基于数据仓库的数据挖掘技术研究现状与进展*

朱玉颖，刘宏伟，张 岩
（西南科技大学，四川 绵阳 621010）

摘 要：随着时间的推移，社会的进步，越来越多的数据被海量积累下来，如何合理处理数据，并利用相关数据获取人们所需的知识，是进入21世纪以来人们一直深入研究的方向。以此为出发点，从数据仓库与数据挖掘的诞生谈起，详细介绍了数据仓库的构建、几种数据挖掘算法以及数据挖掘过程，分析提出了数据挖掘技术的进一步发展和研究方向。

关键词：SQL Server 2000；数据仓库；数据挖掘
中图分类号：TP311.131 **文献标识码**：B

Research and progress of data mining based on data warehouse

ZHU Yu Ying, LIU Hong Wei, ZHANG Yan
(Southwest University of Science and Technology, Mianyang 621010 ,China)

Abstract: With rapid development of society, human has accumulated more and more data information magnanimously. How can human deal with the data and make use of it to get the knowlegde that they want? It is still the way that human search for. This article take this as a starting point. Then it mentions from the data warehouse and data mining's birth, and also introduces in detail data warehouse's construction, several kind of data mining algorithm,as well as the data mining process. The ultimate analysis proposed further development and the research direction for the data mining technology.

Key words: SQL Server 2000; data houseware; data mining

随着计算机应用技术的快速发展，令各行各业收集数据的能力大力提升，随之也就带来了“数据爆炸”现象。如何将这些海量数据存储与分析，令其转换成信息和知识，辅助决策管理，成为亟待解决的问题。由此，数据仓库与数据挖掘技术应运而生。

20世纪90年代初期INMON W H.在其里程碑式的著作《Building the Datahouse》中提出了“数据仓库”的概念[1]，而后随着数据库与计算机技术的不断进步，数据仓库技术也得以快速发展，并逐渐渗透到生物医学、零售、医学信息系统、移动通信等行业中。

数据挖掘技术自从1989年8月在底特律召开的研讨会上提出后迅速发展，该研讨会组委会在1997年开始拥有了自己的杂志“Knowledge Discovery and Data Mining”，并且在数据仓库的基础上，在保险业务、金融风险预测、基因工程研究、产品产量和质量分析等领域中得到了成功应用。

1 数据仓库技术

1.1 数据仓库介绍

数据仓库不仅包含分析所需的数据，而且包含处理数据所需的应用程序，这些程序包括将数据由外部媒体转入数据仓库的应用程序，也包括了将数据加以分析并呈现给用户的应用程序。

根据该定义，一个数据仓库包括了数据以及负责管理与分析工作的程序管理器，其主要目的是提供可用的数据，使分析人员可以取得所需的正确统计信息，以作为管理决策的参考依据。

* 基金项目：四川省重点学科技术研究项目(01GY051-37)

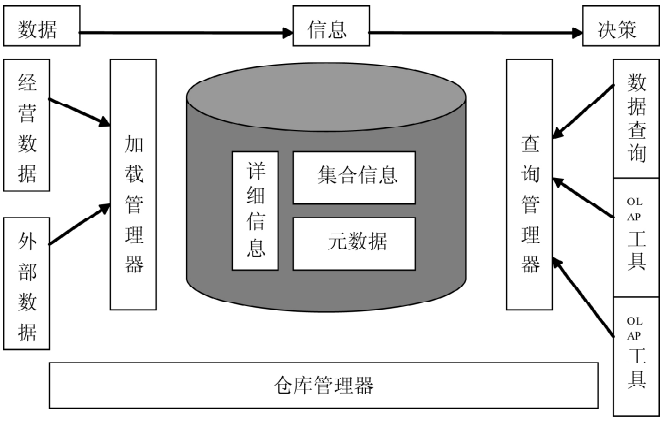

图1 数据仓库系统的架构

1.2 数据仓库系统的架构

一个数据仓库的大小一般都是在100 GB以上，因为传统的关系数据库技术是针对OLTP而发展的，并不适用于数据量大而且复杂度高的数据仓库系统，需要用不同的方式来设计和开发一个数据仓库系统。

