项目编号 201610486094

武汉大学国家大学生创新创业训练计划项目中期报告

商场中移动群组消费基于安卓手机传感器的数据采集及聚类算法研究

院（系）名 称：国际软件学院

专 业 名 称 ： 软件工程

学 生 姓 名 ： 许琳 张泽宇 胡浙捷

胡成 杨耀航

指 导 教 师 ： 朱卫平 副教授

二〇一七年三月

INTERIM REPORT OF PLANNING PROJECT OF INNOVATION AND ENTREPRENEURSHIP TRAINING OF NATIONAL UNDERGRADUATE OF WUHAN UNIVERSITY

The Data Acquisition and Clustering Algorithm Research Based on Android Sensor of Mobile Group Consumption in Mall

College ： Wuhan University

Subject ： Software Engineering

Name ： Xu Lin Zhang Zeyu Hu Zhejie

Hu Cheng Yang Yaohang

Director ： Zhu Weiping Associate Professor

March 2017

郑 重 声 明

本项目组呈交的初期报告，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我们所知，除文中已经注明引用的内容外，本报告的研究成果不包含他人享有著作权的内容。对本报告所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本报告的知识产权归属于培养单位。

项目组签名：_____ 日期：_____

导师签名：_____ 日期：_____

摘　　要

随着社会经济快速的发展，人们的消费方式也越来越多样化，而在公共场合中群组消费占着较大的比重，群组消费的决定相对单独的消费更依赖与群组成员的交流沟通。本文主要以基于移动设备交流的进行的群组性消费活动为研究方向，利用数据收集系统采集商店和人的数据，通过计算机算法识别群组信息，得到群组中最大影响力的领导者以及每位成员角色，从而有针对性的向商场实行个性化推送。

在本项目的中期，我们完成的任务是设计了基于群组内人员交互行为的聚类方法，基于 wifi 信号强度的室内定位的实现，并设计了实验进行实际数据的采集验证了提出的聚类算法。关于聚类算法，首先我们定义了动作种类的集合，基于行为聚类算法就是将用户每个窗口的动作进行行为动作分类，并将得到动作序列进行编辑距离计算得到的分组结果。基于行为的聚类方法的准确度达到了 96.6%。数据采集是使用的安卓手机进行的传感器以及定位的数据采集，采集系统主要包括安卓客户端以及服务器。

关键词：交互行为；数据采集；传感器；RSSI； Wifi 定位； 行为识别； 编辑距离；聚类

ABSTRACT

With rapid development of social economy, people's consumption is becoming more and more diversified, and in public consumption accounts for a larger proportion in the group, group decision relatively separate consumption more dependent on communication with group members.This article mainly based on mobile communication group of consumer activity as the research direction, using the data store and data collection system, via computer algorithms to identify group information, so as to get the group leader of the biggest influence and role of each member, and targeted to the stores to implement personalized push.

In the middle of the project, The task we finished is that designing method based on personnel in the group interaction of clustering method, implementing localization based on the wifi signal strength indoor and designing the experiment for actual data collection are presented to verify the effectiveness of the proposed clustering algorithm. As for clustering algorithm, we define a collection of communication action types. The content of clustering based on behavior consists of two steps. First, we need to classify users' behaviors in each window which make an action sequence for each user. Then we use the editing distance to compute the dissimilarity of each two user thus we get a distance matrix. The accuracy of clustering method based on behavior was about 96.6%.Data acquisition is using the android mobile phone to get sensor localization including the android client and server.

Key words: Interaction behavior; Data acquisition; Sensor; RSSI; Wifi positioning; Behavior identification; Editing distance; clustering

目　录

第 1 章　绪论

1.1　研究背景

移动群组消费是指一群人，例如家人，伴侣，同事朋友等，基于移动设备交流的进行的群组性消费活动。群组性消费的决定相对个人单独的消费更依赖与群组成员的交流沟通，往往群组中一个主要人物会决定这个群组最终的走向，因此想要研究群组消费要关注群组内交流行为并且找出群组中影响力最大的人。群组性消费在生活中并不少见譬如大学生社团班级朋友聚餐、情侣闺蜜逛街、家长孩子出游等。据调查，当人在公共场合时，百分之七十的时间是和其他人一起的 [1]，也就是说大部分情况下他们是进行的群组性消费。所谓“众口难调”，往往这时候总会遇到吃什么，玩儿什么，买什么，什么便宜，哪家团购等一系列决策问题。

近年来，越来越多的人使用手机等移动设备进行群组消费。他们使用手机等移动设备扫描，搜索并获得商品售卖信息和预定各种服务和商品。InMobi Insights 团队的的一个调查显示，46%的人已经用他们手机进行过购买，80%的人表示会在未来一年进行此项活动 [2]。在这种趋势下，许多公司也开始对移动群组消费加大市场力度，2014 年全球此类广告投入已达 32.7 亿美元，占互联网广告总数的 1/4 [3]。

当前应用市场上比较受欢迎的美团，百度糯米，大众点评，以及喵街等只考虑了向用户推送优惠，但不具有即时性，也不具有群组针对性。但是地点是影响移动消费的重要因素[4,5]，群组性消费这一概念也是十分重要，[8]因为单独的个人会因为他们是社会角色一部分进行托管式的群组行为。

