

中南大學
本科生毕业论文(设计)

题 目 轨迹挖掘在机会网络异常发现中的应用研究

学生姓名 李众力

指导老师 陈再良

学 院 信息科学与工程学院

专业班级 物联网工程 1002 班

完成时间 2014.06

中文摘要

最近这几年，个人智能终端设备迅速普及，搭载了谷歌的 Android、苹果 ios 的智能手机已经成为了市场主流。在这些设备上有很多的传感器，如何利用这些传感器获得的数据更好地为人们进行服务成为了当今的研究热点。

本次毕业设计围绕“轨迹挖掘在机会网络异常发现中的应用研究”这个问题展开，我们可以利用轨迹数据进行数据挖掘来进行异常发现，而轨迹数据挖掘的前提是获取轨迹信息，获取轨迹数据之后才能进行进一步的挖掘工作，所以本次毕业设计我将就如何获取用户的轨迹信息进行研究。本次毕业设计我采用 Wi-Fi 无线定位的方式进行用户轨迹数据的采集，本文也将比较详细的介绍 Wi-Fi 定位技术。

本文也设计并实现了一款在 Android 平台上的 Wi-Fi 定位的原型程序。Android 平台良好的开放性，支持应用程序的并行特性，强大的计算能力和数据库技术，都对实现本文的目标提供了强有力的支持。通过这个程序得到的结果来分析 Wi-Fi 定位的现状以及以后的发展方向。

关键词： Wi-Fi 定位 机会网络 轨迹数据挖掘 异常点发现 Android 系统

Abstract

In recent years, the personal electronic device has a great deal of development. With the Apple's IOS and Google's Android operating systems, smart phones began a large-scale popularization. Smart phone has many functions, how to use these functions help people to make a easy life become a research hotspot nowadays.

My topic is "The application of trajectory mining in opportunity network". I can carry out data mining for anomaly detection using trajectory data. How to get the trajectory data is the problem before all questions so in this paper I will introduce how to get user's location with Wi-Fi.

This thesis also designs and implements a Wi-Fi positioning prototype program in Android platform . Android platform support parallel application, has powerful computing ability and support database technology, provides a strong support to the realization of the goal of this paper. Get through this procedure results to analyze the development status of Wi-Fi positioning and future.

Keywords: Wi-Fi Positioning, Opportunistic Networks, Trajectory Data Mining, Abnormal Points Found, Android System

目 录

中文摘要.....	I
Abstract.....	II
第一章 绪论.....	1
1.1 课题研究背景.....	1
1.2 相关领域研究现状.....	2
1.3 本文主要研究内容.....	4
1.4 本文的章节安排.....	5
第二章 基于 Wi-Fi 的定位技术.....	6
2.1 Wi-Fi 信号特性分析.....	6
2.1.1 不确定性分析.....	6
2.1.2 非线性特性分析.....	9
2.3 基于 Wi-Fi 指纹定位的理论基础及一般过程.....	9
2.4 本章小结.....	12
第三章 异常轨迹的发现及应用.....	13
3.1 异常点的定义.....	13
3.2 异常点与数据挖掘.....	13
3.3 异常点挖掘的主要方法.....	14
3.4 轨迹异常点挖掘的应用.....	14
3.5 本章小结.....	15
第四章 Wi-Fi 定位的原型系统的设计与实现.....	16
4.1 系统设计.....	16
4.1.1 架构设计.....	16
4.1.2 数据库设计.....	17
4.2 系统实现.....	19
4.2.1 服务器端搭建平台.....	19
4.2.2 Android 系统编程简介.....	20
4.2.2 训练客户端与定位客户端实现.....	22
4.2.3 服务器端实现——定位算法.....	26

4.3 Wi-Fi 定位中的问题.....	29
4.3.1 城市基础设施建设不完善.....	29
4.3.2 需要大量的现场勘测工作.....	29
4.4 一种“不需要”现场勘察的 Wi-Fi 定位方法.....	30
4.4.1 轨迹数据聚类——逻辑平面图创建.....	30
4.4.2 逻辑平面图与物理平面图匹配.....	31
4.5 本章小结.....	34
第五章 总结及展望.....	35
参考文献.....	37
致谢.....	38

第一章 绪论

1.1 课题研究背景

近些年个人电子设备的发展速度可谓是突飞猛进，几年前智能手机还只是存在于科幻电影中，而最近随着苹果ios和谷歌Android智能操作系统的出现，智能手机开始大规模普及，智能手机上集成了很多其他的功能，Wi-Fi（WIreless-FIDelity，无线宽带）和GPS（Global Positioning System，全球定位系统）成为智能手机的标配，在这种情况下，如何利用好智能手机的这些功能来为人们提供更方便快捷的服务成为时下的一个研究热点。

一个人的位置信息是十分重要的。在道路上，得到你的位置信息可以为你指明方向，在仓储物流管理中，得到物品的位置可以方便进行快速有效的管理，在公司中，得到每个人的位置可以方便管理者了解员工的工作状况。连续的位置信息就是轨迹，基于轨迹数据的挖掘可以更好的获得用户的行为习惯，兴趣爱好等更加详细的信息^[1]。对于轨迹数据而言，有效的信息大部分都是轨迹的异常点^[2]，例如在老年人监护系统中，通过平时的数据积累可以获得用户的正常轨迹信息，若检测到和平常不一样的轨迹信息那就有可能会有意外事件的发生，这样可以及时的通知监护用户。

GPS（Global Positioning System，全球定位系统）是现在获取室外位置信息的最常用方式。随着移动网络的发展，结合手机信号信息而获取位置的技术AGPS（Assisted Global Positioning System）也被应用到各种LBS（Location-Based Services，基于位置的服务）应用中，但是由于人们的一般活动基本都在室内，而GPS信号不能传到室内，所以研究一个实用性强的室内定位系统十分必要。本文没有使用一般的GPS芯片作为位置信息获取的工具，因为GPS有启动速度慢，容易受干扰等缺点，而随着数字化智能化城市的建设，越来越多的城市开始覆盖了Wi-Fi信号，而且现在Wi-Fi早就成为了智能手机的标准配置，本文就将设计一个Wi-Fi定位系统（Wi-Fi Positioning System，WPS）。

Wi-Fi是20世纪发展起来的一种无线局域网技术，技术标准组为IEEE 802.11，目前应用最广泛的标准是IEEE 802.11b和IEEE 802.11g^[3]。Wi-Fi网络具有高速通信、部署方便的特点，切合了现代社会对移动办公、移动生活娱乐的需

求。室内环境和人们活动的热点地区(如机场、写字楼、大型超市、校园、酒店和家庭)是 Wi-Fi 主要的应用环境^[4]。加上近些年智能手机大量普及,智能手机上面几乎都有蓝牙, Wi-Fi 等短距离的无线通信设备, 基于这些技术的室内定位技术发展迅速, 各大科技公司, 学校及研究机构都在竞相研究无线室内定位技术, 在众多技术中利用 Wi-Fi 进行室内定位几乎不需要用户花费额外的购置设备费用, 所以是更有竞争力的是内定为解决方案。

基于 Wi-Fi 的室内定位需要经常采集 Wi-Fi 接收信号强度(RSS, Received Signal Strength)数据并且这些信息随着时间变化可能会改变, 所以需要经常更新, 这就需要连接网络, 传统的通用分组无线服务技术(GPRS, General Packet Radio Service)需要用户支付额外的流量费用, 为了减少用户在流量费用上的开销需要一种更廉价而且可靠的数据传输方式。正是由于智能手机等具有短距离无线通信能力的设备大量普及, 在人群中大规模的部署机会网络成为了可能。轨迹信息属于实时性要求比较低的数据, 可以使用机会网络进行传输, 同时机会网络需要得到用户的运动规律来进行更高效的数据分发, 两者结合可以相辅相成, 发挥更好的效果。

1.2 相关领域研究现状

本课题研究的内容不是一个全新的技术领域, 而是将以前有的技术和方法有机的结合起来, 本节将对本课题中要用到的技术的研究现状进行一个简单的介绍。

Wi-Fi 定位是一个很热门的研究问题, 国内外很多学者专家对其进行过研究。根据位置指纹表示的不同, 基于 Wi-Fi 定位的技术可以分为两大类。

第一类方法的特点是指纹信息来自每个 AP (Access Point, 访问接入点) 的信号平均强度表示, 然后使用确定的推理算法来估计用户的位置。在位置指纹数据库里找出与实时信号强度样本最接近的一个或多个样本, 将它们对应的采样点或多个采样点的平均作为估计的用户位置。他们利用曼哈顿距离来度量信号强度样本间的相似度, 然后以表中最匹配的样本的位置作为估计的位置。^[9]这也是一种基于庞大的数据量的室内定位的方法, 采取指纹匹配的位置确定的基本方案。其主要思想是将指纹周围有用信息, 然后建立一个指纹数据库, 然后通过映射对数据库中的指纹测量估计^[10]。研究人员一直在努力挖掘现有设备的不同的特征或

减少的映射工作。大多数的这些技术利用射频信号。使用这些技术的早期系统是雷达、Horus、改进后雷达，采用的 RSS 的位置关系的随机描述，并使用一个最大似然法来估计位置。OIL 结构有机的室内定位系统使用的 Voronoi 区域输送的不确定性和采用聚类方法识别潜在的错误的用户数据。研究表明，手机信号基站的 GSM 信号也可以用于室内定位。PlaceLab 采用无线电信标本地化在野外移动设备。Active-Campus 项目采用类似的技术，但承担的 AP 位置是已知的。有些系统，如 LANDMARC，利用 RFID 技术的室内定位。近日 Surround-Sense 基于环境功能，包括声，光，色，WiFi 等进行逻辑位置估计，利用 FM 收音机，音响背景光谱（ABS）和地理为指纹也可以进行室内定位预测。所有这些方法都需要现场勘查，建立指纹数据库。需要相当大的人工成本和工作量，而且对动态环境的不敏感也是基于指纹的方法的主要缺点^[11]。

另一种类型的本地化方法使用几何模型计算出的位置。在这些方法中，位置的计算，而不是从已知的参考数据搜索。例如，对数距离路径损耗（LDPL）模型是根据所测量的 RSS 值来估计 RF 传播距离。由于在室内环境中的不规则的信号传播增加了测量的难度，所以这些方法在实际应用中降低定位精度。为了减少费力的测量工作，并避免使用 AP 的位置，EZ 与 LPDL 模型无线传播模型的物理约束，并使用遗传算法来解决他们的定位。然而，EZ 在门口或窗户附近仍然依赖于偶然的 GPS 信息。此外，EZ 涉及复杂的计算和物理定位方案，可能会导致很多房间的检测出错。

其他的 RSS 相关模型，其他几何模型也利用用于表征信号的发射机和接收机之间的关系。这些系统包括的 PinPoint 基于到达时间（TOA），Cricket 基于到达时间（TDOA）的时间差，和 VOR 基于到达角度（AOA）。基于模型的方法通常需要安置额外的基础设施，现成的产品，或硬件配置的修改。

在轨迹异常点发现方面，也有很多研究成果，2000 年，Knoor 等人最先针对轨迹数据的异常检测做出了研究，首先，他们把轨迹信息中的点提取一部分出来表示轨迹，接着使用基于距离的异常检测的方法来检测异常的轨迹。这个方法把轨迹当做一个整体，没有考虑子轨迹，轨迹检测的基本单元是整条轨迹。如果每条轨迹的主要特征完全不一样，这个方法可以检测出这一整条轨迹都是异常轨迹。^[16]但是我们知道，一般来说估计是很长的而且有的时候只是在某一段出现了

异常如图 1.1 所示，比如说在开车过程中为了躲避一个路面上以前没有的障碍物而仅仅是在有障碍物的那一小段路上绕远，这样上面的方法就很可能检测不到这个异常。

为了能够有效地检测异常子轨迹，2008 年 Lee 等人提出了基于划分和检测的框架用于轨迹异常检测，并基于该框架设计了轨迹异常检测算法 TRAOD。^[16] 顾名思议，基于划分和检测框架的异常检测可分为两个阶段：①划分阶段：将轨迹划分成一些轨迹分段用于下阶段的检测；②检测阶段：检测异常的轨迹分段。TRAOD 的主要优点在于，检测外围子轨道的能力。它已被广泛地用在许多种应用程序。但它也有一定的缺陷。首先，在检测阶段中，TRAOD 主要采用了一些基于距离的方法来测量的 t-分区的距离。这意味着用户必须选择一个全局性的距离阈值，因此该算法检测到的外围 t-分区可以被视为全局离群值。然而，在实践中，一些有趣的数据集目前复杂而分散的特点。他们的离群值是相关的与他们的邻居的密度。这些可以被视为本地异常子轨迹。但是随着轨迹变得越来越长，线段并不能有效地表示轨迹的局部特征^[19]。

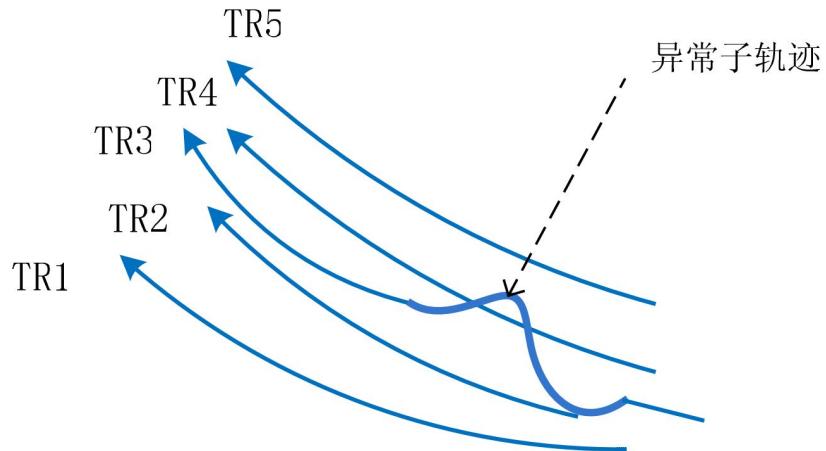


图 1.1 一小段的异常轨迹往往表示着重要信息

1.3 本文主要研究内容

本课题为综合性课题，本次毕业设计针对其中以个子课题进行研究，主要研究轨迹数据挖掘轨迹数据的异常点发现中的应用以及研究一种新的获得轨迹数据的方法——WiFi 定位。

