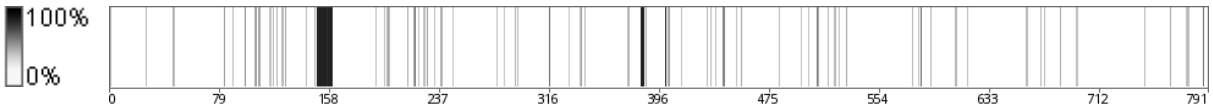
PaperPass旗舰版检测报告
简明打印版

比对结果(相似度):

总　体： 6% (总体相似度是指本地库、互联网的综合对比结果)
本地库： 5% (本地库相似度是指论文与学术期刊、学位论文、会议论文、图书数据库的对比结果)
期刊库： 2% (期刊库相似度是指论文与学术期刊库的比对结果)
学位库： 3% (学位库相似度是指论文与学位论文库的比对结果)
会议库： 0% (会议库相似度是指论文与会议论文库的比对结果)
图书库： 1% (图书库相似度是指论文与图书库的比对结果)
互联网： 2% (互联网相似度是指论文与互联网资源的比对结果)

编　号：59C8B250D3680AN5I
版　本：旗舰版
标　题：基于交通大数据的商圈可视化研究
作　者：李柯林
长　度：27133字符(不计空格)
句子数：791句
时　间：2017-9-25 15:37:52
比对库：学术期刊、学位论文、会议论文、书籍数据、互联网资源
查真伪：http://www.paperpass.com/check

句子相似度分布图:



本地库相似资源列表(学术期刊、学位论文、会议论文、书籍数据):

1.相似度： 1% 篇名：《零售理论与实践》
来源：书籍数据 电子科技大学出版社 2008-12-1

互联网相似资源列表：

1.相似度： 2% 标题：《零售商圈吸引力：基于雷利法则和赫夫模型的实证研究...》
http://www.docin.com/p-516089127.html
2.相似度： 2% 标题：《零售商圈吸引力：基于雷利法则和赫夫模型的实证研究...》
http://www.wenku1.com/view/A56379C849323627.html
3.相似度： 2% 标题：《本文在对商圈理论进行回顾和梳理的基础上-教育学/...》
http://www.tubaobei.com/show-4f336261-0-6570063e.html
4.相似度： 2% 标题：《零售商圈吸引力_基于雷利法则和赫夫模型的实证研究...》
http://www.docin.com/p-228700627.html
5.相似度： 2% 标题：《零售商圈吸引力_基于雷利法则...》
https://wenku.baidu.com/view/445669bf5727a5e9856a619f.html
6.相似度： 1% 标题：《零售商圈吸引力_基于雷利法则和赫夫模型的实证研究...》
http://www.docin.com/p-633992194.html?docfrom=rrela

全文简明报告:

摘要

{42%：由于公共地铁系统的便捷性，使它成为大多数上班族的首选出行方式。} 然而，近年来随着交通数据在数量和种类上的急剧增加，传统的基于统计抽样以及专家经验进行分析的方法已经不再适用， 同时随着数据存储能力的提高以及数据分析与挖掘技术的发展，进行大数据的分析和可视化对城市交通的研究变得很重要。

商圈是零售商店聚集所产生的商业范围，商圈是近些年来商业领域和经济学领域的研究重点之一， 零售商店最关心的是利润，而利润的多少与人流量呈直接关系。 公共交通的便利能够为商圈带去庞大的人流，但是传统关于商圈的研究不能很好地利用大数据的优势，而如何清晰的表达商业与交通数据间隐藏的规律，这就需要我们使用可视化的相关技术进行可视分析。

可视化是通过一系列视觉手段将数据间的关系和数据隐藏的规律清晰的展示出来的一种方式。 本文首先进行了基于地铁刷卡数据的人群移动行为分析，并设计了交互式的可视分析系统，旨在展示时序的交通流量信息以及分析不同群体的移动行为规律。 之后对人流数据和商业数据进行了深层次的研究，并通过可视化的方式从多角度进行分析。 本文通过挖掘销售数据，商圈数据和多维地理空间数据来进行可视分析。 同时本文对商圈吸引力模型进行了深入研究，通过对比已有模型的优势，以及多次相关性分析， 提出了适用于大型城市的商圈吸引力模型，并指出了影响因素与城市的关联性。 同时，本文从新的角度提出了一种规范商圈辐射范围的方法。

本文最后一项工作是关于零售商店选址问题的研究，大型零售商店在我国发展迅速，如何选择合适的新店位置，对企业获得更大的利润至关重要。 在大数据时代到来的今天，零售业海量数据的产生，使得选址问题能够更加客观更加科学的进行分析。 但是同样，由于数据量巨大以及跨学科研究的成本过高，没有一个明确的模式来对选址问题进行全面的研究。选择正确的位置需要大量复杂的信息，例如商业区的属性，客户流程和当前业务绩效。 本文构建了一个交互式视觉分析系统并提供数据驱动视觉比较的方法，用于商业历史数据，客户流数据，选址推荐和可视比较的交互式的查询方式。

本文的研究有很好的应用背景，能够为政府和企业制定相应策略提供很好的辅助。

关键词： 可视分析，交通大数据，引力模型，人群行为，选址推荐

第一章 绪论

1.1 研究背景与意义

在当今时代，大数据已经深入人心，企业和政府都对大数据的研究抱有很大期望，那么如何才能利用大数据来完成各项研究成为了人们关心的重点。 其中随着大数据的发展同样发展迅速的是信息可视化领域。 {43%：可视化的合理应用不仅仅能够更好地展示数据间的联系，也能通过可视分析挖掘隐藏在数据内部的规律。} 信息可视化包括数据可视化、视觉设计、可视分析等内容，这些内容的合理应用以及与数据分析、数据挖掘知识的结合成为了各行各业关注的重要课题。

商业数据由于其本身的特性，即社会性，连续性，综合性以及其获取的困难性和私密性，对其进行可视分析十分困难。 传统的商业分析方法大多是基于统计学的分析，即使是今天仍有很大规模的研究是基于抽样调查和问卷调查来进行。 这样小样本的数据分析很难应用

数据挖掘和机器学习的方法来研究数据深层次的规律，同样也会造成分析结果的误差。 本文通过使用交通数据结合商业数据的方法对商圈进行了一系列的研究。

商圈研究是商业研究中很重要的一部分，商圈是零售业聚集的区域，通常是一个地理位置范畴。 广义上来说就是城市中的各类零售商店的聚集而成的商业街区，包含餐饮，服饰，金融等各式各样的店铺； 而狭义上来说是一家或者多家店铺的覆盖范围。 传统商圈分析主要考虑人口特征，经济基础特点，竞争状况和市场饱和度等因素，但是在大型城市，商圈遍布整个城区， 经济基础特点、市场竞争等因素已经没有很大的区分度，这就要求我们根据实际情况来进行研究。

如何选择合适的位置开设新店对零售企业的发展至关重要，选址的研究同样具有悠久的历史。 但是由于行业差别、政府政策不同，对选址的研究十分困难。 当今企业中对其的研究还大多停留在专家经验以及用户调研的基础上，虽然能够一定程度上代表了选址位置的优劣，但是大多数情况下会有很大的偏差。 而如今已经步入大数据时代，基于海量数据的商圈选址研究已经成为趋势，因此本文基于交通数据与 商业数据进行了可视分析并设计了交互式的可视化系统来进行商圈选址推荐。

1.2 研究现状

计算机存储性能与运算性能的飞速提升，使得大数据走进了日常的生活中，近些年，大数据的飞速发展带动了相关技术的发展。 {49%：大数据分析、数据挖掘、深度挖掘等技术能够有效地解决前些年无法解决的问题，} {41%：同样，信息可视化作为展示数据并且从图形的角度挖掘数据信息最有效的手法，} 也得到了迅猛发展。

交通数据分析与可视化是当前十分热门的研究方向之一，因为无论是个人、企业甚至是政府，交通问题都是其必须重视的。 在国内，北京大学最早进行交通数据可视化的研究，他们做了一系列关于交通拥堵、交通轨迹的可视分析工作， 引领了国内关于交通数据可视化研究的潮流。 在他们的研究工作中，主要集中在对拥堵的研究，有关于十字路口交通堵塞问题的可视化分析，拥堵传递问题的研究以及拥堵解决方案的研究等等。

本文中，我们研究了上海市轨道交通刷卡数据，进行了数据匹配、数据清洗等多项工作，之后进行了轨迹匹配， 完成了基于轨道交通的人流流动趋势研究，并设计了可视化系统来展示我们的研究成功。 最后基于人流流动趋势，我们进行了聚集地的研究与划分。

在可视化领域同样有很多关于商业的研究，但是由于商业数据具有其独特的性质，譬如说私密性以及低精确度，对研究造成了很大困扰。 现有的对商业数据的研究大多停留在企业内部，但是由于技术水平不能和科研机构相比，导致研究结果不能很深入到数据的深层次。同时，现有的关于商业数据的研究大多只是单纯的统计分析，并且没有很好地推广价值。除此之外，对于商业数据有着深入研究的机构和个人绝大部分属于经济学或者商学， 他们能够很好地运用前沿理论，但是无法很好地使用大规模数据分析与挖掘的方法。 抽样调查和统计分析能够很好地对商业数据进行分析研究，但是在大数据时代，有着更准确更便捷的方式， 这就要求我们结合多种学科的优势，进行系统化的商业数据研究。

