文章编号: 1000-5862(2013)04-0382-05

一种分布式网络爬虫的设计与实现

杨　瑞,胡弘思,张文波,姚天昉*
(上海交通大学计算机科学与工程系,上海　200240)

摘要: 利用用户指定的关键字和搜索引擎生成 URL 种子,通过分布式网络爬虫抽取符合用户需求的网页作为研究所用的语料. 实验结果表明: 分布式网络爬虫可以较好地解决在短时间内抽取大量语料的需求.

关键词: 分布式系统; 网络爬虫; 设计
中图分类号: TP 391　　**文献标志码**: A

0　引言

网络爬虫,英文名称为 Spider 或 Crawler,是一种功能强大的自动提取网页的程序,它为搜索引擎从互联网上下载网页,是搜索引擎的重要组成部分. 此外,它可以完全不依赖用户干预实现网络上的自动"爬行"和"搜索". 网络爬虫工作过程一般是从一个或若干个初始网页的 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,直到满足系统一定的停止条件为止[1].

根据网络爬虫的网页搜索策略,网络爬虫可以分为基于传统图算法的深度优先、广度优先和最佳优先 3 种方法[2]. 这 3 种方法都是通用网络爬虫的爬取策略,从理论上来说,它可以通过一定的优先级对整个网络中的所有页面进行访问. 在网页搜索中,深度优先策略并不大适用,很多情况下无条件的深度优先遍历,会导致爬虫程序进入一个无穷尽的深度,即爬虫的陷入(Trapped)问题. 广度优先搜索策略的缺陷在于,在每个层次,爬虫尽可能多的覆盖当前页面中的链接,这样会导致爬虫漫无目的地扩散,带入大量无用链接. 最佳优先搜索虽然可以更有效地抓取目标网页,但是页面解析算法是该算法是否高效的关键.

本文采用广度优先搜索算法,通过页面分析算法,将 URL 的解析与广度优先结合起来,在广度优先搜索时,过滤页面中无关的 URL,从而提高广度优先算法的效率.

分布式网络爬虫实质上是多台机器上的不同网络爬虫的整合及合理调度,每个爬虫需要完成的任务和单个爬虫类似,通过分布式系统的调度,使得系统作为一个整体有效均匀地完成爬取任务. 分布式爬虫的多台机器可能分布在同一个局域网,也可能分布在不同地理位置的不同局域网. 按照分布式网络爬虫通信方式的不同,分布式网络爬虫大致可分为主从、自治和混合 3 种模式:

(ⅰ) 主从模式: 主从模式下的分布式网络爬虫一般由一台机器上的控制程序作为控制节点,其它机器的控制程序作为爬虫节点. 控制节点负责爬行节点之间的调度、控制节点与爬行节点之间的通信及整个系统的维护; 爬行节点只负责具体 URL 的爬取,而无需关心爬行节点之间的通信.

(ⅱ) 自治模式: 自治模式下分布式系统一般没有专门的控制节点,而是由节点之间的协作完成系统的调度及维护. 自治模式的通信方式可以是全连接通信或者环形通信,自治模式的调度算法较主从模式更复杂.

(ⅲ) 混合模式: 混合模式为主从模式和自治模式相结合的折衷方案. 该模式所有的爬虫都可以相互通信同时都可以进行任务分配; 特殊爬虫节点会对经过爬虫分配任务之后无法分配的任务进行集中分配.

收稿日期: 2012-11-15
基金项目: 国家自然科学基金(60773087)资助项目.
通信作者: 姚天昉(1957-),男,上海人,副教授,博士,主要从事 web 挖掘,信息抽取和机器学习等方面的研究.

1　分布式系统及网络爬虫设计

1.1　分布式系统构架

由于硬件条件的限制,本文的分布式网络爬虫采用局域网和广域网的折衷,并采用地理位置不同(IP 地址不同)的机器进行无差别地爬取;而本文涉及的系统不是大规模分布式系统,故采用简单易行的主从模式作为分布式通信模式,控制节点与爬行节点之间的通信不会成为整个系统的瓶颈.