1.4 本文的章节安排

本论文共分为五章。

第一章：绪论。将轨迹挖掘在机会网络异常发现中的应用研究这个课题分解为轨迹数据获取（Wi-Fi 定位），数据传输（机会网络）以及数据分析挖掘（轨迹异常点发现）三个部分并分别叙述了三个方面的技术背景及研究现状。最后说明全文的章节安排。

第二章：基于 Wi-Fi 的定位技术。对 Wi-Fi 定位技术进行较为详细的介绍，先是分析 Wi-Fi 信号的特征，列举影响 Wi-Fi 信号强度的因素，再对时下 Wi-Fi 定位技术进行分类说明，最后详细介绍了基于 Wi-Fi 指纹的定位方法。

第三章：异常轨迹的发现及应用。通过 Wi-Fi 定位获得轨迹信息后需要进行归集数据的处理，本章对数据挖掘中的轨迹异常点挖掘进行较为详细的描述，先解释什么是异常点，再说明异常点挖掘与数据挖掘的关系，最后列出几个主要的轨迹数据挖掘的方法。

第四章：Wi-Fi 定位的原型系统的设计与实现。介绍了一个自己制作的 Wi-Fi 定位系统原型程序，从系统结构设计到算法设计详细地阐述了该程序并且针对该系统的不足提出改进方案。

第五章：结论与展望。对本文的研究工作进行总结，指出了轨迹数据挖掘在机会网络异常发现中的更多应用并且对当前 Wi-Fi 定位技术当中的一些需要去继续研究的问题做一个简单的说明并且提出了一些可能的研究思路。

第二章 基于 Wi-Fi 的定位技术

2.1 Wi-Fi 信号特性分析

Wi-Fi 属于无线信号，有很多影响的因素，本届通过实测数据来说明影响 Wi-Fi 信号的因素并分析接收信号强度的不确定性和非线性特性。

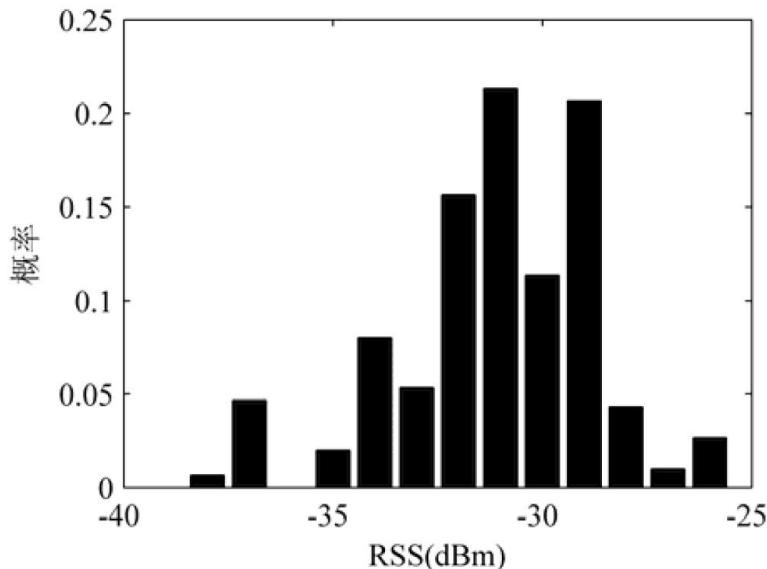


图 2.1 固定位置上某个 AP 的 RSS 样本直方图

2.1.1 不确定性分析

在室内环境下，无线电的传输环境很复杂，即使把手机固定在一个位置，同一时段测量的来自某个 AP 的 RSS 也在不断变化。如图 2.1 所示为固定位置上某个 AP 的 RSS 样本直方图；图 2.2 为在中南大学升华学生公寓 14 栋某房间测量的 Wi-Fi 信号强度；图 2.3 为在中南大学新校区图书馆的同一个地点 60 秒内测到的不同信号的强度，我们可以看到，由于多径、无线电干扰等影响主要起信号衰减作用，Wi-Fi 信号幅度波动比较大而且有的时候还时有时无。尽管如此，大多数情况下仍然可以将 RSS 的时间分布看作是近似高斯分布的。RSS 的不确定性，即时改变的特性主要由以下几个方面的因素产生：

1. 多径效应。

无线电波在空间中传播时会像音波一样反射折射，会从很多不同路径到达目的地，这种现象称之为多径效应。每条路径的长短不同，所经过的介质也有可能不同，所以到达的时间以及强度都不一样。传播路径还有可能随着

时间而变化，接收方有任何微小的移动都有可能造成接收信号强度的大幅度波动。

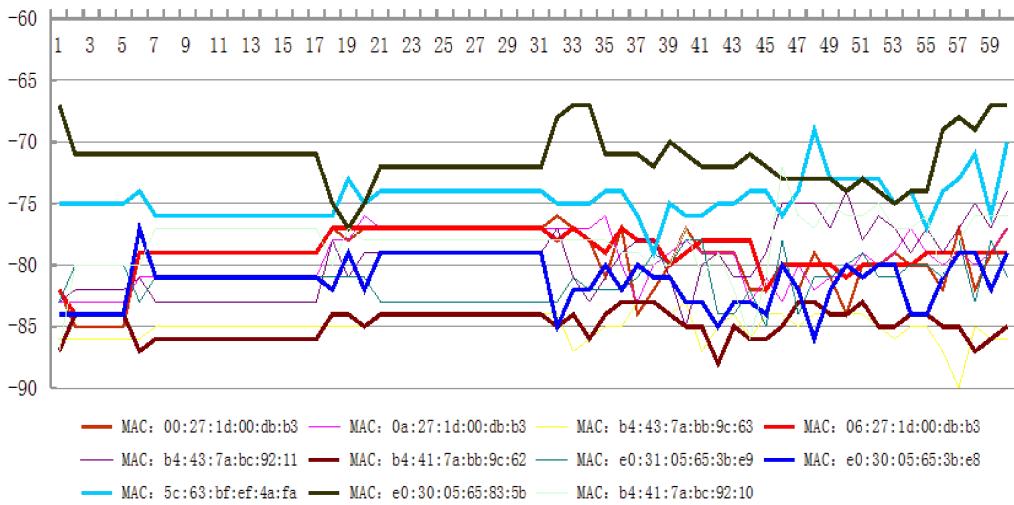


图 2.2 Wi-Fi 信号在 60 秒内的变化（同一地点）

2. 同频无线电干扰

现在人们的无线设备越来越多，又因为 Wi-Fi 所在的 2.4GHz 频段是全球免费使用的波段，很多如蓝牙、ZigBee、微波炉、无线键盘鼠标等无线设备都工作在这个频段，导致 Wi-Fi 的无线信号极易受到其它同频无线电的干扰。

macid	pointid	level	maxlevel	minlevel	times	d
1	tsg34	-890	-89	-89	7	106
2	tsg34	-890	-89	-89	6	108
3	tsg34	-860	-86	-86	2	189
4	tsg34	-870	-87	-87	7	149
5	tsg34	-832	-77	-87	58	19
6	tsg34	-861	-85	-87	9	164
7	tsg34	-890	-89	-89	7	106
8	tsg34	-932	-91	-95	21	31
9	tsg34	-880	-88	-88	6	129
10	tsg34	-854	-85	-87	60	0
11	tsg34	-907	-90	-91	20	57
12	tsg34	-890	-88	-90	14	93
13	tsg34	-850	-81	-87	24	137
14	tsg34	-880	-88	-88	13	112
15	tsg34	-869	-86	-88	20	114
16	tsg34	-870	-87	-87	6	152
17	tsg34	-924	-90	-93	15	43

图 2.3 Wi-Fi 信号的出现次数不稳定

3. 人体吸收

一般情况下是室内有人员走动的。构成人的 70% 是水，而对 Wi-Fi 信号的衰减是很大的干扰。无论是在学习阶段建立指纹数据库的时候，还是在定

位阶段用户使用的时候，人体对于信号强度的影响都是很大的。

4. 室内环境

室内环境复杂多变，有的时候开门或者关门，开窗或者关窗都会对 Wi-Fi 信号强度造成影响。还有室内的墙有的是承重墙，由钢筋混凝土构成，对于 Wi-Fi 信号的阻隔作用就很大，有的非承重墙就对于信号的阻隔作用比较小。室内环境的温度以及湿度也都会对 Wi-Fi 信号强度造成影响。

5. 天线方向性

天线是决定接收 Wi-Fi 信号强度额的重要因素，一般不能保证每次使用手机时 Wi-Fi 天线的方向是固定的。经过测试发现，改变手机的方向或者天线的方向，可以改变大约 6dB 的接收信号强度。

6. Wi-Fi 接收设备的差异

不同的 Wi-Fi 接收设备的接收信号的能力又不同，这将会导致指纹信息不能通用，这又会增加采集指纹信息的工作量。图 2.4 展示了两个不同的手机之间的 Wi-Fi 接收能力的差异。如何减少或消除这些不稳定因素对于 Wi-Fi 定位精度的影响也是一个问题。

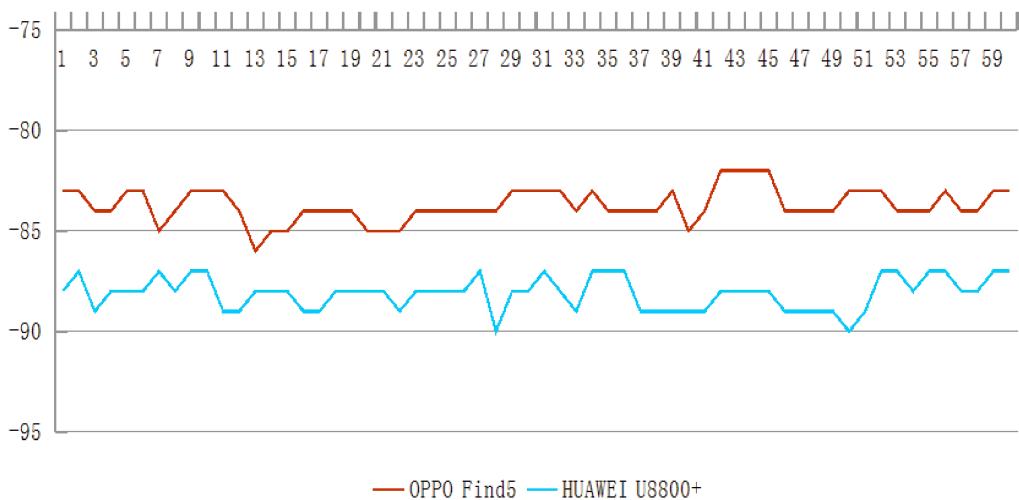


图 2.4 不同手机对于同一个信号的接收能力

一般情况下，我们无法人为地去控制上述造成信号不稳定的因素，这些影响因素是由复杂的室内无线电传播环境造成的。在设计基于 Wi-Fi 指纹的定位系统中，为了减少上述不确定性因素带来的影响，可以采取以下方法：首先在设计系统的时候要充分考虑这些不确定因，并加以描述。例如本次毕业设计实现的程序中就在一个地点测试了很多次，取平均值还有记录下来出现次数一边对信号

的波动进行分析，在 AP 选择环节，去除波动较大不稳定的 AP。然后，在定位阶段，通过前面手机的信息来确定稳定与不稳定的节点，并尽量通过未定的节点来估算位置。最后，采用带有纠错功能的方法将物理位置与指纹映射起来。

2.1.2 非线性特性分析

RSS 的非线性表现在 RSS 与物理位置呈现为复杂的非线性映射关系，如上节所述，即使在非常简单的室内环境下信号强度也不是完全线性的，也会根据很多的外界因素而发生改变。更大的非线性变化发生在有明显物理界限，如两个房间隔着一面墙，这样在房间内 Wi-Fi 信号还是基本成一个线性关系，但是由于两个房间中间隔了一面墙，这样两个房间的 Wi-Fi 信号就会有明显不同，如图 2.5 所示，我们可以利用这个差别来区分不同的房间，在本文中就将介绍一种利用这个特征实现虚拟房间划分的自动识别房间的 Wi-Fi 定位方法。

