Journal of Computer Applications
计算机应用,2016,36(1):44 -51
ISSN 1001-9081
CODEN JYIIDU
2016-01-10
http://www.joca.cn

文章编号:1001-9081(2016)01-0044-08
DOI:10.11772/j.issn.1001-9081.2016.01.0044

基于手机轨迹数据的人口流动分析

孔扬鑫,金澈清*,王晓玲
(华东师范大学 数据科学与工程研究院,上海 200062)
(* 通信作者电子邮箱 cqjin@sei.ecnu.edu.cn)

摘　要: 随着通信技术的发展和智能手机的普及,运营商基站所采集的大规模手机轨迹数据在城市规划、人口迁移等领域中发挥了重要价值。针对城市人口流动问题,提出一种利用手机轨迹数据的基于轨迹行为特征的人口流动判定(MF-JUPF)算法。首先,可对手机轨迹数据进行数据预处理,以提取用户活动轨迹;然后根据进出城市的行为模式提取重要特征,再根据真实标注数据集合利用多种分类模型进行参数训练;最后,根据模型训练结果判定用户轨迹是否为进出城市行为。所提系统使用 MapReduce 框架进行数据分析,以提高性能和可扩展性。基于真实数据集合的实验结果表明,对于进出城市的判定,该方法的准确率和召回率可达 80% 以上,与基于信号消失时长的人口流动判定(SD-JUPF)算法相比,在判定进入城市的准确率上提高了 19.0%,召回率提高了 13.9%;在判定离开城市的准确率上提高了 17.3%,召回率提高了 6.1%。相比非过滤算法,根据手机轨迹数据特点进行的数据过滤算法可减少处理时间 36.1% 以上。理论分析和实验结果表明 MF-JUPF 方法精度高,可扩展性好,因此对城市规划等领域有重要应用价值。

关键词: 基于位置服务;手机轨迹数据;人口流动;城市规划;MapReduce
中图分类号: TP311　　**文献标志码:** A

Population flow analysis based on cellphone trajectory data

KONG Yangxin, JIN Cheqing*, WANG Xiaoling
(*Institute for Data Science and Engineering*, *East China Normal University*, *Shanghai* 200062, *China*)

Abstract: With the development of communication technology and popularization of smartphones, the massive cellphone trajectory data gathered by base stations plays an important role in some applications, such as urban planning and population flow analysis. In this paper, a Movement Features-based Judging Urban Population Flow (MF-JUPF) algorithm utilizing cellphone trajectory data was proposed to deal with the issue about the population flow. First, users' activity trajectories were mined from cellphone trajectory data after data preprocessing. Second, the movement features were extracted according to the pattern of entering and leaving a city, and the parameters of these features were trained using various classification models upon real data sets. Finally, trained classification models were used to judge whether a user came in/out of the city. To enhance the efficiency and scalability, a MapReduce-based algorithm was developed to analyze massive cellphone trajectory data sets. As reported in the experimental part upon real data sets, the precision and recall of the proposed solution to judge the entering and leaving behaviors were greater than 80%. In comparison with Signal Disappears-based Judging Urban Population Flow (SD-JUPF) algorithm, the precision and recall of entering city judgment increased by 19.0% and 13.9%, and the precision and recall of leaving city judgment increased by 17.3% and 6.1%. Compared with the non-filtering algorithm, the time cost of the improved filtering algorithm was reduced by 36.1% according to the traits of these data. The theoretical analyses and experimental results illustrate the high accuracy and flexibility of MF-JUPF which has applicable values in urban planning and other fields.

Key words: location-based service; cellphone trajectory data; population flow; urban planning; MapReduce

0　引言

城镇化的持续发展使得城市人口剧增、城市人口迁移更加显著。对于城市规划者而言,实时掌控监测人口迁移情况非常重要,掌握进出城市的人口数量,合理分配社会资源,以应对交通压力、维护社会公共治安等问题。

然而,准确估算进出城市的人口流量并不容易。传统方法包括:人工统计、基于超声波或红外线方法以及基于视频图像方法等。人工统计是指在特定区域内统计流动情况,该方法受限于人力,不够精确;后两者分别利用信号反射机制和图像分析算法捕捉人类移动行为。然而,当人流量较大时,信号反射机制无法统计人群中间的人员数量,视频图像中的画质也将急剧下降,影响识别效果,因此亟需开发新的方法以解决上述问题。

收稿日期:2015-09-15;修回日期:2015-10-12。
基金项目:国家 973 计划项目(2012CB316203);国家自然科学基金资助项目(61170085,61472141,61370101)。
作者简介:孔扬鑫(1991 -),男,河北辛集人,硕士研究生,主要研究方向:位置服务技术及应用、数据挖掘;　金澈清(1977 -),男,浙江文成人,教授,博士生导师,博士,CCF 会员,主要研究方向:数据流管理、基于位置服务、不确定数据管理;　王晓玲(1975 -),女,山东烟台人,教授,博士生导师,博士,CCF 会员,主要研究方向:面向数据密集型计算的数据管理、位置服务技术及应用。

智能手机的迅速发展与普及为流动人口估算提供了可能。从全世界范围来看,2015 年全球的移动电话保有量将超过全球的人口总量(http: //tech. huanqiu. com/comm/2014-06/5011776. html) ; 从我国来看,截至 2015 年 6 月,我国手机网民规模达 5. 94 亿(http: //www. cnnic. net. cn/gywm/xwzx/rdxw/2015/201507/t20150723_52626. htm) 。当手机用户在主叫、被叫、短信收发或者访问互联网时,运营商均会记录相关的轨迹数据,因此,用户每天使用手机将产生大量轨迹数据,该信息中包含基站的位置,即可从侧面反映用户所处的位置。此外,由于基站覆盖范围大,用户可能处于某基站方圆几百米甚至几千米的信号覆盖范围内任意一点,因此位置精度远不及全球定位系统(Global Positioning System,GPS) 技术[1] 。在基站边缘处的信号不稳定,手机信号可能存在频繁切换的现象。虽然手机轨迹数据总量大、覆盖面广,但低质量数据为城市人口流动的监测带来了较大挑战。

由于移动运营商多以城市为单位管理手机数据,难以同时掌握周边城市的手机连接数据,因此常基于信号消失时长的人口流动判定(Signal Disappears based Judging Urban Population Flow,SD-JUPF) 算法来进行统计。图 1 所示为利用手机轨迹数据统计人流量的一种常用方法。图中 A、B、C、D、E 为 5 座城市,C_1、C_2 为基站,其中 C_2 位于火车站,P_1、P_2 为相邻两条手机连接记录,连接时刻分别为 t 和 $t+\delta$ 时刻,现需判定该用户是否有进出 A 城市行为。若 δ 超过较长时间(例如 3 天) ,且用户在 P_2 处与火车站范围内的基站进行了交互,由此该方法认为用户在 t 时刻离开了 A 城市,并于 $t+\delta$ 时刻进入 A 城市。

