

5-9,42

# 智能计算中的热点问题\*)

Some Focus Questions in Intelligence Computation

董 聪 郭晓华

TP18

(清华大学 北京 100084) (中国科学院模式识别国家重点实验室 北京 100080)

**Abstract** An introduction is made on some focus questions in intelligence computation, and a systematic analysis is done about the researching advances and existing common questions in the fields of the map and representation, generalization, topologic structure learning, global optimum approximation to discrete point set of multilayer feedforward networks and the genetic algorithm and simulated annealing algorithm. It is a significant feature of the present paper to find out differences by means of comparing the actual biology process, physics process or computing process with the corresponding simulating process or mathematical proof process and so to discover the cause of the focus questions about which a widespread discuss has being made. One hand, some interesting questions is presented, on the other hand, some preliminary researching thoughts and solving mehtods to these questions are also suggested for reference and deeper discussing.

**Keywords** Feedforward networks, Map, Presentation, Discrete point set, Generalization, Topology structure, Global optimization, Genetic algorithm, Simulated annealing algorithm

## 1 引言

人工神经网络研究热潮的再度兴起有其客观的历史背景。50年代以来,以符号机制为代表的经典人工智能形式体系取得了巨大的成功。80年代,当人们对过去30年的成就与问题进行反思时,却不得不承认,智能系统如何从环境中自主学习的问题事实上并未很好地解决。从逻辑的角度讲,以演绎逻辑为基础的算法体系可以发现新的定理,却无法发现新的定律。换句话说,基于符号推理的经典人工智能形式体系在机器定理证明方面的成功和在规则提取方面的失败同属必然。事实上,早在50年代,那些热衷于研究知识发现内在逻辑的人们就已经隐约地意识到,归纳逻辑,尤其是不完全归纳逻辑是通往知识发现的合理途径。Rosenblatt[1957]提出了感知机模型,首次将自组织和自学习的思想引入智能系统;Nilsson[1965]提出了含有隐节点的多层网络的构想;Minsky和Papert[1969]对感知机在处理非线性问题方面的局限性做了系统的分析,对多层网络是

否存在有效的学习方法持怀疑态度,在60年代末至80年代中期的10余年间,虽然也出现过一些新的人工神经网络模型,但终究未掀起狂风巨澜。80年代中期,经典人工智能的发展由辉煌转入相对停滞,这是一个充满惶惑、渴望变革的特定历史时期。Rumelhart, Hinton及Williams[1986]发现十余年前由Werbos发明的误差反传算法(Back Propagation error, BP)能够有效地解决多层网络中隐节点的学习问题,证明Minsky对多层网络可能不存在有效的学习方法的断言并不正确<sup>[1]</sup>,BP算法简洁的数学结构和明确的逻辑关系,以及Hinton等人在对算法进行诠释时使用的富有想象力和充满隐喻的语言,所有这一切在很大程度上有力地推进了整个人工神经网络研究热潮的再度兴起和迅速蔓延。

## 2 由精确表达到近似描述的转变及其问题

用一元函数的复合表示及逼近多元函数是数学中的一个重要问题。Hilbert[1900]猜想,多元连续函数不能用一元连续函数的复合表示。Hilbert的猜想

\*)国家自然科学基金(59505011,59778039)、863计划(863-2-443)、国家重点实验室基金和攀登计划重大项目资助课题。

在被著名数学家 Arnold 和 Kolmogonov 证伪。Kolmogonov [1957] 证明: 定义在紧致子集上的多元连续函数可用有限一元连续函数的复合精确表示。Kolmogonov 的工作奠定了多层前向网络映射能力数学证明的基础。Kolmogonov 工作的意义还不止于此, 他实际上也证明了, 借助于广义坐标变换, 偏微分方程可用常微分方程的有限复合精确表示, 后者正是力学界熟知的有限元法法和 Kantorovich 法及其延拓形式的数学依据。在 BP 算法被重新发现后的一段时间里, 仿真计算对推进人工神经网络的研究和应用起了十分关键的作用。大量的结果向人们显示: 多层前向网络可能具有实现任意非线性连续函数逼近的潜力。

