

硕 士 学 位 论 文

文献综述

**基于数据挖掘的出租车候客站点布置算法的研究与实现**

|  |  |
| --- | --- |
| 作 者 姓 名 | 马 伟 建 |
| 作 者 学 号 | ZY1321130 |
| 学 科 专 业 | 软件工程 |
| 专 业 方 向 | 软件工程与管理 |
| 学院指导教师 | 王海泉 副教授 |
| 企业实习导师 | 周礼元 |
| 企业实习导师单位 | 北京三源合众科技有限公司 |
| 培 养 院 系 | 软件学院 |
| 论文起止时间 | 2014年11 月 ~ 2015年 12月 |
| 文献综述提交时间 | 2014年11 月 |
| 学院指导教师签字： |  |

##### 摘 要

站点候客是出租车提供服务的一种重要方式，相对于巡航类方式而言，它具有节省资源，缓解交通压力的优势。目前，北京市的出租车大部分是巡航类服务，仅在机场、火车站等地方有少数的候客站点。如果在合适的地方引入候客站点，将会有助于提高出租车的运行效率和服务质量。另一方面，基于海量出租车数据集的挖掘可以发现出租车的行为规律，从而有助于候客站点的部署。本文首先介绍了目前出租车的交通服务方式，分析了关于出租车交通服务研究的热点问题及其研究方法，并指出基于大数据的方法对于出租车交通服务研究是比较有效的；然后介绍了目前基于大数据的研究状况，分析了基于大数据挖掘的技术可行性并指出了当前存在的问题。

**关键词：**大数据，数据挖掘，出租车服务，候客站点，站点模型

##### Abstract

Taxi rank segment is an important way of taxi services, which can save more resource and release traffic pressure rather than cruising segment. But nowadays, most of taxi services of Beijing are cruising segment, and there only few taxi station in some important places such as airport, train station etc. If the government set more taxi station in some exact places in the city, the efficiency and quality of taxi service will be improved. On the other hand, data mining of big taxi trace data will discover the law of taxi behavior, which will contribute to deploy of taxi stations. Firstly, this paper introduces ways of taxi services and analyzes hot problems of research based taxi services and its method, which tells us the method based big data will contribute to the research based taxi service. Secondly it introduces the situation of research based big data and analyzes the technology feasibility, and then it points out existing problems.

**Key words**: Big data, Taxi service, Taxi rank station, station model.

目录

[第一章 出租车交通服务研究 5](#_Toc402466588)

[1.1 出租车交通服务的方式 5](#_Toc402466590)

[1.1.1 巡航类服务 5](#_Toc402466591)

[1.1.2 站点候客类服务 6](#_Toc402466592)

[1.1.3 电话预约类服务 6](#_Toc402466593)

[1.2 出租车交通服务研究的热点问题 7](#_Toc402466594)

[1.2.1 热点挖掘 7](#_Toc402466595)

[1.2.2 交通瓶颈问题 8](#_Toc402466596)

[1.2.3 路径规划 8](#_Toc402466597)

[1.2.4 移动模型 9](#_Toc402466598)

[1.2.5 出租车派遣算法 9](#_Toc402466599)

[1.2.6 出租车调度系统的设计 9](#_Toc402466600)

[1.2.7 站点规划 10](#_Toc402466601)

[1.3 出租车交通服务的研究方法 11](#_Toc402466602)

[1.3.1 海量数据挖掘 11](#_Toc402466603)

[1.3.2 图论 11](#_Toc402466604)

[1.3.3 数学模型 11](#_Toc402466605)

[1.3.4 GIS地图数据匹配 11](#_Toc402466606)

[1.4 本章小结 12](#_Toc402466607)

[第二章 基于大数据的研究 12](#_Toc402466608)

[2.1 大数据类型 13](#_Toc402466611)

[2.2 大数据研究的体系结构 13](#_Toc402466612)

[2.3 大数据研究的热点问题 15](#_Toc402466613)

[2.3.1 微博用户社会网络特性分析 15](#_Toc402466614)

[2.3.2 城市交通网络分析 16](#_Toc402466615)

[2.3.3 居民行为建模 16](#_Toc402466616)

[2.3.4 大数据可视化 16](#_Toc402466617)

[2.4 本章小结 17](#_Toc402466618)

[第三章 当前存在的问题 17](#_Toc402466619)

[参考文献 17](#_Toc402466620)

# 出租车交通服务研究

出租车是城市的重要组成部分，如何规划好出租车的交通策略，是影响城市交通状况的一个重要因素。本章首先论述了目前国际上出租车常见的交通服务方式，然后综述了目前在出租车交通服务研究领域的一些典型的研究问题及其常用的研究方法。



