

文章编号:1673-0062(2013)03-0056-05

生产者—消费者二维队列模型在舆情监控系统中的应用

雷龙艳,万亚平,徐 强,阳小华

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:生产者—消费者模式是程序设计中的经典模式.为了满足舆情监控系统 Web 数据挖掘的实时高效性,提出基于生产者—消费者模式的二维矩阵队列链模型,得到二维长度设定准则,并引入线程池进行管理,提高了系统资源的利用率,实时高效,且能够十分方便的对其队列进行日志监控,为 Web 数据抓取分析系统提供了可供借鉴的经验.

关键词:生产者—消费者;二维队列;舆情;线程池;

中图分类号:TP391

文献标识码:B

Application of Producer-consumer Model with Two-dimensional Queues in Public Opinion Monitoring System

LEI Long-yan, WAN Ya-ping, XU Qiang, YANG Xiao-hua

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: Producer-consumer model is classic in program design. To satisfy real time and efficiency in web data mining of public opinion monitoring system, a two-dimensional matrix queues chain based on producer-consumer model is presented. We used multi-thread controlled by thread pool and improved the resource utilization rate. Thus, the system was running for real-time and efficiently. Moreover, we can monitor the queues data conveniently. This design provides a thought to web data mining and analysis system.

key words: producer-consumer; two-dimensional matrix queues; public opinion; thread pool

0 引 言

随着网络成为继报纸、无线广播和电视三大

传统的传播媒体之后的“第四媒体”,网络舆情对社会的影响与日俱增,舆情信息的挖掘与发现^[1-5]已成为研究热点.面对 Web 海量数据,舆情监控

收稿日期:2013-04-11

基金项目:湖南省自然科学基金资助项目(11JJ6047);湖南省科技计划基金资助项目(2011FJ3087);衡阳市科技局基金资助项目(2012KG68)

作者简介:雷龙艳(1988-),男,湖南岳阳人,南华大学计算机科学与技术学院硕士研究生.主要研究方向:信息检索,数据挖掘,机器学习.

系统要求做到实时高效,以能够快速的获取互联网数据并发现敏感舆情信息. 多线程技术可以广泛地应用于解决各种高并发和请求低效问题^[6-7], 基于多线程的生产者—消费者模型是提高连续性和实时性常用的一种设计,有不少研究开发人员已应用这一模式进行生产和实验,淘宝网在模块耦合性问题上选取生产者—消费者模型^[8], Web 服务环境中的复杂事件处理中使用应用通知生产者和通知消费者来构造通用的底层事件通知模块^[9],视频会议系统的设计中,在生产者—消费者模型基础之上,实现了一套集中式的多点视频会议系统^[10].

但是,在某些环境下,比如在舆情监控系统中需要按网站域名分类抓取与分析,及时的回收已完成任务的资源,实时便捷的监控和诊断队列的运行情况等,传统的生产者—消费者的多线程设计模式的应用并不理想. 本文探索将基于二维生产者—消费者队列模型应用于舆情监控系统中,为 Web 数据抓取分析系统提供思路和可供借鉴的经验.

1 二维生产者—消费者队列模型

经典的生产者—消费者问题是一个多线程同步问题的经典案例,该问题描述了两个共享固定大小缓冲区的线程——即所谓的“生产者”和“消费者”. 生产者的主要作用是生成一定量的数据放到缓冲区中,与此同时,消费者也在缓冲区消耗这些数据. 程序要保证生产者不会在缓冲区满时加入数据,消费者也不会不在缓冲区中空时消耗数据. 经典生产者—消费者模式的明显不足之处在于数据存储能力有限,数据量大时显得臃肿而有失灵活性.

二维生产者—消费者队列模型是基于经典生产者—消费者问题的扩展设计. 分为两步:横向扩展和纵向扩展.

