文本复制检测报告单(全文标明引文)

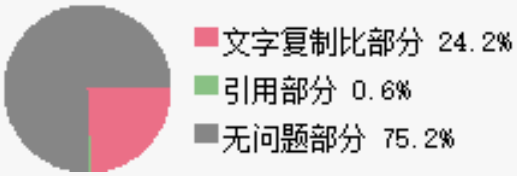
№:ADBD2018R_2018050917371820180518091825428659975004 检测时间:2018-05-18 09:18:25

检测文献: 1526567738385_陶超权_出租车营运大数据分析
作者: 陶超权
检测范围: 中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
大学生论文联合比对库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
CNKI大成编客-原创作品库
个人比对库
时间范围: 1900-01-01至2018-05-18

检测结果

总文字复制比：24.8% 跨语言检测结果：0%
去除引用文献复制比：24.2% 去除本人已发表文献复制比：24.8%
单篇最大文字复制比：9.1%（基于Web的校园办公自动化系统设计与实现）

重复字数： [4842] 总段落数： [6]
总字数： [19543] 疑似段落数： [5]
单篇最大重复字数： [1783] 前部重合字数： [272]
疑似段落最大重合字数：[2097] 后部重合字数： [4570]
疑似段落最小重合字数：[142]



指 标：☐疑似剽窃观点 ☑疑似剽窃文字表述 ☐疑似自我剽窃 ☐疑似整体剽窃 ☐过度引用

表格： 0 公式： 0 疑似文字的图片： 0 脚注与尾注：0

7.7%（142） 中英文摘要等（总1856字）
8.1%（192） 第一章引言（总2385字）
44.9%（2035） 第二章大数据隐私保护及数据挖掘相关技术（总4534字）
46.6%（2097） 第三章数据获取及预处理（总4500字）
7.7%（376） 第四章基于MapReduce框架的热点区域挖掘及可视化（总4884字）
0%（0） 第五章总结与展望（总1384字）

（注释：无问题部分 文字复制比部分 引用部分）

1. 中英文摘要等

总字数：1856

相似文献列表 文字复制比：7.7%(142) 疑似剽窃观点：(0)

1	交通与物流工程学院-交工132班-220130752-张起明-乌鲁木齐市火车南站出租车载客率调查班 -《大学生论文联合比对库》- 2017-05-11	7.5%（139） 是否引证：否
2	关于出租车载客地点序列推荐技术的研究 陈轶非;李治军;姜守旭; -《智能计算机与应用》- 2013-12-01	1.6%（30） 是否引证：否

原文内容 红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

出租车营运大数据分析

摘要

随着国民经济的发展以及人民生活水平的提高，出租车越来越成为人们出行不可缺少的工具，无论是周内的上班工作还是周末的出行娱乐，出租车都在扮演着越来越重要的角色。在万物联网的今天，基本每辆出租车都安装了GPS终端，这些终端装置会定时向数据库发送出租车实时的行驶状态，营运状态等。这些数据将会成为我们用来挖掘有用信息的重要资源。

通过对出租车营运数据进行处理和分析，可以挖掘出租车热点载客区域，为出租车司机进行智能推荐，优化出租车公司的资源调度，以达到更高的效益。

本文借助MongoDB数据库提供的MapReduce框架，首先对营运数据进行预处理，剔除空数据和重复数据；然后利用MongoDB对MapReduce的支持，通过DBSCAN聚类算法，按照是否是工作日，以及每天不同的时间段，将数据分类，对营运数据进行聚类分析；最后得到每天不同时段内的乘车热点区域，为司机和乘客提供个性化推荐。

关键词：出租车，大数据，聚类，MapReduce,MongoDB，DBscan

Abstract

With the development of the national economy and the improvement of people’s living standards, taxis have become an indispensable tool for people to travel. Taxis are playing an increasingly important role every day. Every taxi has a GPS terminal installed with the time of Internet of things coming. The data of each operation will be submitted to the database for storage. These terminal devices will periodically send taxis the current driving status and operating status of the taxi. This data will become an important resource for us to mine useful information.

Through the processing and analysis of taxi operating data, we can get taxi hotspot areas, which can be recommended to taxi drivers, and also, taxi company's resource scheduling can be optimized to achieve higher efficiency.

Using the MapReduce framework provided by the MongoDB database, this thesis preprocess with the operational data firstly, eliminate null and duplicate data, and then use MongoDB's support for MapReduce, through the DBSCAN clustering algorithm, according to whether it is a working day and a different time period each day. , Classifying data, clustering analysis of operational data, and finally obtaining hotspots for driving within different time periods each day, providing personalized recommendations for drivers and passengers.

Key Words：Taxi GPS Data; big data; MapReduce; MongoDB; DBSCAN

目录

第一章引言

研究背景及意义

随着社会经济的发展与人民生活水平的提高，城市中的交通路网越来越完善，人们出行也越来越方便。作为城市中的主要交通工具之一的出租车，其所占比例也在逐年上升，在人们出行方式中所占的比重越来越高。近几年由于“滴滴打车”等平台的出现，除了传统的路边叫停出租车外，“网约车”也应用而生，极大地方便了市民的出行。各种类型的出租车基本都装有GPS设备，在营运的同时向服务器发送轨迹等相关数据，这为我们提供了海量的可使用数据。但同时，出租车的大量普及和对客源时空分布全面认识的缺乏，极有可能造成出租车过度空驶，同一区域出租车数量过多，或因无法确定未来一段时间内热点乘车区域而不能获得更大收益等问题。

在此背景下，如果能够充分利用出租车GPS终端所提供的轨迹数据和营运数据，对其进行分析和处理，宏观上得到某一时段内乘客的热点乘车区域，微观上结合数据库以及特定的算法对车辆信息进行搜索，然后集成为合多功能于一体的数据和信息可视化出租车信息管理查询平台。还可以利用大数据挖掘的方法，通过对历史数据的分析，挖掘出某段时间内的热点乘车区域，向出租车司机个性化推荐，整合出租车资源，提高司机收益，方便乘客出行。谷歌在2004年提出的MapReduce[1]框架通过映射和化简两个步骤将一个大作业分解为若干并行的小作业，借助此框架，通过具体的聚类算法，可以更方便的挖掘出热点区域。与此同时，分析和处理过程中所使用的海量数据极有可能包含用户敏感信息。在大数据研究如火如荼的今天，如何在数

据挖掘的同时保护好用户的隐私成为了备受关注的问题，而隐私保护又具体分为位置隐私保护、标识符匿名保护、连接关系匿名保护等[2]。本文将在出租车营运数据挖掘过程中，对车牌号进行隐私保护，以防止司机和乘客信息泄露。

研究现状

1.1.1. 大数据及其隐私保护研究现状

大数据尚未有一个公认的定义，目前比较有代表性的是3V[3]定义，即高容量(Data Volume)、高速度(Data Velocity)、多类型(Data Variety)。其处理过程包含数据采集与预处理，数据分析，数据解释三部分。

近年来数据挖掘方法被广泛应用在各个领域，尤其是医疗健康和公共卫生领域。鲁淳欣通过数据挖掘和统计学方法分析针灸治疗乳腺增生症的选穴规律[21],最后得出的结论和中医理论基本符合，为中医疗法提供了数据支持。孙轶轩通过对我国道路交通事故的分析和处理，采用分类思想，基于ARIMA和SVR的预测模型对交通事故中财产损失、死亡人数、受伤人数和事故起数进行了预测和证实[22]。熊亚军等人基于KNearest Neighbor数据挖掘算法对北京地区雾霾等级进行了预测[23]。

