数据挖掘技术标准综述*)

刘明亮 李雄飞 孙 涛 许晓晴
(吉林大学计算机科学与技术学院 长春 130012)

摘 要 随着数据挖掘技术应用日趋广泛,涌现出各种挖掘工具和系统。为规范相应的软件开发和数据交换方法,制订数据挖掘技术规范和标准成为当务之急。在将数据挖掘标准划分为过程标准、接口标准、语言标准和Web标准等四类进行分析介绍后,给出一个综合多种标准的应用程序框架,最后总结出数据挖掘标准化领域面临的问题和挑战,并对发展趋势予以展望。

关键词 数据挖掘,技术标准,紧密耦合,应用程序架构

Survey of Data Mining Technology Standards

LIU Ming-liang LI Xiong-fei SUN Tao XU Xiao-qing
(College of Computer Sci. & Tech., Jilin University, Changchun 130012)

Abstract With the increasing use of data mining techniques, various data mining tools and systems have come forth. It is urgent to constitute data mining technology criterions and standards in order to standardize the methods of software development and data exchange. We have consolidated all the current popular data mining standards and divided them into four kinds: process standards, interface standards, language standards and Web standards. In this paper, we first illustrate these four categories of standards respectively, and then design an application framework using these standards, at last summarize the problems and challenges that exist in the domain of data mining standards and view the future development of data mining standards.

Keywords Data mining, Technical standard, Close-coupled, Application architecture

1 引言

数据挖掘(Data Mining, DM),又称为数据库中的知识发现(Knowledge Discovery in Database, KDD),是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程[1]。它融合数据库系统、人工智能、统计学、机器学习、信息科学等,是一个新兴的多学科交叉应用领域。数据挖掘技术已经在各行业的决策支持活动中扮演着越来越重要的角色。

从早期解决简单问题的独立数据挖掘软件,到解决通用问题的、集多种数据挖掘算法于一体的横向数据挖掘系统,迄今已经涌现出大量的数据挖掘工具,如,DBMiner、MLC++、Rosetta、49er、Clementine、Intelligent Miner、Enterprise Miner、MineSet、Darwin等。其中,比较著名的有IBM的Intelligent Miner, SPSS的Clementine, SAS的Enterprise Miner, SGI的MineSet, Oracle的Darwin等。

由于数据挖掘技术是多学科技术的综合体,加之各企业和科研部门均按照各自的标准开发数据挖掘软件,导致在数据挖掘领域出现了以下问题:

(1)各种模型和技术难于集成。

数据挖掘技术是面向问题的,不同的问题往往采用不同的模型和技术,且彼此相互独立,联系很少。这样由数据挖掘工具开发商们提供的工具之间难以交互,不容易集成到同一个应用中。

(2)缺少简明精确的问题描述方法。

语义通常是由实现方法决定的,很难用统一的原语言描述数据挖掘问题。特别是急需用形式化语义精确地描述数据挖掘问题。

(3)挖掘系统仅提供孤立的知识发现功能,难于嵌入大型应用。

大多数数据挖掘工具采用独立的数据挖掘模型,不能同操作环境中的预言模型无缝集成。

(4)缺少与数据库系统耦合的通用API或原语。

数据挖掘引擎和数据库系统是松散耦合的,T. Imielinski和H. Mannila称其为"文件挖掘系统"[2]。数据挖掘系统缺乏统一的对数据库系统的高性能访问接口,也没有支持与数据库紧密耦合的原语。

标准化是解决上述问题的出路。制定统一标准可以规范数据挖掘工具开发过程、方法、接口等,有利于这些工具的维护、升级、集成和数据交互。目前已经出现一些数据挖掘标准,部分挖掘工具和应用程序也已开始遵从标准开发,如SPSS Clementine(采用CRISP-DM标准和PMML标准),IBM Intelligent Miner(采用PMML标准)[3],Oracle ODM(采用JDM标准),Microsoft SSAS(采用OLE DB for DM标准)[4]等等。此外,数据网格、网络服务、语义网等也建立了与数据挖掘相关的框架和服务标准。数据挖掘标准以及基于这些标准的框架和服务将在很大程度上影响未来数据挖掘的应用方式。

就目前国外数据挖掘标准的研究进展、现状和问题将在随后的章节中讨论。第2节分类总结了数据挖掘的业界标准,第3节讨论基于这些标准开发的应用程序框架,第4节分析标准之间的关系和存在的问题,最后给出急需解决的问题

*)本文得到国家自然科学基金(60373097)和中国博士后科学基金(2004035170)资助。刘明亮 硕士研究生,主要研究方向为数据挖掘;李雄飞 博士,教授,博导,主要研究方向为数据挖掘、网格计算、信息融合。

和发展趋势。

2 数据挖掘业界标准

数据挖掘标准化问题还处于初级阶段，迄今没有形成公认的统一标准。按目前各种标准所解决问题方法和侧重点的不同，将数据挖掘标准划分为四类。

(1)过程标准：定义数据挖掘模型产生、使用和部署的过程标准，如 CRISP-DM 和 Fayyad 过程标准等。

(2)接口标准：为方便客户应用程序调用，针对具体编程语言和系统提供的数据挖掘 API 接口，如，JDM，SQL/MM 等。

(3)语言标准：针对数据挖掘问题定义，用于问题描述、知识发现和表达的数据挖掘语言标准。用统一的语言标准规范数据挖掘平台和应用程序开发。与 SQL 语言类似，已经设计出数据挖掘查询语言(如，DMQL、MSQL 和 Mine Rule 等)、数据挖掘定义语言(如，PMML、CWM for DM 等)和集查询、定义和操纵于一体的通用数据挖掘语言(如，OLE DB for DM)。

(4)网络标准：用于解决网络上分布式和远程数据挖掘问题的数据挖掘 Web 标准，如，XML for Analysis，Data Space，Semantic Web 等。

