发明名称：

基于北京地铁IC卡出行记录的乘客画像设计方法

摘要：

本法明公开了基于北京地铁IC卡出行记录的乘客画像设计方法，乘客画像内容主要包括：乘客出行模式的刻画；乘客个人偏好的分析；乘客住址与工作地以及两者之间距离的判断。设计方法包括根据乘客IC卡出行记录，对记出行录进行数据清洗，填充，格式转换，特征抽取及缩放等预处理，结合kmeans算法与轮廓系数法刻画乘客出行模式；基于乘客IC卡出行记录和时空分析相结合的方法判断乘客住址，工作地，住址与工作地的距离，个人偏好；本发明能刻画每张IC卡背后客户的出行规律，个人偏好等，可应用于未来城市轨道公共交通领域推出的针对乘客的智能个性化服务，有很高的实用性。

主要内容：

(一).根据乘客IC卡出行记录，对记出行录进行数据清洗，填充，格式转换，特征抽取及缩放等预处理，结合kmeans算法与轮廓系数法刻画乘客出行模式

(二).基于乘客IC卡出行记录和时空分析相结合的方法判断乘客住址，工作地，住址与工作地的距离，个人偏好等。

|  |  |
| --- | --- |
| 主要内容实现大致步骤 | |
| 乘客出行模式的分析 | 乘客个人偏好，住址，工作地及其距离的分析 |
| 1.开始  2.获取数据源：地铁乘客IC出行记录  3.数据存储：存储在非关系型数据库hbase中  4.数据处理：异常值，空值，极端值，不符合对应规则的记录删除  5.乘客出行链条的生成：每次出行记录都包括进站与出站，按时间排序，删掉只有入站或出战记录的数据  6.乘客出行特征抽取：通过数据库编程语言sql查询得出首次进站时间，终次离站时间，总出行记录数，日均出行记录数，周均出行记录数，单次出行记录平均时长等  7.数据特征缩放处理  8.导入乘客出行特征数据，使用k-means算法对乘客出行模式的分析：算法簇类中心个数k由轮廓系数(SilhouetteCoefficient)法得出  9.乘客出行模式的分类结果展示  10.结束 | 1.开始  2.根据主要内容跟（一）抽取的乘客出行特征数据：首次出行最大概率的进站点与终次出行最大概率的出站点相同判定为居住地范围，首次出行最大概率的出站点与终次出行最大概率的进站点判定为工作地范围。  3.使用百度地图api返回的经纬度，推理距离计算公式，得到工作地与工作地的大致距离。  4.根据主要内容（一）得到的用户出行链记录与时空分析方法相结合，判定乘客参与过某种活动，设定阈值，给乘客添加个人偏好标签  5.结束 |

主要内容（一）具体实现：乘客出行模式刻画

**1源数据获取**：

本发明使用数据源来自于北京地铁的乘客IC卡出行记录（2019年x月x日—2019年x月x日）。地铁IC卡部分数据格式如下所示。

|  |  |  |
| --- | --- | --- |
| 列名 | 描述 | 示例 |
| ID | 卡号 | 10007511274081734059 |
| LINE | 线路 | 8 |
| TIMESTAMP | 时间 | 20190930103015 |
| IN\_OUT | 进出站标识 | OUT |
| STATION | 站点 | 朱辛庄 |
| PRICE | 价格 | 4 |
| GATE | 闸机号 | Xxxxxxxxx |

上表含义为为卡号为10007511274081734059的乘客于2019年9月30号10点30分15秒在地铁8号线朱辛庄站刷xxxx号闸机下车，本次出行消费4元。

**2.数据预处理**

1）数据存储：将IC卡出行记录存入数据库中，相比关系型数据库，非关系型数据库数据存储时不会受制于关系型数据库的各种完整性规则，易于操作，存储便利。

2）异常，缺失字段，极端数据清理：设定规则进行相应处理，以8号线为例，其运营时间为05：00—22：05，如果表数据时间属性有不在此范围内，应该将此条记录删除

3）乘客出行链条生成：乘客的一次出行记录应该为进站和出站两次刷卡记录的结合，将以存储在Hbase数据库的乘客IC卡出行记录，按时间戳大小从小到大便利，正常条件下首条进出站标识为in，次条为out。对于只有入口或出口的乘客记录，应当将词条记录删除。执行sql自连接操作，生成用户出行链视图结构如下表所示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | TIME\_IN | STATION\_IN | TIME\_OUT | STATION\_OUT | PRICE |
| 001 | 20190601093010 | A | 20190601103020 | B | 4 |

