基于 Wang−Landau 抽样的主题爬虫方法

刘景发[1]，陈靖岚[1]，赵　鹏[2*]

(1. 广东外语外贸大学信息科学与技术学院　广州　510006；2. 南京信息工程大学计算机与软件学院　南京　210044)

【摘要】针对传统爬虫方法存在搜索易陷入局部最优，且很少考虑结合历史爬行经验对爬行路径进行修正的缺陷，提出一种基于 WL 抽样的主题爬行方法。该方法分别使用向量空间模型 (VSM) 和 PageRank 算法对链接的相关性和重要性进行评价，采用区域竞争策略从具有主题相关或潜在价值的链接集合中选出目标链接。基于概率密度函数，WL 抽样算法对侯选集中选出的目标链接进行抽样判断，根据历史统计经验指导爬虫的后续爬行，从而优化搜索路径。实验结果表明，提出的基于 WL 抽样的主题爬虫方法比其他主题爬虫方法能搜索到更多主题相关的网页，其爬准率和所有下载网页主题相关度的标准差具有明显优势。

关　键　词　网络爬虫；　信息检索；　暴雨灾害；　台风灾害；　Wang-Landau 抽样

中图分类号　TP391　　**文献标志码**　A　　doi:10.12178/1001-0548.2022183

Focused Crawler Method Based on Wang−Landau Sampling

LIU Jingfa[1], CHEN Jinglan[1], and ZHAO Peng[2*]

(1. School of Information Science & Technology, Guangdong University of Foreign Studies　Guangzhou　510006;

2. School of Computer & Software, Nanjing University of Information Science & Technology　Nanjing　210044)

Abstract　Aiming at the problem that the traditional crawler methods are easy to fall into local optima of the search and rarely consider modifying the crawling path based on historical crawling experience, a focused crawler method based on Wang−Landau (WL) sampling is proposed. This method uses the vector space model (VSM) and PageRank algorithm to evaluate the relevance and importance of links, respectively. Regional competition strategy is used to select the target link from the link set containing the topic−related links and links with potential value. Based on probability density function, the WL algorithm is used to sample the selected target links in the set, and guides the subsequent crawling of the crawler according to the historical statistical experience, so as to optimize the search path. The experimental results show that the WL-based focused crawling method can search more topic-relevant webpages than other methods in the literature, and the climbing accuracy and standard deviation of topic relevance of all downloaded pages are also significantly improved.

Key words　focused crawler; information retrieval; rainstorm disaster; typhoon disaster; Wang-Landau sampling

当今，有效信息变得越来越重要，而网络爬虫正是穿梭于互联网中帮助人们自动获取信息的一把利器。据中国互联网络信息中心第 49 次调查报告显示[1]，截至 2021 年 12 月，仅中国网页的数量就已经达到 3 350 亿个，全球互联网的网页数量更是难以估计，这使得传统搜索引擎很难满足特定用户的需求。因此，如何利用主题爬虫[2] 抓取特定领域的数据，从而为特定人群提供个性化推荐已经成为许多学者研究的热点。

主题爬虫是一种旨在搜集某一领域数据的网络爬虫，它将使用某种策略和网页分析算法来降低传统网络爬虫因大量网页遍历而带来的代价，从而更有效地获取主题相关的信息。目前，主题爬虫的研究主要集中在链接的评价和搜索策略上。链接评价是指对链接相关性和重要性的评价。链接评价方法主要有基于网页内容分析的方法和基于链接结构分析的方法。前者的代表主要有 Fish Search[3] 和 Shark Search[4]；后者的代表主要有 PageRank 算法[5] 与

收稿日期：2022－06－13；修回日期：2023－01－30

基金项目：广东省基础与应用基础研究 (2021A1515011974, 2023A1515011344)

作者简介：刘景发 (1972－)，男，教授，博士，主要从事信息检索、计算智能、多目标优化等方面的研究.

*通信作者：赵鹏，E-mail：zhaopeng_159@163.com

HITS(hyperlink−induced topic search)[6] 算法。由于基于网页内容分析的主题爬虫方法仅考虑网页、锚文本等文本信息，缺乏对网页链接结构的考虑，所以容易导致爬虫陷入局部最优。基于链接结构分析的主题爬虫应用较为广泛，近年来一些学者进行了深入研究。文献 [7] 通过在 Shark Search 算法的基础上引入相关链接块的概念来解决链接的锚文本较为短小、不能准确表明链接指向主题相关性的缺陷。文献 [8] 以主机目标网站中若干相同目录路径中的页面聚合相关性作为领域相关性特征，提出一种拥有“领域功能”的主题爬虫方法。但该类方法由于未考虑链接的主题相关性，所以爬虫容易产生主题漂移现象[6]。鉴于此，一些学者尝试综合这两类方法来提高链接评价的准确性[9]。另外，搜索策略决定不同价值链接被访问的次序。传统的纯粹基于链接相关性高低的爬行策略难以使爬虫通过具有潜在价值的链接进行网页有效搜索。为提升主题爬虫的隧道穿越能力[10]，引导主题爬虫以最小的代价从相关度低的区域跳转到相关度高的区域，文献 [11-14] 采用智能优化方法来改进爬虫的全局寻优能力。文献 [11] 为提取和揭示网络犯罪，使用基于蚁群优化的主题爬虫，为 Web 犯罪挖掘开发一种基于本体的优化方法。文献 [12] 将多目标优化算法引入主题爬虫，提出了一种基于网页空间进化 (WSE) 算法的主题爬虫方法。文献 [13] 进一步将 WSE 算法和领域本体相结合，提出了一种新的基于 WSE 的主题爬虫算法 (focused crawler combining Web space evolution and ontology, FCWSEO)。在该算法中，通过构建领域本体建立主题模型，使用一种最近最远候选解法挑选 Pareto 最优链接。文献 [14] 为克服人工构建本体过程中构建者知识储备和主观意识的限制，提出了一种结合潜狄利克雷分布和 Apriori 算法的领域本体半自动构建方法。他们基于链接评估的多目标优化模型和改进的多目标蚁群优化算法来指导爬虫的搜索方向。文献 [15] 通过加入时间因子，动态调节适应度函数和遗传算子，在一定程度上避免了爬虫陷入循环爬行的陷阱。文献 [16] 结合爬虫记忆历史主机信息和模拟退火算法设计了暴雨灾害主题爬虫。文献 [17] 应用多目标蚁群算法来选择主题爬虫过程中的 Pareto 最优链接，进而引导爬虫搜索更多主题相关的页面。本质上，基于智能优化算法的主题爬虫以一定规则或概率接收主题相关度低的页面，具有全局优化的优点。但上述算法在链接搜索次序选择上缺乏对已爬取网页的统计，导致历史经验难以指导后续的网页搜索。特别是当爬虫距离主题相关页面较远时，常常出现主题偏移现象。

