淘宝数据抓取需求规格说明书

# 流程概述

1. 登陆淘宝首页www.taobao.com，选择“淘宝服务 -> 更多>”；
2. 在淘宝更多服务中选择“淘宝排行榜”；



1. 进入淘宝排行榜后，选择商品的种类，目前我们需要抓取的种类有：

* 数码家电 -> 手机/配件 -> 热门手机；
* 数码家电 -> 电脑/配件 -> 笔记本；
* [化妆品](http://top.taobao.com/level2.php?cat=TR_HZP) -> [美容护肤](http://top.taobao.com/level3.php?cat=TR_MRHF) -> 洁面



1. 选择其中的一个商品种类如热门手机，点击“销售热门”，选择top10的热门手机；



1. 选择其中的一款热门手机，如“Sony Ericsson/索尼爱立信 LT26i“，点击进入后，需要抓取 page1-4的卖家的信息，针对每一个卖家需要抓取的信息有：

* 卖家名称（下图绿色标出）
* 抓取卖家是否有全球购、金牌卖家等等的标志（下图绿色标出）
* 卖家报价；
* 运费；
* 信用卡支付；
* 卖家地址；
* 最近成交笔数；
* 评价条数；
* 消费者保障；
* 假一赔三；
* 七天退换；
* Page rank；(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)

注意：不抓取tmall的卖家；

1. 选择其中的一个卖家如“正港 Sony Ericsson/索尼爱立信 LT26i/LT26 质量保证+惊喜4重奏”，点击进入后，需要抓取的信息主要有以下方面：
2. 卖家提供的宝贝基本信息，主要是抓取以下方面的信息：

* 卖家报价最低价；
* 卖家报价最高价；
* 物流运费；
* 30天售出；
* 评价；
* 宝贝类型；
* 浏览次数；
* 支付类型；
* 服务；
* 如果是“化妆品“类别的，需要抓取化妆品的规格和容量；



1. 点击页面右侧的钻石标记，进入该卖家的信用等级详细页面，此页面基本所有的信息我们都需要抓取下来；



1. 页面下侧的评价详情，我们需要获取第一条评价和最后一条评价的时间；

时间的格式是 **（年.月.日）**



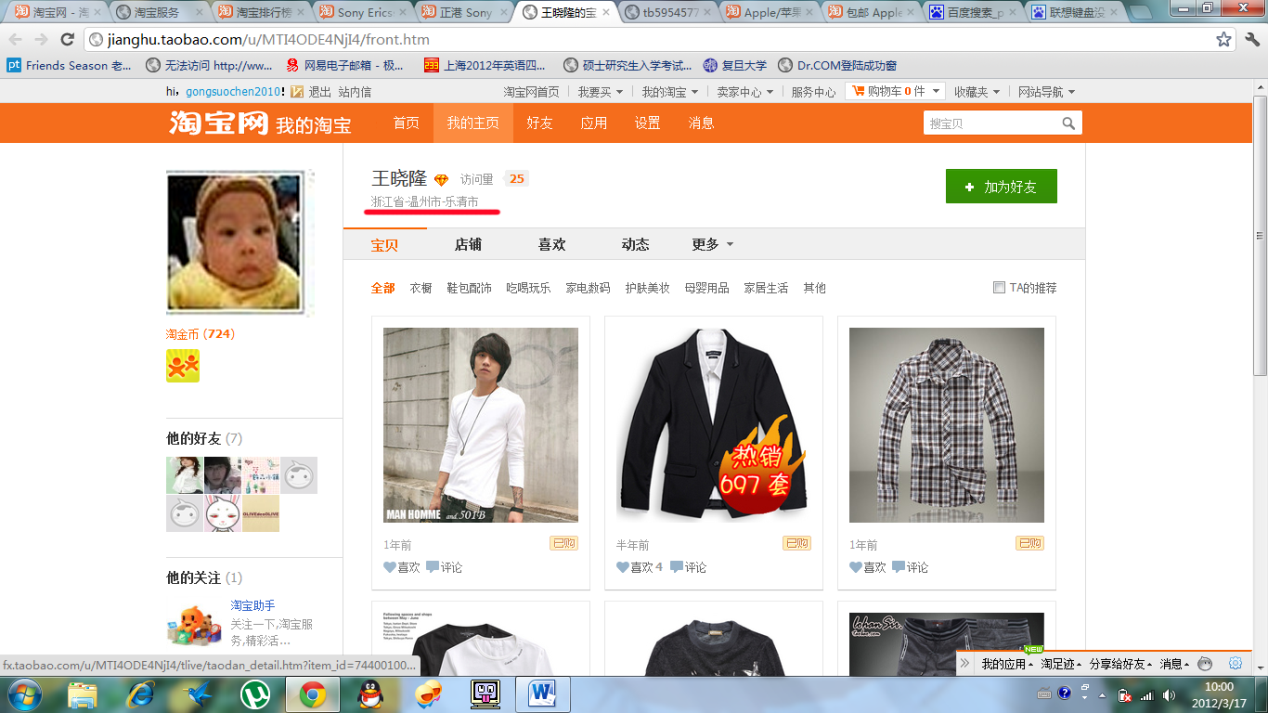
1. 页面下侧的成交记录，我们选择其中的一条成交记录如“tb5954577”，点击进入后，需要抓取**买家**昵称下方的地理位置信息，如果没有位置信息，做标0处理；

