

基于移动大数据的用户职住特征分析

谢三山

院 (系): 经济与管理学院 专 业: 信息管理与信息系统

学 号: 1171000514 指导教师: 叶强教授

2021 年 5 月 24 日

哈爾濱工業大學

毕业设计（论文）

题 目 基于移动大数据的

用户职住特征分析

专 业 信息管理与信息系统

学 号 1171000514

学 生 谢三山

指 导 教 师 叶强教授

答 辩 日 期 2021 年 5 月 24 日

摘 要

摘要的字数（以汉字计），硕士学位论文一般为 500 ~ 1000 字，博士学位论文为 1000 ~ 2000 字，均以能将规定内容阐述清楚为原则，文字要精练，段落衔接要流畅。摘要页不需写出论文题目。英文摘要与中文摘要的内容应完全一致，在语法、用词上应准确无误，语言简练通顺。留学生的英文版博士学位论文中应有不少于 3000 字的“详细中文摘要”。

关键词是为了文献标引工作、用以表示全文主要内容信息的单词或术语。关键词不超过 5 个，每个关键词中间用分号分隔。（模板作者注：关键词分隔符不用考虑，模板会自动处理。英文关键词同理。）

关键词：大数据；职业；家庭；用户画像分析

Abstract

An abstract of a dissertation is a summary and extraction of research work and contributions. Included in an abstract should be description of research topic and research objective, brief introduction to methodology and research process, and summarization of conclusion and contributions of the research. An abstract should be characterized by independence and clarity and carry identical information with the dissertation. It should be such that the general idea and major contributions of the dissertation are conveyed without reading the dissertation.

An abstract should be concise and to the point. It is a misunderstanding to make an abstract an outline of the dissertation and words “the first chapter”, “the second chapter” and the like should be avoided in the abstract.

Key words are terms used in a dissertation for indexing, reflecting core information of the dissertation. An abstract may contain a maximum of 5 key words, with semi-colons used in between to separate one another.

Keywords: bigdata, occupation, family, user potrait analysis

物理量名称及符号表

表1 国际单位制中具有专门名称的导出单位

量的名称	单位名称	单位符号	其它表示实例
频率	赫[兹]	Hz	s ⁻¹

目 录

摘要	I
Abstract	II
物理量名称及符号表	III
第1章 绪论	1
1.1 研究课题的来源、背景和意义	1
1.2 国内外与课题相关研究领域的研究进展及成果	1
1.2.1 大数据及其相关理论和应用的发展概况	1
1.2.2 手机信令数据与职住特分析的相关理论和发展概况	3
1.2.3 城市功能区与用户行为目的和职住特征分析研究进展及成果	4
1.2.4 大数据下的职住空间聚类算法的分析和改进	5
1.2.5 基于手机信令的用户特征画像分析相关研究进展及成果	6
1.2.6 存在的不足或有待深入研究的问题	6
1.3 本课题的主要研究内容概述	6
1.3.1 数据分析与可用性判断	6
1.3.2 大数据下 K-Means 聚类方法的分析和优化	7
1.3.3 哈尔滨市栅格化处理以及对栅格化后的街区类型分析	7
1.3.4 大数据系统及其套件的分析和比较	8
1.3.5 基于手机信令数据的用户驻留点判定	8
1.3.6 基于手机信令数据的不同时间段用户访问街区类型偏好	8
1.3.7 基于手机应用数据的不同时间段用户手机应用类型偏好	9
1.3.8 基于手机通信数据的用户社交习惯和偏好分析	9
1.3.9 基于手机信令数据的用户职业特征与家庭特征分析	9
第2章 原始数据统计性描述、可用性分析和冗余与错误清理	10
2.1 原数数据分析与统计性描述	10
2.1.1 用户与基站连接记录数据分析	10
2.1.2 基站基本信息表分析与描述性统计	10
2.2 哈尔滨市 POI 兴趣点爬取、处理和描述分析	11
2.3 用户手机应用使用流量记录表分析	12

2.4 用户通话和短信使用记录数据表分析.....	13
2.5 本章小结.....	15
第3章 哈尔滨市城区POI兴趣点数据获分析和栅格化街区聚类分析	16
3.1 哈尔滨市城区POI兴趣点数据具体分析.....	16
3.2 基于路网和栅格化的功能区聚类分析模型对比	17
3.3 功能区的栅格化参数的分析和对比.....	17
3.4 本章小结.....	17
第4章 大数据领域下聚类方法的对比和在哈尔滨城市功能区的聚类结果分析	18
4.1 DBSCAN聚类算法概述和分析.....	18
4.2 K-Means聚类算法概述.....	18
4.3 K-Means初始质心选择优化K-Means++算法	18
4.4 大样本优化Mini Batch K-Means算法.....	19
4.5 哈尔滨市功能区聚类结果分析和展示.....	19
4.6 本章小结.....	19
第5章 大数据系统不同计算框架的分析与对比	20
第6章 基于手机信令数据的用户驻留点、手机应用偏好以及社交地位分析	21
第7章 基于驻留点、应用偏好和社交地位分析结果的用户职住特征分析	22
结 论	23
参考文献	24
哈尔滨工业大学本科毕业设计（论文）原创性声明	26
致 谢	27
附录1 外文资料原文	28
1.1 Single-Objective Programming.....	28
1.1.1 Linear Programming	29
1.1.2 Nonlinear Programming	30
1.1.3 Integer Programming	30
附录2 外文资料的调研阅读报告或书面翻译	32
2.1 单目标规划	32
2.1.1 线性规划	32
2.1.2 非线性规划	33

2.1.3 整数规划	33
附录 3 其它附录	34

第1章 绪论

1.1 研究课题的来源、背景和意义

本研究课题来源于哈尔滨工业大学和中国移动联合实验室承接的基于黑龙江省移动用户数据开展多方面的研究和分析。本研究需要解决的问题是通过移动提供的数据，对哈尔滨市用户的职业、家庭特征进行分析，绘制出用户画像，为后续精准营销、城市规划等研究打下数据基础。

1.2 国内外与课题相关研究领域的研究进展及成果

1.2.1 大数据及其相关理论和应用的发展概况

全球互联网时代的到来，不管是早期的个人电脑、移动电话，甚至是目前发展势头正盛的各种接入互联网的设备形成的物联网，虽然带给人们越来越多的便携性，但各方面人士对偌大的全球性网络中产生的数据提出相当多的问题。除了个人隐私、数据归属等使用权以及相关法律层面的问题，开发人员不能逃避的一个重要问题则是超大规模数据的存储、处理和查询。

根据国际数据公司（IDC, International Data Corporation）的预测^[1]，由于5G的商用化在中国国内逐渐铺开，企业和用户的设备产生的数据将会成为中国数据的主流，数据量和数据市场将产生跨越式增长，在2023年左右达到40ZB（ $1ZB \approx 10^{12}GB$ ）。

飞速增长的数据量，无论单台计算机的存储能力、计算速度达到何种水平，在数以PB计的数据量面前，都不足以应对。同时，单机的价格相对于其性能并不是线性增加的，昂贵的单机服务器使团体或者公司无法负担处理如此庞大数据量的成本，二十世纪初的大数据也只能被丢弃。

首先解决该问题是谷歌，2003年谷歌发表了三篇技术论文^[2]，分别是解决了存储大规模数据的GFS（Google File System，谷歌文件系统），解决了大规模计算的MapReduce方法，以及解决了实时查询的BigTable系统。之后雅虎公司推出了Hadoop平台及其生态，被Apache Software Foundation公司引入开源应用。其中，最有价值的GFS相关的论文使得HDFS（Hadoop Distributed File System，Hadoop分布式文件系统）的开源实现成为了目前绝大部分大数据平台的基石。