## 出租车交通服务的方式

目前国际上的出租车交通服务类型主要分为巡航类服务、站点候客类服务和电话预约类服务，下面分别对其进行简单的介绍。

### 巡航类服务

巡航类服务(Cruising/Hailing Segment)是指出租车不在固定站点候客，而是四处巡游，应沿路乘客的招呼，随时停车载客并将乘客送至指定地点的一种服务[1]。巡游服务的随机性很强，招呼出租车的乘客和寻觅乘客的出租车是在事先没有任何约定的时间和地点购买与出售服务。当然，有经验的司机知道在哪些地点、哪些时段乘客较多，而有经验的乘客也知道哪些地方更容易招呼出租车。在出租车市场提供的三种主要服务类型中，北京市场普遍采用的是巡游类的服务方式。这种招手即停的服务方式在我国大中型城市被广泛采用，其最大的优点就是便于消费者叫车，有乘车需求的乘客只要看到路面上有空乘的出租车，便可招手示意停车，驾驶员停车询问目的地确认可以前往后，出租车服务即告形成。这种简单便捷的服务方式，经多年传承，己被广大消费者所普遍接受[2]。

然而，无论是从社会效益还是从模式本身而言，巡游类服务都存在着一些固有的缺陷。首先，由于巡游服务的特点，乘客与招手即停的出租车存在严重的信息不对称。一般而言，乘客对于出租车的安全性能、是否购买责任保险、司机的服务水平、司机的驾驶经验与技术、司机是否疲劳驾驶等信息一无所知，而且乘客不知道需要等待多长时间会有可供选择的其他车辆出现，更无法预期下一辆车是否比当前的车辆更令自己满意。其次，成千上万的出租车涌入交通主干道漫无目的地游曳，必将占用大量的公共道路空间，加剧道路的拥挤程度，这一点在一些繁华地带，特别是客流量比较大、人口密度高、打车乘客数量多的地段表现尤为明显。最后，出租车在道路上巡游揽客，其成功载客的或然性很强，部分时间处于空载状态，如此一来，便会导致出租车空驶里程增加，运营成本上升，并不利于自身的服务经营[2]。

### 站点候客类服务

站点候客类服务(Taxi Rank Segment)是指出租车在机场、车站、宾馆或者其他固定站点等候不特定乘客租乘的一种服务方式[2]。如果在某一站点候客的出租车不止一辆，他们通常会按照到来的顺序排队侯客。在许多国家，出租车站点都有一个约定俗成的规矩：乘客只能依次选乘排在最前列的车辆。这便意味着先到来的出租车相对于后到来者具有载客优先权，这种规则被称之为“先来先走”(First—in，First—out)规则。“先来先走”规则使得在站点排队候客的出租车和排队候车的乘客可以依次出发，井然有序；但是，该规则也是一种出租车经营者的集体保护机制，它排除了乘客挑选出租车的可能，从而避免了出租车之间的价格和服务竞争。

与巡游式不同，站点候客类服务出于自身特点恰好拥有节省能源、缓解交通压力的优势。一方面，由于该种“守株待兔”的服务模式要求驾驶员静止不动，等待乘客前来询问，故其可以节省路面空载对出租车及燃油等成本的不必要消耗，另一方面，由于不必四处行车招揽生意，站点候客也减少了对道路空问的占用，在一定程度上减轻了交通负担。但与之相反，这类服务方式的缺点则是对候车地点的选择要求较高，如果选择不当，司机的经济收入可能会大打折扣。同时，由于经营者自发地向客流量较多的区域集中，也在某种程度上导致了欠发达或较偏僻地区的出租车供给，从而使得这些地区的乘客在打车时需要花费更多的时间成本[2]。

### 电话预约类服务

电话预约类服务(Telephone—booking Segment)也称电召服务，它是指出租车接受乘客预约，在约定的地点和事件接送乘客[2]。电话预约服务既可以与巡游类、站点候客类服务并存，也可以单独采用。预约服务大多是通过电话联系的一次性预约，但也有提供长期包车服务的。乘客通常不直接与出租车司机联系约车，而是通过出粗车调派中心或派车公司(Dispatch Center or Company)。有些地方，派车公司是法律上独立于出租车经营者的营业机构，他们与出租车公司或个体出租车订立长期合作协议，派车公司接听预订，出租车听候调派；也有地方由出租车经营者组成合作社，共享派车服务；还有些地方，出租车公司或出租车加盟的车行本身具有调派功能。派车公司对预约服务市场的运行具有十分关键的作用[2]。

电话预约类服务的特点十分鲜明。一方面，其极大方便了乘客的出行安排，消费者如想乘坐出租车，可以提前通过通讯中心进行预约，与附近的出租车迅速取得联系，如此一来，可以大幅减少乘客的等车时间，提高其出行效率；另一方面，由于该方式是被动等客，每次出行都有目的，里程利用率很高，这一特点大大降低了出租车的空驶率，避免了无谓的成本消耗，同时也缓解了交通压力。当然，电话预约服务的缺点就是前期投入较高，毕竟建立一个标准化的呼叫中心，需要投入巨额资金，且该种服务类型需要形成规模效应，否则收入难以支撑日常的经营成本[2]。

