# 数据挖掘大作业

## FIMI accidents数据集关联规则算法及结果分析

姓名 孙振义 学号 GS1506103

## 数据集基本描述

该数据集来自比利时国家统计协会。该数据集包含了比利时弗兰德斯地区1991-2000年间共计340184起交通事故的详细记录。数据集下载地址 <http://fimi.ua.ac.be/data/accidents.dat>

该数据集中平均每起交通事故数据填充了 45 种属性，全部为序数属性，共有572 种属性值，全部为整数类型。

|  |  |
| --- | --- |
| 事故是否在居住区发生 | 1 是, 2否 |
| 道路类型 | 3 高速公路; 4其他 |
| 车道 | 5 单车道; 6 双车道 |
| 是否位于交叉路口 | 7 是; 8 不是 |
| 路口通行规则 | 9 交警指挥;10 交通灯指挥; 11 路口闪灯; 12 交通信号; 13 右侧优先 |
| 道路环境 | 14 普通公路; 15桥梁; 16 隧道; 18 铁路; 19 弯道 |
| 路况 | 20 路况较差; 21 错误的交通信号; 22 错误的信号灯; 23 路况正常; 24 道路拥挤; 25 下山路; 26 转弯; 27 能见度较差 |
| 杂项 | 28 次生事故; 29 路面湿滑; 30 阳光刺眼; 31 32学校附近; 33游乐场所附近;34公交车站附近; 35 人员被甩出车外; 36 无安全带; 37 无头盔 38 无儿童座椅; 39 事故前货物散落于路面; 39 事故后货物散落于路面;40 事故后起火 |

由于属性较多，其他不再一一翻译，其他属性和属性值可到 <http://fimi.ua.ac.be/data/accidents.pdf> 查阅。

## 数据预处理

将数据从文件中读取出来插入mysql数据库，方便进行简单的统计和数据观察。

## 算法及开源工具介绍

1. Java

Java语言是一种跨平台、适合于分布式计算的、面向对象的编程语言。Java 技术具有简单性、稳健性、平台移植性和安全性，广泛应用于桌上型应用程序、电子设备的嵌入式软件开发、建设大型的分布式企业级应用程序。

2．JDBC

JDBC（Java Data Base Connectivity,java数据库连接）是一种用于执行SQL语句的Java API，可以为多种关系数据库提供统一访问，它由一组用Java语言编写的类和接口组成。JDBC提供了一种基准，据此可以构建更高级的工具和接口，使数据库开发人员能够编写数据库应用程序，同时，JDBC也是个商标名。

1. MySQL

MySQL是一个关系型数据库管理系统，由瑞典MySQL AB 公司开发，目前属于 Oracle 旗下公司。MySQL 最流行的关系型数据库管理系统，在 WEB 应用方面MySQL是最好的 RDBMS (Relational Database Management System，关系数据库管理系统) 应用软件之一。

3．频繁模式树方法

(1)频繁模式树(Frequent Pattern tree)简称为FP-tree，是满足下列条件的一个树结构：它由一个根节点(值为null)、项前缀子树(作为子女)和一个频繁项头表组成。

(2)项前缀子树中的每个结点包括三个域：item\_name、count和node\_link，其中：

item\_name记录结点表示的项的标识；count记录到达该结点的子路径的事务数；

node\_link用于连接树中相同标识的下一个结点，如果不存在相同标识下一个结点，则值为“null”。

(3)频繁项头表的表项包括一个频繁项标识域：item\_name和一个指向树中具有该项标识的第一个频繁项结点的头指针：head of node\_link。

拟使用频繁模式树算法得到包含车祸死亡人数项在内的强模式，分析车祸中人员死亡的伴随因素，希望能得到一些新的知识从而一定程度上减少车祸中的死亡悲剧。

## 挖掘结果分析

王老师，由于最近猎豹移动公司开始试行996的加班政策，而且由于和我小组英语演讲、英语写作作业冲突，另对挖掘算法掌握也不牢固等原因，挖掘尚未完成，但是为了避免拉班级后腿，先将前三部分当作作业提交，其他代码和结果我会在一周之内更新到github，<https://github.com/sunzy993/DM_HIMI_ACCIDENT>，并发送到您的邮箱，给您带来不便，非常抱歉