如图1-1所示，最底层的HDFS分布式文件存储系统是整个大数据领域的

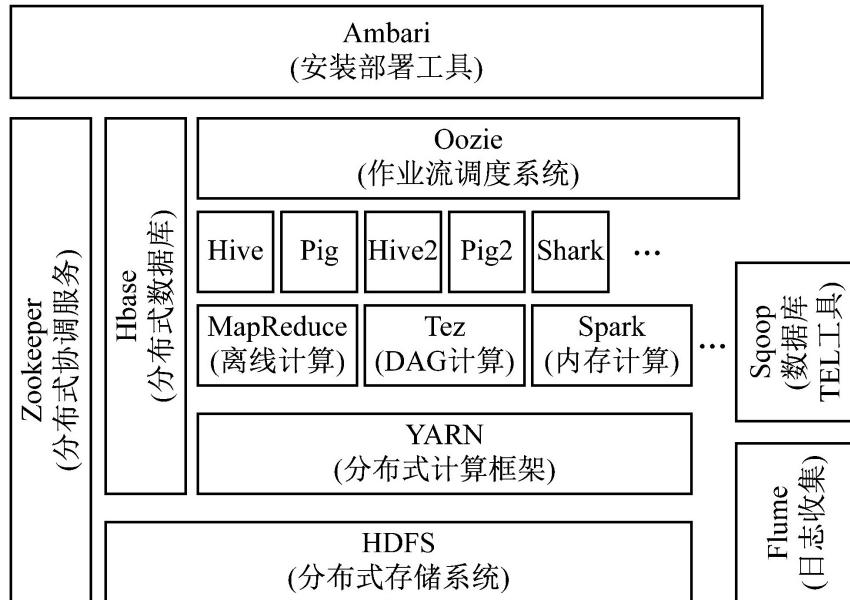


图 1-1 Hadoop 生态系统

基础，通过切分超大文件，实现了一个多机保存、备份以及高容错的大数据文件系统。HDFS 只上的是 Yarn (全称 Yet Another Resource Manager)，主要由于 Hadoop 1.0 版本中的重大架构缺陷，导致任务调度系统 (JobTracker) 承担了太多诸如资源调度、异常监控、接受任务等不同且复杂的任务，于是行业将 JobTrakcer 的任务拆分开来，形成了建立在 HDFS 上的新的调度系统，能够兼容更多的计算框架，如本研究使用的 Spark 以及同样用 DAG 优化的 Tez 等。

MapReduce 则是一种计算模型，通过一次 Map 操作生成 key-value 键值对，以及第二次 Reduce 操作对所有键值对进行规约，得到最终的结果。而 MapReduce 如此暴力的实现方法被数据库领域的专家 David J. DeWitt^[3] 在一篇著名论文 *MapReduce: A Major Step Backwards* 中指责谷歌在大数据的计算上即 MapReduce 方法放弃了数据库领域中的优秀理论和方案，选择了简单粗暴的解决办法。不久，加州大学伯克利学院的 AMP 实验室则推出取 MapReduce 和数据库中精华的 Spark 平台，通过 DAG (有向无环图) 解决了多次机器学习、聚类算法中多次迭代的麻烦问题，以及稳定的 Spark Streaming，MLlib 等拳头产品，使 Spark 平台及其套件，成为众多公司的选择。

如图 1-2 所示，整个 Spark 生态系统中的核心是 Spark Core，从基于 HDFS 的 YARN 框架、HBase、Hive 数据库等中读取数据，通过独立调度器 (Standalone)、

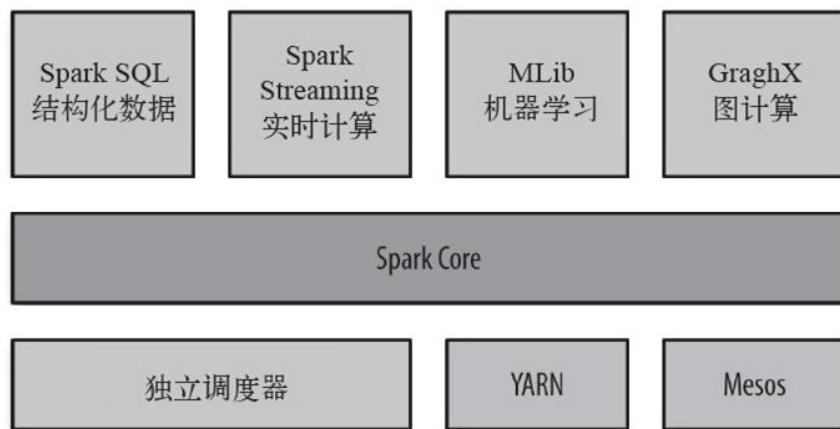


图 1-2 Spark 生态系统

YARN 和 Mesos 调度任务 (Job)，然后完成 Spark 程序，或者通过 spark-shell/spark-submit 等命令行工具，批量运行或提交任务，再利用基于 Spark Core 的 Spark Stream 完成实时计算、MLLib 实现机器学习以及 GraphX 进行图形计算提高程序运行速度。

除此之外，与只推荐用 Java 编写 MapReduce 程序不同，Spark 能够和多种动态或静态语言配合，如 Python、Scala、Java，甚至能够启用 spark-shell 命令行工具，使用 Scala 语言进行交互式查询，灵活便捷。Spark 还支持更复杂的操作，除开类似 MapReduce 的固定的生成 key-value 键值对的方法，Spark 基于 RDD 模型 (Resilient Distributed Datasets，抽象弹性分布式数据集)，只需要承担非常小的代价，就能实现如 SQL 中的 join、union 等高阶方法，以及流式查询功能。

1.2.2 手机信令数据与居住特征分析的相关理论和发展概况

2020 年 9 月，中国互联网络信息中心 (CNNIC) 第 46 次发布了《中国互联网络发展状况统计报告》^[4]，截止 6 月底，中国已经形成了 9.4 亿规模的网民群体，移动互联网用户规模更是达到了 13.19 亿人，同时“中华人民共和国国民经济和社会发展第十三个五年规划纲要”中首次明确指出 2020 年的互联网普及率计划，要将移动宽带的普及率提升到 85%，如此庞大的用户数量，必然会产生超大规模的用户数据，客户在日常的手机使用中，会产生诸如基站连接记录、手机应用流量使用记录、点对点通话记录、短信记录等。

其中，用户手机接受信号产生的基站连接记录的数据成为一个重点分析对象，连接数据能够模糊推断出用户的位置，根据这一点，2017 年孔扬鑫^[5]采用

了轨迹行为特征的判定算法，基于 MapReduce 分布式计算模型，基于用户的基站连接数据，分析基站连接的消失点、消失时长、用户停留点、用户工作地点与家庭住址的平均距离，计算出用户进出城的人口流动的统计数据，再基于基站与县区之间的关系，判断用户的通勤起终点所属区域，最终得到用来反映城镇间、工作地居住地间的人口流动的状态转移序列。

