在决策支持系统中应用数据仓库技术的研究

赵玉勇　吴永明　（同济大学计算机科学与工程系　200092）

摘要：传统的决策支持技术发展到现阶段，已成为具有四库结构的智能型决策支持系统。而九十年代出现的数据仓库技术更加有力地推动了决策支持的发展。本文说明了两种技术系统的概念、结构及发展状况，提出了以数据仓库为核心的综合决策支持系统，并对其功能框架、逻辑结构及内部各部件之间的支持、集成加以论述。

关键词：决策支持系统　数据仓库　模型　知识　联机分析处理　数据挖掘

一、前言

1. 决策支持系统

决策支持系统（DSS）是由电子数据处理系统（EDPS）、管理信息系统（MIS）逐步发展而来的，是支持半结构化和结构化决策，允许决策者直接干预并能接受决策者的直观判断和经验的动态交互式计算机系统。

自从 Scott Morton 等人在 70 年代初提出计算机对于决策的支持作用和决策支持系统的概念后，二十年来，随着决策理论、计算机技术、人工智能、信息技术的发展，DSS 无论在概念、结构方面还是在应用方面都取得了较快的发展。

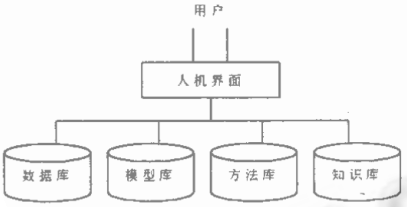

图 1

对于 DSS 的结构，最初由 R. H. Sprague 提出了基于人机对话系统、数据库与模型库的两库结构。而后出现的三库结构则实现了模型与方法的分离存储，即添加了方法库。在近年来，把人工智能技术、专家系统、知识工程的思想方法引入 DSS 后，即在原来的结构基础上，增加了知识库，并引入了推理机制，就形成了 DSS 的四库结构框架（图 1），从而使 DSS 成为智能决策支持系统（IDSS）。

有人认为 DSS 是由语言系统 LS、问题处理系统 PPS 和知识系统 KS 三部分组成，这三种系统实际上是由上面提到的四库结构的基本部件发展而来的。所以，本文还是按照四库结构的决策支持系统进行讨论。

2. 数据仓库技术

90 年代初，W. H. Inmon 提出了“数据仓库”的概念：数据仓库就是面向主题的、集成的、稳定的、不同时间的数据集合，用以支持经营管理中的决策制定过程。

数据仓库概念提出的意义在于，使数据操作型环境与数据分析型环境分离开来，建立一种数据存储体系结构，把分散的、不利于访问的数据转换成集中、统一、随时可用的信息，从而可以集成不同形式的数据，并为数据分析产品提供系统开放性。

数据仓库为不同来源的数据提供了一致的数据视图，一经与数据挖掘（DM，Data Mining）、联机分析处理（OLAP，On Line Analytical Processing）等数据分析技术相结合，即实现了为用户提供灵活自主的信息访问权力、丰富的数据分析与报表功能的目的，使企业数据得到充分的利用。

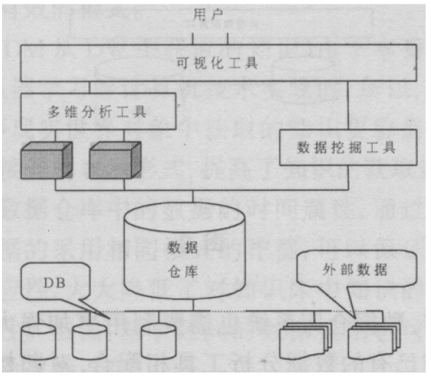

图 2

数据仓库系统由数据仓库(DW)、数据仓库管理系统(DWMS)、数据仓库工具三个部分组成,结构如图 2。

数据仓库管理系统负责管理整个系统的运转,包括从 OLTP 数据库、市场报告及各种文档等数据源进行数据抽取、清理与转换,划分维数及确定数据仓库的物理存储结构,以及对数据的安全、备份、恢复等工作,是整个系统的引擎;

数据仓库则包含了早期细节级、当前细节级、轻度综合级、高度综合级的数据,是整个数据仓库系统的核心;

而数据仓库工具则通过使用 OLAP 分析工具、数据挖掘工具及查询检索工具,实现各种需求,是整个系统发挥作用的关键。

目前,大型企业几乎都在建立或计划建立自己的数据仓库系统,数据库厂商也纷纷推出自己的数据仓库软件。已经成功建立和使用的数据仓库应用系统都取得了明显的经济效益。

二、功能框架的提出

决策支持系统为了更有效的实现对企业高层管理人员的支持,需要掌握充分的信息,从而经常需要访问大量的、不同数据源的、当前或历史的数据,即使得到所需的数据,还需要对其中具体的、细节的数据进行综合、总结、概括。而这些正符合数据仓库内数据的特点。

