数据挖掘系统研究

王 冠，司建辉，杨昌锋
（北京工业大学 计算机学院，北京 100022）

摘 要：结合当前数据挖掘系统发展状况，分别介绍了集中式和分布式的数据挖掘系统，并着重对集中式数据挖掘系统中的各个组成部分及其具体实现技术做了详细的介绍，同时对集中式和分布式的数据挖掘系统目前的发展状况分别做了总结. 提出了数据挖掘系统的研究方向和发展趋势 增强可视化和交互性，提高可扩展性，与特定行业应用相结合，遵循统一标准和支持移动环境中数据的挖掘. 对数据挖掘系统的发展做了简要的总结和展望.

关键词：数据挖掘；数据挖掘系统；体系结构
中图分类号：TP 311 文献标识码：A 文章编号：0254 – 0037(2005)04 – 0383 – 05

数据挖掘(data mining)又称为数据库中的知识发现(KDD)，是指从存放在数据库、数据仓库或其他信息库中的大量数据中挖掘出有趣知识的过程[1]. 近年来为了推动数据挖掘在实际中的应用，许多研究者对数据挖掘系统的体系结构做了大量的研究工作. 一个结构合理的数据挖掘系统应该具有以下几个特点[2]：1) 系统功能和辅助工具的完备性；2) 系统的可扩展性；3) 支持多种数据源；4) 对大数据量的处理能力；5) 良好的用户界面和结果展示能力. 当前出现的数据挖掘系统主要包括集中式的和分布式的数据挖掘系统，而每种系统的具体结构及其各个组成部分却有多种不同的实现技术和实现方式.

1 集中式的数据挖掘系统

单一数据库/数据仓库的数据挖掘系统是当前发展得较为成熟的数据挖掘应用系统，许多商业性的数据挖掘应用软件都是基于这种结构. 通过对当前主要的数据挖掘系统进行分析可以发现，这种集中式的结构如图 1，但各个不同产品对各个不同功能模块的具体实现技术又不尽相同.

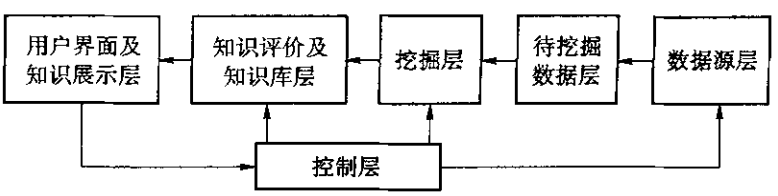

图 1 集中式数据挖掘系统的体系结构

Fig.1 Framework of the centralized data mining system

1.1 用户界面及知识展示层

在该层通过提供友好的用户界面及利用数据可视化技术展示挖掘结果，可以大大提高系统的易用性. 数据挖掘的可视化是指利用可视化技术从大量的数据集中发现隐含的和有用的知识. 数据挖掘的可视化主要包括数据的可视化、挖掘过程的可视化和挖掘模型的可视化. 当前的可视化技术主要包括传统的几何学方法(如曲线图、直方图、散点图、饼图等)、SOM 网可视化技术、平行坐标系技术、面向象素的可视化

收稿日期：2004-10-30.
作者简介：王冠(1968 –)，男，山西长治人，副研究员.

技术等. 基于 SOM 网络和基于平行坐标系的可视化技术是目前应用较多的 2 项技术,它们的原理都是通过把高维数据映射为二维数据从而将数据显示在二维平面上. 如汪加才等设计的一个基于 SOM 网的可视化挖掘系统 VISMiner[3],刘勘等研究了平行坐标系技术在数据挖掘系统中的具体应用[4].

1.2 控制层

控制层用于控制系统的执行流程,协调各功能部件间的关系和执行顺序,主要包括对数据挖掘任务进行解析,并根据任务解析的结果判断挖掘任务涉及到的数据和应该采用的数据挖掘算法.

数据挖掘任务一般是通过数据挖掘语言定义和解释的,当前许多研究者提出了自己的数据挖掘语言,这些语言从结构上看都是类 SQL 语言,如 DMQL 语言[1]等,但是并没有实现挖掘语言的标准化. 2000 年 3 月,微软推出了一个新的数据挖掘语言规范 OLE DB for Data Ming,向着数据挖掘语言标准化又迈进了一大步. Amir Netz 等[5]详细介绍了如何将 OLE DB for DM 规范应用到数据挖掘系统之中.