近几年，一系列打车软件的出现逐渐地充实了出租车的电召类的服务市场，也从一定程度上减少了出租车的空载率，提高了出租车的服务质量。

## 出租车交通服务研究的热点问题

本节主要论述了目前出租车交通研究领域的热点问题，包括热点挖掘、交通瓶颈问题、路径规划、移动模型、出租车派遣算法、出租车调度系统的设计、站点规划等问题。

### 热点挖掘

热点问题是出租车服务研究的一个重要问题，对海量出租车浮动数据进行分析可以发现其背后存在的规律，热点现象就是一个典型的问题，对热点的挖掘是出租车行为分析的一个重要前提。目前，研究者们针对热点问题做了大量研究，例如刘盼盼等人[3]通过研究目前国内外出租车调度系统的应用发展状况，结合空间聚类技术，设计了一种基于出租车载客热点区域的分布式、动态调度方案，提出带有范围控制的空间聚类算法 R-FDBSCAN,对出租车 GPS 历史数据进行均匀、细粒度聚类;并在 Weka 平台中集成 R-FDBSCAN 算法,针对北京市出租车数据进行热点区域挖掘。Kang等人[4]为了解决出租车的分布规律问题，提出了一种改进的DB-SCAN算法将出租车的轨迹数据进行分簇，然后通过卡方统计方法识别出了一些热点区域。为了使得出租车乘客能更高效地打到出租车，Hu等人[5]开发了一个面向服务的在线系统Taxi-Viewer来帮助用户寻找热点区域，即更容易打到车的地方，并且根据用户提供的实时位置给出一个临近的出租车的位置信息。Gui等人[6]提出了一个基于MapReduce的二阶段分布式并行算法从出租车轨迹数据中挖掘城市的热点区域。该算法的第一阶段从原始的轨迹数据中导出了站点信息，并排除了其中的噪声点，第二阶段并行地执行了基于密度的稀疏分簇算法和DBSCAN降噪算法，并发先了其中的热点区域数据。为了将出租车位置数据和地图可视化数据结合起来，Wang等人[7]基于聚合技术构建了一个交互式可视化系统，并通过内核密度估计和聚集层级分簇方法来发现热点区域。**总的来看，目前对热点挖掘的研究已经很成熟，这在一定程度上反映了出租车的行为规律，对出租车的站点规划具有指导意义。**

### 交通瓶颈问题

交通瓶颈问题是出租车行为研究的另一个重要方面。Wei-Hsun等人[8]提出了一个三阶段的时空交通瓶颈挖掘模型，该模型通过出租车的位置数据发现城市交通网络里的时空交通瓶颈。为了检测城市交通条件，Lu等人[9]通过出租车的GPS数据来估计城市交通拥塞的特征，包括信号控制，驾驶员异常行为，不同车辆性能，城市交通的速度分布等等特征。作者首先使用内核密度估计技术对速度分布的概率密度进行估计，然后使用高斯模型分析拥塞的特征。为了提高出租车的运行效率，Yao等人[10]在考虑城市交通拥堵的情况下分析了出租车的调度情况，并基于Dijkstra算法对出租车路径进行优化，设计了一个出车派遣系统。**对交通瓶颈的准确识别有利于城市交通的规划，也是影响出租车站点设计的一个重要因素。**

### 路径规划

对于巡航类服务的出租车而言，驾驶员如何制定行驶路线往往影响着他的收益，因此，路径规划问题也就成为了研究者们比较关注的一个问题。为了对交通资源进行优化，谭卫等人[11]针对出租车的路径规划问题，提出了一种改进的蚁群算法，根据同一蚁群的信息素相互激励，不同蚁群之间信息素相互抑制的原理，该算法实现了出租车资源的合理分布。为了解决出租车的路径优化问题，Hou等人[12]通过使用GIS地理信息系统，并建立了SMS平台和合理的出租车派遣中心，然后结合蚁群算法寻找与行人最近的空载出租车，并提供了一个到行人的最优路径。Li等人[13]通过对海量GPS轨迹数据的统计分析得到了一个分层路网交通经验信息，根据这个信息，道路被划分为频繁路段，第二频繁路段和不频繁路段，最后在此基础上构建了一个出租车的分层路由规划算法。为了更加准确的对出租车路径进行规划，Li等人[14]在考虑能源约束的条件下，提出了一个多目标的路径规划算法，该算法把能源约束下的路径规划问题转换为一个传统条件下的能量可达图问题。

### 移动模型

为了准确反映出租车背后所隐藏的行为规律，研究者们通常用移动模型来描述出租车的运动。例如，为了使用位置服务应用来推断人类的移动模式，Ganti等人[15]基于图论-拉伸因子的概念和隐马尔科夫模型提出了一种识别人类移动模式的算法，并基于上海市的真实出租车数据集做了仿真，证明该算法的准确率能达到90-94%。为了从微观和宏观层面上捕捉城市车辆网络的特性，Huang等人[16]设计了三个参数并且从GPS轨迹数据中抽取出了它们的值，该移动模型可以生成城市出租车的合成轨迹并进行仿真。为了从已有的出租车知识中找出快速到达目的地的算法，Yuan等人[17]从历史出租车的GPS数据中挖掘了智能驾驶方向信息，并基于时间依赖的地标图和可变熵分簇算法提出了一个移动模型，并在该模型上提出了一个二阶段的路由算法来计算实际行程中的最快路径。**移动模型是从一定程度上反映了出租车的行为规律，对站点模型的设计具有一定的指导意义。**