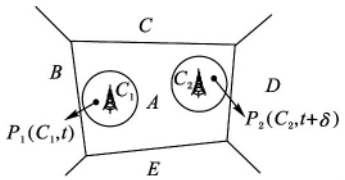


图 1　常用方法

该方法忽略了不常使用手机(如老年人) 且在车站周边生活或工作的人群。虽然他们手机的相邻两条连接记录的时间间隔很长,并且存在与火车站范围内基站的连接记录,但他们却一直处在城市内部,未有进出城市行为,因此,上述方法在进出城市人流量估计上差错率较大。

本文提出基于轨迹行为特征的人口流动判定(Movement Features based Judging Urban Population Flow,MF-JUPF) 算法,通过分析用户基站记录间的关系来生成用户活动轨迹,提取进出城市行为的轨迹特征,利用分类模型判定用户是否有进出城市行为。

本文的主要贡献包括:

1) 充分考虑手机轨迹的特点,根据原始轨迹数据挖掘用户的行为活动,力图避开低质数据的影响。

2) 根据多数人的行为轨迹,挖掘用户进出城市过程中的行为模式,提取重要特征,并利用分类模型进行参数训练。

3) 基于真实数据集进行了验证分析,证明了本文所提方法的正确性和有效性,提高了判定进出城市行为的准确度。

1　相关工作

近年来,针对手机轨迹数据的研究得到空前发展,多数研究工作主要从个人行为和城市分析的角度开展研究。在个人行为研究方面,文献 [2-4] 不断深入判定移动行为的起始地点,利用 OD(Origin-Destination) 矩阵描述个人移动轨迹。其中文献 [4] 将基站的连接记录与特定区域内的交通流量相结合,建立用户基站到基站的 OD 矩阵,从而进一步分析用户轨迹。

为挖掘用户行为模式,文献 [5] 从个人手机轨迹数据中发现其可能活动地点,再将以时间先后顺序串联的活动地点序列进行聚类分析,挖掘用户活动的转移模式。文献 [6] 将地图网格化,根据 POIs(Points of Interest) 数据将格子聚类成不同功能区域(餐饮、购物、娱乐等) ,再将用户每天不同时段所处的功能区域串联,得到行为模式。另外,重要地点挖掘是个人行为模式分析的重要一环,文献 [7] 根据连接时间和连接频率,将城市中的基站聚类,可找到城市中较为重要的区域,根据人们生活习惯,分析个人数据在不同时段的聚类情况,即可得到用户的居住和工作地。用户行为模式的挖掘有助于进行社交活动识别[8] 以及社交活动推荐[9] 。

对城市的分析更多是建立在个人轨迹分析的基础上进行的,文献 [10] 分析实时采集的移动手机数据,模拟城市交通情况,监测城市交通情况。除了应用于交通方面,该数据还可结合其他数据源,帮助人们更好理解城市的发展。文献 [11] 根据城市经济水平分布,将城市划分成不同收入水平的区域,挖掘用户轨迹与收入有关的特征,利用支持向量机(Support Vector Machine,SVM) 和随机森林模型建模,预测个人经济水平。文献 [12] 结合匿名手机数据和 Flickr 的带有位置信息的图片数据来展现城市中不同地区吸引力的发展趋势。

在城市人口迁徙领域,文献 [13-16] 通过分析人类移动模式,论证了人类轨迹的规律性,为人口流动分析提供理论基础。然而由于手机轨迹数据质量低以及移动运营商数据限制等因素,近年来难有对该领域上的研究突破,因此本文研究进出城市人流量估算问题,提出了 MF-JUPF 方法,可为今后的研究工作提供了一种新的思路。

2　问题定义

当用户使用手机进行通话、短信、上网的服务时,均需要与基站进行交互,基站会记录交互日志,并汇总到某个数据中心。日志记录包含用户 ID、基站编号和连接时间等。本文通过该三元组建立数据模型。

定义 1　原始轨迹 $Otraj$(Original Trajectory) 。$Otraj = \{P_i\}$,包含多个轨迹点,其中: $P_i = (Cid, T)$,T 表示时刻,$Cid = (x, y)$ 表示基站的经纬度。

例如某用户在某时段内多次和城市内基站进行交互,每次交互产生的连接记录均包含 Cid 和时间戳,则可确定一个轨迹点 P_i,该时段内的所有 P_i 按时序排列即为一条原始轨迹 $Otraj$。

定义 2　枢纽区域 TR(Transportation Region) 。该区域为能够覆盖某交通枢纽区域(汽车站、火车站、飞机场以及重要码头) 内所有基站的空间范围,用圆表示,$TR = (TR_i, x, y, r, C)$,其中: TR_i 为枢纽区域标识,x、y 分别为圆心的经纬度坐标,r 为半径,$C = \{Cid\}$ 涵盖了该区域内的所有基站。

定义 3　边境区域 ER(Edge Region) 。该区域为用户进出城时经过城市内某些特定区域中的基站,$ER = \{Cid\}$。特定

区域包括:飞机场、高速公路口、国道口、省道口以及客运港口。

定义4 停留位置 L(Location)。$L=(x,y,T_{in},T_{out})$,x 和 y 为基站的经纬度坐标,T_{in}、T_{out} 为在该位置停留的起止时刻。若时序上相邻的两轨迹点 P_i、P_{i+1} 的连接基站相同,则将两轨迹点合并为一个停留位置,$T_{out}-T_{in}$ 为用户在该位置的停留时长。

图2中 L_1 至 L_7 均为停留位置(图中 t_i 仅表示各停留位置在时序上的先后次序),若用户在某 P_i 处仅有一次记录,则此停留位置的 $T_{in}=T_{out}$。

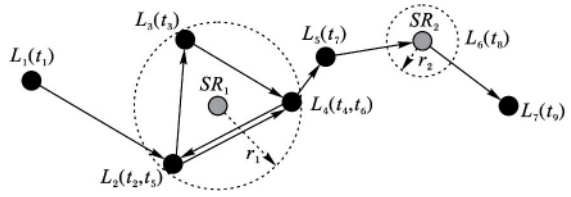


图2 个人移动轨迹

定义5 移动轨迹 $MTraj$(Moving Trajectory)。用户的一条移动轨迹 $MTraj$ 由一系列时序的停留位置 L 组成,$MTraj=\{L_i\}$。

