基于网格概率的离群点检测算法①

彭艳兵[1], 冯利容[1,2]

[1](烽火通信科技股份有限公司 IAO, 南京 210019)
[2](武汉邮电科学研究院 电信系, 武汉 430074)

摘 要: 随着移动网络、智能终端的迅猛发展, 基于位置的服务 LBS(Location-based Service)越来越热门, 因此基站位置信息的正确与否成为关注的重点. 针对基站地理位置存在部分错误这一现象, 提出了基于网格概率的离群点检测算法来核查错误的基站. 首先, 根据基站分布的规则将数据空间分成若干网格单元; 其次, 根据用户轨迹签到信息关联出其在动态时间范围内经过的基站序列, 将基站序列映射到网格中, 计算出临近网格单元集合; 最后, 根据基站分布特点对网格单元内目标基站的临近基站求隶属概率, 筛选出离群点, 即错误的基站. 实验表明, 该算法的时间复杂度低且核实准确率较高.

关键词: 基于位置的服务; 网格划分; 隶属概率; 离群点检测

Outlier Detection Algorithm Based on Grid Probability

PENG Yan-Bing[1], FENG Li-Rong[1,2]

[1](FiberHome Communications Science & Technology Development Co., Ltd. Nanjing 210019, China)
[2](Wuhan Research Institute of Posts and Telecommunications,Wuhan 430074, China)

Abstract: With the rapid development of the mobile networks and intelligent terminals, location-based service has become more and more hotter on the internet, therefore the correction of the base stations' position becomes a critical factor. For the wrong base stations are uncertain, it proposes a new detecting algorithm based on the probability of the near grids, which is used to verify the wrong base stations. Firstly, it divides the data space into some grids. Secondly, combining with the users' attendance location information, it gets the track of the base stations in a short dynamic time and maps them to the corresponding grids. Finally, referring to the position characteristics of the base stations, it could give the membership probabilistic and filter the outliers, that are the wrong base stations. The results show that the algorithm has low complexity and high accuracy of detecting the wrong ones.

Key words: location-based service; grid plot; membership probabilistic; outlier detection

近些年, 随着定位技术的日趋成熟以及定位设备的大量普及, 面向不同的应用领域的移动终端产生了大量的轨迹数据. 这些数据里面蕴藏着非常丰富的知识信息, 它能准确地反映人们的移动规律, 能生动地体现交通情况, 能正确地揭示道路结构[1]. 目前, 与之相关的数据挖掘的研究备受关注, 其中一项比较有现实意义的研究就是通过移动轨迹数据来检测核实错误的定位设备(本文指的是基站).

在移动通信发展的前几年, 各大运营商对于基站的建设没有统一的规划, 导致许多基站的维护变得非常困难. 目前随着网络时代的飞速发展, 移动通信进入了高速发展的通信时代. 因此, 移动基站坐标的正确对于网络发展来说, 显得越来越重要. 但是, 由于存在基站信息的变更、人工录入失误等因素, 导致基站的坐标数据可能存在 3%左右的错误, 同时运营商每年更新的基站信息不会实时同步到位置信息服务商. 所以高效地检测出错误的基站是非常有必要的. 由于错误基站的数量不会很多, 可以将其看做异常点、离

① 收稿时间:2015-08-17;收到修改稿时间:2015-10-26

群点，这种研究非常适用于离群点检测这一应用场景.

为了有效的检测出离群点，很多研究人员已经开发了大量的离群点检测算法，包括基于统计的离群点检测算法、基于距离的离群点检测算法、基于密度的离群点检测算法和基于深度的离群点检测算法等，其中基于距离的离群点检测算法包含并且扩展了基于统计的思想，需要首先确定参数，然后将非数值型属性转换成数值型数据，计算对象之间的欧式距离，最终确定离群点. 这算法比较容易理解，而且具有比较直观的意义，故在实际场景中的应用很多[2]. Knorr[3,4]等人最先提出了基于距离的离群点的概念. Ramasmawy 对基于距离的离群点的定义做了改进，通过对与对象距离最近的第 K 个对象之间的距离排序，将数据集中距离排在前面的 m 个对象标记为离群点[5]. FAngiulli 等提出离群点是数据集中与其 K 个最近邻居的平均距离最大的前 m 个对象[6]，主要通过比较对象与 K 个最近邻居的平均距离来检测离群点. 这些算法将需要 N^2 次的数据对象之间的距离计算，当 N 很大时就不适用了. 因此，本文在基于距离的思想上，提出基于网格概率的离群点检测算法，直接通过数据点处于邻近网格的概率来确定离群点，避免了计算 N^2 次的数据对象之间的距离，然后结合实际基站位置的空间位置关系，给出一种高效识别错误基站的方法.