由于城区的范围都比较大，因为基站的覆盖面积广而导致位置的推算的误差并不能显著的影响用户所处城镇的结果。但如果需要精确到城区内部的定位，或者对用户未来位置的推断，就必须要把基站范围导致的误差考虑在内。北京邮电大学 2019 年刘奕杉^[6]则是基于用户连接基站的记录产生的状态转移序列，构建了位点转移的语义序列，再利用 TS-RNN (Three-layer Symmetrical Recurrent Neural Network，三层对称卷积神经网络) 提取出转移特征值，实现了更精准的用户活动区域获取以及位置预测。

常在春^[7]在延庆区职住空间关系分析一文中同样利用 Spark 大数据平台，首先对延庆地区的手机信令数据做一个初步筛选，主要针非过境、常驻、非固定设备用户的工作以及居住地位置的分析，然后使用核密度法得到多个工作地和居住地对应的基站的信息，最终实现用户对工作地点及居住地点的信息推算，并且对延庆区的通勤、工作地使用情况、居住地使用情况等进行描述和分析。

1.2.3 城市功能区与用户行为目的和职住特征分析研究进展及成果

需要分析用户的职业居住的特征分析，就必然不能绕开对用户所处位置的功能语义分析，即用户处于什么目的而移动到该区域。为了能够大致描述用户目的，就需要对城市进行功能区划分。由文献^[8]可知，杨振山等人融合北京市移动电话信令数据以及网络地图 API 提供的 POI (Point of Interest, 兴趣点)，通过计算单位数量 POI 的人口密度、标准化的单位面积 POI 密度以及类内散度矩阵等特征，对栅格化后的北京市推断所有区块的主导功能类型，得到如北京市功能区比率分布，居住和餐饮、娱乐、服务等工作区块分布特点，以及区块的日夜活跃等特点。

除了对城市进行栅格化城区对每个区块进行分析，还有根据城市主要路网对市区进行一个大致上的划分，福建农林大学毋亭在对泉州市^[9]按照路网信息划分出不规则的网格作为研究单元，同时提出按照多个属性如群众对区域类型的认知度（如对大学的认知度较高、对一般商店认知度较低）、该类型覆盖的面积对区块进行权重赋值，基于该种划分和权重法，再加上核密度估计法，完成对不同功能区的识别和划分，最后通过对泉州市地图进行对比，验证准确率。

本文研究用户的职住画像，需要根据城市中职住区域的分布进行参数上的调整，文献^[10]对北京市的职住空间分布进行定量的研究，通过分析北京市三十余天一亿多条的基站的链接数据，根据北京市的路网结构情况，综合运用职住偏离度、空间错误指数和职住分离率等特征，研究城市功能区分布规律和匹配特点，为本研究对哈尔滨市的研究提供了参考。

1.2.4 大数据下的职住空间聚类算法的分析和改进

本研究需要对城市的功能区进行聚类，得到用户移动语义，针对哈尔滨市 67 万余条城市兴趣点信息，需要一个高效、准确的聚类算法。如章节 1.2.1 所言，基于 Spark Core 这个核心，以及拆分成不同组件的调度器、Mesos 等结构，Spark 能够灵活的提供非常多种无监督聚类以及有监督的机器学习等操作。比如 Spark MLlib 库中提供的 K-Means 算法，通过基于 HDFS 的 Yarn 框架，读取分布式数据，采用类似 Map 的方法生成若干对键值对，Spark Core 再将这些键值对分配到各个节点上，分布式地完成 K-Means 算法，该算法充分利用了 Spark 的高弹性分布式数据集，在高扩展、速度快的条件下，完成多种无监督聚类算法。

文献^[11]，东北大学张景奇等人指出，中国 2010 年到 2019 年截止，知网上有 625 篇关于 POI 兴趣点数据研究的文章，特别是在 2017 年之后发生了爆炸式的增长，而关于 POI 兴趣点的分析方法主要是空间自相关分析、核密度估计以及 DBSCAN 聚类算法。作者的研究表明，兴趣点大数据对城市功能区布局、空间结构以及发展规律的分析都起到一个非常重要的影响。

比如文献^[12]就通过背景和上海的兴趣点大数据以及不同的聚类算法，分析北京市和上海市的街道元素，从而得到两市的城市舒适度评价。文献通过路网数据将城市进行语义上的分隔，通过街景元素的可视数量的占比加上 K-Means 聚类分析，将街景分成 29 类并进行相关性分析以及舒适度评价。利用路网数据进行语义分割加上 K-Means 街景分类为本研究的街区分类提供了聚类路线上的参考。

不过本研究的街区划分颗粒度较细，POI 兴趣点的数量较多，使用一般的 Kmeans 聚类算法在时间复杂度上不够优秀，需要寻找优化方法。比如文献^[13]中，为了提高室内 WLAN 定位的算法性能和准确度，希望通过先聚类再采用 XGBoost 分类算法确定准确位置，为了解决超大规模数据量的问题，文献采用大数据优化的 Minibatch K-Means 聚类算法，降低了数据维数、提高了运算速度的同时，仍然保持了较高的准确率，不失为一个大数据下的聚类优化算法。

1.2.5 基于手机信令的用户特征画像分析相关研究进展及成果

文献^[14]没有着眼于兴趣点驻留偏好，而是通过手机信令数据，刻画出用户的高聚集点，刻画出两点一线、双核心、均匀分布等轨迹类型，再根据用户的其他特征，描述和推测用户的职业类型（如学生、退休老人、技术工种等）、爱好属性，以及年龄层分布等相关信息。文献中值得借鉴的方面，包括数据的预处理，作者首先对原始数据进行均匀间隔采集处理、清楚无效和重复数据、分隔时间段，然后基于 DBSCAN 聚类算法改进出一种高簇聚类法得到用户的居住点、工作地、娱乐场所等核心点。

除了基站连接数据的应用，文献^[15]中利用游戏类 APP 流量数据与其他 APP 流量数据的比率，同时处理用户的短信费用、流量套餐及其费用以及手机连接流量时发送的请求推算出手机的大致平台，通过上述处理后的数据，划分出不同年龄层、不同消费等级、不同的消费行为以及不同的消费心理等，细分用户群，准确把握客户价值从而实现精准营销。其中对不同 APP 相关数据运用对本研究处理用户 APP 流量数据的处理有一定的引导和启发效果。

1.2.6 存在的不足或有待深入研究的问题

章节 1.2.3 中提到了基于路网的城市区块划分，如文献^[16]中以辽宁省本溪市为例，配合核密度法、公众认知度以及区块大小等因素推算出城市功能区分类，如图 1-3。但是路网的颗粒度太粗的同时，一些区块如小学校、商场等建筑，并没有用街道去划分界限，导致并不能够较为准确地判断该区域的功能区类型。

如果采用文献^[8]中提到的通过栅格化城市来描述区块的类别，虽然可以在较细颗粒度的情况下分析功能区，但是同样需要注意的是，该方法需要考验研究人员对栅格长宽的确定，不同大小的栅格以及栅格兴趣点 POI 的数据量都会成为影响区块分类的因素。

上面提到的文献中，描述的用户画像多为工作地、居住地、通勤时间等交通、城区规划方向的标签，而缺少对用户是否有工作、工作时长、是否经常加班、使用 APP 偏好等画像的分析，这些都是有待本研究深入研究的问题。