图2所示情况下,按定义4中的方法处理原始轨迹即可得到该移动轨迹 $MTraj=\{L_1,L_2,L_3,L_4,L_2,L_4,L_5,L_6,L_7\}$。

定义6 停留区 SR(Stay Region)。某一停留区表示为一圆形区域,$SR=(SR_i,x,y,r,SR\_T_{in},SR\_T_{out})$,$SR_i$ 为停留区标识,x 和 y 分别为圆心的经纬度坐标,r 为半径,SR_T_{in} 为用户进入该停留区的时刻,SR_T_{out} 为离开时刻,该区域表示用户的一个活动区域。

例如图2轨迹 $MTraj$ 中,停留位置 L_2、L_3、L_4 之间满足一定时间和空间限制(将在3.1节详细说明),说明用户在该区域进行某种活动的可能性较大,确定为 SR_1,其范围为能够覆盖该组位置点的最大圆,半径为 r_1,且 $SR\_T_{in}=L_2.T_{in}$,$SR\_T_{out}=L_4.T_{out}$;用户在 L_6 处停留时间较长,满足时间约束但与 L_5 和 L_7 不满足空间约束,因此可确立 SR_2 为以 L_6 为圆心,r_2 为此处基站覆盖范围的半径,$SR\_T_{in}=L_6.T_{in}$,$SR\_T_{out}=L_6.T_{out}$。

定义7 活动轨迹 $ATraj$(Activity Trajectory)。一条活动轨迹 $ATraj$ 由一系列时序的停留区 SR 组成,$ATraj=\{SR_i\}$。例如图2中时序串联的 SR_1、SR_2 即为用户的活动路线 $ATraj=\{SR_1,SR_2\}$。

定义8 行为标识 $MFlag$(Moving Flag)。$MFlag=(uid,flag,T)$。其中 uid 为用户标识,$flag$ 取值范围为 -1、1 或 0,T 为时刻。$flag=1$ 或者 -1 表示用户 uid 在时刻 T 进入或者离开城市;$flag=0$ 表示无进出城市情况。

定义9 进出行为查询。给定轨迹集合 S,从中找出轨迹集合 S',$S'\subset S$,且对于 $\forall MTraj\in S'$ 需满足:$\exists T$ 时刻的 $MFlag.flag=1$ 或 -1。

例如,给定多名用户的轨迹,根据本文的处理方法,返回用户 u_1 的某段轨迹 $MTraj$,且该段轨迹中用户 u_1 在 T_1 时刻的行为标识 $MFlag=(u_1,-1,T_1)$ 或 $MFlag=(u_1,1,T_1)$,$flag=-1$ 表示用户 u_1 在 T_1 时刻离开城市,$flag=1$ 表示用户 u_1 在 T_1 时刻进入城市。

3 MF-JUPF 解决方案

图3为本文所涉及的 MF-JUPF 方法的处理框架,主要分为数据预处理和进出城市流动分析两部分。其中数据预处理部分处理用户原始轨迹,生成用户移动轨迹和活动轨迹,并根据基站信息表得到枢纽区域和边境基站表。进出城市流动分析主要包括模型训练和轨迹判定,X_i^{in} 为用户进入城市时的轨迹特征向量,X_i^{out} 为用户离开城市时的轨迹特征向量。

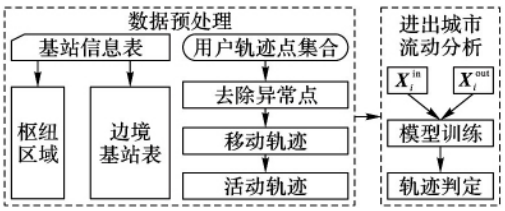


图3 处理框架

3.1 数据预处理

为解决基站定位精度低和手机信号在基站间频繁切换所造成的数据低质问题,本文通过挖掘用户活动的重要区域,来减少用户在区域内定位不准和信号频繁切换所带来的影响。

3.1.1 去除异常点与建立移动轨迹

数据中存在基站日志异常情况,如图4所示。图4中包括5个基站,即 A、B、C、D 和 E。某用户沿 $a\rightarrow b\rightarrow d$ 轨迹移动,手机与基站日常切换次序为 $A\rightarrow B$,但实际记录显示在用户移动到 b 点后,下一条的基站记录为 c 点,之后再到 d 点,即基站切换次序为 $A\rightarrow C\rightarrow B$。由于基站 C 与 A、B 较远,且相隔多个基站,因此基站 C 日志异常。

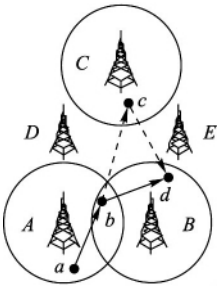


图4 信号异常切换

针对上述情况,本文提出基于时间和距离的轨迹去噪算法 processOTraj。

算法1 processOTraj。

输入: 原始轨迹 $otrajList$,时间阈值 θ_{t1},距离阈值 θ_{dis}。

输出: 去噪轨迹 $cOtrajList$。

```
1)   i = 2; cOtrajList = otrajList;
2)   while (2≤i≤|cOtrajList|-2) do
3)     if (Dis(i, i+1) > θdis and Time(i,i+1) < θt1) then
4)       if (Dis(i+1,i+2) < θdis and Time(i+1,i+2) < θt1)
           then
5)         cOtrajList.remove(i);            /* i 为异常点* /
6)       end if
7)       if (Dis(i-1,i) < θdis and Time(i-1,i) < θt1) then
8)         cOtrajList.remove(i+1);          /* i+1 为异常点* /
9)       end if
10)    end if
11)  end while
12)  return cOtrajList;
```

由于存在手机信号在两个邻近基站之间频繁切换的情况,因此将 θ_{dis} 设置为基站覆盖范围直径,以过滤上述情况,并且利于发现信号频繁切换时用户较为可能的位置。基于实验结果分析,多数用户在 10 min 内位置移动概率较小,因此 θ_{dis} 可被设置为 10 min,以捕捉到用户可能的停留位置。算法 1 首先初始化,令去噪后的轨迹 $cOtrajList = otrajList$,然后第 3) 步判断轨迹点 P_i 和 P_{i+1} 是否有异常点,若有则在第 4) ~ 9) 步根据相邻记录的距离和时间因素确定两记录中的异常点,并去除。其中 $Dis(i, i+1)$ 返回 P_i 与 P_{i+1} 间的欧氏距离,$Time(i, i+1)$ 返回 P_i 与 P_{i+1} 的相隔时间。算法在第 4) ~ 6) 步和第 7) ~ 9) 步通过比较预先设定好的距离阈值 θ_{dis} 以及时间阈值 θ_{t1} 判断当前位置点前后是否为异常点,若为异常点则去除。

