**中国科学技术大学软件学院**

**软件工程实验项目环节**

**开题报告**

**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**

**开 题 时 间： 2019.10.29**

**中国科学技术大学软件学院**

**填表日期：2019 年 11 月 1 日**

**一、 简况**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **中文** | **基于 Spark 的景点推荐系统** | | | |
| **英文** | **A spot recommender system based on Spark** | | | |
| **项目组成员名单** | **姓名** | | **学号** | 项目中的分工 | 签 章 |
| **林靖清** | | **SA19225242** | **主要负责推荐系统** |  |
| **王炜康** | |  | **主要负责爬虫** |  |
| **褚思源** | | **SA19225123** | **主要负责网站** |  |
|  | |  |  |  |
| **中 英 文 摘 要** | **（中文 400 字左右）**  **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。**  **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.** | | | | |
| **主题词** | **主题词数量不多于三个，主题词之间空一格（英文用“/ ”分隔）** | | | | |
| **中文** | **Spark 引擎 景点推荐** | | | |
| **英文** | **Spark/Spot recommendation** | | | |

**二、选题依据**

1. **阐述该选题的研究意义，或工程设计的价值和意义，国内外概况和发展趋势，选题的先进性和实用性，技术难度及工作量。**

**随着 Internet 的发展，因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。**

**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。**

**基于以上原因，我们计划开发一个景点推荐系统。系统以网站的方式呈现，通过 Scrapy 框架爬取数据，并通过 Spark 引擎进行数据处理、实现推荐算法。**

1. **参考文献。**

**[1]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.**

**[2]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.**

**[3]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.**

**三、课题内容及具体方案**

|  |
| --- |
| **1．课题内容**  **本课题计划开发一个景点推荐系统。系统以网站的方式呈现，通过 Scrapy 框架爬取数据，并通过 Spark 引擎进行数据处理、实现推荐算法。用户可在网站上浏览景点，对景点打分并获取景点推荐；管理员拥有对用户和景点信息进行增、删、查、改的权限。**   1. **系统需求分析**   **在网站中，用户只需在系统中对他们已有印象的景点进行打分，系统便可推荐其他景点。以下为用户用例图：**    **以下为管理员用例图：**     1. **系统概要设计**   **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（https://www.qunar.com/）和携程（https://www.ctrip.com/），数据字段包括省、市、景点名称、图片 url、景点地点等等。**  **网站是一个基于 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）。**   1. **拟采用的开发方法、环境，测试方案等**   **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. **技术难度及特色分析**   **此课题的难点与特色为推荐系统的实时推荐和离线计算。离线计算使用基于物品 ALS 算法，数据来源为会员注册时选择喜欢的景点，收藏的景点；实时推荐的数据来源为用户对景点的评分情况。** |

**四、工作进度的大致安排**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 应包括文献调研，理论分析，数值计算，理论分析，软硬件设计，仪器设备的研制和调试，撰写结题报告、技术论文等，要给出各个阶段的成果形式。   |  |  | | --- | --- | | 起止时间 | 内容 | | 2019 年 10 月 – 2019 年 11 月 | 熟悉课题，阅读相关文献，了解推荐系统现状，原理以及应用 | | 2019 年 11 月上旬和中旬 | 完成开题报告 | | 2019 年 12 月 | 系统开发 | | 2020 年 3 月 – 2020 年 4 月 | 系统开发基本完成 | | 2020 年 5 月上旬 | 论文初稿完成，系统验收 | | 2020 年 5 月 | 答辩 | | |
| **预期成果及特色**  **预期成果截图参考 MovieLens 网站。**  **网页主页面：**    **景点详情页：** | |
| **导师意见（对选题和工作过程及成果进行说明，并给出成绩。）**  **导师签名：**  **年 月 日** | |
| **答辩小组意见**  **答辩组长签名：**  **年 月 日** | |