主题搜索引擎网络爬虫搜索策略的研究与实现①

刘淑梅 夏 亮 许南山 (北京化工大学 信息研究院 北京 100029)

摘 要: 根据网络页面结构的特点，提出通过页面之间的主题传递来预测页面主题相关性的方法，解决了主题爬虫通道堵塞，抓取遗漏的问题。首先根据锚文本传递一个相关性信息值，如果锚文本给出的信息是相关，相关阈值就直接传递；如果是不相关，就乘以遗传基因比例之后传递。传递的过程中如果遇到相关的网页就恢复链接的相关性信息值到初始值。最后根据实验结果验证了算法的查全率与查准率，查全率有显著的提高。

关键词: 网络爬虫；搜索引擎；主题相关；遗传；抓取

Search Strategy and Achieve of the Topic Search Engine Spider

LIU Shu-Mei, XIA Liang, XU Nan-Shan

(Academe of Information, University of Chemical Technology, Beijing 10029, China)

Abstract: According to the characteristics of the cyber page structure, this paper proposes the theme which predicts the correlativity by delivering the theme among the pages, and solves the problems of channel jamming and capture omission. Firstly, a correlative information value is delivered according to the anchor text. If the information given by the anchor text is correlated, the correlative threshold will be delivered directly. Otherwise, it will be multiplied by the genetic ratio before delivery. In the process of the delivery, correlative information value may be reset to the initial value if it encounters the correlative Web page. At last, the recall ratio is proven to be greatly improved based on the experimental result.

Keywords: cyber worm; search engine; theme correlativity; genetic algorithm; crawl

1 引言

随着互联网的飞速发展，网上信息量大量增加，传统的搜索引擎已经不能满足人们的搜索信息需求，这就促进了主题搜索引擎的发展。主题搜索引擎的信息量是针对某一个特定领域，主题搜索引擎网络爬虫抓取网页的时候只抓取主题相关的网页，过滤不相关网页的抓取工作。为了提高主题爬虫的工作效率，过滤工作就相当重要。目前的主题爬虫在抓取网页时，往往会遗漏大量的相关网页，因此本文提出一种在权值传递过程中权值恢复的策略弥补主题爬虫的这种不足。

2 主题爬虫的模型框架

如模型框架图所示(见图 1)，在系统框架中，主过程由控制器，解析器，资源库三部分组成。过滤处理部分由页面处理，主题判断两部分组成。控制器主要的工作是给多线程中的各个线程分配工作任务。解析器主要的工作是下载网页，页面处理，主题计算等工作，爬虫的基本操作由解析器完成。资源库用来存放下载到的网页资源，并对其建立索引。 主题网络爬虫与通用的网络爬虫[1]相比，多了一个主题判断筛选的过程，引导爬虫的抓取方向，缩减爬虫的工作量。初始 url 地址通过控制器分配给解析器，解析器根据 url 地址从 web 互联网上抓取网页，将其放进资源库，随后将 url 地址放进等待队列。，抓取时分析页面，提取 url，根据 url 的链接文本判断主题，将相关 url 放进等待队列。

