基于 Multi-Agent 的数据挖掘模型的研究

庄新鹏,赵建民,朱信忠
(浙江师范大学 信息科学与工程学院,浙江 金华 321004)

摘 要:数据挖掘是人们长期对数据库研究的结果,但是传统的数据挖掘存在低效性和非智能化等不足。随着具有自主性和社会性的智能计算实体 Agent 的出现和发展,文中将 Multi-agent 技术应用到数据挖掘中,并提出了基于 Multi-agent 智能化的数据挖掘模型,讨论了模型的运行过程。这一模型弥补了传统数据挖掘的缺陷和不足,而且在很大程度上提高了数据挖掘的智能性和高效性,减少了人工的参与。

关键词:数据挖掘;数据库;Multi-agent 技术
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2006)07-0129-03

Research of Data-Mining Model Based on Multi-Agent

ZHUANG Xin-peng, ZHAO Jian-min, ZHU Xin-zhong
(School of Information Science and Engineering, Zhejiang Normal University, Jinhua 321004, China)

Abstract: Data-mining is the result of the researches of database by people for a long time, but there are some inefficiencies and non-intelligence. With the performance and development of intelligent computing entity, agent, which has the capabilities of self-determination and socialization, this paper applies the multi-agent technology into data-mining, and puts forward an intelligent model based on multi-agent, and discusses the running process of this model. The model makes up the limitation and deficiencies; moreover, to a great extent it improves the capabilities of intelligence and efficiency, and reduces the participation by people.

Key words: data-mining; database; multi-agent technology

0 引 言

近十几年来,随着信息技术的飞速发展,信息的数量也在不断呈指数增长。面对如此海量的数据,人们也研究了许多办法来存储和管理这些数据,但要是从中找到满足人们需要的、有价值的数据,那要花费人们很多时间。往往这些结果并不尽如人意,许多内在的、潜在的信息并未被人们所发现,为了解决这个问题,数据挖掘技术也就应运而生了。传统的数据挖掘技术需要人们的大量参与,而且往往一次要处理很多数据,还有可能因为低准确性而造成多次重复操作,这就大大降低了它的智能性和低效性,浪费了用户的时间和精力。

随着 Agent 技术的出现和研究,人们也尝试将其应用到数据挖掘中来,并取得了很好的效果,文中就是在这个基础上,提出了基于 Multi-agent 技术的数据挖掘模型,这个模型在很大程度上弥补了传统数据挖掘技术中出现的不足和问题。

收稿日期:2005-11-06
基金项目:国家自然科学基金资助项目(60473050);浙江省自然科学基金资助项目(ZD0108)
作者简介:庄新鹏(1982-),男,山东日照人,硕士研究生,研究方向为模式识别、数据挖掘和软件 Agent;赵建民,硕士,教授,研究方向为模式识别与图像处理、网络安全与软件 Agent。

1 数据挖掘

随着计算机科学技术和经济的快速发展,大量的数据被积累下来,人们希望对这引起数据进行分析,找出数据之间的内部联系,得到有价值的数据,从而有助于进行决策。这是摆在人们面前的一大课题,引起许多学者的兴趣,并对其进行研究,数据挖掘技术也就随之被提出。

数据挖掘,也称为数据库中的知识发现(KDD,knowledge discovery in database),是从大量原始数据中挖掘出有用的、潜在的信息和知识(如知识概念 concepts、规律 regulations、规则 rules、限制 constraints、可视化 visualization[1]等),由一组操作组成,它被认为是解决知识爆炸、数据丰富但信息贫乏(Data Rich and Information Poor)的一种有效方法[2]。这里所说的数据有结构化的,比如像数据库中的表格,也有非结构化的,比如图像、声音、文字等。因此数据挖掘涉及到许多领域,也有许多方法和算法对数据进行挖掘,特别是近几年来数据库、并行计算、人工智能等领域的学者也纷纷加入到研究数据挖掘的行列中来。

目前国内外在数据挖掘方面的发展和研究主要有:对知识发现方法的研究进一步发展,如 Bayes 方法以及 Boosting 方法的研究;传统的统计学回归法在 KDD 中的应用以及一些学习算法的研究等,但国内的数据挖掘产品很少[3]。

