式数据集RDD(Resilient Distributed Datasets)，它是一个容错的、并行的数据结构，可以让用户显式地将数据存储到磁盘和内存中，并能控制数据的分区。同时，RDD还提供了一组丰富的操作来操作这些数据。下面具体介绍spark 执行流程：

如图所示，当RDD算子遇到Action方法时，DAG图即确定了边界，DAG图形成，随后将DAG提交给DAGScheduler。DAGScheduler 会将DAG切分成多个stage,切分依据(宽依赖—shuffledRDD—即数据需要网络传递) 。多个stage封装到TaskSet后提交给TaskScheduler ，随后TaskScheduler把任务提交给worker执行。其中DAGScheduler 和TaskScheduler都在Driver端，main函数创建SparkContext时会使得driver和Master建立连接，Master会根据任务所需资源在集群中找到符合条件的worker。随后Master对worker进行RPC通信，通知worker启动Executor ，Executor会和Driver 建立连接，随后的工作worker和Master不再有关系。 随后Driver会向Executor提交Task。启动Executor。



图2-？ Spark任务执行流程

### 2.1.2 Spark SQL

基于Spark Core，SparkSQL构建了一套基于SQL的引擎。该引擎有两层构成，最底层是查询优化器Catalyst Optimizer, 可以优化程序，使得SQL的运行更加高效。比如两个表的join操作，有多种方式，其一、轮询操作，具有O(n^2)的的时间复杂度。其二、先排序再遍历，该时间复杂度为O(nlogn)。其三则是基于Hash的一种实现，其时间复杂度是O(m+n)。Catalyst Optimizer则会根据其数据分布，数据的特点等应用场景生成合适的执行计划，从而达到SQL高效运行的目的。



图2-？ Spark SQL整体架构

在Catalyst Optimizer之上，Spark SQL提供了两种编写应用程序的方法，一种是经典的SQL方式，另一种是DataFrame/DataSet API 。由于Spark SQL相比于Spark Core显著的优势，使大部分程序的开发不再基于RDD API （如MLlib）而转向基于DataFrame/DataSet API进行开发。

### 2.1.2 Spark Streaming

Spark Streaming 主要功能是根据时间片的长短，将流数据累积为RDD, 然后再将每个RDD进行批量处理，从而达到大规模处理流数据的作用。由于这样的一种模式，使它的吞吐量比现在主流框架Storm还大，同时，它还提供了丰富的API 用于流数据计算中。

### 2.1.3 Spark MLlib 技术

机器学习和统计学技术是从复杂的海量数据中挖掘有用的信息，对于一般开发者来说，即使了解其中的相关算法，但是要想实现，也是有一定困难的。因此MLlib的愿景就是解决开发者的这些问题，为开发者提供便捷的API接口。这样只要具备一定的理论知识，便可以进行机器学习及统计学方面的开发了。

### 2.1.4 Spark与Hadoop区别

Hadoop和Apache Spark都是大数据框架，但它们并不真正具有相同的用途。 Hadoop本质上是一个分布式数据基础架构：它在大量服务器群集中的多个节点之间分发海量数据集，这意味着您不需要购买和维护昂贵的定制硬件。 它还可以索引并跟踪数据，从而大大提高了数据处理和分析效率。 另一方面，Spark是一种数据处理工具，可以对这些分布式数据集合进行操作; 它不做分布式存储。

Hadoop不仅包括称为Hadoop分布式文件系统的存储组件，还包含一个名为MapReduce的处理组件，因此您不需要Spark来完成处理。 相反，您也可以使用Spark而不使用Hadoop。Spark是一个多阶段的具有RAM能力的计算框架，具有机器学习，交互式查询和图形分析的库。 Spark并没有自己的文件管理系统，所以它需要与一个HDFS（而不是另一个基于云的数据平台）集成。

Apache Spark非常适合批处理和流处理，这意味着它是一个混合处理框架。 Spark通过内存计算和处理优化加快了批处理速度，因此其速度通常比MapReduce快得多。 这是流式传输工作负载，交互式查询和基于机器学习的不错选择。 Spark也可以与Hadoop及其模块一起使用。 实时数据处理能力使Spark成为大数据分析的首选。

Hadoop和Spark都是开源项目，因此免费。 但是，Spark使用大量的RAM来运行内存中的所有内容，RAM比硬盘更昂贵。 Hadoop是磁盘绑定的，因此节省了购买昂贵的RAM的成本，但需要更多的系统来分配多个系统上的磁盘I / O。

就成本而言，如果要处理大量的大数据，Hadoop将会更便宜，因为硬盘空间比内存空间要低得多。

## 2.2 HDFS 分布式存储

## 2.3 推荐算法

### 2.3.1 推荐系统介绍

推荐系统将知识发现技术应用于在实时互动中的信息，产品或服务制定个性化建议的问题。在这些系统，特别是基于k-最邻近法（k-nearest）协同过滤的系统已经在网络上取得了广泛的成功。近年来，可用信息量和网站访问量的巨大增长形成了对推荐系统的一些关键挑战。这些挑战是：产生高质量的建议，为数百万用户和项目在每秒钟内提供大量的建议，并在面对较少的数据的情况下实现高覆盖率。在传统的协同过滤系统中，工作量随着系统参与者的数量而增加。因此需要一个即使在非常大规模的问题的情况下，也可以快速产生高质量建议的新的推荐系统技术。为了解决这些问题，我们探讨了基于项目的协同过滤技术。基于项目的技术，首先分析用户项目矩阵以识别不同项目之间的关系，然后使用这些关系间接计算用户的建议。在章节中，我们分析了不同的推荐算法，已经其各自的优缺点。

### 2.3.2 协同过滤概述

协同过滤算法的目标是根据用户以前的喜好和其他志同道合的用户的意见来建议新项目或预测特定用户的某个项目的效用。 在典型的CF场景中，有一个m个用户U={u1,u2,…..um}列表和n个项目I={i1,i2,….in}的列表。 每个用户ui都有一个项目列表Iui，来表达了他/她的意见。 用户可以通过分析定时日志，通过挖掘Web超链接等来将用户的意见明确地进行评分操作，通常这些分数在一定数值范围内，或者可以隐含地从采购记录中导出。 请注意，和Iui可能是一个空集。 存在一个被称为活动用户的专用用户，协同过滤算法的任务是为了找到可以是两种形式的项目可能性。

预测是一个数值，Pa, j，表示对于活跃用户Ua的项目ij ∉Iua的预测可能性。 该预测值与Ua提供的意见值在相同的比例（例如，1至5）内。

推荐是N个项目的列表，Iτ∈I,是活跃的用户最喜欢的项目。 请注意，推荐列表必须是活动用户尚未购买的项目，即 。 CF算法的这个接口也被称为N类（Top-N）建议。

