

面向博客的智能网络蜘蛛技术及其系统实现

原舒雨, 蔡皖东, 李勇军, 彭 冬

(西北工业大学 计算机学院, 陕西 西安 710129)

摘 要: 从博客网站的结构特点出发, 抽象出博客网站的链接分级模型, 并在此基础上提出了基于蚁群算法的“目录链接—博文链接”自适应采集策略, 设计并实现了针对博客网站的智能网络蜘蛛系统. 实验表明, 采用该策略的系统在采集效率上取得了显著提高.

关键词: 博客; 网络蜘蛛; 采集策略; 蚁群算法

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2011)01-0193-04

The Research and Implementation of Intelligent Spider in Blog

YUAN Shu-yu, CAI Wan-dong, LI Yong-jun, PENG Dong

(College of Computer Science, Northwestern Polytechnic University, Xi'an 710129, China)

Abstract: This paper analyzed the architecture of blog site, presented a Link-Rank model and also the “Catalog Link-Article Link” crawling strategy based on ant colony algorithm, then an intelligent web spider system face web blog was designed and implemented. In practice, this solution performed remarkably better on crawling efficiency.

Key words: Blog; web spider; crawling strategy; ant colony algorithm

1 引言

面对网络博客中的海量信息, 如何高效、准确的采集有效信息, 已成为搜索引擎中网络蜘蛛程序 (Web Spider, 或称网络爬虫、采集器) 所面临的新课题.

将目前网络蜘蛛所采用的一些通用采集策略应用到面向博客网站的信息采集过程^[1-4], 在采集准确性和覆盖率等判断采集策略优劣的性能指标上难以令人满意, 其原因是这些采集策略没有考虑到博客网站自身的结构特点, 缺乏针对性.

文中通过分析博客网站的结构特点, 抽象出面向博客网站的“目录链接—博文链接—其他链接”这一逻辑上的链接分级模型, 并在此基础上提出了一种基于蚁群算法的自适应采集策略. 该策略使网络蜘蛛模拟人的思维方式, 通过目录链接进而采集博文链接, 以此达到提高采集效率的目的^[5].

2 博客网站的特点与采集难点

(1) 链接种类繁多. 除了文章链接之外, 博客中还存在着大量的功能性链接和噪声链接.

(2) 博文等有效链接所处层次不固定. 有些博文可能置于某一博客板块的索引页甚至博客网站的首页之上; 而另一些博文则置于某一博主自己的分档之中.

(3) 链接冗余现象显著. 链接冗余即同一页面与多个 URL 相对应的现象.

目前采用的采集策略往往没有考虑到博客网站自身的结构特点, 使得很大一部分无效链接被采集; 同时由于有效链接往往位于不同深度层次, 影响了采集覆盖率; 另外大量的链接冗余现象也可能导致网络蜘蛛陷入采集陷阱之中.

3 基于链接分级模型的采集策略

3.1 链接分级模型

对于博客网站中各种形式的链接, 首先抽象并

定义如下几种链接类型:

定义 1 博客网站一般包含许多注册博主,每一博主有其唯一的链接,定义该类链接为博主链接。

定义 2 博客网站一般可以划分出若干板块,如人文板块,体育板块等,每一板块有其唯一的链接,定义该类链接为板块链接。

定义 3 每一博主可以撰写若干文章,收录于不同页面中,每一篇文章都有其唯一的链接,定义该类链接为博文链接。

定义 4 由于博主链接与板块链接有一定的相似性,其对应的页面中都包含了若干博文链接,因而可以把两者抽象为目录链接(即目录链接 = 博主链接 $\cup$ 板块链接)。

定义 5 把一些功能性链接、噪声链接等与采集主题无关的一类链接定义为其他链接。

在此基础之上,文中把博客网站中的链接归纳划分为三种级别:(1)目录链接;(2)博文链接;(3)其他链接。将上述概念抽象为博客网站的链接分级模型,如图 1 所示。

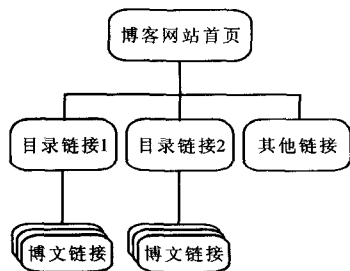


图 1 链接分级模型

3.2 “目录链接—博文链接”采集策略

网络蜘蛛是围绕各种各样的链接展开工作的,而博客网站中的链接又是有规律可循的,如果能够充分挖掘利用链接中包含的信息,可以用于改进网络蜘蛛的采集效率。

根据链接分级模型,网络蜘蛛所重点关注的博文链接往往包含在目录链接下。这样首先区分出各种级别的链接,进而找到目录链接,再采集其中的博文链接,称之为“目录链接—博文链接”采集策略。该采集策略有如下一些优点:

(1) 将注意力集中在目录链接与博文链接上,网络爬虫可以有效的规避一些无效链接。

(2) 在逻辑上规定了两级采集深度,有效的避免了采集陷阱的发生。

(3) 这一采集策略更加符合人的思维方式。

3.3 URL 的相似性及目录索引

定义 6 URL 的相似性:把两个 URL 以其中

的“.”“/”“=”等字符划分为不可再细分的完全相同的字符串 strSame 和两个 URL 中不相同的字符串 strDiff,如果两个 URL 划分出 strSame 和 strDiff 的个数以及出现顺序完全相同,则认为这两个 URL 是相似的。

定义 7 依据 URL 的相似性,把 URL 相似的目录链接归纳为同一个目录索引,并规定目录索引的形式为“strSame_{i1} + strDiff_{i1}”[+ “strSame_{i2} + strDiff_{i2}”,...]。

这样,URL 相似的目录链接就可以映射到同一个目录索引上。而对于一个新的链接,如果它可以映射到某个已知的目录索引上,那么该链接就很可能属于该类型的目录链接,这样就为判断目录链接提供了一种先验方法。

进一步将该目录索引与一个反映其可信程度的数值构成一个二元组,称之为索引二元组:

$$E = (e_1, e_2) = (\text{目录索引}, \text{权值})$$

其中, e_1 唯一的标明了—一个目录索引; e_2 反映了该目录索引的可信度,即一个可以映射到该目录索引的链接确实属于目录链接而非其他链接的置信程度。该数值越大,其置信度就越高。

3.4 基于蚁群算法的目录链接采集模型

蚁群算法^[6]是受蚂蚁觅食行为的启发而提出的一种路由算法。仿生学研究发现蚂蚁在行进过程中能够释放出一种可挥发的分泌物——信息素。所有蚂蚁都能够感知这种物质的存在及其强度。当蚂蚁遇到一个路口的时候,选择信息素残留浓度较大的路径前行的概率相对较大。这样就形成了一个正反馈现象:某条路径上走过的蚂蚁越多,后来的蚂蚁越有可能选择该路径。蚁群通过这种集体自催化行为最终找出最佳觅食路径。

将蚁群算法应用于网络蜘蛛采集目录链接的采集策略中,可以抽象出如下采集模型:一个链接的采集线程可以看作一只实体蚂蚁;目录链接相当于蚁群所要寻找的食物,而其他链接则不是食物;目录链接的相似性可以理解为食物较为集中的出现;蚁群所经过的存在食物的路径可以抽象为目录索引;反映目录索引可信度的权值则相当于信息素的浓度。若采集线程采集到有效的目录链接(即模拟蚂蚁找到食物),则增加对应目录索引的权值(即模拟提高信息素浓度),另外该模型定期降低所有索引权值的大小(即模拟信息素的挥发)。后续的采集线程会优先采集目录索引权值较大的链接(即模拟蚁群沿着信息素浓度较大的路径觅食)。该模型如图 2 所示。

