

文章编号:1001-9081(2009)04-1114-03

基于 Web 页面平均质量的 Web 搜索模型和优化算法

付国瑜, 黄贤英

(重庆工学院 计算机科学与工程学院, 重庆, 400050)

(studywork2008@yahoo.com.cn)

摘要: 针对 Web 搜索引擎的特点, 提出了一种基于量子遗传克隆挖掘 (QGCMA) 的搜索策略。该算法将用户的查询描述为 Web 页面的平均质量, 并通过克隆, 变异, 交叉的操作获取具有高亲和度的抗体 (Web 页面)。通过实验结果分析得出, 在 Web 搜索中该方法比标准的遗传算法 (GA) 具有较明显的优势。

关键词: 搜索引擎; Web 搜索; 遗传算法; 克隆选择算法; 量子计算

中图分类号: TP391 **文献标志码:** A

Web search model and optimal algorithm based on mean quantity of Web pages

FU Guo-yu, HUANG Xian-ying

(School of Computer Science and Engineering, Chongqing Institute of Technology, Chongqing 400050, China)

Abstract: This paper proposed a search strategy based on Quantum Genetic Clonal Mining Algorithm (QGCMA) for Web search. The user query was used to mathematically define a mean quantity of Web pages, and evolved a population of Web pages for maximizing the affinity by clonal, mutation and crossover operator. The analysis and experimental results show that the proposed method is superior to standard genetic algorithm in Web search.

Key words: search engine; Web search; Genetic Algorithm (GA); Clonal Selection Algorithm (CSA); Quantum Computing (QC)

0 引言

随着 Web 技术的飞速发展, 人们越来越依靠网络来查找他们所需要的信息, 但是, 由于网上的信息源多不胜数, 也就是我们经常所说的“Rich Data, Poor Information”。所以如何有效地发现我们所需要的信息, 就成了一个很关键的问题。为了解决这个问题, 搜索引擎就随之诞生。搜索引擎以一定的策略在互联网中进行搜索、发现信息, 对信息进行理解、提取、组织和处理, 并为用户提供检索服务, 从而起到信息导航的目的^[1]。目前, 75% 的 Web 用户经常进行搜索, 64% 的 Web 用户以搜索作为寻找信息的主要方法, Web 搜索引擎已经成为当今信息技术领域研究的热点和焦点问题。

一个优秀的搜索引擎必须处理以下几个问题: 1) 网页的分类; 2) 自然语言的处理; 3) 搜索策略的调度和协作; 4) 面向特定用户的搜索。所以很多搜索引擎不同程度地使用了一些人工智能的技术来解决这些问题。近年来, 遗传算法 (Genetic Algorithm, GA) 因其高效的优化性能在 Web 搜索中得到了广泛的应用。在这些研究成果中, GA 算法通过用户提供的一些主页, 能自动搜索并获取其他相关主页。文献 [2] 描述了一种基于 HTML 标签的遗传学习机制的 Web 文档检索方法; 文献 [3] 提出了一种基于 GA 的信息检索方法, 通过 GA 可以对 Web 页进行自动分类和更新^[4]。文献 [5-6] 主要针对 Web 页检索, GA 用来预测用户的参数选择、动态优化和 Web 页的动态链接。文献 [1] 利用遗传算法建立用户的 Profile 应用于个性化系统。文献 [7-8] 提出了一种基于 GA 的 Web 关联规则挖掘算法。GA 是一类模拟生物进化的智能优化算法, 它在解决组合优化问题中具有明显的优势和特色,

但在 Web 挖掘问题上, 最优个体并不能代表问题的最优解, 问题的解要通过一组染色体来表示, 因此该方法存在染色体集成的问题。在研究中发现, 并不是适应高的染色体集成在一起形成的解越好, 这就要求算法不仅要能找出适应度高的个体, 还要能找出适应度不高但能提高最终挖掘结果准确性的那部分个体。能否找出这些适应度不高的个体并在进化过程中保留这些个体, 是决定挖掘算法性能好坏的关键。这不仅要求算法具有很好的全局搜索能力和局部搜索能力, 而且能在进化过程中维持多样性的有效探索。