目前国内外对于大数据隐私保护主要有以下几个方面的研究。基于位置的服务LBS给人们提供了许多便利，但同时也威胁着用户的位置隐私，其中典型的位置隐私保护技术有基于k-匿名的位置隐私保护技术，通过对用户C在某一时段内的位置信息进行匿名和泛化处理，使得在该时段内至少有k个用户都在C所处的位置范围内；侯士江***基于用户—匿名器—LBS架构，分别针对欧氏空间和路网环境的位置隐私保护、常见查询等问题进行了研究[4]；***车延辙***围绕着改进算法效率和提高抵御攻击能力的目标，研究移动点对点体系结构下的用户位置隐私保护关键技术[5]；***Gupta针对k-匿名中存在可信第三方的主要限制问题，提出了通过若干个同等级用户相互共享一些位置信息的方法来扩展可信第三方[6]。

1.1.2. 出租车数据应用研究现状

近年来国内外基于出租车数据所做研究主要针对乘车热点分析以及居民出行特征分析等方面。周洋通过分析出租车的起始点和终点，提取乘客的出行热度和及其空间规律，将居民区进行聚类，得出不同居民区的不同背景信息[7]；王郑委基于Hadoo平台通过K-Means聚类算法，挖掘出租车载客热点区域[8]；Putri F K 基于MongoDB和Spark提出了分布式的热点乘车区域搜索系统[9]；Deri J A利用纽约市的出租车轨迹信息，基于Dijkstra算法在时间和空间上优化了其对轨迹的计算的复杂度[10].

1.1. 论文内容

本文应用南京市出租车的营运信息，对不同天不同时间段内热点乘车区域进行挖掘，主要包括数据获取，数据预处理及隐私保护，聚类算法实现和热点区域可视化四个步骤。首先需要将出租车公司的营运数据从关系型数据库导入到非关系型数据库MongoDB中；由于原始数据包含许多空值字段和重复数据，需要对数据进行清洗，并在清洗过程中对用户敏感信息进行保护；为挖掘出南京市一天内不同时段乘车热点区域，本文采用基于MapReduce框架的DBSCAN聚类算法，将一周时间划分为工作日和非工作日，同时将一天划分为不同时段，通过使用不同的阈值，来找到最佳乘车热点区域；最后通过高德地图API将乘车热点区域可视化。

论文结构

论文一共分为五个章节。

第一章引言——主要介绍了论文的研究背景和意义，以及国内外研究现状，包括大数据及其隐私保护的研究现状和出租车数据利用的研究现状，最后介绍了本文的内容和结构。

第二章大数据隐私保护及数据挖掘相关技术——主要介绍了大数据隐私保护的常见方法，并对数据挖掘做了概述并介绍了其挖掘过程。

第三章数据获取及预处理——首先介绍了数据来源和数据格式，然后介绍了数据预处里过程，包括脏数据的剔除、数据得到导入导出以及用户敏感信息保护。

第四章基于MapReduce框架的热点区域挖掘及可视化——首先实现了DBSCAN聚类算法，然后通过不同阈值的取值分析，得到最佳热点区域，最后通过高德地图API将热点区域呈现出来。

第五章总结与展望——对本文所做工作、不足之处以及将来可改进之处进行了总结。

2	DBSCAN聚类算法原理 - xuanyuansen的专栏 - CSDN博客 - 《网络（http://blog.csdn.net）》- 2017	是否引证：否
3	2220132708-杨俊波-基于浏览器数据的文本聚类程序设计 杨俊波 - 《大学生论文联合比对库》- 2017-06-10	5.9%（269） 是否引证：否
4	1_王震_一种基于密度最大值聚类算法的研究与应用 王震 - 《大学生论文联合比对库》- 2017-05-11	5.6%（254） 是否引证：否
5	基于密度的聚类方法研究 陈雪儿 - 《大学生论文联合比对库》- 2014-05-23	5.0%（225） 是否引证：否
6	基于密度的聚类方法研究 陈雪儿 - 《大学生论文联合比对库》- 2014-05-27	5.0%（225） 是否引证：否
7	基于DBScan算法的网页聚类分析 沈逸帆 - 《大学生论文联合比对库》- 2017-05-21	4.8%（218） 是否引证：否
8	探索推荐引擎内部的秘密，第3部分：深入_sofiafighting - 《网络（http://blog.sina.com）》- 2012	4.5%（202） 是否引证：否
9	推荐引擎相关算法 - 聚类_W老泉 - 《网络（http://blog.sina.com）》- 2013	4.5%（202） 是否引证：否
10	探索推荐引擎内部的秘密，第 3 部分: 深入推荐引擎相关算法 - 聚类_SomeThere - 《网络（http://blog.sina.com）》- 2014	4.5%（202） 是否引证：否
11	3130931034-王杰-k-modes聚类算法的实现与应用 王杰 - 《大学生论文联合比对库》- 2017-06-12	4.4%（200） 是否引证：否
12	数据挖掘中基于遗传算法的聚类方法应用研究 吴多比(导师：朱晓红) - 《重庆大学硕士论文》- 2009-04-01	4.3%（195） 是否引证：否
13	数据挖掘中基于遗传算法的K-means聚类算法的研究及应用 赵松(导师：何熊熊) - 《浙江工业大学硕士论文》- 2014-04-08	4.3%（193） 是否引证：否
14	基于划分的聚类算法研究 郑柏杰(导师：杨小帆;何国辉) - 《重庆大学硕士论文》- 2005-10-01	4.3%（193） 是否引证：否
15	数字版权管理策略研究 马军军 - 《大学生论文联合比对库》- 2017-04-27	4.1%（185） 是否引证：否
16	网格聚类算法的研究 程伟想(导师：孟建良) - 《华北电力大学（河北）硕士论文》- 2008-12-18	4.1%（185） 是否引证：否
17	瞿安国_1120132189_学术研究趋势分析系统的设计与实现 瞿安国 - 《大学生论文联合比对库》- 2017-05-31	4.0%（180） 是否引证：否
18	基于密度的建筑物聚类分析 王安东 - 《大学生论文联合比对库》- 2017-05-29	3.9%（179） 是否引证：否
19	49-3_基于微博的舆情分析与热点推荐 基于微博的舆情分析与热点推荐 - 《大学生论文联合比对库》- 2017-04-06	3.5%（157） 是否引证：否
20	色谱指纹图谱的智能聚类分析在中医湿证辨别方面的研究 胡琳(导师：黄振国) - 《东华大学硕士论文》- 2003-12-01	3.4%（156） 是否引证：否
21	基于特征向量的个性化推荐算法研究 杜定宇(导师：王茜) - 《重庆大学硕士论文》- 2011-04-01	2.9%（131） 是否引证：否
22	新闻聚合系统设计 徐宇航 - 《大学生论文联合比对库》- 2017-05-06	2.4%（109） 是否引证：否
23	基于支持向量机的电信话务量预测方法 陈电波(导师：徐福仓;吴敏) - 《中南大学硕士论文》- 2008-06-30	2.4%（109） 是否引证：否
24	00343186681013071_王震_一种基于密度最大值聚类算法的研究与应用 郑浩 - 《大学生论文联合比对库》- 2017-05-12	2.2%（101） 是否引证：否
25	高考志愿填报的数据分析研究 杨浩杰(导师：陈志国;刘刚) - 《河南大学硕士论文》- 2011-05-01	2.0%（89） 是否引证：否
26	1495515502393_夏浪伟_20170523查重 夏浪伟 - 《大学生论文联合比对库》- 2017-05-23	1.9%（88） 是否引证：否
27	1495515502393_夏浪伟_20170523查重 夏浪伟 - 《大学生论文联合比对库》- 2017-05-23	1.9%（88） 是否引证：否
28	改进的模糊聚类算法在入侵检测中的应用研究 邹翔(导师：张玉芳) - 《重庆大学硕士论文》- 2015-04-01	1.5%（70） 是否引证：否