横向扩展思想:在很多程序设计场景中,数据流需要经过多道程序进行按次处理,而且要求连续性和实时性强. 因此,可将一个共享的缓冲区变成多个缓冲区,缓冲区按数据流处理次序依次首尾相连得到缓冲区链,如图 1,队列 1 和队列 2 之间的处理线程既作为队列 1 的消费者也作为队列 2 的生产者,其余队列类同. 在数据流的处理过程中,将会有线程的休眠和唤醒.

纵向扩展思想:面对数据流较大,处理时间较长的情景,一条如图 1 的生产者—消费者链往往

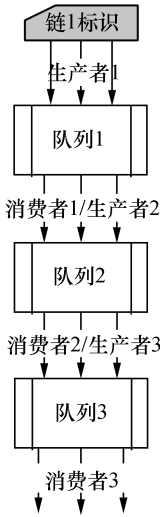


图 1 生产者—消费者队列链

Fig.1 Queue chain of producer-consumer

会显得过于臃肿而不便于控制管理. 二维生产者—消费者队列模型基于分而治之的思想,将如上生产者—消费者链进行纵向扩展而得到的,如图 2.

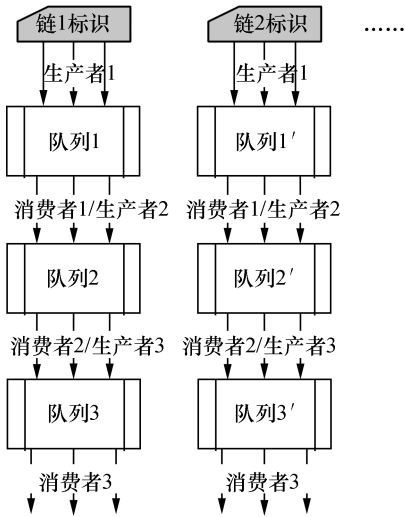


图 2 二维生产者—消费者队列链

Fig.2 Queue chain of two-dimensional producer-consumer

扩展成多条链后,每条链根据业务找一个最合适的数据流分类标准作为链标识,这样可将一个大的数据流按照合适的规则进行分流. 横向分析,数据处理流程不变,纵向分析,队列实例会成倍增长,但队列大小将因为分流而成倍减小,生产

者消费者线程会根据链标识对相应的队列进行生产和消费,这样带来的优势有三:

- 1) 将大数据流分解成小数据流进行处理,有利于数据流的分类管理,如较细粒度分类记录日志,便于更准确的调试诊断;
- 2) 链之间互不干扰,一个链阻塞了不会影响另一个链. 另外,有些数据流小的分流将会先结束处理,先释放资源;
- 3) 有利于多线程的并发执行,提高执行效率.

2 二维生产者—消费者队列模型在石鼓舆情监控系统中的应用

2.1 石鼓舆情监控系统简介

石鼓(SCool)舆情监控系统是南华大学计算机学院知识科学与信息检索团队自主研发的校园舆情实时监控系統. 系統將实时收集南华 BBS, 红网, 新浪新闻等站点网页, 然后对下载网页进行内容获取、分词和索引, 最后进行舆情分析过滤, 整个过程实时连续, 如图 3 系統结构图左部分所示.

