

蚁群算法在分布式智能答疑系统中的应用研究

郑耿忠¹, 刘秋梅²

(1 韩山师范学院 数学与信息技术系, 广东 潮州 521041; 2 韩山师范学院 图书馆, 广东 潮州 521041)

摘 要: 针对分布式智能答疑系统的特点, 在对基于蚁群优化的分类规则发现算法研究的基础上, 提出了分布式蚁群智能分类算法, 并对该算法在分布式智能答疑系统中的应用进行了分析和测试. 实践证明, 该算法能有效地对分布式智能答疑系统的范例库进行数据挖掘, 进一步提高智能答疑系统检索的效率和智能性, 使分布式智能答疑系统的应用进入一个新台阶.

关键词: 分布式智能答疑系统; 蚁群算法; 数据挖掘; 范例库

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2010)06-0050-04

The Study of Ant Colony Algorithm's Application in Distributed Intelligent Question Answer System

ZHENG Geng-zhong¹, LIU Qiu-mei²

(1 Department of Math and Information Technology, Hanshan Normal University, Chaozhou 521041, China;

2 Library of Hanshan Normal University, Chaozhou 521041, China)

Abstract: According to the characteristic of Distributed Intelligent Question Answer System(IQASD), the paper afford intelligent classify algorithm of distributing ant colony based on the study of ant colony optimization for classify rule discovery, analyzing and testing the algorithm's application in case base's data mining of IQASD. Practices prove that the algorithm can mining data of IQASD's case base in effect, make IQASD's searching more efficiency and intelligently, promote IQASD's application enter a new step.

Key words: IQASD; ant colony algorithm; data mining; case base

1 引言

随着远程教育的不断发展, 答疑系统已从单一方式向分布式智能答疑系统(Distributed Intelligent Question Answer System, IQASD)方式发展, IQASD是采用 Agent 技术, 构建分布式环境下多站点间的智能答疑平台, 即通过构建一个分布式联合虚拟答疑平台, 建立有多部门参加的、具有实际服务能力的、可持续发展的分布式联合虚拟答疑服务体系, 以本地化运作为主, 结合分布式、合作式的运作, 实现范例库、学习中心共享共建的目的.

在 IQASD 中通常数据是分布的、自治的、异构

和动态的, 若采用已有的集中式的数据挖掘方法对答疑系统范例库中的数据进行挖掘是不能满足分布式环境要求的. 目前, 针对 IQASD 中范例库进行数据挖掘的分布式算法有许多, 但不足之处在于各范例库之间缺乏相互联系, 从而使得分析结果不准确, 最终导致智能答疑系统并不智能^[1]. 为此, 文中将蚁群算法和数据挖掘算法相结合, 利用移动的、分布的、相互协作的蚂蚁智能体对 IQASD 范例库中的数据进行挖掘, 从而进一步提高 IQASD 的检索效率.

2 分布式智能答疑系统体系结构

由于 IQASD 中存在多个范例库, 为了提高

IQASD 的智能性和开放性,对于大量分散范例库中的数据挖掘,IQASD 采用基于 Agent 的分布式结构模型,通过以 Agent 为主要组件构造满足分布式智能答疑需要的分布式数据挖掘系统. IQASD 体系结构如图 1 所示,整个体系结构可以分为 4 个部分:用户群、用户接口、数据挖掘系统和范例库^[2]. 用户通过用户接口提交数据挖掘请求,数据挖掘系统规划数据挖掘任务并创建装配数据挖掘 Agent,依据要求,数据挖掘 Agent 迁移到范例库中进行数据挖掘,并提交结果给蚁群 Agent 进行综合,最后按照用户要求返回结果^[3].