1　基本理论知识

1.1 网格划分和数据映射

把地球看成一个平面图，选择一个中心点，中心点选择“赤道与本初子午线交叉点”，然后以这个中心点同时向上下左右按步长 0.01 度进行扩展，每扩展一次可以得到一个长和宽都为 0.01 度的正方形，此正方形则为一个栅格. 通过此算法划分出的每个栅格的地理面积约为 1.24 平方公里左右(地球两极点除外)[7,8].

按照上述划分方式，对本文实验数据空间进行划分. 任意给定一地理区域，将其表示成二维空间 M，按照经纬度方向分别划分为 a、b 等份，a>0 ,b>0 且划分的单位网格经纬度均为 0.01 度，这样区域被划分为 $a*b$ 个网格单元，

$$M=\{G_{1,1}, G_{1,2}, G_{1,n}, G_{2,1} \ldots, G_{a,b}\}, a>1$$

按照这种方式划分之后，每个网格都有自己唯一的编号标识. 各个维度划分的网格数可以按式(1)、(2)计算且向上取整:

$$a=\left\lceil \frac{a_{\max}-a_{\min}}{0.01} \right\rceil \quad (1)$$

$$b=\left\lceil \frac{b_{\max}-b_{\min}}{0.01} \right\rceil \quad (2)$$

其中，$a_{\max}$ ，$a_{\min}$ ，$b_{\max}$ ，$b_{\min}$ 为每个维度的最值，读取数据后，可形成如图 1 所示的网格.

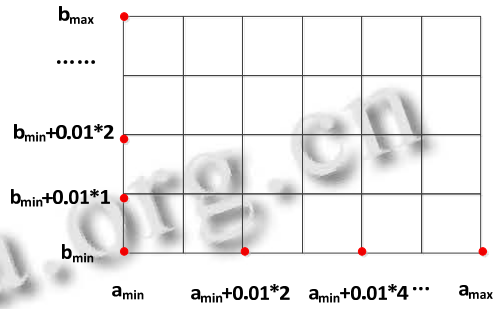

图 1　网格单元示意图

网格划分和数据映射的算法如下:

输入: 所有基站点坐标(数据集 D)、边长 r
输出: 数据点映射到网格坐标的数据集

```
Begin
    DataScan(D, r)    //数据读取，网格划分
 for  D 中的每个对象 p(a,b)
    amax=Math.max(a)
    amin=Math.max(a)
    bmax=Math.max(b)
    bmin=Math.max(b)
 end for
    row=(int)(( amax -amin)/r) +1 //网格行号
    colum=(int)(( bmax -bmin)/r) +1//网格列号
    DataGrid(D, r)   //数据对象映射到网格单元
 for  D 中的每个对象 p(a,b)
    r=(int)(( a -amin)/r) +1    //计算 P 的行号
    c=(int)(( b -bmin)/r) +1    //计算 P 的列号
    add  p(a,b) to G(r,c)
 end for
end
```

1.2 隶属概率

每个网格单元包含有两层邻居，第一层为紧邻数据点 $O(a,b)$，所在网格单元的周围的外部网格单元，第二层为紧邻第一层邻居周围的外部网格单元. 将这

两层邻居称为 *O(a,b)*所在单元网格的邻近网格单元.

对于数据点 *O(a,b)*, 假设 $O \in G_{i,j}$, 则 *O(a,b)*所在单元网格单元的邻近网格单元集为

$$G=\{G_{x,y} \mid x=i\pm n, y=j\pm n, 0\le n\le 2, x>0, y>0\}$$

数据集中 *S* 一个对象 *O(a,b)*为 DB*(P,N)*-Outlier, 如果它满足以下性质: 数据集 *S* 中至少 *q**100%的对象处于临近单元网格集 *G* 之外. 这里隶属概率为 *p*=1-*q**100%. 换句话说, 如果存在少于 *n* (*n*=*N**p*100%,*N* 为数据集的总数)个邻居数据点位于 *G* 集合以内, 则 *O(a,b)*是关于隶属概率为 *p* 的 DB*(P,N)* 离群点.

