doi:10.3969/j.issn.1000-1247.2017.04.006

基于手机信令数据的大客流监控应用研究

胡忠顺　王　进　朱　亮
上海理想信息产业（集团）有限公司

摘要　首先分析处理全市用户位置的大数据所需的架构、特点以及当前存在的问题，然后从各个数据源的应用场景和算法特征分析能覆盖2G/3G/4G用户各种应用场景的数据源算法。为了更好地验证所采用的基于手机信令数据的各种算法对大客流监控能力的提升，结合试验结果给出中国电信应用项目场景的成功案例，便于基于手机信令数据的大客流监控在电信信息化的应用实施进行参考。最后对基于基站的定位在高密度人群等大客流监控中的位置和角色以及对此可能带来的变化进行了探讨。

关键词　手机信令数据　手机信令数据算法　平均定位精准度　大客流监控

1 引言

目前，传统使用的大客流监控方式主要有人工客流监控、闸机客流监控、视频客流监控，这些方式的缺点如下。

人工客流监控需投入大量人力物力、费时费力、无法数字化、精确化，信息再利用率低。

闸机客流监控获取的信息仅限于出入闸的客流，机械工作方式效率低，无法实时获知或预测大客流信息，对突发性大客流的管理缺乏手段，存在较大安全隐患。

视频客流监控需安装维护大量设备，成本投入大，同时视频监控仅限于可视范围，并常受天气、光线等因素影响，监控效率不高。

鉴于以上方式的不足，引入基于运营商移动通信手机信令数据（以下简称“手机信令数据”）的大客流监控方式。

手机信令数据是指移动终端用户在发生通话、短信、上网及变换寻呼区时在运营商网络中产生的大量手机信令数据，移动终端数据会反馈如时间、基站信息、场强和时延等关于用户位置的有效信息，对用户数据产生的时刻进行精准位置定位，从而判断用户所在的区域范围。手机信令数据的生成催生了地域区域性统计分析的应用，如区域人口统计分析、旅游景点客流分析预测和用户人群画像等。

基于手机信令数据的大客流监控方式是通过电信数据中的位置信息算法定位用户经纬度位置，对监控管理和人口统计进行分析，与人工客流监控、闸机客流统计、视频客流监控等传统方式相比，除了有先天性的优势之外，还可减少大量人工成本、大量设备的安装和维修以及监控区域局限等一系列常规监控方式存在的弊端，大大降低了经济成本和客流安全隐患问题。

手机信令数据拥有实时、快速、精准三大优势，能更高效地实现对大客流的监控需求，通过电信大数据平台支撑，使用户手机信令位置数据得以长久保存，并结合其位置轨迹行为数据、用户基础画像数据、用户互联网行为数据等，可实现城市常住人口分析、区域实时客流监控、区域精准营销等应用。因此，基于手机信令数据的大客流监控的应用研究具有很高的实用价值。

2 基于手机信令数据的大客流监控架构搭建

随着监控的应用场景愈发普及，建立大客流监控平台显得愈加必要。大客流监控平台基于实时采集的运营商手机信令数据，文中主要使用运营商的PCMD数据。通过这三类数据实现大数据监控平台，充分利用运营商数据资源和大数据分析技术，比以前传统方式的客流统计，无论在科学性、时效性还是投资成本效应方面都有大的提高和跨越。

首先阐述运营商的PCMD信令数据的含义。PCMD数据：PCMD-1X表示当手机用户发生通话、短信等行为时，记录接入基站编号、基站扇区、主寻呼基站周边基站编号、周边基站扇区、时间、场强、时延、手机号码、手机IMSI等信息；PCMD-DO表示当2G、3G手机用户发生上网等行为时，记录接入基站编号、基站扇区、主寻呼基站周边基站编号、周边基站扇区、时间、场强、时延、手机IMSI等信息。

手机信令数据系统的数据也有其局限性，最大的瓶颈在于数据的实时性。手机信令数据系统融合了多个基站信令数据，各个数据的更新周期相同，势必要等到各数据源都到达计算得到的最优定位数据结果才最为准确。这样不仅加大了计算处理集群的处理压力，也降低了实时数据的准确性。因