### 出租车派遣算法

Seow等人[18]基于一个多代理架构，提出了一个多代理自动派遣出租车的算法，该算法可以以分布式的方法进行出租车派遣。Yao等人[19]基于城市道路交通状态分析了出租车的调度，并且在考虑交通拥堵的情况下基于Dijstra算法对出租车调度进行优化。为了提高出租车的载客效率，减少出租车的随机巡航，Zou等人[20]建立了一个动态机会模型，该模型考虑了时间对乘客的出现和不同路线的空车载客之间的影响，最后基于该模型提出了一个可以提高载客效率的出租车派遣算法。为了减少乘客的等待时间，Zhang等人[21]基于主成分分析和小波神经网络开发了一个自适应的机场出租车派遣系统，并通过短时预测技术对出租车的供需关系进行捕捉。Ding等人[22]利用实时收集的出租车数据和小波神经网络对目前的乘客进行预测，并构建了一个中心控制，实时管理的出租车派遣系统。

### 出租车调度系统的设计

出租车调度系统的设计是提高城市交通效率的一个重要手段，因此也成为研究者所关注的一个问题。Wei等人[23]提出了一种可以满足城市出租车呼叫应用的轻量路由协议，该路由协议可以预测和更新路由信息，并使用基于跳数距离的贪心算法来快速的寻找到达目的地的路径。为了解决城市居民呼叫出租车的困难，减少出租车的空载率，Zhao等人[24]通过使用无线通信和定位技术设计并实现了一个基于信息自匹配的城市出租车呼叫服务平台，该平台可以实现整个服务区域的实时监测和管理。通过使用服务终端，乘客可以预定最近的空载出租车来等候驾驶。Yu等人[25]基于GPS移动电话和GIS显示算法，设计并实现了一个出租车呼叫和派遣系统。为了解决出租车的不合理分布问题并提高城市出租车的服务效率，Xie等人[26]基于智能手机设计并实现了一个出租车驾驶员和乘客的双向搜寻系统，该系统包括地图匹配、热点分析、出租车路径规划和双向搜寻UI。

### 站点规划

对不同服务方式下的出租车交通问题的研究将有利于提高城市交通的运行效率，提高服务质量。为了逐步完善国内的出租车候客站点模式，很多研究者们已经做出了很大的努力。例如同济大学的张欣环等人[27]通过对出租车候客站设置分级，根据不同道路条件和相应的交通状况，对其设置要求进行了研究；对其设置位置、类型做了详细的分析说明，并给出了推荐值，为城市出租车候客站点的合理设置提供了理论依据。陈仁春[28]以福州市区出租车服务网点为研究对象利用三种方法预测福州市出租车发展规模,以此为基础确定各个交通小区的出租车服务需求量;根据出租车服务网点选址的原则和影响因素,在现有选址模型的基础上,以出租车司机的效益最优为目标,以出租车服务需求、土地限制、乘客出租车需求为约束,在选定备选网网点的基础上,构建了约束型选址模型,并设计了遗传算法对其求解;针对福州市现有出租车服务站点特点,将该模型和算法用于福州市区一级出租车服务网点选址规划,确定选址方案并对方案做出评价,同时还定性给出了二级、三级服务网点的规划。为了提高出租车的运行效率，东南大学的Jia等人[29]通过分析影响出租车运行效率的主要因素，建立了一个概率模型来对潜在的乘客进行预测，通过蚁群算法对其进行优化，从而改进了站点候客类服务的出租车的运行效率。Zhao等人[30]提出了一种基于分布式出租车站点交通策略模式的出租车管理系统，该系统由四个主要部分组成：出租车站点、车辆GPS、移动电话网络和调度中心。该系统放弃了传统的“扫街寻客”的模式，乘客可以通过电话或者网络获取空载出租车的信息，或者直接到出租车站点乘车，这种模式有效地消除了出租车驾驶员和乘客之间的信息障碍，提高了出租车的运行效率。为了方便市民出行，北京市交通部门于2013年7月初，根据民意调查和实地调研的方式设置了600处出租车扬招站[31]。**从目前的研究状况可以看出，研究者们针对国内的出租车站点服务模式已经做出了一定研究，但就北京市而言，出租车的候客站点的选址主要还是通过民意调查的方式选取的，并没有考虑出租车的行为规律，使得站点的选址存在一定的不合理性，需要进一步研究。**