隶属概率公式为: *p=n/N* (其中 *n* 为处于的邻近网格单元集合 *G* 内的数据点个数, *N* 为数据空间中数据点的总个数.)

1.3 基站空间位置关系

参考文献[9]提出的结论可知, 无论城市区域还是乡村区域, 所有基站的位置分布并不是相互独立的, 城市内的基站分布表现了比较强烈的聚类特点. 而本文主要对密集城区的基站进行分析, 核查离群点基站. 显而易见, 密集城市中的基站与它近邻的基站之间的距离应该处于大致均匀的分布中, 而且不会相差很远.

考虑到用户通过蜂窝网进行 LBS 服务时, 需要与用户所处地理位置的基站进行交互, 日志信息中保留了基站的编号信息、上下线的用户经纬度信息、交互的时间信息等. 虽然基站的地理位置获取困难, 但是用户的签到信息获取相对容易. 对于基站密度比较密集的城区, 由于基站的覆盖范围较小, 用户的签到信息过于密集、越区切换较频繁, 可以利用用户一段时间的签到轨迹信息关联出一个基站 ID 网格序列(即临近基站), 然后根据这些临近基站的位置特点来核实目标基站的位置.

1.4 用户签到信息

如图 2 所示, 从海量日志文件中提取关键的特征信息来表示用户签到轨迹 *Tr*[10]. *Tr* 是具有时间戳的空间位置序列数据, *Tr*=(<A_1,Lat_1,Lat_1,T_1,$BSID_1$>,..,<A_i,Lat_i,T_i,$BSID_i$>,…,<A_n,lat_n,Lgt_n,T_n,$BSID_n$>)1 # *i* *n*, 其中 A_i 表示用户的手机账号信息, <Lat_i,Lgt_i>表示空间地理位置信息, T_i 为对应位置<Lat_i,Lgt_i>的时间戳信息, $BSID_i$ 为 A_i 在 T_i 时刻所处的基站覆盖范围内对应的基站 ID.

account,Latitude,Longitude,Time,base_station_id				
A1:	Lat1,	Lngt1,	T1 ,	******
A2:	Lat2,	Lngt2,	T2 ,	******
…				
An:	Latn,	Lngtn,	Tn,	******

图 2 海量日志提取的特征信息网格单元

1.5 基站的网格序列

每个基站 ID 也有唯一的地理位置<Lat_i,Lgt_i>. 通过用户的签到轨迹信息可以关联出该轨迹中基站所在的网格序列 *GridIDs*, *GridIDs* 是一个无序的网格集合. *GridIDs* 满足以下性质:

(1) A_i 在时间范围 Δt 内所处的地理位置<Lat_i,Lgt_i>对应的基站 ID 所映射的网格.

(2) 在 $\Delta t/2$ 时刻及很小的邻域范围内, $A_i(1\leqq i\leqq n)$ 经过了相同的基站, 该基站即为目标基站, 而根据时间 Δt 关联出的其他基站均为临近基站.

(3) *GridIDs* 中的元素均为临近基站所在的网格单元, 包括目标基站所在的单元网格.

如图 3 所示, 虚线表示用户 *A* 的签到轨迹信息, 实线表示用户轨迹对应的基站序列信息, 目标基站位于 $G_{2,2}$ 中, $G_{2,2}$ 的 $GridIDs=\{G_{1,1}, G_{2,3}, G_{1,3}, G_{2,2}\}$, 实际情况中经过目标基站的用户可能不止用户 *A*, 应该把所有符合情况的临近基站补充完整, 并且将这些临近基站所属网格也全部添加到 *GridIDs* 集合中.

