**nextlinks.txt文件内容：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 类型 | 描述 | 值域 | 示例 |
| link id | int | 道路小段id | 多为4到6位数字组成的字符串 | 639878 |
| nextlinks | int | 与该道路小段直接相连的下一个道路小段id集合 | 现有道路小段id集，每个道路小段的nextlinks数量不超过七个 | 539133,639879 |

**行程信息：每一条记录为一条行程信息，这一条记录中包括三部分，分别为以下三个表格所示的head部分、link部分、cross部分。每个部分之间以“;;"号分割，部分内部以” “空格或者逗号分割。**

1）行程的订单信息部分由head部分组成，该部分由下表的字段依次以空格为间隔组成

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| head部分 | 字段 | 类型 | 描述 | 值域 | 示例 |
|  | order id | string | 订单id | ---- | 8854706 |
|  | ata | float | 真实行程总时间 | 订单结束时间-乘客上车时间，单位秒 | 823 |
|  | distance | float | 路线距离 | 全程的路面距离，即子路段长度之和，单位米 | 3782.3359 |
|  | simple eta | float | 出发时刻平均通行时间加和 | Link time和cross time的累加 | 493 |
|  | driver id | int | 司机id | ---- | 26392 |
|  | slice id | int | 出发时刻时间片id(每5分钟一个) | 对出发时刻做了分桶，粒度为每5分钟一个时间片，起始时间片为上午上午8点，24小时循环一次 | 41 |

2）行程的轨迹部分由link部分和cross部分组成，因此这两个部分都是序列形式，序列的每个元素由表格内的字段构成。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| link部分 | 字段 | 类型 | 描述 | 值域/说明 | 示例 |
|  | link id | int | 道路小段id | 多为4到6位数字组成的字符串 | 1829 |
|  | link time | float | 出发时刻道路小段的平均通行时间 | link time是出发时刻道路小段的平均通行时间。在订单出发时刻前十分钟内，根据通过该道路小段的历史数据进行求平均统计值，该值作为该订单中该道路小段的link time，即同一个道路小段在不同订单中的link time不同。此外，link time的大小还会受到link ratio的影响 | 2.5503 |
|  | link ratio | float | 道路小段通行比例 | link ratio是指实际轨迹覆盖该小段的占比，在行程的首端和末端是小于1的值 | 0.8855 |
|  | link current status | int | 出发时刻道路小段的路况状态 | 0（未知）  1（畅通）  2（缓行）  3（拥堵） | 0 |
|  | link arrival status | int | 到达时刻道路小段的路况状态 | 0（未知）  1（畅通）  2（缓行）  3（拥堵） | 0 |
| cross部分 | 字段 | 类型 | 描述 | 值域 | 示例 |
|  | cross id | int | 红绿灯路口id | cross id标识一个红绿灯id，用路口入link和出link加"\_"拼接而成，出link和入link可能不相邻。同一个路口因为出link和入link不同会有多种表示方法 | 527206\_223528 |
|  | cross time | float | 挖掘路口通行时间 | 该红绿灯的挖掘红灯周期时间(单位秒)，统计方法同link time | 6 |

**对已有数据进行梳理后，对数据（已涵盖以上所有数据）进行划分和归类：**

1. 描述订单信息的数据：

订单id编号（列名：order id）

真实行程总时间（列名：ata）

路线距离（列名：distance）

出发时刻平均通行时间加和（列名：simple eta）

司机id编号（列名：driver id）

出发时刻时间片id编号（列名：slice id）;

1. 描述路段信息的数据：

道路小段id编号（列名：link id）

出发时刻道路小段的平均通行时间（列名：link time）

道路小段通行比例（列名：link ratio）

出发时刻道路小段路况状态（列名：link current status）

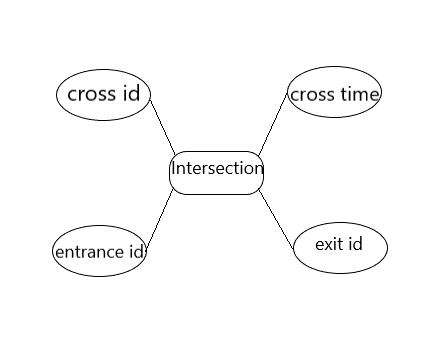
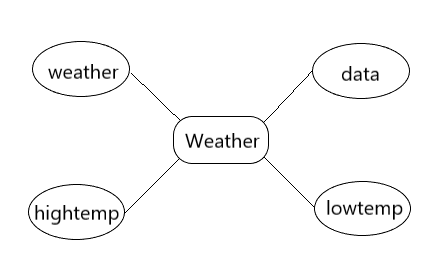
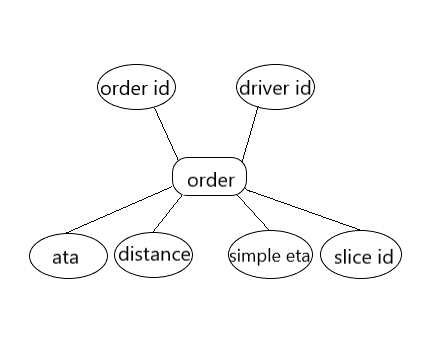
到达时刻道路小段路况状态（列名：link arrival status）;