图2-1显示了协同过滤过程的示意图。 CF算法将整个m x n个用户项目数据表示为评级矩阵A。A中的每个条目ai, j表示第i个用户在第j个项目上的偏好分数（评级）。 每个单独的评分都在数字范围内，并且也可以为0，表示用户尚未评估该项目。 研究人员设计了一些协同过滤算法，可以分为两大类：基于内存（基于用户）和基于模型（基于项目的）两种算法。 在本节中，我们提供了基于CF的推荐系统算法的详细分析。

图（2-1）协同过滤过程

### 2.3.3 基于SVD的推荐

在评分系统中，如果建立一个二维评分矩阵，则这将是一个巨大的矩阵，假设用户有10000个，而物品也有10000个，那么这个矩阵的规模就是10000\*10000=1亿。对于大型电商动辄数以亿计的用户及物品而言，这个矩阵的规模更是无法衡量，更重要的是，该矩阵很可能非常稀疏。例如Netflix数据集中非零元素个数只占所有元素个数的1%左右，这种稀疏性给存储和计算带来了巨大的挑战。

奇异值分解(SVD)是一种分解技术。假设A是一个m行n列()的矩阵，则对A进行奇异分解如下：

(公式2.1)

其中∑是一个m\*n的矩阵，除了对角线的元素外，其他的值都是0，而对角线上的非零元素成为奇异值。令s=(), s中的每一个元素为一个奇异值，且按由大到小排列，则一般s中元素会减小的很快，当我们去前r个奇异值进行降维时，即得到：

(公式2.2)

其中r远远小于m和n，这样 , , 三个矩阵的计算会比基于A的运算快很多。

基于SVD的推荐就是先将评分矩阵进行SVD分解，选取适当的奇异值个数r进行低阶近似，得到，,三个矩阵，然后基于这三个矩阵进行推荐。

基于SVD的推荐本质上还是利用用户行为数据，只是在推荐前后先离线对评分矩阵进行分解，然后进行降维，从而大大减小存储的数据量和在线计算量。但它也有以下两个主要缺点：

1) 降维时r维不好确定。r太大，则降维效果不好，r太小又会损失太多信息。在实际使用过程中往往根据奇异值的衰减程度确定r值。

2）SVD本身计算量就非常大，当数据量比较大时，消耗的时间会难以忍受。不过现在Spark已经实现了SVD的并行化计算，当r值选取的较小时，Spark的计算速度还是比较快的。

### 2.3.4 基于人口统计学的推荐

基于人口统计学的推荐是利用人的静态描述信息进行推荐的方法。它简单的根据人口统计信息发现用户的相似度，然后给当前用户推荐与他相近的人喜欢的物品。

具体来说，该方法首先会利用用户的统计信息，如年龄，性别，住址等，然后采用特定的相异度计算策略，计算用户之间的距离。实际计算时用户属性类别往往是不同的，例如性别是二元属性，年龄是序列属性，职业是名称属性等。假设数据集含有p个混合属性，对于对象i和j，我们采用下面的公式计算它们的距离：

(公式2.3)

公式中是属性f的指示符，是含有属性f的记录i和j的相异度。

如图2-2所示，左边是用户甲、乙、丙的信息，根据公式（2.3），用户的属性均是名称类或者二值类属性，计算后得到的用户丙和用户甲的距离为0，而和乙的距离为1，因此用户丙和用户甲更加相似，故我们就能将甲喜欢的东西向用户丙推荐。这里的邻居数只有一个，在实际应用中，可以通过调整邻居数目进行覆盖率更高的推荐。



图2-2基于人口统计学的推荐

### 2.3.5 基于内容的推荐

基于内容的推荐是利用物品的内容信息，先算出物品的相似度，然后根据用户过去喜欢的物品，来向用户推荐与这些记录相似的物品。物品间相似度计算方法很多，这里采用Jaccard公式：

(公式2.4)

或者采用简化的余弦相似度公式：

(公式2.5)

公式（2.4）和公式（2.5）中的表示电影u与电影v的相似度，表示电影u的内容，表示电影v的内容。

如图2-3所示，电影A类型为“爱情，浪漫”，电影B类性为“恐怖，惊悚”，电影C类性为“爱情，浪漫”使用公式2.5计算的到电影A与B的相似度为0，电影A与C的相似度为1，用户甲喜欢电影，故系统会将C推荐给甲。



图 2-3基于内容的推荐

# 第三章 基于Spark 的推荐系统设计

## 3.1 推荐系统实时流计算

### 3.1.1 业务需求分析

  对于推荐系统而言，随着应用场景的多样性，以及许多业务场景的特殊性，对推荐系统实时性的要求日益提高，如果推荐系统不能根据用户的行为变化和商品信息的变化做出及时的结果更新，推荐结果的有效性可能会大打折扣，用户可能会因此无法获取有效的符合自身需求的推荐结果。

       传统的推荐方式采用定时离线计算来更新推荐结果，该方式计算周期往往以小时、天为单位。推荐给用户的商品或信息往往是滞后的、过时的，是基于用户前几个小时或者前一天的行为得到的结果，往往无法满足用户当前的需求，因此这就迫使实际的推荐系统实时性有本质的提高。

### 3.1.2 功能需求和系统整体结构

 一、功能需求

（1）数据收集

      要完成用户实时推荐的计算任务，第一步，要对数据进行采集和清洗操作。对于实时推荐系统的计算需要两部分数据集，一部分是用户之前的离线的数据，第二部分是用户最新的实时数据。其中实时数据主要包括用户ID,用户最新的行为日志(如：用户对商品和评价操作，即用户-商品信息，包括用户ID，商品ID，商品分类，评分等级等)，时间戳字段等。而离线数据如要是用户历史行为和信息构成的用户画像的相关内容。

（2）数据计算

实时流计算是指对源源不断的数据流进行系统的连续计算，并实时的返回结果。为了确保计算的精确性和有效性，就要求系统提供实时的用户行为信息，然后通过相关计算，将结果写入hbase中，为后续推荐操作提供查询。

二、性能需求

(1) 低要求，高效率，可扩展

对于该系统使用的大数据模型采用分布式框架，支持良好的横向扩展性，对于并发的提高只需要横向扩展机器即可。

(2) 高容错性，高可靠性

对于分布式数据集的容错方式一般来说有两种，数据检查点和记录数据的更新。Spark采用记录更新的方式。但是如果更新粒度太细太多，则记录更新成本也会相应的增加。因此spark提供的弹性分布式数据集RDD具有数据流模型的特点：自动容错，位置感知调度和良好的可伸缩性。与分布式共享内存系统需要付出高昂代价的检查点以及回滚机制不同，RDD通过Lineage来创建丢失的分区：RDD记录了单个块上执行的单个操作，每个RDD都记录了他变换的所有操作，以便恢复丢失的分区，这种容错机制有着较低的资源开销和一定的性能提升。

