

注：教师姓名后留有一个空格，后面填写教师职称。下面加下划线。

阅后删除此文本框。

**本科毕业设计**

**房源数据分析与推荐系统设计与实现**

|  |  |  |
| --- | --- | --- |
| **学生姓名** | **：** | 黄浩星 |
| **学号** | **：** | 141544225 |
| **学院** | **：** | 互联网金融与信息工程程 |
| **专业** | **：** | 计算机科学与技术 |
| **指导教师** | **：** | 郭艺辉 **职称：** 讲师 |
| **提交日期** | **：** | 2018 年 2 月11 日 |

**本科毕业论文（设计）诚信声明**

本人郑重声明：所呈交的本科毕业论文（设计），是本人在指导老师的指导下，独立进行研究工作所取得的成果，成果不存在知识产权争议，除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

 学生签名：

时间： 年 月 日

**关于论文（设计）使用授权的说明**

本人完全了解广东金融学院关于收集、保存、使用学位论文的规定，即：

1.按照学校要求提交学位论文的印刷本和电子版本；

2.学校有权保存学位论文的印刷本和电子版本，并提供目录检索与阅览服务，在校园网上提供服务；

3.学校可以采用影印、缩印、数字化或其它复制手段保存论文。

本人同意上述规定。

学生签名：

时间： 年 月 日

摘 要

最近几年来，大数据一词在人们的视线范围内出现得越来越频繁，大数据已经逐渐深入各个行业和业务领域之中，从信息爆炸时代产生的海量数据中分析，提取有价值的信息已经成为重要的生产要素。

房源数据在市场的交易中日益积累了海量数据，房地产市场交易的对象一般以房源作为基本单元，同时房源也是房地产的基本物理构成单元，房源数据主要包括：面积信息，周边环境，房价信息，地理位置，配套设施等。

如今人们越来越依赖于互联网信息，所以购房者在互联网的帮助下得到良好的服务质量的同时一定也留下了重要的数据痕迹，例如，兴趣取向，需求选择，环境考量等用户行为信息，当然，购买者同时也关心房源的历史发展信息，历史房价波动，房价升值等信息。这些动态的用户行为数据蕴含着巨大的商业价值，而且对房地产市场起着导向与反馈的作用，开发者可以从中挖掘丰富的数据信息从而拓展更大的市场信息。

大数据分析是用户交互，用户共识，信息交互之间的重要桥梁，大数据分析典型的应用就是通过海量数据检索，统计，比较，发现潜在的关联，然后获取关联的信息进行模式的预测和匹配分析。

**[关键词]**：大数据分析；用户行为分析；房源信息；用户推荐

Abstract

In recent years, the word BigData appears frequently in people's eyes. BigData has gradually penetrated into various industries and business fields. Analyzing and extracting valuable information from the massive data become an important factor of production.

In the market Housing transactions have accumulated huge amounts of data. Generally housing as the basic unit in the market transactions, while housing is also the basic physical units. Housing data include area information, the surrounding environment, price information, location information, facilities and so on.

Nowadays, people always depends on Internet information, therefore, home buyers, with the help of the Internet, get good service quality and leave important data traces while they buy housing in the Internet. User behavior information such as interest orientation, demand selection and environmental considerations are very important. Buyers are also concerned about the historical development information, fluctuations in historical prices, prices appreciation and other information. These dynamic user behavior data contain huge commercial value witch plays a guiding and feedback role in the market. Developers can explore huge data information to expand the market information.

BigData analysis is an important bridge between user interaction, user consensus and information exchange. The typical application of big data analysis is to find out the potential correlation through massive data retrieval, statistics, comparison, and then obtain the associated information to predict and match the patterns analysis.

