



(12)发明专利申请

(10)申请公布号 CN 109840555 A

(43)申请公布日 2019.06.04

(21)申请号 201910051785.8

(22)申请日 2019.01.21

(71)申请人 华南理工大学

地址 511458 广东省广州市南沙区环市大道南路25号华工大广州产研院

(72)发明人 黄紫林 许伦辉

(74)专利代理机构 广州市华学知识产权代理有限公司 44245

代理人 李君

(51)Int.Cl.

G06K 9/62(2006.01)

G06Q 10/10(2012.01)

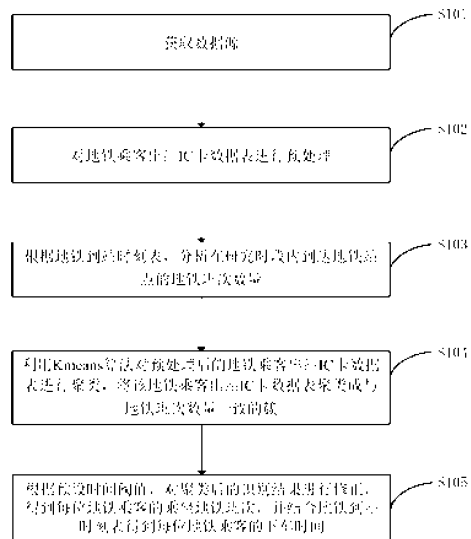
权利要求书2页 说明书8页 附图4页

(54)发明名称

地铁乘客下车时间识别方法、系统、计算机设备及存储介质

(57)摘要

本发明公开了一种地铁乘客下车时间识别方法、系统、计算机设备及存储介质,所述方法包括:获取数据源,包括地铁到站时刻表和地铁乘客出站IC卡数据表;对地铁乘客出站IC卡数据表进行预处理;根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量;利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将其聚类成与地铁班次数量一致的簇;根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。本发明通过对地铁到站时刻表和地铁乘客出站IC卡数据表进行数据挖掘,运用修正后的K-means聚类方法进行地铁乘客下车时间识别,识别精确度较高。



1. 一种地铁乘客下车时间识别方法,其特征在于,所述方法包括:

获取数据源;其中,所述数据源包括地铁到站时刻表和地铁乘客出站IC卡数据表;

对地铁乘客出站IC卡数据表进行预处理;

根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量;

利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇;

根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

2. 根据权利要求1所述的地铁乘客下车时间识别方法,其特征在于,所述对地铁乘客出站IC卡数据表进行预处理,具体包括:

从地铁乘客出站IC卡数据表中选择一个地铁站点作为研究对象,筛选该地铁站点的地铁乘客出站IC卡数据;

选择该地铁站点在某个研究时段内的地铁乘客出站IC卡数据;

对该地铁站点在研究时段内的地铁乘客出站IC卡数据中时间重复、有误的数据进行过滤,确保每张IC卡对应研究时段内的一个出站刷卡时间。

3. 根据权利要求1所述的地铁乘客下车时间识别方法,其特征在于,所述根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量,具体包括:

根据地铁到站时刻表中研究时段的起点和终点,统计该研究时段内到达地铁站点的地铁班次数量,记录该研究时段内的每趟班次及其到达时间。

4. 根据权利要求1所述的地铁乘客下车时间识别方法,其特征在于,所述利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇,具体包括:

将研究时段内到达地铁站点的地铁班次数量作为K-means算法的聚类数,先随机选取K个聚类质心点,重复以下过程直到收敛,使该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇:

将该地铁乘客出站IC卡数据表中的每个地铁乘客出站IC卡数据作为一个样例,对于每个样例,计算该样例应该属于的类,如下式:

$$c^{(i)} := \operatorname{argmin}_{(j)} \|x^{(i)} - \mu_j\|^2$$

其中, $c^{(i)}$ 表示样例*i*与K个类中距离最近的那个类,值为1到K中的一个, μ_j 表示对属于同一个类的样本中心点的猜测, $x^{(i)}$ 表示样例*i*的数据;

对于每个类*j*,重新计算该类的质心,如下式:

$$\mu_j = \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}$$

其中, μ_j 表示重新计算后的质心。

5. 根据权利要求1所述的地铁乘客下车时间识别方法,其特征在于,所述根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间,具体包括:

将预设时间阈值作为聚类质心的邻域半径,将属于邻域内的地铁乘客归为同一地铁班次,得到每位乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

6. 根据权利要求1-5任一项所述的地铁乘客下车时间识别方法,其特征在于,所述地铁到站时刻表包括地铁线路名称、地铁站点、到达地铁站点的时间和地铁班次编号;所述地铁乘客出站IC卡数据表中的每个地铁乘客IC卡数据包括IC卡号、地铁出站时间和出站地铁站点,以IC卡号作为乘客的唯一标识,匹配地铁乘客的一次地铁出站时间和出站地铁站点。

