第 36 卷 第 9 期 计 算 机 工 程 2010 年 5 月
Vol.36 No.9 Computer Engineering May 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)09—0285—03 文献标识码：A 中图分类号：TP311.52

Web 论坛数据源增量爬虫的研究

蔡欣宝，郭若飞，赵朋朋，崔志明
(苏州大学智能信息处理及应用研究所，苏州 215006)

摘 要：针对 Web 论坛站点结构复杂、内容更新快等特点，提出一种针对论坛的增量信息采集算法，使用站点地图重建技术及网页更新频繁度估计方法，根据站点地图选择有效的链接，按照网页更新频度确定网页的采集频度。实验结果表明，该方法是有效的。
关键词：Web 论坛；增量爬虫；站点地图；泊松模型

Research on Web Forum Data Source Incremental Crawler

CAI Xin-bao, GUO Ruo-fei, ZHAO Peng-peng, CUI Zhi-ming
(Institute of Intelligent Information Processing and Application, Soochow University, Suzhou 215006)

【Abstract】According to the characters of Web forum site such as the complex structure and quickly updating contents, an algorithm of forum incremental information sampling is presented. The technologies of site map rebuilding and estimating the frequency of page update are used. According to the site map, the crawler selects effective links. According to the frequency of Web page update, the crawler determines the crawling frequency of the Web page. Experimental results indicate this method is effective.
【Key words】Web forum; incremental crawler; site map; Poisson model

1 概述

随着 Web 2.0 的发展，论坛作为 UCC(User-Created Content)的典型代表，已成为网络上重要的资源。对论坛信息的采集是搜索引擎、数据挖掘以及商业智能等其他应用的基本步骤[1]。因此，针对论坛数据源的信息采集也具有越来越重要的意义。总的来说，论坛的内容是存储在数据库里的，当服务器接受到用户请求时，它根据模板动态生成应答页[2]。采集 Web 论坛数据源的信息不同于采集普通网站，它存在 2 个方面的难点：(1)论坛站点结构非常复杂；(2)论坛信息更新非常快。为了使用户访问方便，论坛中往往完全相同的网页会有多个不同的 URL 地址，同一主题的帖子会被分成多个不同页来显示，且很多链接是交叉相连的，这使得论坛站点的结构非常复杂，也给传统的采集方法带来了前所未有的困扰。同时，论坛由于用户参与非常频繁，几秒钟内就有大量新的内容出现，并且网页之间更新频繁度的差别很大。传统的周期性爬虫通常是定期采集站点的网页直到一定的数量来更新已收集的网页集，对所有的网页作相同的处理。因此，周期性爬虫只能通过缩短采集的周期来提高已收集网页的新鲜度。但是这种方法可行性很小，一方面由于系统的带宽有限；另一方面被访问的站点不允许爬虫过于频繁地采集，爬虫过于频繁地访问某个站点，可能导致该站点的管理员拒绝该爬虫的访问。对于网页更新速度快和网页之间更新频繁度差别很大的数据源，采用增量的采集方法是尤为必要的[3]。

本文针对论坛结构复杂以及内容变化快的特点，提出一种针对论坛数据源的增量爬虫算法。该算法通过对样本网页集的离线分析从而得到论坛的站点结构地图，再根据站点结构地图提供的信息选择有效链接，从而避免下载大量重复页面和无效页面。同时，该算法通过对有效页面更新频繁度的估计，合理分配采集资源下载有更新的网页，从而避免对没有更新的页面的频繁采集。

2 相关研究

增量爬虫是根据网页变化的频繁度不停地采集网页，以提高已收集网页的新鲜度和及时下载新的网页，用更重要的网页取代那些次重要的网页。

增量爬虫的目标主要是保持本地文档的时新性和增加本地数据库的网页质量。一类方法是采用非线性设计来增加爬虫的更新效果。另一类方法是根据网页更新频繁度来确定对网页采集的频度[3-4]。文献[5]提出根据一组网页最后更改日期来估计网页更新频繁度的方法。还有一类方法是利用 RSS 协议来帮助用户获得新的信息。

普通爬虫通常采用宽度优先策略，会下载大量的重复页面和无效页面同时也不能很好地平衡成本和效率，这在爬取论坛时是无效的。Deep Web 数据和论坛数据类似都存储在数据库中，根据用户要求把信息动态呈现给用户。但是 Deep Web 爬虫自动采集隐藏在表单背后的网页，而论坛爬虫需要跟踪下载有效链接。基于 DOM 树的爬虫也不能有效解决论坛中无效页面和重复页面的问题。

