

基于量子行为进化算法的聚焦爬虫搜索策略*

刘丽杰¹, 李盼池², 张强²

(1. 黑龙江八一农垦大学 信息技术学院, 黑龙江 大庆 163319; 2. 东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318)

摘要: 针对单一价值评价的聚焦爬虫搜索策略存在主题漂移等问题进行了研究, 充分利用量子进化算法所具有的智能性, 提出一种新的聚焦爬虫爬行算法。该算法充分结合网页在互联网上的分布特点, 利用立即价值和未来价值两类评价标准的优势, 根据聚焦爬虫实际运行过程中的搜索情况, 在线调整这两种标准在综合价值中的比重。实验仿真结果表明, 相对于单一价值的搜索策略, 量子进化算法获得较高的页面查全率和信息查准率, 能较好地解决现存问题, 具有一定的自适应性。

关键词: 聚焦爬虫; 主题相关度; 立即价值; 未来价值; 量子进化算法

中图分类号: TP393 **文献标志码:** A **文章编号:** 1001-3695(2012)11-4280-04

doi:10.3969/j.issn.1001-3695.2012.11.071

Search strategy of focused crawler based on Bloch quantum evolutionary algorithm

LIU Li-jie¹, LI Pan-chi², ZHANG Qiang²

(1. College of Information Technology, Heilongjiang Bayi Agricultural University, Daqing Heilongjiang 163319, China; 2. School of Computer & Information Technology, Northeast Petroleum University, Daqing Heilongjiang 163318, China)

Abstract: According to the single value evaluation focused crawler search strategy has the topic drift problem, and make full use of the intelligence of the Bloch quantum evolutionary algorithm (BQEA), this paper proposed a new algorithm of focused crawler. The algorithm integrated Web distribution on the Internet fully, used the advantages of two types of evaluation criteria of the immediate value and the future value adjusted to the proportion of two standards online in the integrated value, according to focused crawler search on the actual process. The experimental result by simulation show that, compared with the search strategy of a single value, the BQEA obtains a higher recall rate, and precision rate and can solve the existing problems with certain self-adaptive.

Key words: focused crawler; topic relevancy; immediate value; future value; Bloch quantum evolutionary algorithm

聚焦爬虫(focused crawler)又称为主题爬虫,它是专为查询某一领域或主题信息而出现的网页抓取工具。与通用网络爬虫不同,聚焦爬虫只需要索引小部分万维网数据,即用很少的服务器资源覆盖绝大部分某一特定行业、主题的数据,为面向主题的用户查询准备数据资源^[1]。但相对于通用网络爬虫,聚焦爬虫还需要解决的关键问题就是如何判断一个网页是否与主题相关并且如何根据主题相关度来决定待采集URL的访问顺序以获取更高的查全率和查准率。

1 相关研究现状

1.1 量子行为进化算法的基本思想

基本的量子遗传算法的主要问题是测量个体量子位的状态来生成所需要的二进制解,这是一个概率操作过程,具有很大的随机性和盲目性,因此在整体进化的同时,个体将不可避免地产生退化现象。此外,关于二进制编码,虽然适合某些优化问题如旅行商问题等,但对于数值优化问题,由于需要

频繁编码、解码,无疑加大了计算量;同时算法在进行粗略搜索时容易丢失最优解,使得过早地收敛到某个局部极值点。因此,需要引入一种既能保证种群多样性、又能保证算法高效性的机制,相应地产生了量子行为进化算法。量子行为进化算法(Bloch quantum evolutionary algorithm, BQEA)的基本思想^[2]是:a)直接采用量子位的Bloch球面坐标对量子个体进行编码,采用量子旋转门更新量子位;b)对于量子旋转门转角方向的确定,提出了一种简单实用的确定方法,对旋转和变异操作,提出了基于量子位的Bloch球面的坐标的新算子;c)将量子位的三个Bloch球面坐标都视为基因位,因此每个个体有三条基因链,分别对应于Bloch球面的x轴、y轴和z轴坐标。这样就能够避免通过对量子位测量来生成二进制编码带来的随机性,同时也能够扩展全局最优解的数量,提高获得全局最优解的概率。

