DLB & OSS 2009论文选登

基于 Nutch的 Web网站定向采集系统 *

徐　健[1,2]　张智雄[1]

[1] (中国科学院国家科学图书馆　北京 100190)

[2] (中山大学资讯管理系　广州 510275)

【摘要】在对目前具有代表性的开源网络抓取软件 Nutch、Heritrix、WCT、Web - Harvest进行比较分析的基础上,提出基于 Nutch的 Web网站定向采集系统,并对种子站点的选取、抓取过程管理、网页去噪、新种子站点的发现等关键问题进行重点探讨。

【关键词】网站定向采集系统　Nutch　网站抓取　网页去噪

【分类号】G250.76

Targeted Websites Harvest System Based on Nutch

Xu Jian[1,2]　Zhang Zhixiong[1]

[1] (National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2] (Department of Information Management, Sun Yat - Sen University, Guangzhou 510275, China)

【Abstract】The paper analyzes typical open source Web crawl software, such as Nutch, Heritrix, WCT, and Web - Harvest. Following the analyzed result, it puts forward a targeted websites harvest system based on Nutch. Four key issues of this system are discussed emphatically, which are the initial seed websites selection, the harvest process management, the web page content denoising, and discovering of new seed websites.

【Keywords】Targeted websites harvest system　Nutch　Website crawl　Web page denoising

1 引　言

利用网络信息更新速度快、获取方式灵活等特点,可以实现对特定领域、学科的实时监测和有效分析。开展此类任务的第一步,就在于如何将相关网络科技信息内容存储到本地,继而为后续的信息抽取、链接分析、知识库构建、可视化等工作提供重要的基础数据。为了实现对 Web上特定领域英文科技信息的采集,笔者对目前具有代表性的开源网络抓取软件进行了分析,并最终选择在 Nutch基础之上进行多种扩展和改进的专题网站定向采集方案。通过对种子站点的选取、抓取实施、网页去噪、发现新的候选种子站点等关键问题的攻克,基本实现了网站科技信息定向采集系统,为后续工作的开展提供了质量较高的处理语料。

收稿日期: 2009 - 02 - 17

收修改稿日期: 2009 - 04 - 01

*本文系国家"十一五"科技支撑计划子课题"网络科技信息监测与评价"(项目编号: 2006BAH03B05)的研究成果之一。

2 Web抓取开源软件比较分析

通过对 SourceForge以及网络搜索引擎的检索,笔者认为比较具有代表性的网站抓取开源软件有:Nutch、Heritrix、WCT以及 Web - Harvest。上述 4种软件在功能上各有特色(见表 1)。Nutch开源软件不仅提供了抓取网页的功能,还提供了解析网页、建立链接数据库、对网页进行评分、建立 Lucene索引和提供检索界面等丰富的功能[1,2]。Heritrix开源软件提供了丰富的抓取设置选项,主要被用来获取完整的、精确的站点内容深度复制,包括获取图像以及其他非文本内容[3]。WCT(The Web Curator Tool)开源软件的主要特点在于能够对目标网站进行采集授权、采集调度、资源描述、资源收割、质量检查、采集结果提交等 Web收割过程进行有效管理[4]。Web - Harvest开源软件能以用户所指定的网页为抓取起始页,通过规则表达语法进行多层抓取,并抽取网页中以 XPath表达的内容片段,形成 XML文档[5]。

表 1 4种 Web抓取开源软件的特征比较

比较项目	Nutch	Heritrix	WCT	Web - Harvest
操作方式	命令行	Web控制界面	Web控制界面	界面/命令行
Web抓取功能	有	有	有	有
集群扩展能力	有	无	有	无
抓取内容完整性	只对可索引内容进行抓取	完整	完整	对网页特定字段进行抓取
内容索引功能	有	无	无	无
搜索功能	有	无	无	无
内容解析	有	无	无	针对特定字段
链接解析	有	无	无	无
网页评分	有	无	无	无
采集过程管理	无	无	有	无