三、系统设计

        实时流计算作为加速层主要用于实时处理，缺点是不能进行复杂计算。

        实时计算层的特点主要是对数据进行流式处理，持续性计算，存储和分析某个窗口内的数据，该方式通常是基于实时性和准确性的一种折中，因此准确性有时并不高(临时解)，后期最终的结果，会通过批处理获取，批处理会覆盖掉流式处理的结果。具体架构图如下：



图3-？ 实时计算系统架构图

其中对于新产生的数据会收集用于离线计算部分，因为实时计算不允许进行耗时的复杂计算。，因此其中一小部分数据会选择性的提交给kafka消息总线，用于实时计算。对于数据收集部分会通过节点上部署的flume上进行，而且flume与kafka也有很好的兼容性，而flume的数据收集部分在离线计算章节会有较为详细的介绍。

## 3.2 推荐系统离线计算

### 3.1.1 业务需求分析

    推荐系统在数据基础构建阶段有大量的耗时操作比如模型训练，中间数据计算如：线性分类器参数，协同过滤中item的相似矩阵的计算，用户画像构建，关联规则的发现等，这些计算通常比较复杂并且有大量的数据库查询操作，计算周期长，有且是算法的瓶颈，需要通过离线计算获得。然后为实时推荐提供数据基础，通过实时数据校正计算结果，从而达到近乎即时的推荐。

### 3.1.2 功能需求和系统整体结构

新数据产生时，我们通过flume/sqoop（用于关系型数据库）将其收集到hdfs中，每半小时存放到hdfs的一个目录中，然后进行批处理(可以使用mapReduce或者spark)处理完后倒入impala, impala是一个sql引擎，可以通过sql语句进行查询。

批处理层特点，因为写到了hdfs系统中，因此数据不可修改，只可以追加，可通过spark或map-reduce进行任意计算，支持水平扩展（通过扩展机器），缺点是高延迟，根据计算量或者数据量的不同，运行时间可能是几分钟到几个小时。



图3-？ 离线计算系统架构

 一、功能需求

(1) 数据收集

数据采集我们使用的是flume框架，该框架主要用于采集非关系型数据-如log，flume为收集、移动大量的日志数据提供了一种可靠的可用的分布式框架，它具有基于流式数据简单灵活的架构，并具有故障转移和恢复机制，为系统提供一定的可靠性和容错能力。flume需要在产生日志的agent结点上部署,用于监听节点产生的数据，agent节点可以将收集的数据传输给collector, 然后collector对数据流进行汇总(主要是为了减少hdfs中小文件的个数，减轻hdfs的并发压力)，并将其写入hdfs中。



图3-？ flume数据采集流程图

（2） 数据计算

## 3.3 相关详细介绍

### 3.3.1 分布式日志框架

基于Spark平台的分布式日志框架如图3-2所示，该框架用于采集访问各个业务系统的日志信息，通过可靠的消息传送中间件作为模型训练与数据源采集之间的纽带，系统构建了基于Kafka集群的消息分发中间件，实现了日志数据的统一下发。由于日志数据中包含着各个业务系统的日志以及用户点击流的日志，在进入离线或实时推荐节点前，需要进行统一的数据清洗。本系统采用Spark Streaming技术按时间分片，对固定时间间隔内收到的数据进行统一的批处理，能达到实时处理的效果，并具有很高的吞吐量。



图3-2 基于Spark分布式日志框架

数据收集主要会经历如下阶段：日志数据采集，日志数据聚合和日志数据过滤。

1)分布式日志实时采集

实时推荐系统需要大量的隐式的用户行为作为基础数据,而且没种用户的行为的源日志信息分布在不同的业务系统中，需要构建分布式日志汇总系统将日志进行收集。以备后续流程使用。分布式日志采集如图3-3所示



图3-3 分布式日志采集模块

2)基于Kafka集群的数据传输

通过构建分布式日志实时采集模块，完成用户行为日志的采集。但是在进入日志过滤阶段之前，由于日志流并发产生且数据量大，为了保证数据的实时性以及尽量减少数据丢失，使用分布式消息订阅和发布系统-Kafka,图3-4 为基于Kafka集群的数据分发架构。



图3-4基于Kafka集群的数据传输

### 3.3.2 推荐引擎组

本系统在线计算部分与具体的推荐目标相关，需要在离线计算的基础上，以最快的速度给出推荐结果。在线计算部分计算量不是很大，数据量也不多，引入spark后，先将输入载入内存，来提高推荐速度。在线计算的结果就是推荐引擎的输出。在经过一系列处理后，就可以返回给前台页面展示。

****

图3-6推荐系统整体架构

### 3.3.3 推荐系统整体架构

推荐系统整理架构主要包含批处理层、实时处理层，数据存储层、应用层。其中批处理层用户进行复杂的耗时的数据处理，实时处理层主要是从消息队列中选择部分数据通过流计算快速获取一个临时解，从而实现准实时的推荐效果。数据存储层用于读取批处理层和实时处理层通过一定的合并策略(择优选择，累加等) 将数据合并，支持随机读，可以在较短的时间内返回查询结果。应用层则是用于推荐结果的展示等一系列的业务处理。



# 第四章 关于推荐算法的相关问题研究

电子商务在过去十年中的巨大增长已经导致了推荐系统的日益普及，帮助客户找出最符合个人品味的产品。近年来提出了许多推荐系统方法，其中基于协同过滤的算法被发现是非常有效的。

## 4.1 基于item相似度计算

我们基于Item的相似度计算做了一定的改进，并用于向用户生成预测。基于item的方法会根据目标用户已评估的项目集合，计算它们与目标项目i的相似程度，然后选择k个最相似的项目{i1，i2，…..，ik}，并且他们之间的相似性{si1，si2，…. sik}也被计算。 一旦发现最相似的项目，则通过对这些类似项目采用目标用户的评级的加权平均来计算预测。 我们在这里详细描述这两个方面，即相似度计算和预测生成。基于项目的协同过滤算法的一个关键步骤是计算项目之间的相似性，然后选择最相似的项目。 两个项目i和j之间的相似度计算的基本思想是首先隔离对这两个项目进行评级的用户，然后应用相似度计算技术来确定相似度si，sj。 图2说明了这个过程; 这里矩阵行表示用户，列表示项目。

计算项目之间的相似性有许多不同的方法。 在这里我们提出三种这样的方法。 这些是基于余弦的相似性，基于相关的相似性和调整余弦相似度。

### 4.1.1 余弦相似度计算

在这种情况下，两个项目被认为是m维用户空间中的两个向量。 它们之间的相似性通过计算这两个向量之间的角度的余弦来测量。 形式上，在图2中的m x n个等级矩阵中，项目i和j之间的相似度表示为sim(i,j)，sim(I,j)由以下公式给出