针对 GA 算法在 Web 搜索中存在的问题, 我们在 GA 算法的基础上, 融入了量子计算和克隆选择算法的思想, 并提出一种新的量子遗传克隆挖掘算法 (Quantum Genetic Clonal Mining Algorithm, QGCMA)。克隆选择算法 (Clonal Selection Algorithm, CSA) 是模拟免疫系统对病菌的多样性识别能力而设计出来的多峰值搜索算法^[9], 其对父代进行克隆复制的策略, 能有效地保持了解的多样性并扩大空间搜索范围。量子计算 (Quantum Computing, QC) 是信息科学和量子力学相结合的新颖交叉科学。QC 的多样性、全干扰性, 可克服 GA 的早熟收敛现象^[10]。通过仿真实验证明, QGCMA 算法能有效弥补遗传算法在 Web 搜索中的不足, 是一种解决 Web 搜索问题行之有效的快速方法。

1 Web 搜索问题描述

在 Web 搜索过程中, 每个 Web 页面都有作为权威 (Authority), 并被指向的属性 $A(P)$; 同时具有作为资源中心 (Hub), 并指向其他页面的属性 $H(P)$ 。针对上述 Web 搜索的特点, 本文将以 Web 页面权威属性 $A(P)$ 和 Web 页面资源属

收稿日期: 2008-10-27; 修回日期: 2008-12-09。

作者简介: 付国瑜 (1973-), 女, 重庆人, 讲师, 硕士, 主要研究方向: 数据挖掘、信息安全; 黄贤英 (1968-), 女, 重庆人, 教授, 硕士, 主要研究方向: 信息安全。

性 $H(P)$ 两个性能指标来描述 Web 搜索问题。

定义1 Web 页面权威属性 $A(P)$:

$$A(P) = \sum_{i=1}^n ZK_i \quad (1)$$

其中, K_1, K_2, \dots, K_n 为用户输入的关键词, n 为输入关键词的总数, ZK_i 为每个关键词所链接的 Web 页面平均数。

定义2 Web 页面资源属性 $H(P)$:

$$H(P) = \sum_{i=1}^m A_i(P) \quad (2)$$

其中 m 为每页的链接总数。

定义3 Web 页面平均质量 $M(P)$:

$$M(P) = \frac{H_{\max}(P) + H_{\min}(P)}{2} \quad (3)$$

其中: $H_{\max}(P)$ 和 $H_{\min}(P)$ 分别指应用 QGCMA 算法后 Web 页面资源属性的最大值和最小值。

2 QGCMA 算法的应用

QGCMA 挖掘算法将 QC、GA 和 CSA 的优点充分进行结合, 下面是该算法的设计模型。

定义4 QGCMA 模型:

$$QGCMA = (E, F, C_l, C_r, M) \quad (4)$$

其中, E 表示量子编码, F 是亲和度函数, C_l 是克隆操作, C_r 是量子交叉操作, M 是遗传变异操作。

2.1 量子编码及初始抗体群

在量子计算中, 最小的信息单元用量子位表示。量子位又称为量子比特。一个量子比特的状态可表示为:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, |\alpha|^2 + |\beta|^2 = 1 \quad (5)$$

把满足式(5)的一对复数 α 和 β 称为一个量子比特的概率幅, 因此量子比特可以用概率幅表示为 $[\alpha \beta]^T$ 。

在 QGCMA 中, 在第 k 代的抗体种群为 $A(k) = [a_1(k) a_2(k) \dots a_n(k)]$, $a_i(k)$ 定义如下:

$$a_i(k) = \begin{bmatrix} \cosh k_1 & \cos k_2 & \dots & \cos k_m \\ \sinh k_1 & \sin k_2 & \dots & \sinh k_m \end{bmatrix} \quad (6)$$

其中, $k_i = 2 \times \pi \times r$, $r = \text{random}[0, 1]$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$; m 为抗体(Web 页面)的长度, n 为抗体群大小(Web 页面的数量), k 为进化代数。

在式(6)中, $\cosh k_i, \sinh k_i$ 都被初始化为 $1/\sqrt{2}$, 并产生规模为 n 初始抗体群 $A(0)$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$ 。