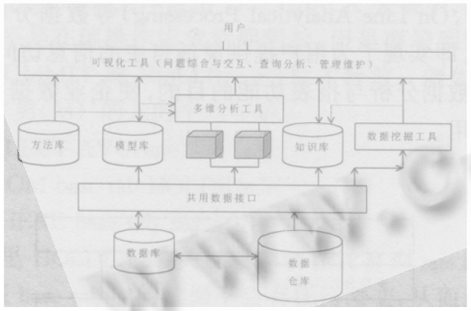

图 3

同样,数据仓库系统也需要利用更加强大的决策支持技术与已有的数据分析工具相配合,对归档数据及多维分析处理数据进行更利于决策的处理。而事实上,建立以数据仓库为核心的综合决策支持系统(图 3),将决策支持系统技术与数据仓库技术有机的结合在一起,必将会更大限度的发挥对决策的支持作用。

数据库、数据仓库及共用数据接口是系统的数据管理部分,构成了整个系统的核心与基础,向上层应用提供所需数据。其中数据库与数据仓库的管理分别通过的数据库管理系统(DBMS)与数据仓库管理系统(DWMS)进行。

模型库与方法库、知识库、数据挖掘工具、联机分析处理构成了工具层,相互配合协调,完成用户的决策处理任务。其中模型库、方法库及知识库的管理则通过各自的管理系统进行。可视化工具,包括了问题分析与综合、查询分析、管理维护等功能,实现了用户与系统的动态交互。

本文将对数据管理部分及工具层各部件之间的集成与支持加以论述。

三、系统各部件的集成与支持

1. 数据库、数据仓库及共用数据接口

决策支持系统中的数据必须可以满足各种层次、各种类型、不同决策者的要求,并且要求数据管理系统应能根据决策活动的需要,把有关的数据面向决策过程组织起来。

数据仓库中的数据是面向主题组织的,主题对应了针对企业中某一宏观分析领域所涉及的分析对象,面向主题组织从而确保在较高层次上实现企业信息系统中数据的综合、归类并进行分析利用的抽象。另外,其他三个特点保证了数据仓库的数据是不同时间期间的、各种层次的综合数据和细节数据。从而满足了驱动一个决策过程的数据要求。数据仓库是整个系统的基础和核心。

模型库做为决策支持系统的重要部件,要求数据与模型的有机结合。模型必须与所需要的数据相匹配,才能被用于决策过程。所以对于数据仓库中不满足要求的数据,首先需要组织整理。决策过程是一个与用户动态交互的过程,在这个过程需要一些预先输入的数据、会产生一些中间数据及结果数据,要求可以快速的维护及查询。还包括系统本身的一些数据管理工作,都由数据库来完成。

共用数据接口则协调工具层对数据的需求,完成工具层、数据库、数据仓库之间必要的数据调度,有效的对数据检索、查询及操作处理。根据工具层部件在一个决策过程中对数据的需求信息,可以实现不同的功能:通过对数据仓库及数据库的检索查询实现对数据的直接利

用，通过在数据库中实现数据的重新组织实现对数据的预处理，通过在数据库插入删除更新等操作实现对决策过程中产生的中间数据及系统维护数据的管理。

数据仓库及数据库自身的管理与维护工作由各自的管理系统完成。实际上，对于在决策过程中，根据经常进行的数据仓库数据的重新组织的需求，可以在数据仓库中按所面对的主题进行组织实现，进而提高决策支持的效能；另外，数据库中的数据，也可以作为数据仓库的外部数据来源，组织到数据仓库中去。

2. 共用数据接口对模型库的支持

相对于独立的决策支持系统，本系统中共用数据接口对模型库的支持，主要来源于数据仓库的大量数据。表现在以下几个方面：

(1)建立模型。在数据仓库系统的大数据量的支持下，可以用选择样本的办法。但样本的选择必须合理，它对于大多数业务问题来说起码不会损失信息。当合理选择样本，同时实现了真正的随机性选择时，建立在所有数据上的多个模型通常并不比建立在一个样本上的多个模型更准确有效。因为，事实上，所有的全部数据通常也只是实际数据的一个样本。因此，在把数据分成训练数据组和测试数据组时，也必须进行抽样。

在进行数据抽样时，最关注的是能够保留关键信息，其实这一点可以从统计理论和实践中得到保障。从统计论上来讲，由于数据仓库能够提供足够多的数据，搜寻的对象也足够普遍，那么，抽样出来的数据不会损失关键的和重要的信息，在这些抽样数据基础上建立起来的模型的准确性，也会得到最大限度的保障。