### 4.1.2 基于相关的相似性计算

在这种情况下，通过计算Pearson-r相关关系j来测量两个项目i和j之间的相似度。 为了使相关性计算准确，我们必须首先隔离共同的情况（即用户对i和j进行评级的情况），如图2所示。将评价i和j的用户集合用U表示 相关相似性由下式给出

这里Ru,i表示用户u对项目i的评级，R̅i是第i项的平均评级。

### 4.1.3 调整余弦相似度

基于用户的CF和基于项目的CF之间的相似度计算之间的一个根本区别在于，在基于用户的CF的情况下，沿着矩阵的行计算相似度，而基于项目的CF的情况下，是沿着矩阵列计算相似度，即共同设定中的每对对应于不同的用户（图4-1）。 在基于项目的案例中使用基本余弦测量的计算相似度有一个重要缺点，不考虑不同用户之间的评级量表的差异。 调整的余弦相似性通过从每个共同对中减去相应的用户平均值来尽可能 抵消这个不足。 一般来说，使用该item i和j之间的相似性由下式给出:

这里R̅u是第u项用户评级的平均值。



图 4-1 共评项目的隔离与相似度计算

## 4.2 预测计算

协同过滤系统中最重要的一步是根据预测生成输出接口。 一旦我们基于相似性度量来隔离最相似项目的集合，下一步是研究用户评级，并使用一种技术来获得预测。 这里我们考虑两种这样的技术。

### 4.2.1 加权求和

顾名思义，该方法通过计算用户对类似于i的项目给出的评分的总和来计算用户u的项目i的预测。 每个评级由项目i和j之间的对应相似度si,j加权求和。 使用图4-2所示的概念，我们可以表示预测Pu, i为

基本上，这种方法尝试了解活动用户如何评估类似项目。通过相似项的加权和来缩放，以确保预测在预计范围内。



图4-2 Item-based 协同过滤算法预测产生过程

### 4.2.2 回归

这种方法类似于加权求和的方式，使用基于回归模型的评级得到近似值，而不是直接使用类似项目的评级的方式。 实际上，使用余弦或相关性度量计算出的相似度可能会产生误差，因为两个评估矢量可能是相差甚远的（欧几里得），但可能具有非常高的相似度。 在这种情况下，使用所谓的“相似项目”的原始评级可能导致较差的预测，基本思想是使用与加权求和技术相同的公式，而不是使用类似项目N的“原始”评级 Ru, N的值，该模型基于线性回归模型使用其近似值Ru, N。 如果我们用Ri和RN表示目标项目i和类似项目N的各个向量，则线性回归模型可以表示为

回归模型的参数α和β的值是通过两个评估向量来确定。 是回归模型的误差。

### 4.2.3 性能影响及意义

最大的电子商务网站以强调直接实施协同过滤的规模运作。 在基于邻域的CF系统中，邻域形成过程尤其是用户相似度计算步骤成为性能瓶颈，这可能会使得整个推荐系统无法完成实时推荐。 确保高可扩展性的一种方法是使用基于模型的方法。 基于模型的系统有可能有助于推荐系统高效运行。 这里的主要思想是隔离邻居的生成和预测生成。

本文中，我们提出了一种基于模型的预测计算项目相似度分数的方法。 相似度计算的约束仍然是基于相关的，但是在项目空间上执行计算。 在典型的电子商务方案中，我们通常会有有一组item，它与用户数量相比而言是变化不频繁的。 项目的静态性质使我们产生预先计算项目相似性的想法。 预先计算项目相似度的一种可能的方式是计算所有相似性，然后执行快速表查找以检索所需的相似度值。 这种方法虽然节省时间，却需要一个O(n2)空间用于n个项目。

事实上，我们只需要一小部分相似的项目来计算预测，这就导致了一个替代的基于模型的方案。 在这个方案中，我们只保留少量的类似物品。 对于每个项目j，我们计算k个最相似的项目，其中k<<n并记录这些项目编号及其与j的相似度。 我们将k作为模型大小。 基于此建模步骤，我们的预测生成算法工作如下。 为了为项目i上的用户u生成预测，我们的算法首先检索与目标项目i相对应的预先计算的k个最相似的项目。 然后统计k个项目中多少是由用户u购买的，然后使用item-based的协同过滤算法来计算预测。

我们在这里观察质量表现的折衷点：为了确保质量，我们必须具有较大的模型尺寸，从而导致上述性能问题。 在一个极端，我们可以有一个模型大小为n，这将确保与原始方案完全相同的质量，但会具有高的空间复杂度。 然而，我们的模型构建步骤确保我们保留最相似的项目。 在产生预处理之前，这些项目对预测分数贡献最大。 因此，我们假设这种基于模型的方法将提供相当好的预测质量，甚至具有小的模型尺寸，因此提供了良好的性能。 我们在本文后面实验验证了我们的假设。 特别地，我们通过改变要存储的类似项目的数量来实验模型大小。 然后，我们进行实验来计算预测和响应时间，以确定模型大小对整个系统的质量和性能的影响。

## 4.3 基于矩阵分解的协同过滤的异步分布随机梯度算法

### 4.3.1 算法概述及引入原因

该方法的主要思想是将用户评级矩阵分布在不同的机器上，每个机器只能访问一部分信息，并且通过网络异步的传播随机梯度优化的更新。每次机器接收到一个参数向量时，它会将其当前参数向量与所接收到的向量进行平均，并从该新点继续进行迭代。此外，我们还引入了基于相似性的正则化，约束用户和项目因子接近其相似用户的平均因素，以及在分布式用户评价矩阵的子部分中找到的项。我们分析正则化术语对Movie-Lens（100K，1M，10M）和NetFlix数据集的影响，并且显示它导致在均方根误差（RMSE）和平均绝对误差（MAE）方面更有效的矩阵分解， 并且与等效同步分布式方法相比，异步分布式方法在收敛时间显着提高。

这些算法的基本假设是，如果两个用户对一组项目具有相同的意见，则它们可能在新项目上具有相同的意见。基于这个想法开发的两种流行的协同过滤技术是: 1）邻近方法，使用类似用户或项目的过去评级生成未知物品的预测; 和 2）矩阵分解技术，其近似于等级的矩阵与对应于用户的两个矩阵的乘积和一些潜在空间中的项目。

