

分类号 TP393

学号 14063020

U D C

密级 公开

工学硕士学位论文

基于智能手机传感器数据的用户关系计算研究

硕士生姓名 王峰

学科专业 计算机科学与技术

研究方向 移动感知

指导教师 刘东波 研究员

国防科学技术大学研究生院

二〇一六年十月

Users Relationship Strength Measurements Based on the Mutuilayer Hybird Model

Candidate: Feng Wang

Advisor: Professor Dongbo Liu

A dissertation

**Submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
in Computer Science and Technology**

Graduate School of National University of Defense Technology

Changsha, Hunan, P. R. China

October 5, 2016

独 创 性 声 明

本人声明所呈交的学位论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表和撰写过的研究成果，也不包含为获得国防科学技术大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文题目： 基于智能手机传感器数据的用户关系计算研究

学位论文作者签名： _____ 日期： 年 月 日

学 位 论 文 版 权 使 用 授 权 书

本人完全了解国防科学技术大学有关保留、使用学位论文的规定。本人授权国防科学技术大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档，允许论文被查阅和借阅；可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

(保密学位论文在解密后适用本授权书。)

学位论文题目： 基于智能手机传感器数据的用户关系计算研究

学位论文作者签名： _____ 日期： 年 月 日

作者指导教师签名： _____ 日期： 年 月 日

目 录

摘要	i
ABSTRACT	iii
第一章 绪论	1
1.1 研究背景	1
1.2 研究现状	2
1.2.1 关系强度的研究现状	2
1.2.2 基于移动数据人际关系度量研究	3
1.2.3 基于移动轨迹数据的关系度量研究	3
1.3 研究内容	4
1.4 论文结构	6
第二章 相关技术研究	9
2.1 用户轨迹预处理	9
2.1.1 滤波算法	9
2.1.2 轨迹停留点检测	12
2.1.3 聚类算法	12
2.2 时间序列相似度度量方法	15
2.2.1 轨迹中心距离	17
2.2.2 Dynamic Time Warping	18
2.3 自然语言处理方法	19
2.3.1 LDA 主题模型	20
2.3.2 word2vec	20
2.3.3 快速 Hash 算法	23
2.4 WiFi 和蓝牙数据的度量方法	23
2.5 小结	23
第三章 RSMHD 用户关系强度计算框架模型	25
3.1 RSMHD 模型框架描述	25
3.1.1 基于轨迹数据的关系强度计算模块概述	25
3.1.2 基于 WiFi 和蓝牙数据的关系强度计算模块概述	27
3.2 RSMHD 模型计算流程概述	27
第四章 RSMHD 框架的相关技术研究	29
4.1 GPS 轨迹处理计算技术	29

4.2 GPS 轨迹预处理与语义化	30
4.2.1 剔除轨迹中的异常点	30
4.2.2 轨迹中停留点检测	32
4.2.3 用户轨迹停留点聚类	34
4.2.4 对用户轨迹的语义位置添加语义标签	40
4.3 用户 GPS 轨迹数据结构化表示	43
4.4 WiFi 感知数据处理计算技术	45
4.5 蓝牙感知数据处理计算技术	46
4.6 小结	48
第五章 RSMHD 用户关系强度计算方法概述	49
5.1 用户关系强度计算方法概述	49
5.2 输入数据准备	49
5.2.1 轨迹数据的处理与准备	50
5.2.2 语义位置数据的处理与准备	51
5.2.3 语义标签数据的处理与准备	51
5.3 关系强度计算	52
5.3.1 基于原始轨迹数据的关系强度计算	52
5.3.2 基于主题模型的关系强度计算	53
5.3.3 结果投票	54
5.4 小结	55
第六章 数据集、评估方法及实验结果	57
6.1 数据集	57
6.2 评估方法	61
6.2.1 构造真实结果	62
6.2.2 评估方法	63
6.3 实验结果与分析	65
6.4 小结	75
第七章 结束语	77
7.1 工作总结	77
7.2 工作展望	78
致谢	81
参考文献	83
作者在学期间取得的学术成果	87

表 目 录

表 6.1 调查问卷 ^[3]	58
表 6.2 基站区域号与对应的语义标签	61
表 6.3 真实结果	64
表 6.4 无优化 DTW 方法得到好友列表	67
表 6.5 优化 DTW 方法得到好友列表	68
表 6.6 基于语义位置行为模式相似度得到好友列表	71
表 6.7 LDA 模型学习到的主题	72
表 6.8 基于语义标签行为模式相似度得到好友列表	73
表 6.9 三层结果投票得到好友列表	74

图 目 录

图 2.1 用户 GPS 轨迹示例图	12
图 2.2 聚类效果示意图	13
图 2.3 DBSCAN 密度直达和密度可达示意图	14
图 2.4 关于 decision graph 的示例 ^[42]	16
图 2.5 聚类中各点的 ρ 和 δ 取值 ^[42]	16
图 2.6 轨迹相似计算示例 ^[43]	17
图 2.7 基于 Center of mass 轨迹计算示例 ^[43]	18
图 2.8 DTW 算法匹配序列结果示意图	19
图 2.9 LDA 的图模型表示	21
图 2.10 word2vec 神经网络结构图 ^[48]	21
图 2.11 word2vec 神经网络结构图	22
图 3.1 RSMHD 用户关系强度计算模型框架图	26
图 3.2 RSMHD 中基于用户轨迹计算模块详细框架图	26
图 3.3 RSMHD 中基于 WiFi 和蓝牙计算模块详细框架图	27
图 4.1 基于多层次多粒度轨迹的用户关系强度计算	29
图 4.2 轨迹中值滤波实验结果 1	30
图 4.3 轨迹中值滤波实验结果 2	30
图 4.4 轨迹均值滤波实验结果 1	31
图 4.5 轨迹均值滤波实验结果 2	31
图 4.6 卡尔曼滤波实验结果 1	32
图 4.7 卡尔曼滤波实验结果 2	32
图 4.8 分段卡尔曼滤波轨迹结果 1	32
图 4.9 分段卡尔曼滤波轨迹结果 2	33
图 4.10 现实生活中的停留点示意图	33
图 4.11 停留点检测实验结果 1	33
图 4.12 停留点检测实验结果 2	34
图 4.13 基于 K-means 的轨迹聚类实验结果 1	35
图 4.14 基于 K-means 的轨迹聚类实验结果地图展示 1	35
图 4.15 基于 K-means 的轨迹聚类实验结果 2	36
图 4.16 基于 K-means 的轨迹聚类实验结果地图展示 2	36
图 4.17 基于 DJ-Cluster 的轨迹聚类实验结果 1	37
图 4.18 基于 DJ-Cluster 的轨迹聚类实验结果地图展示 1	37

图 4.19 基于 DJ-Cluster 的轨迹聚类实验结果 2	38
图 4.20 基于 DJ-Cluster 的轨迹聚类实验结果地图展示 2	38
图 4.21 轨迹聚类实验结果展示 1	39
图 4.22 轨迹聚类实验结果地图表示 1	39
图 4.23 轨迹聚类实验结果展示 2	39
图 4.24 轨迹聚类实验结果地图表示 2	40
图 4.25 位置语义标签示意图	41
图 4.26 语义标签识别示意图	42
图 4.27 语义标签结果示意图	42
图 4.28 用户轨迹的抽象表示	44
图 4.29 用户语义轨迹描述模型	44
图 4.30 语义轨迹示意图	45
图 4.31 WiFi 数据结构化表示示意图	46
图 4.32 WiFi 上下文抽象化表示示意图	46
图 4.33 WiFi 图结构抽象化表示	47
图 4.34 蓝牙数据结构化表示示意图	47
图 4.35 蓝牙数据上下文抽象化表示示意图	47
图 4.36 蓝牙数据图结构抽象化表示	48
图 5.1 URSHV 模型框架	49
图 6.1 志愿者采集基站数据可视化	60
图 6.2 朋友关系可视化	62
图 6.3 DTW 实验结果	69
图 6.4 加权前后实验结果对比	69
图 6.5 基于语义位置实验结果	70
图 6.6 基于语义标签实验结果	72
图 6.7 投票结果	75

摘要

近几年，随着传统手机逐渐被智能手机所取代，搭载了智能操作系统(如 iOS、Android 和 Windows Phone)的智能手机已经成为了人们生活中集日常通信、娱乐游戏、商务办公、感知计算等于一体的移动掌上终端平台。通过搭载了更多传感器设备的智能手机能够随时随地的获取到用户的位置、通信记录、短信记录、日常轨迹分布情况等各种体现用户与用户之间的隐藏社会关系的感知信息。人们之间的轨迹相似性、日常轨迹的共现频率和时长，用户连接 Wi-Fi 的共现使用情况以及用户蓝牙之间的交互信息都能够分析得出人与人之间的交互关系以及他们之间的关系强度。通过对用户之间的相似性和关系强度的探究和计算有利于进一步发展个人的社交网络和探究社会群体结构的发展以及演变过程等。

传统的度量用户之间的社交关系大多采用的是基于社交关系数据来分析用户之间的社会关系(社交关系数据如通话记录，短信记录，社交软件的交互等)，本文基于智能手机所采集的非社交关系数据针对如何使用非社交关系数据来分析度量人们社交活动之间的相似性以及关系强度问题展开了进一步的深入研究。设计并且实现了一个基于用户轨迹数据、Wi-Fi 感知数据、蓝牙感知数据抽象出的多个层次的度量人与人之间在社交活动中相似性以及关系强度的计算模型 RSMHD(Relationship Strength based on Multiple Hierarchy Dimension)。整个计算模型的主要涵括的内容如下：

为了便于研究的顺利开展，我们假设陌生人之间的关系强度应该是无限趋近于零的，同时我们为了便于对计算结果的准确性进行判断和数据的采集，本文研究对象主要限于在校学生群体。社会心理的理论支持以及研究成果指出：相似的人更容易产生交集成为朋友，现实生活中人们也更加倾向于同自己相似的人做朋友，也就是说相似导致喜欢，从而发展为友情或者爱情。这也就预示着相似度人与人之间的关系强度要高于非相似着，这也就为我们的研究提供了理论依据。

首先，针对用户日常轨迹数据我们分别从空间移动轨迹和现实语义轨迹出发，对用户之间的相似性进行计算，本文采用位置漂移修正算法对部分因无信号原因导致的 GPS 位置缺失的轨迹进行预测补全；针对补全后的用户轨迹，采用时间片式的卡尔曼滤波算法剔除用户 GPS 轨迹中的异常点；采用基于时间 - 密度的聚类算法得到用户空间轨迹中的停留点，在此基础上，采用基于语义规则的语义标签标注机制，通过用户参与，反地理编码等手段对停留点进行现实语义标注，将用户的空间轨迹转变为符合现实意义的日常语义轨迹，为利用位置感知数据来

计算用户的关系强度做好准备。然后针对用户空间语义轨迹，本文采用了快速 DTW 计算方法来计算用户空间轨迹之间的相似性，并采用动态加权算法对计算结果进行加权处理；针对日常语义轨迹，我们分别从两个层次出发，一方面采用了 Word2vec 来计算用户在切片时间内的语义轨迹的相似性；另一方面，根据用户的语义轨迹挖掘出用户的轨迹运动模式，通过快速计算相似度算法 Simhash 来计算出用户的日常轨迹运动模式之间的相似性并对结果采取加权处理得出用户之间的关系强度，最后对计算结果进行融合来得出用户的关系强度。

其次，针对用户 WiFi 感知数据，采用关联拓扑图表示某个时刻的用户 WiFi 上下文环境信息，然后提取出用户 WiFi 的交互情况作为计算用户关系强度的特征之一；针对时间片上的拓扑图创新的提出了通过计算图与节点之间的相似性作为某时刻 WiFi 上下文环境信息之间的相似性。针对用户的蓝牙感知数据，通过结合脸呀交互时长、交互频率、蓝牙上下文环境相似性等来计算出用户之间的关系强度。

最后，利用集成学习的思想对以上三种不同感知数据源计算结果进行融合处理，得到基于不同非社交数据源所计算得到的最终用户关系强度。在该模型算法研究的基础上，本文基于自主开发的用户上下文信息收集系统 StarLog 收集了多名学生成长时间的智能手机感知数据，利用该数据源来计算出用户之间的关系强度，并结合用户的相似度问卷调查表对结果进行有效性验证，表明该模型能够有效地利用非社交关系度量出用户之间的关系强度。

关键词: 关系强度；非社交数据；轨迹模式；word2vec；集成学习

ABSTRACT

Smart phones have become an integral part of daily life communication tools, we can collect intelligently location, call logs, text messages, WeChat anywhere which reflect a variety of information daily interactions and social relations between people. People interaction frequency, time, location, distance and the similarity of trajectory information reflects the strength of the relationship and the relationship between people. Relationship strength reflects the degree of intimacy between two different persons, which is of great importance in analyzing human's social relationship as well as social network.

In this paper, we propose a URSHV(User Relationship Strength Hierarchy Vote), which can measure the relationship between people in daily life through GPS data from three levels, namely: daily trajectory, semantic locations and semantic labels. To sum up, the main research contents and contributions are as follows:

First of all, the strength of the relationship between users and semantic location with semantic labels are closely related, this paper uses segmented kalman filtering algorithm on GPS trajectory data to de-noising; using the clustering algorithm based on density of position trajectory data clustering, and form a semantic position; on this basis, the semantic annotation mechanism based on Rules, the semantic annotation of semantic encoding, geographic location by anti semantic label inference and input auto completion etc; the GPS position trajectory data sequence clustering into meaningful semantic position and semantic labels, which laid the foundation for the semantic location and semantic labels based on the strength calculation of the relationship between users.

Secondly, in order to calculate the strength of the relationship between users from the three levels of trajectory data, semantic location and semantic labels, using DTW model of spatial distance calculation between users to measure the similarity between the users, the use of trajectory sequence similarity of users every day to track the entropy weighting processing, and the strength of relationship between users the LDA were calculated using the topic model; semantic locations similarity and semantic labels based on behavior patterns among users, as the strength of the relationship between users; measurement results of three levels of the ensemble learning theory to vote, to vote as the strength of relationship between end users.

Finally, on the basis of the above study, based on the MIT reality mining project of publicly available data sets, similarity between users by using the data set of the questionnaire, users construct between a real relationship strength as a benchmark for testing proposed an inverse logarithmic induced score measurement method to measure the strength of the relationship between users based on, and effectiveness model of URSHV for experiments, the results show that the model can effectively measure the strength of the relationship between users.

Key Words: relationship strength; trajectory similarity; DTW; entropy; LDA; vote

第一章 绪论

近年来，随着传统手机逐渐被智能手机取代，搭载了智能操作系统（如ios,Android,Windows Phone）的智能手机已经成为人们日常生活中集通信，娱乐游戏，商务办公，感知计算等于一体的移动个人终端平台。随着传感器工艺界的发展以及内置手机传感器种类的增加，如今智能手机已经能够基于传感器为用户提供良好的用户体验与服务，因此智能手机被科技赋予了更多的责任，智能手机不仅仅是简单的日常通讯工具，还是能够在记录用户的日常活动轨迹，了解用户的使用习惯，为用户提供实时信息推荐，俨然已经成为了一种新型的可穿戴计算^[1]的载体，随着传统社交方式逐渐转变为手机上的社交方式，借助现有的数据分析和数据挖掘技术，通过分析手机丰富的上下文信息（Context）来研究用户社交关系以及人与人之间的社会关系已经成为移动计算中一个研究热点。

1.1 研究背景

近年来，随着传统手机逐渐被智能手机取代，搭载了智能操作系统（如ios,Android,Windows Phone）的智能手机已经成为人们日常生活中集通信，娱乐游戏，商务办公，感知计算等于一体的移动个人终端平台。通过丰富的内置传感器，我们可以收集到更加丰富的上下文信息（context） 分析出人与人之间的交互情况以及相似性^[2-5]，以及通过用户轨迹数据分析用户的社交动态^[6]。关系强度在本文中表现为用户在现实生活中的亲密程度，通过掌握用户之间的关系强度有利于合理的拓展社交网络和提升社交质量。文献 [3] 通过对手机传感器数据分析得出不同社会关系群组的用户在日常社交活动中的规律。CenceMe^[7] 通过收集智能手机传感器的上下文信息以及用户的社交应用（如 Facebook，MySpace 和即时通讯工具 Skype）的使用信息，分析用户的位置、活动、情绪和周围环境并将所有信息通过社交应用来分析用户的社会交互情况但是并未根据社交情况去推测用户之间的社交关系等。Dartmouth 大学学者们的研究中^[8,9]，通过对班级 48 名学生成达 10 周所采集的手机传感器数据进行研究从中分析出学生之间的社交活动有助于减轻个人压力、保护精神健康和提升学业成绩的作用，研究人与人之间的社交活动和关系强度也因此具有重要的研究意义。美国社会学家格兰洛维特在研究过程中^[10] 首次提出了关系强度这一概念，研究将关系强度分为强关系和弱关系。同时指明能够充当信息传递载体的纽带关系必定是弱关系，而强关系只是存在于那些你真正充分信任的人之间；强关系存在与和你更加相似的用户之间。Caroline^[11] 等进一步证实了用户交互的频率以及交互的持久度决定了用户之间的关系强度，如具有较高亲密度和交互频率用户之间的关系强度要高于偶然交互和非持久交互

的用户。文献 [12–14] 基于在线社交网络 (Facebook、Twitter、LinkedIn、Instagram 等) 利用社交数据、个人资料、状态互动等分析用户之间的亲密度，结合机器学习方法采用决策树算法、MLP 算法、SVM 算法等进行用户之间关系强度的预测以及分类。以上提及的研究大都是基于用户社交数据进行的关系强度计算研究，而基于智能手机收集的非社交的传感器数据进行人与人之间关系强度的度量仍是一个值得研究的问题。

随着各种定位技术（如全球定位系统 GPS 和 WLAN 和移动网络）的发展和内嵌的定位模块，通过智能手机可以准确的获取用户的位置信息，提供各种基于位置的服务（LBS），并将用户的活动以轨迹的形式记录下来。我们阅读了大量社会心理学相关的论文书籍，从中证实得到在现实生活中关系亲密的两个用户会更加倾向于一起进行面对面的交流、共同进行社交活动等，因此通过对手机传感器数据的处理分析能够从中挖掘出人们现实生活中的关系强度。文献 [15] 立足于空间距离，提出了在空间距离上非常接近的人们在现实生活中就越可能发生面对面的交互，该文献通过调研一个小区的住户发现人们在现实生活中越是接近，就越容易成为朋友。文献 [16, 17] 进一步通过研究用户的轨迹数据发现由于在空间距离中接近的用户在现实生活中更可能产生交互行为，也就是说拥有相似日常生活轨迹的用户更可能产生交互活动。郑宇 [18] 以及其他学者 [19–21] 尝试了基于用户轨迹计算用户相似性的研究，得出相似轨迹的用户更可能成为朋友，因此，针对相似用户进行朋友关系推荐。基于以上的理论基础和研究经验本文通过收集智能手机获取到的用户位置、WiFi、蓝牙、通话记录等手机上下文数据，并且从这些数据中分析计算出现实生活中用户的交互位置、交互时间、交互频率以及交互的持续性等一系列能够反射出用户间关系强度的信息特征。

目前智能手机增加了多种传感器用于实现更加丰富的用户交互功能，由以前单一的加速度传感器，距离传感器逐步集成了压力传感器，温度传感器，心率传感器等。这使得智能手机能够更加准确地感知到更加丰富多样的周围环境信息如：用户位置，社交通讯记录，WiFi 和蓝牙连接记录等体现人与人之间交互情况以及轨迹交互情况等体现用户关系强度的数据，通过收集分析这些非社交数据，我们能够进一步得出人与人之间的关系相似性以及关系强度。

1.2 研究现状

1.2.1 关系强度的研究现状

Granovetter 关于弱关系的研究奠定了社会关系强度理论的基础^[10] 被认为是社会关系理论研究的开始的标志，紧接着 Burt 根据 Granovetter 的弱关系理论研究提出了结构洞理论^[22]。Granovetter 针对弱关系和强关系的度量方法提出了基于四

种维度的度量准则，即用户之间的日常互动次数、人与人之间的亲密程度、双方投入感情的程度以及用户日常的交换程度，基于这四条衡量标准就可以将用户的关系强度划分为弱关系和强关系；在 Wegner^[23] 的研究中，更是对这四种度量标准进行了进一步的研究推进，采用数值化来衡量四种维度的标准使得能够以指标数值化来区分强关系和弱关系；Muncer 等^[24] 提出并验证了用户的关系数量以及任意关系之间的交互频率对人与人之间的关系有影响；Paolillo^[25] 从日常用户交流的角度出发，发现人与人之间关系的亲密程度与日常交流中昵称使用的频率有关。随着研究的进一步深入，度量用户的关系强度逐步形成了基于感知用户的社交数据出发，以用户交互、亲密度等出发为度量标准的研究观点^[26]。

1.2.2 基于移动数据人际关系度量研究

Hsu 通过采集到的志愿者的手机日常 WiFi 数据^[27]，将位置与 WiFi 信息关联起来得到用户的 WiFi 关联向量，并用关联向量来表示用户的行为轨迹，同时基于提出的 AMVD 模型用来计算人与人之间的距离，最后根据距离对社交关系进行聚类对社会关系进行划分，文中认为当连接过相同 WiFi 的情况下可以认为两个人在现实生活中有较强的社交关系。并且根据 WiFi 所对应的语义位置（如图书馆、教室、咖啡厅、会议室等）推测他们之间可能的社交关系。基于蓝牙感知信息，研究者通过分析手机收集的蓝牙数据对用户的社交圈进行划分，将用户的社交圈划分为室友，好朋友，工作伙伴等^[28-30]。Mtibaa 等^[31] 通过收集分析了 28 位参加同一个计算机国际会议参会者的手机蓝牙数据，根据分析结果绘制了关于 28 位作者的社交网络关系图。

1.2.3 基于移动轨迹数据的关系度量研究

在现实生活空间中的用户交互能够更加直接的反映出社会关系的情况，如面对面的交流，共同用餐等，这些现实生活中的用户交互相对于用户的传统社交数据能够更加真实的反映出用户二者之间的关系。但是目前在这方面的研究还仅仅局限于某个局部的方面，文献 [32] 创造性的提出了根据用户的日常轨迹来衡量用户之间的关系强度。Ma 等人^[32] 提出了一种根据多层基于用户轨迹的层级熵关系度量方法 HERMA(Hierarchical Entropy-based Relationship Measurement Approach)，该模型根据手机收集的用户的 GPS 位置轨迹信息进行处理，从用户轨迹中提取出共同的位置记录来推断用户之间的物理交互，进一步使用用户之间的物理交互来度量用户之间的社会关系强度，最后在仿真的数据集上进行验证。

1.3 研究内容

本研究课题针对如何通过智能手机所收集到的用户之间的非社交数据来计算度量人与人之间的社交关系强度展开了一系列的研究，寻求通过建立一个能够同时计算处理多种不同非社交数据源数据的用户关系度量框架。以此为基础来衡量用户之间的关系强度。经过前期的研究，决定将手机感知数据中的非社交关系数据（用户轨迹数据、用户蓝牙数据、用户 WiFi 数据）作为衡量用户关系强度的数据源。为了实现基于不同感数据多维度计算用户之间的关系强度，本文分别针对每一个独立数据源展开研究，最终采用集成学习的思路将结果进行融合。

(1) 如何基于用户日常轨迹度量关系强度

随着各种定位技术（如全球定位系统 GPS 和 WLAN 和移动网络）的发展和内嵌的定位模块，通过智能手机可以准确的获取用户的位置信息，提供各种基于位置的服务（LBS），并将用户的活动以轨迹的形式记录下来。我们阅读了大量社会心理学相关的论文字籍，从中证实得到在现实生活中关系亲密的两个用户会更加倾向于一起进行面对面的交流、共同进行社交活动等，因此通过对手机传感器数据的处理分析能够从中挖掘出人们现实生活中的关系强度。在获取用户位置的过程中，其结果既可以由智能手机内置的 GPS 传感器提供，还可以通过基站和 WiFi 定位获取。用户的每一个位置点都是由一个三元组结构 (Latitude,Longitude,Time) 组成的，用户在一段时间内连续的位置记录就构成了一条完整的连续轨迹。根据用户的日常语义轨迹，采用滤波技术手段剔除用户轨迹中的异常点，同时利用轨迹预测算法对短时间窗口内的用户位置进行有效地补充。针对用户的轨迹从不同层次出发，计算不同轨迹形态下用户之间的关系强度加以合并。

a) 基于空间轨迹的用户关系度量

用户的空间轨迹中包含了一串由 (Latitude,Longitude,Time) 所组成的，寻找出用户空间轨迹中的特殊点即停留点，停留点^[33](Stay Point) 在现实生活中并不是指用户轨迹中速度为零的点，而是由一组 GPS 点构成区域，表示用户在这段轨迹中在某一个区域停留的时间超过了设定的阈值。它比传统的 GPS 位置点蕴含了更加重要的信息在，如该用户去过图书馆、体育馆等。因此用户的空间轨迹的距离现在一定程度上体现了用户之间的亲密关系，本文正是基于此尝试采用空间轨迹距离来度量人与人之间的社交关系强度。传统的距离计算主要使用欧氏距离、曼哈顿距离、马氏距离等。但是，在使用欧氏距离来度量用户空间轨迹的相似性过程中，轨迹之间的距离会受到轨迹长度的影响，导致结果出现较大偏差。通过结合传统语音识别算法 DTW^[34]，降低用户空间轨迹长度对计算结果的影响；考虑到

DTW 计算结果受到序列长度的影响，轨迹序列越长，得到的轨迹距离结果越大，因此对 DTW 计算结果采用合理的归一化处理以消除不同用户轨迹长度所带来的差异性影响。

b) 基于语义轨迹的用户关系度量

基于空间轨迹的度量能够反映出用户在地理空间上的相遇或者相邻，基于语义轨迹的度量能够进一步得出用户之间的关系强度。借助自然语言处理思想，把用户每天的语义轨迹作为一条自然语句，用户所有时间段内的 n 条轨迹记录就生成了一篇文档，最后通过计算用户语义轨迹生成的文档之间的相似性来表示用户之间的关系强度。通过将语义标签作为分词后的结果，利用 hash 函数计算每个分词特征向量的 hash 值，然后根据词频对每个 hash 特征向量进行加权合并，最后经过计算两条语义轨迹的海明距离即可表示为语义相似度。比较了此方法和传统的主题模型以及 word2vec 模型之间的性能和效率结果。

c) 基于日常轨迹模式的用户关系度量

人们的日常活动具有很强的时间和空间的规律性，在计算用户相似度或者关系强度的研究中，还没有出现过从用户轨迹模式角度出发计算用户相似度和关系强度的研究。González 等人通过对大量用户轨迹数据的分析发现用户在日常生活中常有规律的访问相似的路径，说明用户的轨迹运动模式能反映出用户之间的相似性^[35]。从用户的历史轨迹数据中所挖掘出的频繁模式和序列模式能够反映出用户的日常轨迹运动习惯和行为规律，运动模式在现实生活中表现为用户经常行走的路径序列，是用户轨迹数据规律的抽象表示。用户的日常运动模式在一定程度上代表了用户的个人喜好、意图以及活动模式，例如用户 A 经常下午去操场跑步，B 周末经常去市区逛街等，当从一个更细的粒度甚至能够根据用户的用餐地点推测出用户的口味喜好等。本研究首先采用频繁模式和序列模式挖掘算法来探寻用户的日常轨迹运动模式，然后根据计算用户运动模式之间的相似性得出人与人之间的关系强度。

(2) 如何基于用户 WiFi 感知数据度量关系强度

随着无线网络的覆盖普及，WiFi 已经覆盖了我们日常生活中的每个区域，例如在家中、在商场中、公交车等都能够随时便捷的接入 WiFi。在使用 WiFi 进行网上活动的时候，WiFi 所记录的信息也能够从网络空间维度反映出人们的活动轨迹，同时 WiFi 的数据交也是能够有助于探究人们现实生活中的交互行为，因此通过用户的日常 WiFi 数据也能够了解到人与人之间的社交关系以及计算用户之间的关系强度。在基于用户 WiFi 感知数据度量关系强度中，本文提出了从用户 WiFi 感知上下文环境出发，通过计算 WiFi 感知环境的差异性计算用户之间的相似性进而得到用户之间的关系强度，采用了一种区别于传统意义的计算方式，通过对 WiFi 数据进行数据结构的构建和模型化表示，结合图形学的分析理论，将