## 出租车交通服务的研究方法

### 海量数据挖掘

出租车的GPS数据中通常包含了很多信息，通过对海量的出租车数据进行挖掘，可以发现出租车的行为规律、驾驶信息，甚至可以推断出路网结构和路况信息，因此对于海量出租车数据的挖掘目前成为了出租车交通服务研究中的一种重要方法。例如文献[4]的热点挖掘、文献[9]中的交通拥堵特征分析、文献[13]中的路网交通经验信息挖掘、文献[16]中的车辆特征参数提取、文献[17]中的智能驾驶信息挖掘等等，均是通过海量出租车数据完成的，而且基于出租车数据集还可以对所提出的理论和模型等等进行仿真验证[15, 16]。

### 图论

图论的方法常常成为出租车路径规划中的一种重要方法，研究者们在对出租车进行路径规划或者派遣算法和调度系统设计时，通常需要将问题抽象为在某种约束下的图上寻径或者优化问题，进而采用图论的方法进行解决[11-15]。

### 数学模型

在解决一些实际的问题中，研究者也会将问题抽象为一些经典的数学问题，从而采用数学模型的方法进行解决。例如文献[15]中的隐马尔科夫模型和文献[20]中的动态机会模型。

### GIS地图数据匹配

通过将出租车的GPS数据和GIS地图数据进行匹配，从而对出租车的行为规律或者城市的道路情况进行研究也是一种重要的方法，因此这种方法也常常被用于热点挖掘[7]、路径规划[12]和出租车的调度系统的设计[25]中。

## 本章小结

综上所述，在北京出租车市场中，站点候客类服务主要集中在机场、车站、地铁及大型写字楼、商场等人口众多的公共场所。比如，经常有数百辆出租车在首都机场、北京西站的出租车通道中排起长龙。而对于其他人口密度不高的地区，采用站点候客方式的出租车则明显减少。因此，为了提高北京出租车的服务质量，加入新的出租车站点将会是一个有效的方法。

目前针对出租车的行为规律的研究已经取得了一定的成果，为基于出租车的交通服务研究提供了一定的研究基础。虽然国内的研究者就站点候客模式问题进行了一定了研究，但缺乏对出租车行为规律的考虑，其准确性还有待进一步的提高，同时，针对北京市的出租车站点模式的研究比较缺乏，因此如何基于站点候客模型对北京市的出租车服务进行改进将有待进一步研究。另一方面，通过对现有的出租车服务研究的常用方法的分析来看，基于大数据的方法可以有效的发现出租车的行为规律，有利于指导出租车站点的设置。

# 基于大数据的研究

在信息技术革命的推动下，以计算机、网络通信等信息通信技术为动力的信息化浪潮席卷全球．经过近10年的信息化与数字化建设，现代城市的运行方式与城市居民的生活环境已经发生了根本性的改变．城市的经济、文化、交通、娱乐等方面都已经和信息化的数字空间紧密融合，网络空间（cyberspace）成为城市居民生活的组成部分．完善的信息基础设施以及丰富的数字化应用成为现代数字化城市的基本特征之一．丰硕的城市信息化建设成果在为人类生活带来极大便利的同时，也为现代城市形态的进一步演进奠定了技术与数据的基础。进入21世纪，以大数据、数据活化为代表的数据科学与技术开始受到人们的广泛关注．以数据为中心的研究方法与技术理念在信息、生物、能源、医药、社会学等不同的学科领域都得到了广泛应用与认可，并促成了大量科研成果的诞生．以信息技术为支撑的数据分析与研究方法正深刻地改变着传统科学探索的工作方式，成为人类科技发展与知识获取的一种新兴模式。

本章主要从大数据类型、大数据研究的体系结构和大数据研究的热点问题三个方面对当前基于大数据的研究进行综述。



## 大数据类型

目前研究者们比较关注的大数据的类型有很多，包括地图与兴趣点数据、GPS数据、社交网络数据、客流数据、手机数据、LBS位置服务数据、视频监控数据、环境与气象数据、社会活动数据等等，这里简单介绍GPS数据和LBS位置服务数据。

安装有GPS接收芯片的移动设备可以收集城市中人、车等流动物体活动信息．例如目前应用比较广泛的浮动车技术[32]就是将出租车、公交车等公共交通工具上安装GPS设备，将其作为传感器对于城市的交通情况进行采样．安装有GPS接收芯片的智能手机也可以当作个人行为轨迹的收集设备．但是由于隐私、安全等诸多问题，手机GPS数据很难大规模收集应用，目前只能依靠志愿者进行小范围收集和研究[33]．

LBS(location based service)位置服务是移动互联网时代一种新兴的网络服务方式，通过LBS应用所收集到的数据具有明确的地理位置坐标并兼具传统Web服务的语义特性．LBS数据是对POI数据的一种深度的描述和补充，与地图和POI等简单的城市地理数据相比，LBS数据包含有大量的语义信息，可以帮助人们更加深刻地理解城市运行动态[33]。

