计算机科学 1997Vol. 24 №. 2

应用的深化推动数据库技术发展

——VLDB'96 纵横

杨冬青　唐世渭

（北京大学计算机科学技术系　北京 100871）

第 22 届超大型数据库国际会议（VLDB'96）于 1996 年 9 月 3 日至 6 日在印度孟买举行，来自 20 余个国家和地区的 450 多位代表出席了会议。会议收到 340 篇论文，从中选出 48 篇收入论文集并在会上报告。会议安排了 4 个特邀报告、7 个技术讲座、3 场专题讨论、16 场研究论文报告，和 6 场工业/应用论文报告。

从本次会议接受的论文和安排的报告反映出一种面向实际、面向应用的倾向。对于为满足实际应用的需求而发展起来的新技术，和对应用的发展将会产生重大影响的新技术，以及数据库系统实现技术接受了较多的论文，安排了较多的报告，而对于纯理论性的研究和与实际应用距离较大的技术，则论文和报告都很少。

本文第一部分对 VLDB'96 安排的报告和讨论作一总的概述，第二、三、四部分分别对其中的一些议题进行更为详细一些的介绍。

一、面向实际、面向应用是总的趋向

数据库系统是实用性非常强的系统。数据库技术的进步推动数据库应用领域的扩展和应用层次的提高。反过来，新的应用需求又促进了数据库技术的进一步发展。从近年的情况看，有广泛应用需求的新技术，和对应用模式和应用的发展将会产生重大影响的技术问题很容易成为技术的热点，也成为学术研究的热点。VLDB'96 的热点之一是面向对象技术与数据库技术的结合，及其最新进展--对象-关系数据库系统。VLDB'96 程序委员会组成了一个特别委员会，从 VLDB'86 的论文中挑选一篇 10 年来在数据库领域中产生了最大影响的论文。挑选的结果，中选的是 Michael J. Carey 等的论文"EXODUS 可扩充数据库系统中的对象和文件管理"。并且，Michael J. Carey 被邀请到 VLDB'96 来做了题为"对象和数据库：十年混战"的特邀报告。除了他的报告外，会议还安排了两个技术讲座和一场研究论文报告，讨论的都是面向对象与数据库技术结合的问题，及与此相关的新的 SQL 标准问题。（本文第二部分介绍有关的内容）。VLDB'96 的另一个热点是数据仓库和 OLAP，这是随着数据库应用深化，为了解决如何有效地将数据库中的大量数据变成有用的信息、为决策支持服务而提出的问题。随着过去几年中数据仓库方面大量的市场活动，有关的研究工作也活跃起来了。关于这方面的内容，会议安排了两场研究论文报告，1 个技术讲座，和 1 次专题讨论。本文第三部分将对有关的内容进行讨论。

除上述两个技术热点外，会议安排的其他特邀报告，技术讲座，专题讨论所涉及的也都是数据库领域中近年来迅速发展，具有重要的实际意义的新研究领域、新技术，以及对数据库领域的未来发展有重要意义的标准和技术。例如，数据库管理系统和 Internet，图象和影象数据库，数据挖掘，大型协同式信息系统的设计和实现，商业应用环境中的超大型数据库，数据库互操作性和可移植性标准，变化中的软件工业，未来数据库系统的性能等。各场研究论文报告所介绍的研究工作除上述的一些技术外，还较多地涉及了数据库管理系统的实现技术，例如查询处理、查询优化、并行查询处理、空间存取方法、I/O 优化、数据完整性/安全性等。本文第四部分将对上述内容做简要介绍。

VLDB'96 的安排中还有一个比较突出的特点，那就是对计算机工业界的重视。特邀报告和技术讲座的讲演者和专题讨论的发言人中，有一半以上是来自计算机软件、硬件公司的研究人员和工程技术人员。这样的安排在一定程度上反映了计算机工业界对技术发展的重要作用。工业界与学术界相比，他们更面向实际，更接近用户，更了解实际应用的需求，因此在应用的深化推动数据库技术发展的形势下，整个数据库界对计算机公司的技术动向有了比以往更多的关注，这也是很自然的事情。

二、面向对象与数据库

1. 回顾与展望

IBM Almadem 研究中心的 Michael J. Carey 所做的题为"对象和数据库:十年混战"的特邀报告是对面向对象技术引入到数据库领域以来 10 年中的研究项目和商业软件开发的回顾,对当前状况的评估和对未来 10 年研究和发展前景的展望。

