

# 出租房源采集客户端系统 V1.0

## 用户手册

方宏

(1010402827@qq.com)

### 1 系统介绍

随着网络技术的广泛深入应用，越来越多的人开始在网络上发布出租、求租信息，网络上也出现了为数众多的地产信息相关网站，主要分成房地产信息专业网站（如搜房网、365 等）和大型综合网站的房地产频道（如赶集网、58 同城等）。对于找房者和房产中介而言，一个重要的任务就是从这些网站上检索到所需的信息，但面对众多地产网站上大量的、动态变化的出租信息，手工检索费时费力，且难以及时搜索到所需信息。为了减轻房产中介、找房者上网找房的负担，提高找房的效率，开发出了出租房源采集系统，该系统分为两个子系统，一个是服务端系统，主要是一个垂直搜索引擎，负责从各个房地产信息网站采集出租信息并存储到数据库；另一个是客户端系统，负责响应用户检索请求，从房产信息数据库中提取出出租信息。本手册主要是针对第二个系统即出租房源采集客户端系统（整个系统简称搜房宝）。

出租房源采集客户端系统的主要功能是根据用户要求，查询出租房信息。提供建立查询、保存查询、浏览查询结果、修改查询到的记录、启动数据采集等功能。系统的主要特性如下：

- （1）查询全，响应快。较之手工逐个网站查询，系统从存储了搜索引擎采集的出租信息的数据库（包含了主要地产信息网站的相关信息）中查询数据，相当于一次检索了多个房地产网站，找到的信息更多，速度也更快。
- （2）不同于一般的根据查询条件，立即检索相关信息的普通查询，系统支持延时自动查询。若当前未检索到符合条件的信息，可以保存用户查询，以后会自动检索符合查询条件的出租信息，保证最新采集到的符合查询条件的出租信息能够及时反馈给客户。即使程序退出，重启后，仍能恢复先前的查询数据。该功能可以有效提高客户的找房能力。
- （3）系统支持多条件查询，模糊查询，尽可能多地匹配到请求的信息。
- （4）系统提供手工修正补全出租记录信息的功能。
- （5）系统性能良好，即使有大量的定制查询，仍能保证及时的响应。

该系统在实际使用中显示出优良的性能，较之手工检索网站，查找出租信息，找房速度提高了数十倍，节省了地产中介业务员、找房者的大量时间，找到的信息更全、更准。

系统使用 Java 语言开发，提供 GUI 界面。程序执行后的初始界面如图 1 所示，基本使用过程是：先点击“新建查询”按钮，在弹出的查询窗体中，选择查询条件，浏览查询信息，必要时保存查询；然后在主界面查看保存的查询，系统会自动将最新数据更新到相应查询，期间若有需要，用户可以修改相应信息，或上信息的来源网站直接查看相关信息；最后，如果找到满意的信息，可以将相应查询删除。（为了保护客户信息，下面示例中涉及到的租房数据均是历史抽样数据，仅



图 1    出租房源采集客户端系统主界面

2 建立客户查询

建立客户查询的步骤如下：

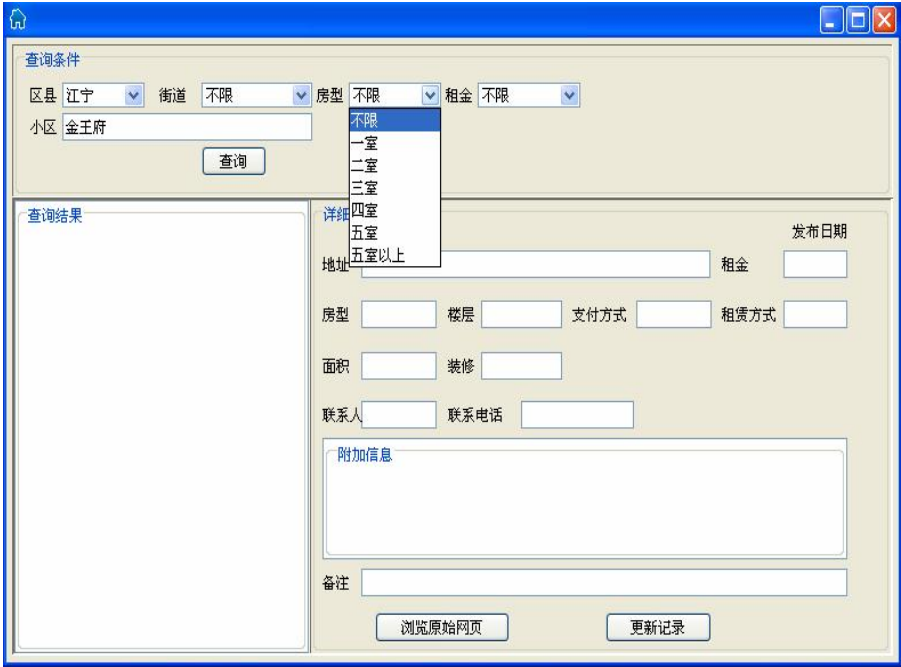


图 2    客户查询对话框

（1）在主界面点击“新建查询”按钮，弹出图 2 所示的对话框，在查询条件栏，根据需要，分别选择填写区县、街道、房型、租金、小区等信息，点击“查询”按钮，一般会在左侧的查询结果列表中出现查询结果，可以點選列表中的出租信息记录，在右侧会显示出该记录的详细信息，结果如图 3 所示。如果查询时提示查不到记录，则可以重新筛选条件，查询需要的信息；如果未发现符合要求的信息，但希望系统能够在以后自动提供新的信息，则可以先保存查询。