29	基于电网运行数据集的电力系统运行评估及优化研究	1.5%（69）
	刘柏林(导师：穆钢) -《华北电力大学(北京)博士论文》- 2017-06-01	是否引证：否
30	个人信息去标识化框架及标准化	1.0%（45）
	谢安明;金涛;周涛; -《大数据》- 2017-09-20	是否引证：否
31	基于深度学习的车辆检测和车牌定位	0.9%（40）
	封晶(导师：任克强) -《江西理工大学硕士论文》- 2017-05-24	是否引证：否
32	面向数据挖掘的隐私保护算法研究	0.7%（33）
	郑少飞(导师：李玲娟) -《南京邮电大学硕士论文》- 2011-03-01	是否引证：否
33	加载隐私保护的网络安全综合管理关键技术研究	0.7%（33）
	马进(导师：李建华) -《上海交通大学博士论文》- 2012-05-01	是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第二章大数据隐私保护及数据挖掘相关技术

大数据隐私保护

1.1.3. 大数据隐私保护概述

隐私[11]指主体没有公开的知识、信息等，根据不同主题，隐私又可分为公共隐私和个人隐私。公共隐私指群体的共同信息或者模式信息，主要针对企业和政府部门。个人隐私指个人独立的基本资料，如医疗信息，电子档案信息等。大数据在分析过程中如果不对用户敏感信息进行保护，极有可能造成用户信息泄露。如何在不泄露用户隐私的前提下，提高大数据的利用率，是目前大数据研究领域的关键问题之一。

1.1.4. 大数据隐私保护方法

（1）数据发布匿名保护技术

为防止攻击者在数据发布时使用链接攻击获取用户敏感信息，常常需要在数据发布对敏感信息进行匿名处理。其中最常见的是k匿名技术[12]；通过对用户敏感信息使用隐匿和泛化两种方法，使得对用户的同一属性字段至少有1-k个用户与之相同，从而迷惑攻击者。由于k匿名的敏感属性在分类时可能会缺乏多样性，导致匿名表产生信息泄露，因而又有人提出了l-多样化技术[13]，使得匿名组里敏感属性的多样性大于或等于l。

（2）数据水印技术

数据水印技术[14]指将特定的信息嵌入数字信号中，数字信号可以是音频、图片或视频等。若要拷贝有数字水印的信号，所嵌入的信息也会一并被拷贝。数字水印可分为浮现式和隐藏式两种。一般来说，浮现式的水印通常包含版权拥有者的名称或标志，其所包含的信息可在观看图片或视频时同时被看见。隐藏式的水印是以数字数据的方式加入音频、图片或视频中，但在一般的状况下无法被看见。隐藏式水印的重要作用之一是保护版权，期望能借此避免或阻止数字媒体未经授权的复制和拷贝。

数据挖掘

1.1.5. 数据挖掘的概念

数据挖掘[15]是一个多学科交叉领域，它融合了数据库技术、人工智能、机器学习、统计学、知识工程、面向对象方法、信息检索等最新技术的研究成果，从大量数据中提取知识和模式。数据挖掘的实际工作是对大规模数据进行自动或半自动的分析，以提取过去未知的有价值的潜在信息，例如数据的分组（通过聚类分析）、数据的异常记录（通过异常检测）和数据之间的关系（通过关联式规则挖掘）。

1.1.1. 数据挖掘的过程

（1）数据预处理

在使用挖掘算法计算之前，必须收集数据集。由于数据挖掘只能发现存在于数据中的模式，所以数据集应该足够大，以包含这些模式。对于分析多元数据集来说，数据预处理是必不可少的一个过程，通过预处理过程来剔除噪声数据和修补缺失数据，以提高挖掘结果的准确性。

（2）数据挖掘

数据挖掘过程通常涉及六种常见的任务：异常检测，用来识别异常数据；关联学习规则，用来搜索变量之间的关系；聚类，在数据集中发现具有某些相同特征的数据子集；分类，将已知的分类或者结构应用于未知的数据，对其进行分类；回归，寻找用最小错误对数据建模的函数；汇总，提供更紧凑的数据表示方法，包括生成可视化和报表。

（3）结果验证

如果数据挖掘方法被误用，可能会产生一个看似很重要实则在其他新的数据集上不能复现的结果，这种结果通常是使用了太多假设或者不符合假设统计规律所造成的，在机器学习中的一个例子就是过度拟合。所以在进行数据挖掘之后，往往需要在更大的数据集中验证所得模式是否正确。数据挖掘所得到的模式并不是一直都是有效的，很多挖掘算法所得到的结果并不能够在更宽泛的数据集中复现。

聚类算法

1.1.6. 聚类算法概述

聚类分析是目前大数据分析的常用手段之一。聚类分析也称群集分析，将本没有类别参考的数据进行分析和划分，把相

似的对象通过静态分类的方法分成不同的组别或者更多的簇。聚类分没有指定特定的算法，但它可以根据不同的需求被各种不同的算法所实现。簇指的是具有某些共同特征的数据的集合，不同的实现算法，不同的聚类标准都会得到不同的簇，这也是为什么会有很多不同的聚类算法的原因之一。

1.1.7. 聚类算法分类

按照聚类算法所采用的思想不同，大致可分为五类[16]。

（1）层次聚类算法

通过计算不同类别数据点之间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的叶子节点，树的顶层是一个聚类的根节点。创建聚类树分为自上而下的合并聚类和自下而上的分解聚类两种方法。自上而下聚类先将所有原始数据点看作一个聚类，然后根据每个原始节点之间的差异性将其不断分解。自下而上聚类先将每个原始数据点看作一个原子聚类，然后根据彼此之间的相似性不断进行聚合。

（2）分割聚类算法

首先将数据集分割为k个划分，然后通过迭代某特定算法，使得某个指标最优以达到最终结果。这种聚类算法又可详细划分为基于图论的聚类、基于网格的聚类、基于平方差的聚类和基于密度的聚类。

基于图论的聚类将聚类转换为组合优化问题，然后利用图论相关知识，并结合启发式算法来求解原问题。其一般做法是，先构造数据集的最小生成树，然后逐步删除最小生成树中最大长度的边，形成更多聚类。

基于网格的聚类算法[17]采用空间驱动的方法，把嵌入空间划分成独立于输入对象分布的单元。这种方法使用一种多分辨率的网络数据结构。它将对象空间量化成有限数目的单元，这些网格形成了网格结构，所有的聚类结构都在该结构上进行。这种方法的主要优点是处理速度快，其处理时间独立于数据对象数，仅依赖于量化空间中的每一维的单元数。基本思想就是将每个属性的可能值分割成许多相邻的区间，创建网格单元的集合（我们假设属性值是连续的，序数的，区间的）。每个对象落入一个网格单元，网格单元对应的属性空间包含该对象的值。

基于平方差的聚类算法的主要思想是逐步优化聚类结果，通过迭代想目标数据集重复向聚类中心进行聚类以得到最优解。该聚类方法又可详细分为最邻近聚类算法概率聚类算法、K-medoids[18]算法和K-means算法。其中K-means是目前应用最多的一种聚类算法，给定划分数量 k并创建一个初始划分，从数据集中随机地选择 k 个对象，每个对象初始地代表了一个簇中心（Cluster Centroid）。对于其他对象，计算其与各个簇中心的距离，将它们划入距离最近的簇。采用迭代的重定位技术，尝试通过对象在划分间移动来改进划分。所谓重定位技术，就是当有新的对象加入簇或者已有对象离开簇的时候，重新计算簇的平均值，然后对对象进行重新分配。这个过程不断重复，直到各簇中对象不再变化为止。

基于密度的聚类算法根据数据的分布密度，合并密度超过某个阈值的相邻区域为一个区域，可以在有噪音的数据中发现各种形状和各种大小的簇，常用于对空间数据进行聚类。DBSCAN算法[19]就是该类方法中最典型的算法之一。

