淘宝数据抓取概要设计说明书

# 引言

根据对方的需求，我们将从淘宝网上抓取指定格式的数据，并将其保存在excel文件中。利用httpclient来完成登录和抓取网页，利用jsoup来解析html dom树。关于数据抓取基本的思想就是先用get请求获得该页面，再用jsoup来解析dom树，从而获得我们想要的数据。

# 总体设计

系统架构图如下所示：

系统主要由3个模块构成，分别是登陆认证模块，页面解析模块和excel输出模块。

1. 登陆认证模块主要完成系统的自动登陆，获得服务器端向客户端写入的cookie信息，以后的每次http请求中都将携带该cookie信息；
2. 页面解析模块主要是完成对各页面的解析，获得我们需要的数据；
3. Excel输出模块主要是将我们获得的数据写入excel文件中；

主要模块的处理流程图为：



## AutoLogin Module

### 模块描述

实现系统的自动登录功能，获得服务器端向客户端写入的cookie信息，以便之后的请求能够携带该cookie信息；

### 输入

用户认证信息和登录页面post表单信息

### 输出

服务器端向客户端写入的cookie信息

### 处理

利用httpclient完成淘宝网的自动登录，登录验证通过后，将相应的cookie信息写入httpclient.cookiestore，以后的http请求都将携带此cookie信息；

## CrossDomainCookie Module

### 模块描述

跨域获得taobao的cookie信息；

### 输入

无

### 输出

无

### 处理

## TopTenPageParser Module

### 模块描述

解析淘宝网top10页面，获得我们想要的数据

### 输入

各类产品的Top10 page url即CategoryInfo.Href

### 输出

List<TopTenItemInfo>

### 处理

TopTenPage页面是显示需要抓取的某一个类别的前10位的产品 ，根据传入的CategoryInfo.Href获得页面，利用jsoup解析得到List<TopTenItemInfo>

## SearchResultPageParser Module

### 模块描述

解析淘宝网Search Result页面，获得我们想要的数据

### 输入

TopTenItemInfo.href

### 输出

List<SellerInSearchResult>

### 处理

该页面是用户点击top10的某一项商品之后转向的页面，抓取的信息将保存在Sheet1 - 宝贝搜索结果页面上的卖家相关信息；根据传入的TopTenItemInfo.ItemHref，先通过http get请求获得该页面，然后再利用jsoup来解析页面，获得我们想要的List<SellerInSearchResult>

## ItemDetailPageParser Module

### 模块描述

解析淘宝网ItemDetail页面，获得我们想要的数据

### 输入

SellerInSearchResult.href

### 输出

List<ItemInfo>

SellerRateUrl 卖家信用地址url

BuyersUrl 所有的买家页面url

### 处理

该页面是商家提供的关于该宝贝的详细信息页面，抓取的信息将保存在Sheet2 - 卖家提供的关于该宝贝的详细描述信息；

## ItemBuyersPageParser Module

### 模块描述

解析淘宝网宝贝买家页面，获得我们想要的数据

### 输入

BuyerUrl所有用户买家地址

### 输出

List<BuyerInfo> 大部分字段

### 处理

根据传入的url地址，通过get请求获得该页面，解析得到指定的数据

## UserRatePageParser Module

### 模块描述

商家信用页面解析模块，解析商家信用页面，获得指定的数据

### 输入

SellerRateUrl

### 输出

List<SellerRate>

### 处理

该页面是商家信用页面，从该页面抓取的信息将保存在Sheet3 - 卖家的信用等级信息中。根据商家信用页面的url地址，首先通过http get请求，获得该页面，然后利用jsoup解析页面，得到指定格式的数据。

## ItaobaoPageParser Module

### 模块描述

买家信息页面解析模块，解析得到指定的数据

### 输入

BuyerUrl

### 输出

List<BuyerInfo>中某些字段

### 处理

该页面是购买某商家提供的宝贝的用户的详细信息页面，抓取的信息将保存Sheet4 – 买家相关信息中；根据买家的url地址信息，解析页面，获得我们想要的数据；

## XML解析模块

### 模块描述

Xml解析模块主要是用来解析系统的配置文件，获得相应的变量的值，格式如下所示：

<configuration>

</configuration>

|  |  |  |
| --- | --- | --- |
| **字段名称** | **代码** | **备注** |
| 浏览器地址 | BrowserPath | 配置浏览器的路径 |
| 抓取文件保存位置 | ExcelFileSavePath |  |