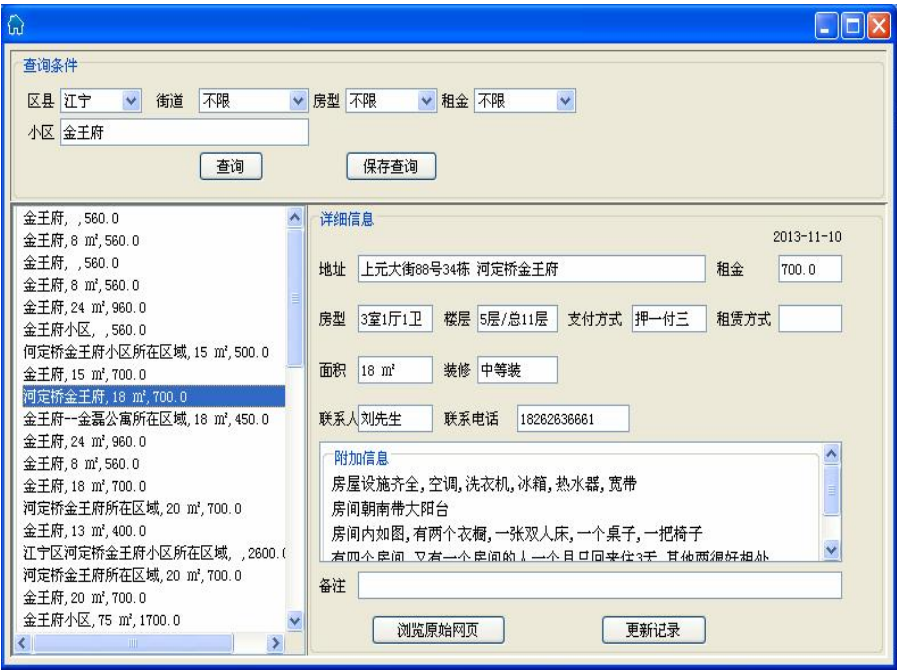


图 3 查询结果

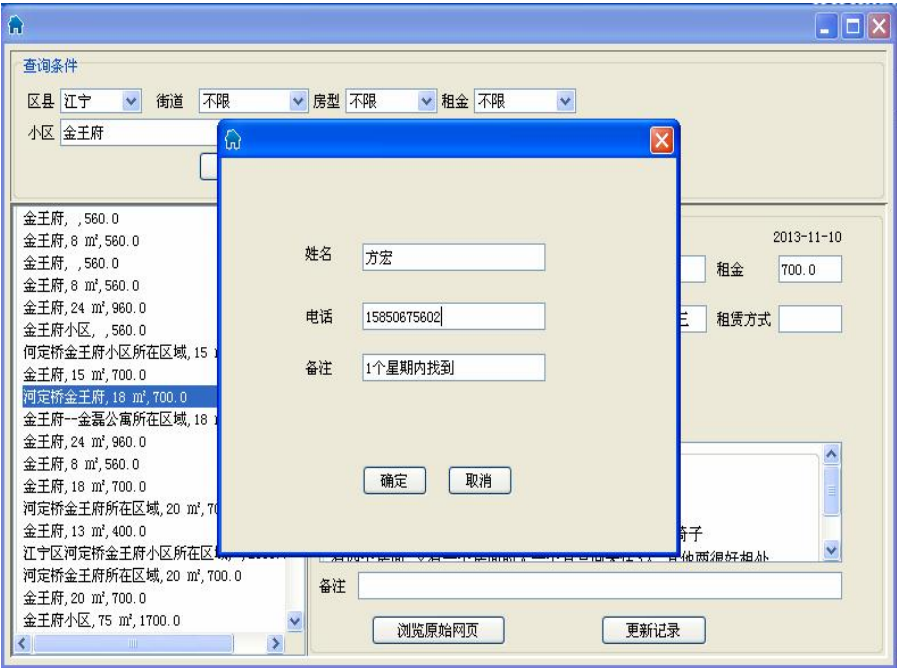


图 4 保存查询对话框

- (2) 保存查询。点击“保存查询”按钮，出现图 4 所示对话框，填入客户的名称、联系电话及备注信息，点“确定”按钮，则系统一般会提示保存成功，同时，系统会在主界面“客户查询”列表中，生成查询记录节点。
- (3) 在主界面浏览最新信息。回到主界面，点击右侧“客户查询”栏中的客户节点（树形结构），会出现刚才保存的查询节点，通过该节点，可以浏览到最新检索到的符合查询要求的记录。（系统会定期更新查询结果），结果如图 5 所示。
- (4) 重复 1-3 步，可以建立多个客户查询，结果如图 6 所示。

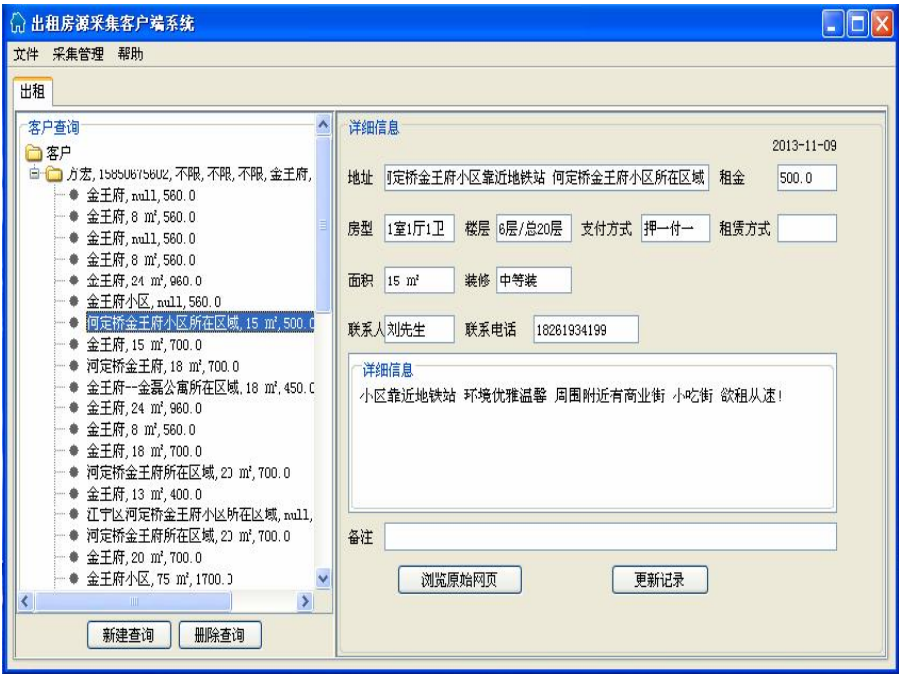


图 5 主界面客户查询列表中的记录

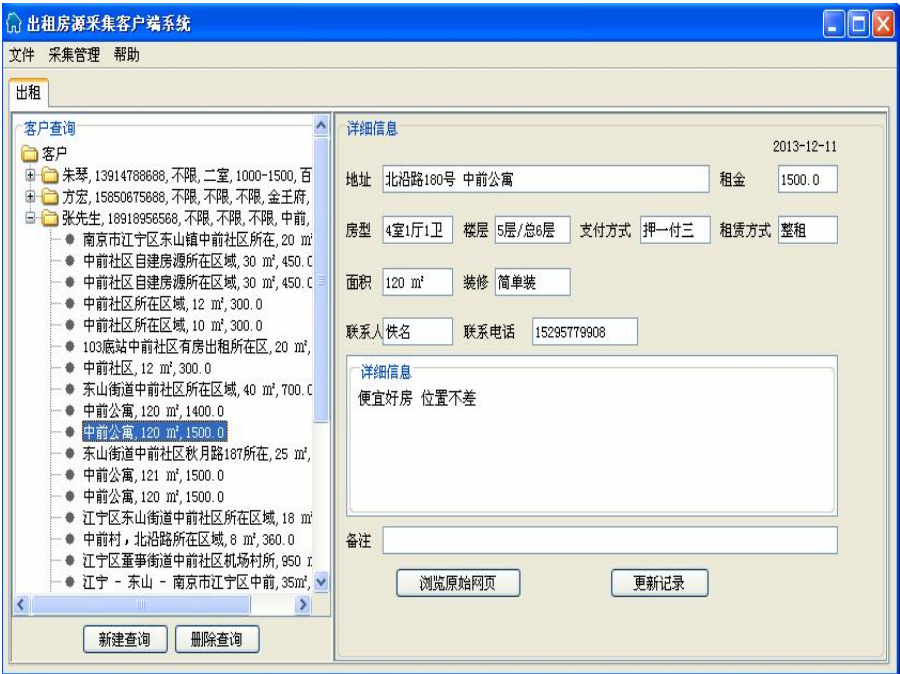


图 6 主界面中的多条客户查询记录

3 修改出租记录信息

出于各种原因需要人工修改补充出租记录信息：如搜索引擎从网页提取的信息未必百分百准确；有些网站的电话号码是图片表示的，程序无法自动识别，需要人工识别输入；与户主联系后，需要修正信息，备注信息等。修改信息可以在主界面实现，也可以在查询界面实现，两者的操作类似，下面以在查询界面修改信息为例，说明如何修改出租记录。

