

**内蒙古师范大学计算机科学技术学院**

**毕业设计（论文）开题报告**

**题 目： 热门城市旅游酒店的分析与推荐**

**专 业： 信息管理与信息系统（云计算大数据）**

**学 生 陈爽**

**学 号 20161104552**

**指导教师 苗忠义**

**日 期 2019年11月22日**

1. 课题来源及研究的目的和意义

旅游业是当今世界发展最快、前景最广的新兴产业之一，也是国家政策重点扶持和鼓励发展的产业。近年来越来越多的人喜欢采取自由行方式外出游玩，传统的旅行社模式，由旅行社统一规划团队的住宿安排，基本不允许游客自己选择酒店，这与游客对旅游的个性化需求背道而驰，因此自由行成为了更多游客的选择。自由行的方式让游客可以根据需要对酒店进行安排，但旅游地往往不是自己所熟悉的地方，因此对酒店的相关信息不会太了解，需要花费大量的时间精力去查询[[1]](#footnote-1)。

为了制订合适的住宿方案，游客们常通过各大旅行社和互联网来获取信息。但是旅行社所能提供的信息比较单一，都是通用化的方案，网络上的信息又太过庞大。为解决这一难题，我通过对相关数据的可视化分析，构建热门酒店模型，进行智能酒店推荐，极大的方便了游客的住宿选择。

1. 国内外在该方向的研究现状及分析

旅游电子商务于二十世纪九十年代初开始出现在欧美等发达国家。随着互联网技术的不断发展，旅游电子商务的应用模式和理论体系不断完善，给社会、企业和个人带来重大的影响。近几年，网上旅游份额增长迅猛。艾瑞咨询研究发现，在2009年，进行过在线预定的用户达2750万，同比增长95. 0 %，在线旅游预定市场规模达38 . 9亿元，同比增长32 . 3%。按照这种增长速度，到2013年，进行在线预定的用户人数将达到16800万。从营收结构上看，网上旅游预定市场主要以网上酒店预定业务为主，2009年网上酒店预定市场规模占整个在线预定市场规模的45.4%[[2]](#footnote-2)。

信息技术和电子商务对于各行各业的重要性已经不言而喻了，酒店行业也不例外。传统的企业在拥有自己的网站后，将更多的精力逐渐投入到网站功能的优化上，提高用户网络预定体验。一些在线代理商，如携程、艺龙等，则提供用户对酒店的评论和打分，借助 SNS的影响，扩大网站和酒店业务的影响。酒店提供的是一种体验式的服务。因此，消费者喜欢将自己的消费体验写在论坛、旅游博客中，而这些信息就是旅客的真实体验，与体验相关的这些评论的可信度较高，基本能够反映商家的服务质量，对那些潜在的预订者来说，这些评论具有很好的参考价值。相关研究[[3]](#footnote-3)表明，大部分在线酒店的访问者在预定酒店之前都会参考网站上的评论信息。在这些访问者中，84%的人在选择酒店时都会受到他们所看到的评论信息的影响。由此可见，酒店电子商务的推荐服务应该更加重视用户的评论和体验。然而遗憾的是，国内的酒店电子商务网站主要是通过酒店基本信息(包括星级、房型、价格等) 的介绍、酒店图片和酒店的综合排名等信息将酒店推向网民和客户，无法根据用户的个人体验和喜好向用户推荐满足其兴趣的酒店，也谈不上是个性化的服务。因此，研究用户的个性化体验打分，结合酒店的特征，向用户提供个性化的推荐，从而提高浏览者的对推荐信息的满意度。

1. 主要研究内容

推荐系统在电子商务领域已经得到广泛应用，如电影推荐、视屏推荐、书本推荐等等，然而在对于酒店的推荐却很少有学者研究。并且现有的推荐算法往往只用到用户对商品的总体打分，虽然总体打分在一定程度上能表示用户的兴趣爱好，却忽略了用户对商品的关注面是不同的。因此，总体打分无法最为准确的表示出用户的真实兴趣，也就无法构建准确的用户模型，也就无法为用户提供更为精确的推荐。而传统的协同过滤算法正是基于用户的单个评分，主要结合携程网上用户对酒店多维度的打分，在传统协同过滤算法的基础上进行改进，提出基于酒店多维度属性打分的推荐算法。改进方法如下：

