Agent技术在Web数据仓库结构中的应用研究

张 谦 俞集辉

(重庆大学高电压与电工新技术教育部重点实验室 重庆400044)

摘 要 针对基于Web的数据仓库体系结构存在的问题,在该体系结构中引入软件Agent技术,运用移动Agent技术来解决传统方法难以解决的一些主要问题,提出了一种基于Agent的Web数据仓库系统体系结构。文中发挥数据仓库技术和Web技术结合的优势,重点研究了Agent技术在Web服务器端的应用,在把有用的Web数据集成并入到数据仓库中这一目前的研究热点问题中,提出了基于Agent技术将HTML页面转化为XML数据源的解决方案。此外,文章分别研究了应用服务器端和数据仓库系统中的Agent技术的应用,并提出了将Agent技术引入后实现基于Web的数据仓库体系的关键技术。

关键词 软件Agent,Web,数据仓库,移动Agent

Study on Application of Agent Technology in the Web-Based Data Warehouse System

ZHANG Qian YU Ji-Hui

(The Key Laboratory of High Voltage Engineering and Electrical New Technology under the State,Ministry of Education,Chongqing University,Chongqing 400044)

Abstract In order to resolve the shortages and problems of the Web-based data warehouse system that traditional method can't solve,by introducing the software Agent technology to every part in the configuration,the author proposes a Web data warehouse system configuration based on the Agent. Firstly,by taking the advantage of the combination of Web technology and data warehouse technology,the application of Agent technology in the Web server terminal is deeply researched in this paper. To the important issue that how to integrate the useful Web data to data warehouse,the method is proposed to transfer the HTML page into XML data source combined with Agent technology. In addition to,the application of Agent in the application server terminal and in the data warehouse system are analyzed respectively,and the key technologies to realize the Web-based data warehouse system based on Agent are proposed.

Keywords Software agent,Web,Data warehouse,Mobile agent

1 引言

随着数据仓库技术和Web技术的发展,数据仓库技术和Web技术的结合是大势所趋。基于Web的数据仓库体系结构较好地解决了C/S结构对数据仓库使用的局限性,大大扩展了数据仓库的应用范围。但是,同时也带来了关于异构数据源中的数据如何向数据仓库转化问题、系统安全性问题、Web信息提取、网络信息的协调性、大量信息拥塞网络造成的网络瓶颈和传输速度等许多有待解决的问题[1,2]。尤其是如何把有用的Web数据集成并入到数据仓库中为自己所使用这一问题,是目前的研究热点,但一直没有一种很好的解决方案。如果不能很好地解决基于Web的数据仓库体系结构中的这些主要问题,就无法真正实现数据仓库技术的作用,更加不能充分发挥数据仓库技术和Web技术结合后的优势。这就正是本文认真探讨,着力提出解决方案的中心问题。基于这个思想,本文提出将Agent技术引入基于Web的数据仓库体系结构中,面对传统方法难以解决的问题,提出有效解决方案,使得用户可以更加有效地管理分布的、异构的集成环境,同时保持信息源的自主性和独立性。

2 基于Web的数据仓库体系结构与软件Agent技术引入

2.1 基于Web的数据仓库体系结构

基于Web的数据仓库体系结构与传统的C/S结构的数据仓库最大的区别在于改变了最终用户对数据仓库的使用模式,人们不再局限于通过局域网(LAN)使用数据仓库,对数据仓库的建立、维护和使用都是在Internet/Intranet/Extranet环境下进行的,所得的分析结果也可以借助Web服务器迅速发布[3~5]。图1即为基于Web的数据仓库的体系结构。

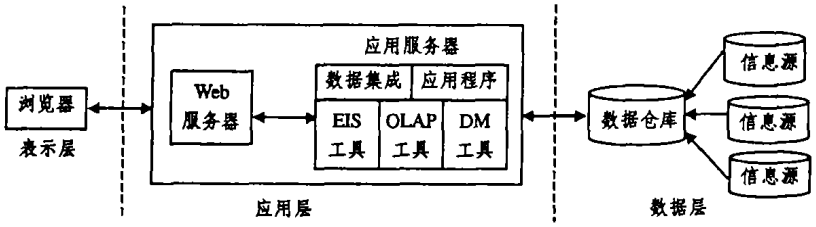

图1 基于Web的数据仓库系统的典型结构

张 谦 博士研究生,主要从事电力工程管理中数字化与优化技术的研究。俞集辉 教授,博士生导师,主要从事电力工程管理中数字化与优化技术的研究。

由图1可知，基于 Web 的数据仓库体系结构的核心就在于 Web 服务器和应用服务器这个中间层。

这种结构更易于构造基于三层逻辑结构的应用系统。它允许同一层的不同应用交互使用，并且相互支持相邻层的相同接口。表示层、应用层和数据层三者相互协作，每一层既作为上一层的服务器端，同时又作为下一层的客户端，共同完成分布在不同地域、不同平台的用户对异地数据仓库的信息访问。