本文的研究重点关注与大型商圈，基于商业数据与交通数据进行了商圈吸引力模型的研究。我们提出的商圈吸引力模型所计算出的结果是一个概率值，代表了某一地点到此商圈的可能性。同时基于模型计算结果进行了商圈辐射范围的划分，这种划分方式是于传统的商圈辐射范围划分方式不同的， 本文是通过概率值划分，因为模型计算出的结果与实际结果具有很好地相似性。

除此之外，商圈吸引力模型还能够一定程度上推算商圈的客流量，因为其代表了一个地点

去商圈的概率，　那么如果我们知道了这个地点的出行人数，就可以知道这里去商圈的总人数。　幸运的是，我们对交通数据的分析能够很好地帮助我们推测客户流。

香港科技大学进行了多个关于选址问题的研究，选址问题是通过多维数据分析，提出一系列相对适合的选址位置的研究。　本文是对零售商店选址问题进行了研究，我们有幸能够和企业经理合作，获取了相对私密的商业数据，　因此很好地提取了相关影响因素，并进行了统计分析和数据挖掘。　在此之后，我们设计了具有良好使用价值的零售商店选址推荐可视化系统，能够很好地为用户提供统计分析和选址推荐。

1.3　研究目标和内容

本篇论文针对交通数据、商业数据对商圈吸引力和零售商店选址进行深入研究，主要从三个方面入手。　第一是关于居民出行行为和聚集地划分的研究，这项研究能够很好地为商圈辐射范围和吸引能力提供研究基础。　第二是商圈吸引力模型的研究，本文提出了一种新的，适用于大型商圈的吸引力计算方法。　最后本文设计了一个可视化视图来为用户提供零售商店选址功能。

本文的主要贡献有如下几点：

（1）提出了基于交通卡刷卡数据的人群划分方式，并结合可视分析技术判断出居民聚集地与工作地。　设计一个可视化系统来展示上海市轨道交通站点之间的联系，并分析不同人群的移动行为特征。

（2）构建了对大型城市具有很好普适性的商圈吸引力模型，能够有效计算核心商圈对城市任意区域的吸引程度。　{51%：提出了以人为中心的商圈辐射范围规划方法。}　并通过可视分析手段与用户调查方式对其进行了验证。

（3）提出了基于多变量的利润驱动的选址推荐方法。　设计了零售商店选址推荐可视化系统，提供了统计分析，商业分析，选址推荐等功能，并支持交互式查询与智能化展示。

1.4　论文组织结构

本论文的组织结构如下：

{43%：第一章主要介绍了论文的研究背景和意义，分析了商圈研究的重要意义、现状和存在问题，介绍了论文的研究目标、研究思路以及论文的组织结构。}

第二章主要是对本文的相关工作进行介绍，包括交通人流研究，商圈吸引力模型分析，零售商店选址研究等三方面。　介绍了介绍了交通数据和商业数据在可视化领域上的应用以及数据分析方法和统计分析方法。

第三章主要是关于居民出行行为与聚集地划分，首先基于交通卡数据对人群进行划分；其次通过对上班族群体的研究，评估了此群体的人群行为，并提出了聚集地划分方式；　最后设计可视化界面对不同人群的人流情况以及聚集地划分提供了可视分析视图。　并为后面的研究提供了理论基础。

第四章主要讲述了关于商圈吸引力模型的研究，首先通过使用统计分析方法进行相关性与相关系数的研究；　之后通过对经典模型的计算与优化提取出适用于大型商圈的影响因素；{58%：最后进行误差分析验证模型的有效性和可靠性。}

第五章主要根据第三、四章的工作结果设计零售商店选址推荐可视化系统。　主要包括

四个视图的设计与实现，其中商业影响力视图和统计分析视图主要展示商圈信息的统计分析结果； 提出一个多变量的利润驱动的选址推荐方法来完成选址推荐视图的设计； 最后使用可视比较视图来实现推荐结果的交互式对比与分析。

第六章主要对本文的三项研究进行了实验并分析实验结果。 第一部分是关于居民出行行为与聚集地的案例研究，第二部分提供了商圈吸引力模型的案例研究以及用户调查研究，{48%：第三部分提供了可视化系统的使用情况研究与案例研究。}

{54%：第七章对论文内容进行总结，对未来的研究工作进行展望和规划。}

1.5 本章小结

{50%：本章论述了本文的研究背景及研究意义，介绍了当前国内外的研究现状，阐述了交通与商圈结合研究的意义、主要难题和解决方法，} 本文研究目标、研究思路，以及论文的组织结构。 之后的章节会详细地论述本章所提出的内容。

第二章 相关工作

本章主要介绍了交通流量、居民行为与聚集地划分标准的相关研究，以及商圈研究方法，商圈吸引力模型以及商业可视化技术，可视交互技术的相关研究。

2.1 交通大数据

地铁在城市生活中扮演着中重要的角色，它是能够保证整个城市正常运转的关键部分，并能降低整个城市的交通成本[1]。 公共交通服务在城市内部快速有效地移动大量的人群，由于交通拥堵问题，越来越多的人，尤其是上班族，选择公共交通方式出行。 {41%：大多数公共交通系统采用射频识别卡(RFIDCard)记录乘客的行程。} 在这个过程中会产生大量的刷卡数据，这些数据包括进站、出站、进站时间和出站时间等。 {42%：比如上海的公共交通系统每天会产生上千万条的刷卡记录。} 如何从这些数据中分析整个地铁系统的交通流量变化和探索不同人群在城市地铁系统中的移动行为成为了一项新的挑战。 一些研究工作应用聚类方法分析大量的移动轨迹数据并从这些数据中发现潜在的移动模式[2, 3, 4]。{45%：Tominski[5]等提出了一种基于堆栈的可视化方式来分析轨迹数据的属性。}Crnovrsanin[6]等提出了一种近似的方法研究移动轨迹数据。 这些研究主要集中在如何清晰地展示轨迹数据并没有着重分析轨迹数据背后的移动行为特征。

{43%：近几年来，如何利用可视化技术对庞大而复杂移动轨迹数据进行分析成为了可视化领域的研究热点。} {46%：Andrienko[7]等提出了针对轨迹数据三种可视化研究和分析方法：} 直接描述、总结和模式提取。 {49%：一些研究工作针对不同类型的移动轨迹数据进行分析。} TripVista[8]主要研究交通工具和步行者的个体移动轨迹，并运用ThemeRiver[9]和平行坐标分析移动轨迹数据。 T-watcher[10]主要分析出租车轨迹数据，并设计一个交互式可视化系统对城市交通情况进行分析。 FromDaDy[11]主要针对飞行器的轨迹数据提出了轨迹可视化工具，方便用户对其进行灵活的操作。 上述工作中处理的不同类别的轨迹数据比较散乱，没有相对固定的轨迹，不同于本文用到的地铁交通流量数据，并且较少涉及到分析不同人群的移动行为分析。 此外，还有一些研究工作侧重于分析路径规划和乘客行程分析。 TrajectoryLenses[12]允许用户选取特定的区域和时间范围对起始地到目的地的轨迹数据进行分析。 Zheng[13]等提出了一系列的可视化技术对公共交通系统中乘客乘车路线和交通效率进行分析。

Kieu[14]等提出了一种从智能卡数据中区分不同群体的方法，能够帮助交通管理者了解每个智能卡使用者的出行特征。 但该项研究没有充分结合可视分析技术来帮助用户分析人群

移动行为。 Ma[15]等和Yang[16]等利用可视化技术和聚类方法从移动电话通信数据中分析人群移动行为。 但他们采用的数据不同于本文着重分析的有固定轨迹的刷卡数据。 本文主要是结合可视分析技术从大量的地铁刷卡数据中分析不同上班族群体的移动行为特征以及展示城市地铁系统不同时段的流量变化。 首先，从地铁刷卡数据中区分出上班族群体并推测出上班族的居住地点和工作地点； {42%：其次，用户可以利用本文设计的三个可视化模块解决提出的分析任务。}

2.2 智慧商圈

商业区吸引力模型在经济学领域中已有广泛的研究。 研究者们建立了很多适用于不同情况的吸引力模型，例如有关于城市与郊区商业吸引力的研究[17， 18]，零售商店[20]和购物中心[17，19]的吸引顾客能力的研究。 这些模型通常是在雷利法则[21]和哈夫模型[22]的基础上延伸而来。 {45%：与雷利法则相关模型相比，哈夫模型与其延伸模型可以提供更详细的信息。} 因此，我们的研究之一是优化哈夫模型来计算商业区的吸引力。

{96%：雷利法则的启示在于它内含一个临界的概念，当然雷利法则的延伸模型--康威斯模型能够更清楚地体现这一点：} {100%：在两个商店的直线距离之间，总能够找到一点，在该点的左边属于一个商店的商圈范围，该点的右边属于另一个商店的商圈范围，} {100%：而在这点上的消费者去这两个商店购物的概率其实是一样的。}

{98%：相对于雷利法则，哈夫模型的进步意义体现在它是站在消费者的一个比较微观的角度考察商圈，并且模型的计算结果是概率的形式。} {91%：哈夫意识到消费者对距离的感知是存在差异的，} {98%：并不是完全客观的，所以在公式中消费者到商店的时间距离或空间距离 T的右上角会有一个调解指数λ□} {100%：用来反映消费者对距离的敏感性。} {87%：哈夫模型的商圈吸引力因素只有两个：} {100%：一个是距离 ，另一个是商店规模，一般用商店营业面积来表示。} {100%：但在现代零售业已经展开激烈的全方位竞争的背景下，仅仅用这两个因素来反映商圈吸引力显然已经不够了。} {97%：基于此，布莱克模型对哈夫模型进行了改进，提出了多因素的作用模型。}