由表 1可以看出,Nutch具有突出的功能特征和性能指标。本文的任务重点在于获取 Web上文本信息进行分析,而不是进行网络资源长期保存,因此 Nutch在抓取过程中对于图片、视频等无法解析的内容不进行保存的特征不但没有劣势,还可以节省大量的存储空间。另外,Nutch所具有的内容解析、链接解析、网页评分等功能为后续的 Web内容分析提供了更多便利。Nutch对分布式文件系统的支持,使多节点并行抓取和索引成为可能。基于上述分析,在任务中选择 Nutch进行 Web站点抓取任务。

3 Web网站定向采集系统整体设计

尽管 Nutch具有较丰富的功能和相对完备的结构体系,直接将其应用于 Web网站定向采集系统是不适宜的。在 Nutch现有结构基础上,还需要针对特定任务需求进行修改和完善工作。

经过改进和扩展后的抓取系统整体设计如图 1所示:

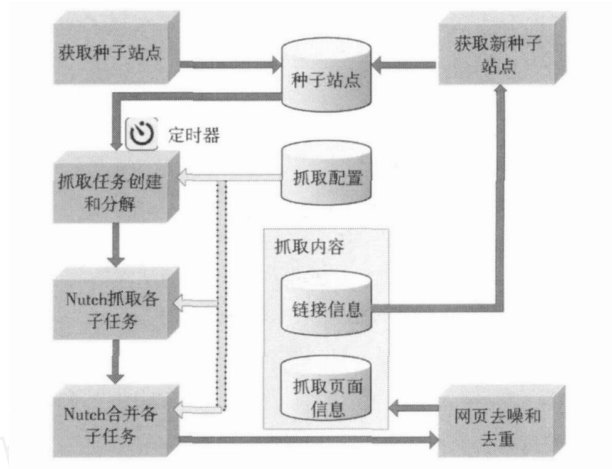

图 1 基于 Nutch的 Web网站定向采集系统整体设计

整个系统的运行流程可以分解为以下 6个步骤:

(1)获取种子站点。Nutch没有提供种子站点的获取机制,因此需要通过半自动或人工方式获得具有目标特征的站点相关信息,作为抓取种子存储于种子站点库中,以备抓取过程使用。

(2)抓取任务创建和分解。定时器周期性地激活抓取流程后,根据预先设定的抓取配置信息(例如种子站点类型、子任务所含最大种子站点数等),自动将种子站点进行分组,形成抓取子任务集合。对种子站点进行分组的好处在于能够有效解决单个抓取任务过于庞大造成的抓取过程意外/人为中断问题。

(3) Nutch抓取各子任务。根据预先设定的抓取参数(例如网站的抓取深度、抓取线程数、存放路径等),调用 Nutch的网站抓取接口,依次对各个子任务进行抓取。

(4) Nutch合并各子任务。在所有子任务抓取完成后,调用 Nutch的抓取目录合接口,将各个子任务对应的数据合并为一个数据集合。

(5)网页去噪和去重。通过 Nutch抓取的网页,除含有有效的正文内容外,还携带有广告信息、客户端运行代码、版权声明、栏目设置等噪音信息。为了给后续的信息抽取、分析步骤提供高质量语料,减少噪音信息的干扰,本文在网页抓取阶段设置网页去噪模块,对网

页内容进行过滤。网页去重功能保证了抓取内容数据库中存储的网页是不重复的,也可以识别新发布的页面。

(6)获取新的种子站点。通过对已抓取页面链接信息的分析,能够发现一些频繁被引用,但未在种子站点库中登记的网站,可以对这些网站作进一步的甄别,从中发现新的相关种子站点,并存入种子站点库,纳入新的抓取任务中。

基于 Nutch的 Web网站定向采集系统整体设计具有以下几个特点:

(1)对种子站点的动态管理。本系统实现了半自动地获取初始种子站点,以及根据已抓取网页的共链分析发现新的种子站点的机制,使种子站点库能够随着 Web网络的动态发展而得到相应更新。

