



(12)发明专利申请

(10)申请公布号 CN 110111140 A

(43)申请公布日 2019.08.09

(21)申请号 201910339482.6

(22)申请日 2019.04.25

(71)申请人 华南理工大学

地址 510640 广东省广州市天河区五山路
381号

(72)发明人 黄紫林 许伦辉

(74)专利代理机构 广州粤高专利商标代理有限
公司 44102

代理人 何淑珍 黄海波

(51)Int.Cl.

G06Q 30/02(2012.01)

G06Q 50/26(2012.01)

G06F 16/215(2019.01)

G06F 16/2455(2019.01)

G06F 16/2458(2019.01)

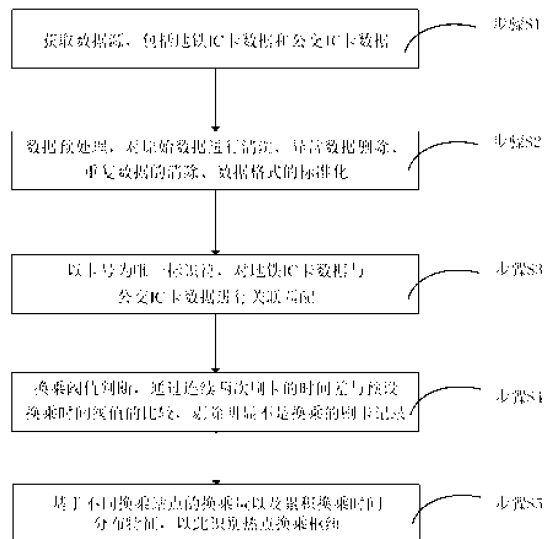
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种基于IC卡数据的热点换乘枢纽识别方法

(57)摘要

本发明公开了一种基于IC卡数据的热点换乘枢纽识别方法,包括:步骤S1:获取数据源,包括地铁IC卡数据和公交IC卡数据;步骤S2:对原始数据进行清洗,异常数据删除、重复数据的清除、数据格式标准化;步骤S3:通过“卡号”对地铁IC卡数据与公交IC卡数据进行关联匹配;步骤S4:比较连续两次刷卡的时间差与预设换乘时间阈值,剔除不是换乘的刷卡记录;步骤S5:基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的空间分布特性来识别热点换乘枢纽。本发明基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的空间分布特性以识别热点换乘枢纽,充分体现出不同换乘枢纽之间的换乘能力差异。



1. 一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,包括:

步骤S1:获取数据源,所述数据源包括地铁IC卡数据和公交IC卡数据;

步骤S2:数据预处理,对原始数据进行清洗,异常数据删除、重复数据的清除、数据格式的标准化;

步骤S3:以“卡号”为唯一标识符,对地铁IC卡数据与公交IC卡数据进行关联匹配;

步骤S4:换乘阈值判断,通过连续两次刷卡的时间差与预设换乘时间阈值的比较,剔除明显不是换乘的刷卡记录;

步骤S5:热点换乘枢纽识别,基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的空间分布特性,以此识别热点换乘枢纽。

2. 根据权利要求1所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,步骤S1中,所述地铁IC卡数据和公交IC卡数据以卡号作为乘客的唯一标识,匹配乘客的地铁刷卡行为和公交刷卡行为,每一张IC卡对应一位持卡乘客。

3. 根据权利要求1所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,步骤S1中,所述公交IC卡数据包括以下6个字段:卡编号、交易类型、交易时间、公司名称、线路号和车牌号。

4. 根据权利要求1所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,步骤S1中,所述地铁IC卡数据包括以下5个字段:卡编号、交易类型、交易时间、地铁线路和站点名称。

5. 根据权利要求1所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,步骤S2具体为:

数据字段选取,主要选取字段:IC卡号、交易时间、公司名称、线路号、车牌号;

错误、重复数据的清除,主要包括删除折前金额、折后金额为“0”,以及下车站点跟车牌号为“无值”的记录。

6. 根据权利要求1所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,所述步骤S3包括:

选取某一天的IC卡数据,并以卡号Card为唯一标识,匹配地铁IC数据库和公交IC数据库的IC卡数据,将匹配成功的IC刷卡记录记为相应数据库。

7. 根据权利要求6所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,

当IC卡数据中含车载机编号时,通过IC卡数据与GPS数据关联匹配,得到公交线路GPS数据,识别规则为车载机号的后5个字符与公交车辆车牌号匹配。

8. 根据权利要求7所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,若存在部分公交匹配不上GPS信息时,则删除该部分数据;部分公交车载号是字符“A”结尾,识别规则为:先删除字符“A”,再将删除字符“A”后的车载机号的后5个字符与公交车辆车牌号匹配。

9. 根据权利要求1所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,步骤S4中,所述连续两次刷卡的时间差为:

$$t_{M_i, B_j} = t_{B_j} - t_{M_i}$$

t_{M_i, B_j} 为地铁与公交的换乘时间, t_{B_j} 为公交上车刷卡时刻, t_{M_i} 为地铁出站刷卡时刻;

所述换乘的刷卡记录具指地铁出站刷卡时刻 t_{M_i} 与公交上车刷卡时刻 t_{B_j} 的时间间隔应满足预设换乘时间阈值的刷卡记录,即:

$$t_{M_i, B_j} \leq T_{M_i, B_j},$$

其中, t_{M_i, B_j} 表示两次刷卡时间间隔; T_{M_i, B_j} 表示预设换乘时间阈值。

10. 根据权利要求1所述的一种基于IC卡数据的热点换乘枢纽识别方法,其特征在于,步骤S5中,基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的空间分布特性,以此识别热点换乘枢纽时,主要指标包括:

概率密度函数可表示为:

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2 / 2\sigma^2}$$

式中, x 为不同影响因素下换乘时间从小到大排列的有序集合; μ 和 σ 分别为变量对数的平均值与标准差;

期望可表示为:

$$E(x) = e^{\mu + \sigma^2 / 2}$$

方差可表示为:

$$Var(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}。$$

一种基于IC卡数据的热点换乘枢纽识别方法

技术领域

[0001] 本发明涉及时空数据挖掘及交通技术领域,尤其涉及一种基于IC卡数据的热点换乘枢纽识别方法。

背景技术

[0002] 近年来,城市公共交通系统逐渐由单一的常规公交系统向多模式公共交通系统转变。作为衔接城市公共交通系统中不同模式交通系统和不同层次交通网络的关键节点,换乘枢纽在现代城市公共交通系统中的地位愈发重要。利用IC卡数据挖掘城市居民交通出行特征规律,分析热点换乘枢纽,并以此有的放矢指导城市交通管理与规划成为热门趋势。

[0003] IC卡系统近年来在我国得到广泛应用,很多大中城市持卡乘客达到60%~80%。但IC卡刷卡系统每天所产生的海量数据,在完成其费用结算之后,往往就成为沉淀信息或被删除,造成城市交通信息资源的极大浪费。其实,IC卡数据隐含着城市居民出行客流量信息及时空分布信息等。通过公共交通IC信息采集分析,可以得到城市过去任何一天的时空分布特征,这对城市公共交通系统的规划设计有着重要意义。

[0004] 目前,对热点换乘枢纽的识别方法常用进站刷卡量、出站刷卡量的数理统计方法,工程量巨大、具有区域局限性且代价昂贵,也无法准确体现出不同换乘枢纽之间的换乘能力差异。利用IC卡中隐含的大量信息,对地铁与常规公交间的换乘行为进行识别,进一步基于换乘数据区分热点换乘枢纽是一种行之有效的思路。

[0005] 因此,针对上述技术问题,有必要提供一种基于IC卡数据的热点换乘枢纽识别方法,为城市公共交通热点换乘枢纽区分提供科学依据。

发明内容

[0006] 有鉴于此,本发明的目的在于提供一种基于IC卡数据的热点换乘枢纽识别方法。

[0007] 为了达到上述目的,本发明实施例提供的技术方案如下:

[0008] 一种基于IC卡数据的热点换乘枢纽识别方法,包括:

[0009] 步骤S1:获取数据源,所述数据源包括地铁IC卡数据和公交IC卡数据;

[0010] 步骤S2:数据预处理,对原始数据进行清洗,异常数据删除、重复数据的清除、数据格式的标准化;

[0011] 步骤S3:以“卡号”为唯一标识符,对地铁IC卡数据与公交IC卡数据进行关联匹配;

[0012] 步骤S4:换乘阈值判断,通过连续两次刷卡的时间差与预设换乘时间阈值的比较,剔除明显不是换乘的刷卡记录,此处的阈值可以根据居民出行调查或长期的数据分析基础设置;

[0013] 步骤S5:热点换乘枢纽识别,基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的时空分布特性,以此识别热点换乘枢纽。

[0014] 作为本发明的进一步改进,步骤S1中,所述地铁IC卡数据和公交IC卡数据以卡号作为乘客的唯一标识,匹配乘客的地铁刷卡行为和公交刷卡行为,每一张IC卡对应一位持

卡乘客。

[0015] 作为本发明的进一步改进,步骤S1中,所述公交IC卡数据包括以下6个字段:卡编号、交易类型、交易时间、公司名称、线路号和车牌号。

[0016] 作为本发明的进一步改进,步骤S1中,所述地铁IC卡数据包括以下5个字段:卡编号、交易类型、交易时间、地铁线路和站点名称。

[0017] 作为本发明的进一步改进,步骤S2具体包括:数据字段选取,主要选取字段:IC卡号、交易时间、公司名称、线路号、车牌号;

[0018] 错误、重复数据的清除,主要包括删除折前金额、折后金额为“0”,以及下车站点跟车牌号为“无值”的记录。

[0019] 作为本发明的进一步改进,所述步骤S3具体包括:选取某一天的IC卡数据,并以卡号Card为唯一标识,匹配地铁IC数据库和公交IC数据库的IC卡数据,将匹配成功的IC刷卡记录记为相应数据库。

[0020] 作为本发明的进一步改进,当IC卡数据中含车载机编号时,通过IC卡数据与GPS数据关联匹配,得到公交线路GPS数据,识别规则为车载机号的后5个字符与公交车车辆车牌号匹配。

[0021] 作为本发明的进一步改进,若存在部分公交匹配不上GPS信息时,则删除该部分数据;部分公交车载号是字符“A”结尾,识别规则为:先删除字符“A”,再将删除字符“A”后的车载机号的后5个字符与公交车车辆车牌号匹配。

[0022] 作为本发明的进一步改进,步骤S4中,所述连续两次刷卡的时间差为:

$$[0023] \quad t_{M_i, B_j} = t_{B_j} - t_{M_i}$$

[0024] t_{M_i, B_j} 为地铁与公交的换乘时间, t_{B_j} 为公交上车刷卡时刻, t_{M_i} 为地铁出站刷卡时刻;

[0025] 所述换乘的刷卡记录具指地铁出站刷卡时刻 t_{M_i} 与公交上车刷卡时刻 t_{B_j} 的时间间隔应满足预设换乘时间阈值的刷卡记录,即:

$$[0026] \quad t_{M_i, B_j} \leq T_{M_i, B_j}$$

[0027] 其中, t_{M_i, B_j} 表示两次刷卡时间间隔; T_{M_i, B_j} 表示预设换乘时间阈值。

[0028] 作为本发明的进一步改进,步骤S5中,基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的空间分布特性,以此识别热点换乘枢纽时,主要指标包括:

[0029] 其概率密度函数可表示为:

$$[0030] \quad f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

[0031] 式中, x 为不同影响因素下换乘时间从小到大排列的有序集合; μ 和 σ 分别为变量对数的平均值与标准差;

[0032] 期望可表示为:

$$[0033] \quad E(x) = e^{\mu + \sigma^2/2}$$

[0034] 方差可表示为:

$$[0035] \text{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu \cdot \sigma^2}。$$

[0036] 相比现有技术,本发明具有以下有益效果:

[0037] 本发明通过关联匹配公交IC卡数据与地铁IC卡数据,给出乘客在地铁与公交两种交通方式之间换乘所应具备的条件,基于换乘数据提出一种针对“地铁-公交”系统新的热点换乘枢纽识别方法。本发明基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的空间分布特性,以此识别热点换乘枢纽,该方法能充分体现出不同换乘枢纽之间的换乘能力差异。

附图说明

[0038] 图1是本发明一种基于IC卡数据的热点换乘枢纽识别方法的流程图;

[0039] 图2是基于换乘数据的热点换乘枢纽站点识别图;

[0040] 图3(a)是基于换乘数据的热点换乘枢纽区域识别图(立体)。

[0041] 图3(b)是基于换乘数据的热点换乘枢纽区域识别图(平面)。

具体实施方式

[0042] 为更好地理解本发明,下面依据附图结合实施例进一步阐明本发明的内容,应理解这些实施例仅用于说明本发明而不适用于限制本发明的范围,在阅读了本发明之后,本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

[0043] 如图1所示,一种基于IC卡数据的热点换乘枢纽识别方法,包括:

[0044] 步骤S1:获取数据源,所述数据源包括地铁IC卡数据和公交IC卡数据;

[0045] 步骤S2:数据预处理,对原始数据进行清洗,异常数据删除、重复数据的清除、数据格式的标准化,从而改善数据质量、提高分析速度;

[0046] 步骤S3:以卡号为唯一标识符,对地铁IC卡数据与公交IC卡数据进行关联匹配;

[0047] 步骤S4:换乘阈值判断,通过连续两次刷卡的时间差与预设换乘时间阈值的比较,剔除明显不是换乘的刷卡记录,此处的预设换乘时间阈值根据居民出行调查或长期的数据分析基础设置;

[0048] 步骤S5:热点换乘枢纽识别,基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的空间分布特性,以此识别热点换乘枢纽。

[0049] 在本发明的一具体实施方式中,一种基于IC卡数据的热点换乘枢纽识别方法,包括:

[0050] 步骤S1:获取数据源,所述数据源包括公交IC卡数据和地铁IC卡数据。两种数据都以卡号作为乘客的唯一标识,匹配乘客的地铁刷卡行为和公交刷卡行为,每一张IC卡对应一位持卡乘客。

[0051] 本发明实施例选取深圳市地铁3号线(龙岗线)作为研究对象,它连接了福田、罗湖和龙岗三个行政区,由双龙站至益田站,总长41.7公里,共有30个地铁站点。

[0052] 乘客地铁IC卡数据表主要包含五个字段,即:CardID、StationID、TrnsctTime、TrnsctyType、Statiom,其中CardID是智能交通卡的唯一标识;StationID是地铁站的唯一标识;TrnsctTime是交易(刷卡)时间,TrnsctTime有两种类型,进站和出站,分别用21,22标

识,Station是交易站点,详情如表1所示。

[0053] 表1:

[0054]

卡编号	交易类型	交易时间	地铁线路	站点	车载机号
361955655	22	2016/5/23 0:01	地铁三号线	爱联	AGM-105
667356942	22	2016/5/23 0:01	地铁三号线	双龙	AGM-122
684144991	22	2016/5/23 0:01	地铁三号线	爱联	AGM-119

[0055] 公交IC卡数据表记录着每一位乘客乘坐公交的出行信息,其中卡编号作为是被公交卡的唯一识别码,每一张IC卡对应一位持卡乘客,CardID代表卡编号,TradeTime代表交易时间,LineID代表线路号,BusID代表车牌号,详情如表2所示。

[0056] 表2:

卡编号	交易类型	交易时间	公司名称	线路号	车牌号
326091092	31	2016/5/23 1:05	巴士集团	216路	DN2471
20733829	31	2016/5/23 0:00	巴士集团	63路	DCM425
80464253	31	2016/5/23 0:09	巴士集团	10路	DCJ411
.....
296270453	31	2016/5/23 0:22	巴士集团	10路	DCJ411