在我们的工作中，我们需要获得准确的客户流程，以进一步评估最大可能的利润。 传统商业与交通的研究中很大程度上是基于抽样与专家经验来完成[23， 24， 25]，但是在如今复杂的商业环境下，已经无法满足需求。 传统商圈分析主要考虑人口特征，经济基础特点，竞争状况和市场饱和度等因素，但是在大型城市，商圈遍布整个城区， 经济基础特点、市场竞争等因素已经没有很大的区分度，这就要求我们根据实际情况来进行研究。

但是在经过验证之后，我们发现现有模型很难适用于我们的研究场景，并且会产生较大的误差。 因此为了能够更好地对商圈引力进行研究，我们综合了零售交易区域的辐射范围的研究方法[27， 28]和人群流动趋势的研究方法[26]， 并结合统计分析、数据挖掘和机器学习的方法对商圈吸引力模型进行研究。 最后本文中提出并验证了适用于大型商圈的吸引力模型。

零售商店选址是基于地理空间数据的一项研究，而地理空间数据大多是由交通网络产生。Bartram[29]确认用户可以在交通网络的导航的帮助下有效、准确地得到需要的示意图。Meilinger等人[30]进一步的探索和确认了图解地图在发现和定位方面的价值。 此外，这些研究还表明，在PTS中用户仍然可以获得许多有用的信息即使确切的地理信息没有很好的展示出来。

Andrienko[31，32]等人讨论了运动数据的各种特征，同时总结出了可视化的三个类别：方向描述，概括和模式提取。 之后，Zhong[33]等人和Goncalves[34]等人总结了在运动数

据可视化中的常见技术，例如静态地图、时空概念、动画地图和低维概念等。 最近，一些工作的关注点在运动数据与城市交通的可视分析。 Archambault[35]等人通过测量特定机场周围的距离和路径探索了在全球机场间的飞行数据的连接模式。 Wang[36]等人通过对出租车数据的分析设计了一种多视图的可视化分析系统来研究交通拥堵状况。 Ferreira[37]等人提出了一种新颖的可视分析模型，允许用户从数据百万计的出租车数据中交互式的探索。 Zeng[38]等人提出了交换圆环图的可视化技术来揭示大规模公共交通出行的模式。

关于商业选址方面的研究也有很长的历史，例如Foursquare[39，40]和Heckman-style[41，42]模型。 然而，现有方法不足以支持解决方案之间的比较。 在我们的研究中，我们使用k-location[40，43]方法。 与其他人不同，我们首先量化位置影响因子并计算每个因素的权重，然后获得每个可能位置的优势，最后推荐最佳位置。 此外，我们的研究具有良好的适用性和扩展性。

2.3 可视化技术

视觉比较是一个基本的可视化任务[44，45]。 研究人员探索了[44，46]有效的方法，为快速直观的比较提供了清晰的观点[43]。 大多数视觉比较采用并置方法或叠加方法[47，48]。我们在我们的工作中使用并置方法，如以前的工作[43，49]，并且将所有推荐的并排放在一起，以便用户可以在不同的条件下获得每种方法的优缺点。 {43%：在我们的系统中，我们为用户提供了三个交互视图，以获得更好的体验。}

商业智能已经在零售商店[50]，客户行为分析[51]和市场研究[52，53]中被广泛研究和应用[54，55]。 然而，这些大部分研究通过一系列现场调查完成，具有一定的主观性。 我们的研究重点是解决市场准入策略中的问题，特别是通过利用视觉分析技术推荐零售店的位置。基于热图[56，57]和流图[58]，我们开发基于时间序列的客户流可视化技术，直观地揭示消费者行为模式。

2.5 本章小结

{41%：本章首先介绍了交通数据分析与可视化的相关工作，包括人流数据分析方法、轨迹数据研究方法等。} 然后介绍了商圈研究的相关工作，包括商圈吸引力模型、拓展模型研究现状等。 {42%：最后介绍了数据可视化的相关工作，包括可视比较方法、交互式界面设计等等。}

{52%：第三章 居民出行与聚集行为可视分析}

{42%：由于公共地铁系统的便捷性，使它成为大多数上班族的首选出行方式。} 然而，随着交通流量数据在数量和种类上的急剧增加，使得设计有效的可视分析方法成为挑战。本章着重分析基于地铁刷卡数据的人群移动行为，并提出交互式的可视分析视图，旨在分析不同群体的移动行为和展示时序的交通流量信息。 在流量快照可视化模块中，允许用户选择不同时段的地铁流量信息和分析上班族的居住地点和工作地点。 地铁站之间流量变化展示在流量关系视图中，并可以分析不同人群的出行特征。 流量时序视图展示了整体的地铁流量数据。 最后和交通研究者做了两个案例分析验证该系统。

3.1 分析思路

本章的目标是结合可视分析技术从大量的地铁刷卡数据中分析不同上班族群体的移动行为特征以及展示城市地铁系统不同时段的交通流量变化。 首先，从大量的地铁刷卡记录中发现上班族群体，按照出行持续时间将他们划分为常规上班族和非常规上班族。 其次，根据上班族的刷卡数据推测近似的居住地点和工作地点。 最后，本章设计了三个可视化分析模

块来解决本章提出的三个分析任务。 流量快照可视化模块展示地铁系统中不同时段的流量变化以及进站和出战流量展示，方便用户根据流量分布来推断上班族的居住地点和工作地点，{47%：用户可以选择感兴趣的地铁站进一步分析。} 站点流量关系可视化模块展示了选定站点之间的流量变化，该模块可以呈现上班族群体不同时段的出行特征， 其中弦图表示选定站点与其他地铁线路的整体流量关系。 地铁流量时序可视化模块展示了每一个地铁站为期一个月的交通流量情况， {47%：能够让用户针对不同的地铁站比较工作日和非工作日的地铁交通流量变化。}

{68%：本章的贡献主要包括以下三个方面：}

1)系统化的定义: 如何从地铁刷卡数据中发现上班族群体，并结合可视分析技术推断出上班族群体的近似居住地点和工作地点。

{41%：2)本章提出了一个交互式的可视化系统，帮助用户分析地铁系统中不同时段的流量变化。}

3)本章设计新的可视化模块展示地铁站之间的流量变化，能够清晰地展示不同人群的移动行为特征。

3.2 研究任务

{41%：本章中使用的是上海市公共交通系统中的地铁刷卡数据。} {43%：这些刷卡数据记录了每一个乘客的行程，为期一个月，其中包括工作日和非工作日。} 在公共交通系统中，乘客通过刷卡来进站或者出站，读卡机记录乘客的每一次刷卡行为。 {50%：其中，每一条刷卡记录包括一个匿名的卡号、进站、出站、进站时间、出站时间和价格。} 每条记录仅有出站和进站，为了尽可能准确地获取每个乘客的完整行程，即从始发站到目的站途经的所有站点， 本章利用上海市地铁交通网络数据来推断乘客的完整行程。 一部分乘客的形成会涉及到换乘，换乘是指乘客在车站内进行跨线乘坐列车的行为。

本章把上海市地铁交通网络数据看作一个有向图。 {41%：一个交通网络包含一些站点和路线信息，交通网络中的站点可以看作有向图中的节点，路线可以看作连接节点的边。} 利用构造的有向图能够推测出乘客的完整行程，从始发站到目的站可能存在多条线路， 我们选用时间成本最少路线作为乘客的选择，上海市地铁系统按照乘坐里程计费，时间成本较少等同于价格较低， 这比较符合人们的实际乘车经验。

地铁调度数据包括上海地铁系统中各个线路的进站和出站信息。

{47%：由于地铁的便捷性，乘坐地铁成为了大多数上班族的首选出行方式。} 从地铁刷卡记录数据中能够探索不同人群的移动行为特征。 此外，不同时段的地铁交通流量需要以合适的方式展示出来，以便用户能够了解流量变化。 本章的目标是结合可视分析技术来发现和分析不同人群的移动行为特征以及地铁交通流量变化。 本章的分析任务总结如下:

1)如何分析和展示地铁系统中不同时段的流量变化。 这有助于用户了解不同地铁线路和站点的流量变化。

2)如何从刷卡记录数据中发现上班族的居住地和工作地。 {41%：确定的位置信息能够为分析不同人群的移动行为带来便利。}

3)如何从刷卡记录数据中区分出不同的人群并能够多维度地展示人群的行为特征。

3.3 交通卡数据分析

地铁刷卡数据中的每一条记录包含多个维度的信息，从这些信息中我们能够发现不同的群体以及各自的移动特征。 在此部分，我们介绍了如何发现上班族群体以及推断上班族的居住地点和工作地点。

3.3.1上班族行为分析

每天生成的上千万条刷卡记录中包含着不同的群体，如上班族，老人和游客等。 本文的目标是分析上班族群体的移动行为特征和对应的交通流量变化。 与其他群体相比，上班族群体的出行往往更有规律。 他们的行程在工作日有着连续性，比如从周一到周五都有刷卡记录，并且在工作日有着相同的始发站和目的站。 如果一个人的刷卡记录满足

其中，W表示所有满足条件的上班族集合； W_ i表示对第 i个乘客若满足至少在工作日连续四天存在刷卡记录并出现在相同的始发站和目的站， 则认为该乘客属于上班族群体。