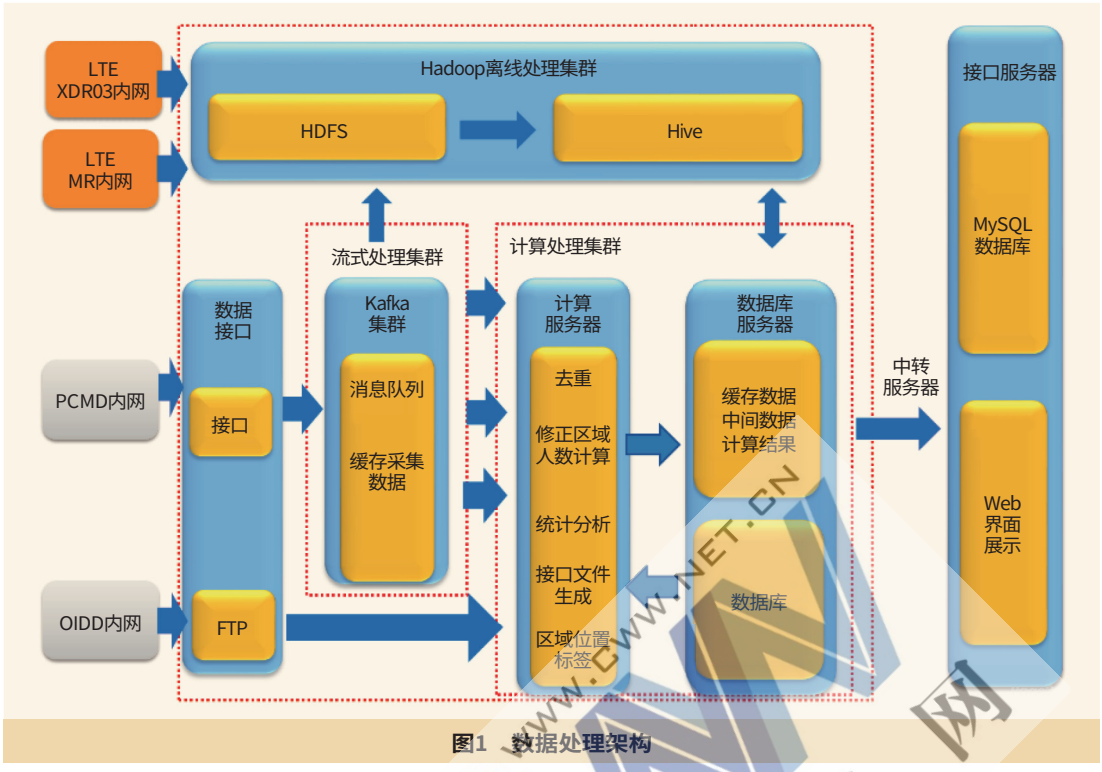

图1 数据处理架构

此对于实时性要求较高的数据，只能对实时效果和数据准确率进行综合考虑。

考虑到用户位置结果的实时性要求，以及全市用户多种位置数据处理大数据流的能力，基于手机信令数据的大客流监控的数据处理架构搭建如图1所示，说明如下。

(1)基站信令数据单位时间段内生成文件，通过采集服务器进入大数据平台。

(2)采集到PCMD信令数据文件缓存进入分布式队列Kafka集群。

(3)在Storm流式处理平台进行实时的数据筛选与预处理，过滤无效数据。流式处理结果缓存进入队列Kafka集群。

(4)Kafka集群缓存数据通过队列进入计算处理集群，由计算服务器进行定位处理。

(5)多数据汇总入计算处理集群后，筛选择优后得到单位时间力度的最优定位结果，最终的定位结果存入Hadoop离线处理集群，作为统计分析的基础数据。

(6)对于高实时性数据，经计算处理集群处理后直接由中转服务器传递至接口服务器进行实时展示，以保证快速响应的性能。

3 手机信令数据算法的建立、优化及验证

3.1 大客流系统数据处理

原始手机数据从产生到进入大数据平台，需要经过数据采集、数据清洗、数据标准化处理和数据入库，之后进行定位算法计算，结合数据模型分析后可以使用在不同业务场景中，例如区域客流统计、流入流出分析等。

(1)数据采集

原始数据通过接口方式（FTP文件，HTTP接口等）直接对接系统平台。OIDD及PCMD话单数据采用流式处理集群Kafka接口传输。对每个时间片段传输过来的数据，进行实时的数据过滤等预处理，便于后续定位计算，写入分布式队列Kafka中，入库计算处理集群，和其他定位数据汇总；同时数据写入大数据平台的HDFS，并进一步建立Hive数据表进行历史保存。

