**中国科学技术大学软件学院**

**工程实践**

**项目技术论文**

**项 目 名 称：基于 Spark 的景点推荐系统**

**成 员 名 单： 褚思源 林靖清 王炜康**

**导 师： 袁宇丹**

**工 程 领 域： 推荐系统**

**研 究 方 向： web 开发、个性推荐**

摘要

随着因特网的快速发展，用户可以享受更加便利地服务。同时，面对海量数据，精准地获得目标信息也变得更加困难。一方面，用户面临对大量的信息束手无策，经常会迷失在大量的信息空间中；另一方面，网站也失去了与用户的联系，不能与用户建立稳固的合作关系。在这种背景下推荐系统（Recommender Systems）应运而生，它根据用户的趣爱好推荐符合用户兴趣爱好的对象。

本课题计划开发一个景点推荐系统。系统是一个基于 SSM 框架的 web 项目，通过 Scrapy 框架爬取景点数据，并通过 Spark 引擎进行数据处理、实现推荐算法。

其中后台管理系统对用户信息和景点信息进行管理，如添加删除景点信息和完成用户信息的完善。推荐系统通过在网站系统获取到用户的点击事件（如用户对某各景点的评分、用户的对某个景点详情页的浏览次数）记录信息，并根据该信息做出相应的处理，将推荐结果存入到数据库中，web 前端通过查询数据库将推荐的景点展示给用户。

**关键词**：爬虫，web 开发，个性推荐

Abstract

With the rapid development of Internet, users can enjoy more convenient services. At the same time, in the face of massive data, it is more difficult to obtain target information accurately. On the one hand, users are at a loss to deal with a large amount of information and often get lost in a large amount of information space; on the other hand, the website also loses contact with users and cannot establish a stable cooperative relationship with users. In this context, recommender system arises. It recommends the objects according to the interests of users.

This project plans to develop a scenic spot recommender system. The system is a web project based on SSM framework. It crawls the data of scenic spots through scrapy framework, processes the data through spark engine, and realizes the recommendation algorithm.

The background management system manages user information and scenic spot information, such as adding and deleting scenic spot information and completing the perfection of user information. The recommender system obtains the user's click event (such as the user's rating for each scenic spot, the user's browsing times for the details page of a scenic spot) record information in the website system, and makes corresponding processing according to the information, stores the recommendation in the database, and the web front end displays the recommended scenic spot to the user through the query database.

**Keyword：crawler, web development, personalized recommendation**

目录

[摘要 2](#_Toc46521041)

[Abstract 2](#_Toc46521042)

[第一章 绪论 5](#_Toc46521043)

[1.1 研究背景及意义 5](#_Toc46521044)

[1.2 研究现状 5](#_Toc46521045)

[第二章 推荐技术概述 5](#_Toc46521046)

[第三章 系统设计 6](#_Toc46521047)

[3.1 需求分析 6](#_Toc46521048)

[3.2 概要设计 8](#_Toc46521049)

[3.3 详细设计 9](#_Toc46521050)

[第四章 参考文献 12](#_Toc46521051)

# 绪论

* 1. 研究背景及意义

随着 Internet 的发展，英特网为用户提供越来越多信息和服务的同时，其结构变得更加复杂，在网络上的海量信息中，用户及时地获得所需要的信息也更加困难。一方面，用户面临对大量的信息束手无策，经常会迷失在大量的信息空间中；另一方面，网站也失去了与用户的联系，不能与用户建立稳固的合作关系。在这种背景下推荐系统（Recommender Systems）应运而生，它根据用户的趣爱好推荐符合用户兴趣爱好的对象。推荐系统模拟商店销售人员向用户提供商品推荐，不仅帮助用户找到所需信息，而且通过推荐系统，网站可以有效保留用户，提高网站的点击率和用户的忠诚度。