他首先追溯到 10 年前面向对象技术引入到数据库领域时的情况。如何将面向对象技术与数据库技术相结合,当时的研究工作所采用的方法主要可以分为以下四类:

1)扩充关系数据库系统。主要是将关系数据库系统的类型系统开放,允许用户定义自己的抽象数据类型。

2)持久的程序设计语言。即采用面向对象程序设计语言的类型系统和编程模式,并使数据成为持久的,程序执行成为原子的。

3)面向对象的数据库系统。将数据库系统的特性与面向对象程序设计语言的特性结合起来,产生面向对象的数据库系统。

4)数据库系统工具包/部件。提供数据库管理系统核心和工具包,以便在各个层次上对数据库管理系统进行扩充,以满足不同应用领域的需求。

1986 年前后,在这四个方面上都有研究项目在进行。计算机工业界落后于学术界,当时只有两家 OODB 小公司,努力将它们的早期产品推向市场。

十年后的 1996 年,再来回顾这四个方向上的工作。方法 2 和方法 4 产生了若干有趣的研究成果,但在商业实践的意义上并不成功。原因很多,包括灵活性不够,太繁琐,一些其他方法提供的可扩充性已能满足需要等等。方法 3 产生了许多研究成果,并且一些新兴的规模不大的公司推出了面向对象数据库系统产品,但它的商品化现状显然比以前预期的要差,因为 OODB 产品缺乏统一标准,在许多特性上不如 RDB 产品,应用开发工具不足等等。方法 1 现在换了个名字,称作对象-关系数据库系统,看来是现在最占优势、最能满足数据库应用需求的系统。另外,在数据库市场上还出现了一种新的方法,面向对象的客户包装层。即对关系数据库加一个对象包装层,以支持开发传统数据库上的面向对象的客户端应用。

展望十年后的 2006 年,对象-关系技术将成为成熟的技术,向用户提供的应是高度集成的,客户/服务器结构的,可伸缩的,健壮的对象-关系的数据库系统。为使这一光明前景成为现实,在这十年中还有许多研究工作需要做,包括集成问题,对象-关系查询的并行处理问题,对遗产数据源的存取问题,查询语言的标准化问题等。

2. 对象-关系数据库(ORDB)

Michael J. Carey 的特邀报告认为,在面向对象技术与数据库技术相结合的几种方法中,对象-关系数据库系统是当前的优胜者和今后的发展方向。ORACLE 公司的 Anil K. Nori 的技术讲座对对象-关系数据库及有关的技术进一步做了详细的阐述。

对象-关系数据库系统是扩充关系数据库方法走向成熟的产物,并且在发展过程中吸取了面向对象数据库方法的许多长处。当前的关系数据库系统和当前的面向对象数据库系统都不能满足新的多媒体应用,迅速发展的 Web 应用,以及新的商业应用的需求,这些应用需要对复杂数据进行复杂查询。对象-关系数据库系统将关系数据库管理系统的功能和特性与面向对象的建模能力结合起来,从而提供对复杂数据进行复杂查询的支持。对象-关系数据库系统具有许多面向对象的特征,但它与面向对象的数据库系统不同,它是从关系数据库模型和其查询语言 SQL 出发,在此基础上建立起来的。对象-关系数据库模式的最顶层仍然是命名的关系(表)的集合,但是关系中的对象可以很丰富,象面向对象数据库系统支持的一样。SQL 语言关于对象查询的扩充包括路径表达式,类似于方法的函数调用语法,对于 FROM 子句中嵌套集合的支持等。

现在已有几家厂商推出了具有对象-关系数据库特征的产品,例如 IBM DB2/CS V2,Illustra, UniSQL 等。可以预见,大多数的数据库厂商都将会推出可靠的对象-关系数据库管理系统(ORDBMS)。

不同的厂商推出的对象-关系数据库系统在抽象数据类型定义、复杂对象构造、查询语言扩充等方面存在着诸多差异,这对于系统的开放性非常不利,因此标准化是一个亟待解决的问题。好在新一代的数据库语言标准 SQL3 正在制定这些方面的标准,而大多数的数据库厂商对 SQL3 的进展都持非常积极的态度,并在其产品中遵循 SQL3 标准。

