文章编号:1000 - 582X(2004)03 - 0036 - 05

联机分析挖掘(OLAM)技术的现状与发展*

蒲晓湘,刘文才
(重庆大学 自动化学院,重庆 400030)

摘 要:从联机分析处理技术与数据挖掘技术的互补性出发,介绍了联机分析挖掘(On-Line Analytical Mining, OLAM)技术的形成原因、功能特征、分析操作方法以及模型结构,分析了目前该技术存在的主要问题以及实现的关键技术,并展望了 OLAM 未来的发展方向。

关键词:数据库;数据仓库;联机分析处理;数据挖掘;决策支持
中图分类号:TP311.138 **文献标识码**:A

联机分析挖掘,又称为 OLAP Mining[1]。它是联机分析处理技术与数据挖掘技术在数据库或数据仓库应用中的结合,是联机分析处理技术的新发展,也是近年来数据库领域的研究重点和热点。

而数据仓库、联机分析处理和数据挖掘则是 20 世纪 90 年代中期国外兴起的 3 种决策支持技术[2]。

数据仓库(Data Warehouse, DW)是在数据库的基础上发展起来的。1992 年,W. H. Inmon 首次提出数据仓库概念:"支持管理决策过程的、面向主题的、集成的、稳定的、不同时间的数据集合形式"[3]。数据仓库为复杂分析、知识发现和决策提供数据访问。数据仓库在规模、历史数据、数据集成和综合性、查询支持等方面都和传统数据库有着本质区别[4]。作为一种新型的数据存储地,数据仓库为数据挖掘和数据分析提供了新的支持平台。基于数据仓库的数据挖掘,面对的是经加工和概括的数据,简化了数据挖掘过程的某些步骤,大大提高了数据挖掘的工作效率。同时,数据仓库事先收集、归纳、处理了整个企业范围内的数据,为数据分析提供高质量的数据源,更好地支持管理决策[5-6]。

数据挖掘(Data Mining, DM)是在人工智能、机器学习中发展起来的,也称为数据库中的知识发现(KDD)。1995 年在美国计算机年会(ACM)上首次提出数据挖掘概念:是从大型数据库或数据仓库中提取隐含的、未知的、具有潜在使用价值的信息或模式的过程[7]。DM 通过分析大量的原始数据,作出归纳性的推理,挖掘出潜在模式并预测客户的行为,为企业的决策者提供决策支持[8]。

联机分析处理(On-Line Analytical Processing, OLAP)是由关系数据库之父 E. F. Codd 于 1993 年提出的[10]:是共享多维信息的、针对特定问题的联机数据访问和分析技术。具有快速性、可分析性、多维性、信息性、共享性 5 个特点[9-10]。OLAP 具有灵活的分析功能、直观的数据操作和分析结果可视化表示等突出优点,从而使用户对基于大量复杂数据分析变得轻松而高效,以利于迅速做出正确的判断,辅助决策[11]。

OLAM 正是在这 3 种技术的基础上建立起来的。它的出现为企业管理和决策活动提供了一个新的工具,也为决策支持系统的研制提供了新思路。为了让大家对 OLAM 技术有一个全面的认识,笔者就目前 OLAM 的发展现状进行了归纳分析,并展望了它未来的发展趋势。

1 OLAM 技术的现状

1.1 OLAM 技术的形成原因

OLAP 与 DM 虽同为数据库或数据仓库的分析工具,但两者侧重点不同[11-12]。同时,随着 OLAP 与 DM 技术的应用和发展,数据库领域在 OLAP 基础上对深层次分析的需求与人工智能领域中数据挖掘技术的融合最终促成了联机分析挖掘技术[13]。

一方面,分析工具 OLAP 功能虽强大,能为客户端应用程序提供完善的查询和分析,但它也存在以下

* 收稿日期:2003 - 10 - 12
作者简介:蒲晓湘(1973 -),女,四川西充人,重庆大学硕士研究生,研究方向为计算机监控与管理信息系统。

不足[14-15]：

1)OLAP 是一种验证型分析工具，是由用户驱动的。即在某个假设的前提下通过数据查询和分析来验证或否定这个假设，这很大程度上受到用户假设能力的限制。

2)OLAP 分析事先需要对用户的需求有全面而深入的了解，然而用户的需求并不是确定的，难以把握。所以 OLAP 分析常常采用试凑法在大型数据库或仓库中搜索，不仅花时间，而且可能产生一些无用的结果。