此外，在技术层面上，也有许多相关研究。为了分析移动群组消费，首先需要的是一个有效的数据收集系统。典型的要被收集的数据包括在一段时间内人群的轨迹和行为。有多种可用于此的技术，例如 wifi 接入点到用户手机的 wifi 信号可以被收集。每个手机附属的 RFID 标签也可以用来分析。当然移动设备中的传感器也可以用于测定群组移动轨迹和行为[9]。还有就是采集商场中摄像头数据，通过视频流进行识别[10]。此外通过这些数据能否很好的识别出群组关系以及人与人之间交流行为也是实现移动群组消费的关键。[11]提出了移动社交网络环境下的真实社会关系估计，它基于用户 GPS 轨迹挖掘语义化访问地点，然后结合语义化地点以及临近特征估计用户间的社会关系类型，结果表明是可以准确估计家人同事

朋友等社会关系。这些研究均表明了商场移动群组消费系统设计具有它的可行性。

1.2 研究意义

通过对移动群组消费的研究，一方面我们能够对商场中人的群组的交互和消费行为有更加深刻的了解，在行为学方面；另一方面通过移动设备研究群组消费将线上线下人的社会关系结合，促进了基于互联网的移动社交。

移动群组消费是一个新的概念，通过对移动消费者群组的识别和影响力研究。商家能够针对群组消费提出更好的营销策略，减少广告投放随意性，降低商家广告投入成本。

对于群组中消费者本身来说，我们研究的群组性移动消费的推送能够更好的解决“众口难调”的问题，促进用户的理智消费。

1.3 研究方法

通过对相关研究的总结我们的项目主要分为这样几个部分：

(1)调查当前商场中人群信息数据收集方法

(2)设计商场人群及商店模型模拟器软件

(3)基于模式识别聚类设计群组聚类算法并使用模拟数据验证准确度

(4)选择实验场景，设计实验情节，完成实际数据的采集并验证提出的算法准确度

(5)基于复杂网络知识建立群组影响力模型识别群组人员影响力

(6)将群组聚类算法和影响力算法应用到团购推销中

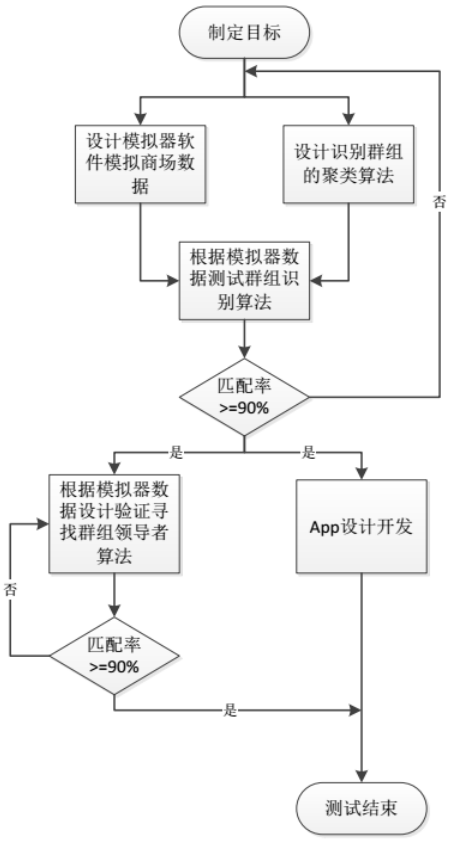

图 **1.1** 项目方案流程

针对于算法我们针对数据的一些特征进行算法设计，然后将实际的数据用于算法的验证。在对算法进行评估之后再根据数据的特征进行进一步的改进，以获得较高的准确率。

1.4 研究内容

中期阶段我们提出了我们自己的基于群组内交互行为的聚类算法，并通过实际实验我们采集了数据，通过运行实际数据我们得到了较高的准确度，主要研究内容如下：

(1) 设计出基于群组内人员交互的聚类算法

(2) 针对设计出的算法，在实际环境中进行了实验设计并采集到了数据

(3) 将实际数据用于算法的验证

(4) 研究基于 WIFI 信号强度的室内定位技术，并在同一实验地点完成了指纹库的建立和测试数据的采集，最后得到定位结果。

第 2 章　传感器数据采集实验

为了进一步研究在商场中人群的聚类算法设计，并提出自己的聚类算法。我们考虑了从多个方面设计算法。由于之前关于人群聚类的算法大致分为几类。首先是空间距离的聚类，考虑人员当前位置和轨迹，通过基于密度的算法得到聚类结果[17]。但是由于商场中也许存在比较拥挤的状况，加上定位数据本身存在误差较大的局限性，用户的位置很大程度上存在重叠且十分不可靠，所以仅仅依靠用户位置进行聚类是不够的。所以就有使用用户行为数据进行分组的算法出现[18]，使用佩戴的传感器采集到加速度和方向等数据计算实验人员的行为相似性，再进行分组。但是实际情况下，人员同一群组中人员的运动模式会存在相似但是更多的时候会有一定的交互，造成行为的不相似性，这个时候只针对一致性比较强的场景下的算法就不适用了。所以本项目的中期阶段，我们完成的是基于人员行为识别的聚类算法，这个算法的考虑了人员之间的交互行为。本章的主要内容是介绍为算法验证做的实验过程。下面主要从采集程序和采集过程两个方面来介绍。

