毕业设计说明书

作　　者： _____ **学　号：** _____

学　　院： 计算机科学与工程学院

专业(方向): 计算机科学与技术

班　　级： _____

题　　目： 电商爬虫与观点挖掘系统

_____

指导者： _____

_____

评 阅 者： _____

2017　年　6　月

声　明

我声明，本毕业设计说明书及其研究工作和所取得的成果是本人在导师的指导下独立完成的。研究过程中利用的所有资料均已在参考文献中列出，其他人员或机构对本毕业设计工作做出的贡献也已在致谢部分说明。

本毕业设计说明书不涉及任何秘密，南京理工大学有权保存其电子和纸质文档，可以借阅或网上公布其部分或全部内容，可以向有关部门或机构送交并授权保存、借阅或网上公布其部分或全部内容。

学生签名：

年　月　日

指导教师签名：

年　月　日

毕业设计说明书中文摘要

随着科技发展，信息技术对社会和经济都产生了巨大的影响。电子商务作为互联网技术对社会和经济产生影响的重要形式，成为了一种网络化的新型经济活动，人们也习惯了在互联网上发表自己对于所购产品的看法和观点，并利用商品的评论来帮助自己在网购时进行决策。与此同时，针对这些评论的分析也将对商家或生产者发挥越来越重要的作用，它不光能帮助人们改进产品，及时发现问题，也能帮助设计者发掘消费者的潜在需求，创造更大的市场。

本文介绍了互联网电子商务的发展，阐述了电商数据中所蕴含的巨大经济价值，分析了现有爬虫技术，针对国内综合性 B2C 互联网零售电商平台——天猫、京东、亚马逊中国，设计并实现了一种以商品评论为主的数据抓取、聚合和管理方案，在所获得的数据基础上，提出了一种用统计方式对电商评论进行观点挖掘的方法。

关键词 电子商务 网络爬虫 观点挖掘

毕业设计说明书外文摘要

Title E-commerce crawler and Opinion Mining

Abstract

With the development of science and technology, information technology has had a great impact on society and economy. E-commerce is an important form of Internet technology that has an impact on society and economy. It has become a kind of new online economic activity. People are used to express their views and opinions on the purchased products on the Internet and use the reviews of the products to help themselves in the online shopping decision-making. At the same time, the analysis of these comments will also play a more important role for businesses or producers. It can not only help people improve their products, find problems in time, but also help designers to explore the potential needs of consumers to create greater market.

This paper introduces the development of Internet e-commerce, expounds the huge economic value implied in the data of e-commerce website, analyzes the existing crawler technology, and designs for the domestic major B2C Internet retail platform - Tmall, JD, Amazon China and implements a kind of data capture, organization and aggregation program based on the comments and other data obtained from the above websites. This paper put forward a statistical opinion mining method to analyze the comments.

Keywords E-commerce ; crawler ; opinion mining

目 次

1 绪论 …… 1
1.1 工程背景及意义 …… 1
1.2 国内外研究现状 …… 1
1.3 总体技术方案及其社会影响 …… 1
1.4 技术方案的经济因素分析 …… 2
1.5 论文章节安排 …… 2
2 爬虫技术和反爬虫技术介绍 …… 4
2.1 网络爬虫技术调研…… 4
2.2 反爬虫技术调研 …… 5
2.3 关于 Html 中的字符编码 …… 6
3 电商爬虫设计概述 …… 7
3.1 全站商品目录解析 …… 7
3.2 商品、目录之间的关系分析 …… 8
3.3 不同商品关联度分析…… 9
3.4 数据库设计 …… 10
4 电商爬虫系统概述与具体实现…… 13
4.1 电商商品目录解析模块 …… 13
4.2 页面抓取模块 …… 14
4.3 作业调度模块 …… 16
4.4 页面解析与数据清理模块 …… 16
5 针对电商评论的观点挖掘 …… 18
5.1 出现频率最高的商品属性提取…… 18
5.2 评论跟所描述的对象之间的联系 …… 20
5.3 针对属性的评论情感分析 …… 21
6 系统运行结果…… 23
结论 …… 25
致谢 …… 26
参考文献…… 27

1 绪论

1.1 工程背景及意义

近年来，信息技术的飞速进步对社会和经济产生了日益重大的影响，互联网经历了急速的成长，知识和信息以爆炸性的速度不断以数字化的形式积累，网络成为了人类最重要的信息获取和存储场所。Web2.0 时代[1]的一系列技术使得普通用户也方便地在网络中产生内容、发布信息，网民们通过各种渠道在网上产生的各种形式的数据，成为了互联网信息的主要增长点。

电子商务是互联网技术对社会和经济产生影响的重要形式，作为一种网络化的新型经济活动，正以前所未有的速度迅猛发展，成为了主要发达国家增强自身经济竞争力，获得全球资源配置优势的重要手段。在中国，电子商务的发展受到了国家的高度重视，出台了一系列政策鼓励和引导这一新型产业的发展.第一财经商业数据中心预计 2016 年中国 B2C 零售电商市场总额将达 5.2 万亿元，是全球最大的电商零售市场。。

中国大陆的电子商务凸显了其巨大的经济价值，以天猫、京东、亚马逊为代表的 B2C 零售电商平台品类最多，其中天猫和京东更是以巨大的优势占据了绝大部分的 B2C 零售市场份额。这些电商网站上的评论内容，很大程度上包含了消费者对商品的看法和评价。对这些电商网站上发布的消费者的评价和观点进行深入的分析和研究，将能为决策和生产提供重要的参考依据。

1.2 国内外研究现状

文献[2]讲述了应该从哪些方面提升爬虫性能，为提高抓取信息的效率提供了一些思路。

文献[3]中讲述了一种增量式爬虫，通过设计抓取规则，来应对多变的电商信息。虽然本文中爬虫的主要抓取对象时商品评论等文本，商品的价格等易变动的信息并非抓取的重点，仅是设置了可供扩展的字段用于储存商品的其他信息，但这种增量式的设计用于处理易变动的电商信息给本文提供了参考。

文献[4][5]提到的基于特征的观点挖掘、对象特征提取等方法对评论信息进行分析，给本文的观点挖掘提供了思路上的启迪。

1.3 总体技术方案及其社会影响

通过对各个电商网站结构的分析，提出了一种用三级商品目录对商品进行组织的方法，并以最细目录——三级目录为单位对商品信息及其评论进行遍历抓取。通过商品 ID 在商品

和评论之间建立联系，并通过目录在商品之间建立联系。最后通过自然语言处理的方式对评论样本进行统计分析，找出用户最关心的商品属性以及对这些属性的评价内容。爬虫程序可以将互联网上大量的非结构化文本信息进行整理收集，为数据分析提供样本，进行数据分析或者观点挖掘，继而从中提取出有价值的信息。

不加约束的爬虫程序会大大增加网站的服务器负担，甚至降低其他用户对网站的访问速度。爬虫程序通过自动化手段获取网站对外公开的信息，但如果将数据用于商业用途就很可能会侵犯网站的权益。针对这些评论的分析也将对商家或生产者发挥重要作用，它不仅能帮助商家改进服务、及时发现问题，也能帮助生产者发掘消费者的潜在需求，创造更大的市场。

1.4 技术方案的经济因素分析

爬虫程序选择用 Python 语言来编写，借助了一些现有模块来对 HTML 进行 DOM 解析。利用驱动浏览器的方法来进行网页脚本进行渲染获取异步加载的页面信息。Python 作为一种动态解释性语言，以其兼顾简洁和运行效率的特点在近些年受到了广泛的关注和应用，利用 Python 的语言特点可以快速地开发出灵活易用的程序，同时开源的特点也使得其拥有一个庞大的技术社区和支持队伍。爬虫程序本身运行时间开销主要集中在进行互联网访问这种 I/O 操作，而非需要大量运算的 CPU 密集型计算，同时爬虫程序所面临的环境经常变化，所以选择 Python 作为爬虫程序的开发语言也能使得程序能更快速地适应变化。同时 python 也拥有丰富的第三方模块库，里面有例如 lxml、Beautiful Soup 等 DOM 来解析 HTML 文本的工具。综合各种因素，Python 已经成为当下最热门的爬虫程序开发语言。

选择 MySQL 5.7 作为数据库，MySQL 是一个成熟和流行的 SQL 数据库，技术兼容性优秀，开源免费的特点也使得其成为很多工业界开发商的选择。

方案所用软件技术皆为开源免费，且都能方便地实现跨平台运行。

1.5 论文章节安排

本文针对国内市场上主要的 B2C 零售电商网站的结构进行了分析，根据各网站特性，设计了一种将商品和评论信息进行聚合的爬虫策略并加以实现，并提出了一种针对评论进行观点挖掘的方法。