1.2 聚焦爬虫研究现状

聚焦爬虫爬行的基本思想是计算爬行队列中待爬行URLs的访问优先级,每次选出主题相关度最大者进行爬行。因此,

收稿日期: 2012-04-10; 修回日期: 2012-05-20 基金项目: 中国博士后科学基金资助项目(20090460864); 黑龙江省教育厅科学技术研究资助项目(11551015)

作者简介: 刘丽杰(1982-),女,讲师,硕士,主要研究方向为网络信息系统、数据挖掘(bynd_llj@163.com);李盼池(1969-),男,河北大城人,教授,博士,主要研究方向为量子智能优化、过程神经网络;张强(1982-),男,讲师,硕士,主要研究方向为神经网络、智能算法。

聚焦爬虫的核心计算问题是在下载索引一个网页前如何选择合适的搜索策略来预测从网页中解析到的 URL 以及爬行下来的网页是否与主题相关^[3,4]。目前常用的聚焦爬行策略主要有两类^[5],即基于网页内容分析的搜索策略和基于超链分析的搜索策略。

基于网页内容分析的搜索策略主要是利用网页内容(文本、数据等资源)特征进行网页评价,网页的内容从原来的以超文本为主发展到后来动态页面(hidden Web)数据为主,它是通过计算这些信息与预设主题之间的主题相关度来评价爬行价值的高低并指导网络蜘蛛的爬行方向。基于网页内容分析的搜索策略的典型代表是 BestFirst 算法^[6],该类算法的优点是具有较好的理论基础且计算简单。但由于这类方法忽略了链接结构信息,因此在预测链接价值方面的准确性方面存在一些不足。基于超链分析的搜索策略是指互联网中网页的存在不是相互独立的,网页中的各个链接表明了网页间的相互引用关系,通过分析关系来确定爬行的网页的价值。通常认为当页面具有较多的人链或者出链时,则该页面具有较高的爬行价值,网页的价值又决定了爬虫的爬行顺序。目前基于链接结构的搜索策略的代表算法有 PageRank 算法和 HITS 算法^[7]。这类算法的优点就是考虑了链接的结构和页面之间的引用关系,缺点就是忽略了页面与主题的相关性,在某些情况下会出现搜索偏离主题的问题。

2 基于量子行为进化算法的聚焦爬虫方案设计

2.1 算法设计思想

2.1.1 算法思想

根据互联网信息资源分布的群聚性特点,可将搜索过程分为探索和发掘两个部分。在相关 Web 区域中爬行时,由于其中网页内容和主题相关,蕴涵了较多的相关信息,网页链接具有较大的立即价值,这时适合选用注重发掘的基于立即价值的搜索策略;同时互联网信息分布又具有隧道特性,使用基于立即价值评价的标准爬虫很容易迷失搜索方向,而基于未来价值评价的标准中爬虫可以预测链接的未来价值,也就是说,这种方法更注重探索,可以引导搜索从无关区域穿过隧道过渡到相关区域。因此网络爬虫在爬行时,必须能够自己决定什么时候需要发掘信息资源,什么时候需要探索信息资源。

2.1.2 算法相关定义

下面给出相关算法的定义^[8]。

定义 1 立即价值。针对某一检索主题 t , 假设 l 是网页 p 中的一个网页链接, 如果 l 所链接的网页 q 与主题 t 相关, 采用立即价值评价方法来预测 q 与 t 的相关程度, 称 l 具有与主题 t 相关的立即价值, 用符号 $It(l)$ 表示。

定义 2 未来价值。针对某一检索主题 t , 假设 l 是网页 p 中的一个网页链接, 如果 l 所链接的网页 q 与主题 t 无关, 但通过 q 访问若干网页后可检索到与主题 t 相关的页面 r , 估测 l 与主题 t 的相关度通常采用计算未来价值的方法, 称 l 具有与主题 t 相关的未来价值, 用符号 $Ft(l)$ 表示。

定义 3 综合价值。针对某一检索主题 t , 设页面 p 中有一链接 l , l 关于 t 的立即价值表示为 $It(l)$, 未来价值表示为 $Ft(l)$, 则 l 关于 t 的综合价值表示为

$$St = \alpha \times It(url) + \beta \times Ft(url), \alpha + \beta = 1 \quad (1)$$

其中: α, β 为动态权值, 根据网络爬虫在线获得的 Web 状态信

息动态调整。计算综合价值的公式表明, 聚焦爬虫对立即价值和未来价值的分配由 α, β 决定, 搜索每次选择综合价值函数最高的链接爬行。本文算法即以综合价值作为个体适应度函数。

