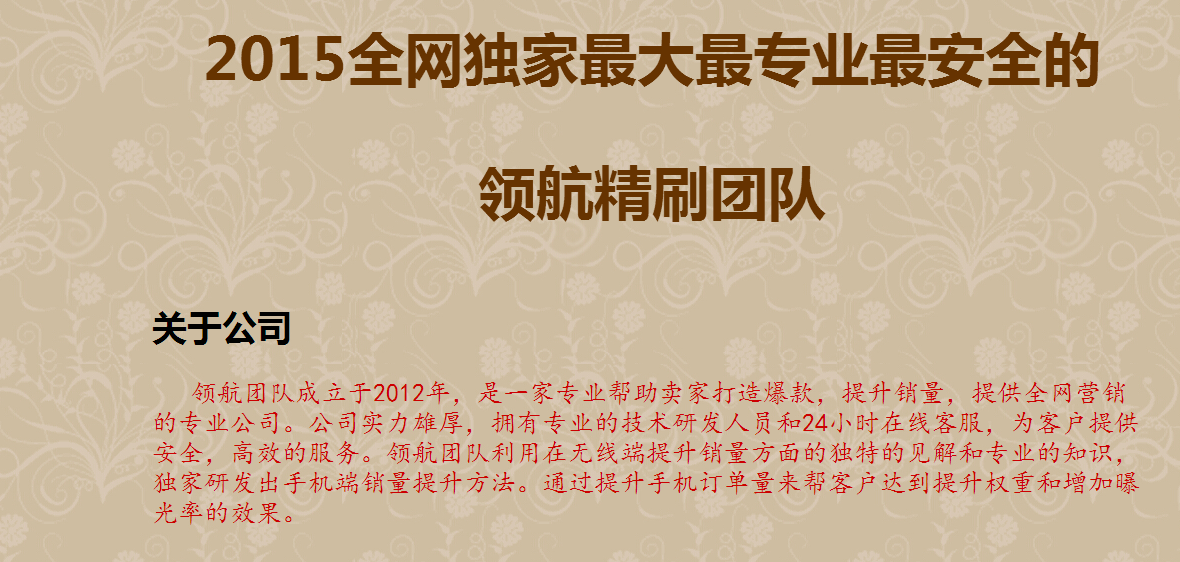
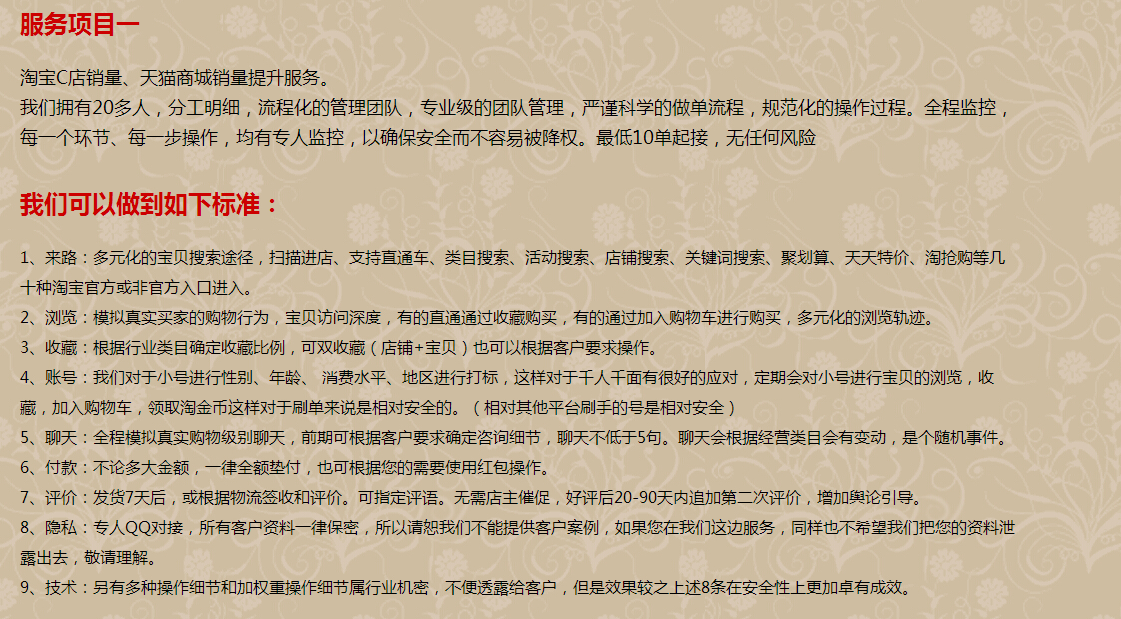
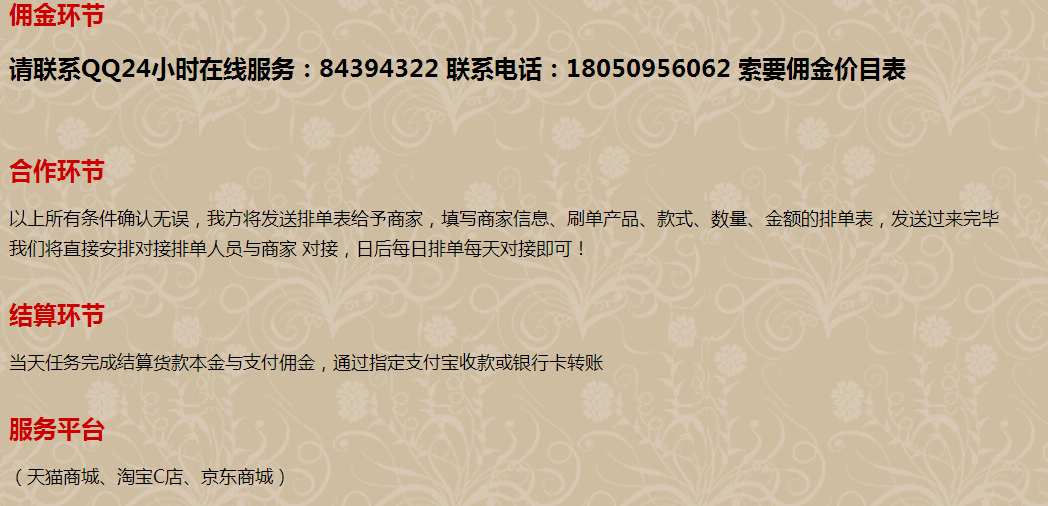
1、欺骗性域名

<http://www.jjrtvu.com/lh/a/>是爬到的一个精刷团队的宣传主页，如下图：







但是其对应的二级域名：[www.jjrtvu.com](http://www.jjrtvu.com)却是晋江广播电视大学的主页：



这种情况如何判断？

2、脏数据

爬到的url有的是不正确的形式，如：http://空包代发[www.5180kongbao.com](http://www.5180kongbao.com)

但是却包含有效的网站，需要处理一下

解决办法：爬网页的link的时候不要处理，爬完之后用正则提取有效url

3.有的网站URL为IP地址的形式：

<http://123.56.103.156/>对应的是易空包网，里面有交易号、交易金额、交易时间,可以作为一些参考。

但是通过网站首页显示的域名[www.eKongbao.cn](http://www.eKongbao.cn) 访问的却是另外一个空包模板网站可以考虑其他黑灰产网站是否都使用相似的模板。例如刷单网站模板http://www.topaaa.com