WiFi 相似性分析与图分析相结合，通过计算用户之间的日常 WiFi 数据之间的相似性分析人们之间的关系。

(3) 如何基于用户蓝牙感知数据度量关系强度

随着通信技术的发展以及与局域网通信之间的融合，出现了许多支持近场通讯的标准如 NFC、蓝牙等，而蓝牙模块在手机历经了几代更迭之后依然保留在手机中。用户手机之间的近场交互情况，对研究用户之间的交互以及用户的关系有着重要的作用。在现实生活中，用户设备之间的近场交互能够暗示出用户之间的交互，如蓝牙连接传输文件，NFC 交互交换名片等。通过将收集到的用户蓝牙感知数据序列化，并且按照时间片分割进行特征向量的构建和数据结构化表示计算蓝牙交互频率以及交互时长，对用户之间的关系强度进行计算。

(4) 如何对三种独立数据源结果进行融合

前文分别概要描述了针对每种不同的非社交感知数据做如何以及采用的计算方法来得出用户之间的关系情况，在分别对每种感知数据进行计算的基础上，如何对每种计算方式内部的结果进行融合以及最后对三维数据的最终结果进行融合是一个非常重要的问题。基于 DTW 加权得到的空间距离表示用户轨迹的空间距离越小，用户之间的相似度越高关系强度越强；而基于运动模式的计算方式表明用户的轨迹运动模式越相似，说明用户之间的关系强度越强。参照集成学习的算法，对三维数据计算结果进行加权融合，作为最终的计算结果。

1.4 论文结构

本文的主要结构分为六章节，各章节的主要内容描述如下：

第一章为绪论，主要讲述了本课题的研究背景、然后介绍了智能手机感知数据同社会关系以及感知数据同人际关系强度研究所解决的问题，最后在提出了本文的研究目的，以及本文所研究的基于三种主要非社交关系数据研究的主要内容，最后总结归纳，基于这三种主要研究内容，提出本文的组织结构。

第二章主要描述了与本课题研究内容密切相关的前人研究成果，从最开始的轨迹数据处理中的滤波技术、停留点检测算法到最后的轨迹停留点聚类主要方法技术，并对相关技术进行了进一步分析；其次介绍了针对用户轨迹数据所采用的时间序列相似度计算所采用的常用距离算法以及各自的意义和优缺点；紧接着介绍了利用自然语言处理用户语义轨迹以及在自然语言处理用常用的计算文本相似性的相关技术，同时介绍了本文所采用快速计算文本相似性的方法，并且和 word2vec 进行了效率对比；最后针对剩下的非社交数据介绍了关于 WiFi 和蓝牙在挖掘用户社会关系中的应用。

第三章详细阐述多层混合模型 (RSMHD) 的整体框架，主要包含了基层的轨迹数据预处理，补全模块、轨迹语义化模块、以及三维轨迹数据计算模块和 WiFi 蓝牙数据处理计算模块。每个模块的详细描述将放在后面几章进行。

第四章主要讲述了如何对用户的空间轨迹数据进行预处理，建模等技术。首先描述了轨迹滤波模块；其次详细讲述了如何针对短时间窗口内的缺失轨迹数据进行智能补全；然后对用户轨迹进行语义化；针对用户的 WiFi 数据和蓝牙数据主要描述了如何借助图形学的思路对数据进行建模表示规整。

第五章主要详细描述了针对三种非社交感知数据进行用户的关系强度计算。首先描述了 RSMHD 的计算框架；其次分别针对各个数据模块的计算流程工作进行有针对性的说明。最后分别描述了基于各个感知数据源的计算结果和结果的融合。

第六章描述了本次研究所收集的用户数据集；最后展示了本研究的实验结果并进行分析。

第七章对本研究课题进行了总结，并提出了对未来工作的展望。

第二章 相关技术研究

上一章的内容中详细描述了本研究课题的研究背景，并且进一步分析了国内外的相关研究进展和现状，最后提出了本课题说研究的三个主要数据源和预计采用的研究方法。依据研究对象不同讲课题研究分为三个主要部分。本章将会详细的分析和阐述一些针对感知数据预处理的关键技术和方法以及论文中所涉及的快速语义轨迹计算、缺失轨迹补全等算法。

2.1 用户轨迹预处理

在用户的轨迹信息的采集过程中，因受到地形、气候、GPS 传感器和 SA 干扰误差的影响，会导致用户位置的跳跃移动等称之为“GPS 漂移现象”^[36]，GPS 的位置漂移使得用户的轨迹数据中存在大量的噪音数据，影响后续对数据的处理和分析。因此对采集到的用户轨迹数据采取滤波处理，消除轨迹中所蕴含的噪音数据。GPS 位置的采集还受到地形环境的影响，在室内无法获取 GPS 位置信息的时候会导致用户轨迹的缺失，一旦用户轨迹出现缺失，对后续的度量工作也会产生影响。因此本小节首先描述 GPS 轨迹的噪音消除算法以及我们所采用的轨迹补全算法，最后描述各种常用聚类算法。

2.1.1 滤波算法

滤波的主要目的是消除特定频率波段的噪音，用户的日常运动是连续的，所以采集到的用户 GPS 轨迹数据也应该是由连续的位置点构成的，但是由于 GPS 采集过程中受到漂移现象的影响，导致用户的 GPS 轨迹中存在位置点跳跃现象，因此要通过采用滤波算法对用户轨迹进行异常点剔除工作。接下来主要描述了一些经常使用的滤波算法：均值滤波、中值滤波、卡尔曼滤波这三种滤波算法。

a) 均值滤波：均值滤波也被称之为线性滤波，主要思想是采用结合中心点周围的数值，采取邻域平均法来表示这个邻域。数学公式表示如2.1所示，假设当前点为 p ，则设置点 p 为采样中心。将 p 前 m 个采样点和后 m 个采样点的平均值作为当前点 p 的取值。

$$g(p) = \frac{1}{2m} * \sum_{j=i-m, j \neq i}^{i+m} g(j) \quad (2.1)$$

b) 中值滤波：中值滤波是一种基于排序统计理论的提出信号噪声点的非线性的信号处理方法，其基本原理就是点 p 的取值是由其邻域内各个点取值的中值来

决定的，让数据能够更加接近于真实的取值，从而有效地减少噪声数据点。中值滤波的具体数学公式见2.2，其中函数 $median()$ 表示求中值。

$$g(p) = median(g(p - m), \dots, g(p - 1), g(p), g(p + 1), \dots, g(p + m)) \quad (2.2)$$

c) 卡尔曼滤波：卡尔曼滤波是卡尔曼于 1960 年提出的^[37]，卡尔曼滤波器由一系列递归数学公式描述。它们提供了一种高效可计算的方法来估计过程的状态，并使估计均方误差最小。卡尔曼滤波器应用广泛且功能强大：它可以估计信号的过去和当前状态，甚至能估计将来的状态，即使并不知道模型的确切性质。接下来将介绍卡尔曼理论和实用方法。

在此之前需要引入离散随机过程，卡尔曼滤波器用于估计离散时间过程的状态变量 $x \in \Re^n$ ，这个离散随机过程的方程如2.3描述：

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1} \quad (2.3)$$

定义我们所观测到的变量 $x \in \Re^m$ ，在此基础上得到卡尔曼滤波的测量方程见公式2.4：

$$z_k = Hx_k + v_k \quad (2.4)$$

其中随机的变量 w_k 和 v_k 分别表示计算过程中的激励噪声和观测到的噪声，我们假设他们二者之间相互独立，则可以得到正太分布的白色噪声如公式2.5和2.6描述：

$$p(w) \sim N(0, Q) \quad (2.5)$$

$$p(v) \sim N(0, R) \quad (2.6)$$

在实际随机过程中，激励噪声 w_k 的协方差矩阵 Q 和观测到的噪声 v_k 的协方差矩阵 R 是会随着每次迭代计算而改变的，因此为了便于推演我们假设它们都是固定的常数。当函数 u_{k-1} 等于 0 时或者噪声函数 w_{k-1} 等于 0 时，随机过程方程2.3中的 $n * n$ 阶矩阵 A 将 $k - 1$ 时刻的状态通过线性映射到 k 时刻的状态， $n * l$ 阶矩阵 B 表示变量 $u \in \Re^l$ 的增益，为了便于计算，这些变量在此都假设为常数。

我们定义 $\hat{x}_{\bar{k}} \in \Re^n$ 为在第 k 项之前的状态下计算得到的第 k 项的先验状态估计值，设 $\hat{x}_k \in \Re^n$ 表示已经得到的变量 z_k 时，第 k 步的后验状态估计值。由此根据以上描述我们可以定义出如2.7和2.8表示的先验估计误差后后验估计误差：

$$e_{\bar{k}} \equiv x_k - \hat{x}_{\bar{k}} \quad (2.7)$$

$$e_k \equiv x_k - \hat{x}_k \quad (2.8)$$

进一步得出先验估计和后验估计的协方差为：

$$P_{\bar{k}} = E[e_{\bar{k}} e_{\bar{k}}^T] \quad (2.9)$$

$$P_k = E[e_k e_k^T] \quad (2.10)$$

基于以上的理论准备，构建出卡尔曼滤波算法的数学表达式2.11，其含义为：先验估计 $\hat{x}_{\bar{k}}$ 和测量得到的变量 z_k 同其预测值 $H\hat{x}_{\bar{k}}$ 之差的线性组合共同组成了后验状态估计 \hat{x}_k

$$\hat{x}_k = \hat{x}_{\bar{k}} + K(z_k - H\hat{x}_{\bar{k}}) \quad (2.11)$$

根据2.11中表示可以进一步得到 K 的具体表达公式，其中真实变量与其预测之差 $(z_k - H\hat{x}_{\bar{k}})$ 被称之为残差，该指标有效地反映了预测值与实际值之间的不一致的程度，如果残差为零，则表示二者完全相吻合。 $n * m$ 阶矩阵 K 称之为剩余增益或者混系数，其主要作用是使得2.10中所得到的后验估计误差协方差最小，其计算步骤为：首先根据2.11代入2.8中求得 e_k ，再将 e_k 代入2.10求得期望后对 K 进行求导，并令一阶导数为零就可以求得 K 的值如2.12：

$$\begin{aligned} \hat{x}_k &= \hat{x}_{\bar{k}} + K(z_k - H\hat{x}_{\bar{k}}) \\ &= \frac{P_{\bar{k}}H^T}{HP_{\bar{k}}H^T + R} \end{aligned} \quad (2.12)$$

从2.12中能够得知观测到的数据的噪音的协方差 R 越小，残余增益 K 越大，特别地当 R 趋向零时有：

$$\lim_{P_k \rightarrow 0} K_k = H^{-1} \quad (2.13)$$

另一方面，先验估计误差的协方差越小，残余增益 K 越小，特别当 $P_{\bar{k}}$ 趋近于零时有：

$$\lim_{P_{\bar{k}} \rightarrow 0} K_k = 0 \quad (2.14)$$

综合2.12、2.13、2.14针对单个模型的测量利用上述说描述的所有公式就能够通过不断的迭代计算得出最优的估算结果。对于上述介绍的三种滤波技术而言，均值滤波在剔除数值中的随机噪音表现良好，但是不易消除脉冲误差；中值滤波能够减少偶然数据波动带来的误差影响，但是对变化快速的数据不宜使用；卡尔曼滤波最终的结果会优于前两者，但是模型较为复杂，所要求的运算时间复杂度也高于前面的滤波算法。

2.1.2 轨迹停留点检测

现实生活中用户的日常轨迹通常是一系列包含了地理坐标和时间戳的 GPS 位置点构成，每个坐标点包含了详细的经纬度、时间戳信息、海拔高度、移动速度等信息。如图2.1所示。我们可以通过一些算法检测出用户在一段轨迹运动中停留过的地方称之为停留点^[38]，本文的停留点并不是指运动速度静止的点，而是由一系列 GPS 点构成的。如图2.1中 $p_4 \sim p_8$ 构成了一个停留点 stay point 在图中由红色点表示。这个点表示用户在该区域内的停留时间超过了一个设定的时间阈值 ΔT ，相比于用户的其他轨迹位置点，这些计算得出的停留点蕴含了更重要的信息，通过语义标签甚至可以得出用户去过某家餐馆和电影院等。基于以上理论，用户的 GPS 位置轨迹可以转换为一组由停留点所组成的序列，如 $sp_1 \xrightarrow{\Delta t_1} sp_2 \xrightarrow{\Delta t_2} \dots \xrightarrow{\Delta t_{m-1}} sp_n$ ，这样由停留点所表示的序列不仅对原有数据维度进行了压缩，同时也保存了用户的重要信息。



图 2.1 用户 GPS 轨迹示例图

停留点的计算过程如2.1所示。

2.1.3 聚类算法

由于现实生活中人们可能会经常多次访问同一个空间地点，但是所计算得到的停留点却可能并不完全相同（坐标的变差，计算的误差等因素影响），因此采用传统直接比较停留点的方法不具备可行性。我们采取对停留点进行聚类的处理方法，这样地理位置非常相近的点就会被划分为同一个类别中。接下来介绍一些常用的聚类算法。

聚类是属于无监督学习中的一种重要的方法，在其他的机器学习方法中如：回归分析、朴素贝叶斯网络等的数据都是带有类别标签 γ ，也就是说在训练集中的样例就已经给出了样例的类别，而聚类数据样本却没有给出样本类别 γ 聚类的目标是根据组内元素距离最小，组间距离最大将原始数据划分为若干组如图2.2所示。本节主要介绍几种常用的聚类分析方法：K-Means 聚类算法、基于密度的 DJ-Cluster 聚类算法以及近年发表的一种改进的基于密度的聚类算法。

K-means 聚类算法是比较经典的聚类方法之一，由 J.MacQUEEN 在 1967 提出^[39]。该算法执行效率高，在大规模的数据处理聚类时被广泛使用，该算法输入

算法 2.1 停留点检测算法

已知: 用户 GPS 轨迹 $Tra, \Delta T, \Delta_{distance}$

求: 用户停留点序列 SP

```

1:  $i = 0, PointNum = |Tra|, SP = Null$ 
2: while  $i < PointNum$  do
3:    $j = i + 1$ 
4:   while  $j < PointNum$  do
5:      $distance = Dis(Tra_i, Tra_j)$ 
6:     if  $distance > \Delta_{distance}$  then
7:        $\Delta_t = Tra_j.T - Tra_i.T$ 
8:       if  $\Delta_t > \tau_{time}$  then
9:          $p.Lat = avg(Tra_k.Lat)$ 
10:         $p.Lng = avg(Tra_k.Lng)$ 
11:         $p.T = Tra_i.T(arv|lev)$ 
12:         $SP.add(p); j++;$ 
13:       end if
14:     end if
15:   end while
16:    $i = j; break$ 
17: end while
18:
19: return  $SP$ 

```

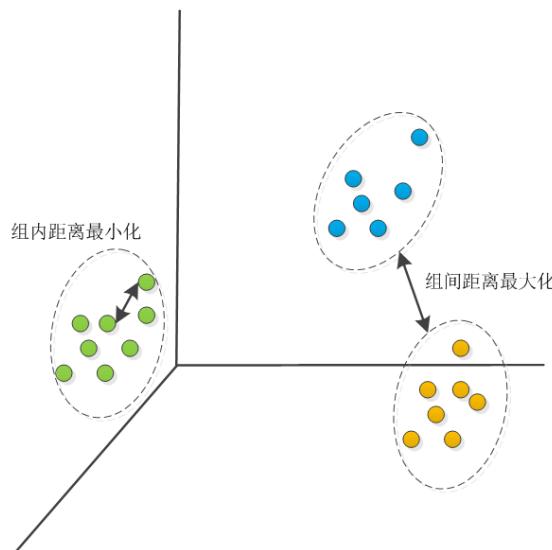


图 2.2 聚类效果示意图

k 作为最终聚类的个数，将待分类的 n 个数据分成 k 个簇，使得簇内数据具有高的相似度而簇间的数据存在较低的相似度。K-means 聚类算法的执行过程如下：

首先根据输入的参数 k 随机从原始数据中选择 k 个对象，每个初始化的对象代表了一个簇的中心；其次，剩余的每个对象计算与簇中心的距离将它们赋予距离最近的簇；第一轮结束后将重新计算每个簇的中心，这个过程不断重复知道准则函数收敛或者簇没有新的变化为止，通常采用平方误差的度量准则如2.15：

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2 \quad (2.15)$$

K-means 算法虽然简单，易于实现，但是在实际使用过程中需要用户指定 k 的取值，而 k 的取值是难以估计的，针对不同的数据事先并不可能确切的知道这些原始数据应该划分为多少个类别才正确；同时该聚类算法对异常数据非常敏感，一旦出现离群点将容易导致簇中心点出现漂移，对计算结果影响巨大。

相比于前面描述的基于距离的聚类方法，研究者于 2007 年提出了一种基于密度的聚类算法 DBSCAN^[40] 以及日后基于该算法改进的 DJ-Cluster 聚类算法^[41]，DBSCAN 算法的基本思想是扫描整个待分类的原始数据集，当扫描到数据对象 P 时，计算 P 的 Eps 邻域内所密度可达的数据对象个数是否大于等于定义的 $MinPts$ ，如果为真，则设立以 P 为核心的簇，然后尽可能的寻找与该簇密度相连的最大集合或者不断迭代查找该簇中每个数据对象的直接密度可达点，加入到该簇中。如图2.3中所示，设 $MinPts = 3$ ，从图中我们可以看出原始数据点 M, P, O 和点 R 的 Eps 邻域内说包含的点均大于等于 $MinPts$ 因此都可以把它们标记为核心；M 是 P 的直接密度可达，Q 是 M 的直接密度可达。因此可以得知：Q 是 P 的密度可达，但是 P 不是点 Q 的密度可达，点 O, R 和点 S 是密度相连的。通过实验证明，DBSCAN 会受到 Eps 和 $MinPts$ 取值的影响。

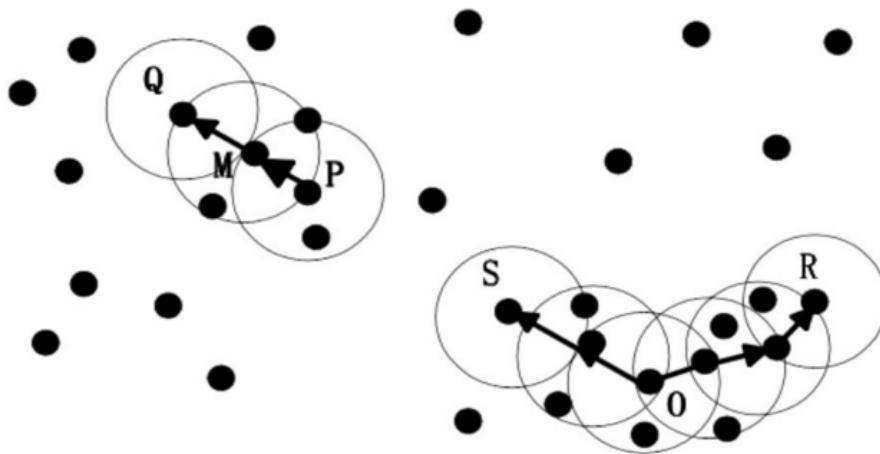


图 2.3 DBSCAN 密度直达和密度可达示意图

Rodriguez A 等人 2014 年在 Science 上发表了一种新的基于密度的聚类算法^[42]。该算法相比较于之前的聚类算法，具有对参数不敏感便于输出正确的聚

类结果的有点。该算法的主要改进想法是针对所有待聚类的组标点基于它们之间的距离，提出了两个新的标准属性： ρ 和 δ ，基于聚类中心点具有：中心点自身的密度大，即它的密度超过邻域集合点的密度同时距离其它密度大的中心点之间的距离也足够大这样的特征，其中局部密度 ρ 采用 Cut-off kernel 计算方式，公式如2.16所示，其中 p_i 表示 S 中与数据点 x_i 距离小于 d_c 的点的数量； d_c 为截断距离需要用户事先设定并且保证 $d_c > 0$ 。

$$\rho_i = \sum_{j \in I_s \setminus i} \chi(d_{ij} - d_c) \quad (2.16)$$

公式2.16中的 $\chi(x)$ 的计算方式为：

$$\chi^x = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases} \quad (2.17)$$

算法中另一个指标距离 δ 的定义为设 $\{q_i\}$ 表示 $\{p_i\}$ 的降序的下标序列，因此该序列满足：

$$\rho_{q_1} \geq \rho_{q_2} \cdots \geq \rho_{q_N} \quad (2.18)$$

则可根据公式2.19 计算出 δ :

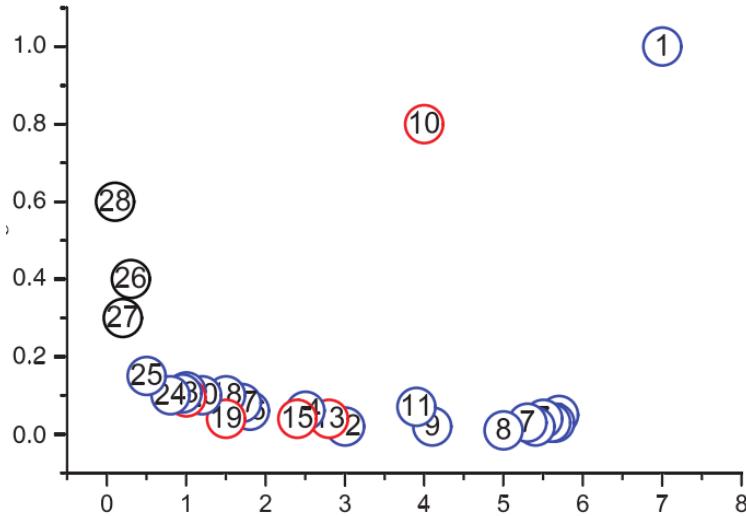
$$\chi^x = \begin{cases} \min_{q_j, j < i} \{d_{q_i q_j}\} & i \geq 2 \\ \max_{j \geq 2} \{\delta_{q_j}\} & i = 1 \end{cases} \quad (2.19)$$

该算法原理示意见图2.4和图2.5, 从图2.5可以看出，编号为 1 和 编号为 10 的坐标点具有较大的 ρ 和 δ 取值，因此在原始数据中我们将这两个点设置为簇的中心，而在图2.4中这两个坐标点恰好是分类簇的中心点；同时还出现了编号 26-28 三个“离群点”它们的特点是 δ 值很大而 ρ 取值很小。

总的来说，上述三种聚类算法通过精心调整参数都能取得非常好的聚类效果。不同的算法拥有不同的优缺点，在第四章我们将通过实验来展示各种算法在不同参数下的聚类结果。

2.2 时间序列相似度度量方法

我们阅读了大量社会心理学相关的论文学籍，从中证实得到在现实生活中关系亲密的两个用户会更加倾向于一起进行面对面的交流、共同进行社交活动等，因此通过对手机传感器数据的处理分析能够从中挖掘出人们现实生活中的关系强度。文献 [37] 立足于空间距离，提出了在空间距离上非常接近的人们在现实生活中就越可能发生面对面的交互，该文献通过调研一个小区的住户发现人们在现实

图 2.4 关于 decision graph 的示例^[42]图 2.5 聚类中各点的 ρ 和 δ 取值^[42]

生活中越是接近，就越容易成为朋友。文献 [16][17] 进一步通过研究用户的轨迹数据发现由于在空间距离中接近的用户在现实生活中更可能产生交互行为，也就是说拥有相似日常生活轨迹的用户更可能产生交互活动。在现实生活中亦是如此，我们很容易和同自己经常出没同一个地点、行走在同一条轨迹上的人发展友谊关系。

用户的轨迹序列由带有时间戳的位置数据构成，位置数据可能是 GPS 也可能是基站号。因此我们可以将轨迹序列看作时间序列，从而使用一些时间序列相似度量模型来度量轨迹的相似度，下面依次描述编辑距离和 DTW 这两种相似度量方法以及序列熵值的计算方法。前面章节描述到用户的轨迹是由 GPS 位置点所构成的，包含了 (,,) 等详细信息，因此我们将用户的轨迹看作一条由带有时间

截所构成的坐标点序列，采用序列相似度计算方式来处理轨迹之间的相似度。接下来详细描述两种常用的序列相似度计算方法。

2.2.1 轨迹中心距离

传统的计算序列相似度所采用的方法如：编辑距离、汉明距离和夹角余弦等有其优缺点，如果贸然用来计算用户轨迹序列的相似性可能不太符合实际问题的需求，在此基础上，Hechen Liu 等学者提出了一种基于用户地理轨迹的相似度度量方法^[43]，根据用户的轨迹形状以及时间片内轨迹的中心点之间的距离来作为度量用户轨迹之间相似度的标准。如图2.6中(a)图所示，现有两条用户轨迹 $Tra_1 = < a_1, a_2, a_3, a_4, a_5 >$ 和 $Tra_2 = < b_1, b_2, b_3 >$ 从观测来看两条轨迹是非常相似的，但是从整条轨迹对比来看，就很难评判两条轨迹是否相似了，因为 Tra_1 中有一处急转弯点 a_3 ，称为转折点。因此算法首先检测出轨迹中的转折点如图2.6中(b)再针对每段轨迹中中心点 (Center of mass) 来计算轨迹之间的相似性。算法中的 Center of mass 是轨迹 Tra 的质量中心，计算公式如2.20所示，轨迹计算结果示意图如图2.7所示，最后再计算轨迹之间的距离，采用余弦相似度来衡量轨迹之间的相似性。

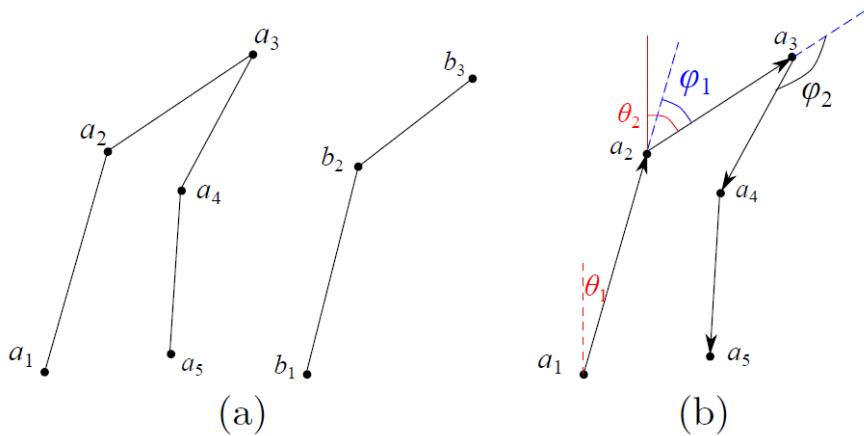
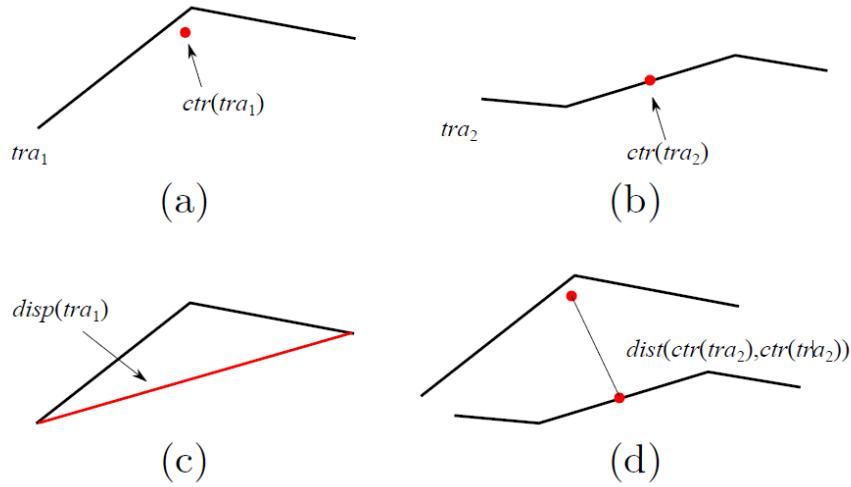


图 2.6 轨迹相似计算示例^[43]

$$ctr(Tra) = (\bar{x}, \bar{y}) = \left(\frac{\int xf(x)dx}{\int f(x)dx}, \frac{\int yf(y)dy}{\int f(y)dy} \right) \quad (2.20)$$

采用该算法来计算用户轨迹之间的相似度同直接计算轨迹之间的欧氏距离相比较，能够将轨迹的形状考虑在内，这样能够结合避免直接计算轨迹之间的距离所因为离群点造成了计算结果的巨大偏差。

图 2.7 基于 Center of mass 轨迹计算示例^[43]

2.2.2 Dynamic Time Warping

DTW(Dynamic Time Warping) 算法最初是由 Itakura 提出的的一种新型的计算距离的方法^[34]，在最初的一段时间是被应用于语音识别领域，在语音识别中即使是同一个词语，由不同的人说出口但是他们的语速、语气等不同造成了单词音频的差别。DTW 正由于其计算距离的特殊处理使其能够胜任这一工作。DTW 算法采用动态时间规整的思想，将需要比较的两个序列在横轴时间维度上进行压缩、拉伸等操作，使得两条序列具有相同的长度具有更有效的匹配度，同时也消除了传统基于欧式距离计算带来的弊端。