将研究重点由存在性的精确表示转向构造性的近似描述是逐步实用化的体现, 学术界往往忽略了这种转变的存在。事实上, 这种转变对后续研究的影响极其深刻, 目前引起广泛关注的许多问题均由此而生。在 1989 年至 1994 年间, 由于以 Hornik 为代表的一批杰出数学家的加盟, 使研究工作取得了一系列令人鼓舞的进展。Hornik 等人证明<sup>[2]</sup>: 在一个相当宽的范围, 3 层前向网络具有以任意精度逼近定义在紧致子集上的任意非线性连续或分段连续函数的能力。与 Kolmogonov 不同, Hornik 等人的研究工作有两个突出的标志: 一是不断放松对隐节点函数的限定条件; 二是对隐节点的数目不做限制。Hornik 等人逐渐向人们展示: 在隐节点的数目不受限制的条件, 对函数逼近而言, 隐节点函数的有界性要求是必要的, 单调递增的限制条件并非必要。放松对隐节点函数的限定条件及对隐节点的数目不做限制意味着生物学约束的弱化, 这一点并不是生物学家们所期望的, 也不是以探索智能形成机制为目标的一些 AI 研究者的初衷。不论生物学家是否对 Hornik 等人的研究结果感兴趣, 也不论一些 AI 研究者的初衷到底是什么, 有一点是可以肯定的, 即人们对 Hornik 等人研究结果的正确性和适用性并未产生过怀疑。将函数逼近论作为前向网络数学逼近理论基础的思想从此开始盛行。

90 年代, 与 Hornik 等人的证明工作并行展开的是不断增多的工程应用, 于是, 函数逼近论中从未出现过的一个概念开始在前向网络的应用中频繁地出现, 这个概念就是泛化 (Generalization)。前向网络泛化问题的出现使人们认识到, Hornik 等人的证明曾经带给他们的信心看来多少有些脆弱。人们开始反思, 函数逼近论是否足以构成前向网络数学逼近

的理论依据? 我们认为, 经典函数逼近论事实上并不足以构成前向网络数学逼近的理论依据, 原因是: 经典函数逼近论研究的主要是三类问题。第一类问题是线性赋范空间  $X$  中的一个确定的元素  $x$ , 借助于  $X$  的某个给定的子集  $F$  来逼近的问题; 第二类问题是线性赋范空间  $X$  中具有某些已知性质的  $X$  的子集  $\mu$ , 借助于  $X$  的某个给定的子集  $F$  来逼近的问题; 第三类问题通常涉及寻找 (在一定意义下的) 最佳逼近集和最佳逼近方法。很明显, Hornik 等人研究的实际上是第二类问题。上述三类问题的共同点在于, 被逼近对象 (无论它是一个元素  $x$ , 还是一个子集  $\mu$ ) 的特性是已知的。而在多层前向网络的应用中, 多数情况下被逼近对象的特性是未知的。众所周知, AI 中的联结机制 (Connectionism) 优于物理符号机制 (Symbolism) 的一个最主要的特征就是它所具有的学习功能, 学习功能在人工神经网络研究热潮的再度兴起中起了决定性的作用。从认知的角度讲, 所谓学习, 就是对未知现象或规律的发现和归纳。因此, 经典函数逼近论的研究所基于的数学空间和多层前向网络的实际逼近过程所基于的数学空间的特性是不同的, 它们是两类不同性质的逼近问题, 具有完全不同的逼近机制。

Kolmogonov 研究的是精确表示而非逼近问题, Kolmogonov 只是改变了研究对象的表达形式, 但未改变它所具有的特性, 因此不存在泛化问题。Hornik 等人研究的是经典函数逼近论框架下的函数逼近问题。经典函数逼近论中, 虽然新的逼近表达式中的元素和被逼近对象中的有关元素的特性不同, 但由于被逼近对象的完整特性是已知的, 逼近过程在整个定义域内实现, 无延拓现象发生, 因此也不存在所谓的泛化问题。从数学的角度讲, 前向网络的泛化问题是在被逼近对象的完整特性未知的条件下, 由有限维离散点集间的映射关系向无限维连续点集间的映射关系延拓时产生的。