(1) 在查询界面左侧查询记录列表中，点选出租记录，右侧会显示该记录的详细信息，如图 7 所示。

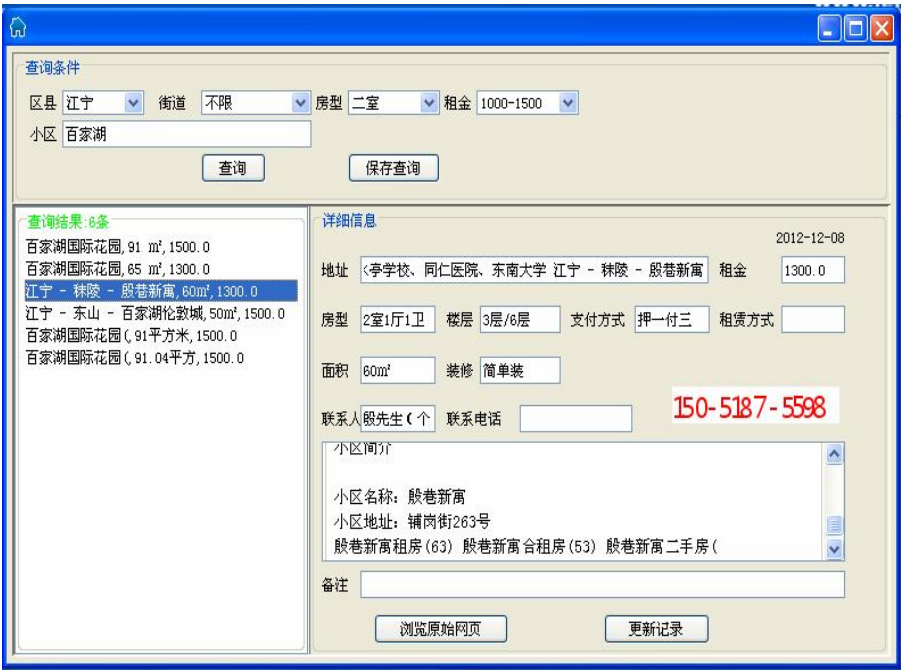


图 7 在查询界面选定修改的记录

(2) 图 7 中需要修改的信息首先是电话号码，该电话号码在源网站上是图片表示的，需要人工识别，并在联系电话输入框输入数字电话号码；另外联系人一栏的信息提取的也不够精确，需要修正；在备注栏，也可以输入一些信息，修改的结果如图 8 所示。

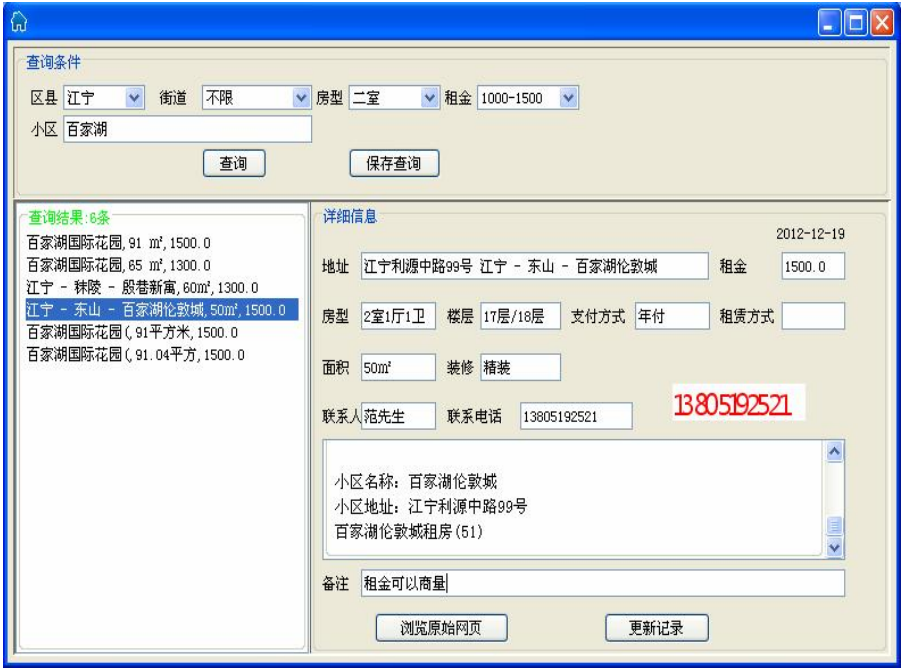


图 8 修改记录

(3) 点“更新记录”按钮，系统会更新记录，一般会提示更新成功。更新成功后，再次选定该记录，可以看到记录的相关信息已做了修改，结果如图 9 所示。注意到电话号码图片已被数字电话号码所取代。

在修正记录信息时，可以点击“浏览原始网页”按钮，系统会自动启动浏览器，加载该记录对应的原始网站网址，通过浏览原始信息，可以更有效的进行人工修正。由于一般网站上显示出租记录信息的时间有限（约 1 个月）以及信息发布者可能删除了信息，所以并不能保证总是能查看到原始信息。



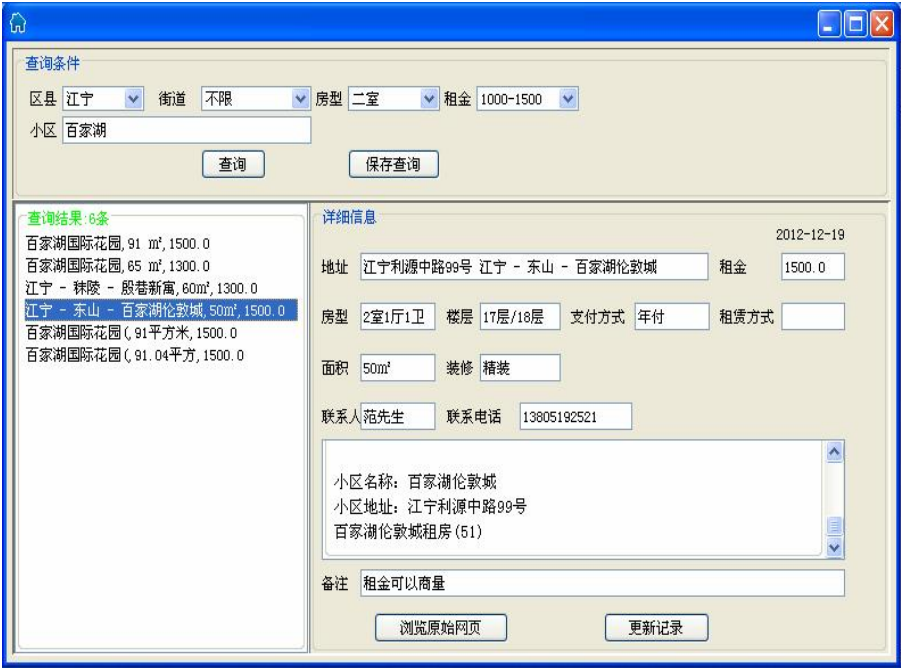


图 9    记录修改后的结果

4 删除客户查询

在主界面客户查询列表中，如果客户查询已查到满意的信息或已失效，就需要将其删除。一方面可以提高系统查询性能，另一方面也利于查询记录的浏览。删除客户查询的步骤如下：

- (1) 在主界面，客户查询列表中，点击选择需要删除的客户查询，然后点击“删除查询”按钮，会弹出确认删除对话框，如图 10 所示
- (2) 在确定删除对话框中点“确定”按钮，系统会“删除”该查询，结果如图 11 所示。实际上并未真的从数据库中删除，只是做了被删除的标记。