设待计算距离的两个序列，序列 Q 和序列 C （两条序列的表示见公式2.21、2.22）。如果两条序列的长度相同，则计算变得很简单。但是如果 $m \neq n$ ，就需要拉伸变形两条序列，使得它们的长度能够尽量对齐同时保留原有序列的信息，算法首先构造一个 $n * m$ 矩阵 d ，其中 $d[i, j]$ 表示 q_i 和 c_j 之间的距离。采用动态规划的方法来找出序列 Q 和 C 之间的最佳匹配，其中转移方程为找出一条路径使得所有总和的距离 $\gamma[i, j]$ 最小化，具体公式描述见公式2.23。路径及矩阵可视化见图2.8。其中：A) 为两个待计算比较的序列；B) 为通过 DTW 动态规划求解后两条序列计算距离时所对应的点；C) 将原有序列所对应的位置进行拉伸展开后更加直观的 DTW 匹配计算结果示意图。

$$Q = q_1, q_2, \dots, q_i, \dots, q_n \quad (2.21)$$

$$C = c_1, c_2, \dots, c_j, \dots, c_m \quad (2.22)$$

$$\gamma[i, j] = d[i, j] + \min\{\gamma[i - 1, j - 1], \gamma[i - 1, j], \gamma[i, j - 1]\} \quad (2.23)$$

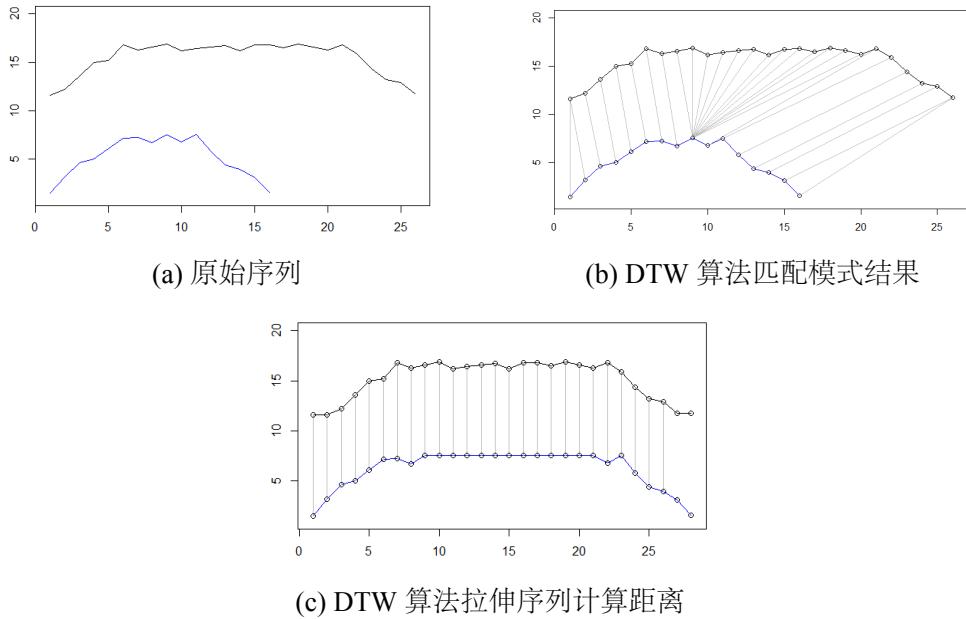


图 2.8 DTW 算法匹配序列结果示意图

通过对 DTW 算法的原理分析可以得知，序列 Q 和序列 C 的长度越长，则最终计算结果得到的距离就会越大。因此，后者研究中，采用了多种归一化的加权处理方法对结果进行加权处理 [44]，获得最优的 DTW 计算结果。

2.3 自然语言处理方法

相对于用户的空间轨迹度量，用户在日常生活中的语义轨迹蕴含了更丰富的上下文信息同时人与人之间在语义轨迹层次上特别是好友之间可能表现出惊人的一致性，如经常去某些地方，在一天中总是先从某个地点出发，再经过某些地点，最后在某个咖啡店相遇等等。基于轨迹的用户模式的交互能够反映出用户在某个空间中的相遇；而基于用户语义轨迹的分析能够在一定程度上展现出用户在社会活动上的相似性，在绪论部分已经详细从社会心理学的研究中描述了人们在现实生活中相遇频率能够反映出用户之间的关系强度，文献 [45] 研究发现人们会更加喜欢那些在兴趣、价值观、人格上和自己相似的人。因此通过用户的语义轨迹来在更高的层次上对用户的行为进行相似度计算，进而推测出用户的关系强度。

通过将用户的语义位置序集合同自然语言处理中的文档进行对比，可以借鉴自然语言处理的算法来处理用户的语义轨迹序列。用户每天的活动轨迹通过语义标签标注后，得到的是一个现实生活中语义轨迹的序列集合，如将一个用户的活动轨迹表示为 $\langle \dots \rangle$ ，通过结合自然语言文档相似性计算的思路，把用户每

天所计算得到的语义轨迹作为一条原始自然语句，用户在 n 天内所采集的所有轨迹记录就生成了一篇文档，最后基于自然语言处理方法通过计算用户语义轨迹生成的文档之间的相似性来表示用户之间的关系强度。

在以往的研究中，通常使用编辑距离来衡量语句的相似性。在用户轨迹相似分析上，文献 [46] 借助了 LDA 主题模型来计算用户语义轨迹之间的相似性，后来的研究在其基础上对计算方法进行了改进^[47]。其算法改进在于基于 LDA 主题模型，将用户的所有语义轨迹看作一篇文档来训练出若干个主题。在计算用户轨迹相似性的时候，将用户输入到训练好的 LDA 模型中，然后采用余弦相似度分析比较二者输出的主题分布集合之间的相似度，作为这两个用户之间的关系强度。

word2vec 是 Google 在 2013 年开源的一种词向量计算算法^[48]，word2vec 借鉴深度学习方法，将文本通过训练的得到文本的 K 维向量表示，通过词与词之间的距离来计算它们之间的相似度。

2.3.1 LDA 主题模型

2003 年文献 [49] 提出了 LDA(Latent Dirichlet Allocation) 模型对自然语言进行建模，可以用来识别文档语料中潜在的主题信息。整个计算模型采用了词袋的计算方法，计算出文档的向量表示，每一篇文档计算出一部分主题的概率分布，而每一个主题内又可以表示为很多词语的一个概率分布，LDA 的训练过程为：遍历每一个文档，在主题分布袋中随机抽取一个主题；然后在被随机抽取到的主题中再随机抽取一个单词；最后重复上述步骤直到遍历完文档中的所有词语。整个生成的过程可以用图2.9表示：对于每一份文档与设定的 T 个主题之间的概率分布对应 θ ，每个主题与词袋中的 V 个词语之间的概率分布 ϕ ，其中 θ 和 ϕ 分别有一个参数 α 和 β 的狄利克雷的先验分布，这样对于任意一份文档 d 中的每一词语，我们从与文档相对应的概率分布 θ 中选取一个主题 z ，随即在根据与主题 z 对应的概率分布 ϕ 中选取一个词语 w ，最后重复上述过程 N_d 次，就能够产生文档 d ，公式2.24为 LDA 的核心公式。

$$p(w|d) = p(w|t) * p(t|d) \quad (2.24)$$

2.3.2 word2vec

word2vec 是 Google 在 2013 年开源的一种词向量计算算法^[48]，word2vec 借鉴深度学习方法，将文本通过训练的得到文本的 K 维向量表示，也就是说把特征转换到了 K 维空间进行表示，通过词与词之间的距离来计算它们之间的相似度。word2vec 采用了三层的神经网络结构即为输入层 - 隐含层 - 输出层构成，其

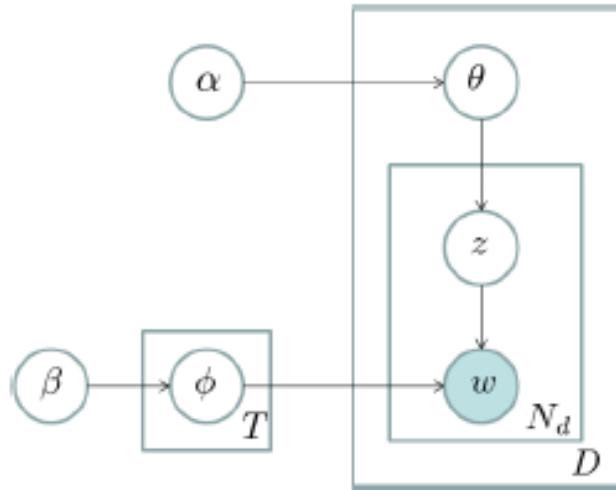


图 2.9 LDA 的图模型表示

主要方法是借助哈夫曼编码针对相似词频的词语构建出相似的隐藏层激活函数，算法采用了层次化的 Log-Bilinear 模型，其中一种是基于 CBOW(Continuous Bag-of-Words Model) 的计算模型，在 CBOW 模型中，根据上下文信息可以预测下一个词语 w_t ，其公式如2.25所示，CBOW 的计算模型如图2.10所示。

$$p(w_t|context) = p(w_t|w_{t-k}, w_{t-k+1}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k}) \quad (2.25)$$

现在的 CBOW 计算采用层次的 Softmax 算法，每个单词 w_i 都可以有一条从根节点出发被唯一访问到的路径，这条路径就形成了词语的编码。

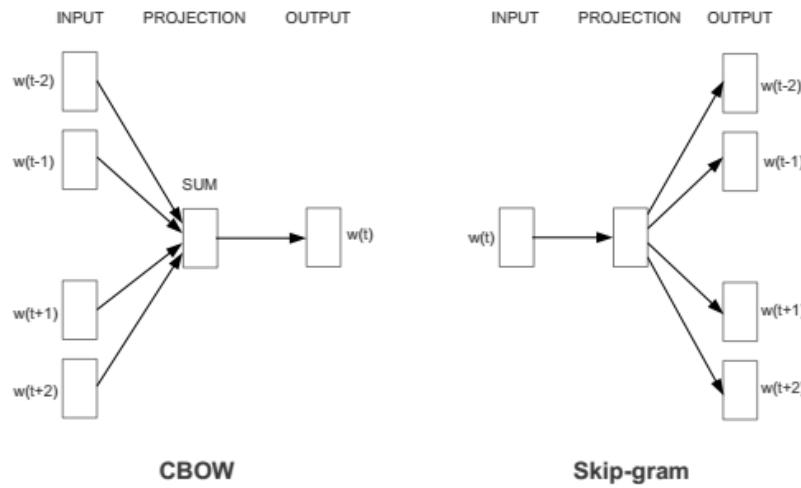


图 2.10 word2vec 神经网络结构图[48]

在图2.11中，第一层也称之为输入层，它的输入是词向量；中间的层次为隐含层，得到的的输入是输入层若干个词向量的向量累加结果；第三层则是由二叉

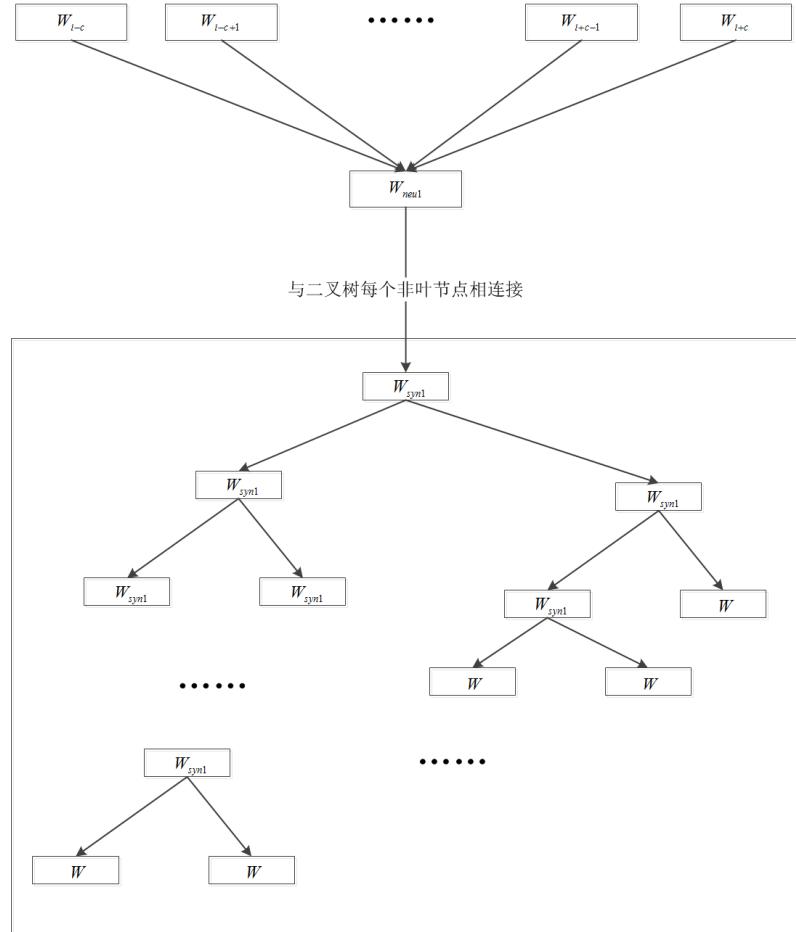


图 2.11 word2vec 神经网络结构图

树所构成的霍夫曼树所组成的输出层，其中每个非叶节点是一个计算后的向量但是不同于输入层的词向量，这里计算后的向量不代表某个词语，而是表示一个类别的词语，同时这棵霍夫曼树的所有叶子节点包含了输入词袋中的所有词。

对于词袋 BG 中的单词 w ，在图2.11中一定能够找到有且仅有一条从 Root 节点到叶子节点 w 的路径。路径中的每一次分支都是一个概率问题，因此得到2.25中的数学表达式，将其中的 w_i 用词向量替换，得到该三层神经网络的目标公式如2.26， Num^w 表示从 Root 节点到词向量节点 w 路径中包含的节点个数， w_j 表示该路径上的第 j 个词， v_w 表示词 w 对应的词向量， σ 表示 SIGMOID 函数，通过求解目标函数的最大值，就可以得到每个词所对应的向量。

$$L = \sum_{w \in C} \sum_{j=2}^{Num^w} \{(1 - w_j) \log[\sigma(v_w^T \theta_{j-1}^w)] + w_j \log[1 - \sigma(v_w^T \theta_{j-1}^w)]\} \quad (2.26)$$

2.3.3 快速 Hash 算法

基于 word2vec 的计算方法针对文档的分词结果，再根据神经网络模型计算得到每个词的向量表示，但是这样无疑增加了整个计算过程的时间复杂度，本文研究目的是能够为众多用户提供一种计算关系强度的框架，因此如果采用传统的 Hash 算法进行处理，则能够减少程序的运行时间，来自于 GoogleMoses Charikar 针对海量文档快速计算相似计算提出了一种局部敏感 hash 算法^[50]，其核心内容恰巧和 word2vec 相反：希望将数据进行降维处理，将原始的高维词向量通过一系列运算映射到低维的词向量，最后来计算相似性。

以上介绍了几种方法在自然语言处理中都得到了广泛的应用，因此综合考虑使用最后一种方法来作为计算方法，在后面还会详细讲解如何使用 hash 来度量用户之间的关系强度，以及在不同处理方法之间进行横向和纵向的比较。

2.4 WiFi 和蓝牙数据的度量方法

WiFi 和蓝牙作为一种主动对外界感知和探索所得到的感知信息，已经证实能够被用来分析用户之间的社交关系。文献 [27] 通过收集到用户的 WiFi 数据，然后将用户的 WiFi 数据同现实语义位置进行关联，得到了 WiFi 的语义向量，这样就能够得到基于 WiFi 的用户活动轨迹序列，提出了一种新的模型 AMVD(average minimum vector distance) 来计算人与人之间的距离，其计算公式如2.27所示，其中 a_i 和 b_j 表示用户 A 和用户 B 之间的联合向量， $d(a_i, b_j)$ 表示计算二者向量之间的曼哈顿距离。

$$AMVD(A, B) = \frac{1}{|A|} \sum_{\forall a_i \in A} \arg \min_{\forall b_j \in B} d(a_i, b_j) \quad (2.27)$$

蓝牙数据和 GPS 轨迹数据都能够推测出用户之间的交互行为，蓝牙作为一种近距离的无线通讯手段，收到波长影响使得通讯距离只能是 5-10 米内，因此蓝牙感知数据能够进一步描述了用户之间在现实生活中的物理接触。前人的研究中，有根据蓝牙信息来简单计算出用户在现实空间中的面对面交互次数，从而推测出用户的社会关系情况^[28, 29]。在本文中，我们使用了新的数据结构来表示用户的 WiFi、蓝牙感知数据，同时直接计算其感知上下文环境信息的相似度作为用户之间环境的相似度的参照。

2.5 小结

本章详细的从本课题研究的三个数据源出发对不同数据源的研究方法、现状进行了详细的分析和讨论。2.1 节讨论了在用户原始 GPS 轨迹数据处理中要解决的三个主要问题：剔除异常点、停留点检测以及空间轨迹的聚类。2.2 节则是介绍

了计算用户轨迹序列相似度的常用方法，并以此结果作为用户的关系强度计算结果。2.3 节从自然语言处理的角度介绍了常用的文档相似性检测计算方法，为计算用户语义轨迹的相似度提供了可行性算法。2.4 节从 WiFi 和蓝牙数据处理计算角度出发，介绍了前人研究的计算方法理论，为本研究中所采取的基于上下文环境的计算方式作出铺垫。

第三章 RSMHD 用户关系强度计算框架模型

上一章详细介绍了和本研究数据源相关的处理技术与方法，接下来这一章将主要介绍 RSMHD 的多维据源多维度的关系度量模型的整体框架结构。

3.1 RSMHD 模型框架描述

本研究的主要内容是通过采集到的用户感知数据（GPS 轨迹信息、WiFi 数据、蓝牙数据）来开展用户之间的关系强度度量工作。因此本文希望能够提供一种基于此类非社交的数据建立的通用的用户关系强度计算框架。非社交数据主要是指不是来自于用户直接社交活动中所收集到的数据，在本研究中主要包含了用户轨迹数据、用户 WiFi 感知数据和用户蓝牙感知数据。因此本课题基于以上三种不同的感知数据源提出了能够统一处理多种数据源的计算框架，其整体概要结构图如图3.1所示，图中描述整体架构由以下几部分组成：上下文感知收集模块、基于用户日常轨迹的关系强度计算模块、基于用户 WiFi 感知数据的关系强度计算模块和基于用户蓝牙的关系强度计算模块。用户的日常语义轨迹是有连续的 GPS 位置点所组成的轨迹的集合，在用户的轨迹信息的采集过程中，因受到地形、气候、GPS 传感器的干扰误差的影响，会出现 GPS 漂移现象，GPS 的位置漂移使得用户的轨迹数据中存在大量的噪音数据，影响后续对数据的处理和分析。因此对采集到的用户轨迹数据采取滤波处理，消除轨迹中所蕴含的噪音数据。GPS 位置的采集还受到地形环境的影响，在室内无法获取 GPS 位置信息的时候会导致用户轨迹的缺失，造成计算结果产生巨大偏差。第四章部分将重点针对上述问题展开 GPS 轨迹的处理研究工作，以获得好的信息效果。对于剩余的 WiFi 和蓝牙感知数据来讲，主要的工作是针对二者数据源的特点以及后续的算法输入对原始数据进行提取和结构化建模表示存储。第四章的其余部分将会针对 WiFi 和蓝牙感知数据的预处理和结构化表示进行详细的描述。

3.1.1 基于轨迹数据的关系强度计算模块概述

在本研究中所提出的 RSMHD 计算框架模型中，第二层需要基于用户的日常语义轨迹分别通过计算地理轨迹相似性、用户的语义轨迹相似性以及日常轨迹运动模式三个层次关系强度融合为轨迹强度结算结果，其详细框架图如图3.2所示。在这个模块中，通过计算用户日常地理轨迹的相似性推测出用户之间的关系强度；针对用户活动产生的日常语义轨迹，RSMHD 模型采利用自然语言处理的思想，通过快速编码变换计算出用户基于语义轨迹的关系强度；第三层中，RSMHD 进一步挖掘出用户轨迹中更高一层的上下文信息 (High Context) 得出用户日常轨

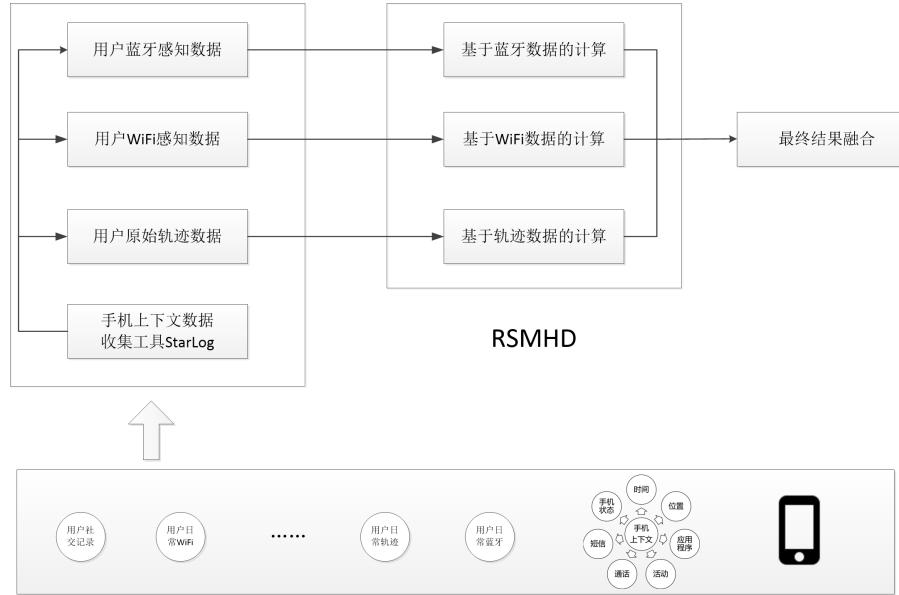


图 3.1 RSMHD 用户关系强度计算模型框架图

迹模式如：用户 A 经常喜欢到校外某地用餐、用户 B 经常下午在操场跑步等这些代表用户日常轨迹活动模式的信息，计算用户之间基于运动模式的关系强度。

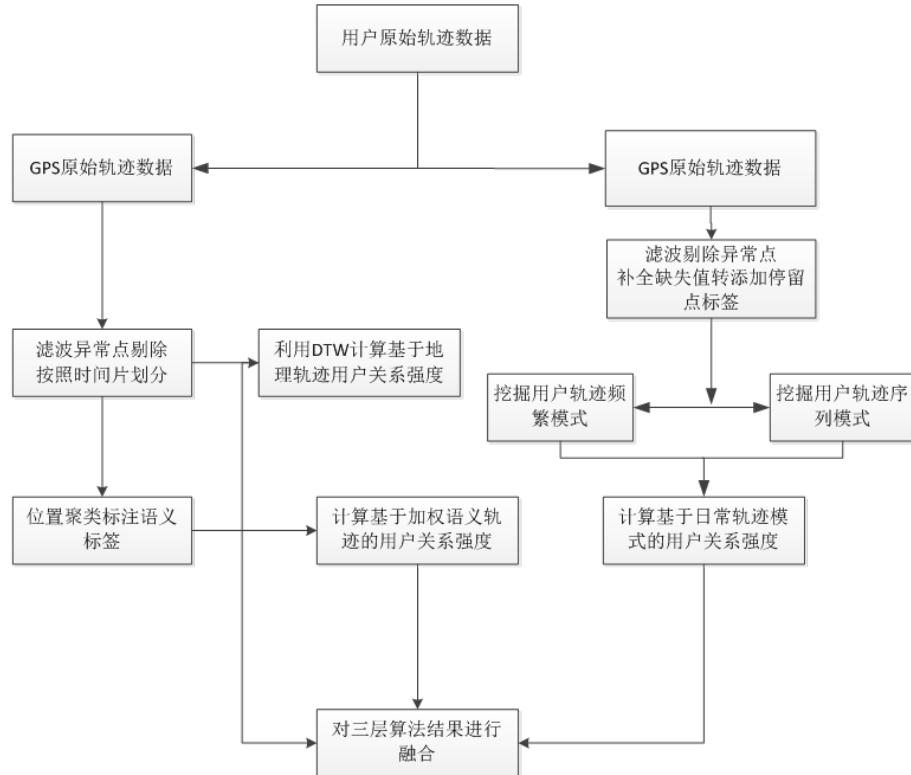


图 3.2 RSMHD 中基于用户轨迹计算模块详细框架图

3.1.2 基于 WiFi 和蓝牙数据的关系强度计算模块概述

在 RSMHD 计算框架模型中我们针对 WiFi 和蓝牙原始数据的特点以及现实生活中 WiFi 和蓝牙上下文环境的特征，采用了区别于原有基于 WiFi 和蓝牙计算社交关系的方法，整体的详细计算框架如图3.3所示。在 WiFi 和蓝牙数据处理计算框架中，分别针对底层手机的上下文感知信息进行数据信息的提取的规整，从复杂冗余的信息中萃取出关键的富有价值的信息；然后结合 WiFi 和蓝牙各自的上下文环境特征信息，将整理后的数据用图的数据结构进行结构化表示，使得抽象表示后的数据结构更加符合现实中的含义；最后基于结构化后的感知数据进行关系强度计算，将结果输出到下一层。

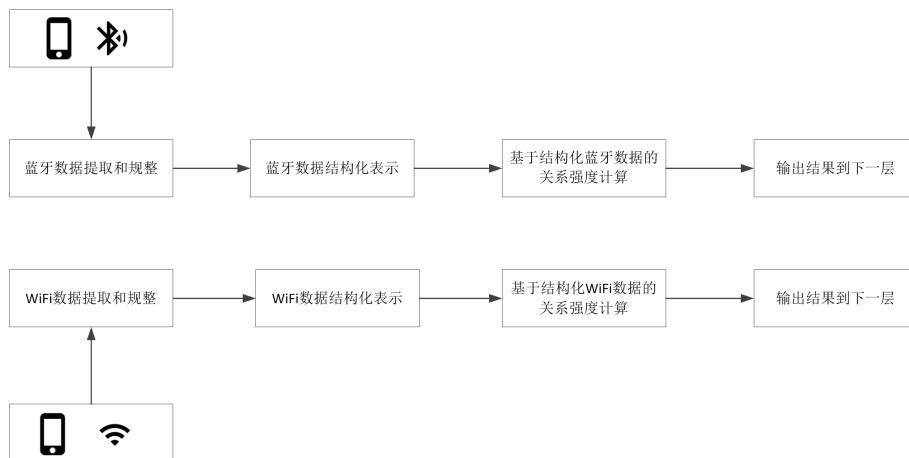


图 3.3 RSMHD 中基于 WiFi 和蓝牙计算模块详细框架图

3.2 RSMHD 模型计算流程概述

整个信息采集是采用我们自己编写的用户手机上下文信息收集工具 StarLog 来记录保存用户手机的上下文感知数据如：GPS 位置信息、WiFi 感知信息、蓝牙感知信息、脱敏的通话记录、短信记录、APP 使用记录等。

针对用户的 GPS 位置数据我们首先对轨迹数据进行预处理，即采用滤波算法检测出用户日常轨迹中的漂移点减少轨迹中的噪音；然后根据停留点检测算法识别出用户轨迹中的停留点，针对缺失的轨迹点进行预测补全；第一层次中针对用户日常空间轨迹相似性计算时，可用将每天的轨迹数据划分为若干时间片的数据，然后采用 DTW 加权算法对时间片内的空间轨迹相似度进行计算并按照一天的轨迹进行处理。对于基于用户语义轨迹推倒关系强度这层我们首先将得到的停留点进行聚类分析，得到聚类后的停留点集合因此对于每一个簇都是同一个语义位置，然后将聚类 GPS 转换为现实社会中的语义标签，在计算用户语义轨迹的相似性时，采用快速 hash 算法将用户分片后的语义轨迹序列作为输入，得到用户之

间每天的语义轨迹相似度，以及最终的语义轨迹相似度并将结果作为用户自己的关系强度。在上一层的基础上，我们针对得到的用户语义估计进行挖掘，寻找出用户的日常运动模式，然后计算用户运动模式之间的相似度并作为关系强度之一进行输出，综合三层的计算结果得到用户基于轨迹数据计算得到的关系强度。

其次针对用户的 WiFi 和蓝牙数据，首先从原始数据中提取出重要的信息，并将 WiFi 和蓝牙数据进行结构化存储处理，采用图的存储结构，将 WiFi 和蓝牙数据还原为现实生活中的存在环境，然后针对用户切片时间内每刻的感知环境进行相似度计算，并将一天中的计算结果进行汇总进一步得到所有时间段内的用户相似度，最后将结果作为基于两种数据源推测出的用户关系强度。

