Web新闻自动采集发布系统的设计与实现

张春元,康耀红,伍小芹
(海南大学 信息科学技术学院,海南 海口 570228)

摘 要:针对新闻网站通过人工方式采集发布来自其它网站的Web新闻费时费力、易重采与漏采这一问题,综合运用Web信息采集技术、网页去噪技术、文本文档消重技术以及文本自动分类技术设计并实现了一种基于网络爬虫的Web新闻自动采集发布系统。在给出系统总体结构的基础上,对其各个模块的功能、设计与实现方法进行了详细介绍。实验表明,该系统设计合理,具有采集效率高、消重准确、集成方便、运行费用低等优点,可作为新闻网站的采编工具加以推广使用。

关键词:网络爬虫;网页去噪;文档消重;Web新闻发布
中图分类号:TP391.3 **文献标识码**:A **文章编号**:1673-629X(2009)09-0250-04

Design and Implementation of Web News Automatically Gathering and Publishing System

ZHANG Chun-yuan, KANG Yao-hong, WU Xiao-qin
(Institute of Information Science and Technology, Hainan University, Haikou 570228, China)

Abstract: News sites manually gather and publish Web news from other sites, which is inefficient and easy to repeatedly collect or miss some news. To solve this problem, using Web information fetching technology, Web pages noises eliminating technology, replicated text documents eliminating technology and automatic text classification technology, a Web news automatically gathering and publishing system is designed and implemented. The whole structure of the system is presented, and then the main function and design method of its each module are introduced. The experiment shows its design is reasonable, and crawling efficiency is high, and eliminating replicated documents is accurate, and integrating into a new site is easy, and operation cost is low, and it can be used as the gathering and editing tool of news sites widely.

Key words: Web crawler; Web pages noises elimination; replicated documents elimination; Web news publishing

0 引 言

目前,不少新闻网站仍然依靠网络编辑从其它网站手工采集各种新闻信息来丰富自己的内容,这种方式发布的Web新闻虽然具有较高的质量,但是费时费力。为了即时转发有价值的新闻,网络编辑们需要花费大量时间频繁浏览各大新闻网站;多名编辑一同采编时,由于缺乏统一的协作平台,还比较容易出现新闻信息的重复采集或者漏采。针对这一问题,笔者参与设计并实现了一种基于Crawler的Web新闻自动采集发布系统。该系统通过网络爬虫对用户指定网站区域下的Web新闻网页增量采集,经解析、消重、分类处理后保存至新闻网站的数据库中,可以7×24小时自动采集发布Web新闻,也可作为新闻网站编辑人员的日常采编工作平台。

收稿日期:2008-12-23;**修回日期**:2009-03-21
基金项目:海南省自然科学基金项目(80638)
作者简介:张春元(1973-),男,湖北武汉人,讲师,硕士,研究方向为信息检索与数据挖掘。

1 系统设计

1.1 系统结构

图1为Web新闻自动采集发布系统结构图,整个系统由虚线框内各模块组成。也可将系统处理后的新闻文档保存到系统外部待集成新闻网站的数据库中,完成Web新闻的自动发布。下面将逐一介绍系统各模块的功能和设计思想。

1.2 增量式Crawler

Crawler即网络爬虫,又称网络蜘蛛(Spider)、网络机器人(Robot)等,主要用来采集各种Web信息资源,本系统所采用的增量式Crawler是一种介于主题网络爬虫和个性化网络爬虫[1,2]之间的轻量级采集系统。我们注意到,大多数新闻网站或者包含新闻信息服务的综合性门户网站设计比较规范,分类也比较清晰,同类别的新闻网页一般放在同一子目录下;另外,