VLDB'96 安排了一个 SQL3 技术讲座,由 Sybase 公司的 Jim Melton 和 IBM Santa Teresa 实验室的 Nelson M. Mattos 主讲。他们对 SQL 标准的发展、SQL3 的现状、SQL3 中对关系数据和操作的进一步支持、SQL3 的面向对象特性以及 SQL3 的对象-关系扩充等做了详细的阐述。

三、数据仓库与联机分析处理(OLAP)

对于数据仓库和 OLAP 这个与数据库应用密

切相关的议题，会议安排了一个专题讨论，而且邀请的发言人全部是计算机工业界人士。此外，还安排了1个技术讲座和两场论文报告。这些讲座和报告对数据仓库、OLAP的有关概念、模型等做了较全面的介绍，对某些方面的技术进行了较深入的讨论。

1. 概念

数据仓库技术和OLAP技术是数据库技术和OLTP技术发展、数据库应用深化的产物。其目的是把数据库中的大量数据转化为有用的信息，为企业更好地进行决策服务。

OLAP（联机分析处理）与传统的OLTP（联机事务处理）相比较，有许多不同的特点。其用户主要是企业的分析、决策人员，而不是业务处理人员；它的功能是业务分析和决策支持，而不是日常的事务处理；它所需要的数据是历史的、总括的、集成的、多维的数据；而不是当前的、详细的、孤立的、纯关系的数据。所以，它的数据量很大，其数据库设计是面向分析决策的，而不是面向事务处理的，OLAP的数据访问方式通常是即席的、复杂的查询，而不是重复的，简单查询和更新。

如果在支持OLTP的数据库上支持OLAP应用，则有很多矛盾。因为要支持数据的多维视图和典型的OLAP操作需要特殊的数据组织，存取方法和实现方法，例如多维数据库组织、多维聚集计算等，这些是OLTP数据库所不支持的。另外，OLAP的复杂查询花费的时间长，如果放到OLTP数据库上做会降低事务处理的性能。因此，需要将数据库中的数据抽取出来，加以集成和重新组织，成为单独存储的数据仓库，以支持OLAP应用。

按照W. H. Inmon在《建立数据仓库》一书中的定义，“数据仓库是面向主题的、综合的、不同时间的、稳定的数据集合，主要用于支持经营管理中的决策制定过程”。数据仓库是面向主题，即按某种客观分析领域来组织的，它综合了企业网络不同信息点上的数据库中的不同时间段的数据，从而能更好地支持分析决策。

2. 技术

数据仓库技术源于数据库技术，但由于其组织结构和操作类型的特色，又形成了自己的技术分枝。

数据仓库的数据库模式适应多维数据分析的需要。比较典型的是星形模式（Star Schema）。星形模式由一个事实（fact）表和多个维表构成。事实表（或称主表）中包括业务事件的信息，这些信息有多个维度。每个维度对应一个维表（或称辅表），它包括一个维度的描述信息。事实表中的一个事实指向每个维表中的一个元组。星形模式结构简单，表的数目少，易于理解。除星形模式外，还有事实星座模式（Fact Constellation Schema，它有多个事实表，这些事实表共享多个维表），雪花模式（Snowflake Schema，它将维表规范化，直接表示维中属性的层次）等。

数据仓库和OLAP的主要技术问题还包括数据仓库体系结构，数据仓库的数据库服务器实现技术等。数据仓库的数据库服务器实现技术大体上分为三类：①专用的关系DBMS，即在索引技术、扫描方法、复杂查询处理等诸方面加以特殊处理的关系DBMS。②关系的OLAP（ROLAP）DBMS，即扩充RDBMS，将多维数据上的操作映象到标准的关系操作。③多维OLAP（MOLAP）DBMS，即直接实现多维数据管理和操作。

数据仓库和OLAP的操作问题包括数据抽取和建模、数据装入、数据刷新、构造导出数据和视图、查询处理、多维聚集计算、数据仓库监控等。有几篇研究论文报告了在其中一些问题上较深入的研究工作。例如Stanford大学的题为“数据仓库的高效的快照差异算法”的论文，德国RWTH Aachen的题为“外部的物质化的视图的增量维护”的论文讨论的都是数据仓库的刷新问题。Wisconsin大学的两篇论文反映出他们在多维聚集计算方面做了很深入的工作。

