**基于用户画像的自适应推荐系统的设计与实现**

软件工程162班 唐稳稳（10161873）

摘要：随着信息化社会的发展，推荐系统成为了解决信息过载的有力工具。本课题将使用用户画像来改善目前推荐系统存在的用户-评分矩阵数据稀疏的问题。设计且实现了基于用户画像的自适应推荐系统。系统的功能包括数据初始化、结合用户操作推荐相关产品等。

关键词：推荐系统，自适应，用户画像

1研究背景与意义

从互联网的诞生到现在的全面普及，数以万计的信息通过不同途径进入人们的视野，但是用户在面对大量的信息时，从中获得自己真正所需要的信息是十分麻烦的。为此，近些年来，推荐系统应运而生，具有巨大市场需求和研究价值。近年来，搜索引擎技术和信息检索技术发展迅速，在一定程度上缓解了信息过载的问题，如Google，Baidu等。然而这些系统并没有考虑单个用户所具有的不同兴趣喜好，检索出来的信息仍存在着大量的冗余信息，无法满足不同用户的不同需求。

随着Internet和信息技术的飞速发展，“信息过载”和“信息迷向”的问题愈发凸显，助推了个性化推荐系统的蓬勃发展[2]。推荐系统起源于互联网的沃土，并在这片土壤中蓬勃发展，尤其在电商、新闻、音乐，电影等领域体现出推荐系统的重要性。自适应推荐系统为用户从大量的数据中过滤掉无用信息，只给用户展示真正需要的信息，大大提高了用户体验以及用户对商家的依赖度。同时可以将用户需求转为商家的利润。目前众多知名网站都使用了不同形式的推荐系统来为自己的用户提供推荐服务，如淘宝，当一个用户检索一类商品时，网站根据用户的特点推荐出适合用户购买的商品，用户往往会在网站的推荐商品中购买到自己心仪的商品。这样，推荐系统给用户和商家提供了极大的便利。再如网易云音乐的推荐系统，它在每天会根据每个用户的音乐喜好形成一份用户特有的歌单[12]，极大的提高了用户对于此应用的粘性。等等，众多的应用程序表明了推荐系统在过滤无用信息、针对每个用户推荐适合自己的互联网资源上发挥的重要的作用，具有极高的使用价值和研究热点。

推荐系统一般基于用户的浏览记录、操作行为以及评分等行为，来推测出用户可能所需要和感兴趣的信息。推荐系统建立在海量的数据挖掘上，这些海量的数据我们可以通过系统的后台日志获取或者可以通过网络爬虫获取。

现阶段的推荐系统主要分为基于内容的推荐、协同过滤推荐、基于关联规则推荐、基于效用推荐以及基于知识的推荐和组合推荐。目前，协同过滤推荐算法是推荐系统中使用最多的算法，本文也将使用协同过滤算法，但是，传统的协同过滤算法会随着数据规模的增大出现数据稀疏性的问题，严重影响了该算法的应用，如淘宝等大型电商平台，用户过亿，项目数量也是大量的，但是也只有少部分用户给个别的产品进行过评级，这就造成了用户项目评分矩阵极端稀疏。所以，在本文中，将用户画像引入到协同过滤算法中，以提高推荐系统准确率。

用户画像是大数据的新兴概念，可将用户信息标签化来获得高度提炼的特征标识，本文将在协同过滤算法中结合用户画像，可在协同过滤给出的评分数据基础上根据用户偏好的附加信息给出推荐项目。

本文将使用MovieLens数据集作为推荐系统的数据来源。MovieLens-1m包含了6040名用户对3952部电影的1000000条评分。同时，还包含了用户的其他信息（性别、年龄、职业、邮编）和电影的信息（电影名、类型）。本文将在这些数据集的基础上构建出一个影片推荐系统。

综上所述，本文基于用户画像的自适应推荐系统可以对海量的用户网络行为进行数据分析，挖掘用户的特点，提取用户相关的特征信息，进行动态分析，让系统更好的了解用户喜好，从而给出自适应的推荐。

2文献综述

2.1 用户画像

用户画像（Users’Profile）这个概念是最早由交互设计之父Alan Cooper所提出的，对同一类型的用户进行不同维度的刻画，从不同维度展现用户的信息全貌。他提出“用户画像是目标用户的一个具体代表”，这句话表明了用户原型是真实用户的虚拟代表，是建立在一系列真实数据之上的目标用户模型[5]，通过研究用户画像，分析用户的行为模式，兴趣爱好，挖掘用户的个性化和潜在需求，有助于开展针对用户的个性化知识资源服务，增强用户的黏度和信誉度。