但是，随着数据仓库技术应用的不断深入以及移动计算等新技术的不断发展，图1所示的基于 Web 的数据仓库体系结构开始面临以下4个主要的问题：

(1)因特网对大多数组织机构而言，都是最大的外部数据库。如何在 Web 服务器端实现对 Web 页面信息的有效提取，将有用的 Web 数据集成并入到数据仓库中并为之所用，已成为目前的研究热点；

(2)数据仓库与分布在各处的信息源一般是通过网络联系的，网络拥挤已成为影响集成与下查速度的重要因素，如何减少网络数据传输流量成为人们关注的一个重要问题；

(3)随着移动式数据库的广泛使用，如何把移动信息源中的相关信息高效、便捷地集成到数据仓库中也是亟待解决的问题；

(4)为了维护实视图和响应用户的下查请求，数据源也要参加相应的运算，这给本来就很繁忙的事务数据库增加了额外的负载，如何替事务数据库'减负'日益受到重视。

2.2 多 Agent 协作技术

Agent 具有自主性、交互性、主动性和反应性等，它不仅能作用于自身，而且可以施动作于环境，并能接收环境的反馈信息，重新评估自己的行为；同时，它能与其他 Agent 协同工作[6,7]。因此，将单个 Agent 系统集合起来，通过它们之间的相互作用或相互结合可以产生更高的智能，这就是多 Agent 系统，它与个体 Agent 比较有了质的飞跃。多 Agent 系统通过 Agent 间的合作，不仅改善了每个 Agent 的基本能力，而且，从 Agent 的交互中进一步理解、了解社会行为。如果说独立的 Agent 是模拟个体人员，那么多 Agent 系统则是模拟人类社会。多 Agent 之间通常存在以下几种合作方式[7]：

(1)对等组合作。多个任务 Agent 形成一个任务 Agent 组，每个组成员都有相同的任务，但各自都只能完成其中的一部分，一个组中所有组员独立完成的结果组成了任务的结果。对等组合作中组内各成员都是平等的，它们将自己独立完成的结果发送给同组成员。

(2)主从组合作。和对等组合作一样，主从组合作中每个组成员都有相同的任务，但组内有一个组长，其他为组员。各组员都只能完成任务中的一部分，组长则等待所有组员独立完成的结果，然后将所有结果进行处理，得到任务的最终结果，最后再将结果返回组内其他成员。

(3)直接合作。一个任务 Agent 在运行过程中需要其他任务 Agent 的运行结果。

为了更有效地管理分布的、异构的集成环境，保持信息源的独立性和自主性，减轻事务数据库的负担，我们把合理利用软件 Agent 技术的诸多优势，引入到基于 Web 的数据仓库系统之中[8~10]，提出了一种新的切实可行的基于 Agent 的 Web 数据仓库系统体系结构，以达到：

(1)运用移动 Agent 减少、平衡网络负荷。移动 Agent 可以使大部分的集成操作在信息源端执行，无须把大量的原始数据先通过网络传送到数据仓库后再进行集成。这样就使信息源与数据仓库传输的数据量大为减少。

(2)运用移动 Agent 为移动信息源提供更好的支持。移动 Agent 断续地与网络连接，并在给定目标的驱动下，智能地移动到合适的网络位置进行工作，在网络连接时返回任务结果，使数据集成效率能得以有效提高。

(3)运用 Agent 技术保证信息源的独立性。现实环境中信息源的异构性和分布性的特点，增大了数据集成的复杂性。利用 Agent 的自主性和协作性的特点，让相应 Agent 去完成数据集成任务，既可以保证信息源的独立性，又可以提高数据集成的效率。

(4)运用 Agent 技术降低用户工作的复杂程度。Agent 具有目标驱动的特性，在用户指定一定任务如数据清理、数据转换、数据集成等后，Agent 可以在此任务的驱动下，利用内部知识和能力持续主动地产生面向目标的行动，自主地、高效地完成任务，从而降低了用户工作的复杂程度。

