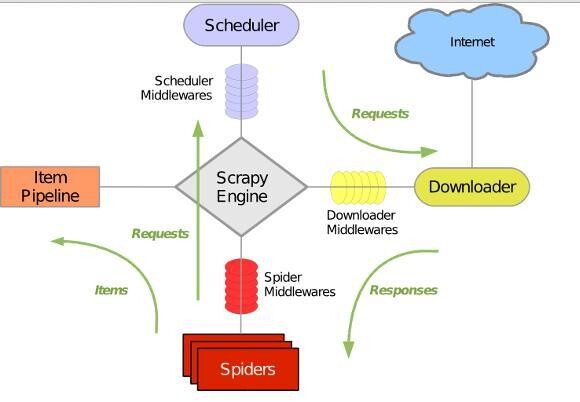
2**相关知识及技术介绍**

**2.1网络爬虫**

**2.1.1 什么是网络爬虫**

简单的说，网络爬虫就是一个在互联网上抓取信息的程序，一个爬行程序，一个抓取[网页](http://baike.baidu.com/view/828.htm)的程序。如果我们把互联网形象地比喻成一张蜘蛛网，抓取程序就像是在互联网上爬来爬去的蜘蛛，网络蜘蛛的名称就是由此得来。

网络蜘蛛通过分析[网页](http://baike.baidu.com/view/828.htm)结构，提取出页面包含的URL地址来寻找网页，从一个种子页面开始（导航网站或者门户首页会比较好），下载并解析网页的内容，找到网页中存在的其它超链接地址，再根据这些链接地址继续下载解析，找到下一个网页，这样一直循环下去，直到没有需要抓取的网页为止。整个互联网好比是一个网站，网络蜘蛛就是通过这种方式抓取了整个互联网上的[网页](http://baike.baidu.com/view/828.htm)，如图2-1所示。



**图2-1传统爬虫示意图**

**2.1.2 网络爬虫的分类**

**①通用爬虫**

一般的爬虫框架流程为：从互联网海量页面中先抓取一些高质量页面，抽取其中所包含的url，将这些URL放入待抓取队列中，爬虫依次读取该队列中的url，通过DNS解析，将这些url转化成对应网站的IP地址，网页下载器则通过IP地址下载页面所有内容。

对于已下载到本地服务器的页面，一方面等待建立索引以及后续处理；另一方面将这些已下载的页面记录下来，以免再次被抓取。  
对于刚下载的页面，从页面中抓取页面中所包含未被抓取的URL，放到待抓取队列中，后续抓取过程中将下载该URL对应的页面内容，如此循环，知道待抓取队列为空，则完成一轮抓取。

当然，在如今互联网信息不断海量的时代，为了保证效率，爬虫一般会持续不断的工作。

所以，由此从宏观上我们可以了解，互联网的页面可以划分为以下5各部分：

a)已下载页面集合

b)已过期页面集合

c)待下载页面集合

d)可知的页面集合

e)不可知的页面集合

当然，为了确保页面质量，在以上的一个爬虫抓取过程中还涉及了很多技术手段。

**②聚焦爬虫**

网络爬虫是一个自动提取网页的程序，它为[搜索引擎](http://baike.baidu.com/view/1154.htm)从[万维网](http://baike.baidu.com/view/7833.htm)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。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从[队列](http://baike.baidu.com/view/38959.htm)中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于聚焦爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

相对于通用网络爬虫，聚焦爬虫还需要解决三个主要问题：

(1) 对抓取目标的描述或定义；

(2) 对网页或数据的分析与过滤；

(3) 对URL的搜索策略。

抓取目标的描述和定义是决定网页分析算法与URL搜索策略如何制订的基础。而网页分析算法和候选URL[排序算法](http://baike.baidu.com/view/297739.htm)是决定[搜索引擎](http://baike.baidu.com/view/1154.htm)所提供的服务形式和爬虫[网页抓取](http://baike.baidu.com/view/3692511.htm)行为的关键所在。这两个部分的算法又是紧密相关的。

**③智能爬虫**

垂直搜索的核心技术实际上就是[智能爬虫](http://baike.baidu.com/view/1353563.htm)的技术，也就是说如何将定向或者非定向的[网页抓取](http://baike.baidu.com/view/3692511.htm)下来并进行分析后得到格式化数据的技术。那么衡量一个[垂直搜索引擎](http://baike.baidu.com/view/7491.htm)的好坏主要有以下几个标准。

A.数据的更新频率

顾名思义，就是爬虫从目标网站上爬取数据的频率。

B.覆盖网站个数

覆盖尽量多的网站，对提供的信息数量将是一个保证。

C.单站有效数据抓取率

单个目标网站的有效数据，对数据量的多少有直接的影响。衡量一个爬虫的重要标准之一。

D.[信息抽取](http://baike.baidu.com/view/1133937.htm)完整率和准确率

此项指标的重要度不言而喻。信息的准确率和完整率直接关系到整个搜索引擎搜索结果的质量。

经过发展现有垂直搜索爬虫分为2种基本模式。

一、定向爬虫获取信息，配上手工或者自动的模版进行信息匹配，将信息进行格式化分析存储。

优势：

基于模版的信息提取技术，能提供更加精准的信息。比如价格，房屋面积，时间，职位，公司名等等。

劣势：

目标网站难以大面积覆盖，因为基于模版匹配的信息提取技术，需要人工的参与配置模版，欲要大面积覆盖各个目标网站，需要大量的人力成本，同样维护模板也需要很大的人力成本。