2.2 主题相关性的判定

根据本文给出的算法思想, 在求得页面的相关度时是根据页面的立即价值和未来价值结合起来形成的综合价值而体现的。立即价值又是采用基于网页内容分析和基于超链分析而得出的综合相关度来进行计算的, 在这里采用了 PageRank 算法来确定网页的重要程度, 同时又采用向量空间模型(VSM)来进行页面内容的相关度计算。综合考虑这两项来估测待下载网页的综合价值, 则基于立即价值的网页主题相关度评价为

$$F_p = (1 - \omega)S_p + \omega PR(p) \quad (2)$$

其中: S_p 表示页面内容的主题相关度; $PR(p)$ 表示基于链接结构而表现出的页面的重要性; $\omega (0 \leq \omega \leq 1)$ 表示重要性的加权参数, 它的值只能人为地设定, 可以根据经验和具体的实验情况进行修改。

评价某网页的未来价值通常是网络蜘蛛通过训练获得某些经验信息, 用于对远期回报的预测, 使网络蜘蛛对未来回报的搜索具有一定的倾向性。目前有代表性的方法有基于巩固学习^[9]的方法和基于链接拓扑图的方法, 在这里主要采取巩固学习方法。在巩固学习模型中, 存在一个能够与环境交互的代理体, 代理体对环境状态的感知作为代理体的输入 i , 代理体通过选择行动 a 作为输出并作用于环境状态, 使环境状态发生转移, 伴随着状态转移代理体获得一个回报, 再根据获得的回报选择下一个行动, 如此往复。代理体的目标是通过尝试与错误追踪, 学习到使累计回报达到最大的行动选择策略。代理体的行动选择依赖于一个称为策略 π 的状态随机函数, 一个策略将每个状态映射为动作的概率分布, $\pi: S \mapsto \Pr(A)$ 。对于任一策略 π , 在实际计算中状态的价值函数定义为累积回报的期望值是记录状态/动作对价值的 Q 函数:

$$Q^\pi(s, a) = E\{R_t | s_t = s, a_t = a, \pi\} \quad (3)$$

其中: a 表示执行动作, 而 a_t 表示在 t 时刻执行的动作; s 表示状态, s_t 表示在 t 时刻的状态; 而 $Q^*(s, a) = \max Q^\pi(s, a)$, 由于从 Q^* 可以很容易地得到最优策略 $\pi^* = \arg \max Q^*(s, a)$, 因而求解最优策略的问题通常转换为求解 Q^* 值的问题。

为预测链接的未来回报价值, 搜索过程被划分成训练和搜索两个阶段。在训练阶段, 利用巩固学习算法对典型站点进行学习, 然后计算每一个链接的 Q 价值, 再将链接按 Q 价值大小分为几类, 利用每一类中链接文本训练 Naïve Bayes 分类器; 在搜索阶段, 面对价值未知的链接, 用 Naïve Bayes 分类器计算链接落在每一类中的概率, 并以这个概率为权值计算链接的综合 Q 价值, 以此 Q 价值来表示链接的未来回报价值, 以此决定链接的访问顺序。将网络爬虫经过若干无关页面的访问之后才能获得的主题相关页面称为未来回报, 对未来回报的预测值称为未来回报价值, 用 Q 价值表示。这种方法的核心就是学习如何计算链接的 Q 价值, 根据未来回报价值确定正确的搜索方向。

2.3 量子行为进化算法的聚焦爬行策略

2008 年, Li 等人^[10]提出量子衍生进化算法, 该算法是基于量子位 Bloch 坐标并且显示了良好的优化性能, 但其量子位

的两个自变量采用了分别进化的方式,没有解决两个自变量的协调问题,从而使优化效率受到影响。基于以上问题,通过与李盼池教授多次研讨,提出一种量子行为进化算法 BQEA,该算法采用泡利矩阵建立旋转轴,采用量子位在 Bloch 球面上绕轴旋转的方法实施优化搜索。该方法可实现量子位的两个自变量之间的最佳匹配,即动态调整立即价值和未来价值前面的权重值并将其应用到聚焦爬虫上得到了很好的实验效果。