针对上述问题，本文提出一种基于 Wang-Landau(WL) 抽样的主题爬虫方法。该方法采用一组带有语义权重的向量来描述主题，并使用向量空间模型 (vector space model, VSM)[18] 分别计算链接对应的锚文本和链接指向网页文本的主题相关度；再利用 PageRank 算法从链接结构角度评价链接的重要性，进而选定具有潜在价值的链接。基于 WL 算法，将能量状态密度函数和直方图函数应用于主题爬行策略，通过统计状态密度的变化，决定主题爬虫是否接受目标链接，进而指导爬虫寻找一条合适的主题爬行路径。此外，为防止爬虫接受无关链接而导致主题搜索方向偏移，引入区域竞争策略挑选目标链接，以获得更好的爬行效果。

1　网页主题相关度和重要性计算

在主题爬虫过程中，本文综合考虑链接指向网页的主题相关度和链接锚文本主题相关度，并根据链接所在页面的重要性选择具有潜在价值的页面。

1.1　网页主题相关度的计算

在主题爬虫中，一般采用一组特定术语组成的主题词向量 **TK** 来描述用户选定的特定主题，假设 **TK** 定义如下：

$$\mathbf{TK} = \mathrm{tk}_1, \mathrm{tk}_2, \cdots, \mathrm{tk}_n \tag{1}$$

式中，$\mathrm{tk}_i(i=1,2,\cdots,n)$ 表示第 i 个主题词；n 为主题词集 **TK** 中元素的个数。本文以暴雨灾害和台风灾害为主题，根据领域专家经验和多次对比实验，设置描述暴雨灾害的主题词组分别为暴雨、气象、天气、灾害和预防，描述台风灾害的主题词组分别为台风、风暴潮、暴雨、气压和风力。令 $w_{\mathrm{tk}i}$ 为主题词 tk_i 的权重，其值主要基于特征提取的主题词权重计算方法和领域专家结合经验设定[12]。在主题词向量确定以后，为计算网页主题相关度，需要确定网页文本特征向量。本文采用 IK-Analyzer 对网页文本进行分词、去停用词等预处理。将处理之后的网页文本 P_j 用向量 $\mathbf{DK}_j$ 表示：

$$\mathbf{DK}_j = \mathrm{dk}_{j,1,}, \mathrm{dk}_{j,2}, \cdots, \mathrm{dk}_{j,i}, \cdots, \mathrm{dk}_{j,n} \tag{2}$$

式中，$\mathrm{dk}_{j,i}$ 表示网页文本 P_j 中的第 i 个特征词。对于文本中的某个特征词，它的权重采用词频−逆文档频率 (term frequency−inverse document frequency, TF-IDF) [19] 方法计算：

$$w_{\mathrm{dk}_{j,i}} = \frac{n_{j,i}}{\sum_k n_{j,k}} \lg \frac{D}{1+D_i} \tag{3}$$

式中，$n_{j,i}$ 表示在网页文本 P_j 中第 i 个特征词出现的次数；分母表示文本 P_j 中所有特征词出现的次数之和；D 为当前爬取网页文本总数；D_i 为包含第 i 个特征词的网页文本数。某个词的权重越大，说明该词越能代表文档的特征。当词权重为 0 时，说明该词不出现于该网页。

对于链接 l 所在网页 P_l，本文采用 VSM 计算主题向量 **TK** 和网页文本特征向量 $\mathbf{DK}_l$ 的主题相关度 $R(P_l)$：

$$R(P_l) = \mathrm{sim}(\mathbf{TK}, \mathbf{DK}_l) = \cos < \boldsymbol{W}_{\mathbf{TK}}, \boldsymbol{W}_{\mathbf{DK}_l} >= \frac{\sum_{i=1}^{n} w_{\mathrm{tk}_i} w_{\mathrm{dk}_{l,i}}}{\sqrt{\sum_{i=1}^{n} w_{\mathrm{tk}_i}^2 \sum_{i=1}^{n} w_{\mathrm{dk}_{l,i}}^2}} \tag{4}$$

式中，$\boldsymbol{W}_{\mathbf{TK}}$ 为用户所指定主题词的权重向量；$\boldsymbol{W}_{\mathbf{DK}_l}$ 为网页文本 P_l(或锚文本) 的特征权重向量。本文由式 (4) 分别计算待访问链接 l 所指向网页的主题相关度和其锚文本的主题相关度，再进行线性加权求和得到链接 l 的综合主题相关度：

$$R(l) = \partial R(\mathrm{anchor}_l) + (1-\partial) R(P_l) \tag{5}$$

式中，$R(l)$ 表示链接 l 的综合主题相关度；$R(\mathrm{anchor}_l)$ 和 $R(P_l)$ 分别为链接 l 的锚文本 anchor_l 和链接 l 指向网页 P_l 的主题相关度；∂为权重系数。

由于基于网页文本内容和链接结构分析的主题爬虫方法在各自单独使用时其爬准率往往受限，因此通常相互结合使用，实现互补，进而提高爬虫抓取网页的准确率。此外，上述策略在计算链接优先度后，大多采取最佳优先机制进行网页抓取，该贪心策略往往容易使爬虫陷入局部最优。

1.2 网页重要性计算

从有潜在价值的网页出发，抓取更多主题相关页面，还可以从重要性的角度综合评价网页。一般地，网页的重要性主要由网页链接关系来衡量，本文采用 PageRank 算法来计算网页的重要性。PageRank 算法的思想是通过网页的引用关系来判断网页是否重要，即若某网页被其他网页引用次数越多，就认为该网页越重要，而当某网页被这些重要的网页引用时，该网页也被认为是重要的。本文采用的 PageRank 计算公式为：

$$\mathrm{PR}(P_j) = (1-\alpha) + \alpha \sum_{P_i \in B_j} \frac{\mathrm{PR}(P_i)}{|C_i|} \tag{6}$$

式中，$\mathrm{PR}(P_j)$ 代表网页 P_j 的 PR 值；B_j 为网页 P_j 的所有入链页面集；C_i 为 P_j 的入链页面 P_i 的出链集合；α= 0.85 为跳转因子，主要用于表示 Web 用户有多大概率通过其他网页的出链来访问当前页面，1−α表示用户通过键入 URL 等方式来直接访问该网页的概率。通过式 (6) 可给网页赋予一个重要性的量化值。本文将从主题相关度较低的网页中选取重要性值 PR≥2 的网页作为具有潜在价值的网页，这些网页将有机会在爬行过程中被选中，从而辅助爬虫跳出局部最优，以搜寻更多主题相关的页面。