以下是章节安排：

第一部分：绪论

第二部分：网络爬虫技术和反爬虫技术介绍

第三部分：电商网站结构调研与爬虫设计思路介绍

第四部分：爬虫系统的设计与实现

第五部分：观点挖掘算法设计

第六部分：系统运行结果

2 网络爬虫技术和反爬虫技术介绍

本部分列举并说明了目前流行的网络爬虫技术反爬虫技术，解释了不同网站选择不同编码方式的原因。

2.1 网络爬虫技术调研

互联网带来了海量数据，数据获取也变得更加便利，数据获取的渠道也多种多样。 数据需求方可通过授权合规渠道获取数据，根据数据的价值，往往需要付出一定成本；有些情况下， 比如同行业竞争企业之间， 希望获得对方的一些数据信息， 又不希望透露自己的身份， 其授权方式是行不通的；再有一些情况，发布方希望信息能被最终用户使用，但不希望其他人或者企业利用这些信息做商业用途，也可能不提供授权数据获取的方式，比如法院的执行公示信息，真实存在的各种现实需求推动了网络爬虫技术的发展和应用。

2.1.1 对于网络爬虫的分类

文献[6]根据爬虫对网页的访问方式不同，将网络爬虫分为四类，这四类分别为通用网络爬虫、主题网络爬虫、增量式网络爬虫以及深层网络爬虫。但在实际中，爬虫程序可能兼有几种爬虫的特点。

2.1.2 Python 语言框架下的爬虫技术

添加 HTTP 协议头信息进行网页访问

使用 Python 内置提供的 urllib 模块即可设计包装 HTTP 报文进行网络 URL 访问。

渲染 javascrpt

用到了 selenium 和一个只有命令行界面的浏览器引擎 PhantomJS，尽量减小内存占用。针对异步加载的网页通过渲染 javascript 脚本来获取需要的页面信息。

操纵浏览器模拟人类点击操作

Selenium 是一个用于 Web 应用程序测试的工具，浏览器可以根据脚本代码模仿人的操作做出鼠标点击，输入字符，滚动页面等操作，但仍旧无法直接使其绕过人机验证机制。

2.2 反爬虫技术调研

由于不加限制的爬虫程序会给网站服务器带来很大的压力，或者处于对数据的保护目的，很多网站会采取反爬虫措施，有的甚至会经常变动反爬虫策略来限制爬虫程序获取数据。反爬虫技术的不断更迭，大大提高了爬虫的实现难度。网站所采取的反爬虫策略甚至决定了一个爬虫程序能否如期实现获取所需信息目的。

另一方面，过多的限制可能会引起误伤或者严重影响用户体验，所以网站不会一味对访问进行严格限定，会在采取反爬虫的同时注意用户体验，这也使得完全模拟用户操作的爬虫程序无法被服务器直接识别，比如 selenium 用代码操纵浏览器模拟人类的操作来访问网站。

目前的反爬虫机制大多在 OSI 七层模型[8]的应用层工作，利用 HTTP 报文或基于对爬虫行为的分析，识破伪装，准确锁定爬虫并进行访问限制，具体限制方式如下：

1)对访问行为进行限制：限制访问频率，对于在某一时刻中访问频率过快的 IP 进行限制，或者通过对某机器在一天内对网站的访问的次数进行限制

这种反爬虫方式很可能会误伤用户，考虑到目前大多数访客都是通过动态 IP 来登陆互联网，大多网站一般不会限制某个 IP 的访问次数。

2)通过 HTTP 报文进行识别：对 Header 进行识别，或者对 Cookie 进行识别。

目前拟造访问头的爬虫程序已经普及，所以这种方式只能限制爬虫抓取的速度，无法完全阻止爬虫程序的访问。

3)通过人机验证限制访问：常见的人机验证方式有：验证码，鼠标操作，识别图中物品，以及回答问题等。

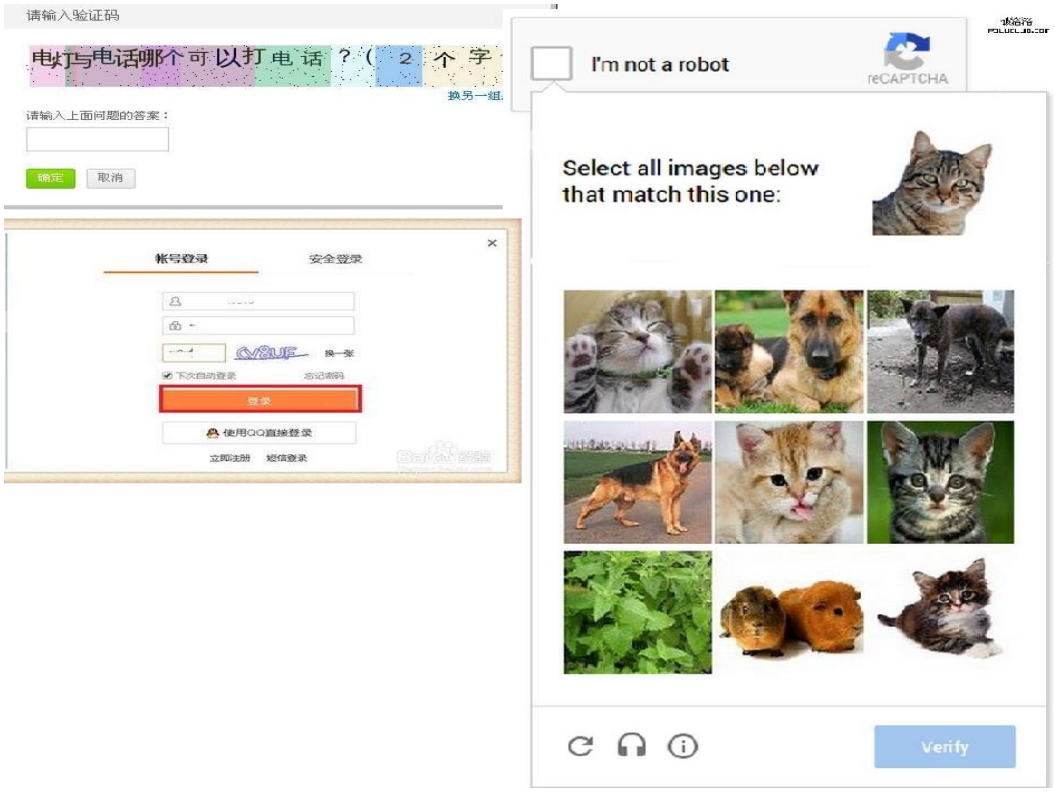


图 2.1 网页人机验证

人机验证是目前应对机器人程序最有效的方法，想要通过这种测试，爬虫程序要付出高昂的代价，反过来这也会使得网站服务器的计算开销提升。

4)协议声明: Robots 协议

robots.txt 是搜索引擎中访问网站的时候要查看的第一个文件。robots.txt 文件告诉蜘蛛程序在服务器上什么文件是可以被查看的，爬虫程序仍然可以选择不遵守这个协议，不过商用搜索引擎公司不会冒着被起诉的风险来忽视它，百度和 360 在 2012 年起展开的一场屏蔽与反屏蔽战是很好的例子[9]。

例如在天猫商城的协议中写到:

*User-agent: **

Disallow: /

天猫的 robots.txt 协议明确指出，不允许机器人访问其网站下的所有页面。

总的来看，在新的爬虫技术不断涌现的同时，反爬虫技术也是马不停碲地更迭换代，有可能一夜过去，花费大量时间和精力才完成的爬虫程序就完全失灵。出于商业对手之间的相互竞争或者研究调查者的数据收集目的，爬虫技术和反爬虫技术之间不断上演着军备竞赛，相互竞争推动了各种新技术的产生和应用，道高一尺魔高一丈的斗争从未停止。

大多情况下，爬虫程序的编写者是处于被动地位的，他们必须在网站运营者做出改变之后尽快做出改变，进行持久抓取的爬虫程序更需要灵活应变。当然网站运营者也是头疼的，在他们发现网站的大多数流量是来自机器人程序而非正常的用户时，可能网站上的所有数据已经被爬取下来了。

2.3 关于 HTML 中的字符编码

UTF-8 是一种被世界广为采用的 HTML 编码传输方式，兼容多国语言。但相比 GBK 编码方式对于中文的处理会多占用 50%的空间——一个中文汉字以 GBK 编码会占用两个字节，若用 UTF-8 方式编码则会占用三个字节，不过 UTF-8 编码的英文和数字以及常用字符只需要占用 1 个字节，所以在 UTF-8 被作为 HTML 编码标准广泛应用的同时，仍有不少中文网站会选择用 GBK 编码传输和处理中文字符信息。UTF-8 和 GBK 都可以解码为 unicode，unicode 中所有字符都占据两个字节，兼容所有世界语言和字符[10]。中文占比高的页面会选择用 GBK 编码以求节省传输流量，英文字母和符号占比高的网页用 UTF-8 编码更为合适。可以推断这也是天猫商城采用了 UTF-8 和 GBK 混合编码的原因。

3. 电商爬虫设计概述

中国市场上商品种类最多的几家 B2C 零售网站（天猫、京东、亚马逊中国），都采用了三级目录的方式对网站上的众多商品进行分类，即用逐层细化的方式将商品组织并将第三级目录中的商品进行视图展示。同一商品可能属于不同的目录，例如：在天猫商城中，平板电脑这一类别属于电脑整机类别，同时电脑整机类别隶属于*手机 /数码 /电脑办公*这一类别目录之下。

