华中农业大学本科毕业论文（设计）开题报告书

题　目	基于 Python 的主题爬虫设计与实现				
姓　名	郑泉石	**学　号**	2013310200408	**专　业**	计算机科学与技术
指导教师	王建勇	**职　称**	讲师	**学　位**	硕士
课题来源	自选				

科学依据（包括课题的科学意义；国内外研究概况、水平和发展趋势；应用前景等）

在网络高速发展，软硬件技术日新月异的今天，各行各业的人们都普遍依赖互联网来获得资讯信息和生活上的便利。同时也有越来越多依赖于网络的电商平台出现，电商平台的优势也日益明显。针对目前电商平台越来越多，商品类型，数量随之越来越繁杂，使用户无法方便“货比三家”的情况，为了方便用户更有效率的了解各电商平台的商品信息，本文提出设计一个面向电商平台的主题爬虫，也是相应电商管理系统的重要组成部分。目前网络爬虫技术已越来越受到重视，但是针对电商平台的主题爬虫存在使用不方便且普遍被掌握在提供付费爬取服务的厂商手中的问题，因此针对电商数据采集的主题爬虫研究是很有意义的。

当前电商平台现状为背景出发，通过分析相关爬虫知识，爬虫算法，电商平台现主要的反爬虫技术，相关应对反爬虫技术的措施与算法，爬取资讯处理等相关技术。设计了一个面向电商平台的网络爬虫，通过该爬虫系统可以对多种主流电商平台商品信息进行高效的采集与识别。

爬虫本身代码和应对反爬虫的技术是爬虫系统关键技术，因此本文主要对爬虫技术和“反反爬虫技术”进行分析和阐述，传统的电商主题爬虫大多通过调取电商平台 API，通过解析平台网页页面的 HTML 解析获得商品信息，但是如今的电商平台越来越注重于保护自己的数据，因此会刻意的在网页中加入动态加载，JS 动态生成信息或者加入认证环节或者发送经过压缩加密过的文本等等多种多样的反爬虫技术，并且反爬虫技术会时常更新变化，在很大程度上加大了爬取信息的代价和成本，因此传统的电商爬虫经常会出现爬取信息不全，爬取信息有误，或是根本无法爬取，或是陷入了爬虫陷阱，出现爬取效率低等等问题。而在经过长时间，持续，反复的对主流电商网站的研究与分析后。本文提出的算法有效解决了传统电商主题爬虫程序的使用不方便，爬取质量低的问题。通过对该电商主题爬虫的测试和与其他商用电商主题爬虫对比，发现本系统在爬取数据的正确率和支持平台数量上都有着明显的提高。

基于以上的研究，设计了基于 python 语言的面向于亚马逊，淘宝，速卖通，ebay 等国内外主流电商平台的主题网络爬虫系统，并在此基础上添加数据库认领模块，web 前端模块，形成一个完整，系统的电商主题爬虫系统。

研究内容

主要是对于以下内容的研究：

（1）本文提到的针对各个主流电商平台网站的网页结构设计的爬虫程序，主要目的是通过调用网站 API 或者模拟浏览器获得网页的的 json，xml 格式数据，或者网站 html 文件进行解析，获得商品的种子列表，进而使用多线程去分别抓取每一个商品的信息

（2）搭建电商爬虫系统，从整体的需求入手进行整体框架和功能模块的设计与实现，整体使用 SQL 数据库+python 后台+web 前端的模式。

（3）针对涉及到的主流电商平台的反爬虫技术的研究，包含但不限于 xhr 动态加载，js 动态生成，xsrf ，cookie 认证，gzip 压缩返回内容

拟采取的研究方法、技术路线、实验方案及可行性分析

（1）对网络爬虫的理论和相关技术的深刻的认识。对基础的网页解析算法，以及浏览器解析页面和渲染页面的机制的了解。

（2）通过对淘宝，亚马逊，ebay，wish 等网站的研究，对这些网站的结构和设计模式的了解。了解了他们从数据库中拿取数据，生成网页的方式与模式，对他们的安全机制的了解。

（3）了解主流电商网站的反爬技术，学习这些网站的安全部门人员为了保护公司的数据想出的各种巧妙的办法，同时提出针对这些反爬技术的措施，锻炼自己的编程思维，提高自己的编程水平。

（4）可行性：根据国内外的研究情况，爬虫是自动获取信息的绝佳选择。同时爬虫性能稳定，python 语言适合于工程编写。总体来看，使用爬虫达到目的，具有较高的可行性。

研究计划及预期成果

具体进度安排：

2017.3.15 - 3.30 进行商城数据采集，网站研究

2017.4.1 - 4.20 实现爬虫构建

2017.4.21 - 4.30 系统前端，数据库的构建

2017.5.1 - 6.1 毕业论文和系统调试

2017.6.1 - 毕业答辩

预期成果：

爬虫能系统能有充分的健壮性，爬取的数据有足够高的质量。信息足够多。

指导教师意见

指导教师签名：

年 月 日