LTE数据从大数据平台的HDFS中获取，定期扫描HDFS，以发现平台上是否有新的数据文件生成。若有新文件生成，则启动数据处理任务，计算用户经纬度位置信息并匹配位置标签信息，数据处理完成后启动文件传输与稽核任务，确保数据尽快汇总。

(2)数据清洗

在数据从操作系统移到数据库的过程中，数据被清洗。一些情况下，为了使输入数据正确，使用简单的规则处理输入数据。复杂情况下，将会采购数据清洗工具，用数据清洗工具把数据转换成可以接受的形式。

当从多个源数据系统集成数据时，必须确保来自不同源系统填充同一目标字段的数据属性相同。例如有的话单用秒记录，而另一话单以分钟为单位记录。

去除重复是指去掉重复的记录。这个处理过程可以合并同一个源系统内重复的数据，或合并一个或多个系统相同的数据。

(3)数据标准化处理

数据标准化处理的主要工作如下。

格式变换，如对不符合日期格式的数据将日期格式统一为yyyy-mm-dd。

赋缺省值，在数据仓库中定义取值不为空的字段在源数据对应的字段可能存在没有取值的记录，这时根据需要，直接赋一个缺省值。

类型变换，如将源系统的Number类型转为varchar2类型等。

长度变换，支持对定长、变长格式数据的格式转换，如将源系统中定义的varchar2(10)转为varchar2(20)等。

代码转换，如源系统的某些字段经过代码升级以后，将老的代码转为新的代码等。

去除空格，去除字符类型的数据中的前后空格。

特定字符转换，如对于用于计算的某些字段不能含有“+、-、*、/”等特殊符号，需要根据业务规则对这些字符进行指定替换。

(4)数据入库

数据入库功能指将人群手机信令数据入到大数据平台中，一般一份数据分别存储于HBase和Hive中，HBase用于实时查询服务，Hive主要用于数据分析和挖掘服务。

3.2 手机信令数据定位算法应用建立

全市手机信令数据系统基于基站信令信息，通过信令信息中包含的基站信息及辅助数据，计算实际用户所处的经纬度，同时，通过利用判断射线法得到的用户经纬度和目标区域范围之间的关系，便于对区域范围内的客流特征进行统计分析。

(1)可行性分析

三个基站或多个手机信令数据，可定位至一个点，且有效基站越多，定位精度越高，但是三点定位可能存在以下问题。

如果有两个正根，只能通过目标与主站之间的距离找到最可能的那个根。由于多径效应的存在，不能保证根的正确性。

多手机信令数据时，如果进行主站轮换，且在轮换过程中至少有两次获得目标的可行解，则可通过对所有可行解的聚合获得目标的唯一最可能位置，准确度很高。

三点定位有一定的不确定性，可能出现无解或获得错误解的情况，但是定位精度比单点和两点定位有了质的提升；多点定位的可行性最高，定位最准确，理论最完备，可排除人为假设的影响，能获得理论上的最佳位置，有条件的情况下应优先实施。

(2)PCMD定位算法的建立

一条PCMD数据中包含了两个关键时间信息，分别为初始和终止时刻的时间戳（Timestamp），这反映了手机接入和断开网络的时间；其次每个时刻都会产生一组信息，其中与定位相关的信息有基站号、扇区号、时延、电磁辐照场强等信息，所以将这两个关键时间信息加入PCMD定位算法中。

通过分析，每组信息可以由一个或多个基站产生，这些基站分为参考基站（Reference Cell）和非参考基站。一个手机同时监听多个基站，为手机提供时间数据的为参考基站，其余为非参考基站。当PCMD采集数据时，可能有1～2组参考基站数据，每组参考基站最多可有5个非参考基站，且这些基站中的一组或多组数据可能均为空或0，这取决于参考基站的个数。因此将有数值的参考和非参考基站关键信息加入PCMD定位算法中。