2.2 亲和度函数

亲和度是用来表明抗体与抗原之间的匹配程度, 亲和度越高, 说明抗体越接近抗原, 也就越接近所求问题的解^[11-12]。本文设计的亲和度函数为:

$$f(\text{antibody}) = e^{M(p)} \quad (7)$$

亲和度函数表明, Web 页面平均质量 $M(P)$ 越高, 则该页面被搜索到的概率就越大。

2.3 克隆

克隆是依据抗体与抗原的亲合度函数 $f(*)$, 将解空间中的一点 $a_i(k) \in A(k)$ 分裂成 N_c 个相同的点 $a'_i(k) \in A'(k)$, N_c 是克隆规模^[13]。

设抗体群 $A(k) = [a_1(k), a_2(k), \dots, a_n(k)]$, 克隆算子 C_l 定义为: $C_l(A(k)) = [C_l(a_1(k)), C_l(a_2(k)), \dots, C_l(a_n(k))]$ 。其中, $C_l(a_i) = I \times a_i$, $i = 1, 2, \dots, n$, I 为 N_c 维行向量, 而 $N_c = g(\beta, f(a_i(k)))$ 。一般取 $g(\beta, f(a_i(k))) =$

$\beta \times f(a_i(k)) / \sum_{i=1}^n f(a_i(k))$, $i = 1, 2, \dots, n$, β 是放大系数。克隆

后抗体群变为: $A'(k) = \{A(k), A'_1(k), \dots, A'_{N_c}(k)\}$ 。其中, $A'_i(k) = \{a_{i1}(k), a_{i2}(k), \dots, a_{i(N_c-1)}(k)\}$, $a_{ij}(k) = a_i(k)$, $j = 1, 2, \dots, N_c - 1$ 。

2.4 遗传变异

通过克隆扩大了群体的规模后, 对克隆后的临时群体 $A'(k)$ 中每个抗体进行变异, 可以提高群体中抗体的多样性, 扩大搜索范围, 用来寻找更优秀的抗体。遗传变异操作如下:

$$B(k) = A(k)' + P_m \times \exp(-f(*)) \times N(0, 1) \quad (8)$$

其中: $B(k)$ 和 $A(k)'$ 分别是父抗体和子抗体, $N(0, 1)$ 是均值为 0, 方差 $\sigma = 1$ 的高斯变量; P_m 是变异概率; $f(*)$ 是 v 的亲和度。

2.5 量子交叉

在 QGCMA 中, 我们采用量子全干扰交叉操作, 抗体群中所有抗体均参与交叉。这种量子交叉可以充分利用抗体群中尽可能多的抗体信息, 改进普通交叉的局部性与片面性。在抗体群进化出现早熟时, 它能够产生新的抗体, 给进化过程注入新的动力。

这种交叉操作借鉴了量子的相干性, 可以克服 GA 在进化后期的早熟现象。表 1 列出的是抗体群规模为 3, 抗体长度为 5 的一种具体交叉操作。

表 1 全干扰量子交叉

抗体标号		抗体			
1	2	A(1)	C(2)	B(3)	A(4)
2	3	B(1)	A(2)	C(3)	B(4)
3	1	C(1)	B(2)	A(3)	C(4)
					B(5)

在表 1 中, 每个大写字母表示交叉后的一个新抗体, 如:

$$A(1) - C(2) - B(3) - A(4) - C(5)。$$

量子交叉后得到临时抗体群 $C(k)$ 。

2.6 QGCMA 算法的框架

QGCMA 算法描述如下:

1) 在解空间产生初始抗体群 $A(k)$ (样本页面), 初始化种群规模 n 样本页面的数量, 变异概率 P_m , 变异概率 P_e , $k = 0$;

2) 计算抗体群中每个抗体的亲和度 $f(a_i(k))$, $i = 1, 2, \dots, n$;

3) 对抗体群 $A(k)$ 进行克隆操作, 得到抗体群 $A(k)'$;

4) 对抗体群 $A(k)'$ 进行遗传变异操作, 变异后临时抗体群为 $B(k)$;