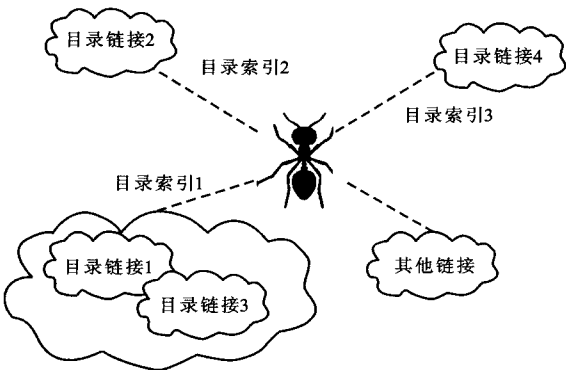


图 2 基于蚁群算法的目录链接采集模型

依据上述的采集模型,目录链接的采集工作流程如图 3 所示。

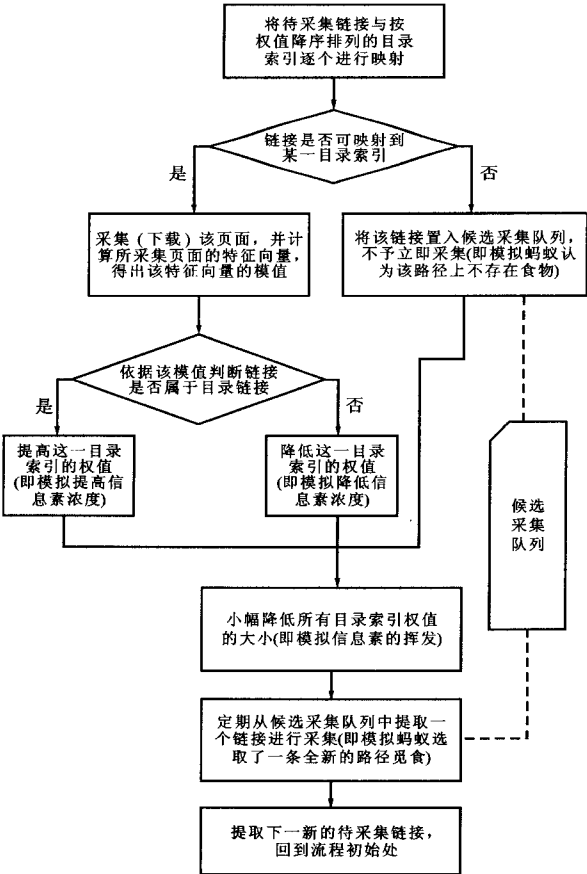


图 3 目录链接的采集工作流程

4 系统设计及实现

4.1 系统模型

基于博客网站的链接分级模型及“目录链接—博文链接”采集策略,文中设计并实现了针对博客的智能网络蜘蛛系统.该系统分为 3 个模块:链接提取模块、链接分级模块、链接下载模块,如图 4 所示.

链接提取与分级模块负责从原始页面的文本中

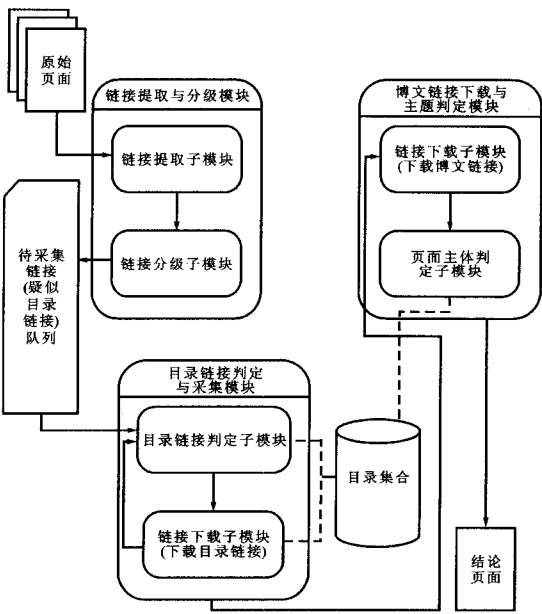


图 4 系统模型图

抽取疑似目录链接并存储;目录链接判定与采集模块负责判定目录链接并收录于目录集合;博文链接下载与主题判定模块负责下载博文链接,得出结论页面。

4.2 关键技术及其实现

(1) 链接提取与分级模块

根据链接分级模型设计网络蜘蛛,首先需要区分各种级别的链接.该模块主要负责这一任务,并将判定的目录链接存入待采集链接队列之中.

在 URL 表现形式上,博文链接通常包含一个博文 ID,该 ID 通常由 6 位或更多位数字构成,同一目录链接下的不同博文链接其 URL 只是该 ID 不同;博文链接在 URL 形式上表现为 htm、html、shtml 等文件类型;博文链接的字符长度较目录链接与其他链接相比,往往长出 50%左右.因而根据 URL 形式在采集之前初步区分出博文链接,称之为“一级先验区分”。

进一步分析下载的目录链接和其他链接其中的统计信息可以发现,目录页面的链接个数、博文链接个数、字符(汉字)个数等统计数据显著高于其他链接中的相应数据.其中页面中是否包含一定数量 URL 形式相似的博文链接是判断该链接为目录链接的必要条件.因而根据链接之中的文本统计信息在采集后判断该链接是否属于目录链接,称之为“二级后验区分”。