2 基于 Wang−Landau 抽样的主题爬虫方法

传统主题爬虫多采用按主题相关度从高到低依次选择链接进行网页抓取。然而，并不是所有的相关页面都聚集在一起，当爬虫始终都抓取相关度高的网页时，容易陷入搜索的局部最优。因此，为实现全局寻优，结合网页主题相关性和重要性评价指标，提出基于 Wang−Landau 抽样的主题爬虫过程。

2.1 Wang−Landau 抽样算法的基本思想

基于频率直方图的 Wang−Landau(WL) 抽样算法[20] 是一种蒙特卡洛 (Monte Carlo, MC) 算法，其理论基础是：状态 X 被访问的概率与其能量状态密度 $g(E(X))$ 的倒数成正比，其中 $E(X)$ 表示状态空间中状态 X 的能量。传统的 MC 方法局限于在温度不变的前提下产生正则分布，因此，想获得高质量的解，算法就必须在不同温度条件下多次运行。如模拟退火算法就需要不断地进行退火操作，十分耗时。Wang−Landau 抽样算法与温度无关，这使得它可以按照一定的概率接受准则漫步于状态空间，进而获取一个平坦的能量直方图 $H(X)$ 并得到概率密度函数 $g(E(X))$，即 WL 抽样算法可从系统中的某个状态出发，逐步统计各个状态的能量，进而发掘整个系统中各个状态能量的分布规律。这与主题爬虫面临的问题场景类似，在网页空间中，各主题相关页面 (对应状态空间中的状态) 分布未知，人们无法通过互联网的结构去直接设计网页抓取策略。因此，本文将整个互联网中的 Web 资源看成是系统的状态空间 S'，主题相关页面分布于 S'中。为了在这个系统中搜寻到更多的主题相关页面，在

计算各链接的主题相关度后，通过将上文给出的链接主题相关度 $R(l)$ 和状态 (链接) 的能量 $E(l)$ 相对应，设计基于 WL 抽样算法的主题搜索策略，即 WL 抽样算法基于状态密度函数行走于网页空间，并根据历史经验的积累实现对低状态密度函数值链接的有效抽样。

2.2　基于 WL 抽样的主题爬虫方法实现

在 WL 抽样算法的开始，网页空间中所有链接 X 及其状态密度函数 $g(E(X))$ 均是未知的。每当访问到一个新的链接 X，这个链接的状态密度函数 $g(E(X))$ 及直方图函数 $H(E(X))$ 就被置为 1。由于本文计算的能量值为链接的主题相关度，所以能量值介于 0～1 之间。不妨将其划分为 50 个能量区间，每个能量区间间隔为 0.02，即 [0, 0.02), [0.02, 0.04),⋯, [0.98, 1]。为便于描述，将能量区间 $[\lfloor E(X)\rfloor, \lceil E(X)\rceil)$ 标记为 $[E(X)]$，其中 $\lfloor E(X)\rfloor$ 是 $E(X)$ 所属能量区间的下边界，$\lceil E(X)\rceil$ 是 $E(X)$ 所属能量区间的上边界，X 为网页空间中某一链接。如 $E(X)=0.621$ 处于能量区间 [0.62, 0.64) 中，记为 [0.621]。将 Wang−Landau 抽样算法应用于主题爬虫，其基本思路如下：首先，从种子链接等待队列中选择一个初始链接 X_1，然后采用区域竞争策略 (见 2.3 节) 产生目标链接 X_2，再根据 $P(X_1\to X_2)=\min(g(E(X_1))/g(E(X_2)), 1)=\min(e^{\ln[g(E1)-g(E2)]}, 1)$ 的值来判断 X_2 是否接受，这里 E_1 和 E_2 分别表示 X_1 和 X_2 的综合主题相关度，即 $E_1=E(X_1)$，$E_2=E(X_2)$。如果 X_2 被接受，则修改 X_2 的状态密度函数 $g(E(X_2))$：令 $g(E(X_2))=f_i g(E(X_2))$，其中，$f_i>1$ 为一个罚因子，同时修改其直方图函数：令 $H([E(X_2)])=H([E(X_2)])+1$；如果链接 X_2 不被接受，则类似修改 X_1 的状态密度函数和直方图函数值。另外，算法的收敛速度由直方图的平坦度控制，而所谓平坦的直方图是指所有 $H(E(X))$ 的值均不小于直方图的平均值 $\langle H([E(X)])\rangle$ 的 $k(0<k<1)$ 倍，k 由系统的复杂度和 $g(E(X))$ 的准确度决定。在实际操作中要获得一个绝对平坦的直方图非常困难。因此，每 1 000 步 MC 迭代后，算法将检查直方图是否近似平坦。若 $H(E)$ 变得近似平坦，即若对于已访问过的能量区间集合 S 中所有 $[E]$，都有 $H([E])\geqslant kH([E(X)])$，则调整修正因子为一个更小的值，即令 $f_{i+1}=\sqrt{f_i}$，$i=i+1$，若迭代步数 $L>10^6$，则提前结束爬虫搜索; 否则，将所有访问过的能量区间 $[E(X)]$ 所对应的 $H([E(X)])$ 置为零，$g(E(X))$ 值保留，开始下一次随机行走。此时，$g(E(X))$ 将以某一精度收敛于真实值。因实际计算中 $g(E(X))$ 的值会变得很大，以至超出双精度数字所能表示的范围，故选用对数式 $\lg(E(X_i))=\ln(f_i)+\lg(E(X_i))$ 进行更新，这里 i 为 1 或 2。基于 WL 抽样的主题爬虫算法伪代码见算法 1。

算法 1　基于 WL 抽样的主题爬虫算法

输入：种子链接

输出：下载的网页

1) 将种子链接加入等待队列 Queue，初始化参数 $i=0$，$f_0=e$，$L=0$，$S=\varnothing$；令 $T=1\,000$，DP=0，LP=0；//DP 表示下载网页数，LP 表示下载相关网页数；

2) 从 Queue 中取出综合主题相关度最高的队头链接 X_1 作为当前链接；将 $[E_1]$ 加入已访问过的能量区间集合 S，即令 $S=S+\{[E_1]\}$，令直方图函数 $H([E_1])=1$，概率密度函数 $g(E_1)=1$；

