

**毕业设计说明书**

**题目：《交通旅游排行数据分析》**

|  |  |
| --- | --- |
| 系 部： | 大数据与电子商务系 |
| 专 业： | 大数据技术与应用 |
| 班 级： | 2020大数据技术与应用（5）班 |
| 学 号： | 202003170519 |
| 姓 名： | 何儒林 |
| 指导教师： | 付颖 |

二〇 年 月 日

论文题目：《交通旅游排行数据分析》

专 业：大数据技术与应用

姓 名：何儒林 （签名） 何儒林

指导教师：付颖 （签名）

摘要

旅游成为人们假期的选择，而交通是旅游的重要组成部分，是旅游活动中的主要时间消耗，旅游的异地性决定了交通是实现旅游的必要手段，交通的便利、全、以及完善程度对旅游者行为以及旅游体验的整体满意度都会产生影响。

交通旅游排行网站的出现让人们足不出乎户便可以知晓想去地方的交通情况，交通旅游网站的排名，根据网站的综合值按照不同的交通旅游网站进行筛选排名结果，通过筛选交通旅游网站可以看到每个交通旅游网站里面的网站排名，从而可以让你跟清楚的进行选择那个网进行浏览，而我用Python从中选取了一个网站进行数据分析。

本文主要从名称、排名、地点、权重、关键词五个维度进行分析，收集官网提供的数据集与Python爬取的日志信息，利用SQL Sever对数据进行预处理与存储，使用python进行分析与可视化呈现，最终对分析结果进行说明并提供相应意见。

**关 键 词**：Python；SQL Sever；交通旅游网站；数据分析

目录

[1 绪论 3](#_Toc113118631)

[1.1 背景及意义 3](#_Toc113118632)

[1.1.1 背景 3](#_Toc113118633)

[1.1.2 目标及意义 3](#_Toc113118634)

[1.2 应用现状 4](#_Toc113118635)

[1.3 主要板块 4](#_Toc113118636)

[1.4 主要方法及技术路线 5](#_Toc113118637)

[1.4.1 主要方法 5](#_Toc113118638)

[2开发环境及技术 6](#_Toc113118639)

[2.1硬件设备 6](#_Toc113118640)

[2.1.1 CPU及操作系统 6](#_Toc113118641)

[2.1.2 其他 6](#_Toc113118642)

[2.2软件及IDE 6](#_Toc113118643)

[2.2.1 MySql 6](#_Toc113118644)

[2.2.2 IDE 7](#_Toc113118645)

[2.3主要技术 7](#_Toc113118646)

[2.3.1 SQL 7](#_Toc113118647)

[2.3.2 Python 7](#_Toc113118648)

[3 可行性分析与模块设计 8](#_Toc113118649)

[3.1 可行性分析 8](#_Toc113118650)

[3.1.1 技术可行性分析 8](#_Toc113118651)

[3.1.2 数据可获得性分析 8](#_Toc113118652)

[3.2 各模块设计 8](#_Toc113118653)

[3.2.1 数据获取方法 8](#_Toc113118654)

[3.2.2 数据预处理设计 8](#_Toc113118655)

[3.2.3 数据分析思路设计 9](#_Toc113118656)

[4 系统实现与结论 9](#_Toc113118657)

[4.1 各模块实现 9](#_Toc113118658)

[4.1.1 获取数据 9](#_Toc113118659)

[4.1.2 数据存储 9](#_Toc113118660)

[4.1.3 分析模块实现 10](#_Toc113118661)

[4.1.4 可视化展示 10](#_Toc113118662)

[4.2 结论 13](#_Toc113118663)

[5总结 13](#_Toc113118664)

[致谢 15](#_Toc113118665)

# 1 绪论

## 1.1 背景及意义

### 1.1.1 背景

交通对旅游业的影响是十分巨大的，交的便捷的地方旅游业将比交通闭塞的地方更加的兴旺。而中国目前的旅游交通公路等级差距大，各种交通设施相对落后，交通的不到位，旅游开展的同时又破坏了环境，所以在去旅游的人要考虑交通的因素，从而产生了许多的交通旅游网，人旅游的人可以更准确的知道去旅游地方的交通好坏以及旅游地的景点人气怎么样。

