人口管理大数据应用级建模技术研究与实践

范英　　康凯　　施一琳　　公安部户政管理研究中心

摘　要：为解决公安户籍人口城镇化率统计和人口迁移情况统计中的问题，基于大数据应用建模技术，建立了地址匹配、城乡属性建模、地址清洗等地址和城乡属性匹配模型，并对人口基本信息、基本信息历史、城乡分类信息和户籍人口业务信息进行数据抽取、清洗、分析，从而形成公安户籍人口城镇化率统计和人口迁移情况统计结果。通过实际应用和优化，户籍人口城镇化率统计结果的准确与效率得到了大幅度提升。

关键词：人口信息　　数据抽取　　数据建模　　城镇化率统计

引言

人是国民经济和整个社会发展的主体，是国家重要的基础性、战略性资源。各级政府、企事业单位、社会团体以及公民个人对人口信息都有着广泛的需求[1]。尽管部级人口信息管理系统的建设和应用已具有一定基础，但由于历史信息缺失，仅靠传统的技术手段，无法满足目前的户籍人口统计业务需求。如何有效利用现有数据资源和技术手段，针对户籍人口管理和数据的特点、应用和服务的具体需求，以城镇化主题算法建模研究为切入点，对户籍人口统计做出准确的统计与推算，攻克户籍人口数据在获取、存储、管理、分析、展现等五个层面的技术难点，是我们亟待解决的问题[2]。本文以户籍人口城乡分类信息为例，通过大数据应用级建模技术的应用，研究与探讨户籍人口城镇化率统计的有效方法。

一、公安人口信息现状和基础条件分析

（一）现状

公安部重视人口管理信息化工作，自 2004 年开始建设“金盾工程”重点建设项目部级人口信息管理系统（一期）（以下简称部库一期系统）暨全国人口基本信息资源库，实现了全国 13 亿人口基本信息的集中存储和管理。部级人口信息管理系统（二期）（以下简称部库二期系统）项目于 2012 年开始建设，2014 年 8 月验收，目前已顺利完成过渡并基本取代了一期系统。

（二）基础条件分析

1．库二期系统实现了自下而上的户籍人口基本信息维护数据获取、传输、入库的全程自动化，做到“当日变动、当日维护”，历史轨迹记录完整，数据完整、准确、鲜活。

2.通过对人口基本信息数据项结构进行扩充，充实完善户籍管理业务信息。2016 年以来的户籍管理业务信息备案完整率已达 99%以上，分析户籍人口城镇化率和农业转移人口变化的基础数据已基本完备。

3.自 2016 年下半年以来，各地按照公安部统一部署，全面开展户籍人口城乡分类属性核实标注和数据上报工作，目前全国人口信息库中户籍人口城乡分类数据与户籍人口基本信息数据一致率已达 99%。

二、技术难点分析

户籍人口统计存在三大技术难点：

1.在不影响部级人口业务系统正常运行的前提下，在短时间内完成对几十亿数据进行多表关联分析，同时产生数十张统计表。需要为户籍人口统计搭建独立的大数据运行系统支撑环境，同时数据访问和运算统计性能需要远大于常规关系型数据库的性能指标。

2．前台人口信息数据维护采取三种机制并行的方式：一是户籍人口基本信息基于传输平台，采取数据视图数据抽取方式汇集全国 389 个地市的户籍人口基本信息；二是户籍管理基本信息采取数据备案机制，由各省级人口管理信息系统通过备案接口向部中心上报业务变动数据；三是户籍人口城乡分类信息由公安各级人口管理机关进行人工标识，由省级人口系统统一上报。由于各省级人口管理信息系统由不同厂商开发，户籍人口基本信息维护与户籍管理业务备案相对独立且尚未建立相互校验机制等诸多原因，使得公安户籍人口城乡分类数据质量存在历史数据缺失、格式不规范、重复记录、不符合业务逻辑等问题，在进行关联分析前，必须对原

始数据进行复杂的数据清洗和处理，数据清洗处理难度大。

3.户籍迁移相关业务中户籍人口城乡分类信息标识完整性、准确性不高，其他户籍管理业务信息城乡分类信息尚未标识，同时户籍地址更新的时效性远远落后于实际地址的变动，急需地址匹配算法模型。通过户籍地址的关联，实现户籍管理业务信息城乡分类标识，是户籍人口大数据分析的关键技术；如何提高地址匹配的准确性是户籍人口大数据分析的技术难点。