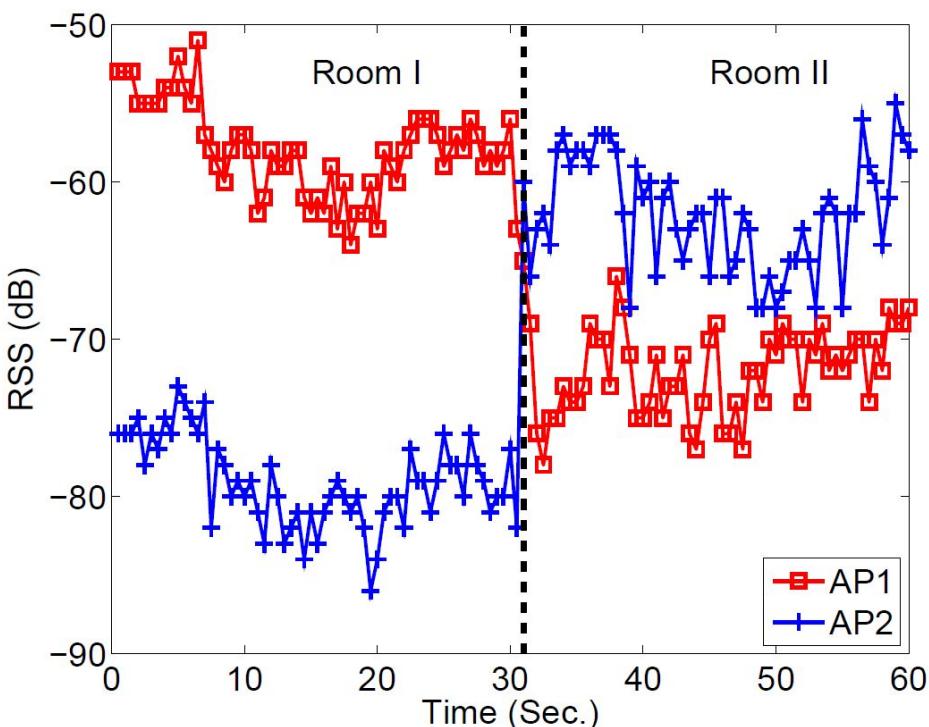


图 2.5 当 Wi-Fi 信号穿过墙壁后会发生明显的变化

2.3 基于 Wi-Fi 指纹定位的理论基础及一般过程

Wi-Fi 指纹定位技术主要依赖于 RSS 的物理位置关联特性，换句话说，就是每个地点上都有不一样的 RSS 信息，就是指纹信息。这些模式通过学习阶段在参考点上的 RSS 指纹采集来描述，并构建指纹数据库，即所谓的 Wi-Fi 指纹数据库。在该数据库中通过机器学习的方法发现不同指纹之间的关系，找出指纹信

息和实际地理位置之间的关系，例如使用 K-Means 聚类方法就可以将数据库中的不同指纹信息聚成一个一个的虚拟房间来分别对应物理位置。在定位阶段时就可以提取这些特征，根据用户提供的指纹信息来估算用户的位置。

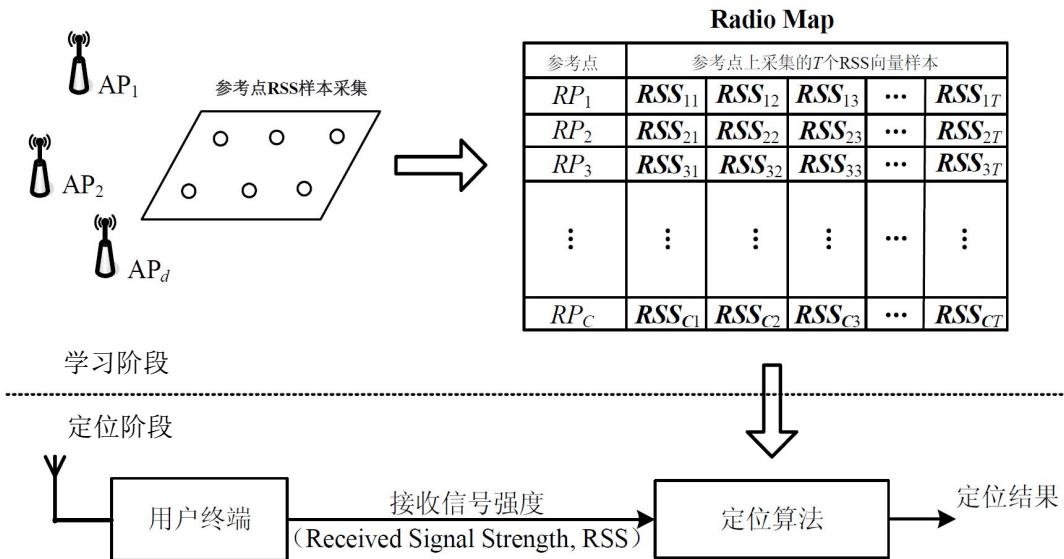


图 2.6 基于 Wi-Fi 指纹定位系统定位流程示意图

基于 Wi-Fi 指纹的定位技术一般可以分为两个阶段：第一阶段为学习阶段，首先需要在目标地区均匀的选取参考点，并采集到 Wi-Fi 指纹以及其他信息上传到数据库，形成这个标志点的特定的指纹信息，如图 2.7 所示就是学习模式的流程示意图。在得到这些指纹数据之后就可以知道指定区域的 Wi-Fi 分布数据，如图 2.8 所示，每个区域都有特定的 Wi-Fi 指纹数据。通过机器学习的方法就可以得到这些轨迹信息与物理位置之间的对应关系。



图 2.7 学习模式示意图

第二阶段是定位阶段，由于定位阶段需要与服务器进行数据通信，所以也称

之为在线定位阶段。首先用户通过他们的客户端采集到他们身边的 Wi-Fi 信息，然后将这些信息转换为 RSS 向量样本上传到定位服务器，与数据库中的指纹数据库进行匹配得到定位结果。指纹定位技术的流程图如图 2.6 所示，在限定为阶段的流程示意图如图 2.9 所示。

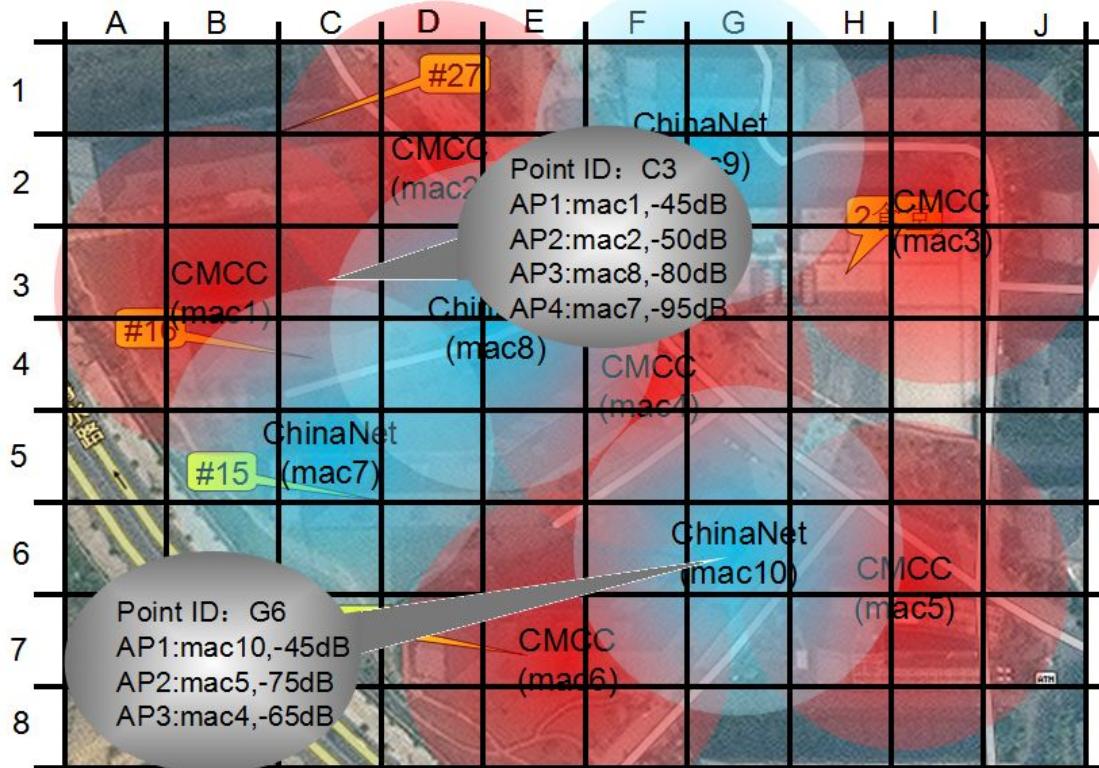


图 2.8 每个地区可以用不同的 Wi-Fi 信号强度来进行划分



图 2.9 定位模式示意图

2.4 本章小结

本章从分析 Wi-Fi 信号特征开始，逐步介绍影响 Wi-Fi 信号的因素、Wi-Fi 定位技术的分类以及详细介绍了基于 Wi-Fi 指纹的定位方法，本文第四章设计并实现了一个基于 Wi-Fi 指纹的定位系统原型，并且会在第四章用实测数据继续对 Wi-Fi 定位技术展开讨论。

第三章 异常轨迹的发现及应用

3.1 异常点的定义

异常点定义有多种，但具有代表性的是 V. Barnett 在统计学研究领域中给出的定义：

一个异常点是这样的数据点，基于某种度量而言，该数据点与数据集中的其他数据有着显著的不同。^[20]

除这个定义以外，还有很多不同的研究者根据他们各自的学术背景对于异常点给出了不同的的定义，尽管它们各不相同，但都反映了异常点的特点：首先，异常点通常是看上去就是使人惊讶的，这是异常点的关键特征之一；其次异常点具有相对性，如果对于一般状态的定义不同，就会得到不同的异常点；最后异常点有较强主观性，几乎所有研究者进行异常点挖掘研究时都定义特有的挖掘背景。

3.2 异常点与数据挖掘

数据挖掘在现在信息爆炸的时代显得尤为重要，在机器学习、数据库技术以及人工智能等方面都起着重要作用。作为数据库技术的研究开发和应用最活跃的分支之一，数据挖掘兴起至今，已经成为许多国际会议关注的焦点。数据挖掘可以处理多种多样的数据，如关系数据库中的结构化数据，文本、图形、图像等非结构化数据甚至是 Web 信息、生物信息等非结构化异构数据。目前数据挖掘的主要研究内容包括挖掘算法研究、基础理论研究、可视化技术、数据仓库、挖掘知识的表示方法、挖掘知识维护，更新和利用、半结构化和非结构化数据挖掘及 WEB 数据挖掘等。

通常，数据挖掘技术以其功能和发现的模式被分为 4 类：依赖性检测、类型识别、类型描述、异常点检测。许多数据挖掘研究例如关联规则挖掘、分类和聚类等都属于前 3 类，他们研究的是数据中的大多数数据对象从而发现规律等信息，而第 4 类由于只占数据集中的较少部分，通常被看作聚类过程的副产品，当作噪声处理，所以在数据挖掘研究领域，异常数据挖掘当初并不是研究主流。但在一些应用中，如保险欺诈、违规交易、信用卡欺诈等检验中人们对其重要性

认识的加深，异常数据挖掘日益受到重视，罕见的事件可能比正常出现的事件更有意义。对异常点的查找过程称为异常数据挖掘，它是数据挖掘技术中的一种。

3.3 异常点挖掘的主要方法

对传统的异常检测方法，国内外都有大量的文献研究。最早的异常点概念是由 Hawkins 提出的，即异常点是在数据集中与众不同的数据，使人怀疑这些数据并非随即偏差，而是产生于完全不同的机制。^[20]换句话说，这些所谓的异常数据代表着一定的特殊意义。常用的检测方法主要有四类：基于统计的方法、基于距离的方法、基于密度的方法、基于深度的方法和基于偏移的方法，如图 3.1 所示。

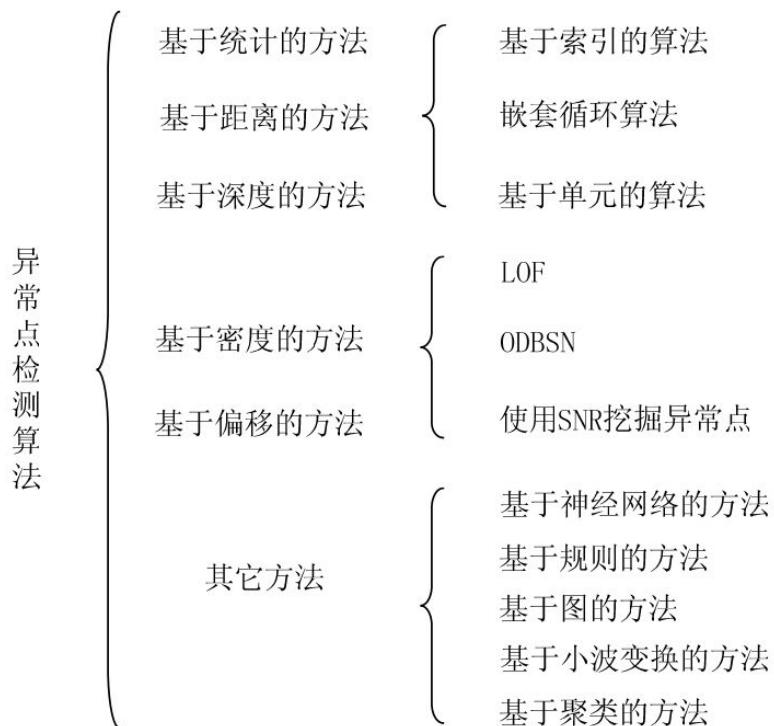


图 3.1 异常点检测算法分类图

3.4 轨迹异常点挖掘的应用

一个人的轨迹信息在一定程度上是有规律可循的，比如一个上班族在工作日都是早上上班晚上下班，这种正常的轨迹信息被定位系统记录下来并保存到数据库中作为参照数据，一旦某一天系统检测到在没有任何特殊情况下轨迹信息发生了变化，可能是用户遭受了一些意外，比如手机被偷或者是本人被劫持，在这种情况下，系统可以自动地发出警报给相关部门，是一种对于用户的人身及财产安全的保障。

