基于灰狼算法的主题爬虫

萧婧婕 陈志云
(华东师范大学计算机科学技术系 上海 200062)

摘 要 为了解决主题爬虫在全局搜索中难以实现最优解的问题,提高主题爬虫的准确率和召回率,文中设计了一个结合灰狼算法的主题爬虫搜索策略。实验结果表明,与传统的广度优先搜索策略以及同样是群体智能算法的遗传算法相比,基于灰狼算法的主题爬虫的性能有了很大的提高,能爬取到更多的主题相关的网页。

关键词 主题爬虫,灰狼算法,主题相关度,网页重要性
中图法分类号 TP301.6 **文献标识码** A

Focused Crawling Based on Grey Wolf Algorithms

XIAO Jing-jie CHEN Zhi-yun
(Department of Computer Science and Technology, East China Normal University, Shanghai 200062, China)

Abstract In order to solve the problem that the focused crawler is difficult to achieve an optimal solution in the global search, and improve the accuracy of the topic crawler and the recall rate, this paper designed a focused crawler search strategy combined with grey wolf algorithm. The experimental results show that compared with the traditional breadth-first search strategy and the genetic algorithm which is also a swarm intelligence algorithm, the performance of the focused crawler based on grey wolf algorithm was greatly improved, and more topic-related web pages can be crawled.

Keywords Focused crawler, Grey wolf algorithm, Thematic relevance, Webpage importance

1 引言

互联网资源随着时间的推移呈现爆炸级的增长,数据更新换代的速度越来越快,同时用户的搜索需求也越来越个性化,传统的爬虫已经不能满足用户的需求,因此主题爬虫随之诞生。主题爬虫将搜索范围限定在特定的领域内,在传统的爬虫的基础上加入网页内容分析以及链接进行分析,尽量减少无关网页的产生,保证了爬取到的相关网页的质量。

主题爬虫的基本流程如图 1 所示。首先,通过网页搜索选取一些 URL 作为初始种子集合,放入待处理队列,分析 URL 与主题的相关度,保存大于主题相关度阈值的网页;然后在下载的页面中提取其中包含的链接,对这些链接进行分析,通过计算网页重要性将符合要求的链接继续放入等待队列中进行下载分析。一直重复这个过程,直到满足一定的条件停止,等待队列为空,此时完成一轮的抓取。

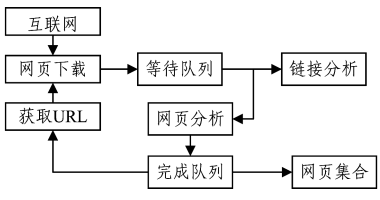

图 1 主题爬虫模型

实现主题爬虫的关键在于对网页的分析,目前主要有两种方法:1)基于链接分析的方法,分析网页之间相互链接之间的关系,从而分析网页的重要性,进而决定网页访问的顺序,常用的算法有 PageRank 算法[1]和 HITS 算法[2]等;2)基于内容分析的方法,通过对文本内容进行相似度等计算,得到相关度较高的网页,并舍弃不符合要求的网页,常用的算法有 shark-search 算法[3]、向量空间模型[4]、贝叶斯分类器算法[5]、SVM 算法[6]等。同时针对相关行业要求的特殊性,发展出了针对某个特定领域的主题爬虫,如针对图书主题的爬虫[7]、针对金融主题的爬虫[8]等。

灰狼算法借鉴生物界中的群体智能,利用其全局搜索最优的特点,将等待搜索的网页的 URL 作为初始的狼群,通过不断包围,得到最适合捕猎的狼群,即为最符合主题的网页。本文通过对主题爬虫和灰狼算法的研究,对传统的搜索策略进行改进,利用灰狼算法分析网页之间的链接,得到主题爬虫的最优解。实验结果证明,该方法可以有效提高爬取网页的准确率与召回率,提高主题爬虫的性能。