三、户籍人口大数据分析实践

（一）技术原理

1.采用 SQOOP 工具，远程连接部库二期系统户籍人口基本信息、历史信息、户籍人口城乡分类信息和户籍管理业务信息数据表，抽取原始数据，保存到本地的 HADOOP 数据库（HDFS 列式数据库）；

2.采用 Hive 工具对 HDFS 数据库进行查询、计数等在线管理；

3.采用 SPARK 内存计算方式，进行数据清洗、标识、统计分析，运算中间结果保存到 HDFS 列式数据库；

4.统计结果一方面导出为报表文件，另一方面转存到分布式数据库；

5. 用基于 EchartS 报表工具对统计结果进行可视化展示。

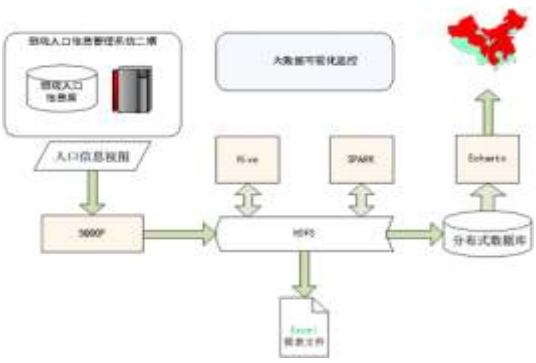

图 1　籍人口大数据分析技术原理示意图

（二）分析环境准备

本研究基于大数据开发平台进行统计分析，支撑环境采用 PC 服务器集群，软硬件配置环境如下：

主节点服务器：1 台，计算节点服务器 9 台；

硬件参数：PC 服务器，CPU：4 颗 12 核，内存：512G，硬盘：2×900G，网卡：2 个万兆端口，操作系统：CentOS 6.6 Linux 操作系统；

软件配置：部署大数据管理平台，包括 Hadoop、Hive、Spark、Sqoop 等组件和分布式数据库。

（三）数据抽取

1.SQOOP 工具在同城跨数据中心异地系统数据抽取过程中工作稳定，无异常情况，数据抽取工作一次性完成；

2.抽取数据后，HDFS 按 3 备份存储，既能保证数据安全，又能保障数据运算时数据集群加载，提高运算性能；

3.HDFS 采取数据压缩格式保存，压缩率达到 200%~300%。

（四）数据清洗

1. 据质量问题

（1）地址问题：户籍地址省市县区、户籍地址区划内详址是户籍人口统计的关键字段，主要存在地址字段缺失、地址不规范、行政区划缺乏历史轨迹等方面的问题；

（2）重复记录：由于户籍人口数据维护机制问题，户籍人口各类原始数据中存在较为严重的记录重复情况；

（3）数据缺失：由于户籍管理业务信息采用备案机制，对各地上报数据的及时性、完整性尚未建立完整的监管机制，存在较为严重的数据缺失情况；

（4）数据错误：存在办理日期错误、逻辑校验不符等情况。

2. 据清洗

（1）对户籍人口各类原始数据中使用的户籍地址进行清洗，包括地址中的全角字符到半角字符的转换、去除空格和非法字符等，然后对户籍地址去除省、地市、县名称信息；

（2）通过关联户籍人口基本信息和历史对出生登记、死亡注销、迁（划）入、失踪注销等户籍管理业务信息进行缺失数据补全；

（3）将不符合逻辑的异常户籍管理业务信息数据从户籍管理业务信息表中去除；

（4）对户籍人口基本信息、户籍管理业务信息按公民身份号码、姓名、办理日期等数据项进行去重；

（5）从户籍人口基本信息和历史中按人口管理注销类别代码，将未备案上报的死亡、出国定居、失踪注销、失踪寻回等注销类业务信息补全至对应业务表。

（五）地址及城乡属性建模

通过地址及城乡属性建模，获取全国户籍地址的城乡分类属性，县区、乡镇、村居委会等地址集合的城乡属性；通过机器学习，分析、获取地址关键字城乡分类属性，为后续城乡分类标识做准备。

（六）地址匹配及城乡分类属性标识

1. 准标户籍地址匹配模型

应用于户籍人口基本信息、户籍人口基本信息历史以

及包含户籍地址的户籍管理业务信息，主要包括标准户籍地址匹配、行政区划城乡分类属性匹配和关键字匹配。

2. 地址信息匹配模型

用于处理死亡、出国定居、服兵役等不包含户籍地址信息的户籍管理业务信息和出生登记、迁（划）入、退伍转业、回国定居等业务表中户籍地址项为空的户籍管理业务信息数据集。通过关联户籍人口基本信息、户籍人口基本信息历史，获取户籍地址信息和城乡分类属性信息。