因此提出一个新名词——系统管理器，它是由多个程序所构成的，而成为层次式的架构，至于一个管理程序的功能与复杂度则会因数据仓库系统而异。系统管理器向下可分为3个管理器：加载管理器（Load Manager）、仓库管理器（Warehouse Manager）和查询管理器（Query Manager）。图1所示架构图实现了一个数据仓库系统的架构，就数据层面而言，图中标示出了3个层次：数据、信息以及决策。而图中各管理器职责为：

（1）加载管理器：程序需实现抽取与加载，功能为抽取并加载数据，在加载数据之前与进行中执行简单的转换。

（2）仓库管理器：程序需实现整理与转换、备份与备存，功能为转换并管理数据仓库数据、备份与备存数据。

（3）查询管理器：程序需实现查询功能，从而可引导并管理数据仓库的查询。

1.3 数据仓库设计

数据仓库的设计和创建是一个分布实施的连贯过程，在确定用户需求的基础上，完成数据仓库的设计和建立、提取和加载，最后进行长期的使用和维护。从系统的角度看，数据仓库的建立首先必须明确其设计方法，针对解决问题的短期性或长效性，将数据仓库设计方法分为以下3种：

（1）自顶向下的方法：该方法把企业需求作为实现数据仓库的首要任务，其成本、难度和时间花费都远远大于自下向上的方法，一般适合于取得长期效益。

（2）自底向上的方法：该方法设计较小的、更集中的数据仓库应用，可以简化整体处理过程，为兼顾缩短开发时间和可缩放企业应用提供了折中的方案，是快速实现数据集市、部门数据仓库的有效手段。

（3）联合方法：是以上两种方法的合成，企业在保持自底向上方法实现和基于应用的同时，还能利用自顶向下方法的规划和决策，为企业保留建立长远决策方案提供了机会[2]。

2 数据挖掘算法

2.1 数据挖掘理论

数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含的、预先不知道的、但又潜在有用的信息和知识的过程。数据挖掘的相近术语，包括知识发现、数据分析、数据融合（Data Fusion）以及决策支持等。人们把原始数据看作是形成知识的源泉，就像从矿石中采矿一样。原始数据可以是结构化的，如关系数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至可以是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。已发现知识不仅可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。

2.2 数据挖掘基本算法

2.2.1 关联规则

（1）关联规则的定义

设$I=\{i_1, i_2, \cdots, i_m\}$是项的集合[3]。设任务相关的数据D是数据库事务的集合，其中每个事务T是项的集合，使得$T \subseteq I$。每个事务有一个标识符，称作TID。设A是一个项集，事务T包含A当且仅当$A \subseteq T$。关联规则是形如$A \Rightarrow B$的蕴涵式，其中$A \subset I$，$B \subset I$，并且$A \cap B=\varnothing$。规则$A \Rightarrow B$在事务集D中成立，具有支持度s，其中s是D中事务包含$A \cup B$(即A和B二者)的百分比，它是概率$P(A \cup B)$。规则$A \Rightarrow B$在事务集D中具有置信度c，如果D中包含A的事务同时也包含B的百分比是c。这是条件概率$P(B|A)$。即是

$$\text{support}\ (A \Rightarrow B)=P(A \cap B) \quad (1)$$

$$\text{confidence}\ (A \Rightarrow B)=P(B|A) \quad (2)$$

如果项集的出现频率大于或等于min_sup与D中事务总数的乘积，则项集满足最小支持度min_sup；如果项集满足最小支持度，则称它为频繁项集，频繁k-项集的集合通常记作L_k。关联规则分为两步：找出所有频繁项集和由频繁项集产生关联规则。