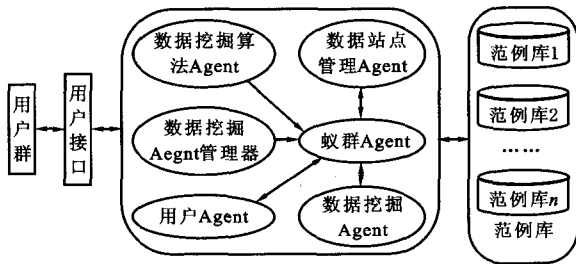


图 1 IQASD 体系结构

3 基于蚁群优化的分类规则发现算法

在传统的答疑系统中,对本地范例库进行数据挖掘时,可以采用基于蚁群优化的分类规则发现算法. 基本思想是将条件项当作蚂蚁智能体的候选路径,按照由信息素和启发函数值所决定的可能性大小,确定选择的条件项,让蚂蚁去对问题的解答做建构和修改,并将其添加到当前正在构建的规则中^[4],其主要算法如下:

```

训练集 = {所有训练范例例};
已发现的规则列表 = []; /* 规则列表在初期为空 */
WHILE (训练集 > Max_uncovered_cases); /* Max_uncovered_cases 为终止条件,即训练集中最多剩下的实例数 */
    t = 1; /* 蚂蚁的索引值 */
    j = 1; /* 收敛测试索引 */
    最初所有路径具有相同的信息素;
    REPEAT
        Antt 从一个空的规则开始,并且通过在当前规则的基础上不断增加规则来构造分类规则 Rt;
        修改规则 Rt;
        修改所爬路径的信息素;
        IF (Rt = Rt-1); /* 更新收敛测试 */
            THEN j = j + 1;
        ELSE j = 1;
    END IF

```

t = t + 1;

UNTIL (i ≥ No_of_ants) OR (j ≥ No_rules_converg); /* No_of_ants 为代理的数目, No_rules_converg 为判断收敛到同一条规则的代理数 */

从 ants 所构建的规则库中选择一条质量最好的;

将规则加入已发现规则的列表中;

训练集 = 训练集 - {被规则所正确覆盖的范例集};

END WHILE

4 分布式蚁群智能分类算法

在传统的远程教育答疑系统中,答疑系统的范例库是孤立的,数据源只有一个,蚂蚁智能体是一个接一个地构建规则. 但在 IQASD,存在着多个范例库,即有多个数据源,因此,需采用不同的处理过程^[5]. 假设 IQASD 中存在 n 个结构相同的范例库,分布在各个远程教育平台中,则分布式蚁群智能分类的具体过程如下所示:

(1) 统计各范例库的信息

① k : 所有范例类型的数量;

② term_{ij} : 条件项, 即 $A_{ij} = V_{ij}$;

③ $|T_{ij}|$: 满足 term_{ij} (即 $A_{ij} = V_{ij}$) 的范例数;

④ $\text{freq} T_{ij}^a$: 满足 term_{ij} (即 $A_{ij} = V_{ij}$), 且类型为 a 的范例数;

⑤ a : 范例库中属性的数量;

⑥ b_i : 属性 i 所有可能取值的数目;

⑦ $\tau_{ij}(t)$: 表示条件项 term_{ij} 在时间 t 的信息素

总量,其初始值为 $\tau_{ij}(t=0) = \frac{1}{\sum_{i=1}^a b_i}$

(2) 汇总各范例库的统计信息

信息服务器及相关信息素服务器对各范例库统计的范例数和信息素总量进行汇总,并计算出如下信息^[6]:

① η_{ij} 表示条件项 term_{ij} 基于密度的启发函数值,可按式(1)计算.

$$\eta_{ij} = \frac{\max(\sum_n \text{freq} T_{ij}^1, \sum_n \text{freq} T_{ij}^2, \dots, \sum_n \text{freq} T_{ij}^k)}{\sum_n |T_{ij}|} \quad (1)$$

② $R_{ij}(t)$ 表示条件项 term_{ij} 在时间 t 的相关函数值,其值可按式(2)计算.

$$R_{ij}(t) = \frac{\tau_{ij}(t) \cdot \eta_{ij}}{\sum_{i=1}^a \sum_{j=1}^k \tau_{ij}(t) \cdot \eta_{ij}}, \forall i \in I \quad (2)$$

式中, I 表示当前规则中未出现的属性.

