**基于蜂窝网络大数据的用户行为分析及预测**

**摘 要**

近年来，随着信息技术的高速发展，移动互联网、智能物联网等相关技术的进步, 人们产生和采集数据的能力越来越强，数据的规模、维度、类型也随之增长而日益复杂。大数据早已在经济、政治、文化、军事领域以及金融、通信等行业发挥着越来越重要的作用。信息的数字化也推动了数据量的大量增长，在移动互联网飞速发展的今天，社交网络、在线音乐、流媒体直播等平台产生的数据量疯狂增长，无处不在的商业、工业和社会活动正在源源不断的产生各种各样的数据，一个数据爆炸的时代早已来临。在这样的背景下，随着越来越多的手机用户从主要使用传统的语音和短信业务转而投入到数据业务的阵营中，数据业务占比越来越大。 蜂窝网络中业务的增长伴随着流量的增长，产生了大量的用户行为数据，从这些数据中挖掘出用户的行为模式将有效地促进蜂窝无线网络的设计、规划与优化，并有效提升网络性能。

目前国内外对蜂窝网络下的用户行为分析已进行了大量的研究，但仍存在两大问题：一方面，大部分的研究工作都是基于3G的蜂窝网络数据， 对于4G蜂窝下用户行为的研究工很少；另一方面，蜂窝网络下用户行为的研究虽然很多，但大多数都是针对用户行为的某一个维度， 比如说单独从移动性的维度或者从业务使用量的维度出发，没有综合考虑， 从而忽略了它们之间的关联性。针对上述两大大问题，本论文基于真实的蜂窝网络4G用户行为数据， 从数据量使用模式、 业务使用模式以及移动性三个角度出发，探索三者的规律及关联性，并基于这些规律进行用户行为的预测 。本论文的主要工作以及主要贡献为以下三方面：

一、 基于蜂窝网络的用户数据量模式挖掘：首先基于真实的数据，利用spark大数据处理平台，进行亿级别的大规模数据清洗、转换、聚合，提取出用户的行为轨迹。利用用户的行为轨迹，进一步得到数据量的使用规律，发现少部分的用户使用的绝大部分的蜂窝网络流量。基于这个规律，定义的激活概率和激活状态转移概率，得到数据量使用模式并进行用户的群体划分，将用户划分成“重要用户”和“普通用户”。更进一步的，比较了“重要用户”和“普通用户”在移动性和业务使用上的差异。最后使用聚类算法，找到“重要用户”的使用流量占比最显著的一些应用。

二、 基于蜂窝网络的的业务模式挖掘及预测：首先研究了业务模式的基本特征， 例如用户在基站下的出现次数， 访问基站数， 访问业务类型数等。其次，进行用户业务轨迹建模，具体的，将用户一天的业务行为建模成一个向量，即用向量取表示用户的业务轨迹。利用聚类算法，根据轮廓系数确定最佳的聚类个数，挖掘出用户的业务模式，统计不同类别的用户数，并比较不同业务模式之间的差异。最后，基于挖掘出的用户模式，使用改进后的CNN模型，进行用户业务模式预测，并与传统的模型进行比较。

三、用户的移动模式与业务模式的关联分析及预测：首先探索业务模式也和移动模式之间的关系。主要从一下两方面进行分析：第一，分析访问基站数和业务使用之间的关系；第二，主要基站的定义，及“舒适区”的提取，分析“舒适区”的业务使用规律。其次对每个用户，识别出家和工作地点，计算家和工作地点的距离，并比较不同发达程度的城市家和工作地点的距离，挖掘出城市的发展程度和工作距离之间的关系。最后，我们使用基于深度学习的预测算法，预测用户使用的业务类型。具体的，根据每条记录的时空信息以及上下行流量，上下行发送包等参数，进行特征的提取。利用DeepFM算法进行点击率的预估，并与传统算法进行对比，实现用户业务点击预测。

**关键词**：用户行为分析 业务使用模式 移动模式 深度学习 用户行为预测

目录

[目录 III](#_Toc514323867)

[第一章 绪论 1](#_Toc514323868)

[1.1 研究背景与意义 1](#_Toc514323869)

[1.1.1 研究背景 1](#_Toc514323870)

[1.1.2 研究意义 1](#_Toc514323871)

[1.2 国内外研究现状 1](#_Toc514323872)