(2)抓取配置的集中管理。Nutch本身是通过其 conf目录下的 nutch - site. xml以及 crawl - urlfilter. txt等文件进行配置管理的。将 Nutch原有的抓取配置项目和为实现任务管理而设置的配置项目进行了整合,通过集中管理,提高了配置效率。

(3)基于子任务的断点续传。通过实验发现,当种子站点数较多、抓取深度较深时,长时间的抓取容易发生因系统资源不足、网络中断或人为因素导致的抓取中断,而重新启动抓取程序意味着已抓取内容的重复抓取。通过抓取任务的自动分解和抓取管理步骤,按照预先设定的子任务大小,将抓取任务分解为若干子任务,基本实现了基于子任务的断点续传机制。

(4)网页自动去噪和去重。借助 Html Parser[6],根据噪音信息的一般特征,对网页进行去噪处理,为后续对网页的信息抽取、分析等操作提供高质量语料。去重步骤以 Nutch为每个网页计算得到的 MD5码为依据,通过 MD5码的比较来判断网页是否已存在。

4　核心问题的解决思路

在基于 Nutch的 Web网站定向采集系统中,以 Nutch原有架构和接口为基础,针对新任务需求提出新的抓取架构,重点解决了获取种子站点,实现抓取任务管理,进行网页去噪,以及获取新的种子站点这 4个方面的核心问题。

4.1　获取种子站点

对特定领域的网站进行定向采集,首先需要解决的问题是确定要采集的网站列表,也就是种子站点列表。对种子站点的要求是:该站点内容与要抓取的目标领域大体一致、具有较强的网络影响力、能够有效访问等。在通过半自动方式获得符合以上要求的站点相关信息后,作为抓取种子存储于种子站点库中。具体的实施思路参见图 2。

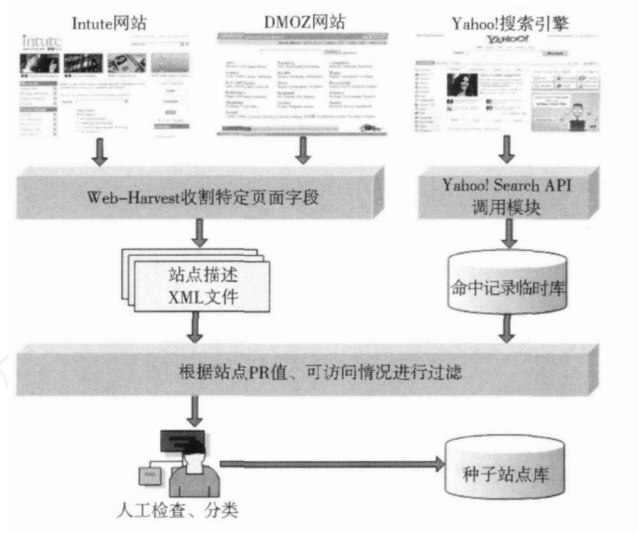

图 2　获取种子站点模块设计

在图 2中,首先借助 Web - Harvest开源软件对 intute[7]、DMOZ[8]等开放式分类目录中的特定领域站点列表进行抽取,形成站点描述 XML文件。为了获得更多的种子站点,可以使用能够代表领域特征的语词,通过 Yahoo! Search API[9]进行检索来获得更多的候选站点 URL。经过上述途径获得的候选站点,需要根据 PageRank值和连通情况指标进行初步过滤,将影响力不大和难以访问到的站点排除,最后通过人工方式进行核查和分类,最终获得高质量的站点种子。

4.2　抓取任务管理

Nutch提供了对个别企业网进行爬行和对整个互联网进行爬行这两套运行方案[10]。为了将爬行范围限制在特定领域种子站点范围内,进行定向采集,本系统选择使用 Nutch对个别企业网进行爬行的运行方案。以 Nutch提供的接口为基础,结合 MySql数据库设计,本系统实现了对种子站点和抓取过程进行管理的功能。在数据库中,每一个种子站点的内容类型、连通情况、PageRank值、站点来源、所属学科领域等信息均被管理。用户还可以对每个站点的抓取方案 (例如抓取深度、过滤条件等)进行选择配置。