3) If DP<15 000 && Queue !=∅ then

令 $m=0$，转至步骤 4)；//m 用于记录不接受目标链接的次数；

Else 计算爬准率等评价指标，算法结束；

End If

4) 采用区域竞争策略 (如 2.3 节) 产生目标链接 X_2 并计算其综合主题相关度 E_2，令 $L=L+1$；

5) If $[E_2]\notin S$ then

令 $g([E_2])=1$，$H([E_2])=1$，$S=S\cup\{[E_2]\}$，计算接受概率：

$P(X_1\to X_2)=\min(e^{\ln g(E_1)-\ln g(E_2)}, 1)$

End If

6) If random(0,1)< $P(X_1\to X_2)$ then

令 $X_1=X_2$，$E_1=E_2$，$\lg(E_2)=\ln f_i+\lg(E_2)$，

$H([E_2])=H([E_2])+1$; 转步骤 8)；

Else 令 $\lg(E_1)=\ln f_i+\lg(E_1)$, $H([E_1])=H([E_1])+1$，$m=m+1$；转步骤 7)；

End If

7) If $m\geqslant 5$ then

从 Queue 中取出综合相关度最大的链接作为新的队头链接 X_1；转步骤 8)；

Else 转步骤 4)；

End If

8) 将 X_1 移出 Queue，取出 X_1 指向页面的子链接中满足 $R(l)\geqslant 0.2$ 的链接 (去重)，将这些链接 (假设其个数为 k_1) 加入 Queue，令 DP=DP+k_1；

9) For $j=1$ to k_1 do

If $R(P_j)\geqslant\beta$ then //如果子链接 j 指向网页的主题相关度 $R(P_j)$ 大于或等于阈值 β

令 LP=LP+1；//保存该网页 P_j 到主题相关网页集中
End If
End For
10) If $L\%T==0$ then
If 任给 $[E]\in S$，都有 $H([E])\geqslant \ln 2/\ln f_i$ then
令 $f_{i+1}=\sqrt{f_i}$，$i=i+1$，转步骤 11)；
Else 转步骤 3)；
End If
End If
11) If $L<10^6$ then
对于 S 中所有 $[E]$，恢复 $H([E])=0$，而 $g([E])$ 值保持不变，转步骤 3)；
Else 算法提前结束；
End If

从上述算法不难看出，算法将以 $P(X_1\rightarrow X_2)$ 的概率接受目标链接 X_2，并以一定的概率接受当前最佳链接，这既避免了算法陷入局部最优，也防止了因跳坑次数太多而增加时间损耗。在将目标链接 X_2 指向页面的子链接加入 Queue 之前，需要进行去重操作。此外，为提升等待队列中链接的质量，算法将保留子链接中综合主题相关度 $R(l)\geqslant 0.2$ 的链接。目前，WL 抽样算法已经成功运用于数字水印中的参数优化[21]、统计物理方向中的相变[22]等问题，这表明 WL 算法具有适应于求解不同问题的能力。本文将 WL 抽样算法应用于主题爬虫，使爬行路径得到进一步优化。

2.3 应用区域竞争策略生成目标链接

在爬行过程中，爬虫需不断地获取新的链接，以便搜索更多主题相关页面。本文提出启发式的区域竞争策略以从链接等待队列中挑选出目标链接。该策略的主要思想为：将等待队列中的链接按链接指向的网页域名归属划分为不同的区域，并从竞争获胜区域中采用轮盘赌策略选定某链接作为目标链接。

区域竞争策略的具体步骤如下。

1) 按照域名将链接等待队列 Queue 中的链接划分为不同的区域。

2) 选取上述区域中具有最高平均主题相关度的区域 R_i (R_i 表示按主机信息划分后的第 i 个区域，$i=1,2,\cdots,u$) 作为竞争 (competition) 获胜区域 (winer)。

3) 采用轮盘赌选择 (roulette wheel selection, RWS)[23] 策略挑选目标链接，基本过程为：

①将获胜区域 (winer) 中所有链接 X_j ($j=1,2,\cdots,k$) 构成候选链接集合 C，显然，候选链接集合 C 中既包含主题相关链接，也包含具有潜在价值的链接 (重要性值 PR≥2 的链接)；②对候选链接集合 C 中的所有链接，根据式 (5) 采用轮盘赌，选出一个链接为目标链接 X_2。

本文提出的区域竞争策略与传统的最佳优先搜索 (OPS)[24] 最大的区别在于将链接等待队列按区域划分后再选择目标链接。这样算法既不会因仅接受主题相关度高的链接，从而导致爬虫易陷入局部最优；也不会因接受一些主题相关度很低的链接，从而导致搜索方向发生主题偏移。另外，该策略也保证 WL 算法能够在具有主题相关度高的区域进行集中搜索后，依旧能够跳出该区域进行分散搜索，进而提高主题爬虫抽样的多样性。

3 实验与结果

实验环境：处理器 Intel(R) 为 Core(TM)i7-7700 HQ，CPU 为 2.80 GHz，RAM 为 8 GB；开发工具为 MyEclipse+JDK1.7+Java 语言；分词工具为 IK-Analyzer；HTML 解析器为 Jsoup。

爬行主题：暴雨灾害和台风灾害。

实验数据：暴雨灾害的主题词组为暴雨、气象、天气、灾害和预防，权重依次设为 0.8、0.1、0.1、0.5 和 0.3；台风灾害的主题词组为台风、风暴潮、暴雨、气压和风力，权重依次设为 0.8、0.5、0.3、0.2 和 0.2；对暴雨灾害和台风灾害，使用百度进行主题词搜索，对得到的搜索结果结合人工经验分别筛选 18 个和 30 个 URL 作为初始种子链接。

实验参数：令调节因子 $\partial=0.3$；对暴雨灾害，设置主题相关度阈值 $\beta=0.62$，对台风灾害，设置主题相关度阈值 $\beta=0.67$；最大爬取网页数 DP=15 000，其余参数均已在前文给出。