[1.3 研究内容和创新点 1](#_Toc514323873)

[1.4 论文结构安排 1](#_Toc514323874)

[第二章 基于蜂窝网络大数据的用户数据量模式挖掘 1](#_Toc514323875)

[2.1 概述 1](#_Toc514323877)

[2.2 用户行为轨迹 1](#_Toc514323878)

[2.2.1 数据预处理 1](#_Toc514323879)

[2.2.2 用户行为轨迹提取 1](#_Toc514323880)

[2.3 用户数据量模式挖掘 1](#_Toc514323881)

[2.3.1 问题描述 1](#_Toc514323882)

[2.3.2 用户数据量使用规律的发现 1](#_Toc514323883)

[2.3.3 用户数据量使用基本特征的提取 1](#_Toc514323884)

[2.3.4 用户数据量模式挖掘 1](#_Toc514323885)

[2.3.5 “重要用户”VS “正常用户” 1](#_Toc514323886)

[2.3.6 探索“重要用户”的成因 1](#_Toc514323887)

[2.4 本章小结 2](#_Toc514323888)

[第三章 基于蜂窝网络大数据的用户业务模式挖掘及预测 2](#_Toc514323889)

[3.1 概述 2](#_Toc514323891)

[3.2 业务模式的基本特征 2](#_Toc514323892)

[3.2.1 问题描述 2](#_Toc514323893)

[3.2.2 用户出现次数 2](#_Toc514323894)

[3.2.3 访问应用数 2](#_Toc514323895)

[3.3 用户业务模式挖掘 2](#_Toc514323896)

[3.3.1 问题描述 2](#_Toc514323897)

[3.3.2 用户业务轨迹建模 2](#_Toc514323898)

[3.3.3 用户业务模式挖掘 2](#_Toc514323899)

[3.4 用户业务模式预测 2](#_Toc514323900)

[3.4.1 问题描述 2](#_Toc514323901)

[3.4.2 使用传统的方法进行业务模式预测 2](#_Toc514323902)

[3.4.3 使用改进后的CNN进行业务模式预测 2](#_Toc514323903)

[3.4.4 预测性能对比分析 2](#_Toc514323904)

[3.5 本章小结 2](#_Toc514323905)

[第四章 用户移动模式与业务模式关联分析及预测 2](#_Toc514323906)

[4.1 概述 3](#_Toc514323908)

[4.2 移动模式和业务使用模式关联分析 3](#_Toc514323909)

[4.2.1 问题描述 3](#_Toc514323910)

[4.2.2 访问基站数和业务模式之间的关联 3](#_Toc514323911)

[4.2.3 主要基站的定义及分析 3](#_Toc514323912)

[4.2.4 用户“舒适区”提取 3](#_Toc514323913)

[4.2.5 “舒适区”业务模式分析 3](#_Toc514323914)

[4.3 家和工作地点提取及行为模式分析 3](#_Toc514323915)

[4.3.1 问题描述 3](#_Toc514323916)

[4.3.2 家和工作地点的识别 3](#_Toc514323917)

[4.3.3 家和工作地点的距离 3](#_Toc514323918)

[4.3.4 比家和工作地点业务模式的差异 3](#_Toc514323919)

[4.4 基于深度学习的业务点击预测 3](#_Toc514323920)

[4.4.1 问题描述 3](#_Toc514323921)

[4.4.2 数据预处理及特征工程 3](#_Toc514323922)

[4.4.3 算法描述 3](#_Toc514323923)

[4.4.4 预测结果分析 3](#_Toc514323924)

[4.5 本章小结 3](#_Toc514323925)

[第五章 总结与展望 3](#_Toc514323926)

[5.1 研究工作总结 4](#_Toc514323928)

[5.2 未来工作展望 4](#_Toc514323929)

[参考文献 4](#_Toc514323930)

[致谢 4](#_Toc514323931)