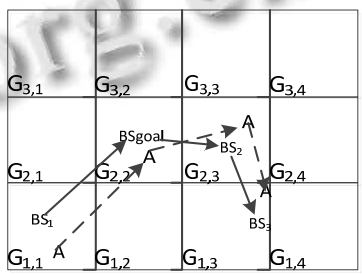

图 3 用户签到轨迹对应的基站的网格转换图示

2 基于网格概率的离群点识别

2.1 算法描述

首先按照数据集 D(所有的基站)的维度将数据空间划分为多个相邻的网格单元, 遍历数据集将其映射到所属的网格单元中, 从而将无序的数据集合转换成一种有序的数据结构. 然后根据海量日志文件中的用户轨迹信息关联出空间数据点之间的联系, 即目标基

站点和临近基站点的关系，进而转换为基站所在网格之间的关系．根据隶属概率的概念，判断该目标基站是否为离群点基站．

算法流程如图 4 所示．

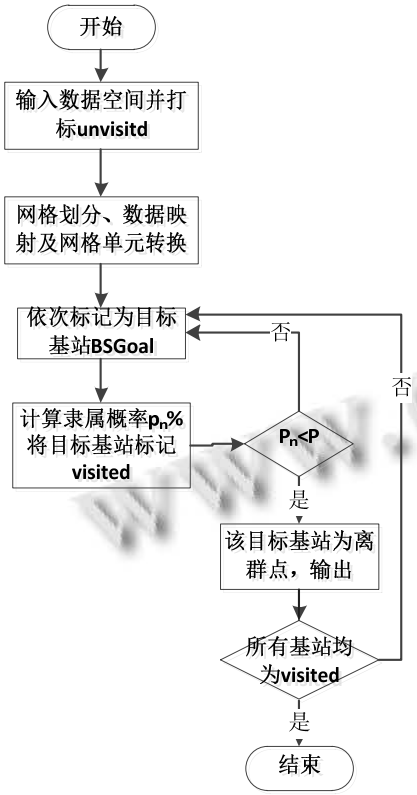

图 4　算法流程图

2.2 算法说明

(1) 将待核实的所有基站最初标记为 unvisited，在遍历数据集中的基站点时，将其依次打标为 BSGoal，通过参照临近基站位置计算出隶属概率之后将目标基站标记为 visited．算法结束的条件为所有待核实的基站标号均为 visited．

(2) 算法在每次确定目标基站之后，必须根据时间戳计算出该目标基站的临近基站，即重新确定临近网格集合的范围，这是由用户的签到轨迹信息决定的．

3 实验与性能分析

3.1 实验数据源

实验数据采用两个数据集，做对比分析的原始测试基站数据来自采集到的国内某收费网站，数据量为 7334825；用户轨迹数据为某 LBSN 网站华中某省 2015 年 2 月 5 日到 2015 年 2 月 11 共 7 天的签到数据，数据总量为 128163870．本文只对用户的群体特征进行分析，不对用户的敏感信息进行挖掘，对基站信息只用于核实正误，不用于其他不安全的行为．

实验开始前对两个数据集进行预处理[11]，如针对基站编码不符合编码规范中国移动的基站不以“46000”开头，经纬度倒置，经纬度明显错误，用户位置信息重复等问题进行过滤．经过预处理之后的基站基础表中的数据为 7152425 条，用户数据为 128152794 条，明显减少了数据的运算量．

3.2 基站离群点检测

根据网格划分的规则，每个基站都会被映射到唯一的一个网格单元中，基于网格概率算法的网格划分可以复用网格，为了设置算法中的参数，即隶属概率 p，对每个网格的基站数量做统计分析，抽样华中某省的原始测量基站总数为 611440，其对应的地理位置划分网格为 54607，平均每个网格中有 11.2 个基站．图 5 为网格中所含基站的统计情况

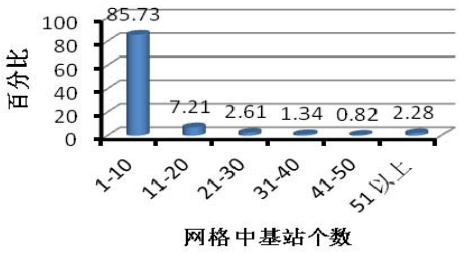

图 5　网格包含基站个数的百分比示意图

由于不同样本的基站分布情况不同，故隶属概率取值也会有所不同．在前面先抽取的样本中，有 85.73%的网格中所含的基站在 10 个以下，结合网格的面积约为 1.21 平方公里，显而易见，基站与基站之间的距离不会很远，这里分别取 p=60%，p=70%，p=80%，p=90%对样本数据进行离群点检测，抽样样本总数据量为 611440．检测结果如表 1 所示．

表 1　基于网格概率的离群点检测测试结果

隶属概率 p(%)	离群点数据量	离群点检测率(%)
60	11617	1.9
70	14674	2.4
80	20116	3.29
90	20422	3.34

从表 1 可以明显的观察到，隶属概率的值越小，满足离群条件的离群点数量越少．由于隶属概率与临近基站点(上文提到)的数量成正比，隶属概率越大，说明满足

离群点的条件的精度要求高，导致离群点数越多，但是结果不准确. 所以必须选择一个适中的隶属概率，结合表中的结果可知，隶属概率选择 p=80%比较适合.

