可配置网络爬虫开题

1. **研究背景**

近年来随着网络、数据挖掘、人工智能等技术的快速发展，人们对如何快速高效的从互联网里提取有价值的数据变得越来越感兴趣。这给现代的爬虫技术带来巨大的挑战。传统的爬虫技术对各种不同类型的网站进行分析，进而实现不同的数据抓取策略，实现不同的爬虫逻辑。这种方法不但使得爬虫的开发耗费大量的时间，而且爬虫的监控与维护也变得十分麻烦。

根据不同的应用，一个成熟的爬虫系统存在着多方面的差异，大体而言，可以将爬虫分为如下三种类型：

1. 批量型爬虫：批量型爬虫有比较明确的抓取范围和目标，当爬虫达到这个设定目标后，即停止抓取过程。至于具体目标可能各异，也许是设定抓取一定数量的网页即可，也许是设定抓取消耗的时间等。
2. 增量型爬虫：于批量型爬虫不同，会保持持续不断的抓取，对于抓取到的网页要定期更新，因为互联网的网页处于不断地变化之中，新增网页、网页删除或者网页内容的更改都很常见，而增量型爬虫要及时反映出这种变化，所以处于持续不断的抓取过程中，不是在抓取新网页，就是在更新已有的网页。通用的商业搜索引擎爬虫基本属于此类。
3. 垂直型爬虫：垂直型爬虫关注特定的主题内容或者属于特定行业的网页，比如对于健康网站来说，只需要从互联网页里找到与健康相关的内容页面即可，其他行业的内容不再考虑范围内。垂直型爬虫一个最大的特点和难点就是：如何识别网页内容是否属于指定行业或者主题。从节省系统资源的角度来说，不太可能吧所有互联网页面下载下来之后再去筛选，这样浪费资源就太过分了，往往需要爬虫在抓取阶段就能够动态的识别某个网址是否与主题相关，并尽量不去抓取无关的页面，以达到节省资源的目的。垂直搜索网站或者垂直行业网站往往需要此类型的爬虫。

网络爬虫是一个自动提取网页的程序，它为搜索引擎从web上下载网页，是搜索引擎的重要组成。

  传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列，直到满足系统的一定停止条件。搜索引擎是基于传统爬虫技术建立的，但其存在着一定的局限性，例如：

(1) 不同领域、不同背景的用户往往具有不同的检索目的和需求，通用搜索引擎所返回的结果包含大量用户不关心的网页。

(2)通用搜索引擎的目标是尽可能大的网络覆盖率，有限的搜索引擎服务器资源与无限的网络数据资源之间的矛盾将进一步加深。

(3)万维网数据形式的丰富和网络技术的不断发展，图片、数据库、音频、视频多媒体等不同数据大量出现，通用搜索引擎往往对这些信息含量密集且具有一定结构的数据无能为力，不能很好地发现和获取。

(4)通用搜索引擎大多提供基于关键字的检索，难以支持根据语义信息提出的查询。

因此，垂直领域的搜索引擎越来越引起大家的关注，而垂直领域的搜索引擎的数据源就来自于垂直领域的爬虫，但是，由于传统的垂直领域的网络爬虫开发模式，对各种不同类型的网站进行分析，进而实现不同的数据抓取策略，实现不同的爬虫逻辑将导致很多的重复性工作，因此我们计划开发一个可配置化的网络爬虫，来帮助爬虫开发者乃至一些无爬虫开发经验的用户使用爬虫获取数据。

**2.研究意义**

针对以上问题，本文采用HTTPClint，Jsoup等技术， 开发自己的框架。实现一个可配置化的垂直领域爬虫。可以很好的解决传统爬虫存在的一些问题。将传统爬虫仅限于爬取固定网站改成为可配置的网络爬虫。可配置的网络爬虫可以很好的解决在编写爬虫是一些繁杂的重复操作，只需要修改配置化的方法就可以爬取一个网站。采用xml进行配置文件，后期将考虑转用javascrip，前端页面将实现可视化图形配置和资源监控，显示出CPU占比和资源的占比等，不同的网站类型显示出不同的选项。

**3.技术分析：**

（1）HTTPClient简介

HTTP 协议可能是现在 Internet 上使用得最多、最重要的协议了，越来越多的 Java 应用程序需要直接通过 HTTP 协议来访问网络资源。虽然在 JDK 的 java.net 包中已经提供了访问 HTTP 协议的基本功能，但是对于大部分应用程序来说，JDK 库本身提供的功能还不够丰富和灵活。HttpClient 是 Apache Jakarta Common 下的子项目，用来提供高效的、最新的、功能丰富的支持 HTTP 协议的客户端编程工具包，并且它支持 HTTP 协议最新的版本和建议。HttpClient 已经应用在很多的项目中，比如 Apache Jakarta 上很著名的另外两个开源项目 Cactus 和 HTMLUnit 都使用了 HttpClient。

（2）HTTPClient功能

实现了所有Http的方法，支持自动转向，支持HTTPS协议，支持代理服务器等。应用HTTPClient来对付各种顽固的WEB服务器。可以用来读取网页内容，使用post方式提交数据，用于重定向机制，提交XML格式参数。最重要的一点是它可以在多线程模式下使用，为程序增加了并发。

1. Jsoup

Jsoup是一款java的HTML解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法和操作数据。他的主要功能有：从一个URL，文件或字符串中解析HTML，使用DOM或CSS选择器来查找、取出数据，是可操作性的HTML元素、属性、文本。Jsoup是基于MIT协议发布的，可放心使用于商业项目。

