基于 Heritrix 视频资源抓取的研究与实现

徐 枫[1,2] 归伟夏[1]

[1](广西大学计算机与电子信息学院 南宁 530004)
[2](广西银行学校信息与管理教学部 南宁 530007)

摘 要 教学视频资源是教学资源库的重要组成部分，对视频资源的添加是系统平台的一项重要工作。目前很多教学资源库对视频资源的添加采用手工方式进行，效率不理想且工作量极大。通过引入网络爬虫，利用 Heritrix 的扩展功能，可以定制相应的模块，使其自动抓取网络上的课程视频资源。而通过优化其抓取算法，可以提高资源库中视频的抓取效率和准确率。

关键词 视频资源；Heritrix 抓取；主题爬虫；垂直搜索
中图分类号 TP 311.52 TP 302.1 **文献标志码** A

Research and Implementation of Video Resource Capture Based on Heritrix

XU Feng[1,2] GUI Weixia[1]

[1](*School of Computer*, *Electronics and Information*, *Guangxi University*, *Nanning* 530004, *China*)
[2](*Department of Computer*, *Guangxi Banking School*, *Nanning* 530007, *China*)

Abstract The video teaching resource is an important part of the teaching resource library, and it is important to add video resources for the system platform. At present, the adding of video resources for many teaching resource libraries is done by hand, which is of low efficiency and produces heavy workload. By introducing the network crawler and using the extended function of Heritrix, the corresponding module was customized to make it automatically grasp course video resources from the network. And it could improve the video grasping efficiency and accuracy of the resource library by optimizing its grasping algorithm.

Keywords video resources; Heritrix grasp; the theme crawler; vertical search

1 引 言

为了实现教学资源的知识积累和共享、辅助教师进行备课、学生进行协同学习，越来越多的院校建立了视频资源库。面对网络上越来越多的、大量膨胀的视频数据资源，如何从中找到所需的数据资源，是当前视频检索领域研究的一个热点和难点。

能否快速、准确地查找到所需视频资源是教学资源库的一个重要问题。现有的许多视频资源的来源往往采用手工添加或者使用通用的搜索引擎进行检索。通用搜索引擎的特点是大而全[1]，但对只需要查找特定资源的用户来说，这样往往

收稿日期：2014-4-4
作者简介：徐枫(通讯作者)，工程硕士，讲师，研究方向为数据挖掘，E-mail：wangluo0301@126.com；归伟夏，博士，副教授，研究方向为电子商务。

影响工作的效率。我们希望能够精准地查找并下载到所需的资源，因此教学视频资源主题爬虫搜索是一个值得研究的课题。Heritrix 的强大组件扩展功能正好能够满足个性化、专门化抓取的要求。本文先介绍了网络爬虫的工作流程[2]，通过研究 Heritrix 爬虫的相关技术，定制 Heritrix 的各个组件，以实现面向特定主题网页信息的抓取，提高抓取的准确率。同时引入 ELFHASH-x1 算法[3]，采用多线程抓取网页来提高抓取的效率。

2 相关领域研究

从技术角度来看，网络资源的检索与下载可分为两种方式：通用搜索引擎和主题搜索引擎。从上个世纪 90 年代开始，国内外就着力研究通用搜索引擎，有许多产品至今仍在互联网上广泛使用，如 Google、百度等搜索引擎，以及优酷、土豆等提供搜索的视频网站等。

在主题搜索引擎[4]研究方面，国外的研究起步较早，也有较多的成型系统。如 1999 年美国卡内基梅隆大学的 A.K.McCallum、M.Niga 等人针对计算机领域设计的 CORA；斯坦福大学设计了用于 Google 的爬虫；North Carolina 大学专门建立法律资源的系统 LIBClient——IRISWeb[5]，以及由Elsevier 公司在科技信息检索领域内建立的 Scirus 系统[6]。

国内关于主题搜索的研究也较多，且在不同的行业具有不同的应用。如针对旅游行业建立的去哪儿网和赛迪网，为搜索房地产信息而建立的房老大。国内研究高性能爬虫的产品有北大天网等。而利用主题爬虫来构建教学视频资源的研究则较少涉及。因此本文的研究具有一定的现实意义。