1) 权重值初始化 在未访问的 URL 队列中选取一个 URL 进行相关度计算,根据式(2)计算得出 URL 的立即价值,根据巩固学习的算法得出该 URL 的未来价值。分别赋予 URL 的立即价值和未来价值 n 组不同的权重值,根据式(1)进行计算,得出 n 组不同的综合价值,选择综合价值最高的称为最优价值,同时最优价值所对应的权重值称为最优权重值。

2) 使用量子进化算法进行权重值调整 根据得到的最优价值和最优权重值,将其他 $n-1$ 组给出的权重值进行调整,下面给出具体的调整过程:

a) 权重值的量子表示。将需要调整的每一组的权重值都表示成为量子态的形式。由于在量子计算中,采用 $|0\rangle$ 和 $|1\rangle$ 表示微观粒子的两种基本状态,称它们为量子比特 (quantum bit, qubit),一个量子比特的任意状态都可以表示为这两个基本状态的线性组合,则所求出的 URL 对应权重值的量子态可以表示为

$$|\varphi\rangle = \alpha |I_t(\text{url})\rangle + \beta |F_t(\text{url})\rangle \quad (4)$$

其中: $I_t(\text{url})$ 表示 URL 的立即价值; $F_t(\text{url})$ 表示 URL 的未来价值; α 和 β 称为量子态的概率幅,即量子态 $|\varphi\rangle$ 因测量导致或者以 $|\alpha|^2$ 的概率塌缩到 $|I_t(\text{url})\rangle$, 或者以 $|\beta|^2$ 的概率塌缩到 $|F_t(\text{url})\rangle$, 且满足

$$|\alpha|^2 + |\beta|^2 = 1 \quad (5)$$

式中: $|\alpha|^2$ 表示立即价值的权重值, $|\beta|^2$ 表示未来价值的权重值。因此,权重值的量子态也可由概率幅表示为 $|\varphi\rangle = [\alpha, \beta]^T$ 。同时,一个量子比特也是一个可以在二维复希尔伯特空间中描述的两能级量子体系。根据叠加原理,权重值的量子态又可以写为

$$|\varphi\rangle = \cos \frac{\theta}{2} |I_t(\text{url})\rangle + e^{i\phi} \sin \frac{\theta}{2} |F_t(\text{url})\rangle \quad (6)$$

其中: $0 \leq \theta \leq \pi, 0 \leq \phi \leq 2\pi$, 不同于经典比特,量子比特属于一个由连续变量 θ 和 ϕ 所刻画的矢量空间。

量子比特可以借助 Bloch 球描述,该球为量子比特及对量子比特所进行的变换提供了几何图像。具体而言,量子比特可以用嵌入三维笛卡尔坐标系中的 Bloch 球面上的一个点来描述。如图 1 所示,图 1 中 $x = \cos \phi \sin \theta, y = \sin \phi \sin \theta, z = \cos \theta$ 。这样,量子态 $|\varphi\rangle$ 可以写为

$$|\varphi\rangle = \left[\sqrt{\frac{1+z}{2}}, \frac{x+iy}{\sqrt{2(1+z)}} \right]^T \quad (7)$$

b) 权重值量子位的投影测量。一个权重值对应量子位的 Bloch 坐标 (x, y, z) 可以利用量子位 $|\varphi\rangle$ 在计算基矢上的泡利矩阵 $\sigma_x, \sigma_y, \sigma_z$ 通过投影测量获得。

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

具体的量子位坐标的投影测量计算式为

$$x = \langle \varphi | \sigma_x | \varphi \rangle = \langle \varphi | \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} | \varphi \rangle$$

$$y = \langle \varphi | \sigma_y | \varphi \rangle = \langle \varphi | \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} | \varphi \rangle$$

$$z = \langle \varphi | \sigma_z | \varphi \rangle = \langle \varphi | \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} | \varphi \rangle \quad (8)$$