## 大数据研究的体系结构

随着大数据研究问题的深入，研究者们针对大数据的技术体系结构也进行了大量的研究，智慧城市技术体系研究是其中的一个比较典型的问题。作为对整个智慧城市技术研究工作的顶层设计，对于指导领域技术的发展方向、明确研究工作的内涵与外延、优化现有研究资源的配置与分布等均具有非常重要的意义。

浙江大学的潘纲教授在2013年《IEEE Communication Magazine》的Smart Cities专题中阐述了基于轨迹数据分析与挖掘的智慧城市技术体系框架[34]。如图 1所示，该框架将基于轨迹数据的智慧城市技术体系分为轨迹感知(trace)、知识发现(knowledge)和具体应用(applications)3个层次。该框架从概略的层面上系统描述了以数据为中心的智慧城市技术的整体技术路线。

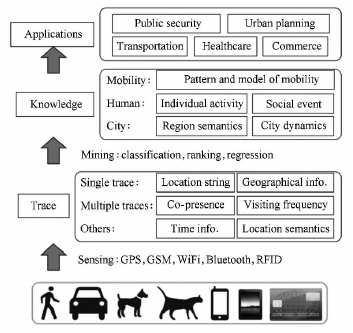


图 1基于轨迹挖掘的智慧城市技术体系框架[34]

微软亚洲研究院的郑宇团队在2013《CCF通讯》城市计算专题中提出了一种“四层反馈”结构的城市计算技术体系框架[35, 36]，如图２所示，该技术体系将城市计算的技术框架细分为“城市感知与数据捕获”、“城市数据管理”、“城市数据分析”、“服务提供”等4个层次。该框架的一个特色在于其引入了“服务提供”层对于真实物理世界的反馈回路，更加完善地考虑了智慧城市技术对于城市生活的影响。

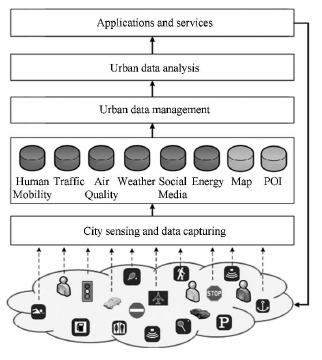


图 2 MSRA城市计算技术框架[35]

## 大数据研究的热点问题

基于大数据的研究目前已经成为一个比较流行的研究趋势，本节简单介绍了几个目前基于大数据研究的热点问题。

### 微博用户社会网络特性分析

自2006年Twitter获得广泛应用后，微博这一新兴社会网络逐渐引起学术领域的关注，研究人员对其社会网络特性进行了相关分析。Java等人[37]对Twitter的基本功能及特点进行了详细介绍，并对其社会网络特性进行了初步分析，数据集包括76,177个用户和1,348,543条微博信息，结果表明Twitter表现出一定的幂律分布和小世界等特性；同时还研究了Twitter用户社会网络的拓扑结构和地理位置等特征，并从个人和社区两个不同层次对用户使用Twitter的意图进行了分析，结果表明用户一般通过Twitter讨论日常事件或共享信息。Kwak等人[38]对整个Twitter进行了定量分析，数据集包括Twitter上的4,170万用户、14.7亿用户社会关系、4,262个热点话题和1.06亿微博等大量数据信息，通过分析Twitter用户间“关注—被关注”的拓扑结构，对Twitter用户社会网络统计特性进行了分析，统计结果表明Twitter在一定程度上表现出用户间的互惠性，但其社会网络特性较一般社会网络存在一定的偏差，例如，用户的follow数并不呈现幂律分布以及分割度更小等不同于一般社会网络的基本特性。总的来说，对微博用户社会网络特性的分析已经成为热点，也是目前社会计算的研究领域中的一个重要组成部分。

### 城市交通网络分析

城市交通网络分析也是大数据研究的一个热点问题。1998年Watts和Strogatz在《Nature》上发表文章提出了小世界(small world)网络模型[39]，该模型描述了从完全规则的网络到完全随机网络的网络转变。小世界网络既具有与规则网络类似的聚类特性，又具有与随机网络类似的较小直径。随后，1999年Barabasi和Albert在《Science》上发表文章指出许多实际的复杂网络的连接度分布都具有幂律形式，由于幂律分布没有明显的特征长度，该类网络又被称为无标度(scale free)网络[40]。在两篇经典网络研究论文的推动之下，复杂网络理论开始在各个学科显现出巨大的能量，并逐渐成为交叉学科研究的热点之一。

### 居民行为建模

城市作为人类活动最为密集的区域之一，在该研究领域中必然不会缺席．在众多城市数据当中，最先被用来分析人类时空行为的是手机数据．文献[41]使用志愿者3 ~ 6个月的短消息通信记录进行了人类活动的时间分析，发现其行为符合幂指数在1.2 ~ 1.7之间的幂律分布．文献[42]利用2006年新年期间600万用户的3000万条短信数据进行统计分析，发送时间间隔和回复时间间隔均符合截尾幂律分布．文献[43]对来自3家公司14万用户的150万条短信记录进行了分析，并发现其符合双模分布，即前段为幂律分布，后段为指数分布。

