遗传算法在主题 Web 信息采集中的应用研究

唐 志[1] 王成良[2]

(重庆大学计算机学院 重庆 400044)[1] (重庆大学软件学院 重庆 400044)[2]

摘 要 传统的基于本地搜索算法的信息采集系统存在诸如主题漂移和采集结果局部最优等问题。在深入研究 Web 拓扑结构基础上,利用网络蜘蛛的在线状态,提出了基于全局信息的、动态综合了链接的立即回报价值和未来回报价值的遗传算法。通过此算法,利用元搜索技术可进一步提高网络蜘蛛的性能,具有更高的查全率和查准率,能够较好地解决现存问题。

关键词 网络蜘蛛,遗传算法,Web 社区,信息采集

Research of a Focused Crawler Using Genetic Algorithm

TANG Zhi[1] WANG Cheng-Liang[2]

(College of Computer Science, Chongqing University, Chongqing 400044)[1]

(School of Computer Software, Chongqing University, Chongqing 400044)[2]

Abstract Traditional focused crawler uses local search algorithms. It causes the problems of 'topic drift' and 'partially most superior'. Based on the knowledge of Web structure and web crawler's online status and meta-search technology, we proposed a new global search algorithm—genetic algorithm, which synthesizes the linkage's immediate value and future value dynamically. Our experiments show that the new algorithm has better recall rate and precision.

Keywords Genetic algorithm, Web spider, Web community, Information retrieve

1 引言

据统计,目前 Web 上的文档个数已超过 30 亿,并且以每天 750 万个的速度增长。面向所有用户建立的通用搜索引擎已不能满足人们的需要。因此,针对特定领域、面向特定用户的专业搜索引擎应运而生。如何全面和准确地搜集到特定领域内的相关内容,是主题 Web 信息采集领域的研究重点。传统的网络蜘蛛(Web spider)[3]一般使用一些启发式的规则,每次选择"最有价值"的链接进行优先访问(Best-N-First)。链接价值评价算法可以分为两类:基于立即回报价值的评价算法与基于未来回报价值的评价算法。

基于立即回报价值(以下简称立即价值)的评价方法主要是依据搜索时"在线"获得的文本或 Web 结构信息来对链接所指页面的重要性(或称立即价值)进行预测。依据评价对象的不同,可将其分为基于内容的评价算法[4,5]和基于链接的评价算法[6,7]两类。基于立即价值的评价方法的缺陷是不能预见以后要搜索页面的重要性,容易导致局部最优。

基于未来回报价值(以下简称未来价值)的评价方法主要利用经验信息来预测链接在未来搜索中的回报,并选择未来价值最大的链接进行优先搜索,其中使用最广泛的是利用巩固学习(reinforcement learning)的方法[8,9]实现链接未来价值的评价。这类评价方法的问题是忽略了网络蜘蛛在实际搜索时的"在线"状态,欠缺灵活性,容易引起主题漂移(topic drift)。

本文在分析了上述基于本地信息的搜索策略的不足后,提出了一个遗传算法,它利用了 Web 的拓扑结构,能在信息采集的过程中根据网络蜘蛛的在线状态动态地改变立即价值和未来价值在搜索策略中的权重,并利用元搜索技术(meta-search)改进网络蜘蛛的性能。

2 Web 拓扑结构

在 Web 中存在着一种被称为 Web 社区(Web communities)的结构,即在 Web 上的网页自然地组成各种不同的链接结构,每一个链接结构称为一个"Web 社区",其中的成员页面都大致与某一主题相关[10]。这样,进行主题信息采集的过程可以被看作是对所有主题相关的 Web 社区中的页面进行采集。在 Web 中存在着两种主要的 Web 社区结构。

第一种 Web 社区结构:某些领域型网站内的页面很少有外链接指向其它的相关页面,比方说各种 B2B 或 B2C 的商务型网站。对这些网站来说,拥有共同或相似主题的其它网站往往是它们商业上的竞争对手,提供链接,不符合它们的商业利益。我们借鉴了信息检索领域内的"协同引用(co-citation)"的概念来发掘这样的相关性——被协同引用的一系列网页拥有相同的一组父结点,被相同的一组父结点引用的次数越多的网页之间越有可能具有同一主题。