最后，在基于前面三种数据源的用户关系强度计算结果的基础上，采用集成学习的思想对计算结果进行融合处理，得到最终的用户关系强度计算结果。

第四章 RSMHD 框架的相关技术研究

上一章节主要是对本课题研究的 RSMHD 框架进行了详细的描述，本章接下来将针对每一个模块中的具体面对的问题以及采取的方法进行详细描述，包括噪音点的剔除、如何语义轨迹、如何计算轨迹相似性以及如何对 WiFi 和蓝牙数据进行结构化处理等。

4.1 GPS 轨迹处理计算技术

用户的日常 GPS 轨迹数据包含了用户的生活工作以及娱乐等丰富的上下文信息，如图4.28所示，我们基于用户的日常轨迹分别从地理位置相似性、语义轨迹相似性以及轨迹模式相似性出发度量人与人之间的关系强度。模块的输入是由 StarLog 收集的 GPS 感知数据，通过对初始数据的解析提取得到原始的 GPS 位置点构成的用户轨迹序列。预处理阶段主要是通过滤波算法对原始轨迹序列进行异常点检测，并通过轨迹预测补全缺失的部分 GPS 轨迹序列；第二层中将第一层得到的停留点赋予语义得到用户的语义轨迹序列，采用自然语义处理思想计算基于语义轨迹的相似性；再往上一层，从用户的历史轨迹数据中挖掘出的频繁模式和序列模式能够反映出用户的日常轨迹运动习惯和行为规律，运动模式在现实生活中表现为用户经常行走的路径序列，是用户轨迹数据规律的抽象表示，最后计算用户之间的关系强度。

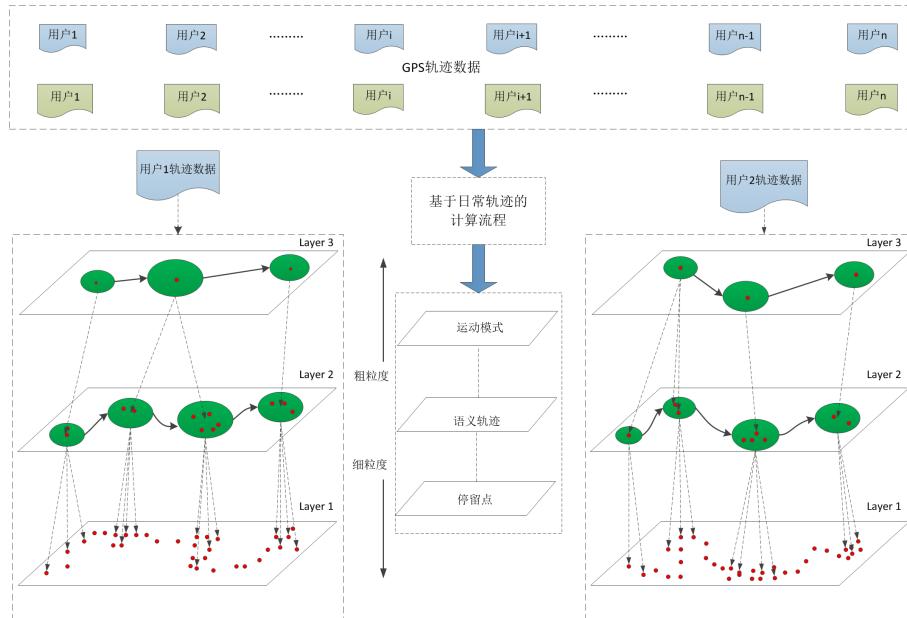


图 4.1 基于多层次多粒度轨迹的用户关系强度计算

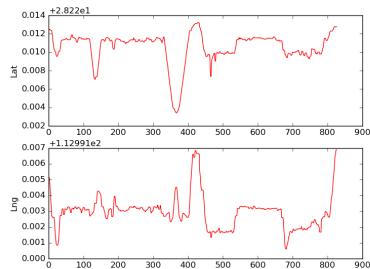
4.2 GPS 轨迹预处理与语义化

本节主要讲述如何对 GPS 轨迹数据进行清洗规整，以及得到用户的语义轨迹进行结构化存储。

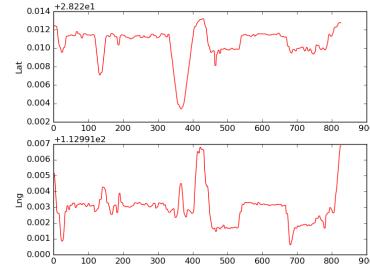
4.2.1 剔除轨迹中的异常点

前文已经提到，由于在获取 GPS 位置信息的时候会受到位置漂移的影响，导致在获取用户实际位置的时候可能产生采样的误差甚至跳跃，为了使得最终计算得到的关系强度结果更为准确我们需要对 GPS 轨迹数据进行滤波分析。在第二章中描述了常用的三种滤波方法，在本章中将会针对三种滤波算法进行最后结果展示并根据结果分析最终采用此滤波算法的原因。

首先接下来使用我们自己开发的用户感知数据收集软件 StarLog，来分析观察各种滤波算法对用户轨迹中异常点检测剔除的效果，如图如图4.2、4.3所示。

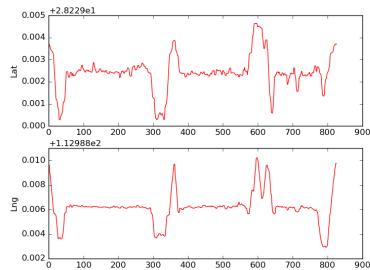


(a) 原始轨迹数据

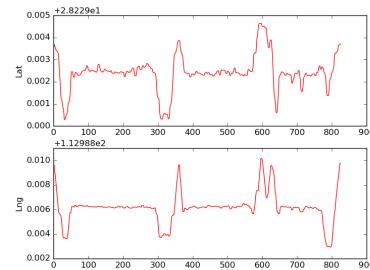


(b) 中值滤波后的轨迹数据

图 4.2 轨迹中值滤波实验结果 1



(a) 原始轨迹数据

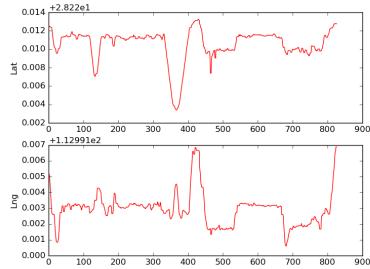


(b) 中值滤波后的轨迹数据

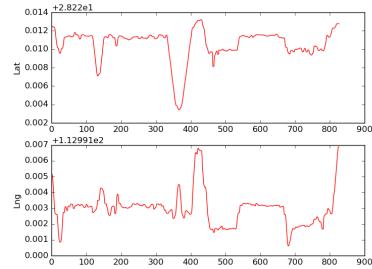
图 4.3 轨迹中值滤波实验结果 2

从实验结果中我们可以观察到虽然中值滤波能够过滤掉其中少部分的位置漂移点，但是针对于一些明显的轨漂移点却没有能够很有效的识别过滤。

接下来我们再使用均值滤波算法对用户轨迹进行分析，观察均值滤波算法对用户轨迹中异常点的检测情况，部分实验结果见图图4.4、4.5。

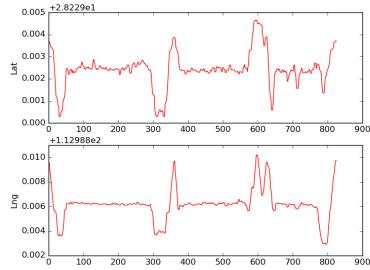


(a) 原始轨迹数据

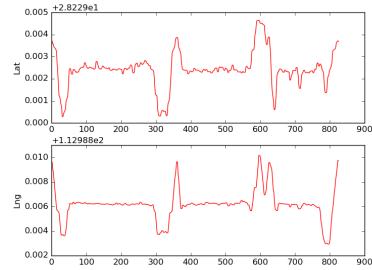


(b) 均值滤波后的轨迹数据

图 4.4 轨迹均值滤波实验结果 1



(a) 原始轨迹数据



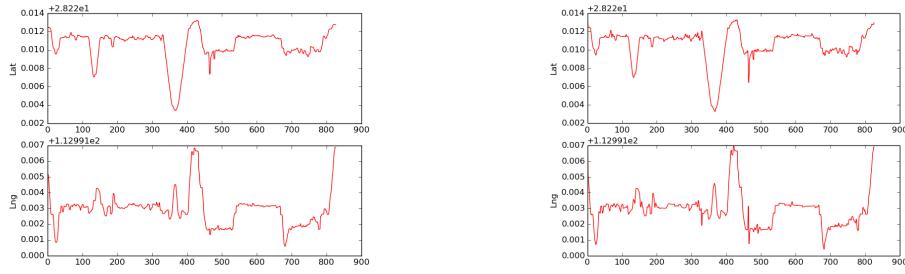
(b) 均值滤波后的轨迹数据

图 4.5 轨迹均值滤波实验结果 2

通过对比观察滤波结果图可以发现，相对于中值滤波，均值滤波能够较好的平滑用户的轨迹变化曲线，剔除 GPS 漂移点。但是，同样无法很有效的处理漂移偏差大的位置点。

通过采取卡尔曼滤波算法得到的用户轨迹如图4.6、4.7，根据观察图中滤波后的用户轨迹数据，我们可以发现虽然图形变得平滑了许多，但是却使得原有的轨迹信息收到了模糊，难以有效的将两个用户轨迹进行相似度计算。

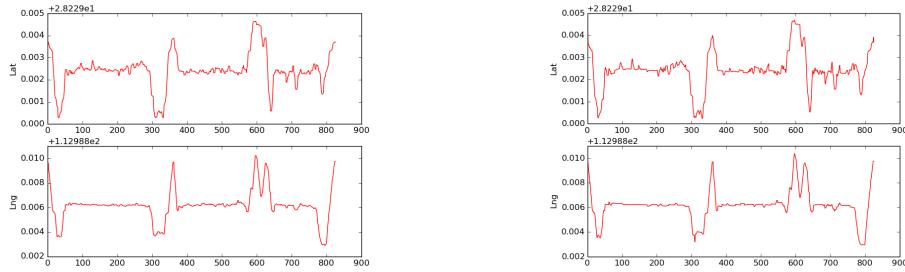
在本研究中，我们采用了一种基于速度的分段卡尔曼滤波方法。考虑到在一个时间片内，如果当前位置点的速度和它之前的位置点的速度绝对差大于 Δv 时 (Δv 作为一个未知的参数，需要我们在实际使用中给出) 采用这样的方法将原有用户轨迹切分为 n 段然后针对每一段轨迹采用卡尔曼滤波算法，最终的部分轨迹滤波结果见图4.8、4.9，可见经过按照速度分段后使用卡尔曼滤波能够比较好的过滤掉漂移点。



(a) 原始轨迹数据

(b) 卡尔曼滤波后的轨迹数据

图 4.6 卡尔曼滤波实验结果 1



(a) 原始轨迹数据

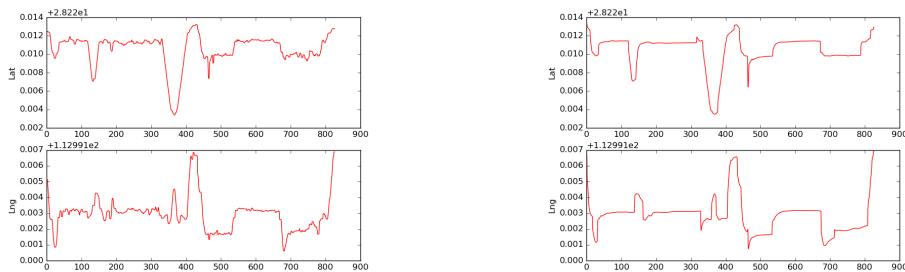
(b) 卡尔曼滤波后的轨迹数据

图 4.7 卡尔曼滤波实验结果 2

本小节主要针对前文中提及的三种常用的滤波算法进行了实验结果展示和效果分析，初步得出了使用卡尔曼滤波更加可行的判断结果。

4.2.2 轨迹中停留点检测

在实际生活中用户的轨迹是由一系列 GPS 点构成的，剔除其中的噪音点和用户在路上的点，能够从中挖掘出进一步信息的位置信息即为轨迹中的停留点，通常停留点并不是指用户轨迹中速度为零的点，而是由一组 GPS 点构成区域，一个



(a) 原始轨迹数据

(b) 分段卡尔曼滤波数据

图 4.8 分段卡尔曼滤波轨迹结果 1

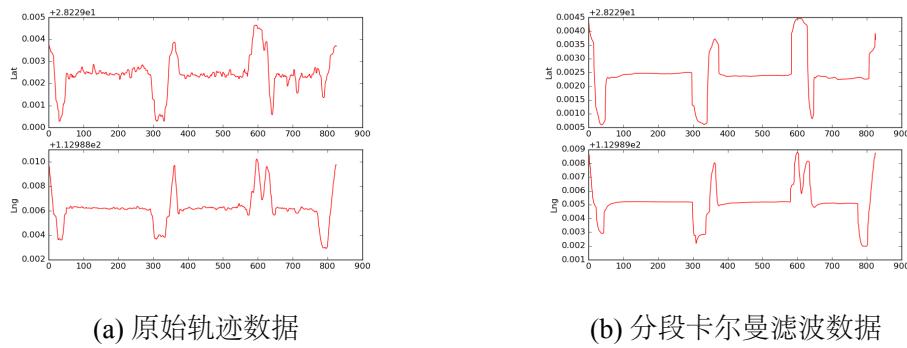


图 4.9 分段卡尔曼滤波轨迹结果 2

停留点通常对应现实生活中富有具体意义的点，能够更好的反映出用户之间轨迹的相似程度。如图4.10所示，在现实生活中的停留点可以是一个具体的建筑名称，也可以是代表一个固定的区域，甚至可以是一个具有特殊意义的点。

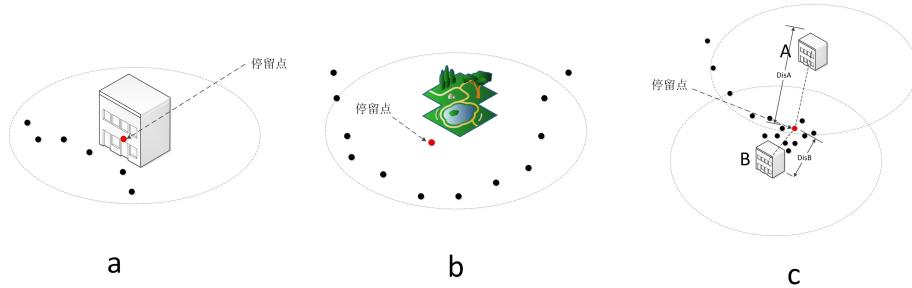


图 4.10 现实生活中的停留点示意图

通过对停留点的识别，能够让我们更加深入的了解和认识用户的日常轨迹，同时从一个更加细粒度的层次来分析用户之间的轨迹相似度，停留点检测实验结果如图4.11、4.12所示，图中所得到的每一个停留点都具有丰富的现实意义，能够有效地表示用户访问的地点、位置。

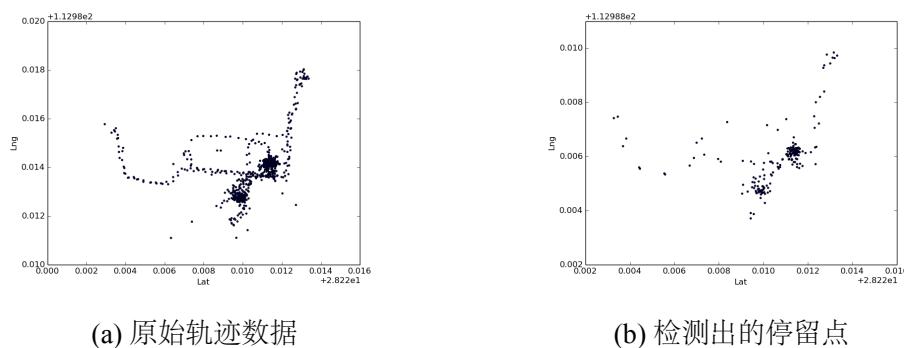


图 4.11 停留点检测实验结果 1

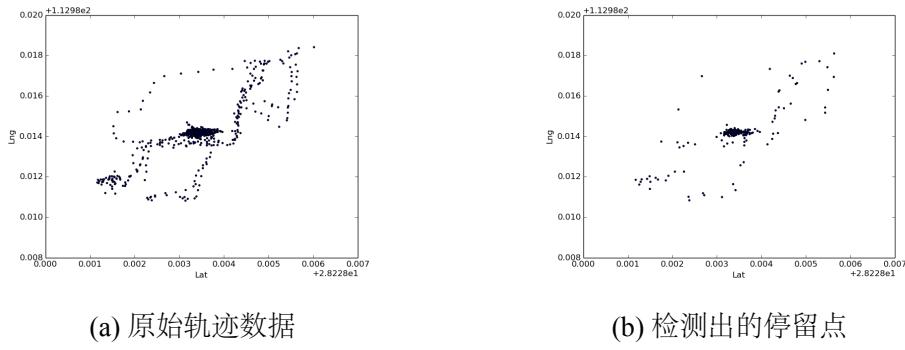


图 4.12 停留点检测实验结果 2

4.2.3 用户轨迹停留点聚类

本节针对上面得到的用户轨迹中的停留点，采用不同的聚类算法所得到的不同的聚类结果进行描述。在前面的章节中我们详细介绍了聚类中常用的三种算法 K-means 聚类、DJ-Cluster 的密度聚类和最新发表的改进的密度聚类算法，接下来我们将在实验中对这三种聚类算法进行比较。

首先我们对用户轨迹数据采用 K-means 算法进行聚类实验，但是 K-means 聚类算法具有以下的明显的缺点：首先是 K-means 聚类结果依赖于参数 k 的初始化，其中 k 是指聚类个数。 k 个数的确定往往是靠经验值设定；其次是 K-means 算法初始化时聚类中心的选择，因为该算法采用的随机初始化聚类中心点，中心点选择不同会导致聚类的结果也出现差异甚至影响聚类的时间复杂度，使得聚类结果容易得到局部最优解而非全局最优解；最后的一个缺点是 K-means 聚类算法对原始数据中的噪音点和离群点非常敏感，聚类结果很容易受到离群点的影响从而导致簇的偏移。

在本研究中因为是针对用户的空间轨迹进行聚类分析，而现实生活中用户的停留点主要是以相隔距离比较大的建筑等，因此在采用 K-means 聚类过程中，对噪音点的影响基本可以忽略不计，最主要的就是考虑不同的参数 k 对最终聚类结果的影响，图 4.13、4.15 主要展示不同的参数 k 取值对最终轨迹聚类结果的影响。图 4.14、4.16 在地图上标记了在不同参数 k 取值情况下的簇中心使得能够有更加直观的观察。

通过对轨迹数据的聚类分析我们可以得知，如果在聚类前我们能够确切的知道用户的轨迹数据应该划分的簇的个数，那么 K-means 聚类算法将会得出非常合理的结果，我们在观察图 4.14、4.16 时发现，当簇的个数越接近符合用户访问位置个数的时候，那么聚类点将更能够反映出真实的情况，在图 4.18 中，(a) 所对应的聚类得到的簇包含了实际上没有正确的将用户访问过的位置区分开，因为这两个访问的位置距离相隔太近，导致 K-means 算法将这两个簇合并为一个簇，但是

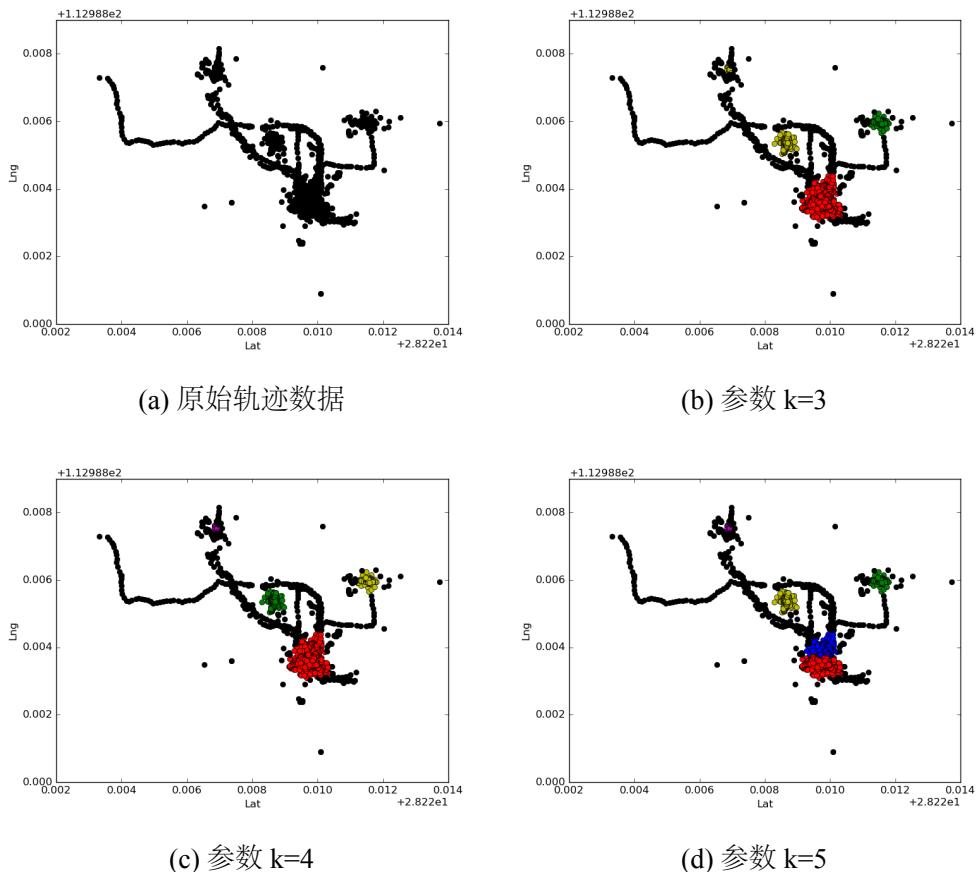


图 4.13 基于 K-means 的轨迹聚类实验结果 1



图 4.14 基于 K-means 的轨迹聚类实验结果地图展示 1

如果我们将聚类个数设置为 5 就能够成功的识别出这两个位置。在另一实验结果图 4.20 中也出现了类似的情况。但是根据图中实验情况分析发现对于一部分在路上的点被划分到了聚类簇中，同时有些应该被单独聚类成一个独立的簇的点被划

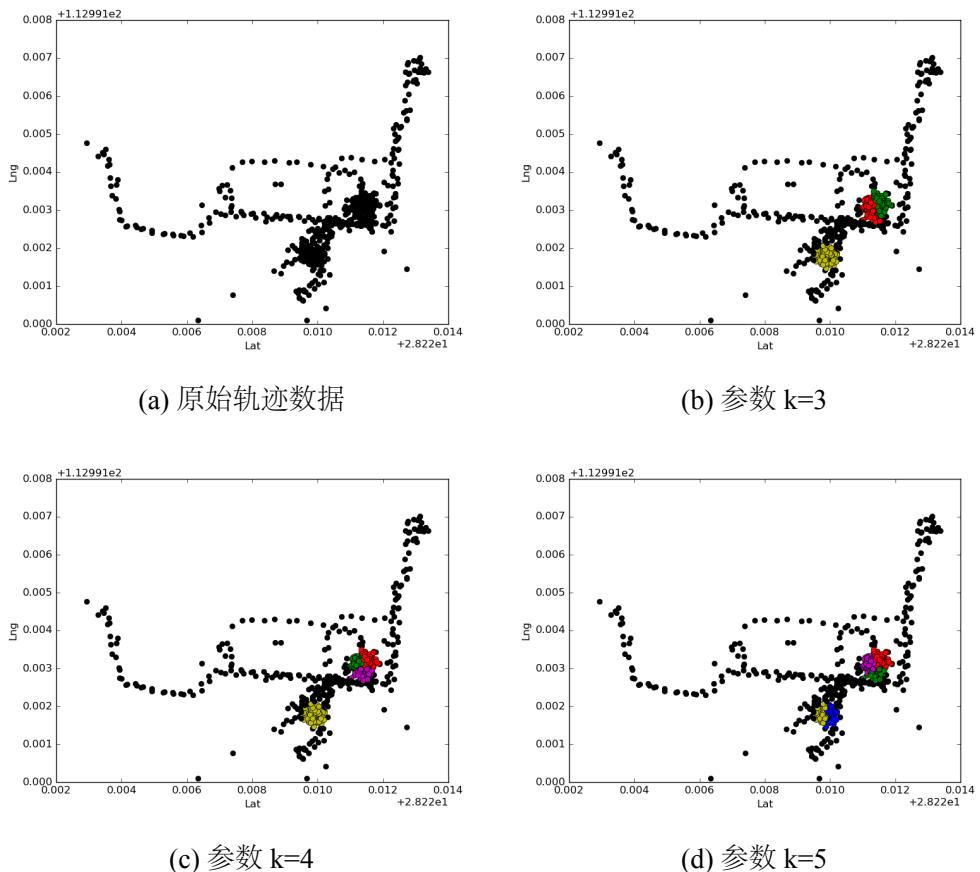


图 4.15 基于 K-means 的轨迹聚类实验结果 2



图 4.16 基于 K-means 的轨迹聚类实验结果地图展示 2

分到了在路上的点中，这些都是因为我们无法正确的给出符合当前轨迹访问点个数的聚类个数所造成的。

接下来将会基于 DJ-Cluster 算法来对轨迹数据进行聚类并对结果进行分析。DJ-Cluster 算法是基于密度聚类算法 DBSCAN 而改进的一种密度聚类算法，通过利用密度相连的原理可以发现任意形状的簇，但是算法存在有一个局限性：算法的性能和准确性依赖于参数 Eps （邻域半径）和 $MinPts$ （最少点数量），如果邻域半径的取值过大那么所有的点将会划分为同一个簇，因此在实际的使用过程中，需要不断地对参数进行调整才能得到比较理想的结果，如图4.17、4.19结果所示，实际地图中结果展示如图4.18、4.20所示。

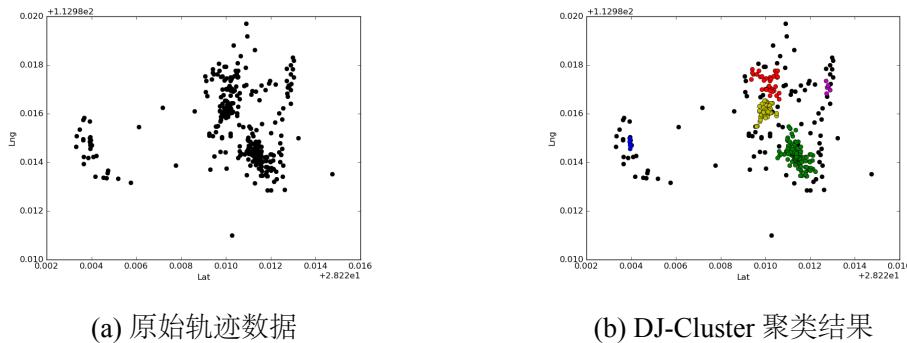


图 4.17 基于 DJ-Cluster 的轨迹聚类实验结果 1



图 4.18 基于 DJ-Cluster 的轨迹聚类实验结果地图展示 1

通过观察结果地图展示和原始数据对比能够发现，相比于 K-means 算法 DJ-Cluster 能够得到比较正确的结果，但是结果依赖于对前面所描述的参数的不断调整。从图4.18中 (a) 中描绘的用户位置点集合 (b) 中得到的聚类结果进行分析，右上角是宿舍区域，因为当前点比较少，未能够正确的将这个区域聚类成一个单

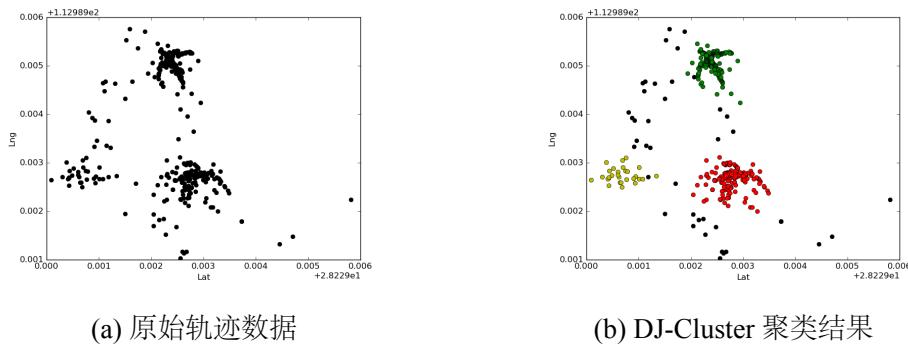


图 4.19 基于 DJ-Cluster 的轨迹聚类实验结果 2



图 4.20 基于 DJ-Cluster 的轨迹聚类实验结果地图展示 2

独的簇进行标记。但是从图4.20中发现 DJ-Cluster 基本得到了非常好的聚类结果。受限于 DJ-Cluster 算法对参数的敏感性，接下来将介绍一种新的聚类算法，该算法将大大降低对参数和数据的依赖性。