（2）Apriori算法：使用候选项集找频繁项集

Apriori算法[4]是一种最有影响的挖掘布尔关联规则频繁项集的算法。Apriori使用一种称作逐层搜索的迭代方法，k-项集用于探索(k+1)-项集。首先，找出频繁1-项集的集合。该集合记作L_1。L_1用于找频繁2-项集的集合L_2，而L_2用于找L_3，如此下去，直到不能找到频繁k-

项集。找每个 L_k 需要一次数据库扫描。

为提高频繁项集逐层产生的效率,用 Apriori 性质来压缩搜索空间，该性质称为 Apriori 性质：即频繁项集的所有非空子集都必须也是频繁的。Apriori 性质基于如下观察：根据定义，如果项集 I 不满足最小支持度阈值 min_sup，则 I 不是频繁的，即 $P(I)$<min_sup。如果项 A 添加到 I，则结果项集(即 $I \cup A$)不可能比 I 更频繁出现。因此，$I \cup A$ 也不是频繁的，即 $P(I \cup A)$<min_sup。该性质属于一种特殊的分类,称作反单调，意指如果一个集合不能通过测试，则它的所有超集也都不能通过相同的测试。

(3)由频繁项集产生关联规则

一旦由数据库 D 中的事务找出频繁项集，由它们产生强关联规则是直截了当的，对于置信度，可以用下式，其中条件概率用项集支持度计数表示：

$$\text{confidence}(A \Rightarrow B)=P(B|A)=\frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)} \quad (3)$$

其中 support_count($A \cup B$)是包含项集 $A \cup B$ 的事务数，support_count(A)是包含项集 A 的事务数。根据该式，关联规则可以产生如下：

对于每个频繁项集 l，产生 l 的所有非空子集；

对于 l 的每个非空子集 s，如果

$$\frac{\text{support\_count}(l)}{\text{support\_count}(s)} \geqslant \min\_conf \quad (4)$$

则输出规则为“$s \Rightarrow (l-s)$”，其中，min_conf 是最小置信度阈值。

Apriori 算法在剪枝步中的每个元素需在交易数据库中进行验证来决定其是否加入，这里的验证过程是算法性能的瓶颈，这个方法要求多次重复扫描可能很大的交易数据库，还会产生大量的候选项集，这是 Apriori 算法的两大缺点。

2.2.2 遗传算法

遗传算法是进化计算方法的实例,是优化型算法。遗传算法是一个计算模型，由 5 部分组成：个体的初始集合 P、杂交技术、变异算法、适应度函数以及对 P 反复应用杂交技术和变异技术的算法。该算法用适应度函数确定 P 中应保留的最优个体。算法每次迭代都从种群中替换许多预先定义的个体，直至达到某一阈值为止。遗传算法的优点是容易并行化，但它也存在许多缺点：遗传算法对于最终用户来说很难理解和解释、问题抽象和个体表述十分困难、最佳的适应度函数难以确定以及杂交和变异过程难以确定。

2.2.3 决策树

决策树[5]是一个类似于流程图的树结构，其中每个内部节点表示一个属性上的测试，每个分枝代表一个测试输出，而每个树叶节点代表类或类分布。决策树根据不同的特征，以树型结构表示分类或决策集合，产生规则和发现规律。决策树的算法主要有：ID3 算法、C4.5 算法、SLIQ 算法和 SPRINT 算法。

3 数据挖掘的过程

3.1 确定业务对象

清晰地定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结构是不可预测的，但要探索的问题应是可预见的，为了数据挖掘而数据挖掘则带有盲目性，是不会成功的。

3.2 数据准备

（1）数据的选择，搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用的数据；

（2）数据的预处理，研究数据的质量，为进一步分析作准备，并确定将要进行的挖掘操作的类型；

（3）数据的转换，将数据转换成一个分析模型，这个分析模型是针对挖掘算法建立的，建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