此外，为了进一步分析上班族的地铁移动行为特征，本文定义了常规的上班族和非常规的上班族。 常规的上班族是指那些一周最多连续工作五天的人； 非常规的上班族是指连续工作大于五天的人，通常会在周六或者周日加班。

3.3.2居民聚集行为分析

确定的位置信息包括上班族的居住地点和工作地点，有助于进一步分析该群体的出行行为特征。 对于上班族居住地点和工作地点的推断基于这样的假设： 上班族群体总会选择距离家较近的地铁站作为起始站； 并会选择距离公司较近的地铁站作为终点站。 该假设与我们的实际经验吻合。 关于发现上班族的居住地与工作地的公式如下：

��_ r表示所有上班族的居住地集合，对每一个上班族 w_ i，如果其进站时间在上午五点到十点之间， 那么进站 S_ in就近似为 w_ i的居住地。 L_ w表示所有上班族的工作地集合，对每一个上班族 w_ i，如果出站时间在下午五点到八点之间，那么出站 S_ out就近似为 w_ i的工作地。

3.3.3章节架构

本文设计面向基于地铁刷卡数据的可视分析系统，帮助交通管理者快速分析不同上班族群体的移动行为特征和不同时段的地铁流量数据。 为清晰地表述本文可视分析工具的设计与实现过程，图3.1为该系统的架构图。

首先，在数据预处理阶段，为了进一步分析不同群体的移动行为特征，本文整合了地铁刷卡数据、地铁交通网络数据和地铁线路调度数据， {41%：为发现上班族以及推断居住地和工作地做准备。} 其次，在可视分析阶段，在上一阶段处理的数据结果展示在三个可视化模块中： 地铁流量快照可视化模块展示了某一时刻的地铁流流量信息，其中，饼图中展示的进出站数量信息能帮助用户发现潜在的居住地和工作地； 地铁站流量关系可视化模块展示了不同地铁站之间的流量关系，内层的弦图表示该站到其他线路的流量， 外层的流图表示该站到其他站点的流量变化； {40%：流量时序图展示了一个月的整体地铁流量变化。}

3.4 可视分析系统

如图3.2所示，本文针对提出的分析任务有针对性的设计三个可视化模块： 地铁流量快

照可视化模块(图3.2 A)、流量关系可视化模块(图3.2 B)和流量时序可视化模块(图3.2 C)。

3.4.1流量快照模块

针对第一个和第二个分析任务，本文设计地铁流量快照可视化模块。 如图3.3所示，地铁流量快照可视化模块展示不同时间段的地铁交通流量变化， 用户通过选择位于图3.3上部的蓝色圆圈展示不同时间段的流量信息。 {46%：每个圆代表一个长度为半小时的时间片。}具体的流量信息展示在地图中，同一条地铁线路的两个相邻站点被带颜色的线连接， {44%：不同的颜色表示不同的地铁线路，线的粗细表途经两站的流量大小。} 如图3所示，与其它线路相比，绿色的上海地铁2号线负载了较多流量。 此外，散布在地图上的饼图展示了上班族群体在某一时刻的进站和出站数量，饼图中红色部分表示出站的数量，蓝色部分表示进站的数量。 通过图中饼图分布情况，可以判断出大规模的居住地点和工作地点。 如果一个区域进站的数量较大，即红色部分较大的饼图较多，表明该区域有较多的居民区。 如果一个区域出站的数量较大，即蓝色部分较大的饼图较多，表明该区域可能是分布着大量公司的科技园区。

3.4.2站点流量关系模块

针对上文提出的任务三，本文设计了站点流量关系可视化模块，如图3.4所示，该模块展示了不同站点之间的流量。 当用户在流量快照视图中选择一个站点，该站点到其他各个连接的站点的流量细节信息会展示在该模块中。 图3.4a展示了与选定地铁站连接的流量较大的所有站点。 其中的每一个弧段代表一个站点。 图3.4b展示了站点之间的交通流量，并能够区别出不同上班族群体的地铁交通流量。 如图3.5所示，本文运用流图展示两个站点之间不同时刻的流量，其中上半部分表示进站流量， 下半部分表示出站流量，外层的浅色部分代表非常规的上班族，内层的深色部分代表常规的上班族。 该视图能够不同站点之间的交通流量信息，并直观地展示不同上班族群体之间的出行特征。 图3.4 c中展示的弦图表示选定的站点到其他地铁线路的流量关系，其中颜色与上海地铁线路的官方颜色一致， 该图能够呈现出整体的连接趋势，方便用户分析不同群体的移动行为。

3.4.3时序流量模块

图3.6展示了为期一周的流量变化，在该图中，我们使用较大的正方形映射流量信息，时间间隔为15分钟，横轴为时间轴，从早晨六点到晚上十一点。 从该图中可以看出前三天和最后一天早晚高峰流量聚集明显，中间三天流量分布与其它四天不同， 是因为这三天是法定假日，使得整天的交通流量比较分散。 为了探究流量变化的细节，我们采用了更细的力度来展示流量信息，每一条线表示每一分钟， {67%：用颜色来映射流量大小，颜色越深表示流量越大。} 如图3.7所示，流量时序可视化模块展示了每个地铁站为期一个月的进站和出站流量信息。 这里用颜色的深浅映射一个地铁站在某个时刻经过的流量，颜色较深表示此时的流量较大，颜色较浅表示此时的流量较小。 每一个长条代表每天从早上六点到晚上十一点的流量分布。 通过该视图，用户能够直观地了解该地铁站一个月的流量分布情况。在图3.7中，从上午八点到九点和下午五点到七点有明显的地铁交通流量聚集现象。 这种现象只发生在工作日。 在周末或者假日，整体的出行比较分散，并且大部分地铁交通流量分布在中午十二点以后。 在该视图中，我们还可以发现一些有趣的现象，比如从下午五点到七点，在流量聚集区中会有一些间隔。 这可能是因为不同的下班时间引起的。 关于这些，本文会在案例分析部分详细讨论。

3.5本章小结

我们的案例分析验证了可视化系统的有效性，能够帮助用户分析地铁系统中不同时段的流量变化信息，　并结合可视化模块发现上班族的居住地点和工作地点以及不同上班族群体的交通流量。　{40%：本文提出的可视化系统主要围绕着上述问题，但仍存在许多问题值得仅以研究。}　我们的系统侧重于分析和展示整天和整月的流量信息，但并不支持更粗时间力度的地铁交通流量分析，比如每个月之间的地铁交通流量对比。　在不同群体的移动行为方面，用户可以发现常规上班族和非常规上班族的移动行为特征，　但更加多样的群体信息没有被展示，这是因为地铁刷卡数据中的群体信息维度较少，　需要结合其他的数据源来挖掘出更丰富的信息。

第四章　商圈引力模型研究

在本章中，我们通过分析不同模型的优劣以及相关因素的影响程度进行商圈引力模型的构建。　具体做法是通过实际人流数据计算出商圈对所有地点的吸引力程度，用概率表示。{42%：之后使用经典模型进行计算，分析实验结果，找出实验结果产生误差的原因。}　{52%：最后分析相关影响因素的影响因子，构建引力模型。}

4.1　分析思路

{44%：商圈是零售业聚集的区域，通常是一个地理位置范畴。}　广义上来说就是城市中的各类零售商店的聚集而成的商业街区，包含餐饮，服饰，金融等各式各样的店铺；　而狭义上来说是一家或者多家店铺的覆盖范围。　本文基于城市的轨道交通数据和商业数据，对大型城市的核心商圈（广义）来进行的研究，探索了商圈的吸引力与辐射范围。

商圈的研究有着久远的历史，从最早雷利进行商圈吸引力程度的划分，到哈夫进一步提出商圈的影响力模型，　再到如今各种适用于不同类型商圈的模型的构建，商圈的研究更趋向于定制化。　但是，由于各个国家的经济基础与发展情况不同，很难有一个通用的模型来解释商圈的优劣。　今天，大城市的交通迅速发展，曾经孤立的零售商圈串联起来，曾经用来对商圈吸引力，商圈辐射范围与程度的模型已经无法适应如今的商圈研究。　本文主要通过对轨道交通数据的分析，提取出顾客群体，并根据这个群体的形成特征，计算出相对意义上的商圈的真实吸引力与辐射能力。

传统商圈研究中很大程度上是基于抽样与专家经验来完成，但是在如今复杂的商业环境下，已经无法满足需求。　传统商圈分析主要考虑人口特征，经济基础特点，竞争状况和市场饱和度等因素，但是在大型城市，商圈遍布整个城区，　经济基础特点、市场竞争等因素已经没有很大的区分度，这就要求我们根据实际情况来进行研究。

商圈吸引力指的是一个商圈吸引顾客来此购物的能力，这是基于万有引力定律而产生的。商圈的吸引力不是一成不变的，由于城市发展，交通变迁等因素，商圈的吸引能力也是有所变化。　但是在大型城市商圈的发展相对稳定，成熟商圈的变迁缓慢，这就为我们的研究提供了便利。　商圈的吸引力主要受商业面积、商圈等级和知名程度等因素影响，这也可以称为商圈的魅力属性，　而商圈对某一地点的吸引力还要加上这一地点到商圈的阻力因素。　本文中通过多种形式的研究，提取出了最可能产生影响的因素来进行商圈引力模型的构建，并验证了它的可行性和有效性。

