面向网页结构特征的 Hopfield 算法①

李光敏，陈年生，许新山

(湖北师范学院 计算机科学与技术学院，黄石 435000)

摘 要：针对目前互联网信息资源广泛、网页结构复杂、噪音信息较多的现状，主题爬虫获取有效信息过程中精确度低、耗费时间长等问题。结合经典的 Hopfield 算法，提出了针对网页结构特征进行分块的主题爬行改进算法，实验证明该改进算法在一定程度上能有效地解决目前信息获取过程中所面临的问题。

关键词：垂直搜索；网页分块；主题爬行；相关度计算

Hopfield Algorithm Orienting for Web Page Structure-Feature

LI Guang-Min, CHEN Nian-Sheng, XU Xin-Shan

(College of Computer Science and Technology, Hubei Normal University, Huangshi 435000, China)

Abstract: As the Web continues to grow, it has become increasingly obvious that information overload and is terribly noisy. In this paper, to address such issues as low precision, much time-consumption, we present an improved Hopfield algorithm orienting web page structure feature. The experimental results show that the proposed approach is practical.

Key words: vertical search; page segmentation; topic crawling; relevance computation

1 引言

Internet 的出现使得互联网的信息容量按指数规律飞速增长,对于学术研究方面的信息增长同样也不例外。目前通用搜索引擎很难满足学术研究人员对特定领域范围内高质量、个性化、即时化信息检索的需求，但垂直搜索引擎能很好的解决这一窘境，它对抓取的信息进行分析、挖掘、筛选，精准的定位，从而确保了学术研究人员能够迅速准确地了解最新学术研究动态、分享交流研究经验。为了保证垂直搜索引擎信息资源的充足有效，而主题爬虫的高效爬取方式则担当其重要角色。

2 相关工作

早期主题爬虫的爬行策略主要分基于网页链接分析和基于网页内容分析，前者经典的算法有 Pinkerton[1]提出的宽度优先、De Bra[2] 和 Post 提出的深度优先、De Bra 提出的 Fish-Search 算法、Hersovici[3] 等人在此基础上提出的 Shark-Search 算法。后者主要包括有 Chakrabarti[4]等人提出的分层主题分类的方式来选择待爬行的链接、Aggarwal[5]等人建立的统计分析主题特征的学习模型来供爬行。这两种爬行策略在 Web 信息获取的召回率和精确度上都有很好的效果，但也存在一些不足之处，即它们都是对整个网页内容和链接综合评价，这样导致有大量不相关的内容和链接会被优先分析爬行[6]。同时 Michael Chau 在文献[7]中对比分析的 Hopfield 爬行算法要明显优于传统的网络爬虫算法。因此本文提出的专门针对在线期刊论文的网页结构特征且结合应用 Hopfield 算法的爬行策略避免了对整个网页进行评价的粒度过粗问题，又在保证信息的召回率和精确度的同时，提高爬行效率。

针对网页结构特征进行分块的算法，许多学者[8,9]致力于从 HTML DOM 树中提取出结构化信息，但由于 HTML 语法的不灵活性和网页结构的不规范性导致 DOM 树结构错误，更重要的是 DOM 树无法准确描述

① 基金项目:湖北师范学院教研资助项目(2009051)
收稿时间:2010-10-27;收到修改稿时间:2010-12-20

Web 页面的语义结构[9]。同时 Jie Zou[10]等人通过隐马尔可夫模型和 viterbi 算法构造区域树的方法对网页结构进行分块在实验中取得了不错的效果。其中 Deng Cai[11] 等人提出基于视觉的分块方法（VIPS，vision-based page segmentation）在 Web 页分块方面效果显著，该方法充分利用 Web 页面的视觉提示结合 DOM 树来重组 Web 页的语义结构。本文的分块算法正是基于这一思想。

3　页面分块技术及相关度计算

3.1 页面分块

针对在线的期刊论文网页结构布局进行深入分析，发现与爬行主题相关的内容和链接总是成块出现。以全球最大的学术期刊数据库 ScienceDirect（www.ScienceDirect.com）为例，如图 1 所示，页面的左中上部显示的是与论文相关元数据信息，右下部显示的是与该论文主题相关的其他文献链接，并且以文字标识开始相关链接。其他的噪音内容和链接（如：广告、导航链接等）则与爬行的主题毫无相关。基于如上页面结构规律假定页面由内容块 CB 和链接块 LB 构成，其中内容块(CB)包括论文标题块(CBt)、作者姓名块（CBau）、工作单位块（CBaf）、论文摘要块（CBab）、关键词块（CBk）；链接块（LB）包括导航块（LBn）、噪音块（LBs）、相关链接块（LBr）。为了计算页面内容块和链接块与爬取主题的相关度，本文采用信息分类效果较好的向量空间模型（VSM）中的向量夹角余弦相似度的计算方法，主要从两个方面入手。

