

文章编号: 2096-1472(2017)-11-18-03

基于缓存技术的多线程下载软件的改进研究

谌 颀

(广东技术师范学院天河学院, 广东 广州 510540)

摘要: 本文通过对WEB浏览器访问网络资源特性的研究, 将浏览器缓存技术应用到下载软件系统, 提出一种新的基于图片资源缓存的下载软件设计方法, 可实现多线程、断点续传、限速等功能, 改进了传统软件下载图片时的不足。测试结果表明, 该设计具备良好的稳定性和可行性, 图片资源下载速度较传统软件有了较大提升。

关键词: 多线程; 断点续传; 下载; 图片缓存

中图分类号: TP301 **文献标识码:** A

The Improvement Research of Multi-Threaded Download Software Based on the Cache Technology

CHEN Hang

(Tianhe College, Guangdong Polytechnic Normal University, Guangzhou 510540, China)

Abstract: Based on the characteristics of web browsers to access network resources, the study applies the browser cache technology into the download software and puts forward a new download software design method based on image resource cache, which can achieve the functions of multi-thread transferring, broken-point continuingly-transferring and download rate restriction, etc., and improve the image download through the traditional software. The test results show that the design has good stability and feasibility, and the image download rate is greatly increased.

Keywords: multi-thread; broken-point continuingly-transferring; download; image cache

1 引言(Introduction)

互联网的飞速发展促进了网络资源爆炸式成长。用户通常使用浏览器获取网络资源信息, 而通过下载软件下载资源则成为用户最普遍的应用^[1], 主流的下载软件国内有迅雷、快车, 国外有Orbit、TuoTu和电驴等。

研究浏览器访问网络资源的特点发现, 为了提高网页访问速度, 通常它会采用累积式加速的方法, 即将曾经访问过的网页内容, 如图片、cookie等存放到本地的缓存区域, 当有访问请求时, 浏览器将首先搜索该目录。如果内容已存在, 则直接从缓存中读取内容, 以此达到提高访问速度的目的。本课题基于IE浏览器缓存机制, 将该技术应用到下载软件系统的设计中, 提出了一种基于图片资源缓存的下载软件设计方法, 改进了传统软件下载图片资源的不足。

2 主要技术原理分析(Main technical principle analysis)

2.1 线程与线程同步

本系统采用了线程技术, 与进程相比, 线程的主要特点有: 轻型实体, 独立调度, 可并发执行, 共享进程资源^[2]。由于线程很“轻”, 故线程的切换非常迅速而且开销小, 并发执行和共享资源方面优于进程。

线程同步的主要任务是对多个相关进程在执行次序上进行协调, 以使并发执行的诸进程能有效地共享资源和相互合

作^[3], 从而使程序的执行具有可再现性。

2.2 图片缓存下载机制

基于浏览器缓存机制, 用户下载网页图片资源时, 可以先查找缓存区域。如果目标图片已经存在, 则从缓存中下载; 如果不存在, 则从远程服务器下载。本文所述方法采用基于IE浏览器缓存机制实现, 具体步骤在第五部分详细阐述。

2.3 下载限速原理

本文提到的限速是指下载客户端的限速, 而非服务器端的限速。由于网络资源下载会占用较大带宽, 必须对下载流量进行限制。本设计采用线程休眠来减少流量请求, 从而实现流量限制。

3 多线程下载软件系统设计(Multi-threaded download software system design)

3.1 系统整体框架

系统基于.NET平台开发, 采用面向对象的方法与双层架构并行的设计理念。双层结构包括表示层与逻辑层。逻辑层提供一切逻辑上的处理, 如下载、切割文件等; 表示层是用户的功能界面。在分层架构的思想指导下, 逻辑层与表示层具有较小的关联性, 易于后期的维护和修改。

3.2 系统核心类交互关系及功能

本系统采用面向对象的编程思想, 所有核心的代码都由类构成。系统下载类主要有下载模型DownloadBase、

5.2 图片缓存下载技术性能分析

本文4.4节介绍了IE图片缓存下载技术的实现方法, 本节就对该技术的性能与迅雷软件进行对比分析。本实验的数据源是从互联网上任意选取20个网页, 实验中对比了两种下载软件的性能, 结果见表2。在全部缓存到本地的情况下, 缓存下载技术下载速度较传统下载技术有了很大的改进。

