OcrKing 配合火车采集器简易说明

应管理 真视觉 之邀请，特做此简单说明文档，时间比较仓促，不详之处请多多见谅！首先请在本地或远程都搭建好 php 环境（推荐 xampp 或 wamp）,安装即可用，并下载 OcrKing.php 脚本，请用文本编译器 notepad++或系统记事本打开 OcrKing.php，文件内会有以下内容，请获取 key 后替换下面的 你的 apiKey 并保存脚本

```
//本识别服务为免费，没有 apiKey 可以 key 为标题
// 任意内容为正文发邮件到 ok@ocrking.com 获取
//可能会有延迟，请匆重复发送
//授权 apiKey,请注意区分大小写
define ( 'API_KEY', '你的 apiKey' );
```

将保存好的脚本上传到你的 php 空间，然后访问

http://domain.com/OcrKing.php?u=http://www.e-fa.cn/extend/image.php?auth=EXcQdVFjIkcOPxd1DW4wAF5iNAUJPQ

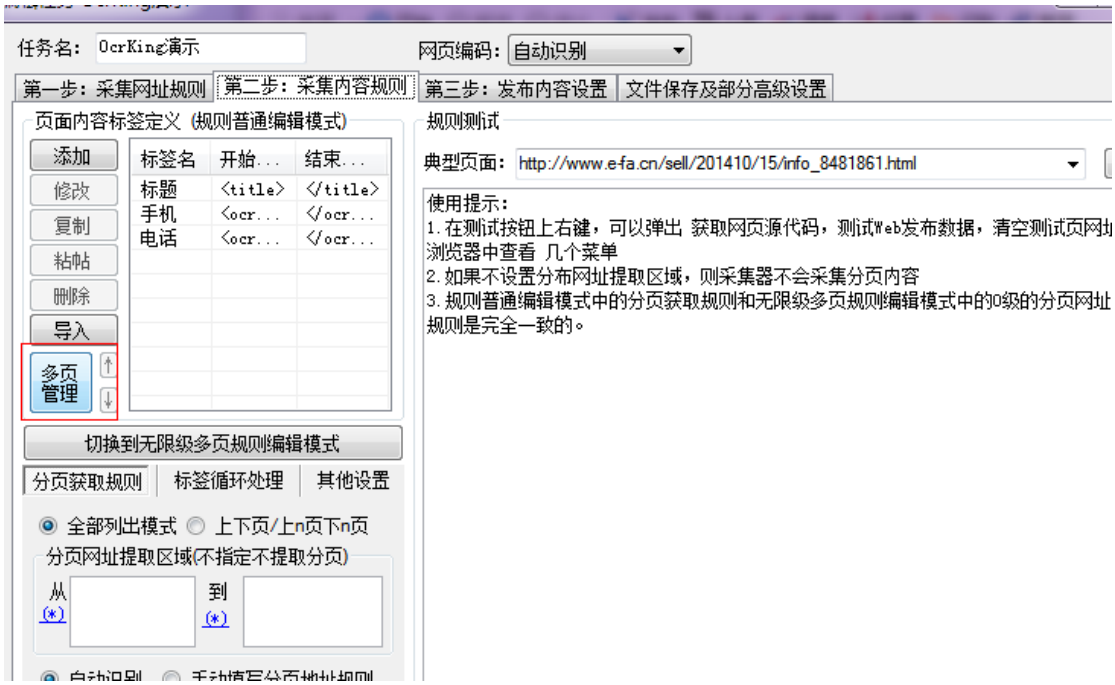
如果返回结果为 15201124022 就说明成功了

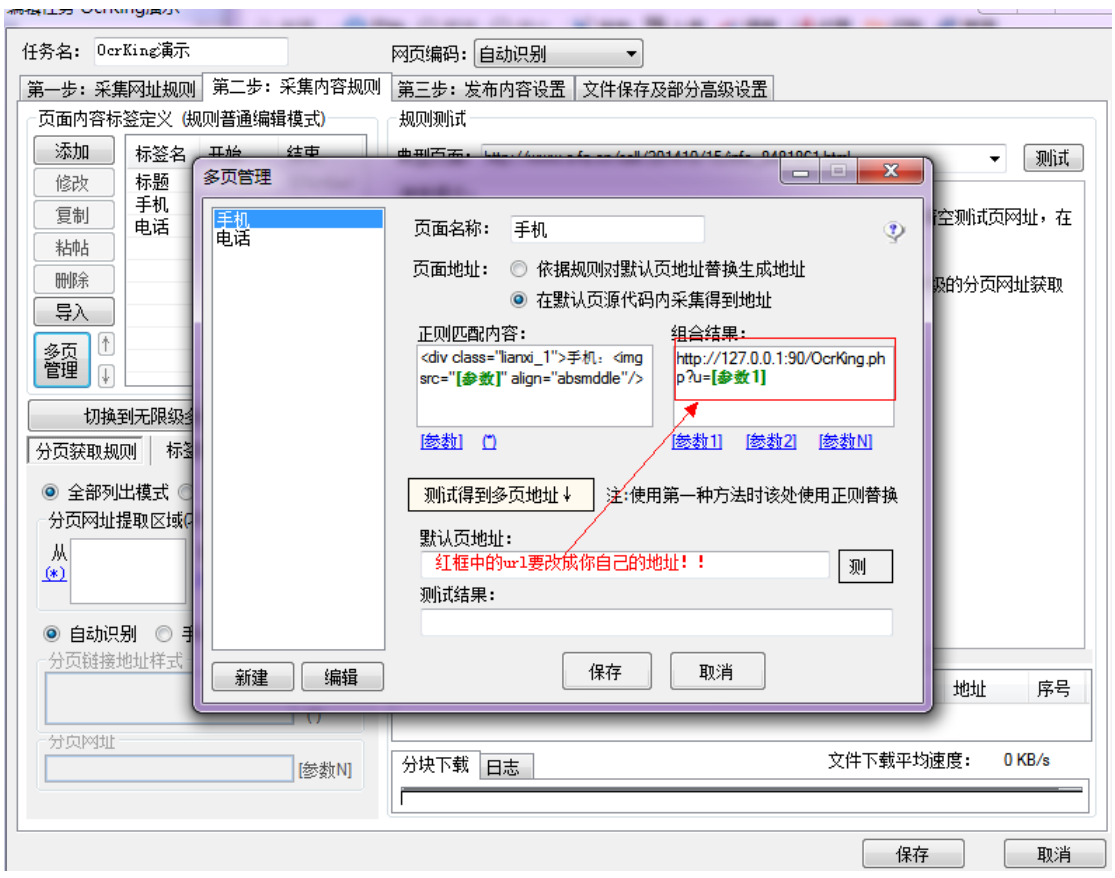
配合火车,适合各个版本的火车采集器，主要利用了多页功能进行识别

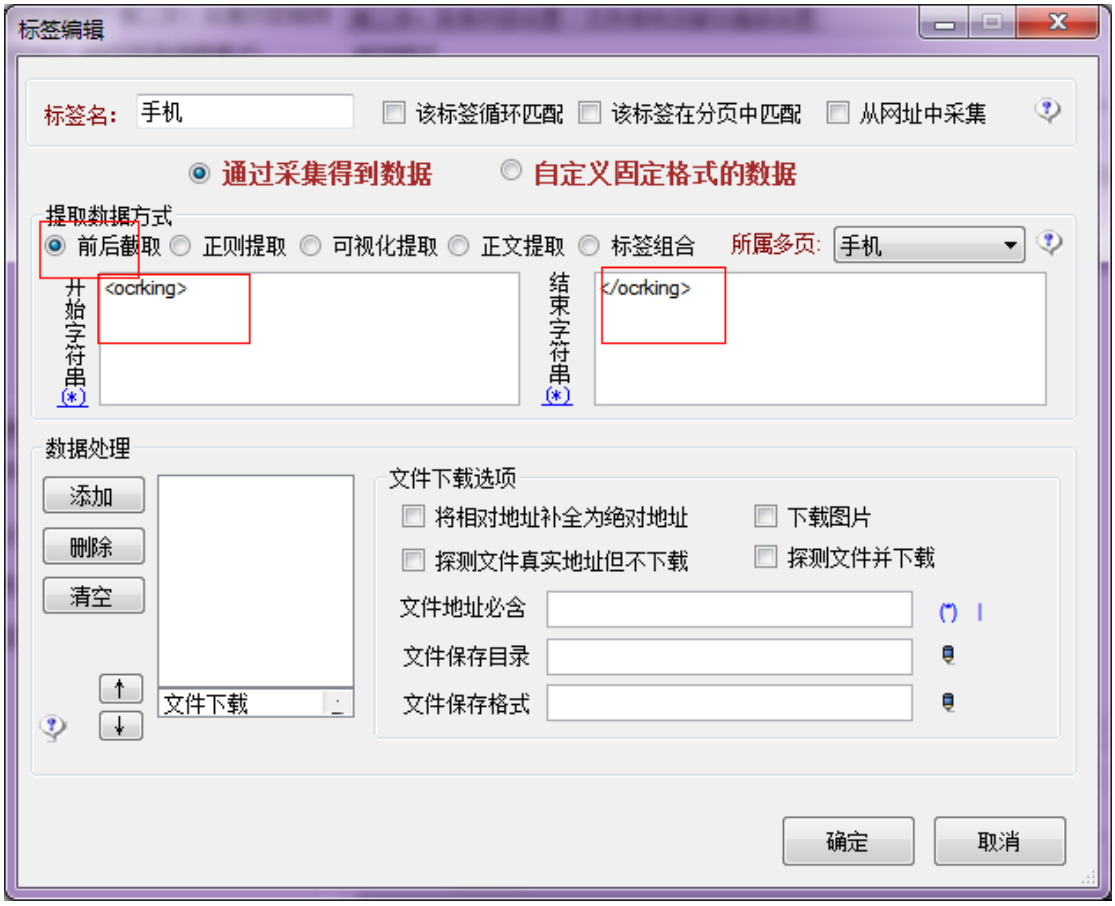
请下载 OcrKing 演示（v8 规则，其它版本类似） 规则导入，查看多页设置仿照设置即可，如要采集的内容如下图