爬虫的设计以此为线索，将第三级商品目录作为单位，对商品基本信息和评论进行抓取，并记录商品所属目录，期望通过对商品目录关系的分析，构建起不同商品之间的关系，通过数据库和外围函数接口的构建，提供一种针对特定类别商品的查询方法。

3.1 全站商品目录解析

通过对网站结构的分析搞清商品目录层次关系，为数据抓取和数据存储提供参考，并分析评论以及商品之间存在的关联，得到将不同种类的商品和评论数据聚合关联起来的方法。

3.1.1 商品目录总结：

三家电商平台都将商品目录分为了三个层级，从中列举了一些不同层级的目录如图 3.1 所示：

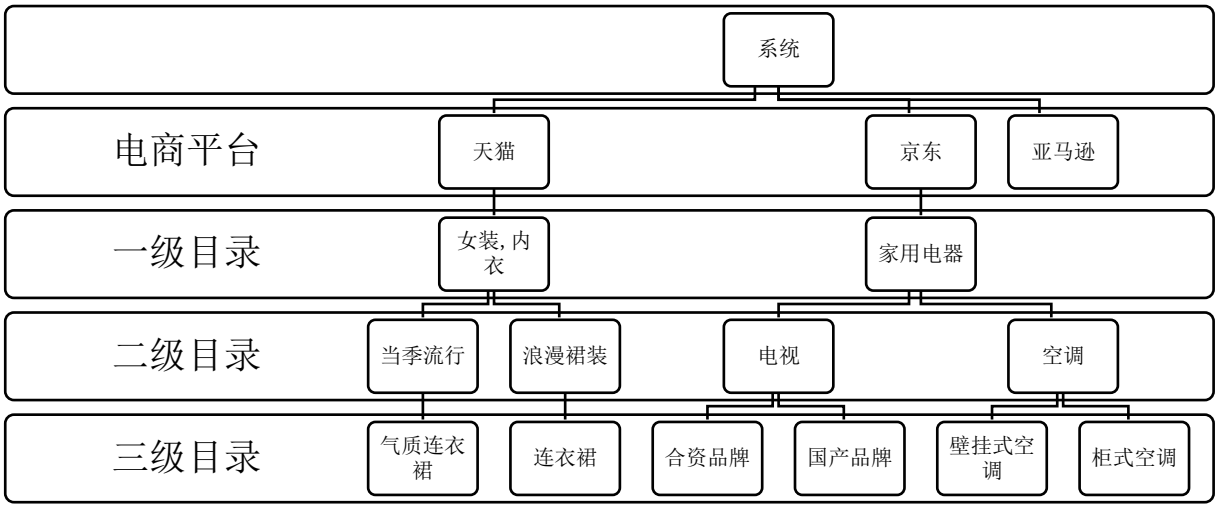


图 3.1 三级目录示意图

在一个平台电商中，一件商品可能归属于不同的目录，如上图中天猫平台下的三级目录连衣裙，其目录下的商品跟二级目录*当季流行*目录中的*气质连衣裙*会有交集。对于类似的分类条件模糊的目录，可以选择手工从抓取目录中去除，但这会增加工作量，也使得系统自动

化程度降低。但为了更好地建立商品之间的关系，对于抓取规则必要的人力干预是重要且有效的。

需要注意的是，目录关系不是一成不变的，有的商品目录是季节性的，例如天猫商城“当季流行”这一 2 级目录：

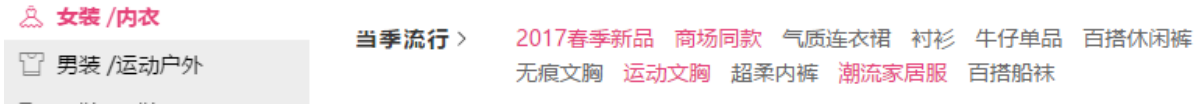


图 3.2 天猫商城目录截图

对待这种动态变化的目录要注意保持数据库中目录关系字段的动态可扩展性，允许存储同一商品的多个目录层级关系数据。

3.1.2 各平台目录数据

天猫商城上共 16 个一级目录，126 个二级目录，1509 个三级目录，网页 Html 编码方式 GBK 和 UTF-8 混合编码，其中首页为 UTF-8，评论 API 返回的中文数据为 GBK 方式编码；亚马逊中国共 14 个一级目录，共 98 个二级目录，867 个三级目录，网页 HTML 中中文编码方式为 GBK；京东商城共 16 个一级目录，共 134 个二级目录，1352 个三级目录，网页 HTML 中编码方式为 UTF-8。

3. 2 商品、目录之间的关系分析

3.2.1 目录之间的关系：

同一平台

1) 下级目录归属于上级目录
2) 同级目录可能归属于同一上级目录。
3) 不同三级目录可能不属于同一二级目录但归属于同一一级目录。

不同平台：

1) 不同平台的同级目录存在着直接对应关系，此关系建立依赖于语义分析，对二级目录或一级目录因为数目不多，可以依靠人工标记建立联系。
2) 不同平台的下级目录可能会通过上级目录的直接对应关系建立起间接对应关系。

3.2.2 商品之间的关系列举

同一平台下：

1) 通过商品名称直接建立对应关系，比如不同商家销售的同一商品。方法：数目巨大，需要通过机器比较商品名称，这里需要有可行的商品名称对比算法。
2) 归属于某一级别的同一目录。

3) 通过其他筛选条件建立联系。

不同平台：

1) 通过商品名称直接建立对应关系

2) 所属目录之间存在直接或间接的联系。

3) 通过其他筛选条件建立联系。

商品通过目录建立的联系又有一定的集合包容，同属于一个三级目录的商品肯定也在同一个二级、一级目录之中。

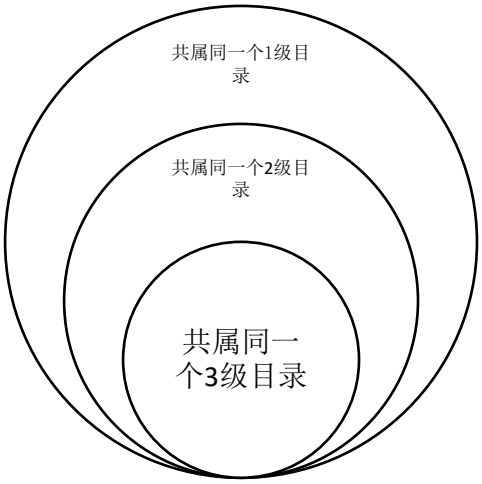


图 3.3 不同目录层级共属关系

3.3 不同商品关联度分析

根据是否为同一种商品，是否属于同一目录以及所共属的目录层级，以及是否属于同一平台来区分商品关联度等级，如：

不同店家的同一种商品，通过商品名称全文对比建立直接联系：

1) 通过商品名称和这 id 判断为同一种商品

2) 同一平台下，同一三级目录下的商品

3) 不同平台下，归属于具有直接联系的三级目录的商品

4) 同一平台下，归属于同一二级目录的商品

5) 不同平台下，归属于具有直接联系的二级目录的商品

6) 同一平台下，归属于同一一级目录的商品

7) 不同平台下，归属于具有直接联系的一级目录的商品

其中数字越小关联度越强。

3.3.1 商品与评论的关系：

评论通过商品 id 与商品建立直接联系，同其他商品的联系可以通过继承商品所在目录的关系来获得。

3.3.2 在不同平台的目录之间建立映射关系

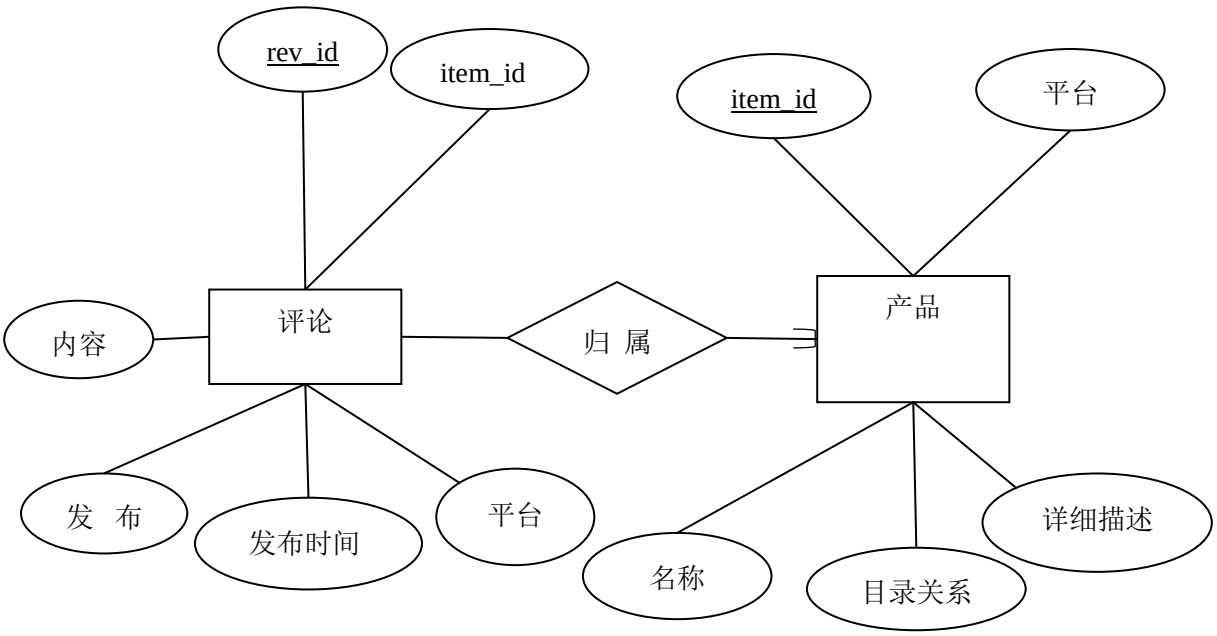
在通过爬虫成功将数据抓取下来之后，需要额外建立目录关系，将不同平台或者不同目录下的商品通过目录联系起来。

