2017 年深圳市交通仿真系统日常更新

深圳市规划和国土资源委员会（市海洋局）
深圳市规划国土发展研究中心

二〇一八年 九月

目 录

插图目录

表格目录

1 数据挖掘系统更新

在上节的数据更新的基础上，深圳市交通仿真系统（二期）分析挖掘的交通指标数据也需要重新进行运算，获得最新的各类交通指标。因此，需要在深圳市交通仿真系统（二期）已建成的数据挖掘系统基础上，按照设计的计算机流程对交通指标进行重新运算，并更新原有系统后台的数据库。

§1.1 数据挖掘系统结构

深圳市交通仿真系统（二期）建成的数据挖掘系统是一个标准的计算机软件系统，按照严格的数据流处理架构将数据挖掘的整个流程分为：数据载入、数据入库、数据汇总四个部分。具体的系统结构图如下所示。

§1.2 更新后台数据库

1.2.1 数据载入

在第??章的数据更新工作中，已经完成了原始数据的采集和预处理工作，得到了动态交通数据和静态 GIS 数据两个数据集。数据载入的主要工作是将以文件形态存在的原始动态交通数据转移（载入）到关系数据库中，为下一步处理步骤（数据挖掘）做数据准备。在数据转移的过程中，会对数据进行基本的清洗过滤，剔除那些格式不正确的数据。

由于动态数据预处理阶段只是对数据的格式进行标准化操作，数据本身仍然是原始数据，其中包含有数据传输过程中产生的错误、冗余和无效数据。因此，针对原始动态交通数据，需要根据每种数据类型特点设定的清洗规则进行清洗，获得可以用于数据挖掘计算的有效数据。

清洗后的有效数据存入数据挖掘后台数据库的数据表中，数据表全部以 RAW 开头，以下是所有清洗后数据表信息的说明。

表 1.1: 清洗后的数据表结构

表名	表说明	字段名	字段说明
RAW_IC_DATA	存储公交和轨道深圳通 IC 刷卡数据	CARD_ID	IC 卡 ID 编号
		TRADE_DATE	刷卡时间
		TRADE_TYPE	交易类型
		TERMINAL_ID	刷卡终端 ID 号
RAW_GPS_DATA	存储公交 GPS 数据	LINE_NAME	公交线路 ID 编号
		BUS_NAME	公交线路名称
		GPSTIME	GPS 时间戳
		STATION	公交到站站点 ID 号
RAW_LP_DATA	存储车牌识别数据	PASS_TIME	识别时间戳
		DETECTOR_ID	识别点 ID 号
		CAR_ID	被识别车牌
		LANE	车道

1.2.2 数据挖掘

数据挖掘主要功能是以数据载入生成的数据为输入，根据交通规划的具体业务需求和各类交通数据的特点设计专门的数据挖掘算法，对动态交通数据进行全方位的挖掘运算，生成各种专题挖掘数据，主要包括 15 分钟、60 分钟和全天的分时段指标数据。

数据挖掘的工作在深圳市交通仿真系统（二期）中已经开发了专门的数据挖掘程序，可以进行全自动化作业。数据挖掘的基本思路是：从表 10 中将数据库后台的有效数据和相关静态 GIS 数据读取处理，转换为 JAVA 对象并在内存中存储，然后按照数据挖掘的算法进行挖掘运算，最后将运算结果按照一定的数据结构存储在内存中，等待后续的数据入库操作。

虽然数据挖掘步骤实现了全自动化，但是在工作中也需要特别注意：

- **更新 GIS 相关数据表**

数据挖掘系统后台数据库中有大量和 GIS 相关的数据表，表名以 GIS 开头。由于前面已经进行了 GIS 数据图层的更新，因此在进行数据挖掘前要首先把 GIS 相关数据表也同步更新。

更新的方法是：首先删除后台数据库中原有 GIS 表的数据，然后将第**??**节中更新后的 shapefile 文件的属性数据导出为文本格式，最后再将文本格式文件导入后台空的

GIS 表。

- **防止重复处理**

输入数据表已经有被处理过的数据，例如，输入数据表中有 1 季度的数据，并且已经处理过，现在载入 2 季度的数据进行处理，就要指定数据挖掘的时间范围，可以在数据挖掘程序中通过设置 from 和 to 参数来实现。

- **防止内存占用过高**

有些类型的数据在处理时，占用的内存资源会随着被处理数据数量的增加而递增，例如出租车数据。为了避免一次处理太多的数据而导致处理程序内存溢出，需要对数据进行分片多次处理。根据之前的更新经验，一次处理的数据不要超过每种数据类型各一周。

- **并行处理**

在系统资源允许的情况下，可以将原始数据按天为单位划分为多个任务进行并行处理，可以大大加快数据的处理速度。

- **错误异常处理**

如果在程序执行过程中发生异常，程序中断，导致只有部分数据被处理，结果数据不完整，这时候需要重新执行任务。在重新执行任务时，必须将上一次处理生成的部分结果数据删除，否则在后续的数据入库流程中会产生数据库数据冗余错误。

1.2.3 数据入库

上面的数据挖掘步骤将挖掘运算后的结果存储在内存中，本节的数据入库工作就是将内存中的计算结果，插入到中间数据库中预先设计的数据表中。这部分工作全部由程序自动化完成。