其中: $|\varphi\rangle$ 是用向量形式表示的权重值量子态,而式(6)中具体的计算首先将泡利矩阵中的三个分量分别与向量形式的权重值进行叉乘,将得到一个新的向量;再与前面的向量做内积操作,结果是一个数值,这样就可以得到一个权重值对应量子位的 Bloch 坐标 (x, y, z) 的值。

c) 旋转轴的确定。根据上面得出的最优价值和对应的最优权重值 p_{best} ,将基于 Bloch 球面建立搜索机制,即使其他权重值对应的量子比特在 Bloch 球面上绕着某一固定轴向最优权重值的方向进行旋转。这种旋转可同时改变量子比特的两个参数 θ 和 ϕ ,从而可更好地模拟量子行为,并可提高优化能力和优化效率。因为两点间的球面距离,以过该两点的球大圆上的劣弧最短。所以,为了使某一权重值 p_i 旋转之后能够逼近最优权重值,应该使这一权重值沿着球大圆的劣弧移动。假设向量 r 为 p_i 和 p_{best} 的向量积,由向量积的定义, r 的方向为垂直于 p_i 和 p_{best} 组成的平面,且其指向与 p_i 和 p_{best} 的指向满足右手定则,即右手四指以小于 π 的角度从 p_i 握向 p_{best} ,拇指指向即为 r 的方向,如图 2 所示。因此,若使 p_i 绕着轴 r 旋转,其轨迹恰好为 Bloch 球大圆上的劣弧,即旋转轴为 r 。

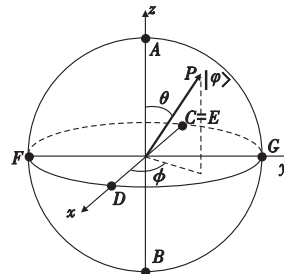


图1 一个量子比特的 Bloch球面表示

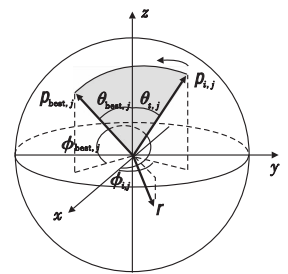


图2 量子比特的 Bloch球面旋转表示

d) 旋转操作。在量子行为进化算法中,提出一种量子比特在 Bloch 球面上的绕轴旋转方法,该方法可同时改变量子比特的两个参数。根据量子计算原理,在 Bloch 球面上,绕一个沿单位矢量 $n = [n_x, n_y, n_z]$ 的轴旋转角度 δ 的旋转矩阵为

$$R_n(\delta) = \cos \frac{\delta}{2} I - i \sin \frac{\delta}{2} (n \cdot \sigma) \quad (9)$$

其中: $\sigma = [\sigma_x, \sigma_y, \sigma_z]$ 。因此,根据确定下来的旋转轴,当前量子比特 $|\varphi\rangle$ 在 Bloch 球面上绕轴 r 向 $|\varphi_{\text{best}}\rangle$ 旋转 δ 的旋转矩阵为

$$R_n(\delta) = \cos \frac{\delta}{2} I - i \sin \frac{\delta}{2} \left(\frac{r \cdot \sigma}{\|r\|} \right) \quad (10)$$

当前量子比特 $|\varphi\rangle$ 在 Bloch 球面上向目标量子比特 $|\varphi_{\text{best}}\rangle$ 旋转 δ 角度的旋转操作为 $|\varphi\rangle = R_n(\delta) |\varphi\rangle$, 其中 $r = p_i \times p_{\text{best}}$, 旋转角度 δ 一般可取为 $\delta \leq 0.05 \pi$ 。旋转矩阵与原来权重值所对应的向量叉乘则会得到一个新的向量,再将这个向量转换为权重值,则该权重值就是得到的调整之后的新的权值,如式(11)所示。

$$|\varphi_{\text{new}}\rangle = \langle R_n(\delta) | \varphi \rangle = \left\langle \cos \frac{\delta}{2} I - i \sin \frac{\delta}{2} (n \cdot [\sigma_x, \sigma_y, \sigma_z]) \right| \varphi \rangle \quad (11)$$

3) 权重值更新 将除最优权重值之外的其他所有权重值按照调整过程进行调整,则都会得到新的权重值,将这些新的权重值与原来的最优权重值一起构成一组新的已知权重值,再重新计算它们对应的综合价值,找出得值最高的为新的最优价

值,再将其其他的权重值进行调整。如此反复,直到满足结束条件,即前后相邻两次得到的最优价值之差的绝对值小于 10^{-6} 即可。

4) 算法的复杂度 算法中的第 1) 步的时间复杂度为 $O(n)$, n 为分别赋予立即价值和未来价值的权重值数。第 2) 步中表示一个权重值的调整过程,由于调整步骤只是有限的几步计算,故时间复杂度为 $O(1)$ 。第 3) 步转第 2) 步是循环操作,它是先将 $n-1$ 个非最优权重值全部进行调整,共需时间复杂度 $O(n)$,然后再将所有的权重值进行下一次调整。所以,算法的最坏时间复杂度为 $O(N \times n)$,其中 N 表示权重值调整次数。