第一章 绪论

* 1. 课题研究背景

近年来，随着信息技术的高速发展，移动互联网、 智能物联网等相关技术的进步，人们产生和采集数据的能力越来越强，数据的规模、维度、类型也随之增长 而日益复杂。大数据早已在经济、政治、文化、军事领域以及金融、通信等行业发挥着越来越重要的作用。 信息的数字化也推动了数据量的大量增长，在移动互联网飞速发展的今天，社交网络、在线音乐、流媒体直播等平台产生的数据量疯狂增长，无处不在的商业、工业和社会活动正在源源不断的产生各种各样的数据，一个数据爆炸的时代早已来临。在大数据时代，随着无线网络智能化进程的加快，蜂窝无线网络数据逐渐成为了通信行业的重要研究对象。随着近年来3G、4G网络的推广普及，蜂窝无线网络显露出了越发强劲的活力，各种各样的新型业务形态逐渐进入移动终端。同时伴随着无线网络智 能化的发展，越来越多的智能终端推陈出新，进一步扩大了蜂窝无线网络的用户规模。截至 2015年12 月，中国网民规模达到了6.88 亿，互联网普及率已经达到了50.3%，其中手机上网的比例不断增长，已经达到了90.1%，通过3G或4G上网的比例为88.8%。在这样的背景下，随着越来越多的手机用户从主要使用传统的语音和短信业务转而投入到数据业务的阵营中，数据业务占比越来越大，移动蜂窝无线网络的服务质量和优化建设面临着前所未有的冲击和挑战。 蜂窝网络中业务的增长伴随着流量的增长， 产生了大量的用户行为数据，从这些数据中挖掘出用户的行为模式将有效地促进蜂窝无线网络的设计、规划与优化，并有效提升网络性能。

蜂窝无线网络大数据具有很大的应用价值，根据分析的对象差异，有不同的研究意义。对用户分析，蜂窝网络大数据可产生用户个体精准画像，为网络优化和资源配置提供支持，为内容提供商提供业务匹配和满意度优化支持，实现“端-管-源”的协同优化，包括：用户的360°画像+用户估计移动性+网络质量诊断+企业用户投诉解决方案，保证业务的满意度最大。 对业务分析，蜂窝网络大数据可以评价业务质量，分析各类业务特征和质量，对内容提供商优化业务具有重大意义；也可以建立业务感知模型，基于用户感知评价体系进行业务优化。 对用户群体行为分析，蜂窝网络大数据也可实现用户群体行为分析，对网络特定业务评估和优化、 资源分配提供支持，并有利于分析用户潜在的消费能力，具体包括用户行为分析、行业流量分析、群体行为分析等。

用户行为分析可以帮助用户更好地理解自己。首先，通过智能手机感知数据理解用户有很强的商业价值，可以用来改善推荐、广告精准投放等个性化服务。例如，通过检测用户所在的地理位置信息，可以改善基于地理位置信息的服务；通过理解用户的兴趣爱好、属性等基本信息更好地提高应用的个性化，如个性化网页搜索和个性化推荐；通过分析某些应用软件的长期使用用户，可以对这些用户做广告投放，提高广告投放的精确性；通过统计分析用户对某些应用软件的使用频率，可以针对性的推广个性化定制的应用服务等，从而提高商业价值蜂窝网络下的用户行为可以从移动性和业务使用两个方面进行分析。从海量的蜂窝网络数据中挖掘出用户的行为模式，找到用户的移动规律和业务使用规律，并探索出两者之间的关系，对于运营商而言，可以更好的进行无线资源的调度与优化，提升网络性能

* 1. 国内外研究现状
  2. 研究内容和创新点

本文利用蜂窝网络大规模DPI业务数据，对用户的行为进行全面的分析，挖掘用户的业务使用模式，移动模式，数据使用量模式，探索三者之间的关系，并进行群体用户的划分，比较不同群体下用户的行为区别。论文的重要研究内容如下：

1. 用户行为轨迹提取及数据量模式挖掘

研究用到的数据是中国某运营商4G数据面HTTP业务数据，字段大致可分为3类：用户相关字段、业务 相关字段和网络服务相关字段。利用Spark大数据平台，进行数据的清洗。首先根据业务的HTTP记录，即url地址对业务进行分类，有即时通信类，阅读类，博客类，导航类，视频类，音乐类，应用商店类，游戏类，电子商务类，邮箱类，社交网络类，新闻类，下载类，搜索类，其他。总共15类应用。然后进行用户粒度的数据聚合，提取出每个用户的行为轨迹。具体来讲，用户行为轨迹的每条记录有四元组来描述，（时间，地点，数 业务类型， 数据量）。通过对用户的每条记录进行聚合，得到用户的行为轨迹。基于用户的行为轨迹，可以进行用户行为模式的挖掘。对于数据量的使用模式，由于蜂窝网络中用户生活工作等行为，不同的用户的数据量消耗是不一致的，从而导致用户数据量使用在时间上的不均匀分布。基于二八定律，我们猜测，可能少部分的用户使用了绝大多数的流量。基于这一假设，统计出用户的流量使用分布，并挖掘出“重要用户”，获取用户的数据量使用规律。