3.4 数据库设计

3.4.1 数据库 E-R 图

在数据库中直接体现的关系是商品与评论的关系，更为复杂的关系不能直接通过字段建立，考虑到数据之间普遍存在着关联关系，选择了使用关系型数据库存储经过解析后的结构化数据，关系如图 3.4 所示：

图 3.4 产品信息与评论 E-R 图



评论通过商品 item_id 同产品建立多对一关系并产生外键依赖。

3.4.2 数据库关系设计

存储评论的数据库表字段信息如表 3.1 所示：

表 3.1 评论数据表

字段	类型	可否为空	缺省值	注释
id	int(11)	No		主键，自增
rev_id	varchar(150)	No		评论 id,唯一
item_id	varchar(100)	No		评论所属商品 id
content	text	No		评论内容
Author	varchar(150)	YES	NULL	评论作者
date	datetime	No	NULL	评论发布日期
rating	float	YES	NULL	评论打分标准值
Platform	varchar(150)	NO	NULL	评论所属平台

存储商品信息表，用于存储商品的名称以及目录关系，其字段信息如表 3.2 所示：

表 3.2 商品信息表

字段	类型	可否为空	缺省值	注释
id	int(11)	No	NULL	主键，自增
item_id	varchar(150)	No	NULL	商品 id，主键
title	varchar(500)	No	NULL	商品名称
platform	varchar(50)	No	NULL	所属电商平台
category	text	No	NULL	目录信息
detail	text	Yes	NULL	商品的具体描述

为了使网络爬虫支持更多抓取方式，在商品信息表中设置了 detail 作为拓展字段用于存储商品检索关键字等信息。

用于存储 URL 页面信息的字段信息如表 3.3 所示：

表 3.3 URL 内容存储表

字段	类型	可否为空	缺省值	注释
id	int(20)	No	NULL	主键，自增 id
url	varchar(500)	No		页面 URL，唯一
content	mediumBlob	YES	NULL	存储原始页面内容
detail	text	YES	NULL	补充描述信息：包括所在目录层级
status	int	No	0	状态码

3.4.3 触发器

商品评论对商品具有依赖关系，当删除某种商品时，通过级联[11]的方式触发将评论中对于该商品的评论也一并删除。

4 电商爬虫系统概述与具体实现

整个系统分为三个模块，分别是商品目录解析模块、信息抓取和存储模块以及数据清理和分析展示模块。

其中，商品目录解析模块分析各个电商网站的页面，自动地抽取出各个电商平台的三级商品目录关系树，并提供查询接口供其他模块获取需要的信息；信息抓取和存储模块将接受三级目录节点，根据其 url 链接进行单个三级目录下商品和对应评论的遍历抓取，所有返回的页面信息会以原始二进制数据的形式存储在 MySQL 相应表格(url_content)的 blob 型字段中，同时存储其他附加描述信息；数据清理和分析模块借助以上信息，利用 lxml 等工具对数据库中的页面信息进行解析，抽取出格式化的数据条目存储在数据库中的相应表格，并根据目录层级等字段建立不同商品和评论之间的关系，提供几种有实际应用背景的扩展查询方式，根据已经建立关系的评论进行分析，在 web 上展示出分析结果，各个模块之间的交互关系如图 4.1 所示：

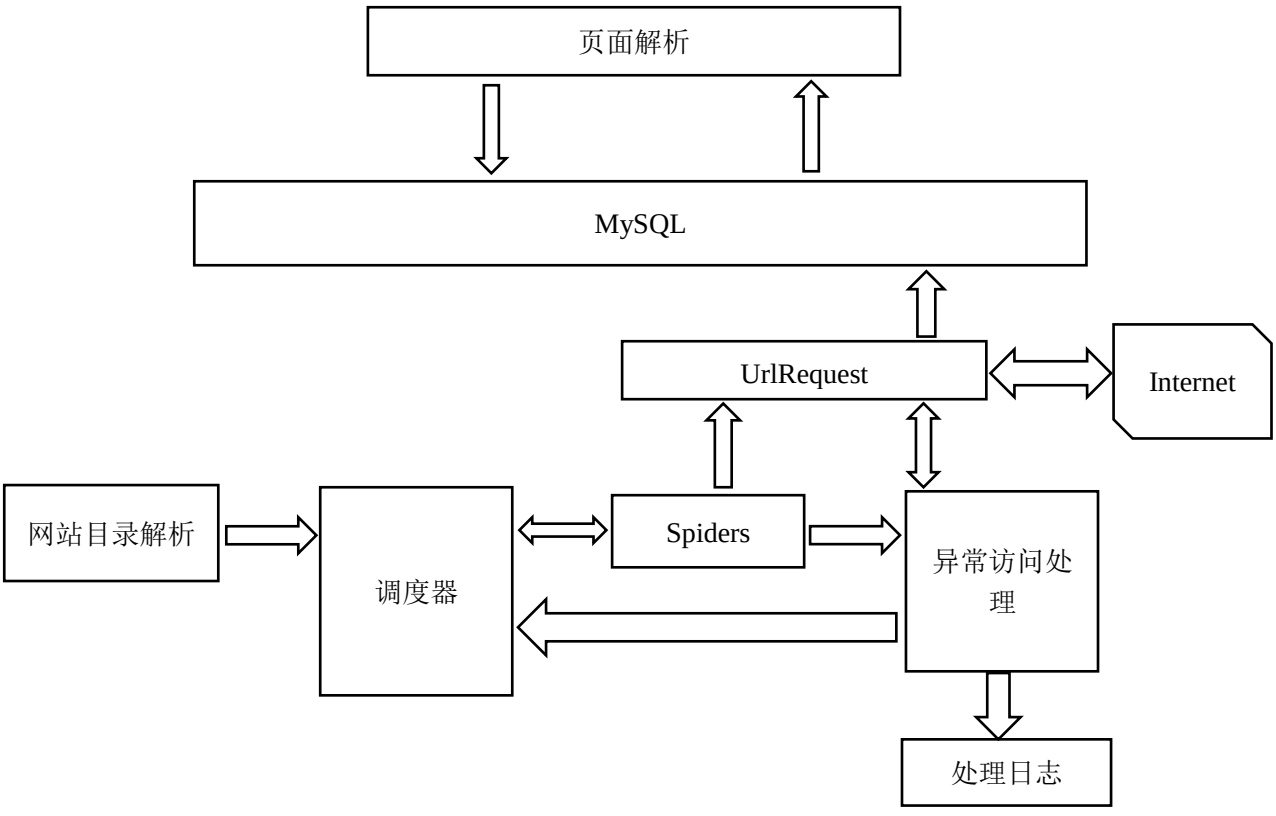


图 4.1 爬虫系统结构

4.1 电商商品目录解析模块

目录解析模块通过对于电商网站页面 html 结构的分析，抽象提取出了网站的商品目录结构，在不同目录之间建立了关系。并提供了一种贴近实际需求的查询接口，功其他模块按需获得想要的商品目录数据，如图 4.2：

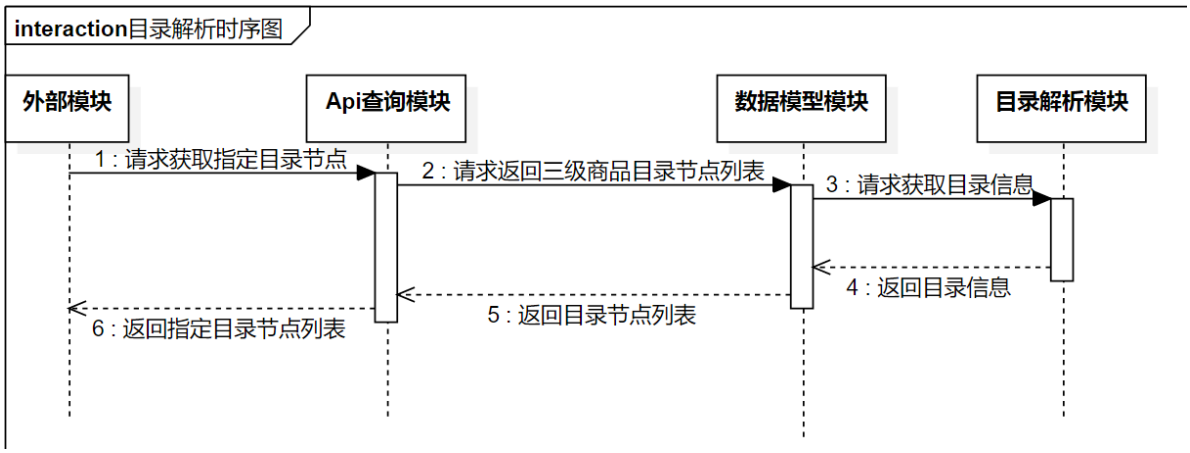


图 4.2 目录解析时序图

目录数据对象建模：对数据和关系进行抽象，用面向对象的方式定义一级目录(CateLV_1)、二级目录(CateLV_2)、三级目录(CateLV_1)节点，将各个相互联系，加强各个层级目录之间的联系，使得对数据的解析和管理变的方便容易，理解起来更为直观。