3 Web 服务器端 Agent 技术的应用

虽然数据仓库技术与 Web 技术的结合是一个强大的非常成功的组合，但是，人们不再满足于只在因特网上展示数据仓库中的数据，而越来越关心怎样把 Web 数据集成到数据仓库中。而 HTML 作为 Web 页面信息的主要载体，存在着很严重的缺陷：HTML 无法提供管理数据的标准方式，在数据管理方面的功能明显不足。并且，由于 HTML 标记几乎不含任何数据信息，因此很难支持对数据的搜索，即 HTML 只是描述了页面的外观形式，而不能显示其数据。Web 中有大量丰富的数据：文本、图片、声音、图像等，这些数据多存在于 HTML 文件中，没有严格的结构及类型定义，被称为半结构化的(Semi-structured)数据。对于 Web 数据仓库，用户感兴趣的往往是这些半结构化的数据。随着 XML 的快速发展，针对这一问题，已有许多组织采用了将 HTML 页面转换为 XML 数据源的方法[11]。将 HTML 页面转化为 XML 数据源结构图，如图2所示。

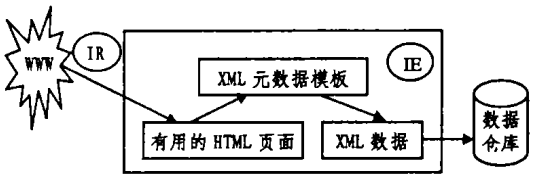

图2 HTML 页面转化为 XML 数据源结构图

在图2所示的结构中，信息检索(即 IR)的主要任务是搜索到与用户需求相关的 HTML 页面，因为并非所有的 HTML 页面的信息对用户而言都是有用的信息。信息抽取(即 IE)的主要任务是，首先定制生成 XML 元数据模板，然后将 HTML 页面的相关数据填充到 XML 元数据模板中，形成 XML 数据源。

XML 文档不仅包括了用户感兴趣的数据，同时也包括了数据的结构信息和其它相关的元数据。XML 可看作是一种半结构化的数据模型，它可以方便地将 XML 的文档描述与关系数据库中的属性一一对应起来，实现精确的查询与模型抽取，因此，XML 数据源中的数据可以通过自动提取，导入到数据仓库中，即通过将 HTML 页面转化为 XML 数据源，可实现将 Web 数据集成到数据仓库中[5]。在此结构中，难点在于如何实现信息检索(即 IR)和信息抽取(即 IE)。由于软件 A-

gent 是具有一定智能的软件，并具有诸多优势，尤其是多个 Agent 可共同协作完成共同的任务这一特点，将软件 Agent 引入如图2所示的结构中，可将 HTML 页面转化为 XML 数据源，实现 Web 页面信息提取的功能。其具体结构如图3所示。

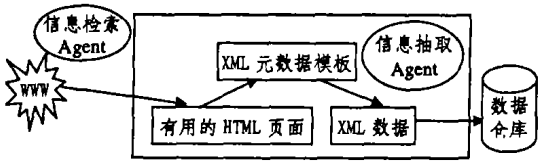

图3　引入软件 Agent 后的结构框图

信息检索 Agent 实现搜索到与用户需求相关的 HTML 页面。信息抽取 Agent 实现信息抽取的功能，即首先定制生成 XML 元数据模板，然后将 HTML 页面的相关数据填充到 XML 元数据模板中，形成 XML 数据源。可见，在该结构中，信息检索 Agent 与信息抽取 Agent 之间采用的是对等组合作方式以实现将 Web 数据集成到数据仓库中的共同目标。

4　应用服务器端 Agent 技术的应用

引入 Agent 技术后应用服务器的结构和工作流程如图4所示。在图4中，需要建立3种类型的 Agent：信息分析 Agent，数据处理 Agent 和结果输出 Agent。