1. 用户业务模式挖掘及预测

在众多的智能手机感知数据类型中，手机应用（App）使用信号与用户的个性化信息也密切相关联，手机App成为用户使用手机各种服务的主要入口。本文结合现网中采集的用户业务使用记录，对用户的业务轨迹进行建模, 使用聚类算法对用户的业务使用模式进行挖掘，观察不同业务模式之间的差异。同时，基于挖掘好的用户业务模式，设计了基于CNN的用户分类算法，即对于每一个新用户，都可以确定其所属的用户业务模式。

1. 用户移动模式与业务模式的关联分析及预测

移动性和业务使用模式是我们研究的重点，可以用移动性的各个指标与业务使用进行联合分析，挖掘出群体用户行为的内在规律。比较访问基站数不同的用户和业务量使用之间的区别可以发现那些移动范围比较大的用户和移动范围比较小的用户在在业务使用之间的区别。除此之外，用户在重要地点的用户行为也是一个值得研究的问题。用户在几个重要地点停留的时间最长，例如家和办公地点。识别出这些重要的地点对理解用户的行为模式有很大的帮助。我们将一个重要的地点定义为用户停留时间较长并频繁访问的地点。重要地点的例子包括：家，办公地点，健身房，百货商店，教堂等。这里，主要关注家庭地点和办公地点。识别出家庭地点和工作地点对于理解用户的行为模式是很重要的一方面。可以比较用户在家中的行为模式和工作地点的行为模式，更进一步的，可以比较不同发展程度的城市家庭地点和工作地点的距离 ，从而对用户行为有更加立体的刻画。

* 1. 论文结构安排

论文具体内容安排如下：

第一章：绪论部分。本章对该研究的选题背景及研究内容进行了简单的介绍，阐述了蜂窝无线网络中对用户行为分析的研究意义和现状。

第二章：系统介绍了蜂窝无线网络中数据类型，数据字段及用户行为轨迹的提取过程，并对用户的数据量使用模式进行挖掘。

第三章：定义了用户业务模式的基本特性，并对用户的业务模式进行挖掘，比较不同业务模式之间的区别。使用基于CNN的用户分类模型对新用户进行业务模式的判断，并与传统的分类方法进行比较。

第四章：探究用户移动模式和业务模式之间的关系。包括访问基站数与业务使用量之间的关联，“主要基站”的定义及分析，用户“舒适区”的提取，用户家庭地点和工作地点的提取，计算出家庭地点和工作地点的距离，并且对不同发达程度的城市对工作距离的影响。最后基于DeepFM算法进行用户业务点击预测，即基于用户本身的属性和时空特征，判断用户点击某一类App的概率。

第五章：本研究的总结和展望。总结了该论文取得的工作成果，提出了下一步研究计划，并对研究计划的可行性进行了分析。

第二章 用户行为轨迹提取及数据量模式挖掘

随着移动互联网的高速发展，用户的上网数据量呈现爆炸式的增长。网络数据结构复杂且数据量非常大。 从这些庞大的数据中进行有效的数据处理及模式挖掘，提取用户的行为轨迹，并进行群体用户的数据量模式挖掘，对于运营商而言，可以有效地促进蜂窝无线网络的设计、规划与优化，并有效提升网络性能，具有重要意义。

1. 1. 用户行为轨迹
      1. 蜂窝网络的数据类型

端到端的蜂窝移动网络由用户端、无线接入网络、核心网和内容提供云端组成。蜂窝网络数据主要从网络接口进行收集和提取，大致分为4类：流记录数据、网络性能数据、移动终端数据和额外的数据信息[1]。