2.1 传感器数据采集系统

图 **2.1** 安卓端采集 UI

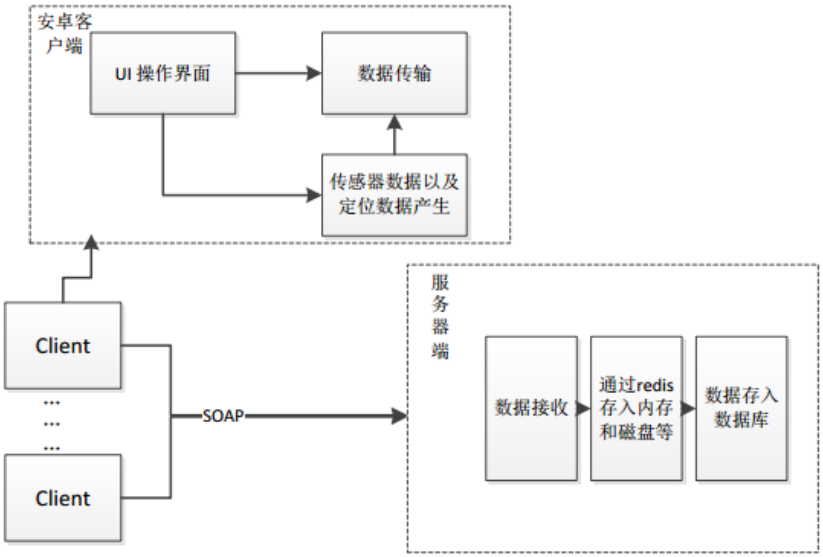

图 **2.2** 数据采集系统图

传感器采集系统是一个典型的 Client-Server 模型的系统，其中在安卓端实现了对加速度传感器，方向传感器，陀螺仪，磁传感器以及 wifi 信号强度 RSSI（接收信号强度指示器 Received Signal Strength Indicator）的采集。这些数据能够实时地存入服务器端的数据库，以便今后实时地分析当前运动人员地分组情况。

根据 IEEE 802.11 系统定义，RSSI 是 WIFI 衡量接收信号强度的一个相对值。RSSI 表示的是在信号传播丢失之后的无线信号接收强度。因此 RSSI 数越高，信号越强。一般 RSSI 以负值形式表示，值越接近于 0，接受信号越强。早在 2000 年，就已经使用 RSSI 进行较为粗粒度的位置估计，而近年来越来越多更先进的技术用在了基于 wifi RSSI 的定位上，精度也相对提高，但是由于信号的不稳定性等因素，RSSI 并不总是能够提供足够准确的定位结果[13]。

2.2 基于行为的数据采集实验

2.1.1 行为定义

在日常生活中，群组之间的交互行为丰富多样，交互行为分为语言（verbal）和非语言（nonverbal）交互。我们将人的行为指的是人在运动过程中的各种动作，由于人的动作十分多，如果要对人的行为进行分类的话，将会有无数种。本实验主要考虑的行为是在商场情景下与群组分组相关的动作。考虑到实验环境的限制：视频信息难以获取，并且视频信息处理比较复杂，音频信息容易受到周围对象和环境

噪声的影响；并且该交互信息又要能够被传感器比较好的识别，因此在我们的实验中我们选取了几种常见的交互行为动作，如挥手、握手、拥抱、挽手、勾肩、行走、奔跑、坐下、起立、静止作为群组成员之间的交互行为。

2.1.2 实验设计

实验环境为实验室大厅，实验设备分别为三星和小米手机，手机型号和安卓版本号并不做限制。在实验对象的手机上安装数据采集客户端程序，用来获得手机的加速度数据和方向数据，一次实验结束后将本次实验数据上传到服务器。

实验过程中，分别将实验对象分为两组、三组和四组分别实验，在实验过程中，实验对象手持手机，并不限制手机的摆放位置。每个组有一名动作执行决定者，同时有一名记录员记录下执行各种交互行为动作的时间区间，作为最后的行为识别结果的比较，以确保行为的识别的正确性。实验过程中，每个组的执行的行为种类和次数不限。同一群组的对象会有分开活动再合并的过程。一次实验持续时间约为 9 分钟。图 2.3 是实验过程中的记录。

实验4 1-5 6-10分为两组				
1到5				
挥手	15:04:33-15:04:47		6到10	
握手	15:05:00-15:05:29		握手	15:04:00-15:05:09
拥抱	15:05:40-15:06:07		拥抱	15:05:16-15:05:42
挽手	15:06:35-15:07:46		勾肩	15:05:52-15:08:22
勾肩	15:08:10-15:09:10		挽手	15:08:29-15:10:43
奔跑	15:09:29-15:10:35		分开	6-8 9-10
坐下、起立	15:11:00-15:11:55		挥手	15:11:10-15:11:28
挽手	15:12:42-15:13:56		跑步	15:11:42-15:12:31
结束	15:14:14		坐下、起立	15:12:38-15:13:29
			结束	15:14:14

图 **2.3** 数据采集设计

2.1.3 实验结果

实验编号	实验说明	分组情况
3	两组　有交互	（ 1 2 3 4 5）（6 7 8 9 ）
7	三组　有交互	（1 2 3 4）（5 6 7 ）（9 11）
10	四组　有交互	（1 2）（3 4）（5 6 7）（8 9 11）
11	三组　有交互	（1 2 3 4）（5 6 7 8）（9 11）

图 **2.4** 数据采集设计图

上图为 2016 年 12 月 18 日进行数据采集的结果记录，在实验过程中 ID 为 10 的手机出现故障没有采集到数据，因为排除 10 号手机。实验 7 中 ID 为 8 的手机没有数据，因为在实验 7 中排除 ID 为 8 的手机。

