# 提取策略

一、输入

ACCESS/Excel的一行记录

二、行分析

分三种模式进行数据提取

2.1 精确提取：

可以使用正则表达式对手机号码，邮箱，固话，身份证进行精确提取

出生年月由于表达方式各样，有两种办法：一个是对现有数据的出生年月进行统计，得

到其各种表达方式，然后再用几种不同的正则表达式进行提取；另一个是根据19\*\*来

判别

2.2 模糊提取：

对姓名、单位、地址没有统一的模式来匹配，无法做到精确提取，需要对其

进行模糊匹配和根据表头相结合的方式来做，尽量做到数据的准确性，提高提取的质量。

第一步：表头是可信赖的优先级较高，首先根据表头来提取这部分信息

第二步：对姓名，长度在2-4个关键字，姓氏匹配中国姓氏，就认为其姓名

第二步：单位/地址：中文长度大于4个字以上，中间含省市地区镇等关键字信息

2.3 隐含信息提取：

对身份证信息可以提取省份，性别，出生年月这三个信息

对手机号码可以提取省份地区这个信息。