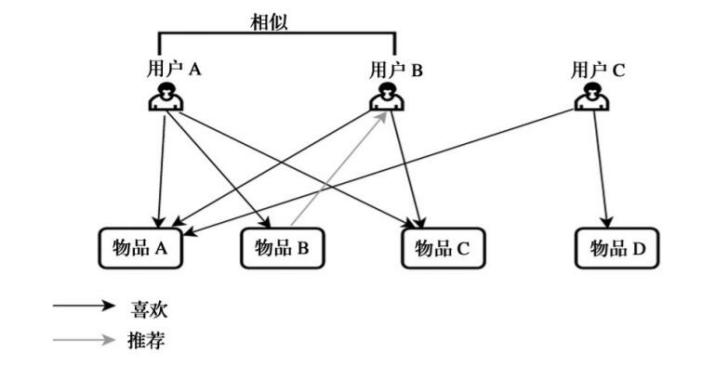
用户画像是任何一个推荐系统中必不可少的一部分[6]，推荐系统可以根据用户画像得出较为准确的用户行为偏好，从而更好的为用户提供推送服务。不同的业务系统同一个用户具有的用户画像是不同的，比如微信中的用户画像偏向于社交，而同一个人在支付宝中则更加偏向于支付，用户画像是由数据依据模型构建而成的。所以我们需要结合具体的业务来制定相关的标签，通过对标签数据的分析来达到预测用户偏好的目的[7]。

李丹、高建忠在基于用户画像的图书馆推荐服务初探一文中[15]展示了如何构建用户画像，通过读者入馆行为数据为读者提供推荐服务。在这篇文章中，讨论了基于用户画像的图书推荐过程：首先通过显式或隐式方式相结合读取读者信息，构建出读者的用户画像；再次将具有相同特征的用户划归为一类，分析用户借阅等行为信息，给出个性化推荐，采用协同过滤算法进行Top\_N推荐，向读者推荐兴趣度排名前N的内容项。

2.2 推荐系统

推荐系统能够有效的解决信息超载的问题，能够在海量的数据上，结合用户之前的行为模式等信息用户信息，为用户推荐其感兴趣的内容。现阶段的推荐系统算法主要包括协同过滤推荐、基于内容的推荐、基于二部图的推荐、基于关联规则的推荐和基于社交网络的推荐[8]，以下将逐一介绍各个推荐算法。

（1）协同过滤算法是由Goldberg等[9]所提出，最初的协同过滤算法用于过滤邮件，经过多年发展，协同过滤算法成为了推荐系统中最重要的算法之一。协同过滤算法利用大量用户与产品之间的关联历史数据来计算用户与产品之间的相似度，查找与用户相似度较高的近邻集，并通过近邻集用户对其他产品的评分来预测目标用户对该产品的预测评分，从而产生推荐的产品集合，再根据Top\_N来产生推荐项。如图2-1所示。可通过计算得知，用户A与用户B相似，用户A的兴趣偏好为物品A、B、C，用户B的兴趣偏好为物品A、C，所以可将物品B推荐给用户B。



**图2-1 协同过滤算法简化过程**

（2）基于内容的推荐算法考虑了协同过滤算法中忽略的用户信息（如年龄、性别等）和产品信息。基于内容的推荐算法通过提取用户-产品特征，学习用户兴趣模型，考量用户信息与待选产品间的匹配度，再将匹配度最高的产品推荐给用户[10]。基于内容的推荐算法第一步是建立起推荐对象模型，然后建立用户兴趣模型，再计算推荐对象模型和用户兴趣模型的相似度，最后按照匹配程度的高低进行相关产品的推荐。

（3）基于二部图的推荐算法有Aggarwal由1992年提出[11]，基于二部图的推荐算法只关注用户是否选择了某项，而不关心用户与产品是何种形式。在二部图算法中，用户与产品都被看作为二部图的节点，若用户选择了某产品，用户与某产品之间则存在一条边。所以通过研究节点之间的相关性，可以得出用户感兴趣的产品。

（4）基于关联规则的推荐算法的核心是关注用户的行为数据，从大量的数据中获取潜在有用的关联规则，从而产生推荐产品。关联规则的推荐算法广泛应用于电商领域，它通过用户的操作日志分析用户的每次购买之间的联系，形成用户的购买模式，当用户在进行购买活动时，可进行相关的推荐，达到交叉销售的目的。现阶段，基于关联的推荐算法主要使用Apriori算法以及FP-Grouwth算法，这类算法具有不可避免的缺点：需要对数据集进行频繁的扫描，计算量大，耗时长。同样也具有冷启动和数据稀疏的问题。