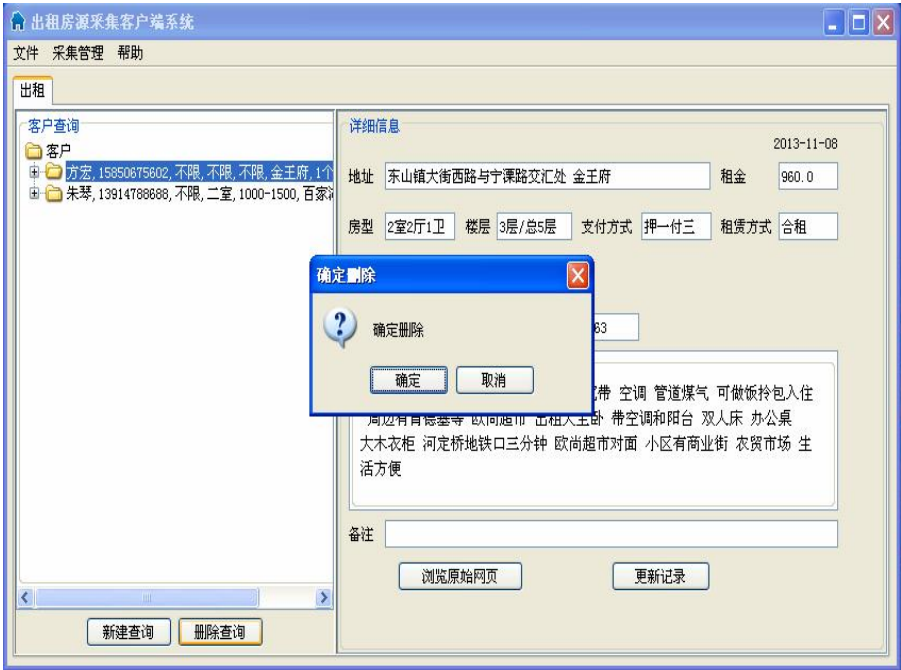


图 10    删除对话框



图 11 查询删除后的主界面

### 5 启动和退出数据采集

在客户端和服务端安装在同一台机器的情况下，可以通过主界面“采集管理”菜单启动出租信息采集服务，如图 12 所示。



图 12 启动采集

如果出现如图 13 所示的情况，则说明网络不通，需要检查、恢复网络的连通性后，再启动数据采集。实际上即使不能联网采集数据，只要能连上数据库，客户端仍能使用，只不过不能查询到最新的出租房信息。

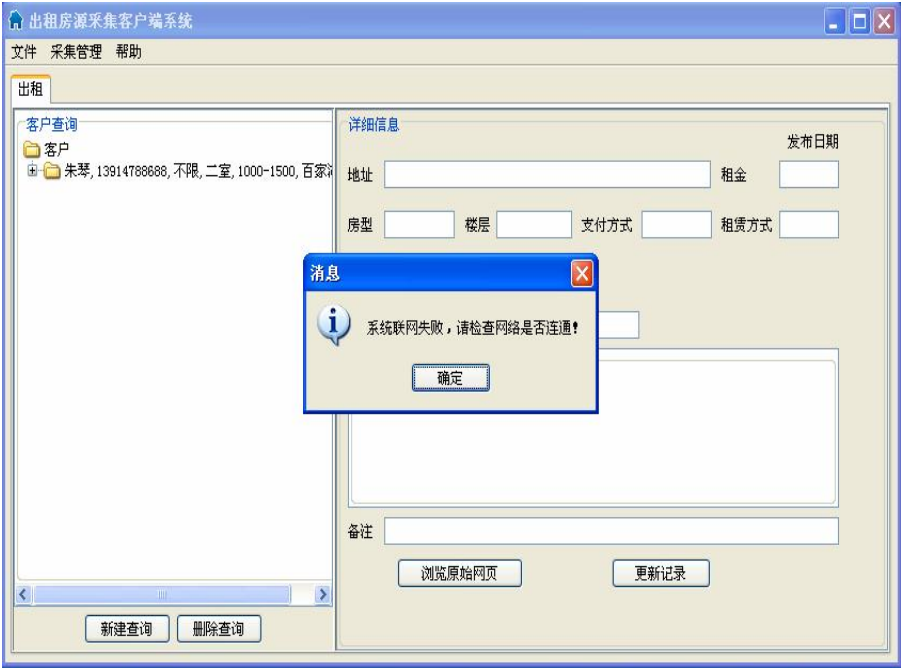


图 13 联网失败对话框

退出采集需要点击主界面菜单文件菜单的退出项，在出现如图 14 所示的退出对话框后，点击确认，则整个系统退出。

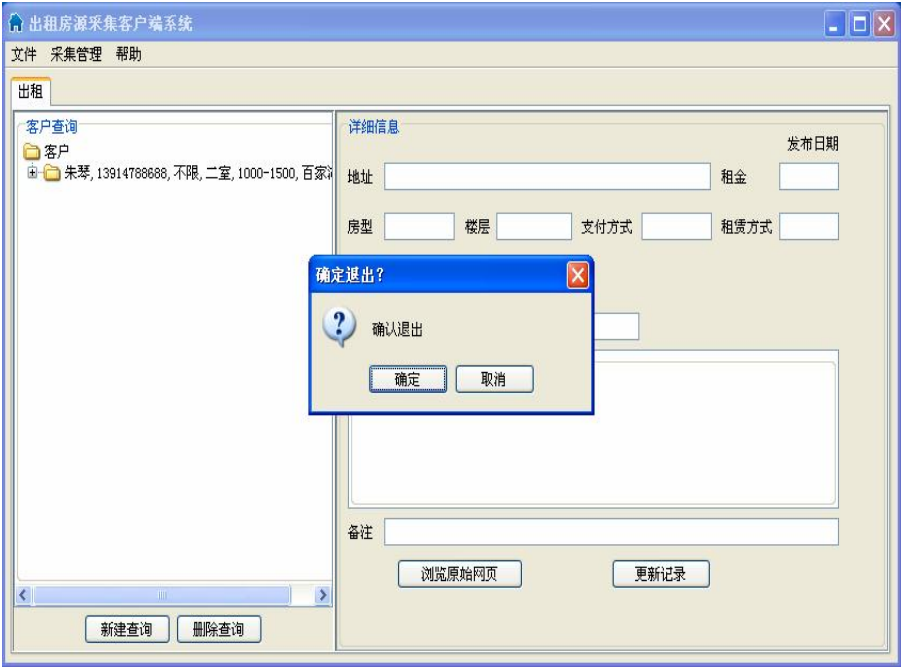


图 14 系统退出对话框

## 6 系统安装配置

出租房源采集客户端系统（搜房宝）的发布形式是一个可执行文件，基本上点击执行，一路按下一步，即可完成安装，但由于系统使用 java 语言开发，程序运行离不开 java 运行时环境，所以需要预先安装 jdk。