另例如中国老龄人口逐渐增多，更多的多老年人需要不断地监测和照顾，但是有些时候不能做到有人时时刻刻陪护在老人身边，因此当老人外出活动的时候有一个随身的定位系统是很有必要的，系统通过人为设定或者是平常的数据收集找到老人的一般行动规律。假如有一天老人上错了公交车，系统可以检测到老人的行动路线出现异常，这样他的家人可以得到通知，这样就防止了一起老人迷路事件的发生，既保障了老人的人身安全也减轻了家庭的担心。图 3.2 描述了一个典型的案例，蓝色路线是平常走的路线，红色路线是一条异常数据，系统会自动将这个异常通知老人的家人。

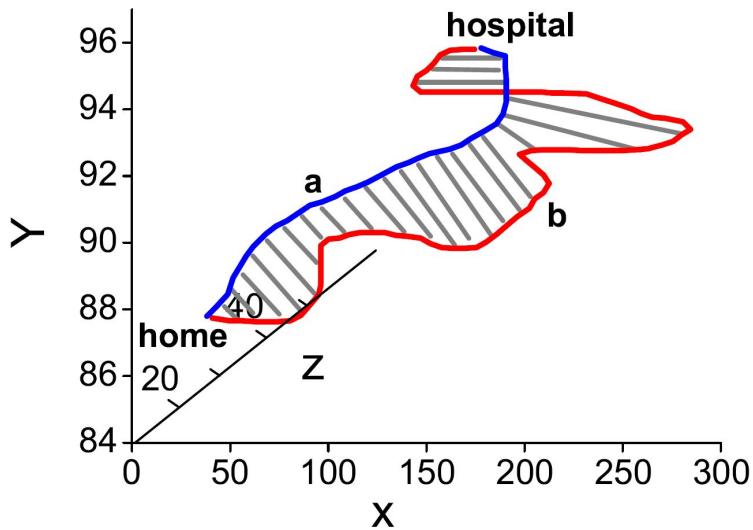


图 3.2 正常轨迹与异常轨迹

在本次毕业设计中，我主要设计了一个基于 Wi-Fi 指纹的定位系统，而 Wi-Fi 定位主要用于室内环境下的定位，在室内环境下有本系统得出的轨迹结合异常点挖掘会有更大的用处。比如在公司部署了基于 Wi-Fi 的定位系统，可以通过检测员工手机来确定员工的位置从而实现智能打卡，工作情况记录等活动，方便了企业管理层对于员工的管理。

3.5 本章小结

本章首先介解释了什么是异常点并写出了异常点的定义，接着对异常点挖掘的发展作了较为详尽的叙述，最后介绍了异常点挖掘方法以及针对于轨迹数据异常点挖掘的方法。最后举了几个例子来说明轨迹异常点挖掘在现实生活中的作用。

第四章 Wi-Fi 定位的原型系统的设计与实现

4.1 系统设计

4.1.1 架构设计

本系统有两个版本，第一个版本采集数据存储数据与数据处理都在手机上完成，这个版本为了验证定位算法，在实际应用中，我们不能让用户在使用这个定位系统之前还要花费很多时间去采集各种样本数据，这样不利于用户使用也不利于软件的推广。所以本系统采取的是客户端服务器端分离的方式，手机作为客户端只是用来采集 Wi-Fi 指纹信息，采集后将数据通过手机网络发送到服务器，服务器端处理这些数据并返回位置信息，客户端接收位置信息并显示给用户。其中客户端与服务器通讯采取的是最普遍的 HTTP 通讯，服务器为普通的 WEB 服务器，这样可以增加系统的通用性也可以在未来根据业务需要自由的扩展不同的功能。整个系统的结构如图 4.1 所示。

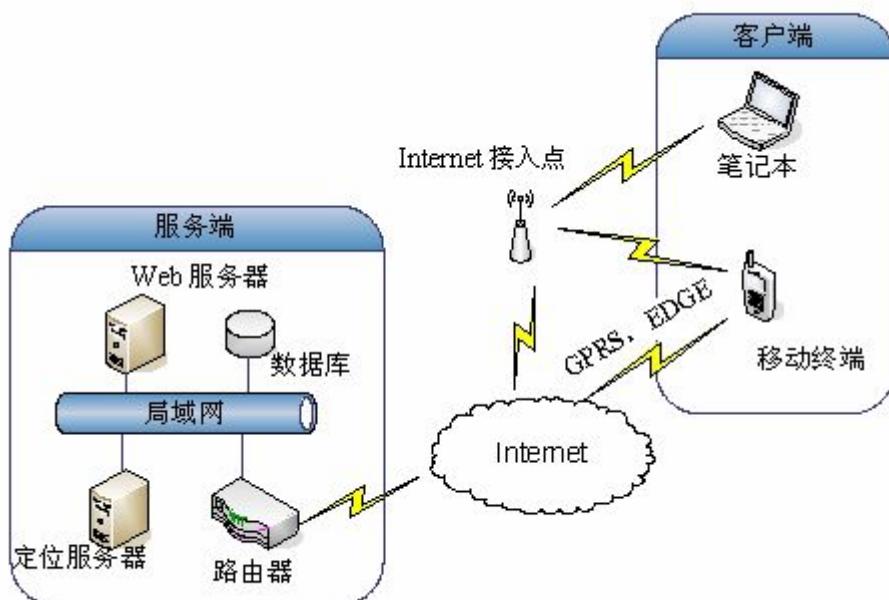


图 4.1 定位系统网络结构

图 4.2 为本定位系统的信息交互流程图。移动端向定位服务期提出定位请求，并且将需要的数据通过 HTTP 方式上传，由于数据量比较大可能会超过 POST 的上限，所以我们采用 GET 的方式传输数据。定位服务器接收到数据后对数据库中信息进行比对并进行定位计算，最终将位置信息以相同的方式发送回客户

端。

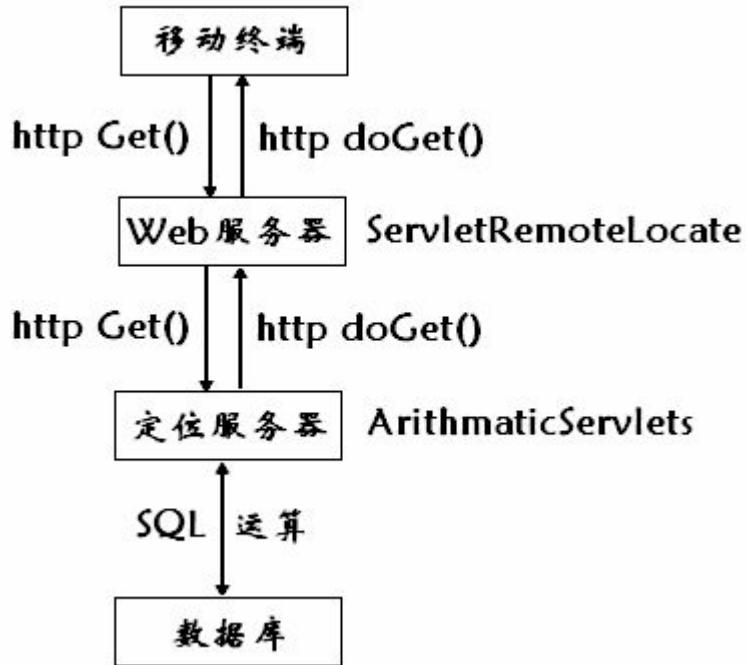


图 4.2 移动终端与服务器间的信息交互

4.1.2 数据库设计

本系统数据库一共三张表，分别为记录 Wi-Fi 物理地号址与编的对应关系的 macIDs 表，记录所有地点的 allPoints 表和记录所有 Wi-Fi 指纹信息的 allMacs 表。

1. macIDs 表——设备物理地址与编号对应表。

本 Wi-Fi 定位系统采用的是基于 Wi-Fi 指纹的定位技术，需要存储大量指纹信息，而区分 Wi-Fi 使用的是设备的物理地址，为了方便比较，不同的地址对应为编号，如图 4.3 所示。这些对应关系存储在 macIDs 表（物理地址与编号对应表）中。

macid MAC编号	macaddr MAC地址
511	c0:61:18:12:f0:4c
512	2e:c8:97:9a:06:20
513	fc:c8:97:9a:06:20
514	c8:3a:35:1a:00:c8
515	20:dc:e6:60:d1:7c

图 4.3 部分 MAC 地址与编号的对应关系

2. allPoints 表——指纹数据与地点对应表。

本 Wi-Fi 定位系统只是一个简单的原型，尚不能自动根据 Wi-Fi 自动的划分每个指纹信息对应的地点，所以需要在学习阶段进行标记，本文将会提到一种采用了聚类方法自动进行划分的方法，将会在本章靠后的小结中讨论。

pointid 地点ID	x	y	pointname 地点名称	macnum 搜索到的AP数
sj-304m	-1	-1	三教304	47
sj-304b	-1	-1	三教304	51
sj-303f	-1	-1	三教303	59
sj-303b	-1	-1	三教303	68
sj-305m	-1	-1	三教305	65
sj-305f	-1	-1	三教305	74
sj-301f	-1	-1	三教301	68
sj-301b	-1	-1	三教301	57
sj-302f	-1	-1	三教302	60
sj-302b	-1	-1	三教302	67

图 4.4 指纹与地点对应表

图 4.4 展示了该表中的部分数据，pointid 为地点 ID，是每一条采集的指纹数据集的唯一标识，xy 是该数据集在地图上所对应的 XY 坐标，但是本系统还没有涉及到地图对应部分，所以在这里统一填-1.pointname 为真实地点名称，在未来的聚类算法中这个键值作为不同的聚类中心，macnum 表示本指纹数据集所包含的 Wi-Fi 指纹数据数量。

macid	pointid	level	maxlevel	minlevel	times	d
444	sj-302b	-707	-68	-72	10	2
428	sj-302b	-780	-74	-80	10	7
429	sj-302b	-894	-88	-90	10	0
494	sj-302b	-900	-90	-90	5	0
44	sj-302b	-795	-79	-80	10	0
477	sj-302b	-860	-85	-87	8	1
431	sj-302b	-703	-68	-72	10	3
432	sj-302b	-700	-68	-72	10	2
539	sj-302b	-900	-90	-90	9	0

图 4.5 指纹数据表

3. allMacs 表——指纹数据表。

这个表里面保存着所有的指纹数据，如图 4.5 所示。如上文所述 Wi-Fi 信号强度很不稳定，为了获得一个地点的比较准确的 Wi-Fi 指纹信息，本系统在同一个位置采集十次，将每次出现的值相加再除以出现次数作为强度值，为了保证精度，数据库中存储的强度值乘以 10 来存储在 level 键值中。maxlevel 和 minlevel 分别表示最大以及最小强度，times 表示出现次数，d 表示这些数据的方差。此表中的数据直接作用于定位程序，也是以后数据挖掘的数据源。

4.2 系统实现

4.2.1 服务器端搭建平台

本系统的主要定位算法以及数据可操作都是在服务器端完成，为了方便测试，我首先设计了一个只在手机端运行的程序，但是为了方便以后的后续开发和加强可扩展性，便把核心算法都搬到了网络服务器上，由于宿舍条件有限，不能在宿舍搭建能被广域网访问的服务器，于是我便将服务器程序部署在了 SAE(Sina App Engine，新浪应用引擎)上。

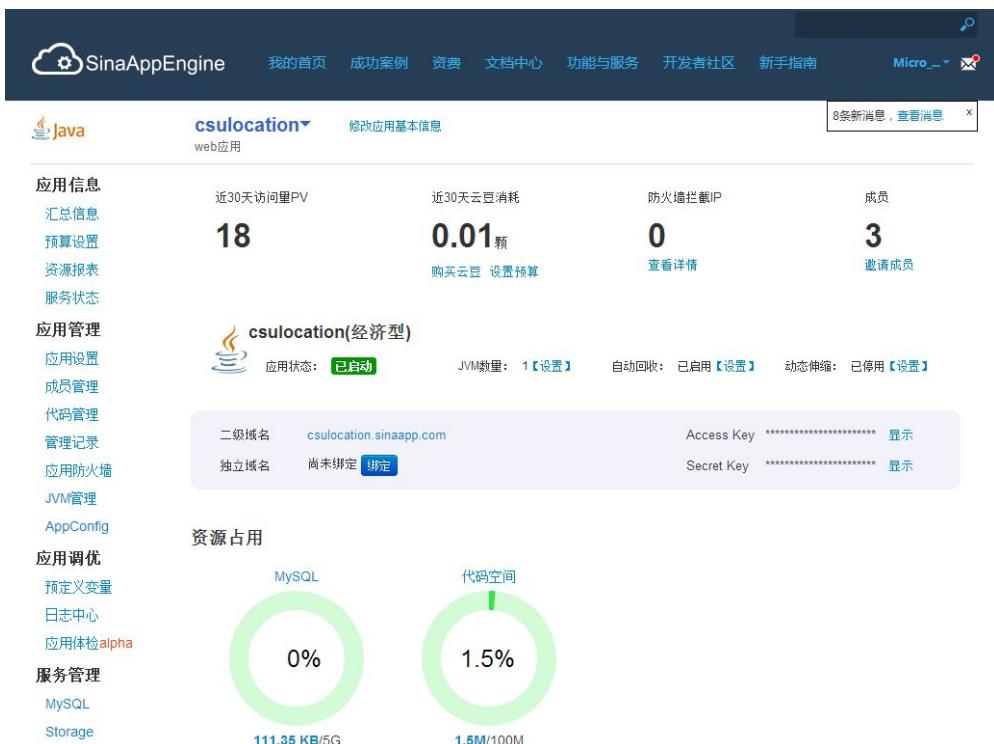


图 4.6 SAE 控制界面截图

SAE 是新浪研发中心于 2009 年 8 月开始内部开发，并在 2009 年 11 月 3 日

正式推出第一个 Alpha 版本的国内首个公有 App Engine，SAE 是新浪云计算战略的核心组成部分。而且 SAE 对于非商业开发者来说几乎完全免费，很适合像我这样的学生进行使用。