二、语义爬虫全网爬取，爬虫根据语义识别，自动进行信息格式化分析，并存储。

优势：

1、全网非定向抓取目标网站，有效的保证信息数量。

2、不需要人工参与定制和维护模板，有效的保证了自身的人力和维护成本。

劣势：

相对于第一种模板匹配，根据语义来进行数据抓取，准确率略有下降。

**2.2 网页解析技术**

**2.2.1 网页解析技术的基本概念**

信息抽取 （Information Extraction: IE）是把[文本](http://baike.baidu.com/view/300107.htm)里包含的信息进行结构化处理，变成[表格](http://baike.baidu.com/view/899068.htm)一样的组织形式。输入信息抽取系统的是原始文本，输出的是固定格式的信息点。信息点从各种各样的文档中被抽取出来，然后以统一的形式集成在一起。这就是信息抽取的主要任务。信息以统一的形式集成在一起的好处是方便检查和比较。 信息抽取技术并不试图全面理解整篇文档，只是对文档中包含相关信息的部分进行分析。至于哪些信息是相关的，那将由[系统设计](http://baike.baidu.com/view/170106.htm)时定下的领域范围而定。

因特网提供了一个巨大的信息源。这种信息源往往是半结构化的，虽然中间夹杂着结构化 和自由文本。网上的信息还是动态的，包含超链接，以不同的形式出现，而且跨网站和平 台，全网共享。因此，因特网是一个特殊的挑战，直推动着从结构化和半结构化文本中 抽取信息的研究向前迈进。

网页的结构化程度总是取决于用户想要抽取的属性是什么。通常，机器产生的网页是非常 结构化的，手工编写的则结构化程度差些，当然有很多例外。 传统的NLP技巧对抽取半结构化文本的信息并不是很有用，因其缺少规范的语法结构，而且 ，NLP方法的处理速度通常比较慢，这对于网上海量信息来说是一个大问题。 网上大部分内容都以属性列表的形式呈现，例如很多可搜索的网页索引。这种外观上的规律性可被利用来抽取信息，避免使用复杂的语言学知识。

**2.2.2 网页解析技术的常用方法**

**①基于树结构的网页信息抽取技术**

利用TagNode算法构造的网页结构树，可以把网页信息的抽取映射为在网页结构树中信息的查找。这里设计了抽取手机信息的网页信息抽取器。通过对不同网站的手机信息网页的分析，发现具有这样的规律：要抽取的手机型号和手机价格通常在同一个table中，并且手机型号和手机价格间存在一对一或一对多的关系。因而可采用如下启发式规则进行网页信息抽取。启发式规则为：待抽取的信息的各部分通常在同一个table中，并且它们之间存在一对一或一对多的关系。

在进行信息抽取前，首先建立模式库。模式库包含待抽取信息的表述、特征项等。抽取信息时，在网页结构树中定位模式库中的待抽取信息，用模式库中的待抽取信息和网页结构树的叶结点对应的网页信息进行匹配，如果匹配成功，则找到了一部分要抽取的信息。利用上述启发式规则，其他待抽取信息对应的结点和这个已匹配的叶结点在同一个table中。因此，模式库中的待抽取信息或特征项只需和已匹配的树叶结点的兄弟结点对应的网页信息比较，通常就可匹配成功，完成信息的抽取。

**②基于包装器的信息抽取方法**

基于包装器的信息抽取包装器由一系列的抽取规则以及应用这些规则的程序代码组成。通常,一个包装器只能处理一种特定的信息源。从几个不同信息源中抽取信息,需要一系列的包装器程序库。形式化地,每一类Web页面对应一个包装器。

与自然语言处理方式比较,包装器较少依赖于全面的句子语法分析和分词等复杂的自然语言处理技术,更注重于文本结构和表格格式的分析。但是由于一个包装器只能处理一种特定的信息源,当出现一类新的Web页面或旧的页面结构发生变化,原来的包装器就会失效,无法从数据源中获得数据或得到错误的数据,缺乏对数据语义的主动理解。

**③基于Ontology的信息抽取**

基于Ontology的方法主要依赖一个完整的知识库。知识库定义了各个元素的抽取模式,还有它们之间的联系。在抽取之前,需要将包含数据的记录块分隔开来,然后依次对每个记录块进行信息抽取。抽取模式没有使用依赖于特定文档的分隔符或者词性这样的自然语言理解技术,而是主要使用通用的词法模式。这种方法不依赖于任何结构和表现形式,它使用Ontology来定位关键信息并使用这些元素构造对象。不过,这事先需要构造一个完整的Ontology库,而这要由专家花费很长时间。因为,基于自然语言理解的方法和基于包装器的方法都需要人为给定样本来学习抽取规则。而给定样本是一项费时的工作。