2 灰狼算法

Seyedail 等[9]受到狼群之间分工明确、协作捕猎食物的启发,于 2014 年提出了灰狼优化算法(GWO)。这是一种新的群体智能算法。灰狼优化算法模拟了狼群中的等级制度以及狼群的捕猎行为。如图 2 所示,狼群中等级最高的是 α 狼,它位于食物链的最顶端,负责领导、决策等行为。接下来的 β 和 δ 虽然不是最好的狼,却能在狼群缺失领导时,接替 α 狼,成为新的领导者。

本文受基于 MOOC 的计算机课资源建设项目资助。

萧婧婕(1994—),女,硕士生,主要研究方向为信息检索;陈志云(1967—),女,副教授,主要研究方向为多媒体技术、教育技术,E-mail:13611947576@163.com(通信作者)。

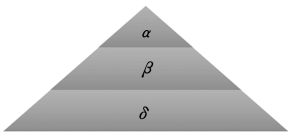

图 2　狼群等级结构

GWO 算法将每一只狼作为一个潜在的解[10]，其中 α 狼是最优解，β 狼和 δ 狼则是优解和次优解。GWO 算法是一个不断优化的过程，在这个过程中，α,β,δ 的位置在不断更新。狼群通过式(1)和式(2)更新距离，完成对猎物的搜索接近。

$$D=|C\times X_P(t)-X(t)| \tag{1}$$

$$X(t+1)=X_P(t)-A\times D \tag{2}$$

其中，D 表示灰狼与猎物之间的距离；t 表示迭代次数；X_P 表示猎物的位置，X 表示灰狼的位置。A 和 C 表示系数，$A=2a\times r_2-a$，$C=2r_1$。当 $|A|>1$ 时表示全局搜索，即灰狼群体扩大搜索范围，寻找更好的猎物；当 $|A|<1$ 时表示局部搜索，灰狼群体将包围圈缩小，在附近搜索猎物。$a=2-2(\frac{t}{\max})$，其中收敛因子 a 随着迭代次数从 2 线性递减到 0，t 是当前迭代次数，max 是最大迭代次数，r_1 和 r_2 均是[0,1]内的随机数。

狼群在狩猎时，α 狼、β 狼和 δ 狼对猎物有不同的适应度值，通过计算不同的适应度值，得到最优解、优解和次优解，并保存当前的位置信息；同时狼群根据这 3 个位置信息判断猎物的移动方向并逼近猎物完成狩猎。然后再次更新灰狼的位置，直到输出最优解。其表达式为：

$$D_\alpha=|C_1\times X_\alpha(t)-X(t)| \tag{3}$$

$$D_\beta=|C_2\times X_\beta(t)-X(t)| \tag{4}$$

$$D_\delta=|C_3\times X_\delta(t)-X(t)| \tag{5}$$

$$X_1=X_\alpha-A_1\times D_\alpha \tag{6}$$

$$X_2=X_\beta-A_2\times D_\beta \tag{7}$$

$$X_3=X_\delta-A_3\times D_\delta \tag{8}$$

$$X(t+1)=\frac{(X_1+X_2+X_3)}{3} \tag{9}$$

3　主题爬虫的设计

3.1　设计思想

研究表明，互联网上的信息是按照主题进行分类的，相关或重要的网页都是相互链接的，基于链接的搜索方法就是在此基础上进行分析，但是容易出现主题漂移的问题，最终导致结果中包含一些链接密度高但在内容上又与查询的主题无关的网页。而基于内容的搜索方法虽然根据主题相关度(VSM)算法[6]分析网页之间的相关性，却遗漏了分析网页之间的链接。本文分析以上两种方法，结合灰狼算法的优化特点，设计了如下算法：基于“从优质网页链接出去的 URL 可能还是优质网页”和“链接目标是优质网页的 URL 可能也是优质网页”的原则[11]，在初始 URL 集合中选择符合一定规则的新的 URL 作为新的集合，然后通过搜索操作，将不符合要求的 URL 淘汰，并不断缩小 URL 集合，得到主题相关度较高的网页集合。