7. 根据权利要求1-5任一项所述的地铁乘客下车时间识别方法,其特征在于,所述对地铁乘客出站IC卡数据表进行预处理之后,还包括:

对预处理后的地铁乘客出站IC卡数据表中的地铁乘客出站IC卡数据按照时间进行升序排序。

8. 一种地铁乘客下车时间识别系统,其特征在于,所述系统包括:

获取模块,获取数据源;其中,所述数据源包括地铁到站时刻表和地铁乘客出站IC卡数据表;

预处理模块,用于对地铁乘客出站IC卡数据表进行预处理;

分析模块,用于根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量;

聚类模块,用于利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇;

修正模块,用于根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

9. 一种计算机设备,包括处理器以及用于存储处理器可执行程序存储器,其特征在于,所述处理器执行存储器存储的程序时,实现权利要求1-7任一项所述的地铁乘客下车时间识别方法。

10. 一种存储介质,存储有程序,其特征在于,所述程序被处理器执行时,实现权利要求1-7任一项所述的地铁乘客下车时间识别方法。

地铁乘客下车时间识别方法、系统、计算机设备及存储介质

技术领域

[0001] 本发明涉及一种地铁乘客下车时间识别方法、系统、计算机设备及存储介质,属于时空数据挖掘及交通技术领域。

背景技术

[0002] 近年来,利用IC卡数据挖掘分析城市居民交通出行特征规律,并以此指导城市交通管理与规划成为热门趋势。针对地铁系统,无论分析地铁乘客的时空分布,还是模拟地铁的运营状况或者实时预测地铁密度,都会用到两个时间常量,那就是在假设乘客步行速度一定的前提下,离开地铁门后步行到达收费闸机的时间以及从地铁收费闸机到站台的步行时间。

[0003] 目前还没有针对地铁乘客地铁门后步行到达收费闸机的时间做计算的方法,但这个步行时间对于模拟地铁运营状况以及乘客的时空分布很重要的,例如可以通过识别乘客所乘坐的地铁班次,进一步计算出不同时间段地铁乘客出站密度和速度,这个是模拟地铁运营状况的基础。

[0004] 针对如何识别出乘客所乘坐的地铁班次,传统的方法主要是通过志愿者统计、人工调查问卷等方法,耗费大量的人力和物力。基于IC卡数据进行数据挖掘是新的方向,但IC卡数据中并不包含乘客下车时间。

发明内容

[0005] 有鉴于此,本发明提供了一种地铁乘客下车时间识别方法、系统、计算机设备及存储介质,通过对地铁到站时刻表和地铁乘客出站IC卡数据表进行数据挖掘,分析其时空特性关系,并运用修正后的K-means聚类方法进行地铁乘客下车时间识别,识别精确度较高。

[0006] 本发明的第一个目的在于提供一种地铁乘客下车时间识别方法。

[0007] 本发明的第二个目的在于提供一种地铁乘客下车时间识别系统。

[0008] 本发明的第三个目的在于提供一种计算机设备。

[0009] 本发明的第四个目的在于提供一种存储介质。

[0010] 本发明的第一个目的可以通过采取如下技术方案达到:

[0011] 一种地铁乘客下车时间识别方法,所述方法包括:

[0012] 获取数据源;其中,所述数据源包括地铁到站时刻表和地铁乘客出站IC卡数据表;

[0013] 对地铁乘客出站IC卡数据表进行预处理;

[0014] 根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量;

[0015] 利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇;

[0016] 根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

[0017] 进一步的,所述对地铁乘客出站IC卡数据表进行预处理,具体包括:

[0018] 从地铁乘客出站IC卡数据表中选择一个地铁站点作为研究对象,筛选该地铁站点的地铁乘客出站IC卡数据;

[0019] 选择该地铁站点在某个研究时段内的地铁乘客出站IC卡数据;

[0020] 对该地铁站点在研究时段内的地铁乘客出站IC卡数据中时间重复、有误的数据进行过滤,确保每张IC卡对应研究时段内的一个出站刷卡时间。

[0021] 进一步的,所述根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量,具体包括:

[0022] 根据地铁到站时刻表中研究时段的起点和终点,统计该研究时段内到达地铁站点的地铁班次数量,记录该研究时段内的每趟班次及其到达时间。

[0023] 进一步的,所述利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇,具体包括:

[0024] 将研究时段内到达地铁站点的地铁班次数量作为K-means算法的聚类数,先随机选取K个聚类质心点,重复以下过程直到收敛,使该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇:

[0025] 将该地铁乘客出站IC卡数据表中的每个地铁乘客出站IC卡数据作为一个样例,对于每个样例,计算该样例应该属于的类,如下式:

[0026] $c^{(i)} := \operatorname{argmin}_{(j)} ||x^{(i)} - \mu_j||^2$

[0027] 其中, $c^{(i)}$ 表示样例i与K个类中距离最近的那个类,值为1到K中的一个, μ_j 表示对属于同一个类的样本中心点的猜测, $x^{(i)}$ 表示样例i的数据;