### 6.1 JDK 安装配置

#### 1、 下载安装程序



JDK需要1.6以上版本，下载地址如下：

<http://www.oracle.com/technetwork/java/javase/downloads/index.html>

这里使用的是 jdk-6u22-windows-i586.exe Window 安装版本。

## 2、进行安装

双击安装程序进行安装，初始安装界面如图 15 所示，以后按照提示，一路点“下一步”按钮即可。其中安装路径使用默认路径。



图 15 JDK 安装界面

## 3、配置path和classpath环境变量

对java程序进行编译和运行，主要用到两个java处理程序：javac.exe(用于编译java源程序)、java.exe(用于运行编译好的.class文件)，其所处位置如下，也就是jdk安装目录下的\bin子目录下面，如图16所示：



图 16 JDK 安装 bin 路径下文件

安装完毕jdk后，java.exe和javac.exe还不是系统内部处理命令，打开命令界面（Win+R在弹出对话框中输入cmd然后回车就进入命令界面，如图17、18所示）



图17 运行命令对话框

在命令界面下输入javac命令，发现不是内部命令，即无法进行java源程序编译。



图18 命令窗口

下面要进行path和classpath配置，使得javac命令能在命令界面下运行。依次点击：“我的电脑”—(右键)“属性”（图19）—“高级”—“环境变量”，弹出Windows的“环境变量”配置窗口（图20所示）。

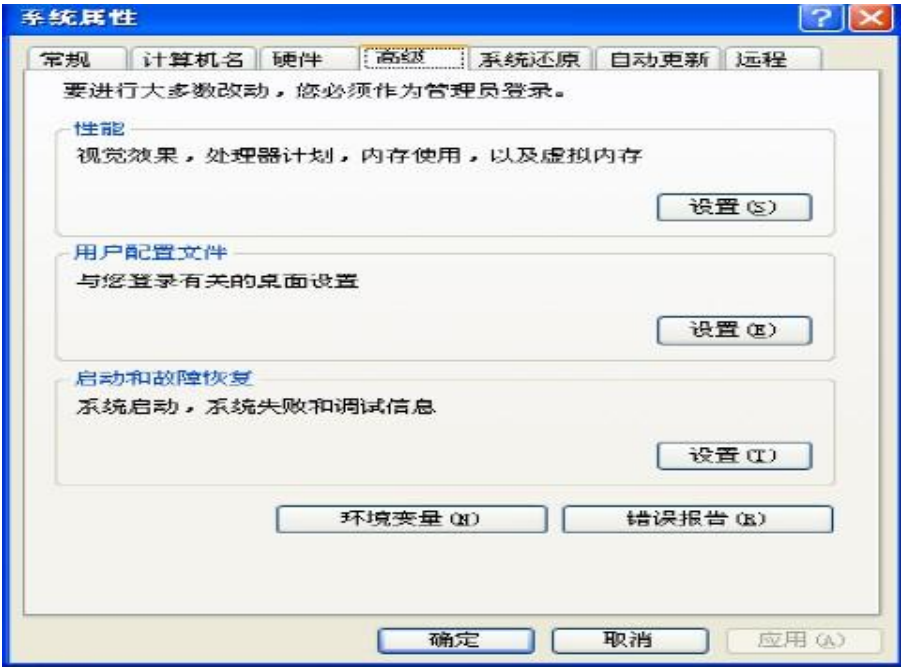


图19 系统属性窗口

在“系统变量”栏下执行三项操作：



图20 环境变量设置窗口

一是新建“Java\_Home”，设置其值为 JDK所在的绝对路径(例如 D:\Program Files\Java\jdk1.6.0\_29)。二是新建“Classpath”(如果已有，则直接编辑)，设置其值为 .;%Java\_Home%\lib (若值中原来有内容，用分号与之隔开)。注意路径前的符号为. ;不能漏掉。三是新建“Path”(如果已有，则直接编辑), 值: %Java\_Home%\bin; (若值中原来有内容，用分号与之隔开)。重新打开一个字符命令界面发现javac命令可用了(图21所示)。

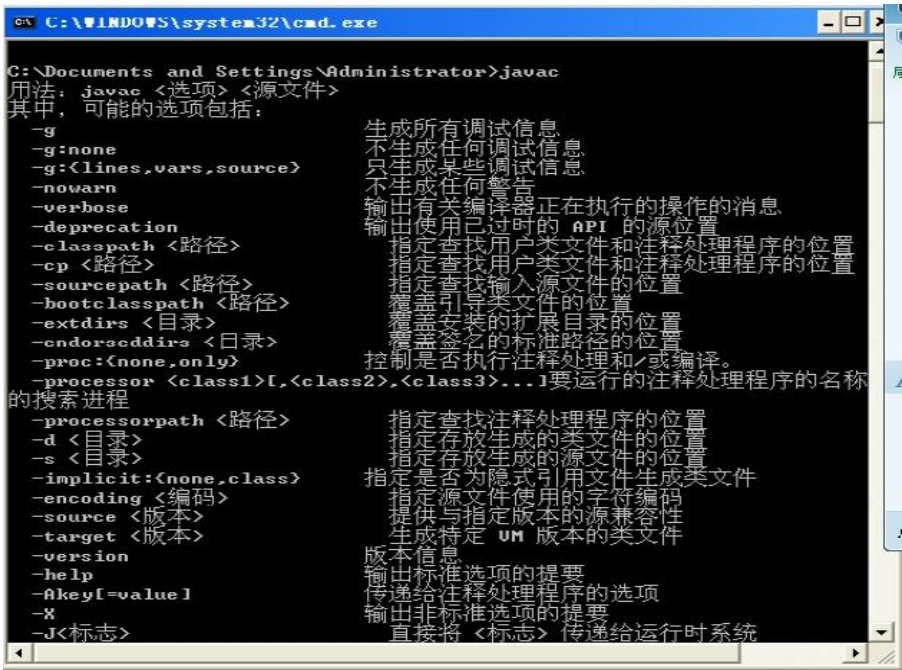


图21 命令窗口执行 javac

## 6.2 应用程序配置

系统安装完毕后，典型的目录结构如图 22 所示，一般安装完成后，会在桌面和开始程序栏有启动程序的图标，也可以到安装目录下，双击执行“搜房宝.exe”程序。一般情况下均能正常执行，无需再配置参数。但对于一些高级客户，需要根据自身情况定制程序，下面介绍系统提供的可供配置的参数，主要有数据库配置和查询条件配置。