上述这些实验数据和实验参数的设置对爬虫算法的效果有一定影响，其值是参考文献 [12, 14] 以及通过多次对比实验使爬虫算法达到最好结果而确定的。

3.1 评价指标

为评价爬虫方法的效果，通常使用爬准率 (Accuracy) 和爬全率 (Recall) 作为评价指标。因爬全率一般难以计算，本文选用爬准率作为评价主题爬虫效果的主要指标：

$$\text{Accuracy} = \text{LP}/\text{DP} \tag{7}$$

式中，LP 表示下载的主题相关网页数量；DP 表示下载的总网页数。本文还使用平均主题相关度 (average relevance, AR) 和标准差 (standard deviation, SD) 来测试爬虫的性能，包括所有下载网页的平均主题相关度 (average relevance of all downloaded pages, AR_{DP})、所有下载网页主题相关度的标准差 (standard deviation of topic relevance of all downloaded pages, SD_{DP})、所有下载主题相关网页的平均主题相关度 (average relevance of all downloaded topic-relevant pages, AR_{LP}) 和所有下载主题相关网页主题相关度的标准差 (standard deviation of topic relevance of all downloaded topic-relevant pages, SD_{LP}):

$$AR_{DP} = \frac{1}{DP}\sum_{l=1}^{DP} R(P_l) \tag{8}$$

$$SD_{DP} = \sqrt{\frac{1}{DP}\sum_{l=1}^{DP}(R(P_l) - AR_{DP})^2} \tag{9}$$

$$AR_{LP} = \frac{1}{LP}\sum_{l=1}^{LP} R(P_l) \tag{10}$$

$$SD_{LP} = \sqrt{\frac{1}{LP}\sum_{l=1}^{LP}(R(P_l) - AR_{LP})^2} \tag{11}$$

式中，$R(P_l)$ 表示网页 P_l 的主题相关度。

3.2　实验结果与分析

针对暴雨灾害主题，本文实现了基于 WL 抽样的主题爬虫算法，并与文献中的广度优先搜索 (breadth first search, BFS)[25]、最佳优先搜索 (best first search, OPS)[24]、多目标蚁群优化 (multi-objective ant colony optimization, MOACO)[17]、记忆历史主机信息退火爬虫算法 (focused crawler based on simulated annealing with host information, FCSA-host)[16] 及结合本体的网页空间进化主题爬虫算法 (focused crawler combining webpage space evolution and ontology, FCWSEO)[13] 进行了对比，如表 1 所示。评价指标包括爬准率 (Accuracy)、下载的主题相关网页数 (LP)、所有下载网页的平均主题相关度 (AR_{DP}) 及其标准差 (SD_{DP})、所有下载相关网页的平均主题相关度 (AR_{LP}) 及其标准差 (SD_{LP})，以及算法运行时间 (Time)。由于文献中 MOACO 和 FCWSEO 两个算法没有测试 AR_{LP} 和 SD_{LP} 两个评价指标，故在表 1 中未给出相应的结果。

由表 1 可知，在爬虫初始阶段 (DP=1 000)，因采用区域竞争策略对种子链接附近中的高度相关区域进行了访问，故 WL 方法在 Accuracy、LP 和 SD_{DP} 上均取得最优结果。OPS 算法因每次均接受相关度最高的链接，所以具有最高的 AR_{DP} 和 AR_{LP}，且在 Accuracy 和 LP 上的表现仅次于 WL 方法，而 FCSA−host 算法获得最优的 SD_{LP}。当 DP=5 000 时，FCSA−host 方法在 Accuracy、LP 和 SD_{LP} 这 3 个指标上取得最优结果，在 AR_{DP} 上 MOACO 方法对 OPS 算法实现反超。此后，OPS 获取相关页面的能力逐渐降低，并因贪心策略而陷入搜索局部最优。当 DP=10 000 时，WL 方法在 Accuracy 和 LP 两个指标上再次反超 FCSA−host，同时高于其余对比方法。此时，FCWSEO 方法结合本体的优势逐渐显现，在 AR_{DP} 指标上取得最优效果。随着爬行的深入，当 DP=15 000 时，WL 方法在 Accuracy、LP 和 SD_{DP} 这 3 项指标上均高于其他方法。这是因为，WL 方法的随机抽样能力使其始终能稳定地穿越一些潜在价值较高的链接并获取更多的主题相关网页。FCWSEO 算法通过最近最远候选解法选择一组 Pareto 最优链接来更新种子链接库中的链接，这使其在 Accuracy 和 LP 两项指标上超过了除 WL 外的所有方法，且在 AR_{DP} 上表现最优。MOACO 策略通过选择主题爬虫过程中的 Pareto 最优链接，从而使更多可能指向主题相关页面的链接被接受，这使得它在 AR_{DP} 和 SD_{DP} 指标上分别取得了仅次于 FCWSEO 和 WL 方法的结果。FCSA−host 方法给予了次优链接优先访问的机会，且具有历史记忆能力，这使得它的 6 项指标值均取得了较好的结果，且在 AR_{LP} 指标上取得最优结果。当 DP=15 000 时，本文 WL 方法的 Accuracy 值分别比 BFS、OPS、FCSA-host、MOACO 和 FCWSEO 的结果高出 335.09%、44.56%、17.72%、11.58% 及 6.1%。表 1 显示，本文所提方法在爬准率 Accuracy、下载相关网页的数量 LP 和下载网页主题相关度的标准差 SD_{DP} 这 3 个指标上效果最好。不过，在 AR_{DP}、AR_{LP} 和 SD_{LP} 指标上，WL 方法表现一般，这主要是因为 WL 方法采用修改的状态密度函数和直方图策略 (即每访问一次链接就给予该链接状态密度函数一个惩罚，并将其能量区间直方图函数+1) 引导爬虫访问态密度低的区域，有利于爬虫扩大搜索范围，抓取更多主题相关网页，但并不能确保每次都只接受主题相关度高的链

接。另外，在运行时间方面，WL 方法跟文献中的其他方法相比也具有明显优势或相当。

表 1 β=0.62 时，暴雨灾害主题下各种主题爬虫算法在不同爬行阶段的实验结果比较

DP	算法	Accuracy	LP	AR_{DP}	SD_{DP}	AR_{LP}	SD_{LP}	Time/h
1 000	BFS	0.320 0	320	0.497 7	0.229 0	0.755 1	0.078 0	—
	OPS	0.823 0	823	**0.714 3**	0.157 5	**0.772 2**	0.064 2	—
	FCSA-host	0.701 0	701	0.672 0	0.150 9	0.749 2	**0.047 8**	—
	MOACO	0.629 1	629	0.695 7	0.281 0	—	—	—
	FCWSEO	0.454 2	454	0.710 9	0.164 3	—	—	—
	WL	**0.852 0**	**852**	0.661 3	**0.071 6**	0.682 7	0.083 5	—
5 000	BFS	0.269 0	1 345	0.436 3	0.237 8	0.742 1	0.057 2	—
	OPS	0.770 4	3 852	0.680 2	0.161 0	**0.749 4**	0.058 6	—
	FCSA-host	**0.803 0**	**4 015**	0.695 3	0.119 9	0.742 2	**0.045 2**	—
	MOACO	0.704 2	3 521	**0.780 1**	0.186 5	—	—	—
	FCWSEO	0.657 2	3 286	0.759 0	0.176 7	—	—	—
	WL	0.779 6	3 898	0.677 9	**0.092 9**	0.679 1	0.080 2	—
10 000	BFS	0.134 5	1 345	0.296 6	0.222 6	0.742 1	**0.057 2**	—
	OPS	0.653 3	6 533	0.626 9	0.217 5	0.758 9	0.062 2	—
	FCSA-host	0.762 8	7 628	0.700 4	0.137 1	**0.764 0**	0.059 5	—
	MOACO	0.787 2	7 872	0.775 6	0.153 7	—	—	—
	FCWSEO	0.807 0	8 070	**0.790 4**	0.161 2	—	—	—
	WL	**0.850 4**	**8 504**	0.674 9	**0.082 1**	0.686 2	0.077 4	—
15 000	BFS	0.202 9	3 043	0.370 5	0.243 8	0.747 1	0.063 7	8.6
	OPS	0.610 7	9 160	0.616 9	0.233 9	0.765 0	**0.063 3**	9.0
	FCSA-host	0.749 9	11 248	0.700 3	0.147 1	**0.771 2**	0.066 5	11.8
	MOACO	0.791 2	11 868	0.778 0	0.137 7	—	—	15.8
	FCWSEO	0.832 2	12 483	**0.820 1**	0.155 8	—	—	12.2
	WL	**0.882 8**	**13 242**	0.691 7	**0.078 4**	0.712 2	0.082 0	**8.1**