# 数据结构设计

## CategoryInfo

|  |  |  |
| --- | --- | --- |
| **字段名称** | **代码** | **备注** |
| 类别名称 | CategoryName |  |
| 链接地址 | CategoryHref |  |

## TopTenItemInfo

|  |  |  |
| --- | --- | --- |
| **字段名称** | **代码** | **备注** |
| Top Rank | TopRank |  |
| 产品名称 | ItemName |  |
| 链接地址 | Href |  |

## 宝贝搜索结果页面中卖家相关信息 - SellerInSearchResult

|  |  |  |
| --- | --- | --- |
| **字段名称** | **代码** | **备注** |
| 卖家编号 | SellerId | 卖家编号，通过卖家编号，实现4个sheet的关联，即通过编号可以看到卖家关于宝贝的描述信息，卖家信用信息，买家信息等 |
| 链接地址 | Href | 根据该链接地址，进入卖家关于该宝贝的详细描述页面 |
| 类别名称 | CategoryName | 宝贝所属的类别，本需求中只要3种类别，分别是：热门手机、笔记本、洁面 |
| 卖家名称 | SellerName |  |
| 全球购 | isGlobalBuy | 是否有全球购标识 |
| 金牌卖家 | isGoldSeller | 是否有金牌卖家标识 |
| 卖家报价 | Price |  |
| 运费 | FreightPrice |  |
| 信用卡支付 | CreditCardPay |  |
| 卖家地址 | SellerAddress |  |
| 最近成交笔数 | SaleNum |  |
| 评价条数 | Reviews |  |
| 消费者保障 | isConsumerPromise | 是否有消费者保障标识 |
| 假一赔三 | isLeaveACompensableThree | 是否有假一赔三标识 |
| 七天退换 | isSevenDayReturn | 是否有七天退换标识 |
| 正品保障 | isQualityItem | 是否有正品保障标识 |
| 30天维修 | Is30DaysMaintain | 是否有30天维修标识 |
| Page | Page | 在搜索结果中的页数(1,2,3或者4)以及在其所在页的排名位置，格式为页数，排名 |
| Rank | Rank | 所在页的排名位置 |

## 卖家关于宝贝的详细描述信息 - ItemInfo

|  |  |  |
| --- | --- | --- |
| **字段名称** | **代码** | **备注** |
| 卖家编号 | SellerId | 通过卖家编号，实现4个sheet的关联，即通过编号可以看到卖家关于宝贝的描述信息，卖家信用信息，买家信息等 |
| 价格区间 | PriceRange |  |
| 物流运费 | FreightPrice |  |
| 30天售出 | SaleNumIn30Days |  |
| 评价 | Reviews |  |
| 宝贝类型 | ItemType | 如 全新  259001次浏览 |
| 支付 | PayType | 如信用卡分期，货到付款 |
| 服务 | ServiceType | 如七天退换，30天维修，公益宝贝，集分宝 |
| 规格 | Spec | 主要针对化妆品类宝贝 |
| 容量 | Capacity | 主要针对化妆品类宝贝 |
| 第一条评价时间 | FirstReviewDate |  |
| 最后一条评价时间 | LastReviewDate |  |
| 卖家信用页面链接地址 | UserRateHref | 该链接进入卖家信用页面UserRatePage |
| 宝贝买家链接地址 | ItemBuyersHref | 该链接进入购买该宝贝的所有买家页面 |
| 宝贝评价链接地址 | ItemReviewsHref | 该链接进入评价该宝贝的评价页面 |