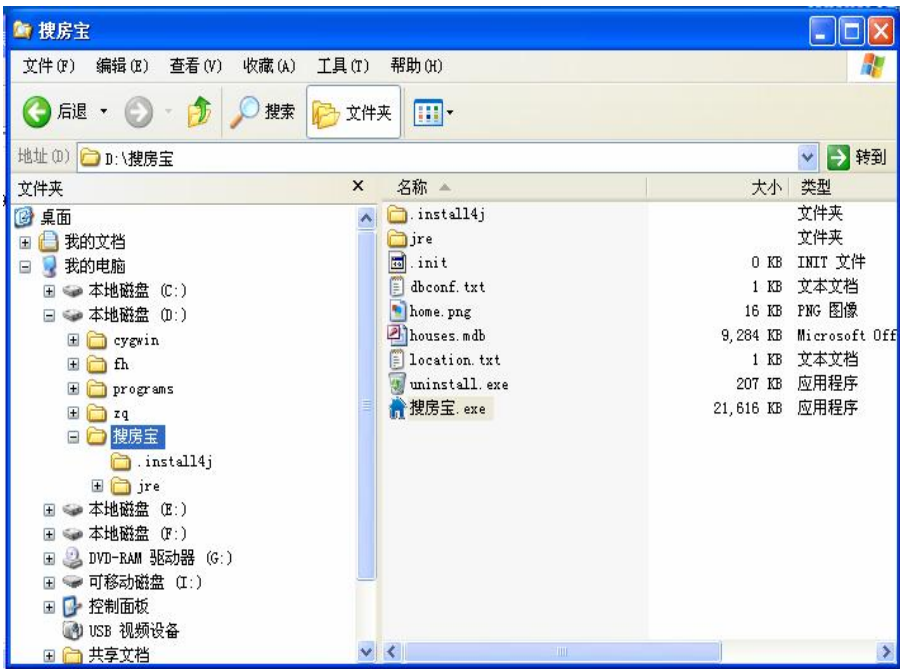


图 22 程序安装的目录结构

### 1、数据库配置

用文本编辑器打开安装路径下的 dbconf.txt 文件，默认情况下，配置的是本地的 access 数据库（如图 23 所示）。如果需要更改数据库，如改为 oracle 的，则需要修改相应的参数：driver、url、密码和用户名。如图 24 所示。相