### 3 泛化问题的逻辑根源

上面我们解释了泛化问题的数学根源, 接下来进一步分析泛化问题的逻辑根源。

无论是联结机制还是物理符号机制, 其学习过程均表现为广义的归纳过程。联结机制以隐含的方式表示其知识, 而物理符号机制的知识表示是显现的。在归纳过程中, 不存在泛化问题, 泛化是在应用过程中出现的, 无论是联结机制还是物理符号机制, 其应用过程均表现为广义的演绎过程。联结机制以

隐含的方式进行推理,而物理符号机制以显现的方式进行推理。当演绎推理的结果超出了早先用于归纳学习的原始知识的具体描述时,泛化问题便产生了。

归纳有两种,即完全归纳方式和不完全归纳方式。由完全归纳方式得出的结论是事实而不是猜想(假说),由不完全归纳方式得出的结论是猜想(假说)而未必是事实。以完全归纳方式得出的结论为基础的演绎推理不会出现泛化问题,因为其演绎推理的结果必然是早先用于归纳学习的原始知识形式的某一种或某几种具体的描述,这是一种封闭的体系,用 AI 的术语来说,就是这种体系只有有限的记忆功能,而没有创造功能,泛化问题来源于以不完全归纳方式得出的结论为基础的演绎推理过程,这是一种开放的体系,它不仅具有有限的记忆功能,同时还具有创造功能。网络的泛化问题是伴随着网络的创造功能而产生的,泛化的结果既可能是新的发现,也可能是错误的推论。因此,泛化问题是任何开放体系所具有的一个最基本的问题,泛化问题并不局限于联结机制,物理符号机制同样存在泛化问题。

Brooks 题为“没有推理的智能”(Intelligence without Reason)的论文在 AI 领域产生了广泛的影响<sup>[3]</sup>,赞赏和批评接踵而至。Brooks 的文章使人们普遍产生了误解,似乎人工神经网络所表现出的智能行为仅仅源于反馈。我们认为,反馈在智能形成机制中起了重要作用,但不是全部作用;同时,不以显现的方式进行推理和没有推理过程发生是两码事。事实上,人工神经网络不是不进行推理,只不过推理过程未以显现的方式表达而已。原因很简单,没有推理就不会出现有目的的行动,没有行动就无法产生反馈信号,没有反馈信号的引导,所谓的自组织行为便无法完成。从这个意义上讲,我们认为 Nilsson 对 Brooks 工作的批评准确而恰当。Nilsson 在评价 Brooks 的工作时非常巧妙地引用了 Anden 的诗句:“不会推理的将在行动时消亡,不会行动的则因此而消亡”。

人工神经网络提供了一种揭示智能和了解人脑工作方式的合理途径,但两者之间并不存在一一对应的关系,要真正揭示智能的形成机制和了解人脑的工作原理,深入研究神经生物学的基本规律并在构建相应的人工模型时考虑生物学的基本约束无疑十分必要。

#### 4 新的启示与猜测

把目光从数学家虚拟的背景中移开,再次投向真实的生物系统时,我们惊奇地发现,在神经网络的研究中,生物学家们正在进行的工作比人们想象的更具有启发性和创造性。

Shatz 对发育中的大脑进行了深入的研究,他发现成熟的神经连接形式的出现在于轴突有选择地收缩及长出新的分枝的重新安排,轴突在其靶结构(靶位)内伸到不同地址(靶址),然后再通过竞争逐步消除初始选址的错误。视觉系统中的靶位对应于人工神经网络中的隐层,而靶位内的靶址对应于隐层中隐节点的位置。Shatz 研究的是视网膜节细胞的轴突在外侧膝状核内发育及成型的过程,同样的过程也发生在外侧膝状核内神经元的轴突在初级视觉皮层(亦称 V1 区)内的发育成型中<sup>[4]</sup>。我们将 Shatz 等人的发现以计算机模拟程序的方式再现出来,提出了多层前向网络中跨层连接的概念及网络拓扑结构压缩的通用算法<sup>[5]</sup>。对于著名的 2~7 位 Parity Mating 问题,我们的仿真结果表明,采用通用算法压缩后的网络结构为 N-N-1。也就是说,当学习样本以指数方式增长时,压缩后的网络结构,其隐节点的增长方式是线性的,人们原先担心的 NP 难问题并未出现。我们还发现,最大限度地压缩网络的拓扑结构是实现网络泛化功能的有效途径。