③ 平均密度按式(3) 确定.

$$\theta = \frac{\sum_{i=1}^a \sum_{j=1}^{b_i} \eta_{ij}}{\sum_{i=1}^a b_i} \quad (3)$$

式中, θ 表示蚂蚁智能体选择条件项的阈值.

(3) 构建分类规则

建立一条规则的关键是产生某一类型的条件项. 一个条件项 term_{ij} 添加到当前规则中的可能性主要由式(4) 决定.

$$P_{ij}(t) = \frac{R_{ij}(t)2}{R_{ij}(t)^2 + \theta^2} \quad (4)$$

(4) 信息素的修改

当 n 只蚂蚁构建完一条规则后, 可根据式(5) 更改此规则的条件项信息素, 而其他不在此规则中出现的条件项信息素仅进行规范化处理.

$$\tau_{ij}(t) = (1 - \rho) \cdot \tau_{ij}(t - 1) + (1 - \frac{1}{1 + Q}) \cdot \tau_{ij}(t - 1) \quad (5)$$

式中, Q 表示所构建规则的质量.

(5) 规则质量的评估及筛选

衡量分类规则质量的标准如式(6), 经过若干蚂蚁智能体对同一训练集挖掘后, 产生了若干条分类规则, 从中挑选质量最好的规则加入到最终分类规则集中^[7].

$$Q = (-\frac{TP}{TP + FN}) \cdot (\frac{TN}{TN + FP}) \quad (6)$$

式中, TP (TruePos) 是满足规则条件, 并且与规则的预测类型相同的实例个数; FP (FalsePos) 是满足规则条件, 并且与规则的预测类型不同的实例个数; FN (FalseNeg) 是不满足规则条件, 并且与规则的预测类型相同的实例个数; (TN) TrueNeg 是不满足规则条件, 并且与规则的预测类型不同的实例个数.

在对 IQASD 中范例库的数据进行挖掘时, 根据以上的过程, 我们设计了分布式蚁群智能分类实现算法, 如下:

训练集 = { n 个范例库中的范例};

WHILE (存在任一训练集中的实例数目 $> \text{Max_uncovered_cases}$); /* $\text{Max_uncovered_cases}$ 为终止条件, 即训练集中最多剩下的实例数 */

初始化信息素;

For n agents in parallel

统计各范例库的有关信息;

End for

计算启发函数值;

$i = 0$;

Repeat

$i = i + n$;

For n agents in parallel

计算相关函数值;

构建一条规则;

更新刚建立的规则;

修改 agents 所爬路径的信息素;

End for

Until ($i \geq \text{No_of_ants}$) OR (Agent_i 与 $\text{No_rules_converg} - 1$ agent 构建的规则相同); /* No_of_ants 为代理的数目, No_rules_converg 为判断收敛到同一条规则的代理数 */

从 ants 所构建的规则库中选择一条质量最好的;

从各范例库训练集中移走满足此条新规则的实例;

END WHILE

5 仿真测试

从 IQASD 中选取几组数据集进行测试, 如表 1 所示. 将每个数据集分为 10 个部分, 假设它们分布在 10 个不同的范例库中, 选择其中一个部分作为测试集, 而其余 9 个部分作为训练集. 实验参数为:

$\text{Max_uncovered_case} = 10$;

$\text{No_of_agents} = 3\ 000$;

$\text{No_Rules_Coverg} = 10$;

$\rho = 0.1$

表 1 IQASD 数据集

数据集	范例	绝对属性	连续属性	分类
DB1	282	9	—	2
DB2	683	—	9	2
DB3	958	9	—	2
DB4	366	33	1	6
DB5	155	13	6	2
DB6	303	8	5	5

仿真测试通过使用训练集产生的规则对测试集进行分类, 通过仿真测试所产生的数据集分类准确率的对比情况见表 2, 从仿真结果中可以看出算法 1 与算法 2 的分类准确率十分接近, 这说明了改进后的分布式蚁群智能分类算法在对 IQASD 中范例库的数据进行挖掘时是可行和有效的.

