第 36 卷 第 8 期 计 算 机 工 程 2010 年 4 月
Vol.36 No.8 Computer Engineering April 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)08—0275—03 文献标识码：A 中图分类号：N945

面向垂直搜索引擎的 Web 站点划分方案

李学凯，许 笑，孙春奇，张伟哲，李 斌
(哈尔滨工业大学计算机学院，哈尔滨 150001)

摘 要：分析传统搜索引擎分配任务的方式及存在的问题，根据垂直搜索引擎的特点，提出一种比传统方法粒度更细的任务分配方式——网站划分。该分配方式将较大规模的网站切分为若干较小规模的子集，并将子集交给若干爬虫节点并行抓取，以加快爬虫系统的整体获取速率，作为对传统方法的有效优化。将网站划分算法应用于样本数据集，验证其有效性。

关键词：垂直搜索引擎；任务分配；网站划分；爬虫

Web Site Partition Scheme for Vertical Search Engine

LI Xue-kai, XU Xiao, SUN Chun-qi, ZHANG Wei-zhe, LI Bin
(College of Computer, Harbin Institute of Technology, Harbin 150001)

【Abstract】 In allusion to the problem of traditional search engines' task allocating methods, a new fine-grained method called Web site partition is presented, which is as an effective optimization of the traditional method adopted by vertical search engines. This method divides large-scale Web sites into a number of smaller subsets, so that several crawlers can parallel crawl each subset in order to accelerate the overall downloading progress. The proposed algorithm is proved to be effective against the sample data sets.

【Key words】 vertical search engine; task allocation; Web site partition; crawler

为了协调多机爬虫并行工作，提高抓取效率并均衡爬虫之间的负载，搜索引擎需要好的任务分配策略。通用搜索引擎对任务的分配主要有 2 种方式[1-2]：(1)基于 URL 的 Hash 算法；(2)基于网站的 Hash 算法。

1 垂直搜索的特点[3]

不同于通用搜索引擎，垂直搜索引擎的信息来源只是一小部分网站，如新闻搜索的信息来源主要是新闻网站，而视频搜索的信息来源主要是视频网站。同时，垂直搜索对信息的实效性要求也较高，新闻搜索要求几分钟之内做出更新，火车票转让信息更是如此。

通用搜索引擎中采用的基于网站的 Hash 算法并不能很好地适合于垂直搜索引擎。在基于网站 Hash 的调度算法中，任务的划分粒度是网站，即同一个网站的内容只能归某一个爬虫节点抓取。通用搜索引擎的爬虫系统，由于其处理的 Web 站点数量庞大，因此能够从全局上弱化站点规模差异造成的影响；但是对于像垂直搜索引擎这样面向较小 Web 站点集合的爬虫系统，由于其获取的 Web 站点数量有限，因此站点规模差异造成的负载不均就会突出显现出来。另一方面，垂直搜索引擎为了追求信息的实效性，通常的做法是加大从 Web 站点获取信息的力度，而 Web 站点为了避免自身遭受攻击，又会屏蔽这种频繁抓取的行为，需要设计一种新的任务分配方式来解决这个矛盾。本文提出网站划分的概念，在容忍一定的网页缺失以及容忍一定的网页重复抓取的前提下，将较大规模的网站切分为若干较小规模的子集，并将子集交给若干爬虫节点并行抓取，以加快爬虫系统的整体获取速率，作为对传统方法的有效优化。为了便于研究，本文只关心网页的站内链接(指向本网站的链接)。因为处理站外链接的方法无非是爬虫与调度中心通信，由调度中心来调度，跟本文研究没有关系，只会增加表述的复杂性。如果没有特别说明，下文所指的网页链接均指站内链接。