2 Agent 技术及 Multi－agent 系统

2.1 Agent 技术

Agent 是由分布式人工智能发展而来的一种新型计算机模型[4]。Agent 是处在某个环境中的计算机系统，该系统有能力在这个环境中自主行动以实现其设计目标(引自 Wooldridge 和 Jennings(1995)的定义)[5]。Agent 是协作系统中的独立行为实体，它能够根据内部知识和外部激励决定和控制自己行为[6]。

Agent 一经出现，就引起了广大学者的兴趣，人们纷纷对其进行研究，使 Agent 技术涉及到许多不同领域，因此不同领域的学者对 Agent 的定义和特性持有不同的观点。一种较普遍的观点认为，软件 Agent 是具有一定智能体的软件，它应具备以下一些基本特性[7]：

(1)自主(治)性：Agent 能自行控制其状态和行为，能在没有人或其它程序介入时操作和运行。

(2)反应性：Agent 能及时感应和响应其所处环境的变化。

(3)能动性：Agent 是目标驱动的。

(4)持续性：Agent 是持续或连续运行的过程。

笔者认为，除了以上这四点基本特点之外，对于一个 Agent 来说，它还应该具有：

①协调性：多个 Agent 之间可以就某个问题进行通信，协调合作，共同完成任务。

②移动性：Agent 可以在分布式的网络中进行移动，它可以携带其所处的环境状态和参数到另外一个地方完成任务[8]。同时，这也就减少了网络中的多次通信。

③自适应性：Agent 可以根据其所具备的知识和环境的变换，使自己达到另一个状态来适应新的要求，完成新的任务。

当然这些特性并不是所有 Agent 必备的，没必要在设计模型时全都用到，只是选择几个有利于模型研究的特性就可以了。具备不同特征的 Agent 可以应用到不同的领域，解决不同的问题，具体的选择需要根据需要解决问题的状况来确定。

随着系统的智能化和 Agent 技术的发展，数据挖掘系统可以用有用的信息去指导人们的工作，监视企业运行过程，及时把企业支行过程中每个人们的环节所需要的信息提交给用户[4]。

2.2 Multi－agent 系统

所谓 Multi－agent 系统(MAS)就是指多个智能 Agent 通过协作完成任务或达到某些目标的计算机系统[2]。在 Multi－agent 系统中，各个 Agent 实体也具有以上所提到的特性，可以单独地完成任务，在必要时可以通过通信合作共同完成某些任务。可以将某些任务进行化分或细分，并且根据各个 Agent 的特性和状态来分配任务，使它们并行执行任务，这种合作机制可以减少工作的时间。

Multi－agent 系统的主要特征就是任务共享(task－sharing)和结果共享(result－sharing)。

①任务共享就是指单个 Agent 可以花费较少的资源和较少的通信就可以完成子问题，这是建立在对整个任务进行适当的分解之上的。

②结果共享就是指 Agent 之间通过共享部分结果的形式互相协调合作，这样可以适用于在多个 Agent 之间有结果交互的情况。

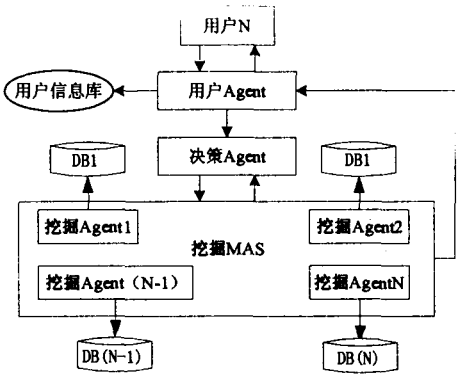
在 Multi－agent 系统中，环境可以分为两部分：外部环境和内部环境。外部环境就是指除 Agents 以外的环境，比如 Agents 所处的位置、所处环境的状态变化等；内部环境就是指 Agent 自身的一些特性，除上面提到的还有比如 Agent 的学习能力、对外界的感应能力，自我状态调整的能力等，这些特征也与它所要完成的任务有关。

在文献[9]中提出了基于访问、感知和知识的多 Agent 系统模型，引入了环境状态划分函数、感知函数以及状态转移函数等对 Agents 和环境进行研究，并将其应用到清理垃圾实例中，取得了很好的效果。