为生成用户活动轨迹,在去除轨迹异常点后需进行位置合并,即建立移动轨迹。若去噪轨迹 $cOtrajList$ 相邻记录的连接基站相同,且相隔时间小于 θ_{t1},则合并该相邻记录为一个停留位置,只需遍历一遍轨迹序列即可建立用户移动轨迹 $mtrajList$(该过程的处理算法较为简单,此处不再赘述)。

3.1.2　生成用户活动轨迹

在数据预处理之后,可以尝试生成活动轨迹。该步骤的重点是发现用户停留区 SR,需要考虑时间和空间两方面因素。停留区内的所有停留位置之间需满足以下公式:

$$L_{i+1}.T_{in} - L_i.T_{out} < \theta_{t1} \quad (1)$$

$$Dis(L_i, L_j) < \theta_{dis}; \quad \forall i, j \quad (2)$$

$$L_{first}, L_{last} \in SR_i;\ L_{last}.T_{out} - L_{first}.T_{in} > \theta_{t2} \quad (3)$$

上述公式中,L_i 和 L_{i+1} 分别表示两个相邻的停留位置,$Dis(L_i, L_j)$ 返回两停留位置间的欧氏距离,式(3)中 L_{first} 为 SR 中最早进入的停留位置,L_{last} 为最晚离开的停留位置,该时间跨度用以约束 SR 的最小停留时长,若用户仅在某一停留位置 L 的停留时间较长,且与其相邻停留位置不满足式(1)或(2),则可构成一个单点停留区,即 $SR = \{L\}$。基于实验统计,发现用户在某区域活动时持续的时间多数超过 1 h,因此可将 θ_{t2} 设置为 1 h。活动轨迹生成算法 findATraj 描述如下:

算法 2　findATraj。

输入: $mtrajList$,时间阈值 θ_{t1}、θ_{t2},距离阈值 θ_{dis}。

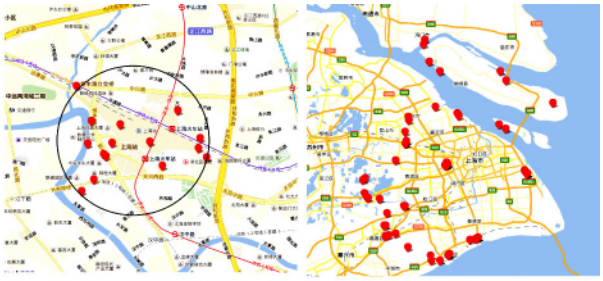
输出: $atrajList$。

```
1)   i = 1; SR = atrajList = ∅;
2)   while (i ≤ |mtrajList|) do
3)     j = i;
4)     locationList.add(L_j);        /* L_j 组成临时停留区 SR* /
5)     while (j+1 ≤ |mtrajList|) do
6)       if (L_{j+1}.T_in - L_j.T_out < θ_t1 and
7)         disCondition(locaitonList, L_{j+1}, θ_dis)) then
           /* 式(1) ~(2) * /
8)         locationList.add(L_{j+1});
9)         j = j + 1;
10)      else
11)        break;
12)      end if
13)    end while
14)    if (L_{j-1}.T_out - L_i.T_in < θ_t2) then        /* 式(3) * /
15)      SR = buildSR(locationList);
16)      atrajList.add(SR);
17)    end if
18)      i = i + 1;
19)    SR = ∅;
20)  end while
21)  return atrajList;
```

算法 2 针对单个用户历史轨迹生成活动轨迹,因此初始时该用户的活动轨迹中无停留区域,即第 1) 步。第 2) ~ 20) 步判断 3.1.1 节中得到的移动轨迹中的每个停留位置是否属于某停留区。为防止遗漏单点停留区,第 4) 步首先将起始判断位置 L_j 列入临时停留区队列。第 5) ~ 13) 步保存满足式(1) ~ (2) 的停留位置,之后在第 14) 步判定是否构成停留区或临时停留区。在算法中,buildSR($locationList$) 函数返回 $locationList$ 中所有停留位置的最小覆盖圆;disCondition($locationList, L_j, \theta_{dis}$) 函数将 L_j 与临时停留区 $locationList$ 中所有的停留位置进行距离测算,判断是否满足式(2),满足则返回 true;否则返回 false。本文略去这两个算法的细节描述。

3.1.3　建立枢纽区域和边境区域

本文利用基站表信息表中各个基站地理坐标,并根据定义 2 和定义 3 区分枢纽区域和边境区域,其中将各枢纽区域(火车站、机场等)表示为圆形区域,例如图 5(a)为上海火车站的枢纽区范围,而边境区域的分布如图 5(b)所示,并维护一张属于边境区域的基站信息表。(出于保护用户隐私和数据安全的目的,图 5 中基站位置仅作为示意图,与真实位置无关。)



(a) 某枢纽区域示例　　(b) 边境区域

图 5　两类区域示例

3.2　特征提取与模型应用

用户行为轨迹复杂,本文通过分析用户进出城市的轨迹行为特点,对照非进入非离开城市行为,挖掘轨迹特征。下面将对各个特征作具体说明。

3.2.1　信号消失时长(T_{miss})

该特征计算为 $T_{miss} = L_i.T_{out} - L_j.T_{in}$,其中 L_i 与 L_j 为相邻的两个停留位置。在用户离开城市而后再进入城市的过程中,必然在一定时间范围内其手机信号缺失,即在 $L_i.T_{out}$ 时刻与 $L_j.T_{in}$ 之间缺失。

3.2.2　枢纽区域出现概率(P_{TR_i})

P_{TR_i} 描述手机信号消失之前或再次出现之后用户在枢纽区域 TR_i 出现的可能性。现实生活中,人们大多通过现有的交通工具(汽车、火车、飞机等)进出城市,而且用户以某种交通方式离开后,更倾向于以同样的交通方式回到城市。本文通过分析用户停留区域与枢纽区域之间的面积关系,建立枢纽区出现的概率模型。