2.1 数据挖掘过程标准

数据挖掘是分步骤、多角度数据分析和知识获取的过程。为使数据挖掘过程与具体应用开发过程相结合，成为商业开发的关键步骤，需要建立统一的过程标准。这将有助于形成一个可以有效记录工作经验的统一体系，能够加强项目计划和项目管理；有助于新手顺利地完成数据挖掘的整个工作流程；有利于详细规划和设计过程标准中的每个步骤，很好地控制和降低项目的成本。在此背景下数据挖掘过程标准应运而生。1996 年出现了 Fayyad[5] 标准，1998 年提出 Cabena[6] 标准，1999 年提出 CRISP-DM[7] (Cross Industry Standard Process for Data Mining)标准，2001 年的 Cios[8] 标准以及 SAS 的 SEMMA(Sample，Explore，Modify，Model，Assess)模型等相继出现。

最初，Fayyad 标准是应用最广泛的标准，如 IBM Intelligent Miner，DBMiner 等都采用此标准。它包含数据选择、数据预处理、数据转换、数据挖掘和解释评估等五个步骤。但是，Fayyad 模型是一个偏向于技术的模型，对商业问题的理解、挖掘模型应用等缺乏指导性，因此不适合现代商业项目开发。Cabena 标准虽然包含商业理解和模型部署阶段，但是没有数据理解阶段，不利于数据挖掘模型和工具的选择。Cios 是与 CRISP-DM 最为接近的标准，仅在过程中的反馈和每个阶段有微小差异，但是后者更适用于解决商业问题。SEMMA 是 SAS 公司提出的一套行之有效的数据挖掘方法论，主要是从探测和挖掘具体数据集的角度来实行，强调在最终确定模式和模型前，要经过充分的探索和比较。它也是一个更偏向于技术的模型[9]。目前，最有影响力的数据挖掘过程标准是 CRISP-DM，已得到 20 多家公司和企业认同，如 IBM，SPSS，SGI，NCR 等都采用该标准，SPSS 公司的数据挖掘工具 Clementine 就是该标准的典型应用之一。目前，CRISP-DM 已经成为应用最广泛的、事实上的工业标准[10]。

CRISP-DM 是一个分级的过程模型，它将数据挖掘过程分解成 6 个阶段和 4 个层次。

(1)理解商业背景：确定商业目标，评估目前的情势，明确数据挖掘目标并建立项目计划。

(2)理解原始数据：收集并描述原始数据，检查和确认数据的质量。

(3)数据准备：选择、清理数据，数据综合并做数据标准化。

(4)建立数据挖掘模型：选择建模算法，产生测试模型，建立模型和评估模型。

(5)评估：评估数据挖掘的结果，监视数据挖掘过程并确定下一步工作。

(6)部署：制定数据挖掘实施计划，制定监控该计划实施的方法，完成最终报告，最后回顾整个工程。

CRISP-DM 标准的阶段和层次关系如图 1 所示。

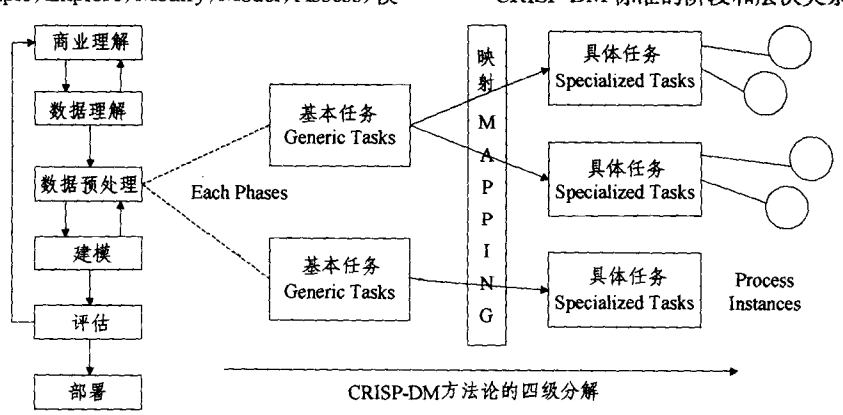

图 1 CRISP-DM 过程模型

CRISP-DM 的成功之处在于它是面向数据挖掘项目开发的，并且与行业、背景、数据挖掘工具无关。它可以将整个数据挖掘过程同标准的商业过程相结合，把具体的商业目标映射为数据挖掘目标，从而保障数据挖掘的结果能更好地指导商业决策。

2.2 数据挖掘接口标准

在早期，数据挖掘服务提供商只能为应用程序或终端用户图形界面提供后台算法，用户应用程序来调用数据转换，建立模型，并完成测试和评估等过程，但是由于这些过程都与专门的提供商有关，当用户拟采用多种数据挖掘解决方案时就会遇到问题，不得不重新修改代码，调用工具商提供的服务。为了使不同开发商的数据挖掘工具互连成为可能，各个数据挖掘工具在不大量修改代码的情况下直接为终端用户提供服务，提出了数据挖掘的接口标准。

目前主要有两种接口标准，SQL/MM(SQL/Multimedia)[11] 和 JSR-073/JDM (Java Specification Request 073 / Java Data Mining)[13]。

2.2.1 SQL/MM 标准

SQL语言是广泛应用的关系数据库语言。数据挖掘可以看成是对SQL语言基本数据操作的延伸，因此，可以通过扩展SQL语言来标准化数据挖掘功能，为数据库系统和数据挖掘系统集成提供一种标准接口，解决它们之间的松散耦合问题。

SQL/MM是一个ISO/IEC的国际化标准项目，主要用于定义纯文本数据、空间数据、静态图像数据和数据挖掘的标准。该标准的第6部分用来解决数据挖掘问题，为数据挖掘模型的生成、测试以及应用等工作定义了标准的SQL API。目前，SQL/MM可以支持分类、聚类、回归和关联规则四种数据挖掘模型，并且允许用户使用自定义的数据类型和方法。如，定义设置类型(Setting Type)来存储用于生成模型的参数，定义挖掘任务类型(Task Type)来控制数据挖掘任务的实现，并且通过统一接口对数据挖掘结果进行测试和应用。