为了满足打折在 SAE 上面的需要，本系统使用的是传统 WEB 服务器（地址为 <http://csulocation.sinaapp.com/>）但是不需要前台界面，只需要后台的数据交换就可以。本程序的服务期为 JSP 服务器，使用 JAVA 作为开发语言，JAVA 运行在虚拟机上，在不同机器上运行的时候不需要修改源代码，方便进行移植，同时也方便在不同的机器上进行部署。为了不让网页显得太单调，本人将第一个单机版运行的室内定位程序上传到网上供大家下载，同时将后台服务器的前端界面修改为了程序的使用说明。图 4.6 为 SAE 管理界面的截图。

4.2.2 Android 系统编程简介

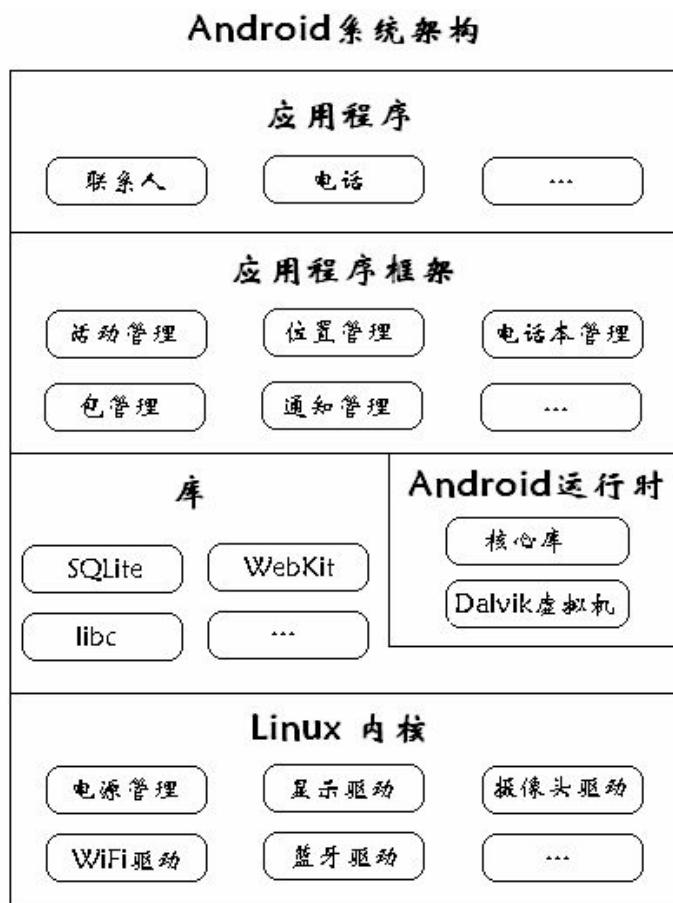


图 4.7. Android 系统架构图

本系统客户端采用 Android（安卓）系统手机。Android 系统是 Google 在 2007 年发布的基于 Linux 平台的开源手机操作系统。由于安卓系统十分自由开

放，而且使用的编程语言为时下非常流行的 JAVA 语言，现在已经有非常多的开发者为其开发应用。Android 系统架构见图 4.7，他的系统内核是 Linux，内核中包含了各种设备驱动和管理模块，安卓囊括了非常齐全的类库和框架，包括轻量级数据库 SQLite、浏览器 Webkit 等，在本系统第一个单机版的测试程序中就是使用的安卓自带的 SQLite 数据库实现数据存储的。安卓系统建立在 Dalvik 虚拟机上，使用 JAVA 作为开发语言。安卓开放了很全面的 API 接口供开发人员使用，整个平台具有良好的开放性和扩展性。

在安卓操作系统中，程序大多数是运行在一个一个的 Activity 中，不同的界面之间的切换就是不同的 Activity 之间的切换，由于手机的硬件运行环境限制，无法让所有的 Activity 同时运行，所以每个 Activity 都有生命周期（图 4.8）。

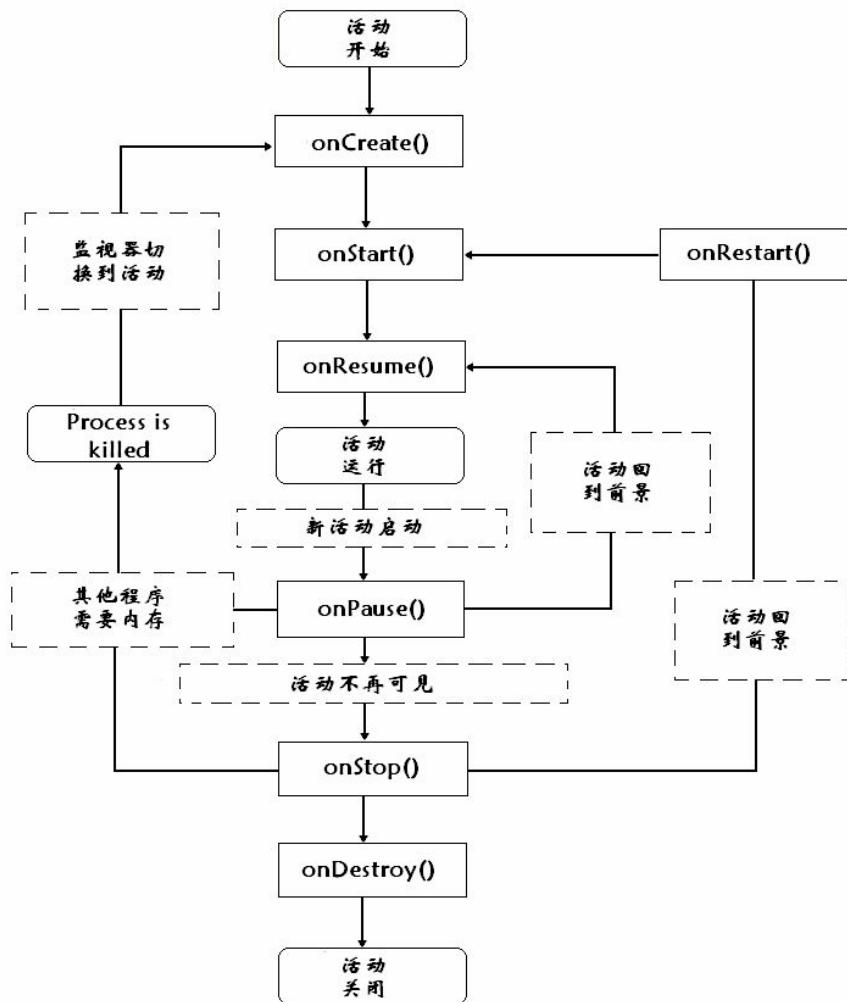


图 4.8 Activity 生命周期

为了实现想要达到的功能必须要了解这个生命周期并且在正确的时间干正确的事。比如在本项目中需要使用到 Wi-Fi 模块，如果在进行 Wi-Fi 扫描的时候

Wi-Fi 模块并没有打开，整个程序就会报错崩溃。所以要在使用 Wi-Fi 扫描功能的 Activity 运行之前先检测 Wi-Fi 模块的状态，如果是关闭的，就需要自动打开 Wi-Fi 模块，这样才能让程序正常工作。

4.2.2 训练客户端与定位客户端实现

现在 Android 的最新版本更新到了 4.4 但是为了保证兼容性，本程序的开发平台为 4.0 版本。本系统有两个客户端，一个是用于 Wi-Fi 定位的第一个阶段，就是数据采集阶段的程序，成为学习模式客户端，他的任务是采集手机周围的 Wi-Fi 信号并且输入地点 ID 和地点名称，再将这些指纹数据上传到服务器。第二个客户端为用户在使用的时候用到的，成为定位模式客户端，他的任务是收集用户周围的 Wi-Fi 信息并且上传到服务器发出定位请求，并等待服务器返回位置信息。

1. 学习模式客户端

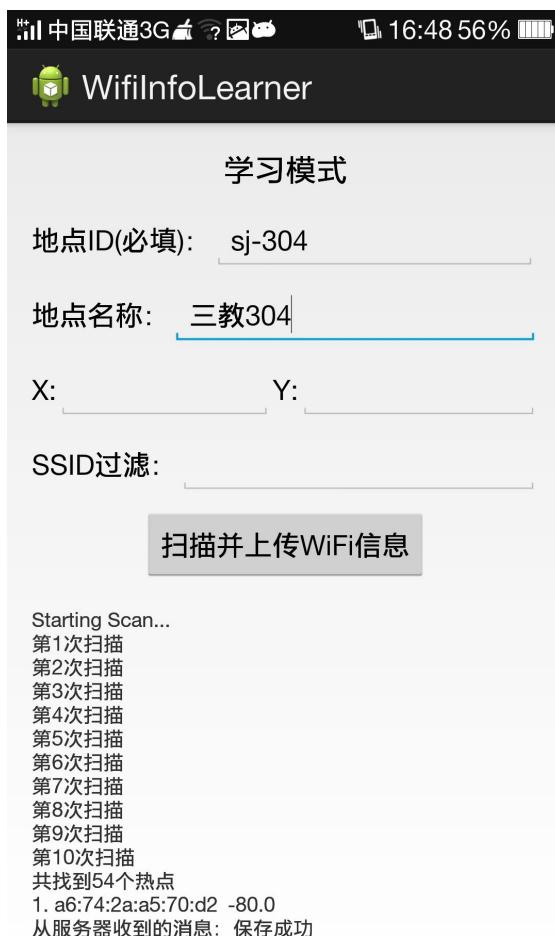


图 4.9 学习模式程序截图

学习模式客户端程序是由工作人员事先进行实地采样时使用的程序，学

习模式程序截图如图 4.9 所示。

在学习模式中，为了降低 Wi-Fi 信号多径效应等外界因素对于 Wi-Fi 信号的影响，需呀连续采集多次获得较稳定数据后再存入数据库。另外由于现在很多人都有了 Wi-Fi 路由器等可以发射 Wi-Fi 信号的设备，这些个人 Wi-Fi 信号发射设备并不是持久存在的，可能会在任意社科内被关闭或者打开，如果在学习阶段检测到这些不稳定的 Wi-Fi 信息应该把它忽略掉，换句话说，在学习阶段应该尽量采集那些比较稳定的 Wi-Fi 热点，例如中国移动架设的 Wi-Fi 网络 ChinaNet，在本程序中我们可以在 SSID 过滤一栏中填入“ChinaNet”，这样搜索到的信号只会有 SSID 中包含 ChinaNet 字符串的 Wi-Fi 热点，这样可以提高学习数据的可靠性。

2. 定位模式客户端

当用户需要查询自己所在位置的时候需要使用定位模式的客户端，首先获取周围的 Wi-Fi 指纹信息，再将这些信息上传到定位服务器，最后接收来自服务器的位置信息。图 4.10 为定位模式程序截图。



图 4.10 定位模式程序截图

在定位模式中，为了缩短定位时间，本程序只采集 3 次周围的 Wi-Fi

信号就去服务器中进行比对，算上数据在网络上传输的延迟，平均定位时间是 1 秒，速度还是相当快的。

3. Wi-Fi 扫描方法及代码实现

上述两个客户端都使用了 Wi-Fi 扫描的方法，在 Android 编程中，我们可以通过操作 WifiManager 对象来获得 Wi-Fi 信息，扫描 Wi-Fi 及上传服务器的关键代码实现如下所示（以学习模式为例）：

```
/*
 * 通过扫描获得当前点WiFi信息，并将详细信息存储如数据库中
 *
 * @param timesValue
 *      扫描次数
 * @return 该点WiFi信息的MAP
 */
private void getWiFiInfo(int timesValue) {
    int count = 0;
    // 存储MAC地址以及扫描的几次的信号强度的队列
    HashMap<String, List<Integer>> tempMap = new HashMap<String, List<Integer>>();
    for (int i = 0; i < 3; i++) {
        wifiManager.startScan();
        try {
            Thread.sleep(500);
        } catch (InterruptedException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }
        wifiList = wifiManager.getScanResults();
    }
    while (count < timesValue) {
        // 获得WiFi扫描结果
        wifiManager.startScan();
        try {
            Thread.sleep(1000);
        } catch (InterruptedException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }
        wifiList = wifiManager.getScanResults();
        count++;
        if (need[0] != "") {
            wifiList = chooseBySSID(need);
        }
        MainActivity.info = "第" + count + "次扫描\n";
        System.out.println(MainActivity.info);
        handler.sendMessage(0x123);
        for (int i = 0; i < wifiList.size(); i++) {
            String tempMac = wifiList.get(i).BSSID;
            int tempLevel = wifiList.get(i).level;
            if (tempLevel > -35 || tempLevel < -95) {
                continue;
            }
            // 将信息存入临时map
            if (!tempMap.containsKey(tempMac)) {
                tempMap.put(tempMac, new ArrayList<Integer>());
            }
            tempMap.get(tempMac).add(tempLevel);
        }
    }
}
```

```

    }
    // 如果存在，则把信号强度相加
    if (tempMap.containsKey(tempMac)) {
        tempMap.get(tempMac).add(tempLevel);
    } else {
        ArrayList<Integer> levels = new ArrayList<Integer>();
        levels.add(tempLevel);
        // 如果不存在，添加新AP
        tempMap.put(tempMac, levels);
    }
}
// 将WiFi信息添加到结果MAP中
Set<String> key = tempMap.keySet();
Iterator<String> itr = key.iterator();
while (itr.hasNext()) {
    // 取得MAC地址
    String BSSID = itr.next();
    // 分析信息，获得最大强度，最小强度，参考强度信息P
    System.out.println("MAC: " + BSSID);
    HashMap<String, Integer> useInfo
    getUseInfo(tempMap.get(BSSID));
    int minlevel = useInfo.get("min");// 最小信号强度
    int maxlevel = useInfo.get("max");// 最大信号强度
    int midlevel = useInfo.get("mid");// 平均信号强度
    int time = tempMap.get(BSSID).size();// 搜索到的次数
    int dx = getDX(tempMap.get(BSSID), midlevel);
    System.out.println("min:" + minlevel + "\tmax:" + maxlevel
        + "\tmid:" + midlevel / 10.0 + "\ttime:" + time + "\td:"
        + dx);
    AP ap = new AP(BSSID, midlevel);
    ap.setMaxlevel(maxlevel);
    ap.setMinlevel(minlevel);
    ap.setTimes(time);
    ap.setD(dx);
    tempPoint.getAps().add(ap);
}
tempPoint.setMacNum(tempPoint.getAps().size());
tempMap.clear(); // 清除临时图
System.out.println("保存该点");
// 显示
MainActivity.info += ("共找到" + tempPoint.getAps().size() + "个热
点\n");
handler.sendMessage(0x123);
for (int i = 0; i < tempPoint.getAps().size(); i++) {
    MainActivity.info += ((i + 1) + ". "
        + tempPoint.getAps().get(i).getBSSID() + "\t"
        + tempPoint.getAps().get(i).getLevel() / 10.0 + "\n");
}
MainActivity.info += "扫描结束，记录的节点信息: \n";
handler.sendMessage(0x123);
}