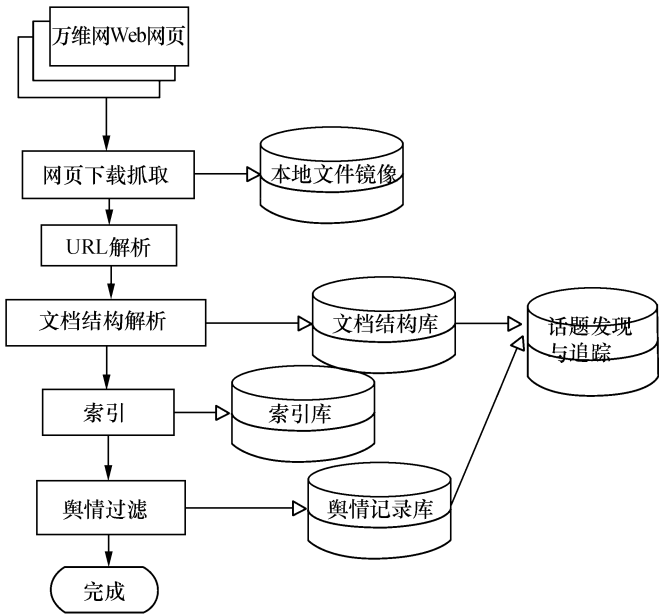


图 3 舆情监控系统结构图

Fig. 3 System structure picture of public opinion monitoring

2.2 二维生产者—消费者队列模型的应用

由于系統要求实时连续, 因此设计出基于生产者—消费者的舆情过滤队列链, 得到如下图 4 所示. 按处理前后顺序看有六个队列, 依次是: 就绪下载队列, 队列中元素为待下载的 URL 标识字符串; URL 待解析队列, 队列中元素为已下载完 URL 标识字符串, 它等待进一步解析以获取其中链接进行进一步下载; 文档结构待解析队列, 队列中元素为已完成 URL 链接解析后待进行页面结构解析的 URL 标识字符串; 待索引队列, 队列中元素为已完成文档结构解析后得到的待索引自定义文档对象; 待过滤队列, 队列中元素为已完成索引的待过滤文档对象; 完成队列, 队列中元素为已完成舆情过滤的页面 URL 标识. 队列元素的出入由生产者—消费者线程完成: 就绪待下载队列初始化

时会加入要抓取的网站根链接 URL, 一般是域名或者某版块地址, 下载线程作为就绪待下载队列的消费者和 URL 待解析队列的生产者, URL 解析线程作为 URL 待解析队列的消费者和就绪待下载队列、文档结构解析队列的生产者, 其他线程类同. 除了完成队列, 其余都是有界队列, 当这个链任务完成后, 完成队列的元素将会得到清理.

以下下载线程为例, 当就绪待下载队列为空时, 该线程将休眠, 直至新的 URL 加入队列时唤醒; 当 URL 待解析队列已满时, 下载线程阻塞等待, 直至队列中有元素出队.

然而面对数据流较大, 尤其是程序需要按站点或模块灵活处理控制时, 一条如图 1 的生产者—消费者链往往会显得过于臃肿而不便于控制管理. 二维生产者—消费者队列模型基于分而治之

思想,将如上生产者—消费者链进行纵向扩展而得到的.如图 5 所示,该图只画出了二维队列链的概要.每条链以抓取网站的域名作为标识,即将抓取的 WEB 数据流以域名为分界点进行分流,各生产者—消费者线程将根据域名标识对相应的队列进行操作.各条队列链互不干扰.

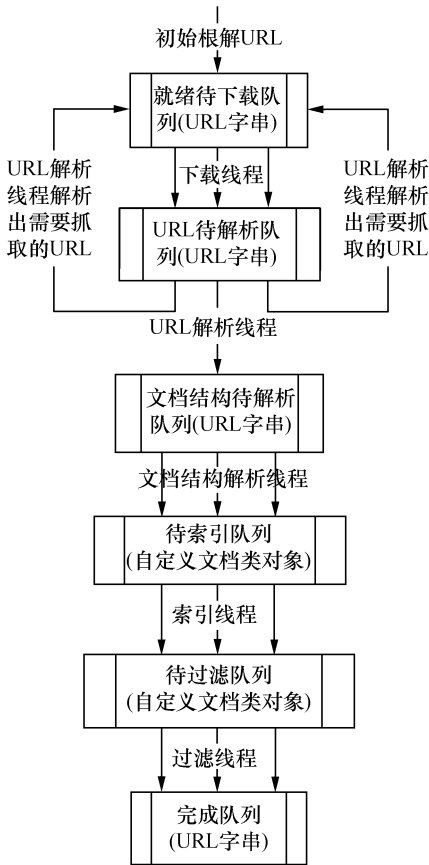


图 4 舆情过滤队列链

Fig. 4 Queue chain of public opinion filtering

为了控制生产者—消费者线程更好的工作,会把这些线程交由他们对应的生产者—消费者控制线程进行管理.这些控制线程将负责生产者—消费者线程的产生和消亡,并将它们抛入线程池进行协调运行,一个生产者—消费者线程实质就相当于一个任务请求.