1.3 本课题的主要研究内容概述

1.3.1 数据分析与可用性判断

本研究首先要对中国移动提供的数据进行统一的描述与判断，比如用户与

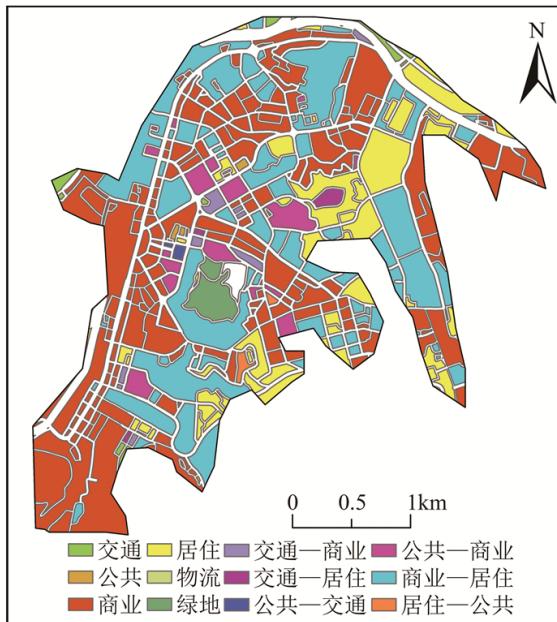


图 1-3 辽宁省本溪市功能区分布图

基站的连接数据，每天有多少条用户，用户唯一标识符使用手机号码还是 ICCID (Integrate Circuit Card Identity，集成电路卡识别码即手机 SIM 卡卡号)，连接时间的数据类型，字符串或是到 1970 年 1 月 1 日零点的毫秒数等，都需要研究人员仔细分析、做下标记，为之后的研究和处理提供数据轮廓。

同时提取重要的信息表，比如研究用户 APP 偏好画像时，用户 APP 使用数据表和 APP 的描述信息表（如类别，名称，唯一标识符等）都是关键数据，需要认真筛选并且做出数据流向图为数据的分析开发奠定基础。

1.3.2 大数据下 K-Means 聚类方法的分析和优化

由于 K-Means 的时间复杂度为 $O(KNTD)$ ， K 表示聚类数量， N 表示被聚类元素数量， T 表示迭代次数， D 表示距离的计算复杂度，具体为样本的特征数量，总体来说是关于被聚类元素的总数量的时间复杂度。为了平衡执行速度和聚类效果，本研究会比较两种 K-Means 优化算法，K-Means++ 和 Minibatch K-Means，从中选出一个各方面均衡的聚类优化方法。

1.3.3 哈尔滨市栅格化处理以及对栅格化后的街区类型分析

综合章节 1.2.6 中提到诸多因素，本研究选择颗粒度更小的栅格化城市处理方式。首先通过爬虫程序获取哈尔滨市区的兴趣点信息表，然后按照不同的标准对城市进行栅格化处理。对于每次栅格化后的城市，根据每一格中 POI 兴趣点的数量和类型，建立不同的模型来表示这一格中的兴趣点分布情况，接着

做一次 K-Means 聚类分析，得到哈尔滨市城市功能区类型的聚类结果。

对于每一次聚类结果，对比哈尔滨市真实的城市地图，分析和判断聚类结果的准确性，调整栅格的长宽，以及描述 POI 兴趣点在每一个栅格中的模型，以求达到一个类型精准、大小合适的城市功能区类型分布描述。

1.3.4 大数据系统及其套件的分析和比较

目前较为常用的大数据计算框架有：基于最早由雅虎开源的 MapReduce 流程；基于 HDFS 和 MapReduce 但内部进行优化同时提供了方便编写成 SQL 语句的 Hive 数据库；以及通过 DAG 和内存优化的 Spark 计算框架。本研究中的数据部分保存在 HDFS 和 Hive 中，为了提高计算速度、降低程序复杂度，该研究会对比这三种工作流程，并做出适合本项目的选择配比。

1.3.5 基于手机信令数据的用户驻留点判定

原始数据中并不包含用户位置的准确信息，只能通过用户对基站的连接序列来确定用户的驻留点偏好。而基站的覆盖范围很广，如一整天待在哈尔滨工业大学一校区十八公寓的学生，能连接到西北方向曲线街附近的一个基站，间隔约 500 米，所以为了更为准确的描述用户的位置信息，需要对基站连接序列进行去震荡处理。对于一个序列 $A \rightarrow B \rightarrow C$ ，通过一定的判断标准，认定 B 点属于漂移点而不是用户真正所前往的点，从而将其剔除，提高用户驻留点结果的可信度。

1.3.6 基于手机信令数据的不同时间段用户访问街区类型偏好

排除掉非个人用户，哈尔滨市一天月 400 万用户产生 4000 余万条约 1.2TB 左右的基站连接数据量，而该研究需要讨论一定时间段，如一个月内的基站连接情况，如此大规模的数据需要研究人员编写大数据计算程序如 MapReduce 或 Spark 任务进行分析。利用章节 1.3.5 提到的驻留点判定方法，筛选出可信的用户位置信息，合并给定时间段内的数据。

同时为了准确表达用户画像，还要将基站连接频次按四个时间段：工作、午休、加班、晚休进行统计，最后得到四个时间段中访问的不同的基站序列，再根据章节 1.3.3 提出的栅格化城市功能区，统计出用户在四个时间段对不同类型区块的访问偏好度，最终得到用户的移动语义。

1.3.7 基于手机应用数据的不同时间段用户手机应用类型偏好

类似于用户基站连接数据，用户的手机应用使用数据条数即使按天计算也是非常庞大的，同样需要编写基于大数据框架下的程序来统计用户在不同时间段对实际应用类型的偏好。首先需要认清楚应用唯一标识符意义，然后根据在应用商城中爬取的数据，一一对应获得不同应用的详细信息，如一级类型、二级类型、评分、用户下载量等。接下来根据用户在不同时段使用某个应用的流量数据，统计用户每天在四个时间段对某类应用的偏好，最终得出用户在一段时间内对某类应用的偏好，由街区访问偏好和应用使用偏好进而刻画用户画像。

1.3.8 基于手机通信数据的用户社交习惯和偏好分析

同样的，还有中国移动公司提供的用户通信数据，包括电话呼叫和短信收发。分析数据结构做出大众的描述性统计，根据结果推算用户在社交中的地位，如接拨电话、收发短信的比率，通话对端归属地比率，通信的总时长，超过某个时长的天数等等，最终得到关于用户社交的习惯和评价。

1.3.9 基于手机信令数据的用户职业特征与家庭特征分析

经过上述一系列处理，最终可以得到用户在四种时间段：工作、午休、加班、晚休对于驻留点的偏好、手机应用使用的偏好，连续一段时间内社交地位（拨入和拨出比率）等，从而刻画用户关于职业的特征。包括用户是否有一个稳定通勤路线的工作，工作地的类型（如果是文教相关，大概推測是哪一所学校），周平均工作时长，是否经常加班，以及加班时长，通过用户的社交地位（如拨入、短信接收占比，或用户 PageRank 在社交圈中的重要性等）可能可以知晓用户的职业地位相关特征。还能大致了解用户的家庭特征，如居住地周边类型，居家时间长短，在家时间对手机应用的偏好。

第2章 原始数据统计性描述、可用性分析和冗余与错误清理

2.1 原数数据分析与统计性描述

2.1.1 用户与基站连接记录数据分析

哈尔滨市移动平台注册 163,348,433 人次（即电话卡数量），仅 2020 年 1 月 1 日一天，就产生了 1,089,986,353 条连接基站的记录，排除掉 400 开头的虚拟号码、座机号码、企业用物联网卡、IMSI 为 13 位的非个人号码之后，4578524 位个人用户当天产生了 412000435 条约 1.2 TB 的基站连接记录。其中，每天有至少 100 位用户，在基站之间切换了 1500 次以上，最高能达到 2189 次，平均每位用户日均切换基站 80.0604 次。