2.4 BQEA 的收敛性分析

在这里使用概率论中马尔可夫链的相关理论证明 BQEA 的收敛性。令 $Q_t = \{r'_1, r'_2, \dots, r'_n\}$ 为量子行为进化算法的第 t 次寻找最优值,其中初始值设定了 n 个不同的权重值,其中 $r'_i = |\varphi\rangle = \begin{bmatrix} \cos \frac{\theta_i}{2} \\ e^{i\phi_i} \sin \frac{\theta_i}{2} \end{bmatrix}$ 。关于量子行为进化算法的收敛性,有如下结论。

定理 1 BQEA 是以概率 1 收敛的。

证明 记 $br_k = \text{fit}_{r'_k \in Q_k} \{r'_k\}, i=1, 2, \dots, n$ 为 $Q_t = \{r'_1, r'_2, \dots, r'_n\}$ 中的最优权重个体,令 $I = \{i | i=1, 2, \dots, n\}$, Q_t^i 表示种群经过 t 次迭代后处于状态空间中的第 i 个状态。下面计算随机过程 $\{Q_t, t \geq 1\}$ 的一步转移概率 $P_t(i \rightarrow j) = P(Q_t^i \rightarrow Q_{t+1}^j)$ 。

因为 BQEA 中采用了精英保留策略,所以 $\text{fit}(Q_{t+1}^i) \geq \text{fit}(Q_t^i)$ 。故当 $i \notin I, j \in I$ 时, $P_t(i \rightarrow j) = 0$, 当 $i \in I, j \notin I$ 时, $P_t(i \rightarrow j) \geq 0$ 。

设 $P_t(i)$ 为种群 Q_t 处于状态 i 的概率,记 $P_t = \sum_{i \in I} P_t(i)$, 由马尔可夫链的性质可知, Q_{t+1} 处于状态 $j \in I$ 的概率为

$$P_{t+1} = \sum_{i \in I} \sum_{j \in I} P_t(i) P_t(i \rightarrow j) + \sum_{i \in I} \sum_{j \notin I} P_t(i) P_t(i \rightarrow j)$$

由 $P_t = \sum_{i \in I} \sum_{j \in I} P_t(i) P_t(i \rightarrow j) + \sum_{i \in I} \sum_{j \notin I} P_t(i) P_t(i \rightarrow j)$ 得 $P_{t+1} = P_t - \sum_{i \in I} \sum_{j \notin I} P_t(i) P_t(i \rightarrow j) \leq P_t$, 故 $\lim_{t \rightarrow \infty} P_t = 0$, $\lim_{t \rightarrow \infty} P(\text{fit}(br_t)) = 1 - \lim_{t \rightarrow \infty} \sum_{i \in I} P_t(i) = 1 - \lim_{t \rightarrow \infty} P_t = 1$, 即 BQEA 是以概率 1 收敛的。

3 实验设计及数据分析

选取某油田科技文档库、项目专题库和 FTP 文档共享系统的网站站点进行搜索系统仿真实验。系统运行参数设置如下:以“石油勘探”为特定主题,初始 URL 的种子站点数量为 15;初始权重值个数 n 为 9,分别赋予其立即价值 0.1, 0.2, ..., 0.9, 则未来价值有其对应的值为 0.9, 0.8, ..., 0.1。反复实践结果证明, n 太大则影响计算效率,太小则又影响结果的精确性;旋转角度 δ 取 0.02 π 。在这里设计的聚焦爬虫分别采用三种不同的搜索策略,即基于立即价值的搜索策略、基于未来价值的搜索策略和本文所提出的 BQEA 搜索策略来在线检索与主题“石油勘探”相关的文档并统计其文档数,然后分别计算出三种搜索策略的检索查准率和查全率并对它们进行比较。仿真评价标准采用查全率和查准率,计算为

$$P = \frac{\sum p_i}{\sum p_i^*}, R = \frac{\sum p_i}{\sum p_i^*} \quad (12)$$

式中: p_i 是指搜索结果中与“石油勘探”相关的文档数, $\sum p_i$ 是针对“石油勘探”主题搜索出来的总文档数, $\sum p_i^*$ 是从三个系统中统计得到的与“石油勘探”相关的总文档数。