## 卖家信用信息 - SellerRateInfo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **字段名称** | | **编号** | **备注** |
|  | 卖家编号 | | SellerId | 通过卖家编号，实现4个sheet的关联，即通过编号可以看到卖家关于宝贝的描述信息，卖家信用信息，买家信息等 |
| 卖家相关信息 | 卖家名称 | | SellerName |  |
| 当前主营 | | MainSale |  |
| 所在地区 | | Location |  |
| 创店时间 | | CreateShopDate |  |
| 卖家信用 | | SellerRate |  |
| 买家信用 | | BuyerRate |  |
| 店铺服务 | 卖家当前保证金金额 | | ChargeNum |  |
| 消费者保障 | | ConsumerPromise |  |
| 7天退换 | | SevenDayReturn |  |
| 店铺半年内动态评分 | 宝贝与描述相符 | | MatchScore |  |
| 卖家的服务态度 | | ServiceScore |  |
| 卖家的发货的速度 | | ConsignmentScore |  |
| 店铺30天内服务情况 | 平均退款速度 | | RefundmentScore | 2.18 天小于2.50 天 |
| 近30天退款率 | | RefundmentRateScore | 7.05%大于3.86% |
| 近30天投诉率 | | ComplaintScore | 0.00%小于0.02% |
| 近30天处罚数 | | PunishmentScore | 0 次小于0.09 次 |
| 卖家信用评价展示 | 最近一周 | 总数-好评 | WeekSumRateOk |  |
| 主营业务-好评 | WeekMainRateOk |
| 非主营行业-好评 | WeekNotmainRateOk |
| 总数-中评 | WeekSumRateNormal |
| 主营业务-中评 | WeekMainRateNormal |
| 非主营行业-中评 | WeekNotmainRateMormal |
| 总数-差评 | WeekSumRateBad |
| 主营业务-差评 | WeekMainRateBad |
| 总数-差评 | WeekNotmainRateBad |
| 最近一月 | 总数-好评 | MonthSumRateOk |  |
| 主营业务-好评 | MonthMainRateOk |
| 非主营行业-好评 | MonthNotmainRateOk |
| 总数-中评 | MonthSumRateNormal |
| 主营业务-中评 | MonthMainRateNormal |
| 非主营行业-中评 | MonthNotmainRateMormal |
| 总数-差评 | MonthSumRateBad |
| 主营业务-差评 | MonthMainRateBad |
| 总数-差评 | MonthNotmainRateBad |
| 最近半年 | 总数-好评 | HalfYearSumRateOk |  |
| 主营业务-好评 | HalfYearMainRateOk |
| 非主营行业-好评 | HalfYearNotmainRateOk |
| 总数-中评 | HalfYearSumRateNormal |
| 主营业务-中评 | HalfYearMainRateNormal |
| 非主营行业-中评 | HalfYearNotmainRateMormal |
| 总数-差评 | HalfYearSumRateBad |
| 主营业务-差评 | HalfYearMainRateBad |
| 总数-差评 | HalfYearNotmainRateBad |
| 半年以前 | 总数-好评 | BeforeHalfYearSumRateOk |  |
| 总数-中评 | BeforeHalfYearSumRateNormal |
| 总数-差评 | BeforeHalfYearSumRateBad |
| 卖家历史信用构成 | 卖家信用 | | SellerRate | 16685 |
| 主营行业 | | MainBusiness | 3C数码  15479 |
| 主营占比 | | MainBusinessPercentage | 92.77% |

## 买家相关信息 - BuyerInfo

主要用来保存购买卖家宝贝的买家相关信息；

|  |  |  |
| --- | --- | --- |
| **字段名称** | **代码** | **备注** |
| 卖家编号 | SellerId | 一个卖家有多个买家，用卖家编号来标识该卖家所有的买家 |
| 拍下价格 | Price |  |
| 数量 | Num |  |
| 付款时间 | PayTime |  |
| 款式和型号 | Size |  |
| 买家信用积分 | RateScore | 把鼠标移到该用户对应的钻或者心上面可以看到该用户的信用分数 |
| 买家地址 | BuyerAddress | 格式为：省份，市，[市区]，如河北省-沧州市-泊头市 |
| 买家性别 | BuyerSex | 根据买家详细页面，头像下方的文字“他/她的好友”来判断性别，如果没有此类信息，标0处理 |