为综合考察 6 种算法在暴雨灾害主题下的实验效果，表 2 给出了 6 种算法在 DP=15 000 时对 4 个典型指标 Accuracy、LP、AR_{DP} 和 SD_{DP} 应用 Friedman 测试[13]的结果，用 Ranks 值表示。其中，Ranks 值为 1 表示对应方法为该指标的最佳结果，6 为最差结果。从表 2 可看出，WL 方法除 AR_{DP} 指标排在第 4 位外，在 Accuracy、LP 和 SD_{DP} 指标上均排在第 1 位。综合来看，WL 方法具有最小的平均 Ranks 值 (Average=1.75)，是所有对比方法中性能最好的方法。BFS 在各阶段的 Ranks 值始终最高，故其性能最差。

表 2 β=0.62 时，暴雨灾害主题下各种主题爬虫算法的 Friedman ranks 比较

算法	Accuracy	LP	AR_{DP}	SD_{DP}	Average
BFS	6	6	6	6	6
OPS	5	5	5	5	5
FCSA−host	4	4	3	3	3.5
MOACO	3	3	2	2	2.5
FCWSEO	2	2	1	4	2.25
WL	1	1	4	1	**1.75**

为进一步测试 WL 方法的性能，在暴雨灾害主题下，设置网页主题相关度阈值 β=0.70，最大爬取网页数 DP=15 000，运行本文提出的基于 WL 抽样的主题爬虫算法，并给出各项评价指标 Accuracy、LP、AR_{DP}、SD_{DP}、AR_{LP}、SD_{LP} 以及运行时间 (Time) 的结果，与文献中的网页空间进化算法 (WSE)[12]、融合本体和改进禁忌搜索策略的主题爬虫方法 (On-ITS)[26]、基于全局与局部本体和模拟退火策略的主题爬虫算法 (FCOSA_LG)[27]、结合本体学习和多目标蚁群优化方法 (OLMOACO) 的主题爬虫方法[14]以及融合本体的网页空间进化主题爬虫算法 (FCWSEO)[13]的结果对比如表 3 所示。由于文献中 WSE、On-ITS、OLMOACO 和 FCWSEO 这 4 种算法没有测试 AR_{LP} 和 SD_{LP} 这两个评价指标，故在表 3 中未给出相应的结果。在 6 种基于智能优化算法的先进主题爬虫方法中，WL 方法在 Accuracy、LP 和 SD_{DP} 这 3 个度量指标上均获得最优结果；而 FCWSEO 在 AR_{DP} 指标上，FCOSA_LG 在 AR_{LP} 和 SD_{LP} 两个指标上得到最好结果。在运行时间上，本文 WL 方法比文献中的算法运行时间

明减少或相当，这主要是因为 WL 方法采用修改的状态密度函数和直方图函数策略更有利于算法爬取主题相关网页，同时 WL 方法不需要通过本体构造得到主题向量。

表 3　β=0.70 时，暴雨灾害主题下各种先进主题爬虫方法的实验结果比较

算法	Accuracy	LP	AR_{DP}	SD_{DP}	AR_{LP}	SD_{LP}	Time/h
WSE(2019)	0.733 4	11 002	0.729 0	0.160 0	—	—	12.2
On-ITS(2020)	0.734 0	11 010	0.729 5	0.161 9	—	—	13.2
FCOSA_LG(2022)	0.765 3	11 479	0.751 1	0.146 2	**0.809 5**	**0.058 1**	13.1
OLMOACO(2022)	0.741 7	11 126	0.778 1	0.137 5	—	—	16.0
FCWSEO(2022)	0.810 8	12 162	**0.821 9**	0.156 6	—	—	11.6
WL(本文)	**0.848 9**	**12 734**	0.682 2	**0.089 7**	0.700 2	0.089 1	**8.4**

为综合考察各种算法效果，表 4 给出了 6 种先进爬虫方法在 DP=15 000 时对 4 个典型指标 Accuracy、LP、AR_{DP} 和 SD_{DP} 应用 Friedman 测试[13]的结果。表 4 显示，WL 和 FCWSEO 在所有对比方法中性能最好，FCOSA_LG 和 OLMOACO 两个算法并列排在第三，WSE 算法排在最后。

表 4　当 β=0.70，时暴雨灾害主题下各种主题爬虫方法的 Friedman ranks 比较

算法	Accuracy	LP	AR_{DP}	SD_{DP}	Average
WSE	6	6	5	5	5.5
On-ITS	5	5	4	6	5.0
FCOSA_LG	3	3	3	3	3.0
OLMOACO	4	4	2	2	3.0
FCWSEO	2	2	1	4	**2.25**
WL	1	1	6	1	**2.25**

针对台风灾害主题，设置网页主题相关度阈值 β=0.67，本文实现了基于 WL 抽样的主题爬虫算法，并与文献中的 BFS[25]、OPS[24]、模拟退火算法 (simulated annealing, SA)[28]、网页空间进化算法 (WSE)[12]、结合本体和改进禁忌搜索策略的主题爬虫方法 (On-ITS)[26] 以及结合本体学习和多目标蚁群优化方法 (OLMOACO) [14] 进行了结果对比，评价指标包括 Accuracy、LP、AR_{DP}、SD_{DP}、AR_{LP}、SD_{LP} 和 Time，如表 5 所示。由于文献中其他算法没有测试 AR_{LP} 和 SD_{LP} 两个评价指标，故在表中未给出相应的结果。本文给出这两个指标的结果供将来其他算法进行比较。

