互联网信息定向采集系统

V1.0

测试报告

2014年06月11日

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

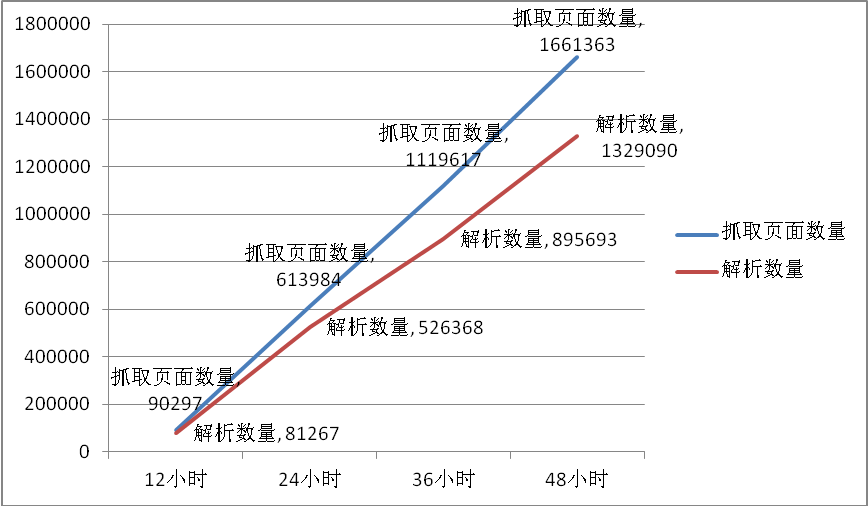
总共解析页面数量：1329090

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

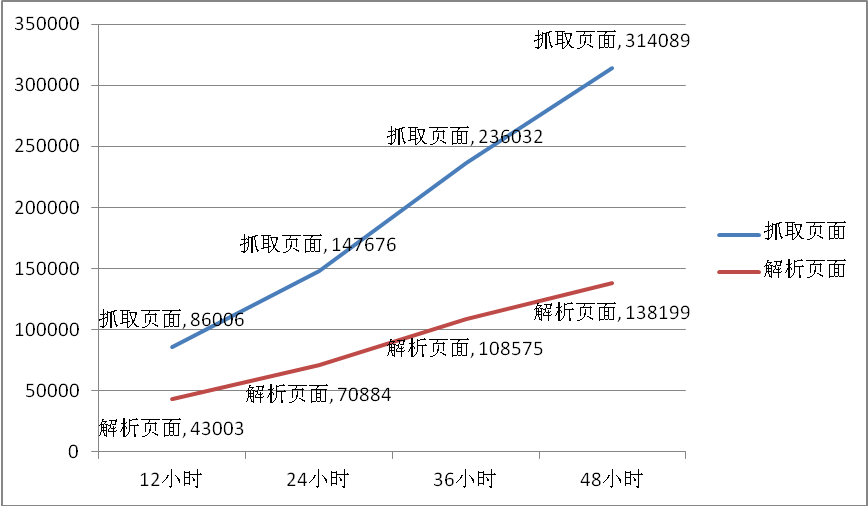
总共解析页面数量：138199

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

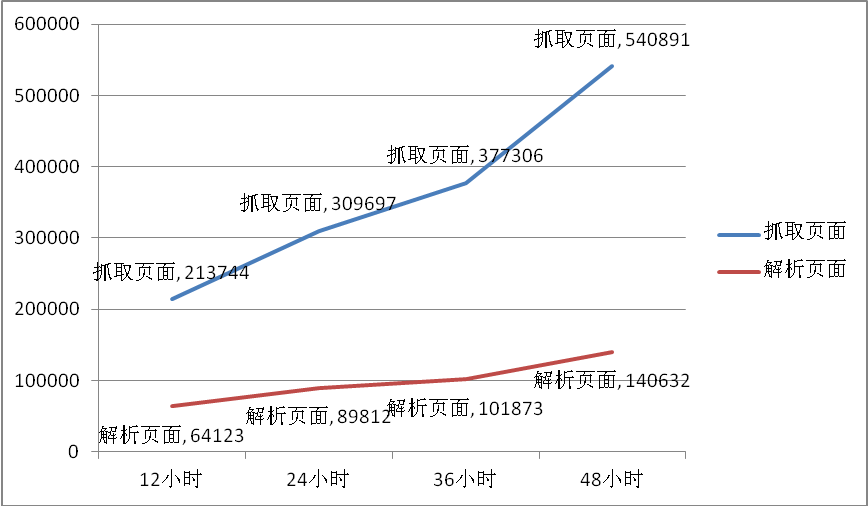
总共解析页面数量：140632

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。… |

测试案例：搜狐

2.3.1 测试目标概述

搜狐公司是中国领先的新媒体、通信及移动增值服务公司，是中文世界强劲的互联网品牌。

作为国内的重要新闻网站之一，也是受到很多爬虫人士的青睐。

2.3.2 测试返回结果记数

开始时间：2014-10-13 16:28:39.010

结束时间：2014-10.15-16:28:48.033

间隔时间：48小时

总共采集页面数量：320891

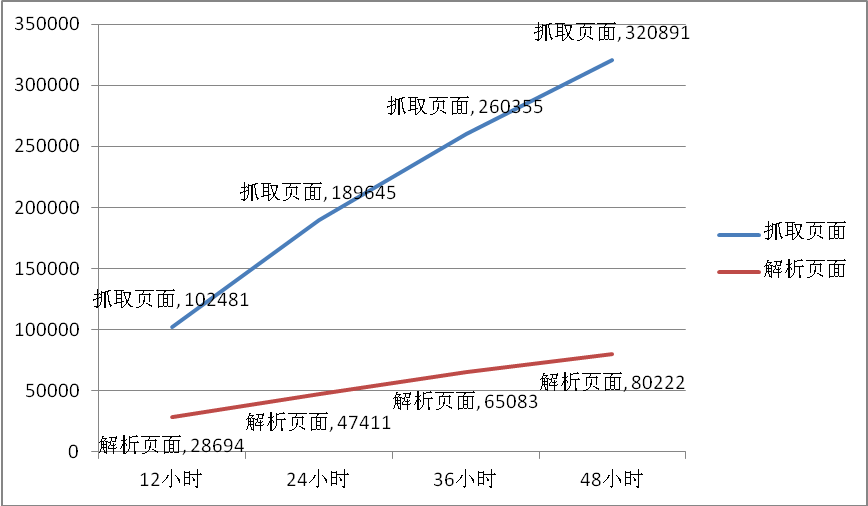
总共解析页面数量：80222

页面解析率：24.99%

单位时间获取页面数：111.42/分钟

单位时间解析数：27.85/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.sohu.com/20141015/n405123960.shtml | 2014-10-15 05:11:36 | 本报北京10月14日电（记者成慧）国家工商行政管理总局今日公布2014年前三季度市场主体发展、市场监管、消费维权基本情况。数据显示，前三季度全国市场主体发展稳步增长，新登记注册企业264.8万户，同比增长52.44%，注册资本（金）13.42万亿… |