(2)模型使用。传统的决策支持系统由于缺乏使用模型所必须的数据，即使在模型准确的情况下，由于使用的数据并不具有普遍性，从而导致计算结果偏离实测结果，甚至会发生极大差异或错误的极端情况。而实际使用许多有效的模型，都需要大量数据作保证，才能真正发挥作用。

数据仓库中大量的集成、统一、综合或细节的、历史或当前的数据，使这种情况发生几率大大减小，为各种模型的计算结果准确、及时支持决策，打下了坚实的基础。

(3)模型维护。模型在使用后，还必须严格考察模型的工作情况。因为无论模型的准确率有多高，仍然不能保证它能够如实的反映世界。一个正确的模型不一定是最佳的，导致这个问题的主要有三个原因：模型中总是隐含着某些假设；另一个主要原因是无论数据收集和准备工作做得多么好，数据本身总会有许多不可避免的问题；建模后，实体或环境发生了变化，而模型未能及时反映这种变化。

为了保证模型的正常运行，必须不断监测模型的运行情况，根据需要进行重新测试、再训练、甚至彻底重构模型。可以采用类似建模时采用的样本方法，并绘出预测值与观测值的差别图，达到监控模型结果的目的，甚至将图表建立在软件当中，实现系统的自我监控。而这些，都是在数据仓库提供批量数据的前提下进行的。

3. 模型库、方法库对联机分析处理的支持

联机分析处理(OLAP)的概念最早是由 E. F. Codd 提出的。OLAP 是针对特定问题的联机数据的访问和分析，通过对信息的很多种可能的观察形式进行快速、稳定一致和交互性的存取。OLAP 是通过建立在 OLAP 服务器中的用户预定义的多维数据库进行，通过多维分析工具与数据仓库打交道。

OLAP 常用的分析采用切片、切块、旋转等基本动作组成。与模型库及方法库中的模型、方法有效的结合，来分析处理多维数据，将会极大的提高 OLAP 的分析能力。

4. 数据挖掘工具对知识库的支持

知识库中存放有经验的决策者的决策知识、推理规则、完整性条件、元知识合语义关系。推理机的主要任务是选择知识合应用知识，按照一定的推理策略，运用启发式方法和各种搜索策略，有条件导出结果，向决策者提出解决问题的方案。

DM 是一种决策支持过程，是从大型数据库或数据仓库中发现并提取隐藏在其中的信息的一种新技术。目的是帮助决策者寻找数据间潜在的关联，发现被忽略的要素，而这些信息对预测趋势和决策行为也许是十分有用的。它主要基于 AI、机器学习，统计学技术，对企业数据进行提取与分析、归纳推理，挖掘出能够被人理解的可信、新颖、有效的模式。

通过 DM 从 DW 中获取的知识，由于本身是采用人工智能、机器学习等计算机技术实现的，所以，要比从专家、规律等现实世界对象中获取的知识更容易转化为知识库所能接受的表示形式，提高了知识的获取速度。

结合数据仓库中的数据的时间属性，通过对不同时间区间数据的采用相同模式的挖掘，可以保证知识库中知识的适应性，大大降低了对知识库中知识的维护和引入的复杂性。当然，基于区间的数据挖掘也可以从地理位置等其他属性入手。

另外，采用针对不同区间的挖掘策略，从而得到基于不同区间的知识，通过采用对比分析等方法，可以获得关

于知识的知识。而这样的知识是建立在更高的认知层次上的，对于从宏观角度去把握决策，对于从全局角度实现包括用户在内的整个的决策支持交互系统的自我调整，都是有意义的。

5.数据分析工具

数据仓库、OLAP 和数据挖掘是作为三种独立的信息处理技术出现的。但都是以解决决策支持分析问题为主要驱动力量发展起来的，由于这三种技术内在的联系性和互补性，三者的结合合本身就是一种基于数据库技术的 DSS 的解决方案。其中数据仓库用于数据的存储和组织；OLAP 集中于数据的分析；数据挖掘则致力于知识的自动发现。

四、结束语

以数据仓库为核心的综合决策支持系统，是更高级的决策支持系统。在这样一种决策支持的系统环境下，由于数据仓库的数据，准确高效的模型、方法、知识及强大的分析工具做保障，用户的决策分析会更加全面、有效、深刻。从另一方面讲，在由决策过程中驱动的，用户对于系统的数据组织、模型重构等技术调整由于系统的强大支持，而更具有实际的意义。从而，形成了人与系统之间协调、支持的良性循环。

参考文献

[1] 王珊等 编著，《数据仓库技术与联机分析处理》，科学出版社

[2] 刘卫东、王诚、周立柱，大型信息系统的数据组织，计算机研究与发展，1997.6

[3] 张宝红、樊惠娟、王能斌，数据仓库的实现技术，计算机科学，1998.2

（来稿时间：1998 年 12 月）