接下来进一步展示在 Science 上发表的聚类算法，该算法通过结合现实理论，假设聚类的中心实质上是有局部密度低的点所环绕的，并且这点和另外密度高的点距离都比较大。这样通过计算得到每个点的当前密度值后根据划分的边界阈值进行密度点的划分，找出每个块中密度最大的点并将其周围密度小于该点密度的点加入到当前簇中得到最终的聚类结果。该算法因为考虑当将点密度排序后进行处理，保持了点密度顺序的相对稳定的同时减小了半径阈值带来的影响。聚类算法的具体的实验结果见图4.21、4.23，将聚类结果通过地图展示如图4.22、4.24所示，从图4.21中可以观察到聚类结果是比较

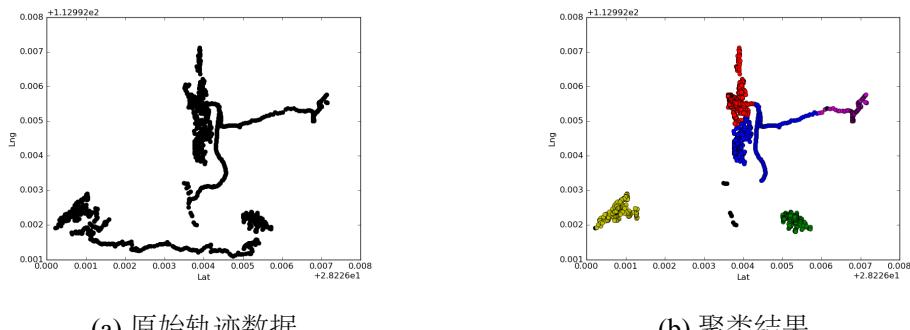


图 4.21 轨迹聚类实验结果展示 1



图 4.22 轨迹聚类实验结果地图表示 1

通过分析图4.21和图4.21中的聚类结果发现该聚类算法能够比较正确的将用户的轨迹数据进行聚类同时也不依赖于太敏感的参数输入。在图4.21中用户正确访问的地方应该是五个，虽然(b)中的聚类结果也得出了五个簇中心位置，但是未能将寝室这一簇识别出来而是将在体育馆周围的点标注为聚类簇。虽然在最终

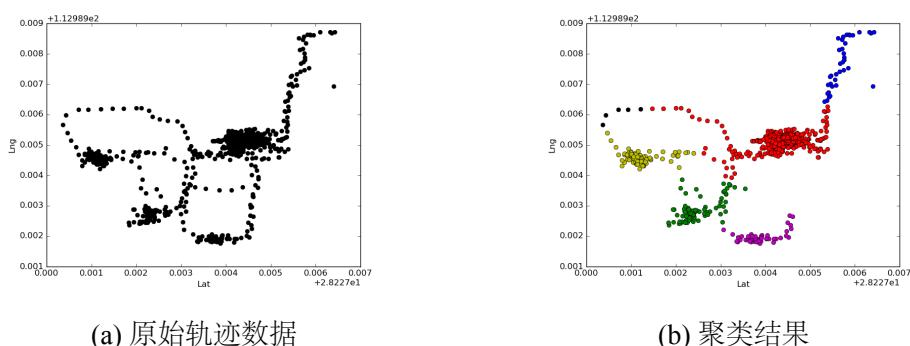


图 4.23 轨迹聚类实验结果展示 2

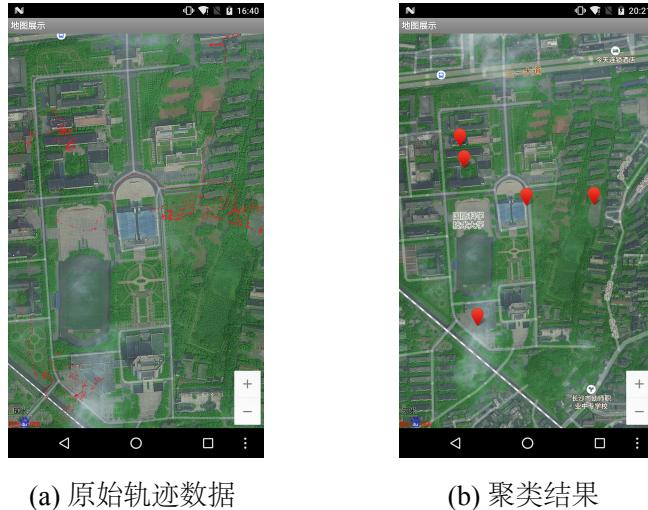


图 4.24 轨迹聚类实验结果地图表示 2

的结果上有的会出现一点误差，但是能够去除对数据和参数的敏感性减少对实验的干预，在本研究中显得更加难能可贵和适用。

本小节主要讨论了前面一章节中介绍的聚类算法在本研究中对用户轨迹聚类的效果以及每种聚类算法对参数的依赖以及影响，然后分析每一种聚类算法各自的优缺点，K-means 聚类结果虽然比较好但是需要预先设定聚类个数，最明显的问题是我们收集的用户轨迹数据中无法预先知晓用户轨迹所应该聚类的个数，导致结果出现偏差；DJ-Cluster 聚类算法同样存在类似的问题。本课题采用针对小数据进行聚类，一方面是因为数据量小比较容易发现较小的聚类簇，能够识别用户更加细粒度的访问位置，如果将用户多天的轨迹数据进行聚类，有可能将会导致较大量簇无法识别。另外采用小数据量时间切片进行分析聚类能够保证当前用户的某个聚类没有识别而下一段数据分析识别出来，形成效果叠加也能够保证最终结果的相对稳定，下一节将简要讨论如何对用户的轨迹点进行语义化。

4.2.4 对用户轨迹的语义位置添加语义标签

在上面的小节中主要讲述了针对用户原始 GPS 轨迹数据进行滤波处理、轨迹停留点挖掘以及针对用户轨迹进行聚类。但是仅仅得到用户的空间轨迹信息是不够的，为此我们需要将用户轨迹中有意义的点添加上语义标签如图4.25所示，为用户轨迹中的停留点加上语义标签，得能有助于我们了解用户去了什么地方，这些地方属于什么类型等。语义轨迹有助于我们进一步分析用户之间的相似性，因此本小节将介绍如何添加语义轨迹。

现有的大多数工作是基于反地理编码技术来将空间轨迹语义化，将用户的位置转换为具体的道路门牌号等（如德雅路 138 号），然而有时候道路地址无法有效

地表示用户位置的现实意义；另外的方法采用了机器学习的方法，首选从用户的轨迹中提取用户的位置点以及位置点的类型，采用关系马尔科夫网模型，利用网中的节点表示用户的访问点，基于条件随机场对用户的访问点进行推测得到用户的访问点的语义标签，但是采用机器学习效率较低，同时最终结果还受限于数据训练集，不适合普适计算的要求。因此我们采用了以语义标签数据库为主，反地理编码为辅的语义化方法。语义标签数据库中覆盖了本研究对象校园内部及其周边大部分语义位置点，假设得到了用户一天的轨迹数据并且进行了停留点切分和聚类得到了 100 个位置，而我们的语义标签数据库中只能够查询到其中的 70 个位置的语义标签，因此剩下的 30 个位置我们将采用反地理编码进行语义化，并将相对于的语义标签存入到语义数据库中避免二次调用反地理编码接口。如图4.25所示，我们采用了三层了轨迹语义化处理模型，在最底层为用户的原始轨迹数据，中间层为上一小节中我们识别出的用户停留点或者聚类簇，最上层为用户的访问地点语义标签。

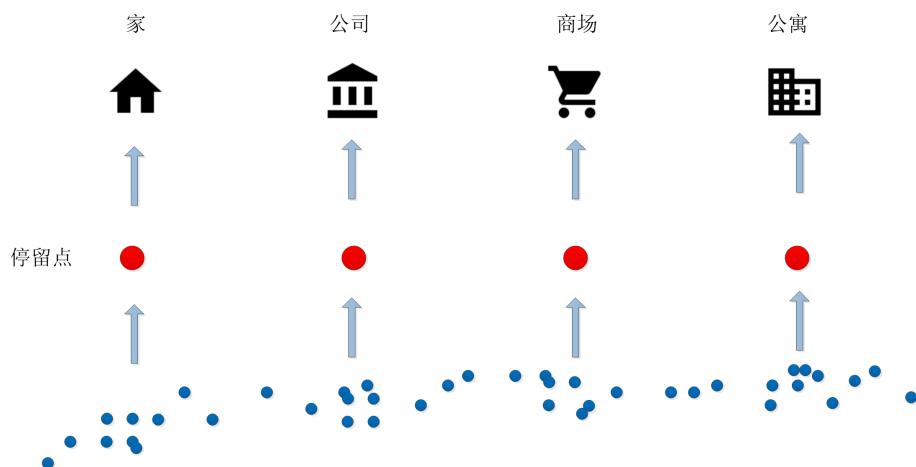


图 4.25 位置语义标签示意图

在语义化的过程中，我们首先构建了一个符合当前范围内的语义标签数据库，里面包含了学校内部各个位置语义点以及学校周边大部分的位置语义信息，在为用户轨迹添加语义标签时，首先查找语义标签数据库内是否有相匹配的位置点。具体的匹配过程如图4.26所示，图中首先我们需要人为设定一个距离阈值，用来衡量当前位置点与语义标签位置点之间的关系。图4.26中 a 所示：当停留点或簇中心点距离语义标签库中有且仅有一个位置点的距离小于设定的阈值，则把当前语义标签赋予给当前位置；若如图4.26中 b 所示，当半径阈值内不止一个符合要求的语义标签点时，我们选择和当前位置点距离最近的语义标签作为当前位置点的标签；当语义标签数据库中无法查询到与当前位置相匹配的标签点时，将会调用反地理编码，获取到当前位置的语义标签，结果示意图如图4.27所示，在

图4.27(a) 和 (b) 中, 因为这两处地理位置的语义标签都已经存在于语义标签数据库中, 所以查询时就返回数据库中的标签, 而 (c) 中的位置因为暂时没有保存在标签数据库中, 因此调用百度地图 API 将得到的中心点坐标转换为百度地图数据库中的标注地址并将结果返回和保存。

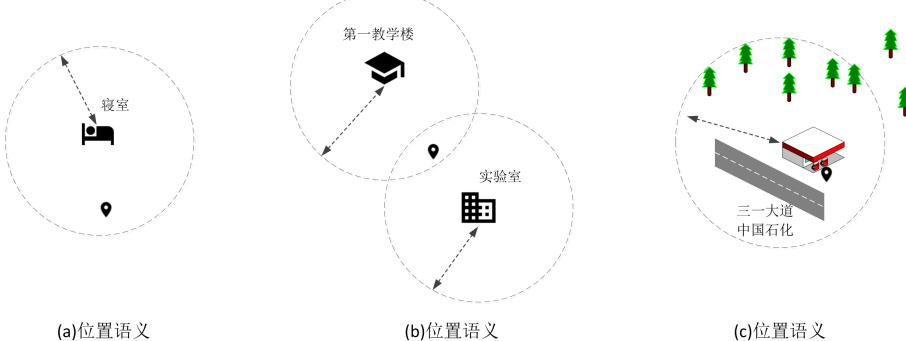


图 4.26 语义标签识别示意图



图 4.27 语义标签结果示意图

在用户轨迹语文化的过程中还存在以下这样的问题: 第一不同用户或者同一个用户访问同一个语义位置 (如食堂、学生宿舍) 等的时候, 由于手机设备不同和 GPS 数据采样的误差, 导致用户当前停留位置的聚类中心并不一定和数据库中记载的完全相同, 因此我们在查询语义数据库的时候, 需要设定一个中心点距离阈值比如说 15 米, 如果当前中心点位置和其中某一个标签的位置距离小于 15 米, 就将该语义标签赋予当前中心点; 如果同时落入了多个语义点的半径范围内, 则将距离最近的语义中心点的标签赋予给当前中心点。

另外一个问题是由于第一个问题引申而来即半径阈值的取值为多少才能给既准确的识别中心点语义标签又能使得临近语义点的距离足够大，这样才能有效的对轨迹停留点进行语义标注，在实验过程中发现基于小样本数据在小范围内（如整个校园内及其周围）的语义标签化得到的数据相对于可靠，可能是由于在小范围内的 GIS 语义标签数据库比较完备，能够有效地识别出用户大部分的停留点。但是如果将数据进一步增加，将活动范围扩大，那么这种方法的可靠性和得到的语义标签的准确性也将受到影响。如果能够得到一个大区域内的 GIS 信息库或者是人工标注好这样的语义标签库，那么就能够应对这样的问题，但是这样显得缺乏可行性，如何针对在大区域内的语义轨迹化也是我们未来工作的研究之一。

4.3 用户 GPS 轨迹数据结构化表示

通过前面章节的停留点检测算法以及实践，能够得到用户轨迹中存在的停留点，接下来我们要构建一个用户表示用户轨迹的模型，针对每个用户利用轨迹表示模型来描述和存储该用户所有时间内的轨迹信息。为了便于在下面章节中对基于空间轨迹的用户关系强度度量，我们采用了一种基于三维的层次化的 GPS 轨迹表示模型来描述用户在纵向时间维度内以及横向空间维度中的活动。如图4.28所示，在此抽象表示结构中，其中第一维表示用户维度，第二维的每一行表示用户在当前日期内所有的轨迹集合，第三维中的每一列表示用户在当前日期内，按照时间分割后的时间片中所有的停留点集合，最后得到的结果中 $Gtra_i$ 表示用户 U_i 所有的轨迹集合。

根据前面章节中获取的用户位置信息多为一个经纬度点的集合或一个几何形状，没有包含任何可以直接识别的信息。通过轨迹语义化，我们基于用户的 GPS 轨迹得到了用户的语义位置轨迹。不同于用户地理轨迹的抽象表示方法，用户的语义轨迹表示更加具有难度和挑战性，根据前面的描述得知：现实生活中同一个用户多次访问同一语义位置所产生的空间位置停留点可能不相同 (GPS 定位存在偏移)，直接对停留点进行语义化并不可行，因此我们在节4.2.3中针对部分停留点进行聚类得到用户的轨迹中停留点的聚类簇，然后用当前的结果替换原有轨迹中对应的停留点。如图4.29所示，我们首先从用户 GPS 轨迹数据中抽象出表示用户空间地理位置的空间轨迹数据结构；然后将聚类后得到的停留点簇按照节4.2.4对停留点添加语义标签得到用户的语义轨迹。现实生活中，人们日常轨迹中的位置移动常常表现为从一个位置移动到另一个位置（如早上从寝室步行到食堂再去图书馆等），在图4.29中我们借鉴这样的想法将用户每天的 GPS 运动轨迹向量序列化后的结果赋值为用户的轨迹向量，再根据时间戳信息结合语义标签构建用户的语义轨迹时间向量，得到用户在一段时间片内的语义轨迹向量，最后将每天的轨

迹向量合并得到用户的语义轨迹序列描述，通过对这种语义轨迹的结构化表示，可以有效地将空间轨迹转换为用户的日常语义轨迹，便于下一步计算用户之间的关系强度。

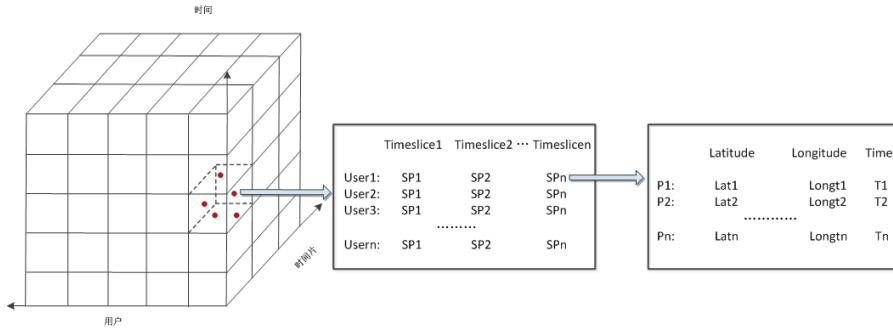


图 4.28 用户轨迹的抽象表示

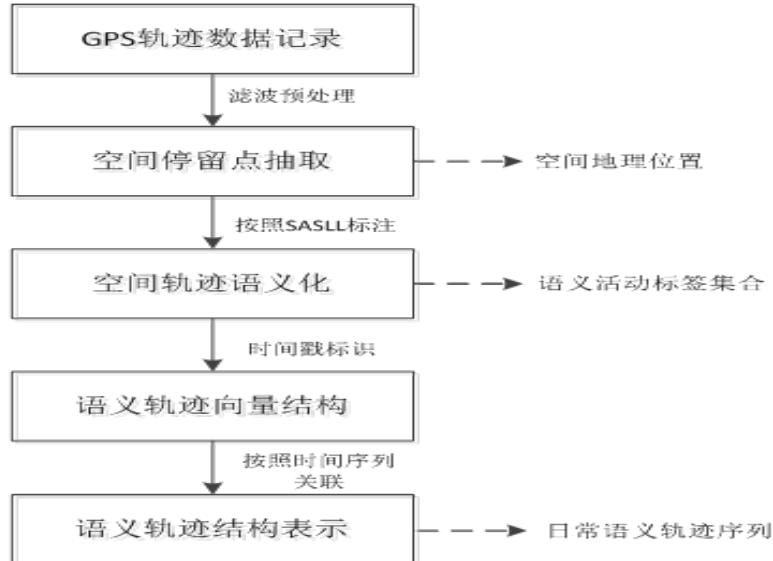


图 4.29 用户语义轨迹描述模型

在获得用户的语义轨迹之后我们将用户的语义轨迹序列按照时间片进行划分，得到形如 $\{Sloc_1, Sloc_2, \dots, Sloc_n\}$ 结构的用户语义轨迹序列集合，最终得到的语义轨迹结构如图4.30中的示例描述。

从用户的历史轨迹数据中所挖掘出的频繁模式和序列模式能够反映出用户的日常轨迹运动习惯和行为规律，运动模式在现实生活中表现为用户经常行走的路径序列，是用户轨迹数据规律的抽象表示。用户的日常运动模式在一定程度上代表了用户的个人喜好、意图以及活动模式，例如用户 A 经常下午去操场跑步，B 周末经常去市区逛街等，当从一个更细的粒度甚至能够根据用户的用餐地点推测

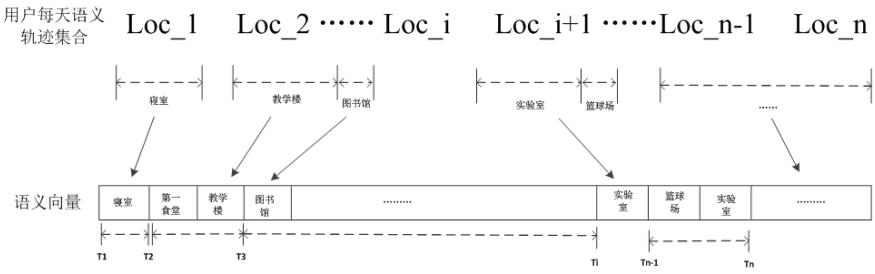


图 4.30 语义轨迹示意图

出用户的口味喜好等。本文首先采用频繁模式和序列模式挖掘算法来探寻用户的日常轨迹运动模式，最后得到用户的运动模式矩阵，矩阵中的行

$$TraFP_{i,j,k} = \{(Sloc_1), (Sloc_2, \dots, Sloc_m) \dots\}$$

表示用户 u_i 在时间周期 k 内的日常语义轨迹中所挖掘出的频繁模式的第 j 项纪录；而

$$TraSQ_{i,j,k} = \{< Sloc_1 >, < Sloc_2, \dots, Sloc_n > \dots\}$$

表示用户 u_i 在时间周期 k 内的日常语义轨迹中所挖掘出的系列模式的第 j 项纪录，这些都将会在下一章中进行详细的描写如何利用用户轨迹运动模式来计算用户之间的关系强度。

4.4 WiFi 感知数据处理计算技术

在日常生活中 WiFi 的感知距离因为受到发射器功率以及周围环境的影响，导致正常 WiFi 的覆盖范围大约在 20 米左右。当用户处在不同的上下文环境中时其周围能够被 StarLog 所感知到的 WiFi 环境也是不同的，不仅如此，对于一个很小的范围内（如教室、图书馆、办公室等）的 WiFi 上下文环境基本是极度相似的甚至相同的，也就是说能够探明检测到的 WiFi 列表应该是相同的，因此我们根据 WiFi 感知数据的这一特性来计算用户之间的关系强度。根据图3.3中的处理流程描述，首先我们将原始 WiFi 感知数据进行处理，从中提取出有用的数据信息（如 WiFi 接入名称、WiFi 强度、WiFi 接入点的 MAC 地址、扫描时间等）；其次在提取出的数据的基础之上，我们根据探测的时间戳信息，将 WiFi 数据序列化。如图4.31所示，我们采用矩阵的方式来表示所有用户在整个数据收集期间的 WiFi 序列集合，其中每一行表示实验数据收集的时间维度从第一天到第 N 天；每一列代表一个独立的用户，其中 $[i, j]$ 表示用户 u_j 在第 i 天的 WiFi 向量数据。每一条 WiFi 向量数据都是由许多切片时间内的 WiFi 上下文（WiFi context）所构成的。

每一次所感知的 WiFi 上下文环境信息是不仅包含了 WiFi 连接名，还包含了接入点 MAC 地址、信号强度、设备类型等复杂的信息，如图4.32所示，WiFi-

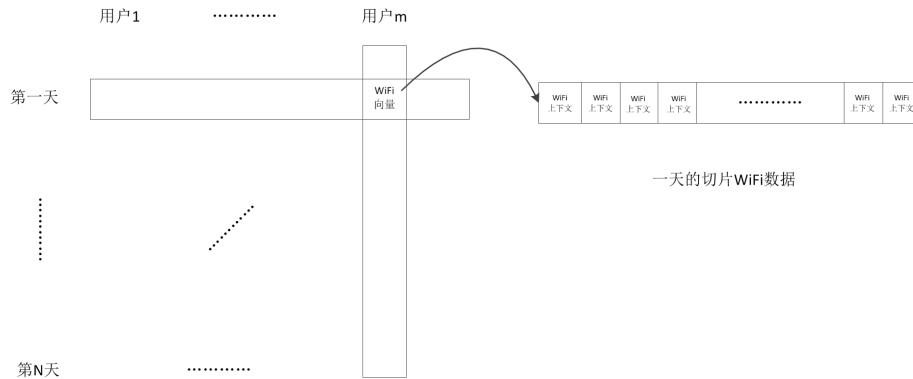


图 4.31 WiFi 数据结构化表示示意图

context 更形象的描述是当前 WiFi 数据收集时刻时，设备周围的感知信息和环境信息，本课题转换视角将 WiFi 感知序列拟物化，看做我们现实中生活的环境，考虑到当用户 1 当前时刻所处的环境和用户 2 所处的环境极度相似或者相同的时候，那么我们就认为二者在同一个环境中也就是产生了现实交互比没有在同一环境中的用户具有更强的关系强度。因此我们将当前时刻的环境用无向带权图来表示，即假设每一个能够探知的 WiFi 源都与设备存在一条边，然后用户在一段时间内的数据就成了一张图如图4.33，最后比较两个用户的在每一个时间切片下的图的相似性推断出用户的关系强度。

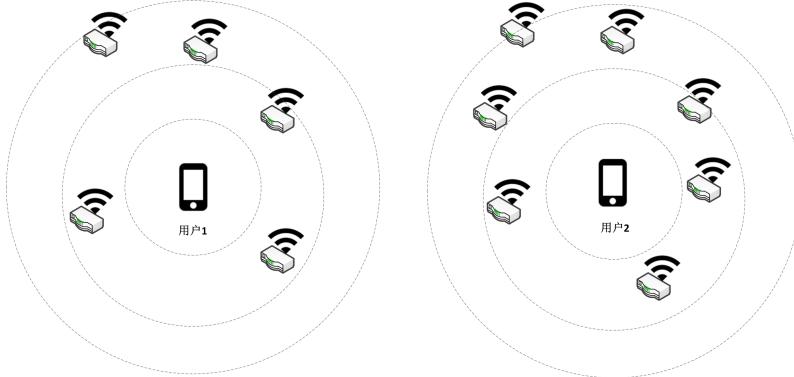


图 4.32 WiFi 上下文抽象化表示示意图

4.5 蓝牙感知数据处理计算技术

相比于 WiFi 设备的感知距离，蓝牙设备的感知距离进一步受限缩减为 5 到 10 米左右，在这样短距离的情况下，则可以考虑二者有现实生活中的交互，然而这种方法是有局限的即如果在一时刻内设备能够搜索到多个蓝牙设备而有未和其中任何一个设备进行连接，那么计算用户之间的相似度就不能采用上述的方法。因此我们采用节4.4中相似的数据抽象化方法，将收集到的用户蓝牙感知数据进行按照图3.3中的流程将蓝牙数据进行结构化清洗并结构化存储如图4.34所示，

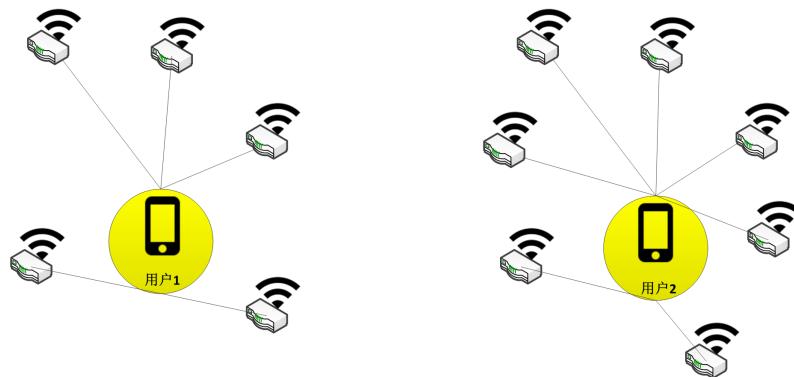


图 4.33 WiFi 图结构抽象化表示

得到蓝牙的上下文感知信息表示如图4.35，最后我们同样采用图的数据结构对蓝牙上下文信息进行存储，如图4.36

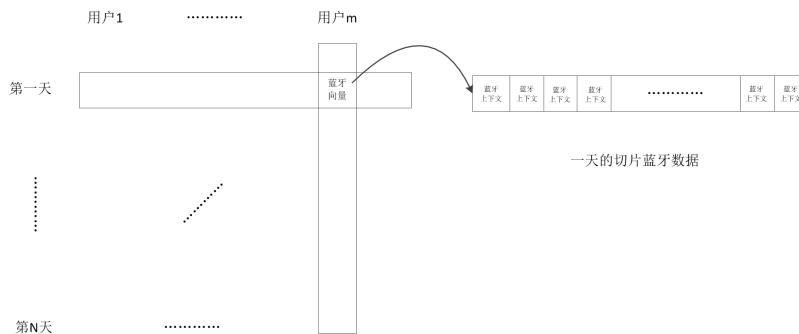


图 4.34 蓝牙数据结构化表示示意图

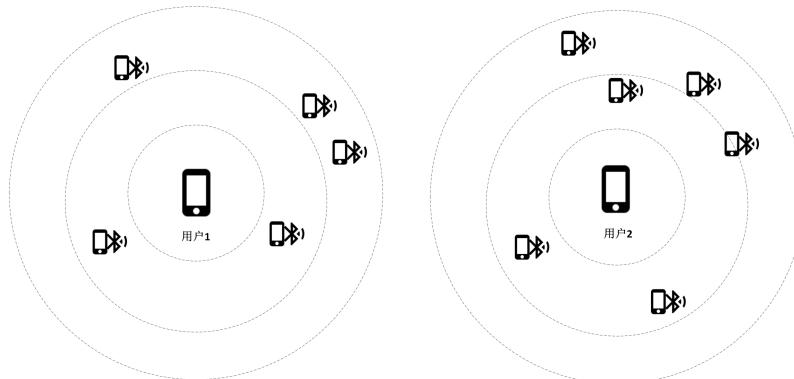


图 4.35 蓝牙数据上下文抽象化表示示意图

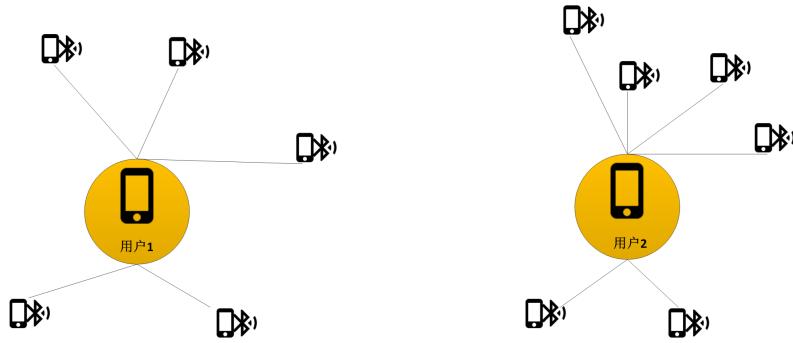


图 4.36 蓝牙数据图结构抽象化表示

4.6 小结

本章主要描述了 RSMHD 计算框架中的三个重要组成部分即基于 GPS 轨迹数据的处理和计算、基于 WiFi 感知数据的处理和计算和基于蓝牙感知数据的处理和计算。在本章中首先对 GPS 轨迹数据预处理方法展开了详细的分析，分别对常用的滤波：中值滤波、均值滤波、卡尔曼滤波进行了 GPS 轨迹滤波和噪音点剔除，在经过对实验结果的对比之后发现采用时间片切分的 GPS 轨迹卡尔曼滤波能够得到比较好的效果；然后描述了用户 GPS 轨迹中停留点的检测，通过停留点检测算法识别出用户轨迹中的停留点并为后续工作做好数据准备；接下来介绍了几种常用的聚类算法：K-means 聚类算法、DJ-Cluster 密度聚类算法以及 Science 改进的密度聚类算法对用户 GPS 轨迹数据中的停留点进行聚类分析，得到用户轨迹的聚类簇，并针对每种聚类算法和实验的结果来分析每一种聚类算法在轨迹点聚类问题上的优势和不足之处；在获取到用户轨迹停留点和停留点聚类簇的基础上阐述了停留点的语义化过程，并进行了相对应的结果展示；在本章的最后三章节中，分别描述了本研究中对用户 GPS 轨迹数据的结构化表示存储、WiFi 感知数据的处理和抽象结构化表示方法和蓝牙感知数据的处理和抽象结构化表示方法，并详细的用流程图和模型图进行了更加详细的过程描述。下一章将会详细描述基于本章中划分的模型进行用户关系强度度量的计算方法。