新闻网页大多具有一定的时效性,一经发布以后基本上不会再被更新。因此,我们的 Crawler 只需对系统管理员所设定的采集区域中新近发布的新闻网页进行增量式采集,主题型网页(具体定义见 1.3 节)采集过后将不再作周期性更新采集。例如打算采集网易发布的体育新闻,只需将 Crawler 的采集对象限定为 URL 以"http://sports.163.com"为前缀的网页即可,如表 1 所示。一般来说,每日上午 9 至 12 时是各大网站发布新闻的高峰时间,其余时间则较少发布甚至零发布。针对这一情形,Crawler 将根据被采集区域上一个工作日的新闻发布情况自适应地调整各种子网页当日的采集时间,具体方法为:刚开始 Crawler 从 6:00 至 24:00 对各采集限定区域每隔 1 小时设定一个采集时间点并采集一次,若某采集限定区域在某采集时间点被采集的网页数量为 0,则该区域该时间点失效;若某采集限定区域在某失效时间点的后一个采集时间点被采集的网页数量大于 0,则恢复该区域该失效采集时间点。

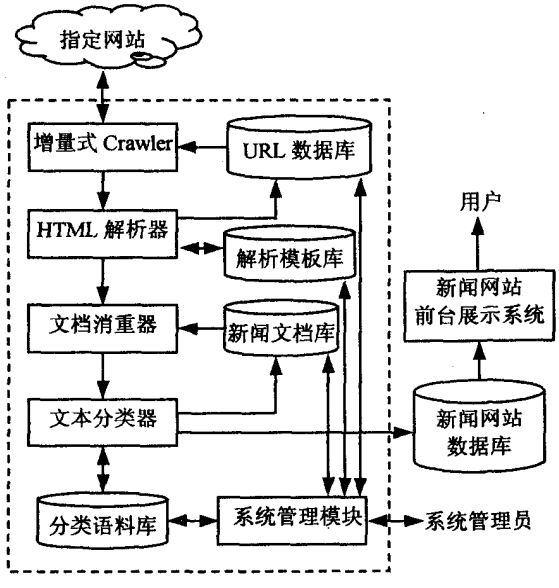

图 1　Web 新闻自动采集发布系统结构图

表 1　Crawler 采集参数设定表

类别	采集限定区域	种子网页	自动采集时间	分类
体育	http://sports.163.com/	http://sports.163.com/	6:00 8:00…	否
体育	http://www.hinews.cn/news/tiyu/	http://www.hinews.cn/news/tiyu/index.shtml	6:00 8:00…	否
热点	http://news.163.com/	http://news.163.com/	6:00 7:00…	是
热点	http://www.hinews.cn/news/system/	http://www.hinews.cn/news/index.shtml	6:00 8:00…	是
⋮	⋮	⋮	⋮	⋮

1.3　HTML 解析器

HTML 解析器实际上是增量式 Crawler 的一个组成部分,为了讨论的方便,将它从 Crawler 中独立出来。HTML 解析器主要用来对 Crawler 采集得到的网页源文件进行解析,从中提取出 URL 与网页正文内容,另外对一些做了分页处理的网页进行内容合成。在解析过程中,将待解析网页分为两种类型进行处理:一类是 Hub 型网页,这类网页主要用来提供网页导航,是超链接聚集的网页,本系统 Crawler 的种子网页就属于此类型;另一类是主题型网页,这类网页大多通过文字或图片描述一件或多件事物来表达一定的主题,本系统采集到的新闻网页绝对多数属于这种类型。Hub 型网页比较容易识别,解析时只需提取出 URL;主题型网页解析除了需要提取出 URL 外,还要提取出正文内容。

HTML 解析器从网页源文件提取 URL 比较简单,提取出来的 URL 经规范化处理后如果在采集限定区域之内且尚未被采集,就将其作为待采集任务加入到 URL 数据库中供 Crawler 采集。

