ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.7, No.33, November 2011.

E-mail: eduf@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

网络舆情分析中智能爬虫的设计

周民[1],邱雅[1],王华彬[2]

(1.南阳理工学院,河南 南阳 473004;2.武汉天和技术股份有限公司,湖北 武汉 430000)

摘要:网络爬虫作为舆情分析系统的信息源收集器,其性能的优良直接关系到舆情分析结果的好坏。该文针对舆情分析要求信息源具有较高的主题覆盖率的要求,同时考虑到现有爬虫在有效利用网络资源方面的薄弱现状,在现有爬虫的基础上加入了爬虫测速模块、主题更改模块,为提高爬虫在舆情信息收集中的性能提出了一些解决方法。

关键词:网络舆情;网络爬虫;网络资源;爬虫测速;主题更改

中图分类号:TP393 **文献标识码**:A **文章编号**:1009-3044(2011)33-8301-02

The Design of Intelligent Iinformation Collector for Analysis of Network Public Opinion

ZHOU Min[1], QIU Ya[1],WANG Hua-bin[2]

(1.Nanyang Institute of Technology, Nanyang 473004, China; 2.Wuhan-Tianhe Technology Company Limited, Wuhan 43000, China)

Abstract: As the information collector for analysis of Network public opinion, the performance of web crawler has an important impact on the results of Network public opinion analysis. The main target of the classical crawler is to enhance the comprehensive of information, but the public opinion demands high efficiency in topic related resources. Meanwhile, the classical crawler is inefficient use of network resources. Based on these considerations, this paper adds two modules to the classical crawler: crawler speed module and change subject module. These policies increase the performance of web crawler.

Key words: network public opinion; web crawler; network resources; web crawler speed; subject to change

网络舆情是网络上围绕中介性社会事件的发生、发展和变化,民众对社会管理者产生和持有的社会政治态度[1]。当今,网络已成为思想文化的集散地和社会舆论的放大器,网络舆情分析变得越来越重要。网络爬虫作为舆情分析系统的信息源收集器,其性能的优良关系到舆情分析结果的好坏[2]。然而传统的网络爬虫在爬行过程中对网页信息的主题相关度不作过多考虑,抓取的网页中存在众多的无用信息,网络宽带利用率低[3]。网络舆情分析因其自身的特点,要求信息源具有较高的主题覆盖率。本文在现有爬虫的基础上加入了爬虫测速模块、主题更改模块,对爬虫的爬行策略进行指导,使其既能满足舆情分析的需要,又能合理有效地利用网络资源。

1 改善爬虫网络利用率的解决方案

为了解决网络爬虫不能有效利用网络资源的问题,本文在传统爬虫中加入爬虫测速模块,对爬虫的抓取速度进行监控,当爬虫的速度慢下来后,采取相应的措施对爬虫的状态进行修改,从而实现对网络资源的有效利用。具体实现步骤如下:

步骤1:爬虫抓取速度监控

为了对爬虫的网页抓取速度进行监控,首先要对该速度进行定义。影响爬虫的网页抓取速度主要有两个因素,分别是抓取页面的大小和抓取这些页面所耗费的时间[4]。本文将爬虫的网页抓取速度B定义为下面的公式:

$$B=\frac{P}{T_{ij}} \tag{1}$$

在公式(1)中T_{ij}表示对爬虫抓取速度进行监控的时间段,P表示在时间段T_{ij}内爬虫所抓取的总的页面大小。

系统可以设定时间段对爬虫的抓取速度进行监控,当系统检测到爬虫的抓取速度低于正常水平的40%后,系统产生报警。

步骤2:爬行策略更改

当系统产生报警后,就要采取相应的措施进行处理。这些措施主要包括:减少爬虫的线程数;暂停当前爬虫的运行,选择适当的时间继续爬行;更换爬行网站。通过这些策略的更改,可以避开网站的访问高峰,选择网站流量小时再进行网页的抓取,从一定程度上利用了有限的网络资源,加快了爬虫的爬行速度。

2 改善爬虫主题覆盖率的解决方案

网络舆情的产生是不可预知的,无法事先确定它的主题。同时,主题爬虫需要计算抓取的网页与主题的相关度[5],造成爬行速度比较慢。基于这些考虑,本文没有采用主题网络爬虫来抓取网页,而是在传统爬虫中加入主题更改模块,来提高爬虫的主题覆盖率。

当需要抓取关于某个主题的网页信息时,根据该主题的关键词,通过调用现有的搜索引擎(百度、google)搜索到关于该主题的

收稿日期:2011-09-19

作者简介:周民(1981-),男,河南南阳人,南阳理工学院软件学院助教,硕士,主要从事数据库相关方向教学研究;邱雅(1981-),女,河南南阳人,南阳理工学院软件学院讲师,硕士,主要从事计算机相关教学研究。