图 1　网页分块示例

为了叙述方便，给出如下符号定义：预先指定主题 T（通过关键词或自然语言描述某学科领域），Uc 表示当前处理的 URL，Pc 表示 Uc 指向的已爬取页面。Ut 表示 Pc 中出现的 URL 链接，其指向待爬取的页面 Pt。主题（包括内容和结构）的相关度计算就是根据 Pc 中的已知信息来预测 Pt 与主题的相关度。经 VIPS 算法分块之后，Pc={ CBt ,CBau ,CBaf,CBab, CBk,LBn, LBs,LBr}，其中非期刊的页面不包括 CBab，CBk 这样的特征块，我们为这些能较好区分期刊论文结构的特征块赋予较高的权值，如果最后的结果超过设定的阈值，就表明待爬取的页面信息符合在线期刊论文特征其 Ut 可加入待爬行队列。

3.2 页面内容块相关度计算

内容块相关度的计算主要通过判断内容块中各子特征块 CBt ,CBau ,CBaf,CBab 和 CBk 与特定主题的相关程度，如公式（1）所示：

$$R_{CB}=\sum_{k=1}^{m}\frac{v_{cb^k}\cdot v_T}{\left|v_{cb^k}\right|\times\left|v_T\right|} \tag{1}$$

其中，m 表示内容块中的子特征块个数，表示第 k 个子特征块的向量空间模型，表示主题 T 的向量空间模型。

3.3 页面链接块相关度计算

页面链接块的计算基于如下前提：导航块(LBn)中的链接上一个（Up）、下一个（Un），一般是与当前页面主题相关的网页链接；噪音块(LBs)中的链接多是广告和图片链接与主题毫不相干；相关文献链接块(LBr)中的链接所指向的页面都是与当前页面的主题相关，实质上锚点的文本内容就是 Pt 中 CBt 的内容，更能反映其主题的相关度。其计算如公式（2）所示：

$$R_{LB}=\begin{cases}\dfrac{v_{LB_n}\cdot v_T}{\left|v_{LB_n}\right|\times\left|v_T\right|}, & \mathrm{U}_t\in LB_n,\ \mathrm{U}_t=\mathrm{U}_n\\ 0, & \mathrm{U}_t\in LB_s\\ \dfrac{v_{LB_r}\cdot v_T}{\left|v_{LB_r}\right|\times\left|v_T\right|}, & \mathrm{U}_t\in LB_r\end{cases} \tag{2}$$

其中，v_{LB_n}、v_{LB_r}、v_T 分别表示导航块、相关链接块和主题 T 的向量空间模型。通过 R_{LB} 值不仅能提高链接块中链接页面与主题相关度，而且还能滤掉与主题无关的噪音块中链接。

3.4 内容块和链接块综合计算

由于同一页面中的内容块和链接块之间也存在相对重要性，为了准确起见，特定义公式(3)：

$$R=\omega_{CB}\times R_{CB}+\omega_{LB}\times R_{LB} \tag{3}$$

其中，ω_{CB} 和 ω_{LB} 是权重因子，$\omega_{CB}+\omega_{LB}$=1，分别代表内容块和链接块之间的相对重要性程度，根据实验统计结果发现 ω_{CB}=0.7，ω_{LB}=0.3 时，爬取的页面信息与主题最为相关，其实也不难发现期刊数据库 ScienceDirect 中期刊论文信息多于图书信息，同时相关链接块中的部分链接指向的是图书信息页面，故页面块的权重值相对链接块的权重值要大。

4 面向网页结构特征的Hopfield算法

Hopfield Net Spider 是将 Web 看作是一个加权单层神经网络，然后利用包含激活扩散算法的爬虫来发现和检索信息[7]。

a.初始化。由专家选择一批权威的种子 URL 作为 0 迭代层的输入节点，并赋予权值为 1，Hopfield Net 爬虫爬取分析 0 迭代层的 Web 页内容。?i(t)表示 t 层迭代的 i 节点权值，因此，该层所有节点的权值可表示为?i(0)=1。每当爬虫从种子网页中找到新的 URL 就把它们作为节点加入到 Hopfield Net 中。

b.激活。继续处理下一迭代层，并计算该层所有节点的权值：

$$\mu_i(t+1)=f_s\left(\sum\omega_{b,i}\mu_b(t)\right) \tag{4}$$

其中，$\omega_{b,i}$ 表示节点 b 和 i 之间的链接权重值，可通过锚点文本的内容与学科主题的相关度来衡量。f_s 是一个规范化转换函数，确保节点的权重值在 0 和 1 之间，其形式如公式(5)所示：

$$f_s(x)=\left(\frac{1}{1+e^{-x}}-0.5\right)\times 2 \tag{5}$$

当爬虫计算出当前迭代层所用节点（URL 链接）的权值并按降序依次加入待爬行队列后，它开始激活权值高于某一设定阈值 θ 的 URL 节点，这样提高了爬行的效率并使目的性更强。

c.权值调整。当主题爬虫访问并下载权值高于设定阈值的 Web 页后，它根据已爬取页面的质量和它与主题的相关度来更新迭代层中各节点的权重值，其计算如公式(6)所示：

$$\mu_i(t+1)=f_s\left(\mu_i(t+1)\times R_i\right) \tag{6}$$

其中，R_i 表示页面 i 与主题相关度，其计算值由公式（3）所得，而常规 Hopfield 算法中的 R_i 计算由页面中各特征词的权值决定。

d.收敛条件。重复上述激活过程直到爬取完预先指定的页面数或者迭代层中所有节点的权值平均数小于某一阈值（该值通常很小）时为止。

5 实验结果与性能分析

我们以数据库 ScienceDirect 中指定 10000 篇关于数据挖掘（Data Mining）方向的期刊论文为训练样本。初始的爬行入口为该在线期刊的 Computer Science 链接，由于该数据库期刊信息量庞大，关于数据挖掘方向的论文具体数目无法准确统计，故我们从精确度（Precision）和爬行时间(Time)来评价性能，实验中我们将主题内容满足数据挖掘方向并且结构也符合期刊论文结构的网页定义为符合要求的网页，精确度的计算公式(7)所示：