1.1.8. DBSCAN算法描述

先介绍一些关键概念的定义：

假设样本集D=(x1,x2,…,xm)

1）邻域：对于xj∈D，其邻域包含样本集D中与xj的距离不大于的子样本集，即,这个样本集的个数记为

2）核心对象：对于任一样本xj∈D，如果其邻域对应的至少包含MinPts个样本，即如果，则xj是核心对象。

3）密度直达：如果且,则称p由q密度直达。当p和q都是核心对象时，密度直达满足对称性。

图2.1

图2.1中，假设MinPts=5,则p是核心对象，q是噪声点，从p到q密度直达，而从q到p不是密度直达

4）密度可达：对于xi和xj,如果存在样本序列p1,p2,…,pT满足p1=x1,pT=xj,且pt+1由pt密度直达，则称xj由xi密度可达。即密度可达满足传递性，但不满足对称性。

5）密度相连：对于xi和xj,如果存在核心对象样本xk,使得xi和xj均由xk密度可达，则称xi和xj密度相连。密度相连满足对称性。

6）簇：簇是样本集的一个子集，当样本点满足下列条件时称样本点包含于簇C。

a) :如果p∈C且q由p密度可达，则q∈C

b) :p,q密度相连

7）噪声点：设C1,C2,…,Ck是样本集D中的簇，则定义噪声点为不属于任何簇的数据集，即

算法流程：

输入：——半径

MinPts——给定点在邻域内成为核心对象的最小邻域点数。

D——集合。

输出：目标类簇集合

方法： Repeat

1）判断输入点是否为核心对象

2）找出核心对象的E邻域中的所有直接密度可达点。

Until 所有输入点都判断完毕

Repeat
针对所有核心对象的E邻域内所有直接密度可达点找到最大密度相连对象集合，中间涉及到一些密度可达对象的合并。
Until 所有核心对象的E领域都遍历完毕

伪码描述：

```
DBSCAN(D, eps, MinPts) {
C = 0
for each point P in dataset D {
if P is visited
continue next point
mark P as visited
NeighborPts = regionQuery(P, eps)
if sizeof(NeighborPts) < MinPts
mark P as NOISE
else {
C = next cluster
expandCluster(P, NeighborPts, C, eps, MinPts)
}
}
}
expandCluster(P, NeighborPts, C, eps, MinPts) {
add P to cluster C
for each point P' in NeighborPts {
if P' is not visited {
mark P' as visited
NeighborPts' = regionQuery(P', eps)
if sizeof(NeighborPts') >= MinPts
NeighborPts = NeighborPts joined with NeighborPts'
}
if P' is not yet member of any cluster
add P' to cluster C
}
}
regionQuery(P, eps)
return all points within P's eps-neighborhood (including P)
```

图2.2

在图2.2中，MinPts=5，红色点为核心对象。所有核心对象密度直达的样本点在以核心对象为圆心的圆中，绿色箭头连起来的核心对象构成了密度可达序列，在这些密度可达序列的邻域内，所有点密度相连。所以DBSCAN算法其本质就是针对核心对象邻域内可密度直达的样本点寻找最大密度相连的集合。图2.2中一共找到两个这样的集合，即最终所得簇。

本章小结

本章首先介绍了数据隐私保护的相关知识及方法。然后对数据挖掘技术进行了阐述，包括其概念和挖掘过程。最后介绍了数据挖掘中的常用方法——聚类算法及其分类，并对其中的经典算法DBSCAN进行了详细介绍，为后面的乘车热点区域挖掘奠定了理论基础。

指　标
疑似剽窃文字表述

1. 信息泄露。如何在不泄露用户隐私的前提下，提高大数据的利用率，是目前大数据研究领域的关键问题之一。
2. 若要拷贝有数字水印的信号，所嵌入的信息也会一并被拷贝。数字水印可分为浮现式和隐藏式两种。一般来说，浮现
3. 隐藏式的水印是以数字数据的方式加入音频、图片或视频中，但在一般的状况下无法被看见。隐藏式水印的重要作用之一是保护版权，期望能借此避免或阻止数字媒体未经授权的复制和拷贝。
4. 聚类算法又可详细划分为基于图论的聚类、基于网格的聚类、基于平方差的聚类和基于密度的聚类。
   基于图论的聚类将聚类
5. 这种方法使用一种多分辨率的网络数据结构。它将对象空间量化成有限数目的单元，这些网格形成了网格结构，所有的

聚类结构都在该结构上进行。这种方法的主要优点是处理速度快，其处理时间独立于数据对象数，仅依赖于量化空间中的每一维的单元数。

6. 每个对象落入一个网格单元，网格单元对应的属性空间包含该对象的值。

基于平方差的聚类

7. 对于其他对象，计算其与各个簇中心的距离，将它们划入距离最近的簇。采用迭代的重定位技术，尝试通过对象在划分间移动来改进划分。所谓重定位技术，就是当有新的对象加入簇或者已有对象离开簇的时候，重新计算簇的平均值，然后对对象进行重新分配。这个过程不断重复，直到各簇中对象不再变化为止。

基于密度的聚类算法根据数据的分布密度，合并密度超过某个阈值的

8. 算法流程：

输入：——半径

MinPts——给定点在邻域内成为核心对象的最小邻域点数。

D——集合。

输出：目标类簇集合

方法： Repeat

1）判断输入点是否为核心对象

2）找出核心对象的E邻域中的所有直接密度可达点。

Until 所有输入点都判断完毕

Repeat

针对所有核心对象的E邻域内所有直接密度可达点找到最大密度相连对象集合，中间涉及到一些密度可达对象的合并。

4. 第三章数据获取及预处理

总字数：4500

相似文献列表　文字复制比：46.6%(2097)　疑似剽窃观点：(0)

1	基于Web的校园办公自动化系统设计与实现 靖雯(导师：陆鑫;卢长军) -《电子科技大学硕士论文》- 2012-03-01	39.6%（1783） 是否引证：否
2	[gb18030] 10055101徐浩程 毕设论文 徐浩程 -《大学生论文联合比对库》- 2014-06-13	38.5%（1731） 是否引证：否
3	[gb18030] 10055101徐浩程 毕设论文 徐浩程 -《大学生论文联合比对库》- 2014-06-13	38.5%（1731） 是否引证：否
4	HMAC-MD5的分析与实现 康四雯 -《大学生论文联合比对库》- 2015-06-01	38.5%（1731） 是否引证：否
5	基于Android端的新闻阅读器的设计与实现 张雨杰 -《大学生论文联合比对库》- 2014-05-02	38.4%（1730） 是否引证：否
6	韩笑_大学生在线求职招聘网站的设计与实施 20142174 -《大学生论文联合比对库》- 2014-10-28	36.6%（1649） 是否引证：否
7	基于WEB的人才招聘系统的设计与实现 王璐瑶(导师：王连平) -《吉林大学硕士论文》- 2013-06-01	36.6%（1649） 是否引证：否
8	现在才知道其实中国对美国军方最具威胁的不是核武也不是打卫星技术.原来是中国的密码破译能力.中国数字家已将西方认为不可能破译的MD5和SHA-1成功破译_标点 - 韶韶 - 记者版_媒体 -《网络（http://www.xici.net/）》- 2013	36.6%（1649） 是否引证：否
9	网络121_2012122631_聂睿 聂睿 -《大学生论文联合比对库》- 2016-06-08	36.6%（1648） 是否引证：否
10	Message-Digest Algorithm 5 - 走马观花 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	36.3%（1632） 是否引证：否
11	超市后台管理系统设计与实现 吴波(导师：张平健;潘勇) -《华南理工大学硕士论文》- 2010-11-10	35.3%（1590） 是否引证：否
12	基于SHA-256算法的嵌入式软件保护技术研究 郑佳敏(导师：沈建华) -《华东师范大学硕士论文》- 2014-03-22	34.3%（1543） 是否引证：否
13	MD5 - yjg428的专栏 - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2013	32.7%（1471） 是否引证：否
14	简要分析用MD5加密算法加密信息（精编版） - Auspicious Cloud's View - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2013	32.4%（1456） 是否引证：否
15	简要分析用MD5加密算法加密信息（精编版） - Auspicious Cloud's View - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	32.4%（1456） 是否引证：否
16	MD5_百度百科	32.2%（1451）