电商网站目录页面解析:针对几个电商网站的目录页进行 DOM 解析，从中抽取出统一格式的中间数据供其它模块调用。这里使用了二级商品目录节点列表作为中间数据，二级商品目录节点数据包含了：电商平台、一级商品目录名称、所含三级商品目录名称和三级商品目录链接。

全站商品目录关系解析：根据前文中关于商品目录层级的分析，对抽象出的节点对象进行组织。将每个电商网站的商品目录关系抽象为一棵高度为 3 的目录树，从而在不同节点之间建立关系。

4.2 页面抓取模块

经过对各个电商网站页面的分析，发现电商网站中评论和商品浏览页中的信息是通过 javascript 延迟加载的，即当用户做出操作时，浏览器会使用异步加载的方式向服务器发送请求，获取需要展示的 JSON 数据，然后经由前段框架渲染出信息。这种方法能提升用户的体验，不用过于频繁地跳转页面，同时也减轻了服务器的压力。使用浏览器开发工具，对于电商网站的后台活动进行分析和调试后，成功获取到了用于访问 JSON 数据的 URL，对其分析后确认了各个参数对于返回数据内容和格式的影响。从而找到了一种对商品目录中的商品及其评论进行遍历访问和数据抓取的方法。

电商爬虫对于不同的商品目录进行的抓取工作是完全平行的，彼此之间共享一个信息队列，用于收集失败的访问操作，最后交予异常处理模块进行操作，从而提高信息抓取的成功率，并及时发现异常情况并及时做出调整，各个模块之间的交互如图 4.3 所示：

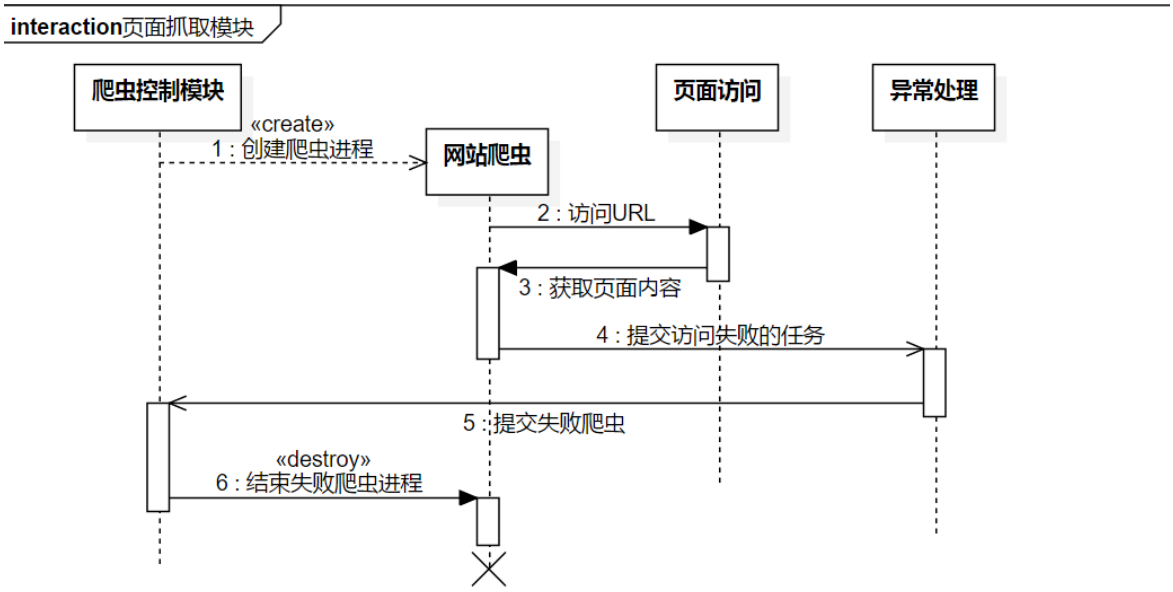


图 4.3 页面抓取时序图

4.2.1 对于原始 HTML 存储

商品的页面信息会随着时间的流逝发生变化，同时由于电商爬虫目前并不关心评论和商品页面的所有内容，解析后的结果并不能包括所有抓取到的数据信息。

比如我们不关心用户的设备，就没有对头像进行解析，但是如果今后本平台对用户所使用的设备分布感兴趣了，那么再去爬取一遍就显得费时费力。倘若我们能将原始 HTML 页面存储起来，那么无论今后有什么样的新增需求，都可以灵活地加以应对。为了保持技术平台一致性，我选择用 MySQL 的 Blob 字段存储原始 HTML。

4.2.2 URL 访问模块

整个爬虫系统对于网络的访问都是通过 UrlRequest 实现的，UrlRequest 负责访问 url 获取页面信息内容。由于这个模块直接面向外部网络环境，如何成功并且高效地获取页面信息对整个系统能否成功运转和运行效率影响重大。这个模块面临的最重要也是最复杂的问题就是如何成功应对各个网站的反爬虫机制。

这个模块提供了两种访问方法：

使用 Python 内建模块 urllib 直接通过 http 报文进行访问和使用第三方模块 selenium 操纵 phantomJS 进行访问。第一种访问方式时间开销和计算资源开销相比第(2)种都要小得多，但

是第(1)种访问方式会丢失异步加载的数据。第一种方式虽然时间和内存开销大，但是可以应对更复杂的反爬虫机制，获取异步加载的页面信息。

4.2.3 京东商城网站爬虫

针对对京东网站的某个三级商品目录，遍历抓取商品信息和评论数据。判断商品目录页面是否符合抓取规范，并探测对商品信息和商品评论页面的访问是否出发了京东商城的反爬虫机制，将异常的页面通过共享队列提交给异常任务处理爬虫。

4.2.4 天猫商城网站爬虫

针对天猫商城网站的某个三级商品目录，遍历抓取商品信息和评论数据。判断商品目录页面是否符合天猫商城爬虫的抓取规则，并通过关键字检测探测对商品信息和商品评论页面的访问是否出发了天猫的反爬虫机制，将异常的页面通过共享队列提交给异常任务处理爬虫处理。

4.2.5 亚马逊中国网站爬虫

针对亚马逊网站的某个三级商品目录，遍历抓取商品信息和评论数据。判断是否成功获取期望的页面信息，判断商品目录页面是否符合亚马逊爬虫的抓取规则，将访问异常的页面通过共享队列提交给异常任务处理爬虫处理。

4.2.6 异常任务处理爬虫

从共享队列中提取网页访问信息，识别关键字对于来自不同网站的任务进行分类处理，根据结果依次采取：

1) 重复访问失败页面，
2) 尝试用反爬虫手段访问页面
3) 分析失败原因，将结果记录在日志系统中。

其中根据日志系统中记录的信息，可以判断网页访问失败的原因，及时对系统访问策略做出调整。

4.3 作业调度模块

作业调度模块是整个系统的控制核心，控制着对电商网站商品目录的解析以及爬虫模块任务的分配和进行，将不同模块通过接口粘合在一起，完成总的作业任务。

调度模块会根据需要从目录解析模块中提取需要的目录节点，再根据类别分配给各个爬虫，爬虫根据接收到的节点数据爬取网站数据并存储数据在本地，随后调度器会操纵页面解析和数据清理模块将存储在本地的页面数据提取出来进行解析，根据数据类别不同，进行不

同的数据清洗和结构化操作，存储在相应的数据库之中。再根据需要是否开启 Django App 在网站上进行数据展示。

4.4 页面解析与数据清理模块

不同电商网站页面结构不同，模块从不同的页面中抽取出统一标准格式化的信息，并将其存储在本地数据库中。

对于通过调用 API 返回的页面，由于本事是 json 格式，经过解码之后可以直接用 python 的 json 模块导入进行页面提取。对于 html 页面，在分析过页面结构之后，既可以通过 lxml 将其抽象成 DOM 进行解析，也可以直接用正则表达时的方法进行解析。

过程中要处理好中文编码问题，有的页面会混用 UTF-8 和 GBK 编码方式。

因为解析方法依赖与 Html 结构，有时网站标签结构乃至于属性的变化都会使得模块无法正确解析出需要的格式化数据，所以模块在进行解析工作时对于获取的数据要进行一些格式判断，对于出现的异常情况要及时给予反馈和记录。