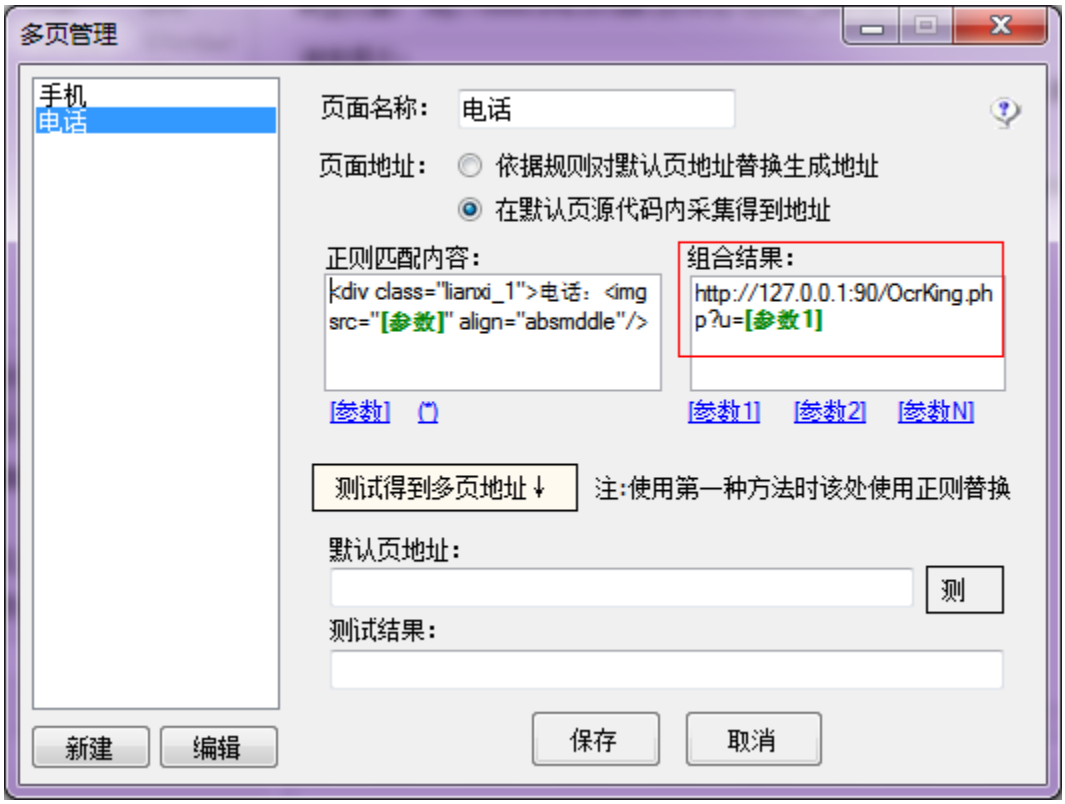
手机电话图片代码为

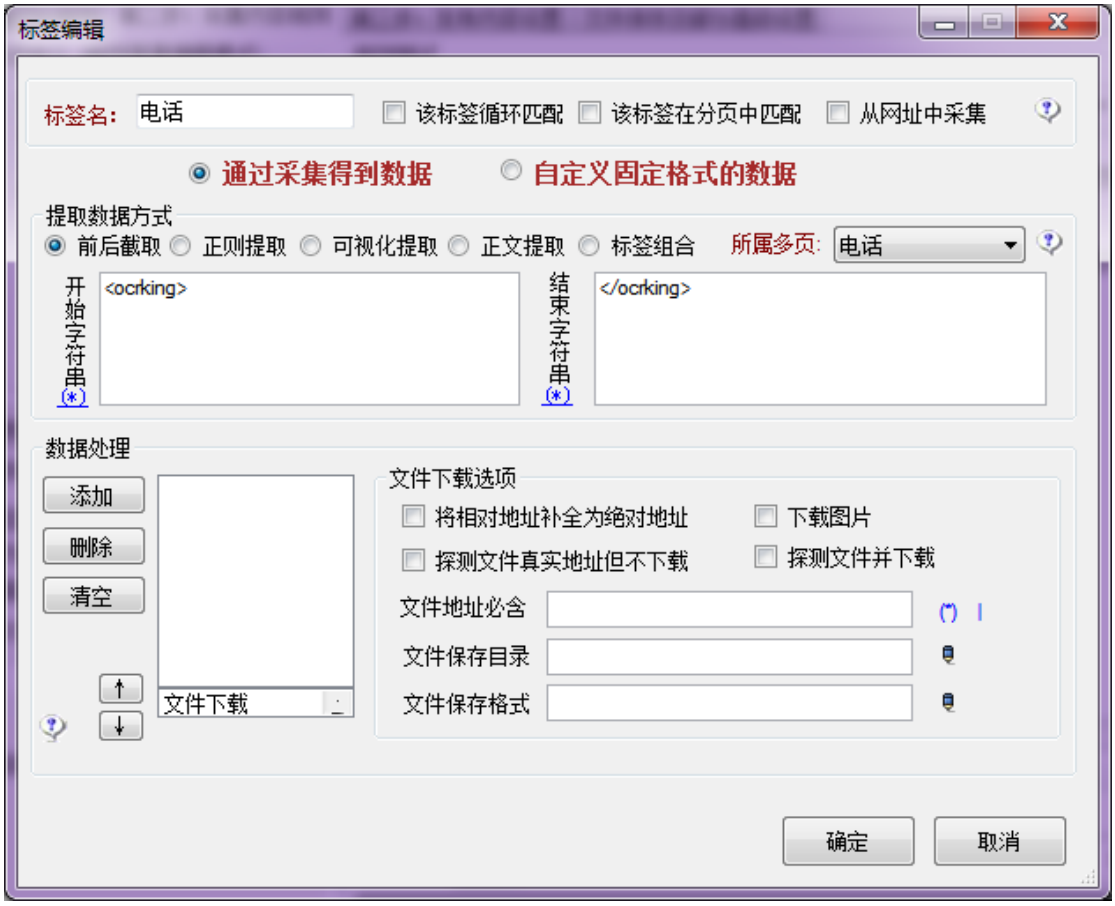
```
手机：<img src="http://www.e-fa.cn/extend/image.php?auth=EXcQdVFjIkcOPxd1DW4wAF5iNAUJPQ" align="absmddle"/>
电话：<img src="http://www.e-fa.cn/extend/image.php?auth=LkksTXtLE2syByZDUTATJH1EGSNDdkQo" align="absmddle"/>
```