表 2-1 基站连接信息

变量名	数据名称	数据类型	样例	备注
id	基站 id	string	3G44	基站唯一标识符
startTime	开始连接时间	Long	1577808000000	到 1970-01-01 00:00 的毫秒数
endTime	结束连接时间	Long	1577958050662	意义同上

数据保存在文件系统基于 HDFS 和 SQL 语句基于 MapReduce 的 Hive 数据库中，数据结构如表 2-1 所示。需要注意的是，由于技术问题，基站 id 并不一定存在于基站基本信息表中（可能非移动公司基站），以及开始和结束连接时间可能为 null 的问题，需要首先对无效的数据进行处理。

2.1.2 基站基本信息表分析与描述性统计

如表 2-2 中所描述的基站基本信息表的数据结构，主要考虑的是基站的唯一标识符 id，以及基站本身的经纬度坐标。

表 2-2 基站基本信息表

变量名	数据名称	数据类型	样例	备注
id	基站 id	string	3G44	基站唯一标识符
lon	基站经度	double	126.624709	绕城高速范围 [126.48, 126.83]
lat	基站纬度	double	46.770206	绕城高速范围 [45.64, 45.86]

基站唯一标识符和用户基站连接记录数据匹配，通过信息表中的经纬度信息，从而得到用户在某一时刻连接的基站的经纬度坐标，则能推算出用户在某时刻的大致位置信息。随机选取一位用户，如图 2-1 表示该用户在一段时间内的基站连接轨迹序列，同时能够得知该位用户连接某基站（图 2-2 中红圈标记附近一基站）3.4 天。

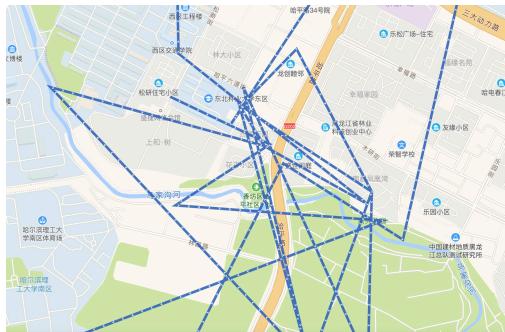


图 2-1 基站连接轨迹图



图 2-2 停留地点

不过由于基站的覆盖面积实在是太大，如图 2-3 中所示，2020 年 1 月 1 日，研究人员本人一天只前往过教化街附近的学生十八公寓、学苑楼、诚意楼，但是却能连接到学校西北部哈尔滨银行附近的一个基站长达 2.1 个小时，以及曲线街（图中未显示，连接次数排名第 5）移动营业厅附近一基站一小时左右。基站的连接轨迹并不能准确的表达用户的真实位置，后面几章将会如何利用基站经纬度坐标推算用户驻留点偏好。



图 2-3 连接的最多的基站

2.2 哈尔滨市 POI 兴趣点爬取、处理和描述分析

首先从高德地图上爬取哈尔滨市所有的 POI 兴趣点的相关信息，包括 POI 兴趣点的序号、名称、类型（如购物中心、餐厅、体育场、学校等）、POI 经纬

度信息，数据结构如表 2-3 所示。

表 2-3 哈尔滨市 POI 兴趣点数据结构

变量名	数据名称	数据类型	样例	备注
id	兴趣点 id	int	123	兴趣点唯一标识符
name	兴趣点名称	string	远大购物中心	名称描述
type	兴趣点类型	string	购物中心	类型描述
lon	POI 位置经度	double	126.556693	
lat	POI 位置维度	double	45.738066	

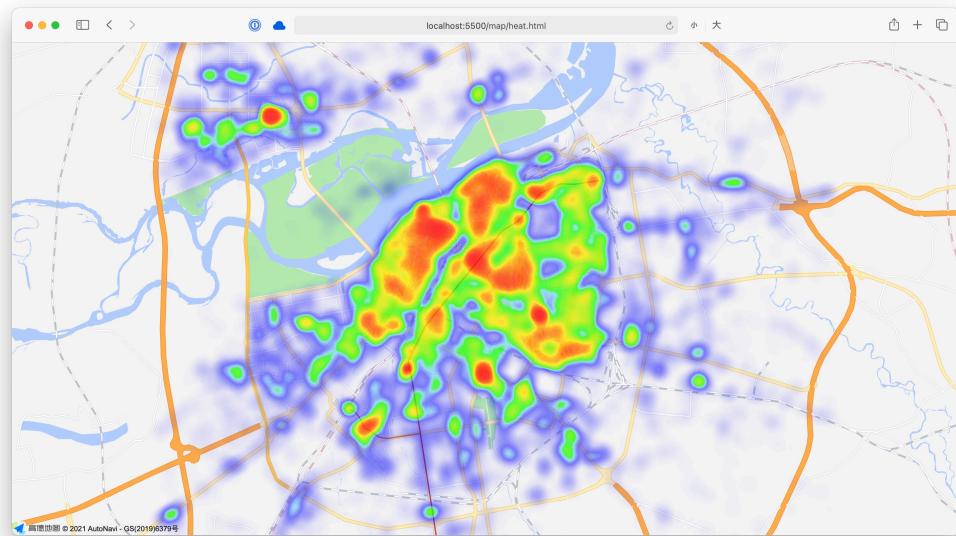


图 2-4 全市 POI 兴趣点分布热力图

一共 477932 个 POI 兴趣点，每个兴趣点间隔最小小于 1m（经纬度描述相同），最远一千米以上。基站分布如图 2-4 所示。根据用户基站连接数据统计数据、POI 分布热力图所描述，用户在环城高速以内的活动更为活跃同时用以描述用户移动语义的 POI 兴趣点也更多，所以我们选择哈尔滨市绕城高速以内的市区进行分析，如图 2-5 所示。

2.3 用户手机应用使用流量记录表分析

由于 4G 网络在中国的快速铺开，网民飞速增长的同时，移动端应用的使用也越来越频繁，哈尔滨市用户日手机应用流量使用记录数据表可达 600MB，包括记录的日期，用户 id（即手机号码），应用 id（应用商店中应用的唯一标识符），应用名称，基站经纬度，以及从 0 时到 23 时的流量使用数据（单位 Byte），数据结构如表 2-4 所示。

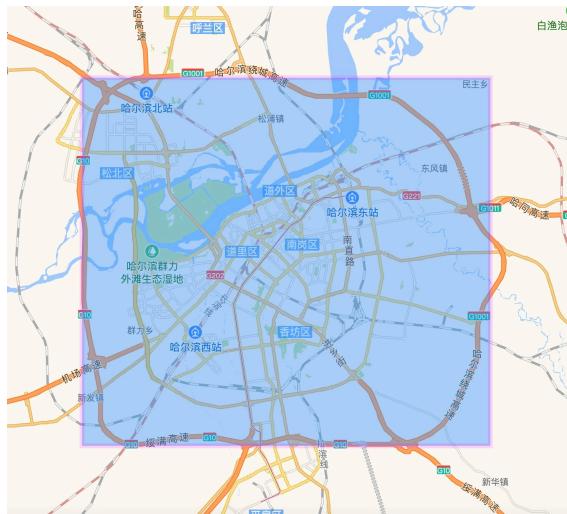


图 2-5 哈尔滨市绕城高速范围

表 2-4 哈尔滨市用户各个基站的分时流量数据表

变量名	数据名称	数据类型	样例	备注
date	日期	string	20200615	
user_id	用户 id	string	13912345678	用户唯一标识符
app_id	手机应用 id	string	C946	手机应用唯一标识符
app_name	应用名称	string	腾讯视频	
lon	POI 位置经度	double	126.556693	
lat	POI 位置维度	double	45.738066	
data	流量数据 (单位 B)	int	578	每个小时一个数据, 共 8 列