我们通过在网络的不同机器上分割评级矩阵来解决这个问题，并提出了一种新颖的矩阵分解异构框架，使用随机梯度优化（SGO）方法进行推荐系统。在现实场景中，由于网络中连接的机器通常加载不同，同步更新可能变得非常慢，因为每台机器必须等待最慢的机器在更新之前完成。此外，为了增强上述基本假设H，我们将经典RMSE目标的两个相似性的正则化项加入到约束用户和项目因素使其在接近于他们最相似的用户和子部分中找到存储在机器中的评级矩阵。在我们的异步框架中，每个机器在评级矩阵的不同部分并行执行SGO，并将其更新的权重发送到网络中的其他机器。一旦机器收到一个更新的权重，它会使用自己的更新对参数权重进行平均，并继续从新的平均应用SGO。 这样可以保证每台机器能够全面了解整个数据。

使用Movielens和Netflix数据集，我们表明，提出的用户和项目正则化项的添加导致在RMSE和MAE方面更有效的矩阵分解。 与同等的同步分布式方法相比，我们还展示了异步分布式更新方案具有吸引力的收敛性。

### 4.3.2 基于Item的矩阵计算

用户和项目分别由集合U和I表示，其中ui和ij分别是第i个用户和第j个项目。此外，我们用rij表示用户ui的项目ij的等级，并且当它们存在时，将等级矩阵R定义为rij的矩阵。用户系数矩阵（相应项）在K维度的潜在空间中K由P（相应的Q）表示，并且pi（和qj）是对应于第i个用户/行（第j项/行）的大小k向量。

随机梯度优化（SGO）的矩阵分解已被证明是推荐系统的成功方法。 这种方法背后的前提是用低维因子矩阵PQT的乘积近似大额定矩阵R.这些因子矩阵试图对具有小数K隐性因子的项目建模用户偏好。 对于存在评级rij的一对用户和项目（ui，ij），该方法是基于最小化的二次正则化二次误差：

其中λ≥0是正则化参数。 因此，整个矩阵分解问题写入

注意，误差仅依赖于P和Q通过pi和qj; 然而，项目ij也可以由用户ui来评估，使得最佳因子qj取决于pi和pi'。 因此，随机梯度优化（SGO）方法如下进行：在每次迭代t中，i选择存在等级的随机用户/项目对（uit，ijt）; ii）对执行梯度步长。

为了加强类似用户具有相似品味的假设，我们强调每个用户（相应项目）的因子向量应该接近于其类似用户（相应项目）的平均因子向量。 为了计算最相似的用户（或项目），我们考虑了两个用户ui和uj写的Pearson相关系数的修改版本：

这里，lc是由两个用户共同评估的项目，ri和rj分别表示ui和uj的平均等级。因此，我们可以找到N个最相似的用户ui，用Ni表示（Nj对应类似于ij的项目）。我们现在提出稍微修改方程式的单个评级目标函数l。 （1）加上另一个正规化术语。 对于存在评级rij的一对用户和项目（ui，ij），相似性规则化的单个目标写成：

其中≥0和≥0分别是与相似用户和相似item相关联的正则化参数。执行与传统SGO相同的更新，但将l替换为l1，我们得到算法l。

以最小化整个矩阵分解问题（方程2），其中l被l1替代，即相似性规则化问题：

### 4.3.3 随机梯度优化

即使随机梯度优化在推荐系统数据集上提供了较高的预测精度，但它也存在与之相关的一些计算挑战。 在单个机器上顺序执行SGO需要较长的收敛时间。 因此，需要以分布式方式对大型数据集进行SGO。 然而，并行化SGO并不是微不足道的。直接实现的一个缺点是因子矩阵的更新可能不是独立的。 例如，对于位于相同行（或列）的训练点，它们同时更新P和Q矩阵的相同对应行（或列）。 因此，计算节点之间的有效通信是必要的，以使因子矩阵上的更新同步。

一种常见的方法是将评级矩阵划分为几个块，并在不同机器上的每个块上运行SGO。在每个机器上更新因子矩阵以获得相应的等级。虽然每个机器上的评级矩阵部分不同，但是因子矩阵之间的更新是同步的。因此，在每个时期之后，每个机器中存在的因子矩阵是同步的。我们将该方法称为“同步SGO”，由于所有机器在每个时期之后同步它们的因子矩阵（在该方法被称为异步SGD，因为参数被异步的传送）。但是，这种方法可能不是有效的，主要是因为SGO的一个阶段的计算时间是根据其执行的机器的不同而不同。因此，在每个时期之后，这种机器之间的同步传播方法可以显着减慢，因为更快的机器必须等待较慢的机器才能同步。

根据上述的不足之处在此提出了一种基于参数矩阵的异步传播的分布式算法。每个机器包含对应于类似用户的集群的评级矩阵的一部分，可以通过大规模图形聚类技术有效地找到。 因此，因子矩阵在每个机器上仅使用其包含的评级矩阵的一部分来更新。与同步方法相反，只要机器完成一个时期，我们就会广播更新的因子矩阵。因此，每当一个新的时期开始，它都会收集其他机器中最新的因子矩阵，并且在完成一个时期之后，将其更新的因子矩阵广播到所有其他机器，并立即开始另一个时期，而与其他机器的状态无关。 这样，所有的机器都是相互独立运行的。 因此，更快的机器将更快地执行，而较慢的机器将在时间上落后，但在完成每个时期之后，它们将从更快的机器接收最新的因子矩阵。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | | SGO | | 基于相似性的正则化的SGO | | | |
| MAE | RMSE | 相似Item()和相似用户 | | 相似用户 | |
| MAE | RMSE | MAE | RMSE |
| ML-100LK | ua | 0.7380 | 0.9469 | 0.7388 | 0.9328 | 0.7396 | 0.9355 |
| ub | 0.7592 | 0.9580 | 0.7555 | 0.9557 | 0.7539 | 0.9580 |
| ML-1M | ra | 0.7314 | 0.9699 | 0.7207 | 0.9514 | 0.7166 | 0.9476 |
| rb | 0.6882 | 0.8859 | 0.6894 | 0.8780 | 0.6510 | 0.8398 |
| ML-10M | rb | 0.6491 | 0.8409 | 0.6479 | 0.8382 | 0.6510 | 0.8386 |
| NF-Subset | NA | 0.6500 | 0.8277 | 0.6459 | 0.8257 | 0.6485 | 0.8265 |

在我们的设置中，我们将等级矩阵R集中到m行的行中，其中m是可用的机器（或进程）的数量。 P矩阵也以相同的方式被阻塞。在我们保持的每个节点处：R的一个块，P的相应块和Q的本地副本。在每个时期，我们基于R的块和各个损失函数来更新P和本地Q矩阵的块（ l或l1）。整个过程如图1所示，显示了三台机器的分布式网络示例。 R1，R2和R3表示每台机器上的额定值。 ep代表了以前时代的结束和不同机器上新纪元的开始。如图所示，在完成一个时期之后，每个机器将其更新的Q矩阵广播到所有其他机器，而在新纪元开始时，每个机器将其Q机更新为来自其他机器的接收更新的Q矩阵（如果有的话）。以这种方式，更快的机器的更新可以帮助较慢的机器更快地收敛，从而增强分布式网络的整体收敛时间。