5 针对电商评论的观点挖掘

本文中的观点挖掘所针对的是，关于某个商品或者某类商品的所有评论，目标是从评论中抽取出用户最关心的商品属性，以及对于属性的评价热点。

评论的来源是数据库中经过清晰、格式化后的电商评论数据，通过查询接口的设计，可以按用户的筛选从数据库中获取指定的评论。对于评论的分析，中文内容的分词[12]借助了结巴分词，以及宾夕法尼亚大学计算机和信息科学系开发的 NLTK[13]工具。其中，结巴分词是一款提供了 Python 接口的中文分词工具，该工具支持自定义词典，能快速地将中文段落切分成一个个单词；NLTK 是一款自然语言分析工具，其将常用的语言分析工具进行了打包，提供了灵活强大的函数接口，对于经过分词的中文语料能高效地进行统计分析。

观点挖掘按分析粒度可以分为：单词级别、句子级、文档级别[14]。由于电商评论篇幅较短，认为更适合单词级和句子级别的分析。

电商评论大多是针对商品的一个或几个属性进行的有一定感情倾向评论，评论所评价的属性是一条评论的重要分类依照，比如一台笔记本商品的属性有价格、性能、商品质量、便携性等。商品属性又有显式属性和隐形属性[15]之分，显式属性会在评论中直接体现，包含描述商品的功能或性能的名词短语，比如价格便宜；隐式属性则没有直接提到某一属性，需要联系上下文才能搞清楚，比如“没钱”、“卖肾”都暗示了商品价格超过了顾客的购买能力。

5.1 出现频率最高的商品属性提取

基于属性的情感分析英文简写为 ABSA(aspect-based sentiment analysis)[16]，对于网上的评论研究非常重要。

用户在网上购买商品之后，能够针对本次购物发表评论，有理由相信这些评论能在很大程度上反应出用户对于商品的看法，其中出现频率最高的内容自然是该商品最受消费者关注的地方。由于商品评论数量巨大且更新较快，所以用肉眼观察的方法不能适应。

需要注意，不是所有评论都能反应消费者的真实想法。有的内容太过简单，例如：“不错”、“还好”，这些评论虽然能看出消费者对本次购物过程比较满意，但并不能从中寻找出评论所针对的主体，为了使分析结果更有价值，这里会将这种过于简单的内容进行剔除；有的评论千篇一律，一句话重复多次，且跟多条评论内容完全重复，有明显的刷单迹象，对于这种评论，对其进行了去重处理，减小其对整个分析样本的影响；对于没有可读性的评论，比如英文数字穿插的乱码，在构建样本时进行剔除。

接下来对样本进行操作，以笔记本电脑为例：

分词：

为了提高分词的准确度，进行了一定程度的人工干预，即从商品网页上寻找针对该商品的一些常用词汇，比如天猫商城中 Macbook air 商品详情页面下，累计评价里面的“大家都写到”栏目中的标签词汇如图 5.1：

5.1 天猫商城评论页面热门标签

可将这些词汇按格式存储在本地，提供给结巴分词分词器。

统计分析：

为了提取出有代表性的关键词，需要采用一种既能展示出热点词汇，又能体现特点的分析算法，经过调研，这里采用了 TD-IDF（term frequency–inverse document frequency）算法。TF-IDF 算法的主要思想是：如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，选取两种商品进行关键词分析如表 5.1：

表 5.1 TF-IDF 算法提取关键词

某款笔记本	爽肤水
小米	不错
笔记本	好好
外观	京东
京东	效果
电脑	正品
开机	感觉
速度	东西
性价比	一直
非常	保湿
屏幕	快递
感觉	没用
颜值	好用

为了进一步凸显商品特点，选取了类别上相差较大的商品作为对照组进行关键词分析，将本商品的热点关键词同对照组进行对比之后，剔除共有的属性关键词，得到商品中更具特点的属性。统计结果如表 5.2 所示：

表 5.2 直接分词统计关键词结果

某本评论关键词	词频	某款爽肤水评论关键词	词频
笔记本	0.103924	爽肤水	0.059102
性价比	0.07211	第一次	0.054374
很漂亮	0.039236	挺好用	0.052009
第一次	0.025451	曼秀雷敦	0.042553
一段时间	0.020148	哈哈哈	0.035461
玩游戏	0.019088	第二次	0.028369
挺不错	0.015907	挺不错	0.021277
全金属	0.011665	男朋友	0.018913
为什么	0.011665	一如既往	0.016548
触摸板	0.011665	挺舒服	0.016548
13.3	0.011665	哈哈哈哈	0.016548
分辨率	0.011665	怎么样	0.014184
第二天	0.010604	价格便宜	0.01182
没什么	0.009544	护肤品	0.01182
充电器	0.009544	看起来	0.01182
挺好用	0.008484	一段时间	0.01182
看起来	0.008484	是不是	0.01182
超轻薄	0.006363	没用过	0.01182
比较满意	0.006363	为什么	0.009456
一分货	0.005302	产品质量	0.009456
总计	0.418876	**总计**	0.465721

5.2 评论跟所描述的对象之间的联系

在找到了受关注的属性词之后，就可以借此对研究顾客对于这种属性的具体看法进行探究。将评论以标点符号进行切割成更小的句子之后组成集合，按是否包含制定关键词对其进行筛选，获取包含关键词的所有句子。

举例：针对关键词“续航”

如果能提升续航就好了

续航给力

电池普通使用续航估计 4 个多小时

电源续航肯定不行

续航也还可以

续航能力正常办公用5 到6 个小时

电池续航还行

电池续航能力没有那么牛逼

续航9 小时是待机吧

电池续航时长时短

电池续航还可以

续航还行吧

续航和重量都非常适合商用

电池续航也不错

续航还不清楚

续航没有宣传的那么厉害

从结果中可以看出针对某款笔记本续航能力的一些有代表性的评论，对这些评论进行分析之后就能够得到顾客对于这款笔记本电脑续航能力的看法和评价。

5.3 针对属性的评论情感分析[17]

5.3.1 基于情感词典[18]的无监督方法

基于词典的无监督机器分类舆情分析主要依靠情感词典中记录的情感词的极性、强度来推算评论的情感倾向，需要考虑的语义现象如下

a) 情感转移：否定词，能够反转情感次的极性，如“不满意”

b) 情感放缩：程度词，能够改变情感词的情感强度，如“不太高兴”

c) 转折句式：转折词，能带来情感变化，如“但是”

d) 虚拟语境：虚拟词，疑问词，条件词，这种语境下的情感可以选择性忽略

e) 情感短语：常用的情感短语，如“有魄力”是一个正面的情感短语

f) 网络用语：一些新流行的网络用语，在情感分析中不可缺少

通过设计相关算法，量化上述因素造成的情感影响，并设定阈值，可对评论进行情感分类

5.3.2 基于统计学习的监督方法

基于统计学习的情感分析方法，要利用经过标注的电商评论预料，通过机器学习算法对语料进行抽象提取特征，然后进行学习计算，得出能将评论的感情倾向很好地分类的分类算法。流程主要包括以下几步