4）乘客出行数据的特征抽取：在4）步骤生成视图的基础上，使用sql基础查询语句及sql的count，avg，max，min等内置函数，初步得到乘客的出行特征数据，每个乘客特征数据形式如下所示：

|  |  |  |
| --- | --- | --- |
| 列名 | 描述 | 示例 |
|  | 首次进站平均时间 | 08:30（8点30分） |
|  | 终次离站平均时间 | 21:20（21点20分） |
|  | 月总出行天数 | 20（天） |
|  | 日均出行记录数 | 5.2（次） |
|  | 周均出行记录数 | 36.4（次） |
|  | 单次出行记录平均时长 | 19.3（19.3分钟） |
|  | 有刷卡记录的站点数 | 20（个） |
|  | 首次出行最大概率的进站点 | 西小口 |
|  | 首次出行最大概率的出站点 | 朱辛庄 |
|  | 终次出行最大概率的进站点 | 朱辛庄 |
|  | 终次出行最大概率的出站点 | 西小口 |

**3.基于k-means算法和轮廓系数法(SilhouetteCoefficient)的乘客出行模式的分析实现**

1)聚类变量的确认：首次进站时间，终次离站时间，总出行记录数，日均出行记录数，周均出行记录数，单次出行记录平均时长，有刷卡记录的站点数。

2）变量数据格式转化：将上述聚类变量转为整形或者浮点型，其中平均进站时间和离站时间转为以分钟为基本单位的格式，例如：08:30转为8\*60+30=510

3）聚类变量的标准化特征缩放：如果各聚类变量的特征范围相差过大，不利于模型计算速度；某个聚类变量值过大，聚类结果会被此变量所主导，会影响模型精度与收敛速度。所以需要进行标准化特征缩放。

求出每个聚类变量x的平均值与标准差，得到新的变量特征值如下：

4）使用k-means算法对聚类变量进行分类，其中初始簇类个数在4-9之间，最终的簇类个数使用轮廓系数法确定。

|  |
| --- |
| **导入经过缩放后的乘客出行特征数据，使用k-means算法具体实现过程** |
| (a).k-means算法的优化目标是要最小化所有的数据点与其所关联的聚类中心点之间的距离之和，k-means算法的代价函数（又称畸变函数 Distortion function）为：  J为代价函数，k为簇类的个数，为簇类中心坐标，m为总的样本点个数，为样本点，，。  (b).k-means算法是一个经典的据类算法，算法接收一组未标记的数据集，然后将数据据类成不同的组。K-均值是一个迭代算法，假设我们想要将数据聚类成 n 个组，其方法为:  <1>首先随机选择𝐾个点，称为聚类中心（cluster centroids）；  <2>对于数据集中的每一个数据坐标点，按照距离𝐾个中心点的距离，将其与距离最近的中心点关联起来，与同一个聚类中心点关联的所有点聚成一类。  min{d(i\_), d(i\_), d(i\_)…,d(i\_)}  <3>计算每一个簇的坐标平均值，将该组所关联的聚类中心点移动到平均值的位置。  =  重复步骤 <1>~<3> 直至簇类中心点不再变化，此时代价函数J也到达最小值。 |
| **k-means算法k值选取的具体过程：** |
| k-means算法的关键在于族类个数k的选取，k过小会使最终的代价函数J较大，k过大虽然代价函数J会非常小，但分类种数太多，导致实际效果不佳，因此k的确认异常重要。首先此k-means算法主要应用到乘客出行模式的分类，不适宜分类过多，应该将k确认在4-9之间。  而k的最终确认需要使用轮廓系数(silhouette cofficient)来确定,选择轮廓系数和最大所对应的作为族类个数。  具体步骤:  (a).计算样本点到同簇其他样本的平均距离a(i)。将a(i)称为样本的簇内不相似度。 a(i) 越小，说明样本点越应该被聚类到该簇。  (b).计算样本到其他某簇的所有样本的平均距离b(ij),称为样本与簇的不相似度。定义为样本的簇间不相似度: b(i)=min{b(i1), b(i2), ..., b(ik)}。  b(i)越大，说明样本越不属于其他簇。  (c).根据样本的簇内不相似度a(i)和簇间不相似度b(i)，计算所有样本的轮廓系数：    轮廓系数范围在[-1,1]之间。该值越大，越合理。s(i)接近1,则说明样本聚类合理;s(i)接近-1,则说明样本更应该分类到另外的簇;若s(i)近似为0,则说明样本在两个簇的边界上  (d)计算所有样本点轮廓系数s(i)的和，选取轮廓系数最大的聚类结果的k值作为簇类个数。 |