[0028] 对于每个类j,重新计算该类的质心,如下式:

$$[0029] \quad \mu_j = \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}$$

[0030] 其中, μ_j 表示重新计算后的质心。

[0031] 进一步的,所述根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间,具体包括:

[0032] 将预设时间阈值作为聚类质心的邻域半径,将属于邻域内的地铁乘客归为同一地铁班次,得到每位乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

[0033] 进一步的,所述地铁到站时刻表包括地铁线路名称、地铁站点、到达地铁站点的时间和地铁班次编号;所述地铁乘客出站IC卡数据表中的每个地铁乘客IC卡数据包括IC卡号、地铁出站时间和出站地铁站点,以IC卡号作为乘客的唯一标识,匹配地铁乘客的一次地铁出站时间和出站地铁站点。

[0034] 进一步的,所述对地铁乘客出站IC卡数据表进行预处理之后,还包括:

[0035] 对预处理后的地铁乘客出站IC卡数据表中的地铁乘客出站IC卡数据按照时间进行升序排序。

[0036] 本发明的第二个目的可以通过采取如下技术方案达到:

[0037] 一种地铁乘客下车时间识别系统,所述系统包括:

[0038] 获取模块,获取数据源;其中,所述数据源包括地铁到站时刻表和地铁乘客出站IC卡数据表;

[0039] 预处理模块,用于地铁乘客出站IC卡数据表;

[0040] 分析模块,用于根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量;

[0041] 聚类模块,用于利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇;

[0042] 修正模块,用于根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

[0043] 本发明的第三个目的可以通过采取如下技术方案达到:

[0044] 一种计算机设备,包括处理器以及用于存储处理器可执行程序存储器,所述处理器执行存储器存储的程序时,实现上述的地铁乘客下车时间识别方法。

[0045] 本发明的第四个目的可以通过采取如下技术方案达到:

[0046] 一种存储介质,存储有程序,所述程序被处理器执行时,实现上述的地铁乘客下车时间识别方法。

[0047] 本发明相对于现有技术具有如下的有益效果:

[0048] 本发明通过对地铁到站时刻表和地铁乘客出站IC卡数据表中的数据进行预处理,利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,修正聚类结果,能够推算出地铁乘客所乘坐的地铁班次,进一步识别出地铁乘客下车时间,为计算地铁乘客离开地铁门后步行到达收费闸门的时间提供数据支撑,且识别精确度较高。

附图说明

[0049] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图示出的结构获得其他的附图。

[0050] 图1为本发明实施例1的地铁乘客下车时间识别方法的流程图。

[0051] 图2为本发明实施例1的地铁乘客出站IC卡数据表预处理的流程图。

[0052] 图3为本发明实施例1的修正后的聚类结果图。

[0053] 图4为本发明实施例2的地铁乘客下车时间识别系统的结构框图。

[0054] 图5为本发明实施例2的预处理模块的结构框图。

[0055] 图6为本发明实施例3的计算机设备的结构框图。

具体实施方式

[0056] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明的一部分实施例,而不是全部的实施例,基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0057] 实施例1:

[0058] 如图1所示,本实施例提供了一种地铁乘客下车时间识别方法,该方法包括以下步骤:

[0059] S101、获取数据源。

[0060] 本实施例的数据源包括地铁到站时刻表和地铁乘客出站IC卡数据表,该数据源可以从数据库查找获取,例如预先在数据库内存储地铁到站时刻表和地铁乘客出站IC卡数据表,从数据库中搜索地铁到站时刻表和地铁乘客出站IC卡数据表即可得到。

[0061] 其中,所述地铁到站时刻表包括地铁线路名称、地铁站点、到达地铁站点的时间和地铁班次编号;所述地铁乘客出站IC卡数据表中的每个地铁乘客IC卡数据包括IC卡号、地铁出站时间和出站地铁站点,以IC卡号作为乘客的唯一标识,匹配地铁乘客的一次地铁出站时间和出站地铁站点。

[0062] 本实施例以深圳市地铁三号线为例,地铁到站时刻表为深圳市地铁三号线地铁到站时刻表,地铁乘客出站IC卡数据表为2018年5月23日深圳市地铁三号线地铁乘客出站IC卡数据表。

[0063] 具体地,深圳市地铁三号线地铁到站时刻表包括四个字段,即:RouteID、StationIDlst、Time、ShiftID,其中RouteID为地铁线路名称,例如三号线,StationIDlst为地铁站点列表,Time表示到达地铁站点的时间,ShiftID表示地铁班次编号。

[0064] 具体地,深圳市地铁三号线地铁乘客出站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是地铁站点的唯一标识;TrnsctTime是交易(刷卡)时间,TrnsctTime有两种类型,进站时间和出站时间,分别用21,22标识,Statiom是交易(刷卡)站点,即出站地铁站点,详情如下表1所示。