3.3 实验结果验证

采用可视化工具 Tableau 来验证上面离群点检测结果. Tableau[12]是一款强大的可视化工具，它能够创建与共享数据可视化内容，快速处理数据，并且能够对多种数据源提供接口，图标展示美观，是一种将数据运算与美观的图标相结合的工具，应用非常广泛. 利用 Tableau 进行地理数据可视化时，只需要导入经纬度数据信息后，选择对应的数据列标注为地理角色，然后选择地图图层信息，就可以将输入坐标信息准确展现. 但是当数据集过大时，视觉上是看不出具体的离群点的位置，故本文的实验是在找出部分离群点之后，对比分析这些离群点在样本中所标识的位置与 Tableau 上所映射的位置是否在同一块区域，比如一个省，或者一个市. 很容易观察出离群点是否跨市或者跨省.

下面将实验检测出的离群点情况分别映射到某省的地图上(实验数据来自网络获取的基站信息). 如图 6，图 7，图 8，图 9 所示，分别对应在不同隶属概率取值的情况下，离群点在 Tableau 上的映射情况.

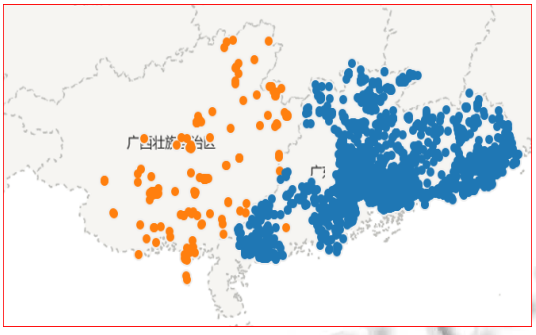

图 6 隶书概率 p 取 60%

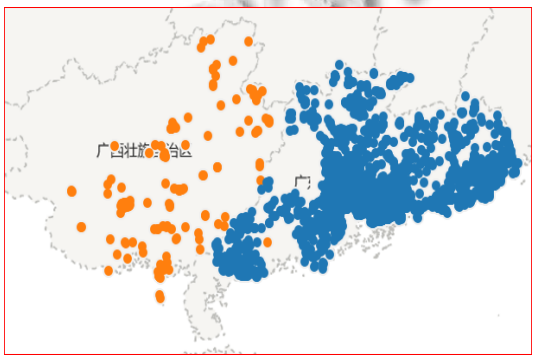

图 7 隶书概率 p 取 70%

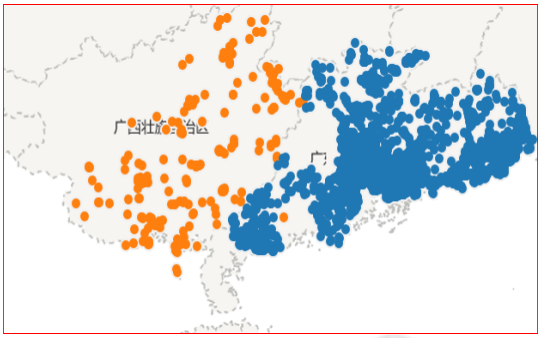

图 8 隶书概率 p 取 80%

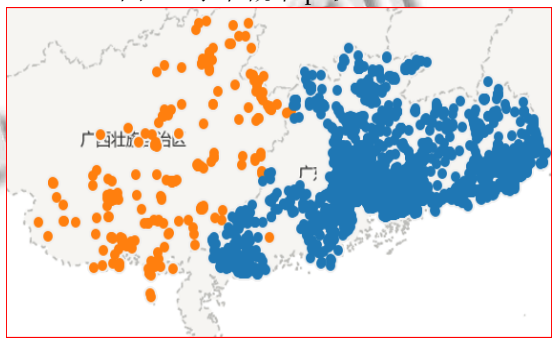

图 9 隶书概率 p 取 90%

图中蓝色的点表示基站所取的正确省份，比较密集，相对而言，红色的点表示偏离该省的基站数，可以明显的观察到错误基站的数量. 将图 6、图 7，图 8，图 9 中的离群点占比以曲线的形式直观展现如下图 10.