```

4.2.3 服务器端实现——定位算法

本系统服务器端主要有两个功能，一是存储指纹数据，二是进行定位计算，存储数据就是用的 MySQL 进行实现，按照 4.1.2 小节中的数据库模型进行设计并编码创建。本节中会对定位算法进行比较详细的介绍。

1. 邻近法描述

关于定位算法，本系统使用的是邻近法。邻近法的作用就是对比计算用户提供的 Wi-Fi 指纹与数据库中指纹的欧氏距离，找出距离最小的一个或者若干个指纹，采用平均或者是加权的方法得到最终的位置信息。

最基本的近邻法是最近邻算法(Nearest Neighborhood, NN)。该算法计算实时测量的 RSS 样本向量与各个指纹对应的 RSS 均值向量之间的欧氏距离：

$$d_i = \sqrt{\sum_{j=1}^d (\overline{RSS}_i^j - RSS^j)^2} \quad (4-1)$$

其中 \overline{RSS}_i^j 是在第 $i(i=1,2,\dots,C)$ 个参考点上来自于第 j 个 AP 的 RSS 均值， RSS^j 是在线阶段实时测量得到的第 j 个 AP 的指纹， C 是全部的参考点个数。NN 方法直接将距离最小的点当做是用户的所在位置，这样容易收到 Wi-Fi 信号波动性的影响而得到偏差较大的结果。但是这种方法的优点是易于实现而且比较速度很快，在本次毕业设计中的定位程序一开始就是用的这种方法来测试 Wi-Fi 定位的可行性以及进行初步的数据分析的。

K 近邻算法(Nearest Neighborhood, KNN)是 NN 法的改进型算法，本方法选取 $K(K \geq 2)$ 个与用户提供的指纹欧氏距离最小的 Wi-Fi 指纹，然后再对指纹对应的位置信息取平均求得最终的位置信息：

$$(\hat{x}, \hat{y}) = \frac{1}{K} \sum_{i=1}^K (x_i, y_i) \quad (4-2)$$

其中， (x_i, y_i) 是第 i 个欧氏距离最近的 Wi-Fi 指纹的位置坐标， (\hat{x}, \hat{y}) 是最终的定位结果。

再进一步，由于考虑到几个邻居与用户位置不同不能简单的取平均来确定最终的定位位置，所以提出了加权近邻算法(Weighted Nearest Neighborhood, WKNN)。WKNN 法与 KNN 法的不同之处在于，计算得出 $K(K \geq 2)$ 个最近邻参考点后，并不是直接对几个参考点的位置坐标取平均数，而是根据他们的距离计算

出一个权值，并且根据权值的不同来算出一个加权后的平均位置作为最终位置：

$$(\hat{x}, \hat{y}) = \sum_{i=1}^K \left(\frac{\hat{\eta}}{d_i + \varepsilon} \cdot (x_i, y_i) \right) \quad (4-3)$$

其中， d_i 是用户提供的指纹样本向量与第 i 个邻居参考点之间的 Wi-Fi 指纹的欧氏距离， $\hat{\eta}$ 为加权系数， ε 是很小的正常数，以防止分母出现为零的情况。

与用户提供的指纹欧氏距离更近的标志点的权值更高，这样可以在一定程度上提高定位的准确度。

由于本系统中没有对地图进行匹配，就没有每个点的 XY 坐标，不好对指纹数据进行 KNN 比较，但是 NN 精度又太低，所以在本系统中，我使用的是加权临近算法中的 $K=1$ 的情况，具体的加权情况被我用“靠谱度”来表示，计算方法是综合考虑一个 AP 热点在训练阶段时的出现次数以及数据的方差来计算的。具体代码将会在下一节中展示。

2. 本项目中的 NN 邻近法代码实现

```
/*
 * 找出权值最大的几个点并且找到这些点的具体信息
 *
 * @param pointMap
 * @return
 */
private ArrayList<Point> findLikelyPoints(HashMap<String, Integer>
pointMap) {
    ArrayList<String> resultIDs = new ArrayList<String>();
    // 排序后的列表
    List<Map.Entry<String, Integer>> infoIDs = new
ArrayList<Map.Entry<String, Integer>>(
        pointMap.entrySet());
    Collections.sort(infoIDs, new Comparator<Map.Entry<String,
Integer>>() {
        public int compare(Map.Entry<String, Integer> o1,
                           Map.Entry<String, Integer> o2) {
            // 根据Value排序
            return (o2.getValue() - o1.getValue());
        }
    });
    int nums = 0;
    System.out.println("infoIDs=" + infoIDs.size());
    if (infoIDs.size() * 0.7 < 10) {
        nums = infoIDs.size();
    } else {
        nums = (int) (infoIDs.size() * 0.7);
    }
    // 取权值最大的几个进行相似度比较
    for (int i = 0; i < infoIDs.size(); i++) {
        if (i < nums) {

```

```

        System.out.println(infoIds.get(i).toString());
        resultIDs.add(infoIds.get(i).getKey());
    } else {
        break;
    }
}
System.out.println(resultIDs.size());
// 通过地点ID列表获取点的所有信息
return dt.getPointsInfoByID(resultIDs);
}

private HashMap<String, Integer> countTimes(
    HashMap<Integer, Integer> scanResult) {
    HashMap<String, Integer> pointMap = new HashMap<String,
    Integer>();
    Set<Integer> key = scanResult.keySet();
    Iterator<Integer> itr = key.iterator();
    List<String> points; // 根据MACID返回的地点列表
    int count = 0;
    while (itr.hasNext()) {
        int MACID = itr.next();
        int pnum = 0;
        points = dt.findPointByMacIDandLevel(MACID,
scanResult.get(MACID),
        YUZHI);
        pnum = points.size();
        if (pnum == 0) {
            // textView.append("找不到可能地点\n");
            System.out.println("找不到可能地点");
            info = "找不到可能地点\n";
        } else {
            count += pnum;
            // 统计各点出现的次数
            for (int i = 0; i < pnum; i++) {
                String pointid = points.get(i);
                // 获取该地点的该MACID的靠谱度，若靠谱，则加入统计，若不靠谱则
                不加入统计
                double KP = dt.getKP(pointid, MACID);
                System.out.println(MACID + "在" + pointid + "的靠谱度为
" + KP);
                // System.out.println(pointid+"\n");
                if (KP > 5) {
                    // 如果重复出现
                    if (pointMap.containsKey(pointid)) {
                        pointMap.put(pointid, (pointMap.get(pointid) +
1));
                    } else {
                        pointMap.put(pointid, 1);
                    }
                }
            }
        }
    }
    // 将总数添加入图中，方便统计概率
    pointMap.put("count", count);
    // textView.append("共有" + pointMap.size() + "个可能地点\n");
}

```

```

info = "共有" + (pointMap.size() - 1) + "个可能地点\n";
System.out.println("共有" + (pointMap.size() - 1) + "个可能地点\n");
return pointMap;
}

```

4.3 Wi-Fi 定位中的问题

Wi-Fi 定位系统相比传统 GPS 全球定位系统有很多优点但是这么多年更多还是停留在实验室或者小规模测试上面，并没有大规模普及，本节将用实测的数据来说明几个阻碍 Wi-Fi 定位系统普及的原因。

4.3.1 城市基础设施建设不完善

Wi-Fi 定位最重要的是有 Wi-Fi 信号，但是中国的城市基础建设还不是很完善，很多地方没有 Wi-Fi 信号，即使有也更多的是一些用户自己架设的 Wi-Fi 热点，这些热点不是很稳定，用户可以随时关闭或者打开他们的热点，这些变化不能被及时反映到数据服务器上就会造成定位系统的误差增大。

不过我们可以看到很多城市现在都在加大智慧城市的建设，中国电信等公司也开始在全城覆盖 Wi-Fi 信号，这为 Wi-Fi 定位技术带来了越来越好的使用条件，相信在不久的未来，基于 Wi-Fi 的定位技术将更进一步走进人们的生活，给人们带来便利。

4.3.2 需要大量的现场勘测工作

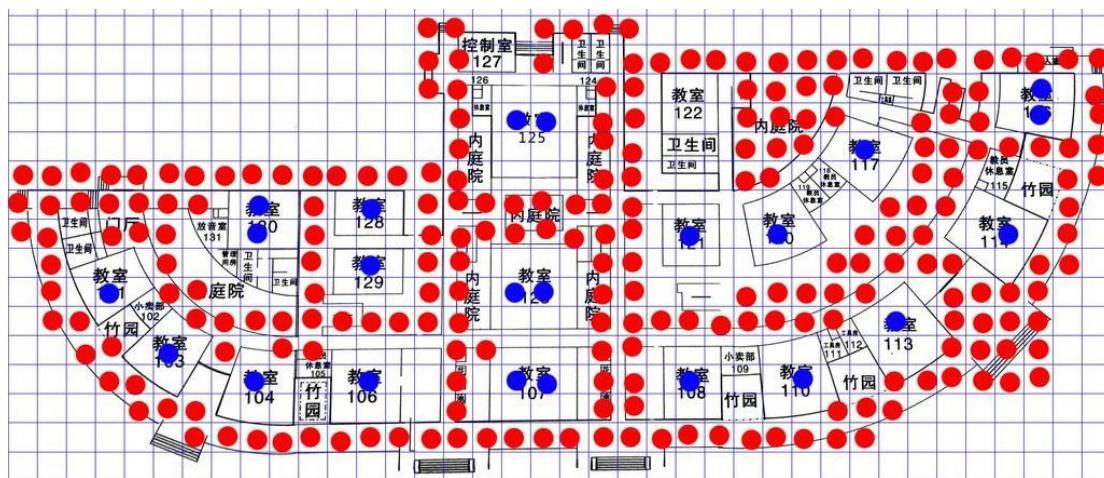


图 4.11 中南大学新校区 D 座 1 楼标志点示意图

使用 Wi-Fi 进行定位的基础是采集到足够多的 Wi-Fi 指纹信息，理论上来说，指纹信息越多越精细，Wi-Fi 定位的精度就越准确，但是这仅仅是在实验室环境下可以达到的，在实际应用中，需要采集指纹的面积远远大于实验室，所以很难

使用认为标定的方式来采集指纹信息,图 4.11 表示了如果想在中南大学新校区 D 座一楼部署 Wi-Fi 定位系统需要采集的指纹信息,其中红色点代表走廊,蓝色的点代表教室。

从图中可以看到如果要人为地去采集这么多的点十分耗费人力,如果要做到 Wi-Fi 定位的普及那么解决这个问题讲是十分重要的。本章下一节将介绍一种不需要用户显式的参与现场勘查的方法。

4.4 一种“不需要”现场勘查的 Wi-Fi 定位方法

上一节说到需要大量的现场调查工作是阻碍 Wi-Fi 定位普及的一个重要原因,在本节中将介绍一个由清华大学软件学院刘云浩等人提出的“不需要”现场勘查的 Wi-Fi 定位方法,本方法不需要用户显式的参与现场勘查工作,用户只需要在建筑物内正常的工作学习即可,系统可以自动地采集需要的数据。本方法详细介绍请参见文献[15],在这里仅针对于他的自动现场调查部份做简单叙述。

4.4.1 轨迹数据聚类——逻辑平面图创建

这个方法使用的是用户的连续轨迹信息,通过手机自带的加速度传感器判断用户的移动状态,当检测到用户移动时,采集 Wi-Fi 信号数据并且在用户移动的时间段内间隔一定时间采集一个数据组成连续的轨迹。系统将这些轨迹信息上传到服务器,并在服务器上,由于 Wi-Fi 信号由于墙壁的阻隔作用在不同房间内的差别很大,可以通过聚类算法将不同的数据点聚类为不同的类,每个类称作一个“虚拟房间”。如图 4.12 所示,是在一个建筑物的内的虚拟房间形成演示。

