多策略的主题集中式万维网爬虫设计

王 超 朱 炜 李 俊 潘金贵
(南京大学计算机软件新技术国家重点实验室 南京 210093)
(南京大学多媒体技术研究所 南京 210093)

摘 要 万物网搜索引擎的建立、操作和维护需要许多的资源,而且在信息时效性和对特定用户的针对性方面还存在着不稳定性。在"主题集中式万维网爬虫"方面的研究希望通过利用主题减少对信息的爬行范围,同时提高信息的利用率。相关的一些研究者已采用不同的方法进行了主题集中式爬虫的设计。本文讨论了多策略的主题集中式爬虫系统的设计,它具有低网宽消耗和容易执行的特点。实验表明:本系统可综合网页的相关性和重要性两方面的需要,并表现出良好的稳定性。同时,本系统在选择优先战略方面是可调和,有很强的灵活性。

关键词 万维网,主题集中,爬虫,多策略,相关性,重要性

Design of A Multi-strategy Topic Specific Web Crawler

WANG Chao ZHU Wei LI Jun PAN Jin-Gui
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)
(Institute of Multimedia Technology, Nanjing University, Nanjing 210093)

Abstract The construction, operation and maintenance of Web search engines require lots of resources while they still have disability in providing timely and topic specific information. The research on "topic specific Web crawler" hopes to use topic to reduce the utility of resource as well as to increase the information usage. Related researchers have adopted various methods in designing the topic specific crawler. In this paper, we discuss our design of topic specific crawler that adopts multi-strategy. Our crawler is characteristic of low cost of network bandwidth and easy implementation. The experiment reveals that it can synthesize the concern of page relevance and importance and have a good performance in stability. Meanwhile, the crawler is adjustable for choosing preferred strategy, demonstrating good flexibility.

Keywords Web, Topic specific, Crawler, Multi-strategy, Relevance, Importance

1 引言

万维网(World Wide Web,简称 Web)高速发展,成为人们现在一个不可或缺的信息来源。快速增长的万维网形成了信息的海量性,这种特性向建筑在万维网信息上的通用型搜索引擎和通用型万维网爬虫提出了挑战,因为要对如此海量的信息进行爬行和索引,需要占用很大的网络带宽,花费很大的存储空间及计算能力,即使这样,也仍然无法很好保证信息的时效性和对特定用户的针对性。

针对以上情形,研究者们进行了主题集中式爬虫(Topic-Specific Focused Crawlers)[1]的研究,希望通过利用主题减小对万维网信息的爬行范围,同时提高对信息的利用率。

"Fish"系统[2]是最早的"主题集中式爬虫"之一。该系统采用深度优先方法对万维网资源进行游历,对网页的相关度评价采用基于关键字和正则表达式的方法。"Shark"搜索算法[3]在"Fish"基础上进行了改进。它采用了向量空间模型来对网页主题的相关度进行评价,同时精化了对需要爬行的超链的评价,充分利用了超链的锚文本信息。IBM 的 Soumen Chakrabati 和他的同事也提出了一种"主题集中式的爬虫"系统[4],这个系统利用分类器(classifier)来评价网页的相关性以决定其取舍,用"蒸馏器"(distiller)来分析网页的重要性以决定扩展的优先度,利用爬虫(crawler)获得指定的网页。另外,还有使用强化学习方法的,如"Cora"搜索引擎[5]的爬虫设计。

以上不同爬虫系统各具特色,但在爬行策略上都有比较固定的倾向性。本文提出了多策略的万维网主题集中式爬虫系统,它构成了 Dolphin 系统--一个面向兴趣、多用户协作的万维网搜索系统--的核心模块。它在爬行策略上具有多样性,可根据需要进行灵活的调节,在使用中有很好的表现。本文首先介绍主题集中式爬虫的常用游历策略以及对策略评价的出发点,然后重点讨论的多策略的构造及实现机制,并利用实验从相关度和重要度两方面对多策略机制下的爬虫进行评价,最后进行了总结。

2 游历策略

2.1 策略介绍

万维网上超文本互相指引,因此,可以将万维网抽象成一个图。用图 $G\langle V,E\rangle$ 表示万维网,其中 V 是所有网页的集合,E 是网页互相指引形成的有向边集合。