所有SQL类型值都存储在数据库表中，不需要标准的函数来操纵模型的复制、更新和删除等操作，数据挖掘工具可以通过扩展的SQL语言，以统一的方式实现数据挖掘功能。此外，SQL/MM还强制将PMML作为模型导入导出的标准，以方便不同数据挖掘工具之间的数据交互。

但是，SQL/MM还没有统一的设置类型的标准格式，因此不同的挖掘模型之间不能共享设置；对挖掘模型的支持也很有限，需要进一步扩充；并且没有定义从挖掘对象中获取元数据的方法等。因此，SQL/MM应该同其它标准相结合，如同JDM标准相结合定义统一格式的设置类型格式，同CWM for DM标准结合定义统一的元数据模型和数据挖掘结构等[12]。

2.2.2 JSR-073/JDM标准

JSR-073又称为JDM，由Oracle，Hyperion，IBM和SUN Microsystems等组织联合提出，是为支持数据挖掘应用而开发的纯Java接口。它支持数据和元数据的创建、存储、访问和维护以及数据挖掘模型的创建和使用。利用JDM，数据挖掘服务的实现者能够将单一、标准的API接口显露于前端的应用程序开发者或者是Java2平台组件的开发者。

JDM主要有以下三个结构组件：

(1)应用程序编程接口。终端用户的可视化部件需要通过此接口调用数据挖掘引擎(DME)提供的数据挖掘服务。一个应用程序开发者可以仅需要掌握此接口即可。

(2)数据挖掘引擎(DME)。提供数据挖掘服务的基础架构，终端用户通过接口调用它提供的数据挖掘服务。

(3)元数据仓库。存储底层的数据挖掘对象，其可以基于CWM框架。

JDM中有两个主要的概念：设置和模型。设置可分为功能设置和算法设置两部分，它是一组用于构造数据挖掘模型的输入参数的集合。功能设置对象定义了支持的数据挖掘的主要功能：分类、回归、属性重要性分析、聚类和关联规则等，它可以在不指明数据挖掘算法的情况下说明将要得到的数据挖掘结果。算法设置是指设置与数据挖掘算法相关的参数。模型是功能设置在算法和数据上应用所得的结果，可用于直接检验、精度测试、数据评分、输入到外部表示如PMML以及作为数据挖掘引擎的模型输入等。

JDM是高度概括的、面向对象的数据挖掘概念模型。它可以与ISO的SQL/MM，以及DMG的PMML等标准配合使用。JDM的具体实现策略由开发商决定，由于JDM的接口分为必须和可选两个部分，开发商可以根据需要实现其中的一部分接口来实现自己的功能。不过，JDM目前还缺乏网络服务、数据转换、非结构化数据挖掘、多目标模型以及更多的数据挖掘功能(如，特征提取和预测等)的接口，这些将在后续的版本中添加[14]。

目前，JDM已经应用到了Oracle10g Data Mining(ODM)中。ODM是Oracle 10g数据库提供数据挖掘功能的可选组件，其API采纳了JDM标准的概念和方法，允许用户利用统一的Java API编制挖掘程序，而模型的建立、测试和评估等功能在数据库内部完成。通过ODM客户端、PL/SQL和标准的Java API，开发者可以将数据挖掘功能无缝集成到具体应用中[15]。

2.3 数据挖掘的语言标准

普遍认为关系数据库的成功在很大程度上取决于关系查询语言SQL，尽管商品化的关系数据库都拥有各自的图形用户界面，但是每个界面下面的核心都是关系查询语言。人们可以借鉴SQL语言，建立数据挖掘语言标准，从而实现数据挖掘系统的标准化，支持统一的和交互的数据挖掘，便于灵活有效的知识发现。

数据挖掘语言从早期的各个企业和研究单位自行研究和开发，到现在的大的组织和联盟提出的各种不同类型的标准，已经出现了很多不同的数据挖掘语言和标准，根据数据挖掘语言的功能和侧重点不同，可以将它分为数据挖掘查询语言，数据挖掘定义语言和通用数据挖掘语言三种。

2.3.1 *数据挖掘查询语言*

目前，在数据挖掘语言方面已经取得一些研发成果，如JiaWei Han等人定义的面向文本数据挖掘的查询语言DMQL[16]，Imielinski和Vermani提出的用于发现面板(Discovery Board)数据挖掘系统的语言MSQL[17]，以及由Meo，Psalia和Ceri设计的用于查询关联规则的MineRule[18]等。

多数数据挖掘查询语言采用类似SQL语言的语法，提供一些数据挖掘原语，用户通过这些原语制定数据挖掘任务。数据挖掘原语通常从五个方面描述问题，即待挖掘的数据、挖掘知识的类型、背景知识、兴趣度度量、模式的表示与可视化。用户在数据挖掘过程中使用这些原语从不同的角度或深度与系统进行交互式通信，达到知识发现的目的。

现今的数据挖掘查询语言没有形成统一的标准，不能解决数据挖掘系统之间各自独立，难于嵌入大型应用系统的问题。

2.3.2 *数据挖掘定义语言*

2.3.2.1 *预言模型标记语言(PMML)*

现有的描述数据挖掘结果的模型不能满足异构数据挖掘系统中模型交换和使用的要求。为复用和继承不同数据挖掘系统的模型，需要制订统一的挖掘模型定义标准。

最著名的数据挖掘模型是模型标记语言(PMML)。它由数据挖掘组(DMG)于1999年7月提出，已经被W3C接受。PMML模型采用XML语言的数据分层思想和应用模式，通过XML解析器对输入和输出的数据类型、模型详细的格式、数据挖掘结果模型等进行解析，使统计分析中预测模型具有较强的可移植性。PMML可以解决不同厂商开发的模型之间的兼容性问题。

PMML由以下几个部分组成[19]：

(1)数据字典(Data Dictionary)

用于定义模型输入属性，包含属性的名称，类型和范围。不同的数据挖掘模型可以共用同一个数据字典。

(2)挖掘模式(Mining Schema)