① 收稿时间:2009-06-06

© 中国科学院软件研究所 http://www.c-s-a.org.cn

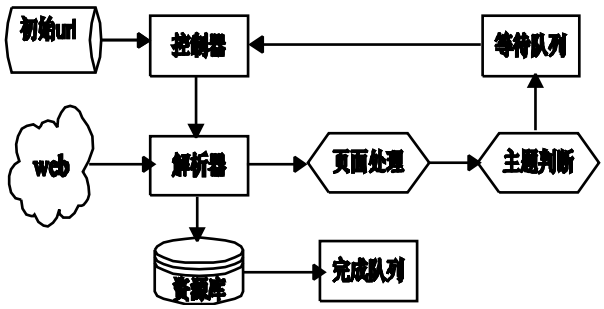

图 1　模型框架图

3　主题爬虫的算法设计

3.1 主题网页链接结构

互联网网页之间的链接是呈蜘蛛网形状，网页与网页之间形成错综复杂的网络通道。对于通用网络爬虫，可以在网络通道里面任意爬行，它们的要求是尽可能多的发现更多的网络通道。对于主题爬虫，网络通道有特定的方向，只能沿着特定的主题方向爬行。目前多数主题爬虫[2-4]在遇到通道堵塞时候的做法是丢弃现有的通道，换另一条并行的通道。这样的做法有一个缺陷就是可能遗弃一些通道深度大的主题相关网页。如主题网页链接图所示，我们假定 A，B，D，F 是相关网页，C，E 是不相关网页。爬虫从 A 网页开始抓取，通道 A-B-D 是无阻塞的，B 与 D 可以很容易抓取到，在通道 A-C-F 中因为有 C 阻隔，按文献[1]的算法当爬虫抓取到这里可能会停止，关闭 A-C-F 通道，结果是 F 就不能够被抓取。下面将要呈现的算法能够使爬虫沿着 A-C-F 通道继续往下爬取，得到相关网页 F，如图 2 所示。

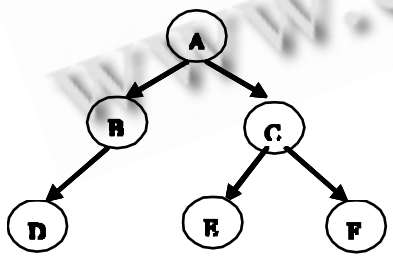

图 2　主题网页链接图

3.2 相关度计算

研究发现在基于 HTML 协议的网页中，每一个 url 的链接文本最能概括表达 url 所指向的网页内容，在网页中有一个链接模型为 text，基于网页结构的明确性，text 往往是一个非常精确的概括性描述文字。在这种结构基础上，我们采用文献[3]中的向量空间模型来计算链接文本 text 的相似度，用它标记 urltext 的相关度。模型公式如公式(1)。

$$SIM(P_j,R)=\frac{\sum_{i=1}^{t}W_{i,j}\times W_{i,r}}{\sqrt{\sum_{i=1}^{t}W_{i,j}^2\times\sum_{i=1}^{t}W_{i,r}^2}} \quad (1)$$

其中 $W_{i,j}$ 表示特征向量在链接文本中的权值，$W_{i,r}$ 表示特征向量 i 在主题特征库中的权值，R 代表主题特征向量，$SIM(P_j,R)$ 表示链接文本 P_j 的相关度。

3.3 主题爬虫抓取算法

在介绍算法的开始需要先做两个定义[5]

定义 1. 父网页：网页 A 中有 url 链接到网页 B，那么网页 A 就是网页 B 的父网页。

定义 2. 子网页：网页 A 中有 url 链接到网页 B，那么网页 B 就是网页 A 的子网页。

爬虫抓取过程中使用了四个队列，分别是等待队列，抓取队列，错误队列，完成队列。

等待队列：爬虫解析到的 url 先保存到等待队列中，在等待队列中的 url 按照特定的排序法则进行排序，等候爬虫的抓取。

抓取队列：url 正在被抓取时放进抓取队列，目的是防止 url 被同时多次抓取。

错误队列：在抓取过程中出错的 url 保存到错误队列。

完成队列：一个 url 被爬虫完全抓取之后就将 url 放进完成队列。

爬虫的抓取算法如下：

① 将初始页面 url 集合放进等待队列，分配每个 url 一个相关性消息值 m，并给每个 url 同样的相关度值，这个相对于后面将要计算到的值较大。初始页面会人为根据主题进行筛选，所以与主题的紧密高。人为的给定一个高的相关度值优点有两个，首先，减少爬虫的计算量，这些种子站点不需要通过相关度的计算。其次，可以在等待队列中置于较靠前的位置，在以后的更新过程中，可以优先更新。