**④基于去除html标签的网页正文提取算法**

基于去除html标签的网页正文提取算法是一种不必建立DOM树的网页正文提取算法 。该算法的主要思想是首先去除html标签，根据去除html标签后的文字密度判断出正文区域。最后将所有的正文区域合并，取得网页正文内容。下面详细介绍具体的算法流程

第一步首先需要去除html标签和javaScipt代码。采用正则表达式与栈的联合使用方法去除html的标签和javaScipt代码。例如对于<script>…</scripte>这种类型的标签，使用正则表达式<script>:?</script>去除标签以及标签中的内容，对于<div>…<div>…</div><./div>这95 种类型，使用栈，完成按顺序取出html中的非标签文字的需求。

第二步根据去除html标签后的文字密度获得正文区域。对于去除html标签后的文字，判断每一行文字的长度，判断每一行的文字数，如果没有两个以上连续的文字长度为零，继续执行该操作，直到遇到至少连续两行出现文字长度为空的情况，这时判断该过程中取出的

文字的总长度是否超过某阈值，如果超过阈值，该部分为网页正文，否则不是。按照上面的100 方法继续执行，直到文件的结尾处结束。

第三步将所有的正文区域合并，取得网页正文内容。具体为按照第二步获得的所有正文区域的先后顺序合并所有得到的正文区，得到最终的网页正文内容。

**2.3相关技术介绍**

## Nutch

Nutch 是 Apache 基金会的一个开源项目，它原本是开源文件索引框架 Lucene 项目的一个子项目，后来渐渐发展成长为一个独立的开源项目。它基于 Java 开发，基于 Lucene 框架，提供 Web 网页爬虫和搜索引擎两部分功能。

Nutch主要分为两个部分：网页爬虫（Crawler）和搜索引擎（Searcher）。Crawler主要用于从网络上抓取网页并为这些网页建立索引。Searcher主要利用这些索引检索用户的查找关键词来产生查找结果。另外很吸引人的一点在于，它提供了一种插件框架，使得其对各种网页内容的解析、各种数据的采集、查询、集群、过滤等功能能够方便的进行扩 展，正是由于有此框架，使得 Nutch 的插件开发非常容易，第三方的插件也层出不穷，极大的增强了 Nutch 的功能和声誉。

## Heritrix

**Heritrix**，是[www.archive.org](http://www.archive.org/) 上的开源产品，出色之处在于它的可扩展性，可以扩展它的组件，来实现自定义的抓取逻辑，但配置比较复杂。Heritrix在抓取中可以获取完整、精确的站点内容的复制。包括文本和非文本信息，抓取并存储内容，不对页面进行内容上的修改。在抓取速度上慢、需要大容量存储空间和宽速网络。重新爬行对相同的URL不进行替换。Heritrix 有 Web 控制管理界面，进行启动、监控。

## larbin

larbin是一种开源的网络爬虫/网络蜘蛛，由法国的年轻人Sébastien Ailleret独立开发，用c++语言实现。larbin目的是能够跟踪页面的url进行扩展的抓取，最后为搜索引擎提供广泛的数据来源。 Larbin只是一个爬虫，也就是说larbin只抓取网页，至于如何parse的事情则由用户自己完成。另外，如何存储到数据库以及建立索引的事情 larbin也不提供。

latbin最初的设计也是依据设计简单但是高度可配置性的原则，因此我们可以看到，一个简单的larbin的爬虫可以每天获取500万的网页，实在是非常高效。 　利用larbin，我们可以轻易的获取/确定单个网站的所有联结，甚至可以镜像一个网站；也可以用它建立url 列表群，例如针对所有的网页进行 url retrive后，进行xml的联结的获取。或者是 mp3，或者定制larbin，可以作为搜索引擎的信息的来源。

# **3 系统设计**

## 3.1 功能结构设计

整个互联网信息定向采集与智能解析系统分三大部件：任务监控管理平台、网页信息定向采集器和网页内容智能解析器，其功能结构如图3-1所示。



**图3-1 系统功能结构图**

（1）任务监控管理平台

任务监控管理平台主要实现对整个采集和解析任务进行监控和管理。

（2）网页信息定向采集器

网页信息定向采集器主要实现网页的采集功能。

（3）网页内容智能解析器

网页内容智能解析器主要对采集的网页进行智能分块、去噪，抽取网页正文和其它关键信息等。

## 3.2 任务监控管理平台设计

（缺图）

## 3.2.1任务监管平台与网页采集（解析）器间的通信模式设计

任务监管平台与任务采集（解析）器间通过共享数据库数据进行信息通信，其通信模式如图4-7所示。



**图4-7 通信模式**

（1）任务监控平台将任务配置信息、控制指令等数据写入任务监管数据库，并从任务监管数据库中读取任务状态信息和网页采集的统计数据，并显示在任务状态页面中；

（2）采集解析监管服务程序是任务监管平台和网页信息采集和解析程序间的桥梁。采集解析监管服务程序定期（每隔3秒）扫描任务监管数据库中其对应任务的控制信息，并分析其控制指令，根据控制指令操作（启动、停止）网页信息采集和解析程序，并将相应的状态信息写入任务监控数据库。