**[Key Words]:** BigData analyzing; user behavior analyzing; housing data; user recommended

目 录

注意：（1）目录必须通过引用自动生成，不能手工生成；（2）论文修改后，重新生成目录，更新目录的标题及页码

样式：一级标题

字体：黑体，居中，加粗

字号：小三

段落：段前0行

段后0.5行

阅后删除此文本框

一级目录

字体：黑体、加粗

字号：小三

段落：1.5倍行距

阅后删除此文本框

[摘 要 I](#_Toc495071423)

[Abstract II](#_Toc495071424)

[目 录 III](#_Toc495071425)

[1 绪论 1](#_Toc495071426)

二级目录

字体：宋体

字号：四号

段落：1.5倍行距

阅后删除此文本框

[1.1 研究背景 1](#_Toc495071427)

[1.2主要研究工作 2](#_Toc495071428)

[2 相关技术 2](#_Toc495071429)

[2.1分布式系统 2](#_Toc495071430)

三级目录

注意：目录最多为三级

字体：宋体

字号：小四

段落：1.5倍行距

首行缩进2个字符

阅后删除此文本框。

[2.1.1 概述 2](#_Toc495071431)

[3 分布式移动 3](#_Toc495071432)

[4 结论 4](#_Toc495071433)

[参考文献 5](#_Toc495071434)

[致 谢 6](#_Toc495071435)

[附录 内容名称 7](#_Toc495071436)

注意：附录根据实际来确定是否需要

阅后删除此文本框

**房源数据分析与推荐系统设计与实现**

2008年末，“大数据”得到部分美国知名计算机科学研究人员的认可，业界组织计算社区联盟 (Computing Community Consortium)，发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。它使人们的思维不仅局限于数据处理的机器，并提出：大数据真正重要的是新用途和新见解，而非数据本身。此组织可以说是最早提出大数据概念的机构。

1 绪论

1.1 研究背景

数据时代飞速发展，近几年大数据的发展和研究更是达到了高峰，尤其在2017年人工智能的推动下数据的采集和建模显得尤为重要，表1.1所示。各互联网公司都致力于提供更好的数据分析服务，数据挖掘服务，产品推荐服务等，从而使得人们在生活等各个方面得到更好的服务体现。数据分析及推荐服务的应用领域主要有电子商务，销售，金融，并深入应用于房地产行业中。房地产市场积累了海量的用户数据，应用数据分析手段可以从海量低价值密度的数据中提取对房地产商有潜在价值的信息，并且对新用户可以更好的推荐一款适合的楼房。目前，市场上对房源数据研究调研的公司比较少，长期以来，用户一般购房都会花大量的时间在对比价格，地段，交通，增值等方面，这种方式不但效率低下而且往往还不是最适合自己的购房意向。在房源多元化的现今用户很难不出门就一一对比房源的各种信息，包括历史信息和未开放信息。为了提高房地产商的投资准确性和用户购房的良好体验，迫切需要一个给予房源数据分析与推荐系统。

表1.1 大数据的发展历程

|  |  |
| --- | --- |
| 2005年 | Hadoop项目诞生，提供了大数据的存储方案 |
| 2008年 | 美国发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。它使人们的思维不仅局限于数据处理的机器，并提出：大数据真正重要的是新用途和新见解，而非数据本身 |
| 2011年 | IBM的沃森超级计算机每秒可扫描并分析4TB（约2亿页文字量）的数据量，并在美国著名智力竞赛电视节目《危险边缘》“Jeopardy”上击败两名人类选手而夺冠 |
| 2013年 | 英国政府宣布注资6亿英镑发展8类高新技术，其中，1.89亿英镑用来发展大数据技术 |
| 2015年 | 五中全会的“十三五”规划中将大数据作为国家级战略 |

1.2国内外研究现状

在国外，通过数据的采集，分析与统计房源的应用网站已经相对比较成熟，国外存在很多一些地产开发商推荐网站。在我国其实也有一些房源推荐应用网站，例如链家网，安居乐等，他们聚集了众多的房地产商的信息和资料，他们的网站也各自有优缺点，这些网站大多数只是展列房源的一些详细信息就完事了，在一定成度上没有往用户的更好体验上提供比较好的解决方案，例如很少为用户分析某个地产近几年的发展趋势，增值趋势，或者价格变动，人均购买量等信息，更没有一个比较完善的用户体验良好的推荐系统为购买用户做到引导作用。

**现今，数据分析成为数据时代的核心，广泛采用实时性的数据处理方式。**大数据的价值体现在对大规模数据集合的智能处理方面，进而在大规模的数据中获取有用的信息。要想逐步实现这个功能，就必须对数据进行分析和挖掘。而数据的采集、存储、和管理都是数据分析步骤的基础，通过进行数据分析得到的结果，将应用于大数据相关的各个领域。未来大数据技术的进一步发展，与数据分析技是密切相关的。