(3)三点及多点定位方法的采用

如果包含3个或更多基站的数据，则可以根据该组基站进行较为准确的定位。基站越多，定位精度越高，因此使用最小二乘TDOA算法。

(4)定位算法改进内容

经过实际路测数据验证，在多点定位算法中对不同的基站类型按照不同类型进行计算，能够得到更加精确的定位结果，因此最终对于定位算法进行改进优化。

将PCMD分成初始化时刻与结束时刻两部分数据，分别使用两时刻数据定位。

由于定位基站数目从1～6不定，算法实现中按基站类型和时延数据对定位结果赋予一定的权重，按加权平均进行计算，若属于室内站或时延较短，则权重较高，对定位结果影响较高。

计算方法见公式（1）。

$$\sum\left(\frac{1}{1+R^2}+is\_indoor\right)\times X \quad (1)$$

其中若是室内站is_indoor取值为1，非室内站is_indoor取值为0。

(5)定位算法改进前后路测对比

为了验证定位算法的准确性，测试人员通过手机GPS工具记录实际经纬度位置，与定位算法的计算结果比较，得到算法准确性的判断结论。

定位准确度分析如下。

根据每条数据的时间戳，以及GPS工具箱导出数据中的时间戳，匹配并计算每个PCMD数据的GPS坐标。

匹配规则：一条PCMD数据时间上下浮动5s内，并与之匹配的所有GPS数据。如果匹配出多条，按均值计算。

得到具体的定位结果见表1。

根据定位结果来看，算法优化后的定位结果精确度得到显著提升。

表1　定位结果

误差范围	50m以内	100m以内	150m以内	200m以内	250m以内	300m以内	350m以内	400m以内
V2.0	10.1%	18.6%	43.6%	69.7%	85.6%	88.3%	98.2%	100.0%
V1.0	0.5%	1.4%	2.3%	4.4%	6.9%	8.6%	10.0%	12.2%

3.3 数据分析模型

在人群手机信令数据和定位算法的基础上，还需要结合数据分析模型，采集计算出区域中人群数量、流向和趋势等。

(1)区域人数计算模型

根据预先设定的监控区域范围，如网格、商圈、监控地块等，对当前周期内的区域人数、流入流出人数进行计算。

对于区域人数的计算，重点是区域内部位置的判别。使用计算几何中的射线法，对在任意多边形内的基站进行检出。

区域人数计算的另一个重要方面，是去重和修正。在当期周期内有些数据是重复数据，是由于某些目标多次产生数据且这些目标均位于该区域内。因此，对于这些数据需要去重处理。

对于区域流入和流出的人数计算，提取出可以反映当前时刻与上一时刻人群移动模式的、可计算的、具有边界意义的特征量，根据该特征量构造合适的统计量，从而可以正确反映人群的移动情况。

计算得出的停留人数、进入人数和离开人数等统计结果保存在数据库中。

(2)人群流动模型

人群流动运行状态分析是进行大客流监控的核心，为了模拟在有限空间中大量人群的流动分布，拟采用基于面向对象的技术，在交通流模型和行人流模型的基础上，建立一个在有限空间内高密度人群流动的元胞自动机模型。采用面向对象思想的建模方法，使模型具有很好的适用性、扩展性和复用性。人群流动运行状态分析的主要工作包括以下两点。

基本变量计算：核心是统计和计算各个采集点的人流速度（加速度）、密度和流量。

人群流动可表达为速度、流量、密度三者之间的关系，见公式（2）。

$$Q=k\times v \qquad (2)$$

其中Q为人流的流率，k为人流密度，v为人流速度。

人流密度和人流速度主要依靠手机用户在感知设备定位的位置移动进行计算；也可以通过视频的人流识别计算得到。

人流的区域分布分析：将海量的用户手机信息位置数据，按照区域、时间段、移动方向进行分类和统计，计算出区域网格内的人群流入量、流出量、存量与密度，以及人群流动的方向和速度。通过人流的区域分布分析，可以实现热点区域的识别。

3.4 手机信令数据算法应用范围

通过定位算法及其优化，保留了大量有效及准确的手机信令数据，定位的准确度可以达到300m范围内，在此基础上，通过区域划分及手机信令数据分析可以得到多个复合维度的基础数据。例如高密度区域人群监控；区域（商圈、景区等场景）内的实时客流监控；事后客流分析，包括人群密度变化趋势、人群来源/去向区域分布、人群归属地分析、人群基本画像分析等。