	- 《网络（http://baike.baidu.c）》- 2010	是否引证：否
17	MD5算法Qt实现 - tandesir的专栏 - CSDN博客	32.2%（1451）
	- 《网络（http://blog.csdn.net）》- 2017	是否引证：否
18	什么是MD5 - 豆丁网	32.2%（1451）
	- 《互联网文档资源（http://www.docin.com）》- 2017	是否引证：否
19	常数ti是4294967296*abs_caoaugt	31.4%（1415）
	- 《网络（http://blog.sina.com）》- 2014	是否引证：否
20	md5算法及其C语言的实现 - 网络与安全,mayu8758,拥抱开源，把握未来！	31.4%（1411）
	- 《网络（http://blog.opendige）》- 2011	是否引证：否
21	电子印章的设计与实现	31.0%（1393）
	龙宇 - 《大学生论文联合比对库》- 2015-05-27	是否引证：否
22	MD5加密算法详细分析	4.0%（181）
	- 《网络（http://blog.csdn.net）》- 2017	是否引证：否
23	基于C/S与B/S混合模式下教务系统的研究和分析	3.9%（175）
	刘立(导师：王长波) - 《华东师范大学硕士论文》- 2011-04-01	是否引证：否
24	13刘佳文--2014年度结项报告_移动支付安全问题研究—以微信支付为例	3.9%（175）
	- 《大学生论文联合比对库》- 2015-10-20	是否引证：否
25	13刘佳文--2014年度结项报告_移动支付安全问题研究—以微信支付为例-新	3.9%（175）
	- 《大学生论文联合比对库》- 2015-10-21	是否引证：否
26	20123264从海云_温蜜_毕设论文	3.6%（160）
	从海云 - 《大学生论文联合比对库》- 2016-06-07	是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第三章数据获取及预处理

数据来源及数据格式

本文采用南京市中北股份有限公司的出租车营运数据作为研究对象，数据采集起止时间为2018年4月25日到2018年5月2日，共有出租车1639辆，数据总量17MB,约包含170418条记录。数据源来自数据库名为T_BUSI_RUN的数据表，每条记录包含车牌号，上车原始经维度，上车高德经纬度，上车时间等字段。数据字段及其含义对应关系如表3.1所示。

表3.1 营运数据与字段含义表

字段名称类型

FACT_ID 厂商标识 NUMBER(2,0)

DEV_ID 设备ID NUMBER(10,0)

CPHM 车牌号 VARCHAR2(9 BYTE)

SJDM 司机代码 VARCHAR2(20 BYTE)

DWDM 单位代码 VARCHAR2(20 BYTE)

SCSJ 上车时间 DATE

SCJD 上车高德经度 NUMBER(9,6)

SCWD 上车高德纬度 NUMBER(8,6)

XCSJ 下车时间 DATE

SCJD 下车经度 NUMBER(9,6)

XCWD 下车纬度 NUMBER(8,6)

DHSJ 等候时间，单位：秒 NUMBER(5,0)

YYSJ 营运时间，单位：秒 NUMBER(5,0)

YYLC 营运里程，单位：公里 NUMBER(6,1)

KSLC 空驶里程，单位:公里 NUMBER(4,1)

YYJE 营运金额，单位：元 NUMBER(6,1)

JYLX 交易类型，0：现金，1，2，3，5：市民卡，6：支付宝，7：微信，4，144，145，146：银联卡 CHAR(3 BYTE)

YYID 营运ID NUMBER(11,0)

PJID 评价ID NUMBER(10,0)

PJXX 评价选项，0：未评价；1：满意；2：一般；3：不满意；4：投诉 VARCHAR2(2 BYTE)

FJF 附加费，单位：元 NUMBER(6,0)

SCSJGPS GPS的上车时间 DATE

XCSJGPS GPS的下车时间 DATE

数据预处理

1.1.9. 数据清洗

通常由于设备故障，司机误操作或者天气原因，GPS所采集到的原始数据并不是百分之百准确，会存在一些异常数据。为了使数据挖掘的结果更加准确，需要在数据挖掘之前进行数据预处理。本文的数据预处理主要包括两方面，剔除重复数据和有空字段的数据，剔除经纬度越界数据。由于设备异常，同一条记录有时会被多次提交，这种重复记录应该在分析之前被剔除。乘车热点区域挖掘主要针对南京市，但由于出租车没有市级范围的限制，有些出租车会出现在南京市以外的地区，为了提高挖掘的准确性和后期可视化的美观性，需要将南京市以外的记录剔除掉。南京市经纬度大概范围为118.35-119.23，31.236-32.611，预处理时剔除上车经纬度在这个范围以外的数据。

1.1.10. 数据导出导入

本文所述的数据分析基于非关系型数据库MongoDB进行，由于原始数据存储于关系型数据库Oracle数据库中，需要进行数据的移植工作。考虑到数据挖掘不会用到营运数据表中的全部字段，为节省存储空间，提高挖掘效率，本文在数据移植时只移植了对数据挖掘有用的相关字段：车牌号码，上车高德经度，上车高德纬度，上车时间。MongoDB数据库中是以文档的形式来存储数据，MongoDB中文档相当于关系型数据库中的一条记录，集合相当于关系型数据库中的表。

为实现数据的导出与导入，首先在sqlplus中通sql语句剔除重复数据和有空值字段的数据，然后将查询结果保存在csv文件中，最后通过mongoimport命令将csv文件中的数据导入到MongoDB中事先创建好的数据库中，移植后所得文档结构如下。

{
"_id" : ObjectId("5add799dd9c9d696a38ed0a9"),
"CPHM" : "苏A70007",
"SCWD" : 31.967239,
"SCJD" : 118.795548,
"SCSJ" : "2018-04-22 12:05:00"
}

_id是一个12字节长的十六进制书，它保证了每个文档的唯一性，其余字段含义如本章第一节所示。

1.1.11. 用户隐私保护

通过上节所述可以发现，如果直接将数据存入MongoDB中，车牌号将以明文方式显示。对出租车司机和乘客来说，车牌号码具有隐私性，如果数据分析就被不可信的第三方执行，它将会获得车牌号码以及营运数据，结合链接攻击等方式，极有可能造成司机或者乘客的个人隐私泄露，因此，需要对车牌号码进行隐私保护处理。

美国密码学家罗纳德·李维斯特于1992年公开了MD5[20]摘要加密算法，输入任意长度的信息，经过处理，输出128位的信息。经过程序流程，生成四个32位数据，最后联合起来成为一个128-bits散列。基本方式为，求余，取余，调整长度，与链接变量进行循环运算，得出结果。图3.1展示了其运算过程，一个MD5运算—由类似的64次循环构成，分成4组16次。F 一个非线性函数；一个函数运算一次。Mi 表示一个 32-bits 的输入数据，Ki 表示一个 32-bits 常数，用来完成每次不同的计算。

图3.1 MD5运算流程

摘要算法流程如下：

1）填充：如果信息长度（以bit为单位）对512求余的结果不等于448，则对信息进行填充，在信息尾部添加一个1和若干个0，使得填充后信息长度为N*512+448位.