对于现在人们日益追求高效，可靠的信息发展时代，购买者不再想一一地去现场进行比量，销售人员也不再想一一地进行各式各样的推荐销售。所以这款房源数据分析与推荐系统不但可以给购买者推荐他们满意的，理想的，合适的，定制化的房源信息，而且可以给销售人员以及房地产商提供更好的服务。此系统从信息来源处做到了多元化，多渠道，可靠化，以为此系统会采集不但只采集一家房地产销售网站的数据，而是采集各城市，各地区，不同的房地产商销售网站的数据信息，尽量做到数据来源的真实性。

发现知识，寻找关联，挖掘价值是数据时代信息分析的真正需要。大数据背景下，对于投资商来说，海量的房地产数据一直推动着房地产市场的相关研究，基于数据挖掘算法，合理应用大数据分析技术，可以有效地帮助我们从历史与一直更新的数据中提取有价值的信息，及时掌握市场发展动态，对于市场的供需提供有用的引导，合理定制产业调整策略。对于消费者来说，如何从大量信息中找到对自己感兴趣的信息是一件非常困难的事情，推荐系统的任务就是关联用户和信息，一方面帮助用户发现对自己有价值的信息，另一方面让信息能够展现在对它感兴趣的用户面前，从而实现信息消费者和信息消费者的双赢。

1. 主要研究工作
   * 1. 主要研究内容
2. 本论文研究的房源数据分析与推荐系统的设计与实现是基于B/S架构的。采用Python进行数据的采集和去重以及初步的数据清洗，把过滤后的数据进行存储，主要采用Mysql数据库进行存储。数据的处理逻辑业务则由Java进行实现，其中涉及到Spring boot, Spring-mvc, Spring, Mybatis, Redis等技术，系统界面则利用Vue.js等进行开发。在开发过程中，采用分模块开发提高可维护性和可扩展性。此系统的主要内容是针对目前市场上的房源信息第一系列数据的汇总统计与分析，以及针对购买用户进行准确的合适的推荐。系统分为数据采集模块，数据分析模块，推荐服务模块等。
   * 1. 主要解决问题

围绕从当前各个房源售卖网站采集来的多样化，低价值密度，海量数据中通过数据分析的手段寻找数据中各种隐藏的关联关系，挖掘价值信息，从而分析房地产市场的周期波动变化，寻找波动周期的规律和变动的时间点以及分析影响其变动的主导关系和主要因素。进一步探索房地产市场的发展可能因素和潜在的市场价值，在结合用户的各种需求，推荐一款用户合适的，满意的，性价比高的房屋信息，同时借助现代计算机的技术快速为房地产销售提供准确的，高效的，可靠的定位不同用户的需求。

2 相关技术

2.1 开发工具与开发技术简介

2.1.1 开发工具与环境介绍

开发环境：Windows 7

开发工具：Intellij Idea Community

数据库　：Mysql

开发语言：Java(JDK1.8) Python(2.7.3)

Web服务器：Apache Tomcat8.0

IDEA 全称 IntelliJ IDEA，是Java语言开发的集成环境，IntelliJ在业界被公认为最好的Java开发工具之一，尤其在智能代码助手、代码自动提示、重构、J2EE支持、各类版本工具(Git、Svn、Github等)、JUnit、CVS整合、代码分析、 创新的GUI设计等方面的功能可以说是超常的。IDEA是Jetbrains公司的产品，开发人员以严谨著称的东欧程序员为主。它的旗舰版本还支持HTML，CSS，PHP，MySQL，Python等。免费版只支持Java等少数语言。

Tomcat是一个小型的轻量级应用服务器，在中小型系统和并发访问用户不是很多的场合下被普遍使用，是开发和调试JSP程序的首选。对于一个初学者来说，可以这样认为，当在一台机器上配置好Apache服务器，可利用它响应对HTML页面的访问请求。实际上Tomcat部分是Apache服务器的扩展，但它是独立运行的，所以当你运行Tomcat时，它实际上作为一个与Apache独立的进程单独运行的。

Java 8允许我们给接口添加一个非抽象的方法实现，只需要使用 Default关键字即可。新增lambda表达式，提供函数式接口，Java 8 允许你使用，:: 关键字来传递方法或者构造函数引用我们可以直接在Lambda表达式中访问外层的局部变量。