[0058] 步骤S2:数据预处理,对原始数据进行清洗,异常数据删除、重复数据的清除、数据格式的标准化。

[0059] 数据预处理主要包括两步:

[0060] (1) 数据字段选取,公交IC卡原始数据中包含了大量有关乘客消费信息的数据字段,而在用于换乘行为识别和OD推导过程中,乘客的消费数据对结果没有影响。为此,在数据预处理阶段,可以剔除这些对后期研究没有影响的字段。在进行IC卡字段选取时,主要选取字段:IC卡号、交易时间、公司名称、线路号、车牌号。

[0061] (2) 错误、重复数据的清除。IC卡数据中,有一小部分记录由于收到数据移动环境的影响,造成数部分据重复或数据字段为空的记录,需要对这部分记录进行清理和修正,进而保证数据的质量。主要是:删除折前金额、折后金额为“0”,下车站点跟车牌号为“无值”的记录。

[0062] 步骤S3:以“卡号”为唯一标识符,对地铁IC卡数据与公交IC卡数据进行关联匹配。

[0063] 当IC卡记录中含车载机编号时,可以通过IC卡数据与GPS数据关联匹配,得到公交线路GPS数据,识别规则为车载机号后5位与公交车辆车牌号匹配,巴士集团也存在部分公交匹配不上GPS信息,应删除该部分数据;部分公交车载号是字符“A”结尾,识别规则为:先

删除字符“A”，再将删除字符“A”后的车载机号的后5个字符与公交车牌号匹配。

[0064] 步骤S4:换乘阈值判断,通过连续两次刷卡的时间差与预设换乘时间阈值的比较,剔除明显不是换乘的刷卡记录。

[0065] 根据居民的出行特点,当明确目的地后,在选择出行路径时,乘客一般选择直达、换乘步行时间最少的站点进行换乘。因此地铁-公交的刷卡时间差不应太大,地铁出站刷卡时刻 t_{M_i} 与公交上车刷卡时刻 t_{B_j} 的时间间隔应满足设定的时间阈值,即 $t_{M_i, B_j} \leq T_{M_i, B_j}$ 。其中, t_{M_i, B_j} 表示两次刷卡时间间隔; T_{M_i, B_j} 表示设定的时间阈值。换乘的时间必要条件如下式所示:

$$[0066] \quad t_{M_i, B_j} = t_{B_j} - t_{M_i}$$

[0067] 式中, t_{M_i, B_j} 为地铁与公交的换乘时间; t_{B_j} 为公交上车刷卡时刻, t_{M_i} 为地铁出站刷卡时刻。

[0068] 因此,本发明实施例以30min作为换乘时间阈值,进行地铁换乘公交行为识别。

[0069] 步骤S5:热点换乘枢纽识别,基于不同换乘站点的换乘量以及累积换乘时间分布特征,统计乘客换乘行为的时空分布特性,以此识别热点换乘枢纽,主要指标包括:

[0070] 其概率密度函数可表示为:

$$[0071] \quad f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\ln x - \mu^2 - 2\sigma^2}$$

[0072] 式中, x 为不同影响因素下换乘时间从小到大排列的有序集合; μ 和 σ 分别为变量对数的平均值与标准差。

[0073] 期望可表示为:

$$[0074] \quad E(x) = e^{\mu + \sigma^2/2}$$

[0075] 方差可表示为:

$$[0076] \quad Var(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

[0077] 2016年5月23日深圳地铁运营时间(6:00-23:00)内,深圳市地铁三号线热点换乘枢纽站点识别结果如图2所示。可以发现:发生换乘行为极少的站点为3-购物公园、8-通新岭、9-红岭、14-水贝。经过调查分析可知,原因有几点:地理位置过于偏僻;周围无常规公交站点设置;前后站点为较大的公交换乘枢纽。