4 基于手机信令数据应用实例

实时采集三类运营商信令数据，通过搭建大数据监控平台架构，利用手机信令数据算法进行分析，为政府提供实时景区客流监控等应用案例。

4.1 顾村公园实时客流监控

在顾村公园运营管理综合信息平台上，引入各手机位置信令数据作为客流监控与预测的重要数据补充，具有明显的实用价值。

(1)实时客流分布

用热力图显示顾村公园及附近多个主要区域的实时人群密度，更新周期为5min一次，如图2所示。

图2展示的是顾村公园各区域的客流分布情况，不同的客流密集程度显示不同的预警颜色，能帮助管理人员更快速地给出客流疏散方案。

(2)历史客流统计分析

按天统计历史客流，可以对比在樱花节时周末和平时的客流情况，结果如图3所示。

按小时统计每天各个时段的客流分布曲线，可以任选2天进行对比分析，结果如图4所示。

通过实时基站的定位算法实现人流监控平台，可以实时统计顾村公园内外各个地块的人群数量、人群密度和热门集中地分布情况。通过人群流动情况，计算出不同区域内的人群数量变化情况，可以对比分析得出网格区域内人群数量变化情况。此外充分利用中国电信DPI互联网大数据，探索线上与线下的互动模式，实现更长周期、更大范围的客流预测。

图2　顾村公园实时客流监控示意

图3　按天客流统计分析示意

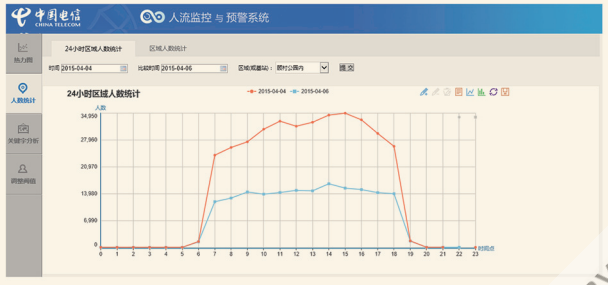

图4　按小时客流统计分析示意

5　结束语

文中研究了基于手机信令数据的大客流监控应用，并在实际应用案例中实现了特殊区域、特殊时期的客流监控分析与预测，有效支持了大客流的及时管控，实现了对特定区域的人群特征及行为的画像分析，可有效应用于社会公共安全、区域商业洞察、城市交通规划、商业规划等方面，具有良好的社会效益和经济效益。

因此，基于基站的定位作为一种移动通信定位技术，在定位时效性和可扩展性方面存在较大优势，适用于需结合用户画像的大客流监控分析应用。该方法可成为其他大客流监控方式（如Wi-Fi探针）的有益补充，对提升目标客户的感知度、降低运营分析成本起到良好的效果。

（上接20页）

直接从后台网管平台获取。回归建模可以采用更复杂但效果可能更好的曲线回归模型。

参考文献

[1]　张文彤，钟云飞.IBM SPSS数据分析与挖掘实战案例精粹[M].北京：清华大学出版社，2013

参考文献

[1]　YD/T 2232-2011.cdma2000数字蜂窝移动通信网基于用户平面的定位系统技术要求[S].2011

[2]　屠晓东.基于UWB信号的多基站与单基站定位算法的研究与性能分析[D].青岛：中国海洋大学，2012

[3]　姚金杰.基于地面基站的目标定位技术研究[D].太原：中北大学，2011

[4]　聂颖，易强，江红，等.CDMA无线定位系统的基站选择算法[J].电讯技术，2004(1)

[5]　夏林元，吴东金.多基站模式下的实时与自适应室内定位方法研究[J].测绘通报，2012(11)

作者简介

胡忠顺

硕士，毕业于上海海事大学，长期从事IT领域的技术研究、产品研发及推广工作，先后负责多项中国电信及其他政府企业的重大信息化工程，如中国电信大数据平台、上海市12345政府服务热线等，获得过多项省部级科技进步奖，近期，主要从事互联网技术、大数据、分布式技术的研究，及相关产品研发。

王进

本科，毕业于东南大学，现就职于中国电信上海理想信息产业（集团）有限公司大数据业务部，主要研究方向为大数据分析、数据建模等。

朱亮

本科，毕业于上海师范大学，现就职于中国电信上海理想信息产业（集团）有限公司大数据业务部，主要研究方向为大数据分析、大数据平台架构等。

作者简介

刘凯凯

硕士，毕业于重庆邮电大学，现就职于中国移动通信集团设计院有限公司重庆分公司，高级工程师，长期致力于无线网络规划设计技术的研究和跟踪。