（七）户籍人口变动统计

分别进行户籍人口基本情况统计、户籍人口增量统计、户籍人口减量统计。根据户籍基本情况统计、增量统计、减量统计等结果，推算各时期人口总数、城镇人口数和农村人口数，生成包括行政区划、自 2014/1/1 以来的半年日期、人口总数、市辖区人口数、城镇人口数、人口增加数、城镇人口增加数、农村人口增加数、城镇出生人口数、农村出生人口数、城镇死亡人口数、农村死亡人口数、城镇迁入人口数、农村迁入人口数、城镇迁出人口数、农村迁出人口数等数据项的基础性多维表。按行政区划标识信息，抽取、分类、汇总多维表，生成全国、省级、地市级以及商圈、方位等各类统计表。

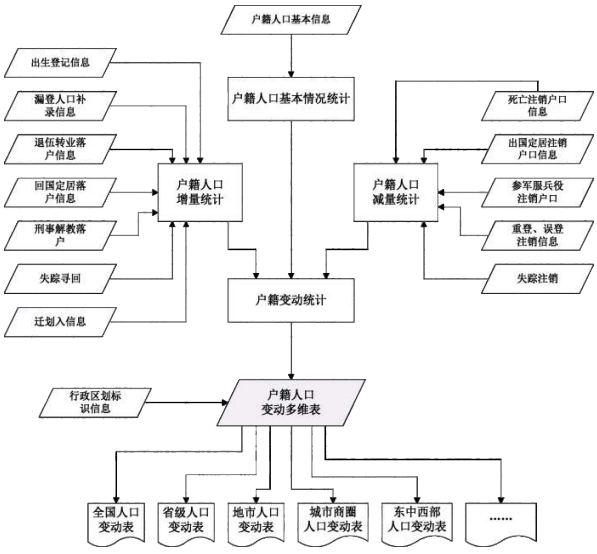

图 2　籍人口变动统计分析流程图

（八）模型性能验证

与传统的统计分析相比，本研究采用的大数据平台在计算性能、性价比和工作效率上具有明显优势。

1. 算性能

在计算性能方面，大数据平台采取分布式内存计算，数据从 HDFS 读入内存后并行计算，性能远高于传统技术架构。

本研究共采用 9 计算节点，每个节点包括 48 核CPU，共计 432 核CPU。以地址城乡分类统计标识为例，算法计算相当于 70 亿次常规数据库访问操作，按部库二期系统关系型数据库服务性能 1000 万tpc-c 计算，数据库满性能计算至少需要 12 个小时，而大数据平台型验证系统在 6 分钟内完成算法计算，数据访问性能提升 200 倍

2. 作效率

对海量数据的抽取、分析、挖掘、输出，大数据平台建立了一套敏捷的开发模型，工作效率远高于传统模型。

本研究在 1 个月内，对 15 个表完成了数据质量分析，9 轮数据清洗和 10 轮全表统计。

后期算法模型建立后，从 15 个表、60 亿条记录中建立多维表，生成全部户籍人口变动、迁移、城镇化、农村转移城镇等 100 多张统计报表，并将统计报表输出报表文件，最后对报表文件处理等全过程，在 2.5 小时内全部完成。

3. 价性比

本次研究内容原型验证系统采取了 10 台PC 服务器，总投入约为传统架构硬件成本的 1/10。因此，从数据统计、分析角度，采用大数据平台技术路线进行人口统计应用，具有超高的高性价比。

四、总结

通过本文统计方法，能够满足户籍制度改革中期评估对人口信息统计分析的工作要求，同时由于基于大数据分析管理平台，在抽取全量户籍人口信息、建立地址匹配和城乡分类属性标识算法模型时，比基于传统业务算法的统计方法在投入资源的性价比和工作效率上具有明显优势。

综上所述，人口管理大数据应用级建模技术研究是公安部部级人口大数据分析管理平台建设的前瞻性研究，为全国户籍人口数据统计常态化工作机制的建立起到了推动作用，具有指导性意义。同时，本研究中的户籍人口数据清洗、地址匹配等中间成果，能够为人口信息统计分析提供有效的技术手段。

参考文献

[1] 占玉宏，明睿. 口信息数据集成模型研究 [J]. 算机应用与软件，012, 8(2): 53-1154.

[2]武洁，校伟 .我国人口城镇化率统计与推算方法探讨 [J].调研世界，013, 7): (4-46.

[3] 超王. 口统计信息化建设探讨 [J]. 代经济信息，015(7).