挖掘模式用于描述数据挖掘对象，是数据字典中所有属性的子集。为适应数据挖掘模型需要，挖掘模式还包含特定的属性描述信息。如，针对属性的使用特性可描述为测试属

性(输入属性)、类属性(预测属性)或附加属性等。

(3)数据转换字典(Transformation Dictionary)

利用转换函数,将数据转换成适合特定模型的数值,包含数据的离散化、正则化和聚集等。

(4)模型统计(Model Statistic)

记录模型使用属性的固有统计信息。

(5)挖掘模型(Mining Model)

PMML 定义了多种数据挖掘的模型,如关联规则模型、聚集模型、回归模型、朴素贝叶斯模型、决策树模型等。虽然各种数据挖掘模型的定义方法不同,但是在模型名称、功能描述、挖掘算法定义,以及挖掘模式等属性上均有相同之处。因此,在同一个 PMML 文件中可以包含多个数据挖掘模型,这些模型之间也可以组合使用。

PMML 模型中各部分之间的联系参见图 2。

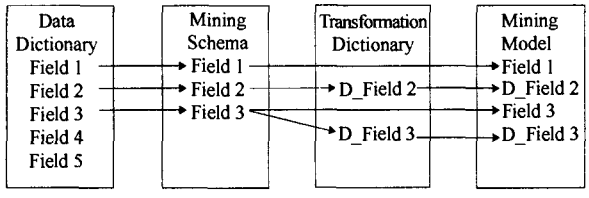

图 2 PMML 模型示例

从数据输入到模型生成的整个过程可以看出,PMML 挖掘模型中的属性分为两种类型:一种是挖掘模式描述的基本属性,另一种是由数据转换字典导出的属性。

PMML 模型有两种应用形式:将 PMML 接口作为挖掘工具接口 API 的一部分或者是用专门的 PMML 结果模型的导入导出组件将挖掘工具的内部模型转换成 PMML 模型。由于后者可以在内部生成符合自身算法的模型,因此应用更加广泛。

PMML 是使用最为广泛的数据挖掘标准之一。国内外很多公司已经采用 PMML 标准,如 IBM DB2 DWE(Data Warehouse Enterprise),SAS 的 Enterprise Miner 和 SPSS 公司的 Clementine 等。作为数据挖掘模型结果的定义标准,PMML 还可以同其它标准相结合作为应用的基础。

PMML 模型的提出和广泛应用促使数据挖掘系统向数据挖掘中间件转变,出现了专门用于产生静态模型的模型生产者(Model Producer)和使用静态模型的模型消费者(Model Consumer)[20]。这些中间件可以灵活地应用于应用系统,有利于嵌入式系统的开发。

PMML 的表现能力有很大的局限性,描述的内部数据挖掘模型并不完整,需要利用 XML 的可扩展性来丰富 PMML 的内涵;该模型还只是对数据挖掘结果的静态描述,缺乏对于数据挖掘过程本身的描述;此外,在方便模型部署、增添数据转换和内置函数以及添加附加信息等方面还需要进一步改进[21]。

2.3.2.2 通用数据仓库元模型(CWM for DM)[22]

元数据是关于数据的"数据",用于描述数据的含义。利用元数据可以更好地理解、管理和使用企业拥有的数据,也是数据仓库项目取得成功的关键因素之一[23]。

2001 年 2 月,OMG 颁布了 CWM1.0 标准。目的是在异构环境下,实现数据仓库工具、平台和元数据知识库之间的元数据交换。CWM 模型既包含元数据存储,也包含元数据交换。

CWM 为数据仓库和商业智能工具之间共享元数据提供了一整套的语法和语义规范:

(1)CWM 元模型:描述数据仓库系统的元模型。

(2)CWM XML:CWM 元模型的 XML 表示。

(3)CWM DTD:DW/BI(Business Intelligence)共享元数据的交换格式。

(4)CWM IDL:DW/BI 共享元数据的应用程序访问接口。

CWM for DM 是 CWM 元模型中针对数据挖掘服务的分析子包,包含从数据输入到挖掘模型的建立、测试和应用的数据挖掘任务的整个过程。它由 6 个 CWM 数据挖掘接口的基本包组成:

(1)挖掘功能设置包:定义特定数据挖掘功能的参数对象。

(2)挖掘模型包:定义为基本挖掘模型对象,用于所有模型对象继承,作为挖掘任务的构建结果。

(3)挖掘结果包:定义基本的挖掘结果对象,用于被所有的结果对象继承,作为特定数据挖掘任务的结果。

(4)挖掘数据包:定义描述输入数据、输入数据处理方式、输入数据和挖掘算法能理解的内部表示映射等对象。

(5)挖掘任务包:定义表示特定挖掘操作任务的对象。

(6)入口点包:定义应用编程入口点的顶层对象。

数据仓库工具通过声明文档类型定义(DTD)来标准化 XML 元数据交换(XML Metadata Interchange,XMI),实现不同工具之间的元数据互操作。建立 DTD 的过程分为三步,首先,利用 UML 对数据建模;然后,根据 UML 模型生成统一的 CWM 的交换格式 MOF(Meta-Object Facility)和 XMI;最后,将 MOF 和 XMI 自动转换成 DTD。

CWM for DM 已经应用在一些数据挖掘工具中,用来支持数据挖掘元数据的交换和管理,如,Oracle 10g 集成的商业智能设计工具 OWB(Oracle Warehouse Builder)[24],Prudsys 的与平台和数据源无关的嵌入式数据挖掘工具 XELOPES Library[25]等。

采用 CWM for DM,可以为不同的数据仓库和数据挖掘服务提供统一的元数据管理环境,方便数据仓库和数据挖掘工具之间的系统集成、数据理解,可以提高系统的灵活性。

2.3.3 通用数据挖掘语言:OLE DB for DM[26]

2000 年 3 月,微软提出的 OLE DB for DM 是数据挖掘语言标准向前迈出重要的一步。它既可以像数据挖掘查询语言那样,通过类似 SQL 语言同数据挖掘系统进行交互,又可以集成所有符合其标准的数据挖掘软件,为挖掘服务提供者和消费者提供统一的接口。因此,也可以将 OLE DB for DM 看成是一种接口标准。OLE DB for DM 的目的有两个:

(1)可以使不同服务提供商开发的数据挖掘算法很容易地集成到用户应用程序中,解决了数据挖掘模型部署、预测和浏览问题。

(2)给出的数据挖掘解决方案的基础结构与数据库开发环境相一致,采用统一的数据库访问接口,使企业应用程序开发者可以参与到研发数据挖掘解决方案的过程中去。

在 OLE DB for DM 中,属于单一实体的数据叫做案例,相关的一系列案例称为案例集,所有输入的数据均以案例(集)形式给出。另外,还提出概念数据挖掘模型(DMM),用于分类预测。

数据挖掘客户可以使用以下三种语句获取数据挖掘服务。

(1) Create 语句:创建数据挖掘模型对象,定义 DMM 中的列(如,在挖掘过程中需要分析的属性)和数据挖掘算法。

(2) Insert 语句:装入训练数据,训练数据挖掘模型对象,

生成模型。

(3) Select 语句:查询 DMM 的内容,做出预测或浏览由模型得到的统计结果。

在微软的 Microsoft SQL Server 2005 Analysis Services 中应用了 OLE DB for DM 标准,它通过统一的接口为用户和开发者提供服务。用户可以通过 DMX(Data Mining Extension: SQL)语言获取数据挖掘服务而不必了解内部的模型结构和工作机制。其它开发者也可以使用统一接口将其它的挖掘算法添加其中,使新算法成为挖掘工具的一部分,并且享受原有的数据集成、分析和报告服务[4]。

OLE DB for DM 使得数据挖掘市场从私有的、分散的标准变成公有的、开放的标准,利用通用接口将来源不同的数据挖掘算法方便地集成到应用程序中。从某种程度上讲,它是目前解决数据挖掘平台独立性问题的最全面的解决方案。

2.4 数据挖掘的 Web 标准

网络已经成为最大的"数据仓库"。针对 Web 数据的挖掘工作急需建立统一的 Web 数据挖掘标准。目前,与此相关的标准有:

2.4.1 XMLA(XML for Analysis)[27]

XMLA 是一种数据挖掘的 Web Service 标准,由 Microsoft 和 Hyperion 于 2001 年 4 月提出,已经得到 SAP 和 SAS 等公司的支持。XMLA 是一种基于 SOAP(Simple Object Access Protocol)的 XML 应用程序接口,用于标准化客户端应用程序和数据分析服务提供者之间的数据传输,具有跨平台性和编程语言无关性。服务器和客户端之间是松散耦合的,两者之间是基于 SOAP 的 XML 格式的数据传输。避免了紧密耦合结构带来的平台、系统、编程语言、用户应用程序和挖掘平台之间的版本依赖性。

XMLA 规范定义了两个方法:

(1)Discover:用于从网络服务上获取信息和元数据,返回的数据由参数决定。Discover 允许用公用的方法说明,不需要重写现有的函数就可以进行扩展。

(2) Execute:用于向服务器端发送命令请求,执行 MDXML 表达式或服务提供者专门的指令。MDXML 是基于 MDX 的一种语言,由单一的(Statement)元素组成,以后会陆续增加其他元素。

为了降低挖掘工具与应用程序之间的耦合度,可以通过 WSDL(Web Service Description Language)[28] 描述数据挖掘工具的挖掘服务,用 UDDI(Universal Description Discovery and Integration)[29] 协议发布、发现和使用这些挖掘服务,用 XML 描述数据和挖掘结果,SOAP [30] 作为挖掘工具之间的通信协议。将这些 Web 标准和协议相结合,能够把挖掘工具提供的算法作为一种网络服务来使用,可以解决不同挖掘工具的异构性所带来的问题。文献[31, 32] 中给出基于网络的数据挖掘服务架构的详细论述。

2.4.2 其它的 Web 标准

(1)语义网(Semantic Web)[33]:是 W3C 提出的用于应用程序、企业和社区之间数据共享和复用的通用框架标准。其中定义的 OWL(Web Ontology Language)是便于机器自动处理的信息描述语言,RDF(Resource Description Framework)是数据共享和复用的基础框架。语义网定义的框架用于存储数据挖掘结果信息,目前尚处于理论研究阶段。

(2)数据空间(Data Space)[34]:是基于网络服务的架构,用来搜索、分析和挖掘远程和分布式的数据[35]。包含用于数据搜索、查询和获取的数据空间传输协议 DSTP(Data Space Transfer Protocol)和用于实时评估的 PSUP(Predictive Scoring and Update Protocol)协议。目前,Data Space2.0 中已经应用到开源数据挖掘包 R 中。

数据挖掘的 Web 标准可以解决数据挖掘系统紧密耦合问题,通过挖掘网络上的分布式和远程数据,扩展挖掘服务的应用范围。作为 Web 标准,仍然存在数据安全性、异步网络服务、数据挖掘的会话管理状态等问题[36]。

3 基于数据挖掘标准的应用程序框架

一个统一、通用的数据挖掘系统的数据源可以来自网络中不同的存储介质,系统需将不同数据源中的数据集中到统一的数据仓库中,执行数据的清洗和转换操作。为方便不同数据仓库工具之间的数据交换,采用统一的数据挖掘元数据模型。数据挖掘工具利用统一的驱动程序存取数据仓库中的数据,并且采用统一的结果模型表示形式。应用程序通过统一的接口访问数据挖掘服务。