HTML 解析器从主题型网页源文件提取正文内容则比较复杂,主要是由于其中往往夹杂着导航信息、广告信息、评论信息等噪声内容。国内外关于网页噪声内容的去除已提出了许多方法[3~9],在去噪声过程中大多采用了基于块和 DOM 树的分析方法。在文献[7~9]的基础上,通过机器自学习方式构建解析模板库来完成主题型新闻网页正文内容的提取。解析模板库建立的具体思路是:首先,将来自同一网站的网页按 URL 组织成一棵 URL 树,如图 2 所示,底层黑色结点为网页结点,其余为目录结点。需要说明的是,对于诸如"http://www.hainan.com/app/news/view.jsp? Id=I87"这种类型的 URL,其网页结点为"Id=I87",而非"view.jsp? Id=I87"。然后,从 URL 树中选取两张互为兄弟网页结点的网页(以下简称"兄弟网页"),根据网页中的 HTML 标记生成两棵 DOM 树(如图 3 所示)。对主题型兄弟网页而言,它们的 DOM 树剔除内容结点(如图 3 中的黑色结点)后通常具有相同的组织结构。比较兄弟网页的 DOM 树结构,如果相同,对其中一棵 DOM 树深度优先遍历得到一个标签序列(剔除内容结点),记为 structureTagSeq,再采用基于块分析的网页去噪算法[8]和 DSE 算法[9]对两棵 DOM 树作进一步分析,确定正文内容提取的标签序列,记为 contentTagSeq。structureTagSeq 和 contentTagSeq 一起构成了兄弟网页的解析模板,将其保存在 URL 树中相应的目录结点上。如果两棵 DOM 树不相同,则意味着兄弟网页的父目录下存在多个解析模板,需重新选取兄弟网页来生成解析模板。最后,检查 URL 树,如果兄弟目录结点的解析模板一致,就将其解析模板保存到它们的父目录结点上,兄弟目录结点及其网页结点

予以删除。URL 树经修剪后,可能仍存在部分叶子结点为网页结点,将其仍旧保存在 URL 树中,当 Crawler 为其采集到合适的兄弟网页后,再生成新的解析模板。解析模板库随着 HTML 解析器对主题型新闻网页的不断解析而逐步得以建立与完善。在 HTML 解析器中,一张新闻网页的解析过程为:HTML 解析器首先根据其 URL 在 URL 树中查找是否有相应的解析模板存在,如果存在,将该网页转换成 DOM 树,进一步利用 structureTagSeq 查找是否存在标签结构相同的模板,如果存在则利用该模板的 contentTagSeq 完成网页正文内容的提取;如果 HTML 解析器在 URL 树中没有为待解析网页找到合适的解析模板,则将该网页作为网页结点添加到 URL 树中,直到建立起新的模板再进行正文内容的提取。

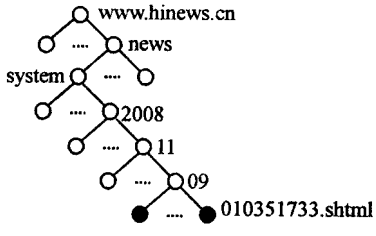

图 2 海南新闻网 UTL 树

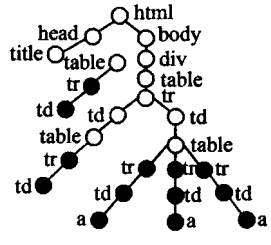

图 3 HTML 文档生成 DOM 树示例

1.4 文档消重器

互联网上一些相对重要的新闻通常会被众多网站转载,故 Crawler 采集到新闻网页中存在不少内容雷同的页面,文档消重器就是对这些网页进行过滤。

当前基于文档内容的消重算法大多是在对文档内容进行分析的基础上,从中抽取出一定数量的特征项,按一定规则组成特征串进行数字签名,然后以此来判断文档内容是否雷同;或者直接利用两篇文档的特征项集合的相似度实现文档查重[3,10~11]。文档消重器具体设计思路是:在待查重文档中的逗号、句号前后各取 2 个汉字或字符作为特征项,按它在文档中出现的顺序连接起来组成特征串,然后与新闻文档库中的文档进行比较,如果某文档的特征串与该特征串连续相同的字符数超过一定比例,就认为该文档与已有文档内容重复,停止后续处理并且不予对外发布。这种方法在构建特征串时具有较高效率,但在查重时需与新闻文档库中的文档逐一比较,因而计算量较大。考虑到新闻具有一定的时效性,转载大多集中在一两天内,超过一周的转载非常少;部分网页在转载过程中虽然作了一些修改,但主要集中在标题和首段上,文档字符数变化并不是太大,因此可对文档消重器做如下效率优化:从新闻文档库中按时间顺序由近及远取一周内对外发布的文档进行比较,且只比较特征串长度差值在 20 以内的文档。