[0065] 表1深圳市地铁三号线地铁乘客出站IC卡数据表

[0066]

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
660896555	22	2016/5/23 0:01	地铁三号线	爱联	AGM-113
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

[0067] S102、对地铁乘客出站IC卡数据表进行预处理。

[0068] 该步骤如图2所示,具体包括:

[0069] S1021、地铁站点M选择:从地铁乘客出站IC卡数据表中选择一个地铁站点M作为研究对象,筛选该地铁站点M的地铁乘客出站IC卡数据。本实施例以深圳市地铁三号线的益田站作为研究对象,筛选益田站的地铁乘客出站IC卡数据。

[0070] S1022、研究时段h确定:选择该地铁站点M在某个研究时段h内的地铁乘客出站IC卡数据。本实施例以早高峰时间段(7:00-9:00)为研究时段h,选择益田站在研究时段h内的地铁乘客出站IC卡数据。

[0071] S1023、数据清洗:对该地铁站点M在研究时段h内的地铁乘客出站IC卡数据中时间重复、有误的数据进行过滤,使每张IC卡对应研究时段h内的一个出站刷卡时间。本实施例

对益田站在研究时段h内的地铁乘客出站IC卡数据中时间重复、有误的数据进行过滤,使每张IC卡对应研究时段h内的一个出站刷卡时间。

[0072] 为了方便后续对地铁乘客出站IC卡数据进行研究,本实施例在对地铁乘客出站IC卡数据表进行预处理之后,还可包括:对预处理后的地铁乘客出站IC卡数据表中的地铁乘客出站IC卡数据按照时间进行升序排序。

[0073] S103、根据地铁到站时刻表,分析在研究时段h内到达地铁站点的地铁班次数量r。

[0074] 具体地,根据地铁到站时刻表中研究时段h的起点和终点,统计该研究时段h内到达地铁站点的地铁班次数量K,包括常规车和区间车,记录该研究时段h内的每趟班次及其到达时间,每趟班次记为k ($k=1, 2, \dots, K$),每趟班次的到达时间记为 t_k ($t_k=t_1, t_2, \dots, t_k, \dots, t_k$)。

[0075] 2018年5月23日为工作日,经过查询深圳市地铁三号线到站时刻表,当天经过益田站的地铁班次共有145趟,其中区间车8班(益田-塘坑),常规车137班。在26路公交早高峰(7:00-9:00)期间,地铁三号线益田站共有17辆地铁班次到达,总换乘人数为1298人次。本实施例中定义7:00.00 ($0s, i=1$)为研究时段h的起点,此时第1辆地铁到站;研究时段h的起点为8:59.50,即第17辆地铁到站,选择换乘26路公交的地铁乘客出站IC卡数据,出站时刻表详情如下表2所示。

[0076] 表2地铁乘客出站时刻表

序号	到站时间 T_{Ri}	间隔	序号	到站时间 T_{Ri}	间隔
1	7:03:49	-	10	8:10:08	426
2	7:07:11	202	11	8:17:14	426
3	7:15:11	480	12	8:24:20	426
4	7:23:11	480	13	8:31:26	426
5	7:31:11	480	14	8:38:32	426
6	7:39:09	478	15	8:45:38	426
7	7:47:09	480	16	8:52:44	426
8	7:55:09	480	17	8:59:50	426
9	8:03:02	473	-	-	-

[0078] S104、利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇(即r个簇)。

[0079] K-means (K均值) 算法是经典的基于划分的聚类方法,是十大经典数据挖掘算法之一,其基本思想是:以空间中k个点为形心进行聚类,对最靠近他们的对象归类。通过迭代的方法,逐次更新各簇的形心的值,直至得到最好的聚类结果。

[0080] 具体地,将研究时段h内到达地铁站点的地铁班次数量作为K-means算法的聚类数,先随机选取K个聚类质心点,重复以下过程直到收敛,使该地铁乘客出站IC卡数据表聚类成r个簇:

[0081] 将该地铁乘客出站IC卡数据表中的每个地铁乘客出站IC卡数据作为一个样例,对于每个样例,计算该样例应该属于的类,如下式:

$$c^{(i)} := \operatorname{argmin}_{(j)} \|x^{(i)} - \mu_j\|^2$$

[0083] 其中, $c^{(i)}$ 表示样例i与K个类中距离最近的那个类,值为1到K中的一个, μ_j 表示对

属于同一个类的样本中心点的猜测, $x^{(i)}$ 表示样例 i 的数据;

[0084] 对于每个类 j , 重新计算该类的质心, 如下式:

$$[0085] \quad \mu_j = \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}$$

[0086] 其中, μ_j 表示重新计算后的质心。

[0087] 本实施例在 Matlab 中利用 K-means 聚类算法对益田站地铁乘客在 2018 年 5 月 23 日的研究时段 h 的地铁乘客出站 IC 卡数据聚类, 将 r 设置为 17, 通过上述聚类过程得到聚类结果。