2）记录信息长度：用64位来存储填充前的信息长度。这64位追加在第一步所得结果后面，这样信息长度就变成N*512+448+64=(N+1)*512位。

3）初始化MD数组：A=0X67452301L，B=0XEFCDAB89L，C=0X98BADCFEL，D=0X10325476L

4）进行循环运算

伪码描述：

```
//Note: All variables are unsigned 32 bits and wrap modulo 2^32 when calculating
var int[64] r, k
//r specifies the per-round shift amounts
r[ 0..15] : = {7, 12, 17, 22, 7, 12, 17, 22, 7, 12, 17, 22, 7, 12, 17, 22}
r[16..31] : = {5, 9, 14, 20, 5, 9, 14, 20, 5, 9, 14, 20, 5, 9, 14, 20}
r[32..47] : = {4, 11, 16, 23, 4, 11, 16, 23, 4, 11, 16, 23, 4, 11, 16, 23}
r[48..63] : = {6, 10, 15, 21, 6, 10, 15, 21, 6, 10, 15, 21, 6, 10, 15, 21}
//Use binary integer part of the sines of integers as constants:
for i from 0 to 63
k[i] := floor(abs(sin(i + 1)) × 2^32)
//Initialize variables:
var int h0 := 0x67452301
var int h1 := 0xEFCDAB89
var int h2 := 0x98BADCFE
```

```
var int h3 := 0x10325476
//Pre-processing:
append "1" bit to message
append "0" bits until message length in bits ≡ 448 (mod 512)
append bit length of message as 64-bit little-endian integer to message
//Process the message in successive 512-bit chunks:
for each 512-bit chunk of message
break chunk into sixteen 32-bit little-endian words w[i], 0 ≤ i ≤ 15
//Initialize hash value for this chunk:
var int a := h0
var int b := h1
var int c := h2
var int d := h3
//Main loop:
for i from 0 to 63
if 0 ≤ i ≤ 15 then
f := (b and c) or ((not b) and d)
g := i
else if 16 ≤ i ≤ 31
f := (d and b) or ((not d) and c)
g := (5×i + 1) mod 16
else if 32 ≤ i ≤ 47
f := b xor c xor d
g := (3×i + 5) mod 16
else if 48 ≤ i ≤ 63
f := c xor (b or (not d))
g := (7×i) mod 16
temp := d
d := c
c := b
b := leftrotate((a + f + k[i] + w[g]),r[i]) + b
a := temp
Next i
//Add this chunk's hash to result so far:
h0 := h0 + a
h1 := h1 + b
h2 := h2 + c
h3 := h3 + d
End ForEach
var int digest := h0 append h1 append h2 append h3 //(expressed as little-endian)
```

为充分保护用户隐私，本文在进行数据导出时对车牌号码这一关键字段采用MD5算法进行映射。Oracle提供了DBMS_OBFUSCATION_TOOLKIT.MD5方法来对字段进行加密，但是此时加密得到的是乱码，并非十六进制数字，需要使用utl_raw.cast_to_raw将字符串转换为RAW类型。

本章小结

本章首先介绍了本文进行数据挖掘所需数据的来源和格式，以及各个字段的含义；然后介绍了数据挖掘的第一步——数据预处理，主要包括重复数据和空数据的剔除；紧接着介绍了数据从数据关系型数据库到非关系型数据库MongoDB的移植；最后阐述了使用MD5对用户隐私字段车牌号进行了映射保护的方法。本章为下一章的数据挖掘做了数据准备。

指　标
疑似剽窃文字表述
1.　经过程序流程，生成四个32位数据，最后联合起来成为一个128-bits散列。基本方式为，求余，取余，调整长度，与链

接变量进行循环运算，得出结果。

2. 运算过程，一个MD5运算—由类似的64次循环构成，分成4组16次。F 一个非线性函数；一个函数运算一次。Mi 表示一个 32-bits 的输入数据，Ki 表示一个 32-bits 常数，用来完成每次不同的计算。

5. 第四章基于MapReduce框架的热点区域挖掘及可视化

总字数：4884

相似文献列表　文字复制比：7.7%(376)　疑似剽窃观点：(0)

1	聚类之层次聚类、基于划分的聚类（… - bluewater的专栏 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	5.8%（281） 是否引证：否
2	基于大数据分析的集装箱箱修点合理性分析 汪子君 -《大学生论文联合比对库》- 2017-05-26	5.7%（278） 是否引证：否
3	201224060201_李丹宁_基于出租车GPS数据的居民出行热点方法研究 李丹宁 -《大学生论文联合比对库》- 2016-06-06	5.7%（278） 是否引证：否
4	201224060201_李丹宁_基于出租车GPS数据的居民出行热点方法研究 李丹宁 -《大学生论文联合比对库》- 2016-06-13	5.7%（278） 是否引证：否
5	基于集装箱维修数据的集装箱箱修点合理性分析 汪子君 -《大学生论文联合比对库》- 2017-05-31	5.7%（278） 是否引证：否
6	数据挖掘技术中聚类算法的研究 许晨俊 -《大学生论文联合比对库》- 2017-05-31	5.7%（278） 是否引证：否
7	【数据挖掘】聚类算法总结 -- 数据科学自媒体 -- 传送门 -《网络（http://chuansong.me/）》- 2016	5.6%（274） 是否引证：否
8	聚类算法总结 - 南山牧笛的博客 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	5.6%（274） 是否引证：否
9	基于密度的聚类 - suichen1的专栏 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	5.6%（274） 是否引证：否
10	聚类算法 - 皮皮blog - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	5.6%（274） 是否引证：否
11	11668144_马晨阳_基于密度的数据挖掘聚类方法研究 马晨阳 -《大学生论文联合比对库》- 2017-06-10	5.3%（260） 是否引证：否
12	基于云计算分布式技术的海量AIS数据挖掘系统设计与实现 尚斯年(导师：杨家轩;孙霄峰) -《大连海事大学硕士论文》- 2017-04-01	2.6%（129） 是否引证：否
13	Mongodb中Mapreduce特性与原理 - shift_alt -《网络（http://blog.csdn.net）》- 2017	1.9%（95） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第四章基于MapReduce框架的热点区域挖掘及可视化

基于MapReduce框架的聚类算法实现

1.1.12. 实验环境搭建

框架：Node.js 8.9.4,MapReduce

语言：JavaScript,Java

数据库：MongoDB 3.4.10.

1.1.13. 基于MongoDB的MapReduce框架实现

作为数据分析的常用数据库之一，MongoDB自身集成了MapReduce框架的接口。下面对其进行介绍。

图4.1 MapReduce命令格式

在图4.1中，map函数用于将键和值进行映射，将同一个键对应的值组合成数组，交由reduce函数处理。Reduce函数对同一键所对应的值构成的数组进行化简处理，最后返回与输入结构相同的结果。

图4.2 MapReduce流程

图4.2描述了MapReduce的工作流程，query阶段通过筛选得到status为A的文档，map阶段以cust_id为键，以amount为值进行映射，每个键都会对应一个由值组成的数组，reduce阶段对每个键对应的值组成的数组进行求和化简，最后返回一个键值对。

1.1.14. 聚类算法实现

借助于MongoDB对MapReduce框架的集成，算法实现分为五个步骤：读取数据，映射，确定，化简，聚类，结果输出。

图4.3 聚类流程

读取数据：node.js为mongodb数据库操作提供了mongodb模块，通过require语句引入后即可连接数据库，将数据读入内存并进行操作。

映射：为突显出周内和周末不同时间段乘车热点区域的不同，本文按照每条数据是否是工作日isWorkDay（即周一到周五

），以及所在的不同时间段timeID进行划分。如表4.1所示，将数据按照上车日期划分为工作日和非工作日。在工作日与非工作日中，又按照不同时间段划分为不同的timeID，然后以isWorkDay和timeID为键，以车牌号，上车时间，上车经度，上车维度为值进行映射，得到一个由这四个变量组成的数组。

isWorkDay time TimeID

false 0:00 - 8：00 0

false 9:00 - 23:00 1

true 0:00 - 7:00 0

true 8:00 – 10:00 1

true 11:00 – 13:00 2

true 14:00 – 16:00 3

true 17:00 – 20:00 4

true 21:00 – 23:00 5

表4.1 时间与timeID对照表

确定：对于映射得到的六个分组，对每一组中的每个样本点计算离它第k近的样本点的距离k-dist，并将最终结果降序排序，取其中变化最剧烈的点所对应的距离为，取MinPnts为k.