多个商圈的辐射范围是可能会产生重叠的，在之前的研究中，大部分的工作是根据移动设备（手机）信号来进行的研究，　这样做能够很有效的得到顾客行程轨迹等结论，但是会出现一些问题。　首先，大型城市核心商圈附近总会提供很多工作岗位，这就造成了大量从业人员的产生，移动设备信号是无法区分这些人群的区别；　然后，许多商圈是依托于交通枢纽而建成，一些具有很高低位的交通枢纽会产生很多路过人群，这就对实际的研究造成了很大困扰。

本文中，我们基于轨道交通数据根据不同人群特征进行了人群类型分离，然后基于购物人群进行进一步研究。 除此之外，之前研究中，商圈辐射范围大多是基于人数来进行划分，本章中，我们根据某一地点到商圈购物的可能程度进行划分， 这两种方式各自有各自的优势，会在后面详细介绍。

本章的工作和研究一共有以下几点贡献：

1.构建了对大型城市具有很好普适性的商圈吸引力模型，能够有效计算核心商圈对城市任意区域的吸引程度。

2.提出了新的商圈辐射范围规划方法（以人为中心），并使用可视分析的手段展示了商圈辐射范围内不同类型居民与企业，为决策提供帮助。

3.提供商圈预测的方法，能够很有效的对新商圈的可能效益与发展方向进行预测。

4.2 研究现状

零售商圈是零售交易区域的辐射范围，但是商圈的概念并没有很明确的定义，本文中，我们认为零售商圈的商业企业聚集所形成的空间范围。 {44%：同样，在商圈级别划分标准下，本文中所研究的商圈是指由核心商业圈和次级商业圈组成的空间范围。} {46%：商圈理论中应用最广的是雷利法则和哈夫模型以及其的演化模型。}

由于大多数情况下，企业很难获取详细的商业信息，那么如何选择投资地成为了一个难题，而雷利法则最早为企业提供了容易实现的决策指导。 雷利法则认为商业也具有相互吸引的特性，它以万有引力定律为核心，来确定商圈吸引力临界范围。 但是雷利法则是以商圈为中心的研究，并且需要有较严格的前提才能使结果有效。 在我们对上海十九个大型商圈进行研究之后，我们发现使用雷利法则是很难确定商圈范围的，由于商圈差异和人群行为等因素。

与雷利法则有所不同，哈夫模型是从顾客的角度进行研究，通过模型计算出一个概率值，这个概率值能够代表当前用户去往某个商圈的概率， 但是模型计算概率的时候用到的阻力和魅力因素仅仅包含距离和商店规模。 在大都市商圈引力研究中，这两个因素依旧很重要，但是其他因素，例如地理位置，商品档次等，对概率的计算同样占据很重要的位置。

雷利法则是W.J.Reilly最早在1931年提出，它的核心观点是： {100%：具有零售中心地机能的两个都市，对位于其中间的一个都市或城镇的零售交易的吸引力与两都市的人口成正比，} {100%：与两都市与中间地都市或城镇的距离成反比。} 模型(康帕斯—Reilly变形)如下：

{49%：其中D_ab是A城市的辐射范围（与B相比），d为城市A和B之间的距离，P_a和P_b分别是两个城市的人口。} 在计算过程中，为了更加符合实际以及方便计算，我们用商圈代替城市，时间成本代替距离，商圈所在行政区人口总数进行计算。

雷利法则在一定程度上能够确立商圈辐射范围，但是由于它考虑的因素过少，同时它认为某地在选择商圈购物时具有唯一性， 因此导致误差十分大，我们使用康帕斯法则得到了商圈辐射范围，但是和实际有很大误差，它的误差如图1所示， 其中选择的四个商圈分别具有不同的特征，徐家汇同时具备交通枢纽和金融中心的地位； 南京东路拥有相当多的小型商场，同时吸引了大量游客； 中山公园是重要的交通枢纽； 龙阳路连接了上海浦东区域的郊区与市区，本文中的研究大多使用了这四个商圈。 同样，应用雷利法则还需要具备几个前提： （1）两个城市（商圈）交通情况类似； （2）两个城市（商圈）属性类似； （3）两个城市（商圈）人口（人群类型）类似。 而在实际的研究中，这些因素很难测定，并且具有很大的差异性，我们对满足这些前提的商圈进行研究。

哈夫模型是由David.L.Huff于1963年提出，它认为： {98%：从事购物行为的消费者对商店的心理认同是影响商店商圈大小的根本原因，商店商圈的大小规模与消费者是否选择该商店进行购物有关。} 我们对商圈吸引力的研究以及对大都市零售商圈引力模型的建立主要是基于哈夫模型，模型如下：

{48%：其中 P_ ij为 i地区顾客到商圈 j消费的概率， S_ j是商圈 j的魅力， T_ ij i地区顾客到商圈 j的阻力，} {51%□μ和λ是以经验为基础所估计的修正值， n是互相竞争的商圈数。}

哈夫模型从某种程度上得到和实际值很相近的商圈吸引力概率值，但是如果相互影响的商圈数目过多， 那么它的计算精度会有一定程度的下降（如图4.2所示），同时在当前经济背景下， 更多的因素会对商圈魅力产生影响，而阻力的测定也不仅仅是基于空间距离。 在本文中，我们根据文献并且与企业市场经理多次交流，总结了十几个可能对商圈魅力和阻力造成影响的因素， 并进行了相关性研究（5.1），之后对商圈吸引力模型进行了设计（5.2）。

4.3 引力模型分析

4.3.1相关性与相关系数

我们计算了哈夫中两种主要变量对概率值的显著性与相关性，如表1， {41%：我们可以看到商业面积与时间成本对商圈吸引力来说都具有显著地相关性，同时时间成本的相关系数为-0.489，} 商业面积的相关系数为0.149。 在与零售企业经理讨论后，我们认为所得系数是相对可靠的，因为在上海， 交通相对发达，时间成本的影响程度已经远没有十几年前那么大，而由于研究样本都为核心商圈， 商业面积的区别程度不大，因此得出的影响因子过小。 在对大型商圈的研究中这也是可信的。

由公式4.2可以看出，在哈夫中，μ和λ是模型调节指数，由于在商业方法中， 这两个指数是由相关领域专家通过经验得到，为了对商圈吸引力的研究更加深入，我们邀请了相关领域专家， 帮助我们给出两个调节指数值，作为主观指数值，同样，我们通过大样本相关分析，得到了相关系数， 把相关系数作为一组调节指数，作为客观指数值。 我们加入了一个约束条件，μ+λ=2□经过归一化处理和放大处理之后，我们得到了两组调节指数值。 我们使用具有两种指数值得模型进行了商圈吸引力概率的计算，得到了商圈的辐射范围。 两种调节指数值如下表：

我们通过加入调节指数，使用哈夫模型进行计算，辐射区域划分结果如图4.3所示，其中我们使用的是进行归一化与放大操作之后的指数值。

通过模型计算结果的可视化对比之后，我们可以清晰地看出，经过指数调节后的模型精度有了明显的提高， 但是两种指数调节方法并没有很明显的优劣性，经过讨论后，我们认为这是由于哈夫模型仅仅使用商业面积和距离来进行计算的原因， 而实际中，魅力和阻力的确定更加复杂。 为了能够得到更准确地吸引力值，我们使用机器学习的方式对数据进行了训练， {43%：得到了一组影响因子的值，可能影响因子与训练结果如表4.3所示：}

使用机器学习方法得到的影响因子都没有很高，我们讨论后认为导致这个原因的可能因素是因为商圈的吸引力影响很复杂，同时主观情绪占有一定比重。

4.3.2引力模型设计

{43%：商圈是具有吸引力的，我们模型设计的基础同样是万有引力定律。} 面积越大，商品种类越多的商圈自然而然的吸引更多的消费者，但是在本文的研究中，商圈选取的都是大

城市的核心商圈， 商圈属性的差异比较小，在我们的相关性分析中，也能得到同样的结论：{42%：商业面积，商圈等级等因素对吸引力结果的影响因子都没有很高。} 但是一些因素，例如时间成本，换线次数等，对顾客选择商圈产生的影响要高得多。

在经过多次验证和分析之后，我们提出了一个适用于大城市大型商圈的商圈吸引力模型：

其中，α□β□γ和δ是基于数据分析和挖掘所得出的调节指数。 我们可以得到一个能够一定程度上表示商圈魅力的Attraction和一个能够表示商圈对某地吸引程度的Attractive。

我们认为影响顾客选择商圈的阻力因素主要为时间成本，尽管其受到一定的主观情绪影响（地铁线路图中的距离等因素）， {61%：但是这是影响顾客选择商圈的一个很重要因素。}其次，商圈商品档次，商圈内商场数，商圈内商业面积，商圈知名度为主要魅力因素。

而某地顾客到各大商圈的概率值则为：

因为在我们现阶段的工作中，无法验证调节因子的值是否适用于所有相似的大城市商圈。因此，我们通过多种分析方式所做的调节因子研究暂时只适用于上海市。

本文中，我们使用了两种方式对调节因子进行研究与确立，统计学相关性分析的方法与机器学习训练因子的方法。 我们分别求出了不同因子的对应值，并对其进行对比研究。 结果如图所示：

但是我们发现在某些情况下，计算出的值会有较大的误差（图4.5， {41%：图4.6），我们经过讨论认为，可能产生这些误差的因素主要有两个，一个是公交车对数据统计的影响，}另外一个是换线次数对模型精度的购物阻力的影响。 为了能够更加准确地预测出商圈的吸引力与辐射范围，我们对实验样本进行了分割，并对阻力值进行了优化。 得到公式：