重新审视一下如式(1)表示的含有 n 个输入、1 个输出、M 个隐节点的 3 层前向网络:

$$f(X) \sim \sum_{m=1}^M \beta_m h_m(Z_m) = \sum_{m=1}^M \beta_m h_m\left(\sum_{i=1}^n a_{im} x_i\right) \quad (1)$$

如果把隐层的节点输出看作是经过变换后的广义坐标的话,则 3 层前向网络对离散点集的最佳逼近,其实质就是选择最合适的广义坐标  $h_m(Z_m)$ 。从式(1)可以看出,广义坐标的形成依赖于输入层到隐层的连接权矩阵  $a_{im}$  及隐层节点函数  $h_m$  的特定选择。 $a_{im}$  是无级可调的,因此我们得出结论,广义坐标的特性本质上主要取决于隐层节点函数的特定选择。White 认为,重要的是大量神经元之间的相互连接和作用,而神经元特性的特定选择在网络逼近中的作用是无关紧要的。Hornik 等人将 White 的观点在经典函数逼近论的框架下做了数学证明。上文说过, Hornik 等人的证明是基于被逼近对象的完整特性已知和隐节点的数目不受限制的理想条件下得到的,这一证明所基于的数学空间并不是网络的实际学习过程所基于的数学空间。现在则进一步发现,在

网络的实际学习过程中,由于学习样本是有限的,因此,网络中有效隐节点的数目存在上限( $\square N$ )。换句话说,不可能通过无限扩充隐节点数目的方式改进逼近效果,这意味着由 White 开始的、目前甚为流行的、认为神经元特性的特定选择在网络逼近中的作用不重要的观点事实上并不正确。

由于著名心理学家 Hebb 在 40 年代初期的杰出工作,科学家们已在下述问题上达成共识:动物的某些学习和记忆过程可通过突触特性的变化加以解释。在人工神经网络中,生物神经系统中的突触变化可用节点间连接权的变化加以表征。Hebb 在 40 年代初期的发现对后来人工神经网络学习方法的研究方向产生了深远的影响,突出表现是,目前所谓的神经网络学习方法,大多数研究的实际上是网络中连接权的合理变化规律。70 年代以来,现代生物学的研究成果进一步揭示,动物的学习和记忆过程不仅表现为突触的变化,而且突出地表现为神经细胞内部构造的变化,即人工神经网络中节点函数的特化。生物系统发育过程中神经细胞内部构造的变化已被哈佛医学院 Selkoe 的研究结果所证实。

上文证明,从数学上讲,多层前向网络对离散点集的最佳逼近能力主要取决于隐层节点函数的特定选择;现在则进一步证实,真实的生物系统确实存在调整神经细胞内部构造以适应环境变化的内在机制。

研究智能的形成机制,考虑生物学的基本约束是十分必要的。除此之外,许多人仅把前向网络作为一种合适的逼近工具来研究和利用,这种工具主义的思潮目前在工程界甚为流行。

从工具主义的角度讲,前向网络的学习过程更类似于多元非线性回归分析。其中,投影寻踪算法(Projection Pursuit Algorithm, PPA)中的一些思想值得借鉴。投影寻踪算法中的岭函数、岭函数总数、岭函数自变量的系数、岭函数权重和 3 层前向网络中的隐节点函数、隐节点总数、输入层至隐层的连接权系数和隐层至输出层的连接权系数之间存在一一对应的关系,在投影寻踪算法中,岭函数和岭函数总数自身也是学习的对象。

上文给出的网络拓扑结构压缩算法实际上就是一种隐节点总数的学习算法,而以径向基函数(Radical Functions)、特别是以小波函数(Wavelets Functions)作为隐节点函数的研究,也证明了隐层节点函数特性的特定选择对网络的逼近效果具有十分关键的作用。在工具主义思潮的影响下,我们给出

了 BP 算法的广义描述,证明基于广义 BP 算法的任何全样本前向网络权值修改方式都是收敛的,除此之外,我们还给出了前向网络学习的多种方式<sup>[6]</sup>。