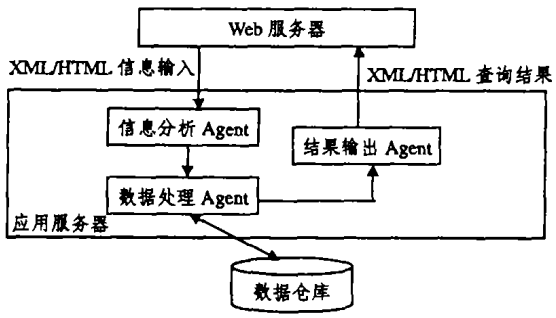

图4　引入 Agent 技术后应用服务器的结构图

在图4所示结构中，各个 Agent 具有相同的任务，即实现 Web 服务器中的数据与数据仓库中数据进行数据交换，同时，又各自完成其中的一部分，并将独立完成的结果组成了共同任务的结果，具体工作如下：信息分析 Agent 的主要任务是对从 Web 服务器发送给应用服务器的信息进行分析，首先确定用户是要从数据仓库提取数据信息，还是要将 Web 数据集成到数据仓库中，然后把这些信息进行细化、归类，最后提交到数据处理 Agent。数据处理 Agent 的主要任务是接收到信息分析 Agent 提交的数据后，对这些内容进行进一步的处理。如果是需要从数据仓库中提取数据，数据处理 Agent 会在数据仓库中进行相关内容的查询和其他操作，完成信息搜索操作，并将搜索结果提交给结果输出 Agent，结果输出 Agent 再将接收到的信息以 XML/HTML 形式的查询结果发送给 Web 服务器。如果接收到的信息是要将 Web 数据集成到数据仓库中，则数据处理 Agent 将这些数据进行分析、归类，再将其存入到数据仓库中对应的位置。因此，该结构中，各 Agent 之间也是采用对等组合作方式。

5　数据仓库系统中 Agent 技术的应用

图5表示了基于 Agent 的数据仓库的基本结构。在图5所示的基于 Agent 的数据仓库结构中，引入了4种类型的 Agent：监控 Agent、分析 Agent、方案 Agent、转换 Agent。

其中，监控 Agent 负责检测信息源中数据的变化，并及时通知方案 Agent。分析 Agent 对进入数据仓库的数据按照一定的数据质量标准进行分析，并对这些数据进行净化、分类。方案 Agent 针对信息源中的数据变化检索字典（对每个信息源而言，字典是特定的），找出与监控 Agent 提到的更新相对应的数据仓库更新方案，并将方案发送给转换 Agent。如果该方案没有涉及到其它信息源，那么转换 Agent 就使用该方案访问信息源，并把需要更新的信息发送给数据仓库；如果该方案中涉及到其它信息源的数据，转换 Agent 就迁移到其它 Agent 群体中，以集成 Agent 的身份与其中的转换 Agent 一起完成集成工作。可见，集成 Agent 先等待所有组员独立完成的结果，然后将所有结果进行处理，得到任务的最终结果，最后再将结果返回组内其他成员。故在该结构中各 Agent 之间采用主从组合作方式，集成 Agent 起到一个组长的作用。

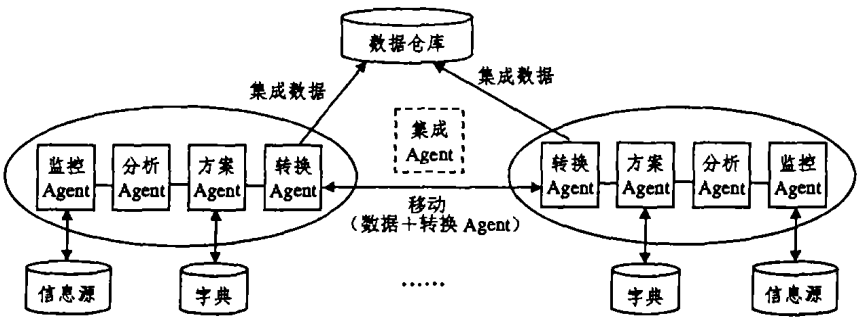

图5　基于 Agent 的数据仓库的体系结构

6　几个关键技术

将 Agent 技术引入基于 Web 的数据仓库系统的实现过程中，主要涉及到以下几个关键技术：

（1）将 HTML 页面转化为 XML 数据源。在系统中，将 HTML 页面转化为 XML 数据源的工作由信息抽取 Agent 来完成。信息抽取 Agent 首先定制生成 XML 元数据模板，然后将 HTML 页面的相关数据填充到 XML 元数据模板中，形成 XML 数据源。

（2）软件 Agent 的构建与实现。软件 Agent 的结构主要有慎思主体结构、反应主体结构和混合主体结构[6]。慎思主体是一个显式的符号模型，包括环境和智能行为的逻辑推理能力，是一个基于知识的系统。它需要解决两类问题：a. 转换问题，如何在一定的时间内将现实世界翻译成一个准确的、合适的符号描述；b. 表示/推理问题，如何用符号表示负责的现实