爬取手机应用商店中的数据（此处选择华为手机应用商店），经统计，一共 4563 款下载量明显的手机应用，数据维度过多，此处选择使用手机应用的二级类型数据进行降维处理，转为统计用户在某个时间段对某类应用的偏好。爬取的应用数据中包含：如为了和用户应用使用信息表匹配的应用唯一标识符 id，应用一级类型，应用二级类型，下载量，评分等，数据结构如表 2-5 所示。

表 2-5 手机应用基本信息表

变量名	数据名称	数据类型	样例	备注
app_id	手机应用 id	string	C946	手机应用唯一标识符
type_1	应用一级类型	string	社交软件	
type_2	应用二级类型	string	聊天	
download	下载量	int	50912769	截止到爬取时
score	评分	double	3.5	满分 5 分

2.4 用户通话和短信使用记录数据表分析

为了描述用户的社交地位和偏好，还需要使用用户通话和短信使用记录的

相关数据，通话数据包括统计的月份，拨出手机号码，接收手机号码，对端归属地，通话时长，通话次数，通话天数等，数据结构如表 2-6 所示。

表 2-6 手机应用基本信息表

变量名	数据名称	数据类型	单位
stat_mon	统计月份	string	
serv_no	手机号码	string	
opp_serv_no	对端号码	string	
opp_type	对端类型	string	
opp_region_code	对端归属地	string	
call_dur	通话时长	long	秒
calling_dur	通话时长（主叫）	long	秒
called_dur	通话时长（被叫）	long	秒
call_cnt	通话次数	long	次
calling_cnt	通话次数（主叫）	long	次
called_cnt	通话次数（被叫）	long	次
busy_call_cnt	忙时通话次数	long	次
idle_call_cnt	闲时通话次数	long	次
busy_dur	忙时通话时长	long	分钟
idle_dur	闲时通话时长	long	分钟
weekday_sum_call_cnt	工作日通话次数	long	次
weekday_work_sum_call_cnt	工作日上班时间通话次数	long	次
weekday_offwork_sum_call_cnt	工作日非上班时间通话次数	long	次
weekend_sum_call_cnt	周末通话次数	long	次
weekday_sum_call_dur	工作日通话时长	long	分钟
weekend_sum_call_dur	周末通话时长	long	分钟
call_days	通话天数	long	天
first_call_date	末次通话时间	string	
last_call_date	首次通话时间	string	

同理，用户点对点短信记录信息表中包含统计日期、用户类型、发送方号码、发送方归属区号、接收方号码、接收方归属区号、短信条数等信息，数据结构如表 2-7 所示。

表 2-7 手机应用基本信息表

变量名	数据名称	数据类型
stats_mon	统计月份	string
user_type	用户类型	string
send_serv_no	发送方手机号码	string
home_city_code	发送方归属区号	string
reci_serv_no	接收方手机号码	string
opp_city_code	接收方归属区号	string
gms_count	短信条数	long

2.5 本章小结

本章通过对原始数据的统计性描述和分析，首先确定了可用的数据库和能用以分析用户画像的数据的结构，统计数据中可能有误和缺失的信息，做好标记用以支持之后的正式处理中采用剔除或默认值的办法替换原值；同时统计数量大小和有效数据范围，如每日信息数量、统计区域范围（哈尔滨市绕城高速）等，为后续小样本的计算模型测试提供理论基础；

第3章 哈尔滨市城区POI兴趣点数据获分析和栅格化街区聚类分析

3.1 哈尔滨市城区POI兴趣点数据具体分析

为了能够描述用户的移动语义（即用户前往目的地的原因），同时由于基站的覆盖面积过于宽广，并不能直接将基站的位置认为是用户的目的位置，文献^{[17][18]}中 Phithakkinukoon 和 Dashjor 指出，用户的移动语义是由某一个范围内的兴趣点的分布情况决定的，如果得知了该区块中各类型兴趣点的分布，那么我们就能大致推算前往该区域的用户的目的。

如图 3-1，为哈尔滨工业大学一校区周边 POI 兴趣点分布地图，若一个区域中分布着教学楼、公寓、食堂、超市等兴趣点，



图 3-1 哈尔滨工业大学一校区周边 POI 兴趣点分布图

3.2 基于路网和栅格化的功能区聚类分析模型对比

3.3 功能区的栅格化参数的分析和对比

3.4 本章小结

第4章 大数据领域下聚类方法的对比和在哈尔滨城市功能区的聚类结果分析

4.1 DBSCAN 聚类算法概述和分析

4.2 K-Means 聚类算法概述

K-Means 算法的思想非常简单，对于样本集，按照距离划分为若干个簇，让簇内距离最小化，簇间距离最大化。如果有集合 G ，假设划分的簇为 (U_1, U_2, \dots, U_m) ，则我们的目标则是最小化误差 E ：

$$E = \sum_{i=1}^m \sum_{x \in U_i} \|x - \mu_i\|_2^2$$

其中， μ_i 是每个簇的质心，通常是一个均值向量。传统的算法通常是随机选择 k 质心（即要分成多少类），经常会根据先验经验做一个 k 值选择，对于每一次迭代，求每个点到每个质心的距离，最小距离则为改簇，重复迭代，直到质心没有变化。

4.3 K-Means 初始质心选择优化 K-Means++ 算法

在最基础的 K-Means 聚类中，最初的质心的选择通常是随机选择，而在 K-Means++ 中，则是优化最初的质心选择，优化策略也比较简单，方法如下：

1. 随机选择一个点作为第一个质心 μ_1 ;
2. 对于寻找第 i 个质心，则计算数据集中没有作为质心的点，分别到前 $i-1$ 个质心的最小距离 $D_j = \min(Distance(j, i))$;
3. 有概率的选择第 i 个质心，原则是 D_j 越大的点作为第 i 个质心的概率越大;
4. 重复，选出 k 个质心;

但是 K-Means 的优化只能优化迭代的次数，而 K-Means 的算法复杂度是 $O(MNKD)$ ， N 为样本数量， K 为聚类个数， D 为数据维度， M 与数据集本身的分布情况和中心点有关。此处有引用，K-Means++ 的优化对于本研究 4 百万的原数样本数量来说，并不是特别的高效。

4.4 大样本优化 Mini Batch K-Means 算法

在上上节所述的 K-Means 的传统算法中，每次迭代都需要计算数据集中每个节点到质心的距离，当数据量达到 10 万以上时，即使加上诸如 elkan K-Means 优化此处有引用也无法应付。此时，就要用到 Mini Batch K-Means 算法。

如同字面意思，Mini Batch，原理十分简单，从数据集中随机抽取一部分样本，做普通的 K-Means 聚类，得到的质心就被认定为最终的质心。由于 N 缩小的很快，导致算法每次跌打的速度大大加快，当然，精度也会降低，不过经过证明此处有引用，仍然在我们的可接受范围内，而且通常会进行多次 Mini Batch K-Means 聚类，选择其中最优的聚类结果。

如图的一次 K-Means 与 Mini Batch K-Means 对比，二者的运行时间相差两倍多，但最终的结果差异确非常小，图中第三张（Difference）的粉色错误点。