1. 描述路口信息的数据：红绿灯路口id编号（列名：cross id）

路口通行时间（列名：cross time）;

1. 描述路段拓扑信息的数据：直接相连的下一个道路小段集（列名：nextlinks）
2. 概念模型设计

思路或准则：在对以上数据进行说明后，进行分类和概括，能将以上数据集进行准确概括且易于理解的实体：订单、路段、路口、天气。



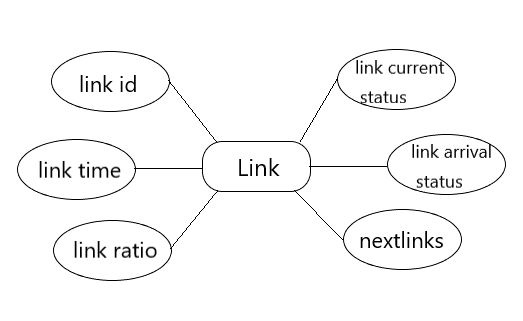


图1：实体属性图

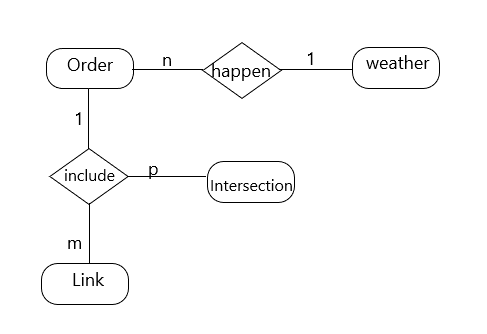


图 2：实体联系图

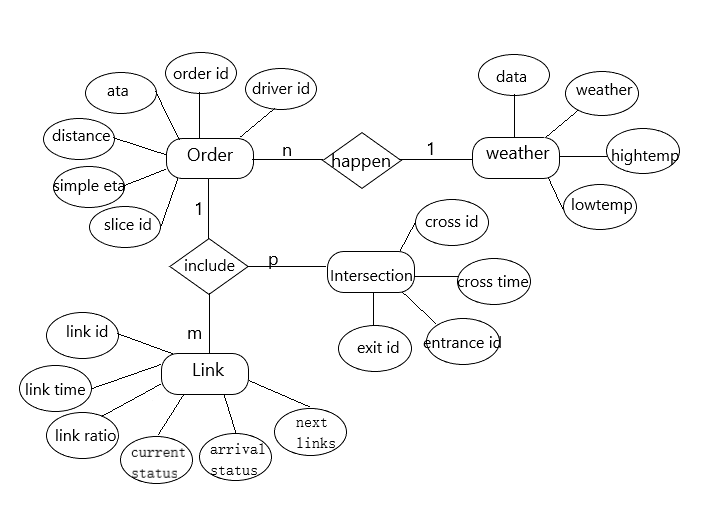


图3：E-R图

1. 逻辑结构设计

共有4个实体，共有20个属性。已覆盖原数据格式字段和信息，且增加2个字段(表3的入link id列和出link id列)。原数据格式中的字段用黑色字体，新增加的属性列用红色字体。现将e-r图进行关系表转化。

在关系表设计之后，共有24个属性。增加的原因：1）外键；2）将nextlinks属性列与link id属性列拿出单独再形成路网拓扑信息表

表 11：order表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **order id**  **(主键）** | ata | distance | simple eta | driver id | slice id | date  (外键，参照表2） |
| 8854706 | 823 | 3782.3359 | 493 | 26392 | 41 | 20200801 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

表 22：weather天气信息表

|  |  |  |  |
| --- | --- | --- | --- |
| date  （主键） | weather | hightemp | lowtemp |
| 20200801 | rainstorm | 28 | 26 |
|  |  |  |  |

表 33：订单下的路小段路况表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **link id**  **（主键）** | link time | link ratio | link current status | link arrival status | **order id**  **（外键参照表1）** |
| 1829 | 2.5503 | 0.8855 | 0 | 0 | 8854706 |
| 483120 | 13.2261 | 1.0000 | 1 | 1 | 8854706 |
| 452455 | 15.3818 | 1.0000 | 1 | 1 | 8854706 |
|  |  |  |  |  |  |

表4 4：订单下的路口信息表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **cross id**  **（主键）** | 入link id | 出link id | cross time | **order id**  **（外码参照表1）** |
| 527206\_223528 | 527206 | 223528 | 6 | 8854706 |
|  |  |  |  |  |
|  |  |  |  |  |

|  |  |
| --- | --- |
| link id  （主键） | nextlinks |
| 876328 | 278684,839685,783268 |
| 635882 | 3728211 |