### 1.1.2 目标及意义

目标：

本文从交通旅游网站的人气、网上查询的浏览都进行综合分析，通过整体访问和浏览量从而知晓那些旅游地方的人气最高，对所去旅游地方有一个大概的了解，让旅游的人更加便捷知晓交通情况，而交通旅游网站的排行。是为了进一步了解交通旅游网站排名的整体情况，从交通网站的、人气、关键词，以及用户的浏览情况，并通过交通旅游网站上的各种景点及交通情况进行排名。

## 1.2 应用现状

近些年，随着社会的发展越来越多的人喜欢在假期进行旅游，从而产生了许多交通旅游网站的诞生。旅游交通是国内外旅游文献长期以来的研究热点，如董亚娟（2016）以及黄柯、祝建军和蒲素（2007）分别对国外和国内旅游交通研究文献进行叙述，但尚未有学者对国内外旅游交通研究文献进行对比研究。因此，对比分析国内外旅游交通研究现状，在此基础上对未来旅游交通的重点研究领域和研究方向进行展望，以期为国内交通旅游实现系统性和准确性。

## 1.3 主要板块

（1）数据获取板块

数据获取板块功能[1]主要是依据分析目的及要达到的目标，确定获取的数据种类，并使用直接获取数据文件方式或爬虫方式获取原始数据。

（2）数据预处理板块

数据预处理板块功能是对获取到的数据进行预处理操作：选择重要字段，标准化处理，异常值处理等预处理操作。

（3）数据存储板块

数据存储板块主要功能是把经过预处理的数据持久化存储，以便于后续分析。

（4）数据分析板块

数据分析板块主要功能是根据分析目标，找出数据中字段之间的内在关系，与规律。

（5）数据可视化板块

数据可视化板块主要功能是使用适当的图标展现方式，把数据的内在关系、规律展现出来。

## 1.4 主要方法及技术路线

### 1.4.1 主要方法

（1）数据获取；本文中数据的获取方式是直接获取交通旅游网站排行进行数据分析（2）数据清洗[2]：本文数据清洗使用的是python技术，通过python语言读取数据文件，并入库操作。在入库之前，剔除无用字段、标准化处理。

（3）数据分析：在数据库中使用sql语言多维度排序分析。

（4）分析结果展示与说明：将分析结果通过python绘制趋势图可视化展现。

# 2开发环境及技术

## 2.1硬件设备

### 2.1.1 CPU及操作系统

要求P41.8 GHZ 以上；内存：8 G 以上；操作系统windows7及以上；

### 2.1.2 其他

网络、交换机、防火墙、输入输出设备等。

## 2.2软件及IDE

### 2.2.1 MySql

MySQL 是一个小型关系型数据库管理系统，关系型数据库是目前最常用的数据库类型，它的访问速度更快、数据表示更灵活。MySQL 支持结构化查询语言 SQL，数据查询简便快捷。MySQL 软件采用了 GPL，即 GNU 通用公共许可证，保证了我们共享和修改自由软件的自由。虽然与其他的大型数据库如 Oracle、SQL Server 等相比，MySQL 规模相对较小，功能也不如其竞争对手强大，可以说有诸多缺点，但是对于中小型企业来说，MySQL 的规模和功能已经足够。而 MySQL 的优点在于体积小、速度快、跨平台，特别是开放源码这一特点，可以获得大量免费资源，促使许多中小型网站都选择My SQL 作为网站后台数据库支持，以降低网站总体成本。

MySQL 的核心程序采用完全的多线程编程，服务灵活，支持多线程，可以充分利用 CPU 资源。MySQL 支持多种接口，JDBC、ODBC、PHP、C 语言、Python、Perl、Ruby、VB 都可以与 MySQL 数据库直接连接提出数据。MySQL服务器提供了连接池，也就是多线程，可以同时为多个客户端提供服务。 常用的 MySQL 管理工具有 MySQL Workbench、MYSQL Front和 PHPMy Admin 等，其中最为出色的是基于Web的PHPMyAdmin 工具。PHPMyAdmin 是一个用 PHP 编写的，免费的数据库前台管理工具，可以通过web方式控制和操作MySQL 数据库。它支持常用的数据库操作，包括管理数据库、表格、字段、联系、索引、用户、许可等增、删、改、查各种操作。PHPMyAdmin 的缺点是必须安装在Web服务器上，所以访问权限如果设置的不正确，就有可能遭到黑客攻击使数据受损。