[0088] S105、根据预设时间阈值, 对聚类后的识别结果进行修正, 得到每位地铁乘客的乘坐地铁班次, 并结合地铁到站时刻表得到每位地铁乘客的下车时间。

[0089] 本实施例的预设时间阈值为 3min, 将预设时间阈值作为聚类质心 μ_j 的邻域半径, 将属于邻域内的地铁乘客归为同一地铁班次, 此处理即为对聚类后的识别结果进行修正, 修正后的聚类结果如图 3 所示, 通过修正后的聚类结果可以得到每位乘客的乘坐地铁班次, 并结合地铁到站时刻表得到每位地铁乘客的下车时间。

[0090] 本领域技术人员可以理解, 实现上述实施例的方法中的全部或部分步骤可以通过程序来指令相关的硬件来完成, 相应的程序可以存储于计算机可读存储介质中。

[0091] 应当注意, 尽管在附图中以特定顺序描述了上述实施例的方法操作, 但是这并非要求或者暗示必须按照该特定顺序来执行这些操作, 或是必须执行全部所示的操作才能实现期望的结果。相反, 描绘的步骤可以改变执行顺序。附加地或备选地, 可以省略某些步骤, 将多个步骤合并为一个步骤执行, 和/或将一个步骤分解为多个步骤执行。

[0092] 实施例 2:

[0093] 如图 4 所示, 本实施例提供了一种地铁乘客下车时间识别系统, 该系统包括获取模块 401、预处理模块 402、分析模块 403、聚类模块 404 和修正模块 405, 各个模块的具体功能如下:

[0094] 所述获取模块 401, 用于获取数据源; 其中, 所述数据源包括地铁到站时刻表和地铁乘客出站 IC 卡数据表。

[0095] 所述预处理模块 402, 用于对地铁乘客出站 IC 卡数据表进行预处理; 该预处理模块 402 如图 5 所示, 具体包括:

[0096] 地铁站点选择单元 4021, 用于从地铁乘客出站 IC 卡数据表中选择一个地铁站点作为研究对象, 筛选该地铁站点的地铁乘客出站 IC 卡数据;

[0097] 研究时段确定单元 4022, 用于选择该地铁站点在某个研究时段内的地铁乘客出站 IC 卡数据。

[0098] 数据清洗单元 4023, 用于对该地铁站点在研究时段内的地铁乘客出站 IC 卡数据中时间重复、有误的数据进行过滤, 确保每张 IC 卡对应研究时段内的一个出站刷卡时间。

[0099] 所述分析模块 403, 用于根据地铁到站时刻表, 分析在研究时段内到达地铁站点的地铁班次数量。

[0100] 所述聚类模块 404, 用于利用 K-means 算法对预处理后的地铁乘客出站 IC 卡数据表

进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇。

[0101] 所述修正模块405,用于根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

[0102] 进一步地,在预处理模块402之后,还可包括:

[0103] 排序模块,用于对预处理后的地铁乘客出站IC卡数据表中的地铁乘客出站IC卡数据按照时间进行升序排序。

[0104] 本实施例中各个模块的具体实现可以参见上述实施例1,在此不再一一赘述;需要说明的是,本实施例提供的系统仅以上述各功能模块的划分进行举例说明,在实际应用中,可以根据需要而将上述功能分配由不同的功能模块完成,即将内部结构划分成不同的功能模块,以完成以上描述的全部或者部分功能。

[0105] 实施例3:

[0106] 本实施例提供了一种计算机设备,该计算机设备可以是计算机,如图6所示,其包括通过系统总线601连接的处理器602、存储器、输入装置603、显示器604和网络接口605,该处理器用于提供计算和控制能力,该存储器包括非易失性存储介质606 和内存存储器607,该非易失性存储介质606存储有操作系统、计算机程序和数据库,该内存存储器607为非易失性存储介质中的操作系统和计算机程序的运行提供环境,处理器602执行存储器存储的计算机程序时,实现上述实施例1的地铁乘客下车时间识别方法,如下:

[0107] 获取数据源;其中,所述数据源包括地铁到站时刻表和地铁乘客出站IC卡数据表;

[0108] 对地铁乘客出站IC卡数据表进行预处理;

[0109] 根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量;

[0110] 利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇;

[0111] 根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

[0112] 所述对地铁乘客出站IC卡数据表进行预处理之后,还可包括:

[0113] 对预处理后的地铁乘客出站IC卡数据表中的地铁乘客出站IC卡数据按照时间进行升序排序。

[0114] 实施例4:

[0115] 本实施例提供了一种存储介质,该存储介质为计算机可读存储介质,其存储有计算机程序,所述程序被处理器执行时,处理器执行存储器存储的计算机程序时,实现上述实施例1的地铁乘客下车时间识别方法,如下:

[0116] 获取数据源;其中,所述数据源包括地铁到站时刻表和地铁乘客出站IC卡数据表;

[0117] 对地铁乘客出站IC卡数据表进行预处理;

[0118] 根据地铁到站时刻表,分析在研究时段内到达地铁站点的地铁班次数量;

[0119] 利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,将该地铁乘客出站IC卡数据表聚类成与地铁班次数量一致的簇;

[0120] 根据预设时间阈值,对聚类后的识别结果进行修正,得到每位地铁乘客的乘坐地铁班次,并结合地铁到站时刻表得到每位地铁乘客的下车时间。

[0121] 所述对地铁乘客出站IC卡数据表进行预处理之后,还可包括:

[0122] 对预处理后的地铁乘客出站IC卡数据表中的地铁乘客出站IC卡数据按照时间进行升序排序。

[0123] 本实施例中所述的存储介质可以是磁盘、光盘、计算机存储器、只读存储器 (ROM, Read-Only Memory)、随机存取存储器 (RAM, Random Access Memory)、U盘、移动硬盘等介质。

[0124] 综上所述,本发明通过对地铁到站时刻表和地铁乘客出站IC卡数据表中的数据进行预处理,利用K-means算法对预处理后的地铁乘客出站IC卡数据表进行聚类,修正聚类结果,能够推算出地铁乘客所乘坐的地铁班次,进一步识别出地铁乘客下车时间,为计算地铁乘客离开地铁门后步行到达收费闸门的时间提供数据支撑,且识别精确度较高。

[0125] 以上所述,仅为本发明专利较佳的实施例,但本发明专利的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明专利所公开的范围内,根据本发明专利的技术方案及其发明构思加以等同替换或改变,都属于本发明专利的保护范围。