第五章 RSMHD 用户关系强度计算方法概述

在上一章详细描述了面向 GPS 数据的语义标签标注技术，这一章重点描述我们自己提出的 URSHV 用户关系强度计算方法。将从计算方法概述、输入数据准备以及关系强度计算三方面来描述 URSHV 计算方法。

5.1 用户关系强度计算方法概述

层级用户关系度量计算方法 URSHV 从三个不同的抽象层次，从不同角度采用不同的方法来度量用户之间的关系强度。第一层基于用户日常的原始轨迹数据度量用户日常轨迹之间的相似度；第二层度量的是基于语义位置的用户行为模式之间的相似度，其抽象层次比第一层更高，其含义比第一层更加丰富；第三层度量的是基于语义标签的用户行为模式之间的相似度，其抽象层次比第二层更高，语义更加精确。URSHV 模型从轨迹、物理位置以及语义位置等三个由低到高的抽象层次，从三个反映人们日常活动和行为模式的方面来度量人们之间的关系强度，并基于这三个层次的度量结果，采用集成学习的思想进行投票，以投票结果作为人们之间的关系强度，因而能够全面真实地反映日常生活当中人们之间的关系强度。

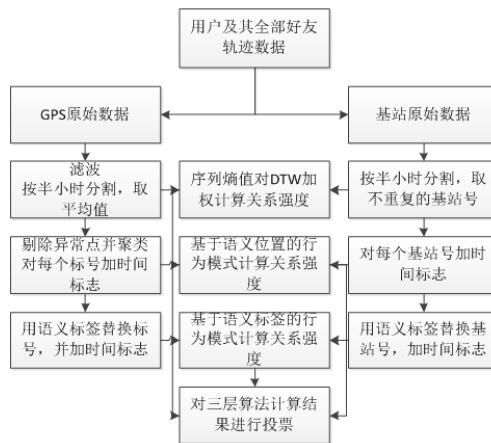


图 5.1 URSHV 模型框架

5.2 输入数据准备

上一节概述了计算方法，这一节将具体描述如何对 GPS 数据和基站数据进行预处理以得到需要的输入，下一节将具体描述如何计算用户之间的关系强度。

在日常生活中，用户的位置既可以通过智能手机内嵌的 GPS 传感器获取其位置信息，又可以通过用户所处区域内的通信基站进行定位。通过 GPS 获取的位置

信息相对与通过基站获取的位置信息要精确，但是长时间通过 GPS 传感器采集用户的位置信息将消耗大量的电量，会对用户手机的日常使用造成一定的影响。虽然基于基站的定位方式相对于 GPS 定位方式获取的位置信息精度要低，但其更有利于用户隐私的保护。因此，为了满足不同用户的不同需求，URSHV 模型既能够对 GPS 位置数据进行处理，同时又能够对基站位置数据进行处理。但是，无论是基于 GPS 的位置数据还是基于通信基站的位置数据都包含着大量的噪音，因此，为了更加精确地度量用户之间的关系强度，我们首先对这些数据进行去噪处理，而后采用不同方法来计算用户之间的关系强度。

设用户集合为 U , $U = \{u_1, u_2, \dots, u_n\}$, 其中 n 表示用户个数, D_i 表示用户 u_i 采集数据的日期的集合, 表示为 $D_i = \{d_1, d_2, \dots, d_{m_i}\}$, 其中 m_i 表示用户 u_i 采集数据的总天数。 F_i 表示用户 u_i 的全部朋友组成的集合, 表示为 $F_i = \{u_{k_1}, u_{k_2}, \dots, u_{k_{f_i}}\}$, 其中 f_i 表示用户 u_i 的好友的个数。 $Trace$ 表示所有用户所有天的轨迹数据的集合, 表示 $Trace = \{Trace_1, Trace_2, \dots, Trace_n\}$, 其中 $Trace_i$ 表示用户 u_i 所有天采集的轨迹序列的集合, 表示为 $Trace_i = \{Trace_{i,k} | k \in D_i\}$ 。 $Trace_{i,k}$ 表示用户 u_i 在 k 这一天的轨迹序列, 表示为 $Trace_{i,k} = \{l_1, l_2, \dots, l_{n_{i,k}}\}$, 其中 $n_{i,k}$ 表示用户 u_i 在 k 这一天采集的轨迹数据的条数, l_b 表示 b 时刻采集的位置数据记录, 可以为 GPS 经纬度, 也可以是基站号。

对于 GPS 数据和基站数据表示的用户轨迹序列进行预处理时，我们在下面三小节中分别依次描述模型的三层输入。

5.2.1 轨迹数据的处理与准备

处理 GPS 数据：对每个用户每天的数据 $Trace_{i,k}$ 进行滤波，目的是减少数据噪声；对每个用户每天的数据按半个小时进行切割，即将用户 u_i 的每天数据 $Trace_{i,k}$ 按时间均分为 48 份，表示为 $Sep_trace_{i,k} = \{Sep_trace_{i,k,1}, \dots, Sep_trace_{i,k,48}\}$ ，其中每一份数据表示为 $Sep_trace_{i,k,s} = \{l_{a_i} | l_{a_i} \in Trace_{i,k} \wedge a_i \in s\}$ ；对 $Sep_trace_{i,k,s}$ 按经纬度计算平均值，并将新的轨迹序列表示为 $Ntrace_{i,k}$, $Ntrace_{i,k} = \{Ntrace_{i,k,1}, \dots, Ntrace_{i,k,48}\}$ ，其中 $Ntrace_{i,k,s} = \frac{\sum(Sep_trace_{i,k,s})}{len(Sep_trace_{i,k,s})}$ ；其中 $sum(A)$ 表示对序列 A 中的元素求和， $len(A)$ 表示序列 A 的长度。将 $Ntrace_i$ 作为用户 u_i 使用第一层算法计算其与全部好友关系强度的输入。

处理基站数据：对每个用户每天的数据按半个小时进行切割，即将用户 u_i 在 k 这一天的数据 $Trace_{i,k}$ 按时间均分为 48 份，表示为 $Sep_trace_{i,k} = \{Sep_trace_{i,k,1}, \dots, Sep_trace_{i,k,48}\}$ ，其中每一份数据表示为 $Sep_trace_{i,k,s} = \{l_{a_i} | l_{a_i} \in Trace_{i,k} \wedge a_i \in s\}$ ；对每半个小时内数据计算依次不重复的基站号序列，即对每一份数据计算其对应的集合 $set(Sep_trace_{i,k,s})$ ，确保基站号不重复；再将每天 48 份数据重新拼成一个序列，目的是对每天轨迹序列降维，否则计算量太大而实

际无法计算，新序列记为 $Ntrace_{i,k}$ 。其中 $Ntrace_{i,k} = \bigcup_{s=1}^{48} set(Sep_trace_{i,k,s})$ ；将 $Ntrace_i$ 作为用户 u_i 使用第一层算法的输入。

5.2.2 语义位置数据的处理与准备

GPS 数据准备：采用上一章中讨论的聚类方法对所有用户的轨迹数据进行聚类，得到全部语义位置表示为 $Loc = \{pl_1, \dots, pl_g\}$ ，其中 g 表示总共的语义位置的个数。通过聚类得到用户 u_i 在 k 这一天的语义位置序列表示为 $Ltrace_{i,k} = \{loc(l_1), loc(l_2), \dots, loc(l_{n_{i,k}})\}$ ，其中 $loc(l_j)$ 表示位置数据记录 l_j 对应的语义位置标号。所有用户的所有语义位置序列表示为 $Ltrace = \{Ltrace_1, \dots, Ltrace_n\}$ ，其中用户 u_i 的全部语义位置序列表示 $Ltrace_i = \{Ltrace_{i,k} | k \in D_i\}$ 。对每个用户每天的数据按半个小时进行切割，即将用户 u_i 的每天数据 $Ltrace_{i,k}$ 按时间均分为 48 份，表示为 $Sep_ltrace_{i,k} = \{Sep_ltrace_{i,k,1}, \dots, Sep_ltrace_{i,k,48}\}$ ，其中每一份数据表示为 $Sep_ltrace_{i,k,s} = \{loc(l_{a_i}) | l_{a_i} \in Trace_{i,k} \wedge a_i \in s\}$ 。在准备 word2vec 模型的输入数据以及对应的模型输入数据时：我们需要首先计算每份数据不重复的语义位置序列，即 $Lsep_ltrace_{i,k,s} = set(Sep_ltrace_{i,k,s})$ ，然后将每天的 48 分数据合并成一个序列得到 $LLtrace$ ， $LLtrace_{i,k} = \bigcup_{s=1}^{48} Lsep_ltrace_{i,k,s}$ 。将 $LLtrace$ 作为 word2vec 模型的输入，训练得到对应模型 $LW2V(M)$ ， M 表示每个语义位置对应的实数值向量的长度。将 $Lsep_ltrace_i$ 作为用户 u_i 在第二层使用 word2vec 模型计算关系强度时的输入。其中 $Lsep_ltrace_{i,k} = \{Lsep_ltrace_{i,k,1}, \dots, Lsep_ltrace_{i,k,48}\}$ 。在准备 LDA 模型的输入数据以及对应的模型输入数据时：在已得到 Sep_ltrace 的基础上，对每份数据计算不重复出现的语义位置，并对每个位置加上时间标记。用户 u_i 在 k 这一天第 s 时间段语义位置序列表示为 $Tltrace_{i,k,s} = \{TT(sl_b) | sl_b \in set(Sep_ltrace_{i,k,s})\}$ ，其中 $TT(sl_b)$ 表示对 sl_b 添加时间标记，表示该语义位置在该时间段出现。 $set(A)$ 表示计算序列 A 对应的集合，即 A 中无重复元素。将 $Tltrace_i$ 作为用户 u_i 在第二层算法使用 LDA 模型计算关系强度时的输入。将 $Tltrace_{i,k}$ 中的 48 份数据合并成一个序列 $LTltrace$ ，其中 $LTltrace_{i,k} = \bigcup_{s=1}^{48} Tltrace_{i,k,s}$ 。将 $LTltrace$ 作为 LDA 模型的输入，训练得到对应的 LDA 主题模型 $LLDA(K)$ ， K 表示主题的个数。

基站数据准备：将每一个基站视为一个物理位置，即 $Ltrace=Trace$ 。其余处理与 GPS 处理完全相同。

5.2.3 语义标签数据的处理与准备

GPS 数据准备：对前文得到的 Loc 中每一个语义位置采用上一章中讨论的方法标记其语义标签，标语义标签后用户 u_i 第 k 天的语义标签序列表示为 $Strace_{i,k} = \{Label(ll_b) | ll_b \in ltrace_{i,k}\}$ ，其中 $Label(ll_b)$ 表示 ll_b 对应的语义标签。所有用户的所有语义标签序列表示为 $Strace = \{Strace_1, \dots, Strace_n\}$ ，其中用户 u_i 的

全部语义位置序列表示 $S\text{trace}_i = \{S\text{trace}_{i,k} | k \in D_i\}$ 。对每个用户每天的数据按半个小时进行切割，即将用户 u_i 的每天数据 $S\text{trace}_{i,k}$ 按时间均分为 48 份，表示为 $Sep_strace_{i,k} = \{Sep_strace_{i,k,1}, \dots, Sep_strace_{i,k,48}\}$ ，其中每一份数据表示为 $Sep_strace_{i,k,s} = \{Label(l_{a_i}) | l_{a_i} \in L\text{trace}_{i,k} \wedge a_i \in s\}$ 。在准备 word2vec 模型的输入数据以及对应的模型输入数据时：我们需要首先计算每份数据不重复的语义标签序列，即 $S\text{sep_strace}_{i,k,s} = set(Sep_strace_{i,k,s})$ ，然后将每天的 48 份数据合并成一个序列得到 $S\text{Ltrace}$ ， $S\text{Ltrace}_{i,k} = \bigcup_{s=1}^{48} S\text{sep_strace}_{i,k,s}$ 。将 $S\text{Ltrace}$ 作为 word2vec 模型的输入，训练得到对应模型 $SW2V(M)$ ， M 表示每个语义位置对应的实数值向量的长度。将 $S\text{sep_strace}_i$ 作为用户 u_i 在第三层使用 word2vec 模型计算关系强度时的输入。其中 $S\text{sep_strace}_{i,k} = \{S\text{sep_strace}_{i,k,1}, \dots, S\text{sep_strace}_{i,k,48}\}$ 。在准备 LDA 模型的输入数据以及对应的模型输入数据时：在已得到 Sep_strace 的基础上，对每份数据计算不重复出现的语义位置，并对每个位置加上时间标记。用户 u_i 在 k 这一天第 s 时间段物理位置序列表示为 $T\text{trace}_{i,k,s} = \{TT(sl_b) | sl_b \in set(Sep_strace_{i,k,s})\}$ ，其中 $TT(sl_b)$ 表示对 sl_b 添加时间标记，表示该语义位置在该时间段出现。 $set(A)$ 表示计算序列 A 对应的集合，即 A 中无重复元素。将 $T\text{trace}_i$ 作为用户 u_i 在第三层算法使用 LDA 模型计算关系强度时的输入。将 $T\text{trace}_{i,k}$ 中的 48 份数据合并成一个序列 $S\text{Ttrace}$ ，其中 $S\text{Ttrace}_{i,k} = \bigcup_{s=1}^{48} T\text{trace}_{i,k,s}$ 。将 $S\text{Ttrace}$ 作为 LDA 模型的输入，训练得到对应的 LDA 主题模型 $SLDA(K)$ ， K 表示主题的个数。

基站数据准备：计算每一个基站对应的语义标签，其余处理与 GPS 数据处理完全相同。

准备好模型各个层次的输入数据后，在下一节我们将详细描述如何使用输入数据计算用户之间的关系强度。

5.3 关系强度计算

上一节我们描述了如何准备模型对应的三层输入数据，这一节我们将分别描述基于轨迹数据的关系强度计算方法和基于主题模型的关系强度计算方法。

5.3.1 基于原始轨迹数据的关系强度计算

我们计算每一个用户 u_i 与其每一个朋友 $u_k (u_k \in F_i)$ 之间的关系强度，并对 F_i 中的每一个朋友，按照其与 u_i 的关系强度大小按降序排列，使此序列中任意两个朋友与 u_i 的关系强弱顺序尽可能与实际情况一致。

基于 DTW 及序列熵值加权计算用户之间的关系强度。对用户 u_i 的每一个好友 u_k ，利用上一小节得到的 $N\text{trace}_i$ 和 $N\text{trace}_k$ 计算其轨迹序列相似度。 $N\text{trace}_{i,a}$ 表示用户 i 在 a 这一天的数据，其中 $a \in D_i$ ， $N\text{trace}_{k,b}$ 表示用户 k 在 b

这一天的数据，其中 $b \in D_k$ 。 $S(i, j)$ 表示若 $a = b$ 则取值为 1，否则取值为 0。 $DTW(Ntrace_{i,a}, Ntrace_{k,b})$ 表示用户 u_i 在 a 这一天的轨迹和用户 u_k 在 b 这一天的轨迹的相似度， $Entropy(Ntrace_{i,a})$ 表示用户 u_i 在 a 这一天的轨迹序列的熵值。则用户 u_i 和用户 u_k 的基于轨迹序列的关系强度计算方法见公式 5.1。DTW 计算的是距离，距离越小相似度越大，即该公式值越小，两个用户关系强度越强。

$$Ent_{DTW}(u_i, u_k) = \frac{1}{\sum_{a \in D_i, b \in D_k} S(a, b)} \sum_{a \in D_i, b \in D_k} S(a, b) \frac{DTW(Ntrace_{i,a}, Ntrace_{k,b})}{Entropy(Ntrace_{i,a})} \quad (5.1)$$

5.3.2 基于主题模型的关系强度计算

LDA 模型对应的关系强度计算方法： $Tltrace_i$ 表示用户 u_i 根据上一小节得到的语义位置序列， $Tltrace_k$ 表示用户 u_k 根据上一小节得到的语义位置序列。 $T(a, p, b, q)$ 表示若用户 u_i 在 a 这一天第 p 个时间段和用户 u_k 在 b 这一天第 q 个时间段数据均存在则为 1，否则为 0。 $LLDA(K).inf(Tltrace_{i,a,p})$ 表示对 $Tltrace_{i,a,p}$ 推断得到的主题分布，通常表示为 K 维的向量，其中 K 表示主题的个数。基于用户语义位置的行为模式的关系强度计算方法见公式 5.2，其中 \cos 表示余弦相似度。

$$\begin{aligned} LocLDA(u_i, u_k) = & \frac{1}{\sum_{a \in D_i, b \in D_k} S(a, b)} \sum_{a \in D_i, b \in D_k} S(a, b) \frac{1}{\sum_{p=q=1}^{48} T(a, p, b, q)} \\ & \sum_{p=q=1}^{48} T(a, p, b, q) * \cos(LLDA(K).inf(Tltrace_{i,a,p}), LLDA(K).inf(Tltrace_{k,b,q})) \end{aligned} \quad (5.2)$$

基于用户语义标签的行为模式的关系强度计算公式与基于语义位置的关系强度计算公式相似，见公式 5.3。

$$\begin{aligned} SemLDA(u_i, u_k) = & \frac{1}{\sum_{a \in D_i, b \in D_k} S(a, b)} \sum_{a \in D_i, b \in D_k} S(a, b) \frac{1}{\sum_{p=q=1}^{48} T(a, p, b, q)} \\ & \sum_{p=q=1}^{48} T(a, p, b, q) * \cos(SLDA(K).inf(Tstrace_{i,a,p}), SLDA(K).inf(Tstrace_{k,b,q})) \end{aligned} \quad (5.3)$$

word2vec 模型对应的关系强度计算方法： $Lsep_strace_i$ 表示用户 u_i 根据上一小节得到的语义位置序列， $Lsep_strace_k$ 表示用户 u_k 根据上一小节得到的语义位置序列。 $T(a, p, b, q)$ 表示若用户 u_i 在 a 这一天第 p 个时间段和用户 u_k 在 b 这一天第 q 个时间段数据均存在则为 1，否则为 0。 $DTW(Lsep_strace_{i,a,p}, Lsep_strace_{k,b,q})$ 表示 $Lsep_strace_{i,a,p}$ 和 $Lsep_strace_{k,b,q}$ 之间的 DTW 距离。计算 DTW 距离时需要知道两个语义位置之间的距离，我们用这两个语义位置对应的实数值向量之间的

余弦距离作为这两个语义位置之间的距离。即由用户语义位置的行为模式得到的关系强度计算方法见公式5.4。

$$\begin{aligned} LocW2V(u_i, u_k) &= \frac{1}{\sum_{a \in D_i, b \in D_k} S(a, b)} \sum_{a \in D_i, b \in D_k} S(a, b) \frac{1}{\sum_{p=q=1}^{48} T(a, p, b, q)} \\ &\quad \sum_{p=q=1}^{48} T(a, p, b, q) * DTW(Lsep_strace_{i,a,p}, Lsep_strace_{k,b,q}) \end{aligned} \quad (5.4)$$

基于用户语义标签的行为模式的关系强度计算公式与基于语义位置的关系强度计算公式相似，见公式5.5。

$$\begin{aligned} SemW2V(u_i, u_k) &= \frac{1}{\sum_{a \in D_i, b \in D_k} S(a, b)} \sum_{a \in D_i, b \in D_k} S(a, b) \frac{1}{\sum_{p=q=1}^{48} T(a, p, b, q)} \\ &\quad \sum_{p=q=1}^{48} T(a, p, b, q) * DTW(S sep_strace_{i,a,p}, S sep_strace_{k,b,q}) \end{aligned} \quad (5.5)$$

我们更关注的是用户和好友 A 的关系强度大于或小于用户与好友 B 的关系强度。因此我们实际计算结果为用户与其全部好友按关系强度降序排列得到的好友序列。

5.3.3 结果投票

对于用户 u_i ，我们对其全部好友 F_i 中的每一个朋友 u_k 使用 $EntDTW(u_i, u_k)$ 计算用户 u_i 和用户 u_k 之间的关系强度，并对 F_i 中的每一个朋友按照计算得到的关系强度降序排列得到 $E_i = \{u_{d_1}, \dots, u_{d_{f_i}}\}$ ，其中 $EntDTW(u_i, u_{d_a}) > EntDTW(u_i, u_{d_b})$ 如果 $a < b$ 。在此基础上，我们使用 $LocLDA(u_i, u_k)$ 或者 $LocW2V(u_i, u_k)$ 计算用户 u_i 和用户 u_k 之间的关系强度，并对 F_i 中的每一个朋友按照计算得到的关系强度降序排列得到 $L_i = \{u_{l_1}, \dots, u_{l_{f_i}}\}$ ，其中 $LocLDA(u_i, u_{l_a}) > LocLDA(u_i, u_{l_b})$ 或者 $LocW2V(u_i, u_{l_a}) > LocW2V(u_i, u_{l_b})$ 如果 $a < b$ 。最后我们使用 $SemLDA(u_i, u_k)$ 或者 $SemW2V(u_i, u_k)$ 计算用户 u_i 和用户 u_k 之间的关系强度，并对 F_i 中的每一个朋友按照计算得到的关系强度降序排列得到 $S_i = \{u_{s_1}, \dots, u_{s_{f_i}}\}$ ，其中 $SemLDA(u_i, u_{s_a}) > SemLDA(u_i, u_{s_b})$ 或者 $SemW2V(u_i, u_{s_a}) > SemW2V(u_i, u_{s_b})$ 如果 $a < b$ 。我们采用集成学习的思想对三个层次的计算结果 E_i 、 L_i 、 S_i 进行投票，投票规则为：对于与用户 u_i 关系第 k 强的好友 u_{v_k} ($k \leq 1$ 且 $k \ll f_i$)，我们使用三个层次对应的方法分别计算得到 u_{d_k} 、 u_{l_k} 和 u_{s_k} ，若这三个用户都不相同，则我们认为 $u_{v_k} = u_{d_k}$ ，若某个用户比如 $u_{l_k} = u_{s_k}$ 出现两次及以上，我们认为 $u_{v_k} = u_{l_k}$ 。以 $V_i = \{u_{v_1}, \dots, u_{v_{f_i}}\}$ 作为投票的最终结果。

5.4 小结

本章就如何使用轨迹数据度量用户之间的关系强度进行了深入讨论，首先描述了URSHV的计算方法，该方法能同时处理GPS数据和基站数据，并使用轨迹数据计算用户之间的关系强度；其次我们从GPS数据和基站数据两方面描述了如何准备URSHV的输入数据；最后我们从基于轨迹数据计算用户关系强度和基于用户行为模式计算用户关系强度两方面详细描述了我们如何使用轨迹数据度量用户之间的关系强度。下一章我们将主要描述实验用到的数据集，评估方法以及在数据集上的实验结果。

第六章 数据集、评估方法及实验结果

上一章我们首先概括描述了 URSHV 层级模型；然后描述了如何准备数据作为模型的输入；最后从基于原始轨迹数据的用户关系强度计算、基于主题模型的用户关系强度计算以及结果投票三方面重点描述了基于轨迹数据的用户关系强度度量模型。本章我们将描述如何在真实数据集上对第四章提出的模型进行验证以及对实验结果进行分析。

6.1 数据集

我们采用真实场景下采集的数据作为验证数据集。数据集由 MIT 媒体实验室在 2004-2005 年主持的 RealityMining 项目收集整理得到。RealityMining 项目追踪了 94 个使用安装预装软件的手机的用户，这些预装软件能够记录并发送用户数据，比如：通话记录、近似 5 米范围内的蓝牙设备、基站塔编号、应用使用以及手机状态。该项目追踪观察了包括学生和来自同一个研究机构的两个课题组的职员总共九个月对手机的使用情况。与此同时，该项目收集了每个志愿者提供的关系数据比如谁和谁是朋友等。

每个志愿者使用 Nokia6600 在后台运行一个称为 ContextLog 的程序采集数据。在该项目中期，研究者组织了一次在线调查问卷，106 个志愿者中共有 94 个人完成了该调查问卷。调查问卷内容见表 6.1。除此调查问卷外，采集的数据有每个志愿者手机的蓝牙 MAC 地址、每个志愿者开始参与该项目的日期、每个志愿者隶属的机构、每个志愿者隶属的研究小组、每个志愿者收集的 IMEI、每个志愿者的邻居、每个志愿者自己告知的工作时间、每个志愿者是否有一个规律的工作计划、每个志愿者自己告知的常去的聚集地、每个志愿者是否有一个可预言的日程安排、每个志愿者是不是把手机忘在家里或工作的地方、每个志愿者手机电量是不是经常耗光、每个志愿者生病频率、每个志愿者最近是否生病、每个志愿者是否经常出去旅游、给每个志愿者提供通话服务的运营商、每个用户每个月买东西花费的时间、每个志愿者发短信的频率、每个志愿者是否经常被描述给其他人、每个治愈者对他所处团体的评价、每个志愿者的通信时间记录、每个志愿者收集充电记录、每个志愿者收集使用日期记录、每个志愿者手机采集数据的记录时间、每个志愿者开机关机时间记录、每个志愿者的轨迹数据记录（由基站号、区域号以及时间戳构成）、用户经过的唯一的位置（由基站号和区域号表示）、每个志愿者周围的蓝牙设备名称、每个志愿者周围的蓝牙设备的 MAC 地址、每个志愿者扫描周围蓝牙设备的时间、基站号及区域号对应的位置语义标签、每个志愿者使用应用的开始时间及使用时长、每个志愿者手机每天记录数据的时长、

每个志愿者家对应的基站号和区域号托等。基站号及区域号与对应的语义标签见表6.2。

表 6.1 调查问卷^[3]

问题及答案选项

(1)Have you travelled recently?

1 Very often - more than a week/month 2 Often - week/month 3 Sometimes - several days/month 4 Rarely - several days/term 5 Never

(2)Do you own a car?

1 Yes 2 No

(3)How many miles to you live from MIT?

1. less than 1 2. 1-3 3. 4-10 4. more than 10

(4)How do you daily commute to MIT?

1. By foot 2. By bike 3. By T/bus 4. By car

(5)How much has your social network evolved since the start of Fall term?

1. A lot 2. Somewhat 3. Slightly 4. None

(6)Have you been sick recently?

1. Yes, in the last week 2. Yes, in the last two weeks 3. Yes, in the last month 4. No

(7)How long into the term did it take for your social circle to become what it is today?

1. Still evolving 2. 2 months into term 3. 1 month into term 4. Several weeks into term 5. First couple of days here

(8)I use my phone:

1. exclusively for work/school related matters 2. primarily for work/school related matters, but occasionally for personal/social use
3. equally for work/school and for personal/social use 4. primarily for personal /social use 5. exclusively for personal/social use

(9)How often do you send text messages?