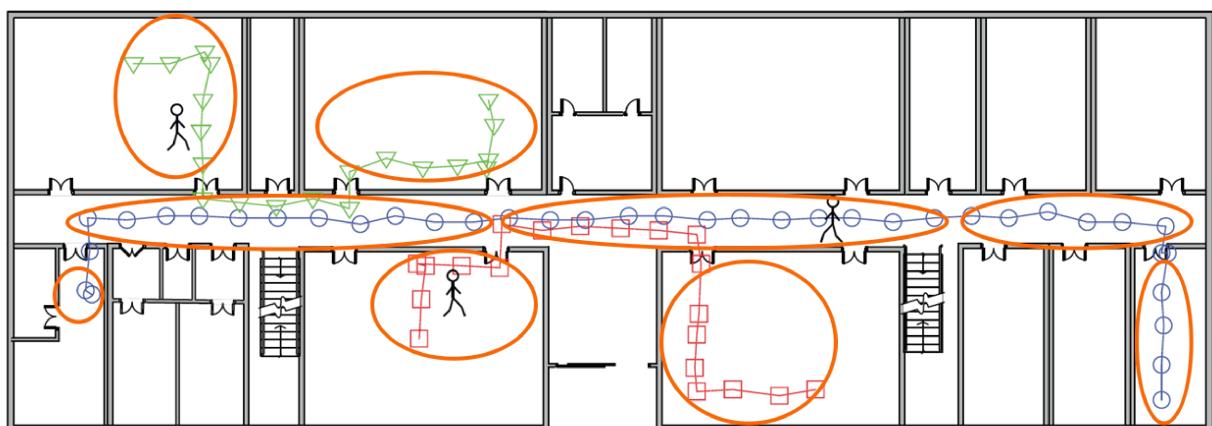


图 4.12 对 Wi-Fi 数据的聚类示意图

由于本方法使用的是轨迹信息,所以在使用聚类方法把数据点划分到不同的

虚拟房间中后可以根据用户的轨迹信息确定每个虚拟房间的连通情况,如图 4.13 为图 4.12 中虚拟房间的连通情况。例如用户从房间 A 经过走廊 C 到房间,说明 A 和 C 互相连通, C 和 B 互相连通,但是 A 和 B 不能直接连通。

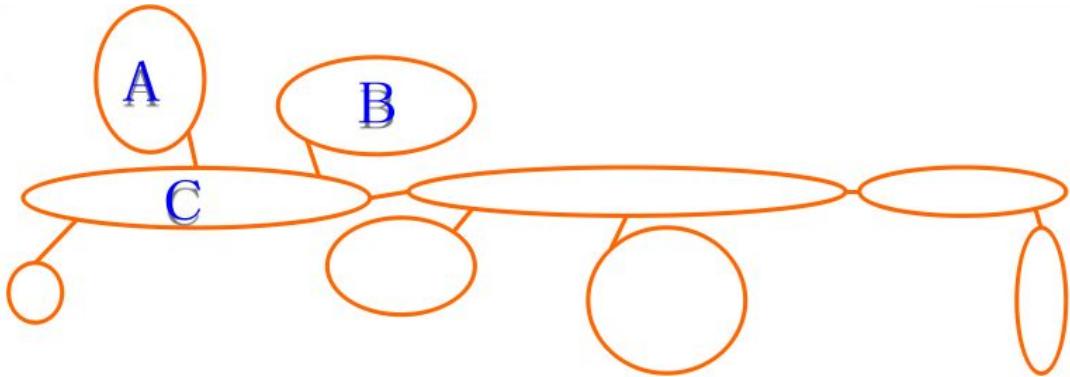


图 4.13 虚拟房间的连通示意图

4.4.2 逻辑平面图与物理平面图匹配

物理楼层平面图被建模成一个无向图 $P' = (V', E')$, 即物理图, 其中每个顶点 $v \in V'$ 表示一个房间(或功能区), 每条边 $(U, V) \in E'$ 是指 u 和 v 两个房间的可达性。根据这个方案, 当房间与房间之间没有门相连的时候走廊会与大部分房间相连。在物理平面图, 走廊可以根据不同房间的对应情况被分为若干段。具体来说, 对应着一个房间的走片段域通常被划分为一个区域, 这在物理图中表示为一个顶点(如图 4.14 所示)。因此, 每个段的长度大约是在与它连接的最大房间一致。本方法的实验模拟物理平面图见图 4.15, 走廊被分割为四部分。由于逻辑平面图 $P = (V, E)$ 和地面真实平面图 $P' = (V', E')$, 我们定义平面图映射为一个函数 $p: V \rightarrow V'$ 。再本方法中, 其设置的虚拟房间的数目等于或大于的物理区域的数量, 即, $|V| \leq |V'|$ 。

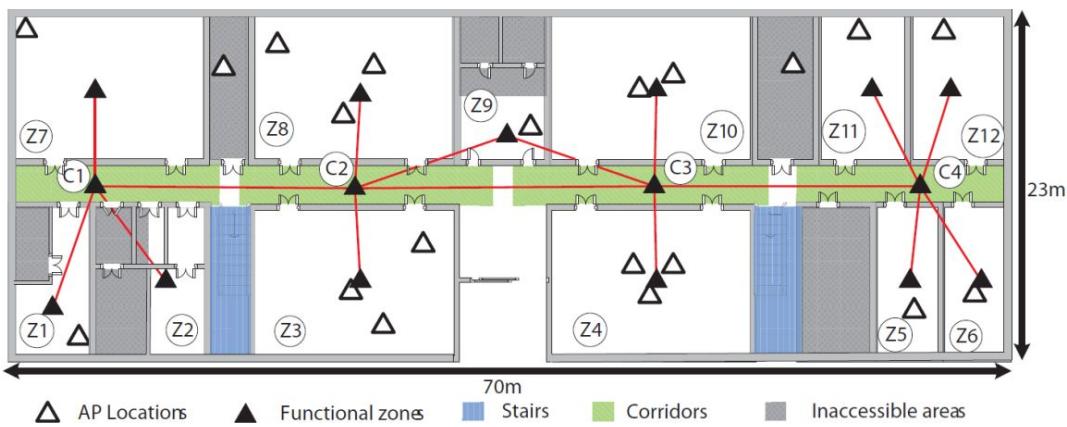


图 4.14 清华大学建筑平面图。不明位置的 AP 没有标出

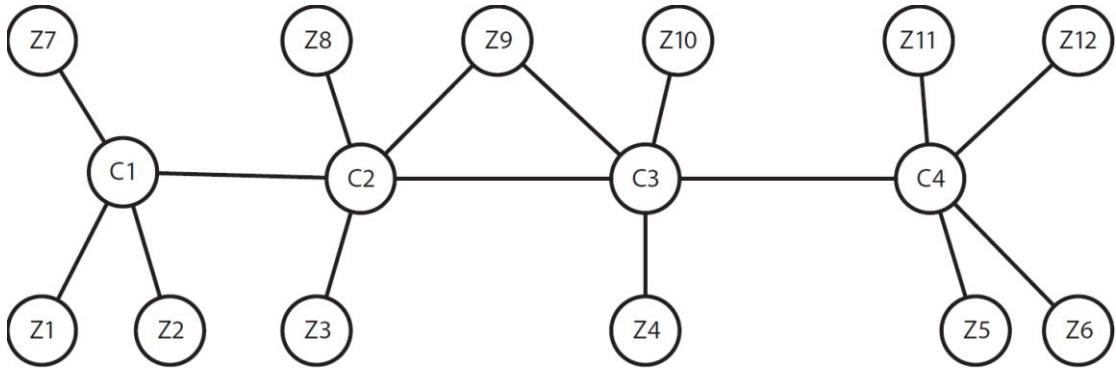


图 4.15 物理平面图。顶点 Z 代表房间，顶点 C 代表走廊片段

这个方法提出了一个分段映射方法 (SSMM)，其中包含三个阶段：骨架映射，分支节点映射和修正。具有较高的中介性虚拟房间都在骨骼映射映射之前，其余都使用分支结映射二分匹配的映射。初始映射结果在校正阶段调整。由于篇幅所限，只在这里介绍的简单框架。对于映射算法的细节，读者可参阅文献[22] 和[23]。

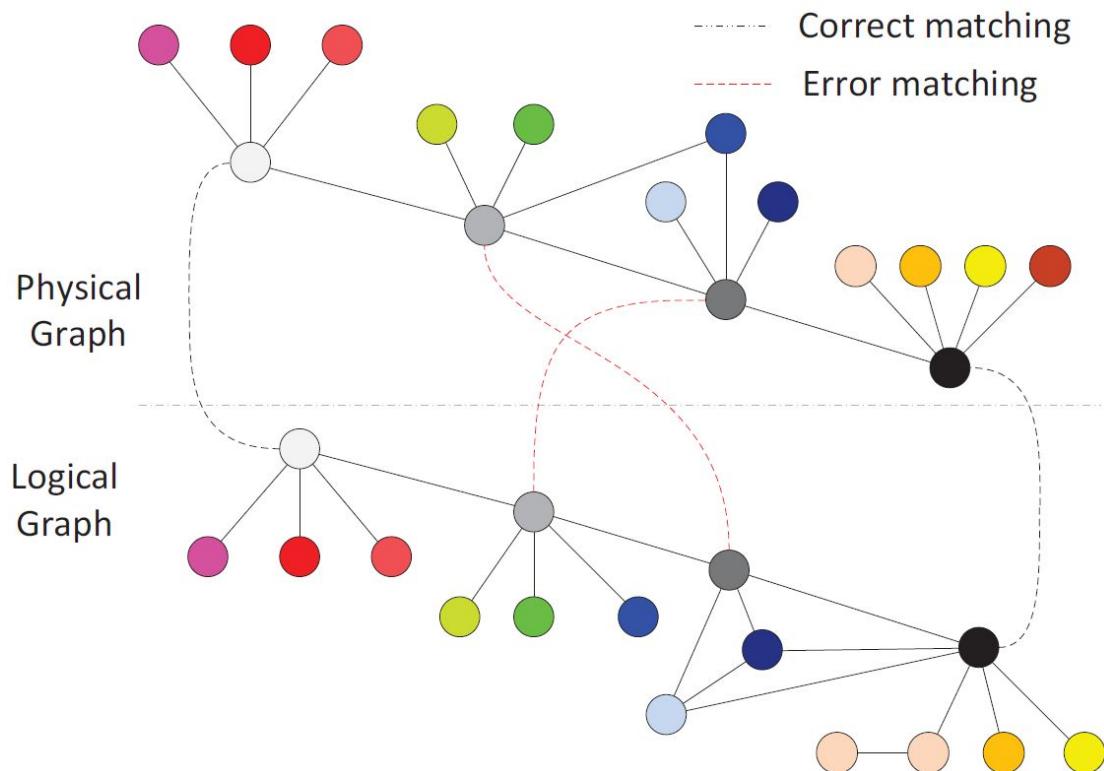


图 4.16 SSMM 第一步：骨架映射

1. 骨架映射。

中介中心性用来表示一个顶点在一张图上的位置。那些于其他节点之间有很多最短路径相邻的阶段比那些没有的具有更高的中介性。如图 5.15 所示，在中心（标记为 C1, C2, C3, 和 C4）顶点的中介性显然有比其他的

高。基于这一观察，其具有在 P 上的最高中介性顶点被映射到那些在 P' 中有最高的中介性的点上。这里映射的目的是尽量减少所有的匹配对之间的总差异。

2. 分支节点映射。

P 中的顶点的其余部分使用最短路径长度之和作为权重映射。换句话说，对于每一个顶点 v 在图 P 中，它的权重 $w(v)$ 等于所有最短路径的长度总和从 v 到 P 中的所有其他顶点，即 $w(v) = \sum_{u \in P, u \neq v} d(v, u)$ 。其中 $d(v, u)$ 是从 v 到 u 的最短路径的长度。每个顶点在 P' 的权重的计算方法相同。然后映射目标是最小化总重量差异，也就是说， $W(p) = \sum_{v \in V} |w(v) - w(p(v))|$ 。

本方法将分支节点映射形式化为一个 P 中每个顶点都与在 P' 中的另外的顶点完美匹配的加权最小二分匹配问题 (WMBM)。该 WMBM 问题应用的是 Kuhn-Munkras (KM) 算法。

结合骨架映射和分支节点映射的结果是，获得一个原始映射。图 4.16 和图 4.17 表示的骨架的映射和分支结映射的结果。由图可见最初的映射结果可能存在错误的匹配。我们进行 SSMM 的修正阶段修正一些错误的映射。