若干网页,然后获得搜索到的这些网页的URL列表,并将其导入到URL队列中,从而实现网络爬虫抓取主题的更改。主题更改模块的具体实现步骤如下:

步骤1:提供主题关键词

提供能够反映该主题特征的关键词,关键词可以是一个,也可以是多个。

步骤2:根据关键词确定搜索引擎返回的搜索结果中第一页对应的URL;

以百度搜索引擎为例,利用关键词确定搜索引擎返回的搜索结果中第一页对应的URL的方法如下:

http://www.baidu.com/s?wd=%CB%EF%CE%B0%C3%FA%20%BE%C6%BC%DD&pn=0&tn=baidudg

其中wd的值为%CB%EF%CE%B0%C3%FA%20%BE%C6%BC%DD,这个值是关键词的utf-8编码,pn的值为0,表示这是第一页。

步骤3:获取第一页中与主题相关的URL

与主题相关的URL的获取主要通过对第一页进行解析获得,主要利用正则表达式提取源码中span标签之间的内容。具体的正则表达式如下:(.*?)

步骤4:确定搜索引擎返回的搜索结果中第二页对应的URL

以百度搜索引擎为例,搜索结果通常是每页包含10个相关的URL,利用这一特征,并根据搜索结果中第一页对应的URL便可确定第二页对应的URL。具体实现是,将第一页的URL的pn值设为10。

步骤5:利用和步骤3相同的方法,提取第二页中与主题相关的URL,重复这样的过程,直到获得与主题相关的指定数目的URL。

步骤6:将获取的与主题相关的URL放入爬虫的URL队列中,实现对该主题的爬行。

3 系统设计与测试

基于上述分析,本文给出了爬虫系统的设计方案,如图1所示。该爬虫在传统爬虫的基础上添加了爬虫测速模块和主题更改模块。

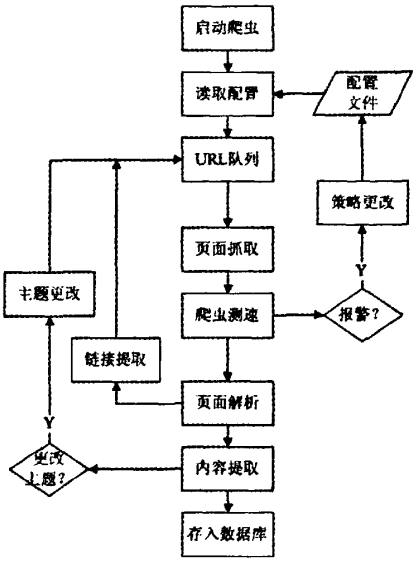

图1 系统设计方案

系统的具体工作流程如下:

1)启动爬虫

2)读取配置文件

配置文件中主要包括爬虫开启的线程数目、任务列表、爬行时间等配置信息。

3)开始爬行

爬虫根据任务列表,从种子URL开始对网页进行抓取。

(爬虫在抓取网页的同时,对爬虫抓取网页的速度进行监控,当抓取速度低于正常水平的40%时产生报警,此时就要更改爬虫的爬行策略,把新的爬行策略信息写入配置文件,当爬虫处理完该条URL后,根据新的策略进行爬行。

5)主题更改

爬虫在抓取页面后对页面进行解析,提取其中的URL放入URL队列中,并将提取的网页主体内容存入数据库中。在这一过程中,如果网络舆情分析需要某一特定主题的大量网页信息,爬虫根据该主题的关键词调用百度、谷歌等搜索引擎获取关于该主题的URL列表,然后,该URL列表放到URL队列中。

4 系统测试

系统的软硬件环境如下:2G内存,2.66GHzCPU,320G串中硬盘,WindowsXP操作系统。采集的数据来自2011年6月至2011年9月间各主流网站。试验结果表明,传统爬虫在抓取1.2G数据时耗时9800s,而改进后爬虫在抓取相同数量的数据耗时仅耗时4268s,抓取速度有显著的提高。爬虫在对特定主题的网页进行抓取时,主题覆盖率也达到了86%。

5 小结

本文针对现有网络爬虫网络资源利用率不高的现状,并结合网络舆情分析对信息源具有较高的主题覆盖率的要求,在传统爬虫中加入了爬虫测速模块和主题更改模块。试验结果表明,本文提出的这些解决方案对网络舆情分析中的信息收集器进行了一定程度的改善。

参考文献:

[1] 金晓鸥.互联网舆情信息获取与分析研究[D].上海:上海交通大学,2008.
[2] 张祥.网络舆情分析中智能信息收集器的设计实现[D].武汉:华中科技大学,2008.
[3] 尹江.网络爬虫效率瓶颈的分析与解决方案[J].计算机应用研究,2008,28(5):1114-1119.
[4] 孟祥乾.一种新的网络爬虫带宽控制策略[J].网络与通信,2008,24(11):76-78.
[5] 刘金红.主题网络爬虫研究综述[J].计算机应用研究,2007,24(10):26-29.