1. 流记录数据

流记录是由网络中两个端点之间的数据包组成的，这些数据包之间具有一些共同的特定属性，如TCP/IP协议中相同的属性字段。可以在核心网的一些重要网络端口采集数据，再通过深度包解析技术（DPI）来获取蜂窝无线网络中的流数据。这些数据包含了用户一次连接会话中的许多重要参数，如基站位置（小区编号、小区经纬度）、用户标示（手机号、IMSI、IMEI）、会话的起始和结束时间、数据包的数量和字节大小、请求地址的URL和URI、访问主机名和IP地址、客户端操作系统类型等。这些数据是刻画用户行为的重要信息，可以用来刻画用户360画像、描述用户轨迹、分析用户偏好等。对于一个人口约100万的城市来说，一周的流记录约为几十TB到几百TB[2]。

1. 网络性能数据

网络性能数据可以反映出网络的性能和服务质量，一般以一定的时间粒度进行采集（5分钟粒度、10分钟粒度等），主要字段包括寻呼信号占用率、上下行吞吐量等。网络性能数据主要分成两类，一类是关键性能指标数据，另一类是测量的报告结果。其中，关键性能指标数据主要用来检测网络性能，并为提高网络质量、资源利用率提供支持；测量报告主要包括信道质量的相关信息，用来协助网络进行切换和功率控制。对于一个人口为100万的城市来说，采集粒度是10分钟，一周的网络性能指标数据约为几TB[2]。

1. 移动终端数据

随着智能终端的快速发展、蜂窝网络覆盖范围的增长以及4G网络的全面来临，越来越多的智能设备接入蜂窝无线网络，从而产生了大量的数据。终端数据可以通过终端上的APP进行采集，然后将采集结果传入云端，这些采集的数据可以包含网络多个层次的属性字段，如应用层、网络层、链路层等。这些数据可以包含了服务质量、终端属性、网络参数（信号强度、信噪比、干扰比、上下路速率）等[2]。

1. 基础属性数据

基础属性数据主要包括用户信息、基站地理信息、手机终端信息等，这些数据对分析蜂窝网络中用户行为非常重要。其中，用户信息主要包含用户计费信息、流量套餐、通话短信使用量等；地理信息主要包含基站位置、基站编号、基站区域的兴趣点（POI）等；手机终端信息包含手机型号、操作系统等。与上述三种数据相比，基础属性数据相对静态，一般不会发生变化[2]

* + 1. 用户行为轨迹的提取

本文用到的蜂窝无线网络DPI数据面向4G HTTP业务数据，当用户使用移动端设备接入基站并上网时，将会产生flow记录，记录字段可以分为三类，用户相关字段，业务相关字段和网络服务相关字段。采集点在Gn、Gb、Gi等关键接口处[3]，处理时间为3天，每天的数据量大约700G，包含的基站小区约5400个，数据的基本信息如下表3-1所示。

表2-1 蜂窝网络数据信息

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 时间 | 上行流量 | 下行流量 | 用户数 | 小区数 | Flow记录数 |
| 20150403 | 0.72TB | 20.37TB | 1015328 | 1806 | 10.6e+8 |
| 20150404 | 0.72TB | 21.42TB | 984524 | 1806 | 10.7e+8 |
| 20150413 | 0.72TB | 19.07TB | 1021516 | 1806 | 10.6e+8 |

每条数据中包含了用户接入网络的起始时间，终止时间，接入基站，使用的具体App和上下行流量等。我们通过规则筛选的方式将App分成十五大类，分别是即时通讯、阅读、博客、导航、视频、音乐、应用商店、游戏、电子商务、邮箱、社交、新闻、下载、搜索、其他。业务分类示例如下：

表2-2 业务类别信息

|  |  |
| --- | --- |
| 类别 | 应用 |
| 即时通信 | MSN，QQ，微信 |
| 阅读 | iReader，QQ阅读 |
| 博客 | 新浪博客，QQ空间 |
| 导航 | 谷歌地图，百度地图 |
| 视频 | 优酷，CCTV |
| 音乐 | 网易云音乐，QQ音乐 |
| 应用商店 | Apple Store，应用宝 |
| 游戏 | 4399，腾讯游戏 |
| 电子商务 | 亚马逊，天猫，京东 |
| 邮箱 | 163邮箱，网易邮箱 |
| 社交 | 领英，新浪微博 |
| 新闻 | 网易新闻，新浪新闻 |
| 下载 | 百度云，迅雷 |
| 搜索 | 百度，搜狗 |
| 其他 | 广告，SWF |