精确度(Precision)=符合要求页面数/总爬行页面数 (7)

表 1 BFS、HF 和 HFB 算法的爬行性能比较

	总爬行页面数	符合要求页面数	精确度	耗费时间（分钟）
BFS 算法	10000	2272	23%	3.1
Hopfield 算法	10000	2617	26%	2.9
Hopfield Block 算法	10000	3758	38%	1.8

由表 1 统计结果可知在面向网页结构特征的 Hopfield 算法在精确度和耗费时间上较之以前都有所提升，在实验中还发现采用 Hopfield 算法爬取符合主题内容的页面数实际是 3915 张，但是部分页面来自图书系列不符合期刊论文的主题结构特征(BFS 算法中也同样存在此问题)，符合主题内容且是期刊论文的页面是 2617 张。同时面向网页结构特征的 Hopfield 算法耗费时间明显减少，因为在爬行过程中无需下载非期刊论文，但是对网页结构的分析需花费一定时间，如何提高网页的分块速度这是后期项目改进的方向。

6 结语

本文针对专业项目需求，在分析网页结构特征的基础上采用基于视觉网页分块 VIPS 算法计算爬行主题相关度，最后结合神经网络的 Hopfield 算法实现主题网页的爬取，通过精确度和爬行时间两项衡量指标实验证明该方法的有效性。

参考文献

1 Pinkerton B. Finding what people want: Experiences with the WebCrawler. Proc. 1st International World Wide Web Conference (Geneva). 1994.

(下转第 232 页)

```
public int getr()
{
return r;
}          //返回当前圆的半径，根据每一个实例的不同，返回不同的值。所以此时不能使用静态方法，只能使用实例方法
}
```

5 结语

变量和方法是 java 语言中重要的概念，恰当的使用变量和方法可以保证程序的简洁和高效。本文详细讨论了变量和方法的分类及其使用情况，并给出了部分实例。希望本文的研究对于 java 程序设计人员准确使用 java 语言中变量和方法具有一定的帮助。

参考文献

1 李占波,姬莉霞,王海玲,欧研.程序设计基础(java 版).北京:中国铁道出版社, 2007.80–81.

2 CAD 教育网.Java 中的方法和变量在继承时的覆盖.北京:http://www.cadedu.com/编程开发.java 语言编程, 2010.02.06.

3 李尊朝,苏军.Java 语言程序设计.第 2 版.北京:中国铁道出版社, 2009.97–98.

4 刘培文.java 程序设计教程.北京:北京科海电子出版社, 2009.88–89.

5 张素珍,耿磊.Java 语言静态变量和静态方法的分析及其应用研究.计算机系统应用, 2006,15(5):84–86.

(上接第 227 页)

2 De Bra PME, Post RDJ. Information retrieval in the World Wide Web: Making client-based searching feasible. Proc. 1st International World Wide Web Conference (Geneva), 1994.

3 Hersovici M, Jacovi M, Maarek YS, et al. The shark-search algorithm An application: Tailored Web site mapping. Proc. 7th Intl. World-Wide Web Conference, 1998.

4 Chakrabarti S, van den Berg M, Dom B. Focused crawling: A new approach to topic-specific Web resource discovery. Computer Networks 1999,31:1623– 1640.

5 Aggarwal CC, Al-Garawi F, Yu PS. Intelligent crawling on the World Wide Web with arbitrary predicates. Proc. 10th International World Wide Web Conference.2001.96–105.

6 Michelangelo Diligenti, Frans Coetzee, Steve Lawrence et al. Focused crawling using context graphs. Proc. Very Large Data Bases 2000 (VLDB 2000).

7 Michael Chau, Hsinchun Chen . Comparison of three vertical search spiders. IEEE Computer, 2003,36(5).

8 Chakrabarti S. Integrating the Document Object Model with hyperlinks for enhanced topic distillation and information extraction. 10th International World Wide Web Conference, 2001.

9 Chen J, Zhou BY, Shi J, et al. Function-Based Object Model Towards Website Adaptation. Proc. of the 10th International World Wide Web Conference, 2001.

10 Zou J, Le D, Thoma GR. Online medical journal article layout analysi. Proc. SPIE, 2007, 6500: 1–12.

11 Cai D, Yu SP, Wen JR, et al. VIPS: A Vision-based Page Segmentation Algorithm. Microsoft Technical Report, 2003, MSR-TR-2003–79.