1. Several times / day 2. once / day 3. once / week 4. once / month 5. never

(10)The majority of my daily work communication is done through: (you can select more than one) face-face discussion

1. Yes NaN. No

(11)The majority of my daily work communication is done through: (you can select

续下页

续表 - 调查问卷

问题及答案选项

more than one) email

2. Yes NaN. No

(12)The majority of my daily work communication is done through: (you can select more than one) phone

3. Yes NaN. No

(13)The majority of my daily work communication is done through: (you can select more than one) text-messaging

4. Yes NaN. No

(14)The majority of my daily personal communication is done through: (you can select more than one) face-face discussion

1. Yes NaN. No

(15)The majority of my daily personal communication is done through: (you can select more than one) email

2. Yes NaN. No

(16)The majority of my daily personal communication is done through: (you can select more than one) phone

3. Yes NaN. No

(17)The majority of my daily personal communication is done through: (you can select more than one) text-messaging

4. Yes NaN. No

(18)I am satisfied with my experience at MIT thus far I am satisfied with my current social circle

1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

(19)I am satisfied with my current social circle

1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

(20)I feel I have learned a lot this semester

1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

(21)I am satisfied with the content and direction of my classes and research this semester

1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

(22)I am satisfied with the support I received from my circle of friends

续下页

续表 - 调查问卷

问题及答案选项

1 – Strongly Agree 2, 3, 4, 5, 6, 7 – Strongly Disagree

(23) I am satisfied with the level of support I have received from the other members in my Media Lab research group/Sloan core team.

1 – Strongly Agree 2, 3, 4, 5, 6, 7 – Strongly Disagree

(24) I am satisfied with the quality of our group meetings

1 – Strongly Agree 2, 3, 4, 5, 6, 7 – Strongly Disagree

(25) I am satisfied with how my research group interacts on a personal level

1 – Strongly Agree 2, 3, 4, 5, 6, 7 – Strongly Disagree

我们对每个志愿者每天采集的基站编号的个数进行统计，发现所有志愿者总共在连续 374 天采集了数据，采集了基站数据的志愿者共有 88 人。图6.1是一个 88*374 的图像，每一个像素点表示一个志愿者在某一天采集数据，颜色越黑表明用户采集的基站数据条数越多。



图 6.1 志愿者采集基站数据可视化

该数据集提供了志愿者之间的朋友关系，提供朋友关系的志愿者共 94 人，其中有 6 个人没有轨迹数据，我们剔除掉这 6 个人后，对剩余 86 人的关系进行可视化。图6.2是一个 172*172 的图像，每相邻 4 个像素点若全部为黑色表示一个志愿者和另一个志愿者是朋友关系，否则不是。

在对数据集的分析过程中，我们发现朋友关系信息表中存在如下问题：1) 部分用户自己和自己是好朋友，另外一部分用户自己和自己不是好朋友；2) 某用户和另一个用户是好朋友，另一个用户和该用户不是好朋友。我们认为用户之间的好友关系应该满足反自反和对称，即自己和自己不是好友，如果用户 A 和用户 B 是好朋友，则用户 B 和用户 A 是好朋友。经过这样处理后，我们得到好友数大于 1 的用户共有 34 个，若用户好友数为 1，则使用模型计算得到的关系强弱顺序与实际必定一致，故剔除这部分用户。为此，在后面的实验中，我们使用这 34 个用户及其全部朋友的数据来对 URSHV 模型进行验证。

表 6.2 基站区域号与对应的语义标签

基站号	区域号	语义标签
5119	40811	T-Mobile Media lab 1
5119	40332	TMO Tech sq 2
5123	40763	TMO MIT / Ashdown 3
5119	40342	TMO Ashdown 4
5119	40801	T-Mobile East campus / hyatt 5
5119	40342	T-Mobile Inf corr 6
5119	40802	T-Mobile Tang 7
5131	43861	T-Mobile Tang 8
5119	40793	T-Mobile Mit 9
24127	132	AT&T Wirel 1-115
24127	131	AT&T Wirel 1-115
24127	2421	AT&T Wirel 2-103/ ML / End Inf cor
24127	2353	AT&T Wirel Build 3
24127	2833	AT&T Wirel Student center
24127	111	AT&T Wirel ML / Mass Ave/ Infinite
24127	182	AT&T Wirel Mass ave bridge 310 smoots New house
24127	2832	AT&T Wirel ML
24127	113	AT&T Wirel Ml
24127	2422	AT&T Wirel Ml
24127	2833	AT&T Wirel Ml
24127	112	AT&T Wirel Ml
24127	2413	AT&T Wirel Ml
24127	133	AT&T Wirel Ml
24127	2433	AT&T Wirel Ml
24123	261	AT&T Wirel Ml
24127	2832	AT&T Wirel Medical
24127	182	AT&T Wirel Mass ave bridge 310 smoots

该数据集中采集的位置信息是基站信息，虽然基站定位方式的精确度比 GPS 定位方式低，但更有利于用户隐私的保护，这也是我们选择该数据集进行实验的主要原因之一。

6.2 评估方法

上一节我们主要描述了验证我们算法需要用到的真实数据集，但是数据集中并未给出志愿者之间关系强度的数值或者大小关系，因此需要我们自己构造志愿

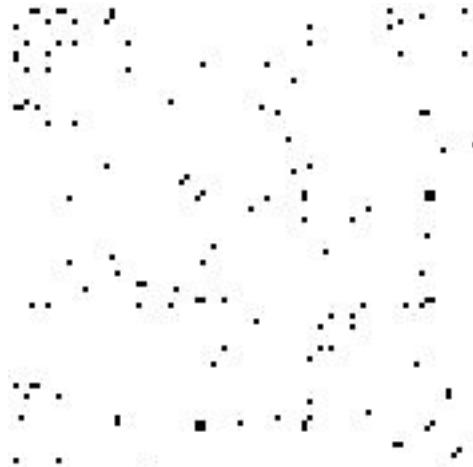


图 6.2 朋友关系可视化

者之间的关系强度作为真实结果，且目前信息检索方面的评估方法不太适合本课题，我们自己提出了一种评估方法作为对我们自己提出的模型的性能的评价。

6.2.1 构造真实结果

根据上文提到的社会心理学一些研究成果，态度、兴趣、价值观、背景和人格等方面更相似的人关系更亲密，尤其是对生活在一起的一个群体来说，如果在这些方面类似并且对某些问题的看法相似，则其关系可能就更加紧密。在现实生活中，通常通过问卷调查方式来获得这这些方面的信息，问卷调查结果是这些方面的一种真实体现和反映，因此，我们可以认为问卷调查结果越相似的用户关系越亲密，为此，我们根据上一节描述数据集中问卷调查回答结果的相似性作为朋友之间真实的关系强度。

经过对上一节描述的数据集中的问卷调查的仔细分析，我们发现问卷调查中的所有问题基本上可以分为两类：第一类问题可以用“是”或“否”来回答，另一类问题答案多选，但是每个选项按顺序呈现强度增强、次数增加或者次数减少。为了计算朋友之间的真实的关系强度，针对这两类问题，我们采用不同的评分方法。针对第一类问题当中的每一个问题，如果两个朋友的答案相同，则评分为 1，否则评分为 0；针对第二类问题当中的每一个问题，如果两个朋友的答案越接近，则评分越高，并且将评分归一化到 0-1 之间，使得每个问题在总的关系强度评分中占有相同的权重。在完成对所有问题评分基础上，对所有评分进行累加求和，以此作为两个朋友之间的关系强度。依次对每个用户及其所有朋友按上述方法计算其与每个朋友之间的关系强度，并对其所有朋友的评分按降序排列，得到一个用户与其所有朋友之间的关系强度序列，以此序列作为该用户与其朋友之间真实的关系强度。真实结果见表 6.3。在此基础上，使用 URSHV 模型计算出来

的用户之间关系强度序列与真实的关系强度序列进行对比，从而验证 URSHV 模型的有效性。

6.2.2 评估方法

通过前面的描述，我们可以知道，经过投票之后得到的用户之间的关系强度是该用户的全部好友按照与该用户的关系强度由强到弱排列的一个用户序列，而我们上一小节计算得到的用户之间的关系强度也是该用户的全部好友按照与该用户的关系强度由强到弱排列的一个用户序列，所以我们度量方法最关键的问题是如何度量两个有完全相同元素组成的有序序列，而这两个序列仅有的差别在于所有元素的排列可能不同。我们这个问题看起来很像一个信息检索问题，就是一个结果的排序问题，但是主要区别在于，信息检索对应的问题有很多无关的结果，这样我们只需要计算排在前面的正确的结果就可以得到准确率；而对本课题而言，所有结果都是准确的，只是应该按照一定的顺序。这个原因使得我们通过深入分析发现信息检索相关的一些度量方法不满足我们度量的要求，我们通过查阅相关资料和文献，发现逆序对数是度量两个有序序列是否一致很合理的一个指标，因此下面将主要描述如果使用逆序对数度量两个有序序列的一致性。

为了度量使用 URSHV 模型计算出来的用户与朋友之间关系强度序列 V_i 与真实的关系强度序列 G_i 的一致性，我们参考文献 [51]，提出一种基于逆序对数的有序序列一致性度量方法。设 A 为一个有 N 个数字的有序集 ($N > 1$)，且所有数字均不相同，如果存在正整数 i, j ，使得 $1 \leq i < j \leq N$ ，而 $A[i] > A[j]$ ，则称 $\langle A[i], A[j] \rangle$ 为 A 的一个逆序对。 A 中全部的逆序对的个数称为逆序对数。我们把序列 G_i 作为有序集，来计算序列 V_i 的逆序对数。设该用户共有 f_i 个好友，若逆序对数为 0，说明实验结果与实际结果完全一致，若逆序对数为 $\frac{f_i * (f_i - 1)}{2}$ ，则说明实验结果恰好是实际结果的逆序。因此，我们提出的有序序列一致性度量公式见公式6.1。其中 f_i 为用户 u_i 的全部好友的个数， k_i 为 V_i 相对于 G_i 的逆序对数。根据该公式我们可以发现，若实验结果序列的逆序对数为 0，则评分为 1，若实验结果与实际结果完全相反，则评分为 0。对每个用户可计算得到一个一致性评分，在此基础上，对所有用户的一致性评分取平均值，以此作为对模型对朋友关系强度度量好坏程度的度量，见公式6.2。

$$score(u_i) = 1 - \frac{k_i}{f_i * (f_i - 1)/2} \quad (6.1)$$

$$Score = \frac{1}{n} \sum_{i=1}^n score(u_i) \quad (6.2)$$

表 6.3 志愿者和其全部好友 (按关系强度递减排列)

志愿者	朋友 (按关系强度递减排列)
1	9, 19, 85, 71, 10, 4, 5
2	77, 19
3	18, 12, 7, 73
4	71, 1, 56
7	12, 22, 9, 3, 56
9	1, 85, 7, 73
11	48, 36
12	22, 7, 3
18	3, 30
19	1, 47, 2, 5
20	77, 50, 78
22	12, 7
30	18, 56
35	79, 55, 78, 36
36	55, 79, 35, 48, 78, 11
38	67, 45
40	64, 55
48	11, 36
50	77, 20
52	25, 24
53	31, 14
55	79, 36, 35, 78, 40
56	74, 77, 7, 66, 30, 4, 24
58	60, 64
60	58, 64
64	40, 60, 58
67	38, 76
71	1, 4, 5
73	82, 3, 9
77	2, 50, 20, 56
78	43, 35, 79, 55, 20, 36
79	35, 55, 36, 78
82	73, 72
85	1, 9

6.3 实验结果与分析

实验环境为 windows 7 64 位, 4 核, 3.2GHz 主频, 8G 内存, 使用 Python 编码实现。

为了使用第三章提到的基于原始轨迹的用户关系强度度量方法, 即对基站数据使用 DTW 方法时, 首先要确定任意基站之间的距离, 在此基础上使用扩展 DTW 方法计算用户之间的物理距离。欧式距离是最常见的一种度量方法, 对于 GPS 形式的轨迹数据我们就可以使用欧式距离来度量, 但是基站号只是不同基站之间为了区分生成的一个标号, 并无实际物理意义, 因此无法直接使用欧式距离, 所以我们需要采用一些方法使用基站号来定义两个基站之间的距离。

我们采取如下方法来定义基站之间的距离, 我们把每天用户手机连接过的基站视为一条基站序列, 对于基站 A 和 B, 我们从所有用户所有天的基站序列中找到同时出现 A 和 B 的序列, 计算每个序列中 A 和 B 中间不同的基站号的个数, 取最小值加一作为基站 A 和基站 B 之间的距离。例如, 假设找到全部同时出现 A 和 B 的序列有 ACDEEB、ADCCB 以及 AECFDEB, 则第一个序列计算得到 A 和 B 的距离为 4, 第二个序列计算得到 A 和 B 的距离为 3, 第三个序列计算得到 A 和 B 之间的距离 5, A 和 B 之间的距离取所有距离的最小值, 即 A 和 B 之间的距离为 3。若通过上述方法能够计算出两个基站之间的距离, 则称为这两个基站之间的距离存在。若 A 和 B 从未在同一个基站序列中出现过, 则定义 A 和 B 之间的距离为所有两个基站距离存在且最大的距离的 K 倍, K 为一个正实数参数, 在后面实验中我们能够看到该参数对实验结果的影响。如果对任意两个基站都从所有用户所有天的基站序列中找出同时出现这两个基站的序列, 然后按照上文所述的方法计算这两个基站之间的距离, 则其时间复杂度非常大, 因此, 我们通过对每个基站号建立倒排索引来减少计算量。倒排索引是指对每个基站号, 我们可以找到它在哪个用户那一天的轨迹数据中哪个位置出现。这样对于计算两个基站号之间的距离, 我们可以依次通过查找是否有相同用户, 是否在同一天, 以及同一天的位置来计算距离, 最后取最小值。使用倒排索引, 整个数据集只需要遍历一遍。如果不使用倒排索引, 本文使用数据集中不同的基站区域号总共有 30991 个, 则需要计算的基站距离共有 480205545 个, 若对全部数据遍历四亿多次, 可以想象时间复杂度将会特别大。

实验 1: 基于轨迹相似性计算用户之间的关系强度通过上面的处理方法, 可以计算出任意两个基站之间的距离, 因而就可以使用 DTW 方法来计算每一个用户 u_i 与其所有朋友 F_i 中每个人之间的关系强度, 进而得到每个用户与其所有朋友之间的关系强度序列, 记为 $W_{0,i}$ 。将该序列与 G_i 进行对比, 并按公式6.1对两者

的一致性进行评分，进而对所有用户使用公式6.2计算最终的一致性评分，验证结果的有效性。与此同时，一方面，第二章描述 DTW 算法时指出可以使用三种正则化方法对 DTW 计算结果进行优化处理来提升算法的效果，为此，我们使用这三种方法对 DTW 计算结果进行优化来获得每一个用户与其所有朋友之间的关系强度，进而得到优化后的每个用户与其所有朋友之间的关系强度序列，记为 $W_{1,i}$ 、 $W_{2,i}$ 以及 $W_{3,i}$ 。将 $W_{1,i}$ 、 $W_{2,i}$ 以及 $W_{3,i}$ 与 G_i 进行对比，并按公式6.1对两者的一致性进行评分，进而对所有用户使用公式6.2计算最终的一致性评分。另一方面，前面提到，如果两个基站 A 和 B 从未在同一个基站序列中出现过，则定义 A 和 B 之间的距离为所有两个基站之间距离最大值的 K 倍， K 为一个正实数参数， K 的设置对两个关系强度序列的一致性评分具有一定影响，图6.3描述了参数 K 的不同设置对 $W_{1,i}$ 、 $W_{2,i}$ 以及 $W_{3,i}$ 分别和 G_i 一致性评分的影响情况。观察6.3可以发现，当 K 为 2.5 时，通过 DTW 计算方法得到的 $W_{0,i}$ 与 G_i 更加接近一致。通过对经过三种正则化方法优化后的 DTW 计算结果，可以发现，通过使用 DTW 结果除以最优序列长度这种优化方法，得到的 $W_{3,i}$ 与 G_i 更加接近一致。不进行任何优化的 DTW 方法计算得到的用户好友序列见表6.4，使用最优序列长度归一化 DTW 距离并使用序列熵值加权得到的用户好友列表见表6.5。

在使用 DTW 及其经三种优化方法获得 $W_{0,i}$ 、 $W_{1,i}$ 、 $W_{2,i}$ 及 $W_{3,i}$ 的基础上，我们对每个用户每天的轨迹序列进行熵值加权，进而得到每个用户与其所有朋友之间的关系强度序列 E_i ，再使用公式6.1对 E_i 与 G_i 的一致性进行评分，进而对所有用户使用公式6.2计算最终的一致性评分，验证其有效性。图6.4描述了加权前后计算得到的 E_i 与 G_i 对应的全部用户的一致性评分结果。经验证，对于不同的 K 值，使用熵值加权后得到的一致性评分均好于不加权得到的一致性评分，图6.4 中仅列出当 $K = 2.5$ 的实验结果。通过对实验结果的进一步分析，我们发现对于编辑距离，使用熵值加权后计算得到的用户与其所有朋友关系强度序列与 G_i 更加一致，因此我们可以认为使用熵值加权的确能够更好的度量用户之间的关系强度。

实验 2：基于语义位置用户行为模式相似性计算用户之间的关系强度

为了进一步基于语义位置的相似性来度量用户与其所有朋友之间的关系强度，我们将每天每个用户其手机连接的所有基站号加上时间标记，例如 ‘5119.40332_24’ 表示用户在 11:30 到 12:00 期间（最后 2 位表示时间段）连接过基站 5119.40332。在此基础上，我们将每个经过这种方法处理后的基站号视为一个单词，用户每天连接过的基站号序列视为一个句子，每个用户连接过的全部基站号序列视为文档，使用所有用户的全部文档对 LDA 模型进行训练。在进行 LDA 模型训练时，首先需要确定主题的个数，主题个数是 LDA 模型的一个参数，主题个数不同，实验结果亦不相同。在计算关系强度的过程中，我们使用 LDA 模

表 6.4 无优化 DTW 方法得到好友列表

用户编号	真实关系强度对应好友列表	DTW 关系强度对应好友列表
1	9, 19, 85, 71, 10, 4, 5	4, 85, 71, 19, 5, 9, 10
2	77, 19	77, 19
3	18, 12, 7, 73	12, 7, 73, 18
4	71, 1, 56	1, 56, 71
7	12, 22, 9, 3, 56	12, 56, 22, 3, 9
9	1, 85, 7, 73	1, 7, 73, 85
11	48, 36	36, 48
12	22, 7, 3	7, 22, 3
18	3, 30	30, 3
19	1, 47, 2, 5	47, 1, 5, 2
20	77, 50, 78	77, 78, 50
22	12, 7	12, 7
30	18, 56	56, 18
35	79, 55, 78, 36	79, 36, 55, 78
36	55, 79, 35, 48, 78, 11	55, 79, 11, 78, 48, 35
38	67, 45	67, 45
40	64, 55	55, 64
48	11, 36	36, 11
50	77, 20	77, 20
52	25, 24	25, 24
53	31, 14	14, 31
55	79, 36, 35, 78, 40	79, 36, 78, 35, 40
56	74, 77, 7, 66, 30, 4, 24	74, 66, 7, 30, 77, 4, 24
58	60, 64	60, 64
60	58, 64	58, 64
64	40, 60, 58	60, 58, 40
67	38, 76	76, 38
71	1, 4, 5	1, 5, 4
73	82, 3, 9	3, 82, 9
77	2, 50, 20, 56	2, 20, 50, 56
78	43, 35, 79, 55, 20, 36	20, 36, 55, 79, 35, 43
79	35, 55, 36, 78	35, 55, 36, 78
82	73, 72	73, 72
85	1, 9	1, 9

表 6.5 熵值加权最优序列长度归一化 DTW 方法得到好友列表

用户编号	真实关系强度对应好友列表	优化 DTW 关系强度对应好友列表
1	9, 19, 85, 71, 10, 4, 5	4, 71, 85, 19, 5, 9, 10
2	77, 19	77, 19
3	18, 12, 7, 73	12, 18, 7, 73
4	71, 1, 56	1, 56, 71
7	12, 22, 9, 3, 56	12, 22, 56, 3, 9
9	1, 85, 7, 73	1, 7, 85, 73
11	48, 36	36, 48
12	22, 7, 3	7, 22, 3
18	3, 30	30, 3
19	1, 47, 2, 5	47, 1, 2, 5
20	77, 50, 78	77, 78, 50
22	12, 7	12, 7
30	18, 56	18, 56
35	79, 55, 78, 36	79, 36, 55, 78
36	55, 79, 35, 48, 78, 11	48, 55, 79, 35, 11, 78
38	67, 45	67, 45
40	64, 55	55, 64
48	11, 36	36, 11
50	77, 20	77, 20
52	25, 24	25, 24
53	31, 14	14, 31
55	79, 36, 35, 78, 40	79, 36, 78, 35, 40
56	74, 77, 7, 66, 30, 4, 24	74, 66, 7, 30, 77, 4, 24
58	60, 64	60, 64
60	58, 64	58, 64
64	40, 60, 58	60, 58, 40
67	38, 76	76, 38
71	1, 4, 5	1, 5, 4
73	82, 3, 9	3, 82, 9
77	2, 50, 20, 56	20, 2, 50, 56
78	43, 35, 79, 55, 20, 36	20, 79, 36, 55, 43, 35
79	35, 55, 36, 78	35, 55, 36, 78
82	73, 72	73, 72
85	1, 9	1, 9

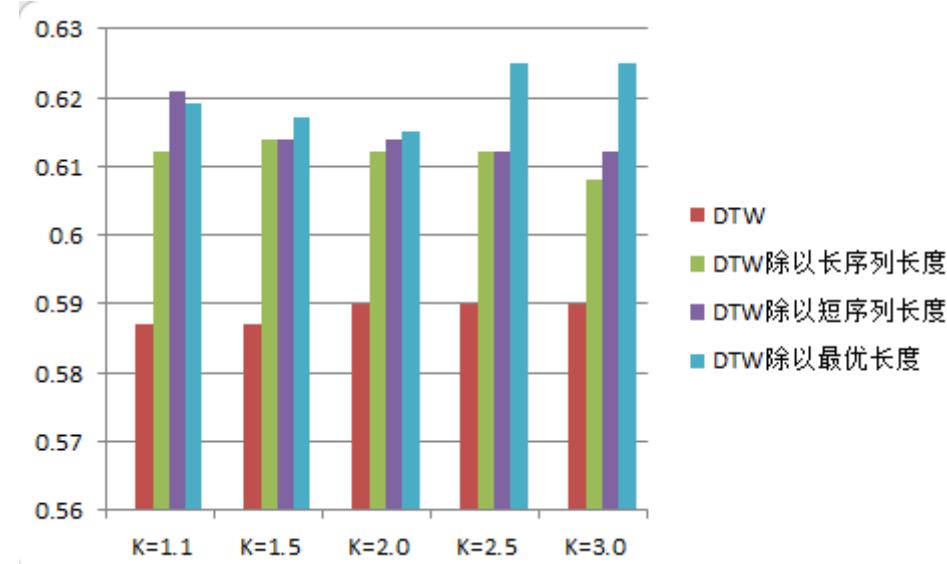


图 6.3 DTW 实验结果

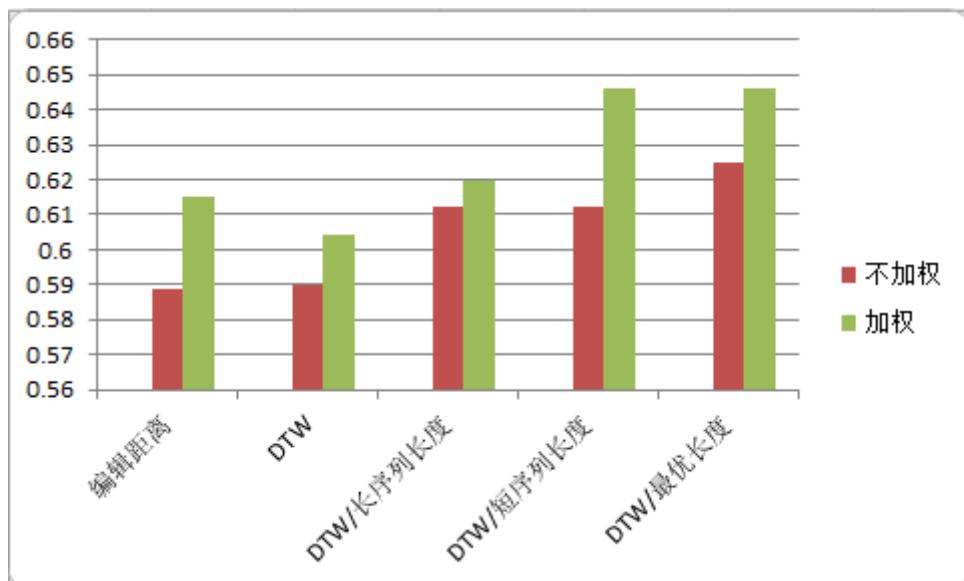


图 6.4 加权前后实验结果对比

型进行推断，因为推断过程进行随机初始化，从而使得 LDA 模型的每次执行结果不一定完全相同，因此，在实验中，针对每个不同的参数值（即主题个数）执行 10 次，并将每次计算获得的 L_i 与 G_i 进行一致性评分，对所有用户按公式6.2 计算最终的一致性评分。进而取这 10 个一致性评分的中位数作为该参数对应的一致性评分，如图6.5所示。

对图6.5进行分析，可以发现对于不同的主题个数有两个峰值：当主题个数为 50 时，一致性评分为 65.6%；当主题个数为 90 时，一致性评分为 66.2%。若主题个数太少，则区别能力太小，两个用户不管是非常相似还是比较相似都拥有相同

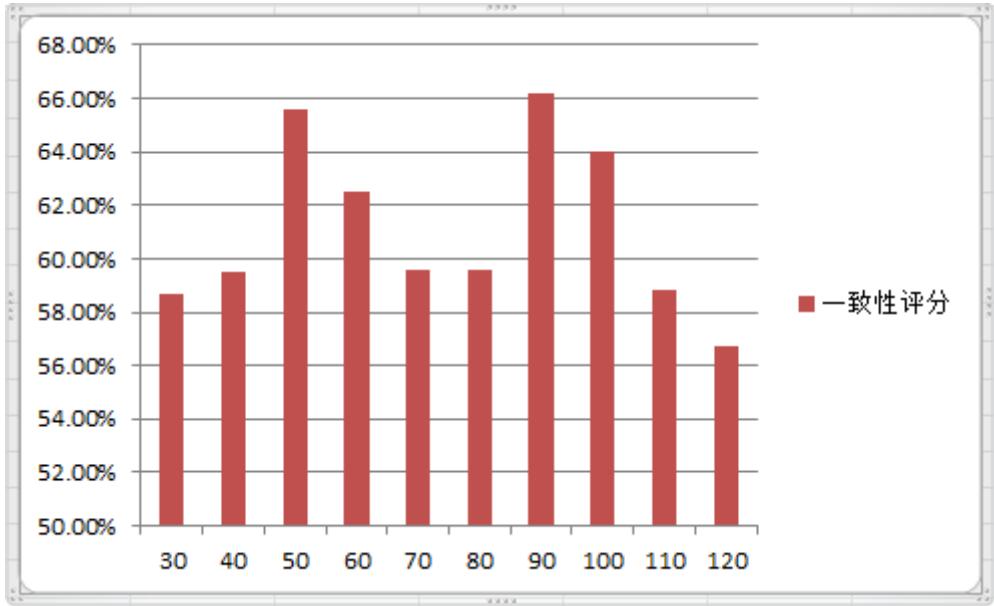


图 6.5 基于语义位置实验结果

的行为模式，则因为行为模式完全相同，因而无法区分这两个用户是非常相似还是比较相似。若主题个数太多，区别能力也将降低，每个用户分别对应不同的行为模式，即使两个用户实际上非常相似，当时因为行为模式不同，导致计算结果表明两个用户不相似。有两个峰值有可能是因为主题其实是一个层级概念，在某个抽象层次上可能主题数在 50 个左右，在另一个抽象层次上，主题数可能在 90 个左右。50 个主题对应的 LDA 模型计算得到的好友序列见表 6.6。

实验 3：基于语义标签用户行为模式相似性计算用户之间的关系强度

本章第一节我们描述数据集时说到该数据集提供了基站号和区域号对应的位置的语义标签，包括实验室以及每个用户的家庭住址对应的基站号和区域号，例如 5123.40811 对应 Media lab。为了进一步基于语义位置的相似性来度量用户与其所有朋友之间的关系强度，我们将基站号转换成对应的语义标签，形成一个基站号与语义标签相对应的映射表，如果一个基站号没有对应的语义标签，则其映射 Unknown。在此基础上，对每个语义标签加上时间标记，例如 ‘Media lab_27’ 表示用户在下午 2:00 到 2:30 期间（最后 2 位表示时间段）在 Media lab 出现过。对所有语义标签加上时间标记后，我们将每个带时间标记的语义标签视为单词，每天的语义标签序列视为句子，每个用户所有语义标签序列视为文档，使用所有用户的全部文档对 LDA 模型进行训练，其实验过程与上面的基于物理位置的实验过程一样，并将每次计算获得的 S_i 与 G_i 进行一致性评分，对所有用户按公式 6.2 计算最终的一致性评分。图 6.6 描述了在主题个数取不同值时所对应的一致性评分结果。