1.3 数据源层

为了提高数据的一致性和完整性,进行数据挖掘前首先应将分散存储在多个数据源中的数据通过数据清理和数据集成等预处理操作集成到一个统一的数据库/数据仓库中. 为了提高系统的可扩展性,屏蔽数据源采用的具体数据库产品,数据库接口应该采用 ODBC、JDBC 或 OLE DB 等技术,以便于更改数据源. 赵志宏、钱卫宁等分别提出了基于数据仓库[6]和大规模数据库[7]的数据挖掘系统框架及其应用.

数据库可以通过 4 种形式集成到数据挖掘系统中:无耦合的,松耦合的,半松耦合的和紧耦合的[1]. 最理想的是紧耦合方式,即通过把数据挖掘查询优化成循环的数据挖掘和检索过程从而将 2 者结合起来,这样可以充分利用数据库所具有的查询、汇总等数据处理功能,减少数据挖掘系统开发负担,提高系统的效率. Rosa Meo 提出了一种使用数据挖掘语言 Mine Rule 实现与数据库紧耦合的数据挖掘系统框架[8].

1.4 待挖掘数据层

该层为数据挖掘层提供符合数据挖掘算法要求的待挖掘数据集,待挖掘数据集是由数据源层中与挖掘任务相关的数据经过数据变换和数据规约等数据预处理操作形成的.

除了直接基于数据库/数据仓库中的数据进行挖掘外,数据挖掘还可以基于联机分析处理(OLAP)进行,称作联机分析挖掘(OLAM). 由于 OLAM 将 2 者结合了起来,充分发挥 2 者的优点,所以可以使数据挖掘具有较高的效率和良好的交互性. Jia-wei Han 教授等提出了一种 OLAP 和 DM 集成的 OLAM 系统的结构框架,并且开发出了基于这种结构的一个数据挖掘系统 DBMiner[1]. Sanjay Goil 等研究了一种基于并行处理技术的可扩展的 OLAP 和数据挖掘集成的系统体系结构[9].

1.5 挖掘层

该层是数据挖掘系统的核心,该层的具体实现直接关系到整个系统的功能性和可扩展性. 数据挖掘主要包括概念/类描述、关联规则分析、分类及预测、聚类分析、孤立点分析和演变分析等几种类型的模式的挖掘[1],针对各种类型的模式人们又都提出了多种不同的实现算法,对于一个特定的数据挖掘系统应该包括哪些类型的模式挖掘算法则要由该系统的开发目的及其面向的具体应用领域来决定.

为了提高系统的可扩展性,许多系统采用了组件技术来实现数据挖掘算法及其管理. 当前比较成熟的组件技术主要有 COM/DCOM、EJB/Java RMI 和 CORBA/IIOP. 组件(component)是指应用系统中可以明确辨识的、具有一定功能的构成模块. 一个组件的典型结构包括组件接口和组件实现 2 部分,组件接口和组件实现是相互分离的,只要在应用程序中保持统一的接口标准,就可以方便地在系统中加入或替换组件. 如刘君强等设计的 SmartMiner 数据挖掘系统[10]中的算法模块采用了组件对象模型 COM 技术进行构造,并通过算法描述库为组件提供注册机制,任何符合 COM 标准的算法模块可方便地加入到系统中. 在史忠植等人研究开发的 MSMiner 系统[11]中各种数据挖掘核心算法以动态链接库 DLL 的形式加以实现,并可以在系统运行过程中动态加载. 该系统中还提供了专门的算法管理模块,通过挖掘算法库管理各种挖掘算法,

并通过元数据的形式提供算法的注册机制.

1.6　知识评价及知识库层

在将挖掘结果呈现给用户之前通过知识评价可以有效地去除冗余的、无用的挖掘结果，对提高系统的可用性有着重要的意义. 知识评价的度量标准主要包括有效性、新颖性、潜在有用性和最终可理解性. 綦艳霞等详细介绍了知识评价与数据挖掘过程结合的 4 种方式[12].

数据挖掘系统挖掘的知识模式经过知识评价后可以存储在知识库中以便重用. 为了便于不同数据挖掘系统间知识模式的共享，DMG 组织(the data mining group)提出了预言模型标记语言 PMML(predictive model markup language)[13]. PMML 是一种基于 XML 的语言，为数据挖掘产生的预言模型提供了一种统一的定义和描述标准，使得遵循该标准的不同厂商的数据挖掘系统之间可以方便地共享预言模型，提高了模型的可重用性和系统的可扩展性. Wettschereck 等介绍了 PMML 在模型交换中的应用[14].