最初的抓取实验证明,在单服务器环境下,Nutch对少量种子站点进行一次性抓取时成功率比较高。但随着种子站点逐步增加和抓取深度逐步增长,目标网页数量成级数比例增长,抓取周期动辄长达数天。此时,由于系统资源耗尽、网络意外中断、系统意外死机、意外断电、人为重启系统等因素都可能会造成当前爬行任务失败,而重新开启 Nutch程序则意味着要从头开始这一抓取过程。针对这一问题,系统加入了任务管理模块和相应配置管理功能,起到了较好的效果。具体运行机制如下所述。

(1)根据用户事先设定的子任务规模,任务创建模块将自动把要抓取的若干种子站点进行分组,每一组对应一个抓取子任务。

(2)调用 Nutch抓取接口,逐个运行上一步产生的子任务。运行过程中对各子任务运行状态进行记录,这样当意外 /人为退出程序后,再次启动程序时能够继续上次的抓取过程,而不用对已抓取的内容进行重新抓取。

(3)当各个子任务都运行完毕时,调用 Nutch合并接口,将各子任务对应数据进行合并。

基于上述机制,特定领域大规模种子站点的个性化抓取配置和定向采集得以有效实施。

4.3 网页去噪

网页去噪的目的是在尽量保留能体现网页实质性内容的前提下,去除客户端脚本、广告、导航栏、HTML标签等非实质性内容或通用构件。借助 Html Parser开源软件,根据噪音信息的一般特征,对网页进行去噪处理,为后续对网页的信息抽取、分析等操作提供高质量语料。对于一个网页,去噪过程包括以下步骤。

(1)获取网页正文题名、作者 /发布者以及网页发布 /修改日期。网页题名、作者 /发布者和网页发布 /修改日期等信息通常独立成行,内容较短,在去噪过程中容易被当作广告等信息而被误删除。因此在去噪第一步,可以根据题名字号较大、经常出现在 < h1 >标签中的特点抽取出正文题名;根据作者 /发布者和网页发布 /修改日期的字体字号通常较为特别、经常跟随标志词“by”、“last modified”出现等特征,抽取出作者 /发布者和网页发布 /修改日期信息。

(2)使用 Html Parser去除脚本、图片以及其它标签,获得只有链接和文本的字符串。

(3)根据导航栏的一般性特征 (例如“|”符号的数量,词的数量,空格数量等)去除导航栏文字,例如 Local News | Skilled at code, he wins a load | Seattle Times。

(4)去除广告,例如 http://seattletimes.nwsource.com/html/foodwine/ Food & Wine。可以通过两个规则来判断某一行是不是广告:

①一行中的链接数不为 0,且词数小于某个阈值 (实验中取经验值 10);

②一行中的链接数和词数之比大于某个阈值 (实验中取经验值 0.35)。

(5)去除所有以“<”和“>”标识的链接文字,例如: < http://blog.seattletimes.nwsource.com/allyoucaneat/ > All You Can Eat。

(6)去除版权声明信息。通过观察发现,版权声明信息行通常含有“Copyright”、“Statement”、“E-mail”、“Company”等特征语词,且词数少于某个阈值 (实验中取经验值 80),应去除。

经过执行上述步骤,能够基本去除广告、导航信息、客户端代码等对后续抽取分析无用的信息,获得较高质量的网页内容文本。

4.4 获取新的种子站点

种子站点的动态更新机制是通过 Web网络客观反映一个学科领域动态变化情况的一种途径。在系统用户需要扩充特定学科领域对应的种子站点数量时,可以通过新种子站点获取模块来获得新的候选种子站点,经过人工核查后加入到种子站点库。