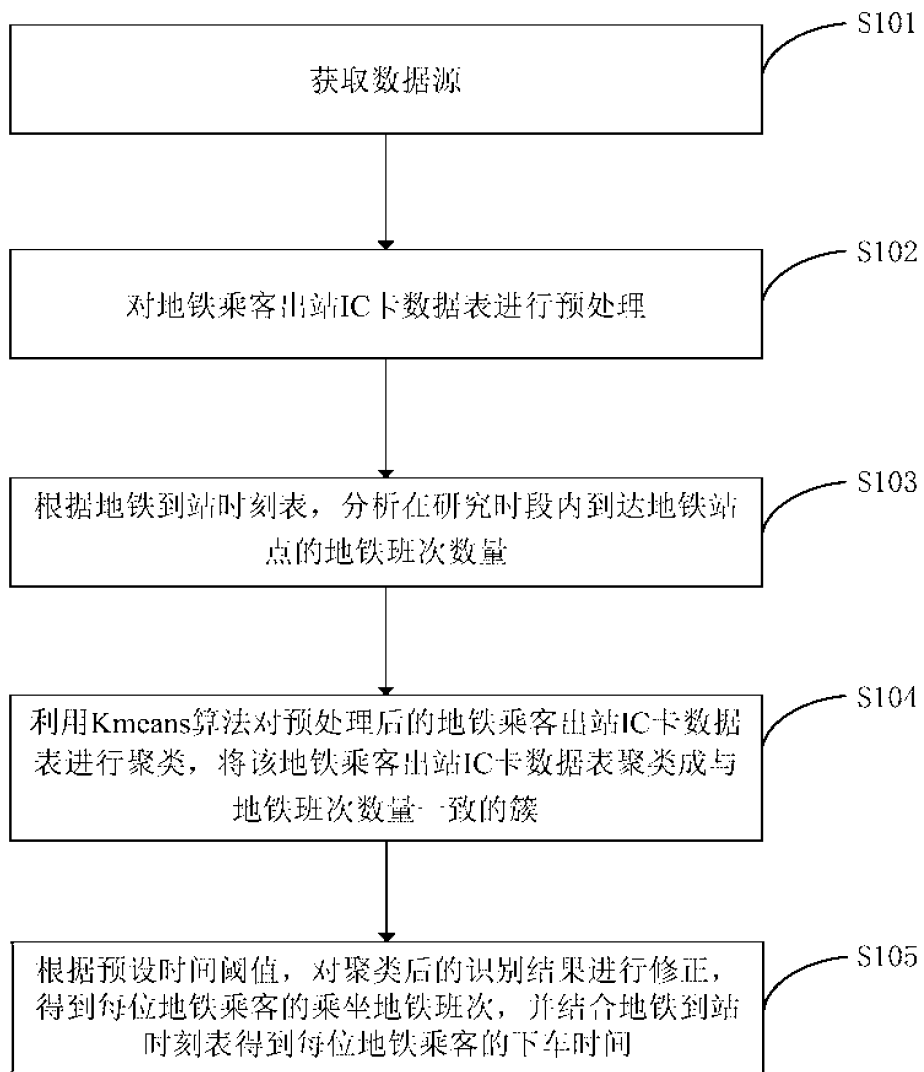


图1

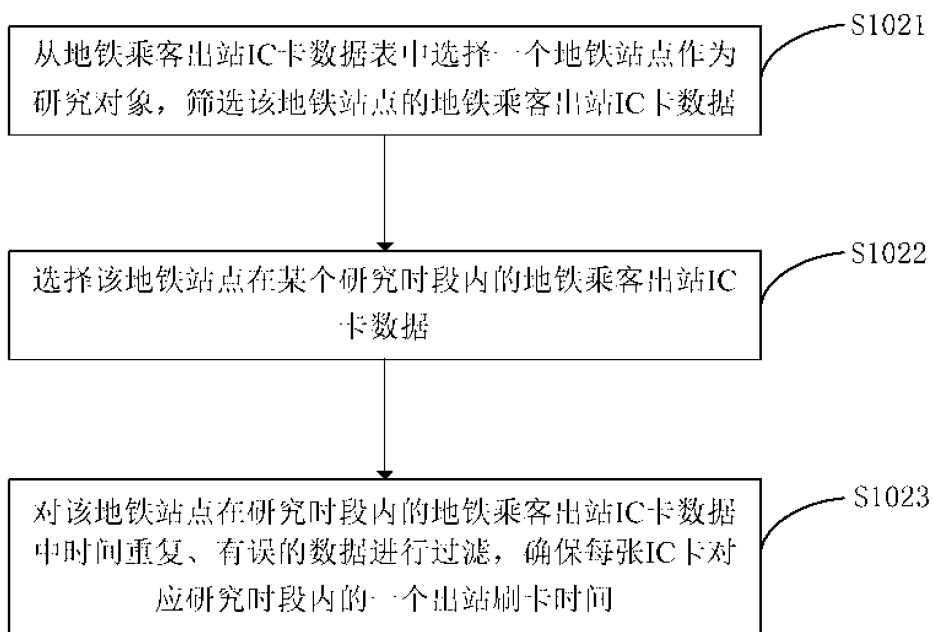


图2

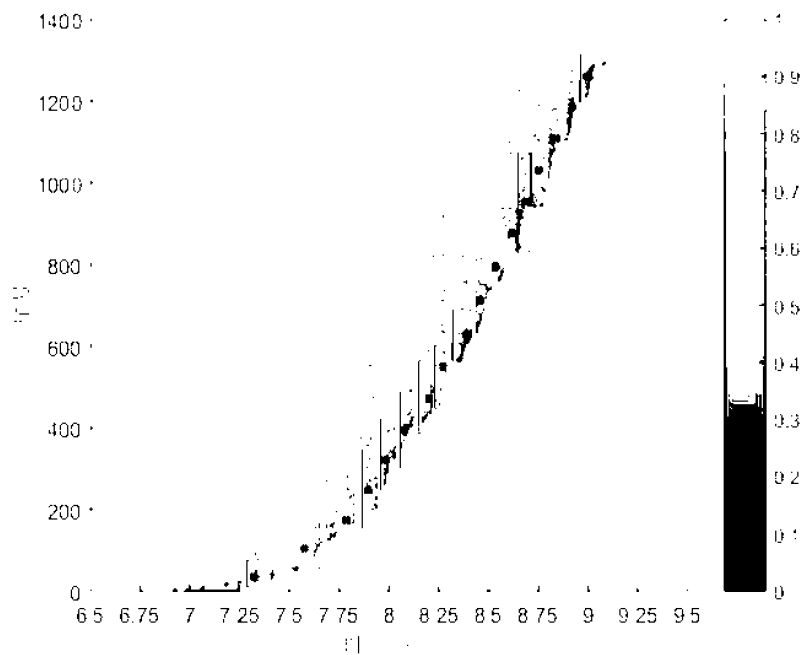


图3