化简：reduce这一阶段接收由map阶段传过来的key和values变量，如果map阶段一个key对应的values条数很多，可能会多次调用reduce方法，即前一次reduce的结果可能被包含在values中再次传递给reduce方法，这也要求，reduce的返回结果需要和value的结构保持一致。因此本文不在reduce中进行聚类分析，只是将values递归原样返回。

聚类：finalize方法在reduce函数结束后接收其key和values做最后的处理并输出，本文聚类算法将在这个阶段进行。在上述的map阶段，将所有的营运数据按照isWordDay和timeID进行了分类，在finalize中将对每一类别进行聚类，阈值取第二阶段所得到的阈值。具体流程如图4.2所示。在判断样本点是否是核心点时，本文通过样本点的经纬度来计算两点之间的距离，公式如下。其中R为地球半径，取6367000m,lat,lng分为别样本点的维度和经度的弧度大小。

在实验中作者发现，对于十七万多个样本点，cpu计算其两两之间的距离特别耗时，而且十分占用cpu资源，因此本文提出一种改进方案。由于所有样本点基本都是在南京市内采集的，即两点之间的距离不会超过200千米，在这个范围内，本文认为经线和纬线是垂直的。如图4.3所示，如要计算A(116.8,39.78)和B(116.9,39.68)之间的距离，可以先计算AM,BM的距离，最后通过勾股定理计算AB之间的距离。

图4.3

通过上述优化，可以使距离计算公式中只含有一个三角函数。为了进一步加快cpu计算速度，本文使用泰勒公式将余弦函数展开，经过尝试本文将其展开到四次。

通过取不同的点对并计算之间的距离，本文验证了简化计算的性能，如表4.2所示。可以看出，在误差允许范围内，这种简化算法的精确度是可以接收的。

P1 P2 Original algorithm Reduce algorithm difference

(39.941,116.45) (39.94,116.451) 140.02855253140535 140.03920028866003 0.011

(39.96,116.45) (39.94,116.40) 4804.4211529412405 4805.200681310494 0.780

(39.96,116.45) (39.94,117.30) 72444.54071515072 72459.75352928363 15.212

(39.26,115.25) (41.04,117.30) 263508.55921885674 263550.1567337578 41.580

表4.2

图4.4 DBSCAN算法流程

结果输出：MongoDB文档的存储结构有两种设计模式，内嵌和引用。内嵌模式即在文档中继续嵌入其他文档对象，可以将数组等对象嵌入到文档中，能够直观的体现对象模型的结构。但是在MongoDB中文档大小不能超过16M，一旦超过mongo server就会报错。一般的一对一、一对多关系，比如说一个人多个地址多个电话等都可以放在一个文档里用内嵌来完成。引用模式类似关系型数据库中主外键的关系，将内嵌的文档拆分成多个文档放在一个独立的集合中，然后通过id来访问。比如，一个明星的博客可能有几十万或者几百万的回复，这时如果把comments放到一个数组里，可能会超出16M的限制。这个时候就需要考虑使用引用的方式，在主表里存储一个id值，指向另一个表中的 id 值。

本文所用数据集大约17万条，在map阶段通过emit函数分为8组，平均每组有两万多条数据，极有可能超出16M，导致程序报错。这里采用引用模式的一种变种方法，在finalize中首先将聚类结果通过mapreduce中的scope选项保存到局部变量中，等到mapreduce过程结束后再将其插入到新的集合中，访问时直接通过访问新的集合来得到聚类结果，而finalize在返回时直接返回null.

1.1.15. 阈值取值分析

通过上述分析可知，DBSCAN算法有两个阈值的输入，和MinPts。聚类结果的好坏直接取决于这两个阈值。决定了每个样本点邻域的大小，MinPts决定了每个邻域的密度大小，聚类簇太小会导致原本应当属于一个簇的样本点分散到多个簇中，簇太大会导致聚类结果密度稀疏，使得本应属于多个簇的样本点集中在一个簇中。如何恰到好处地选择阈值，使聚类结果最优，是决定聚类结果性能的关键因素。下面描述一种确定两个阈值的启发式算法。

1) k-dist: 给定数据集P={p(i); i=0,1,…n}，对于任意点P(i)，计算点P(i)到集合P的子集S={p(1), p(2), …, p(i-1), p(i+1), …, p(n)}中所有点之间的距离，距离按照从小到大的顺序排序，假设排序后的距离集合为D={d(1), d(2), …, d(k-1), d(k), d(k+1), …,d(n)}，则d(k)就被称为k-距离。即点p(i)到所有点（除了p(i)点）之间距离第k近的距离。对聚类集合中每个点p(i)都计算k-距离，最后得到所有点的k-距离集合E={e(1), e(2), …, e(n)}。

2) 对于任意一点p，如果令=k-dist，MinPts=k，那么所有k-dist小于等于p的点都将是核心点。如果能在集合D中密度最小的簇中找到最大的k-dist，这时令=簇密度，MinPts=k-dist，即可得到阈值。为了直观寻找阈值，对所有点的k-dist进行降序排序，拟合出一条曲线，如下图所示。曲线第一次变化最剧烈的地方就是阈值点，其所对应的k-dist就是，MinPnts=k.

3) 在第二步的基础上如果能够得到k的具体取值，就可以完全得到两个阈值。通过尝试取值发现，k>4时，所得到的k-dist曲线和k=4时的曲线相差不大[19]，甚至会花费更多的计算量。如何选择k值，对于不同的场景有不同的要求，本文针对不同时段内的出租车密度，结合实际情况，认为一定范围内出租车数量超过5即为热点，选择k=5，即MinPts=5。

noisek-distCore pointsThreshold pointpoints

O

图4.5

基于上述算法，本文首先根据isWorkDay和timeID将营运数据分为八组，然后对每组数据中的点计算其k-distance，并将结果降序排列，得到如下图表。

图4.2 workDay:false,timeID=0

图4.3 workDy:false,timeID=1

图4.4 workDay:true,timeID=0

图4.5 workDay:true,timeID=1

图4.6 workDay:true,timeID=2

图4.7 workDay:true,timeID=3

图4.8 workDay=true,timeID=4

图4.9 workDay=true,timeID=5

观察上述图表，在图4.2中，取eps=1000；图4.3中，取eps=3000；图4.4中，取eps=2000；图4.5中，取eps=3000；图4.6中，取eps=2000；图4.7中，取eps=2400；图4.8中，取eps=3500；图4.9中，取eps=3000。

在聚类算法中，按照不同的类使用不同的eps和MinPts值进行聚类，最后得到簇数如下图所示。

图4.10 不同时段的簇数

从上图可以看出，在工作日中，timeID=2，3，4时（即11点到20点）热点簇比较多，而在周末白天时段的热点簇比较多。

为了证明本文所述阈值选择算法的正确性，下面进行横向对比。当MinPts为5时，本文将选择一系列递增的eps值计算clusters，并将结果和图4.10进行比较。可以发现，当每个时间段进行个性化取值时，所得簇数最大。