表 6.6 基于语义位置行为模式关系强度度量方法得到好友列表

用户编号	真实关系强度对应好友列表	基于语义位置行为模式相似度得到好友列表
1	9, 19, 85, 71, 10, 4, 5	85, 5, 19, 9, 10, 71, 4
2	77, 19	77, 19
3	18, 12, 7, 73	12, 18, 7, 73
4	71, 1, 56	1, 71, 56
7	12, 22, 9, 3, 56	12, 22, 56, 3, 9
9	1, 85, 7, 73	1, 85, 7, 73
11	48, 36	36, 48
12	22, 7, 3	7, 22, 3
18	3, 30	30, 3
19	1, 47, 2, 5	5, 1, 2, 47
20	77, 50, 78	77, 78, 50
22	12, 7	12, 7
30	18, 56	56, 18
35	79, 55, 78, 36	36, 78, 55, 79
36	55, 79, 35, 48, 78, 11	79, 55, 11, 78, 35, 48
38	67, 45	67, 45
40	64, 55	64, 55
48	11, 36	11, 36
50	77, 20	77, 20
52	25, 24	25, 24
53	31, 14	31, 14
55	79, 36, 35, 78, 40	79, 36, 78, 40, 35
56	74, 77, 7, 66, 30, 4, 24	66, 30, 74, 77, 7, 24, 4
58	60, 64	60, 64
60	58, 64	64, 58
64	40, 60, 58	60, 40, 58
67	38, 76	76, 38
71	1, 4, 5	1, 5, 4
73	82, 3, 9	82, 3, 9
77	2, 50, 20, 56	2, 20, 56, 50
78	43, 35, 79, 55, 20, 36	20, 79, 43, 55, 36, 35
79	35, 55, 36, 78	36, 55, 78, 35
82	73, 72	73, 72
85	1, 9	1, 9

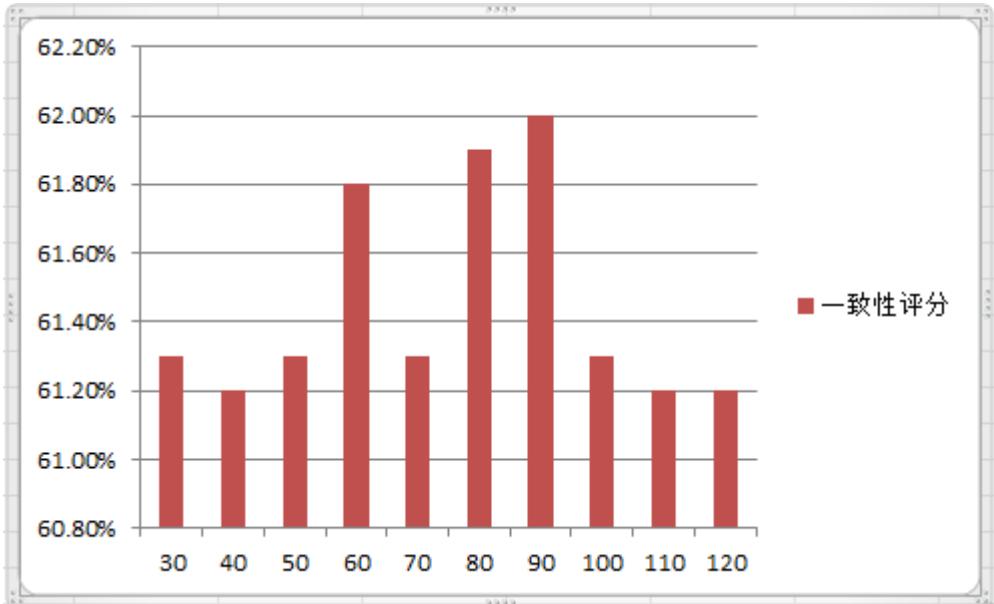


图 6.6 基于语义标签实验结果

对图6.6进行分析，不同的参数值（主题个数）对结果影响不大，原因可能是实验的对象主要是学校教员和学生，大家在日常生活当中的基于语义位置的行为模式非常类似，因而对不同的参数值（主题个数）不敏感。

语义标签有实际含义，以主题个数 75 为例，通过观察 LDA 模型学习到的主题，发现该模型学到了 3 个主题，如表6.7所示，主题 1 表示的是晚上在实验室或教室，主题 2 表示早上和晚上在家，主题 3 表示的上午在实验室。60 个主题对应的 LDA 模型计算得到的好友序列见表6.8。

表 6.7 LDA 模型学习到的主题

主题 1	主题 2	主题 3
Tech sq_47, Tech sq_46	home_14, home_15	Media lab_17, Media lab_16
Tech sq_40, Tech sq_38	home_8, home_6	Media lab_20, Media lab_18
Tech sq_39, Tech sq_42	home_0, home_44	Media lab_19, Tech sq_17

实验 4：对计算结果进行投票

实验 1、实验 2 和实验 3 分别描述了层级模型 URSHV 每一层的实验结果，在此基础上，我们使用前面描述的投票规则对三层实验结果进行投票，得到 V_i ，对三层结果投票的实验结果见图6.7。通过实验结果我们可以发现，使用投票方法后，我们可以更好的度量用户之间的关系强度，投票后得到的用户好友列表见表6.9。

表 6.8 基于语义标签行为模式关系强度度量方法得到好友列表

用户编号	真实关系强度对应好友列表	基于语义标签行为模式相似度得到好友列表
1	9, 19, 85, 71, 10, 4, 5	5, 9, 85, 19, 10, 71, 4
2	77, 19	19, 77
3	18, 12, 7, 73	18, 12, 7, 73
4	71, 1, 56	1, 71, 56
7	12, 22, 9, 3, 56	3, 56, 12, 22, 9
9	1, 85, 7, 73	73, 1, 85, 7
11	48, 36	48, 36
12	22, 7, 3	22, 3, 7
18	3, 30	3, 30
19	1, 47, 2, 5	5, 47, 1, 2
20	77, 50, 78	78, 77, 50
22	12, 7	12, 7
30	18, 56	18, 56
35	79, 55, 78, 36	78, 36, 55, 79
36	55, 79, 35, 48, 78, 11	55, 79, 11, 35, 48, 78
38	67, 45	67, 45
40	64, 55	64, 55
48	11, 36	11, 36
50	77, 20	20, 77
52	25, 24	25, 24
53	31, 14	31, 14
55	79, 36, 35, 78, 40	36, 79, 78, 40, 35
56	74, 77, 7, 66, 30, 4, 24	77, 66, 7, 30, 74, 24, 4
58	60, 64	64, 60
60	58, 64	64, 58
64	40, 60, 58	60, 58, 40
67	38, 76	76, 38
71	1, 4, 5	1, 5, 4
73	82, 3, 9	9, 82, 3
77	2, 50, 20, 56	20, 56, 2, 50
78	43, 35, 79, 55, 20, 36	35, 20, 43, 55, 79, 36
79	35, 55, 36, 78	36, 55, 78, 35
82	73, 72	73, 72
85	1, 9	1, 9

表 6.9 三层结果投票得到好友列表

用户编号	真实关系强度对应好友列表	三层结果投票得到好友列表
1	9, 19, 85, 71, 10, 4, 5	71, 85, 19, 10, 9, 5, 4
2	77, 19	19, 77
3	18, 12, 7, 73	18, 12, 7, 73
4	71, 1, 56	1, 71, 56
7	12, 22, 9, 3, 56	12, 22, 56, 3, 9
9	1, 85, 7, 73	1, 7, 85, 73
11	48, 36	48, 36
12	22, 7, 3	7, 22, 3
18	3, 30	3, 30
19	1, 47, 2, 5	5, 1, 2, 47
20	77, 50, 78	77, 78, 50
22	12, 7	12, 7
30	18, 56	18, 56
35	79, 55, 78, 36	79, 36, 55, 78
36	55, 79, 35, 48, 78, 11	79, 55, 11, 35, 48, 78
38	67, 45	67, 45
40	64, 55	55, 64
48	11, 36	11, 36
50	77, 20	20, 77
52	25, 24	25, 24
53	31, 14	31, 14
55	79, 36, 35, 78, 40	79, 36, 78, 40, 35
56	74, 77, 7, 66, 30, 4, 24	74, 77, 7, 30, 66, 24, 4
58	60, 64	64, 60
60	58, 64	64, 58
64	40, 60, 58	60, 58, 40
67	38, 76	76, 38
71	1, 4, 5	1, 5, 4
73	82, 3, 9	3, 82, 9
77	2, 50, 20, 56	20, 2, 50, 56
78	43, 35, 79, 55, 20, 36	20, 79, 36, 55, 43, 35
79	35, 55, 36, 78	36, 55, 78, 35
82	73, 72	73, 72
85	1, 9	1, 9

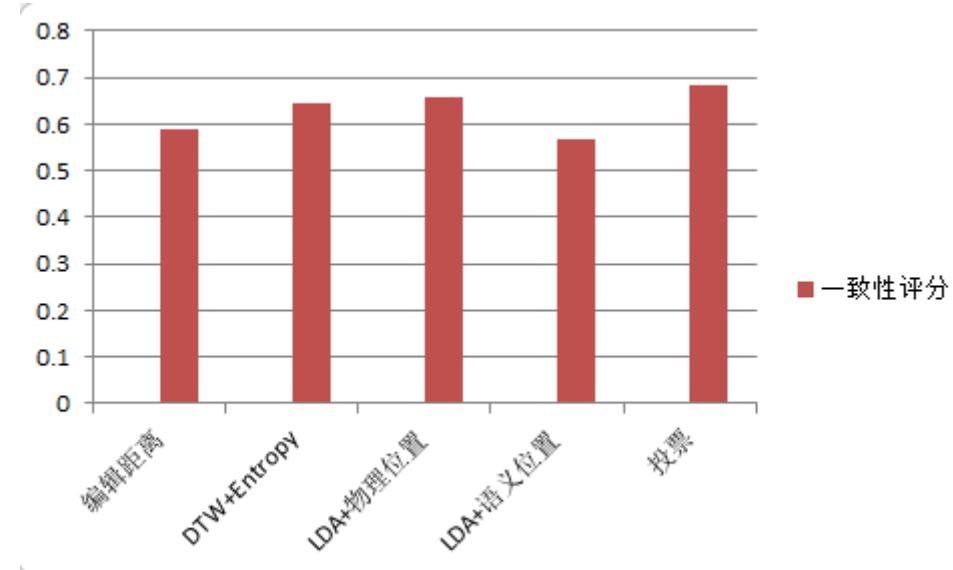


图 6.7 投票结果

6.4 小结

本章我们主要描述了验证我们模型使用的数据集，如何使用数据集构造真实结果，设计适用本课题问题的评估方法以及展示了实验演过并进行了深入分析。

第七章 结束语

7.1 工作总结

随着硬件的迅速发展，智能手机得以嵌入更多的传感器，且拥有更大的内存，更快的处理器，使得我们可以利用智能手机研究好多以前无法解决的问题。通常情况下，手机都会随身携带，从而手机可以基本完整的记录我们每天的生活轨迹，而生活轨迹又能很大程度上反映人和人之间在真实世界的交互，进而使得使用轨迹数据来度量人和人之间的关系强度成为可能。以前的工作只考虑真实世界人和人的交互次数，实际上，使用手机采集的轨迹数据我们能够得到更多的信息，比如用户基于轨迹的行为模式。因此，我们尝试从更多方面去考虑轨迹数据和用户之间关系强度的关系。本课题针对如何度量日常生活中人们之间的关系强度问题展开研究，提出了一个既可以对 GPS 数据进行处理又可以对基站数据进行处理，从日常轨迹、物理位置以及语义位置三个层次度量人们之间关系强度的层级模型 URSHV。概括起来，主要研究内容和贡献如下：

(1) GPS 数据相较于基站数据更杂乱，因此需要许多额外的处理，我们首先研究了对 GPS 数据的一些额外的处理。原始 GPS 数据采样得到结果有较大误差，且采集数据存在大量用户日常活动在路上行进的点，而我们课题只需要考虑用户停留在宿舍、实验室等语义位置相关的点，故需要对 GPS 数据进行降噪并剔除路上的点。在此基础上，我们需要通过一些方法来发现与 GPS 原始数据对应的如宿舍、实验室等语义位置，进而标记每个语义位置对应的语义标签。在本课题中，我们通过实验各种滤波算法，发现分段卡尔曼滤波具有更好的降噪效果；通过对采集的 GPS 数据进行进一步分析，我们发现路上点的密度远小于用户处于语义位置时的点的密度，因此我们采用基于密度的异常点剔除方法，且该方法可以自动学习参数；当前该领域用来发现语义位置的聚类算法存在一些问题，比如需要预先知道类别的个数或者对参数比较敏感，我们采用最新提出的一个基于密度的聚类算法来发现语义位置，该方法对参数更鲁棒，且不需要预先知道类别个数；在得到语义位置的基础上，我们需要通过一些方法标记语义位置对应的语义标签，目前常用的方法是人工手动标注，我们发现可以通过反地理编码，语义标签推断以及输入自动补全来减少用户和语义标签标注系统的交互。

(2) 我们使用原始轨迹数据的相似度，基于语义位置的用户行为模式的相似度以及基于语义标签的用户行为模式的相似度三方面来度量用户之间的关系强度。如何度量原始轨迹数据的相似度、如何度量用户模式之间的相似度以及如何对三层结果进行融合就是本课题最关键的问题。首先，在计算原始轨迹数据的相

似度时，我们发现使用编辑距离计算得到的相似度效果不是很理想，而 DTW 距离更倾向于序列长度较长的序列。因此，我们对 DTW 计算得到的距离使用三种方法归一化。并且我们发现用户每天活动的多样性不同使得该天轨迹数据的相似度对最终的相似度贡献不同，因此，我们使用用户每天轨迹序列的熵值对用户每天的相似度加权。其次，在计算用户行为模式时，我们发现 LDA 主题模型可以很好的用来发现用户基于轨迹的行为模式，且该模型的推断方法能够帮助我们很好的度量行为模式之间的相似度。最后，在得到三个层次的用户关系强度计算结果后，我们使用集成学习的思想对三个结果进行投票，并且以投票结果作为最终的关系强度。

(3) 以前的工作基于仿真数据集进行验证，真实数据集还是存在一些问题，如何对真实数据集进行处理，构造真实结果，以及如何针对我们的问题提供相应的评估方法以及模型中的各个参数对实验结果究竟有什么影响也是本课题急需解决掉一个重要问题。我们使用第五章第一节提到的数据集，对其朋友关系进行处理，使得该关系满足反自反和对称，用朋友之间调查问卷的相似度作为用户之间真实的关系强度。我们对用户关系强度的度量结果其实是该用户全部好友按与该用户关系强度亲密程度降序排列对应的好友序列，因此我们针对有序序列提出基于逆序对数的一致性评分评估标准来评价我们模型的实验结果。最后我们通过实验证明了各个参数对模型实验结果的影响，并且对结果进行了深入分析。

7.2 工作展望

论文针对基于轨迹数据的用户关系强度度量问题展开研究，在对相关技术研究基础上，提出了一个既可以对 GPS 数据进行处理又可以对基站数据进行处理，从日常轨迹、物理位置以及语义位置三个层次度量人们之间关系强度的层级模型 URSHV。虽然取得了一定的结果，但是仍然存在许多问题需要进一步研究和完善。现将这些问题总结如下：

(1) 基于 DTW 的关系强度度量方法虽然能得到一个比较好的度量结果，但是该算法具有比较大的时间复杂度。下一步工作尝试对数据进行一定的优化，使得该算法计算时消耗更少的时间。

(2) 我们只使用了原始轨迹数据的相似度和基于轨迹的行为模式的相似度两方面来度量用户之间的关系强度，根据轨迹数据，我们其实能够得到更多的信息。下一步工作尝试深入理解轨迹数据以及考虑从更多的方面来使用轨迹数据，使得我们对关系强度的度量更加全面。

(3) 虽然轨迹数据和我们在真实世界的交互密切相关。但是，通话记录，短信，蓝牙交互，社交网络交互，这些信息同样能够反映用户之间的关系强度。下

一步工作尝试采集更多的手机传感器数据，分别研究如何基于单个传感器数据度量用户关系强度以及如何综合使用这些传感器数据度量用户之间的关系强度。

致 谢

在本文完成和硕士生涯结束之际，我谨向所有给予我指导、关心、支持和帮助的老师、领导、同学和亲人致以衷心的感谢！

衷心感谢我的导师史殿习教授！感谢史老师对我的无私关怀、精心指导和严格要求。在两年多的硕士学习生活中，史老师在百般繁忙之际仍不忘抽出时间关心我的课题进展，找我聊人生、聊理想。不仅对我的课题提出很多很宝贵的建设性意见，而且对我将来人生的发展方向以及职业规划也提出很多至关重要的建议。我本身不是一个很勤奋的人，史老师也未严格要求每天呆在实验室的时间，但是每次有事找史老师时，史老师都在办公室看论文，史老师对移动感知领域论文了解的广度和深度，我望尘莫及，每次看到老师刻苦的身影，暗暗下定决心，一定要多看论文。史老师严谨的治学精神和精益求精的工作作风在不断的鞭笞着我，让我不断前行。在课题方面，我没有做到史老师希望的那么优秀，辜负了史老师对我的期望，在以后的日子里，我会更加努力，做更好的工作，做更好的自己，在此谨向他表示最衷心的感谢和最诚挚的敬意。

衷心感谢尹刚老师、丁博老师、刘慧老师，感谢你们在我课题的开题阶段给我的宝贵的建议和支持，使我更加明确了研究方向和思路，且为后续课题研究进展提供了很多现在看来确实很有远见的指导，衷心感谢窦勇老师和王晓东老师在预审阶段对我论文撰写提出的珍贵意见，使得该论文能够顺利完成。同时也感谢你们在生活上和学习上对我的帮助。

感谢何炫辰学长、李永谋学长、丁涛杰学长、陈富霞学姐和金星学姐，感谢你们在生活和学习上对我的帮助，和何炫辰学长的每次聊天都能使我烦躁的内心平静好多，每次学术上的问题李永谋学长都不厌其烦的帮我解决，陈富霞学姐给了我很多生活上的经验，受用至今。感谢同门好友吴渊、李寒、陈茜、周荣、谭杰夫、樊泽栋。感谢吴渊一直陪我去五楼实验室打乒乓球，经常陪我去东门吃夜宵；感谢李寒经常带我出去玩，看外面的世界；感谢陈茜帮我报账，提供零食当夜宵；感谢周荣给我平时无聊的生活带来乐趣。非常感谢各位同门好友在学习和生活上给我提供的帮助，因为你们使得平时波澜不惊的生活充满了色彩，非常庆幸、非常感谢！感谢王峰、赵邦辉、陈晓鹏、莫晓赟、李中秋、刘帆、童哲航、颜丙政、魏菁、成瑶瑶学弟学妹，和你们一起讨论学术问题，交流人生经验，我自己也受益匪浅。你们都是富有才华的研究人才，祝你们人生之路丰富多彩。感

谢同一个实验室的张飞、王东升，刘冰珣、古崇明经常打扰你们找你们聊天，你们没有怨言还给了我很多学术和生活的建议。

感谢朱涛政委、孙友佳政委和郑永辉政委。我们是科大第一届地方生，很多规矩都不懂，感谢你们在这两年半中给予我的帮助。感谢室友许名广、伍名、谢飞，感谢对面宿舍孙洪雷、吴平杰、吴茂永，在与你们一起度过的快乐的两年半生活中，感谢你们在学习和生活中给予我的帮助。

感谢学院八队的所有同学，与你们在一起的两年半时光是我人生最宝贵的一笔财富。感谢两年以来悉心指导我的每一位老师和前辈，你们的辛勤汗水让我在成长道路上跨步前进。

感谢 NudtPaper，它的存在让我的论文写作轻松自在了许多，让我的论文格式规整漂亮了许多。

最后，深深感谢生我养我的父母，你们的支持是我前进的最大动力，愿你们健康长寿！

参考文献

- [1] Starner T. Human-powered wearable computing [J]. IBM systems Journal. 1996, 35 (3.4): 618–629.
- [2] Cheng N, Mohapatra P, Cunche M, et al. Inferring user relationship from hidden information in wlans [C]. In MILCOM 2012-2012 IEEE Military Communications Conference. 2012: 1–6.
- [3] Eagle N, Pentland A. Reality mining: sensing complex social systems [J]. Personal and ubiquitous computing. 2006, 10 (4): 255–268.
- [4] Min J-K, Wiese J, Hong J I, et al. Mining smartphone data to classify life-facets of social relationships [C]. In Proceedings of the 2013 conference on Computer supported cooperative work. 2013: 285–294.
- [5] Pentland A, Eagle N, Lazer D. Inferring social network structure using mobile phone data [J]. Proceedings of the National Academy of Sciences (PNAS). 2009, 106 (36): 15274–15278.
- [6] Zhang D, Guo B, Yu Z. The emergence of social and community intelligence [J]. Computer. 2011, 44 (7): 21–28.
- [7] Miluzzo E, Lane N D, Eisenman S B, et al. CenceMe—injecting sensing presence into social networking applications [M] // Miluzzo E, Lane N D, Eisenman S B, et al. Smart Sensing and Context. Springer, 2007: 2007: 1–28.
- [8] Wang R, Chen F, Chen Z, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones [C]. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014: 3–14.
- [9] Wang R, Harari G, Hao P, et al. SmartGPA: how smartphones can assess and predict academic performance of college students [C]. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015: 295–306.
- [10] Granovetter M S. The strength of weak ties [J]. American journal of sociology. 1973: 1360–1380.
- [11] Haythornthwaite C. Strong, weak, and latent ties and the impact of new media [J]. The information society. 2002, 18 (5): 385–401.

- [12] Gustafson S, Moitra A. Extracting and Measuring Relationship Strength in Social Networks [J]. 2012.
- [13] Khadangi E, Zarean A, Bagheri A, et al. Measuring relationship strength in online social networks based on users' activities and profile information [C]. In Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on. 2013: 461–465.
- [14] Zhao X, Yuan J, Li G, et al. Relationship strength estimation for online social networks with the study on Facebook [J]. Neurocomputing. 2012, 95: 89–97.
- [15] Zillmann D, Bryant J. Selective exposure to communication [M]. Routledge, 2013.
- [16] Zajonc R B. Attitudinal effects of mere exposure. [J]. Journal of personality and social psychology. 1968, 9 (2p2): 1.
- [17] Zillmann D. Mood management in the context of selective exposure theory [J]. Annals of the International Communication Association. 2000, 23 (1): 103–123.
- [18] Zheng Y, Zhang L, Ma Z, et al. Recommending friends and locations based on individual location history [J]. ACM Transactions on the Web (TWEB). 2011, 5 (1): 5.
- [19] Lee J-G, Han J, Whang K-Y. Trajectory clustering: a partition-and-group framework [C]. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data. 2007: 593–604.
- [20] Li Q, Zheng Y, Xie X, et al. Mining user similarity based on location history [C]. In Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. 2008: 34.
- [21] Lu E H-C, Tseng V S, Philip S Y. Mining cluster-based temporal mobile sequential patterns in location-based service environments [J]. IEEE Transactions on Knowledge and Data Engineering. 2011, 23 (6): 914–927.
- [22] Burt R S. Structural holes: The social structure of competition [M]. Harvard university press, 2009.
- [23] Carruthers P. The illusion of conscious will [J]. Synthese. 2007, 159 (2): 197–213.
- [24] Burrows R, Nettleton S, Pleace N, et al. Virtual community care? Social policy and the emergence of computer mediated social support [J]. Information, Communication & Society. 2000, 3 (1): 95–121.
- [25] Coles A J, Wing M G, Molyneux P, et al. Monoclonal antibody treatment exposes three mechanisms underlying the clinical course of multiple sclerosis [J]. Annals of neurology. 1999, 46 (3): 296–304.

-
- [26] Petróczi A, Nepusz T, Bazsó F. Measuring tie-strength in virtual social networks [J]. *Connections*. 2007, 27 (2): 39–52.
 - [27] Hsu W-j, Dutta D, Helmy A. Mining behavioral groups in large wireless LANs [C]. In Proceedings of the 13th annual ACM international conference on Mobile computing and networking. 2007: 338–341.
 - [28] Eagle N, Pentland A S, Lazer D. Inferring friendship network structure by using mobile phone data [J]. *Proceedings of the national academy of sciences*. 2009, 106 (36): 15274–15278.
 - [29] Zheng J, Ni L M. An unsupervised learning approach to social circles detection in ego bluetooth proximity network [C]. In Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. 2013: 721–724.
 - [30] Do T M T, Gatica-Perez D. Groupus: Smartphone proximity data and human interaction type mining [C]. In 2011 15th Annual International Symposium on Wearable Computers. 2011: 21–28.
 - [31] Mtibaa A, Chaintreau A, LeBrun J, et al. Are you moved by your social network application? [C]. In Proceedings of the first workshop on Online social networks. 2008: 67–72.
 - [32] Ma C, Cao J, Yang L, et al. Effective social relationship measurement based on user trajectory analysis [J]. *Journal of Ambient Intelligence and Humanized Computing*. 2014, 5 (1): 39–50.
 - [33] Zheng Y. Trajectory data mining: an overview [J]. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2015, 6 (3): 29.
 - [34] Itakura F, Umezaki T. Distance measure for speech recognition based on the smoothed group delay spectrum [C]. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*. 1987: 1257–1260.
 - [35] Gonzalez M C, Hidalgo C A, Barabasi A-L. Understanding individual human mobility patterns [J]. *Nature*. 2008, 453 (7196): 779–782.
 - [36] Wegmann H. Image orientation by combined (A) AT with GPS and IMU [J]. *INTERNATIONAL ARCHIVES OF PHOTOGRAMMETRY REMOTE SENSING AND SPATIAL INFORMATION SCIENCES*. 2002, 34 (1): 278–283.
 - [37] Kalman R E. A new approach to linear filtering and prediction problems [J]. *Journal of Fluids Engineering*. 1960, 82 (1): 35–45.
 - [38] Zheng Y, Zhou X. Computing with spatial trajectories [M]. Springer Science & Business Media, 2011.
-

- [39] MacQueen J, et al. Some methods for classification and analysis of multivariate observations [C]. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967: 281–297.
- [40] Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. [C]. In Kdd. 1996: 226–231.
- [41] Zhou C, Frankowski D, Ludford P, et al. Discovering personally meaningful places: An interactive clustering approach [J]. ACM Transactions on Information Systems (TOIS). 2007, 25 (3): 12.
- [42] Rodriguez A, Laio A. Clustering by fast search and find of density peaks [J]. Science. 2014, 344 (6191): 1492–1496.
- [43] Liu H, Schneider M. Similarity measurement of moving object trajectories [C]. In Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming. 2012: 19–22.
- [44] Ratanamahatana C A, Keogh E. Everything you know about dynamic time warping is wrong [C]. In Third Workshop on Mining Temporal and Sequential Data. 2004.
- [45] Singelis T M. The measurement of independent and interdependent self-construals [J]. Personality and Social Psychology Bulletin. 1994, 20 (5): 580–591.
- [46] Farrahi K, Gatica-Perez D. What did you do today?: discovering daily routines from large-scale mobile data [C]. In Proceedings of the 16th ACM international conference on Multimedia. 2008: 849–852.
- [47] 杨若松. 基于轨迹数据的用户关系强度度量方法研究 [D]. [S. l.]: 万方数据资源系统, 2016.
- [48] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781. 2013.
- [49] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. the Journal of machine Learning research. 2003, 3: 993–1022.
- [50] Manku G S, Jain A, Das Sarma A. Detecting near-duplicates for web crawling [C]. In Proceedings of the 16th international conference on World Wide Web. 2007: 141–150.
- [51] wiki. Inversion. [http://en.wikipedia.org/w/index.php?title=Inversion_\(discrete_mathematics\)&oldid=650315847](http://en.wikipedia.org/w/index.php?title=Inversion_(discrete_mathematics)&oldid=650315847). 2014.

作者在学期间取得的学术成果

发表的学术论文

- [1] Ruosong Yang, Dianxi Shi. SASLL: A System Annotating Semantic Label of Location. The 7th International Symposium on UbiCom Frontiers Innovative Research, Systems and Technologies.(EI 检索)
- [2] Dianxi Shi, Ruosong Yang. Measuring User Relationship Strength Using a Model Based on Hierarchical Voting. 2016 PerCom. (已投稿)