3.2　主题相关度

向量空间模型算法用来计算主题相关度，通过计算文档之间的相似度来选择合适的文档。

向量空间模型算法采用文档中的关键词 n 作为向量空间的维数，每一维分量的大小用每一个关键词的权值 w_i 表示[12]，主题向量则可以表示为：

$$\alpha=(a_1,a_2,\cdots,a_n),a_i=w_i,i=1,2,3,\cdots,n$$

在对网页进行分词处理后，统计页面中的关键词出现的频率，并计算关键词之间的频率之比。然后将出现频率最高的关键词作为基准，将其频率设定为 1，通过频率比求出其他的关键词的相对频率 x_i，该页面对应向量的每一维分量[12]为 x_iw_i，那么页面的主题向量可以表示为：

$$\beta=(x_1w_1,x_2w_2,x_3w_3,\cdots,x_nw_n),i=1,2,3,\cdots,n$$

用两个向量夹角的余弦表示主题相关度，则主题相关度的计算公式为：

$$\begin{aligned}\cos\langle\alpha,\beta\rangle&=\frac{(\alpha,\beta)}{|\alpha||\beta|}\\&=\frac{|(a_1,a_2,\cdots,a_n),(x_1w_1,x_2w_2,\cdots,x_nw_n)|}{|(a_1,a_2,\cdots,a_n)||(x_1w_1,x_2w_2,\cdots,x_nw_n)|}\\&=\frac{\sum_{i=1}^{n}x_iw_i^2}{\sqrt{\sum_{i=1}^{n}w_i^2}\sqrt{\sum_{i=1}^{n}x_i^2w_i^2}}\end{aligned} \tag{10}$$

3.3　网页重要性

对队列中的 URL 进行网页重要性分析，其决定着主题爬虫的爬取方向。常用的网页重要性分析算法有 PageRank 算法。PageRank 算法是由 Google 的创始人佩奇和布林在 1997 年提出的一种网页排序算法[1]，其基本思想是：1)如果一个网页被很多网页链接到，那么这个网页是很重要的；2)如果一个很重要的网页链接到一个其他的网页，那么这个网页也会很重要。

PageRank 算法预先给每个网页一个相同的初始 PR 值，一般为 $\frac{1}{N}$，N 为网页总数。一个网页的 $PageRank$ 值可以由链接到该网页的所有网页的加权排名之和计算得到。计算公式为：

$$PR_i=\sum_{(i,j)\in E}\frac{PR_j}{L_j}$$

其中，PR_i 代表第 i 个网页的 $PageRank$ 值，作为每一个网页的排名标准，$PageRank$ 值越大，排名越高。PR_j 代表第 j 个网页的 $PageRank$ 值，L_j 代表第 j 个页面所包含的出链数目。

PageRank 算法的优点是在下载了一定数量的网页后，依次计算每一个网页的 $PageRank$ 值，然后使用有向图矩阵进行不断的迭代计算，最终得到一个平稳分布的值，利用这个值对网页进行排序，可以根据一定的网页重要性阈值得到所需的网页。

3.4　适应度函数的确定

适应度函数式是决定网页是否被淘汰的指标，如果直接用主题相关度阈值或网页重要性作为下载网页的标准，就会存在很多问题。例如：直接用网页重要性作为适应度函数，根据互联网的特点，在计算到一定深度时，会形成几个页面之间互相有链接的环形结构，使得这几个页面之间不能向外传递 $PageRank$ 值，网页重要性的计算就会出现误差。

因此，适应度函数值需要同时满足网页主题相关度阈值以及网页重要性。本文使用的适应度函数的计算公式为：$F=\alpha R+PR$，其中，R 代表主题相关度，PR 代表网页的重要性，α 的值根据实验进行计算后再调整。

3.5　灰狼位置的设计

灰狼进行狩猎操作时，需要确定收敛因子 a，这个参数可

以根据文献[13]进行调整，从而优化爬虫。在改进适应度函数的基础上可以进一步提高爬虫的性能.