图4.11 不同eps对应的簇数

热点区域可视化

通过上述算法，按照不同时间段聚类将会得到八个结果集。本文通过高德地图API，将热点区域显示在Web端。在工作日中，当timeID为4时，所得热点区域如下图所示。下图中红点代表热点区域。由于API对海量点在一定比例尺下按照四叉树算法只显示其中的topN个点，所以将局部地图放大后可以看到更明显的聚类效果。可以看到在工作日timeID为4时，南京站和南京南站附近上车点较多，符合常规认识。

图4.12 整体热点区域

图4.13 局部热点区域一

图4.14 局部热点区域二

本章小结

本章主要阐述了乘车热点区域的具体挖掘过程。首先讲述了实验环境以及基于MongoDB数据库的MapReduce框架；然后阐述了DBSCAN算法在此框架下的实现流程，并分析了数据量大于16M时的结果集存储方案；最后针对不同阈值，对聚类结果进行分析，并将最优聚类结果可视化在Web端。

指　标
疑似剽窃文字表述

1. es条数很多，可能会多次调用reduce方法，即前一次reduce的结果可能被包含在values中再次传递给reduce方法，这也要求，reduce的返回结果需要和value的结构保持一致。
2. 对聚类集合中每个点p(i)都计算k-距离，最后得到所有点的k-距离集合E={e(1), e(2), …, e(n)}。
   2)

第五章总结与展望

总结

本文针对出租车的营运数据进行分析，其间涉及对数据库的选择、聚类算法的选择以及存储和计算的优化等问题。从数据预处理到数据挖掘再到最终结果可视化，本文通过比较不同阈值取值所得到的结果，最后得到了每天不同时段的乘车热点区域。本文研究内容主要有以下几个方面。

1) 数据预处理

由于原始Oracle数据库中存在很多重复数据和空数据，为了提高数据挖掘算法的准确性，本文需要在进行数据挖掘之前对脏数据进行剔除。

2) 数据库的选择

GPS终端将数据库传到了Oracle数据库，用传统的关系型数据库来做数据分析费时费力，相比之下，非关系型数据库MongoDB自身支持MapReduce框架，为数据分析提供了良好平台，并且对存储和处理非结构化数据都较为方便，因此本文选择使用MongoDB作为数据分析所用的数据库。

3) 聚类算法的选择

目前常用的聚类算法有K-Means和DBSCAN。对于K-Means算法，首先需要根据先验知识选择一个合适的K值，由于本文所面临的数据较大，选择合适的K值较难，所以不采用这种方法。DBSCAN算法是一种基于密度的聚类算法，通过样本分布的紧密程度来决定每个簇，同一类别的样本紧密相连，它能够有效处理噪声点并且发现任意形状的空间聚类，本文考虑到交通路网形状的多样性，认为所得的聚类簇形状也是依据路网呈现多样性的，因此选用DBSCAN算法。在计算样本点之间的距离时，通过对余弦函数进行泰勒展开，极大的提高了程序的计算效率。

4) 阈值的选择

阈值邻域半径和邻域密度MinPts的选择直接影响DBSCAN算法的优劣性，本文提出了基于样本点k-dist距离降序排序来选择的启发式算法，通过横向对比，证实了该方法的正确性。

5) 聚类结果可视化

为了更加形象直观的观察聚类结果，本文结合高德地图API，将聚类结果通过Web端显示在地图上，可以使司机或乘客更加直观的看到不同时段的热点区域。

展望

本文通过对2018年4月25日到2018年5月2日期间17万条数据的分析，得到周内和周末不同时段内的乘车热点区域并可视化在Web端。但同时由于各方面原因，本文存在以下几点不足之处。

1) 由于数据量较大，在单个机器上进行数据分析耗时良久，为进一步提高运行速率，后期可以在Hadoop平台上进行集群计算，以减少数据挖掘时间。

2) 在对阈值选择进行分析时，本文只提出了如何改进邻域半径使聚类结果达到最优的方法，并未对MinPts的取值进行比较。后期可以通过机器学习的方法来进一步比较不同MinPts取值对聚类结果的影响。

参考文献

[1] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113. DOI:

[2]冯登国,张敏,李昊.大数据安全与隐私保护[J].计算机学报,2014,37(01):246-258.

[3]Douglas, Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety(PDF). Gartner. [2001-02-06].

[4]侯士江. 基于空间k-匿名的位置隐私保护技术研究[D].燕山大学,2014.

[5]车延辙. 基于位置服务中用户位置隐私保护关键技术研究[D].浙江大学,2013.

[6]Gupta R, Rao U P. Achieving location privacy through CAST in location based services[J]. Journal of Communications and Networks, 2017, 19(3): 239-249.

[7]周洋. 基于出租车数据的城市居民活动空间与网络时空特性研究[D].武汉大学,2016.

[8]王郑委. 基于大数据Hadoop平台的出租车载客热点区域挖掘研究[D].北京交通大学,2016.

[9]Putri F K, Kwon J. A Distributed System for Finding High Profit Areas over Big Taxi Trip Data with MognoDB and Spark[C]//Big Data (BigData Congress), 2017 IEEE International Congress on. IEEE, 2017: 533-536.

[10] Deri J A, Franchetti F, Moura J M F. Big data computation of taxi movement in New York City[C]//Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016: 2616-2625.

[11]刘坚. K-匿名隐私保护问题的研究[D].东华大学,2010.

[12]Wei R, Shen H, Tian H. An Improved (k, p, l)-Anonymity Method for Privacy Preserving Collaborative Filtering[C]//GLOBECOM 2017-2017 IEEE Global Communications Conference. IEEE, 2017: 1-6.

[13] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity[C]//Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007: 106-115.

[14] 尹浩,林闯,邱锋,等.数字水印技术综述[J].计算机研究与发展,2005,42(7): 1093-1099.

[15] 邵峰晶, 于忠清. 数据挖掘原理与算法[M]. DynoMedia Inc., 2003.

[16]贺玲,吴玲达,蔡益朝.数据挖掘中的聚类算法综述[J].计算机应用研究,2007(01):10-13.

[17]伍育红.聚类算法综述[J].计算机科学,2015,42(S1):491-499+524.

[18]Hae-Sang Park,Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems With Applications,2008,36(2).

[19]Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Kdd. 1996, 96(34): 226-231.

[20] Rivest R. The MD5 message-digest algorithm[J]. 1992.

[21]Lu C, Lü Y, Ma M, et al. Data mining for points-selection rules in acupuncture treatment of mammary gland hyperplasia 基于数据挖掘探讨针刺治疗乳腺增生病选穴规律分析[J]. Journal of Acupuncture and Tuina Science, 2017, 15(5): 328-336.

[22]孙轶轩. 基于数据挖掘的道路交通事故分析研究[D].北京交通大学,2014.

[23]熊亚军,廖晓农,李梓铭,张小玲,孙兆彬,赵秀娟,赵普生,马小会,蒲维维.KNN数据挖掘算法在北京地区霾等级预报中的应用[J].气象,2015,41(01):98-104.

致谢

此次毕业设计，从最初选题到研读文献到最后做论文定稿，历经数月，其间遇到很多问题，在自己的努力与老师同学的帮助下，一步一个脚印，将其解决。

这次毕业设计可以圆满地完成，首先感谢陈兵老师的谆谆教导，能在百忙之中抽出时间，在生活上和学习上都给予我强烈的鼓励和指导。其次要感谢邓理同学，毕设开启之际，由于本人在前后端设计以及数据处理方面都缺乏经验，邓理同学则不厌其烦的给予本人帮助。最后要感谢本人于偶然之际认识的一位软件工程师，金成，感谢他在毕设中关于MongoDB和MapReduce 方面对我的指导，同时也要感谢王一凡同学，在毕设过程中与我通力合作，共同完成此次课题。