获取新的种子站点的机制如图 3所示:

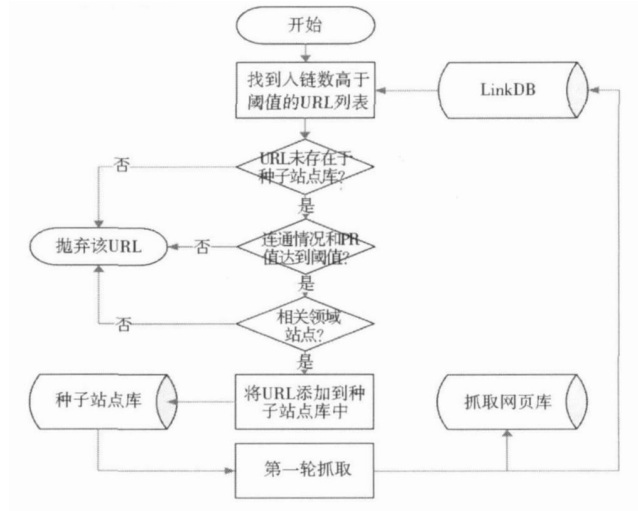

图 3 获取新种子站点流程

在图 3中,LinkDB是 Nutch在对抓取内容进行解析后形成的链接数据库。通过对该链接数据库的访问,能够获得已抓取网页的链接情况,进而通过共链分析等步骤,获得新的候选种子站点。

5　实验效果

在实验中,选取人工智能(Artificial Intelligence)领域的机构类网站作为 Web定向采集目标。实验环境如下:

硬件环境:IBM双核服务器,CPU主频 3.0 GHz,物理内存 4.0 GB;

操作系统:Windows Server 2003企业版;

运行环境:Cygwin 1.5(Linux模拟环境);

Java平台:JDK1.5;

数据库:MySql 6.0;

Nutch版本:Nutch 0.9。

采用 4.1节中的方法,经过半自动筛选后,共获得 376个人工智能领域机构种子站点。以这些种子站点为初始抓取对象,采用 4.2节中子任务管理机制,对这些网站进行了三层抓取,共抓取三个批次。具体抓取情况见表 2:

表 2　抓取系统实验数据表

抓取批次	第一批	第二批	第三批
抓取时间	2008 - 12 - 05	2008 - 12 - 30	2009 - 01 - 04
抓取用时	10小时 41分	15小时 27分	15小时 59分
单批抓取网页记录数	20 005条	39 616条	42 039条
单批有效网页记录数	19 502条	29 647条	20 185条
累计有效网页记录数	19 502条	49 149条	69 334条