3 基于 Heritrix 的 3dmax 的视频资源抓取与实现

Heritrix 是一个功能强大、可扩展性较好的开源网络爬虫系统。Heritrix 工程由 IA 于 2003 年开始研发，源代码采用 java 编写，其最初的目的是开发一个特殊的爬虫，对网上的资源进行归档并建立网络数字图书馆[7]。它可以帮助用户从网络上下载有用的资源。Heritrix 最大的优点是用户可以根据自己需要来定制相应的抓取方式和功能，以实现按用户要求的主题抓取相应的内容。Heritrix 由最初 Heritrix 1.0.0 升级到现在的 Heritrix 3.1.1，它也进行了一系列的改进。本文通过以 3dmax 主题视频资源的抓取为例，采用 Heritrix 中的各种组件定制和扩展，以期达到预期抓取目标。

3.1 基于 Heritrix 视频资源定制

3.1.1 Heritrix 工作流程

Heritrix 主要有三大部件[2]：范围部件、边界部件和处理器链。范围部件：主要按照规则决定将哪个 URI 入队；边界部件：跟踪哪个预定的 URI 将被收集和已经被收集的 URI，选择下一个 URI，剔除已经处理过的 URI；处理器链：包含若干处理器获取 URI，分析结果并将它们传回给边界部件。

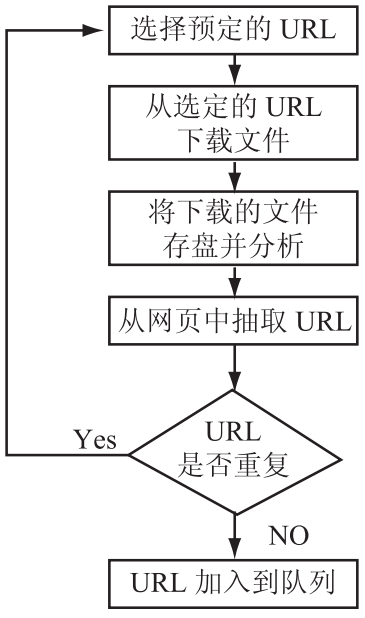

图 1 Heritrix 工作流程图

Fig. 1. Heritrix working flow chart

Heritrix 是一个爬虫框架，它可以根据需要替换相应的组件，其工作流程为递归进行（如图 1），具体流程如下：

（1）选择一个预定的 URL；（2）从选择的 URI 的网址下载文件；（3）分析、归档下载到的内容；（4）从分析到的内容中选择感兴趣的 URI，加入预定队列；（5）标记已经处理过的 URI。

3.1.2 种子的选择

选择好的初始种子，是成功实现主题爬虫抓取目标的基础。要实现高效的主题抓取，我们往往会选择与主题相关的大型网站或行业内有名气的网站。本文选择国内视频教学网站第一视频教程网（http://video.1kejian.com/）为例作为抓取站点。在网站中，3dmax 教学视频资源列表所在网页 URL 为: http://video.1kejian.com/computer/3d/。在该页面中所有与 3dmax 教学视频相关的资源都以列表形式展出，查看页面源文件发现这个页面可以链接到所有 3dmax 教学视频资源二级资源列表所在页面。如其中一个视频资源二级列表页面为：http://video.1kejian.com/computer/3d/64922/，这个页面是爬虫子系统需要抓取的内容页面的资源列表页。可以采用正则表达式来表示 3dmax 教学视频资源二级资源列表页面：http://video.1kejian.com/computer/3d/\d{4,5}/。

3.1.3 3dmax 视频资源的定制

研究并确定了 URL 所采用的种子之后，接下来我们就可以使用 Heritrix 来开始抓取页面的过程。如果不给 Heritrix 设定规则，它会将初始页面中所有有效链接都进行抓取。而种子页面中一般会含有很多与主题无关的链接，会影响抓取的时间和效率。因此，需要对 Heritrix 的抓取类 extractor 进行改进，使其只抓取与主题相关的特定链接。在 Heritrix 项目下建立新建类 Vs3dextractor，该类继承父类 Extractor，采覆写其 extract 方法来实现。具体实现的过程是：在 extract（CrawlURI curi）方法中先判断传入其中的参数是否为第一视频教程网站中的 3dmax 视频资源列表二级页面，如果是，则解析出页面内的链接地址 http://video.1kejian.com/computer/3d/ 后的 id，则生成对应 id 号的 3dmax 教学视频资源列表正则表达式：http://video.1kejian.com/computer/3d/\d{4,5}/。然后将其加到等待 FrontierScheduler 处理的列表中，等待后面进行处理。定制 Extrator 算法思想如下所示：