根据人们制作超文本的习惯,在一个网页中引用其他的

王 超 硕士研究生,研究方向为信息搜索、Web Mining 和 Agent 技术。朱 炜 硕士研究生,研究方向为信息搜索技术、中间件和 Agent 技术。李 俊 硕士研究生,研究方向为 Agent 技术,人机交互技术,多媒体移动网络技术。潘金贵 教授,博士生导师,从事中间件,Agent 技术及多媒体远程教育等的研究。

网页，那么它们之间总存在一些关系，而这种关系就有可能是同一主题或相近主题的关系。这样，对于一个主题 T，与之相同/相近的主题网页的集合也构成了一个图 $G_t(V_t,E_t)$，其中 $V_t\subseteq V$，$E_t\subseteq E$。利用一个"种子"主题网页集 V_0，$(V_0\subseteq V)$，以某种策略游历图 G，获得图 G_t，就是主题爬虫的任务。

现有的游历策略，主要有：广度优先、深度优先、相关度优先、引用度优先。另外也有几种游历策略的综合运用。广度优先和深度优先就是普通意义上图的游历策略，不再赘述。

相关度优先是指对未访问的节点网页，优先考虑那些与主题相关度高的节点。评价网页 P 的相关度可以从如下的角度考虑：

1. P 所包含的内容与主题的相关性 $R_c(P)$，一般可以用向量空间模型(Vector Space Model)来刻画该相关性；

2. 指向 P 的超链中的锚文本(Anchor Text)与主题的相关度 $R_a(P)$；

3. P 与代表主题的"根网页"之间连接的距离 $R_d(P)$。

综合考虑以上评价角度，相关度越高，则越早被游历并分析。

引用度优先策略是以网页超链结构分析为背景的。它是以网页在万维网中的"重要度"为扩展依据的。简单地讲，万维网上有两种网页很重要，一类是被很多网页引用的网页，称为 Authority 网页；另一类是引用了很多 Authority 网页的网页，称为 Hub 网页[6]。从这样的角度看，在网页扩展中，那些 Hub 网页应该被优先考虑。对网页重要度进行评价的，还有 PageRank 方法[7]，因此，也有根据网页的 PageRank 值作为扩展依据的。

2.2 策略评价

不同的游历策略会生成不同的 G_t。根据上文的讨论，一个好的游历策略，应该能够使生成的 G_t 有以下的特性：

- G_t 应该比较大，即包含比较多的网页；
- 对多数网页 $p\in V_t$，$R(p)$ 比较大，其中 R 表示网页和主题的相关度；
- V_t 中包含比较多的重要主题网页，重要主题网页可以通过多次实验和专业人士的评价获得。

另外，好的游历策略还应该有较高的效率，主要表现在尽可能少下载、少分析无关的网页。

3 多策略设计

为了选择一个好的游历策略，我们从相关度、重要度等方面进行了综合考虑。

3.1 相关度分析

如上文所述，相关度评价有三种方式。$R_c(P)$ 表示网页与主题的相关度，当网页还未下载时，$R_c(P)$ 是无法知道的；而如果将网页下载下来对它进行相关度分析，就会增加系统的开销。因为有可能下载下来的很多网页根本就不相关，这样就降低了系统的效率。因此，有必要利用一种预测机制，对网页在下载前就进行评价。利用 $R_a(P)$ 可以建立这样的预测机制，因为 $R_a(P)$ 可以通过分析有超链指向 P 的网页得到。借鉴向量空间模型，可用如下公式来计算它：

$$R_a(P)=\frac{\sum_{k\in t\cap p} f_{kt}\cdot f_{kp}}{\sqrt{\sum_{k\in t} f_{kt}^2\cdot\sum_{k\in p} f_{kp}^2}}$$