（5）基于社交网络的推荐算法利用社交网络数据得到用户兴趣偏好及好友信息，再根据获取到的数据为用户进行产品推荐、好友推荐及会话推荐。

对推荐系统的评价指标包括：平均绝对误差、均分根误差、标准平均误差、召回率、准确率。平均绝对误差用来衡量用户预测评级与实际评级的平均绝对误差；召回率表示推荐列表中用户实际感兴趣的产品占系统中用户感兴趣的所有产品的占比；准确率表示推荐列表中用户感兴趣的产品在所有被推荐的产品中的所占比例。

3技术路线

3.1 技术路线概要

该课题主要采用前后端分离的技术，前端负责用户操作日志采集，推荐可视化等功能，后端主要提供算法支持，Api提供等。

所用的编程语言主要有Golang、Python、JavaScript、Sql、Html等；主要采用Mysql和Redis作为系统数据库；采用的前端框架为VueJs。

其他的后端技术有Kafka、gRPC、Docker等。这些技术应用细节将在下文中详细介绍。

3.2 编程语言

本次项目的编程语言选择了Golang和Python。Golang的轻便以及高并发使得它在过去十年里飞速发展，Golang由Google公司开发完成，开发理念为兼顾开发效率、编译效率以及运行效率，被誉为云计算时代的C语言，结构简单，易学易用。Golang的高性能HTTP Server提供了对Http协议良好的支持，可以使用几行代码开启一个Http服务，十分适合于本项目的后端服务器实现[1]。而使用Python的原因在于Python在科学计算方面有着其他编程语言无法达到的效率和优势，且Python是人工智能领域的首选语言，有着很多性能优越的第三方库供我们选择使用，在本项目中，核心算法部分将由Python完成，再使用RPC的方法与Golang服务器进行通信。所以，在本次项目中，我将使用Golang和Python完成这次的推荐系统项目。

3.3 数据库

本次项目中，将使用Mysql数据库作为主要的数据存储数据库，在伍志聪发表的《Mysql数据库在中小型业务系统的应用》[3]中，Mysql的数据类型、与系统的兼容性以及安全性都相较于其他数据库有较大的优势。Mysql由一个mysqld服务器守护进程和客户程序及库组成，提供了适合于众多编程语言的Api接口，且多平台支持，支持多线程、多用户。

将Redis作为用户Token数据库以及数据缓存。Redis是目前应用十分广泛的键值对数据库，逐渐取代了Memcache成为现在系统中首选的缓存服务器，它支持事务、I/O持久化和订阅发布功能，都对系统数据实时存储提供强有力帮助[6]。在本次课题中，主要是依靠Redis具有的告诉读写，纯内存操作等性能，使用Redis存储用户登录时产生的Token，保证用户的持久化登录，提高用户操作友好度。

3.4 前端框架

本次项目将使用Vue.js作为前端框架。Vue.js是一个轻巧、高性能、可组件化的前端框架，同时拥有者非常容易上手的API，Vue.js不同于其他重量级的框架，它只关注视图层，采用自底向上增量开发的设计，并且学习成本低，非常容易学习，可以很轻松的将其它库与已有项目整合。Vue.js的灵活与渐进式的特点都适合本次项目，因为不管我们最后的前端项目做到何种程度，在代码方面都易于管理。Vue.js提供NPM来安装依赖。Vue.js有着活跃的社区，促进框架的发展，因此学习Vue.js的途径也变得多种多样。

Vue.js采用了MVVC的设计模式。MVC模式是传统软件工程中的一种设计模式，它将项目进行了逻辑层面的划分，其中MVC分别是Model、View和Controller首字母的缩写。View指软件的视图界面，Model指视图中所需要的数据，Controller控制View和Model之间的数据传输。随着互联网的发展，研究人员在MVC的基础上衍生除了MVVM的开发模式。MVVM指Model View ViewModel。重点在于用事件驱动的UI平台开发。MVVM的核心是通过数据绑定链接View和Model，使得数据的变化自动映射为视图的更新。Vue.js对页面中的每个Dom操作都进行了封装工作，让数据发生变化的Dom自动动态做出改变，简化了系统开发的投入，降低了开发难度和日后维护成本[16]。

3.5 核心算法

本项目的核心算法部分将由Python完成。本项目将大量使用Python的一个扩展程序库Numpy。Numpy支持大量的维度数组与矩阵运算，Numpy底层使用C语言编写，数组中直接存储对象而不是对象指针，所有其运行的效率远远高于Python代码。不仅如此，Numpy内置了并行计算功能，当系统有多个核心是，Numpy会自动的进行并行计算，大大提高了运行效率。Numpy的特点，运算效率高以及编码难度低都适合在推荐系统核心算法中使用。