### 2.2.2 IDE

本文使用PyCharm集成开发环境。

## 2.3主要技术

### 2.3.1 SQL

详细阐述SQL的功能、优势以及选择该技术的原因。

### 2.3.2 Python

详细阐述Python的功能、优势以及选择该技术的原因。

# 3 可行性分析与模块设计

## 3.1 可行性分析

### 3.1.1 技术可行性分析

本人掌握python、sql等数据处理技术，并且熟练使用mySql数据库、pycharm集成开发环境，并且在互联网上比较容易获取找到相关技术文档，因此技术思路清晰。

### 3.1.2 数据可获得性分析

本文分析数据由交通旅游网站提供，数据来源真实可靠而且可以轻松获取。

## 3.2 各模块设计

### 3.2.1 数据获取方法

数据集源交通旅游网站，主要记录了各种交通软件以及景点的排名。

### 3.2.2 数据预处理设计

详细说明数据分析中需要的字段以及对数据进行关联、清洗等预处理内容。

例如：

将收集到的信息表导入SQL Sever数据库中进行存储与预处理，因分析内容涉及用户浏览量分析，所以首先需关联交通旅游网站，

（1）选择子集

根据上面的分析思路，只需要用到以下字段：名称、排名、地点、权重、关键词

（2）删除重复值

无重复数据。

（3）数据排序

按购买时间升序排序。

### 3.2.3 数据分析思路设计

交通旅游网站总体分析可以从网站整体访问及交通旅游排行两个大的维度来进行综合分析，通过整体访问和排名情况可以知道网站的排行，对交通旅游网站的排行有所了解，在一定的时间范围内知晓排名情况。为了解交通旅游网站的排名，可以从时间维度分析名称、排名、地点、权重、关键字等。

# 4 系统实现与结论

## 4.1 各模块实现

### 4.1.1 获取数据

输入网址https://top.chinaz.com/hangye/index\_jiaotonglvyou.html。

### 4.1.2 数据存储

清洗后的数据字段与部分内容如图1.1所示，预处理后共有29971条数据。



图4.1 数据处理后的数据集

### 4.1.3 分析模块实现

(截图展示通过mysql实现的数据关联查询，多维度排序，获取最大值，最小值等，并配合文字简要说明)