图 3 是基于数据挖掘标准的应用程序框架。

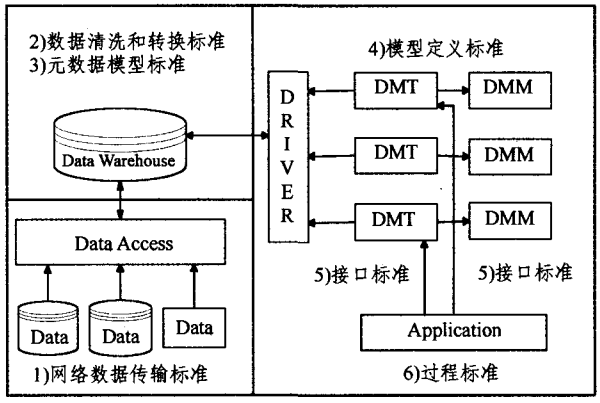

图 3 数据挖掘应用程序构架

Data:待挖掘数据,存放在关系数据库或文件中。

Data Access:获取文件、数据库或视图中的数据,并将数据保存到数据仓库。数据源可以来自分布式和远程的数据库(文件),采用统一的网络数据传输标准。

Data Warehouse:存放待挖掘数据的数据仓库,遵从统一的数据挖掘元数据模型,以支持不同数据仓库工具间的数据交换。

Driver:提供统一的数据库驱动程序连接数据仓库。

DMT(Data Mining Tool):提供不同的算法为应用程序服务。用户可以调用不同提供商提供的服务。

DMM(Data Mining Model):数据挖掘算法在数据上应用所得的结果,不同 DMT 之间可以相互调用数据挖掘模型,用于结果应用、评估和可视化。

Application:客户端应用程序,调用多个(或一个)数据挖掘服务提供商的挖掘服务,得到数据挖掘的结果模型,从而获取决策需要的信息和知识。

在此架构中,通过 Data Space 标准中 DSTP(Data Space Transfer Protocol)协议在网络上传输和获取数据,采用 CWM for DM 定义数据仓库的元数据模型,利用 PMML 标准解决各数据挖掘系统之间知识独立性问题,应用程序通过 JSR-73 或 SQL/MM 等统一的接口标准获取 DMT 的服务。整个系统开发工作遵从 CRISP-DM 标准。

如果要将应用程序部署为网络服务,可以在应用程序客户端建立支持 XMLA 标准的服务器,保障第三方可以直接获取数据挖掘网络服务。

结束语 从早期仅包含单一算法的独立数据挖掘工具,

到现在可以综合集成不同挖掘算法的横向数据挖掘工具集，应用和开发需求推动着数据挖掘技术标准化工作的发展。从数据准备到结果评估的整个数据挖掘过程中，技术标准日趋成熟。在分布式和远程数据挖掘技术中，与网格标准、网络标准相结合，也已取得了长足进步。

目前，数据挖掘标准化建设工作主要实现了以下两个目标：

(1)利用通用功能统一各种数据挖掘工具。

(2)为数据挖掘工具增添新功能留有可扩展的接口。

过程模型定义了从问题理解到模型应用的整个过程，通过实现模型中规定的细节来完成数据挖掘开发项目，它是整个数据挖掘标准的基础。查询语言标准提供描述数据挖掘服务的原语，调用这些原语可以获得数据挖掘服务。PMML 标准用于定义标准的结果模型，解决数据挖掘工具的知识独立性问题。CMW/DM 标准定义用于定义元数据，便于不同数据挖掘工具之间的系统集成、数据理解。接口标准用于数据挖掘工具与应用程序连接，上述标准可以综合使用，互为补充。

现存的技术标准还存在以下一些问题：

(1)数据转换和清洗标准。在 PMML 中已经定义了数据转换和清洗的部分操作，但是还不足以使数据形成一种统一的格式，方便地集成到数据挖掘系统中，这也是影响数据挖掘技术更广泛应用的主要障碍之一[20, 37]。

(2)具有统一接口和运行库的数据挖掘模型评价标准。以统一的方式评价数据挖掘模型，这在同一种算法有几种不同的实现方式时尤其重要。

(3)与特定问题相关的标准。如，针对流数据的挖掘标准[38]，针对网络上异构的半结构化 Web 数据和非结构化数据的挖掘标准[39]，分布式数据挖掘标准以及领域相关的数据挖掘问题的标准等。

(4)结合数据挖掘服务与网格服务的平台标准。将网格标准[40, 41]同数据挖掘标准相结合，利用网格平台的优势(如：资源分配、安全服务、状态监控、远程数据访问和存取等)，使开发者可以专注于挖掘应用本身而不是与分布式处理相关的问题。

虽然数据挖掘的标准化工作已经取得了一定进展，但是这些标准的功能还不完善，并没有被所有开发者和研究单位采用。因此，需要在以下几个方面开展进一步的研究工作：

(1)PMML 标准的完善、应用和模型部署。改进内部支持的数据挖掘模型；增添数据转换机制；增加方便用户跟踪模型的行为以及理解模型创建时环境的模型附加信息；增加异常处理机制等。在解决基于 PMML 标准的应用开发和部署问题时，需要建立结合 PMML 的 KDD 工作流标准，改善部署 PMML 模型的环境描述。

(2)挖掘各种复杂类型的数据挖掘标准。网络上存在大量的非结构化和半结构化的数据，需要对这些数据有统一的描述标准。XML 是解决问题的方向，通过 XML 定义关于数据的元数据，解决数据的输入问题。建立数据的转换和清洗标准，在被挖掘前可以按照规定的方式进行统一的数据转换和清洗。

(3)将数据挖掘标准同网络服务标准、网格服务标准相结合。解决数据挖掘系统的硬件结构、操作系统平台以及编程语言的依赖等问题，挖掘网络上的远程和分布式数据，扩大数据挖掘服务的应用对象，使得企业或个人可以方便快捷地从网络上获取服务。

(4)建立与领域相关的数据挖掘问题平台与框架标准。针对特定的问题开发特定的数据挖掘算法，以及对应的数据挖掘平台和框架。因此，数据挖掘系统应该与特定的应用相结合，建立与特定领域相关的系统框架和平台标准。

(5)数据挖掘标准体系结构的研究。目前主要有三种体系结构：客户端-服务器端结构、基于组件和代理的结构、基于网络服务的结构。良好的体系结构可以更好地利用软件环境，高效完成数据挖掘任务，有利于系统适应用户的各种需求，随时间变化对系统进行改进和优化。这方面的标准制定正在进行中，OMG(Object Management Group)正在制定基于面向对象的数据挖掘标准。