站点地图协议是规定网站提供一个关于网站结构信息的 xml 文档。但是网站并不能保证站点地图文件实时更新以及该文档不能支持超过 50 000 个 URL 链接，因此，对论坛采集没有实际价值。文献[1]提出一种对基于学习的论坛采集方法，通过离线分析论坛的结构特点，发现论坛的有效链接特点和采集网页的顺序。但该方法没有考虑增量采集论坛的方

基金项目：国家自然科学基金资助项目(60673092)；2008 年江苏省重大科技支撑与自主创新基金资助项目(BE2008044)
作者简介：蔡欣宝(1986－)，男，硕士研究生，主研方向：数据挖掘；郭若飞，硕士研究生；赵朋朋，博士研究生；崔志明，教授、博士生导师
收稿日期：2009-12-03 **E-mail：**caixinbao1@163.com

法，同时该文提出的网页采集顺序不适应论坛的增量采集。

3 论坛数据源增量爬虫系统描述

3.1 系统框架

Web 论坛数据源增量爬虫系统框架如图 1 所示。

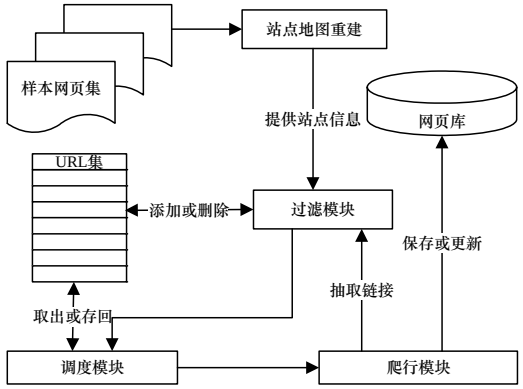

图 1 Web 论坛数据源增量爬虫系统框架

样本网页集是通过采用宽度优先和深度优先算法相结合的策略采集一定量论坛的网页。“站点地图重建”模块通过对样本网页集的聚类分析得到该论坛有效页面和无效页面的特点，以便为过滤模块提供站点信息，来判断哪些链接是有价值的。“调度模块”采用泊松模型对 URL 集里的每个 URL 所代表的网页进行更新频繁度的估计，然后根据网页更新频繁度来确定网页的采集频繁度，从而进行合理的调度。“爬行模块”根据“调度模块”的请求，采集一个网页并更新网页库，同时抽取该网页的所有链接交给“过滤模块”。“过滤模块”根据“站点地图重建”模块得到的站点结构信息，判断每个链接的价值确定是否应该加入到 URL 集的队列当中。同时如果 URL 集已满，需要选择删除一些次重要的链接。

3.2 站点地图重建模块

站点地图重建模块目的是发现有效页面和无效页面的特点，主要通过 3 个步骤完成的：首先根据网页 DOM 树结构特点对网页进行聚类；其次通过比较同类网页的内容和 URL 特点，对每个集合再次聚类；最后根据有效页面的特点发现有效网页集合。以下是具体实现过程：

(1)根据 DOM 树网页的结构特征，通过计算样本网页集中网页之间的距离把网页分成不同集合 S_i。对于网页 s_a 与网页 s_b，可根据式(1)计算它们之间的距离。如果 2 个网页的距离小于预先定义好的阈值，那么这 2 个网页被分配到同一个集合。如此循环，样本网页集的所有网页被分配到不同的集合中。

$$dist(s_a,s_b)=\sqrt{\sum_{i=1}^{\|P\|}(n_a^i-n_b^i)^2} \tag{1}$$

其中，P 表示重复模式(论坛中每个网页都有很多不同的区域块，每个区域块都是用不同的模板生成的，而每个模板就对应一个重复模式，关于模板的检测方法具体见文献[6])集合，$\|P\|$表示重复模式的个数；n_a^i 表示模式 p^i ($p^i \in P$)在网页 s_a 中出现的次数。

(2)对每一个网页集合 S_i，根据网页内容相似度以及链接结构特点把 S_i 分成很多小的集合 $S_{i,1}, S_{i,2}, \cdots$。对于网页 $s \in S_i$，如果在 $S_{i,1}$ 中没有与 s 内容相似的网页 s'，那么把 s 放入 $S_{i,1}$ 中。否则，s 在 $S_{i,1}$ 有相似网页 s'，那么从 $S_{i,2}$ 开始寻找与 s 有类似 URL 结构的集合 $S_{i,j}$。如果找到就把 s 是放入到其中。否则，从 $S_{i,2}$ 开始寻找与 s' 有类似 URL 结构的集合 $S_{i,j}$。如果找到就把 s' 从 $S_{i,1}$ 中删除并放入到 $S_{i,j}$，把 s 放入到 $S_{i,1}$ 当中。如果都没找到就建一个新的集合 $S_{i,k}$，并把 s 放入 $S_{i,k}$ 中。如此重复直到所有网页被分到 $S_{i,j}$ 中。同时计算得到在 S_i 中重复页面的个数 N_i^{dup}。