### 4.1.4 可视化展示

1. 分别从其中一个交通旅游网站中的名称、排名、地点、权重、关键字用可视化进行分析和展示。

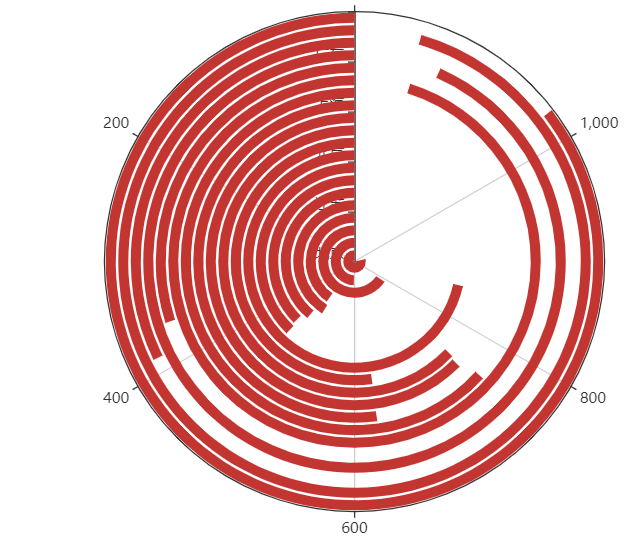


图1.1 名称雷达图

从图里可以看出，浏览天津的用户最多。

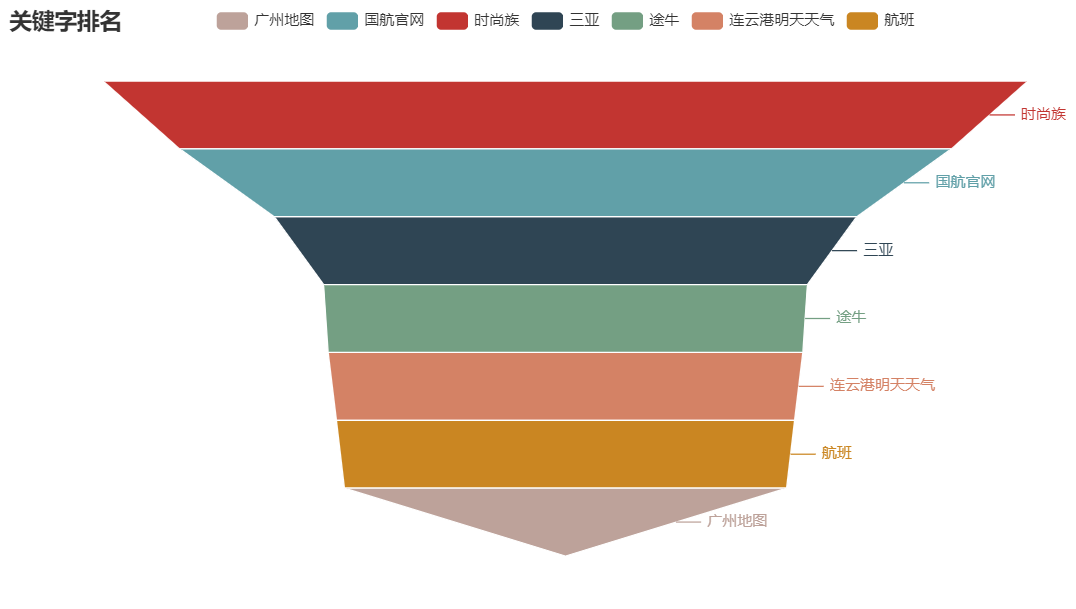


图1.2 地区排名

从图里可以看出，浙江和北京地区的排名最高。

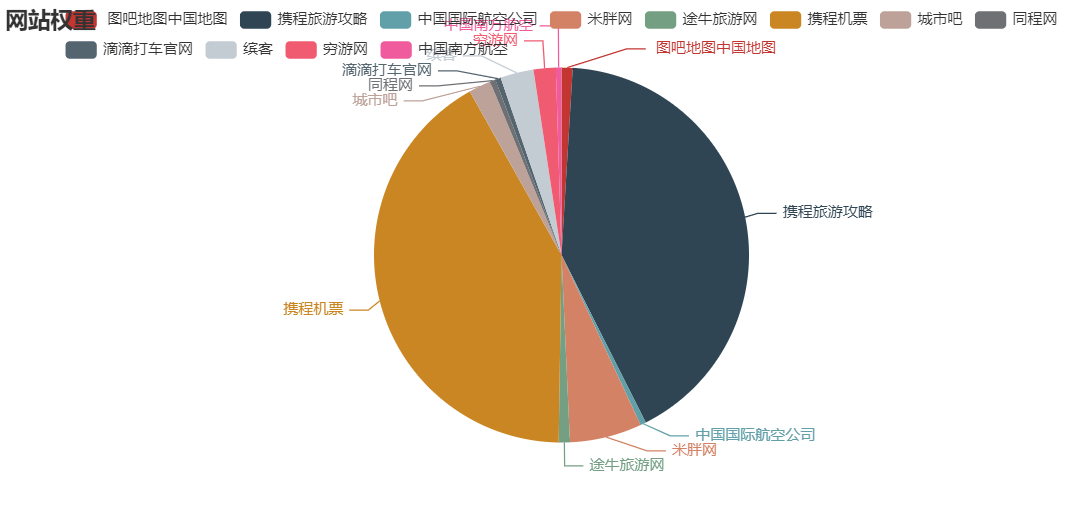


图1.3 浏览软件排名

从图中看出在交通旅游网站中查询旅游地点和交通工具使用携程软件的最多。

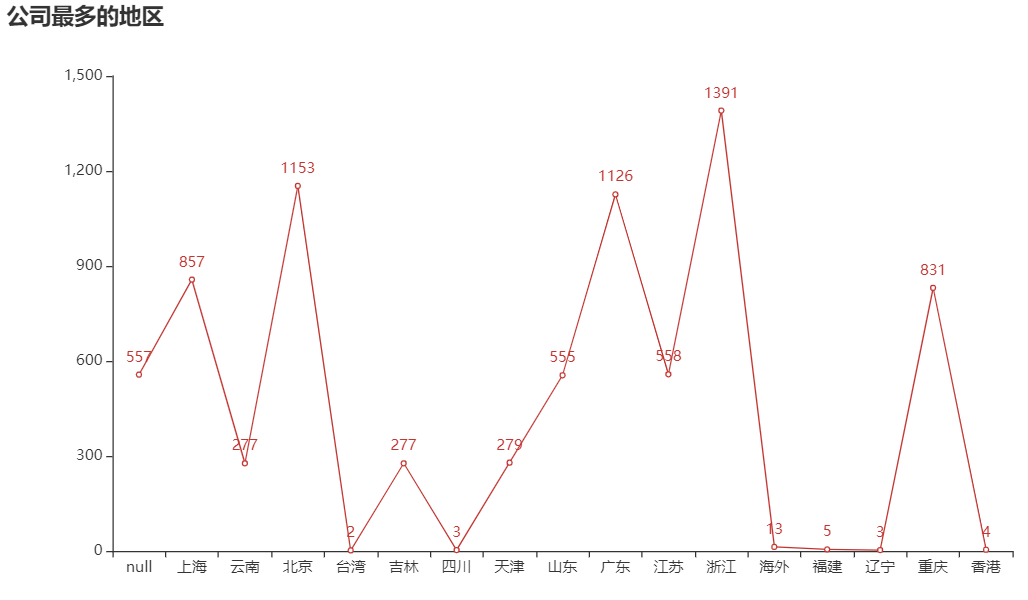


图1.4公司最多的地区

从图中可以看出浙江和北京不仅是地区排名最还是旅游公司最多的地方。

## 4.2 结论

通过可视化的展示可以更快更准确的看出交通旅游网站中所有使用的软件和搜索地区排名情况，从而可以更快的看出你所想去旅游地方的交通情况及旅游景点的火热情况。解决了让旅游的人不用纠结选那个网站来进行查询想去旅游地方的交通情况，可视化可以让你更快的知晓浏览交通软件的排行。

# 5总结

这次选择进行交通旅游网站的分析，因为随着社会的发展，有许多的人在假期喜欢旅游但是不乏去旅游的人不知道所去的地方景点如何交通如何，所去地方人流量如何，更多的是在网上看到的不知道那个地方的排名如何，所以我选择分析交通旅游网站更多的是因为比较常见、也比较实用。在做交通旅游网站的数据分析时也遇许多的难题，虽然克服了，但是也认识了自己的不足，在可视化这一方面还有许多的不足。