## 3.2.2 任务监管平台的功能结构设计

任务监管平台的功能结构如图4-8所示。主要包含任务配置、任务状态浏览和任务控制三大功能模块。



**图4-8 任务监管平台的功能结构**

（1）任务配置

任务配置功能主要实现对采集任务的属性信息和该任务进行采集的网站站点信息进行配置、管理和维护。

任务属性信息主要包括：

任务编号、任务名称、任务描述、采集进程数、解析进程数、数据存放服务器、自动更新服务器、是否保留网页源文件、存放网页源文件服务器、采集间隔时间（单位为分钟）。

采集网站的站点信息包括：

任务编号、站点编号、站点名称、站点类别、站点归属地、站点优先级（高、中、低）、入口网址（可采用表达式匹配方式）、采集深度、采集宽度、用户名、密码

（2）任务控制

任务控制主要实现对指定任务进行启动、停止操作、清空本任务已经采集的数据。

（3）任务状态浏览

任务状态浏览功能主要实现两大功能：

A．任务状态查看：查看当前任务的状态：执行中、停止（完成、错误中断、用户中断）；

B．采集数据统计：统计显示本次任务采集的数据记录数和累计采集的数据记录数，最近采集活动启动时间、最近采集活动结束时间和最近采集活动持续时间。

## 3.2.3 采集解析监管服务程序的功能结构设计

采集解析监管服务程序主要功能包括采集程序控制和任务监听两大模块，其功能结构如图4-9所示。



**图4-9 采集解析监管服务程序的功能结构**

（1）任务监听

任务监听模块定期（每隔3秒）扫描任务监管数据库，分析任务指令数据的变化，根据指令数据调用爬虫控制模块控制网页信息采集与解析程序。指令数据主要分两类：

用户操作指令：启动、停止、清空数据；

自动操作指令：当最近采集活动结束时间已经超过采集间隔时间，自动启动采集任务。

（2）采集程序控制

采集程序控制模块主要控制网页信息采集与智能解析程序，主要控制动作包括：

（1）启动：启动网页采集与解析程序进行网页采集，并在任务监管数据库中记录本次采集活动开始时间，修改任务状态为“进行中”；

（2）停止：停止网页采集与解析程序进行网页采集活动，并在任务监管数据库中记录本次采集活动结束时间；修改任务状态为“已停止”，并记录停止原因；将本次采集的数据量和累计采集的数据量记录到数据库。停止前需备份各个内存数据库。

## 3.3网页信息定向采集器

## 3.3.1 功能描述

网页信息定向采集器是根据任务监控管理平台的采集任务安排，抓取指定URL的网页文件，对网页内的新URL进行提取，作为进一步采集的对象，并将采集到的网页文件传递给网页内容智能解析器进行网页正文和关键信息解析抽取。

## 3.3.2 系统流程设计

系统工作流程如图4-1所示，其中对获取的web网页进行URL解析（新URL抽取）工作的流程如图4-2所示。



**图4-1 系统工作流程**



**图4-2 解析新URL工作流程**

## 3.3.3. 系统功能结构设计

网页信息定向采集器的系统功能结构如图4-3所示。



**图4-3 网页信息定向采集器的功能结构**

系统主要包括以下几大模块：

（1）任务初始化

任务初始化模块的主要功能是对几大URL列表数据的初始化，系统将根据任务启动的类别，初始化网页采集任务：

**系统完全启动：**

待抓URL列表从站点配置数据库中导入；

已抓URL列表和错误URL列表从内存数据库文件中恢复；

定向抓取下限定URL表从站点配置数据库中导入。

**从暂停和错误中断后启动：**

待抓URL列表从内存数据库文件中恢复；

已抓URL列表和错误URL列表从内存数据库文件中恢复；

定向抓取下限定URL表从站点配置数据库导入。

（2）采集任务控制

系统器采用多线程抓取任务分配，进行多线程网页抓取。采集任务控制模块主要实现对各采集进程的任务安排和采集过程监管，及其各进程间的数据共享控制等。系统将根据待抓URL列表数据和多线程限制数进行采集任务分配。当待抓URL列表为空时，并且分配给各进程的采集任务都已完成时，停止网页信息定向采集器工作，并向采集任务监管平台反馈相应信息（任务完成时间和抓取的数据记录数）。

（3）网页编码识别

网页识别模块将自动识别抓取网页的编码方式。系统通过httpclient 模拟浏览器方式访问网页，获取服务器返回的信息，从中获取网页编码方式。如编码为空，则获取网页源文件中head部分中页面编码，作为网页编码方式。

（4）网页下载

网页下载模块的功能是下载指定URL的网页文件。在实现上主要是利用 Java 中的 URLConnection 类来打开 URL 对应页面的网络连接，然后通过 I/O 流读取其中的数据。

（5）URL提取

URL提取模块的重要功能是分析已经下载的网页文件源码，解析出网页中的URL，并提取出来，主要包含URL抽取和URL解析两大功能。URL抽取就是分析网页的HTML源码，抽取其中的URL连接字符串；URL解析就是分析抽取出来的URL连接字符串，将相对地址的URL转换成绝对地址。