3. 产品

在数据仓库和OLAP领域，似乎是市场活动比学术研究更加积极、更加活跃。数据库厂商和其他软件厂商推出了建立数据仓库和OLAP环境的多个方面的软件和工具，包括关系型的数据仓库数据服务器、ROLAP服务器、MOLAP服务器、数据源连接软件、数据抽取工具、数据转换工具、数据刷新工具、查询/报表工具、多维数据分析工具、原数据管理工具、数据仓库管理和监控工具等。选择适当的软件环境和工具，是成功地建立数据仓库和OLAP应用的关键任务之一。

四、其他议题

会议安排的其他论文报告、技术讲座等涉及了数据库技术的发展，数据库应用技术，数据系统实现技术的许多方面。

1. 数据库管理系统与Internet

在Web节点数迅速增加、Java语言逐步普及的形势下，大多数商品化的数据库管理系统已推出了

Web 接口、信关等。那么数据库研究方面有哪些工作要做呢？VLDB'96 安排了一个题为"数据库管理系统与 Internet"的专题讨论会，讨论会的发言人就跨异构数据源的查找、数据库 Web 接口的自动生成、数据库与 Web 的集成问题等展开了讨论。会上还提出了这样的问题：是否已到了建立新的数据管理服务，将适合于当今的 Internet 环境的结构包括进来的时候了？如果数据库界不做，别人就会去做了。

2. 未来数据库系统的性能

并行处理器、64 位处理器、磁盘缓存等技术的发展，数据库技术的发展，以及应用的发展，这些对未来数据库系统的性能评价有什么影响？专题讨论会"未来数据库系统的性能"就此展开了讨论。会上的发言所涉及的问题包括未来数据库系统的性能用哪些标准衡量，是只关心服务器的性能，还是关心客户/服务器性能，数据库引擎的性能和使用数据库的应用系统的性能哪个更重要，顾客和厂商关心的性能一样吗？等等。

3. 异构数据库系统

两场论文报告和 6 篇论文介绍了有关异构数据库系统的研究成果，包括多数据库系统中互操作性语言的设计，数据库互操作中的完整性约束规则，对于数据模式未知的数据进行查询的语言的处理策略，从不完整的数据库获取完整的查询结果的方法等。此外，还有两个技术讲座，分别阐述了大型协同式信息系统的设计和开发方法，以及数据库互操作性和可移植性的多个标准，这些标准的进展、现状和未来趋势。

4. 多媒体数据库

题为"图象和影象数据库"的技术讲座讨论了可按内容寻址的图象数据库的许多基本问题，包括数据库语义、数据模型、索引、查询处理等，介绍了在此新领域中的若干个原型和商品化系统。此外，还有几篇研究论文介绍了在正文数据库检索，医疗图象数据库的相似查找等方面所做的研究工作。

5. 查询处理和查询优化

论文集中收入的关于查询处理和查询优化的论文有十数篇之多，一方面由于这是数据库系统实现的关键技术之一，另一方面也是由于面向对象数据库，多媒体数据库，并行数据库等新的领域为查询处理和查询优化提出了新的研究课题。论文所介绍的研究工作包括对于传统的数据库和各种新型数据库进行查询的处理策略，以及为提高查询效率而设计的各种优化算法。

6. 数据挖掘(Data Mining)

与数据仓库和 OLAP 相类似，数据挖掘的主要目的也是将数据库中的大量数据变成有用的信息。数据挖掘将隐藏在大型数据库中的原先未知的数据关联模式发掘出来。几篇研究论文介绍了在关联规则挖掘方面所做的研究，包括挖掘关联规则的新的操作符，新的算法等。此外，关于数据挖掘的技术讲座以 Quest 数据挖掘系统为背景，介绍了关联、顺序模式、分类、时间序列聚集、分段等数据挖掘技术，以及当前正在进行的一些研究工作。

7. 数据库应用

本届会议鼓励面向实际应用，介绍将数据库技术，尤其是一些新技术应用到各行各业的实践经验和体会的论文，并在论文集中将这些论文特别用星号标记出来。一篇论文介绍了在保健管理系统中采用工作流技术，对一个州范围内的儿童免疫进行跟踪的实践；另一篇论文介绍了在健康保险信息系统中运用数据挖掘技术的经验和体会；等等。会议还安排了一个特邀报告，讨论在传统的商业应用环境中开发非常大的数据库系统所涉及的问题。

其他的议题还包括数据完整性/安全性，I/O 优化，空间数据存取方法等。篇幅所限，不再一一赘述。

参考文献

Proceedings of The Twenty-second International Conference on Very large Data Bases, Morgan Kaufmann Publishers, Inc.