同时，灰狼的位置函数可以利用式(11)－式(13)进行调整。其中 λ 的值可以根据实验结果进行调整，以优化爬虫的链接。

$$X_1 = \lambda X_\alpha - A_1 D_\alpha \tag{11}$$

$$X_2 = \lambda X_\beta - A_2 D_\beta \tag{12}$$

$$X_3 = \lambda X_\delta - A_3 D_\delta \tag{13}$$

3.6 算法流程

在抓取前，首先确定要爬取的主题，利用百度以及 Google 等通用搜索引擎进行搜索，在初步判断筛选搜索后的 URL 种子集后，将其作为最开始的网页种子集合，记为 S。该集合就是初始的狼群。初始化各项参数。

包围操作将集合 S 中未访问过的 URL 依次放入待处理队列 Q 中，分别计算相应的主题相关度，将大于主题相关度阈值的 URL 保留。解析网页，获得链接集合，构成有向图，计算 *PageRank* 值，获得网页重要性排序。

计算同时满足网页主题相关度以及网页重要性阈值的 URL 的目标函数值。根据式(11)－式(13)更新 α 狼、β 狼和 δ 狼的位置。若目标函数值小于 α 狼的目标函数值，则将 α 狼的目标函数值更新为最优目标函数，更新 α 狼的位置为最优位置，即为重要性和主题相关度最好的 URL；若目标函数值介于 α 狼和 β 狼的目标函数值之间，则将 β 狼的目标函数值更新为最优目标函数值，更新 β 狼的位置为最佳位置，为一个 URL；若目标函数值介于 β 狼和 δ 狼的目标函数值之间，则将 δ 狼的目标函数值更新为最优目标函数值，更新 δ 狼的位置为最佳位置，为一个 URL。

迭代完成后计算相应的收敛因子 a 的值。

完成一次网页处理下载后，更新位置，分别获得 α 狼、β 狼和 δ 狼的位置(即 URL)对应的网页链接。然后依次计算每一个链接的 *PageRank* 值并进行排序，得到排序后的集合 X_1, X_2, X_3，集合 $X = X_1 + X_2 + X_3$ 就是它们的并集，放入队列 Q 中。循环迭代，直到最大迭代次数。

算法流程如图 3 所示。

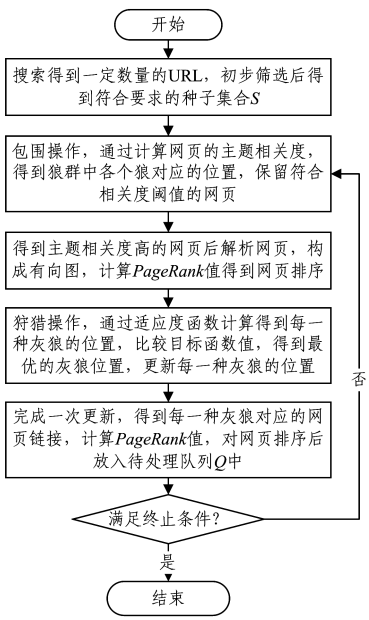

图 3 基于灰狼算法的主题爬虫

4 实验分析

4.1 实验环境

本文实验采用 Python 语言实现，运行在一台装有 Windows 10 系统，CPU 为 Intel Core i5，内存为 4GB，硬盘容量为 1TB 的实验室 PC 机上。

4.2 实验设计

为了验证本文算法的性能，设计了如下两个测试。同时，将“教育技术”作为搜索主题进行爬虫测试，将智慧教学、STEM 教育、慕课、电子书包、自适应学习、翻转课堂、创客学习、泛在学习、VR 教育、学习行为分析等关键词作为主题特征词叙词表。使用 n 维向量 $\alpha=(5,5,5,5,4,4,4,3,3,2)$ 表示各个特征词的权重。

测试 1 采用本文方法(GWOA)、广度优先搜索算法(BFS)、自适应遗传算法搜索方法[14](AGA)进行实验，分别记录这 3 种方法抓取到的网页数据，然后根据文献[14]提出的衡量爬虫性能的指标——准确率的计算方法来进行分析比较。经过人工初步挑选后初始 URL 的数量为 15，通过多次实验调整，将主题相关度阈值设置为 0.90，灰狼位置中 $\lambda=0.6$，适应度函数中 $\alpha=0.75$，迭代 10 次后主题爬虫停止。