第二种 Web 社区结构:一般而言,主题相关的 Web 社区之间总是通过大量主题无关的链接彼此相连。在最近的研究中,Bergmark 分析了在 Web 上的 50 万个页面后发现这样的链接长度在 1～12 之间[11]。当网络蜘蛛处于主题相关的 Web 社区时,采用基于立即价值的链接评价算法能取得较好的效果。而当网络蜘蛛需要从一个主题相关的 Web 社区跨越到另一个主题相关的 Web 社区时,基于未来价值的链接评价算法较为实用[12]。根据网络蜘蛛的在线状态,动态地调整链接的立即价值和未来价值的权重,才能有效地提高网络蜘

唐 志 硕士生,研究方向为网络数据库与智能搜索;王成良 副教授,研究方向为网络及数据库技术、Web 技术等。

蛛的性能。

我们将在遗传算法的框架内实现链接价值的动态调整。

3 遗传算法在主题 Web 信息采集中的实现

3.1 遗传算法的分析研究

遗传算法(genetic algorithm,简称 GA)是 20 世纪 70 年代由美国的 Holland 提出的模仿生物进化过程的优化方法,它的主要思想是基于 C. R. Darwin 的生物进化论和 G. Mendel 的遗传学,它结合了 Darwin 的适者生存和随机交换理论。适者生存理论消除了解中的不适应因素,随机交换理论利用了原有解中的已有知识,从而加速了对优化解的搜索过程。

作为一种通用的自适应随机搜索算法,遗传算法还存在早熟和收敛慢两个问题。造成这两个问题的原因在于"种群多样性"与"选择压力"之间的矛盾[13]。选择压力过大,种群中的次优个体将迅速地被抛弃,容易导致早熟,对应于传统的信息采集领域中局部最优的问题;而若只重视种群多样性,大量次优或不优秀个体的存活则容易造成算法的收敛速度缓慢,使遗传算法收敛不到最优解,对应于信息采集领域内主题漂移的问题。

进一步的研究表明[14],在传统的 GA 中,变异是指对个体少量性质的改变。它是与当前种群状态无关的变异,且变异后的新个体仍保留了原母体中的大量性质。这种变异有其不利于寻找全局最优解的地方:如果全局最优的许多性质都与当前种群中的各个个体的大部分性质不同,那么它往往会在一个局部最优解区域附近徘徊,最终导致求解失败。

为了解决以上问题,本文采用生物学中 ESS(Evolution Stable Strategy,进化稳定策略)[15]的概念,通过引入一个平衡因子来平衡"种群多样性"与"选择压力"之间的矛盾。本质上讲,主题 Web 信息采集可以被看作为一个多峰函数求最优解的过程[16],多峰系数 k 等价于与主题 s 相关的短语集的大小 m。设在第 n 代群体中最优个体总数为 N,引入平衡因子 $b=N/m$,即在最优个体为 N 的群体中加入 b 个次优的个体。这些次优个体的性质不依赖于其母体的性质,这样就扩大了遗传算法寻求最优解的范围。同时,种群中的次优个体数与最优个体数维持一定的比例均衡,保证了一定的选择压力。

由于在实际操作中,Web 文档很难转化为二进制编码,传统的 GA 运算也不能直接使用。这里主要借鉴遗传算法作为一种全局算法的思想,以解决传统的主题 Web 信息采集中面对的问题。本文采用串行运算的遗传算法(sequential genetic algorithm,SGA),使用的遗传算子有如下 3 种:选择算子(selection),交换算子(crossover)和变异算子(Mutation)。

3.2 链接价值的定义

为了表述方便,下面给出几个遗传算法中将要使用的链接价值的相关定义。

定义 1(链接的立即价值) 给定搜索主题 s,设页面 p 中有一链接 a,若 a 所指向的页面 q 与主题 s 相关,则称页面 q 与主题 s 相关;根据立即价值的评价算法来预测链接 a 所指向页面 q 与主题相关的程度,称链接 a 具有与主题 s 相关的大小为 $I_s(a)$ 的立即价值。

可以使用向量空间模型(VSM)[17]和 Hilltop 算法[18]来评价链接的立即价值。

定义 2(链接的未来价值) 给定搜索主题 s,设页面 p 中有一链接 a,若 a 所指向的页面 q 与主题 s 无关,但可由 q 依次访问若干页面后获得与 s 相关的页面 r,则称页面 q 具有与 s 相关的未来价值。根据未来价值的评价算法来预测指向页面 q 的链接 a 与主题相关的程度,称链接 a 具有与主题 s 相关的大小为 $M_s(a)$ 的未来价值。