1.5 文本分类器

考虑到大多数新闻网页已经做了良好的分类,为了提高系统的运行效率,只对一些需要重新分类的网页进行分类,由系统管理员在 Crawler 采集参数设定表中进行设定。

近年来,国内外针对文本自动分类技术的研究一直非常活跃,这方面成果也比较丰富[3]。在文本分类器中,选用搜狐研发中心提供的文本分类语料库[12]作为训练样本集,采用正向最大匹配法进行中文分词,用 CHI 算法[13]提取类别特征向量。对一个待分类的新闻文档,首先对其作分词处理,然后转换成文档特征向量,再采用传统向量夹角余弦公式计算其与各类别特征向量的相似度,以此确定所属类别,最后将其保存至新闻文档库和系统外部待集成的新闻网站数据库中。

1.6 系统管理模块

系统管理模块是整个系统的控制中心,系统管理员在此设置 Crawler 的采集及运行参数、查看解析模板库、管理文本分类语料库、对文本分类器进行训练、查看或修改新闻文档库中的文档。

2 实验结果

整个系统采用纯 Java 实现,分类语料库、新闻文档库选用 MySQL5.1 数据库保存,为了提高系统性能,URL 数据库、解析模板库选用 Berkeley 数据库(Java 版)实现。Crawler 通过线程池来管理活动线程,为了不给被采集网站带来太大的负担,将属于同一采集限定区域的采集任务放在同一个队列中,当此队列中有一个 URL 被取出之后,该队列就会进入阻塞状态,直至被取出 URL 对应的网页被处理完后才从阻塞状态中恢复。

采用 1 台联想 ThinkCentre PC 机(CPU 为奔腾双核 1.6GHz,内存为 2GB,硬盘 7200r/min,Windows XP 操作系统,带宽 1Mb/s)对系统性能进行测试,得到表 2 和表 3 所示测试数据。在多个采集限定区域情形下,系统的平均下载速度达到了 96.39kB/s,基本接近最大带宽;在单一采集限定区域采集情形下,系统的下载速度此时与带宽大小基本没有关系,只与被采集网

站服务器的响应速度和网页平均大小有关。对近 50 家新闻网站各新闻子栏目每日发布新闻数量进行统计，发现大多数在 50 条以下，即便是网易这样的大型网站，其国内新闻频道每日发布新闻的数量平时也不超过 200 条，每小时发布新闻数量一般多在 20 条以下。由此不难推知，新闻采集发布系统在现有条件下，具备对 500 个限定采集区域进行日常采集的能力。从表 3 数据来看，系统的消重性能比较理想，分类性能则有待进一步提高。另外，从系统自动采集发布的 Web 新闻结果中随机抽取了 300 张网页人工检查，发现系统完全正确抽取出正文内容的网页达到了 97%，夹杂有噪声的网页占 3%，没有发现不包含正文内容网页，表明系统能较好地完成新闻网页的正文提取。

表 2　新闻采集发布系统采集性能测试数据

测试内容	活动线程数量	网页平均下载数量	平均下载速度	下载网页平均大小
多个限定区域采集能力测试	50	10573 Pages/h	96.39 kB/s	32.82 kB/Page
单一限定区域采集能力测试	1	1297 Pages/h	8.37 kB/s	23.24 kB/Page