（下转第120页）

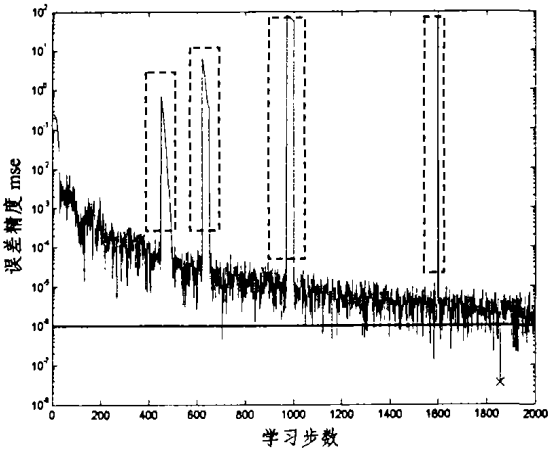

图2 典型收敛过程示意图

结论与展望 综上所述，我们可得出结论：

·将 TS 作为 ANN 的学习算法时，与其它算法相比较而言，它在本质上是一种全局寻优算法，因而具有更高的收敛精度、收敛概率。

·从表2可以看出，在3组不同参数的情况下，最终测试结果相差并不大。这主要是因为，集中性与多样性的自适应搜索策略改善了 TS 的不足，使得：禁忌长度和邻域长度对搜索结果的影响相对较小。

·TS 在训练 ANN 时，不需要计算传递函数的导数。也就是说，除了 sigmoid 函数外，我们可以选择其它非连续可微的函数作为传递函数，来满足一些特殊要求，这正是笔者下一步要进行的工作。

参 考 文 献

1 刘光远，邱玉辉. 基于稳健误差估计器的快速 BP 算法. 计算机科学，1997，24(2)：66～68
2 裴浩东，苏宏业，褚健. 多层前向神经网络的权值平衡算法. 电子学报，2002，30(1)：139～141
3 梁久祯，何新贵，黄德双. 前馈网络的一种超线性收敛 BP 学习算法. 软件学报，2000，11(8)：1094～1096
4 高雪鹏，从爽. BP 网络改进算法的性能对比研究[J]. 控制与决策，2001，16(2)：167～171
5 Engoziner S, Tomsen E. An accelerated learning algorithm for multiplayer prception: Optimization layer by layer. IEEE Transactions on Neural Network, 1995, 6(1): 31～42
6 Scalero R S, Tepedelenlioglu N. A fast new algorithm for training feedforward neural networks. IEEE Transactions on Signal Processing, 1992, 40(1): 202～210
7 Friedrich S, Klaus P A. Clinical monitoring with fuzzy automata. Fuzzy Sets and Systems, 1994, 61(1): 37～42
8 Delgado M, Mantas C, Pegalajar M C. A genetic procedure to tune perceptrons. In: Proc. of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'96), 1996, 2: 963～969
9 Blanco A, Delgado M, Pegalajar M C. A genetic algorithm to obtain the optimal recurrent neural network. International Journal of Approximate Reasoning, 2000, 23 (1): 67～83
10 Blanco A, Delgado M, Pegalajar M C. A real-coded genetic algorithm for training recurrent neural networks. Neural Networks, 2001, 14(1): 93～105
11 Glover F, Laguna M. Tabu Search. Boston: Kluwer Academic Publishers, 1997
12 Glover F, Hanafi S. Tabu search and finite convergence. Discrete Applied Mathematics, 2002, 119 (1-2): 3～36
13 Sexton R S, Alidaee B, Dorsey R E, et al. Global optimization for artificial neural network: A tabu search application. European Journal Operational Research, 1998, 106(2-3): 570～584
14 于志伟. Tabu 机—— 一种新的全局优化神经网络. 电子学报，1999，27(2)：117～119
15 王凌. 智能优化算法及其应用. 北京：清华大学出版社，2001
16 贺一，刘光远，邱玉辉. Tabu Search 中集中性和多样性的自适应搜索策略. 计算机研究与发展，2004，41(1)：162 ～166
17 贺一，刘光远. 基于变异方法的禁忌搜索. 计算机科学，2002，29(5)：115～116
18 方永慧，刘光远，贺一，邱玉辉. 一种基于插入法的禁忌搜索. 西南师范大学学报(自然科学版)，2003，28(6)：887～891
19 Liu Guangyuan, He Yi, Qiu Yuhui, Yu Juebang. Research on Influence of Solving Quality Based on Different Initializing Solution Algorithm in Tabu Search. In: Proc. of Intl. Conf. on Communications Circuits and Systems and West Sino Expositions, Chengdu, China, IEEE Press, 2002. 1141～1145
20 Liu Guangyuan, He Yi, Fang Yonghui, Yuhui Qiu. A Novel Adaptive Search Strtegy of Intensification and diversification in Tabu Search. In: Proc. of IEEE Intl. Conf. on Neural Network and Signal Processing, Nanjing, IEEE Press, Dec. 2003. 428～431
21 Zhang Hongbin, Sun Guangyu. Feature selection using tabu search method. Pattern Recognition, 2002, 35 (3): 701～711
22 Ferland J A, Ichoua S, Lavoie A, Gagné E. Scheduling using tabu search with intensification and diversification. Computer & Operations Research, 2001, 28(11) : 1075～1092