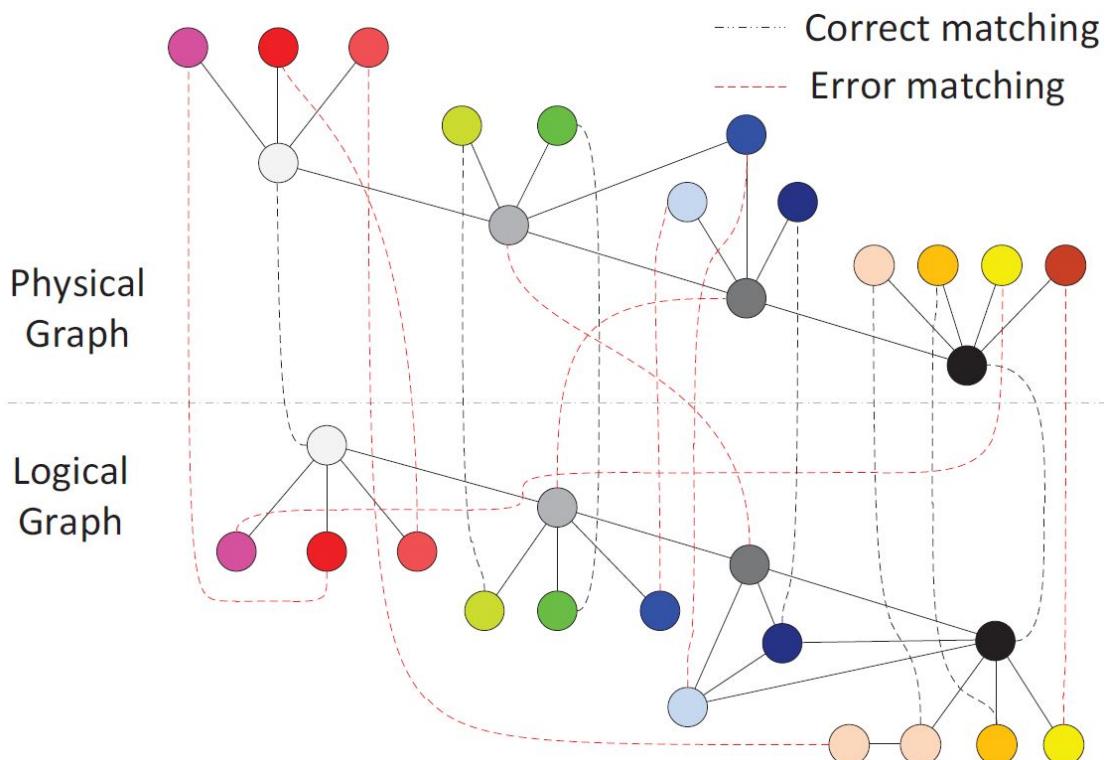


图 4.17 SSMM 第二步：分支节点映射

3. 校正

在最初的映射结果冗余信息被用于校正。通过比较每一个骨架顶点的相邻设置，错误映射可以计算出和纠正。其基本思想是：1) 如果一对映射的骨架顶点有非常不同的相邻集，他们往往是一个错误的链接；2) 如果两个分支结顶点不属于一对映射骨架顶点，它们有可能被错误地映射。SSMM 的在我们的实验校正后的结果示于图 4.18。

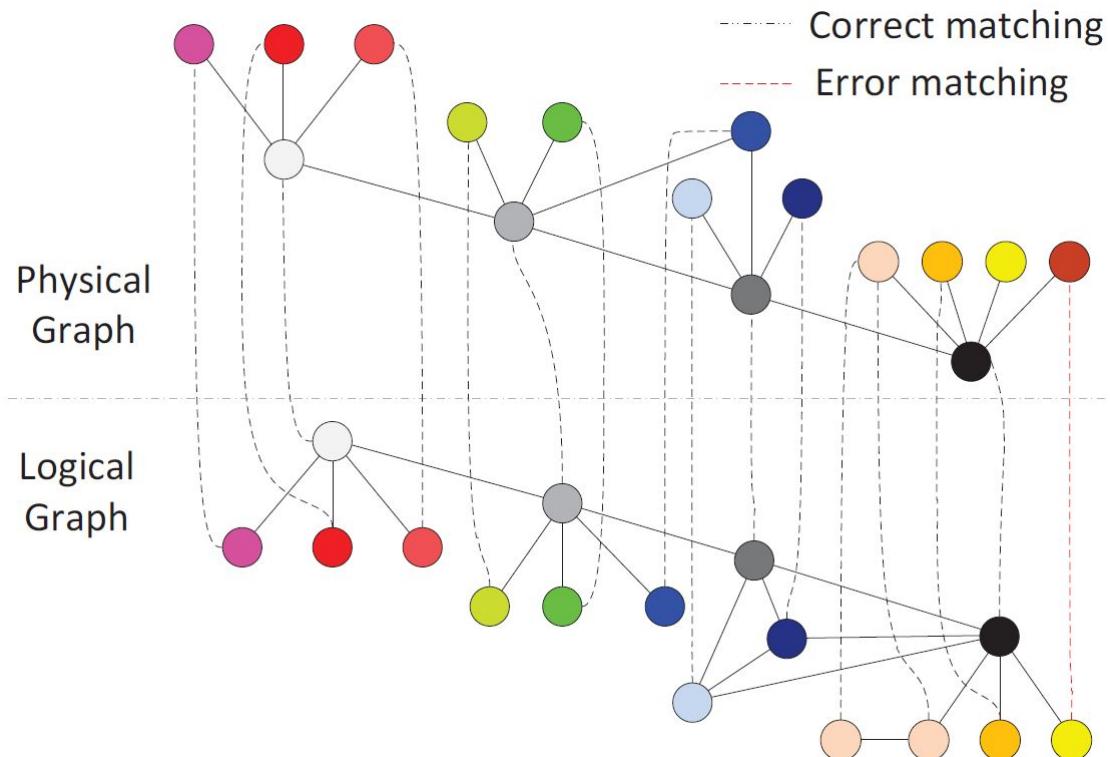


图 4.18 SSMM 第三步：矫正

4.5 本章小结

本章详细介绍了本次毕业设计中设计并实现的一个基于 Wi-Fi 指纹的定位方法的原型程序，描述了数据库建设，数据组织管理策略，定位算法及展示了部分关键代码。接着根据这个原型程序分析了现阶段 Wi-Fi 定位走向普及需要解决的问题，最后着重针对需要大量现场勘查工作这个问题进行具体要论并且介绍了一个由清华大学刘云浩等人提出的“不需要”现场勘察的 Wi-Fi 室内定位的方法，但是由于本次毕业设计时间有限，没能对这个方法进行自己的实现，这也是本项目今后要努力发展的方向。

第五章 总结及展望

本次毕业设计的题目是轨迹挖掘在机会网络异常发现中的应用研究。数据挖掘是现在和未来非常热门的研究领域，随着物联网和云计算概念的提出，会有更多的智能化设备连入到互联网并产生数以亿计的数据，其中人的轨迹信息是十分重要的数据，它蕴藏着巨大的商机。机会网络是靠节点的移动来传输数据的，所以和节点的轨迹信息有着密切的联系。在机会网络的应用中，在人群中应用的机会网络最为受关注，所疑惑的人的轨迹信息是在人群中应用轨迹挖掘技术的基础，而人们的活动大部分时间是在室内，所以本次毕业设计除了介绍轨迹数据挖掘之外还重点对 Wi-Fi 室内定位做了重点介绍并且自己编码实现了一个基于 Wi-Fi 指纹的定位系统。

本文的主要成果如下：

1. 从获取数据，数据传输以及数据分析三个方面阐述了轨迹数据挖掘与机会网路异常点发现的关系；
2. 对基于 Wi-Fi 指纹的定位技术做了较为详尽的描述并且例举了几个得到这些轨迹信息以后的应用；
3. 说明了数据挖掘中轨及异常点挖掘的定义以及介绍了轨迹异常点挖掘的一些方法；
4. 运用 Wi-Fi 指纹定位的方法，进行了 Android 平台下基于 Wi-Fi 指纹的定位系统的设计和实施并分析不足之处。

随着人们对于室内定位技术需求的加大，Wi-Fi 定位技术还将通过和其他技术和应用的不断交际而相互促进。以后的研究工作可以在以下几个方面展开：

1. 动态连续定位技术的研究。如何根据现有的用户位置信息预测用户的未来的运动趋势从而进行实时的定位，是基于 Wi-Fi 指纹的定位技术中的一个难点问题。
2. 融合多种定位技术的研究。如何融合 GPS 信号、运动传感器信号、地磁信息甚至是气压计进行精度更高，实用性更强的定位是一个值得深入研究的问题。随着 Wi-Fi 覆盖率的提升以及智能手机在用户中的不断普及，基于 Wi-Fi 定位，尤其是室内定位的研究和应用会越来越多，获得的轨迹数据将会越来越多，

研究基于 Wi-Fi 定位而获得的轨迹信息进行数据挖掘也将是以后的热门研究方向之一。

3. 运用数据挖掘的方法对获得的位置信息进行分析，在获得大量人群移动规律后通过挖掘发现人群流动规律等信息从而使得在人群中部署机会网络更容易。

随着国内智慧城市的发展以及智能手机的进一步普及，本次毕业设计的题目将越来越具有现实意义。

参考文献

- [1] 袁晶. 大规模轨迹数据的检索、挖掘和应用[D]:[博士学位论文]. 中国科技大学 2012
- [2] 刘大有, 陈慧灵, 齐红, 杨博. 时空数据挖掘研究进展[J]. 计算机研究与发展. 2013(02)
- [3] 姜莉. 基于 WiFi 室内定位关键技术的研究[D]:[硕士学位论文]. 大连理工大学 2010
- [4] 陈典全. LBS 中基于轨迹的用户行为特征分析[J], 全球定位系统 2011(006)
- [5] BAGLIONI, M., DE MACEDO, J. A. F., RENSO, C., TRASARTI, R., AND WACHOWICZ, M. 2012. How you move reveals who you are: understanding human behavior by analyzing trajectory data[J]. Knowledge and Information System Journal (KAIS), special issue on Behavior Computing. To appear.
- [6] 雷地球, 罗海勇, 刘晓明. 一种基于 WiFi 的室内定位系统设计与实现[A]. 第六届和谐人机环境联合学术会议 (HHME2010)、第 19 届全国多媒体学术会议 (NCMT2010)、第 6 届全国人机交互学术会议 (CHCI2010)、第 5 届全国普适计算学术会议 (PCC2010) 论文集[C]. 2010
- [7] Chen, L., M. Lv, and Qian Y. 2011. A personal route prediction system based on trajectory data mining[J]. Information Sciences 181(7), 1264-1284.
- [8] 熊永平, 孙利民, 牛建伟, 刘燕. 机会网络[J]. 软件学报. 2009(01)
- [9] 邓志安. 基于学习算法的 WLAN 室内定位技术研究[D]:[博士学位论文]. 哈尔滨工业大学 2012
- [10] 张明华. 基于 WLAN 的室内定位技术研究[D]:[博士学位论文]. 上海交通大学 2009
- [11] 赵小东. Android 平台下基于无线信号强度的定位系统的实现[D]:[硕士学位论文]. 哈尔滨工业大学 2011
- [12] C. Wu, Z. Yang, Y. Liu, and W. Xi. WILL: Wireless indoor localization without site survey[J]. In Proceedings of IEEE INFOCOM, 2012.
- [13] 姜金凤. 移动对象轨道异常检测算法的研究[D]:[硕士学位论文]. 南京航空航天大学 2010
- [14] 金卫民, 神显豪. 基于 RSSI 的室外无线传感网络自定位算法[J]. 计算机工程 2008, 34(13).
- [15] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition[J]. Machine Learning, 56:9 - 33, 2004.
- [16] DASGUPTA, S. The hardness of k-means clustering[R]. University of California. 2008.
- [17] M. Mahajan, P. Nimborkar, and K. Varadarajan. The planar k-means problem is NP-hard[J]. WALCOM: Algorithms and Computation, pages 274 - 285, 2009.
- [18] 刘良旭, 乐嘉锦, 乔少杰, 宋加涛. 基于轨迹点局部异常度的异常点检测算法[J]. 计算机学报 2011, 10(34)
- [19] 王宏鼎, 童云海, 谭少华, 等. 异常点挖掘研究进展[J]. 智能系统学报, 2006, 1 (1). 67-73.
- [20] 王孝杰, 郑雪峰, 宋一丁, 曲阜平. 面向轨迹数据流的 KNN 近似查询[J]. 计算机工程, 2011, 37(16): 17-20
- [21] C. Wu, Z. Yang, Y. Liu, and W. Xi. Site-survey-free wireless localization using mobile phones[R]. Hong Kong University of Science and Technology, 2011.
- [22] C. Wu, Z. Yang, Y. Liu, and W. Xi, WILL: Wireless indoor localization without site survey[R]. in INFOCOM, 2012 Proceedings IEEE, march 2012, pp. 64 - 72.

致谢

在本论文完成之际，向所有帮助我、关心我的师长、同学、亲人、朋友表示衷心的感谢和诚挚的敬意。

首先，我要诚挚的感谢我的导师陈再良老师和沈海澜老师以及我们的系主任黄东军老师，由于我参加了学校与加拿大渥太华大学的交流学习项目，大四期间都在加拿大渥太华学习，那边与国内的时差正好是十二个小时，尽管如此，在毕业设计以及论文的写作过程中，多次得到老师的督促，并且为我的论文提出了许多宝贵的修改意见。不仅是毕业设计，在平常学习中，老师也经常召集我们开会座谈，对于我们在学习及生活上的一些疑问进行解答。我也要感谢鲁鸣鸣老师，正是他在物联网编程课上的讲解让我对 Wi-Fi 定位产生了兴趣并促使我对其进行了相应的研究。在这里也一并感谢漆华妹老师、高建良老师、任秀老师等老师在大学四年对我的关心和帮助。

其次，要感谢我的班长王皓同学，在我在国外学习期间，各种通知都及时地告知我，让我没有错过学校重要的通知，感谢同组的袁伟同学在我制作 Wi-Fi 定位系统的时候对我的帮助，还有感谢我的朋友及同学们，让我大学四年留下了很多美好的回忆。

另外，我也要感谢我的女朋友李淑芬，正是由于她经常叫我去新校区、图书馆以及三教一起自习让我没有宅在寝室里面浪费时间，而且本次毕业设计中的 Wi-Fi 定位系统的编码实现过程以及 Wi-Fi 指纹数据的采集及分析过程也都是在和她一起自习的时候完成的。

最后，我要特别感谢我的父母大学四年对于我的无私支持，在我想到一个想法并想要付出实践时，我不会为没有足够的经费来采购实验设备而发愁，因为我父母总是无条件支持我的任何有助于学习研究的做法，正是他们的关怀与支持才使我有了不断前进的动力。