输入：待抓取的 URL 列表 Curi

输出：newurl 加到队列 queue

（1）待抓取的 URL 列表 Curi；

（2）从 Curi 中取出要解析的内容 content；

（3）在内容中查找与 3dmax 资源列表页一致的页面；

（4）while（match.find（））

（5）把找到的地址传给参数 newurl;

（6）if（地址 http 开始）

把 newurl 加入到链接队列 queue;

else

If（地址以”\”开始）

在 newurl 之前添加“video.1kejian.com/computer/3d/”并把其加入到链接队列 queue;

在资源二级列表所在的 URL 页如 http://video.1kejian.com/computer/3d/64850/中，我们查找到了类似 http://video.1kejian.com/video/?64850-0-0.html 资源的详细页面。对其进行分析则可以知道，其中的主机部分是 http://video.1kejian.com/，这是第一视频教程网显示页面的域名，“video”表示的是资源类别为视频，而最后的“?64850-0-0.html”是视频资源详细页面的 ID 编号，因此就可以根据 URL 的特点定出一个正则表达式，如下：http://video.1kejian.com/video/\\?\\d{5}-\\d{1}-\\d{1}.html。扩展 FrontierScheduler 实现所抓取网页中符合条件的 URL，重写 schedule（）方法，只有满

足条件的 URL 才允许加入到等待队列中。定制 FrontierScheduler 算法思想如下：

输入：待抓取的 URL 列表 Cauri

输出：抓取的网页

(1)把 Cauri 转化为字符串 uri;

(2) compile(比较)uri 与 3dmax 视频资源页的正则表达式;

(3) if 该 URL 能匹配该正则表达式;

抓取该 URL 对应的网页文件;

else

转到下一个 Cauri;

3.2 抓取效率提升

3.2.1 取消重复访问 robots.txt 文件

robots.txt 是一个用于说明网站中哪些链接或内容禁止爬虫访问的文件。有些网站希望自己的某些内容不被搜索引擎所抓取，因此网站的作者会在网站中放置 robots. txt 文件，在 robots.txt 文件中说明哪些内容或链接不予抓取。如果当一个网站没有放置 robots.txt 文件时，Heritrix 总是要花很多时间试图去找到并访问 robots.txt 文件，并且可能进行多次尝试，消耗过多的时间来判断 robots.txt 文件是否存在。这对提高信息抓取的效率产生很大的不良影响。因此，为了提高网络爬虫的抓取效率，我们可以对访问 robots.txt 的部分进行修改。在 Heritrix 框架中，对 robots.txt 文件的处理是由 PreconditionEnforcer 部件来完成的。在 PreconditionEnforcer 中，有一个 private 类型的方法，它的方法的引用为：private boolean considerRobotsPreconditions(CrawlURI curi)该方法的含义为：在对参数所表示的链接进行抓取前，看一下是否存在一个由 robots.txt 决定的先决条件。如果对每个链接都进行这样的处理，很有可能延阻整个抓取任务。为了避免上述状况，我们需要对其进行相应的修改。利用该方法返回 true 时的含义为需要考虑 robots.txt 文件，返回 false 时则表示不需要考虑 robots.txt 文件，可以继续将链接传递给后面的处理器。所以，最简单的修改办法就是将该方法作为注释来处理，保留 flase 返回值。这样能够使抓取的速度提高了至少 50% 以上。

3.2.2 并行处理

在默认情况下，Heritrix 会根据 HostnameQueueAssignment Policy 策略算法来计算每个 URL 地址的 Key 值，这种方法是采用 URL 的主机域名为 Key 值。因为对一个主机地址只采用一个线程来进行处理，故这种抓取方式速度效率低。本文中的大部分教学视频资源 URL 都在同一个域名下，这就造成了某一个队列太长的情况。当一个线程从该队列中获取一个 URL 链接后，这个队列就会处于阻塞状态，直到该链接处理完才从队列的头部取出下一个链接，这样就导致 Heritrix 的抓取效率低下，会在某个时间段过后处于一直没有进度的状态。因此，需要改变 Key 值的生成方式，使所有的 URL 较平均地散列到不同的队列中，以提高抓取效率。文中在 Heritrix 中扩展 queue-assignment-policy，实现一个继承自 QueueAssignmentPolicy 的类，覆写其中的 getClassKey( )方法。该方法将一个链接对象处理后，再调用散列算法生成一个 Key 值，相同 Key 的链接存于同一个队列中。散列算法有多种，范先爽在文献“基于 Heritrix 网络爬虫算法的研究与应用”中引入 BDKHash 算法[8]进行 URL 散列[9]。本中采用散列程度高 ELFHash-x1 算法生成 Key 值，覆写的 getClassKey( )的算法思想如下:

ELFHash-x1 URL 散列算法

输入：字符串 str

输出：str2

(1)令 hash=0；x=0；

(2)While(str 中的每个字符)

(3) hash 左移 4 位，把当前字符 ASCII 存入 hash 低四位；

(4)判断 hash 值的高 4 位是否非 0，因为非 0 时需要下面特殊处理，否则上面一步的左移 4 位会把这高四位给移走，造成信息丢失；

(5)如果高 4 位不为 0，则把前面 hash 的高 4 位跟 hash 的低 5-8 位异或；

(6)把高 4 位清 0；

(7)返回一个符号位为 0 的非负数，丢弃最高位。

4　性能测试与分析

4.1.1 抓取效率分析

以抓取页面内容 30M 为例，如图 2 采用 Heritrix 默认的单线程抓取，则需要 4 小时 9 分 30 秒。而图 3 引入多线程 ELFHash-x1 算法之后，抓取等量的内容，只需 1 小时 4 分 35 秒。抓取所需时间缩短 74.1%。可见引入多线程相

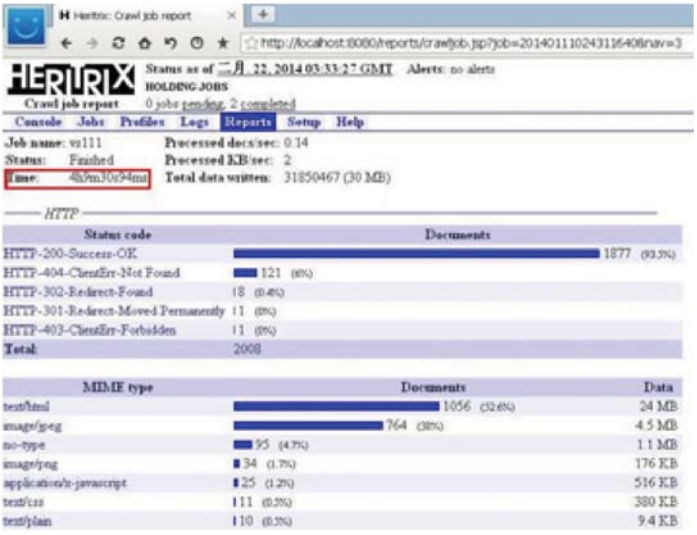

图 2　单线程抓取所需时间

Fig. 2.　Single thread crawling time

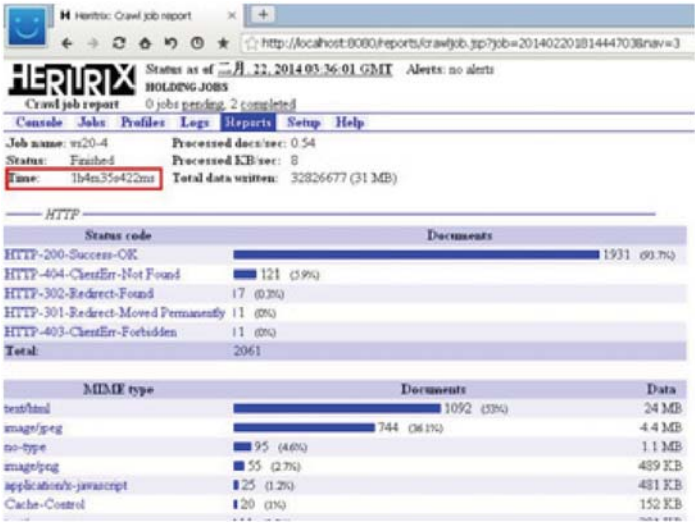

图 3　ELFHash-x1 算法抓取所需时间

Fig. 3.　ELFHash-x1 algorithm crawling time required

表 1 单线程与并行抓取所需时间实验数据

Table 1. Single thread and parallel grasping the time required for the experimental data

实验项目	抓取数据量(单位：M)	抓取所需时间（单位：秒）	（单线程时间-多线程时间）/单线程时间
单线程抓取	30M	14970	74.1%
多线程抓取	30M	3875	
单线程抓取	60M	31849	77.7%
多线程抓取	60M	7094	
单线程抓取	100M	42783	70%
多线程抓取	100M	12768	