（6）URL过滤

URL过滤就是对解析出来的URL连接进行合法性分析，过滤非法和错误的URL，主要操作功能有非法性过滤、去重过滤和非定向过滤。

URL定向过滤就是将待抓URL限定在指定的网站范围内，即信息抓取只在限定站点内进行。为了把信息采集限制在一批站点范围内，需要截断指向站外的链接,保证爬虫总在站内执行,即准确地根据URL判断该链接是否指向站外。

URL去重过滤就是过滤掉已在待抓URL、已抓URL和错误URL列表中的ULR连接，采用布隆过滤器（Bloom Filter）技术进行URL去重过滤操作。

（7）URL存储

为了到达去重、定向抓取（限定抓取范围）、排除错误URL列表和多线程抓取目的，需要保存多个URL数据列表，分别是待抓取URL列表、已抓取URL列表、错误URL列表和定向匹配URL列表。URL存储模块主要实现这个4个URL列表的管理功能，待抓URL列表和已抓URL列表管理采用内存数据库模式进行管理，错误URL列表和定向匹配URL列表采用普通列表管理模式进行管理。

## 3.4网页内容智能解析器

## 3.4.1 功能描述

网页内容智能解析器是对已经获取的网页进行网页内容自动识别和抽取，主要工作包括对网页进行自动分块，智能去噪，抽取网页正文和其它关键信息（标题、文章出处、发布时间、作者等），并将其转换成结构化数据存入到指定数据库，供后续模块调用。

## 3.4.2 系统流程设计

网页内容智能解析器的工作流程如图4-4所示。



**图4-4 网页内容智能解析器工作流程**

其中网页正文抽取采用基于行块分布函数的通用网页正文抽取方法，其工作流程如图4-5所示。



**图4-5 网页正文抽取流程**

## 3.4.3 功能结构设计

网页内容智能解析器的系统功能结构如图4-6所示。



**图4-6 网页内容智能解析器的功能结构**

系统主要功能模块包括：

（1）网页预处理

网页预处理模块主要对网页文件进行去噪处理，过滤掉一些广告等垃圾信息。网页预处理的操作主要包括：

编码转换：自动对网页编码进行识别，并按照统一要求自动进行编码转换；

页面清洗：即页面去噪，利用正则表达式过滤掉网页中一些脚本和特殊字符等“垃圾信息”。

（2）网页正文抽取

网页正文抽取模块主要利用基于行块分布函数的通用网页正文抽取技术抽取网页的正文信息。主要操作包括：计算行块字数的分布函数、计算骤升点和骤降点、提取网页正文。该抽取方法主要针对新闻类网页和博客网页，论坛和贴吧类网页采用其它方法。

（3）网页标题抽取

网页标题抽取模块通过扫描HTML源码文件，提取出网页的标题信息。

（4）网页关键信息抽取

对于一些指定网站的网页信息，需提取一些特殊的关键信息，如新浪新闻网页中的“新闻发布时间”、“新闻来源”、“参与人数”等。对于特殊信息的提取，需要调用网页关键信息提取模块进行详细提取，无需提取特殊信息要求的网页可以直接跳过该步骤。关键信息抽取模块采用解析网页DOM树技术进行关键信息抽取。

## 3.5数据库设计

任务监管数据库TaskManagement，包括两张表：

（1）表（TaskInfo）：存放采集任务信息，表结构如表3-1所示

**表3-1 TaskInfo**

|  |  |  |  |
| --- | --- | --- | --- |
| **字段名** | **类型（长度）** | **是否允许为空** | **说明** |
| TaskID | 自动增长 | 否 | 主键，任务编号 |
| TaskName | varchar(50) | 是 | 任务名称 |
| TaskDescription | varchar | 是 | 任务描述 |
| GatherThead | int | 是 | 最多采集进程数 |
| InfoDB | varchar(50) | 是 | 索引服务器ID |
| SourceFileSave | bit | 是 | 是否保存网页源文件标志 |
| SourceFileDB | varchar(50) | 是 | 存放网页源文件的服务器IP |
| InfoUpdateServer | varchar(50) | 是 | 去重服务器IP |
| TaskBeginTime | longInt | 是 | 采集间隔时间（单位为分钟） |
| TaskRemark | varchar | 是 | 备注 |
| StartTime | DateTime | 是 | 任务最近采集活动启动时间 |
| StopTime | DateTime | 是 | 任务最近采集活动停止时间 |
| OperateType | varchar(50) | 是 | 启动/停止类型 |
| TotalDataSize | longint | 是 | 累计采集数据量 |
| LastGathDateSize | longint | 是 | 最近活动采集数据量 |
| TaskState | varchar(50) | 否 | 任务状态 |