2 网站划分

2.1 网站划分定义

定义 1 假设预期将某特定网站划分为 N 份，由 N 个爬虫协同完成网站的抓取任务。设爬虫集合为 $C=\{c_1,c_2,\cdots,c_N\}$，网站上所有网页的集合为 W。对于 W 的子集的集合 $\beta=\{\beta_1,\beta_2,\cdots,\beta_N\}$，如果同时满足：

$$\frac{|W-\bigcup_{i=1}^{N}\beta_i|}{|W|}\le\varepsilon \tag{1}$$

$$\frac{|\sum_{i=1}^{N}\beta_i|-|\bigcup_{i=1}^{N}\beta_i|}{|\bigcup_{i=1}^{N}\beta_i|}\le\delta \tag{2}$$

其中，ε 和 δ 是一个较小的小数，称一一映射 $\mu:C\to\beta$, $c_i\to\beta_j$, $i=1,2,\cdots,N$, $j=1,2,\cdots,N$ 为网站划分。

网站的结构是一个有向图，网页是图的顶点，链接是图的边，称该有向图为网站的链接关系图。网站划分的目的是把网站链接关系图划分成若干份，使得每一份内部联系比较紧密，外部联系比较疏松。网页的链接一般是指向与自己内

基金项目：国家自然科学基金资助项目(60703014)；国家“973”计划基金资助项目(G2005CB321806)；高等学校博士学科点专项科研基金资助项目(20070213044)；哈尔滨工业大学优秀青年教师培养计划基金资助项目(HITQNJS.2007.034)

作者简介：李学凯(1984－)，男，硕士研究生，主研方向：搜索引擎；许 笑，博士研究生；孙春奇，副教授；张伟哲，副教授、博士；李 斌，教授、硕士

收稿日期：2009-12-01 **E-mail：**meteorlxk@gmail.com

容相关的网页，文献[4]发现用超链接连接起来的 2 个网页比随机挑选的 2 个网页具有更大的相似性。而网站的内容基本都是根据路径来归类的，相同路径下的网页相关性比较大，不同路径之间的网页联系比较疏松。

选定 211 个网站为样本数据集，用爬虫分别抓取样本数据集中每一个网站，抓取策略为广度优先，深度限制为 10 层。根据抓取的结果可以得到样本数据集中每一个网站的链接关系图。严格意义上说，网站的网页数量是未知的，无法确定爬虫是否抓取完了所有的网页，只能假设链接关系图中包含了网站的所有网页。下面将根据这些链接关系图展开。

网页包含的链接数量称为该网页的出度，网页包含的与该网页自身处于相同路径的链接数量称为相同度。对样本数据集的网站链接关系图进行分析，统计每一个网页的相同度与出度的比值，然后求所有比值的平均值，发现平均 74.36% 的网页链接指向与自身处于相同路径的网页。这充分表明，可以把路径作为网站划分的依据。

2.2 URL 路径的确定

URL 的格式为：“网络协议://主机名/一级路径/二级路径/…”，规定“主机名/一级路径”为 URL 的路径，如果 URL 没有一级路径，那么 URL 的路径为“主机名/”。

例如，“http://today.hit.edu.cn/articles/1/2.html”的路径为 today.hit.edu.cn/articles；而“http://www.hit.edu.cn/index.html”的路径为 www.hit.edu.cn/；同样，“http://www.hit.edu.cn”的路径也是 www.hit.edu.cn/。需要指出的是，URL 与路径是多对一的关系，每个 URL 只对映一个路径。

2.3 路径分配算法

确定了 URL 的路径，下一步分别将网站的 URL 按照路径分类，结果表明，大部分 URL 集中在少数路径中。根据样本数据集的统计结果，平均 19.15%的路径包含全网站 90%的 URL。把网站路径按照包含的 URL 数量从大到小排序，从前往后取出包含 90%URL 数量的路径，这些路径称为该网站的关键路径。

如果任务下发时包含网站的所有路径会造成任务描述太繁琐，那么在进行网站划分时，只需要针对关键路径进行任务分配。关键路径包含了整个网站 90%的 URL，因此，即使造成重复抓取，比例也会很小。