# 技术难点说明

## ItemDetailPage抓取说明

关于宝贝详细描述页面即ItemDetailPage中有部分字段的抓取需要特别注意：

1. 关于如何获得物流运费

宝贝描述页面中关于物流运费这块并非直接解析从服务器返回的html页面，而是先通过js构造get请求，获得从服务器端返回的运费的相关信息，其中一个构造get请求的示例如下：

<http://detailskip.taobao.com/json/postage_fee.htm?opt=&catid=50011977&ic=1&id=10425980787&il=%B9%E3%B6%AB%B9%E3%D6%DD&ap=true&ss=false&free=false&tg=false&tid=2866818&sid=70492495&iv=98.00&up=5.00&exp=5.00&ems=0.00&iw=0&is=&callback=TShop.mods.SKU.DefaultShippingInfo.render>

关于get请求中相关的字段说明

catid

sid 用户id

id

从服务器端返回的json数据格式为：

TShop.mods.SKU.DefaultShippingInfo.render({

type:'applyPostage',

location:'广东广州',

destination:'浙江杭州',

carriage:'快递:5.00元 ',

dataUrl:'http://delivery.taobao.com/detail/detail.do?itemCount=1&amp;itemId=10425980787&amp;itemValue=98.00&amp;isSellerPay=false&amp;templateId=2866818&amp;userId=70492495&amp;unifiedPost=5.00&amp;unifiedExpress=5.00&amp;unifiedEms=0.00&amp;weight=0&amp;size=0',

cityId:'330100'

});

从以上json数据中，我们可以获得快递的起始地址和目的地址，以及快递的费用；

1. 关于购买该宝贝的所有买家的信息的获得

如上所示，要想获得该宝贝的所有买家的信息，也不能通过直接解析html页面获得，而是和上面的方法类似，先通过js构造get请求，然后再解析返回的数据，得到买家信息；

以下是一个get示例请求：

http://detailskip.taobao.com/json/show\_buyer\_list.htm?bid\_page=1&page\_size=15&is\_start=false&item\_type=b&ends=1335593146000&starts=1334988346000&item\_id=10425980787&user\_tag=375224336&old\_quantity=2198&zhichong=true&sold\_total\_num=16&seller\_num\_id=70492495&dk=0&title=%C4%A4%B7%A8%CA%C0%BC%D2+%CB%E1%C4%CC%C3%B5%B9%E5%BB%A8%B0%EA%BD%E0%C3%E6%C8%E9150g+%C3%C0%B0%D7%B2%B9%CB%AE%C7%E5%BD%E0%BF%D8%D3%CD+%CF%B4%C3%E6%C4%CC+%D5%FD%C6%B7&callback=TShop.mods.DealRecord.reload&t=1335435871882

然后解析从服务端返回的数据，即可得到我们想要的数据；

## UserRatePageParser相关说明

关于页面中“店铺30天内服务情况”抓取说明：

从chrome开发人员工具中分析可知，店铺内30天内服务情况相关信息是通过ajax请求获得的，请求的一个示例地址如下：<http://ratehis.taobao.com/monthServiceAjax.htm?monthuserid=12573474&userTag=407992336&isB2C>=

KISSY.ready(function(S){

var D = S.DOM, E = S.Event, Anim = S.Anim, pos = [9, 44, 79, 114];

var monthuserid = D.get('#monthuserid').value;

var userTag = D.get('#userTag').value;

var isB2C = D.get('#isB2C').value;

var host = D.get('#hisServerDomain').value;

//console.log(D.get('#hisServerDomain').value);

YAHOO.util.Get.script(

host+"/monthServiceAjax.htm?monthuserid="+monthuserid+"&userTag="+userTag+"&isB2C="+isB2C,{

onSuccess: function(o) {

从上面的url地址以及js如何构造指定的地址中，我们看到，get请求中主要携带3个参数，分别是：

1. Monthuserid

从页面相关分析，我们可以看到，该字段的值我们可以从

<input type="hidden" name="monthuserid" id="monthuserid" value="12573474" />获得

1. userTag：

<input type="hidden" name="userTag" id="userTag" value="407992336" />

1. isB2c：

<input type="hidden" name="isB2C" id="isB2C" value="" />

在页面中通过jsoup解析得到以上三个字段，从而可以完成url请求的构造，然后再解析来自服务器端的json数据。