3)即使搜索到了有用的信息，由于缺乏应有的维度，从不同的视图得到的结果可能并不相同，容易产生误导。

另一方面，数据挖掘虽然可以使用复杂算法来分析数据和创建模型表示有关数据的信息，用户也不必提出确切的要求，系统就能够根据数据本身的规律性，自动地挖掘数据潜在的模式，或通过联想，建立新的业务模型以辅助决策。但它也存在一些缺点[7,16]：

1) DM 是挖掘型分析工具，是由数据驱动的。用户需要事先提出挖掘任务。但对于用户来讲，很多时候预先是不知道想挖掘什么样的知识的。

2)由于数据库或数据仓库中存有大量数据和信息，用户仅仅指出挖掘任务，而不提供其他搜索线索，这样 DM 工具就会遍历整个数据库，导致搜索空间太大。计算机将处于长时间的工作，而且结果中可能会生成很多无用信息。

3)即使挖掘出了潜在有价值的信息，但它究竟用来做什么分析用，用户也可能不太清楚。

可以看出，两种技术各存在不足，但同时也可以相辅相成。如果将 OLAP 同 DM 配合集成，一方面 OLAP 的分析结果给 DM 提供挖掘的依据，引导 DM 的进行；另一方面，在数据挖掘的结果中进行 OLAP 分析，则 OLAP 分析的深度就可拓展。这样用户就可以灵活选择所需的数据挖掘功能，并动态交换挖掘任务，在数据仓库的基础上提供更有效的决策支持。鉴于 OLAP 与 DM 技术在决策分析中的这种互补性，促成了 OLAM 技术的形成，其中所包含的关键技术可用公式(1)来表达[17]

$$OLAM = DW + OLAP + DM \quad (1)$$

但 OLAM 不是这 3 种技术的单纯叠加，而是指多种技术的无缝集成，这种集成将带来 OLAM 技术与其构件技术在基本概念、原理、技术、方法、机制、结构、使用等方面本质上的不同。

1.2　OLAM 概念的提出

正是由于 OLAP 与 DM 技术的相辅相成性，在 OLAM 概念提出之前，实际应用中试图将 OLAP 与 DM 结合起来提供更加优质的数据分析和决策支持的思路早也有之。如文献[7]提出“只有将 OLAP 技术、DM 技术和 DW 中的庞大数据相结合，与企业先进的管理决策方法相结合，才能使 DW 在企业的经营管理决策中发挥巨大的作用”。一些厂商也开始在 OLAP 的基础上添加数据挖掘功能，如 Business Object 公司的 Business Object 产品中的决策树分析、DBMiner 系统中的数据挖掘算法工具箱等，初步实现了两者的结合。而另一些是把数据挖掘算法集成在系统的底层功能中，如 Microsoft 公司的 SQL Server 着 000 中的关联分析方法在数据库端的集成就是实现 OLAP 与 DM 紧密结合的初步尝试。

联机分析挖掘概念正式提出是在 1997 年，由加拿大 Simon Fraser 大学教授 Jiawei Han 等在数据立方体的基础上提出多维数据挖掘的概念，称为 OLAP mining[18-19]。这实际上是在 OLAP 系统的基础上，把数据分析算法、数据挖掘算法引入进来，解决多维数据环境的数据挖掘问题。因此这时的 OLAM 实际上还是 OLAP 和 DM 的松散结合。之后，国内外研发人员在这方面展开了积极的工作，试图将 OLAP 与 DM 技术有机结合起来形成真正的 OLAM 技术和产品。如文献[20]就对 OLAM 的概念进行了扩展，将其定义为联机分析挖掘处理(On - Line Analytical Mining Processing)，其分析和挖掘的数据基础也扩大成包括多维数据模型和关系数据模型等在内的多种模型的异构环境，研究重点在如何实现 OLAP 与 DM 技术紧密集成。

1.3　OLAM 的功能特征

OLAM 融合了 3 种技术，兼有 OLAP 和 DM 的优点，在 DW 上的数据挖掘和分析更具灵活性和交互性。其功能特征如下[1,21]：

1)相对 OLAP 和 DM 技术，OLAM 具有较高的执行效率和较快的响应速度。

2) OLAM 能对任何它想要的数据进行挖掘。OLAM 建立在 OLAP 的基础之上，因此应能方便地对任何一部分数据或不同抽象级别的数据进行挖掘，甚至还可以直接访问存储在底层数据库里的数据。

3)OLAM 中，用户可以动态选择或添加挖掘算法，并可以动态切换挖掘任务。

4)OLAM 中挖掘任务具有多样性、算法具有复杂性，因此应具有标签和回溯功能。标签功能即是标记用户的操作状态功能，回溯指的是退回到上次操作状态。OLAM 这种功能可以避免用户因算法的复杂性而在超立方体中“迷失方向”。