其中关于目录链接的二级后验区分的准确性是一关键点,它直接对目录链接的后续判断起到了指导性作用。

实现“二级后验区分”的基本思路是:首先把已下载页面中的有效链接个数、博文链接个数、字符(汉字)个数等统计数据抽象为链接的“统计特征向量”,即:

$$W = [W_1, W_2, \dots, W_n]$$

其中,每一维 W_i 分别代表不同的特征数据,并设计该向量的模值 $\bar{W}$ 的计算公式:

$$\bar{W} = \frac{\sum_{i=1}^n \epsilon_i \times W_i}{\sqrt{\sum_{i=1}^n \epsilon_i^2 \times \sum_{i=1}^n W_i^2}}$$

其中, ϵ_i 为权重系数,介于 $[0, 1]$ 之间,代表了某一具体的特征数据 W_i 的重要程度. 根据上述公式计算出

这些链接的平均模值 $W_{\text{avg}} = \frac{\sum_{k=1}^m \bar{W}_k}{m}$, 并设定一个阈值:

$$W_{\lambda} = \text{阈值系数 } \delta \times \text{平均模值 } W_{\text{avg}}$$

对于之后新出现的链接,其统计特征向量模值超过上述阈值,则记录其 URL 形式,存入待采集链接队列.

(2) 目录链接判定与采集模块

基于蚁群算法的采集策略中,目录索引的权值相当于信息素浓度,故目录索引权值更新策略的优劣,直接决定了实际搜索效果. 在实际的系统应用中采用了如下关于目录索引权值的更新策略:

对于一个新创建的索引二元组,设置其初始目录索引权值为 W_{ini} . 对于上述提高目录索引权值的一次操作,可采用在原有权值基础上增加一个 W_{ini} 的方式;另一方面,对于降低目录索引权值的一次操作,可采用将当前权值减半的方式. 对于模拟信息素的挥发,可采用在待采集链接队列中设置计数器,每当下载了定量链接后,将所有目录二元组中的权值都减去 W_{ini} 的方式. 同时设置一个权值上限 W_{max} 与一个权值下限 W_{min} , 当现有权值大于 W_{max} 时就不再增加,而当现有权值小于 W_{min} 时就不再减小. 采用该策略,可以较好的消除蚁群算法中信息素的停滞现象,使得常见的目录索引平稳的显现出来,且使非目录索引更快的剔除出去.

5 实验及结果分析

实验平台为 Windows Server 2003 的服务器,

通过 100M 以太网接入教育网,在一些知名的博客网站上进行了实验. 实验证明:相似链接可以映射到相同目录索引,并且链接数量越多,该现象越明显.

采用“目录链接—博文链接”采集策略的网络蜘蛛在采集的起始阶段其效率与通常采集方法相当,然而随着时间的推移,该采集策略的效率明显优于其他. 由于网络蜘蛛的搜索具有“重复性”,而该策略有效利用了先前的“经验信息”实现增量学习,当目录链接以及相应的目录索引累积到一定规模,就为进一步采集有效的文章链接提供了指导性依据,从而提高了整体效率.

6 结束语

文中提出了一种面向网络博客的网络蜘蛛采集策略. 该策略首先抽象出网络博客中三种级别的链接,在此基础上采用由目录链接寻找博文链接的拟人化采集策略,并在实验中取得了良好的采集效果. 下一步研究可侧重于如何更加快捷的发现目录链接以及提高判定目录链接的准确性上.

参考文献:

- [1] Cho J, Garcia Molina H, Page L. Efficient crawling through URL ordering [J]. Computer Networks and ISDN Systems, 1998, 30(7): 161-172.
- [2] Bra D P, Houben G, Kornatzky, et al. Information retrieval in distributed hypertexts[C]// Proc. of the 4th RIAO Conference. France: Paris, 1994.
- [3] 周立柱,林玲. 聚焦爬虫技术研究综述[J]. 计算机应用, 2005, 25(9): 1-3.
- [4] 宋聚平,王永成,尹中航,等. 面向主题的网页搜索系统[J]. 上海交通大学学报, 2003, 37(3): 401-403.
- [5] 张兴华. 智能搜索引擎的机理、实现技术及发展趋势[J]. 现代情报, 2003(12): 66-68.
- [6] 段海滨,王道波,朱家强,等. 蚁群算法理论及应用研究的进展[J]. 控制与决策, 2004, 19(12): 573-576.

作者简介:

原舒雨 男, (1985-), 硕士研究生. 研究方向为网络与信息安全.

蔡皖东 教授, 博士生导师. 研究方向为网络信息安全、网络舆情分析、系统安全测评、网络性能测量等.

李勇军 博士, 讲师. 研究方向为网络信息安全.

彭冬 硕士研究生. 研究方向为网络信息安全.