测试 2 初始种子集的数量设定为 20，从 Google 中搜索得到 30 个 URL，经过人工筛选后选取前 20 个 URL，在多次实验后，将主题相关度阈值调整为 0.95，灰狼位置中的 λ 调整为 0.65，适应度函数中 α 为 0.75，根据文献[6]中提出的另一种衡量爬虫性能的指标——召回率进行计算。分析比较广度优先搜索算法(BFS)、自适应遗传算法的搜索方法[14](AGA)以及本文方法(GWOA)这 3 种不同方法的召回率。

4.3 实验分析

测试 1 的结果如图 4 所示。根据折线可以分析得到：搜索初期，本文提出的基于灰狼算法的搜索方法下载的相关页面的准确率低于采用自适应遗传算法的搜索方法和广度优先的搜索方法，随着爬取的页面数越来越多，使用本文算法的优势越来越明显，下载网页的准确率也越来越高。原因是本文方法使用的适应度函数和灰狼位置函数进一步提升了网页链接的准确性。

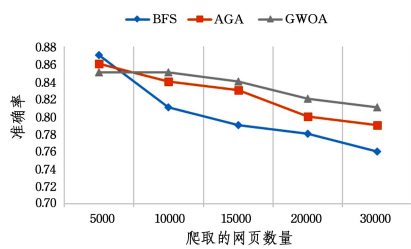

图 4 使用不同搜索方法下载网页的准确率

测试 2 的结果如图 5 所示。从图中可以看出，采用本文方法，主题相关的网页召回率是最高的，其次是自适应遗传算法，最后是 BFS 算法，召回率最低。在网页爬取的数量较少时，3 种方法相差不大。基于灰狼算法的主题爬虫调整了主题相关度阈值以及 PageRank 算法，有效避免了主题漂移问题；利用灰狼位置函数进行调整，避免了主题相关网页的遗漏等问题。因此随着爬取网页数量的逐渐增多，本文方法的召回率逐渐增高，相比其他两种算法，有效性更好。

(下转第 166 页)

[C]// 2013 IEEE 4th International Conference on Electronics Information and Emergency Communication. 2013:238-241.

[9] BHATIA S, VISHWAKARMA V P. Feed forward neural network optimization using self adaptive differential evolution for pattern classification[C]// 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). 2016:184-188.

[10] GOLDBERG D E. Genetic Algorithms in Search, Optimization, and Machine Learning[R]. Addison-Wesley, Reading, MA, 1989.

[11] MENGSHOEL O J, GOLDBERG D E. The crowding approach to niching in genetic algorithms[J]. Evolutionary Computation, 2008, 16(3):315-354.

[12] SHI E C, LEUNG F H F, BONNIE N F. Differential Evolution with adaptive population size[C]// Law 2014 19th International Conference on Digital Signal Processing. 2014:876-881.

[13] BREST J, ZAMUDA A, FISTER I, et al. Large scale global optimization using self-adaptive differential evolution algorithm[C]// Evolutionary Computation (CEC). 2010:1-8.

[14] RONKKONEN J, KUKKONEN J, PRICE K V. Real-Parameter Optimization with Differential Evolution[C]// The 2005 IEEE Congresson Evolutionary Computation. 2005:506-513.

[15] PRICE K V, STORN R M, LAMPINEN J A. Differential Evolution, A Practical Approach to Global Optimization[R]. Springer, 2005.

[16] FEOKTISTOV V. Differential Evolution: In Search of Solutions (Springer Optimization and Its Applications)[R]. Springer-Verlag NewYork, Inc., Secaucus, NJ, USA, 2006.

[17] VOLKOVAS R, FAIRBANK M, PEREZ-LIEBANA D. Diversity maintenance using a population of repelling random-mutation hill climbers[C]// 2017 9th Computer Science and Electronic Engineering (CEEC). 2017:37-42

[18] CHEN L. An Adaptive Genetic Algorithm Based on Population Diversity Strategy[C]// 2009 Third International Conference on Genetic and Evolutionary Computing. 2009:93-96.