其中，t 表示主题的关键字集合，p 表示指向网页 P 的超链的锚文本及周边的文本的关键字集合，f 表示关键字在相应部分出现的频率。

根据上文，在考虑相关度时，还有一个因素，那就是 $R_d(P)$。它可用如下的公式表示：

$$R_d(P)=1/d$$

其中，d 为 P 与"根网页"的最近超链距离，当 P 属于根网页集时，则 $d=1$。

3.2 重要性分析

首先讨论下 PageRank 算法[7]，它是网页重要度分析的经典算法之一，由 Page 等人提出，用来对大规模网页集中的网页做全局的重要性评价[8]。该算法从分析网页超链结构的角度入手，一个网页，如果被若干个网页引用，那么它的重要度就大致由那若干个网页的重要度决定。如果一个网页指向若干个网页，那它就会把自己的重要度分布给那若干个网页。可用公式表示如下：

$$PR(p)=(1-\gamma)+\gamma\sum_{d\in in(p)}\frac{PR(d)}{|out(d)|}$$

其中，$RP(p)$ 表示网页 p 的 PageRank 值，$in(p)$ 表示指向网页 p 的所有其他网页的集合，$out(d)$ 表示网页 d 指向其他所有网页的集合，$\gamma\in(0,1)$，它是个阻尼因子(damping factor)，表示一个随机用户访问下一个随机网页的概率。

从 PageRank 的公式可见，它的计算是一个迭代的过程。首先，每个网页节点被赋予一个初始值，然后，利用一个描述网页节点互联的矩阵进行迭代计算。迭代过程中，节点的值会逐渐收敛，一定程度后，这个值就反映了对应节点的 PageRank。

容易发现，要计算 PageRank，需要构造出比较完备的节点连接图。连接图中的连接信息是计算 PageRank 的关键，一个不完备的连接图，将会导致不完备的 PageRank。

根据以上分析，如果用 PageRank 作为主题网页扩展的依据，将就会有如下的缺点：

1. PageRank 的计算量较多，不可能在每次做扩展决策时进行计算，只能在一定时间间隔后进行，而在这段时间间隔里，有可能会产生一些潜在很重要的新节点无法被排序和扩展；

2. 由于在扩展中，图的连接信息是不完备的，PageRank 反映的也只是不完备图中节点的优先度；

3. PageRank 的计算是从全局出发，针对图中所有节点的，而实际这些节点中，有的节点已经被扩展，不再参与扩展排序，而只是用来传递重要度。如果发现很多重要的节点其实已经被扩展了，那么计算在一定程度上被浪费。

通过分析 PageRank 作为扩展网页依据的不足，我们从提高效率的角度出发，对它进行了简化。简化后的 PageRank，我们称之为 TimelyRank(TR)。TR 在每次分析网页时进行调整，公式如下：

$$TR(p,t_p)=TR(p,t_p-1)+TR(d,t_d)$$

其中，$TR(p,t)$ 表示网页 p 在时刻 t_p 的 TimelyRank 值，$t_i=0,1,2\cdots$，表示网页 i 的逻辑时间，每次对 i 网页进行 TimelyRank 值计算，它的逻辑时间就增加 1，$t_i=0$ 时，网页 i 有初始的 TimelyRank 值；d 表示指向网页 p 的网页。

从公式可见，TR 继承了 PageRank 的思想，但在计算方式上做了变动。每次扩展都将会进行一次计算调整，但计算只发生在所涉及到的网页，排序调整也可控制在未被访问的若干网页中。另外，迭代也在扩展中无形地发生，只是这种迭代不是从全局出发，因此重要性没有很好地在全局图中传递，排

序结果也会比较"短视",这是它的不足之处。

3.3 多策略机制

综合以上对网页相关度和重要度的分析,我们设计了混合游历策略。在该策略中,我们使用如下的公式作为选择未访问网页的依据:

$$D(p,t)=\alpha\cdot R_t(p)+\beta\cdot R_d(p)+\gamma\cdot TR(p,t)$$

其中,$0<\alpha,\beta,\gamma<1$,且 $\alpha+\beta+\gamma=1$,作为对不同评价的权值调整。

4 实验分析

我们设计了实验来验证爬虫的有效性。在实验中,我们主要从两个方面对实验结果进行了评价。一个方面是评价 Crawler 在维持主题相关性方面的效果;另一方面是评价 Crawler 对重要网页的挖掘能力。