网站划分实际上把关键路径分成若干份，由多个爬虫并行完成网站抓取任务。为了确保每个爬虫负责抓取的 URL 数量大致相等，采取平均分配算法进行关键路径的分配。平均分配算法的伪代码如下：

```
struct PathItem {
string path;//路径
int num;//包含 URL 数量
};
List<PathItem> majorPaths;
int N;
int urlNum[N];
for (PathItem path : majorPaths) {
找出 urlNum 数组中，值最小的下标 i;
将 path 分配给第 i 个爬虫;
urlNum[i] += path.num;
}
//end for
```

按照平均分配算法分配关键路径，可以使得爬虫之间分配的 URL 数量尽量相等。

2.4 种子 URL 的选取

单个爬虫抓取网站，种子 URL 只有一个，即网站的首页。在网站划分时，只用首页当作所有爬虫的种子并不合适，因为无法保证网站首页有指向每个关键路径的链接，这样会造成大量的网页被漏抓。解决方法是为每个关键路径选取 2 个有代表的 URL 当作补充的种子。当关键路径被分配给某个爬虫时，连带着该关键路径的补充种子也分配给该爬虫。所以，爬虫的种子有 2 种：(1)网站首页；(2)关键路径的补充种子。

选取有代表的 URL 的原则有 2 个：(1)入度最大；(2)层数最小，优先考虑第一个原则。网页的入度大，表明被其他网页引用的次数多；网页的层数小，易于用户浏览到[5]。研究表明，这些原则是体现网页重要度的特征，优先抓取重要的网页可以提高网页抓取的质量。

3 实验

3.1 网站划分评价指标

定义评价网站划分效果的指标如下：

(1)重复率

下载网页的重复率(*Overlap*)为

$$Overlap=\frac{N-I}{N} \tag{3}$$

其中，N 代表所有爬虫所下载的总网页数；I 代表所有爬虫所下载的网页中不重复的网页数。抓取重复的网页，相应地会增加系统不必要的开销，因此，重复率越小越好。

(2)覆盖率

下载网页的覆盖率(*Coverage*)为

$$Coverage=\frac{I}{U} \tag{4}$$

其中，U 代表所有爬虫应该下载的总网页数；I 代表所有爬虫所下载的网页中不重复的网页数。前文提到严格意义上来说，网站上的网页 U 本身是未知的，所以，很难得到绝对的覆盖率，只能寻求相对覆盖率以作比较之用。在做系统性能评价时，把链接关系图包含的所有顶点当作 U。

(3)加速比

下载网页的加速比(*Acceleration*)为

$$Acceleration=\frac{U}{M} \tag{5}$$

其中，U 代表所有爬虫应该下载的总网页数，它也是单机爬虫应该抓取的总网页数；M 代表所有爬虫中抓取网页最多的那个爬虫所下载的总网页数。

(4)综合指标

下面定义综合指标来衡量网站划分的效果，假设重复率为 O，覆盖率为 C，加速度为 A，可以容忍的最小覆盖率为 ε 以及重复率常数 δ。定义综合指标为

$$\max\left(\frac{A^2\times(C-\varepsilon)}{\sqrt{\max(0,\delta)}},0\right) \tag{6}$$

如果覆盖率 $C\le\varepsilon$，那么综合指标将等于 0，表示网站划分方案不合格，并不是说所有的网站都适合划分，如果网站划分会造成覆盖率降低到不可容忍的地步，还不如不划分。根据系统的不同要求，可以定义符合的最小覆盖率 ε。如果重复率低于 δ，表示重复率很低，不会影响到综合指标。

3.2 网站划分效果

假定可以容忍的最低覆盖率为 ε 为 80%，重复率常数 δ 为 20%，根据爬虫数量 N 从 1~10 这 10 种情况，对网站进行划分，然后根据网站的链接关系图进行模拟抓取，计算出评

