# 网络爬虫可行性分析报告

**1 引言**

* 1. **编写目的**

**本文档旨在现有技术、操作等方面综合研究分析网络爬虫的可实现性，通过分析来预测网络爬虫在现实中的运行情况。**

* 1. **背景**

**说明：**

**项目名称：网络爬虫**

**任务提出者：软件工程课程**

**开发者：刘文进、付立、何发洪、胡海量**

**用户：具有相应基础知识的开发人员**

**网络爬虫可运行于Windows、Linux、Mac、Unix平台上，爬虫使用Python的应用框架Scrapy。**

**定义：**

**爬虫：网络爬虫是一种自动获取网页内容的程序，是搜索引擎的重要组成部分。**

**Python：是一种面向对象、解释型计算机程序设计语言。**

**框架（Framework）：是一个基本概念上的结构，用于去解决或者处理复杂的问题。这个广泛的定义使用的十分流行，尤其在软件概念。**

**结构化数据：简单来说就是数据库**

**数据挖掘：它是数据库知识发现（英语：Knowledge-Discovery in Databases，简称：KDD)中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。**

**API：（Application Programming Interface,应用程序编程接口）是一些预先定义的函数，目的是提供应用程序与开发人员基于某软件或硬件得以访问一组例程的能力，而又无需访问源码，或理解内部工作机制的细节。**

**参考资料**

[**Python Tutorial （Python 官方文档）**](https://docs.python.org/3.5/tutorial/)

[**Scrapy 主页**](scrapy.org)

**Python基础教程 作者:** [**Magnus Lie Hetland**](http://book.douban.com/search/Magnus%20Lie%20Hetland) **出版社: 人民邮电出版社**

**Python Cookbook 作者:** [**Alex Martelli**](http://book.douban.com/search/Alex%20Martelli)**，[Anna Ravenscroft](http://book.douban.com/search/Anna%20Ravenscroft)，[David Ascher](http://book.douban.com/search/David%20Ascher) 出版社: 人民邮电出版社**

**2 可行性分析的前提**

**2.1 要求**

**1.功能**

**爬取网页数据（图片、文字、音视频）**

**批量爬取数据**

**定时自动爬取**

**统计数据信息**

**针对网站的反爬虫策略做出有效应对**

**2.性能**

* **数据的准确性和即时性：**

**数据作为一种信息，准确性和时效性是必须重视的。**

**比如，拉取下来的成绩与姓名不匹配或火车票销售信息不及时都将毫无作用。所以，编写爬虫时特别注意程序的算法准确及算法的高效。**

* **对突发事件处理：**

**网站可能由于服务器的维护或突发故障而暂停服务，所以应考虑爬取的网站不能访问时，爬虫所应该采用的应急策略。**

* **网络爬虫的开放性和爬虫的可扩充性**

**为了方便用户，对于开放性，网络爬虫应该给用户提供一些可修改的参数。而爬虫或许可以支持插件，比如，嗅探器插件。**

* **系统的易用性和易维护性**

**网络爬虫将不会有GUI（图形用户界面），这是考虑到GUI的编写有一定的技术难度，以及本爬虫有一定针对性而导致可能会经常修改代码。**

**易维护性，这是一个小的脚本文件，并不是大型系统，所以易维护性它最大的优势**

* **系统的先进性**

**目前计算机技术发展相当快，作为网络爬虫，使用的是Python这个热门的优秀的脚本语言与Scrapy这个优秀的应用框架，系统的先进性是毫无疑问的。**

**3.输出：**

**拉取下来的数据可以在命令提示符窗口显示，也可以用纯文本文档或图片，视频，音频等形式存储在本地。**

**4．输入：**

1. **命令提示符窗口中输入**
2. **也可以读取纯文本文档**
3. **执行python脚本文件**

**3对现有系统的分析**

**3.1处理流程和数据流**

网页信息

HTTP请求

**目标网站**

**解释器**

**脚本文件**

**本地计算机**

**在网络连通的情况下，解释器读取并解释python脚本文件，然后向目标网站发送HTTP请求，并分析处理来自网站返回的网页信息，以相应文件格式存在本地计算机或数据库。**

**3.2工作负荷**

**系统将运行python解释器，并占用部分网络资源，如果需要将数据存在数据库，还将会运行MySQL或其他类型数据库。由于python支持多线程操作，解释器运行时将会调用多线程执行任务，会占用部分系统资源。**

**3.3费用开支**

**暂时不会产生费用开支**

**3.4人员**

**由于知识水平有限，目前我们写出的python爬虫程序将不会很复杂，所以使用和维护都很容易。如果需要数据库相关操作，将会至少需要两人负责维护，一人负责爬虫程序，一人负责数据库相关操作。如果不需要数据库操作，一个有python基础的人就可以完成维护工作。**

**3.5设备**

**一台正常工作且能联网的Windows，Mac，Linux电脑。**

**3.6局限性**

* **由于本爬虫程序处理的内容是网页，所以必须要在网络通畅的情况下才能正常使用爬虫的功能，且受限于运行环境的网速。**
* **由于爬虫本身的特性，具有很强的针对性，一个爬虫程序只能爬取特定网站的特定类型的网页，不能普遍适用于每一个网站。如果要爬取其他网站的内容，必须修改爬虫程序代码。**
* **需要定期查看目标网站网页布局是否改变，如果改变，将可能会修改爬虫程序**

**4所建议的系统**

**暂时无**

**5可选择的其他系统方案**

**暂时无**

**6投资及效益分析**

**暂时未计算**

**7社会因素方面的可行性**

**7.1法律方面的可行性**

**本爬虫程序将会爬取用户指定的特定网站的特定的内容，且所有内容和操作均在法律允许的范围内，搜集的内容均为网站公开的内容，不会涉及第三方财产及利益纠纷，不会对目标网站和他人带来任何损失。搜集的所有信息不会拿用作任何非法活动。**

**7.2使用方面的可行性**

**本爬虫程序的工作不会为他人或自己带来任何不良影响。可适用与普通用户的简单信息的搜集，供娱乐只用。也可用于商业用途，大量搜集互联网信息并分析挖掘，可实现一定商业价值。**

**8结论**

**本爬虫程序预计编写周期为这一学期，所使用的编程语言为python以及一些python框架，我们将会在一段短暂的python基础学习之后，着手开始编写本项目。**

**其实，能够完成本次网络爬虫项目并不是我们最终的目的，而且，python语言所能实现的功能，远不是一个爬虫程序所能表现的，我们的目的在于通过这次网络爬虫项目的经历，能够了解并深入学习python这门语言，为以后编写更大的项目打好基础。**