

一种改进的 T-Spider 分布式爬虫

金 凡¹, 顾进广^{1,2}

(1 武汉科技大学 计算机科学与技术学院, 湖北 武汉 430065; 2 东南大学 计算机科学与工程学院, 江苏 南京 210096)

摘 要: 为了提高互联网网页的抓取速度, 提出了一个改进的 T-Spider 分布式爬虫模型. 该爬虫在解析 URL 阶段将页面进行切割以并行解析, 在页面调度阶段使用改进的链接优先权计算方法, 提高爬虫的抓取速度和稳定性. 通过实验结果分析, 验证了该方法的有效性.

关键词: 网络爬虫; T-Spider; 页面切割; 链接优先权计算

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)08-0102-03

An Improved T-Spider Distributed Crawler

JIN Fan¹, GU Jin-guang^{1,2}

(1 College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;

2 College of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: To increase the speed of the crawler, this paper proposes a model that is based on the T-Spider. During the time of extracting links from the page content, the crawler takes use of the page cutting algorithm, and then uses a new algorithm of link priority computing to enhance the stability and increase the speed of the crawler. The experiment shows that it is availability.

Key words: web crawler; T-Spider; page cutting; link priority computing

1 引言

随着搜索引擎的广泛应用, 作为其基础的网络爬虫在信息搜集上起着巨大的作用. 当前的网络爬虫研究领域有基于局域网和广域网的分布式爬虫(基于地理位置调度)^[1-2]. T-Spider^[3]作为一种基于网络拓扑结构的站内优先调度的分布式爬虫已经成熟地应用于各种搜索引擎. 本文在 T-Spider 爬虫模型的基础上增加了页面切割算法和改进的链接优先权算法, 运用了当前多核计算和分布式计算的的优势来改进爬虫模型, 不足是对网页内容分析方面不够智能.

2 分布式爬虫

2.1 分布式爬虫的策略

网络爬虫遍历节点有如下四种策略^[4-7]:

- (1) DFS(depth-first search): 深度优先遍历, 使用的策略是后进先出(LIFO), 使用的数据结构是栈;
- (2) BFS(breadth-first search): 广度优先遍历, 使用的策略是先进先出(FIFO), 使用的数据结构是队列;
- (3) DEG(higher degree): 优先度遍历, 最多指向的 URL 将获得优先权, 使用的是优先队列;
- (4) RND(random): 随机遍历, 抽取随机的 URL.

大多数搜索引擎的爬虫所使用的策略居于 BFS 和 Random 之间^[8].

收稿日期: 2011-05-10; 修回日期: 2011-06-27

基金项目: 国家自然科学基金(60803160); 中国博士后科学基金(20060400275); 江苏省博士后科学基金(0601009B); 湖北省自然科学基金重点计划(2009CDA136, 2009CDA034); 湖北省教育厅科研项目(Q20101110, D2009110); 武汉市科技攻关计划(201110821225)

根据无尺度网络的这种站内链接高内聚,大网站聚集大部分链接,不同网站的网页数量差别悬殊等特征,本爬虫使用 BFS 和 DEG 相融合的策略,即在不同的站点之间实行 BFS 算法,在站点内部使用 DEG 算法。

一般网站的首页是一个链接的指南,汇集了站内的大部分链接,提高首页的优先权的策略能够有效地减少页面的缺失率和错误率,并且能迅速获取到更新的页面^[9-10]。

2.2 分布式爬虫的模型描述

本文提出的模型旨在模拟一个改进 T-Spider 分布式爬虫的策略过程,它分两部分,一部分是该分布式爬虫的模型设计原理和算法,另一个部分是爬虫的处理速度的形式化分析。

2.2.1 分布式爬虫的原型与改进算法

爬虫的分布式原型是: ND_i 为爬虫分节点, P_j 为某个爬虫分节点内的爬虫进程, Th_k 为某个爬虫进程内的爬虫线程 ($i, j, k \in \{1, 2, 3, \dots, n\}$), 它们的关系是: $Th_k \in P_j \in ND_i$ 。

该爬虫在处理网页时使用了页面切割算法,见算法 1。

算法 1: Cutting(text)