任务状态：启动、进行中、停止、已停止；

启动/停止类型：用户启动、自动启动、用户停止、任务完成、出错

（2）表（GatherWebsite）：存放定向采集的网络站点信息，表结构如表3-2所示。

**表3-2 GatherWebsite**

|  |  |  |  |
| --- | --- | --- | --- |
| **字段名** | **类型（长度）** | **是否允许为空** | **说明** |
| SiteID | 自动增长 | 否 | 主键，网站编号 |
| TaskID | varchar(50) | 否 | 外键，采集任务ID |
| SiteName | Varchar(50) | 是 | 站点名称 |
| SiteType | Varchar(50) | 是 | 站点类别 |
| SiteArea | Varchar(50) | 是 | 站点归属地 |
| SiteWeight | Int | 是 | 站点优先级 |
| SiteEnterURL | Varchar(50) | 否 | 唯一性，站点入口网址 |
| CrawlerDepth | Int | 是 | 采集深度 |
| FilterDomain | Varchar(50) | 是 | 过滤域名 |
| UserName | Varchar(50) | 用户名 | 需登录网站的用户名 |
| Password | Varchar(50) | 用户名 | 需登录网站的密码 |
| Remark | varchar | 是 | 备注 |
|  |  |  | 模板XML |

归属地：省/地市 形式存放，如浙江省/杭州市

（3）表（WebContInfo）：存放结构化网页内容信息的数据库为：WebContInfoDB，其中表（WebContInfo）用于存放定向采集的网页内容的结构化数据，表结构如表3-3所示。

**表3-3 WebContInfo**

|  |  |  |  |
| --- | --- | --- | --- |
| **字段名** | **类型（长度）** | **是否允许为空** | **说明** |
| ID | 自动增长 | 否 | 主键，信息编号 |
| TaskID | varchar(50) | 否 | 外键，任务ID |
| SiteID | Varchar(50) | 是 | 外键，站点编号 |
| PageURL | Varchar(50) | 是 | 完整网址 |
| ColectTime | DateTime | 否 | 采集时间 |
| TextTitle | Varchar（50） | 否 | 文章标题 |
| Text | Varchar | 否 | 文章内容 |
| PageTime | DateTime | 是 | 文章发表时间 |
| reproter | Varchar（50） | 是 | 文章来源 |
| reproter | Varchar（50） | 是 | 文章作者 |
| Joiner | int | 是 | 文章参与人数 |
| Commenter | int | 是 | 文章评论人数 |
| EmotionClass | int | 是 | 情感等级 |
| Remark | varchar | 是 | 备注 |
| State | varchar(50) | 是 | 状态 |

## 4.系统实现

## 4.1 定向采集器

## 4.1.1 概述

定向采集器按功能大致上可以分为：检查任务状态模块、任务信息初始化模块、多线程集中管理模块以及爬虫线程模块。

1. 检查任务状态模块：

改模块通过每个一定的时间轮循控制台数据库中的状态字段，来对爬虫程序作出相应的指令。



1. 任务信息初始化模块

从控制台数据库导入种子队列以及相关的任务信息

1. **测试环境介绍**

|  |  |  |  |
| --- | --- | --- | --- |
| 软硬件环境 | 控制台服务器 | 采集终端 | 数据库服务器 |
| 硬件环境 | CPU 600 MHz 以上，建议 1GHz  内存 256 MB以上  硬盘 200G | CPU 600 MHz 以上,建议1GHz  内存 256 MB以上  硬盘 200G | Cpu 800 MHz,建议1 .6GHz  内存512 MB以上  硬盘 500G |
| 软件环境 | Windows 2003 server  JDK：1.6.0\_45  Apache:2.2.0  Tomcat:7.0.8 | Windows 2003 server  JDK：1.6.0\_45 | Mssqlserver |
| 网络环境 | 10M LAN | 10M LAN | 10M LAN |

1. **测试案例设计**

为了能够全方位的了解系统的运行情况以及系统的抗压能力，本次测试一共设计了四个实验案例。每一个案例都在同一类网站中具有其代表意义。本次测试将会以12小时为间隔记录数据观察数据的变化情况并随时观察程序输出的日志信息，判断程序是否运行良好。

1. **测试结果**

1.测试背景

1.1 测试对象

本次测试将会分别以如表1.1中所列的网站为起始地址，分时段记录数据，并将整理好的数据以图表的形式附在测试报告后。

表1.1 测试网站名称及地址列表

|  |  |
| --- | --- |
| 网站名称 | 网站地址 |
| 浪新闻 | http://news.sina.com.cn/ |
| 天涯论坛 | http://bbs.tianya.cn/ |
| 百度贴吧 | http://tieba.baidu.com/ |
| 搜狐新闻 | <http://news.sohu.com/> |

1.2 测试目标

本次测试的目的旨在测试软件的性能，在长时间，大规模的情况下程序是否能够高效切稳定运行。测试将从页面采集总体数量、页面总体解析量、单位时间获取页面数、单位时间解析数、以及每个时段的采集效率差等几个方面对软件进行全方面的测试，随后将会附上具体的测试结果报告。

2.测试情况

测试案例1：天涯社区

2.1.1 测试目标概述

天涯论坛（<http://bbs.tianya.cn/>）。天涯社区，创办于1999年3月，自创立以来，以其开放、包容、充满人文关怀的特色受到了全球华人网民的推崇，经过十年的发展，已经成为以论坛、部落、博客为基础交流方式，综合提供个人空间、相册、音乐盒子、分类信息、站内消息、虚拟商店、来吧、问答、企业品牌家园等一系列功能服务，并以人文情感为核心的综合性虚拟社区和大型网络社交平台。