4.5 哈尔滨市功能区聚类结果分析和展示

4.6 本章小结

第5章 大数据系统不同计算框架的分析与对比

第6章 基于手机信令数据的用户驻留点、手机应用 偏好以及社交地位分析

第7章 基于驻留点、应用偏好和社交地位分析结果 的用户职住特征分析

结 论

学位论文的结论作为论文正文的最后一章单独排写，但不加章标题序号。

结论应是作者在学位论文研究过程中所取得的创新性成果的概要总结，不能与摘要混为一谈。博士学位论文结论应包括论文的主要结果、创新点、展望三部分，在结论中应概括论文的核心观点，明确、客观地指出本研究内容的创新性成果（含新见解、新观点、方法创新、技术创新、理论创新），并指出今后进一步在本研究方向进行研究工作的展望与设想。对所取得的创新性成果应注意从定性和定量两方面给出科学、准确的评价，分（1）、（2）、（3）…条列出，宜用“提出了”、“建立了”等词叙述。

参考文献

- [1] 粟翹楚. 分布式存储打开千亿级市场深入推动行业数字化转型[J]. 上海商业, 2021(04): 5.
- [2] Sanjay Ghemawat H G. The Google File System[J], 2003.
- [3] DeWitt J. MapReduce: A major step backwards - Washington[J]. Certified Special Events Professional, 2008, 544(21).
- [4] 孙冰. 谁在为数字经济铺路架桥? 中国 5G 和中国手机[J]. 中国经济周刊, 2020(20): 21-24.
- [5] 孔扬鑫. 基于手机信令数据的人口流动分析[D]. [S.1.]: 华东师范大学, 2017.
- [6] 刘奕杉. 基于运营商数据的用户位置预测系统研究[D]. [S.1.]: 北京邮电大学, 2019.
- [7] 常在春. 基于手机信令数据的延庆区职业空间关系研究[D]. [S.1.]: 北京林业大学, 2019.
- [8] 杨振山, 苏锦华, 杨航, et al. 基于多源数据的城市功能区精细化研究——以北京为例[J]. 地理研究, 2021, 40(02): 477-494.
- [9] 毋亭 . 基于 POI 数据的城市功能区划分和识别[J]. 辽宁大学学报(自然科学版), 2021, 48(01): 28-37.
- [10] 王 , 王 , 刘 , et al. 基于手机信令数据的北京市职住空间分布格局及匹配特征[J]. 地理科学进展, 2020, 39(12): 2028-2042.
- [11] 张景奇; 史文宝; 修春亮;. poi 数据在中国城市研究中的应用[J]. 地理科学, 2021, 41(01): 140-148.
- [12] 邵钰涵; 殷雨婷; 薛贞颖;. 基于街景大数据的北京、上海街景舒适度评价及比较[J]. 风景园林, 2021, 28(01): 53-59.
- [13] 李斌; 张金焕; 封靖川;. 基于 minibatch-kmeans 和 xgboost 算法的大型场所 wlan 室内定位的方法[J]. 数字技术与应用, 2018, 36(10): 139-140.
- [14] 王岩; 范子贤; 李成名; 戴昭鑫;. 利用手机信令数据刻画不同人物画像[J]. 测绘通报, 2021(01): 84-89.
- [15] 郑宝鑫; 周雪松; 李斌; 唐宇;. 基于用户画像、信令挖掘技术的手机游戏产品推广[C] // . 2010: 4.
- [16] 薛冰, Bing X U E, 赵冰玉, et al. 辽宁 POI 功能区识别方法[J], 2020.

- [17] Phithakkitnukoon S, Horanont T, Di Lorenzo G, et al. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data[C/OL] // Salah A A, Gevers T, Sebe N, et al. Lecture Notes in Computer Science: Human Behavior Understanding. Berlin, Heidelberg: Springer, 2010: 14-25. http://dx.doi.org/10.1007/978-3-642-14715-9_3.
- [18] Dashdorj Z, Sobolevsky S, Serafini L, et al. Semantic Enrichment of Mobile Phone Data Records Using Background Knowledge[J/OL]. Knowledge-Based Systems, 2018, 143: 225-235. <http://dx.doi.org/10.1016/j.knosys.2017.11.038>.

哈尔滨工业大学本科毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学攻读学士学位期间，所提交的毕业设计（论文）《基于移动大数据的用户职住特征分析》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名：

日期： 年 月 日

致 谢

衷心感谢导师 XXX 教授对本人的精心指导。他的言传身教将使我终生受益。

.....

感谢哈工大 L^AT_EX 论文模板 HI_THESIS !

附录 1 外文资料原文

The title of the English paper

Abstract: As one of the most widely used techniques in operations research, *mathematical programming* is defined as a means of maximizing a quantity known as *objective function*, subject to a set of constraints represented by equations and inequalities. Some known subtopics of mathematical programming are linear programming, nonlinear programming, multiobjective programming, goal programming, dynamic programming, and multilevel programming^[1].

It is impossible to cover in a single chapter every concept of mathematical programming. This chapter introduces only the basic concepts and techniques of mathematical programming such that readers gain an understanding of them throughout the book^[2,3].

1.1 Single-Objective Programming

The general form of single-objective programming (SOP) is written as follows,

$$\begin{cases} \max f(x) \\ \text{subject to:} \\ g_j(x) \leq 0, \quad j = 1, 2, \dots, p \end{cases} \quad (123)$$

which maximizes a real-valued function f of $x = (x_1, x_2, \dots, x_n)$ subject to a set of constraints.

Definition 1.1 In SOP, we call x a decision vector, and x_1, x_2, \dots, x_n decision variables. The function f is called the objective function. The set

$$S = \{x \in \Re^n \mid g_j(x) \leq 0, j = 1, 2, \dots, p\} \quad (456)$$

is called the feasible set. An element x in S is called a feasible solution.

Definition 1.2 A feasible solution x^* is called the optimal solution of SOP if and only if

$$f(x^*) \geq f(x) \quad (1-1)$$

for any feasible solution x .

One of the outstanding contributions to mathematical programming was known as the Kuhn-Tucker conditions¹⁻². In order to introduce them, let us give some definitions.

An inequality constraint $g_j(x) \leq 0$ is said to be active at a point x^* if $g_j(x^*) = 0$. A point x^* satisfying $g_j(x^*) \leq 0$ is said to be regular if the gradient vectors $\nabla g_j(x)$ of all active constraints are linearly independent.

Let x^* be a regular point of the constraints of SOP and assume that all the functions $f(x)$ and $g_j(x)$, $j = 1, 2, \dots, p$ are differentiable. If x^* is a local optimal solution, then there exist Lagrange multipliers λ_j , $j = 1, 2, \dots, p$ such that the following Kuhn-Tucker conditions hold,

$$\begin{cases} \nabla f(x^*) - \sum_{j=1}^p \lambda_j \nabla g_j(x^*) = 0 \\ \lambda_j g_j(x^*) = 0, \quad j = 1, 2, \dots, p \\ \lambda_j \geq 0, \quad j = 1, 2, \dots, p. \end{cases} \quad (1-2)$$

If all the functions $f(x)$ and $g_j(x)$, $j = 1, 2, \dots, p$ are convex and differentiable, and the point x^* satisfies the Kuhn-Tucker conditions (1-2), then it has been proved that the point x^* is a global optimal solution of SOP.

1.1.1 Linear Programming

If the functions $f(x)$, $g_j(x)$, $j = 1, 2, \dots, p$ are all linear, then SOP is called a *linear programming*.