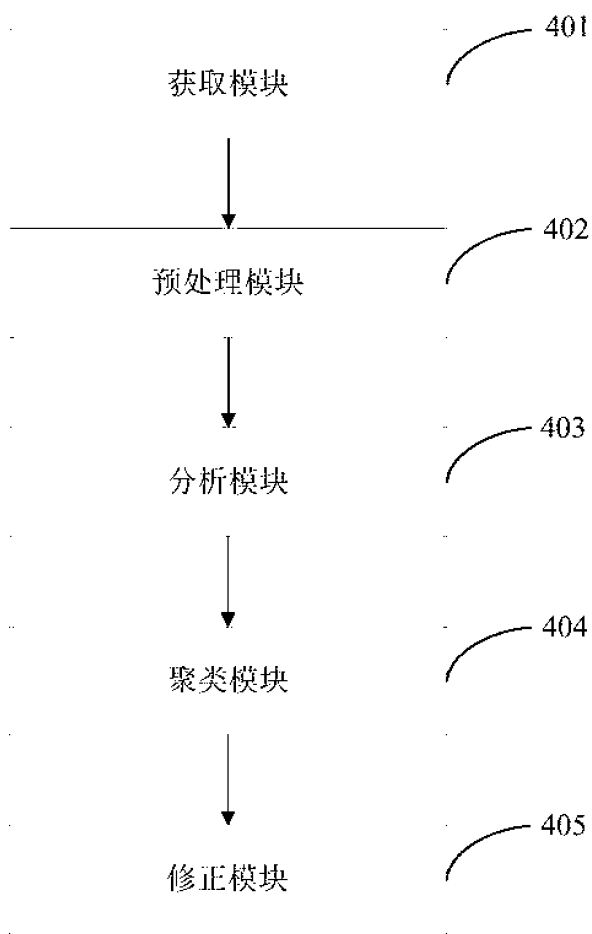


图4

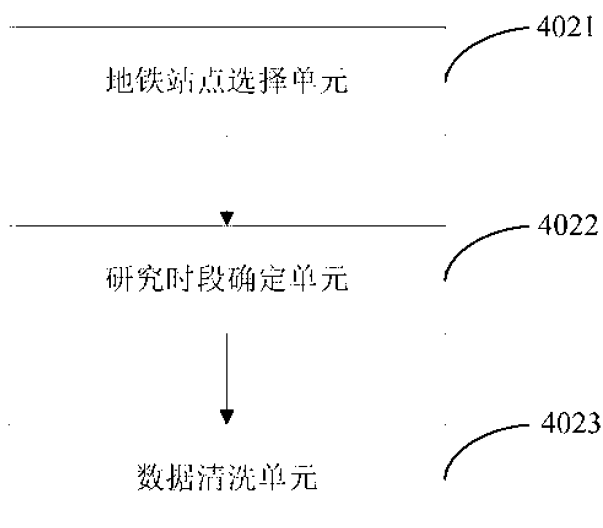


图5

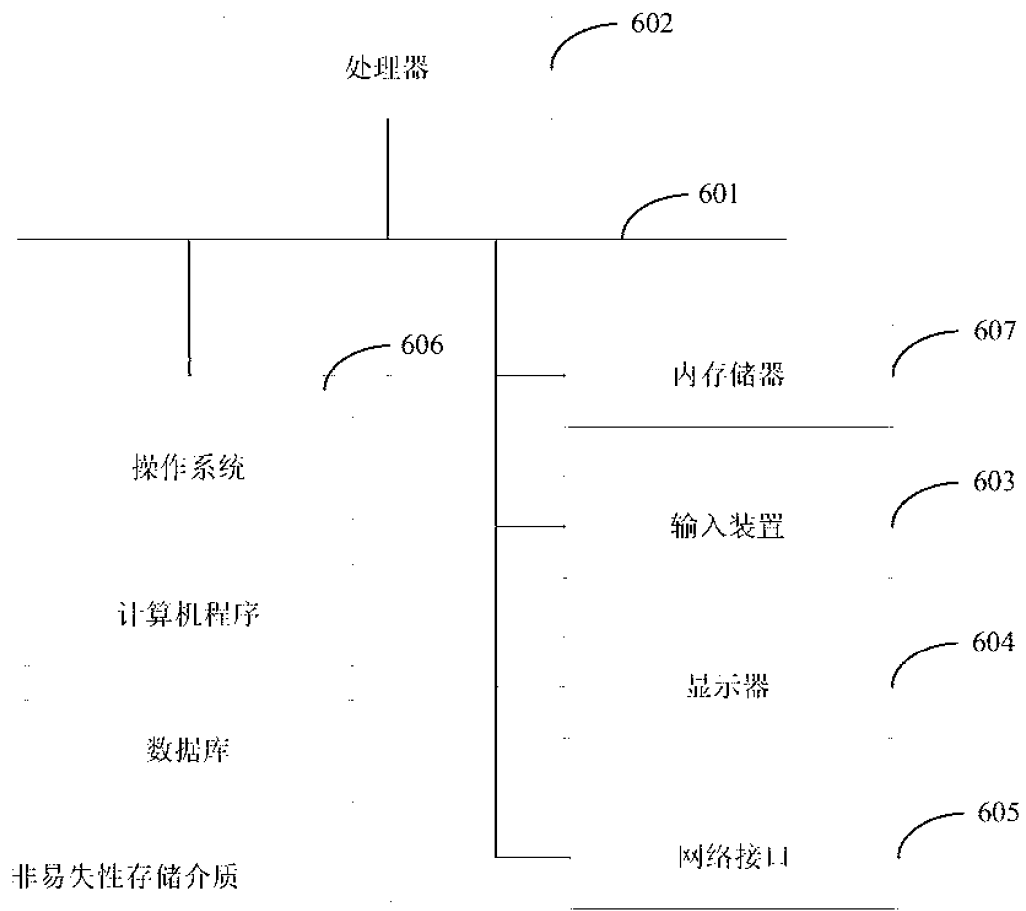


图6