(上接第81页)

世界中的实体和过程，以及如何让一个主体在一定的时间内根据这些信息进行推理并作出决策。在系统中，我们采用慎思主体结构来构建软件 Agent。

(3)各种类型的 Agent 之间通信的实现。将软件 Agent 技术引入到基于 Web 的数据仓库体系结构中的各个部分，需要构建多种类型的 Agent，要求这些 Agent 相互配合，共同完成任务。那么，各种类型的 Agent 之间如何通信是实现我们的这个系统的关键问题之一。目前国际上著名的主体通信语言(Agent Communication Language，简称 ACL)是美国 ARPA 的知识共享计划中提出的两个相关语言：一是 KQML[12](Knowledge Query and Manipulation Language)，另一个是 KIF(Knowledge Interchange Format)。KQML 定义了一种主体间传递信息的标准语法以及一些"动作表达式"，且这些动作从言语行为理论演化而来。KIF 则给信息的内容提供一种语法，它基本上是用类似 LISP 的语法书写一阶谓词演算。在系统中，我们采用 KQML 来表示各个 Agent 间的通信原语，即用 KQML 定义各个 Agent 的通信原语格式。

结束语 随着因特网的日益普及，Web 技术在各个领域得到了广泛的应用。论文深入分析了基于 Web 的数据仓库体系结构存在的不足，尤其是针对如何把有用的 Web 数据集成并入到数据仓库中这一目前的研究热点问题，提出了结合 Agent 技术，将 HTML 页面转化为 XML 数据源的解决方案，从而可实现将 Web 数据转换为数据仓库的数据源，充分发挥数据仓库技术和 Web 技术相结合的优势。文中还分别研究了在基于 Web 的数据仓库体系结构各个层次中软件 Agent 技术的应用，针对该结构目前存在的一些主要问题提供了有效的解决方案，为实现一个更灵活、可伸缩、高效的 Web 数据仓库系统提供了更为广阔的前景。

参 考 文 献

1 吴宏旻，陈奇，俞瑞钊. 关于数据仓库若干问题的讨论. 计算机科学，1999，26(2)：39～43
2 毛国君. 数据仓库的质量管理问题和方法. 计算机科学，2003，30(8)：88～91
3 Inmon W H. Building the Data Warehouse. 2nd ed., New York: John Wiley & Sons, Inc., 1996
4 李秀，廖玉鼻，刘文煌. 基于 Web 的数据仓库系统的研究. 计算机工程，2001，27(11)：44～46
5 许亮，李明，王惠琴. 基于 Web 的数据仓库体系研究. 甘肃工业大学学报，2002，28(1)：68～71
6 Michael R G, Steven P K. Software Agents [J]. Communications of the ACM, 1998, 37(7): 48259
7 Wooldridge M, Jennings N R. Intelligent agents: Theory and practice [J]. Know ledge Engineering Review, 1994, 10(2): 1152152
8 Pham V A, Karmonch A. Mobile Software Agents: An Overview. IEEE Common Magazine, 1998(7): 26～37
9 蒋文伟，许华虎，唐毅. 基于 Agent 的数据仓库的研究. 计算机工程，2001，27(3)：29～32
10 徐忠健，袁捷，杨倩. 基于 Agent 的三层数据仓库系统体系结构的研究. 计算机工程，2003，29(3)：58～60
11 刘振岩，王万森. 基于 XML 的数据挖掘的研究. 计算机科学，2003，30(5)：42～43
12 Finin T, Wiederhodl. An overview of KQML: A Knowledge Query and Mainpulation Language, Stanford University Computer Science Department, 1991