通过这次做毕业设计发现了自己的想法太片面了。毕业设计不仅是对自己所学习知识的检验，而且也是对自己能力的提高，通过这次毕业设计使我明白了原来知识还比较欠缺。自己要学习的东西还很多，在生活中应该不断的学习，努力提高自己的知识。

我相信，在我以后的成长的道路中我一定会铭记在学校里带给我的每一份欢乐与汗水，将它们绘制成只属于我的风画卷。不经历流血的手指，怎能弹出千古绝唱。不经历地狱的磨炼，怎能拥有创造天堂的力量。不经历烈火的焚烧，怎能拥有钢铁的意志。

# 致谢

本论文是在我的导师付颖老师精心指导和亲切关怀下完成的。付颖老师一直关心着我的生活和学习情况，并始终教导我要以认真、求实和创新的态度来对待学习和研究工作，在做毕业设计的这期间老师一直在帮助我，帮助我纠正我在毕业设计中出现的问题，并耐心的为我的论文校正，正是付颖老师的帮助我才能顺利的完成我的毕业设计，老师的认真负责的工作态度和治学严谨使我们即将步入社会的我们受益匪浅。

论文主要内容是对数据获取、存储、分析与可视化呈现进行介绍与总结，在程序及论文的开发与编写中，黎龙珍等老师也给我了及时的帮助。在此，衷心感谢黎龙珍等老师论文实验过程中的悉心指导和耐心帮助，感谢教导我的所有老师，在学校的这一段时间也认识了很多朋友，他们也给与了我很多的帮助，很感谢他们对我的帮助，在做毕业设计的这几周遇到很多的问题是导师给我指点迷津，像海上的灯塔，为我照亮了前行的方向。

最后，还要感谢我的家人给予我的鼎力关怀和支持，为我完成学业付出了辛劳！

由于别人的水平有限，难免在论文中有些许不足之处，望老师们百忙之中抽出时间对我的论文进行评审和斧正，将不胜感激。