应的数据库表也需要从 access 导入到 oracle。

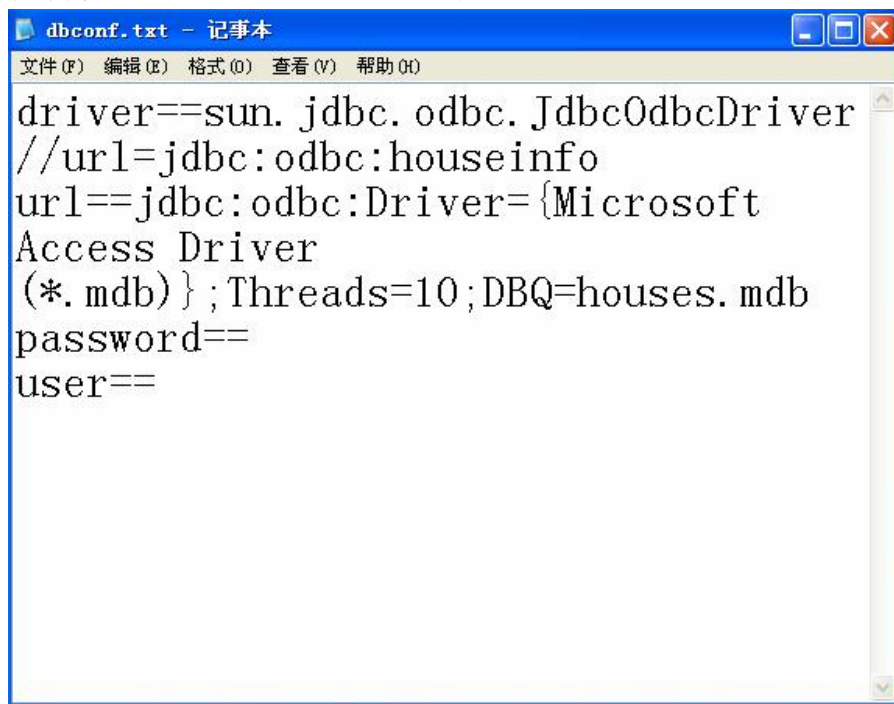


图 23 默认配置数据库参数

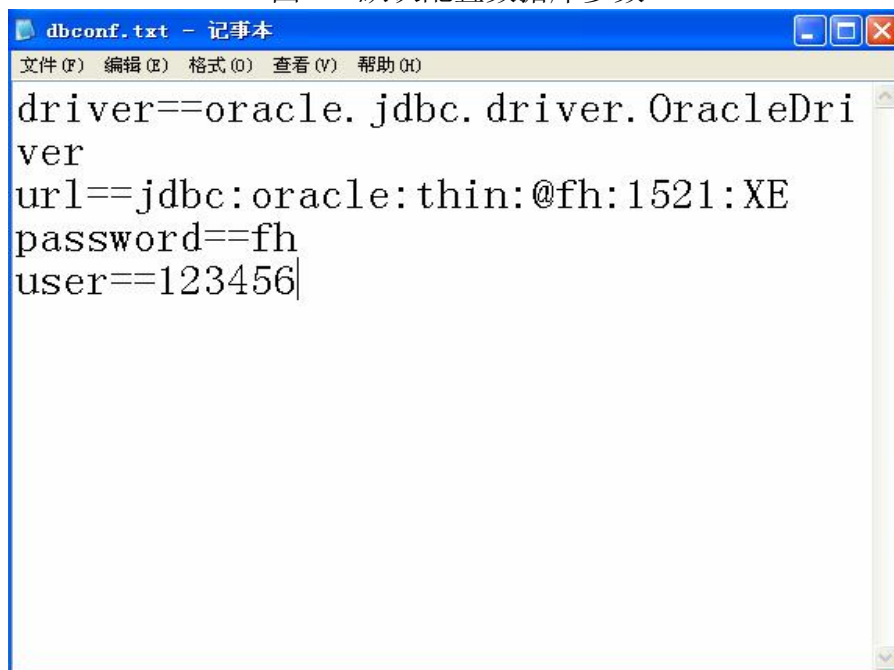


图 24 oracle 配置数据库参数

## 2、查询参数配置

查询参数主要是指区县、街道、房型、租金、时间参数，用文本编辑器打开 location.txt, 如图 25 所示。



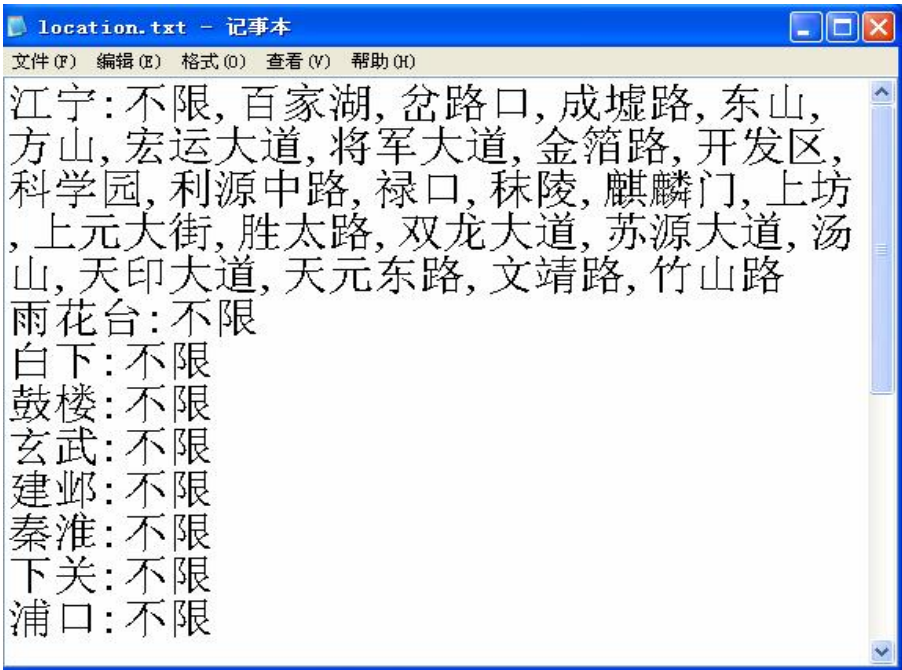


图 25 查询参数

南京市的具体数据如下：

江宁:不限, 百家湖, 岔路口, 成墟路, 东山, 方山, 宏运大道, 将军大道, 金箔路, 开发区, 科学园, 利源中路, 禄口, 秣陵, 麒麟门, 上坊, 上元大街, 胜太路, 双龙大道, 苏源大道, 汤山, 天印大道, 天元东路, 文靖路, 竹山路

雨花台:不限

白下:不限

鼓楼:不限

玄武:不限

建邺:不限

秦淮:不限

下关:不限

浦口:不限

栖霞:不限

六合:不限

溧水:不限

高淳:不限

不限:

租 金 : 不 限 , 500 以  
下, 500-999, 1000-1500, 1500-2000, 2000-3000, 3000-4000, 4000-5000, 5000 以上

户型:不限, 一室, 二室, 三室, 四室, 五室, 五室以上

时间:1 天, 3 天, 7 天, 15 天, 30 天, 60 天。

第 1 部分是区县及相应的街道设置，第 2 部分是租金设置，第 3 部分是户型设置，第 4 部分是时间设置，用于设置查询返回的时间区间，如设置 60 天，则查询返回的是最近 60 天的数据，默认是 30 天。

上述参数设置对应的基本是查询界面中的相应下拉选择项目，程序初始执行时会使用上述数据初始化选择项。

## 7 常见问题

### (1) 为何无法执行程序？

系统安装后，可以通过双击桌面图标来执行程序，如果程序不能执行，可到程序安装路径下，双击“搜房宝.exe”来执行，如果还不能执行，则可能有以下原因：一是Java运行时环境，JDK安装、配置不正确，请参考“系统安装配置”的JDK安装配置部分，正确安装、配置JDK。二是程序使用期限已到，为了保护程序的合法权益，该程序采用了若干安全保护措施，到期程序将无法运行，会弹出如图26所示的对话框。此时若想继续使用，请联系软件开发者，联系方式：[1010402827@qq.com](mailto:1010402827@qq.com)。

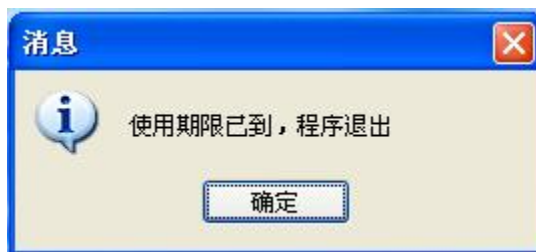


图26 程序到期对话框

三是程序需要的参数遭到了不正确的改动，或丢失。此时可以重新安装程序。

### (2) 为何程序启动后，原先保存的客户查询没有了？

该情况一般出现在数据库存取异常时，程序的主界面如图27所示，而且如果执行查询，会出现如图28所示的对话框。



图27 丢失客户查询时的主界面

这通常是因为数据库参数被改动，改动后导致程序无法正确存取数据库，此时可以重新安装程序。注意在之前要将数据库保存好，安装后，再恢复原先保存的数据库。

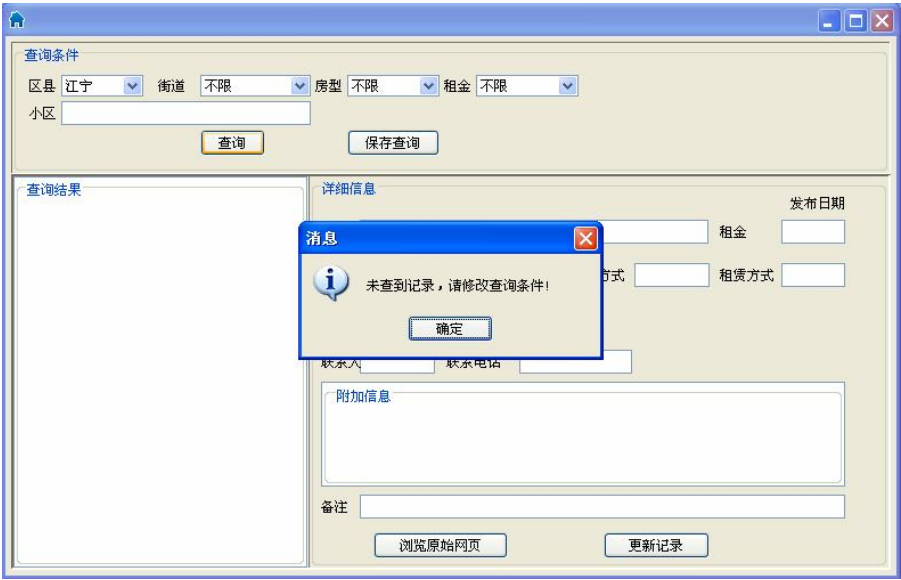


图28 无法查到记录（数据库存取失败）

（3）为何程序不能检索到最新的信息？

出租房源采集客户端是从数据库中检索数据的，而数据库中的数据是通过数据采集服务程序从网站采集的，所以如果出现数据采集服务没有开启、停止运行、不能存取网络、目标网站无法访问等情况，都会导致无法获得最新的信息。此时首先需要确认数据采集服务是否正常运行、网络是否连通、人工访问地产网站是否成功，若排除了以上情况，仍不能解决问题，请与开发者联系。

（4）为何图片电话号码不能自动识别？

地产网站采用图片电话号码，就是为了防止搜索引擎类的程序自动获取用户电话号码，虽然理论上可以借助人工智能技术，从图片中识别出电话号码，但实际中准确识别出11位号码的难度相当大，所以目前只能截取电话号码的图片，然后由人工来识别。

（5）出租房信息来自哪些网站？

在不同地区，系统采集的地产网站是不同的，主要是根据google、baidu搜索到的，排在前2页的地产信息网站，因为即使是人工搜索，一般也就是查看排在第一页的地产网站，且一般不会超过5家。由于不同地区，存在不同的本地地产网站，所以这种方式可以做到检索到最适合本地用户的出租房信息。当然全国性的网站，如58同城、赶集网，一般都会检索到的。