## 5 全局优化的方式和途径

人工神经网络研究中,有 2 种随机优化算法非常流行,即 Metropolis 提出的模拟退火算法(Simulated Annealing, SA)<sup>[7]</sup>和 Holland 提出的遗传算法(Genetic Algorithm, GA)<sup>[8]</sup>。SA 模拟的是金属材料加温后的退火过程,GA 模拟的是生物的自然进化过程。从模拟对象看,两者几乎没有任何共同之处,但从算法本身的形式结构看,两者则极其相象。SA 借助于产生函数确保状态空间各点的概率可达性,通过接受算子模拟平衡态,通过接受算子的定向变化保证算法迭代进程的方向性;而 GA 则通过交换算子和突变算子的协同作用确保状态空间各点的概率可达性,通过选择算子的作用保证算法迭代进程的方向性。

从数学的角度讲,实现全局优化的随机算法可分为两大类:一类是通过遍历搜索的方式,如 SA 和 GA 等;另一类是通过定向推进的方式,如广义遗传算法等<sup>[5]</sup>。

退火是金属材料热处理的一种方法,其作用是使材料的成份更均匀,细化晶粒,增加韧性,因此,退火本身不存在全局最优化问题。SA 将组合优化问题中的状态定义为相格,将各状态的函数值定义为相格所对应的内能,将每次抽样所获得的微粒(微观状态)按照 Boltzmann 统计分布进行取舍,并将留下的微粒向对应的相格中投放,虚拟温度(Boltzmann 常数与温度归并为一体)则定义为系统中所有微粒所含内能的平均值。虚拟温度的定义方式和 Boltzmann 统计分布的引入,其目的就是模拟热平衡状态的形成过程,然后,通过控制虚拟温度的逐步降低,进一步模拟热平衡状态的定向转移过程。因此,SA 希望模拟的实际上是热平衡状态的形成及定向转移过程。

SA 和真实的退火过程之间有明显的差异:1)真实的退火过程中通常伴随有相变,也就是说,有限状态集合会发生变化,作为对比,组合优化问题中的状态集合是恒定不变的;2)真实的退火过程中没有平衡态,只有一系列拟平衡态的渐进推进,平衡态的出现意味着退火过程的终结。SA 通过以下几个相对独立的步骤模拟热平衡状态的形成过程:1)扰动状态的形成;2)将每次扰动所获得的新状态按照

Boltzmann 统计分布进行取舍;3)将留下的微粒向对应的相格中投放;4)使系统的平均内能在算法的反复抽样迭代过程中逼近于当时设定的虚拟温度。热平衡状态的定向转移过程是通过逐步降低虚拟温度,也就是通过降低系统平均内能的方式实现的。

从热力学的角度讲,热平衡状态的定向转移只是使系统各微观状态的内能所服从的随机分布的形态(概率密度函数的形态)发生定向变化,并不能产生新的状态。换句话说,SA 拟解决的是大量抽样中系统平均内能(或峰值状态)的最小化问题,并不是单个状态的最小化问题,而后者才是全局优化方法真正需要解决的关键问题。除此之外,SA 在具体操作上也存在一些问题。

从统计物理学的角度讲,所谓热平衡状态,就是—定温度下所有可能微观状态中出现几率最大的状态,因此,模拟热平衡状态需要两个条件:1)所有可能的微观状态是已知的;2)采用 Boltzmann 统计分布作为接受算子的前提是所有可能微观状态以等概率产生。如果说第2个条件只涉及技术难度的话,则第1个条件事实上无法满足,原因很简单,如果所有可能的微观状态是已知的,则最小值状态自然也是已知的,无进一步求解最小值之必要。鉴于以上原因,本文仅将 SA 作为求解全局优化问题的一种随机算法,而不赋予它相关的物理含意。从这个意义上讲,SA 经改造之后可用来解决系统中单个状态的最大或最小化问题,此时算法成败的关键在于,如何选择产生函数使状态空间各点的概率可达性得到保证,目前普遍关注的虚拟温度下降速率并不是问题的关键。可以证明,只要状态空间各点的概率可达性得到保证,采用确定性的接受算子保证算法迭代过程的方向性,系统中单个状态的最大或最小化问题便可直接解决,模拟热平衡状态的形成及定向转移过程并无必要。