关算法之后从而可以大幅提高抓取效率。表 1 则分别以抓取 30 M、60 M 和 100 M 的数据量为例，比较单线程和多线程的抓取效率。图 4 则对单线程与多线程抓取耗时以图例方式来作比较。

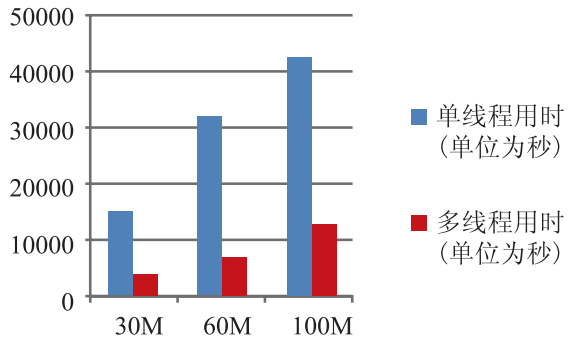

图 4 单线程与多线程抓取耗时对比图

Fig. 4. The crawling time of the single threaded and multithreaded comparison chart

4.1.2 准确率性能分析

表 2 显示采用不同的主题爬虫抓取同一目录下的主题相关性网页正确率实验。经定制扩展后的 Heritrix 以较高的准确率实现相关主题资源的下载，实验达到了预期的目标。

4.1.3 实验的对比及优缺点

从表 1 中可以看出利用 ELFHash 算法合理的分配 Key 值后对一个域名（http://video.1kejian.com/computer/3d/）下的网页抓取速度有了显著的提高，前后时间对比提高了 4 倍左右。实验计算机的网络下载速度对算法实验具有一定的影响，但在不同的时段计算机下载速度具有不确定性。ELFHash 算法分配的线程数并非越大越好，经实验发现，当线程数为 50 时为抓取效率理想，之后抓取效率随线程数加大反而下降。

表 2 三种主题爬虫软件抓取准确率比较

Table 2. The three topic crawler software capture accuracy

软件名称	下载总页数	主题相关性页数	准确率
Loalasam	238	73	30.6%
NwebCrawer	412	258	62.6%
Heritrix	505	492	97.4%

采用 Heritrix 来抓取视频教程网页具有可定制主题的优点，但其配置较为复杂。Heritrix 对并行网站的抓取效率较高，但对单个网站域名下的抓取效率效果并不太理想。

5　结束语

本文首先介绍了教学资源库中视频资源添加的问题，接着分析了搜索引擎目前的研究现状，引入主题爬虫对网络视频资源的抓取的方案。利用 Heritrix 强大的扩展性，定制 extractor 类和 FrontierScheduler 类以抓取学科主题相关的视频网页资源，提高抓取的准确率。在抓取的过程中取消每次的访问 robots.txt 检查，并引入 ELFHash-x1 算法，使其抓取能够并行处理，提高了抓取的效率。

参 考 文 献

[1] Khoo M, Hall C.What would ‘google’ do? Users’ mental models of a digital library search engine [C] // TPDL’12 Proceedings of the Second International Conference on Theory and Practice of Digital Libraries, 2012: 1-12.

[2] Lee HB, Nazareno F, Jung SH, et al. A vertical search engine for school information based on heritrix and lucene [C] // Convergence and Hybrid Information Technology Lecture Notes in Computer Science, 2011, 6935: 344-351.

[3] 赵永鑫, 雷霖. Heritrix 在电子信息垂直搜索平台中的应用 [J]. 成都大学学报(自然科学版), 2013, 32(2): 156-158.

[4] Ojala M. Thinking about search [J]. Online, 2012, 36(3): 5-5.

[5] Dempsey BJ, Vreeland RC, Sumner Jr. RG, et al. Design and empirical evaluation of search software for legal professionals on the WWW [J]. Information Processing & Management, 2000, 36(2): 253-273.

[6] Cummings J. Scirus: for scientific information only [J]. Reference Reviews, 2012, 26(6): 45-47.

[7] Liu DF, Fan XS. Study and application of web crawler algorithm based on Heritrix [C] // Advanced Research on Information Science, Automation and Material System/Advanced Materials Research, 2011, 219-220: 1069-1072.

[8] 朱敏, 罗省贤. 基于 Heritrix 的面向特定主题的聚焦爬虫研究 [C] // 2011 嵌入式技术开发论坛论文集, 2011: 65-68.

[9] 樊多妮, 李禹生. 基于 Heritrix 的网络主题爬虫算法研究与应用——以粮食网站交易信息为例 [J]. 现代物业, 2012, 11(9): 97-100.