第 3 章 基于 wifi 定位算法设计与实现

基于 wifi 的位置指纹定位算法只需 RSSI 作为指纹库的指纹，无需其他特征参数，在定位阶段进行指纹匹配，便可以估计位置。虽然无需测距的定位算法相比于基于测距的定位算法，定位误差较大，但是位置指纹算法降低了对定位系统的硬件要求，采用 RSSI 作为指纹库的特征参数是低成本的并且实现简单。

该算法分为两个阶段(如图 3.1)，数据采集建立指纹库和位置估算阶段。在数据采集阶段，将采集到的数据进行 k-means 聚类得到标准指纹库，降低了在位置估算阶段进行指纹匹配时的计算量；在位置估算阶段，将未知点指纹数据与指纹库匹配，对最近邻算法（最近邻算法：将实测指纹与指纹库中的指纹进行对比，并将相似度最大的指纹作为定位结果，该算法参考位置单一，定位结果不稳定，易产生较大误差）进行改善得到加权 KNNSS 算法（K-Nearest Neighbor in Signal Space），对未知位置进行估算，该算法又称为 KWKLF 算法。

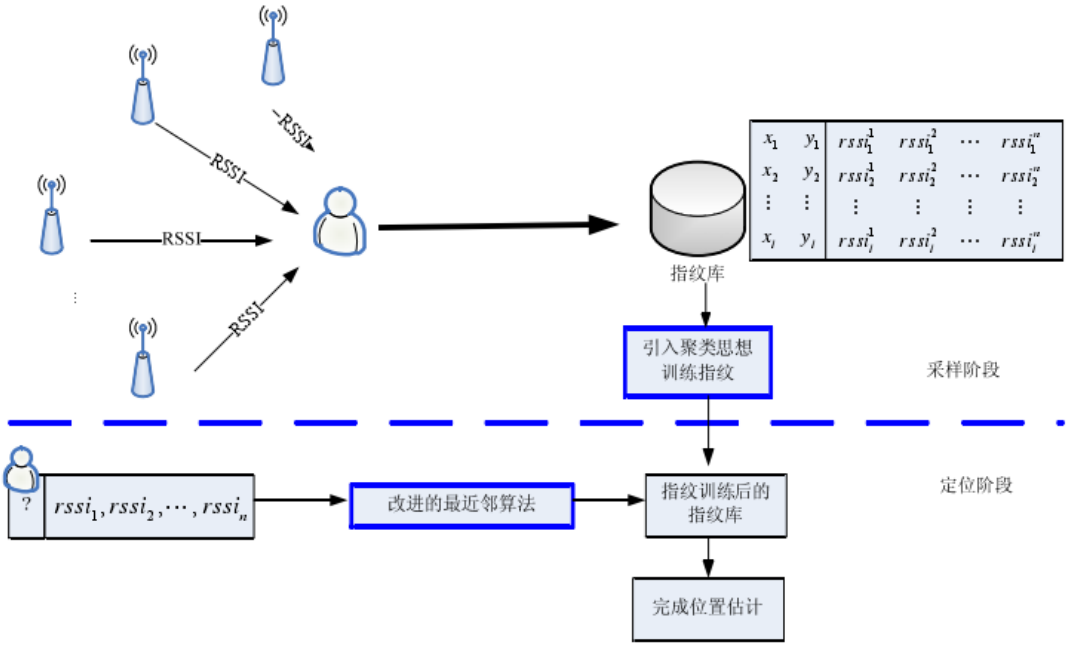

图 **3.1** KWKLF 算法工作原理

3.1 数据采集建立指纹库：基于 k-means 聚类

3.1.1 k-means 聚类

k-means 聚类是基于距离的聚类算法，以距离作为相似性的度量，其算法思想是根据现有的样本之间的相似度划分为 k 个子类，相似度较大的样品聚集在一起，

相似度比较小的样品彼此远离。k-means 聚类可以高效分类，使得整个指纹库中划分为不同的小类，减少位置指纹搜索空间。

3.1.2 数据采集

在实验室大厅，设置 6 个路由器（AP 点），以实验室地面一块瓷砖（瓷砖规格：0.445m * 0.445m）作为单位 1 进行坐标系建立，横纵每隔两块瓷砖进行采集指纹，即每个采样点有 6 个 RSSI 值总共采集 128 个采样点指纹，且记录采样点的位置，用(x ,y)表示。关于指纹的采集，使用了 4 部手机，每个手机采集 2 次，得到 8 个 RSSI 值，去掉最大两个和最小值取剩余四个的平均值将其作为最终指纹，将指纹和位置信息对应的存储在 txt 文件里进行下一步处理。

3.1.3 建立指纹库

将指纹库进行 k-means 聚类，以欧氏距离作为相似度的评价准则，距离相近的指纹聚集在一个子类，距离较大的指纹彼此远离。多次执行该步骤，直到聚类结束，指纹库变成具有 k 个子类的指纹样本空间。 该聚类过程具体如下：