5)OLAM具有灵活的可视化工具。可视化工具以丰富的图文有效地显示分析和挖掘结果给用户,从而实现交互式处理。

6)良好的扩展性。这是指OLAM应该高度模块化,能与其他多个子系统集成。

7)友好的人机交互能力。OLAM的决策分析过程是要在人的指导下进行,人作为系统的组成部分和系统应用密不可分。人与计算机分别承担各自最擅长的工作,实现资源的合理配置。

1.4 OLAM的模型结构

就目前来看,OLAM的结构体系还没有统一的模式。国内一些文献在这方面作了一定的研究,提出了一些OLAM模型。如文献[1]认为OLAM体系结构和OLAP并没有本质区别,结构可以同一。并结合Web技术,提出了基于Web的OLAM模型。文献[20]给出了OLAM概念模型、逻辑模型和物理模型。其中的概念模型指出了必须执行的功能以及这些功能之间的关系。逻辑模型把概念模型中所定义的结构映射到可用软件、过程和体系结构的环境中,它是基于技术类型对基本设计原则的细化,是连接概念模型和物理模型的桥梁。文献[22]提出了基于影响域的OLAM模型,文献[20]将OLAM的体系结构分为4层:数据存储层、多维数据库层、OLAP/OLAM层和用户接口层。这些都是对建立OLAM模型结构的可喜探索。

总之,设计一种高效、优化的OLAM体系结构,是OLAP、DM和DW 3种技术完善集成的重要保证,也是支持OLAM系统提供灵活可靠决策功能的硬件基础,这已成为研究人员正在努力解决的重点问题之一。

1.5 OLAM的分析操作

从OLAM的定义来看,它是建立在多维数据视图基础之上的。因此,对于OLAM的操作应是超立方体计算与传统挖掘算法的结合。这里所说的立方体计算方法一般指切片、切块、上卷、下钻、旋转等操作;而挖掘算法则是指关联规则、分类、聚类等挖掘算法。根据立方体计算和数据挖掘所进行的次序的不同组合可以有以下一些模式[20,23]:

1)先进行立方体计算、后进行数据挖掘。在进行数据挖掘以前,先对多维数据进行一定的立方体计算,以选择合适的数据范围和恰当的抽象级别。

2)先对多维数据作数据挖掘,然后再利用立方体计算算法对挖掘出来的结果做进一步的深入分析。

3)立方体计算与数据挖掘同时进行。在挖掘的过程中,可以根据需要对数据视图做相应的多维操作。这也意味着同一个挖掘算法可以应用于多维数据视图的不同部分。

4)回溯操作。OLAM的标签和回溯特性,允许用户回溯一步或几步,或回溯至标志处,然后沿着另外的途径进行挖掘,这样用户在挖掘分析中可以交互式的进行立方体计算和数据挖掘。

1.6 目前OLAM技术存在的主要问题

OLAM技术现在已取得很大的发展,但总的来说,目前研究工作仍处于起步阶段,很多问题还没有得到解决或重视。其中,OLAM技术面临的主要问题是[20,24]:

1)关于OLAM技术的界定模糊。系统体系结构标准和参照还没有出现,无法区别和衡量现有的所谓的OLAM系统。

2)OLAM模型中信息的表示是OLAM技术的数据基础,目前还没有统一的标准来解决OLAM环境中多种信息如数据、模式等的规范问题。

3)OLAM系统的数据基础是包括多种数据模型在内的异构数据环境。传统的基于关系数据模型或多维数据模型的OLTP和OLAP的任务/事务模型已不适合于OLAM系统。

4)OLAM基于DM和OLAP,但不同于两者的单纯叠加,目前还没有一种优化的管理策略来成功融合这两种技术,实现无缝连接。

总的来说,OLAM目前存在的主要问题是技术理论研究滞后于OLAM产品的开发。关于OLAM技术的基本原理、关键技术,系统整体组织结构、应用开发技术等问题的研究相对较少,也不系统。

2 OLAM技术的发展

2.1 OLAM实现的关键技术

为了成功实现OLAM交互式探索性的数据分析,联机选择数据挖掘功能,动态交换数据挖掘任务,除了解决以上的OLAM技术界定、信息统一表示等规范问题外,以下关键技术尚待解决[9]:

1)OLAM环境中的数据结构是复杂多样的,以支持不同的数据分析方法以及挖掘算法。因此,支持复杂数据环境的数据组织存储是实现OLAM技术的关键之一。

传统的数据仓库已不适合于OLAM。这方面,文献[20]提出了工作仓库的概念,其中的数据和信息称为工作对象。它是数据仓库的扩展,扩展为包括多种数据模型在内的异构环境,以支持不同的OLAM任务,提供灵活的数据类型定义和快速的数据组织方式。工作仓库具有独立性,其数据的存储和组织都由用户

来定义,生存期也是基于该用户权限自定义的。

2)OLAM 建立在多维数据库和 OLAP 的基础之上,因此基于超立方体的高性能挖掘算法应是其核心所在。开发出支持复杂维度和度量的高性能数据立方体技术、以及基于这种立方体的数据挖掘算法应是研究的重点。

这方面,文献[22]提出了影响域(Influence Domain)的概念,影响域是一种广义的数据立方体。立方体上计算的是聚合(Aggregation),而影响域上计算的是蕴涵(Implication),即数据中隐藏的模式。影响域同立方体一样具有属性和值,不同点在于它具有置信度(Confidence)。立方体将维映射至度量,而影响域将维和度量映射至置信度。因此影响域更适合于 OLAM 分析。文献[20]提出了:第一,基于一维数组的高效数据立方体,并由它构建一种 HOLAP,在其基础上提出了关联规则的挖掘算法,这种 HOLAP 实现了快速性和灵活性的平衡,同时也为数据挖掘提供了较好的数据空间;第二,基于数据立方体的关联规则挖掘算法——维内关联规则算法 Intradim_asso_mining,它类似于 Apriori 算法,区别就是它扫描的数据是立方体的一个切片。

3)开发出新的适合于 OLAM 的任务/事务模型,规范数据挖掘分析任务定义语言。

OLAM 中的任务具有多样性和复杂性,兼有数据查询任务、OLAP 任务、DM 任务等多种任务,传统的定义在关系数据库基础上的任务语言如 DMQL(Data Mining Query Language)语言、MINE RULE 操作符等已不适合 OLAM 任务定义的要求。因此需要一套任务定义语言来支持 OLAM 任务的定义和管理。这方面,文献[20]提出了一种基于约束的 OLAM 任务定义语言。它的设计基础是"约束",通过对 OLAM 任务描述相关信息的分类,形成多种"约束",这些约束的联合表示了一个 OLAM 任务和执行方式。

4)OLAM 的挖掘过程是对复杂数据环境不断深入的过程,应具有书签和回溯功能。因此,完善的原数据存储和管理以及中间结果缓存是支持 OLAM 这种功能的基础。

5)OLAM 应具有快速的响应能力和较高的执行效率,这是 OLAM 中最为困难的问题。由于一般挖掘算法都复杂而且耗时,加之 OLAM 与用户频繁交互,因此在执行效率与挖掘的准确性之间应该协调好,选择合适的挖掘算法和数据搜索空间是很必要的。

6)OLAM 应该具有一个通用的标准接口,以便与其他挖掘工具或算法相衔接,以实现在多个数据挖掘功能之间的交互、动态选择或能添加新的挖掘算法[21]。

2.2　OLAM 技术的发展趋势

OLAM 技术实现了 OLAP 和 DM 技术的互补,它的发展趋势是两者更加可靠的集成、融合,有自己合理优化的结构体系和一套完备的技术理论基础,从整体上为决策分析提供完美支持。

OLAM 技术是一门交叉学科,涉及机器学习、模式识别、统计学、智能数据库、人工智能、高性能计算、数据可视化、专家系统等综合技术。这些相关学科的发展,无疑也将会推动 OLAM 技术的发展。特别是,近年来随着数据库技术的发展,出现了不同数据类型的高级数据库,如面向对象数据库、对象关系型数据库、空间数据库、超文本数据库、多媒体数据库、时序数据库等。因此,未来的 OLAM 技术应用应基于这些高级数据库展开。

随着互联网技术的发展,全球信息的共享,基于 Web 的联机分析挖掘(WebOLAM),也将成为 OLAM 技术发展的一个新方向[25]。

3　结束语

OLAM 是 DW、OLAP 与 DM 相结合的产物,它兼有 OLAP 多维分析的在线性、灵活性和数据挖掘对数据处理的深入性,是数据库或数据仓库应用工具未来发展的方向,也为决策支持系统开辟了一条新的途径。笔者主要介绍了目前 OLAM 技术的发展现状,从技术界定、信息规范和异构环境的整合等方面指出了该技术目前主要存在的问题,并就其实现的关键技术提出了一些自己的观点。自 OLAM 概念的提出到成熟完善的 OLAM 技术及应用,应该是一个循序渐进、不断摸索的过程。因此,对于 OLAM 的研究工作仍在继续不断地进行着,还会有很多问题值得大家共同来研究和探讨。

参考文献:

[1]　刘夫涛,张雷,艾波. OLAM 以及基于 WEB 的 OLAM[J]. 计算机工程与应用,2000,9:108 - 109.