图 6 中 TR_i 为某枢纽区的面积,SR_{out} 为用户离开前的最后一个停留区面积,SR_{in} 为用户进入城市后的第一个停留区

的面积。$A(i)$ 表示 i 区域的面积，因此用户在 TR_i 枢纽区出现的概率为：

$$P_{TR_i}^{\text{out/in}} = \frac{1}{2} \times \frac{A(SR_{\text{out}} \cap SR_{\text{in}})}{\min(A(SR_{\text{out}}), A(SR_{\text{in}}))} \times \left[\frac{A(SR_{\text{out}} \cap TR_i)}{\min(A(SR_{\text{out}}), A(TR_i))} + \frac{A(SR_{\text{in}} \cap TR_i)}{\min(A(SR_{\text{in}}), A(TR_i))} \right] \quad (4)$$

$$P_{TR_i}^{\text{wOut}} = \frac{A(SR_{\text{out}} \cap TR_i)}{\min(A(SR_{\text{out}}), A(TR_i))} \quad (5)$$

$$P_{TR_i}^{\text{wIn}} = \frac{A(SR_{\text{in}} \cap TR_i)}{\min(A(SR_{\text{in}}), A(TR_i))} \quad (6)$$

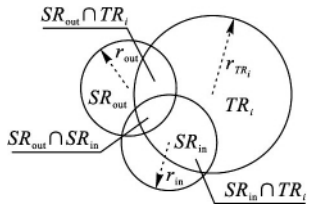


图 6 枢纽区域出现概率计算

$P_{TR_i}^{\text{out/in}}$ 指用户离开后再进入城市的概率，$P_{TR_i}^{\text{wOut}}$ 指用户离开后不再进入城市的概率，$P_{TR_i}^{\text{wIn}}$ 指用户进入后不再离开城市的概率。例如当用户的行为是离开城市后再返回城市，若枢纽区域 TR_i 分别与用户的 SR_{out} 和 SR_{in} 完全重叠，并且 SR_{out} 与 SR_{in} 也完全重叠，则说明用户离开城市和进入城市时都要通过该枢纽区域，即 $P_{TR_i}^{\text{out/in}} = 1$。实际应用中只需计算与用户距离最近的枢纽区的出现概率。

3.2.3 枢纽区域停留指数(N_{TR_i})

该指数描述用户在枢纽区 TR_i 出现的重要程度，如式(7)：

$$N_{TR_i} = (Time_{\text{stay}} + 1) / (Time_{\text{dis}} + 1) \quad (7)$$

其中：$Time_{\text{stay}}$ 为停留时长，即用户在枢纽区域活动的持续时间长度；$Time_{\text{dis}}$ 为相隔时长，即用户离开枢纽区域时刻至信号消失时刻的时间跨度。例如用户乘火车离开城市时，于 T_1 时刻进入等候区等待登乘，于 T_2 时刻登上火车离开车站，于 T_3 时刻离开城市，则 $Time_{\text{stay}} = T_2 - T_1$，$Time_{\text{dis}} = T_3 - T_2$。若在手机信号消失之前，用户在该枢纽区活动的持续时间较长，并且离开枢纽区之后信号便很快从城市中消失，则 N_{TR_i} 值较大，说明该用户通过乘坐该枢纽区的交通工具离开城市的概率较大。

3.2.4 是否在边境区域出现($isER$)

该特征用以说明手机信号消失之前或再次出现之后，用户有是否通过城市边境区域的基站。当用户离开或进入城市时，若用户在移动过程中持续使用手机，那么该用户所持手机较为可能与边境区域内的某基站进行过交互。

3.2.5 与居住地和工作地的平均距离(Dis)

该距离为用户手机信号消失时刻或再次出现时刻的位置与其居住地和工作地之间的平均距离，如式(8)：

$$Dis_{\text{out/in}} = \frac{1}{m+n}\left(\sum_{i}^{m} Dis_H^i + \sum_{j}^{n} Dis_W^j \right) \quad (8)$$

用户可能有多个居住地或工作地，以某用户离开城市为例，该用户有 m 个居住地和 n 个工作地，Dis_{out} 为离开城市时的待求距离，Dis_H^i 为与居住地 i(居住地编号，起始值为 0) 的距离，Dis_W^j 为该用户与工作地 j(工作地编号，起始值为 0) 的距离，因此该特征为用户离开城市时的位置与其各个居住地以及各个工作地的平均距离。用户进入城市时，Dis_{in} 的计算与之类似。

该特征可从侧面反映出用户离开或进入城市的倾向，相距越远，进入或离开城市的可能性越大。对于用户的居住地和工作地的挖掘，本文首先将城市进行栅格化处理，然后利用较长时间跨度内工作日时间(除节假日的周一至周五)的用户基站位置数据，针对不同时段(在家时段：00:00—06:00；工作时段：09:00—11:00，14:00—17:00)采用 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)算法对网格进行聚类，找出在家时段和工作时段内的重要网格，即得用户居住地和工作地的位置。利用 DBSCAN 算法可挖掘用户的多个居住地或工作地。

综上所述，对于进出城市行为而言，用户有 3 种行为模式：离开城市、进入城市、非进入非离开城市，用户的进出城市行为分别发生于 L_i 与 L_j 间 T_{miss} 的前后时刻，因此，在模型应用时，本文分别针对进入和离开城市情况，提取上述特征，得到下列特征向量：

$$\boldsymbol{X}_i^{\text{in}} = (T_{\text{miss}}, P_{TR_i}^{\text{in}}, N_{TR_i}^{\text{in}}, isER^{\text{in}}, Dis^{\text{in}}) \quad (9)$$

$$\boldsymbol{X}_i^{\text{out}} = (T_{\text{miss}}, P_{TR_i}^{\text{out}}, N_{TR_i}^{\text{out}}, isER^{\text{out}}, Dis^{\text{out}}) \quad (10)$$

根据实验的标注数据集(见 4.1 节)，可利用分类模型进行模型训练，将训练参数应用到分布式算法(见 3.3 节)中对进出城市行为进行判定。

3.3 分布式并行判定算法

为提升可扩展能力，本文基于 Map/Reduce 框架将同一用户的基站记录分配到同一个计算节点进行分析，利用多个计算节点并行处理用户的轨迹数据。在 Map 阶段将数据按用户 ID 分组进行输出；在 Reduce 阶段，各个 Reducer 在接收到相同用户 ID 号的数据之后，进行预处理操作，并判定进出城市行为。reduce 方法具体描述如下：