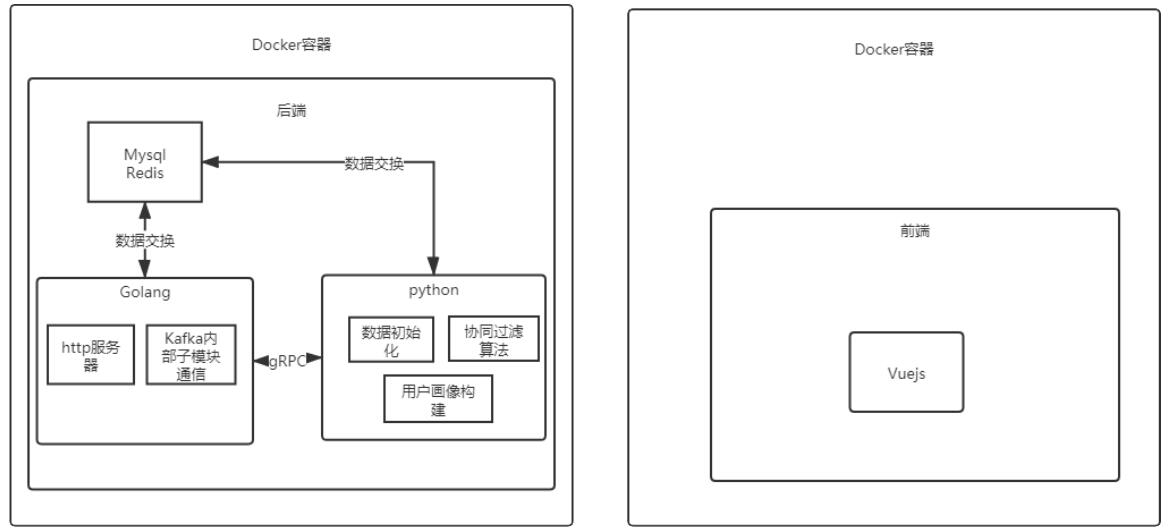
3.6 后端服务器

后端服务器将由Golang实现。Golang原生支持并发编程，可以十分简单的开启并发执行单元，开发人员无需了解执行细节，调度器会自动将这些任务分配给合适的线程执行。Golang在创建时就具有了Python的编码高效、C语言的编译迅速以及C/C++的运行高效。它是编译型的静态语言，相比于解释型语言具有天生的性能优势。

在本项目中还将使用消息队列Kafka来将用户（Producer）与主题（Topic）进行绑定（Subscribe），使得用户数据可以自适应的被推荐系统所接受。同时，Kafka还具有消息的持久化功能，保障了数据传输的稳定性。本课题使用Kafka代替传统的消息代理，Kafka有着更高的数据吞吐量、高性能以及容错机制，都让Kafka成为了大规模消息处理应用软件系统的理想解决方案。之所以Kafka具有高性能的数据传输表现，主要是因为Kafka在发送少量数据时，会将数据暂时缓存在内存中，当超过一定的大小或一定时间后再进行批量发送，同时Kafka还会将数据进行压缩以减小网络负载。在本课题中，使用Kafka传输用户的操作日志给算法程序形成用户画像，分析用户兴趣，再通过Kafka返回推荐项目，取代了传统的http方式，使得数据不易丢失，提高性能[13]。

本课题中，后端系统将会放入Docker容器中进行发布。Docker区别于传统的虚拟机，传统的虚拟机运行在寄主操作系统之上，然后通过对CPU、内存、IO存储设备等进行虚拟化来达到资源隔离和调度的目的。而Docker使用Namespace来进行系统环境的隔离，加上CGroups实现资源的限制，使用了和寄主机相同的内核。Docker轻量化、启动速度快、运行效率高的特点都让Docker成为取代传统虚拟机的条件。使用Docker可以快速的搭建自己想要的环境，快速构建出适合本课题的环境，并且可以打包成镜像文件，方便在寄主机间进行转移 [14]。