[0078] 发生换乘行为较多的站点为1-益田、17-木棉湾、18-大芬、25-大运、30-双龙。经过调查分析可知,5个站点均为大型换乘枢纽,其中益田站和双龙站为首末站。益田站周围设置有2个公交站台,可换乘公交线路有19条,其中2条为首发站;双龙地铁站周围设置有4个公交站台,可换乘公交线路有53条,其中3条为首发站;木棉湾地铁站周围设置有3个公交站台,可换乘公交线路有58条;大芬地铁站周围设置有2个公交站台,可换乘公交线路有40条;大运地铁站周围设置有1个公交站台,可换乘公交线路有29条,其中1条为首发站。

[0079] 深圳市地铁三号线热点换乘枢纽区域识别结果如图3(a)-3(b)所示。从图中看出,右上角密度颜色较深,说明该区域地铁换乘公交人数较多。经查地铁站点GPS数据表,发现该区域为龙岗区(布吉-双龙站)。这与龙岗区地铁系统只有三号线有关(无其他地铁线路),乘客不得不搭乘地铁三号线至最近公交换乘点,随后搭乘公交到达目的地。

[0080] 此外,应当理解,虽然本说明书按照实施方式加以描述,但并非每个实施方式仅包含一个独立的技术方案,说明书的这种叙述方式仅仅是为清楚起见,本领域技术人员应当将说明书作为一个整体,各实施例中的技术方案也可以经适当组合,形成本领域技术人员可以理解的其他实施方式。