算法 3 reduce。

输入：$key = uid$，$values = \{\langle baseID, time\rangle\}$。

输出：$MFlag$。

```
1)  时序排列轨迹数据，生成原始轨迹 otrajList;
2)  执行 processOTraj 算法，生成去噪轨迹 cOtrajList;
3)  遍历 cOtrajList 生成移动轨迹 mtrajList;
4)  执行 findATraj 算法，生成活动轨迹 atrajList;
5)  isER^in = false; isER^out = false; i = 1;
6)  while (i ≤ |mtrajList|) do
7)     T_miss = mtrajList(i).T_out - mtrajList(i+1).T_in;
8)     if (T_miss < γ) then
9)        continue;
10)    end if
11)    if (mtrajList(i).baseID ∈ ER) then
12)       isER^out = true;
13)    end if
14)    if (mtrajList(i+1).baseID ∈ ER) then
15)       isER^in = true;
16)    end if
17)    Dis_out = avgDis(key, mtrajList(i).baseID);
18)    Dis_in = avgDis(key, mtrajList(i+1).baseID);
19)    P^out_TRi = getP(key, atrajList, mtrajList(i).T_out, -1);
20)    N^out_TRi = getN(key, atrajList, mtrajList(i).T_out, -1);
21)    P^in_TRi = getP(key, atrajList, mtrajList(i+1).T_in, 1);
```

22)　$N_{TR_i}^{\text{in}}$ = getN(*key*, *atrajList*, *mtrajList*($i+1$) . T_{in}, 1) ;
23)　赋值特征向量 $\boldsymbol{X}_i^{\text{out}}$ 和 $\boldsymbol{X}_i^{\text{in}}$;
24)　根据模型训练参数判定此刻行为,输出 *MFlag*;
25)　end while

该算法输入 *key* 为用户 ID,*values* 为对应用户的日志记录,每条记录由基站 ID 和连接时间组成。第 1) ~ 4) 步主要利用算法 1 和算法 2 进行数据预处理,得到移动轨迹和活动轨迹。第 7) ~ 20) 步针对用户每相邻两停留位置,计算各个特征值,得到特征向量 $\boldsymbol{X}_i^{\text{in}}$ 和 $\boldsymbol{X}_i^{\text{out}}$。其中: 第 7) 步计算得到 L_i 与 L_{i+1} 之间的信号消失时长; 第 8) ~ 10) 步可减少不必要的运算,加速进出城市行为的判定,具体细节将在 4.3 节介绍; 第 11) ~ 16) 步根据边境区域基站表,判断当前停留位置前后是否处于边境区域; 第 17) ~ 22) 步计算 3.2 节中其他特征维度。avgDis(*key*,*mtrajList*(i) . *baseID*) 函数返回 *mtrajList*(i) . *baseID* 与用户 *key* 的居住地和工作地的平均距离。getP(*key*, *atrajList*,*mtrajList*(i) . T_{out}, - 1) 函数返回用户 *key* 在 *mtrajList*(i) . T_{out} 之前最邻近时刻的 P_{TR_i},getP(*key*,*atrajList*, *mtrajList*($i+1$) . T_{in},1) 函数返回用户 *key* 在 *mtrajList*($i+1$) . T_{in} 之后最邻近时刻的 P_{TR_i}。getN 函数返回离开前或进入后的停留指数 N_{TR_i},与 getP 函数类似(3.2 节中列出了各特征值的计算方法) 。reduce 算法第 24) 步根据 3.2 节中分类模型(本文利用朴素贝叶斯、决策树和逻辑回归模型进行模型训练) 的训练参数进行该点的行为判定,对于进入城市的判定,将 $\boldsymbol{X}_i^{\text{in}}$ 与模型训练的特征权重向量相乘,如果概率大于 0.5,则用户于该位置点的对应时刻进入城市,*MFlag*. *flag* = 1; 否则 *MFlag*. *flag* = 0。离开城市的判定与之类似(判定为离开时, *MFlag*. *flag* = - 1, 否则 *MFlag*. *flag* = 0) 。最终算法输出 *MFlag*,之后可根据 *MFlag* 的时间信息,统计得到各时段进出城市的人流量。

3.4　算法性能分析

本节分析 MF-JUPF 方法的性能。假设单个用户手机轨迹数据记录数为 n,processOTraj 算法只需一次遍历该用户的连接记录,因此时间复杂度为 $O(n)$; 假设去噪后的记录数为 m, 则需 $O(m)$ 的时间建立移动轨迹; 假设单个用户的移动轨迹点有 p 个,在停留区的停留位置数为 l,则 findATraj 算法的时间复杂度为 $O(pl)$; 假设用户的手机轨迹数据可以形成 q 个停留区,则算法 3 中计算各个特征维度值时至多需要 $O(q)$ 的时间,若总共有 k 个用户,则算法 3 的时间复杂度为 $O(k(n+m+pl+pq))$,由于算法执行过程中需保留移动轨迹和活动轨迹,因此空间复杂度为 $O(k(p+q))$。

4　实验结果与分析

4.1　数据集描述

本文所采用的数据集包括基站连接日志和基站信息。表 1 为日志记录的数据结构,表 2 基站信息表结构。数据集为 49 名志愿者的连续 48 天(2014-09-15—2014-11-02) 手机基站连接记录,共 71 525 条。

表 3 为志愿者的性别、年龄分布情况。通过整理出行记录表,构成标注数据集,共计有 99 人次进入上海市,102 人次离开上海市。

表 1　基站连接日志(部分)

匿名后用户 ID	连接时间	基站 ID
U_1	2014-09-15T00:00:23	C1698
U_2	2014-09-15T00:24:15	C1755

表 2　基站信息(部分)

基站 ID	经度	纬度
C1698	121.668 43	31.256 64
C1755	121.681 81	31.264 75

表 3　志愿者情况分布

分类指标	指标值	所占比例/%
性别	男	56.30
	女	43.70
年龄	<25	26.30
	25 ~ 34	39.40
	35 ~ 44	22.10
	>44	12.20

4.2　结果分析

4.2.1　SD-JUPF

引言部分已简要介绍此方法,在此,该方法被用作对比实验。为分析城市人口流动情况,设定区域范围为城市大小,检测是否有 $T_{\text{miss}} > \beta$(参数 β 为设定阈值,以天为单位) 的情况并确定用户再次出现的位置,以判断是否有进出城市行为。通过分析准确率(*Precision*) 、召回率(*Recall*) 以及 F_1 值(F_1-Measure) 验证本文所提方法的正确性和有效性。

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

表 4 为设定不同 β 时的进出城市判定效果,其中取 β = 1.0 时效果较好。

表 4　不同 β 下的进出城市判定结果

天数 β	*Precision*/%	*Recall*/%	F_1/%
0.5	24.5	85.6	38.1
1.0	67.5	76.0	71.5
1.5	75.3	61.5	67.7
2.0	83.3	57.7	68.2
2.5	79.6	41.3	54.4
3.0	75.6	32.7	45.6