4.1 相关性评价

首先,分析 Crawler 在维持主题相关性方面的效果。我们参考了文[9]的一种方法,对主题相关性进行评价。这个方法评价网页集随时间变化的平均相关度,采用如下公式进行计算:

$$sim(q,s(t))=\frac{1}{|s(t)|}\sum_{p\in s(t)}\frac{\sum_{k\in p\cap q}w_{kq}^{tfidf}w_{kp}^{tfidf}}{\sqrt{\sum_{d\in p}(w_{kp}^{tfidf})^2\sum_{k\in p}(w_{kq}^{tfidf})^2}}$$

其中,q 表示某个主题,它由若干个该主题下具有代表性的网页构成;$S(t)$表示在时刻 t 为止爬行到的网页集;w_{kd}^{tfidf}表示单词 k 在文档 d 中的 $tf*idf$ 权重,它用如下公式计算:

$$w_{kd}^{tfidf}=f_{kd}\cdot(1+\ln(\frac{|S|}{n_k}))$$

其中,f_{kd}是单词 k 在文档 d 中出现的频度;$|S|$是网页集 S 的大小;n_k 是单词 k 在网页集 S 中出现的文档频度。

我们选择四个主题进行了实验,每个主题使用了 3 到 4 个网页作为主题的初始网页集,如表 1 所示。我们对每个主题进行不同策略权重的爬行,即选择了四种不同的权重参数向量(α,β,γ),然后对爬行结果进行如上的相关度分析,并按照对应的权重参数对四个主题的结果进行了平均,以减少单个主题结果的随机性。实验结果如图 1 所示。

表 1 主题初始网页集

"Java"主题	"linux"主题
http://java.sun.com	http://www.linuxorg
http://www.javaworldcom	http://www.redhat.com
http://www.microsoft.com/java	http://www.linxdoc.org
http://www.javaboutique.intent.com	
"abortion"主题	"考研"主题
http://www.gynpages.com	http://www.kaoran.net
http://www.naral.ogr	http://www.kaoyan.com
http://www.abortionfacts.com	http://www.cer.net/jiao_yu/kao_yan
	http://www.edu.cn/HomePage/jiao_yu_fu_wu/kao_yan

当 $\alpha=1.0,\beta=0.0,\gamma=0.0$ 时,根据公式,爬行的策略可以认为是锚文本预测相关度优先;当 $\alpha=0.0,\beta=1.0,\gamma=0.0$ 时,可以认为是广度优先策略;当 $\alpha=0.0,\beta=0.0,\gamma=1.0$ 时,爬行的策略可以认为是链接度优先;而当 $\alpha=0.4,\beta=0.3,\gamma=0.3$ 时,可以认为是综合策略。

分析图 1,在爬行初期,广度优先策略对应的网页集,它的主题相关度相对较高,综合策略和锚文本预测相关度优先次之,而链接度优先最后。当爬行到一定程度后,广度优先策略对应的网页集,它的相关度有较大的下降,链接度优先也有一定的下降趋势,综合策略及锚文本预测相关度优先,它们对应的网页集相关度尽管也在下降,但下降趋势相对比较缓慢且稳定。

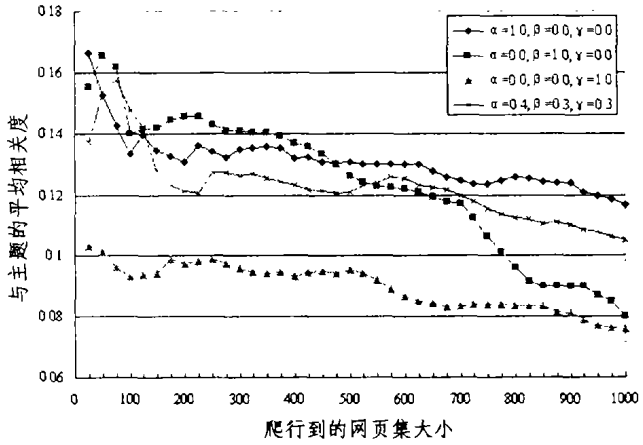

图 1 网页集的主题平均相关性实验比较

广度优先策略在初始时相关度效果较好,而在后期效果下降很大,这是由它的本性决定的。广度优先的本性,就是某段时间集中在一个站点的相关网页上爬行,如果这个站点恰好是与主题相近的站点(这种情况一般发生在开始阶段),那么,由于同一个网站所使用的词汇都有很大相似性,因此它在这个阶段生成的网页集就有很大的相关性。而当爬虫爬出主题网站,爬到一个不相关的网站,这时,网页集的平均相关度呈很大的下降趋势。