[2]　MALLACH E G 著. 决策支持与数据仓库系统[M]. 李昭智,李昭勇译. 北京:电子工业出版社,2001.

[3]　ELMASRI R, NAVATHE S B 著. 数据库系统基础[M]. 邵佩英译. 北京:人民邮电出版社,2002.

[4]　韦洛霞. 数据仓库与 OLAP[J]. 东莞理工学院学报,2000,7(2):20 - 24.

[5]　高洪深. 决策支持系统(DSS):理论·方法·案例[M]. 北京:清华大学出版社,2000.

[6] 张红云. 数据挖掘与决策支持系统的关系[J]. 鞍山师范学院学报, 2001,3(3):86 - 89.
[7] 陈京民. 数据仓库与数据挖掘技术[M]. 北京:电子工业出版社. 2002.
[8] 陈文伟. 决策支持系统及其开发[M]. 北京:清华大学出版社,1994.
[9] 杨光,张雷,艾波. 数据仓库及联机分析处理技术[J]. 计算机工程与科学,2000,22(1):39 - 42.
[10] 王珊. 数据仓库技术与联机分析处理[M]. 北京:科学出版社,1998.
[11] 辛志. 提高 OLAP 系统性能的方法研究[J]. 计算机科学,2003,30(5):8 - 11.
[12] 石磊,石云. OLAP 与数据挖掘一体化模型的分析与讨论[J]. 小型微型计算机系统, 2000,21(11):1 208 - 1 210.
[13] 马丽娜,刘弘,张希林. 数据挖掘、OLAP 在决策支持系统中的应用[J]. 计算机应用研究,2001,11: 10 - 12.
[14] SEIDMAN C. SQL Server 2000 数据挖掘技术提高[M]. 刘艺,王鲁军,蒋丹丹,等译. 北京:机械工业出版社,2002.
[15] 辛志,刘少辉,史忠植. 提高 OLAP 系统性能的方法研究[J]. 计算机科学,2003,30(5):59 - 62.
[16] 朱明. 数据挖掘[M]. 合肥:中国科学技术大学出版社,2000.
[17] 晏凌,王应解. 建立基于联机分析挖掘的信息资源导航系统[J]. 图书情报知识,2000,3:56 - 57.
[18] HAN JIAWEI. OLAP mining: An Integration of OLAP with Data Mining [EB/OL]. http://www-faculty. cs. uiuc. edu/ ~ hanj/. 2003 - 10 - 10.
[19] HAN JIAWEI, CHANG K C C., Data Mining for Web Intelligenc[EB/OL]. http://www-faculty. cs. uiuc. edu/ ~ hanj/, 2003 - 10 - 10.
[20] 曹蓟光. 联机分析挖掘处理技术(OLAM)的研究[D]. 杭州:浙江大学,2001.
[21] HAN JIAWEI, CHEE SONNY H S, CHIANG JENNY Y. Issues for On-Line Analytical Mining of Data Warehouses [EB/OL]. http://www-faculty. cs. uiuc. edu/ ~ hanj/. 2003 - 10 - 10.
[22] 石磊,石云,刘欲晓,等. 基于影响域的 OLAM 模型的研究[J]. 郑州大学学报(自然科学版), 2000, 32(2):16 - 20.
[23] 刘夫涛. 从 OLAP、数据挖掘到 OLAM[EB/OL]. http://www. sqlmine. com/warehouse/htm/7. htm,2003 - 10 - 02.
[24] 周爱广. OLAM 技术及其在财务分析系统中的应用[D]. 济南:山东大学,2001.
[25] 飞思科技产品研发中心编著. SQL Server 2000 数据库和数据仓库[M]. 北京:电子工业出版社,2001.

An Overview of On-line Analytical Mining Technology and It's Development Prospect

PU Xiao-xiang, LIU Wen-cai

(College of Automation, Chongqing University, Chongqing 400030, China)

Abstract: An Overview of On-line Analytical Mining technology (short for OLAM) is given, including it's reasons of developing, characteristics, operating methods and structure mode. Especially, the main problems about OLAM and it's key technologies of implementation are discussed in detail. At last, the future of OLAM is prospected.

Key words: database; data warehouse; on-line analytical processing; data mining; decision support systems

(编辑 张 苹)