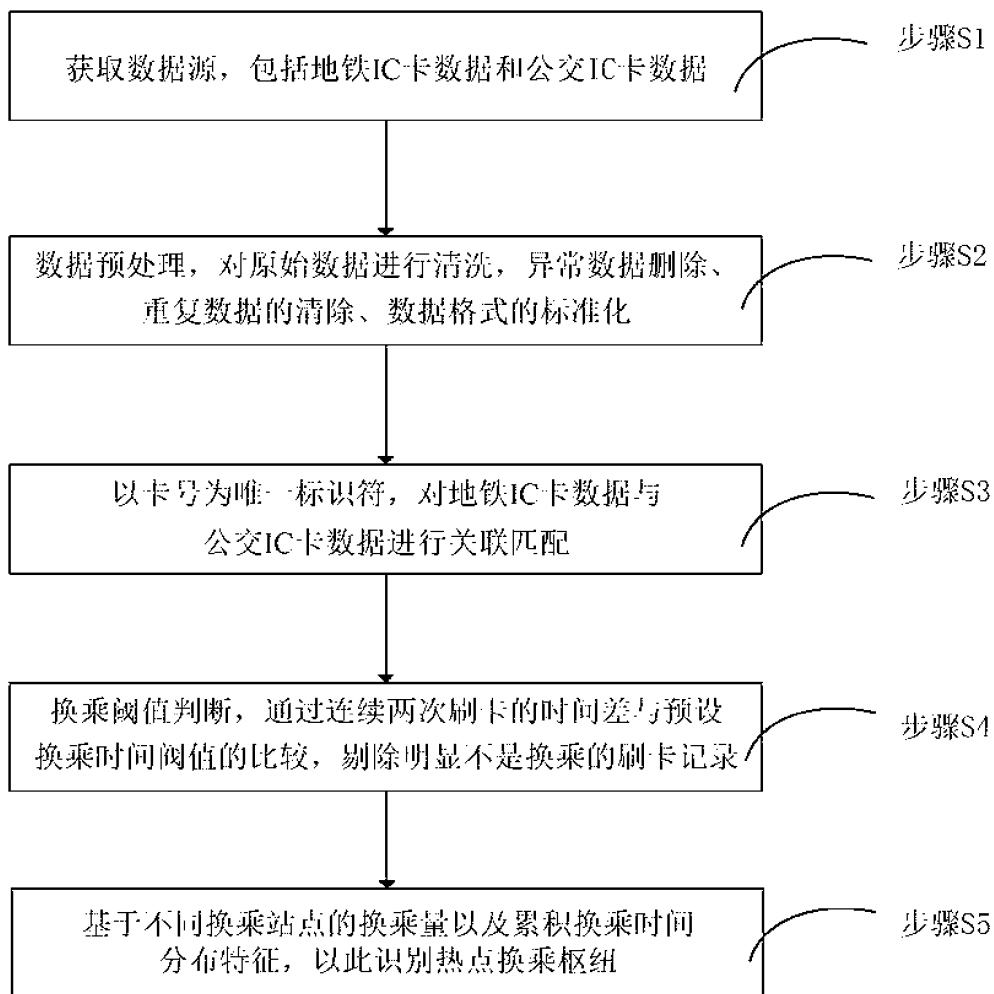


图1

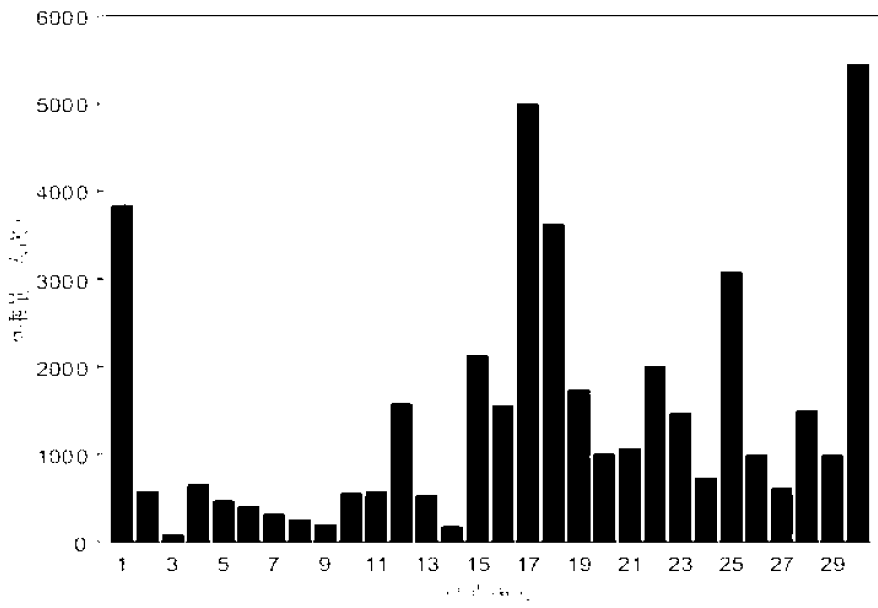


图2

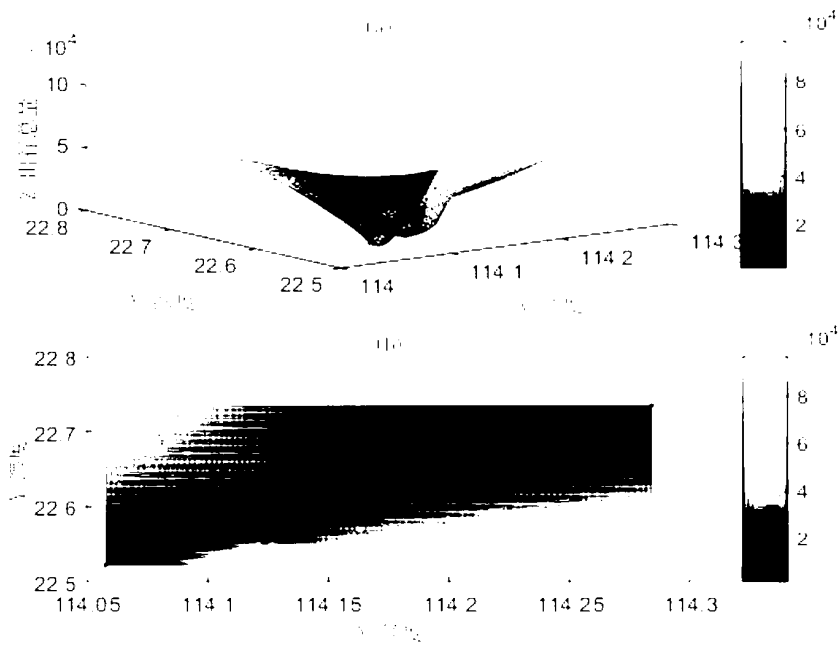


图3