## 4.4 关于算法冷启动问题的研究

### 4.4.1 冷启动问题概述

推荐系统的冷启动问题主要包括新项目冷启动和新用户冷启动两个方面。而对于新用户的冷启动问题一般可以通过利用用户在其他地方已经沉淀的数据和用户兴趣偏好进行冷启动或者制造选项生成粗粒度的推荐等方式进行。这里着重解决新项目的冷启动问题，在此我们引入了用户时间权重的概念，通过对用户评价的时间间隔判断用户活跃度并根据用户对新项目偏爱程度等因素创建构建个性化的预测评分公式。由于新项目不存在任何一个用户对它的评价信息，故无法使用协同过滤算法来实现推荐，因此，本章节提出解决该类冷启动问题推荐算法，该算法不仅仅使用用户对物品的评分信息对来进行推荐，而是要综合考虑项目本身的诸多属性信息、标签对项目的标注信息等较为全面的考虑影响推荐系统的诸多因素。实验证明该方式，在解决推荐系统在新项目的冷启动的问题上是较为有效的。

### 4.4.2 解决方案

现实生活中，有些用户喜欢去关注并评价新事物，我们定义这类用户为活跃用户；而有的用户可能比较喜欢去关注已经被很多用户评价过的事物，我们定义这类用户为消极用户。对于积极用户来说，他们评价项目的时间与项目发布的时间间隔一般都很短；而消极用户则恰恰相反。在推荐系统应用的过程中，这两类用户是普遍存在的，因此在推荐的过程中我们往往会优先将新项目推荐给积极用户。根据这种思想将时间信息加入到推荐算法中，以此来解决冷启动问题。

为了描述时间信息对推荐算法的影响，首先定义用户时间权重如下：用户项目总数与用户所有评价项目的时间与项目发布时间的差值的总和。

其中sum表示用户评价过的项目总数；timeui 表示用户评价的第i个项目的时间，datai 表示项目i发布的时间。则用户u的时间权重Wu 计算公式如下：

(公式4-?)

由此可知，timeui - datai 越小，说明项目评价时间与项目发布时间的距离越近，用户越积极。反之亦然。

与实际项目为例，用户评价时间表如下：

表一 : 用户评价时间表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Item | U1 | | U2 | | U3 | |
| Time | Date | Time | Date | Time | Date |
| i1 | 2016 | 2015 | 2015 | 2014 | 2017 | 2017 |
| i2 | 2016 | 2014 | 2016 | 2013 | 2017 | 2014 |
| i3 | 2015 | 2012 | 2017 | 2014 | 2017 | 2016 |

以表一为例，采用公式1计算得到每个用户的时间权重wui 即：

根据如上计算可知，用户积极程度由大到小依次是Wu1>Wu3>Wu2 ,即用户Wu1更偏爱于新发布的产品。

此外我们还根据项目标签信息预测评分值。

1. 基于项目评分信息预测评分值

这里，我们假定以用户u为研究对象，分析用户-项目评分矩阵的信息，可以得到用户u评价项目的个数和项目j被用户评价的次数。令ruj 表示用户u是否评价项目j,1表示评价过，0表示没有评价过；ICk 表示用户k评价的项目个数，UCj 表示共同评价项目j的用户个数。那么根据用户对项目的评分信息以及用户评价项目个数以及项目被用户评价的次数信息，可以定义基于项目评分信息的预测评分值，从而挖掘他们之间的关系。具体公式如下:

(公式3-2)

2）基于项目属性信息预测评分值

如公式所示表示基于用户u对项目j的预测评分值。如果用户u 对项目 j 做了评价ruj =1, 反之为0。如果项目j具有属性l，则hlj=1，反之为0。

(公式3-3)

3）基于项目标签信息预测评分值

基于项目标签信息的预测评分计算公式如下：

(公式3-4)

公式(3-4) 中， 表示基于标签的用户u对项目j的预测评分值。如果标签g标注的项目j，则bgj=1分之为0。

下面我们基于具体的数据矩阵和公式来围绕研究对象u1计算预测评分值。

表二 用户项目评分矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| rij | i1 | i2 | i3 | i4 |
| u1 | 2 | 0 | 4 | 0 |
| u2 | 0 | 3 | 1 | 0 |
| u3 | 5 | 0 | 2 | 0 |

表三 属性-项目矩阵

表三 标签-项目标注矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| bij | i1 | i2 | i3 | i4 |
| t1 | 2 | 3 | 0 | 0 |
| t2 | 0 | 3 | 1 | 0 |
| t3 | 1 | 0 | 0 | 0 |
| t4 | 0 | 2 | 1 | 0 |
| t5 | 0 | 0 | 2 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| hij | i1 | i2 | i3 | i4 |
| a1 | 1 | 0 | 1 | 0 |
| a2 | 1 | 1 | 1 | 0 |
| a3 | 1 | 0 | 1 | 1 |
| a4 | 0 | 1 | 0 | 1 |
| a5 | 0 | 1 | 0 | 1 |

其中{i1,i2,…in}为n个项目的集合，{u1,u2…um}表示m个用户，rij表示用户ui对项目j的评分。{a1,a2…,an}表示p个属性的集合，hij表示项目j是否具有an属性。{t1,t2…,tq}表示q个标签的集合,bij表示项目j被标注的次数。为了更好的理解公式与表之间的关系，使用两个三分图的形式来表示。



图一 用户-项目-标签信息 图二 用户-项目-属性信息

因此我们可以根据上述理论基础，分别根据用户-项目-标签信息和用户-项目属性信息来挖掘他们之间的潜在关系。从而通过预测评分值来解决新项目的冷启动问题。

# 第五章 实验与测试

## 5.1 实验环境

实验采用Linux（Centos6.4）操作系统，2.9 GHz Intel Core i5，需要安装的软件有：JDK1.7 、Hadoop2.2.0、Spark1.2.0、Scala2.10.4等。编译工具为IntelliJ IDEA, 采用Scala语言编写算法和相关代码。S

## 5.2 数据集

我们使用我们的MovieLens推荐系统的数据。 MovieLens是一个基于网络的研究推荐系统，在1997年秋季首次亮相。每周hundreds的用户访问MovieLens评分和接收电影建议。该网站如今已有超过43000名用户对3500多种不同的电影发表了意见。我们随机选择了足够的用户从数据库中获得100000个评分（我们只考虑了20个或更多电影的用户）。我们把数据库分成一个训练集和一个测试集。为此，我们引入了一个变量来确定用作训练和测试集的数据百分比;我们称这个变量x。 x = 0：8的值将表示80％的数据用作训练集，20％的数据用作测试集。将数据集转换为具有943行（即943个用户）和1682列（即由至少一个用户评级的1682个电影）的用户项目矩阵A.对于我们的实验，我们还考虑了另一个因素，数据集的稀疏程度。对于数据矩阵R这被定义为 。因此，电影数据集的稀疏性程度是 = 0.9369。在整个论文中，我们称这个数据集为ML。