输入: text, n . text 为网页的原始内容, n 为要切割成为的网页片断数目。

输出: Sp . Sp 为分割后得到的网页原始内容片断的集合。

```
begin
text = text - js // js 为网页头部的 JavaScript 脚本
 $Sp_i = \text{text} / n \quad // i \in \{0, 1, \dots, n\}$ 
for each  $p$  in  $Sp$  do
    if ( $p = P$ ) then //  $P$  为链接完整的片段
        continue
    else
         $p = p + \Delta p \quad // \Delta p$  为链接的补全片段.
    endif
endfor
return  $Sp$ 
end
```

该算法首先会去除网页头部的 JavaScript 脚本,然后会对网页内容进行等量分割,最后对得到的切割片段中不完整的链接片段进行补全,以保证链接完整地存在一个片断中。

在站内链接的调度方面,本爬虫使用链接层级和链接内容比来作为链接优先权计算的参数。链接层级越低,则链接的优先权越大,这基于链接层级越

低,则其为站内索引链接的可能也越高,其索引页面也越多,而目标网站对该网页的维护力度也更大,不易产生死链,例如 www.sina.com 层级为 1, www.sina.com/education 层级为 2,前者优先权高于后者;页面的链接内容比越高,页面的优先权也越大,这基于链接内容比越高,页面的外链也越多,则能够索引数量也越多。对链接的优先权值进行排序,从而构成站内页面的优先调度序列。链接优先权计算算法,见算法 2。

算法 2: Priority_Compute(pn, a_{pn}, lv, a_{lv})

输入: pn, a_{pn}, lv, a_{lv} . 其中 $pn = \text{links_len} / \text{content_len}$ 为链接数量, (links_len 为链接数量, content_len 为页面内容的大小). a_{pn} 为平均链接内容比, $a_{pn} = (\sum_{i=1}^n pn_i (i \in \{1, 2, \dots, n\})) / n$, lv 为链接层级, $a_{lv} = (\sum_{i=1}^n lv_i (i \in \{1, 2, \dots, n\})) / n$ 。

输出: p_w . 链接的优先值权重。

begin

$$p_w = \frac{a_{lv} - lv}{a_{lv}} \times u + \frac{pn - a_{pn}}{a_{pn}} \times v$$

return p_w

end

其中 u, v 为调整因子, u, v 分别设为 0.6, 0.4 时效果较好。

2.2.2 改进 T-Spider 分布式爬虫与传统 T-Spider 分布式爬虫抓取速度的比较

下面对传统 T-Spider 分布式爬虫和加入了页面切割和页面优先权计算改进的 T-Spider 分布式爬虫的效率进行形式化比较。

定义 1(爬虫抓取时间计算的基本变量) 互联网上面总共存在 u 个 URL 链接, n 个 page; 平均页面链接数: $k = u/n$, 单位线程间通信时间: T_{th} ; 读取单位页面的时间: T_{down} ; 解析单位页面的时间: T_{proc} ; 完成单位页面迭代过程的总时间: T_{total} ; 查询单位链接是否已访问时间: $T_{visited}$ 。

(1) 传统 T-Spider 分布式爬虫单位页面的迭代的总时间为

$$T_{total1} = k \times T_{th} + T_{down} + T_{proc1} + T_{visited} \quad (1)$$

(2) 改进 T-Spider 分布式爬虫单位页面的迭代的总时间为

$$T_{total2} = k \times T_{th} + T_{down} + T_{proc2} + T_{visited} \quad (2)$$

式(2)中的算法采用页面分割解析的方法,有 $T_{proc2} = T_{proc1}/n$, 根据式(1)和式(2)的比较,很显然改进 T-Spider 爬虫速度性能优于传统 T-Spider 爬虫。

3 实验结果

笔者针对上述问题,做了两组实验,实验一的硬件设备是 4 台 PC 机,实验二的硬件设备是 7 台 PC 机,每组实验中用 1 台 PC 作主节点,操作系统选用 Windows XP, CPU 为酷睿 2 双核,内存 2 GB 所使用的开发平台是 Python 2.5. 本实验的网络环境是华中教育网,带宽是 10 Mb/s,并采用新浪、搜狐、网易作为爬虫入口.

这两组对比实验,一组是 3 个和 6 个分节点的改进 T-Spider 实验,结果见图 1. 另外一组是 6 个分节点的改进 T-Spider 爬虫和传统 T-Spider 爬虫的实验,结果见图 2. 图 1、图 2 中横向坐标表示抓取时间;纵向单位刻度表示抓取的网页总数. 综合的性能统计见表 1.