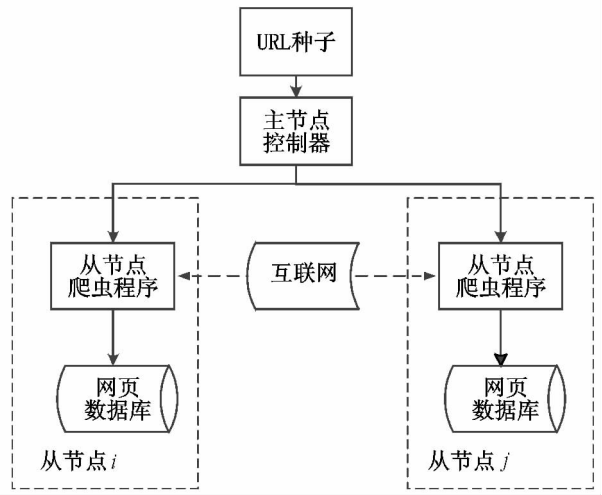

图 1　分布式网络爬虫系统构架

分布式网络爬虫[3-6]由一个控制节点(主节点)和若干爬行节点(从节点)组成. 为了保证分布式系统主从模式的正常运行以及从节点之间的耦合尽可能小.

主节点功能:(a)维护整个系统的 URL 历史记录. 包括 2 个方面工作:(ⅰ)从节点获得其解析的 URL 队列,汇总所有从节点的 URL,将其去重后存于主节点 URL 历史记录数据库;(ⅱ)向从节点分配 URL,保证各个从节点之间尽可能均匀高效地进行页面抓取.(b)保证与从节点之间的通信顺畅,维护整个系统的稳定性.(c)主节点不进行任何页面的抓取与解析工作,只进行调度管理及系统稳定性维护工作.

从节点功能:(a) URL 的爬取:从节点最主要的功能,对主节点分配过来的 URL 进行爬取,不管成功或是失败都要记录爬取结果,有超时及重试机制.(b)页面解析过程:从抓取到的网页中解析有效的 URL,以存入数据库作为进一步网页抓取的 URL 种子. URL 主要有 2 种:(ⅰ)搜索引擎 URL,这类页面只解析搜索引擎返回的有效搜索链接;(ⅱ)普通网页页面信息,这类页面信息只保留页面里的 URL 中与该页面指向 URL 的 host 一致的 URL.(c)从节点不负责任何节点之间的通信及 URL 记录的维护(如 URL 冲突排除),从节点只是将解析出来的 URL 存入数据库,等待主节点完成读取以及去重工作.

主节点的结构相对从节点简单(如图 2 所示),且因为主节点不用涉及网页的爬取操作,故主节点对于网络带宽的消耗主要是与从节点之间的通信. 主节点主要组成部分为数据库及主节点控制器. 主节点控制器完成主节点与从节点的通信及分布式系统稳定性和效率的维护,URL 的去重工作,与从节点数据库保持连接,定期检查连接状态并对失败的 SQL 连接重试.

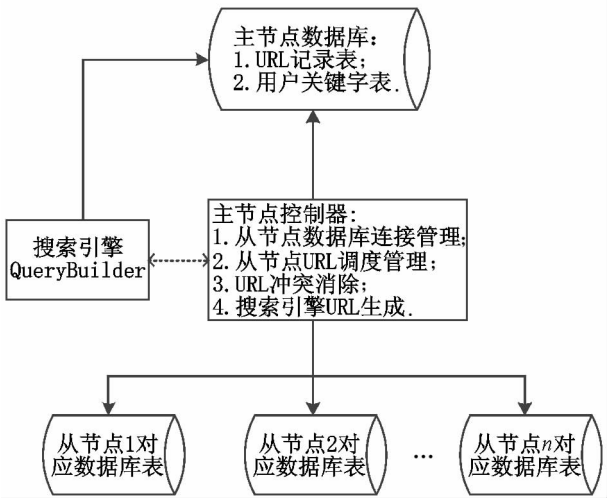

图 2　分布式网络爬虫主节点框架图

1.2　分布式网络爬虫主节点

图 2 中 QueryBuilder 为主节点的一个小模块. 主要功能为:将用户指定的若干关键字进行组合,然后通过一定算法生成搜索引擎的 URL,将这些 URL 存入主节点的数据库表,作为 URL 种子,再分配给从节点进行抓取工作[7-11].