由表 5 可知，当 DP=1 000 和 5 000 时，WL 方法在 Accuracy、LP 和 SD_{DP} 上均取得最优结果，在 AR_{DP} 上其结果略低于 On-ITS。由于 BFS 不对链接进行相关性判断，所以其爬准率始终都很低；而 OPS 因仅接受相关度最高的链接，易使爬虫陷入局部最优，爬准率逐渐降低；SA 根据一定的概率选择主题不相关的网页，从而避免了陷入局部最优，在 Accuracy、LP、AR_{DP}、SD_{DP} 这 4 个评价指标上，其结果均优于 BFS 和 OPS 方法，但由于该算法主题描述不够准确，其结果劣于其他几种算法。当 DP=10 000 时，WL 方法在 Accuracy、LP 上取得最优结果，在 AR_{DP} 上略低于 On-ITS 方法，在 SD_{DP} 上低于 OLMOACO 和 On-ITS 两种方法。当 DP=15 000 时，WL 方法在 Accuracy、LP 上的优势继续保持，最终爬准率稳定在 87% 左右，而 BFS、OPS、SA、WSE、On-ITS 和 OLMOACO 方法的爬准率分别为 4.7%、8.3%、35.8%、70.4%、82.0% 和 75.0%。在 AR_{DP} 指标上，On-ITS 的优势保持，而在 SD_{DP} 指标上，OLMOACO 的优势也继续保持。在运行时间上，当下载 15 000 个网页时，本文 WL 方法比 BFS、OPS、SA、WSE、On-ITS 和 OLMOACO 算法的运行时间明显减少或相当。为综合考察各种算法在台风灾害主题的爬虫效果，表 6 给出了 7 种方法在 DP=15 000 时应用 Friedman 测试[13]4 个典型指标的结果。表 6 显示，WL 方法在 Accuracy、LP 指标上均排在第一位，在 AR_{DP} 和 SD_{DP} 指标上均排在第二位。综合来看，WL 方法具有最小的平均 Ranks 值 (Average=1.5)，是所有对比方法中性能最好的方法，其余 6 种算法按平均 Ranks 值从高到低排序依次是 On-ITS、OLMOACO、WSE、SA、OPS 和 BFS。

表 5 台风灾害主题下各种主题爬虫算法在不同爬行阶段实验结果比较

DP	算法	Accuracy	LP	AR_{DP}	SD_{DP}	AR_{LP}	SD_{LP}	Time/h
	BFS	0.092	92	0.231	0.221	—	—	—
	OPS	0.391	391	0.432	0.283	—	—	—
	SA	0.414	414	0.551	0.212	—	—	—
1 000	WSE	0.571	571	0.672	0.123	—	—	—
	On-ITS	0.707	707	**0.811**	0.152	—	—	—
	OLMOACO	0.625	625	0.681	0.212	—	—	—
	WL	**0.714**	**714**	0.701	**0.121**	0.747	0.112	—
	BFS	0.071	355	0.249	0.222	—	—	—
	OPS	0.248	1 240	0.442	0.241	—	—	—
	SA	0.492	2 460	0.591	0.181	—	—	—
5 000	WSE	0.606	3 030	0.664	0.134	—	—	—
	On-ITS	0.807	4 035	**0.852**	0.128	—	—	—
	OLMOACO	0.767	3 835	0.701	0.126	—	—	—
	WL	**0.868**	**4 338**	0.703	**0.108**	0.721	0.104	—
	BFS	0.045	448	0.212	0.174	—	—	—
	OPS	0.124	1 240	0.431	0.172	—	—	—
	SA	0.422	4 215	0.552	0.169	—	—	—
10 000	WSE	0.707	7 069	0.687	0.124	—	—	—
	On-ITS	0.812	8 120	**0.781**	0.112	—	—	—
	OLMOACO	0.737	7 370	0.700	**0.088**	—	—	—
	WL	**0.893**	**8 927**	0.701	0.121	0.723	0.098	—
	BFS	0.047	705	0.211	0.184	—	—	8.5
	OPS	0.083	1 244	0.436	0.144	—	—	9.1
	SA	0.358	5 370	0.533	0.160	—	—	11.1
15 000	WSE	0.704	10 560	0.684	0.127	—	—	12.2
	On-ITS	0.820	12 300	**0.792**	0.123	—	—	13.2
	OLMOACO	0.750	11 250	0.700	**0.075**	—	—	15.0
	WL	**0.873**	**13 095**	0.702	0.099	0.712	0.095	**8.2**

表 6 台风灾害主题下 7 种主题爬虫算法的 Friedman ranks 比较

算法	Accuracy	LP	AR_{DP}	SD_{DP}	Average
BFS	7	7	7	7	7
OPS	6	6	6	5	5.75
SA	5	5	5	6	5.25
WSE	4	4	4	4	4.0
On-ITS	2	2	1	3	2.0
OLMOACO	3	3	3	1	2.5
WL	1	1	2	2	**1.5**

4 结 束 语

本文提出一种基于 Wang−Landau 抽样的主题爬行方法。该方法改进了传统主题爬虫因缺乏结合历史爬行经验而难以对爬行路径进行修正的缺陷，避免了爬虫过早陷入局部最优的问题。WL 抽样算法根据直方图函数和概率密度函数获取当前网页搜索过程中的统计经验，指导爬虫以一定概率接受目标链接，决定爬虫下一步的搜索方向。同时，考虑到互联网中相关网页大量聚集于具有相同域名站点的特点，本文基于区域竞争策略自适应地生成候选目标链接，以供 WL 算法抽样选择。实验结果表明，本文提出算法可显著提升爬虫的爬准率，在所有下载网页主题相关度的标准差方面也具有优势。因此，WL 方法是一种有效的主题爬虫方法。

参 考 文 献

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[EB/OL]. [2022-01-21]. http://www.cnnic.net.cn/NMediaFile/old_attach/P020220721404263787858.pdf.
China Internet Network Information Center. Statistical report on the development of Internet in China[EB/OL]. [2022-01-21]. http://www.cnnic.net.cn/NMediaFile/old_attach/P020220721404263787858.pdf.

[2] RAWAT A, BANSAL M. Review paper on focused crawler for categorization[J]. International Journal of Computer Trends & Technology, 2014, 12(3): 141-143.

[3] DE BRA P M E, POST R D J. Searching for arbitrary information in the WWW: The fish−search for mosaic[C]// World Wide Web Conference. Geneva: [s.n.], 1994: 1-11.