三个批次累计获得已去重的有效网页记录 69 334条。运行过程中,人为模拟程序意外中断、网络中断等情况,在重启动程序后,能够实现以子任务为最小单元的断点续传。

抓取下来的 HTML网页通过 4.3节中的方法进行了网页去噪。图 4给出了一个原始网页示例,线框内为所需要的网页正文部分,其余部分为需要去除的噪音信息。

经过去噪处理后,获得了该网页的正文部分,如图 5所示。

去噪实验证明,去噪过程去除了正文主要内容外的绝大部分噪音信息,具有较高的去噪准确率。

在第三批抓取数据基础上,采用 4.4节的方法进行新种子站点获取实验。在对 3 797 988条链接数据

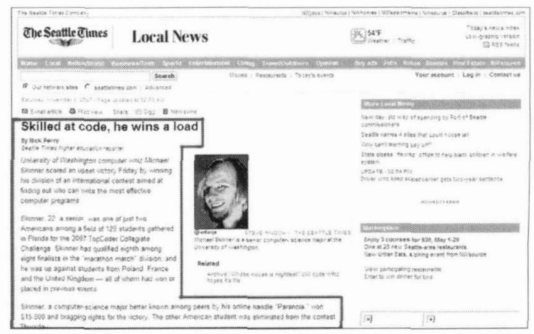

图 4　一个原始网页示例

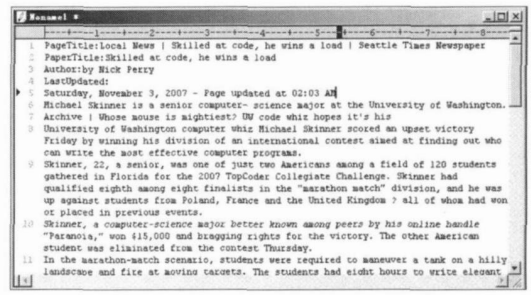

图 5　去噪后的效果

进行共链分析、候选种子站点去重、连通情况和 PageRank值过滤等步骤后,获得较高质量的候选种子站点 1 065个。经人工判断和分类后,共获取人工智能领域新的机构种子站点 43个。被排除的站点多为新闻站点、广告站点、大学站点以及与人工智能领域相关,但不属于机构类型的站点。

上述实验证明,基于 Nutch的 Web网站定向采集系统能够较稳定、快速地获取大量网站页面,对网页进行有效的去噪操作,并能在 Nutch产生的 LinkDB基础上,通过新种子站点获取机制获取更多的新种子站点。该系统功能和性能基本达到了预先设定的目标。

6　结　语

本文在对 Nutch开源软件进行扩展和改进的基础上,提出并实现了 Web网站定向采集系统。针对该系统中 4个核心问题的解决方法,进行了较为深入的探讨。目前,Web网站定向采集系统已被应用于科技领域站点的监测,并取得了较好的效果。尽管如此,随着网站抓取规模的持续增大和网站抓取管理要求的逐步提高,在系统应用过程中,不断出现新的问题。在该系统基础上,下一步将重点解决以下几个问题。

(1)由单机抓取系统扩展为集群抓取系统。Nutch 0.9开源系统软件本身是基于 Hadoop 开发的,可以通过相关配置,利用 Hadoop 进行多节点并行抓取,进一步提高抓取效率。

(2)对 Nutch 原有接口和扩展功能接口进行标准化封装,方便任务调用和管理。例如,对于 Nutch 提供的链接数据库相关功能调用进行开发和标准化,该接口在网站链接分析和新种子站点获取过程中都可能会被调用。

(3)新的种子站点的获取将更加智能化。目前获取新的种子站点仅仅依靠共链分析和人工相结合的方式进行,智能化程度不高。下一步将会融入信息抽取和内容分析的相关研究成果来提高上述过程的智能化程度。通过对候选种子站点的信息抽取和内容分析,能够发现该站点的主题类型,实现站点相关性自动判断和自动分类。

参考文献:

[1] Nutch [EB/OL]. [2009-01-29]. http://wiki.apache.org/nutch/.

[2] Doug Cutting. Nutch, Open-Source Web Search [EB/OL]. [2009-01-29]. http://wiki.apache.org/nutch-data/attachments/Presentations/attachments/www2004.pdf.

[3] Heritrix Introduction [EB/OL]. [2009-01-29]. http://crawler.archive.org/.

[4] The Web Curator Tool Project [EB/OL]. [2009-01-29]. http://webcurator.sourceforge.net/.

[5] Web-Harvest [EB/OL]. [2009-01-29]. http://web-harvest.sourceforge.net/.

[6] Html Parser [EB/OL]. [2009-01-29]. http://htmlparser.sourceforge.net/.

[7] Intute, Best of the Web [EB/OL]. [2009-01-29]. http://www.intute.ac.uk/.

[8] Dmoz Open Directory Project [EB/OL]. [2009-01-29]. http://www.dmoz.org/.

[9] Yahoo! Developer Network [EB/OL]. [2009-01-29]. http://developer.yahoo.com/search/.

[10] Nutch Version 0.8.x Tutorial [EB/OL]. [2009-01-29]. http://lucene.apache.org/nutch/tutorial8.html.

(作者 E-mail: xujian@mail.las.ac.cn)