{47%：我们使用优化模型进行计算，得出的结果对比如图：}

4.3.3模型误差分析

我们使用原始模型进行计算，得出了吸引力值，然后计算其与实际值得平均误差，如图4.4所示， 由于计算结果为0~1之间的概率值，我们使用绝对误差展示：

我们可以发现，在整体情况下，时间成本和误差正负与大小没有必然的联系。 我们认为可能是由于原始模型的不适用以及误差过大导致，之后我们同样对我们模型（公式4.5）的计算结果进行误差研究， 如图4.5所示：

在图中可以清晰的看出时间成本越低的位置产生的误差越大。 同时，我们对具有不同调节指数的哈夫模型计算值进行比较，如图所示：

其中客观指数调节，很明显具有更好的准确度，因为其计算结果中误差值小于0.025的较其他两种要高很多， 而主观指数调节得到的值能够得到更多的误差小于0.01的数据。

在对较大误差地点进行单独分析时，我们发现，这些站点大多去某一商圈的时间成本小于10 min，经过讨论， 我们认为，这是由于数据所产生的误差，因为我们使用的是地铁刷卡数据，对没有加入同样占有公共交通很大比例的公交数据， 而商圈附近居民更偏好于乘坐公交车到最近的商圈购物，这就导致了我们测量的实际概率值有误差， 这种误差主要体现在，过小的估计了最近商圈对居民的吸引力，导致在之后的计算中产生了较大的误差，但是在现阶段的工作中，我们暂时无法解决这个问题，为了再次提高模型的适用程度，并更好的进行优化和改进， 我们去除了这些时间成本小于10 min的地点，再次进行模型计算，

实验结果如图4.6右，其中共有201个位置，共3819组数据。

和上图对比，我们可以看到，在去除这些小时间成本的位置之后，我们得到的结果具有更小误差的位置更多，因此，我们认为数据误差在本文的研究中是客观存在的，但是如果我们暂时剔除这些点，我们能够得到相对误差更小的结果。

在我们通过实际值绘制辐射范围图时，发现了一个很明显的特征，这是在我们之前的研究中没有重视的，那就是在多种影响因素中，换线次数得多少对结果有很大的影响，我们对统计数据进行分析，发现如果两个商圈对某地的吸引力和时间成本大致相等，那么通过更少换线次数能够到达的商圈更具有吸引力，同时，他们的差异是很明显的，我们对一些具有上述特征的位置和商圈进行深入分析，结果如下（选择商圈—中山公园，徐家汇；地点—时间成本差值小于5min，认为时间成本一致）：

{42%：其中中山公园为2，3，4号线交汇，徐家汇站为1，9，11号线交汇。}通过可视分析，我们可以清晰地看出具有相同时间成本的位置与两个相似的商圈之间的联系，其中紫色为更加偏好徐家汇的人，绿色为更加偏好中山公园的人。图中，具有更近地理位置并不意味着具有更好的吸引力，可以很明显的看出，如果没有换线的话，那么顾客会更加偏好这个商圈，尽管成本相同，同样的，换线次数越多，那么这个商圈对顾客的吸引力越差（或者说阻力越大），这样，在我们的研究中需要加入换线次数这一个因素，将会很好地提高准确度。

4.4本章小结

本文从交通数据入手，进行了人群与聚集地的分析，用以辅助零售企业决策，像对群体有目的性的进行零售商铺的规划。同时对商圈吸引力模型进行分析与验证，并根据统计分析与机器学习的方式提炼出相关影响因素，并根据学习出的影响因子进行模型计算。我们的计算结果通过可视化图表的形式展示，同时我们提出了以地区（人）为核心的商圈辐射范围划分方式，与普遍的以商圈为核心的商圈辐射范围划分方式相比，能够有效的解决数据采集不完全的问题。同时以概率为划分标准更能够表示出不同商圈对顾客的吸引程度。

同时本文发现，在大型城市商圈的研究中，是否换线对顾客选择商圈的影响很大，即使时间成本相同。在本文中，为了方式过拟合的发生，在模型中我们只是用了六个变量，但是这六个变量足够进行商圈吸引力的研究和辐射范围的划分。除此之外，在6.2的研究中，我们认为，本文的研究能够很有效的对城市商圈规划进行预测，{40%：为城市规划者提供帮助，同时为零售企业决策者定向销售提供很好的指导。}

本文构建了大型城市（上海）的在现阶段商圈吸引力模型，由于城市和商业的发展相当迅速，许多过去使用的影响因素在如今可能失去了曾经的重要地位。这就要求我们在进行商圈的研究中要时刻关注最新的城市发展细节，例如公共交通的繁荣把传统意义上的距离用时间成本替换，而商业面积由于商圈会不断吸引企业入驻而具有更小的差异。{41%：因此商圈吸引力的研究会根据经济与地区的发展而有所不同。}本文所构建的模型在现阶段只适用于公共交通便利，商业发达的大型城市。在未来我们也会对中小城市进行深入研究。

同时，由于数据限制，本文所做的研究仅仅基于轨道交通数据。但是据我们了解，在上海这种城市，公共车占据公共交通的三分之一，这是不可忽略的。在下一阶段的工作中，我们会着重研究这一部分。

{64%：第五章 零售商店选址可视分析}

{48%：本章介绍了零售商店选址问题的相关可视分析。}零售商店选址是一个跨学科的

研究问题，涉及到社会学、商业、数据分析等领域。 本章的工作主要基于大数据分析与可视化技术，并通过与企业经理的实时交流，完成了以应用为目的的商圈选址推荐可视化系统，旨在为企业提供更加高效的零售商店选址策略。

5.1研究任务

本章通过对业务场景的研究，并结合多学科知识，包括经济学，统计学，营销理论和科学计算，完成了选址工作的研究。 我们邀请的营销领域，企业经理和信息可视化专家共同相关需求的确立以及工作效果的验证。 {44%：我与多位专家达成共识，主要需要解决下列问题：}

1）确定商业区。 商店选址问题所要考虑的第一个问题就是商圈的选择，大型商圈会有更多的客户，但是竞争会更加激烈， 而小型商圈会有更少的顾客光临，这就要求我们寻找一个相对最优的方案。

2）选择合适的商场。 {43%：在确定选址所在商圈之后，下一个需要解决的问题就是商场的选择。} 以徐家汇商圈为例，徐家汇商圈有十几个大型购物商场，但是每个商场的商品等级、面积、地理位置都有所不同， 即使在同一个商圈，不同商场的差别也十分巨大。{41%：企业必须考虑运营成本（店铺租金，劳动力成本等）和竞争环境。} 收集上述信息是一项艰巨的任务，在和专家讨论之后我们认为，客户流对销售和利润至关重要，应在决策中更加关注。

3）为用户提供各种解决方案。 即使是通过模型和算法计算出来的推荐位置也许并不适用与选址。 在讨论后我们认为，许多政府政策、环境变化都可能导致选址出现偏差。{49%：解决这一问题的方法是提供多个备选方案供用户选择。}

4）方案评估。 即使提供了多个备选方案，这些方案的合理性也有待商榷，这就要求我们对方案进行评估。 在没有科学评估方法和直观表现手段的情况下，很难让用户认同。因此我们需要一个方案评估方法。

为了解决上述问题，本文使用大数据分析方法，其中主要数据集为以下三种：

1）交通卡刷卡数据： 包括刷卡时间，刷卡地点，交通工作类型，消费金额等属性，共有4.7亿条记录，记录时间为2016年4月份共30天。

2）服装零售店销售数据： 包含2016年2月到8月共计六个月内关于商品原价，折扣价，类别和销售数量的信息。 共有153家店铺。

3）商圈以及商场数据： 包含118个百货商场和零售商场的确切位置，每间商店的面积，租金成本等。

根据和专家讨论总结出的四个步骤，我们提出了一下任务作为研究重点：

T1： 选址模型设计： 如何调和各因素之间的关系？ 如何判断推荐地点满足预期？如何选择合适的影响因素进行分析？

T2： 商业影响范围： 如何划分商圈的影响范围？ 商圈之间是否有特殊的联系？怎样分析不同商圈间的关联？

上述任务的重点是关于数据挖掘与模型设计。 此外，为了向用户推荐适当的商业位置，我们需要设计一个可视化系统：

T3： 商业信息的显示： 这里的商业发展现状如何？ 竞争是否激烈？ 用户希望获得有用的信息，而对我们来说，统计分析和聚类分析很重要。

T4： 位置推荐： 推荐位置在哪里？ 是否有更大的发展前景？ 视图中将显示各种选项以满足用户要求。

T5： 位置比较，评估和排名： 不同店铺位置有什么好处？ 这些排名的基础是什么？有必要提供详细的说明来说服用户。

5.2客户流预测

我们使用上一章提出的模型（公式4.4）进行关于客户流量的预测，当我们确定某一天商圈对某一地点的吸引力程度， 同时知道这一地点的大致出行总人数，那么我们就可以估算出这一地点到商圈的总人数， 同时计算多个地点就可以得到此商圈这一天的客户流量的预测值。

