计算机科学 1996 Vol. 23 №. 6

数据仓库和数据采掘应用研究

姚卿达 黄晓春 刘向民
(中山大学岭南(大学)学院 广州 510275)

摘 要 随着人们处理的数据量的成指数增长,数据库技术有了新的应用和研究课题——数据仓库和数据采掘。本文介绍了数据仓库和数据采掘的思想和特点,并指出了它们的研究内容和应用前景。

关键字 数据采掘 数据仓库

据统计,19 世纪,人们所需处理的知识量每 50 年翻一倍,20 世纪初,知识量每 30 年翻一倍,20 世纪 50 年代,每 10 年翻一倍,到了 70 年代,知识量每 5 年翻一倍,而当今,不到 3 年就翻一倍,可见信息量增长之快。传统的数据库的检索查询已赶不上社会的需要,许多数据来不及分析就过时了,也有许多数据,因其量大而难以分析数据之间的关联。在这样的背景下,产生了新的数据处理技术-数据采掘(*Data Mining*)和数据仓库(*Data Warehouses*)。

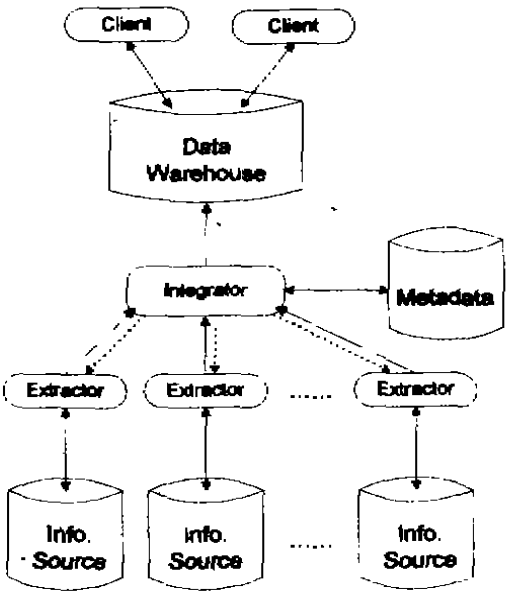
数据仓库和数据采掘是计算机应用领域中的新课题,1995 年 2 月由世界著名的数据库专家 A. Silberschatz, M. Stonebraker 和斯坦福大学的 J. Ullman 等发表了一份权威性报告,题目是《数据库研究:面向 21 世纪的机遇与成就》(美国国家基金 NSF 支持的研究项目),报告谈到今后的研究课题,重点讨论了数据仓库和数据采掘问题,引起了广泛反响。该报告已由中山大学的姚卿达教授和中国人民大学的王珊教授不约而同地编译了出来,发表在《计算机科学》No. 4 1996 pp1-7 上。世界上的大公司(例如 DEC, IBM, ORACLE 等)都声称投入了巨资和重要力量去研究和开发,并有初步的产品。本文将介绍数据仓库和数据采掘的思想和特点,并指出它们的研究内容和应用前景。

一、数据仓库

1. 什么是数据仓库

数据仓库是一个或多个数据库(或数据库的一部分)的数据拷贝,并将来自各个数据库的信息进行集成,供直接查询和分析。因此,数据仓库保存了不同层次粒度的数据——详尽数据到高度汇总的数据,供用户进行数据访问和分析,以便辅助他们进行决策,增强竞争能力。对数据仓库的访问和分析的工具包括采用数据采掘技术,这一技术将在后面详细讨论。

2. 数据仓库的基本体系结构



信息源(Info. Source):可以是异种或异构数据库。这些信息源可以包括非传统的数据,例如:平面文件(flat files)、HTML,知识库等。

提取器(Extractor 或 Wrapper/Monitor):主要进行如下工作:①翻译:把来自信息源的,影响数据仓库信息的数据翻译成数据仓库的数据模式。②监视信息源上数据的变化:当信息源的数据发生变化时,进行数据传播,以便更新和扩充数据仓库;③数据的清洁(Data scrubbing/Data cleansing):保持数据的一致性,减少数据仓库的数据的重复;

集成器(Integrator):初始装载数据仓库,维护数据仓库视图;

元数据(Metadata):关于数据的数据,它包含了关于数据仓库中数据的信息。

数据仓库的特点是：①数据仓库是关系型的；②数据仓库保存的数据通常是历史数据，而且数据仓库保存的数据量十分庞大，可能达到 gigabytes，terabytes，甚至更大。这些数据在数据仓库中很少改动；③对于数据仓库，通常只进行 append 操作；④对信息源中的数据的提取和集成采用批处理的方式进行，通常脱机处理；⑤由于数据仓库是信息源的数据的拷贝，因此，集成器不会访问信息源。