(3)有效页面具有以下的一般特性：有效页面通常占的比例比较高，每个文件的大小比较大以及有效页面的重复率比较低。根据网页集的信息度，判断每个网页集是否是有效网页集。根据式(2)计算每个网页集合 S_i 的信息度。如果 $\inf(S_i)$ 大于预先定义好的阈值那么认为 $S_{i,1}$ 为有效页面集，否则认为 $S_{i,1}$ 为无效页面集，同时把其余的页面集被标识为无效页面集，这样就得到了站点地图。

$$\inf(S_i)=\frac{N_i}{N}\times\frac{S_i^{avg}}{S^{avg}}\times(1-\frac{N_i^{dup}}{N_i}) \tag{2}$$

其中，N 是样本网页集中网页的个数；N_i 是 S_i 中网页的个数；S_i^{avg} 是 S_i 中网页文件的平均大小；S^{avg} 是样本网页集中网页文件的平均大小；N_i^{dup} 是 S_i 中重复页面的个数。

依据站点地图，过滤模块可以根据每个链接的结构特点判断该链接属于哪类网页集，如果该链接是属于有效网页集，那么该链接被加入到 URL 集当中。如果该链接是属于无效网页集，那么该链接被舍弃。如果不属于任何网页集，那么说明站点的结构发生变化，对站点地图重建模块发出警告以便在合适的时候重新进行站点地图重建。

3.3 增量调度设计

调度模块通过估计网页更新频繁度来确定网页采集的频度，选择该采集的网页链接来调度爬行模块。泊松过程(Poisson process)经常被用来描述独立的，随机的具有固定变化频率的事件序列，这些事件以固定频率重复独立发生。例如一个城市发生车祸的情况、到达商店的顾客等，都可用泊松过程来描述。根据泊松模型，事件的发生情况服从分布：

$$f(t)=\begin{cases}\lambda e^{-\lambda t} & t>0\\ 0 & t\le 0\end{cases} \tag{3}$$

其中，t 是事件发生的时间；λ 为该事件发生的固有频率。本文采用泊松模型为网页更新事件建模，估计网页更新频繁度，然后根据网页的更新频度来确定爬虫采集该网页的频率。这并不是说那些更新越快的网页就应该得到更频繁的采集。

网页的最佳采集频度与网页更新频度的关系如图 2 所示。其中，横轴代表网页的更新频度，纵轴代表网页的最佳采集频度。所以对于网页的更新频度小于 λ_h 的网页，更新越快的网页应获得越多的采集次数。相反对于网页的更新频度大于 λ_h 的网页，更新越快的网页应获得越少的采集次数。调度模块根据网页的采集频度以及爬虫带宽等因素选择该采集的网页，把选择好该采集的 URL 传给“爬行模块”。“爬行模块”负责采集“调度模块”传递过来的链接。

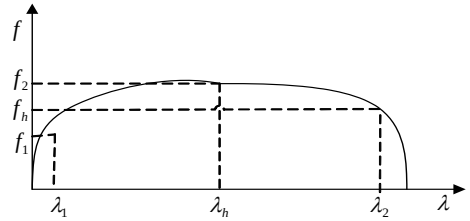

图 2 最佳采集频度与网页固有更新频度关系

3.4 论坛增量爬虫的核心算法设计

论坛增量爬虫的核心算法如下:

输入 Urls——有效的 url 数据集；SampleWebs——论坛的样本网页集

输出 Collection——收集论坛的网页库

```
[1] while (true)
[2] url ← selectToCrawl(Urls)
[3] page ← crawl(url)
[4] newurls ← extractUrls(page)
[5] Collection ← page
[6] if(isexistwebmap) then
[7] Urls ← filter(newurls)
[8] else setupwebmap(sampleWebs)
[9] goto 7
[10] endif
[11] endwhile
```

其中，selectToCrawl()是根据 URL 集中网页的最佳采集频度合理分配采集资源来选择当前该采集的 URL；filter()是根据站点地图信息来判断 URL 是否为有效的 URL，并把有效的 URL 添加到 URL 集中，如果没有站点地图信息，则调度站点地图重建模块 setupwebmap()来创建站点地图。

4 实验设计及结果分析

4.1 实验设计

本系统选取 5 个论坛站点(见表 1)作为实验的数据源。这些站点既有使用了论坛模板的，也有网站自行设计的。其中，站点 bbs.kaoyan.com 和 bbs.hhghost.com 分别使用了非常流行的论坛模板 Discuz!和 PHPWind。同时这些论坛包括教育、生活等各个领域。