作为国内影响力最大的论坛，作为数据收集者的角度而言，是炙手可热的一个网络论坛。

2.1.2 测试返回结果记数

①起止时间

开始时间：2014-10-7-21:30

结束时间：2014-10.9-21:30

间隔时间：48小时

②页面相关信息收集

开始时间：2014-10-11 20:31:57.497

结束时间：2014-10-13 20:32:30.223

间隔时间：48小时

总共采集页面数量：1661363

总共解析页面数量：1329090

页面解析率：85.73%

单位时间获取页面数：396.95/分钟

单位时间解析数：340.32/分钟

测试周期页面抓取情况实时反馈

③单页面解析效果对比图

原网页：



解析结果（数据库记录）：

|  |  |  |
| --- | --- | --- |
| url | time | text |
| <http://bbs.tianya.cn/po>  st-feeling-3689741-1.shtml | 2014-10-9-21:42 | “医生请你无论如何都救救我的孩子”一声悲怆却异常有力的声音划破广西贺州市人民医院烧伤整形外科住院部的宁静。 　　6月21日下午5时，一场灾难降临到广西贺州市八步区信都镇西两村一个… |

测试案例2：百度贴吧

2.2.1 测试目标概述

百度贴吧，是百度旗下独立品牌，全球最大中文社区。贴吧的创意来自于百度首席执行官李彦宏：结合搜索引擎建立一个在线的交流平台，让那些对同一个话题感兴趣的人们聚集在一起，方便地展开交流和互相帮助。

贴吧是一种基于关键词的主题交流社区，它与搜索紧密结合，准确把握用户需求，为兴趣而生。

截止2014年初，贴吧已突跃10亿注册用户，810多万个兴趣贴吧，日均话题总量近亿，浏览量超过20亿次。

作为中国网民最活跃的聚集地，这里能够收集的资源量是相当丰富。百度贴吧从结构上来说，层次框架分明清晰，因此也给抓取数据带来很高的效率。

2.2.2 测试返回结果记数

开始时间：2014-10-07 20:42:41.143

结束时间：2014-10.9-20:42:41.143

间隔时间：48小时

总共采集页面数量：314089

总共解析页面数量：138199

页面解析率：44.62%

单位时间获取页面数：571.59/分钟

单位时间解析数：255.04/分钟

测试周期页面抓取情况实时反馈

③单页面解析效果对比图

原网页：



解析结果（数据库记录）：

|  |  |  |
| --- | --- | --- |
| url | time | text |
| http://tieba.baidu.com/p/3339231875 | 2014-10-11 20:31:53.587 | 贴吧熊孩子十一假期过的爽爽的~~~ 十一期间也有很多小伙伴新加入我们贴吧熊孩子大家庭~~~ 为了使贴吧熊孩子大家庭更加的壮大~~~招募给力的八五小伙伴是必须的~~~ 根据规划，本吧准备招募以下方向吧务：… |

测试案例3：新浪

2.3.1 测试目标概述

新浪网新闻中心是新浪网最重要的频道之一，24小时滚动报道国内、国际及社会新闻。每日编发新闻数以万计。新浪网为全球用户24小时提供全面及时的中文资讯，内容覆盖国内外突发新闻事件、体坛赛事、娱乐时尚、产业资讯、实用信息等，设有新闻、体育、娱乐、财经、科技、房产、等。

新浪作为一个新闻主要发布网站，他的板块分类也是非常的鲜明清楚。

2.3.2 测试返回结果记数

开始时间：2014-10-10 17:18:39.010

结束时间：2014-10.12-17:18:48.033

间隔时间：48小时

总共采集页面数量：540891

总共解析页面数量：140632

页面解析率：26.83%

单位时间获取页面数：187.81/分钟

单位时间解析数：50.40/分钟

测试周期页面抓取情况实时反馈

③单页面解析效果对比图

原网页：



解析结果（数据库记录）：

|  |  |  |
| --- | --- | --- |
| url | time | text |
| http://news.sina.com.cn/c/2014-10-10/103930968478.shtml | 2014-10-10 17:18:39.273 | 2014年3月25日，第三届核安全峰会在荷兰海牙召开。本次会议围绕“加强核安全，防范核恐怖主义”这一主题，就如何维护全球核安全达成新共识，内容涉及削减高浓缩铀核材料数量、加强放射性材料安保监管、推动国际信息沟通等方面。… |

# **6 结论**

研究这个课题已经大半年的时间了，从最初完全无从着手，到查阅文献，上网搜索，总结出几条可行思路，到发现实现起来困难重重，自己掌握的知识还不足等等。接近半年的时间在查阅文献，找资料中度过，最终完成这项课题，感触良多。

此系统的任务就是能够对指定的互联网站点及信息内容进行自动分类采集，并对采集的网页信息进行智能解析生产结构化数据保存到数据库。最终的毕业设计成果已经达到了最初期望的设计要求，完成了各项需求。