通过一个月的数据分析，我们发现除去个别站点，几乎所有站点工作日和节假日的出行人数大致不变，那么我们就可以估算出客户流量多少，公式如下：

我们可以得出预测顾客流的大小，进而完成零售商店选址推荐工作。

5.2选址推荐模型

{43%：研究表明，影响选址的主要因素包括店铺多样性，店铺竞争情况和顾客流量等因素。} 我们研究了选址模型来解决T1提出的问题。

影响店铺收益的最主要因素的顾客的多少，这直接由客户流大小来确定，这部分的研究重点是顾客流量的确定，本节通过交通流量进行预测。

本节首先比较了不同商圈对用户的吸引力，然后研究了同一商圈内不同购物中心或者百货商场的优势与劣势，来进行协助用户制定定位推荐策略。

{42%：在计算过程中，通过对上海商场位置和销售数据的统计分析，} 发现销售量与商场距交通枢纽的距离（销售量与客户流量成正比）存在很大的关系， {68%：距离越近，销量越高。} 另外，商场的规模也有一定的影响，但其影响也很小。 经过讨论，我们相信在研究中我们可以忽略它。

我们总结了影响位置的八个变量，如下： 劳动力成本，竞争压力，交通便利程度，客户流量，租金成本，店铺规模，市场饱和度和店铺知名度。 首先，我们计算这八个变量的影响因子，并将其量化为1到10的值。

在推荐值的计算中，a_i是每个变量的权重，P_i是影响因子的值。

{45%：对于权重的计算，由于数据不足，难以使用数据挖掘模型。} 因此，我们通过专家评分得到权重大小，并通过迭代方式进行校正。

由于商场总数是相对稳定的，因此选址推荐的范围有限。 我们从118个购物中心和百货商场中通过计算得出最符合用户预期的10个方案推荐给用户。

5.3可视化模块

该模块描述了一组可视化技术，可帮助用户开发市场进入策略。 5.3.1和第5.3.2节是为了T3而设计，主要显示包括行政区划，商业区域的影响范围，销售信息和人流信息等。 对于T4和T5，我们在5.3.3和5.3.4节中运用了一些可视分析手段和可视化技术。

5.3.1商业影响力视图

地图视图在多维地理信息可视化研究中是常见的，在我们的研究中，我们使用地铁图来显示商业区和行政区划的影响范围。

在商圈和行政区域影响范围视图中，我们使用颜色的深浅代表影响程度（图5.2 A），其中最深的颜色表示这里的居民有超过30%去商业区购物， 较浅的是20%最浅的是15%。我们通过聚类所有站（聚类中心是城市大型商圈）得到这些数据。 我们发现，居民的消费偏好有明确的规律： 时间成本对居民选择购物场所的影响程度最大，而商场的面积和知名度也有很大的影响。 例如，徐家汇是上海最大的商业区之一，经过分析比较，发现即使来这里需要更大的时间成本，但是居民还是更加偏好来这里购物。

影响范围计算方法如图5.3a所示。 {44%：其中A，B，C为站，A为目标站，a，b，c为两站中点，绿色区域为目标站影响范围。} 首先，我们计算了目标站和相邻站之间的中点，然后我们通过目标站得到中位线。 最后，我们按照两站之间的距离绘制半圆。 如果有多个站，我们也可以使用相同的方法。 在计算大致影响范围之后，我们连接这些图形成一个影响区域，如图5.3b所示。

同时，我们根据行政区域划分所有站点（图5.2B），每个颜色代表一个区域，用户可以通过交互功能来选择想要显示的信息。 此外，此视图还提供了一个缩放功能，以便获得更好地用户体验。

5.3.2统计分析视图

{41%：这部分是一个多层圆形视图，用于展示详细的统计信息以及时序信息。}

在本节中，我们将会对统计信息进行详细的介绍。 图5.4A是在系统中显示的试图。此视图提供两个子视图，一个是商圈信息视图（图5.4D），另一个是行政区信息视图（图5.4C）。

在统计分析视图中，每一个Pizza代表一个商业区或行政区域。 其中展示了多种统计信息，例如一个月内的客户流量（图5.4F）以及一个月内的销售情况（图5.4E）。

此外，我们还通过专家经验（T5）预测当前地区的业务前景，其范围在1到10（图5.4B）。

我们使用另外的三个月的数据进行统计分析准确度测试，发现同样时间的人流情况与顾客流动情况大致相似， 与专家讨论后，我们认为产生这个结果的原因主要是由于上海市的经济繁荣以及人口众多， 导致购物者总数是相对平稳的。

5.3.3选址推荐视图

在这个视图中，我们提供了选址推荐的可视化展示。 用户通过输入预期的利润和成本，通过计算，我们会提供相对最符合用户预期的十个位置。 此外，我们还提供了热力图显示功能以及缩放功能。

热力图显示功能主要显示当前选中商圈内所有商场的繁荣程度，而缩放功能能够为用户提供更好地体验。

5.3.4可视比较视图

为了解决T5和T6所提出的问题，我们设计了可视比较视图（图5.6），主要显示不同推荐位置的优势与劣势。 在图5.6 A中，我们为用户提供了每个解决方案的详细比较，并在图5.6 B中进行了总体比较， 最后，用户可以对这些解决方案进行排序（图5.6 C）。

在这个视图中使用的因素由我们从二十多个因素中选出。 首先，通过查找文献以及与专家研究讨论，我们提出了四个类别的影响因素： 消费因素（商圈繁荣程度，时间成本），商场因素（劳动力成本，租金，广告费用）， 市场因素（行业竞争，市场饱和度）和社会因素（经济基础，经济政策，时尚潮流）。

我们经过研究后发现，在这些因素中，由于上海市大型商圈大多位于中心城区，因此市场饱和程度，经济政策大致相同。 而商品种类和销售风格等因素对我们的研究影响不大。因此最初我们选择了十个因素进行进一步研究。

最后，我们发现环境成本和用户行为的两个因素是无法可用的，这是由于在现阶段的研究中我们很难数值化这两个因素， 因此即使它们对结果有一定程度影响，但是我们必须忽略这些因素。

我们使用一到十来表示这些因素的程度，一个是最差的，十个是最好的。 如一个代表最强的竞争压力或最小的客户流量，十个代表没有竞争压力或无数客户。

5.4本章小结

商铺选址问题一直都是企业所关心的重点之一，然而即使企业能够很好地应对市场的变化，关于顾客流的研究也是很不足的。 本章从多个角度进行选址问题的分析。 设计了一个具有良好交互体验的可视化系统，该系统能够展示关于商业与人流的统计信息，商业影响范围信息以及选址推荐信息。

该系统能够很好地辅助企业经理制定相应的市场进入策略。 但是和企业经理多次探讨之后我们认为只关注与现有商圈与商场是不足的， 只有放眼于新建商场才能更好地进行选址推荐工作，这也是下一步工作的重点。

第六章 实验分析

本章节通过具体实验对本文中提出的可视化系统进行案例分析，同时对选址模型进行用户调查。 对于居民行为可视化系统，我们进行了活跃站点的案例研究与群体移动行为的案例研究。 对于商圈吸引力模型，我们进行了不同模型的对比研究以及用户调查。 对于零售商店选址可视化系统，我们进行了系统有效性评估以及解决方案对比。

6.1居民行为案例研究

{43%：我们与城市交通研究者进行了两个案例分析，以此验证本文提出的可视化系统。}

6.1.1活跃站点的流量趋势分析

通过流量快照可视化模块，用户可以了解地铁网络不同时刻的流量变化。 图3.3展示了上午七点半到八点的流量情况。 可以看到红色的地铁1号线和绿色的地铁2号线比其他地铁线路承载了更多的交通流量。 图6.1展示了中午十二点到十二点半的流量情况，其流量负载明显小于图3.3和图3.2A所示的交通流量负载。 其中图3.2展示的是上午八点到八点半的流量情况。 与早晚交通高峰相比，中午的地铁交通流量明显要小很多。

同时，在流量快照可视化模块中可以通过饼图来验证上班族居住地点和工作地点的分布情况。 城市交通研究者在了解了我们的可视化设计之后能够从地图中找出上班族的居住区域，{44%：如图3.3所示在这些区域中蓝色部分较大的饼图比较多，} 说明该区域进站的人数较多。在地图上，我们发现该居住区域包括东明路新村和南码头新村居民区，这表明专家的发现与实际情况一致。 {42%：在红色部分较大的饼图分布的地方往往聚集着高科技园区，图3.3所示的工作区为张江高科技园区。}

此外，用户可以在该模块中选择感兴趣的地铁站进一步分析。 这里，专家选择了重要的交通枢纽人民广场地铁站作为分析对象。 图3.7展示了该站将近一个月的整体流量分布，可以看到在工作日地铁交通流量主要主机在上午八点到九点和下午五点到七点， 而非工作日的流量比较分散，并从上午十点以后开始密集。 另外，专家注意到从下午五点到七点的流量聚集区中会有间隔的空白格出现，如图6.1所示。 专家认为发生这种现象的原因是不同公司的下班时间可能会不一致。 我们同时查看了其他地铁站的流量情况，这些地铁站周围聚集着一些公司。 这些地铁站的整体流量部分也有同样的现象。 从而表明不同公司的下班时间产生了在流量聚集区中的空白格。 我们还发现在晚上十点左右会有明显的流量聚集，因为在上海地铁系统中大部分地铁线路会在晚上十点左右停运， {45%：乘客为了赶上最后一班地铁，会出现聚集现象。}

6.1.2群体移动行为分析

本文设计的系统支持不同上班族群体的移动行为分析。 如图3.4所示，展示了人民广场地铁站到其他地铁站的流量信息，这里显示的与该站有较大交通流量的20个地铁站。 专家在图3.4中发现与人民广场地铁站相连的地铁站周围分布大量的商业区和科技园区， 这些地铁站与人民广场地铁站之间有较大的交通流量。 结合图6.1，可以发现这种流量聚集现象比较明显，并且非常规上班族在流图中所占的比例较小。 如图3.4中的弦图所示，该站的交通流量主要集中在1号线和2号线，行政区域主要集中在浦东新区和黄浦区。