(6)统一的数据挖掘语言标准的研究。可以建立类似于 SQL 语言的形式化和标准化的数据挖掘语言标准，描述数据挖掘的语义，让用户以统一的方式调用数据挖掘的功能。

参 考 文 献

[1] Fayyad U M, Piatetsky-Shapiro G, Smyth P, et al. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996

[2] 苏卫. 数据挖掘工具的应用和标准化. 计算机工程, 2002, 30(z1): 40-42

[3] IBM. DB2 Intelligent Miner. http://www-306.ibm.com/software/data/iminer/

[4] Microsoft SQL Server 2005 Analysis Services. www.microsoft.com/SQL/techinfo/bi/ analysis.asp, 2007

[5] Fayyad U, Piatetsky-Shapiro G, myth P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996, 1(37): 37-54

[6] Cabena, Hadjinian P. Discovering Data Mining: From Concepts to Implementation. New Jersey: Prentice Hall, 1998

[7] DaimlerChrysler Corporation. Cross Industry Standard Process for Data Mining [EB/OL]. http://www.crisp-dm.org, 2007

[8] Cios K J, Kurgan L A. Trends in Data Mining and Knowledge Discovery// Pal N R, Jain LC. & Teodersesku, N. eds. Knowledge Discovery in Advanced Information System. Spriner, 2002

[9] 潘无名，潘云鹤. 数据挖掘过程的多维视图. 计算机应用研究, 2004, 21(8): 211-214

[10] Shrarer C. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 2000, 5(4): 13-22

[11] Melton J, Eisenberg A. SQL Mutimedia and Application Packages(SQL/MM) ACM SIGMOD Record, 2001, 30(4): 97-102

[12] Llingenfelder C. SQL/MM Data Mining and its Relation to other Data Mining Standards // Proceedings of the DM-SSP 03 Workshop. Washington DC, 2003

[13] JSR-73 Expert Group. Java Specification Request 73: Java Data Mining (JDM) - JDM Final Specification, 2004

[14] Hornick M. Java Data Mining (JSR-73): Overview and Status // Proceedings of the First Annual Workshop on Data Mining Standards, Services, and Platforms. Washington DC, 2003

[15] Oracle. Oracle Data Mining, www.oracle.com/technology/products/bi/odm/index.html

[16] Han J, Kamber M. 数据挖掘：概念和技术[M]. 范明，孟晓峰. 北京：机械工业出版社, 2001

[17] Imielinski T, Virman A. MSQL: A Query Language for Database Mining[J]. Data Mining and Knowledge Discovery, 1999, 3(4): 373-408

[18] Meo R, Psalia G, Crei S. A New SQL-like Operator for Mining Association Rules[C]// Proceedings of International Conference on Very Large Database(VLSB'96). Bombay, 1996

(下转第 14 页)

[9] 伍胜男,伍春洪. 三维集成图像技术及其在发展三维电视上的应用[J]. 江西科学,22(2),2004:110-114
[10] NHK STRL ANNUAL REPORT 2002 —Studies for Future Broadcasting Services and Fundamental Technologies
[11] http://www.extra.research.philips.com/euprojects/attest/. ATTEST2002-2004, European IST-2001-34396 project
[12] Wojciech M, Pfister H. 3D TV: A Scalable System for Real-time Acquisition, Transmission, and Autostereoscopic Display of Dynamic Scenes. Sig, 2004
[13] http://www.ddd.com Dynamic Digital Depth Research Pty Ltd
[14] Stern A, Javidi B. Three dimensional Image Sensing, Visualization and processing using integral imaging [J]// Proc. IEEE. 2006, 94(3):591-607
[15] Okano F, et al. Real-time integral imaging based on extremely high resolution video system [J] //Proc. IEEE. 2006, 94(3): 490-501
[16] Davis N, et al. Design and analysis of an image transfer system using microlens arrays [J]. Opt. Eng, 1994, 33(11):3624-3633
[17] Ives H E. Optical Properties of a Lippmann Lenticulated Sheet [J]// Opt Soc Amer. 1931, 21:171-176
[18] Chutjian A, Collier R. Recording and reconstructing threedimensional images of computer generated subject by Lippmann integral photography, Appl. Opt., 1968, 7(1):99-101
[19] Ren Jinsong. Software Tools for Integral Imaging Virtual Studios. PhD Thesis. De Montfort University, 2003
[20] Igarashi Y, Murata H, Ueda M. 3D display system using a computer genrated integral photograph. Japan J Appl Phys., 1978, 17(9)
[21] Cartwright P. Realisation of computer generated integral threedimensional images. PhD Thesis. De Montfort University, December 2000
[22] Milnthorpe G, McCormick M, Aggoun A, et al. Computer generated content for 3D TV displays // Proc. IBC 2002. Amsterdam, 2002
[23] Milnthorpe G, McCormick M, Davies N. Computer modeling of lens arrays for integral image rendering // Proc. 20th Eurographics UK Conference. Leicester, UK, June 2002:136-141
[24] Nakajima S, Nakamura K, Masamune K. Three-dimensional medical imaging display with computer-generated integral photography. Computerized Medical Imaging and Graphics, 2001, 25:235-241
[25] Jang J, Javidi B. Three-dimensional integral imaging of micro-objects [J]. Optical Letters, 2004, 29(1):230-1232
[26] Javidi B, Moon I, Yeom Seokwon. Three-dimensional identification of biologicalmicroorganism using integral imaging, Optics Express, December 2006, 14(25):12096-12108
[27] Wu C, et al. Depth measurement from unidirectional integral images using a modified multi-baseline disparity analysis algorithm [J]. Journal of Electronic Imaging, 2005, 14(2)
[28] Yeom S, Javidi B. Three-dimensional distortion - tolerant object recognition using integral imaging [J]. Optics Express, 2004, 12(23):5495-5808
[29] Park Jae-Hyeung, Kim Yunhee, Lee Byoungho. Elemental image generation based on integral imaging with enhanced resolution// Proceedings of SPIE. 2005, 5642:186-194
[30] Zalevsky Z, García-Martínez P, García J. Superresolution using gray level coding. Optics Express, 2006, 14(12):5170-5182
[31] Jang Ju-Seog, Javidi B. Three-dimensional projection integral imaging using micro-convex-mirror arrays. Optics Express, 2004, 12(6):1077-1083
[32] Choi Kyongsik, Kim Joohwan, Lim Yongjun, et al. Full parallax viewing-angle enhanced computergenerated holographic 3D display system using integral lens array. Optics Express, 2005, 13 (26):10494-10502
[33] Buckhardt C B. Optimum parameters and resolution limitation of integral photography. J. Opt. Soc. Amer., 1968, 58:71-76
[34] Okoshi T. Optimum design and depth resolution of lens-sheet and projection-type three-dimensional displays. Appl. Opt., 1971, 10:2284-2291
[35] Ren Jinsong, Aggoun A, McCormick M. Maximum viewing width integral image, Journal of Electronic Imaging, 2005, 14 (2):23019-1-9
[36] Park J-H, Min S-W, Jung S, et al. Analysis of viewing parameters for two display methods based on integral photography. Appl. Opt., 2001, 40:5217-5232
[37] Stern A, Javidi B. 3-D computational synthetic aperture integral imaging (COMPSAII). Optics Express, 2003, 11(19): 2446-2451
[38] Jang J S, Javidi B. Three-dimensional synthetic aperture integral imaging. Opt. Lett., 2002, 27:1144-1146
[39] Fukushina R, Taira K, Saishu T, et al. Novel viewing zone control method for computer generated integral 3-D imaging. SPIE - IS&T, 2004, 5291:81-92
[40] Lee Byoungho, Jung Sungyong, Park Jae-Hyeung. Viewing-angle-enhanced integral imaging by lens switching. Optics Letters, 2002, 27(10):818-820
[41] Price M, Thomas G A. 3D virtual production and delivery using MPEG-4 // Proc. of the Int. Broadcasting Convention (IBC 2000). Amsterdam, 2000:8-12
[42] Hoshino H, Okano F, Isono H, et al. Analysis of resolution limitation of integral photography [J]. Opt. Soc. Am., 1998, 15:2059-2065
[43] Brown K, McCormick M, Davies N. The use of computer generated integral images to visualize cyber-sculpture // Proceedings of the 20th Eurographics UK Conference (EGUK. 02). IEEE, 2002:3-8
[44] Liao Hongen, Tamura Hongen, Iwahara M, et al. High Quality Autostereoscopic Surgical Display Ussing Anti-aliased Integral Videography Imaging. MICCAI 2004, LNCS, 2004:464-469
[45] 伍胜男,伍春洪. 全景图像技术及其应用[J]. 科技通报, 2005, 21(6):714-717