基于以上原因，从 20 世纪 90 年代中期推荐系统的概念出现至今，推荐系统在近几年得到了飞速发展，在电子商务、影音娱乐方面有着广泛应用。

* 1. 研究现状

推荐系统是信息检索和信息过滤领域的研究热点。除各大科技公司外，许多研究机构和研究者也开发出各种研究型推荐系统实例。比如：

* GroupLens：自动协同过滤推荐系统，用于新闻组信息荐。GroupLens 系统通过用户的评分信息自动搜索用户的最近邻，根据评分信息产生最终的推荐结果，适合于用户数量比较大的场合。
* MovieLens：Minnesota 大学开发的研究型自动协同过滤推荐系统，用于推荐电影。与 GroupLens 不同，它是一个基于 web 的推荐，系统通过浏览器方式进行用户评分数据收集与推荐结果显示。
* Ringo：由 MIT 媒体实验室开发的研究型协同过滤推荐系统，用于提供化的音乐推荐服务。Ringo 系统可以向用户推荐他们喜欢的音乐，预测用户不喜欢的音乐，也可以预测用户对特定音乐的评分。

综上所述，推荐系统技术大致可以划分为三类：协同过滤推荐（Collaborative Filtering Recommendation）、基于内容推荐（Content-based Recommendation）和混合推荐三种。

# 推荐技术概述

* 基于用户行为推荐：基于用户行为的推荐算法也称为协同过滤算法，是推荐领域应用最广泛的算法，该算法不需要预先获得用户或物品的特征数据，仅依赖于用户的历史行为数据对用户进行建模，从而为用户进行推荐。
* 基于内容推荐：基于内容推荐的基本思想是为用户推荐与他感兴趣的内容相似的物品，比如用户喜欢励志类电影，那么系统会直接为他推荐《阿甘正传》这部电影，这个过程综合考虑了用户兴趣和电影内容，因此不需要提供用户的历史行为数据，这能够很好地解决新用户的冷启动问题。
* 混合推荐：实际应用的推荐系统通常都会使用多种推荐算法，比如使用基于内容或标签的推荐算法来解决新用户的冷启动问题和行为数据稀疏问题，在拥有了一定的用户行为数据后，根据业务场景的需要综合使用UserCF、ItemCF、矩阵分解或其他推荐算法进行离线计算和模型训练。

# 系统设计

* 1. 需求分析

游客出行之前往往会花费大量时间在网络上搜寻景点信息，被各种文章“种草”、“拔草”，这通常会使游客举棋不定。而本工程实践是一个基于 web 的景点推荐系统，用户只需在系统中对他们已有印象的景点进行打分，系统便可推荐其他适合其口味的景点。以下为用户用例图：

图片包含 游戏机

描述已自动生成

网站管理员拥有对景点和用户的增删查改权限。以下为管理员用例图：



* 1. 概要设计

本系统主要由三部分构成：爬虫，web 以及推荐系统。

景点数据来源以爬虫的形式获取，来源网站为去哪儿网和携程网。图示为景点数据的字段：

图片包含 游戏机, 文字, 收据

描述已自动生成

系统是一个基于 SSM 框架的 web 项目。后台管理系统主要对用户信息和景点信息进行管理，如添加删除景点信息和完成用户信息的完善。

推荐系统通过在网站系统获取到用户的点击事件（如用户对某各景点的评分、用户的对某个景点详情页的浏览次数）记录信息，并根据根据该信息做出相应的处理，将推荐结果存入到数据库中，web 前端通过查询数据库将推荐的景点展示给用户。推荐算法使用 ALS 算法（使用最小二乘法求解的协同过滤算法）。

开发环境如下：

**爬虫**

数据库：MongoDB、Redis

框架： Scrapy，使用 scrapy\_redis 和 Redis 实现分布式

**网站**

开发环境： IntelliJ IDEA、maven、git

软件架构： mysql、mybatis、spring、springmvc

**后台管理系统**

开发环境： IntelliJ IDEA、maven、git

软件架构： mysql、mybatis、spring、springmvc

**推荐系统**

开发环境： IntelliJ IDEA、maven、git

软件架构： spark、mysql

* 1. 详细设计
     1. **爬虫**

爬虫使用 Scrapy-Redis 搭建分布式架构，包括一个 Master 端和两个 Slave 端，Master 端负责 url 指纹去重、Request 分配和数据存储，Slave 负责页面数据抓取。

图片包含 游戏机, 截图, 文字

描述已自动生成

Scrapy 框架包括 ScrapyEngine、Scheduler、Downloader、Spiders、ItemPipeline 等组件。调度器负责把请求转发给下载器，再由下载器下载页面给 Spiders 处理。Spiders 通过 xpath 解析网页得到数据后交给管道处理，最终存到数据库中。