### 大数据可视化

可视分析是大数据分析的重要方法,能够有效地弥补计算机自动化分析方法的劣势与不足.大数据可视分析将人面对可视化信息时强大的感知认知能力与计算机的分析计算能力优势进行有机融合，在数据挖掘等方法技术的基础上，综合利用认知理论、科学/信息可视化以及人机交互技术,辅助人们更为直观和高效地洞悉大数据背后的信息、知识与智慧。可视分析领域建立在可视化技术基础上，主要强调认知、可视化、人机交互的交叉与融合。面向大数据主流应用的信息可视化技术，主要包括文本可视化[44]、网络(图)可视化[45]、时空数据可视化[46]、多维数据可视化技术[47]。当前，国内仍十分鲜见信息可视化与可视分析的研究成果，各个机构的研究规模也十分有限，迫切需要更多专家学者的关注与支持。

## 本章小结

综上所述，目前基于大数据的研究已经取得了一定的成果，其研究的体系结构也已渐趋成熟，因此基于大数据的研究已经具备的了一定的可行性，而且从目前的研究成果看来，利用大数据去发现现象及其背后存在的规律是一种非常有效的方法，对于出租车服务站点的研究具有很高的指导意义。

# 当前存在的问题

总的来说，目前北京市的出租车服务方式主要是以巡航类方式为主，容易造成空驶率高、资源浪费和交通拥堵等问题，而站点候客式的服务方式可以解决这个问题。但是由于北京出租车的候客站点的选址主要是通过民意调查的方式进行的，并没有考虑出租车的行为规律，导致站点的布置不够合理，需要对出租车的站点模式进行改进。大数据是目前比较流行的一种方法，对于挖掘行为和现象背后的规律而言也是比较有效的。因此如何基于大数据挖掘去发现出租车的行为规律和城市交通的现象，并结合这种规律去对候客站点模型进行设计是需要重点解决的问题。结合前两章的综述内容，总结目前存在的问题如下：

1. 目前北京市的出租车服务方式主要是以巡航类为主，容易造成空驶率高、资源浪费和交通拥堵等问题。
2. 目前北京市的候客站点选址主要是通过民意调查的方式，没有考虑出租车的行为规律，导致出租车的站点布置不够合理。

# 参考文献

[1] Bernstein M H. Regulating business by independent commission[M]. Princeton University Press Princeton, 1955.

[2] 熊宗鹏. 北京出租车市场法律规制问题研究[D]. 首都经济贸易大学, 2013.

[3] 刘盼盼. 基于空间聚类和Weka平台的出租车载客热点区域挖掘研究[D]. 吉林大学, 2014.

[4] Wan X, Kang J, Gao M, et al. Taxi origin-destination areas of interest discovering based on functional region division[C]. London, United kingdom: 2013.

[5] Hu K, He Z, Yue Y. Taxi-viewer: Around the corner taxis are![C]. Xi'an, China: 2010.

[6] Gui Zhiming，Xiang Yu，Li Yujian. Parallel discovering of city hot spot based on taxi trajectories[J]. Huazhong Keji Daxue Xuebao (Ziran Kexue Ban)/Journal of Huazhong University of Science and Technology (Natural Science Edition), 2012, 40(SUPPL.1): 187-190.

[7] Wang H, Zou H, Yue Y, et al. Visualizing hot spot analysis result based on mashup[C]. Seattle, WA, United states: 2009.

[8] Lee W, Tseng S, Shieh J, et al. Discovering traffic bottlenecks in an urban network by spatiotemporal data mining on location-based services[J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(4): 1047-1056.

[9] Lu S, Mai Y, Liu X. The analysis of characterization of urban traffic congestion based on mixed speed distribution of taxi GPS data[C]. Guangzhou, China: 2013.

[10] Yao Zhong-Min，Long Zhao-Peng，Li Qiang. A taxi dispatch system considering urban traffic congestion[J]. Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/Journal of Transportation Systems Engineering and Information Technology, 2013, 13(4): 42-46.

[11] 谭卫，赖斌. 基于改进蚁群算法的出租车路径规划算法[J]. 微型机与应用; MICROCOMPUTER &amp; ITS APPLICATIONS, 2009, 28(10): 72-74.

[12] Hou Z P, Jin F, Yuan Q J, et al. A methodology and modelling technique for taxi path optimization[C]. Hong Kong, China: 2013.

[13] Li Q, Zeng Z, Yang B, et al. Hierarchical route planning based on taxi GPS-trajectories[C]. Fairfax, VA, United states: 2009.

[14] Li J, Liu Z, Yang F. A multi-objective path planning method for electric taxis with energy-constrained[C]. Wuhan, China: 2013.

[15] Ganti R, Srivatsa M, Ranganathan A, et al. Inferring human mobility patterns from taxicab location traces[C]. Zurich, Switzerland: 2013.

[16] Huang H, Zhu Y, Li X, et al. META: A mobility model of MEtropolitan TAxis extracted from GPS traces[C]. Sydney, NSW, Australia: 2010.