3.3 数据挖掘

对所得到的经过转换的数据进行挖掘，除了完善从选择合适的挖掘算法外，其余一切工作都能自动地完成。

3.4 结果分析

解释并评估结果，其使用的分析方法一般应以数据挖掘操作而定，通常会用到可视化技术。

3.5 知识的同化

将分析所得到的知识集成到业务信息系统的组织结构中去。

总之，数据挖掘过程需要多次的循环反复，才有可能达到预期的效果，如图 2 所示。

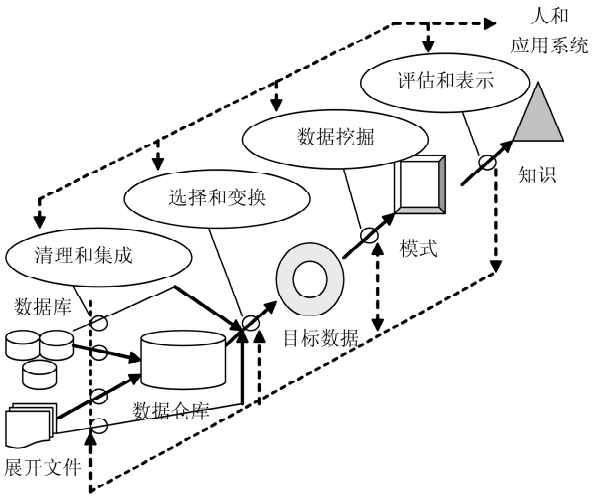

图 2 数据挖掘——数据库知识发现核心步骤

4 研究展望

随着大量算法的完善，数据仓库技术越发成熟，挖掘过程的系统化和规范化、挖掘工具的不断推陈出新，数据仓库与数据挖掘技术已显示了它广泛的应用前景。如：

（1）应用的探索：目前正探索扩大其应用范围，如生物医学、电信等领域。

（2）可伸缩的数据挖掘方法：一个重要方向是基于约束的挖掘，该方向致力于在增加用户交互同时改进挖掘处理的总体效率。

（3）数据挖掘与数据库系统、数据仓库系统和Web数据库系统的集成：数据挖掘系统的理想体系结构是与数据库和数据仓库系统的紧耦合方式。

（4）数据挖掘语言的标准化。

（5）可视化数据挖掘：可视化数据挖掘是从大量数据中发现知识的有效途径。

（6）复杂数据类型挖掘的新方法：复杂数据类型挖掘是数据挖掘中一项重要的前沿研究课题。

（7）Web挖掘：有关Web内容挖掘、Web日志挖掘和因特网上的数据挖掘服务，将成为数据挖掘中一个最为重要和繁荣的子领域。

（8）数据挖掘中的隐私保护与信息安全。

数据挖掘在研究领域和商业领域中越来越多的应用，已经得到人们的关注，促使这一技术得到迅速发展和完善。当看到它给人们带来利益的同时，也不能忽视存在的问题，例如：数据挖掘方法的效率还有待提高，尤其是超大规模数据集中数据挖掘的效率，以及挖掘结果的无效性等等。目前应予综合考虑的是：采用数据挖掘解决的商业问题的类型，为进行数据挖掘所作的数据准备，数据挖掘的各种算法和理论基础。

总之，数据挖掘技术是一个年轻且充满希望的研究领域，如何在数据仓库的基础上，加大力度，促使每年都有新的数据挖掘方法和模型问世，仍然是探究的方向。

参考文献

[1] INMON,W.H.Building the data warehouse,third edition. Copytightc2002 by John Wiley & Sons,Inc:21-24.

[2] 彭木根.数据仓库技术与实现[M]. 北京，电子工业出版社，2000:181-206.

[3] 陈华英,李京.庄成三.构建医疗卫生信息数据仓库研究[J]. 四川大学学报(自然科学版),2001,38(4):505-508.

[4] 石丽，李坚.数据仓库与决策支持.北京.国防工业出版社，2003:149-154.

[5] 朱邵文，胡红银，王泉德，等.决策树数据采掘及发展[J].计算机工程，2000，26(10):1-3,35.

(收稿日期：2008-12-25)