上面对集中式数据挖掘系统的各个组成部分的实现技术做了详细介绍，目前已出现了许多基于集中式结构的商业数据挖掘软件并开始得到广泛的应用. 比较有影响的商业软件主要有 SAS 公司的 Enterprise Miner，IBM 公司的 Intelligent Miner 和 SPSS 公司的 Clementine 等. Enterprise Miner 实现了与 SAS 数据仓库(SAS Warehouse Administrator)和 OLAP(SAS/OLAP Server)的集成，可以实现从提出数据、抓住数据到得到解答的端到端的知识发现. Intelligent Miner for Data 支持对多种数据源的挖掘，如传统文件、数据库、数据仓库和数据中心等. Clementine 采用了数据挖掘过程模型 CRISP-DM(cross industry standard process for data mining)，能让用户轻松、容易且有效地执行与管理整个数据挖掘的工作. 同时这 3 种软件目前都提供了对 PMML 2.1 的支持，实现了挖掘模型的共享.

2　分布式的数据挖掘系统

随着网络技术和分布式数据库技术的发展和成熟，分布式数据库已经得到越来越广泛的应用，原来数据的集中式存储和管理也逐渐转变为分布式存储和管理. 数据存储方式的变化也必然会促进数据挖掘技术及其系统结构的变化. 由于实际应用中数据的安全性、私有性、保密性以及网络的带宽限制，使得首先将分散存储的数据集中到一个数据库中再进行挖掘的方法是不可行的，因此分布式数据挖掘成为在分布式数据库中进行数据挖掘的最为可行的解决办法.

分布式数据挖掘包括以下几个步骤[15]：1) 剖分待挖掘数据成 P 个子集，P 为可用的处理器个数，并把每个数据子集发送到各个处理器；2) 每个处理器运行数据挖掘算法于其局部数据子集，处理器可以运行不同的数据挖掘算法；3) 组合各个数据挖掘算法发现的局部知识成全局、一致的发现知识. 在分布式数据挖掘中有 4 种关键技术：数据集中、并行数据挖掘、知识吸收和分布式软件引擎[16].

分布式数据挖掘的研究主要包括分布式数据挖掘算法和分布式数据挖掘体系结构的研究 2 个方面. 当前已经出现不少分布式和并行的数据挖掘算法，如并行挖掘关联规则的算法 CD(count distribution)、DD(data distribution)，以及 PDM 等. 在分布式数据挖掘系统结构方面，也已出现了许多基于不同技术的体系结构. 如张学明等研究了一种基于 CORBA 技术并采用多线程并行数据挖掘机制的分布式并行体系结构[15]. 陈刚对基于移动 Agent 技术的分布式数据挖掘体系结构进行了研究[16]. 侯敬军等则提出了一种基于 Web Services 的分布式体系结构[17]，可实现分布式异构环境下的大容量数据的数据挖掘. Krishnaswamy 研究了一种用于电子商务应用的基于异构和分布式环境的联邦式数据挖掘系统(federated data mining system)[18]. Omer Rana 等提出了一种基于组件技术的具有良好可扩展性的分布式数据挖掘系统框架[19]，该框架可以方便地集成第 3 方插件和用户自定义组件.

与集中式数据挖掘系统不同，当前分布式数据挖掘系统还主要处在研究阶段，还没有出现成熟的商业产品. 分布式数据挖掘当前的研究热点主要集中在对超大规模数据集的处理以及提高分布式挖掘系统的整体性能. Grossman 等人提出了一种称为 PDS(photonic data services)的集成框架[20]，在该框架中首次集成了支持远程数据分析和分布式数据挖掘的数据服务，设计用于在高性能网络上进行高效数据传输的网络

协议以及设计用于光纤网络的链路服务,该框架可用于进行 Gigabyte 大数据量的分布式数据挖掘.

3 结论

当前已出现的商业化数据挖掘软件进一步推动了数据挖掘技术的普及和发展,但在实际应用中仍存在不少问题和需要继续研究改进之处,当前主要的研究方向和发展趋势包括以下几个方面[21]:

1) 增强可视化和交互性. 一个具有良好的可视化和交互功能的数据挖掘系统可以使用户直观地看到和理解数据挖掘任务的定义和执行过程,减少用户挖掘知识的盲目性和挖掘过程中大量无关模式的产生,提高系统的挖掘效率及用户对挖掘结果的满意度和可信度.