4.2 重要性评价

分析爬虫对重要网页的挖掘能力。在进行此项评价前,有必要对"重要网页"做个解释。如果根据人们的主观评价,来确定一个网页是否重要,以此获得一个重要网页列表,这样显然不够客观。因此,我们是使用 HITS 算法[6]来获得重要网页列表,该算法是通过网页集合中的超链结构信息来计算网页重要性的,因此具有一定的客观性。合并每个主题下各个爬行策略获得的网页集,得到一个综合的网页集,对该网页集进行 HITS 计算,我们就可获得每个主题对应的重要网页列表,作为评价爬虫挖掘重要网页能力的依据。好的游历策略应该可以尽可能早地访问到这些重要网页;并且,在游历过程中,应该可以尽可能多地覆盖到这些网页。

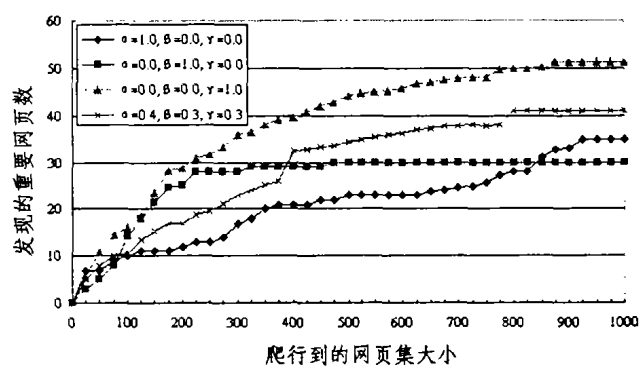

图 2 发现重要网页能力的实验比较

(下转第 208 页)