1.3　分布式网络爬虫从节点

从节点主要由 Frontier、页面爬取线程及数据库 3 大模块组成,如图 3 所示.

页面爬取线程主要进行页面的抓取工作. 一般一台机器的线程数根据机器硬件条件及网络条件的不同可以设置为 10 ~ 50 不等. 对于来自同一个域(即 URL 对应的域,domain)的 URL 会有爬取间隔判断,若爬取间隔小于 1 s,则此爬取线程会等待. 页面解析模块对应网络爬虫的搜索策略,主要负责从 HTML 文件中解析所需的 URL,本文采用的是广度

搜索策略 + Filter 模式. 根据不同的网站给定不同的正则表达式 Filter,页面解析模块只提取满足 Filter 的 URL.

Frontier 是整个从节点的控制模块,Frontier 的功能包括:(a) 新建爬虫线程;(b) 销毁超时爬虫线程;(c) 管理各个队列. 队列包括:URL 等待队列(Pending Queue),处理好的 URL 队列(Processed Queue),已抓取页面队列(Crawled Queue),爬虫线程队列(Crawling Thread List);(d) 完成内存(队列)与数据库的交互操作.

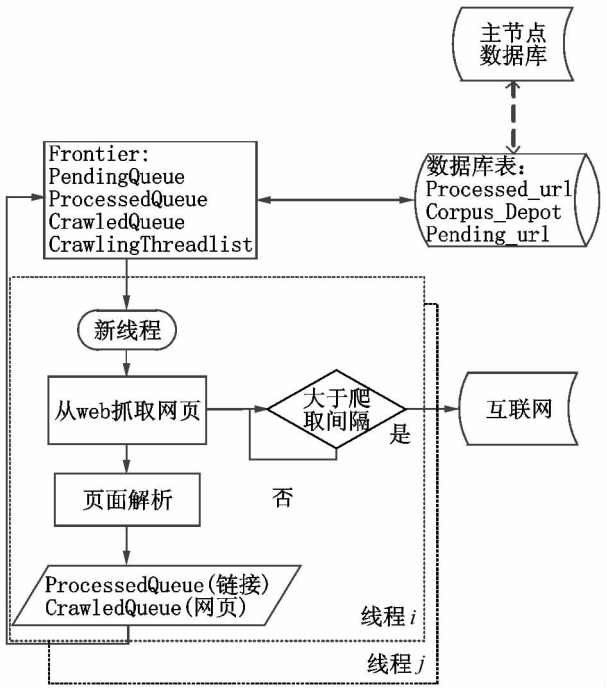

图 3 分布式网络爬虫从节点框架图

2 实验与分析

2.1 测试环境

分布式网络爬虫性能会受到实验机器的网络带宽和机器硬件性能的影响,本节列举测试环境.

表 1 硬件环境

	CPU	RAM
机器 1	Intel Core2 Duo CPU T6670 @ 2.20 GHz 2.20 GHz	2 GB
机器 2	Intel Core2 Duo CPU E4500 @ 2.20 GHz 2.20 GHz	2 GB

由于条件的限制,在测试分布式系统的时候,本文只使用了 2 台机器,即 2 个从节点部署分布式系统.

表 2 软件环境

OS	Windows 7 企业版 32bit
Database	SQL Server 2008 Express Edition
jdk	Jdk1.6.0_10

网络环境:带宽 10Mbps 的网络.

2.2 实验结果及分析

测试按照线程数的不同,单从节点与双从节点的区别,不同主机的 URL 分布和爬取时间对性能影响的 4 个方面进行测试.

2.2.1 线程数的不同对性能的影响 在 30 min、单节点、10 个 URL 种子不变的测试条件下,调节不同线程数,得到如表 3 所示的测试结果.