6 结束语

分布式异构环境下进行答疑系统范例库的数据挖掘与对本地答疑系统的范例库进行数据挖掘相比有很大的不同, 文中将蚁群算法融入答疑系统的范例库数据挖掘中, 对分布式蚁群智能分类算法在答

表 2 算法 1 与算法 2 分类准确率对比情况

数据集	算法 1(基于蚁群优化的 分类规则发现算法)	算法 2(分布式蚁群智能 分类实现算法)
	分类准确率	分类准确率
DB1	75.28±2.24	74.12±2.00
DB2	96.04±0.93	95.00±0.63
DB3	73.04±2.53	74.04±1.42
DB4	94.29±1.20	95.32±1.66
DB5	90.00±3.11	90.00±2.55
DB6	59.67±2.50	57.48±1.78

疑系统中的应用进行了分析和测试,通过实践证明,该算法能有效地对答疑系统特别是分布式智能答疑系统的范例库进行数据挖掘,进一步提高了智能答疑系统检索的效率和智能性,促进了分布式智能答疑系统应用的发展.

参考文献:

[1] 文居峰,朱美琳. 面向商务智能应用的分布式数据挖掘系统设计[J]. 东南大学学报,2005,35(9):858-860.
[2] 刘波,潘久辉. 基于群体智能的分布式数据挖掘方法[J]. 计算机工程,2005,31(8):145-147.
[3] 温蕴. 基于组件式蚁群算法的车辆路径问题研究[J]. 微电子学与计算机,2008,25(6):151-155.
[4] 段海滨. 蚁群算法原理及其应用[M]. 北京:科学出版社,2005:212-332.
[5] 徐克奇. 蚁群优化算法在电镀试验台作业调度的应用[J]. 微电子学与计算机,2008,25(12):102-105.
[6] 林杰,陆耀华. 专家系统辅助的多 Agent 蚁群算法研究[J]. 计算机工程与科学,2007,29(6):47-48.
[7] 秦超,张忠能. 基于改进蚁群算法的客户服务调度方法[J]. 微电子学与计算机,2007,24(7):153-156.

作者简介:

郑耿忠 男,(1975-),博士研究生,副教授. 研究方向为多媒体网络与远程教育、网络优化与传感器网络.

(上接第 49 页)

参考文献:

[1] Sozer E M, Stojanovic M, Proakis J G. Underwater acoustic networks[J]. Oceanic Engineering, 2000,25(1):72-83.
[2] Akyildiz I F, Pompili D, Melodia T. Underwater acoustic sensor networks: research challenges[J]. Ad Hoc Networks Journal, 2005(3):257-279.
[3] Xie P, Cui J H, Lao L. VBF: vector-based forwarding protocol for underwater sensor networks[R]. UCONN CSE, Tech. Rep. UbiNetTR05-03 (BECAT/CSE-TR-05-6), Los Angeles, 2005.
[4] Hai Yan, Zhijie Jerry Shi, Jun-Hong Cui. DBR: depth-based routing for underwater sensor networks[J]. NET-WORKING 2008, 2008(LNCS4982): 72-86.
[5] Casari P, Stojanovic M, Zorzi M. Exploiting the band-

width-distance relationship in underwater acoustic networks[J]. Oceans, 2007(9): 1-6.
[6] Robert J Urick. Principles of underwater sound[M]. 3rd ed. New York: McGraw-Hill,1996.
[7] Stojanovic M. On the relationship between capacity and distance in an underwater acoustic communication channel [C]// Proc. ACM WUWNet. Los Angeles, CA, 2006: 41-47.
[8] Harris A F, Zorzi M. Modeling the underwater acoustic channel in ns2 [C]// ACM International Workshop on Network Simulation Tools (NS Tools). France, Nartes, 2007.

作者简介:

王锦程 女,(1976-),博士研究生,讲师. 研究方向为无线通信与无线传感器网络.