另外,系统中一些辅助线程将会为系统的稳定性提供帮助.比如日志线程将会周期的将其中某些队列中的元素信息按域名分类写入记事本,一是可以为程序的调试提供详尽的依据;二是如果程序因停电等原因意外终止,可将已有的最近日志文件重新读入队列,以避免重复繁杂的网页内容分析过程.

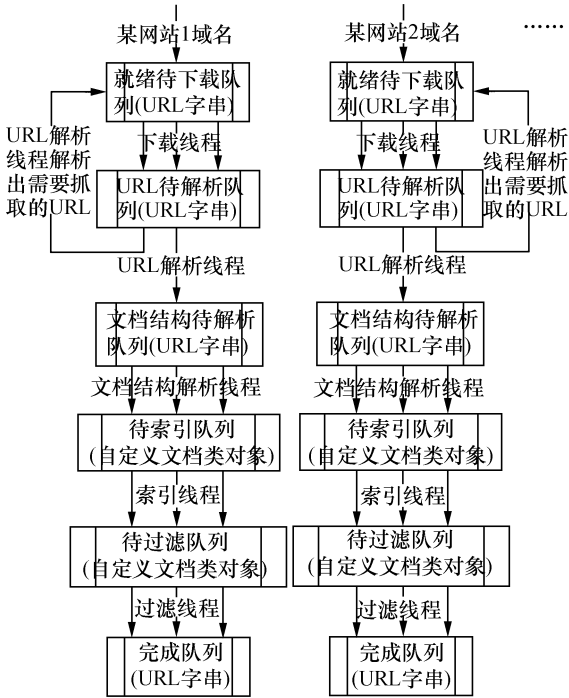


图 5 舆情过滤系统二维生产者—消费者链简图

Fig. 5 Queue chain of two-dimensional producer-consumer in public opinion monitoring system

2.3 二维长度值设定

二维长度值的设定准则基于一组实验分析得出.实验服务器运行测试环境:Windows Server 2003, 4 核 Intel® Xeon® CPU E5506 @ 2.13 GHz, 16 G 内存, Tomcat7.0.

在石鼓舆情监控系统的实现中,分三组进行实验,第一次实验对象为南华 BBS 五大版块,该五大版块页面数相当,分析处理程序相同,第二次为红网三大版块,该三大版块页面数相当,分析处理程序相同,第三次是前两次实验对象的组合,南华 BBS 和红网的分析处理程序不同.三次分别进行抓取分析过滤实验,改变二维队列的链条数,在经过 30 min 的下载分析后,通过日志分别得到处理完成后的页面数(论坛页面数充足),取每分钟的平均处理页面数,得到如图 6 所示.

前两次实验中,对于一条普通生产者—消费者链,即队列链数为 1 时,平均每分钟处理页面数在 32 个左右,随着队列链的纵向扩展,平均处理页面数呈线性增加趋势,当二维长度为 6 时,基本达到峰值,随后平均处理页面数有下降趋势.