表 3　新闻采集发布系统消重及分类性能测试数据

测试内容	平均查全率	平均查准率
文档消重器消重性能	97.5%	98.5%
文本分类器分类性能	75.1%	72.3%

3　结束语

设计并实现了一种 Web 新闻自动采集发布系统，该系统能自动对用户指定网站区域的 Web 新闻进行采集、去噪、消重、分类与发布，可以非常方便地与现有新闻类网站系统集成。从实际运行效果来看，该系统总的来说具有较高的性能，可大幅提高新闻网站的 Web 新闻采集与发布效率，降低网站的运营成本。但是系统在采集种子与采集限定区域的设置上对用户要求较高、对于正文内容下方夹杂有用户评论的网页去噪还不完善、文本分类模块的查全率与查准率也有待进一步提高，这些将是下一步工作的重点。

参考文献：

[1] 李盛韬. 基于主题的 Web 信息采集技术研究[D]. 北京：中国科学院，2002.

[2] 刘金红，陆余良. 主题网络爬虫研究综述[J]. 计算机应用研究，2007，24(10)：26－29.

[3] 李晓明，闫宏飞，王继民. 搜索引擎——原理、技术与系统[M]. 北京：科学出版社，2005.

[4] Gupta S，Kaiser G，Neistadt D，et al. DOM－Based Content Extraction of HTML Documents[C]// Proceeding of the 12th International Conference on World Wide Web. New York：ACM Press，2003：207－214.

[5] CAI Deng，YU Shi－peng，Wen Ji－rong，et al. Extracting Content Structure for Web Pages based on Visual Representation[C]// Proceeding of the 5th Asia Pacific Web Conference. Berlin：Springer－Verlag，2003：406－417.

[6] Zheng Shuyi，Song Ruihua，Wen Ji－Rong. Template－Independent News Extraction Based on Visual Consistency[C]// The 22nd Conference on Artificial Intelligence. Vancouver：AAAI Press，2007：1507－1511.

[7] 欧健文，董守斌，蔡　斌. 模板化网页信息的提取方法[J]. 清华大学学报：自然科学版，2005，4(S1)：1743－1747.

[8] 刘晨曦，吴扬扬. 一种基于块分析的网页去噪音方法[J]. 广西师范大学：自然科学版，2007，25(2)：149－152.

[9] WANG Jiying，Lochovsky F H. Data－rich Section Extraction from HTML Pages[C]// Proceedings of 3rd International Conference on Web Information Systems Engineering. Singapore：IEEE Computer Society，2002：1－10.

[10] 白广慧. 网页排重技术研究与应用[D]. 北京：中国科学院，2006.

[11] 吴平博，陈群秀. 基于特征串的大规模中文网页快速去重算法研究[J]. 中文信息学报，2003，17(2)：28－35.

[12] 搜狐研发中心. 搜狗文本分类语料库[EB/OL]. 2008. http://www.sogou.com/labs/dl/c.html.

[13] 王倩倩，段　震，张燕平. 基于交叉覆盖算法的文本分类[J]. 计算机技术与发展，2007，17(6)：113－115.

关于推荐 2009 年 CCF 优秀博士学位论文的通知

为推动中国计算机领域的科技进步，鼓励创新性研究，激励计算机领域的博士研究生潜心钻研，务实创新，解决计算机领域中需要解决的理论和实际问题，使做出优秀成果的年轻学者获得同行认可并有成就感，中国计算机学会(CCF)自 2006 年起设优秀博士学位论文奖。2009 年度优秀博士学位论文推荐工作已经启动。

具体参评条件和约束条件见“CCF 优秀博士学位论文奖条例”(http://www.ccf.org.cn 之关于 CCF/规则与条例)。CCF 优秀博士学位论文推荐表(必须有作者答辩时所在单位(如系、院、所等)负责人签字、单位盖章，可在 http://www.ccf.org.cn/web/resource/tuijianbiao.doc 下载)。

联系人：韩玉琦　电话：010－6260 1340　　朱征瑜　电话：010－62562503－16

邮寄地址：北京 2704 信箱 中国计算机学会　　邮编：100190