3 基于 Multi－agent 的数据挖掘模型及运行过程

3.1 基于 Multi－agent 的数据挖掘模型

该模型如图 1 所示，每个用户对应一个用户 Agent，而每个用户 Agent 对应于用户的一个用户信息库和一个决策 Agent，接下来就是一个 MAS，它是由 N 个挖掘 Agent 组成：Agent(i)($i=1,2,\cdots,N$)，每个挖掘 Agent(i)对应着一个数据库 DB(i)($i=1,2,\cdots,N$)。



(假设数据已经进行过预处理)

图 1 基于 Multi－agent 的数据挖掘模型

(1)用户 Agent：每个用户都对应着一个确定的用户 Agent，它是人与计算机交互的接口，取代了传统的计算机界面。当用户发出请求或任务时，用户 Agent 就会对用户的请求进行分析和判断，一方面进行自我学习，从而了解到用户的某些特征，分析用户的兴趣和习惯，并把这些特征记录在用户信息库中；另一方面把任务的详细分析交给决策 Agent，由它来决定如何将任务分配给 MAS 中的各个挖掘 Agent。

另外，用户 Agent 还负责将挖掘 MAS 中的各个挖掘 Agent 执行任务的结果进行集中处理，对任务执行的结果进行评估和解释，将冗余和无关信息删除，将最终得到的结果记录在用户信息库中，以备下次请求时使用，同时根

据用户的需求和习惯以友好的形式反馈给用户。

文中的用户 Agent 还有另外一个特点就是在用户没有请求的情况下,隔一定时间后,检查用户信息库中用户经常请求频率较高的几种操作,并自动重新执行。一方面这些用户经常的请求也可能在以后的时间里会多次用到,另一方面,如果数据库进行了更新,也便于及时地向用户反馈经常需要得到的新信息。这也体现了用户 Agent 的自动化和达到了"信息找人"[10]的目标,提高了系统的智能性。

(2)用户信息库:在这个库中,存放着用户经常用的信息,或是多次重复执行的任务结果,如果是下次用户再次请求时,就可以直接从这个库中找到结果,不必再去数据库中检索。另外,这个库中还存有与用户的某些属性,用户请求任务的一些特性,这些特性可以使用户 Agent 在空闲时不断进行自我学习,提高自己的知识库,还可以根据用户的不同请求来改变自己的状态,适应环境的变化。

(3)决策 Agent:它掌握着挖掘 MAS 中所有挖掘 Agents 的信息、特性和属性,以及各自所能完成的任务,存放在自身的挖掘 Agent 状况表中。它接收来自用户 Agent 的任务请求分析报告,并对模糊的信息再次进行分析,将整个问题分解成许多子问题,尽量使这些子问题可以被挖掘 Agent 独立地完成,也就减少了挖掘 MAS 中 Agents 之间的通信和协作。它收分解的子问题,根据各个挖掘 Agent 的自身特性进行分配,使它们尽量独自完成所分配的任务。

(4)挖掘 MAS:在本模型中,它是由 N 个挖掘 Agent 组成的,可以根据所要处理的问题,增加或减少挖掘 Agent 的数量。在这个 MAS 中每个挖掘 Agent 都有一定固有的特性,根据这些特性可以完成不同的任务和请求。每个挖掘 Agent 都对应着一个数据库,它具有对这个数据库操作的所有权限,可以进行读取和写入。当一个 Agent 无法完成所分配的任务时,就需要和其它的挖掘 Agent 进行通信,请求其它的 Agent 来帮助完成任务。当它要与其它 Agent 进行通信,请求帮助时,可以有两种方案来解决它们之间的通信问题:一种方案是每个挖掘 Agent 自身带有一个其它所有 Agent 信息的记录集,而且这个记录集在通信完之后要进行更新,始终保持它记录的是其它 Agent 的最新情况,比如说其它 Agent 的任务分配、自身特性、知识库和其它 Agent 的位置等,这样当某个 Agent 需要进行通信时,就可以查找这个记录集,找到所要请求帮助的 Agent 位置,与它通信,进行合作,完成任务,当无法找到这个需要的 Agent 时就会有报错提示,并将该任务重返给决策 Agent,让其重新进行分析和分配。

另一种方案是所有的挖掘 Agent 在 MAS 中组成一个圆环,在 Agent 进行通信时,所有消息只能沿着一个方向往下一个 Agent 传递,如果找到合适的 Agent,那么它就会主动提供帮助,并有提示回传给原 Agent,如果消息传递了一周回到了原 Agent,那么说明没有 Agent 可以提供帮助,给决策 Agent 发消息,让其重新分析和分配任务。