2) 提高可扩展性. 由于用户的应用环境是不断变化的,因此可扩展性对于数据挖掘系统来说非常重要. 系统应该支持多种数据源的挖掘以及挖掘算法的可扩展性,允许用户根据需要加入新的算法.

3) 与特定行业应用相结合. 随着应用环境的发展,通用的数据挖掘系统已越来越不能满足用户的需要,用户如果不了解挖掘算法的特性就很难得出好的模型. 因此数据挖掘系统应该和特定的应用紧密结合起来,为该应用领域提供一个完整的解决方案.

4) 遵循统一标准. 尽管目前数据挖掘还没有形成一套完整的业界标准,但已出现了一些标准,如数据挖掘过程标准 CRISP-DM、预言模型交换标准 PMML 和 Microsoft 的 OLE DB for DM. 遵循统一标准的数据挖掘系统间可以方便地实现挖掘算法和模型的共享.

5) 支持移动环境. 目前将数据挖掘和移动计算相结合是一个新的研究领域. 因此能够挖掘移动系统、嵌入式系统和普遍存在的计算设备所产生数据的数据挖掘系统是未来的一个新的发展趋势.

参考文献:

[1] HAN Jia-wei, KAMBER M. 数据挖掘:概念与技术[M]. 范明, 孟小峰译. 北京:机械工业出版社, 2001. 305-307.
HAN Jia-wei, KAMBER M. Data Mining Concepts and Techniques[M]. FAN Ming, MENG Xiao-feng, trnsl. Beijing: China Machine Press, 2001. 305-307. (in Chinese)

[2] 周斌, 刘亚萍, 吴泉源. 一个面向电子商务的数据挖掘系统的设计与实现[J]. 计算机工程, 2000, 26(6):18-20.
ZHOU Bin, LIU Ya-ping, WU Quan-yuan. The design and implementations issues of a data mining systems for electronic commerce [J]. Computer Engineering, 2000, 26(6):18-20. (in Chinese)

[3] 汪加才, 陈奇, 赵杰煜, 等. VISMiner:一个交互式可视化数据挖掘原型系统[J]. 计算机工程, 2003, 29(1):17-19.
WANG Jia-cai, CHEN Qi, ZHAO Jie-yu, et al. VISMiner: An interactive visual data mining prototyped system[J]. Computer Engineering, 2003, 29(1):17-19. (in Chinese)

[4] 刘勘, 周晓峥, 周洞汝. 基于平行坐标法的可视数据挖掘[J]. 计算机工程与应用, 2003, 39(5):193-196.
LIU Kan, ZHOU Xiao-zheng, ZHOU Dong-ru. Visual data mining based on parallel coordinates [J]. Computer Engineering and Applications, 2003, 39(5):193-196. (in Chinese)

[5] NETZ A, CHAUDHURI S, FAYYAD U, et al. Integrating data mining with SQL databases: OLE DB for data mining[A]. Pro 17th Int Conf on Data Engineering[C]. Heidelberg: IEEE, 2001. 379-387.

[6] 赵志宏, 骆宾, 陈世福. 基于数据仓库的数据挖掘系统结构框架[J]. 计算机应用与软件, 2002, 19(4):27-30.
ZHAO Zhi-hong, LUO Bin, CHEN Shi-fu. A structure of data mining system based on data warehouse[J]. Computer Applications and Software, 2002, 19(4):27-30. (in Chinese)

[7] 钱卫宁, 魏藜, 王焱, 等. 一个面向大规模数据库的数据挖掘系统[J]. 软件学报, 2002, 13(8):1540-1545.
QIAN Wei-ning, WEI Li, WANG Yan, et al. A data mining system for very large databases[J]. Journal of Software, 2002, 13(8): 1540-1545. (in Chinese)

[8] MEO R. A tightly-coupled architecture for data mining[A]. Pro 14th Int Conf on Data Engineering[C]. Orlando: IEEE, 1998. 316-323.

[9] GOIL S, CHOUDHARY A. A parallel scalable infrastructure for OLAP and data mining[A]. Pro IDEAS '99 on Database Engineering and Applications[C]. Montreal: IEEE, 1999. 178-186.