4.2.2　MF-JUPF

本文根据 3.2 节中各个特征值的提取方法,在用户相邻两停留位置间计算各特征值,得到式(8) 和式(9) 的特征向量,然后分别建立朴素贝叶斯(Naïve Bayesian, NB) 、决策树和逻辑回归(Logistic Regression,LR) 的分类模型进行用户进出城市行为的判定,其中决策树模型选用常用的 C4.5 算法。本文将 80% 的样本作为训练数据集,剩余 20% 为测试样本集。

图 7 为针对进入城市行为的实验结果,平均 *Precision* = 86.5%,*Recall* = 89.9%,F_1 = 87.9%; 图 8 为针对离开城市行为的实验结果,平均 *Precision* = 84.8%,*Recall* = 82.1%, F_1 = 88.5%。从图 7 ~ 8 中可看出,MF-JUPF 在 3 种分类模型下的效果均优于 SD-JUPF 的最好情况。

进入城市行为的判定效果优于离开城市行为的判定的原因是: 当用户进入城市时,手机通常会被动收到城市欢迎短消息,并伴随与边境基站的一次交互; 而用户在离开城市的过程中,不会有类似短信提示,若用户无其他使用手机的行为(接打电话等) ,则不会与边境基站交互,因此会影响 *isER* 特征的效果。

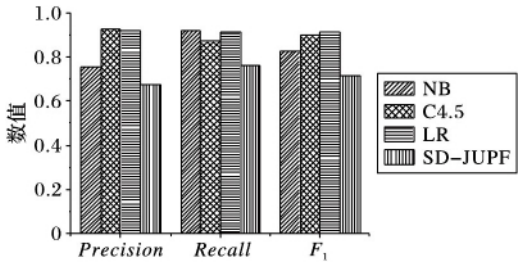


图 7 进入城市行为判定效果

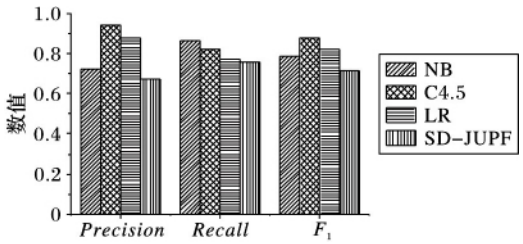


图 8 离开城市行为判定效果

对于进出城市行为的判定,3 种分类模型的效果与基准方法 SD-JUPF 相比有明显提高,其中逻辑回归模型效果相对较好。

4.3 性能优化

针对海量手机轨迹数据的运算,虽然 Map/Reduce 框架可将数据分散到多个计算节点进行分布式计算,但若总时间跨度较长,用户使用手机频度增加,单个用户的基站日志将急剧增长。

图 9 显示,T_{miss} 在 0 ~ 1 h 出现次数的比重占总连接次数的 87.5%。现实中,信号消失时长较短的情况下,用户发生进出城市行为的概率较小,因此算法 3 的第 8) ~ 10) 步对该类情况进行过滤,即排除 $T_{\text{miss}} < \gamma$ 的行为(参数 γ 为消失时长的设定阈值,粒度单位为 h) 。

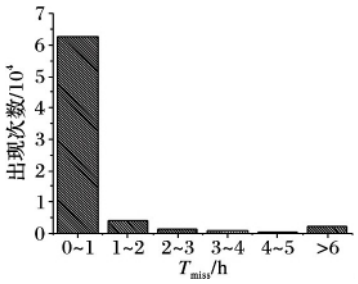


图 9 T_{miss} 分布

在相同实验环境下利用逻辑回归模型分类,图 10 ~ 12 为过滤算法的实验效果,其中图 10 纵坐标表示相比非过滤算法,过滤算法的处理时间减少率。比率越高说明处理速度越快、性能越好。从图中可知执行数据过滤的算法可减少处理时间 36.1% 以上。

图 11 ~ 12 说明,$\gamma < 2$ h,F_1 几乎不变,$\gamma > 2$ h,实验效果明显降低,这是由于 γ 设定较大时,将导致误删真正的进出城市行为,因此取 $\gamma = 2$ h 相对合理,该算法可在保证准确度的同时尽量排除无关行为,提高运行效率。

5 结语

随着移动通信技术的发展,手机等移动端的使用变得普遍,手机轨迹数据的增长为研究者分析个人行为或理解城市发展均提供了有力的数据支持。为减少手机轨迹数据低质的影响,识别用户的进出城市行为,本文从个人移动轨迹出发,建立用户活动轨迹,挖掘进出城市行为特征,提出了在 Map/Reduce 框架下的 MF-JUPF 方法。实验结果表明,该方法在保证运行效率的同时,可有效增强进出城市行为判定的准确度。手机轨迹数据量庞大,难以将大量样本进行标注,未来也可利用无监督的研究方法对此类问题进行探索。

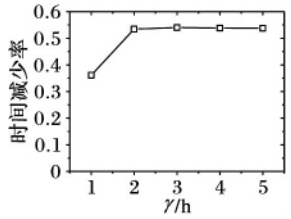


图 10 过滤算法的运行时间减少率

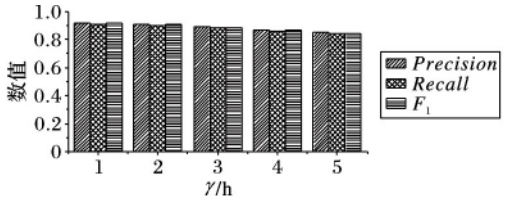


图 11 过滤后进入城市行为判定效果

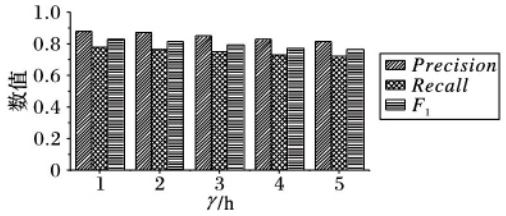


图 12 过滤后离开城市行为判定效果

参考文献:

[1] RATTI C, WILLIAMS S, FRENCHMAN D, et al. Mobile landscapes: using location data from cell phones for urban analysis [J]. Environment and planning B: planning and design, 2006, 33(5): 727.

[2] WHITE J, WELLS I. Extracting origin destination information from mobile phone data [C]// Proceedings of the 2002 Eleventh International Conference on Road Transport Information and Control. Stevenage, UK: IET, 2002: 30 - 34.