图片包含 游戏机, 截图

描述已自动生成

并修改 Scrapy 中间件，添加 IP 动态代理，每个请求都会从 IP 池中随机挑选一个 IP，同时伪装浏览器标识，再发起请求，能够有效防止请求被拦截。

* + 1. **网站**

网站由 Vue + SSM 框架实现。其整体结构可分为表现层、接口层、服务层、计算层和存储层这五部分。用户通过表现层发出请求后，会调用接口层的接口来处理此请求，具体的业务逻辑处理由服务层提供。计算层由 Spark 提供离线计算，它是通过收集到的用户评分数据利用 ALS 推荐算法预测用户可能感兴趣的景点来为用户提供推荐服务。存储层提供数据的存储，数据库中存放着管理员信息、用户信息、景点信息、评论信、评分等信息。

**图片包含 截图, 游戏机

描述已自动生成**

在数据库设计中使用 E-R 图来描述整个数据库的模型，本系统主要包含的实体有：景点、用户、管理员和评论信息等 。

**图片包含 游戏机, 文字, 地图

描述已自动生成**

* + 1. **推荐**

网站推荐算法使用 ALS 算法。ALS 特指使用最小二乘法求解的一个协同过滤算法，是协同过滤中的一种。ALS 算法是 2008 年以来，用的比较多的协同过滤算法。它已经集成到 Spark 的 Mllib 库中，使用起来比较方便。从协同过滤的分类来说，ALS 算法属于 User-Item CF，也叫做混合 CF，因为它同时考虑了 User 和 Item 两个方面，即既可基于用户进行推荐又可基于物品进行推荐。

图片包含 小, 水, 船, 侧面

描述已自动生成

我们假设有矩阵 （用户 – 商品）。ALS 可以将用户（user）对商品（item）的评分矩阵，分解为用户对商品隐含特征的偏好矩阵，和商品在隐含特征上的映射矩阵。我们可以假设用户和商品之间存在若干关联维度（比如用户年龄、性别、受教育程度和商品的外观、价格等），我们只需要将 矩阵投射到这些维度上即可。

ALS 算法的也存在一些缺点：

1. 它是一个离线算法

2. 无法准确评估新加入的用户或商品。这个问题也被称为 Cold Start 问题

网站为单个用户做景点推荐的逻辑如下图：

图片包含 游戏机

描述已自动生成

# 参考文献

[1] Davidson J, Liebald B, Liu J, et al. The YouTube video recommendation system[C]//Proceedings of the fourth ACM conference on Recommender systems. 2010: 293-296.

[2] 蒲姗姗,何燕.个性化学术资源推荐研究:现状、特点及展望[J].图书馆学研究,2019(16):9-17.

[3] 刘鲁,任晓丽.推荐系统研究进展及展望[J].信息系统学报,2008(01):82-90.

[4] 周俊,董晓蕾,曹珍富.推荐系统的隐私保护研究进展[J/OL].计算机研究与发展,2019(10):1-16[2019-11-15].http://kns.cnki.net/kcms/detail/11.1777.TP.20191025.1706.006.html

[5] McDonald D W, Ackerman M S. Expertise recommender: a flexible recommendation system and architecture[C]//Proceedings of the 2000 ACM conference on Computer supported cooperative work. 2000: 231-240.

[6] Errico J H, Sezan M I, Borden G R, et al. Collaborative recommendation system: U.S. Patent 8,949,899[P]. 2015-2-3.

[7] Shaya S A, Matheson N, Singarayar J A, et al. Intelligent performance-based product recommendation system: U.S. Patent 7,809,601[P]. 2010-10-5.

[8] Park M H, Hong J H, Cho S B. Location-based recommendation system using bayesian user’s preference model in mobile devices[C]//International conference on ubiquitous intelligence and computing. Springer, Berlin, Heidelberg, 2007: 1130-1139.

[9] Wei J, He J, Chen K, et al. Collaborative filtering and deep learning based recommendation system for cold start items[J]. Expert Systems with Applications, 2017, 69: 29-39.