5) 对抗体群 $B(k)$ 进行量子交叉操作, 交叉后产生新抗体群 $A(k+1)$ (将平均质量高的 Web 页面保留到下一代继续进化);

6) $k = k + 1$; 当满足终止条件时, 算法结束, 并将结果显示给用户(最终搜索到的 Web 页面信息); 否则, 返回到步骤 2)。

3 仿真实验与结果分析

3.1 实验环境和参数选取

为了验证算法的有效性, 我们用 Java 语言实现了本文提出的 QGCMA 算法, 并在 Pentium 4 CPU 1.80 GHz, 1 GB 内存的 PC 机上进行了测试。实验过程中用到的一些参数定义为: 种群规模 150, 交叉概率 0.85, 变异概率 0.1。在测试前,

我们已经从标准的三个搜索引擎(Yahoo、Google 和 Msn)下载了 500 个样本页面,并存储到硬盘中以备进一步操作。实验测试的结果都是平均运行 10 次以上。

3.2 实验结果与分析

图 1 显示的是搜索算法在迭代 1000 次时,Web 页面平均质量的变化情况;图 2 是样本页面数为 50 的 Web 页面平均质量;图 3 是样本页面数为 250 时的 Web 页面平均质量。在图 1 中,当种群规模太小(例如 60 页),QGCMA 算法和文献[2]中的标准 GA 算法会降低 Web 页面结果的质量。当种群规模增大到一定程度时(例如 140 页)。Web 页面的质量有显著提高,导致执行时间将急剧增加,图 2~3 中明显反映出,当种群规模由 50 页到 250 页时,Web 页面平均质量明显得到改善。同时,QGCMA 算法的性能明显优于 GA,因为在 QGCMA 算法中,克隆操作大大增加了空间解的搜索范围,量子编码和量子交叉丰富了群体的多样性,这些都能充分发挥 GA 的优点和智能算法的优势。

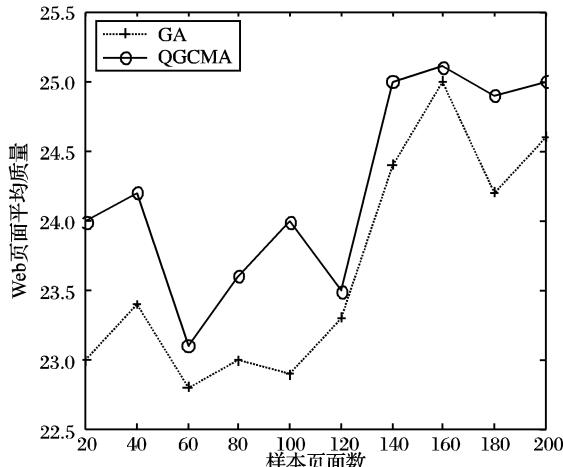


图 1 Web 页面平均质量的变化情况

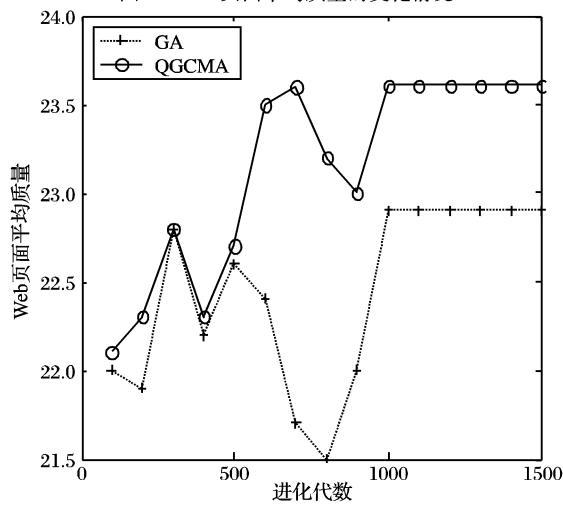


图 2 页数为 50 时 Web 页面平均质量的变化情况

4 结语

随着 Internet 的迅速发展,网上的信息急剧膨胀,要在这样一个浩如烟海的信息空间里查找所需的信息,往往花费了很多时间和精力,却收获甚少。搜索引擎技术的出现,是解决上述问题的主要途径。本文通过那些已下载的原始页码,比较 Web 页面的平均质量,用实验证明了随着迭代数目的增加,在合理的执行时间内仍然可以获得较好的结果。由于一个特定的搜索引擎主要包含某一特定领域的信息,覆盖面有