从宏观来说,一个庞大的任务由许多不同的挖掘 Agent 来完成,这也就是体现了上述提及的任务共享,而挖掘 Agent 也会将自己完成的任务结果给其它 Agent 使用,实现了结果的共享。当某个挖掘 Agent 完成了其所分配的任务时,一方面向用户 Agent 反馈执行结果,另一方面会给决策 Agent 提示,说明自己的任务已经执行完了,同时还会说明自己现在的状况,以使决策 Agent 实时更新它的挖掘 Agent 状况表。

3.2　模型的运行过程

当用户有请求时,首先将该任务交给用户 Agent,由其进行处理和分析。如果问题中有较模糊的问题,那么根据用户 Agent 所掌握的用户信息和兴趣对问题进行分析,同时,检查用户信息库,看是否有与之相关的记录或是相似的请求结果,并推荐给用户选择[11]。否则,就会让用户重新请求。当分析完问题之后,就将用户请求的相关信息和属性记录在用户信息库中,以备下次请求时使用。

用户 Agent 将问题的分析结果和详细报告发送给决策 Agent。决策 Agent 接收到之后,会根据这份详细报告对问题进行分解,尽量使子问题可以由不同的挖掘 Agent 独立完成。当分解好之后,会根据自身所带有的挖掘 MAS 状况表将任务分配下去。分配的原则是:尽量一个挖掘 Agent 完成自己所分配的子问题,如果完不成,那么就要使尽量少的挖掘 Agent 来共同完成这个子问题,如果需要的 Agent 数据量较多,就需要重新再次分解这个子问题。

当把子问题分配下去之后,各个挖掘 Agent 就执行其所分配的子问题,当需要与其他挖掘 Agent 进行通信合作时,就可以按照上述两种方案的其中一种进行协商合作。如果找不到需要合作的挖掘 Agent,那么就将提示决策 Agent,重新将这个子问题进行分解,重表进行分配,并更新挖掘 Agent 状况表。如果找到需要合作的挖掘 Agent,那么就将需要的帮助信息提示给原挖掘 Agent,两者共同来完成任务。若需要,还有可能有第三方挖掘 Agent 的参与。也就是说,执行一次任务,每个挖掘 Agent 并不只与另一个挖掘 Agent 进行合作,也有可能是三者共同来完成的。

当某个挖掘 Agent 完成任务之后,就给决策 Agent 发送提示信息,表明自己已经完成任务,同时将执行的结果反馈给用户 Agent。

用户 Agent 此时一直在等待挖掘 MAS 的执行结果,当用户 Agent 得到所有要得到的结果时,也就说明挖掘 MAS 的所有挖掘 Agents 已经完成了所有的任务,此时决策 Agent 会更新它的挖掘 Agent 状况表,而用户 Agent 会将这些结果进行集中处理,删除一些多余信息,并根据用户的喜好,以最友好的形式反馈给用户。

至此,一次用户请求就完成了。反观整个任务的执行

(下转第 158 页)

(如 SQL)都包含在 DAO 实现中,与业务层隔离开[4]。当业务流程发生变化的时候,只需要改动 Action 中流程控制的代码。

2.3 视图的设计

视图是整个系统的表示部分,用于展现 Model 的内容。本系统是 B/S 模式,视图就是一组 JSP 网页,主要接受用户的请求和显示控制器返回的处理数据。在本系统的 JSP 网页的开发中大量使用了 WebWork 支持的标签库和 OGNL(Object - Graph Navigation Language)。OGNL 是一种功能强大的表达式语言,通过它简单一致的表达式语法,可以存取对象的任意属性,调用对象的方法,遍历整个对象的结构图,实现字段类型的转化等功能[5]。

以流量为例,在接收用户的流量输入时在表单的输入框中这样写:<input name="gbOut. flux"type="text"size="10"maxlength="6">。并在表单对应的执行 Action 中定义一个名为 gbOut,类型为 EnvFlux 的变量(相当于一个模型的对象)。当用户提交这个表单时,OGNL 就被解析成 Java 语句:gbOut. setFlux()。要显示用户输入的流量时,只需要在网页上写 < ww: property value = "gbOut. flux" >, property 标签里的"gbOut. flux"将被解析成 gbOut. getFlux()。