[3] CACERES N, WIDEBERG J P, BENITEZ F G. Deriving origin destination data from a mobile phone network [J]. IET intelligent transport systems, 2007, 1(1): 15 - 26.

[4] IQBAL M S, CHOUDHURY C F, WANG P, et al. Development of origin-destination matrices using mobile phone call data [J]. Transportation research part C: emerging technologies, 2014, 40: 63 - 74.

[5] LIU F, JANSSENS D, CUI J X, et al. Building a validation measure for activity-based transportation models based on mobile phone data [J]. Expert systems with applications, 2014, 41(14): 6174 - 6189.

[6] PHITHAKKITNUKOON S, HORANONT T, DI LORENZO G, et al. Activity-aware map: identifying human daily activity pattern using mobile phone data [C]// HBU'10: Proceedings of the First International Conference on Human Behavior Understanding. Berlin: Springer, 2010: 14 – 25.

[7] ISAACMAN S, BECKER R, CÁCERES R, et al. Identifying important places in people's lives from cellular network data [C]// Proceedings of the 9th International Conference on Pervasive Computing. Berlin: Springer, 2011: 133 – 151.

[8] TRAAG V A, BROWET A, CALABRESE F, et al. Social event detection in massive mobile phone data using probabilistic location inference [C]// Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom). Piscataway, NJ: IEEE, 2011: 625 – 628.

[9] QUERCIA D, LATHIA N, CALABRESE F, et al. Recommending social events from mobile phone location data [C]// ICDM 2010: Proceedings of the 2010 IEEE 10th International Conference on Data Mining. Piscataway, NJ: IEEE, 2010: 971 – 976.

[10] CALABRESE F, COLONNA M, LOVISOLO P, et al. Real-time urban monitoring using cell phones: a case study in Rome [J]. IEEE transactions on intelligent transportation systems, 2011, 12(1): 141 – 151.

[11] SOTO V, FRIAS-MARTINEZ V, VIRSEDA J, et al. Prediction of socioeconomic levels using cell phone records [C]// Proceedings of the 19th International Conference on User Modeling, Adaption and Personalization. Berlin: Springer, 2011: 377 – 388.

[12] GIRARDIN F, VACCARI A, GERBER A, et al. Quantifying urban attractiveness from the distribution and density of digital footprints [J]. International journal of spatial data infrastructures research, 2009, 4: 175 – 200.

[13] GONZALEZ M C, HIDALGO C A, BARABASI A L. Understanding individual human mobility patterns [J]. Nature, 2008, 453(7196): 779 – 782.

[14] GIANNOTTI F, NANNI M, PEDRESCHI D, et al. Unveiling the complexity of human mobility by querying and mining massive trajectory data [J]. The VLDB journal, 2011, 20(5): 695 – 719.

[15] SIMINI F, GONZÁLEZ M C, MARITAN A, et al. A universal model for mobility and migration patterns [J]. Nature, 2012, 484(7392): 96 – 100.

[16] SONG C, QU Z, BLUMM N, et al. Limits of predictability in human mobility [J]. Science, 2010, 327(5968): 1018 – 1021.

Background

This work is partially supported by the National Basic Research Program (973 Program) of China (2012CB316203), the National Natural Science Foundation of China (61170085, 61472141, 61370101).

KONG Yangxin, born in 1991, M. S. candidate. His research interests include technology and application of location-based services, data mining.

JIN Cheqing, born in 1977, Ph. D., professor. His research interests include data stream management, location-based services, management of uncertain data.

WANG Xiaoling, born in 1975, Ph. D., professor. Her research interests include data-intensive computing management, technology and application of location-based services.

(上接第 26 页)

[6] **贾均刚,张炜,高宏**. TIDC: **一种基于属性划分的高频度关系数据压缩存储方法**[C]//**第二十五届中国数据库学术会议**(NDBC2008)**论文集. 桂林**[**出版者不详**],2008: 14 – 22. (JIA J G, ZHANG W, GAO H. TIDC: one kind based on attribute division high frequency relations data compression memory method [C]// Proceedings of the 25th National Database Conference. Guilin: [s. n.], 2008: 14 – 22.)

[7] **王振玺,乐嘉锦,王梅,等. 列存储数据区级压缩模式与压缩策略选择方法**[J]. **计算机学报**,2010,33(8): 1523 – 1530. (WANG Z X, LE J J, WANG M, et al. Row stored datum area level compact model and compression strategy choice method [J]. Chinese journal of computers, 2010, 33(8): 1523 – 1530.)

[8] MÜLLER I, RATSCH C, FRBER F. Adaptive string dictionary compression in in-memory column-store database systems [C]// Proceedings of the 17th International Conference on Extending Database Technology. Athens: [s. n.], 2014: 152 – 158.

[9] FAUST M, SCHWALB D, PLATTNER H. Composite group-keys space-efficient indexing of multiple columns for compressed in-memory column stores [C]// IMDM 2013: Proceedings of the First and Second International Workshops on In-Memory Data Management and Analysis. Berlin: Springer, 2014: 42 – 54.

[10] STONEBRAKER M, ABADI D J, BATKIN A, et al. C-store: a column-oriented DBMS [C]// Proceedings of the 31st International Conference on Very Large Data Bases. [S. l.]: VLDB Endowment, 2005: 553 – 564.

[11] **李超,张明博,邢春晓,等. 列存储数据库关键技术综述**[J]. **计算机科学**,2010,37(12): 1 – 7. (LI C. ZHANG M B, XING C X, et al. Survey and review on key technologies of column oriented database systems [J]. Computer science, 2010, 37(12): 1 – 7.)

[12] **康强强,江舟,金澈清,等**. TPCHSuite: **一个** TPC-H **自动化测试工具的设计与实现**[J]. **计算机研究与发展**,2013, 50(z1): 394 – 398. (KANG Q Q, JIANG Z, JIN C Q, et al. TPCHSuite: the design and implementation of an automatic test tool for TPC-H [J]. Journal of computer research and development, 2013, 50(z1): 394 – 398.)

Background

This work is partially supported by the National Natural Science Foundation of China (81273649), the Natural Science Foundation of Heilongjiang Province (F201434).

DING Xinzhe, born in 1990, M. S. candidate. His research interests include big data compression.

ZHANG Zhaogong, born in 1963, Ph. D., professor. His research interests include massive data mining, bioinformatics.

LI Jianzhong, born in 1950, professor. His research interests include massive data management and computing, wireless sensor network.

TAN Long, born in 1971, Ph. D., associate professor. His research interests include wireless cognitive network, data mining.

LIU Yong, born in 1990, M. S. candidate. His research interests include keyword search.