(1)输入指纹库中的 128 个指纹和聚类个数 k(k≤L)（经实验比较得知，k 设为 15 最为合适）；从 128 个指纹中任意选择 k 个指纹作为初始的聚类中心

$$C=\left(FP_1^T,FP_2^T,\cdots,FP_k^T\right)^T$$

$$Loc=\begin{bmatrix}x_1 & y_1\\ x_2 & y_2\\ \cdots & \\ x_L & y_L\end{bmatrix}_{L\times 2}\quad Fp=\begin{bmatrix}rssi_1^1 & rssi_1^2 & \cdots & rssi_1^n\\ rssi_2^1 & rssi_2^2 & \cdots & rssi_2^n\\ \vdots & \vdots & \vdots & \vdots\\ rssi_L^1 & rssi_L^2 & \cdots & rssi_L^n\end{bmatrix}_{L\times n}.$$

(2)对于剩下的(128-k)个指纹,计算每个指纹到每个类中心的指纹距离（指纹距离：两个指纹之间各对应 RSSI 值差值平法和开方），若第 i 个剩余指纹距第 j 个类中心距离最近（指纹距离），则将第 i 个指纹归到第 j 个类中。

(3)重复第(2)个步骤，将剩下的的指纹分配完成，形成 k 个聚类 $G_1,G_2...G_j...G_k$，每个类 G_j 都包含其聚类中心，和属于该类的指纹成员及其个数 n_j。

(4)根据公式

$$rssi_{ji}^{*}=\frac{1}{n_j}\sum_{rssi_i\in G_j}rssi_i$$

计算新的聚类中心，其中 $rssi_i$ 表示 G_j 类中的第 i 个 RSSI 值。计算每个类的类中心，得到新的聚类中心

$$C^* = \left(FP_1^{T*}, FP_2^{T*}, \cdots, FP_k^{T*} \right)^T$$

(5)若 C*= C，即相邻两次的聚类中心相同，分类趋于稳定，聚类结束，当前的 $G_1,G_2...G_j...G_k$ 代表了最终形成的聚类。否则令 C= C* ，即更新类中心，返回第(2)步骤继续执行聚类过程。

3.2 定位阶段：改进的加权 KNNSS 算法

(1)将实测指纹 IF=($rssi_1,rssi_2, ... ,rssi_n$)与训练之后的指纹库 KFp 进行匹配，计算 IF 与每个类中心的距离记为 DIS=[$d_1,d_2, ... ,d_k$]

(2)寻找 min(DIS)对应的类，记为 $G_{SPECIAL}$。

(3)计算实测指纹 IF 与 $G_{SPECIAL}$ 中的每个指纹的距离，记为 DIS=[d_1*,d_2*, ... ,d_{ng}*], 其中 n_g 表示 $G_{SPECIAL}$ 中指纹的个数。（若 n_g 小于等于 2 则选择 DIS 中第二小的类，直至对应 n_g 大于 2）

(4)将 DIS 按照从小到大的顺序排列，取最小的 $n_g/2$ 个距离，并将这 $n_g/2$ 个距离对应的指纹选定作为参考指纹，其对应的位置坐标作为参考坐标。

(5)分别计算每个参考坐标的平均值和标准差,

$$\overline{rssi_i} = \frac{1}{n}\sum_{j=1}^{n} rssi_i^j \qquad s_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(rssi_i^j - \overline{rssi_i} \right)^2}$$

，令 $v_i = s_i \Big/ \overline{rssi_i}$，则权重系数为 $\omega_i = v_i / \sum_{i=1}^{k} vi$ 其中 i=1,2,..., $n_g/2$

(6)实测指纹位置为 $x_{estimate} = \sum_{i=1}^{k} \omega_i x_i \qquad y_{estimate} = \sum_{i=1}^{k} \omega_i y_i$

3.3 结果分析

在定位阶段，采集了多个实验点的指纹进行匹配分析，其中一些点定位准确度达到 1m 以下甚至正好匹配到其真实点，具体如图 3.2 。由于实验时以地板为单位 1 进行实验，因此该图中 RMSE 的大小是地板的块数，也就是图中 RMSE*0.445 得到的值为以米(m)为单位的误差值；其中传统算法指的是建立指纹库时未进行 k-

means 聚类得到的定位结果。

由于一些噪声点及 k-means 聚类时 k 值的影响，一些实验点的定位准确度较低，导致平均误差约为 2.6m ，可以在下一步进行轨迹预测时对明显偏离点进行对比调整及对寻找自适应 k 值做进一步研究。