(上接第 10 页)

[19] Data Mining Group. PMML 3.1: Predictive Model Markup Language. http:// www.dmg.org /pmml-v3-1.html
[20] Grossman R. Standards, Services and Platforms for Data Mining // Proceedings of the First Annual Workshop on Data Mining Standards, Services, and Platforms. Washington DC, 2003
[21] Pechter R. Data Mining Standards, Services and Platforms 2005 Workshop Report // Proceedings of the Third Annual Workshop on Data Mining Standards, Services, and Platforms. Chicago, 2005
[22] OMG. Common Warehouse Metamodel (CWM) Specification, Version 1.1, 2003, 1(3)
[23] 王强,刘东波,王建新. 数据仓库元数据标准研究. 计算机工程, 2002, 28(12):123-125
[24] OracleWarehouseBuilder. http://www.corba.org/vendors/pages/oracleCWM.html, 2007
[25] XELOPES Library. http://www.prudsys.com/Produkte/Algorithmen/Xelopes, 2007
[26] Microsoft Corporation. OLE DB for Data Mining Specification V1.0, 2000
[27] XMLA(XML for Analysis). http://www.xmla.org, 2007
[28] WSDL(2007). W3C Note. http://www.w3.org/2002/ws/desc/
[29] UDDI(2007). Universal Description, Discovery and Integration specification, v3.0. http://www.uddi.org
[30] SOAP(2007). W3C Note. http://www.w3.org/2000/xp/Group/
[31] Kurgan L, Cios K J, Trombley M. The www Based Data Mining Toolbox Architecture// Proceedings of the 6th International Conference on Neural Networks and Soft Computing. Zakopane, Poland, 2002
[32] Krishnaswamy S, Zaslavsky A, Loke S W. Towards Data Mining Services on the Internet with a Multiple Service Provider Model: An XML Based Approach. Journal of Electronic Commerce Research, 2001, 2(3):103-130
[33] Sementic Web. www.w3c.org/2001/sw, 2007
[34] Data Space. www.dataspaceweb.net, 2007
[35] Grossman R, Mazzucco M. Data Space: A data Web for the exploratory analysis and mining of data // IEEE Computer Society. 2002, 4(4):44-51
[36] Chu R. Web Service Standards for Data Mining // 2nd International Workshop on Data Mining Standards, Services and Platforms. Chicago, 2004
[37] Grossman R L, Hornick M F, Meyer G. Data Mining Standards Initiatives, Communication of ACM, 2002, 45(8):59-61
[38] Amini L, Andrade H, Bhagwan R. SPC: A Distributed, Scalable Platform for Data Mining. ACM SIGKDD Exploration, 2006, 8(2):82-83
[39] Ferrucci D, Grossman R L, Levas A. PMML and UIMA Based Frameworks for deploying Analytical Applications and Services // KDD-2006 Workshop on the Data Mining Standards, Services and Platforms. Philadelphia, PA, 2006
[40] Chien A A, Sun Xian-he, Xu Zhi-wei. Viewpoints on Grid Standards. Journal of Computer Science and Technology, 2005, 20(1):141-143
[41] Baker M, Apon A, Ferner C, et al. Emerging Grid Standards Computer, 2005, 38(4):43-50