**蛛网数据分析——潜在嫌疑人推荐**

### 需求定义

根据犯罪人的通讯录（手机号/微信/QQ）信息推荐出潜在犯罪嫌疑人。

### 研究方向

### 根据需求定义再结合数据，算法方面研究的方向是：

### 联想购物篮数据做关联规则分析（主要计算并分析了支持度和置信度两个指标），分析两两手机号共同发生联系的支持度排序，分析第三者在两两共同发生联系的群体中出现的置信度排名和相对全样本数据的提升度排名；

### 联想网页跳转进出的方向PageRank的思想可视化出社交关系图，联系是有方向的（拨入与拨出），联系人是有权重大小的（不同时间段的联系频次）。

### 具体实现

1.对通讯录数据进行清洗，比如手机号通讯录，只保留11位数字的手机号，清洗的条目大致如下（如下条目是按照顺序进行清洗的）：

如果以0开头则去掉0；

如果以+0开头则去掉+0；

如果以+00开头则去掉+00；

如果以+开头则去掉+

如果有空格则去掉空格

如果有-则去掉-

13位，如果以86开头去掉86；

16位，如果是以10193或17911开头则去掉10193或17911

17位，如果以125831，则去掉125831

11位，只保留11位长度的记录；去掉不是1开头的则去掉这些记录

1. 生成数据集

添加一列新字段，将连续的时间划分成各个时间段，对每个时间段内的相互联系人进行去重（我们认为，相互联系的两个人在一个时间段内的高频联系与判别为潜在嫌疑人是弱相关的，而两个人在不同时间段高频联系则认为其与潜在嫌疑人身份是强相关的）。

用于关联规则分析的数据集样式（在R中是一个list，list的长度是时间段个数×犯罪人个数）：

|  |  |  |
| --- | --- | --- |
| 时间段 | 犯罪人手机号 | 包含犯罪人手机号的以及与其发生联系的手机号（R中的向量表示） |
|  |  |  |

1. 支持度计算

根据支持度计算的逻辑意义，可以计算出这样两个指标：

核心犯罪人，即与其发生联系的人较多的犯罪人

核心相关人，即与之发生联系的人较多的非犯罪人

保存满足条件的支持度数据，条件是：

大于支持度阈值0.001，并按照支持度降序

1. 置信度计算

在计算好的支持度数据（规则集）的基础上，计算置信度，并保存满足条件的规则，条件是：

规则前项必须包含犯罪人，规则后项目不包含犯罪人

置信度大于0.6

1. 穷举尝试

对规则前项中包含的犯罪人个数进行穷举，比如在所有的规则中，规则前项中包含的犯罪人数目最大为n，则穷举n,n-1,n-2,...,1，查看每次穷举后规则后项对应的非犯罪人，即为我们推荐的潜在嫌疑人。

1. 可视化

将以上每次穷举的结果可视化出来，形成社交网络关系图，并通过图可以做适当的过滤以及排序（图中线条粗细表示权重）