传统算法：

MinRMSE: 0.0

MaxRMSE: 17.88854381999832

AveRMSE: 6.581355317729607

RMSE	Number
0<= RMSE <5	55
5<= RMSE <10	40
10<= RMSE <15	29
15<= RMSE	4

注：RMSE为均方根误差

KWKLF算法：

MinRMSE: 0.0

MaxRMSE: 16.1245154965971

AveRMSE: 5.816607856105122

RMSE	Number
0<= RMSE <5	62
5<= RMSE <10	48
10<= RMSE <15	15
15<= RMSE	3

图 **3.2** 实验结果分析

第 4 章 基于行为的分组算法

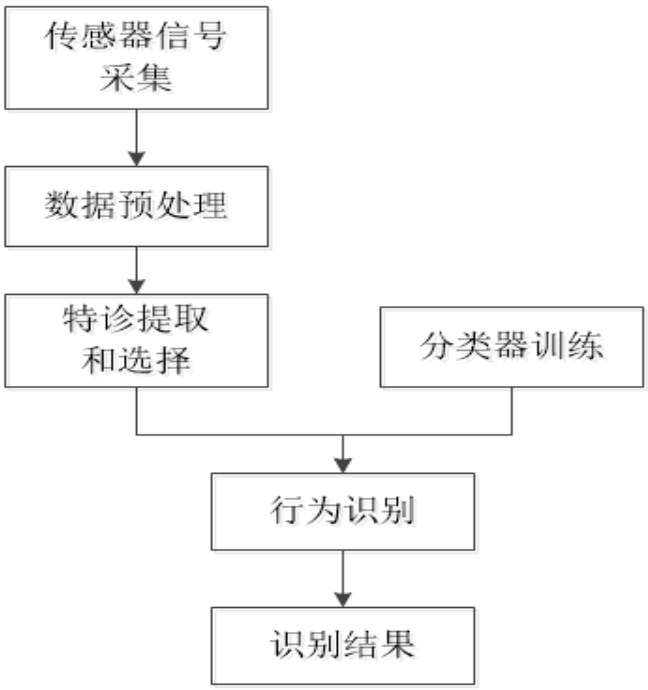

图 **4.1** 基于行为分组算法流程

4.1 基于行为分类的原理概述

商场中消费者的行为可分成多个类别，它们分别有“静止”“慢走”“快走”“奔跑”“坐下”“站起”“挽手”“勾肩搭背”“打招呼挥手”“握手”和“拥抱”。

手机传感器可以获取用户的三轴加速度信息。可以先通过初步的方法计算和分析加速度数据，来判断手机目前处在用户身上的哪一个部位。比如可能在上衣的口袋里，或者用户拿在手上，再或者放在裤子口袋中。接下来，根据推测的不同位置进一步对传感器的加速度数据进行分析，判断当前的用户行为状态。论文[14]中通过早期非智能手机的三轴加速度传感器成功区分人员的静止、行走、跑动、坐下和站立这几个动作。借鉴其中的方法，我们又扩展算法使其能够分辨“挽手”“勾肩搭背”“打招呼挥手”“握手”和“拥抱”这几个动作。

分析出这些动作后，可以根据这些动作进行消费者的分组。因为一个群组中的消费者的行为动作具有相似性或者配对性。相似性表现在，一个群众中的所有成员可能一起静止，一起行走，一起跑动，一起坐下或站立（相近时间范围内），或者

一起挥手、拥抱。配对性表现在勾肩搭背时的动作的配对上，还有挽手的动作的配对上。

对于每个消费者，实时的行为动作确定后，将会在根据距离的初步分组后进行分析判断人与人之间的行为动作是否相似或者匹配，然后再优化分组——不在组内的将被剔除，根据距离重新分组，再判断行为动作的相似和匹配。如此重复，直到算法迭代结束。

4.2 行为识别的特征提取

图 4.1 时是基于行为分组算法的流程，第一步我们通过采集系统得到了实验期间每个实验人员的加速度传感器数据。预处理过程，在基于传感器行为识别中，绝大多数的工作都使用滑动窗口对数据进行分割。在基于滑动窗口的数据分割技术中，如何合理的选择滑动窗口的大小是其核心问题。滑动窗口的大小关系到系统的响应时间和识别精度。交互动作都是周期性动作，单个动作的周期大小为 2 秒，因此本次实验取的滑动窗口为 2 秒，两个相邻窗口之间重叠半个窗口被验证是比较成功的。

然后需要进行特征的提取，选择时域上的特征和频域上的特征作为特征值。我们用 n 来表示一个时间窗口的大小(即窗口内数据个数),用 i 来表示第 i 个数据，我们选取的时域特征有平均值、标准差、最大和最小值。频域特征通过快速傅里叶变换得到，通常被用来发现信号中的周期性信息。频域上的我们选取的特征为均值、标准差、偏度和峰度。

由于加速度分为 x, y, z 三个轴的数据，所以在进行特征提取的时候我们将三个轴的数据以及 x, y, z 的综合量级值（即三维的平方和的平方根）共四个原始数据分别计算了这 8 个特征，最后得到 32 维的特征向量。以下是使用的特征计算公式：

均值： $\mathrm{mean}=\frac{1}{\mathrm{n}}\sum_{i=1}^{n}a_i$

标准差： $std=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i-mean)^2}$

最大值： $\max=\max(a_i),i\in\{1,2,...,n\}$

最小值： $\min=\min(a_i),i\in\{1,2,...,n\}$

设 C(i)是第 i 个窗口的频率幅度值序列，N 表示窗口内数据总数

频域均值：$$\mu_{amp}=\frac{1}{N}\sum_{i=1}^{N}C(i)$$

频域标准差：$$\sigma_{amp}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[C(i)-\mu_{amp}\right]^2}$$

偏度：$$\gamma_{amp}=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{C(i)-\mu_{amp}}{\sigma_{amp}}\right]^3$$

峰度：$$\gamma_{amp}=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{C(i)-\mu_{amp}}{\sigma_{amp}}\right]^4-3$$

4.3 动作分类方法

将采集的交互行为数据提取特征值后，放入分类器中进行训练。在此我们选用了 SVM、随机森林、决策树、KNN 分类器分别进行交互行为的训练和测试。SVM 为台湾大学林智仁(Chih-Jen Lin)博士等开发设计的一个通用 SVM 软件包，在这个包中可以直接调用相关函数对数据进行归一化处理，遍历得到最优的参数值。对于随机森林算法通过调用 python 中的 sklearn 机器学习包中的随机森林算法，参数使用默认值得到分类结果。对于决策树和 KNN 算法，使用 Weka 机器学习算法工具包，实验结果表明，当 k=1 时，KNN 算法能得到比较高的分类准确度。