The feasible set of linear is always convex. A point x is called an extreme point of convex set S if $x \in S$ and x cannot be expressed as a convex combination of two points in S . It has been shown that the optimal solution to linear programming corresponds to an extreme point of its feasible set provided that the feasible set S is bounded. This fact is the basis of the *simplex algorithm* which was developed by Dantzig as a very efficient method for solving linear programming.

表 1-1 *

Table 1

This is an example for manually numbered table, which would not appear in the list of tables

Network Topology		# of nodes	# of clients			Server
GT-ITM	Waxman Transit-Stub	600	2%	10%	50%	Max. Connectivity
	Inet-2.1	6000				
	ABCDEF					

Roughly speaking, the simplex algorithm examines only the extreme points of the

feasible set, rather than all feasible points. At first, the simplex algorithm selects an extreme point as the initial point. The successive extreme point is selected so as to improve the objective function value. The procedure is repeated until no improvement in objective function value can be made. The last extreme point is the optimal solution.

1.1.2 Nonlinear Programming

If at least one of the functions $f(x), g_j(x), j = 1, 2, \dots, p$ is nonlinear, then SOP is called a *nonlinear programming*.

A large number of classical optimization methods have been developed to treat special-structural nonlinear programming based on the mathematical theory concerned with analyzing the structure of problems.

Now we consider a nonlinear programming which is confronted solely with maximizing a real-valued function with domain \Re^n . Whether derivatives are available or not, the usual strategy is first to select a point in \Re^n which is thought to be the most likely place where the maximum exists. If there is no information available on which to base such a selection, a point is chosen at random. From this first point an attempt is made to construct a sequence of points, each of which yields an improved objective function value over its predecessor. The next point to be added to the sequence is chosen by analyzing the behavior of the function at the previous points. This construction continues until some termination criterion is met. Methods based upon this strategy are called *ascent methods*, which can be classified as *direct methods*, *gradient methods*, and *Hessian methods* according to the information about the behavior of objective function f . Direct methods require only that the function can be evaluated at each point. Gradient methods require the evaluation of first derivatives of f . Hessian methods require the evaluation of second derivatives. In fact, there is no superior method for all problems. The efficiency of a method is very much dependent upon the objective function.

1.1.3 Integer Programming

Integer programming is a special mathematical programming in which all of the variables are assumed to be only integer values. When there are not only integer variables but also conventional continuous variables, we call it *mixed integer programming*. If all the variables are assumed either 0 or 1, then the problem is termed a *zero-one programming*. Although integer programming can be solved by an *exhaustive enumeration* theoretically,

it is impractical to solve realistically sized integer programming problems. The most successful algorithm so far found to solve integer programming is called the *branch-and-bound enumeration* developed by Balas (1965) and Dakin (1965). The other technique to integer programming is the *cutting plane method* developed by Gomory (1959).

Uncertain Programming (BaoDing Liu, 2006.2)

References

NOTE: These references are only for demonstration. They are not real citations in the original text.

- [1] Donald E. Knuth. The \TeX book. Addison-Wesley, 1984. ISBN: 0-201-13448-9
- [2] Paul W. Abrahams, Karl Berry and Kathryn A. Hargreaves. \TeX for the Impatient. Addison-Wesley, 1990. ISBN: 0-201-51375-7
- [3] David Salomon. The advanced \TeX book. New York : Springer, 1995. ISBN:0-387-94556-3

附录 2 外文资料的调研阅读报告或书面翻译

英文资料的中文标题

摘要：本章为外文资料翻译内容。如果有摘要可以直接写上来，这部分好像没有明确的规定。

2.1 单目标规划

北冥有鱼，其名为鲲。鲲之大，不知其几千里也。化而为鸟，其名为鹏。鹏之背，不知其几千里也。怒而飞，其翼若垂天之云。是鸟也，海运则将徙于南冥。南冥者，天池也。

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \quad (123)$$

吾生也有涯，而知也无涯。以有涯随无涯，殆已！已而为知者，殆而已矣！为善无近名，为恶无近刑，缘督以为经，可以保身，可以全生，可以养亲，可以尽年。

2.1.1 线性规划

庖丁为文惠君解牛，手之所触，肩之所倚，足之所履，膝之所倚，砉然响然，奏刀砉然，莫不中音，合于桑林之舞，乃中经首之会。

表 2-1 *

表 1 这是手动编号但不出现在索引中的一个表格例子

Network Topology		# of nodes	# of clients			Server
GT-ITM	Waxman Transit-Stub	600	2%	10%	50%	Max. Connectivity
	Inet-2.1	6000				
	ABCDEF					

文惠君曰：“嘻，善哉！技盖至此乎？”庖丁释刀对曰：“臣之所好者道也，进乎技矣。始臣之解牛之时，所见无非全牛者；三年之后，未尝见全牛也；方今之时，臣以神遇而不以目视，官知止而神欲行。依乎天理，批大郤，导大窾，因其固然。技经肯綮之未尝，而况大郤乎！良庖岁更刀，割也；族庖月更刀，折也；今臣之刀十九年矣，所解数千牛矣，而刀刃若新发于硎。彼节者有间而刀刃者无厚，以无厚入有间，恢恢乎其于游刃必有余地矣。是以十九年而刀刃若

新发于硎。虽然，每至于族，吾见其难为，怵然为戒，视为止，行为迟，动刀甚微，**砉**然已解，如土委地。提刀而立，为之而四顾，为之踌躇满志，善刀而藏之。”

文惠君曰：“善哉！吾闻庖丁之言，得养生焉。”

2.1.2 非线性规划

孔子与柳下季为友，柳下季之弟名曰盜跖。盜跖从卒九千人，横行天下，侵暴诸侯。穴室枢户，驱人牛马，取人妇女。贪得忘亲，不顾父母兄弟，不祭先祖。所过之邑，大国守城，小国入保，万民苦之。孔子谓柳下季曰：“夫为人父者，必能诏其子；为人兄者，必能教其弟。若父不能诏其子，兄不能教其弟，则无贵父子兄弟之亲矣。今先生，世之才士也，弟为盜跖，为天下害，而弗能教也，丘窃为先生羞之。丘请为先生往说之。”

柳下季曰：“先生言为人父者必能诏其子，为人兄者必能教其弟，若子不听父之诏，弟不受兄之教，虽今先生之辩，将奈之何哉？且跖之为人也，心如涌泉，意如飘风，强足以距敌，辩足以饰非。顺其心则喜，逆其心则怒，易辱人以言。先生必无往。”

孔子不听，颜回为驭，子贡为右，往见盜跖。

2.1.3 整数规划

盜跖乃方休卒徒大山之阳，脍人肝而**啖**之。孔子下车而前，见谒者曰：“鲁人孔丘，闻将军高义，敬再拜谒者。”谒者入通。盜跖闻之大怒，目如明星，发上指冠，曰：“此夫鲁国之巧伪人孔丘非邪？为我告之：尔作言造语，妄称文、武，冠枝木之冠，带死牛之胁，多辞缪说，不耕而食，不织而衣，摇唇鼓舌，擅生是非，以迷天下之主，使天下学士不反其本，妄作孝弟，而侥幸于封侯富贵者也。子之罪大极重，疾走归！不然，我将以子肝益昼**啖**之膳。”

附录 3 其它附录

前面两个附录主要是给本科生做例子。其它附录的内容可以放到这里，当然如果你愿意，可以把这部分也放到独立的文件中，然后将其到主文件中。