（6）为何有些出租房信息缺失、不准确？

采集到的信息不全、不准确是由多种原因造成的，最常见的出现在地产网站改版期间，由于其展示数据的网址、网页发生了较大变化，导致信息抓取失败，或只能抓取到部分信息，但这种情况一般不会持续太长时间，我们会及时发现，并做针对性的处理，保证及时准确的提供数据；还有些网站上的出租信息本身就是不全、甚至是错误的，这个主要取决于信息的发布者，系统只是忠实的照原样抓取数据，如图29所示，一条信息的“房型”一栏为空，这通常是因为发布数据时忘记输入房型数据，而网站又没有很好地对数据校验的结果；当然，偶尔也可能出现未能抓取到部分信息的情况，如图30所示，其中的“联系人”和“联系电话”都未能提取到，此时可以点击“浏览原始网页”按钮，直接查看在网站上的原始信息，并手工补齐相关信息。我们一直在努力改进程序，尽量做到把出租信息抓全、抓准，上述情况将会越来越少。

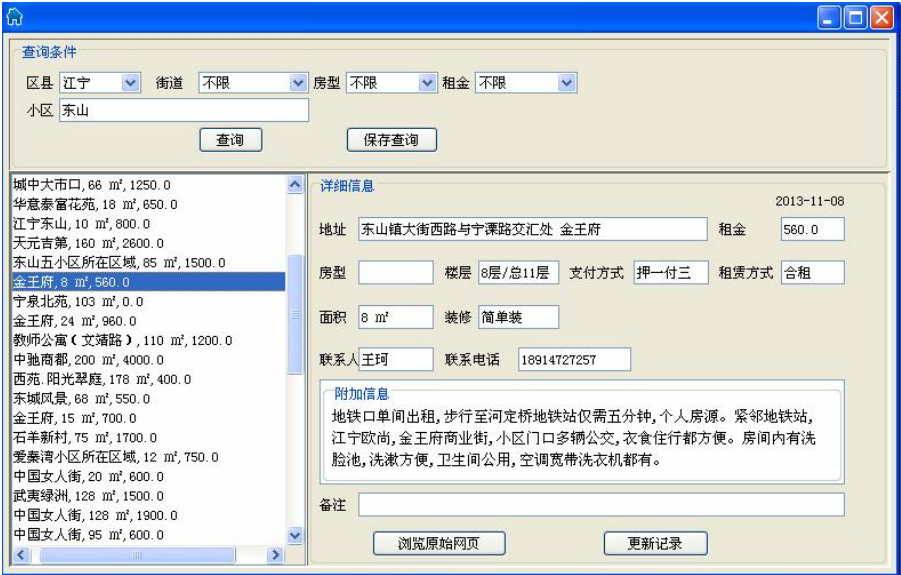


图29 房型一栏为空

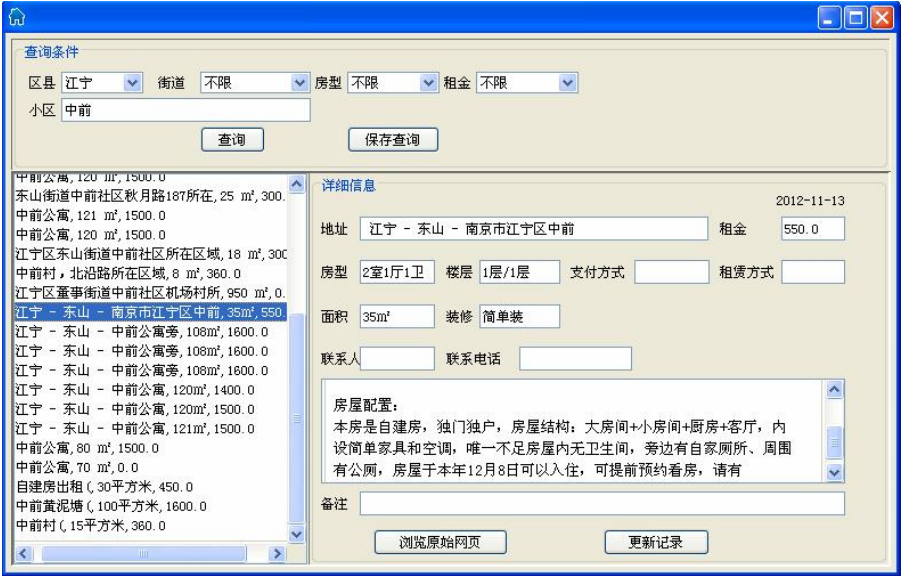


图30 联系人、联系电话为空

(7) 检索不到符合要求的信息，该怎么办？

检索不到需要的信息，原因有很多，常见的有以下两种情况：一种是需要的信息目前确实不存在，但不代表以后没有，此时可以将查询保存，系统会自动查询最新信息，及时反馈给客户；一种是查询条件选择不当，查询条件过于苛刻，导致查不到数据，如图31所示。此时可以将“街道”内容改为“不限”，在“小区”中输入“岔路口”，该项可以进行模糊匹配，这样就有可能找到相关信息，结果如图32所示。这是查询技巧的问题，相信用的时间长了，就会积累出一定的使用经验。

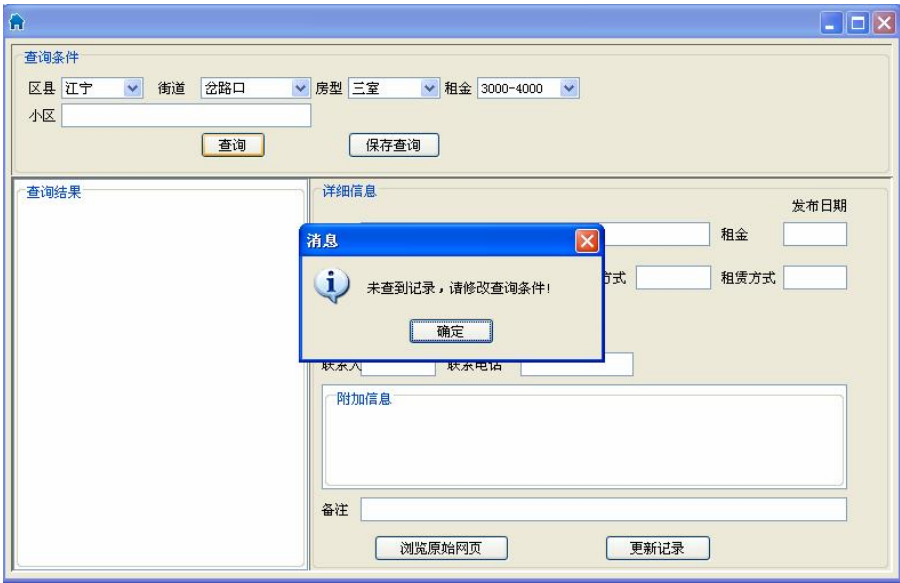


图31 查不到记录的情况



图32 查到记录的情况

- (8) 为何点击“删除查询”按钮，仍不能删除客户查询？  
删除查询前必须选定要删除的客户查询，否则无法执行删除。
- (9) 作为中介，在帮助客户查询数据时，希望一些关键信息如租金、联系电话不能给客户看到，但软件把这些信息都显示出来了，有无办法解决？  
这个可以在显示时将上述信息屏蔽掉，同时为了保证中介能看到上述信息，可以设置某个特殊按键，按键以后，再显示全部信息，具体需要定制程序，请与开发者联系。