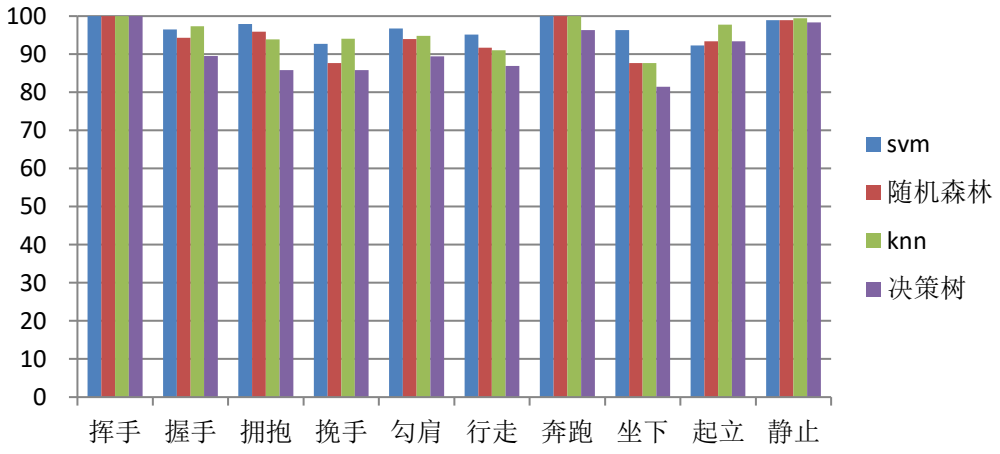

图 **4.2** 分类算法的准确度结果

4.4 行为分组

在对用户每个窗口上动作进行了分类之后，我们得到每个用户在时序上的一个动作序列，所以分组就变成了动作序列的匹配。于是使用“最长公共子序列”尝试寻找不同消费者之间的行为轨迹的公共子序列，以此判断人与人之间行为轨迹的相似性。

实验证实，利用“最长公共子序列”从实验数据中找出的公共子序列太过零散，原因是公共子序列的长度都比较短，使得数量过多。而且如果轨迹 A 有子序列 abc（仅一个），而轨迹 B 有子序列 abc 多个，并假设 A 和 B 的最长公共子序列就是 abc，但并不能说明轨迹 A 和 B 的轨迹就是相似的。

最终，我们选择使用编辑距离（编辑距离指两个字串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。）来比较轨迹序列之间的相似程度。编辑距离越小，两个串的相似度越大。

实验证明，使用编辑距离算法能有效比较轨迹序列之间的相似程度。

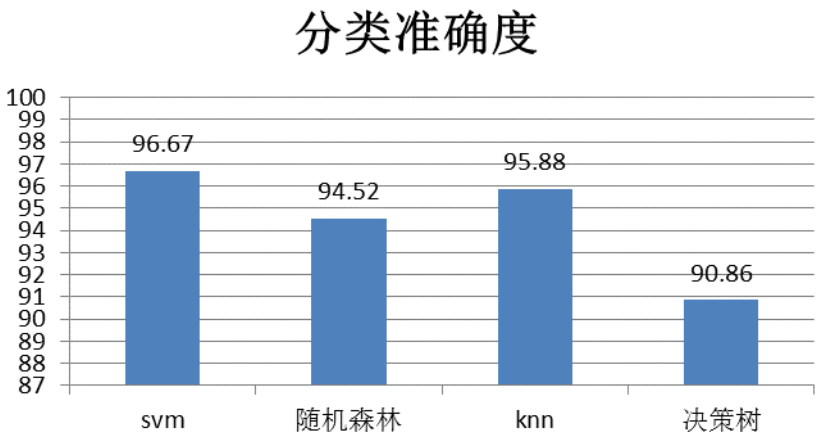

图 **4.3** 聚类算法结果

4.5 准确度结果计算

通过计算序列之间的编辑距离，我们能得到人与人之间的编辑距离的距离矩阵。针对这个距离矩阵我们用了两种方法来计算分组的准确度。

4.5.1 直接阈值准确度计算

通过使用一个阈值，来确定每个窗口里面的两两人员之间是否是一组，如果这

样的窗口数超过了窗口总数的 70%则认为这两个人员是一组，最后再计算准确度。这里群组划分准确度的计算方法为：假设判定为一个群组的两个对象本身就是一个群组的对象的个数为 TP，判定不是一个群组的两个对象本身不是一个群组的对象的个数为 TN，群组划分准确度为 TP 和 TN 之和与可能所有可能关联的个数的比值。

4.5.2 基于密度的聚类方法

我们实验小组考虑使用 DBSCAN 和 DJ-Cluster 算法[15]。DBSCAN（Density-Based Spatial Clustering of Applications with Noise）聚类算法，它是一种基于高密度连通区域的、基于密度的递归型的聚类算法，能够将具有足够高密度的区域划分为簇，并在具有噪声的数据中发现任意形状的簇。DJ-Cluster 是 DBSCAN 的一种变体，这个方法的相对 DBSCAN 是迭代性的而不是递归形式的，所以 DJ-Cluster 更适合大量数据的处理。