[10] 刘君强, 王勋, 孙晓莹. 智能型数据挖掘工具的设计与实现[J]. 计算机工程与应用, 2003, 39(17):195-197.

LIU Jun-qiang, WANG Xun, SUN Xiao-ying. Design and implementation of an intelligent data mining tool[J]. Computer Engineering and Applications, 2003, 39(17): 195-197. (in Chinese)

[11] 秦亮曦，史忠植，刘少辉，等. 多策略数据挖掘平台 MSMiner 的元数据管理[J]. 计算机应用，2003，23(12)：34-36.
QIN Liang-xi, SHI Zhong-zhi, LIU Shao-hui, et al. The meta-data management of a multi-strategy data mining system MSMiner[J]. Computer Applications, 2003, 23(12): 34-36. (in Chinese)

[12] 綦艳霞，杨炳儒. KDD 中知识评价的研究综述[J]. 计算机应用研究，2001，18(12)：1-5.
QI Yan-xia, YANG Bing-ru. An overview of evaluation for discovered knowledge in KDD[J]. Application Research of Computers, 2001, 18(12): 1-5. (in Chinese)

[13] Data Mining Group. PMML2.1 Specification [EB/OL]. http://www.dmg.org/pmml-v2-1.html, 2003-03-25/2004-09-28.

[14] WETTSCHERECK D, MULLER S. Exchanging Data Mining Models With the Predictive Modelling Markup Language [EB/OL]. http://ai.ijs.si/branax/iddm-2001-proceedings/workshop/ Paper26.pdf, 2001-09-06/2004-09-28.

[15] 张学明，施法中. 分布式并行数据挖掘系统的研究与实现[J]. 计算机工程与应用，2002，38(4)：198-200.
ZHANG Xue-ming, SHI Fa-zhong. A data mining method based on VisiBroker and multi-thread[J]. Computer Engineering and Applications, 2002, 38(4): 198-200. (in Chinese)

[16] 陈刚. 基于代理的分布式数据挖掘系统设计[J]. 计算机工程，2001，27(9)：65-68.
CHEN Gang. A distributed data mining system based on agent[J]. Computer Engineering, 2001, 27(9): 65-68. (in Chinese)

[17] 侯敬军，曾致远，向凌. 一种基于 WEB 服务的分布式数据挖掘体系结构[J]. 微机发展，2004,6(14)：48-51.
HOU Jing-jun, ZENG Zhi-yuan, XIANG Ling. An architecture for distributed data mining system based on web services [J]. Microcomputer Development, 2004, 6(14): 48-51. (in Chinese)

[18] KRISHNASWAMY S. Federated data mining services and a supporting XML-based language[A]. Pro 34th Int Conf on System Sciences[C]. Hawaii: IEEE, 2001. 1-10.

[19] RANA O, WALKER D, LI Mao-zhen. PaDDMAS: Parallel and distributed data mining application suite[A]. Pro 14th Int Conf on Parallel and Distributed Processing Symposium[C]. Cancun Mexico: IEEE, 2000. 387-392.

[20] GROSSMAN R, GU Yun-hong, HANLEY D, et al. Photonic Data Services: Integrating Data, Network and Path Services to Support Next Generation Data Mining Applications [EB/OL]. http://www.rgrossman.com/dl/proc-068.pdf, 2004-05-11/2004-10-08.

[21] 朱建秋. 数据挖掘系统发展综述[EB/OL]. http://www.dmgroup.org.cn/zhujianqiu/dmsystem.pdf, 2003-04-20/2004-10-10.
ZHU Jian-qiu. A Survey on Data Mining System Evolution[EB/OL]. http://www.dmgroup.org.cn/zhujianqiu/dmsystem.pdf, 2003-04-20/2004-10-10. (in Chinese)

Research of Data Mining System

WANG Guan, SI Jian-hui, YANG Chang-feng

(College of Computer Science, Beijing University of Technology, Beijing 100022, China)

Abstract: The centralized data mining system and the distributed data mining system are discussed respectively based on the development situation of data mining system at present. We illustrate the components' implementation techniques of the centralized data mining system in detail, and we also make a conclusion about the development of the centralized and the distributed data mining system. Reseatch direction and development trend of data mining system is discussed. Finally the future of data mining system is foreseen.

Key words: data mining; data mining system; framework