表 3 不同线程数对网络爬虫的影响

Threads	10	20	30
成功下载页面数	3 808	6 501	7 522
成功解析 URL 数	62 955	97 456	98 948
失败下载页面数	63	102	167

从理论上来说,当线程较少时,随着线程的增加,网络爬虫的下载速度会随之增加;当线程增加到一定程度时,因为机器网络带宽和硬件资源的限制,爬虫的抓取效率不会继续增加,或者增加得很缓慢,因为上述限制,过多的线程反而会造成爬虫效率的低下.

对比表 3 中 10 线程和 20 线程可以得出,当线程数为 10 时,机器的负载远没达到满负荷;当线程数为 20 时,不论是成功下载还是成功解析的 URL 数都大量增加,下载失败的页面数也在增加. 对比 20 线程和 30 线程,成功下载的页面数有所增加,但增加的幅度远不如 10 线程到 20 线程的增加幅度,但失败下载的页面数却在大量增加. 原因是过多的线程不管是对于机器的硬件资源还是网络带宽都会有很大的副作用,容易造成线程之间的等待以及网络的阻塞,所以导致失败下载页面数大量增加.

2.2.2 单从节点与多从节点的性能差异 表 4 展示的是在 30 min、20 线程、10 个 URL 种子相同测试条件下,单从节点与双从节点测试数据的差异. 理论上,在优秀的无延迟的分布式系统中,2 个从节点的效率应该是 1 个从节点的线性叠加,或者因为优良调度算法的实现,双从节点效率应该比单从节点效率 2 倍更好一些. 但表 4 中数据显示,"成功下载页面数"这一栏,双从节点并没有达到单从节点的 2 倍,分析原因是:在单从节点测试中,主节点和从节

点位于同一台机器上,这样从节点与主节点通信的时间消耗可以忽略不计,但在双从节点测试中,第 2 个从节点与主节点的数据库之间的数据库互相访问的时间消耗,不仅占用了网络带宽,也导致了整个系统效率有了一定的下降. 对于"成功解析 URL 数"这一栏,双从节点解析的 URL 数并没有达到单从节点的 2 倍,分析原因是: 根据网络中局部性原理,当页面数达到一定数量后,很多页面容易指向相同的 URL,故导致不重复的 URL 数量并没有随着解析页面数的增加而线性增加. 对于"失败下载页面数"这一栏,双从节点更容易分担对于同一主机的频繁访问,使得对于同一主机的访问间隔尽量增加,这也是分布式系统设计的一个初衷.

表 4　单节点和双节点对分布式爬虫的影响

	单从节点	双从节点
成功下载页面数	6 501	11 981
成功解析 URL 数	97 456	149 504
失败下载页面数	102	158

2.2.3　对不同主机的 URL 是否爬取均匀　本文在设计调度算法的时候,为了保持对不同主机的访问间隔控制,会尽可能平衡分配来自不同主机的 URL,所以对于不同主机的 URL 是否能够较均匀爬取是验证调度算法是否有效的依据.

表 5 展示的是在 30 min、20 线程、单从节点的条件下,选取来个 10 个不同网站(来自于不同主机) 的 URL 种子. 从表 5 中数据可以看出,从 10 个主机下载下来的页面基本上是平均的,彼此数量上差异不大.

表 5　爬虫对于不同主机的爬取速度

主机名	抓取 URL 数
zhidao. baidu. com	439
baike. baidu. com	593
iask. sina. com. cn	630
ks. cn. yahoo. com	570
www. hudong. com	627
www. chinanews. com	613
news. sina. com. cn	500
tech. sina. com. cn	617
auto. sina. com. cn	439
blog. sina. com. cn	544

2.2.4　不同的爬取时间对爬取效果的影响　表 6 展示的是在 20 线程,双从节点的相同条件下,不同下载时间对爬虫效率的影响. 从表 6 中很明显可以看出,随着时间的增加,爬虫的抓取速度基本线性增长.

表 6　不同爬取时间对爬取效果的影响

时间/min	15	30
成功下载页面数	5 651	11 375
成功解析 URL 数	100 252	149 504
失败下载页面数	75	158

3　结束语

大规模语料库的建立对于自然语言研究至关重要,本文利用用户指定的关键字生成 URL 种子,通过分布式爬虫系统抽取符合用户需求的原始语料. 实验数据表明,本文设计的分布式网络爬虫能够有效地进行大规模语料库的建设. 本文对分布式网络爬虫的测试均是在实验室硬件环境下,如何利用更好的硬件环境更高效率的进行更大规模的语料库的建设是本文下一步的研究内容.