表2 网页图片测试表

Tab.2 Web image test table

网站	URL	图片数量	缓存下载时间/s	迅雷下载时间/s	相差倍数
新浪	http://slide.news.sina.com.cn/c/slide_1_955_11727.html	13	0.01	16.38	1638.00
桌酷1	http://www.zhuoku.com/zhuomianbizhi/design-sport/20100426134638(1).htm	19	0.05	31.40	628.00
桌酷2	http://www.zhuoku.com/zhuomianbizhi/moviecn/20100425145246(1).htm	19	0.01	24.98	2498.00
BBC	http://news.bbc.co.uk/2/hi/asia-pacific/8648970.stm	21	0.04	20.00	500.00
央视	http://photo.cntv.cn/2015/06/06/PHOA1433561649963512.shtml	22	0.03	29.22	974.00
新浪博客	http://blog.sina.com.cn/s/blog_49a1a0250100iyia.html?tj=1	23	0.07	30.61	437.29
猫扑2	http://game.mop.com/wlyx/2027674.shtml	23	0.02	28.13	1406.50
CNN	http://edition.cnn.com/2010/WORLD/asiapcf/04/29/pakistan.mehsud/index.html?hpt=T1	23	0.04	19.58	489.50
BBC	http://news.bbc.co.uk/2/hi/Americas/8650620.stm	24	0.05	47.98	959.60
猫扑1	http://auto.mop.com/data/pic/pic_14283_181_0.html	25	0.05	41.68	833.60
腾讯	http://digi.tech.qq.com/jk.htm	25	0.03	58.18	1939.33

依据表2, 可得两种方法下载时间比较的柱形图, 如图4所示。

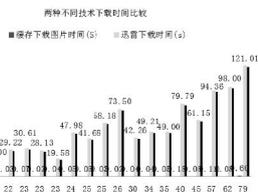


图4 两种不同下载技术时间比较柱形图

Fig.4 Download time comparison between the two methods

从图4中可明显地看到, 相同的图片资源下载, 运用了缓存技术的下载软件下载耗时很小。对比二者, 下载差距如图5所示。

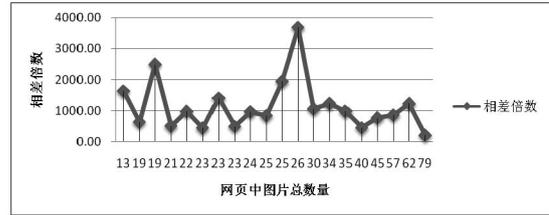


图5 两种不同下载技术相差倍数柱形图

Fig.5 Download time multiplier comparison between the two methods

从中可以看到, 当下载网页中26张图片的时候, 迅雷下载耗时是图片缓存下载技术的3500多倍, 即使在差距最小的情况下, 迅雷下载耗时也是后者的200多倍。由实验分析可知, 图片缓存下载技术无论是在下载速度上, 还是在占用本地资源上, 较传统下载软件有较大优势。

5.3 网页图片识别筛选准确率分析

本实验使用了表2的20个网页进行识别筛选测试, 其结果如图6所示。

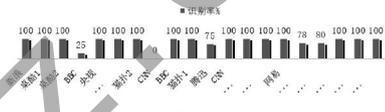


图6 下载识别功能

Fig.6 Download recognition function

由图可知, 20个网页中有15个识别率高达100%, 有4个识别率超过50%, 只有1个识别率为0。从所选用的20个网页来看, 本图片筛选模型平均准确率达到88%。

结合此识别模型与图片缓存下载技术, 可形成一个完整的下载方案: 一方面提高了识别用户目标图片的准确度, 另一方面下载的速度有着质的飞跃。

6 结论(Conclusion)

本设计运用了下载限速、断点续传、多线程、缓存下载等技术, 取得了良好的实现效果, 最大限度利用了带宽, 提高了下载效率。针对目前主流下载软件存在的不足, 提出基于浏览器缓存特性实现的图片缓存下载方法, 显著改善了图片资源下载速度。

参考文献(References)

- [1] 周学威, 闫鑫, 赵祥云, 等. 基于SOCKET的多线程下载工具的开发[J]. 电子测试, 2011(08):104-106.
- [2] 杨代庆, 王志革, 王星, 等. 一种断点续传的多线程新闻组抓取方法及存储结构[J]. 现代图书情报技术, 2011(02):29-33.
- [3] 睦俊华, 刘慧娜, 王建鑫, 等. 多核多线程技术综述[J]. 计算机应用, 2016(06):239-242.
- [4] 王莉敏, 梁正和, 段全锋. 基于HTML5大文件断点续传的实现方案[J]. 计算机与现代化, 2016(03):91-95.
- [5] 王枫, 罗家融. Linux下多线程Socket通讯的研究与应用[J]. 计算机工程与应用, 2004(16):106-109.

作者简介:

谌 颀(1980-), 男, 硕士, 高级工程师. 研究领域: 网络信息系统, 图像识别, 数据挖掘, 云计算, 个性化推荐技术.