[15]提出 DJ-Cluster 是相对于 DBSCAN 更简单，占用的内存较少，在部分参数选取情况下 DBSCAN 使用的内存较多并且会拖慢运行速度。

通过对 DBSCAN 和 DJ-Cluster 进行实验比较，我们发现，两个算法得出的结果基本相似，DBSCAN 相比于 DJ-Cluster 更占一些内存，但是 DBSCAN 相对与 DJ-Cluster 运行时间更少。考虑到我们的群组识别需要进行实时的识别，实验小组最终决定使用 DBSCAN 的方法进行密度聚类。目前还未得出计聚类算法的准确度结果。

第 5 章　结论和展望

在项目中期，我们初步完成了以下几个目标：

(1)设计了基于群组内人员交互行为的聚类算法，目前还未有人提出类似的群组分组算法，相对来说这属于我们的创新。

(2)新提出的聚类算法的准确度达到了 96.6%。

(3)将过去设计实现的聚类算法进行总结，并撰写了论文初稿

在初期阶段我们设计的两种算法分别获得了 65%-85%和 80%-100%的准确率，但是这两种方法主要针对于无交互的数据。本阶段我们提出的算法可以适用于有交互的数据，同时我们的算法还可以避免在分布式情况下大量数据的传输，因为在每个服务器中我们可以计算出每个人员动作序列并将动作序列进行传输，动作序列相对于原始数据大大减少了数据量。这相对于基于 cross-correlation 的方法有了更多的改进。

接下来我们的工作将对新提出的聚类算法进行多个角度的验证，包括使用基于密度的聚类算法对编辑距离矩阵进行聚类，并使用 F-measure 来计算分组准确度，将我们的算法和相关研究的算法进行比较。此外我们还会对聚类算法的继续改进，将室内定位技术加入群组聚类，将人员运动轨迹和行为聚类融合，使我们的人群分组算法适用于更多的场景。

此外，对群组进行影响力建模，是我们后期需要进行的另一个任务，这是对对人群分完组的一个拓展。当然做这些算法更多的是希望可以有实际的用处，所以我们会同时开发出安卓的应用将我们的算法实际运用到生活中。

参考文献

[1] Moussaid M.;Perozo N.;Garnier S.;Helbing D. Theraulaz, G. The Walking Behaviour of Pedestrian Social Groups and Its Impact on Crowd Dynamics[J].PLoS ONE，2010，5，1-7.

[2] InMobi Insights Team. Global Mobile Media Consumption[EB/OL]. http://info.inmobi.com/rs/inmobi/images/Global_Mobile_Media_Consumption_Wave_2_Whitepaper.pdf, 2013.

[3] eMarketer. Mobile Ad Spend in China Hits $7 Billion This Year[EB/OL]. http://www.emarketer.com,2014.Ghose, A.

[4] Goldfarb A.; Han S.P. How Is the Mobile Internet Different?[J]. Search Costs and Local Activities. Information Systems Research, 2013, 24, 613－631.

[5] Molitor D.; Reichhart P.; Spann M.; Ghose A. Measuring The Effectiveness of Location-based Advertising, A Randomized Field Experiment[EB/OL].http://www.fox.temple.edu/cms/wp-content/uploads/2015/01/mksc-crowd.pdf, 2014.

[6] Luo, X.; Andrews, M.; Fang, Z.; Phang, C.W. Mobile Targeting[J]. Management Science, 2013, 60, 1738－1754.

[7] 阿里 O2O 再出新招 “喵街”年内欲完成 500 家商城上线[EB/OL],http://news.xinhuanet.com/tech/2015-05/06/c_127769048.htm.

[8] Fisher, R.J. Group-Derived Consumption: the Role of Similarity and Attractiveness in Identification With a Favorite Sports Team[M]. Advances in Consumer Research, 1998, 25, 283－288.

[9] 张希伟,戴海鹏,徐力杰,陈贵海.无线传感器网络中移动协助的数据收集策略[J].软件学报，2013,24(2):198-214

[10] 李娜，方卫宁. 基于视频流的地铁人群目标识别[J]-北京交通大学学报（自然科学版）,2006(1).

[11] 吕明琪，王琦晖，胡克用. 移动社交网络环境下的真实社会关系估计[J].计算机应用与软件,2015(1).

[12] Mazzon R, Poiesi F, Cavallaro A. Detection and tracking of

groups in crowd[C]//Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on. IEEE, 2013: 202-207.

[13]Received signal strength indication, https://en.wikipedia.org/wiki/Received_signal_strength_indication

[14] Kawahara Y, Kurasawa H, Morikawa H. Recognizing User Context Using Mobile Handsets with Acceleration Sensors[C]// IEEE International Conference on Portable Information Devices, 2007. Portable. IEEE, 2007:1-5.

[15] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. Discovering personal gazetteers: an interactive clustering approach. In Int. workshop on Geographic information systems. ACM, 2004

[16] 王淑婷.基于位置指纹的 WiFi 定位算法研究.吉林大学：通信工程学院，2015

[17] M. Wirz, P. Schläpfer, M. B. Kjærgaard, D. Roggen, S. Feese, and G. Tröster, ″Towards an online detection of pedestrian flocks in urban canyons by smoothed spatio-temporal clustering of GPS trajectories,″ in Proc. 3rd ACM SIGSPATIAL Int. Workshop Location Social Netw., 2011,pp. 17–24

[18]Dawud Gordon, Martin Wirz, Daniel Roggen, Gerhard Tröster, and Michael Beigl. 2014. Group affiliation detection using model divergence for wearable devices. In Proceedings of the 2014 ACM International Symposium on Wearable Computers (ISWC '14). ACM, New York, NY, USA, 19-26