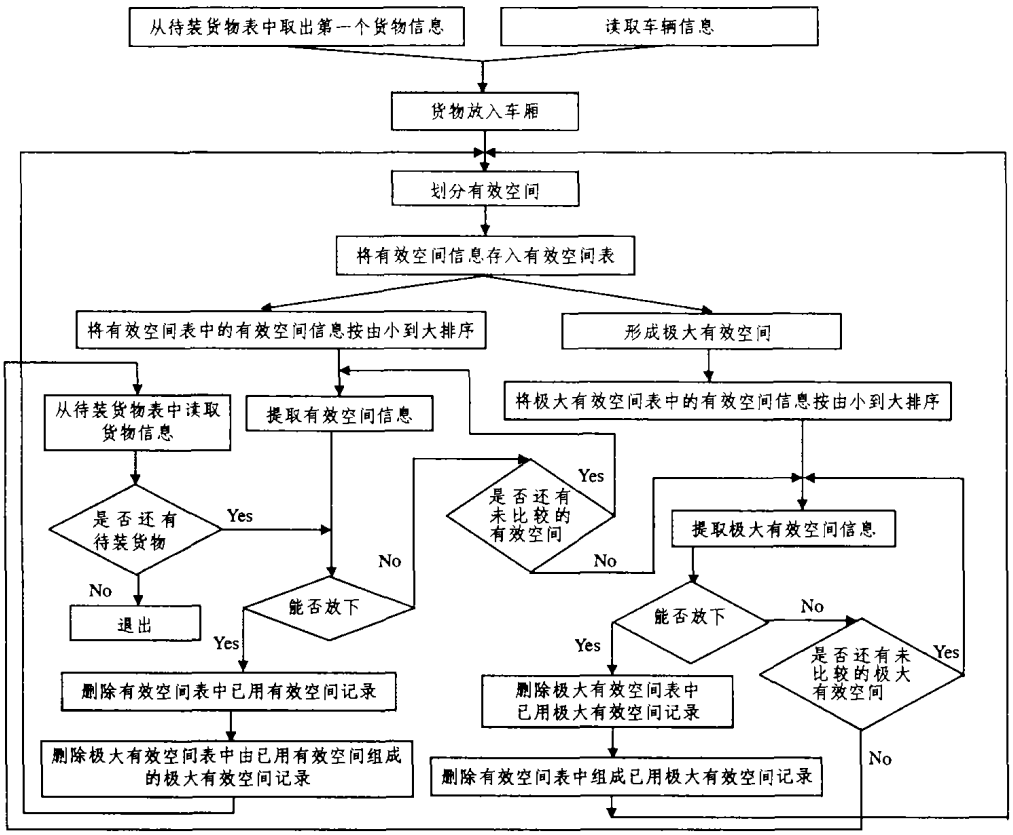

图2 算法基本流程

法和遗传算法等。本文基于有效空间的算法提供了此类问题的一种实际可行的解决方法，其最大特点是可以将货物装载过程与装载顺序分离开来各自独立优化，且由于采用空间解析方法，也便于算法的程序实现，实践证明具有较大的应用价值。

参 考 文 献

1 赵民义. 装箱问题的近似算法[J]. 运筹学杂志，1993(12)：1～6

2 Abdou G, El-Masry M. Three-dimensional random stacking of weakly heterogeneous palletization with demand requirements and stability measures [J]. International Journal of Production Research ,2000,38(14):3149～3163

3 陈迎春，吴晓平，等. 约束装箱问题的遗传混合算法求解[J]. 运筹与管理，2002(4)：21～25

4 阎威武，邵惠鹤，等. 集装箱装载的一种启发式算法[J]. 信息与控制，2002(8)：353～356

(上接第86页)

图2所反映的就是不同策略对重要网页的发现能力。其中可见，链接度优先策略在这方面表现较好，综合策略次之，锚文本预测相关度策略的表现更次，而广度优先策略的表现在初始阶段不错，但到后期效果较差。这个结果与我们设计的预想相符。

结论 本文提出了主题集中式万维网爬虫的多策略设计。结合实验，我们总结了该爬虫设计的若干特色：①多策略。通过多策略的综合使用，我们兼顾了爬行中需要对网页的相关性和重要性两方面的评价。实验表明，单从相关度或重要度评价，综合策略并不一定是最优的，但综合两方面看，它有很好的兼顾性，同时也表现出较好的稳定性。②可调节性。通过调整参数，可以方便地进行策略调整，加强或减弱某种策略对扩展的影响，具有很强的灵活性。③较小的网络开销。爬虫是通过对网页中的超链进行综合评价，以决定是否扩展。因此，评价低的超链被优先扩展的可能性就低了，被优先扩展的超链都是评价较高的网页。这样就提高了网络使用效率，减少了不必要的网络开销。④较小的实现代价。我们没有采用复杂的分类器，避免了采集样本，训练分类器的麻烦。另外，在计算网页重要度时，我们做了简化，尽管损失了一些精度，但是这样避免了定期的做全局的矩阵迭代计算，降低了实现难度。实验表明，经过简化的重要度评价效果亦不错。

参 考 文 献

1 Barfourosh A, et al. Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition:[Technical Report]. http://www.cs.umd.edu/Library/TRs/CS-TR-4291/CS-TR-4291.pdf

2 Bra D,Houben G,et al. Information Retrieval in Distributed Hypertexts. In: Proc. of the 4th RIAO Conf,1994. 481～491

3 Hersovici M,Jacovi M, et al. The Shark-Search Algorithm - An Application: Tailored Web Site Mapping. In: Proc. of 7th Intl. World Wide Web Conf. 1998

4 Chakrabarti S, Van Der Berg M, Dom B. Focused Crawling: A New Approach to Topic-specific Resource Discovery. In: Proc. of the 8th Intl. World Wide Web Conf. 1999

5 McCallium A,et al. Building Domain-specific Search Engines with Machine Learning Techniques. In: AAAI-99 Spring Sym-posium on Intelligent Agents in Cyberspace, 1999

6 Kleinberg J, Authoritative sources in a hyperlinked enviro-nment. In: Proc. of 9th ACM-SIAM Symposium on Discrete Algorithms, 1998

7 Page L,Brin S,Motwani R,Winograd T. The PageRank citation ranking: Bringing order to the web:[Technical Report]. Stanford University, Stanford, CA. 1998

8 Page L,Brin S, The Anotomy of a Large-scale Hypertext Web Search Engine. In: Proc. of the 7th Intl. World Wide Web Conf. 1998

9 Menczer F, Pant G, et al. Evaluating Topic-Driven Web Crawler. In: Proc. 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2001