[19] WAGSTAFF K, CARDIE C. Constrained K-means clustering with background knowledge[C]// The Eighteenth International Conference on Machine Learning. 2001:577-584.

[20] FREY B J, DUECK D. Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315:972-976.

[21] CUI X X, LI M, FANG T J. Study of population diversity of multiobjective evolutionary algorithm based on immune and entropy principles[C]// Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat). 2001:1316-1321.

[22] LIANG J J, QU B Y, SUGANTHAN P N. Problem definitions and evaluation criteria for the CEC 2014 special session and competition on single objective realparameter numerical optimization[R]. Tech. Rep. 201311, Computational Intelligence Laboratory, Zhengzhou University, Zhengzhou, China, 2014.

[23] LIU J, LAMPINEN J. A Fuzzy Adaptive Differential Evolution Algorithm[J]. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 2005, 9(6):448-462.

[24] ZHANG X, YE Z W, YANG J, et al. An approach for learning the optimal "tuned" masks based on differential evolution algorithm[C]// 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). 2017:585-590.

(上接第 148 页)

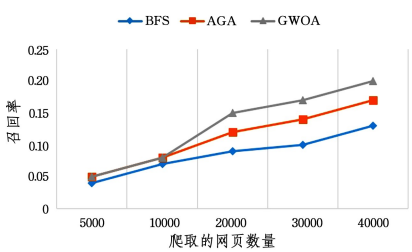

图 5 不同搜索方法下网页的召回率

结束语 基于灰狼算法的主题爬虫能够有效地提高抓取网页的准确率和召回率，提高搜索的精度。但是灰狼算法还存在迭代过程中容易限于局部最优无法完全覆盖网页的情况。改进灰狼算法的位置函数，加入合适的适应度函数使其具有自适应性，从而进一步提升算法性能并运用到主题爬虫中将是下一步的工作。

参 考 文 献

[1] CHO J, GARCIA-MOLINA H, PAGE L. Efficient crawlingthrough URL ordering[J]. Computer Networks and ISDN Systems, 1998, 30(1):161-172.

[2] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM (JACM), 1999, 46(5):604-632.

[3] HERSEOVICI M, JACOV M, MAREK Y S. The Shark-search algorithm an application: Tailored Web site mapping[J]. Computer Networks and ISDN Systems, 1998, 23(1):41-58.

[4] 杨小平，丁浩，黄都培. 基于向量空间模型的中文信息检索技术研究[J]. 计算机工程与运用，2003，15：109-111.

[5] 邹永斌，陈兴蜀，王文贤. 基于贝叶斯分类器的主题爬虫研究[J]. 计算机应用研究，2009，26(9)：3418-3420，3439.

[6] 李璐，张国印，李正文. 基于 SVM 的主题爬虫技术研究[J]. 计算机科学，2015，42(2)：118-122.

[7] 张莉婧，曾庆涛，李业丽，等. 面向图书主题的爬虫算法研究[J]. 计算机科学，2017，44(11)：460-463

[8] 陈黎，李志易，琚生根，等. 基于 SVM 预测的金融主题爬虫[J]. 四川大学学报(自然科学版)，2010，47(3)：493-497.

[9] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimization[J]. Advances in Engineering Software, 2014, 69(7):46-61.

[10] 魏政磊，赵辉，韩邦杰，等. 具有自适应搜索策略的灰狼优化算法[J]. 计算机科学，2017，44(3)：259-263.

[11] 刘国靖，康丽，罗长寿，等. 基于遗传算法的主题爬虫策略[J]. 计算机应用，2007，27(12)：172-176.

[12] 张海亮，袁道华. 基于遗传算法的主题爬虫[J]. 计算机技术与发展，2012，22(8)，48-52.

[13] 郭振洲，刘然，拱长青，等. 基于灰狼算法的改进研究[J]. 计算机应用研究，2017，34(12)：3603-3606.

[14] 荆文鹏，王育坚，董伟伟. 自适应遗传算法在主题爬虫搜索略中的应用研究[J]. 计算机科学，2016，43(8)：254-257.