可以使用巩固学习[19]的方法来评价链接的未来价值。

定义 3(链接的综合价值) 给定搜索主题 s,设页面 p 中有一链接 a,a 的相对于 s 的立即价值为 $I_s(a)$,a 的相对于 s 的未来价值为 $M_s(a)$,则 a 的相对于 s 的综合价值为:

$$S_s(a)=dI_s(a)+(1-d)M_s(a) \quad (1)$$

其中,$d(0\leqslant d\leqslant 1)$为一动态权值,根据网络蜘蛛在线获得的 Web 状态信息动态地调整。

3.3 遗传算法的实现

本遗传算法分为 5 个步骤,其实现过程如图 1 所示。

该算法具体描述如下。

① 初始化:首先确定每一代种群(generation)的大小、选择率、交换率、变异率、搜索主题 s,主题相关的词典 L(lexicon)、以及种子 url 的选择。这里,种子 url 选取与主题 s 相关的领域内符合 Hilltop 算法要求的得分较高的专家文档。初始化完成后,种子 url 所指向的网页被网络蜘蛛采集,它们将作为遗传算法的第 1 代种群而被保留。

② 选择算子:这一步的目的是在第 n 代种群中筛选与主题 s 相关的文档集。使用 VSM 模型来计算每一个文档与主题 s 的相似度,相似度越高的文档其适应性(fitness)就越强,也就越有可能在这一代群体中存活(survive)。计算完这一代中每个文档的适应性之后,使用一个随机过程来选择适应度(fitness value,F)大于给定阈值的 N 个文档。为了避免遗传算法的早熟问题,这里取 $N\geqslant 256$,增强群体中个体的多样性。存活下来的文档作为与主题 s 相关的文档放入数据仓库中,其它的适应性小于给定阈值的文档被当作是与主题 s 无关的文档放入中间数据库,它们将在下一步用于判断有无"协同引用"的情况。

③ 交换算子:这一步的目的是计算存活文档中每一个链接的综合价值,并利用非存活文档判断协同引用。初始的情况下,链接的综合价值 $S_s(a)=dI_s(a)+(1-d)M_s(a)$,这里 d 的值初始化为 0.50,以示无偏向。计算数据仓库中每一个文档 p 的每一个链接 a 的综合价值,然后按照综合价值的大小顺序将链接放入网络蜘蛛的搜索队列当中。接着,取出中间数据库中与主题 s 无关的文档,判断是否存在"协同引用"的情况,将符合协同引用条件的 *url* 集加入到网络蜘蛛的搜索队列中。

④ 变异算子:这一步的目的是利用在线获得的信息判断网络蜘蛛所处的 Web 状态,对综合价值中的权值 d 进行调整,并使用元搜索技术进一步提高网络蜘蛛的性能。设 $\bar{F}(n)$ 为遗传算法中第 n 代页面的平均适应度,当 $|\bar{F}(n)-\bar{F}(n-1)|>\xi(\xi$ 为一阈值)时,就认为网络蜘蛛的"在线"状态发生了改变:若 $\bar{F}(n)-\bar{F}(n-1)>0$,则网络蜘蛛处于与主题 s 相关的 Web 社区。这时,就适当增加权值 d,使基于立即价值的评价算法获得更多的机会;而当 $\bar{F}(n)-\bar{F}(n-1)<0$ 时,网络蜘蛛是处于与主题 s 不相关的 Web 社区或者正处于两个主题相关 Web 社区的过渡区域。这时,适当减少权值 d,使基于未来价值的评价算法获得更多的机会,能够指导主题网络蜘蛛快速越过与主题 s 不相关的 Web 社区。设 $\bar{F}(n)-\bar{F}(n-1)=\lambda$,对权值 d 的调整根据以下公式:

$$d=(1+\frac{\lambda}{1+|\lambda|})d \quad (0<d<1);\text{若 } d\geqslant 1,\text{则 } d=1。 \quad (5)$$