## 5.3 评价指标

推荐系统研究使用了几种类型的评估推荐系统质量的措施。 它们可以主要分为两类：

1、统计精度指标：通过将数值推荐分数与测试数据集中用户项目对的实际用户评级进行比较来评估系统的准确性。 评级和预测之间的平均绝对误差（Mean Absolute Error，MAE）是广泛使用的指标。 MAE是衡量推荐结果偏离其真实用户指定值的度量。 对于每个评级 - 预测对 ˂pi，qi˃这个度量对待它们之间的绝对误差，即|pi –qi| 平等。 通过首先对N个相应的等级 - 预测对的绝对误差求和然后计算平均值来计算MAE。从形式上看，

MAE越低，推荐引擎预测用户评级越准确。 均方根误差（Root Mean Squared Error ，RMSE）和相关性也用作统计精度度量。

2、决策支持精度指标：评估预测引擎在帮助用户从所有项目集合中选择高质量项目的有效性。 这些指标假设预测过程是二进制操作 - 项目是预测（好）还是不好（坏）。 通过这种观察，如果用户仅选择考虑4或更高的预测，则五分尺度上的项目是否具有1：5或2：5的预测分数是无关紧要的。 最常用的决策支持精度指标是反转率，加权误差和ROC敏感度。

我们使用MAE作为评估指标的选择来重新预测实验，因为它是最常用和最简单的直接解释。

## 5.4 实验结果

在本节中，我们介绍了使用基于项目的协同过滤技术生成预测的实验结果。 我们的结果主要分为两部分质量结果和性能结果。 在评估建议质量时，我们首先在运行主要实验之前确定了一些参数的敏感度。 这些参数包括邻域大小，训练/测试比例x的值以及不同相似性度量的影响。 为了确定各种参数的灵敏度，我们仅关注训练数据集，并进一步将其分为训练和测试部分，并用它们来学习参数。

5.4.1相似度算法的影响

我们实现了三个不同的相似度算法，如3.1节所述的余弦相关性，调整余弦和相关性，并对我们的数据集进行了测试。 对于每个相似度算法，我们实现了算法来计算邻域和使用加权和算法来生成预测。 我们对我们的培训数据运行了这些实验，并使用测试集来计算平均绝对误差（MAE）。 图4显示了实验结果。 从结果可以看出，设置余弦相似度计算的用户平均值具有明显的优点，因为在这种情况下MAE显着降低。 因此，我们选择了余下的实验中调整的余弦相似度。

5.4.2训练/测试比的灵敏度

为了确定数据集密度的敏感度，我们进行了一个实验，我们以0：1的增量将x的值从0：2更改为0：9。 对于每个这些训练/测试比值，我们使用两种预测生成技术 - 基本加权和和基于回归的方法来运行我们的实验。 我们的结果如图5所示。我们观察到，随着x的增加，预测的质量增加。 基于回归的方法显示出比低x值的基本方案更好的结果，但是随着我们增加x的质量趋于低于基本方案。 从曲线，我们选择x = 0：8作为我们后续实验的最佳值。

5.4.3邻域大小的实验

邻域的大小对预测的质量有很大影响。为了确定这个参数的灵敏度，我们进行了一个实验，我们改变了要使用和计算出的MAE的邻居数。 我们的结果如图5所示。我们可以观察到邻域的大小确实影响预测的质量。 但是这两种方法显示出不同的灵敏度。 基本项目算法随着邻域大小从10增加到30而提高，之后增加率减小，曲线趋于变大。 另一方面，基于回归的算法表明随着邻居数量的增加，预测质量下降。 考虑到这两个趋势，我们选择30作为邻域大小的最佳选择。

5.4.4质量实验

一旦我们获得了参数的最优值，我们将基于项目的方法与基于用户算法进行比较。我们在图6中给出结果。从图表可以看出，基本项目-项目（item-item）算法在x（邻域大小= 30）的所有值和邻域大小的所有值（x = 0： 8）。例如，在x = 0：5时，用户-用户方案的MAE为0：755，项目方案显示MAE为0：749。类似地，60个用户和项目-项目方案的邻域大小分别显示为0：732和0：726的MAE。然而，基于回归的算法显示出有意思的结果。在x的低值和低邻域尺寸下，它执行其他两种算法，但随着数据集的密度增加，或者随着我们添加更多的邻居，它表现得更糟，甚至与基于用户的算法相提并论。我们还比较了我们的算法与[12]中描述的基本的非个人化算法。

我们从这些结果中得出两个结论。 首先，基于项目的算法在所有稀疏级别上提供比基于用户的算法更好的质量。 第二，基于回归的算法使用非常稀疏的数据集执行得更好，但随着我们添加更多数据，质量下降。 我们认为这是因为回归模型受到高密度水平的数据过拟合的影响。

5.4.5绩效结果

在显示基于项目的算法提供比基于用户的算法更好的预测质量之后，我们专注于可扩展性问题。 如前所述，基于项目的相似性更为静态，并允许我们预先计算项目邻域。 该模型的预先计算具有一定的性能优势。