② 对等待队列中的 url，先根据 m 值大小排序，再根据相关度的大小排序。

③ 根据第二步排好序的等待队列，将排序最前的 url 拿出放进抓取队列，爬虫开始抓取。

© 中国科学院软件研究所 http://www.c-s-a.org.cn

④ 下载网页到本地磁盘，并建立索引，然后将 url 地址放进完成队列。

⑤ 利用解析器解析出网页中的链接与对应的链接文本，利用公式 1 计算链接地址的相关度值。

⑥ 将第四步得到的相关度值与相关度阀值 f 进行比较，其结果分为三种情况：

第一种情况是相关度值大于相关度阀值，且父网页的相关性消息 m 值等于初始值，则直接传递父网页的 m 值给子网页。

第二种情况是相关度值大于相关度阀值，且父网页的相关性消息 m 值小于初始值，则恢复 m 值为初始值，传递 m 值给子网页。

第三种情况是相关度值小于相关度阀值，则将父网页的 m 值乘以遗传基因比例 b 传递子网页的(b 值大于 0 小于 1)，子网页的相关性消息值是 m*b。

⑦ 将 url，m 值，相关度值放进等待队列，重复第二步。

⑧ 算法结束。

在上面的算法中，爬虫的等待队列里面不仅仅有 url 的相关度值，还有一个相关性消息值 m。m 值在爬虫体系中为爬虫指引主题通道，使用算法第六步中的法则传递父网页与子网页之间的 m 值，父网页通过计算链接文本的相关度，与主题切合的就遗传 m 值给相应的链接，与主题不贴合就遗传部分 m 值给相应链接。爬虫首先会沿着 m 初始值主线通道爬行，在爬行的过程中在主线通道的上会开辟很多二层 m 值通道，二层通道上的 m 值是 m 初始值的 b 倍，因为 b 是大于 0 小于 1 的，所以二层通道没有主通道大。m 初始值主线通道上遇到堵塞之后，爬虫会寻找另一条最靠近主题种子团的 m 初始值主线通道，没有 m 初始值主线通道的时候，爬虫寻找二层 m 值通道。一旦发现链接文本相关度大于相关度阀值的链接 url，恢复 url 的 m 值为初始值大小，开辟一条新的主线通道，爬虫跳过二层通道沿着新开辟的主线通道继续爬行。在通道内部，主线通道的 m 值一直不变，永远是初始值，主线通道以外的通道 m 值是呈递减状态的。

如主题网页通道图所示(见图 3)，爬虫由主题种子团出发，沿着 m 初始值主线通道 1，2，3，4 爬行抓取。当已有的主线通道已经没有网页可抓取，爬虫沿着图中的细线通道继续爬行。遇到相关链接，从新链接地址开始恢复其为主线通道。图中 5，6 通道就是由非主线通道发现的主线通道。在爬虫等待都列中的爬虫先根据 m 值 的大小排序，再根据链接文本相关度大小排序，可以保证主线通道上爬虫优先被抓取，而非主题的链接也并没有抛弃。爬虫抓取完主题链接之后再由他们搜索新的主题的链接，再次优先抓取。

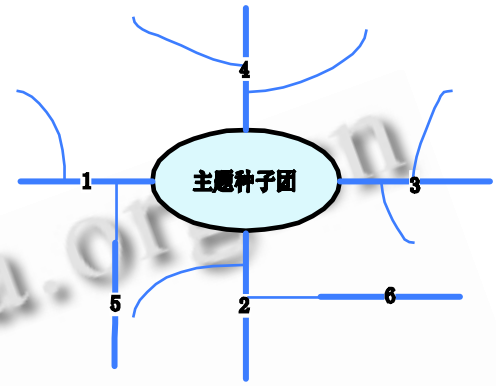

图 3 主题网页通道图

4 实验结果分析

最后在实验室对算法进行了实验分析，实验硬件环境是 dell 台式机一台，奔腾 4 处理器，512M 内存，XP 系统，sqlserver2000 数据库，网络带宽 100.0Mbps。开发语言是 java，开发环境是 eclipse。