4　参考文献

[1] Tripathy A, Patra P K. A web mining architectural model of distributed crawler for internet searches using page rank algorithm [EB/OL]. [2012-08-18]. http: // wenku. baidu. com/view/03181bd084254b35eefd3412.

[2] 周立柱,林玲. 聚焦爬虫技术研究综述 [J]. 计算机应用,2005,25(9):1965-1969.

[3] Radhakishan V, Yaser F, Selvakumar S. CRAYSE: design and implementation of efficient text search algorithm in a web crawler [EB/OL]. [2012-08-19]. http: // libra. msra. cn/Publication/14414792/crayse-design-and-implementation-of-efficient-text-search-algorithm-in-a-web-crawler.

[4] Shekhar S, Agrawal R, Arya K V. An architectural framework of a crawler for retrieving highly relevant web documents by filtering replicated web collections [EB/OL]. [2012-08-19]. http: // dl. acm. org/citation. cfm? id = 1844773.

[5] Zhu Kunpeng, Xu Zhiming, Wang Xiaolong, et al. A full distributed web crawler based on structured network [M]. Berlin: Springer, 2008: 478-483.

[6] 李晓明,李星. 搜索引擎与 Web 挖掘进展论文集 [C]. 北京: 高等教育出版社,2003: 1-8.

[7] Robert C M. Krishna B. SPHINX: a framework for creating personal, site-specific Web crawlers [J]. Computer Networks and ISDN Systems, 1998, 39(1/7): 119-130.

[8] 闵秋应,况庆强. 改进型 BP 神经网络自适应均衡器设计 [J]. 江西师范大学学报: 自然科学版,2012,36(3): 276-278.

[9] 周模,张建宇,代亚非. 可扩展的 DHT 网络爬虫设计和优化 [J]. 中国科学: 信息科学,2010,40(9):1211-1222.
[10] 王珏. 重叠型 P2P 网络中的查询负载均衡策略研究 [J]. 江西师范大学学报: 自然科学版,2012,36(3):292-296.
[11] 姜梦稚. 基于 Java 的多线程网络爬虫设计与实现 [J]. 微型电脑应用,2010,26(7):21-22.

Design and Implementation of a Distributed Web Crawler

YANG Rui,HU Hong-si,ZHANG Wen-bo,YAO Tian-fang*
(Department of Computer Science and Engineering,Shanghai Jiao Tong University,Shanghai 200240)

Abstract: User-specified keywords to generate URL seeds by search engine has been used. Webpage for user's requirements as research corpus through distributed web crawler has been extracted. Experiments show that the distributed web crawler can be good solution to extract a large number of corpora in a short time.
Key words: distributed system; web crawler; design

(责任编辑: 冉小晓)

(上接第 370 页)

Face Detection Algorithm Based on Double Color Space and Facial Rectangular Characteristics

ZHU Zhi-liang[1],XIONG Feng[1],TAO Xiang-yang[1,2*],LIU Xiao-shan[1,2]
(1. College of Physics & Communications Electronics,Jiangxi Normal University,Nanchang Jiangxi 330022,China
2. Key Laboratory of Photo-electronic & Telecommunication of Jiangxi Province,Nanchang Jiangxi 330022,China)

Abstract: An algorithm for face detection and location based on human facial rectangular characteristics in the YIQ and the YCbCr color space has been proposed. The algorithm selects the color space by comparing the average of the input image channels of R,G and B with the average of the channel of B,and segments the skin area with the two-dimensional OSTU and integral figure. Then,the skin segmentation figure by median filter based on the acreage of the skin area has been processed. In the end,an optimal face region rectangle with the acreage of the skin area which has been processed by median filter has been constructed,and the rectangle to locate the face region has been used. The experiments show that the algorithm has higher accuracy than the traditional face detection algorithms using single color space.
Key words: color space; skin segmentation; integral figure; ostu algorithm; face detection

(责任编辑: 冉小晓)