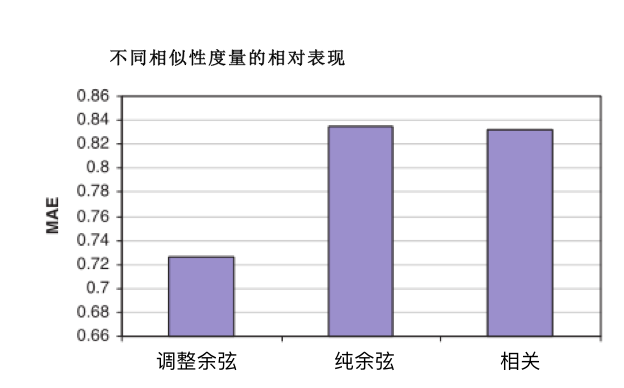


图4：相似度计算对项目协同过滤算法的影响。

# 参考文献

1. 刘建国，周涛.个性化推荐系统的研究进展[J].自然科学进展，2009,19(1):21-15.
2. X. Amatriain and J. Basilico. Netix recommendations: http://techblog.netix.com/2012/04/netix-recommendations-beyond-5-stars.html/.
3. 蔺丰奇,刘益.信息过载问题研究述评[J].情报理论与实践，2007,30(5):710-714.
4. R. Burke, M. O'Mahony, and N. Hurley. Robust collaborative recommendation. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, Recommender Systems Handbook, chapter 25, pages 805-835.
5. Resnick P, Varian H R, Recommender system[J]. Communications of the ACM,1997,40(3):56-58.
6. 刘建国，周涛.个性化推荐系统的研究进展[J].自然科学进展,2009,19(1);1-15.
7. Cao J, Wu Z Wang Y, et al. Hybrid collaborative filtering algorithm for bidirectional Web service recommendation[J]. Knowledge and Information Systems,2013,36(3):607-627。
8. W.-S. Chin, Y. Zhuang, Y.-C. Juan, and C.-J. Lin. A learning-rate schedule for stochastic gradient methods to matrix factorization. In Advances in Knowledge Discovery and Data Mining, pages 442–455. Springer, 2015.
9. J. J. Whang, X. Sui, and I. S. Dhillon. Scalable and memory-efficient clustering of large-scale social networks. In International Conference on Data Mining, pages 705–714, 2012.
10. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In CSCW, 1994.
11. K. Hofmann, A. Schuth, A. Bellogin, and M. De Rijke. Effects of position bias on click-based recommender evaluation. In ECIR’14, pages 624–630. Springer,2014.
12. [Youcong Ni](http://scholar.cnki.net/result.aspx?q=%e4%bd%9c%e8%80%85:(Youcong+Ni)),[Bei Chen](http://scholar.cnki.net/result.aspx?q=%e4%bd%9c%e8%80%85:(Bei+Chen)),[Peng Ye](http://scholar.cnki.net/result.aspx?q=%e4%bd%9c%e8%80%85:(Peng+Ye)),[Chunyan Wang](http://scholar.cnki.net/result.aspx?q=%e4%bd%9c%e8%80%85:(Chunyan+Wang)) A Framework for iOS Application Development[J].Journal of Software,2014 (02):398-403。
13. 郭艳红. 推荐系统的协同过滤其法与应用研究[D]. 大连:大连理工大学,
14. Deshpande M, karypis G. Item-Based top-N recommendation algorithms. ACM Trans. On Information Systems,2004,22(1):143-177.
15. Li G, Li. A new algorithm of cold-start in a collaboration filtering system. Journal of Shandong University (engineering science),2012,2(24):12-17 (in Chinese with English abstract).
16. 梁彦.基于分布式平台Spark和YARN 的数据挖掘算法的并行化研究[D].中山大学,2014
17. 冯琳.集群计算引擎Spark中内存优化研究与实现[D].清华大学，2013
18. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In International conference on World Wide Web, pages 285-295. ACM, 2001.
19. F. M. Harper and J. A. Konstan. The movielens datasets: History and context. ACM Trans. Interact. Intell. Syst., 5:19:1-19:19, Dec. 2015
20. Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems.Computer, (8):30-37, 2009.
21. 项亮.动态推荐系统关键技术研究[D].博士学位论文.中国科学技术大学,2011.
22. Zheng N, Li QD. A recommender system based on tag and time information for social tagging systems. Expert Systems with Applications, 2011,38(4):457-485.
23. 李崇欣. 分布式数据库Hbase: the definitive guide[M]. Sebastopol: O’Reilly Media, Inc. 2015.
24. Huang CG, Yin J, Wang J, Liu YB, Wang JH. Uncertain neighbors collaborative filtering recommendation algorithm. Chinese Journal of Computers,2010,8(33):1370-1377(in Chinese with English abstract).
25. 李改. 一种解决协同过滤系统冷启动问题的新算法[J]. 山东大学学报(工学版),2012
26. T. Tang and P. Winoto. I should not recommend it to you even if you will like it: the ethics of recommender systems. New Review of Hypermedia and Multimedia, 19:111-138, 2016.
27. 徐翔. 协同过滤算法中的相似度优化方法[J]. 计算机工程,2010,36(06):52-56.
28. M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. neural information processing systems, pages 2595-2603,2010.
29. Lops P, De Gemmis M, Semeraro G. Content-based recommender systems: State of the art and trends[M] Springer US,2011:73-102

# 致 谢

首先，感谢我的母校中国科学技术大学给了我读研深造的机会，并且给了我良好的教育。在读研期间，学校的学习氛围以及严谨的治学态度给我留下深刻的印象，同时学校尽可能的为学生提供学习资料与良好的环境，增长了学生的见识，我以我的母校而自豪。

其次，要感谢我的班主任×××老师，感谢他在读研期间对我的生活和学习上的关心及帮助，×××老师在教学上的认真和对学生负责的态度给我留下深刻印象，同时也是我学习的榜样。最重要的是要感谢×××老师作为我的校内导师，感谢×××老师细心指导我的开题报告及论文的撰写，×××老师无论多忙多晚都会耐心为我解决疑惑，同时给予技术上的支持以及论文修改的建议。若没有×××老师的细心指导，论文的完成非常的困难。在这里非常感谢×××老师的指导，×××老师对学术要求严谨，同时又平易近人，给我留下深刻印象，衷心祝愿她永远健康快乐，工作顺利。

再次，要感谢×××公司。感谢×××给了我实习的机会，使我了解了企业工作和学校学习的区别，让我提前进入工作氛围。同时感谢企业导师×××，感谢她在工作中给予我的指导，为我解决了很多工作中的疑惑，同时为我的论文选题以及工作学习提供很多宝贵的建议，感谢她对我的工作和生活的关心。也要感谢同事们在工作上给予的技术指导，使我的技术得到很大提升。

最后，感谢我的家人和朋友，是他们一直在背后默默的支持和鼓励我，在遇到挫折时他们总会为我加油。感谢父母对我的栽培和付出，希望他们永远健康快乐。

××××年××月

**学位论文修改说明**

**意见1：**英文摘要有错误。

修改说明1：针对英文摘要重新做了翻译，纠正错误。详见英文摘要。

**意见2：**存在错字。

修改说明2：通读文章修改错字，如第3页“在加上公司的...”改为“再加上公司的...”。  
**意见3：**总结内容不准确。

修改说明3：重新撰写总结内容，概括本文所做工作。详见第68页7.1节。

**意见4：**英参考文献格式不对。

修改说明4：针对参考文献格式问题做相应调整，如行间距，字体，期刊引用起止页码。详见第82至83页参考文献。

**意见5：**一些流程图不对。

修改说明5：由于流程图中方框只允许有一个出口，第4章概要设计流程图中并行执行的操作流程不对，故将存在问题的流程图改为活动图，能够并行执行某些操作。具体见第4章图4.3，图4.4，图4.5，图4.6，图4.8，图4.11，图4.12。  
**意见6：**在系统实现中，没有系统模块的关联与调用关系。

修改说明6：针对该问题，在第5章添加5.1节系统总体设计与实现，给出系统各模块之间的类图关系，并描述各类作用及调用关系。详见第37至38页5.1节。