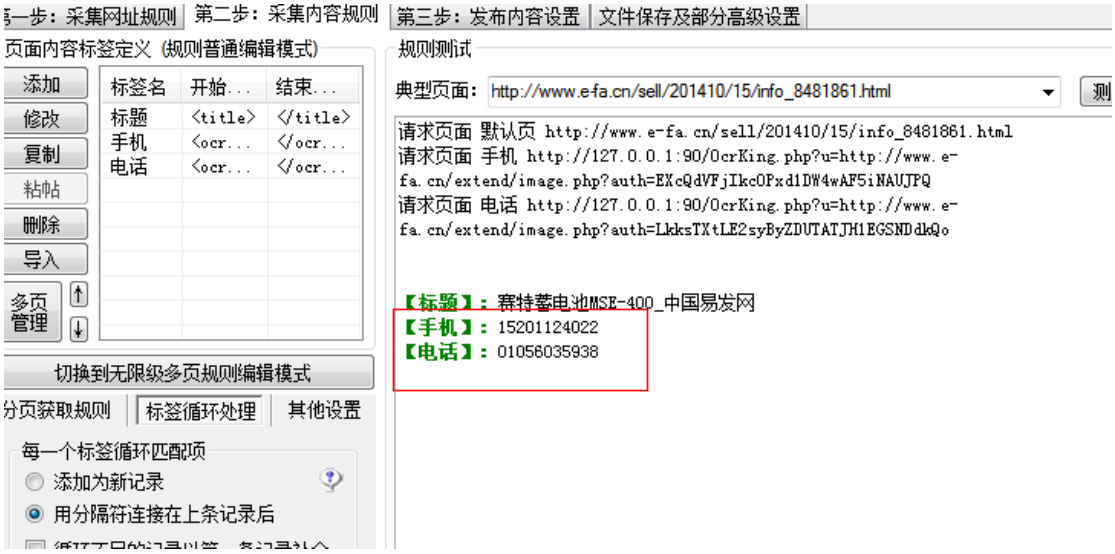

任务名：
OcrKing演示
网页编码：
自动识别
第一步：采集网址规则
第二步：采集内容规则
第三步：发布内容设置
文件保存及部分高级设置
页面内容标签定义（规则普通编辑模式）
添加
修改
复制
粘帖
删除
导入
多页管理
标签名
开始...
结束...
标题
<title>
</title>
手机
<ocr...
</ocr...
电话
<ocr...
</ocr...
切换到无限级多页规则编辑模式
分页获取规则
标签循环处理
其他设置
全部列出模式
上下页/上n页下n页
分页网址提取区域(不指定不提取分页)
从
(*)
到
(*)
规则测试
典型页面：
http://www.e-fa.cn/sell/201410/15/info_8481861.html
使用提示：
1. 在测试按钮上右键，可以弹出 获取网页源代码，测试Web发布数据，清空测试页网址
浏览器中查看 几个菜单
2. 如果不设置分布网址提取区域，则采集器不会采集分页内容
3. 规则普通编辑模式中的分页获取规则和无限级多页规则编辑模式中的0级的分页网址
规则是完全一致的。


任务名：
OcrKing演示
网页编码：
自动识别
第一步：采集网址规则
第二步：采集内容规则
第三步：发布内容设置
文件保存及部分高级设置
页面内容标签定义（规则普通编辑模式）
规则测试
测试
多页管理
手机
电话
页面名称：
手机
页面地址：
依据规则对默认页地址替换生成地址
在默认页源代码内采集得到地址
正则匹配内容：
手机：
组合结果：
http://127.0.0.1:90/OcrKing.php?u=[参数1]
[参数]
[参数1]
[参数2]
[参数N]
测试得到多页地址↓
注:使用第一种方法时该处使用正则替换
默认页地址：
红框中的url要改成你自己的地址！！
测
测试结果：
新建
编辑
保存
取消
地址
序号
分块下载
日志
文件下载平均速度：
0 KB/s
保存
取消


标签编辑
标签名：
手机
该标签循环匹配
该标签在分页中匹配
从网址中采集
通过采集得到数据
自定义固定格式的数据
提取数据方式
前后截取
正则提取
可视化提取
正文提取
标签组合
所属多页:
手机
开始字符串(*)
<ocrking>
结束字符串(*)
</ocrking>
数据处理
添加
删除
清空
文件下载
文件下载选项
将相对地址补全为绝对地址
下载图片
探测文件真实地址但不下载
探测文件并下载
文件地址必含
文件保存目录
文件保存格式
确定
取消


多页管理
手机
电话
页面名称：
电话
页面地址：
依据规则对默认页地址替换生成地址
在默认页源代码内采集得到地址
正则匹配内容：
电话：
组合结果：
http://127.0.0.1:90/OcrKing.php?u=[参数1]
[参数]
[参数1]
[参数2]
[参数N]
测试得到多页地址↓
注:使用第一种方法时该处使用正则替换
默认页地址：
测
测试结果：
新建
编辑
保存
取消





手机：15201124022
电话：010-56035938
所在地：北京