用户的行为轨迹数据主要包含四个方面，即时间，地点，业务类型，数据使用量。可以由以下公式来描述：

其中，表示第个用户的第条记录。U是总用户数，N是记录数。表示第u个用户第i条记录的起始时间，表示第u个用户第i条记录的基站位置，表示第u个用户第i条记录的业务类型，表示第u个用户第i条记录的数据使用量。用户u的行为轨迹可以表示如下：

本文采用的是覆盖100万人口的大规模数据，对于每个用户，经过数据聚合和转换之后，都可以得到行为轨迹。用户行为轨迹的提取是后续所有工作的基础。

* 1. 用户数据量模式挖掘

随着移动互联网的高速发展，人们产生和采集数据的能力越来越强。4G 移动电话用户持续高速增长、移动互联网应用不断丰富，推动移动互联网流量持续高速增长。 移动数据的快速增长要求移动服务提供商有效地管理其无线电频谱资源并发展其基础设施，以满足移动数据用户不断增长的需求。为了应对这一挑战，我们必须分析用户是如何访问蜂窝网络服务以及他们对不同网络资源的需求。

* + 1. 问题描述

在本文中，我们研究了2015年4月13日中国西北某城市的蜂窝流量数据，数据来源于中国最大的通信设备运营商。从海量的数据中，利用spark大数据处理平台，进行大规模的数据清洗、聚合，得到每个用户的行为轨迹数据。这些数据源提供了当今大型蜂窝网络中移动数据用户使用模式的全面视图。 我们提出了用于分析和挖掘此类数据集的新颖且可扩展的统计工具。我们使用二阶马尔科夫模型来表征时域数据的使用模式，并发现了两个不同的用户群，重度用户和普通用户，这两类用户有着截然不同的行为模式。重度用户的数量非常少，但却消耗了大部分的流量，因此，如果能挖掘重度用户的行为模式，得到重度用户的行为规律，对运营商而言，可以有效的进行资源调度以满足不同用户的需求。

* + 1. 用户数据量使用规律的发现

对于每个用户，基于用户行为轨迹，可以统计出一天使用的总数据消耗量，用总字节数来表示。下图展示了所有用于数据消耗量的CDF曲线。

其中x轴（以对数标度表示）表示最高用户的百分比（按其总字节使用量排名）y轴表示底部使用的总字节数的用户。我们观察发现，数据量的使用分布非常不均匀，很少一部分的用户消耗的大多数的流量。具体来说，10%的用户消耗了60%的流量， 20%的用户消耗了80%的流量。

* + 1. 用户数据量模式挖掘

为了研究数据的使用模式，我们从时域维度进行研究。一个值得注意的问题是用户是否会随着时间的推移不断访问数据服务，或者他们是否倾向于使用突发性业务？我们将用户的数据量使用模式用如下时间序列来表示：

我们以5分钟作为时间粒度，所以在一天的时间里，T为288。 表示用户u在t时刻的流量消耗。为了便于分析，将时间序列转换成二进制的序列，如下：

是预先设置的阈值。基于上述公式，可以将用户的数据使用模式用二阶马尔科夫模型来表示。表示的是活跃状态。表示的是休眠状态。更进一步地，我们定义一下特征来表征:

1. 活跃状态概率：
2. 活跃状态转移概率：

其中I是指示函数。同样地，可以得到休眠状态的概率。 值比较大说明用户访问数据服务比较频繁，而比较大的说明用于倾向于保持活跃的状态，即连续的数据量使用，且单位时间粒度的数据用量大于。

我们根据用户一天的流量消耗量，将用户分成四组。分别是大于30MB,、10MB到30MB、5MB到10MB、小于5MB。下图展示了四组用户关于和联合分布不同的数据使用模式。

其中越亮的地方代表用户的密度越低，反之，越暗的地方代表用户的密度越高。

Emm，分析上图。

基于上述不同用户组观察到的不同使用模式，我们使用每天30MB数据使用量作为截止阈值，将用户分为两组：每天使用量超过30MB的用户称为重度用户，剩余用户称为普通用户。 图4显示了这一点重度用户倾向于连续和密集地访问数据服务方式，而普通用户访问数据业务大多数是突发性或者间歇性的。

* + 1. “重度用户”VS “普通用户”