可见使用标签库和 OGNL 使 JSP 网页可读性好,减少了许多重复代码,视图层更加清晰。

3 结束语

本系统采用了基于 MVC 模式的 J2EE 应用框架,具有结构良好、重用性强、易于维护等特点。实际应用证明采用 B/S 模式更能适应三峡 - 葛洲坝联合调度系统的数据维护需要,相比于原系统,数据完整性和实时性都得到了良好的保证。

另外针对模型层中采用 Hibernate 技术已经有成熟的 ORM 自动生成工具,使程序员能够将主要精力集中在业务逻辑的实现。采用 XML 格式映射数据表易于终端的多样化扩展,在客户端还可以采用移动 GPS 终端方便快捷地输入和查询实时数据。

参考文献:

[1] 方 华,郭学俊.基于 MVC 设计模式的水情实时信息系统构建研究[J].计算机与现代化,2005(1):52 - 59.

[2] 魏 勇,唐文彬,郭 梅,等.基于 DAO 模式的 J2EE 应用程序的数据库访问设计[J].计算机应用,2003,23:356 - 357.

[3] Franciscus G. HIBERNATE——符合 Java 习惯的关系数据库持久化[EB/OL]. http://www.narchitecture.net/,2004.

[4] Deepak A. J2EE 核心模式(第 2 版)[M].刘天北等译.北京:机械工业出版社,2005.

[5] 安 子. WebWork 教程—0.9 版[EB/OL]. http://forum.javaeye.com,2005.

(上接第 131 页)

过程。在这个过程中,关键的环节有两点:一个是决策 Agent 对任务进行分解,既要使一个挖掘 Agent 单独解决一个子问题,又要根据各挖掘 Agent 的性能和特征进行任务分配;另一个就是在挖掘 MAS 中,如何使多个挖掘 Agent 进行合理的通信合作,在最大程度上减少通信的时间和占有的资源。文中提供了两个方案可供参考。

4 结束语

文中提出了基于 Multi - agent 的数据挖掘模型,具体说明了各个 Agent 的作用以及整个系统模型的运行过程,将 Agent 技术和思想充分运用到了数据挖掘过程中。从这个数据挖掘过程中也可以看得出,挖掘的智能化得到了很大的提高,减少了人工的参与,挖掘的效率也得到了很大的改善。

当然,在这个系统中也存在着一些问题,比如通信协议及合作安全等,需日后加以完善。

参考文献:

[1] Fayyad U, Piatetsky - Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volume of Data[J]. Communication of the ACM,1996, 39(11):27 - 34.

[2] 殷 燕,白庆华,秦 耕,等. 基于 Multi Agent 技术的信息挖掘系统研究[J]. 计算机应用与研究,1999(12):20 - 22.

[3] 孟晓明.浅谈数据挖掘技术[J].计算机应用与软件,2004,21(8):34 - 36.

[4] 王黎明,柴玉梅,黄厚宽.基于多 Agent 的分布式数据挖掘模型[J].计算机工程与应用,2004,40(9):197 - 200.

[5] Wooldridge M.多 Agent 系统引论[M].石纯一,张 伟,徐晋晖等译.北京:电子工业出版社,2003.

[6] 吴建林,姜丽红,薛华成.专家系统与多 Agent 协作环境[J].计算机科学,1998,25(4):25 - 29.

[7] Lucent Technologies. PSTN - Internet Interworking - An Architecture Overview[Z]. Internet Draft,1997.

[8] 蒋文伟,许华虎,唐 毅.基于 Agent 的数据仓库的研究[J].计算机工程,2001,27(3):29 - 32.

[9] 孙 瑜,夏幼明,李志平. 多 Agent 系统模型及其应用[J].计算机工程与应用,2003,39(12):50 - 53.

[10] Bollacker K D, Lawrence S, Giles C L. Cite Seer: An Autonomous Web Agent form Automatic Retrieval and Identification of Interesting Publications[A]. In: Katia P Sycara, Michael Wooldridge eds. Proceedings of the 2nd International Conference on Autonomous Agents[C]. New York: ACM Press,1998.116 - 123.

[11] 熊忠阳,胡 月,曾令秋,等.一种基于 Agent 的数据挖掘结果模式推荐模型[J]. 计算机应用研究,2003,20(2):71 - 73.