3. 数据仓库方法与联邦数据库

数据仓库和联邦数据库都是为了实现异种数据库互连，能够更容易而有效地访问来自高维异种自治和分布的信息源的信息。数据仓库将异种或异构数据库中的数据按纵向（历史数据）或横向（按主题）分类与集成，具有与数据库不同的模式和存取方法。多数据库联邦也是研究数据集成使用问题，但它是在分布环境下利用数据，数据仓库比联邦数据库的分布处理方法具有更多的优点。以下是数据仓库与联邦数据库的比较：

数据仓库	联邦数据库
对信息源数据的拷贝	不拷贝数据
集中利用信息源数据	分布利用信息源数据
按主题分类（横向）或按长期资料分类（纵向）	不分类
预先提取、翻译和过滤信息（eager/in-advance approach）	接受查询时才处理（lazy/on-demand）
适合静态的、历史的数据的查询	适合动态的变化较快的数据的查询

如上表所述，联邦数据库系统在接受查询时，首先决定解决与查询相关的信息源集，对每一相关信息源产生子查询命令；然后，将从信息源获得的结果翻译、过滤并且合并，以达到最终结果。而数据仓库预先从数据源提取、翻译和过滤信息，并将这些相关信息合并，存于中央数据库（数据仓库）。查询时，只在数据仓库进行，而不必访问信息源。

这两种方法对于数据集成都有各自应用的领域。联邦数据库系统适合多变的、不可预测的客户需求。其缺点是：由于只在提出查询时才从数据源提取信息，导致了查询的低效、延迟，尤其当信息源响应慢时，查询效率更低。相反，数据仓库适用于：用户规范了的可预测的那部分信息；要求有高查询质量，但对信息的最新状态的查询要求不高；信息源的应用环境要求有高性能，不会因为客户的查询而影响其性能，因而需要采用数据仓库的方法，查询时不必涉及信息源。数据仓库的优点是查询效率高，使客户能立即分析信息。其缺点是不适合对变化较迅速的信息的查询。

4. 数据仓库的应用及研究方向

数据仓库主要进行两个问题研究：1）如何从信息源获取信息，并将它们组织，集成到数据仓库中；2）建立了数据仓库后，如何使用它（Warehouse DBMS）。第二个问题是应用方面的问题，主要是使用仓库进行决策支持（DSS），支持多维数据库（multidimensional database），或采用数据采掘技术发现各种数据形式间存在的模式。因此，产生了一些新的研究课题：①查询处理和优化；②数据库的规划；③仓库视图的维护和管理。

而有关建立数据仓库的问题也有许多研究课题，例如：①建立提取器，如何从源提取影响数据仓库的数据，将它们转化为数据仓库模式；②数据清洁的方法，如何解决来自不同信息源的数据的重复和不一致性等问题；③建立和完善元数据字典以向用户提供如何获得数据的信息。

还有许多优化问题，如多视图的优化等，以及数据仓库的并行处理等问题。

二、数据采掘

1. 什么是数据采掘

数据采掘的内容是从大量的数据（VLDB）中采掘（发现）对某种决策有价值的知识和规则，这些规则蕴含了数据库中一组对象之间的特定关系（如同时发生或一方蕴含另一方关系），这些关系可能会揭示一些有用的信息，从而为经营决策提供依据。由于人们要处理的信息量很大，不可能及时发现这许多数据之间的关系，对这些数据往往不能形成精确的查询。数据采掘帮助用户发现数据内部未知的、潜在的关系，以及它们存在的趋向。因此，数据采掘在决策支持、市场策略和金融预测等方面都有广泛的应用。

例如，在市场决策方面，利用数据采掘技术，从超级市场的销售数据仓库中发现：购买面包的顾客众多，而且多数人同时也购买牛油和牛奶。于是，经理把这三种货品摆在同一货架，结果大大提高了这三者的销售量。