（1）是直接对传统的协同过滤算法中相似性的计算进行改进。

（2）是相关方法计算每个用户对酒店不同属性的打分权重，然后根据打分和对应权重得到一个综合打分，然后利用传统的协同过滤算法进行酒店推荐。

1. 研究方案

进行研究主要进行以下几个步骤：

（一）数据爬取

通过python编写爬虫脚本对携程酒店网站进行爬取

（二）数据处理

在Hadoop集群下通过Mapreduce对数据进行清洗、转化、计算、合并和抽取等一系

列处理，得到可以方便使用的数据。

（三）数据探索

引用pyecharts库，绘制热门地区热门酒店散点图，了解数据的结构和特征分布。

（四）数据建模

尝试用多元线性回归及决策树的算法模型进行拟合，用多种模型进行测试，对比模型的

准确度

（五）数据展示

将数据模型拟合出来的未来趋势在Web端用图形化展示，并且利用其展示出来的内容特点提出建议及总结

1. 进度安排，预期达到的目标

**第一阶段（预计11.18-12.31）：**阅读相关书籍，查阅文献，了解课题的研究状况，学习线性回归、决策树算法等

**第二阶段（预计1.2-1.10）：**使用Requests爬取热门城市酒店数据，通过BeautifulSoup方法进行解析，分析出热门城市的热门酒店。

**第三阶段（预计1.11-2.11）**：搭建Hadoop高可用集群，将找到的数据放到Hadoop集群下Mapreduce进行数据清洗，构建线性回归模型，使用决策树方法对数据进行验证。

**第四阶段（预计2.12-2.20）：** 引用pyecharts库，绘制热门地区热门酒店散点图并进行展示。

**第五阶段（预计2.21-3.15）：**论文撰写

1. 课题已具备和所需的条件、经费

系统环境：Windows操作系统、Linux操作系统

研究工具：Pycharm、Excel、Xshell、Idea

相关技术：爬虫：（Requests、BeautifulSoup、fake\_useragent）

数据处理：（Hadoop集群、Zookeeper框架、MapReduce）

数据可视化：（echarts、pyecharts、matplotlib、wordcloud）

相关算法：多元线性回归、决策树

1. 研究过程中可能遇到的困难和问题，解决的措施
2. 相关的建模算法用Python实现困难

通过阅读相关博客，查看建模算法的使用流程，理解原理，然后实现

1. 选择哪种算法进行建模比较合适

阅读相关论文，寻找具有相似性的内容，使用其使用的方法进行测试

1. 挖掘的维度比较局限

尝试在时间的基础上，通过资料的研究发现更多的维度，进行多方面的探索分析

1. 数据指标较多，难以提取的精确扼要

通过深入学习数据处理的相关方法，来实现相关指标的提取

5. Hadoop集群搭建配置过程相对复杂

认真学习Hadoop集群搭建及Linux系统操作

1. 主要参考文献

[1]魏宇.慢旅游与云旅游的对接——新型自由行与半自由行旅游 模式的构建[J].中国外资,2011(16):117-117.

[2]IResearch. 2009-2010年中国在线旅行预订行业发展告[R/OL]. http://www.iresearch.cn.

[3]Fact sheet[OL].2006. http://www.tripadvisor.com/pages/factsheet.html

1. 魏宇.慢旅游与云旅游的对接——新型自由行与半自由行旅游 模式的构建[J].中国外资,2011(16):117-117. [↑](#footnote-ref-1)
2. IResearch. 2009-2010年中国在线旅行预订行业发展告[R/OL]. http://www.iresearch.cn. [↑](#footnote-ref-2)
3. Fact sheet[OL].2006. http://www.tripadvisor.com/pages/factsheet.html [↑](#footnote-ref-3)