1) 评论特征选择和特征提取
2) 用特征重新表达训练样本和测试样本
3) 使用机器学习方法学习训练样本
4) 使用训练好的模型对新的评论进行情感分类

机器学习方法可以选择支持向量机[19]或神经网络等方法，综合多种模型进行集成学习[20]可以得到更好的结果。

6. 系统运行结果

商品目录解析模块得到的天猫商城商品目录树对应的 JSON 格式如图 6.1 所示：

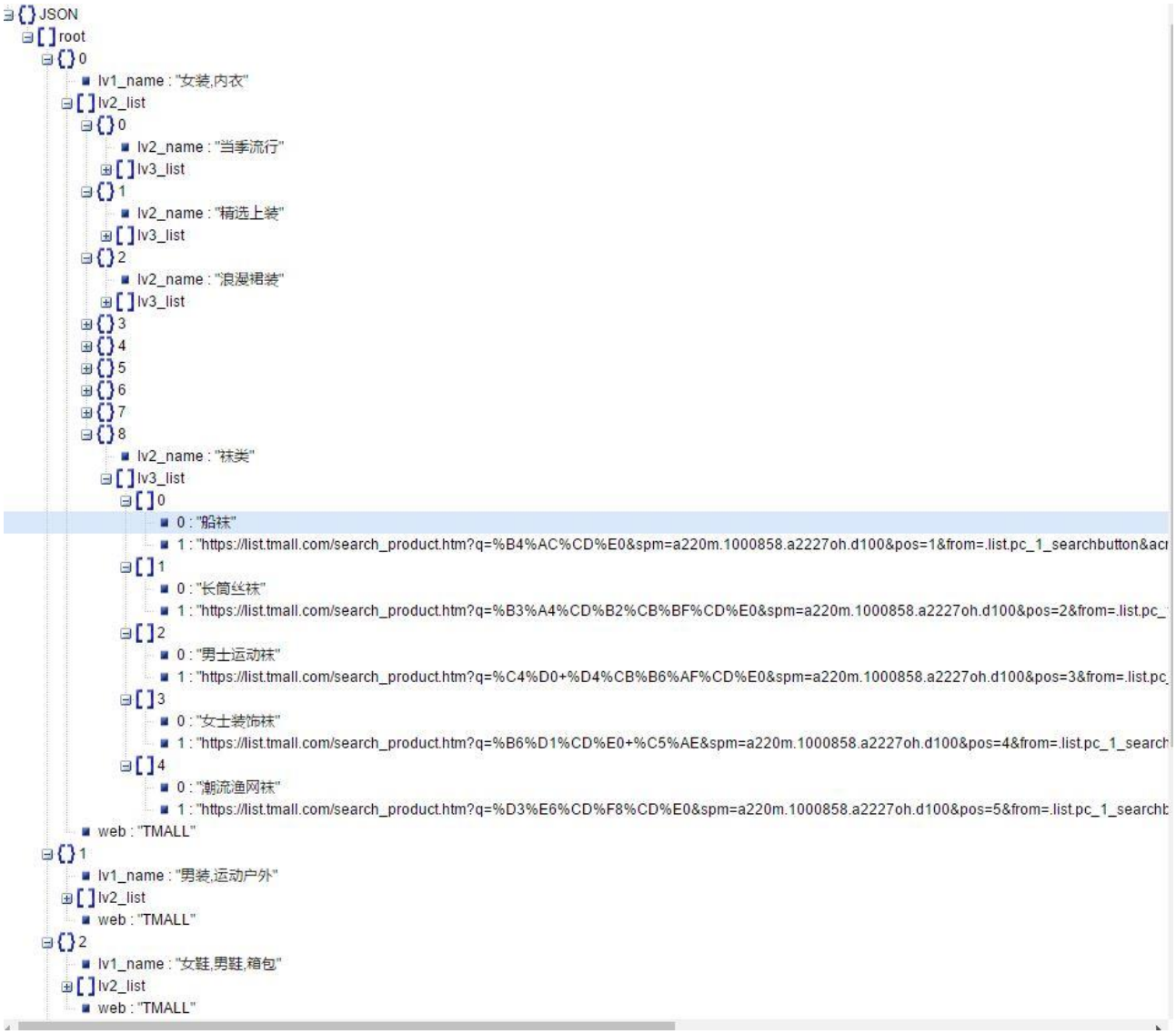


图 6.1 天猫商城商品目录结构

商品信息内容如图 6.2 所示：

item_id	title
TMALL_126446588-544524388480	Lenovo/联想 小新 510S i7
TMALL_133668489-547763460584	HP/惠普 Pavilion 15 AU157TX
TMALL_1647052829-536824039731	Lenovo/联想 小新 700 I5
TMALL_1647052829-544372180064	Lenovo/联想 小新 310-15IKB
TMALL_1669409267-44131265268	Apple/苹果 MacBook Pro MF839CH/A
TMALL_1714128138-539122413896	Xiaomi/小米 小米笔记本Air 13.3吋
TMALL_1714128138-545855523798	Xiaomi/小米 小米笔记本Air 13.3吋 尊享版
TMALL_1714128138-547580364460	Xiaomi/小米 小米笔记本AIR 12.5英寸 M3 4G 256G
TMALL_1730407557-539687125700	MACHENIKE 机械师F117 F6
TMALL_1730407557-543035827310	MACHENIKE 机械师F117 F1K
TMALL_1796610007-534475828739	Lenovo/联想 IdeaPad 310S-14ISK
TMALL_1917047079-530536855103	Apple/苹果 12 英寸 MacBook 256GB
TMALL_1917047079-530592178855	Apple/苹果 13 英寸 MacBook Air 1.6GHz 处理器 128 GB 存储容量
TMALL_1917047079-541510809037	Apple/苹果 13英寸：MacBook Pro Multi-Touch Bar 和 Touch ID 2.9GHz 处理器 256GB 存储容量
TMALL_1996270577-534078622940	Lenovo/联想 YOGA710 -14ISK
TMALL_1996270577-548650899077	Lenovo/联想 拯救者 Y520 I7
TMALL_2111250403-545157787483	THUNDEROBOT 911 Targa
TMALL_2111250403-546655767170	THUNDEROBOT 911 911
TMALL_2182235688-540240591403	炫龙 炎魔T1 Ti
TMALL_2182235688-547576365938	炫龙 炎魔T50 i7
TMALL_2248284179-543241983819	火影 金钢 T2
TMALL_2260266618-543432982068	Lenovo/联想 小新 510S
TMALL_2260266618-548494167433	Dell/戴尔 游匣 5576
TMALL_2484777365-529571326323	Asus/华硕 顽石 –
TMALL_2484777365-533738298376	Asus/华硕 A A
TMALL_2484777365-543706462897	Asus/华硕 R RX
TMALL_255921860-539367366683	Hasee/神舟 战神 Z7-SP5D1
TMALL_255921860-543437409299	Hasee/神舟 战神 Z7-KP7S1
TMALL_255921860-543489254259	Hasee/神舟 战神 Z7M-KP7S1
TMALL_2616970884-530945296812	Apple/苹果 MacBook Air MMGF2CH/A

platform	category	detail
TMALL	TMALL_手机,数码,电脑办公_电脑整机_笔记本	{update: 2017-05-17 15:56:12.231262, price: 4699.00}
TMALL	TMALL_手机,数码,电脑办公_电脑整机_笔记本	{update: 2017-05-17 15:56:11.766238, price: 3999.00}
TMALL	TMALL_手机,数码,电脑办公_电脑整机_笔记本	{update: 2017-05-17 15:56:12.074254, price: 4399.00}
TMALL	TMALL_手机,数码,电脑办公_电脑整机_笔记本	{update: 2017-05-17 15:56:12.199259, price: 4399.00}
TMALL	TMALL_手机,数码,电脑办公_电脑整机_笔记本	{update: 2017-05-17 15:56:11.174210, price: 7988.00}
TMALL	TMALL_手机,数码,电脑办公_电脑整机_笔记本	{update: 2017-05-17 15:56:11.581229, price: 4999.00}
TMALL	TMALL_手机,数码,电脑办公_电脑整机_笔记本	{update: 2017-05-17 15:56:11.101206, price: 5999.00}
TMALL	TMALL_手机,数码,电脑办公_电脑整机_笔记本	{update: 2017-05-17 15:56:11.799240, price: 3999.00}

图 6.2 数据库中商品信息截图

商品评论信息如图 6.3 所示：

id	rev_id	content
3921	TMALL_1669409267-44131265268_309105425641	物流快速 包装严实 价格实惠 质量不错 五星好评啦
3922	TMALL_1669409267-44131265268_309380561902	好评是给宝贝跟技术人员的，正品无疑，技术人员专业又耐心，我是个mac盲，装PS，AI一步步教我还是装的很费劲，最后让技术人员帮忙远程几分钟搞定。这里要特别感谢技术人
3923	TMALL_1669409267-44131265268_309115266066	苹果笔记本电脑外观漂亮上档次，开机发应速度快，发货快，客服一流。
3924	TMALL_1669409267-44131265268_310416387785	包装完好，验证正品！ 售后技术不错，能较快解决问题。对于物流不太满意！
3925	TMALL_1669409267-44131265268_310313335238	电脑用着不错 是新机器 还苏宁送货的 很好
3926	TMALL_1669409267-44131265268_309197396618	第一次买办公用笔记本考虑良久，最后还是选了一个很保值又很fashion的Macbook，杠杠的没得说 菜鸟果然够快啊，两天就到
3927	TMALL_1669409267-44131265268_309410426827	电池循环1次，应该是全新的，试用了2天，很不错，苹果的做工没得说，很喜欢，好评！！！
3928	TMALL_1669409267-44131265268_308858243084	第二次来买了值的过，包装很严实，快递速度不耽搁维啊，售后很棒，很耐心的解答，有问必答，机器没有毛病
3929	TMALL_1669409267-44131265268_309483485498	发票需要联系客服补寄，产品本身没有问题，价格很惊喜，五分好评
3930	TMALL_1669409267-44131265268_309354133332	绝对正品，物流也飞快，最重要是便宜呀
3931	TMALL_1669409267-44131265268_310390643375	在天猫商城买的，应该可以有保障。第一次用苹果电脑还不怎么习惯。话说电话过期是怎么会事，问了客服说过几就好了，这里有鬼。总的来说还是蛮不错的。
3932	TMALL_1669409267-44131265268_307757649468	一直心心念念想买很久了，你们家一直没让我失望。之前是平板电脑，这次是笔记本电脑，一如既往地正品，好用！祝生意兴隆？
3933	TMALL_1669409267-44131265268_309807048302	跟预想的一样，全新未拆封机子，不过发票要找客服才肯给，这里面的奥妙你们应该知道，苹果要有发票才给保修的哦
3934	TMALL_1669409267-44131265268_309671608257	质量有保障，发票也给我邮到了，苹果本做工真不是盖的，就是软件太少，有些操作需要习惯，买之前心里做好了学习成本的准备，还ok
3935	TMALL_1669409267-44131265268_309785661857	电脑小白，现在看看还不错，以后再来评论。
3936	TMALL_1669409267-44131265268_309290930299	正品应该没问题，用了俩天还可以吧，我想问下一条充电线还有一条线拿来干什么？
3937	TMALL_1669409267-44131265268_309414304816	电脑是正品，而且是2017年新机器，还没怎么用，慢慢熟悉中，挺好。
3938	TMALL_1669409267-44131265268_309452673888	Mac用着不错，也激活啥，官方正品，各个地方也查着过啥ok，价格也很美丽
3939	TMALL_1669409267-44131265268_310349355303	电脑没问题很好，也是原装未拆封，但发票怪怪的，希望不影响，物流速度快
3940	TMALL_1669409267-44131265268_309995126654	用了一段时间，还行，更新的时候卡了一下。其他的没什么问题，还有一个2015年的联机分享是什么鬼，客服出来解释一下。
3941	TMALL_1669409267-44131265268_292019281982	快点儿开发票
3942	TMALL_1669409267-44131265268_292461444314	好用 很快，希望发票快点安排 很好
3943	TMALL_1669409267-44131265268_292003752720	放弃新款，买这款。。省钱省钱。。。
3944	TMALL_1669409267-44131265268_291504860410	东西很好，很满意的购物，物有所值了
3945	TMALL_1669409267-44131265268_292119934642	没发现有什么问题
3946	TMALL_1669409267-44131265268_292611719261	不错，服务态度很好
3947	TMALL_1669409267-44131265268_309979466967	一次丰富纠结的购物 1. 本以为天猫商城的东西会相对便宜，所以没有调查就拍了。 比他家贵500-700左右。本想退货，但拿到手里就不愿放手了。 2.屏幕贴与键盘套其实没大
3948	TMALL_1669409267-44131265268_293259866910	宝贝我就不说，服务很不满意，发货慢，物流超慢，一起买的东西，这个最慢，问客服，客服只会敷衍，很差的购物体验
3949	TMALL_1669409267-44131265268_302159154931	此用户没有填写评论!
3950	TMALL_1669409267-44131265268_309196103681	好评！