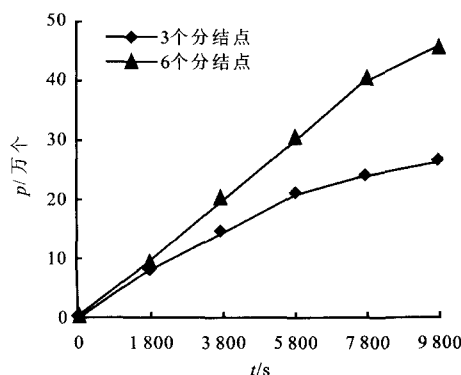


图 1 不同节点结果比较图

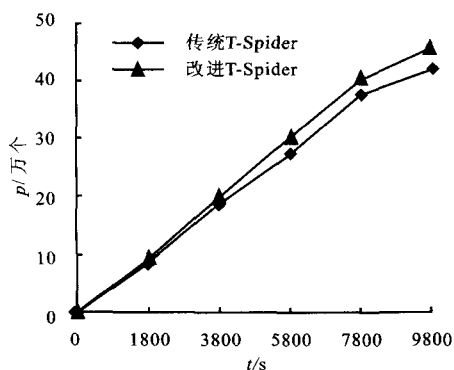


图 2 不同爬虫算法的结果比较图

表 1 爬虫的综合性能比较

类别	分节点 数/个	抓取网页 总数/万个	抓取链接 总数/万个	抓取总时间/ s
改进 T-Spider	3	27.9	2 386	10 000
改进 T-Spider	6	47.7	3 662	10 000
普通 T-Spider	6	42.1	3 286	10 000

通过实验的结果可以看出,分布式爬虫多节点

设计的作用明显,而改进的 T-Spider 分布式爬虫在同等条件下比传统 T-Spider 分布式爬虫多抓取 7.25%网页和 10.1%的 URL 链接.

4 结束语

本文通过对一种增加页面切割和新的链接优先权计算的改进 T-Spider 分布式爬虫的论述和实现,验证了通过对爬虫的分布式设计,和对页面的切割计算,能够充分地利用硬件资源和网络带宽;而基于链接层级和链接内容比的链接优先权算法,能够提高页面链接抓取率.

参考文献:

- [1] 蔡欣宝,陈洪平,赵朋朋,等. 一种 Deep Web 聚焦爬虫爬行策略[J]. 微电子学与计算机, 2009, 26(8): 117-120.
- [2] 许笑,张伟哲,张宏莉,等. 广域网分布式 Web 爬虫[J]. 软件学报, 2010, 21(5): 1067-1082.
- [3] 高克宁,柴桥子,张斌,等. 支持信息分类的高性能蜘蛛程序[J]. 小型微型计算机系统, 2006, 27(7): 1308-1312.
- [4] Fetterly D, Craswell N, Vinay V. Search effectiveness with a breadth-first crawl[C]// Research and development in information retrieval. New York: ACM, 2008.
- [5] Cafarella M J, Madhavan J, Halevy A. Web-scale extraction of structured data[J]. ACM SIGMOD Record, 2008, 37(4): 55-61.
- [6] Broder A Z, Najork M, Wiener J L. Efficient URL caching for world wide web crawling[C]//12th international conference on World Wide Web. New York: ACM, 2003.
- [7] McCown F, Nelson M L. Evaluation of crawling policies for a web-repository crawler[C]//17th conference on Hypertext and hypermedia. New York, 2006.
- [8] Gomes D, Santos A L, Silva M J. Managing duplicates in a web archive[C]//2006 ACM symposium on Applied computing. New York: ACM, 2006.
- [9] Bennouas T, Montgolfier P D. Random web crawls [C]//16th international conference on World Wide Web. New York: ACM, 2007.
- [10] 叶允明,于水,马范援,等. 分布式 Web Crawl 的研究:结构,算法和策略[J]. 电子学报, 2002, 30(12A): 2008-2011.

作者简介:

金凡 男, (1985-), 硕士. 研究方向为分布式计算、自然语言处理、信息检索.

顾进广 男, (1974-), 博士, 教授. 研究方向为语义 Web、服务计算.