[4] 程元堃, 廖闻剑, 程光. 词向量聚类加权 shark−search 的主题爬虫策略研究[J]. 计算机与数字工程, 2018, 46(1): 144-148.
CHENG Y K, LIAO W J, CHENG G. Research on the theme crawler strategy of shark-search weighted by word vector clustering[J]. Computer and Digital Engineering, 2018, 46(1): 144-148.

[5] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.
[6] GAO Q, ZHANG Y P. Study on theme-drift of hyperlink-induced topic search algorithm[J]. Journal of Computer Applications, 2009, 29(11): 3100-3102.
[7] 周雪, 刘乃文. 引入主题链接块因子的候选链接搜索策略研究[J]. 计算机与数字工程, 2018, 46(5): 874-878.
ZHOU X, LIU N W. Research on candidate link search strategy with Topic link block factor[J]. Computer and Digital Engineering, 2018, 46(5): 874-878.
[8] SUEBCHUA T, MANASKASEMSAK B, RUNGSAWANG A, et al. Efficient topical focused crawling through neighborhood feature[J]. New Generation Computing, 2018, 36(2): 95-118.
[9] 库珊, 刘钊. 基于 PageRank 与 HITS 的改进算法的网页排名优化[J]. 武汉科技大学学报, 2019, 42(2): 155-160.
KU S, LIU Z. Webpage ranking optimization based on the improved algorithm of PageRank and HITS[J]. Journal of Wuhan University of Science and Technology, 2019, 42(2): 155-160.
[10] 姜琨, 朱磊, 王一川. 基于动态隧道技术的主题爬行策略[J]. 计算机系统应用, 2020, 29(3): 253-260.
JIANG K, ZHU L, WANG Y C. Topic crawling strategy based on dynamic tunneling technology[J]. Computer System Applications, 2020, 29(3): 253-260.
[11] HOSSEINKHANI J, TAHERDOOST H, KEIKHAEE S. Anton framework based on semantic focused crawler to support web crime mining using SVM[J]. Annals of Data Science, 2021, 8(2): 227-240.
[12] 刘景发, 李新, 蒋盛益. 基于网页空间进化算法的暴雨灾害主题爬虫策略[J]. 计算机工程, 2019, 45(2): 184-190.
LIU J F, LI X, JIANG S Y. Crawler strategy for rainstorm disaster theme based on webpage space evolution algorithm[J]. Computer Engineering, 2019, 45(2): 184-190.
[13] LIU J F, LI X, ZHANG Q S, et al. A novel focused crawler combining Web space evolution and domain ontology[J]. Knowledge-Based Systems, 2022, 243: 108495.
[14] LIU J F, DONG Y, LIU Z X, et al. Applying ontology learning and multi-objective ant colony optimization method for focused crawling to meteorological disasters domain knowledge[J]. Expert Systems with Applications, 2022, 198: 116741.
[15] 荆文鹏, 王育坚, 董伟伟. 自适应遗传算法在主题爬虫搜索策略中的应用研究[J]. 计算机科学, 2016, 43(8): 254-257.
JING W P, WANG Y J, DONG W W. Application research of adaptive genetic algorithm in topic crawler search strategy[J]. Computer Science, 2016, 43(8): 254-257.
[16] 刘景发, 李帆, 蒋盛益. 基于综合优先度和主机信息的暴雨灾害主题退火爬虫算法[J]. 计算机科学, 2019, 46(2): 215-222.
LIU J F, LI F, JIANG S Y. Annealing crawler algorithm for rainstorm disaster theme based on comprehensive priority and host information[J]. Computer Science, 2019, 46(2): 215-222.
[17] 东熠, 刘景发, 刘文杰. 基于多目标蚁群算法的主题爬虫策略[J]. 计算机工程, 2020, 46(9): 274-282.
DONG Y, LIU J F, LIU W J. Focused crawler strategy based on multi−objective ant colony algorithm[J]. Computer Engineering, 2020, 46(9): 274-282.
[18] DU Y J, LIU W J, LV X J, et al. An improved focused crawler based on semantic similarity vector space model[J]. Applied Soft Computing, 2015, 36: 392-407.
[19] CHEN K W, ZHANG Z P, LONG J, et al. Turning from TF-IDF to TF-IGM for term weighting in text classification[J]. Expert Systems with Applications, 2016, 66: 245-260.
[20] WANG F, LANDAU D P. Efficient, multiple-range random walk algorithm to calculate the density of states[J]. Physical Review Letters, 2001, 86(10): 2050.
[21] WANG B W, ZHAO P. An adaptive image watermarking method combining SVD and Wang-Landau sampling in DWT domain[J]. Mathematics, 2020, 8(5): 1-20.
[22] 曹红英, 田华, 侯杰, 等. 使用王朗道蒙特卡罗算法和机器学习研究伊辛模型相变[J]. 山西大学学报 (自然科学版), 2022, 12: 1-9.
CAO H Y, TIAN H, HOU J, et al. Studying the phase transitions for Ising model through Wang-Landau monte carlo algorithm and machine learning[J]. Journal of Shanxi University (Natural Science Edition), 2022, 12: 1-9.
[23] LIPOWSKI A, LIPOWSKA D. Roulette-Wheel selection via stochastic acceptance[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(6): 2193-2196.
[24] RAWAT S, PATIL D R. Efficient focused crawling based on best−first search[C]//2013 3rd IEEE International Advance Computing Conference (IACC). [S.l.]: IEEE, 2013: 908-911.
[25] 王桦. 基于广度优先的主题爬虫的设计与实现[D]. 上海: 复旦大学, 2011.
WANG H. Design and implementation of theme crawler based on breadth first[D]. Shanghai: Fudan University, 2011.
[26] 刘景发, 顾瑶平, 刘文杰. 融合本体和改进禁忌搜索策略的气象灾害主题爬虫方[J]. 计算机应用, 2020, 40(8): 2255-2261.
LIU J F, GU Y P, LIU W J. Focused crawler method combining ontology and improved tabu search for meteorological disaster[J]. Journal of Computer Applications, 2020, 40(8): 2255-2261.
[27] LIU J F, LI F, DING R Y, et al. Focused crawling strategies based on ontologies and simulated annealing algorithm for rainstorm disaster domain knowledge[J]. Frontiers of Information Technology & Electronic Engineering, 2022, 23(8): 1189-1205.
[28] 贺晟, 程家兴, 蔡欣宝. 基于模拟退火算法的主题爬虫[J]. 计算机技术与发展, 2009, 19(12): 55-58.
HE S, CHENG J X, CAI X B. Focused crawler based on simulated anneal algorithm[J]. Computer Technology and Development, 2009, 19(12): 55-58.

编　辑　税　红