5）还原聚类变量值，完成聚类，形成乘客模式

原特征值x=

**4.结果示例**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 出行模式 | 月均出行天数 | 周均出行次数 | 平均出行时长 | 初次出行平均时间 | 末次出行平均时间 | 有刷卡记录的站点数 | 人数占比 |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| …… |  |  |  |  |  |  |  |

**主要内容（二）具体实现：乘客个人偏好，住址和工作地及其距离的大致判断**

本部分主要基于乘客IC卡出行记录和时空分析相结合的方法来实现

**1）个人偏好判断**

在高速发展，生活压力大的繁华大都市，每个人都必须为美好生活而奋斗拼搏，成年人每天的大部分时间是在忙碌的工作中度过的，学生的生活重心在学习上。只有在闲暇之余，我们才可能抽挤出时间用在个人爱好上。

从信息论的角度来看，某件事物发生的频率越低，其中蕴含的信息量越大。

|  |
| --- |
| f(P)对应P事件的信息量，p为该事件的概率。概率越低，信息量越大 |

正所谓偶然中存在着必然，一个人的爱好可以通过其占比频率较低的个人出行记录所对应的事件来判断。通过基于时空分析的方法，例如A时段B地正在举办一场音乐会，而X乘客在该时段，该地附近的P站点有出站记录，且其出现在该站点的概率小于平均站点的概率，可判定该乘客参加了该音乐会，其喜欢音乐。

事件🡪时间，空间，人物🡪个人爱好

判定该乘客参与了某活动及爱好标注的过程：：

1.空间符合：(a).python爬虫，爬取大麦网各种活动的时间，地点，所属爱好类型；

以大麦网为例：可将爱好初步分为音乐，话剧，体育，亲子类，展览，相声等

(b).爬取北京各地铁公交站点和大麦网各活动地点的经纬度

通过经纬度计算直线距离，记录下距离各活动地点前四的地铁站点

(c)将以上内容建表，存储在mysql数据库中

2.人物和时间符合：(a)在活动开始前乘客IC卡有在该活动场地距离前四某站点出站的记录

(b)在活动结束后乘客IC卡有在该活动场地距离前四某站点入站的记录

(c)在活动进行期间乘客IC卡无出行记录

3.空间，时间，人物都满足的乘客最后筛选：

判断该活动地点的附近站点是否是该乘客出行的常用站点，使用数据库操作语言sql查询乘客出行记录里各站点的频度，大于平均频度标记为常用站点，反之为不常用站点。排除掉是常用站点的乘客，为余下的乘客标记上个人爱好。

上述3步骤可用如下公式表示：

|  |
| --- |
| Result=Y&&(  )=0  Result=1则标记爱好，result-0不标记爱好  Y=1为不常用站点，0为常用站点  为进站时间，为出站时间  为活动开始时间，为活动结束时间  )=0表示该时间段没有乘车记录 |

**2)住址与工作点的判断过程：**

根据主要内容（一）获取的乘客出行特征数据得

(a)每日首次出行最大概率的进站点与每日终次出行最大概率的出站点相同🡪住址

例：=西小口

(b)每日首次出行最大概率的出站点与每日终次出行最大概率的进站点相同🡪工作地

例：==朱辛庄

**3).住址与工作地的距离的大致计算：**

（a）注册百度地图开发密钥+python爬虫scrapy框架+地点输入获取json数据，从json数据中读取各地点精度，维度，然后计算距离

（b）距离公式：

|  |
| --- |
| （1）地球球心为原点，球心与赤道上0经度点的连线为X轴，球心与赤道上东经90度点的连线为Y轴，球心与北极点的连线为Z轴：  x=R×cosα×cosβ R为地球半径，约等于6371km  y=R×cosα×sinβ α为纬度，北纬取+，南纬取-  z=R×sinα β为经度，东经取+，西经取- |
| （2）求至直线距离，即弦长：  L=[(x1-x2)^2+(y1-y2)^2+(z1-z2)^2]^0.5 |
| （3）弦长转弧长：  C=arcsin(L/2R)\*Pi\*R/90 |