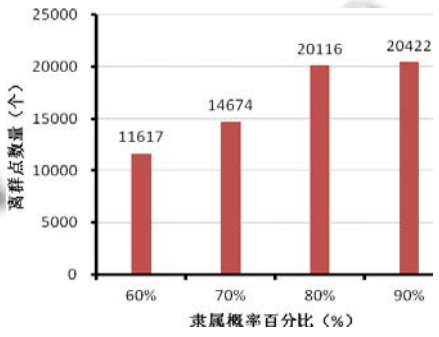

图 10 隶属概率与离群点数量

3.4 实验结论

通过多次实验论证，实验结果表明基于网格的离群点检测算法的准确度比较高，而且实验结论建议隶属概率取值 80%，但是由于不同地区的基站部署情况不一致，导致检测结果会有少许偏差.

3.5 实验应用场景

对中国 34 各省级自治区按 20%抽样，考虑到数据库的样本数量，选取华东六省一市共 7 个电信基站样

本集作离群点检测分析，隶属概率按照 80%来计算，结果如表 2.

表 2 多地区的离群点检测实验结果

样本	样本数据总量	离群点数据量	离群点检测率(%)
样本 1	41348	1240	3.0
样本 2	20399	530	2.6
样本 3	5970	191	3.2
样本 4	13560	420	3.1
样本 5	18950	814	4.3
样本 6	2900	110	3.8
样本 7	611440	20137	3.29

从表 2 可以明显观察到，各个样本检测出的离群点均在 3%左右. 对于全国百万数的基站而言，能检测出 3%的错误基站也是非常有价值的，可以减少人工成本.

4 总结

本文为了找出错误基站的位置信息，提出了一种新的离群点检测方式，通过划分网格和计算隶属概率的方法判定离群点，并且通过取各个地区的基站的做了实验，得出了比较好的结果. 可以将此方法应用到实际的找出错误基站的实践中，很大程度上减少人力劳动.

如果需要得到更精确的结果，可以综合考虑繁华度和单位网格内的基站数量，可以得到更好的精度，同时可以将其扩展出错误基站经纬度数据的自动发现，这方面的工作留作下一步研究的思路供研究者参考.

参考文献

1 吴俊伟,朱云龙,库涛等.基于网格聚类的热点路径探测.吉林大学学报,2015,45(1).

2 韩红霞.基于距离离群点的分析与研究[学位论文].镇江:江苏大学,2007.

3 Knorr EM, Ng RT. Algorithms for mining distance-based outliers in large datasets. Proc. of 24th International Conference on Very Large Data Bases. New York: Morgan Kaufmann. 1998. 392–403.

4 Knorr EM, Ng RT. Finding intensional knowledge of distance-based outliers. Proc. of 25th International Conference on Very Large Data Bases. New York. Morgan Kaufmann. 1999. 211–222.

5 S Ramasmawy RR, Shim K. Efficient algorithms for mining outliers from large dataset. Proc. of 2000 ACM SIGMOD International Conference on Management of Data. Santa Barbara, CA. ACM Press, 2000, (6): 427–438.

6 Angiulli F, Pizzuti C. Fast outlier detection in high dimensional spaces. Proc. of the 6th European Conference on the Principles of Data Mining and Knowledge Discovery in Database. Helsinki, Finland. 2002.19–23.

7 程求江.基于 NGID-DBSCAN 算法与最小包围圆模型的基站位置分析[硕士学位论文].武汉:武汉邮电科学研究院,2015.

8 于浩,王斌,肖刚等.基于距离的不确定离群点检测.计算机研究与发展,2010,47(3):474–484.

9 应倩岚.基于蜂窝网实测数据的基站位置与业务空间分布研究[硕士学位论文].杭州:浙江大学. 2015.

10 王亮,胡坤元,库涛等.位置不确定移动时空轨迹频繁模式挖掘.小型微型计算机系统,2014,35(12):2659–2663.

11 杨小漫.基于位置的服务中数据预处理研究[硕士学位论文].郑州:郑州大学,2013.

12 Nandeshwar A. Tableau 数据可视化实战.北京:机械工业出版社,2014.