在图 3 中分别表示三种爬行算法在爬行阶段具有不同的查全率。基于立即价值的搜索策略由于开始在预设的种子站点集中搜索页面,所以搜索初始获得的查全率与基于未来价值的搜索策略差不多,但在通过隧道网页搜索主题页面时却容易迷失搜索方向。而基于未来价值的搜索策略在通过隧道网页搜索主题页面时具有很好的预见指导作用,可以很快定位主题网页所在的主题区域位置,进而在搜索初期就得到了很好的回报率。基于 BQEA 的智能搜索策略由于其考虑了立即价值和未来价值,在检索过程中动态改变二者所占比重,因此其检索性能增长很快,并明显优于其他两种搜索策略。根据仿真结果有效说明了本文所提智能搜索算法的有效性。

从图 4 中可以看出,基于 BQEA 的网络爬虫的查准率显然高于基于立即价值和未来价值的查准率,主要原因就是该算法在每次选择链接时不仅选择网页的立即价值,还会考虑网页的未来价值,并在系统运行过程中动态改变二者所占比重,能对网页的未来价值进行预测,进而避免了局部最优,克服了主题漂移的缺点。

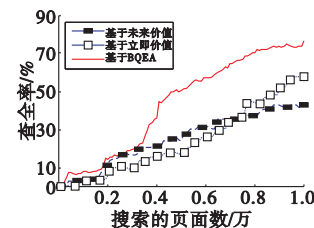


图3 三种搜索策略查全率对比

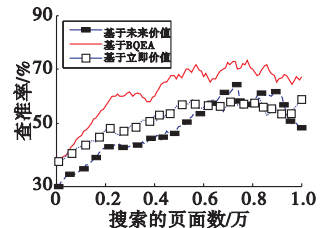


图4 三种搜索策略查准率对比

4 结束语

为了提高对特定主题提取上的查全率和查准率,在对网页链接进行判断时不但要考虑其立即价值,也要充分考虑网页链接的未来价值,在此基础上本文提出了量子行为进化算法,实现动态改变网页链接的立即价值和未来价值在其综合价值中所占的比重,进而指导聚焦爬虫考虑哪种价值的重要程度。通过仿真实验测试该搜索算法在针对某特定主题搜索时,相比较而言有较好的查全率和查准率,具有一定的意义和实用价值。

参考文献:

- [1] 倪贤贵. 聚焦爬虫技术研究 [D]. 无锡: 江南大学, 2008.
- [2] 李士勇, 李盼池. 量子计算与量子优化算法 [M]. 哈尔滨: 哈尔滨工业大学出版社, 2009: 71-92.
- [3] 李学勇, 欧阳柳波, 李国微. 基于模拟退火机制的网络蜘蛛搜索策略 [J]. 湖南理工学院学报: 自然科学版, 2004, 17(2): 60-63.
- [4] 黄胜根. 智能垂直搜索引擎的研究与设计 [D]. 重庆: 重庆大学, 2010.
- [5] 吴安清, 张颖江, 涂军. 主题搜索 ROBOT 综合爬行策略的研究 [J]. 武汉理工大学学报, 2006, 28(2): 74-75.
- [6] CHO J, GARCIA-MOLINA H, PAGE L. Efficient crawling through URL ordering [J]. Computer Networks, 1998, 30(1-7): 161-172.
- [7] 赫枫龄, 左万利. 利用超链接信息改进网页爬行器的搜索策略 [J]. 吉林大学学报: 信息科学版, 2005, 23(1): 59-63.
- [8] 童亚拉. 自适应动态演化粒子群算法在 Web 主题信息搜索中的应用 [J]. 武汉大学学报, 2008, 33(12): 1296-1299.
- [9] 李学勇. 基于巩固学习的网络蜘蛛搜索策略研究 [D]. 长沙: 湖南大学, 2003.
- [10] LI Pan-chi, LI Shi-yong. Quantum-inspired evolutionary algorithm for continuous spaces optimization based on Bloch coordinates of qubits [J]. Neurocomputing, 2008, 72(1-3): 581-591.