验证主题爬虫效率的方法有两个，一个是查全率，一个是查准率。

查全率 = 采集的目标页面数/目标页面总数；

查准率 = 采集的目标页面数/爬行页面总数；

本实验以化学主题，在百度中搜索化学，以前 50 个页面作为主题种子团。相关性信息初始值 m 给定为 100，遗传基因比例 b 为 0.8，相关度阀值 f 为 0.2。与只基于文本内容的 best first search 方法做比较，得到的查准率与查全率数据如查准率图与查全率图，如图 4、图 5 所示。

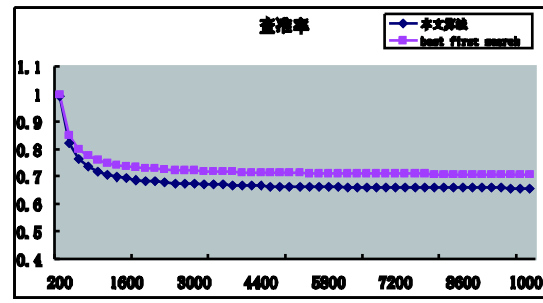

图 4 提案算法与 best first search 方法查准率图比较

© 中国科学院软件研究所 http://www.c-s-a.org.cn

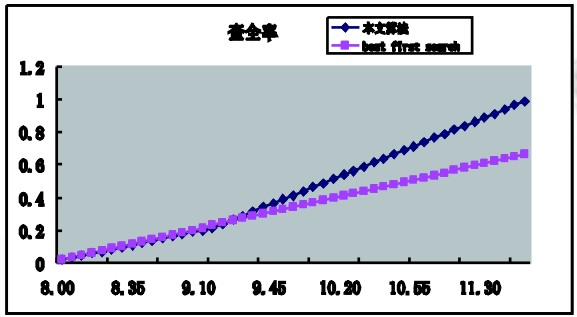

图 5　提案算法与 best first search 方法查全率图比较

由查准率图我们可以看到本文的算法与 best first search[5,6]算法在查准率上几乎是平行相等的，best first search 算法计算了父网页文本的相关度并对子链接加上反馈，所以查准率稍高，缺点是计算量大，影响了速度。但是在查全率上，本文的算法的优势非常明显，抓取前期由于两者都在主通道上抓取，两者没有明显差距，到后期本文算法较之 best first search 算法优先发现大量被后者遗忘的主题网页，开辟了许多新的主题通道，使爬虫的效率明显提高，提案算法的优势得到充分体现。

5　结论

本文的创新点是：链接文本相关度算法与主题信息值遗传恢复的算法相结合，提出一种在权值传递过程中权值恢复的策略，使主题爬虫不断发现新的主题通道，扩大了主题爬虫覆盖度，同时保证了主题爬虫的抓取效率。

不同的个体对同一个互联网信息的需求度不一样，对搜索引擎的后台数据库来说，要满足不同的个体，需要尽可能的让资源库更加完整，所以主题爬虫在相关性获取上就是一个宽泛的计算。本文提出的算法基于这个考虑，直接跳跃父网页文本的计算，只计算链接文本，提高了计算速度，也保证了质量。

参考文献

1　王凤红.简单分布式网络爬虫模型的设计与分析.中国现代教育装备, 2008,4(62):76－78.
2　倪贤贵,蔡明.基于链接结构和内容相似度的聚焦爬虫系统.计算机工程与设计, 2008,7(29):1709－1710.
3　李勇,韩亮.主题搜索引擎中网络爬虫的搜索策略研究.计算机工程与科学, 2008,30(3):4－6.
4　郑健珍,林坤辉,周昌乐,康恺.基于本体语义的定题爬虫.山东大学学报, 2006,41(3):90－94.
5　刘金红,陆余良.主题网络爬虫研究综述.计算机应用研究, 2007,24(10):26－29.
6 Cho J, Garciam H, Page L. Efficient crawling through URL ordering. Computer Networks and ISDN Systems, 1998,30 (107)-161-172.

© 中国科学院软件研究所 http://www.c-s-a.org.cn