为了理解重度用户和普通用户在业务使用上的差异，我们定义了两个指标， 服务流行度（pop）和流量优势（Dom）。给定用户的行为轨迹，表示用户u在业务s上的流量消耗。其中s为定义好的15类业务类型S之一。服务流行度和流量优势的定义如下：

其中表示业务s在所有的业务中消耗的流量占比。I是指示函数，表示在所有y上x的期望值。服务流行度Pop(s)可以表示用户在一天时间内使用业务s的可能性，而流量优势Dom(s)表示业务s平均贡献的流量占比。下图反应了重要用户和普通用户在服务流行度和流量优势之间的差异:

Emm 分析一大堆

* + 1. 探索“重要用户”的成因

在上一小节，我们发现重度用户使用的业务类型非常多，自然引出的一个问题是，到底是哪类用户是他们成为重度用户？在这一节，使用无监督聚类的算法找到重要用户的成因。

* 1. 本章小结

通过大规模的用户上网数据，经过数据的清洗和转换，得到了用户的行为轨迹，并基于用户的行为轨迹，挖掘用户的数据使用模式。定义活跃状态和活跃转移状态，并将用户分成四组，比较不同组用户活跃状态和活跃转移状态的联合概率密度，从而识别出重度用户和普通用户。从数据使用模式看，重度用户倾向于连续和密集地访问数据服务，轻度用户倾向于突发性或者间歇性的访问数据服务。从业务角度去分析，定义的业务流行度以及流量优势，比较两类用户在业务使用上的差异。重度用户倾向于…， 轻度用户倾向于…， 从移动性的角度去分析…

最后通过无监督聚类的方式找到重要用户的成因

第三章 基于蜂窝网络大数据的用户业务模式挖掘及预测

1. 1. 概述

随着手机使用率的不断增长，手机App的普及性以及数目都快速增长。与此同时，近年来越来越多的研究工作尝试去理解手机用户的手机App使用行为手机App的使用信息也在一定程度上反应了用户的生活需求和习惯。不同生活背景的用户，在使用手机应用行为上有一定的差别。在本章节的工作，我们基于用户手机App使用行为的相似性，通过聚类的方法发现不同的用户群体，从而为研究学者、手机制造商、手机运营商、应用开发者等提出针对性的建设性建议。

在本章节的工作里，我们分析了34万手机用户持续两天的手机App的使用记录，基于用户行为轨迹，定义了业务模式的基本特征。并基于用户的业务使用记录，对用户的业务使用情况进行建模，挖掘出用户的业务模式，识别出了6类不同的用户。最后采用基于CNN的分类算法，基于新用户的业务轨迹，对新用户进行分类，并与传统的分类算法进行比较。实验表明， 采样基于CNN的算法在分类性能上比传统方法要好。

* 1. 业务模式的基本特征
     1. 问题描述
     2. 用户出现次数
     3. 访问应用数
  2. 用户业务模式挖掘
     1. 问题描述
     2. 用户业务轨迹建模
     3. 用户业务模式挖掘
  3. 用户业务模式预测
     1. 问题描述
     2. 使用传统的方法进行业务模式预测
     3. 使用改进后的CNN进行业务模式预测
     4. 预测性能对比分析
  4. 本章小结

第四章 用户移动模式与业务模式关联分析及预测

1. 1. 概述
   2. 移动模式和业务使用模式关联分析
      1. 问题描述
      2. 访问基站数和业务模式之间的关联
      3. 主要基站的定义及分析
      4. 用户“舒适区”提取
      5. “舒适区”业务模式分析
   3. 家和工作地点提取及行为模式分析

用户在“舒适区”停留的时间最长，例如家庭地点和工作地点识别出些重要的地点对理解用户的移动模式有很大的帮助。这里，我们只关注家庭地点和办公地点。在这个章节里，我们详细介绍如何从海量的日志记录中识别家庭地点和工作地点，并进而发现相关的生活模式

* + 1. 问题描述
    2. 家和工作地点的识别
    3. 家和工作地点的距离
    4. 比家和工作地点业务模式的差异
  1. 基于深度学习的业务点击预测
     1. 问题描述
     2. 数据预处理及特征工程
     3. 算法描述
     4. 预测结果分析
  2. 本章小结

第五章 总结与展望

1. 1. 研究工作总结
   2. 未来工作展望

参考文献

致谢