author	date	rating	platform	key_words
梅***0	2017-05-02 13:13:53.000000	-1	TMALL	NULL
t***6	2017-05-04 14:09:57.000000	-1	TMALL	NULL
雨***酸	2017-05-02 11:46:25.000000	-1	TMALL	NULL
c***8	2017-05-12 13:53:12.000000	-1	TMALL	NULL
唐***啦	2017-05-11 20:48:25.000000	-1	TMALL	NULL
风***离	2017-05-02 13:23:05.000000	-1	TMALL	NULL
1***n	2017-05-04 11:30:40.000000	-1	TMALL	NULL
喜***忆	2017-05-01 08:21:36.000000	-1	TMALL	NULL
t***1	2017-05-04 11:57:51.000000	-1	TMALL	NULL
x***2	2017-05-04 09:43:11.000000	-1	TMALL	NULL
s***男	2017-05-11 20:52:18.000000	-1	TMALL	NULL
杜***t	2017-04-18 23:24:02.000000	-1	TMALL	NULL
t***9	2017-05-08 10:50:20.000000	-1	TMALL	NULL
淘***侠	2017-05-07 10:53:42.000000	-1	TMALL	NULL
被***啊	2017-05-07 18:12:16.000000	-1	TMALL	NULL
k***a	2017-05-04 12:26:39.000000	-1	TMALL	NULL
j***6	2017-05-03 20:50:15.000000	-1	TMALL	NULL
g***0	2017-05-05 11:35:30.000000	-1	TMALL	NULL
w***g	2017-05-12 20:40:27.000000	-1	TMALL	NULL
1***絮	2017-05-08 16:24:09.000000	-1	TMALL	NULL
黎***云	2016-11-16 20:25:40.000000	-1	TMALL	NULL
m***4	2016-11-19 11:20:42.000000	-1	TMALL	NULL
小***n	2016-11-17 10:19:44.000000	-1	TMALL	NULL
a***a	2016-11-14 16:42:01.000000	-1	TMALL	NULL
木***哥	2016-11-17 18:20:00.000000	-1	TMALL	NULL
t***2	2016-11-19 20:15:06.000000	-1	TMALL	NULL
画***泪	2017-02-09 16:55:49.000000	-1	TMALL	NULL
嘻***虫	2016-11-25 14:09:13.000000	-1	TMALL	NULL
w***4	2017-01-29 02:58:45.000000	-1	TMALL	NULL
g***n	2017-05-02 13:44:50.000000	-1	TMALL	NULL

图 6.3 数据库中商品评论信息截图

结 论

本文中的工作设计了一种能兼容多家 B2C 零售电商网站的数据组织和抓取方式，并用程序进行了一定程度的实现，为实现对电商数据的组织和管理提供了一种思路。同时，在已获取的数据样本上进行了观点挖掘，直观地展示了电商数据中所蕴含的信息的价值，体现了对于电商数据进行分析挖掘的重要性。

电商爬虫直接面临着复杂多变的网络环境，很难确保所获取的数据格式的持久不变，所以网络爬虫程序要能够及时发现异常，并找到问题发生的原因。在设计爬虫的时候，为了让爬虫有更强大的容错能力，并没有在网页获取阶段就对网页内容进行解析，而是优先将页面内容进行存储，这使得爬虫在访问页面时不必受固定的解析规则约束。将页面解析单独作为一个进程，当遇到无法顺利解析的页面时给出提示，这也使得系统能适应对数据库的修改，能有机会从原始界面中获取新的数据加入到数据库之中，而不必重新访问网站，但同时也会增加本地磁盘的存储开销，这在爬虫获取了足够大量的数据后会更为明显，为了减小由此带来的负面影响，可以将数据进行压缩然后再进行存储。

对于爬虫程序，可以适度提高并发量来提高抓取速度；可以设计管理模块来更好地监控爬虫的运行。同时可以修改爬虫同数据库的接口，用更高的抽象方法来构建函数，使爬虫能适应不同种类的数据库。

本文中的观点挖掘方法只是在分词和统计的基础上进行的，想要进一步发现潜藏在大量数据中的信息，可以使用更为高级的机器学习方法，对文本进行分类，获取商品的更多特征，发现消费者对不同商品的选择倾向；当获得更多种类的数据信息之后，针对电商数据的分析也可以从更多维度展开，而不只是仅仅限于商品评论。

参 考 文 献

[1] 提姆·奥莱理, 玄伟剑. 什么是 Web2.0[J]. 互联网周刊, 2005(40):38-40.

[2] 周德懋, 李舟军. 高性能网络爬虫:研究综述[J]. 计算机科学, 2009, 36(8):26-29.

[3] 杨颂. (2010). 面向电子商务网站的增量爬虫设计与实现. (Doctoral dissertation, 湖南大学).

[4] 王辉, 王晖昱, 左万利. 观点挖掘综述[J]. 计算机应用研究, 2009, 26(1):25-29.

[5] 章剑锋, 张奇, 吴立德,等. 中文观点挖掘中的主观性关系抽取[J]. 中文信息学报, 2008, 22(2):55-59.

[6] 孙立伟, 何国辉, 吴礼发. 网络爬虫技术的研究[J]. 电脑知识与技术, 2010, 06(15):4112-4115.

[7] 李盛韬，余智华，程学旗. Web 信息采集研究进展[J]. 计算机科学，2003.

[8] James. 互联网 OSI 七层模型详细解析[J]. 网络与信息, 2009, 23(9):44-44.

[9] 林华. Robots 协议维护互联网秩序[J]. It 时代周刊, 2014(17):50-50

[10] 袁径三. 浅说汉字编码[J]. 绍兴文理学院学报, 2005, 25(9):56-59.

[11]黄欣. 用触发器实现特殊参照约束下的级联操作[J]. 电脑知识与技术, 2010, 6(8):1942-1943.

[12] 龙树全, 赵正文, 唐华. 中文分词算法概述[J]. 电脑知识与技术, 2009, 5(4):2605-2607.

[13] Loper E, Bird S. NLTK: the Natural Language Toolkit[C]// Acl-02 Workshop on Effective TOOLS and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Association for Computational Linguistics, 2002:63-70.

[14]陈旻,朱凡微,吴明晖,应晶. 观点挖掘综述[J]. 浙江大学学报(工学版),2014,(08):1461-1472

[15] 朱卫祥.面向电子商务评论文本的观点挖掘系统研究与实现[D].浙江理工大学,2013.

[16] Pontiki M, Galanis D, Pavlopoulos J, et al. SemEval-2014 Task 4: Aspect Based Sentiment Analysis[J]. Proceedings of International Workshop on Semantic Evaluation at, 2014:27-35.

[17] 魏韡, 向阳, 陈千. 中文文本情感分析综述[J]. 计算机应用, 2011, 31(12):3321-3323.

[18] 柳位平, 朱艳辉, 栗春亮,等. 中文基础情感词词典构建方法研究[J]. 计算机应用, 2009, 29(10):2875-2877.

[19] 张学工. 关于统计学习理论与支持向量机[J]. Acta Automatica Sinica, 2000, 26(1):32-42.

[20] Witten I H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)[M]. 机械工业出版社, 2005.