同时，使用元搜索(meta-search)技术来平衡种群多样性与选择压力之间的矛盾，进一步帮助网络蜘蛛跨越主题无关的Web社区。方法是在词典L中随机选取与主题s相关的1～3个短语，将其发给本领域内的专题搜索引擎进行查询，通过引入平衡因子$b=N/m$，将查询返回的前b个结果加入到主题网络蜘蛛的搜索队列中。这里，由遗传选择与交换得到的个体为最优个体，而由元搜索得到的个体为次优个体，这样的次优个体不依赖于其母体的特性，并且使种群中最优与次优个体的数量保持一定的平衡。

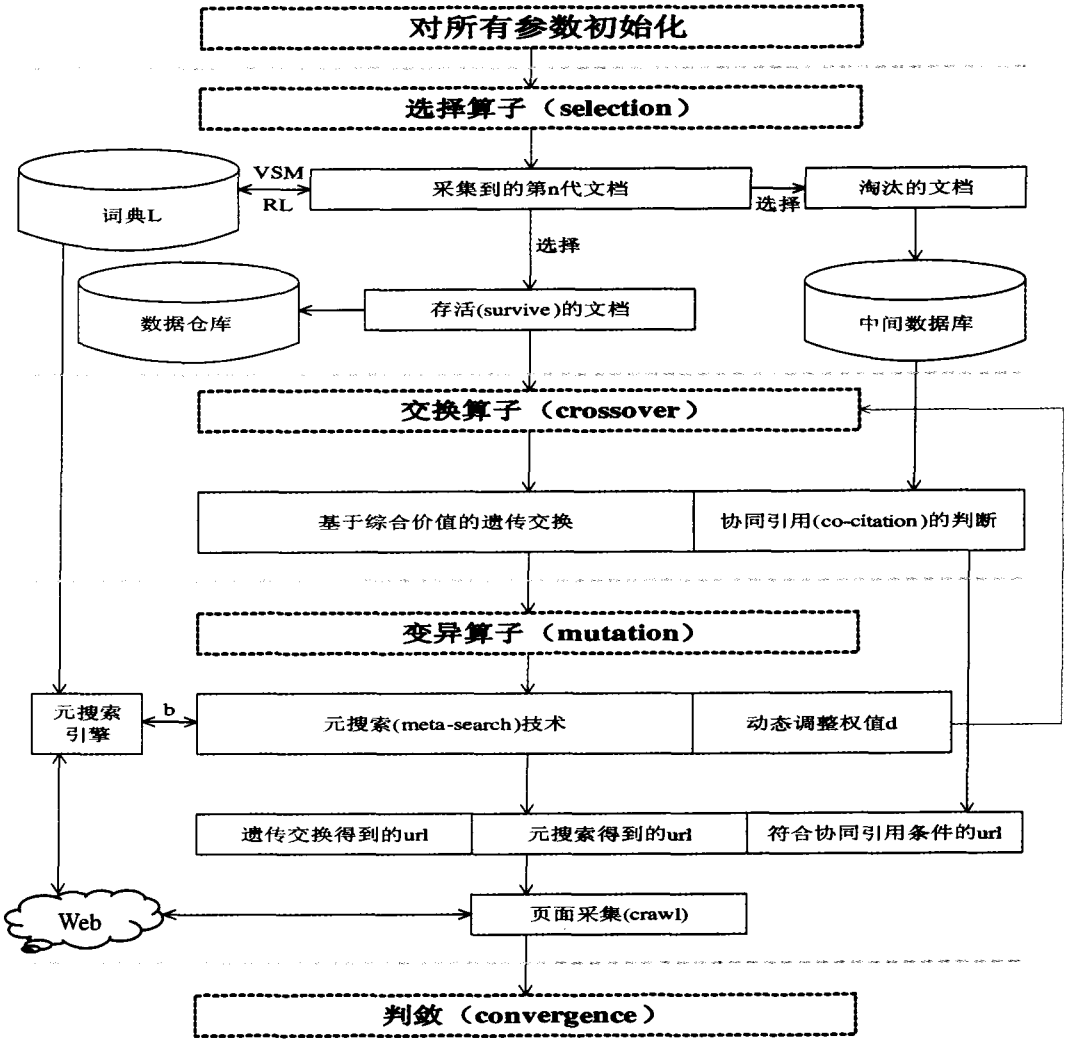

图1　遗传算法的实现步骤

⑤ 判敛(convergence)：循环上述②～④步，直到数据仓库中的文档数量达到给定的上限或$\bar{F}(n)$低于给定阈值。

4　算法性能比较

为了进行算法性能的比较，本文实现了基于立即价值且采用Best-N-First搜索策略、基于未来价值和使用遗传算法的3种网络蜘蛛，选择“天文学”为主题s，评价的标准是文本挖掘领域中的两个经典指标[20]：查准率(precision)与查全率(recall)。实验分为两步：