第三次实验中,由于对于两大站点的处理分析程序不同,如果把二维队列长度设为 1,将很难实现分类分析和管理的,所以二维队列长度至少为 2,以到达分而治之的目的.随着二维长度的加大,

多线程的并发效率提高,基本在长度为6时达到峰值,随后处理效率也有下降趋势。

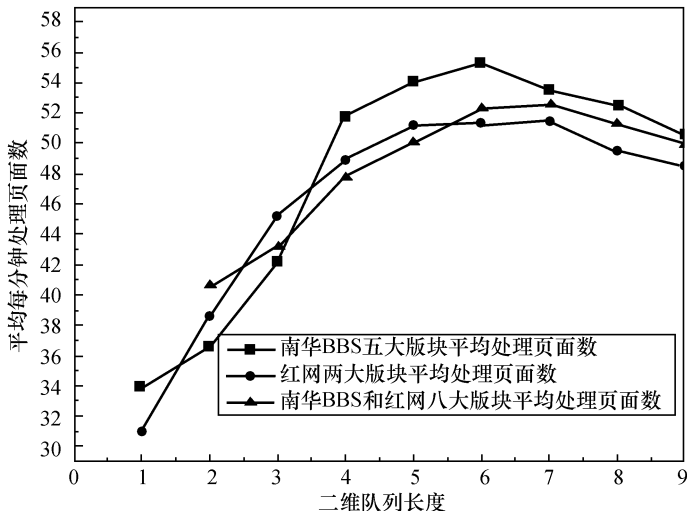


图6 三次实验平均每分钟处理页面数

Fig.6 The number of handled pages per minute on average

因此,基于以上实验分析,对于二维长度值设定,可遵循以下设定准则:

1) 以要抓取站点的模块数或站点数为标准,二维长度值至少大于等于站点数或模块数.这样将为日志记录和管理员的分析带来便利,也为后续的分析处理程序提供分别处理的可能,因为很可能不同的站点或模块所需要的分析处理程序不同;

2) 分块粒度不宜太细,二维长度不宜太大.如果因为准则1)使得二维长度值太大而导致额外需要过多的资源来维护自身的队列资源并最终导致了处理性能显著下降,则可适当将模块聚合以减小二维长度值或增加硬件资源;

3) 二维长度设定后,进入各队列链的数据量(网页数)应基本相当且不能小于队列容量,这样可保持程序的高并发量以提高效率.

3 结束语

基于生产者—消费者模式的二维队列链适合Web数据挖掘与分析系统,它能够有效的利用系统资源,使处理分析过程连续高效,并能够更细粒度的进行监控与调试诊断.本文通过在石鼓舆情监控系统中的应用证明了该方法的有效性,它为以后的web数据挖掘提供借鉴.

参考文献:

[1] Bollen J, Mao H, Pepe A. Modeling public mood and e-

motion; Twitter sentiment and socio-economic phenomena [C]//Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11), 2011. Barcelona, 17-21.

[2] Mahmud A, Ahmed K Z, Khan M. Detecting flames and insults in text [C]//Proceedings of the Sixth International Conference on Natural Language Processing, 2008.

[3] Allan J, Harding S, Fisher D, et al. Taking Topic Detection From Evaluation to Practice [C]//Hawaii International Conference on System Sciences (HICSS 05). Hawaii, MA: Kluwer, 2005: 199-204.

[4] Pena-Shaff J B, Nicholls C. Analyzing student interactions and meaning construction in computer bulletin board discussions [J]. Computers & Education, 2004, 42 (3): 243-265.

[5] 刘德鹏. 互联网舆情监控分析系统的研究与实现 [D]. 成都市: 电子科技大学, 2011.

[6] 骆斌, 费翔林. 多线程技术的研究与应用 [J]. 计算机研究与发展, 2000, 37(4): 407-412.

[7] 杨开杰, 刘秋菊, 徐汀荣. 线程池的多线程并发控制技术 [J]. 计算机应用与软件, 2010, 27(1): 168-170, 179.

[8] 高佳. 淘宝网网络数据处理系统设计与实现 [D]. 北京: 北京交通大学, 2011.

[9] 吴凡, 丁岳伟, 曹健. Web服务环境中的复杂事件处理 [J]. 计算机应用研究, 2009, 26(7): 2550-2553

[10] 褚蓬飞, 张焕强, 方贵明. 视频会议系统中数据缓冲机制的研究 [J]. 计算机应用研究, 2006, 23(5): 67-69.