同样,自然进化过程也不存在全局最优化问题。Jacob 认为,进化就象是个修补匠,它只能从当时所能得到的材料中,有选择地进行调整(使生物体适应环境),其能力受到以前的决定所限制。不仅如此,每个特定的能力看起来都象一个优美的设计,但实际上,它也许不能同更大—些的系统很好地整合在一起,并且可能同系统的整体布局不相匹配。我们认为,生物系统的演化从根本上说是动态的,对于每一个偶然的形态或环境上的变化,生物体皆用已经进化出来的一系列特征去应付生存考验,在此之前并没有一个经过内省的有意识的筹划。换句话说,进化

的产物都是分阶段局部优化的结果,我们不可能从单纯模仿进化过程发现解决全局优化问题的诀窍。GA 成败的关键在于,如何通过交换算子和突变算子的协同作用,确保状态空间各点的概率可达性。选择算子的作用在于保证算法迭代过程的方向性,因此,选择算子为概率算子的限制性要求并无必要。

关于 GA 和 SA 全局优化能力的证明,目前基本上是将算法的迭代过程描述为一个遍历的 Markov 过程,然后证明遍历的 Markov 过程以概率 1 收敛于全局最优解。事实上,如果 GA 和 SA 的迭代过程是一个遍历的 Markov 过程的话,算法的收敛性在算法出现之前百余年来统计学家们早已做了严格的证明。问题是,两类算法的迭代过程是不是,或者是不是必须是一个遍历的 Markov 过程。

从逻辑上讲,一种算法要想具备实现全局最优化的功能,它只需满足两个条件:1)它具有实现局部最优化的能力;2)它具有从一个局部最优状态向下一个更好的局部最优状态定向转移的能力。正是在这样一种思想的指导下,我们创立了广义遗传算法(Generalized Genetic Algorithm, GGA),并解决了多层前向网络的最简拓扑构造问题和全局最优逼近问题<sup>[10]</sup>。GGA 和前向网络最简拓扑构造算法均摒弃了包括经典 GA 和 SA 在内的传统随机优化方法普遍采用的遍历搜索策略,转而采用定向演化模式,定向演化模式不要求算法的迭代过程是一个遍历的 Markov 过程。遍历性要求状态空间的各点是互通的,而定向演化模式只要求状态空间的各点是单向可达的,后者只是前者的必要条件。换句话说,定向演化模式的实现条件宽松得多。下面简要分析一下 GA 的实际操作过程。

设组合优化问题中的状态总数为  $N$ , 状态集合为  $S_N$ , 目标状态为  $S^*$ 。GA 中每一代种群含有的个体总数为  $m$ , 第  $k$  代种群含有的状态集合定义为  $S_k^{(m)}$ , 则经过  $p$  次繁衍后 GA 所历经的状态集合为  $S_k^{(m)} = \bigcup_{i=1}^p S_i^{(m)}$ 。全局优化算法的目的就是尽可能地减少繁衍次数  $P$ , 使得当  $P \ll P_0$  时, 存在  $S^* \in S_k^{(m)}$ 。由于  $m < N$ , 因此  $S_k^{(m)} \subset S_N$ , 即  $S_k^{(m)}$  是  $S_N$  的一个真子集。Markov 过程模型适用的一个前提条件是  $S_k^{(m)} = S_N$ , 换句话说,关于 GA 全局优化能力的数学证明所基于的数学空间并不是 GA 的实际操作过程所基于的数学空间。此外,经典 GA 的证明策略也存在—些问题:经典 GA 实现  $S^* \in S_k^{(m)}$  的方式是通过无限

(下转第 42 页)

式,如果输入的关键字含大写字母,则只匹配当前的形式;跟 Infoseek 一样,AltaVista 也采用全文索引的方式,索引项包括了虚词,支持简单的自然语言查询。AltaVista 采用 Digital 公司开发的 64 位运算技术,其搜索速度比一般的搜索引擎要快。