MySQL是Web世界中使用最广泛的数据库服务器。SQLite的特点是轻量级、可嵌入，但不能承受高并发访问，适合桌面和移动应用。而MySQL是为服务器端设计的数据库，能承受高并发访问，同时占用的内存也远远大于SQLite。此外，MySQL内部有多种数据库引擎，最常用的引擎是支持数据库事务的InnoDB。

* 1. 总体业务架构简介



图2.1 业务架构图

如图2.1为整个系统的总体设计架构图，系统一共分为三大层分别为web层，业务处理服务层，数据持久化层，业务处理服务层又分为两大模块，数据分析服务模块，用户推荐服务模块。

整一个系统首先从获取数据源作为系统运作的支撑，其中time-task-service是一个定时任务，定期地通过数据爬虫的手段向房价网列如：安居乐，链家网等获取重要的源数据例如：房源面积，楼房价格，地理位置，周边环境等数据。获取数据后经过初步的数据过滤和清洗剔除重复数据，剔除无效数据，剔除不完整的数据，进一步对源数据进行分类，建模，然后把有效数据导入数据库。Data-analysis-service从数据库中批量读取数据源，把数据进行提取，归类，组合，分析提供web层进行可视化展示，同时把分析结果存储进数据库已保存现阶段结果。

用户通过登录访问系统，系统access-log-listener通过监听用户的行为，提取用户操作日志中的必要信息，例如：访问源Ip，访问数据，访问次数，点击种类等，对用户行为日志批量入库。Recommend-service通过分析用户行为日志，结合用户兴趣喜好，用户地理位置等推荐多款适合用户的房源信息。

# 3 分布式移动

4 结论

参考文献

[1] 罗建幸, 徐红燕. 低市场集中度行业中的市场机会[J]. 商业研究, 2003, 16(276): 138-138.

注意：编号[ ] 与作者间加一个空格

阅后删除此文本框

一级标题，居中

阅后删除此文本框

[2] 徐齐利.互联网平台电商企业的垄断势力与结构——对百度搜索平台客户集中度的测算[J]. 当代财经, 2017, (3): 91-105.

[3] 王领, 刘融. 基于演化博弈论视角的互联网巨头并购研究[J]. 电子商务, 2016, (9): 13-15.

[4] 张新红, 张贝. 2016年中国分享经济发展报告[R]. 北京: 国家信息中心信息化研究部, 中国互联网协会分享经济工作委员会, 2016.

[5] 程维, 柳青. 滴滴: 分享经济改变中国[M]. 北京: 人民邮电出版社, 2016.

[6] Shabtay D., Gaspar N. and Yedidsion L. A bicriteria approach to scheduling a single machine with job rejection and positional penalties[J]. Journal of Combinatorial Optimization, 2012, 23(4): 395-424.

[7] Pinedo M. Scheduling: theory, algorithms, and systems (4th edition) [M]. New York: Springer, 2012.

注意：（1）参考文献不少于15篇；（2）参考文献以正文中引用次序排列；（3）若标明引用页码，必须标明起始两个页码；（4）期刊论文必须标明页码；（5）必须标明文献类型；毕业论文不能少于2篇英文参考文献

中文字体：宋体

字号：五号

英文字体：Times New Roman

字号：五号

段落：左对齐，固定行距20磅(pt)

阅后删除此文本框

致 谢

在本文完成之际，谨向我的导师郭艺辉老师致以衷心的感谢，本论文是在她的精心指导和关怀下完成的，从论文的选题、方案设计，到论文的撰写和修改，都倾注了郭老师的心血和汗水，在学习期间，她的言传身教将使我终生受益，她认真严谨的治学态度、豁达宽广的胸怀、平易近人的处事风格是我一生的楷模，值此提交论文之时，在此向郭老师表达衷心的感谢！

一级标题，居中

阅后删除此文本框

附录 内容名称

以下内容可放在附录之内：

注意：（1）附录与附录内容之间留有一个空格；（2）附录为可选项，根据实际需要确定，若无，将此页删除

一级标题

阅后删除此文本框

(1) 正文内过于冗长的公式推导；

(2) 方便他人阅读所需的辅助性数学工具或表格；

(3) 重复性数据和图表；

(4) 论文使用的主要符号的意义和单位；

(5) 程序说明和程序全文。

字体：宋体

字号：小四

段落：两端对齐，首行缩进2个字符，1.5倍行距

阅后删除此文本框