在进行毕业设计的过程中，我充分运用到了本科四年所学到的知识。并且在爬虫这一块也有了一定的研究。通过这次毕设，我的编程思想、编写代码的实操能力、文档编写能力有了很大的提高。

因为时间有限，本毕业设计也不是完成的尽善尽美。爬虫在性能方面一直有瓶颈，如果能够投入跟多的时间，相信整套系统会有更大的改善。

# **致谢**

经历了多月的毕业设计以及毕业论文的撰写，终于完成了互联网信息定向采集系统。在进行毕业设计的过程中，遇到了很多的疑惑与困难，最后在老师、同学的悉心指导与帮助下都一一解决与克服了。

在此，需要特别感谢学校的领导与老师的辛勤栽培，因为有了大学四年打下的专业基础，才能顺利的完成毕业设计的各项工作。

同时，还要特别感谢我的毕业论文指导老师——俞定国老师。从最开始的毕业论文选题，到文献综述、外文翻译、开题报告的撰写，再到开展毕业设计以及写毕业论文的整个过程中，老师给予了我充分的指导与帮助，使我对该选题有了更深刻的认识，明确了该系统开发的需求以及具体的实现方式，并且对论文的修改提出了许多意见和建议。在此，谨向俞老师致以我最真诚的谢意。

感谢这篇论文所涉及到的各位学者。本文引用了数位学者的研究文献，如果没有各位学者的研究成果的帮助和启发，我将很难完成本篇论文的写作。

由于我的学术水平有限，所写论文难免有不足之处，恳请各位老师和学友批评和指正！

# **参考文献**

[1]谢希仁.计算机网络[M].第五版.北京:电子工业出版社，2008：236-239

[2]Jeffrey.HTTP协议详解.<http://blog.csdn.net/gueter/article/details/1524447>，2007-03-08

[3][罗刚](http://book.douban.com/search/罗刚%20%20王振东)，王振东.自己动手写网络爬虫[M] 北京：清华大学出版社，2010.

[4]邱哲，王学松，符滔滔.开发自己的搜索引擎--Lucene+Heritrix(第二版)[M] 北京：人民邮电出版社，2010

[5][Aaron M. T. Harmer](http://s.wanfangdata.com.cn/Paper.aspx?f=detail&q=作者:)，[Todd A. Blackledge](http://s.wanfangdata.com.cn/Paper.aspx?f=detail&q=作者:)，[Joshua S. Madin](http://s.wanfangdata.com.cn/Paper.aspx?f=detail&q=作者:)．[High-performance spider webs: integrating biomechanics, ecology and behaviour](http://d.wanfangdata.com.cn/NSTLQK_NSTL_QKJJ0221833967.aspx)[J]，Journal of the Royal Society Interface，2011, 8(57).

[6]金婵鸣，徐东平，JIN Chan-ming，XU Dong-ping．搜索引擎系统中网页抓取模块研究[J] 现代计算机(专业版)，[2010,(3)](http://c.wanfangdata.com.cn/periodical/xdjsj-xby/2010-3.aspx).

[7]Heritrix Home Page，http://crawler.archive.org/index.html，2011.

[8]张君，李保秀．垂直搜索引擎中网络蜘蛛的设计[J] 福建电脑，[2008,24(4)](http://c.wanfangdata.com.cn/periodical/fjdn/2008-4.aspx).

[9]Jeffrey E.F.Friedl．[精通正则表达式](http://book.douban.com/subject/2154713/)[M] 北京： 电子工业出版社，2007.

[10]李盛韬，赵章界，余智华．基于主题的Web信息采集系统的设计与实现[J] 计算机工程，[2003, 29(17)](http://c.wanfangdata.com.cn/periodical/jsjgc/2003-17.aspx).

[11]潘雪峰，花贵春，梁斌.走近搜索引擎[M] 北京：电子工业出版社，2011.

[12][赵晓峰](http://social.wanfangdata.com.cn/Locate.ashx?ArticleId=dnzsyjs-itrzyksb200816037&Name=赵晓峰)．基于Web的网站信息采集系统的设计与实现[J] [微计算机信息](http://c.wanfangdata.com.cn/Periodical-wjsjxx.aspx)，[2007,23(21)](http://c.wanfangdata.com.cn/periodical/wjsjxx/2007-21.aspx).

[13][李伟](http://social.wanfangdata.com.cn/Locate.ashx?ArticleId=bgzdh200707021&Name=李伟)，[黄颖](http://social.wanfangdata.com.cn/Locate.ashx?ArticleId=bgzdh200707021&Name=黄颖)．基于HtmlParser的网页信息提取[J] [兵工自动化](http://c.wanfangdata.com.cn/Periodical-bgzdh.aspx)，[2008, 2(16)](http://c.wanfangdata.com.cn/periodical/dnzsyjs-itrzyksb/2008-16.aspx).

[14]袁津生，李群，蔡岳.搜索引擎原理与实践[M] 北京：北京邮电大学出版社，2008.

[15][陈涓](http://book.douban.com/search/陈涓)，[赵振平](http://book.douban.com/search/赵振平).HTTP权威指南[M] 北京：人民邮电出版社，2012.