在对不同地铁站的流量分析过程中，专家通过地铁站流量关系可视化模块发现了两种不同交通流量模式的地铁站。 如图6.3所示，图6. 3 a中流图有明显的流量高峰，这些地铁站通常分布在上班族的工作地点和居住地点附近， 规律的上下班时间是这些流量高峰形成的原因。 图6. 3 b所示的流量分布较为平稳，这些地铁站主要是不同地铁线路汇集的换乘车站，{41%：整天都有比较密集的交通流量，从而导致较为均匀的交通流量分布。} 图6.4进一步说明了这种现象，我们选择了静安寺地铁站和徐家汇地铁站，其中的大部分流图都有明显的交通流量高峰， 这是因为这两个地铁站周围聚集着大量的商业区。

6.2商圈引力模型评估

6.2.1模型对比分析

本文中我们提出了基于顾客的商圈辐射范围划分方法，其中与基于商圈本身的划分方法对比如下图。

由于数据本身原因，即只使用了轨道交通数据，在以商圈本身的吸引程度为基准进行计算时， 会产生误差，这些误差产生的原因多由无法得到的顾客人流而产生的，因此我们经过讨论之后认为， 我们的数据在以顾客本身为核心计算概率进而划分辐射范围的时候，具有更好的优势。

我们进行了模型对比，如图6.6所示，图中展示了哈夫模型和两种因素调节方式与 我们的模型(公式4.3和4.4)在进行计算后所得结果的对比。 由于发现了换线次数对模型计算结果有很大影响，我们对模型进行了进一步优化，计算结果对比如图6.6所示。

我们可以很清晰的看到，加入换线次数之后，魅力值预期过大（阻力值预期过小）的情况有所解决，和真实值相似程度更加明显。

而以位置为核心计算的误差如图6.7所示：

图中可以看到我们的模型相对来说误差更小，经过换线次数调节之后的误差较为调节时准确程度有所提高。 可以有效的说明我们模型和其余模型相比更加有优势。

6.2.2有效性用户调查

为了验证模型的有效性，我们设计了问卷来进行用户调查，问卷内容如附件。

我们统计了所有问卷，共收到有效问卷65份，其中男43人女22人，关于辐射范围判断统计结果如下图。

从图中我们可以清楚地看出，在大部分情况下本文提出的模型更具有优势，但是在仔细区分问卷结果之后， 我们发现在更加繁荣的商圈，加入换乘因子优化能够得到更好的结果，但是如果在一般繁荣的商圈， 并没有更好地结果，甚至会造成辐射范围划分的失真。 这将是今后工作需要继续研究的问题之一。

6.3零售商店选址效果评估

为了评估零售商店选址可视化系统的使用效果，我们与专家和企业经理进行了研究，他们都熟悉了我们系统的一些部分。

5.2.1有效性评估

我们使用二十二天（16个工作日和6个节假日）公共交通数据，然后再使用另外的九天（6个工作日， 3个节假日）的数据来测试预测客户流量的有效性。 比较结果如图6.8所示，详细内容如表6.1所示（其中A-I分别为实验选取的九个商圈）。

从结果可以看出，我们的优化模型基本上能够实现客户流量预测任务，虽然有时会一些偏差。 我们可以看到商业区D，E和H出现了很大的误差。 关于商业区D，专家认为，产生这样结果的原因可能是由于其独特的地理位置，该区域位于大学城附近。 由于这里的学生人数远远大于其他地区，大学生更倾向于使用个人交通工具（自行车，步行），导致预测结果产生偏差。 此外，另一个可能的原因是地理位置偏远，没有公共交通工具的客户将选择使用个人交通（私家车）。 这种猜测也在另外两个偏远的商业区A和B中有所体现。

商业区E的游客数量太多，使用公共交通工具的情况很复杂，到目前为止，我们还没有一个很好的解决方案。

对上述所产生的偏差，专家给出了可能的几个原因： 1）越接近市中心，结果越准确，2）节假日的准确度高于工作日，以及3）商业区内的商场越多预测结果精度越高。 因此，我们从上面三个角度分析结果。 我们的结果如图6.9所示。

我们把预测结果按商业区的商场数量和距离市中心的距离进行排序。 很明显，是否属于节假日于准确性之间没有明显的相关性，商场的数量和到市中心的距离显示了一个明显的规律。 当然，这不包括商业区E，我们认为它在整个分析过程中十分特别。

对于产生这些结果的原因，专家认为，越靠近市中心，商业区辐射范围越广，而由于城市的交通拥堵问题， 更多的人倾向于选择公共交通工具作为购物方式，因此能够得到更好地

预测结果。 离市中心越远，越多的客户选择自行出行，从而导致预测结果的偏差。 商场数量对预测结果有影响，其中显示了一个明确的阶梯分布。 更多的商场意味着更大的吸引力以及知名度，所以客户流更加稳定，从而导致预测结果会更加准确。

5.2.2解决方案比较

{44%：由于系统推荐结果的准确性难以确定，我们以下列方式进行验证：} 首先与专家讨论后输入预期值； 然后把得到的推荐结果提供给企业经理； 最后验证我们的系统的有效性。

我们输入了50万元/月的成本，70万元/月的利润，其中成本包括劳动力成本，租金等，利润为是税前利润。 我们得到十个解决方案（图5.5A），其中在徐家汇商圈有四个推荐位置，新虹桥商圈有两个推荐位置。 我们可以看到具体位置（图5.5b）和解决方案的排名（图5.6）。

我们与公司经理讨论选址推荐结果，他给了我们一些有用的建议。 他认为，除了徐家汇的四个选址推荐方案外，我们的其余方案是相当可行的。 他提到，由于徐家汇商圈规模很大，他所在公司已有几家门店。

他建议我们考虑同一品牌店铺的密度，例如，两个相邻的耐克专卖店将大大影响利润。他还建议我们重点关注上海市中环（一条划分城市的道路）所在的一个范围。 因为他认为，在市中心开设新店的成本远远要高于预期，因此在相对成本较低的地方开设新店更加符合对利润的要求。

他还建议我们加入对纯利润的分析，因为它可以使系统的表达更加清晰。

6.4本章小结

本章中我们主要进行了三项工作的案例研究和用户调查。 首先，我与城市交通研究者交流，进行了两项关于交通流以及人群行为的案例分析； 其次，对于商圈吸引力模型，本文进行了不同模型的计算结果对比以及用户调查； {46%：最后邀请不同领域人员使用本文设计的可视化系统，并进行案例分析。}

第七章 总结和展望

7.1 总结

关于交通、人流以及商圈的研究具有很大的实际意义，能够很好地为政府和企业的决策提供帮助。 本文从交通数据入手，进行了人群与聚集地的分析，用以辅助零售企业决策，像对群体有目的性的进行零售商铺的规划。 同时对商圈吸引力模型进行分析与验证，并根据统计分析与机器学习的方式提炼出相关影响因素，并根据学习出的影响因子进行模型计算。我们的计算结果通过可视化图表的形式展示，同时我们提出了以地区（人）为核心的商圈辐射范围划分方式， 与普遍的以商圈为核心的商圈辐射范围划分方式相比，能够有效的解决数据采集不完全的问题。 同时以概率为划分标准更能够表示出不同商圈对顾客的吸引程度。

同时本文发现，在大型城市商圈的研究中，是否换线对顾客选择商圈的影响很大，即使时间成本相同。 在本文中，为了方式过拟合的发生，在模型中我们只是用了六个变量，但是这六个变量足够进行商圈吸引力的研究和辐射范围的划分。 除此之外，在6.2的研究中，我们认为，本文的研究能够很有效的对城市商圈规划进行预测， {40%：为城市规划者提供帮助，同时为零售企业决策者定向销售提供很好的指导。}

{43%：在本文最后一项研究中，设计并实现了零售商店选址的可视化系统。} 系统中提供了交互式的查询方式，通过商业影响力视图、统计分析视图、选址推荐视图和可视比较视图为选址工作提供直观的分析与展示。

7.2未来的研究工作

我们的工作还有一些不全面和尚未解决的地方，例如在进行人群行为分析时，我们主要使用了轨道交通数据， 尽管其占公共交通的60%左右，但是还有很大比例的人群使用公交车作为出行工具，因此会造成误差。 在商圈吸引力模型的研究中，由于商业面积、商圈等级等静态因素是由官方网站提供， 更新缓慢，因此会有一些数据滞后，造成了计算结果的误差。在商圈辐射范围的研究中，我们以人为中心计算了选择商圈的概率，这样做能够使，模型计算结果很好的符合实际情况， 但是因为人群分布不均衡，可能会在一定程度上不能符合实际情况。 但是我们的模型是根据概率值进行的设计与构建，在以人数划分辐射范围时会有所偏差。 最后，本文设计的可视化系统只能够基于现有的商圈和商场进行选址推荐，在和多位公司经理讨论之后，我们认为需要提供在建商场的推荐。 {62%：因此未来的研究工作主要应该在以下三方面：} 1）融合公交车数据、出租车数据以及轨道交通数据进行人群行为分析； 2）进一步优化模型，使其能够适用于以人数为核心的商圈引力计算； 以及3）提供包含未来商圈的零售商店选址推荐可视化系统。