价指标。

图 1 是重复率统计积分曲线，其中，*N* 代表爬虫数量。由于在 *N*=1 时，每个网站的重复率都是 0，因此没有在图中标明。

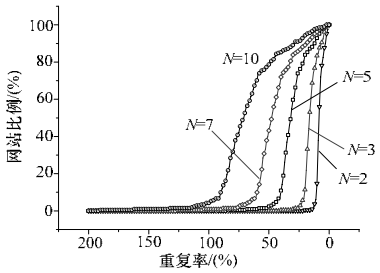

图 1　重复率统计积分曲线

图 2 是覆盖率统计积分曲线，其中，*N* 代表爬虫数量。由于 *N*=1 时，每个网站的覆盖率都是 100%，因此没有在图中标明。

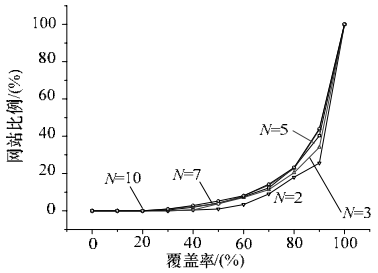

图 2　覆盖率统计积分曲线

图 3 是加速比统计积分曲线，其中，*N* 代表爬虫数量。由于当 *N*=1 时，每个网站的加速度都是 1，因此没有在图中标明。随着 *N* 的不断增大，加速度也在逐渐变大，但是加速度的增长并不是无限制的，当加速度达到一定的极限时就会停止增长；当 *N*=2 时，加速度集中在 1.5~2 之间；当 *N*=3 时，加速度集中在 2.2~2.6 之间；当 *N*=10 时，加速度变得很离散，没有特别突出的区间。同时参照图 1 可以看到，随着 *N* 的增大，重复率会越来越高。当 *N*=2 时，绝大部分的网站重复率不超过 10%；当 *N*=10 时，只有 20%的网站重复率不超过 50%。

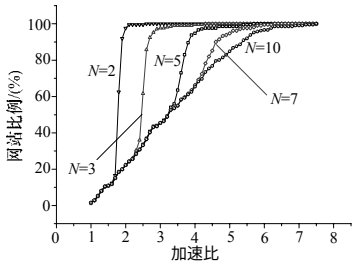

图 3　加速比统计积分曲线

图 4 是统计指标统计积分曲线，其中，*N* 代表爬虫数量。由于当 *N*=1 时，每个网站的综合指标都是 0.45，因此没有在图中标明。综合指标越大表明划分效果越好，相反，若等于 0 表明划分不合格，可以看出，随着 *N* 的增大，不合格的网站数量逐渐增多，而合格的那部分网站的综合指标先逐渐增大，然后逐渐减小，这说明并不是 *N* 越大，划分效果越好。对比图 2 可知，随着 *N* 的增大覆盖率越来越小，但变换趋势不大。覆盖率基本保持在一个较高的比例。

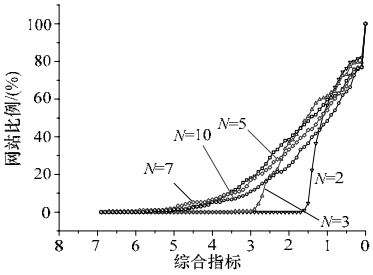

图 4　综合指标统计积分曲线

4　结束语

实验表明，根据路径进行网站划分可以达到很好的效果，尤其是当爬虫数量 *N*=2 或 *N*=3 时，重复率集中在较小的比例，覆盖率基本超过 90%，加速比效果也很明显，绝大部分的网站综合指标超过单机爬虫。这都说明本文提出的网站划分方法能很好地加快网页内容获取的速度，而且造成网页丢失的比例有很少，重复率也控制在可以接受的范围之内。

参考文献

[1] Cho J. Parallel Crawlers[Z]. (2002-05-11). http://www2002.org/CDROM/refereed/108/.

[2] Karger D, Lehman E, Leighton T, et al. Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web[C]//Proc. of STOC'97. New York, NY, USA: ACM Press, 1997.

[3] Chakrabarti S, Berg M, Dom B. Focused Crawling: a New Approach to Topic-specific Web Resource Discovery[J]. Computer Networks, 1999, 31(11): 1623-1640.

[4] Davison B D. Topical Locality in the Web[C]//Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM Press, 2000.

[5] 李晓明, 闫宏飞, 王继民. 搜索引擎：原理、技术与系统[M]. 1 版. 北京: 科学出版社, 2005.

编辑　陈　文