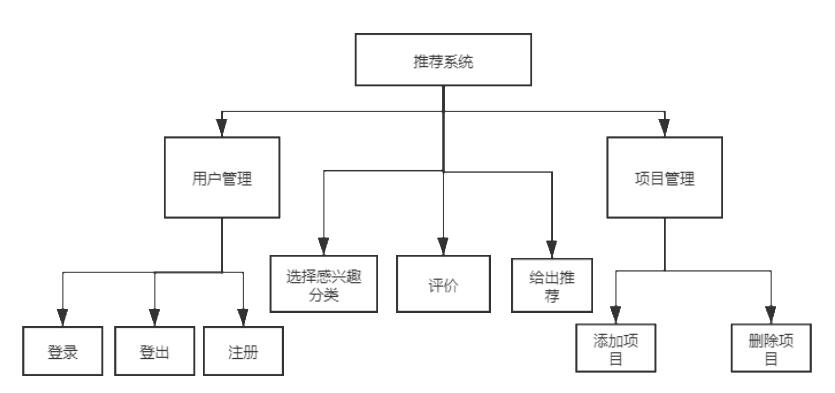
3.7 架构设计



**图3-1 架构图**

本项目的系统架构图由图3-1所示。前后端分别放入Docker容器中进行管理编排。Golang代码负责提供http服务并实现接口部分，内部以微服务进行划分子模块，其子模块间采用Kakfa进行通信。Python代码部分负责初始化数据集，实现协同过滤算法与用户画像的结合。Python代码与Golang代码均可直接与数据库进行数据交换。前端主要以Vuejs为主，负责请求发送数据以及接受响应并动态渲染数据。

3.7 系统设计



**图3-2 系统设计图**

本项目的系统设计图由图3-2所示。系统主要分为用户管理、项目管理两部分。用户可登陆、登出以及注册新用户。此外用户还可以添加新的项目，并给予评价。用户在首次进入系统时要求选择感兴趣的分类，以便解决推荐系统中存在的冷启动问题。由此随着用户给出的评价越多，推荐的准确率也将提高。只有管理员用户才可以删除项目。

4进度安排

4.1 学习安排

项目过程中，将会使用到很多课堂上未获取的知识，在完成项目进行的过程中，需要不断的去学习新的技术，加强自己的编程水平，最终出色的完成毕业设计的所有要求。

4.2 进度安排

本次毕业设计进度安排如下：

2019.12-2020.02研读参考文献、理清实现路线、完成开题报告和文献翻译工作

2020.02-2020.04完成推荐系统前后端搭建以及核心算法实现

2020.04-2020.05完善系统功能

2020.05-2020.06完成毕业设计所有工作，撰写毕业论文、准备毕业设计答辩

5参考文献

[1] 曹廷,付宇.Golang编程语言特性分析[J].电脑迷,2018,11:55-56.

[2] 石雨晴. 基于中图—专业分类映射的高校图书馆图书推荐方法的研究与实现[D].辽宁大学,2017.

[3] 伍志聪.MySQL数据库在中小型业务系统的应用[J].数字技术与应用,2011(11):122.

[4] 刘俊龙,刘光明,张黛,喻杰.基于Redis的海量互联网小文件实时存储与索引策略研究[J].计算机研究与发展,2015,52(S2):148-154.

[5] COOPER A. The Inmates are Running the Asylum [M][. S.l.]:Macmillam Pulihing co.lnc,1999,53:17.

[6] Leung K ,Lee D.Deriving Concept- Based User Profiles Fro Search Engine Logs[J]. IEEE Trans. Knowledge And Data Engg., 2010, 22(7).

[7] 臧勇琪.基于用户画像的精准营销研究[J].全国流通经济,2019(25):14-16.

[8] 崔春生,吴祁宗,王莹.用于推荐系统聚类分析的用户兴趣度研究[J].计算机工程与应用,2011,47(7):226-228.

[9] GOLDBERG D,NICOLS D. Using collaborative filtering to weave an information tapestry[J].Communications of the ACM,1992,35(12):61-70

[10] 常璐.高校图书馆 E-learning支持服务研究[D].上海:东华大学,2013.

[11] 蔡红蕾.二部图网络结构算法在推荐系统中的应用[D].秦皇岛:燕山大学,2014.

[12] 李兰馨.大数据智能算法范式下的用户黏性研究——以网易云音乐为例[J].新媒体研究,2019,5(04):4-6.

[13]刘邦,余华平.Kafka分布式消息队列的高性能研究[J].电脑知识与技术,2019,15(32):4-6.

[14] 赵冠臣,王冬妮,刘至洋,孟振江.浅谈Docker容器技术[J].有线电视技术,2019(09):85-88.

[15] 李丹,高建忠.基于用户画像的图书馆推荐服务初探[J].图书馆,2019(07):66-71.

[16] 徐頔,朱广华,贾瑶.基于VueJs的WEB前端开发研究[J].科技风,2017(14):69.