针对每种数据类型和数据挖掘对数据的依赖关系，具体的数据入库工作分解为 IC 刷卡数据、轨道数据、公交数据三个专题。每类数据的入库流程如下所示，如果在数据入库中发生异常错误情况，可以根据入库流程回溯到错误的初始位置。

表 1.2: 数据挖掘成果入库表

专题	读取表	挖掘指标	入库表
刷卡数据	RAW_IC_DATA	分析换乘	metro_m2b_line_* metro_m2b_station_* bus_b2m_line
		分析刷卡概况	metro_ic_overview bus_ic_overview

轨道数据	RAW_IC_DATA	线路流量	metro_line_flow_*
		站点流量	metro_station_flow_*
		断面流量	metro_segment_flow_*
		小区流量	metro_zone_flow_*
		吸引点流量	metro_sap_flow_*
		站点 OD	metro_od_flow_*
		出行 OD	metro_od_zone_flow_*
		换乘	metro_transfer_line_flow_* metro_transfer_station_flow_*
		平均距离	metro_dst_man metro_net_dst_mean
		距离分布	metro_dst_distance metro_dst_station metro_dst_elapse
		吸引点 OD	metro_sap_od_*
		吸引点出行时长分布	metro_sap_dst_elapse
公交数据	RAW_GPS_DATA RAW_IC_DATA	发车频率	bus_frequency_60
		公交线段车速	bus_link_speed_60
		公交出行行程时间	bus_zone_time_*
		站点流量	bus_station_flow_*
		小区流量	bus_zone_flow_*
		吸引点流量	bus_sap_flow_*
		小区 OD 流量	bus_od_zone_flow_*
		站点 OD 流量	bus_od_flow_*

1.2.4　数据汇总

数据入库工作完成后，数据挖掘系统的工作已经全部完成，但是考虑到在深圳市交通仿真系统（二期）中，数据挖掘系统是数据查询系统的一个前置系统，数据挖掘的数据需要通过数据查询系统来展现，因此数据挖掘后台的中间数据库需要进一步统计汇总，主要将中间数据库中的指标成果进行时段上的集聚，形成日间、夜间、全天、高峰时段等特征时段指标，再将汇总统计后的结果推送至数据查询系统后台的指标数据库中。

由于在深圳市交通仿真系统（二期）已建成的系统中，已经建立了数据挖掘系统后台中间数据库与数据查询系统后台指标数据库相关表的对应关系，因此数据汇总工作可

以由程序自动化完成。数据汇总后，数据查询系统指标数据库中的主要指标表??所示。

表 1.3: 数据汇总后的交通指标

数据类型	基本指标	时段
轨道交通	线路客流、站点客流、轨道断面客流、站点 OD、换乘、出行距离/时间	日间、夜间、全天、高峰时段
公交	总客流、线路客流、换乘	
道路交通	车牌识别点流量、车牌识别点间行程时间	

§1.3 批处理更新操作

由于数据挖掘系统输入数据的数据量大，而且数据挖掘各个流程中涉及的步骤多且结构清晰，因此很适合采用批处理模式针对每种挖掘流程进行自动化运算。

在批处理运行模式下，系统可以将数据载入、数据挖掘、数据汇总等多个处理任务，在批处理任务配置文件中定义，然后将该批处理任务提交给数据挖掘系统，系统将自动执行这些任务，并将执行结果返回并入库。

1.3.1 定义批处理配置文件

批处理配置文件是自定义脚本文件，每行定义一个任务。所有任务都包含以下三个字段：

```
GID;ID;TASK TYPE
```

GID 任务组 ID，由数字组成，使用过的任务组 ID 不能再次使用；具有相同 GID 的任务可以并行执行，具有不同 GID 的任务按照定义的先后顺序串行执行

ID 任务 ID，由数字组成；表示该任务是任务组内的第几项任务；在同一任务组内，任务 ID 不能重复。

TASK TYPE 任务类型，代表不同的任务类型

TASK TYPE 可以从以下字符中选择：

✏ **x 或 X**: 表示是否继续处理标志；如果任务组的最后一个任务的任务类型被设置为 x，表示如果任务组内有任务执行失败，则中断整个批处理的执行

- ✏ **v 或 V**：表示定义变量任务；通常放置在整个批处理配置文件首部，定义多个后续任务可以引用的变量；GID 和 ID 通常都为 0
- ✏ **a 或 A**：表示数据载入任务；需要指定数据源参数；
- ✏ **b 或 B**：表示数据挖掘任务；需要指定被分析数据的起始和截止日期；
- ✏ **c 或 C**：表示数据汇总任务；需要指定汇总的年度参数；
- ✏ **d 或 D**：表示数据库操作任务；对于 V 类型的变量设置任务，包含以下变量设置字段：

其中，V 类型的变量设置任务是一个键值对，格式为“var=value”，等号前后不能有空格，后续任务以“$var”方式引用该变量。例如：

```
0;0;v;year=2016
0;0;v;quarter=s3
0;0;v;day1=${year}0902
0;0;v;day2=${year}0903
...
11;1;a;ic;load.bus;/home/dm/data/${year}/${quarter}/ic/ic_bus_09.csv
11;0;x
...
21;1;b;ic;mine;${day1};${day2}
...
```