(1)搜索的目的是尽可能准确地采集与“天文学”相关的页面和论文等资料，重点测试3种采用不同搜索策略的网络蜘蛛信息采集的查准率；

(2)分别选取了MIT，Princeton，Oxford，Yale4所大学的网站，测试3个蜘蛛的查全率。相关性计算仍然采用向量空间模型VSM，其中查准率(P)与查全率(R)分别按$P=\sum_{p\text{与}s\text{相关}} p/\sum p$和$P=\sum_{p\text{与}s\text{相关}} p/\sum p^*$计算。其中，$\sum_p$是所有已采集的Web文档；$\sum_p{}^*$是从从上述4所大学经统计得到的与“天文学”相关的所有论文总数，一共13046篇。

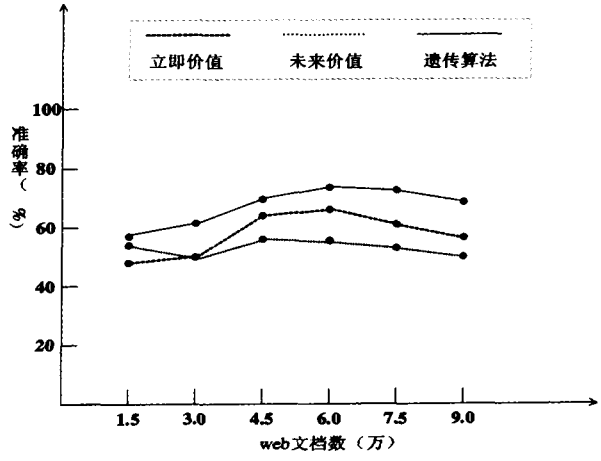

图2　三种蜘蛛的准确率(precision)

从 Yahoo 的分类目录中选取与"天文学"相关的网站和从 Google 搜索到的 PR 值为 9 以上的页面作为 3 个蜘蛛共同的种子 url 集，并利用 ZOOM ASTRONOMY 的在线天文学词典建立主题相关词典 L。设定 Best-N-First 算法 $N=256$，已经训练完成，建立了文本/未来价值映射库的巩固学习算法，遗传算法中的动态调整权值初始化为 $d=0.5$，Hilltop 算法中的阻尼因子 $f=0.85$，Web 文档数量上限为 90000。3 个蜘蛛的查准率如图 2 所示。

另外，以选择的 4 所大学的首页为种子 url 集，屏蔽掉与其非附属的页面，同样利用 ZOOM ASTRONOMY 的在线天文学词典建立主题相关词典 L。3 个蜘蛛的查全率如图 3 所示。

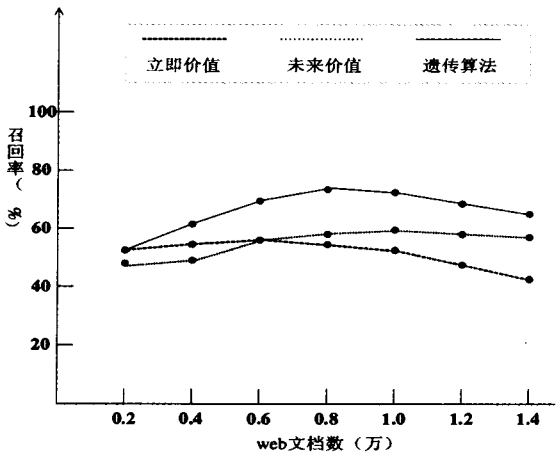

图 3　三种蜘蛛的查全率(recall)

由图 2 和图 3 可以看到，基于未来回报价值的巩固学习算法由于具有一定的跨 Web 社区能力，故其查全率比基于立即回报价值的 Best-N-First 算法高，但其查准率却比较低。这是因为它在跨越与主题无关的 Web 社区时取回了大量的与主题无关的文档，而在主题相关的 Web 社区中其性能又比较低；基于立即回报价值的 Best-N-First 算法虽然在主题相关的 Web 社区其性能较高，查准率也比较高，但由于其跨 Web 社区的能力较差，所以它的查全率较低，容易导致"局部最优"；而使用遗传算法的网络蜘蛛由于可以依靠"在线"信息动态调整立即回报价值和未来回报价值的比重，利用"元搜索"技术，故其查准率和查全率都明显比上述两种算法高。