2. 从数据库中采掘知识的特点

从以上对数据采掘的概述，可以看到数据库中知识采掘有以下特点：

●数据采掘要处理大量的数据，它所处理的数据库的规模十分庞大，达到 gigabytes，Terabytes，甚至更大；

●由于用户不能形成精确的查询要求，因此要依靠数据采掘技术为用户找寻他可能感兴趣的东西；

●在商业投资等应用中，由于数据变化迅速，可能很快就会过时，因此，要求数据采掘能快速做出响应，提供决策支持信息。

另外，数据采掘就是要发现存在于大规模数据库中的潜在的关联规则，此外还要管理和维护规则，因此，这一技术还有如下特点：

●数据采掘中，规则的发现基于统计规律，因此，所发现的规则不必适用于所有数据，而是当达到某一定"门槛"时，即认为具有此规则。由此，利用数据采掘技术可能会发现大量的规则；

●数据采掘所发现的规则是动态的，它只反映了当前状态的数据库具有的规则，随着不断地向数据库中加入新数据，需要不断地更新规则。

3. 数据采掘研究的问题及方向

数据采掘主要进行如下方面的研究：

●设计高效的算法采掘不同的规则。目前已有许多采掘算法，其中 Apriori[2]和 DHP(Direct Hashing and Pruning)[2]是比较成功的两个算法。早期的研究[1]表明可以将挖掘关联规则的问题分解为两个问题：一是通过统计大量事务寻找达到预定"门槛"值(a threshold minimum support)的较大数据项集；二是利用最小信用值(minimum confidence)从发现的较大数据项集产生所有关联规则。

算法 Apriori 和 DHP 都采用大量的循环，从一维开始计算同一维的大数据项集。每一次循环，首先构造一组候选数据项，然后扫描数据库，计算包含了每一数据项的事务量。算法 DHP 是 Apriori 的扩展，它采用了 Hash 技术。(有关数据采掘算法的描述可参看附录。略)

●设计高效的算法对规则进行更新、维护和管理规则。由于数据库的修改会引入新的规则或否定旧的规则，因此，必须对这些规则进行维护和管理。这方面香港大学的张伟荣教授提出了 FUP(Fast Update)算法[4]，当数据库的数据发生变化时，利用原来所做的工作，对规则进行修改。提高了规则维护的效率。

数据采掘的研究领域已扩展到顺序模式的采掘技术[6]、采掘泛化关联规则[8]、多级别的关联规则的采掘[9]等，除了这些顺序的数据采掘技术研究外，还提出了并行采掘关联规则[5,8,10]以提高采掘效率的算法。

同时，数据采掘还进行如下的研究：●复杂的数据查询优化技术包括数据的汇总、分组等；●支持"多维"查询技术；●高级别的查询语言，以支持非专业用户的查询。

4. 数据采掘与数据仓库

前面介绍了数据仓库和数据采掘，它们都是数据库方面的新应用。数据仓库是用采掘等数据访问和分析技术供用户研究和分析数据仓库，进行决策支持；而数据采掘技术可以以数据仓库为基础，充分利用仓库的数据进行知识发现。

采用数据采掘与数据仓库技术，与传统的数据查询的方法有很大的不同：

●传统的数据查询是被动的查询，用户必须清楚要查询的内容；而数据采掘与数据仓库技术可辅助用户发现隐含的信息；

●传统的数据查询需根据用户的查询对数据进行处理，例如进行复杂的过滤和汇总操作，因此对于经常进行的查询效率不高，且开销大；而数据采掘与数据仓库技术则有较高的查询效率；

●传统的数据查询需要与信息源通信，影响了信息源的处理效率；而数据仓库技术由于是多个数据库的拷贝，对它的访问和分析不会影响信息源的事务处理，并可同时对仓库进行复杂的查询。

但是，传统的数据查询技术也有其优点，对于变化迅速的数据或信息源，或用户的需求不可预测时，采用传统的查询技术就有较大的优势。

三、展望

目前，数据仓库和数据采掘在各个领域已引起了很大的注意。随着信息处理量的增加，这两个技术将会有更广阔的应用前景。在市场分析、决策支持，金融预测以及医学诊断等各个方面，它们都可以有广泛的应用领域。运用这两项技术，将会大大地提高信息处理的效率，更有效地发挥数据库的潜在价值。随着数据库应用系统的普及，这两项技术将会给人类带来不可估量的益处。

致谢 感谢香港大学的张伟荣博士，向我们提供了大量的帮助。

参考文献

[1] R. Agrawal et al., Mining Association Rules between Sets of Items in Large Databases. In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data, 1993

[2] J. S. Park et al., An effective hash-based algorithm for mining association rules. In Proc. 1995 ACM-SIGMOD Int. Conf. Management of Data, San Jose, CA, 1995

[3] Jennifer Widom, Stanford University, Research Issues in Data Warehousing