表 1 论坛增量采集算法与广度优先采集算法准确率比较

站点名称	论坛类型	论坛增量采集算法			广度优先采集算法		
		无效的网页数	重复的网页数	准确率/(%)	无效的网页数	重复的网页数	准确率/(%)
bbs.kaoyan.com	Discuz!	534	46	88.4	2 142	643	44.3
bbs.hhghost.com	PHPWind	98	18	97.7	1 657	742	52.0
bbs.chinajavaworld.com	自设计	226	27	94.9	1 753	576	53.4
club.163.com	自设计	67	11	98.4	1 132	269	72.0
club.china.com	自设计	72	12	98.3	1 389	371	64.8

在论坛的采集过程当中，爬虫的任务是采集更多有效的页面，主要体现在准确率和覆盖率上:

准确率=采集到的有效页面数/采集到的页面总数

覆盖率=采集到的主题数/论坛主题数

其中，采集到的有效页面数=采集到的页面总数-无效页面数-重复的页面数。论坛的信息载体是以主题形式体现出来的。同时，论坛网页是由数据库动态生成，在数量上没有确切的限制，因此，覆盖率使用的参数是主题数而不是网页数。

在比较论坛增量爬虫和周期性爬虫的更新效果上，由于很难实现在瞬间比较本地数据库和网站的远程数据库，因此实际测量本地数据的新鲜度是很难实现的，本文采用文献[7]中的新鲜度估计方法进行估计。

4.2 实验结果分析

对本文选取的 5 个站点，分别使用本文提出的论坛增量采集算法和广度优先采集算法进行采集，以采集论坛的 5 000 个网页为停止条件。实验结果如表 1 所示。

对已知论坛主题数的论坛(bbs.kaoyan.com, bbs.hhghost.com, bbs.chinajavaworld.com)，使用论坛采集算法进行采集，实验结果如表 2 所示。

表 2 论坛增量采集算法的覆盖率

站点名称	论坛的主题数	采集到的主题数	覆盖率/(%)
bbs.kaoyan.com	1 493 446	1 309 012	87.7
bbs.hhghost.com	84 772	82 493	97.4
bbs.chinajavaworld.com	183 165	175 457	95.8

分别采用增量爬虫算法和周期性爬虫算法对 5 个已选定的论坛进行采集，并分别估计它们的新鲜度，结果见表 3。

表 3 增量采集算法与周期性采集算法新鲜度比较

站点名称	增量采集算法新鲜度	周期性采集算法新鲜度
bbs.kaoyan.com	0.76	0.38
bbs.hhghost.com	0.65	0.44
bbs.chinajavaworld.com	0.63	0.48
club.163.com	0.61	0.52
club.china.com	0.68	0.43

从表 1~表 3 可以看出，本文的爬虫算法在获取有效页面和提高本地数据库时新性方面都取得了很好的效果。

5 结束语

本文针对论坛数据源提出一种论坛的增量采集算法。在获取有效链接时，考虑论坛的站点地图，在分配页面采集频繁度时，考虑网页的更新频繁度，此方法在实验中取得较好的效果。下一步将要考虑进一步提高爬虫的性能，适应大规模的采集需求。

参考文献

[1] Cai Rui, Yang Jiangming, Lai Wei, et al. iRobot: An Intelligent Crawler for Web Forums[C]//Proc. of the 17th International World Wide Web Conference. Beijing, China: [s. n.], 2008.

[2] 李 魁, 程学旗, 郭 岩, 等. WWW 论坛中的动态网页采集[J]. 计算机工程, 2007, 33(6): 80-82.

[3] Cho J, Garcia M H. The Evolution of the Web and Implications for an Incremental Crawler[C]//Proc. of the 26th Int'l Conf. on Very Large Data Bases. Cairo, Egypt: [s. n.], 2000.

[4] Cho J, Garcia M H. Estimating Frequency of Change[J]. ACM Trans. on Internet Technology, 2003, 3(3): 256-290.

[5] Brewington B, Cybenko G. Keeping up with the Changing Web[J]. IEEE Computer, 2000, 33(5): 52-58.

[6] Zheng Shuyi. Joint Optimization of Wrapper Generation and Template Detection[C]//Proc. of the 13th ACM Int'l Conf. on Knowledge Discovery and Data Mining. San Jose, CA, USA: [s. n.], 2007.

[7] Cho J, Garcia M H. Synchronizing a Database to Improve Freshness[C]//Proc. of 2000 ACM SIGMOD International Conference on Management of Data. Dallas, Texas, USA: [s. n.], 2000.

编辑 陈 文