结论　本文通过研究基于立即价值和未来价值的两种信息采集策略，提出了一个改进的遗传算法，利用网络蜘蛛"在线"的信息动态地调整两种策略的权重，并且采用元搜索技术进一步地增强网络蜘蛛跨越主题无关社区的能力，能在链接的立即价值和未来价值都较低的社区中扩大其搜索范围，解决了传统信息采集系统中存在的局部最优与主题漂移的问题。

参 考 文 献

1 Menczer F, Pant G, Ruiz M, et al. Evaluating Topic-Driven Web Crawlers [A]. In: Proceedings of 24th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval [C], 2001. 241～249

2 Ester M, Grob M, Kriegel H. Focused Web crawling: a generic framework for specifying the user interest and for adaptive crawling strategies [A]. In: Proceedings of 26th International Conference on Very Large Database(VLDB'01)[C], 2001. 527～534

3 Eichmann D. Ethical Web Agents. In: Proceedings of the 2nd International World Wide Web Conference, Chicago, Illinois, USA, 1994

4 Cho J. Crawling the Web: Discovery and maintenance of large-scale Web data [D]. Department of Computer Science, Stanford University, 2001

5 Hersovici M, Heydon A, Mitzenmacher M, et al. The shark-search algorithm -An application: Tailored Web site mapping [A]. In: Proceedings of 7th International World Wide Web Conference [C], 1998. 317～326

6 Borodin A, Roberts G O, Rosenthal J S, et al. Finding Authorities and Hubs From Link Structures on the World Wide Web [A]. In: Proceedings of 10th International world Wide Web Conference, ACM, 2001. 415～419

7 Cho J, Garcia-Molina H, Page L. Efficient crawling through URL ordering [J]. Computer Networks, 198, 30(1～7): 161～172

8 Rennie J, McCallum A. Using reinforcement learning to spider the Web efficiently [A]. In: Proceedings of the International Conference on Machine Learning(ICML 99)[C], 1999. 335～343

9 McCallum A, Nigam K, Rennie J, et al. Building Domain-Specific Search Engines with Machine Learning Techniques [A]. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace [C], 1999

10 Gibson D, Kleinberg J, Raghavan P. Inferring Web Communities from Link Topology. In: Proc. of the 9th ACM Conference on Hypertext and Hypermedia, Pittsburgh, Pennsylvania, USA, 1998

11 Bergmark D, Lagoze C, Sbityakov A. Focused Crawls, Tunneling, and Digital Libraries. In: Proc. of the 6th ECDL, Rome, Italy, 2002

12 Diligenti M, Coetzee F M, et al. Focused crawling using context graphs [C]. In: Proc. of the International Conference on Very Large Database(VLDB'00), 2000. 527～534

13 Dawkins R. The Selfish Gene. Oxford University Press, 1977

14 Pant G, Srinivasan P, Menczer F. Exploration versus exploitation in topic driven crawler [C]. In: Proc. of The WWW-02 Workshop on Web Dynamics, 2002

15 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey [J]. Journal Of Artificial Intelligence Research, 1996, 4: 237～285

16 Zhang J S, Xu Z B, Leung Y. The whole annealing genetic algorithms and their sufficient and necessary conditions of convergence. Science in China(Series E), 1997, 27(2): 154～164

17 Cho J, Garcia Molina H, Page L. Efficient crawling through URL ordering [J]. Computer Networks, 1998, 30(127): 161～172

18 Bharat K, Mihaila G A. Hilltop: A Search Engine based on Expert Documents. http://www.cs.toronto.edu/～georgem/hilltop/, 2004.10

19 Whitley D. The GENITOR algorithm and selection pressure: Why rank-based allocation reproduction trials is best. Schaffer J. ed. In: Proceedings of the 3rd International Conference on Genetic Algorithm. Los Altos: Morgan Kaufmann Publishers, 1989

20 Srinivasan P, Menczer F, Pant G. A general evaluation framework for topical crawlers. Information Retrieval, Submitted, 2003