限。因此,在今后的研究过程中,我们将重点研究如何在 Internet 上获得高检索精度和高检索率。

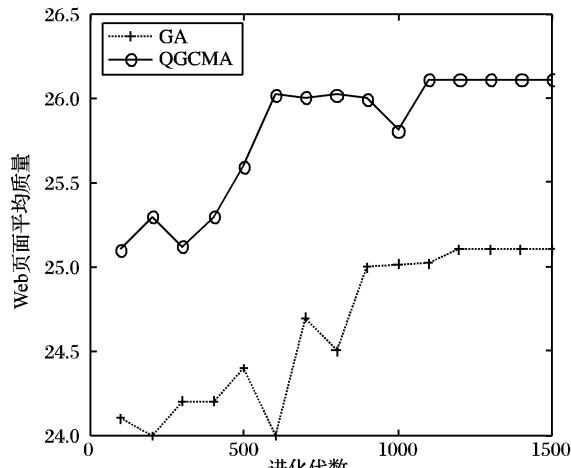


图 3 页数为 250 时 Web 页面平均质量的变化情况

参考文献:

- [1] NASAROUI O, GONZALEZ F, DASGUPTA D. The fuzzy artificial immune system: Motivations, basic concepts, and application to clustering and Web profiling, Fuzzy Systems[C]// Proceedings of the 2002 IEEE International Conference on Fuzzy Systems. Washington, D C: IEEE Computer Society, 2002: 711–716.
- [2] KIM S, ZHANG B T. Web-document retrieval by genetic learning of importance factors for html tags[C]// Proceedings of the 2000 International Workshop on Text and Web Mining. Melbourne, Australia: [s. n.], 2000: 13–23.
- [3] BOUGHANEM M, CHRISMENT C, MOTHE J, et al. Connectionist and genetic approaches for information retrieval[J]. Soft Computing in Information Retrieval: Techniques and Applications, 2000, 50(1): 102–121.
- [4] LOIA V, LUONGO P. An evolutionary approach to automatic Web page categorization and updating[C]// Proceedings of the 1st Asia-Pacific Conference on Web Intelligence: Research and Development, LNCS 2198. London: Springer-Verlag, 2002: 292–302.
- [5] KARGUPTA H. The gene expression messy genetic algorithm[C]// Proceedings of the 1996 IEEE International Conference on Evolutionary Computation. New York: IEEE Computer Society, 1996: 631–636.
- [6] PERKOWITZ M, ETZIONI O. Adaptive Web sites: An AI challenge[C]// Proceedings of the 15th International Joint Conference on Artificial Intelligence: IJCAI97. Nagoya, Japan: [s. n.], 1997: 16–23.
- [7] 汤亚玲, 崔志明. 遗传算法在 Web 关联挖掘中的应用[J]. 微电子学与计算机, 2005, 22(10): 4–6.
- [8] 汤亚玲, 崔志明. 基于遗传算法的 Web 行为挖掘研究[J]. 微电子学与计算机, 2006, 23(8): 168–170.
- [9] 焦李成, 杜海峰. 人工免疫系统进展与展望[J]. 电子学报, 2003, 31(10): 1540–1548.
- [10] 李阳阳, 焦李成. 求解 SAT 问题的量子免疫克隆算法[J]. 计算机学报, 2007, 30(2): 176–183.
- [11] 刘芳, 杨海潮. 一种基于克隆策略的多播路由算法[J]. 电子与信息学报, 2004, 26(11): 1825–1829.
- [12] 胡江强, 郭晨, 李铁山. 启发式自适应免疫克隆算法[J]. 哈尔滨工程大学学报, 2007, 28(1): 1–5.
- [13] 沈艳军, 汪秉文. 基于实数编码的克隆选择算法及其应用[J]. 华中科技大学学报: 自然科学版, 2004, 32(2): 41–42.