STRET-Search 依图 4 所示的框架结构捆绑了上述三个索引系统的能力,在知识库的设计中,参考了 SavvySearch 的 Meta-Index 的方法,建立查询关键字与 Search Engines 的关联,但并不是用这种相关性来实行对 Search Engines 的动态任务分配。STRET-Search 更注重对查询结果的调节。我们也许会注意到,各个索引系统对用户输入的查询关键字常常能得到数千、数万个匹配结果,然后根据相关度的大小分批返回,用户一般没有耐心通过“Next 20”的超连接把所有结果都看完,一般来说,前面 50 条索引匹配项对用户是最有用的。有的查询页面允许用户定制一次返回的数量,但用户并不知道 Search Engines 对当前查询的关键字是否有好的表现,一次返回过多的结果很不利于用户的选择,而且造成时间上的延时。STRET-Search 的任务处理器根据 Search Engines 与查询关键字的相关性,动态控制各个 Search Engines 一次返回结果的数量,使用户更加方便、迅速地选出其所需要的 Web 文档。

在结果返回的处理上,STRET-Search 采用表决的策略,三者同时匹配到的文档被认为相关性最好,最先显示,两两相同者次之,最好分别显示各自不同的查询结果。

**结束语** Web 上的文档查询是一种很不精确的查询,Meta-Searcher 大大增强了这种查询的能力,但并非索引信息系统用得越多就越好,太多的不同“观点”的查询结果往往会干扰用户的选择,而且依当前网络资源的现状,使用过多的索引系统已被认为是一种不道德的行为。本文认为,选择 3~5 个适当的索引系统是合适且有益的。

### 参考文献

- 1 阳小华,周龙骧. World Wide Web 的索引与查询技术. 计算机科学,1997,24(3):29~34
- 2 沈达阳,林作铨. Internet 上的软件 Agent. 计算机科学,1997,24(2):14~19
- 3 Available at: <http://www.albany.net/allinone/>
- 4 Available at: <http://cuiwww.unige.ch/meta-index.html>
- 5 Available at: <http://www.nexor.com/public/cust/cust.html>
- 6 Available at: <http://www.metacrawler.com/>
- 7 Available at: <http://www.cs.colostate.edu/~dreiling/smartform.html>
- 8 Erk Selberg Oren Etzion. The MetaCrawler Architecture for Resource Aggregation on the Web. Available at: <http://www.cs.washington.edu/homes/selberg>
- 9 Dreiling D. Description and Evaluation OF Meta-Search Agent. In partial fulfillment of the requirements for the degree of Master of Science. Colorado State University, Fort Collins, Colorado, Summer 1996
- 10 Available at: <http://www.all4one.com/>

(上接第 9 页)

制地增大  $p$ , 使得  $S_N^{(p)} = S_N$ , 此即所谓的遍历搜索。很显然,用无穷次繁衍来填满一个原本有限的集合  $S_N$  的做法是低效率的。低效率的原因在于,为实现  $S' \in S_N^{(p)}$ , 使  $S_N^{(p)} = S_N$  并无必要。

SA 存在的问题类似,限于篇幅不再赘述。

### 参考文献

- 1 Rumelhart D E, et al. Learning Representation by Backpropagation Errors. Nature, 1986, 323(6188):533~536
- 2 Honik K. Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks, 1991, 4:551~557
- 3 Brooks R. Intelligence without Reason. IJCAI-91, 1991. 42.

- 4 董聪. 大脑、知觉模型和计算机模拟. 科技导报, 1997(7):7~10
- 5 董聪. 多层前向网络的逼近机理与拓扑结构学习方法. 通讯学报, 1998, 19(3):29~34
- 6 董聪, 刘西拉. 广义 BP 算法及网络容错性和泛化能力的研究. 控制与决策, 1998, 13(2):120~124
- 7 Kirkpatrick S. Optimization by Simulated Annealing. Quantitative Studies. J. Statis. Phys., 1984, 34:975~986
- 8 Holland J H. Adaptation in Natural and Artificial Systems. MIT Press, 1992
- 9 董聪, 郭晓华. 广义遗传算法的逻辑结构及全局收敛性的证明. 计算机科学, 1998, 25(6)
- 10 董聪. 前向网络全局最优化问题研究. 中国科学基金, 1997(1):23~29