[17] Yuan J, Zheng Y, Zhang C, et al. T-drive: Driving directions based on taxi trajectories[C]. San Jose, CA, United states: 2010.

[18] Seow K T, Dang N H, Lee D. A collaborative multiagent taxi-dispatch system[J]. IEEE Transactions on Automation Science and Engineering, 2010, 7(3): 607-616.

[19] Yao Zhong-Min，Long Zhao-Peng，Li Qiang. A taxi dispatch system considering urban traffic congestion[J]. Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/Journal of Transportation Systems Engineering and Information Technology, 2013, 13(4): 42-46.

[20] Zou Q, Xue G, Luo Y, et al. A novel taxi dispatch system for smart city[C]. Las Vegas, NV, United states: 2013.

[21] Zhang K, Leng S, Xu S. Adaptive airport taxi dispatch algorithm based on PCA-WNN[C]. Chengdu, Sichuan, China: 2013.

[22] Ding Y, Xiong J, Liu H. The self-adapted taxi dispatch platform based on geographic information system[C]. Hangzhou, China: 2013.

[23] Wei G, Chen Y, Zhu X. A vanet-oriented routing protocol for intelligent taxi-call systems[C]. Chengdu, China: 2012.

[24] Zhao L, Xing J P, Sun C Y, et al. Design of Urban-Taxi accurate call service platform based on information self-matching[C]. Guangzhou, China: 2013.

[25] Yu J, Zhou X, Zhao H. Design and implementation of taxi calling and dispatching system based on GPS mobile phone: A research for LBS application teaching case[C]. Nanning, China: 2009.

[26] Xie R, Yan K, Zhou J. Taxi bidirectional search system based on smart phone[C]. 2014.

[27] 张欣环，赵磊，晏克非. 城市出租车候客站设置方法研究[J]. 交通与运输; TRAFFIC &amp; TRANSPORTATION, 2010(z1): 15-18.

[28] 王两全. 福州市区出租车服务网点规划研究[D]. 2014.

[29] Jia Y. Improvement program of urban taxi stops based on simulated annealing algorithm in the context of China[C]. Jingzhou, Hubei, China: 2008.

[30] Zhao L, Xing J. Taxi operation and management system based on distributed only taxi stations solving strategy[C]. Guilin, China: 2013.

[31] 北京已建成600处出租车扬招站[EB/OL]. http://news.xinhuanet.com/local/2013-07/07/c\_116436483.htm.

[32] Floating car data[EB/OL]. (2013-10-11)http://en.wikipedia.org/wiki/Floating\_car\_data.

[33] 王静远，李超，熊璋，等. 以数据为中心的智慧城市研究综述[J]. 计算机研究与发展, 2014.

[34] Pan G, Qi G, Zhang W, et al. trace analysis and mining for smart cities: issues, methods, and applications[J]. communications magazine, ieee, 2013, 51(6): 120-126.

[35] 郑宇. 城市计算[J]. 中国计算机学会通讯, 2013, 8(9): 6-7.

[36] 郑宇. 城市计算与大数据[J]. 中国计算机学会通讯, 2013, 8(9): 8-18.

[37] Java A, Song X, Finin T, et al. Why we twitter: understanding microblogging usage and communities[C]. ACM, 2007.

[38] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?[C]. ACM, 2010.

[39] Watts D J, Strogatz S H. Collective dynamics of ‘small-world’networks[J]. nature, 1998, 393(6684): 440-442.

[40] Barabási A, Albert R. Emergence of scaling in random networks[J]. science, 1999, 286(5439): 509-512.

[41] Wei H, Xiao-Pu H, Tao Z, et al. Heavy-tailed statistics in short-message communication[J]. Chinese Physics Letters, 2009, 26(2): 28902.

[42] Zhi-Dan Z, Hu X, Ming-Sheng S, et al. Empirical analysis on the human dynamics of a large-scale short message communication system[J]. Chinese Physics Letters, 2011, 28(6): 68901.

[43] Wu Y, Zhou C, Xiao J, et al. Evidence for a bimodal distribution in human communication[J]. Proceedings of the national academy of sciences, 2010, 107(44): 18803-18808.

[44] Havre S, Hetzler E, Whitney P, et al. Themeriver: Visualizing thematic changes in large document collections[J]. Visualization and Computer Graphics, IEEE Transactions on, 2002, 8(1): 9-20.

[45] Herman I, Melançon G, Marshall M S. Graph visualization and navigation in information visualization: A survey[J]. Visualization and Computer Graphics, IEEE Transactions on, 2000, 6(1): 24-43.

[46] Halevi G, Moed H. The evolution of big data as a research and scientific topic: overview of the literature[J]. Research Trends, Special Issue on Big Data, 2012, 30: 3-6.

[47] Keim D A, Kriegel H. Visualization techniques for mining large databases: A comparison[J]. Knowledge and Data Engineering, IEEE Transactions on, 1996, 8(6): 923-938.