在点击进入买家信息页面前，需要记录：

* + 买家自己的信用级别（鼠标移动到名字后面就能显示）
  + 买家的评价得分
  + 拍下价格
  + 数量
  + 付款时间
  + 款式和型号

在点击进入买家信息页面后，需要记录：

* + 买家地址（省份，市，（如果还有市区））
  + 买家性别（在头像下有‘他的好友“或者是”她的好友“。如果没有这类信息，就标”0”）





# 数据保存格式

问：关于抓取下来的数据是全部保存在一个excel文件的一个sheet中，还是一个excel文件的多个sheet中(通过卖家编号来实现sheet关联)或者其他的保存方式？

**一种可供参考的方案**

我们从淘宝上抓取下来的数据将保存在一个excel(2007+) 文件的4个sheet中，分别是：

1. Sheet1 - 宝贝搜索结果页面上的卖家相关信息；
2. Sheet2 - 卖家提供的关于该宝贝的详细描述信息；
3. Sheet3 - 卖家的信用等级信息；
4. Sheet4 – 买家相关信息；

下面将详细介绍具体的数据保存格式：

## Sheet1 – 宝贝搜索结果页面中卖家相关信息

|  |  |
| --- | --- |
| **字段名称** | **备注** |
| 卖家编号 | 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 |
| 卖家名称 |  |
| 全球购 | 是否有全球购标识 |
| 金牌卖家 | 是否有金牌卖家标识 |
| 卖家报价 |  |
| 运费 |  |
| 信用卡支付 |  |
| 卖家地址 |  |
| 最近成交笔数 |  |
| 评价条数 |  |
| 消费者保障 | 是否有消费者保障标识 |
| 假一赔三 | 是否有假一赔三标识 |
| 七天退换 | 是否有七天退换标识 |
| 正品保障 | 是否有正品保障标识 |
| 30天维修 | 是否有30天维修标识 |
| Page rank | 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 |

## Sheet2 – 卖家关于宝贝的详细描述信息

|  |  |
| --- | --- |
| **字段名称** | **备注** |
| 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 |
| 价格区间 |  |
| 物流运费 |  |
| 30天售出 |  |
| 评价 |  |
| 宝贝类型 | 如 全新  259001次浏览 |
| 支付 | 如信用卡分期，货到付款 |
| 服务 | 如七天退换，30天维修，公益宝贝，集分宝 |
| 规格 | 主要针对化妆品类宝贝 |
| 容量 | 主要针对化妆品类宝贝 |
| 第一条评价时间 |  |
| 最后一条评价时间 |  |

## Sheet3 – 卖家信用信息

|  |  |  |
| --- | --- | --- |
|  | **字段名称** | **备注** |
|  | 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 |
| 卖家相关信息 | 卖家名称 |  |
| 当前主营 |  |
| 所在地区 |  |
| 创店时间 |  |
| 卖家信用 |  |
| 买家信用 |  |
| 店铺服务 | 卖家当前保证金金额 |  |
| 消费者保障 |  |
| 7天退换 |  |
| 店铺半年内动态评分 | 宝贝与描述相符 |  |
| 卖家的服务态度 |  |
| 卖家的发货的速度 |  |
| 店铺30天内服务情况 | 平均退款速度 | 2.18 天小于2.50 天 |
| 近30天退款率 | 7.05%大于3.86% |
| 近30天投诉率 | 0.00%小于0.02% |
| 近30天处罚数 | 0 次小于0.09 次 |
| 卖家信用评价展示 | 最近一周 | 好评 中评 差评  总数 57 1 0  3C数码 57 1 0  非主营行业 0 0 0 |
| 最近一月 |  |
| 最近半年 |  |
| 半年以前 |  |
| 卖家历史信用构成 | 卖家信用 | 16685 |
| 主营行业 | 3C数码  15479 |
| 主营占比 | 92.77% |



问：在“卖家信用评价展示”一栏中，卖家信用页面上是以一个表格的形式显示，我们在excel中如何表达这种关系？目前的做法是将这个表格直接保存在一个单元格中，不知这样的做法能否满足需求？

## Sheet4 – 买家相关信息

主要用来保存购买卖家宝贝的买家相关信息；

|  |  |
| --- | --- |
| **字段名称** | **备注** |
| 卖家编号 | 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 |
| 买家的信用级别 |  |
| 买家的评价得分 |  |
| 拍下价格 |  |
| 数量 |  |
| 付款时间 |  |
| 款式和型号 |  |
| 买家地址 | 格式为：省份，市，[市区]，如河北省-沧州市-泊头市 |
| 买家性别 | 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 |

问：买家的信用级别和买家的评价得分具体的是指？针对一般的用户(非淘宝店铺用户)该作如何处理？



# 开发相关

## 开发语言介绍

本系统将采用java开发，采用java主要有以下方面的优势：

1. 跨平台

采用java开发的系统，不仅能够在windows上运行，而且能够在linux或其他任何安装jre(java runtime environment)的操作系统上运行；

1. 扩展性好

采用面向接口编程，使得系统能够很好的扩展，以应对需求的变化；

## 开发计划

根据技术，时间等因素，综合考虑，项目的开发时间大约为两个月。

|  |  |  |
| --- | --- | --- |
| 时间段 | 任务 | 备注 |
| 3.19 至3.25 | 确定项目需求 | 详细描述项目的需求 |
| 3.26 至4.16 | 设计 | 概要设计和详细设计 |
| 4.17 至4.30 | 编码 | 项目编码 |
| 5.1 至5.17 | 调试 | Debug |
| 5.18 至5.19 | 交付 | 交付使用 |