第18卷 第5期
2008年5月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 18 No. 5
May 2008

数据仓库与数据挖掘的关系及其安全性问题

王 预
(安徽财经大学 信息工程学院 信息管理系,安徽 蚌埠 233041)

摘 要:数据仓库与数据挖掘是当今新的技术热点,数据仓库是一种解决数据使用的高效技术,数据挖掘为之提供了更好的决策支持和服务,同时促进了数据仓库技术的发展。从数据仓库的相关技术和体系结构分析展开研究,分析了数据挖掘与数据仓库的关系,重点阐述了数据仓库的安全性问题,展望了数据仓库未来的研究方向。数据仓库和数据挖掘二者既相互结合、共同发展,又相互影响、相互促进。数据仓库越来越广泛地运用到各个领域中,成为企业获得竞争优势的关键武器。

关键词:数据仓库;体系结构; 数据挖掘; 关系; 数据仓库的安全性
中图分类号:TP311.138 **文献标识码**:A **文章编号**:1673-629X(2008)05-0144-03

Relation of Data Warehouse and Data Mining and Its Safety

WANG Yu
(Department of Information Management, College of Information Engineering, Anhui University of Finance and Economics, Bengbu 233041, China)

Abstract: Data warehouse and data mining technology are new hot spots. Data warehouse is a technology to solve the efficient use of information, data mining and decision support provided a better service, while promoting the development of data warehouse technology. Based on data warehouse technology and architecture of the study, analysis of relations between data warehouse and data mining, data warehouse focus on the security issue and looking forward to the data warehouse for future research. Data warehouse and data scoop out two since combine and develop together mutually, and then influence mutually and promote. The data warehouse makes use of each realm more and more and broadly, becoming the key weapon that the business enterprise acquires competitive advantage.

Key words: data warehouse; architecture; data mining; relation; safety of data warehouse

1 数据仓库的相关技术和体系结构分析

1.1 数据仓库的相关技术

数据仓库就是面向主题的、一致的、不同时间的、稳定的数据集合,用于支持经营管理中的决策支持过程。数据仓库是一个处理过程,该过程从历史的角度组织和存储数据,能集成地进行数据分析。简而言之,数据仓库就是一个大的数据库,存储了公司的所有业务数据。数据仓库允许企业的各个部门共享数据,为企业经营决策提供信息[1]。

要拥有一个高效、优良的数据仓库,必须涉及以下技术:管理大量数据、管理多介质存储、索引和监视数据、多种技术的接口、对数据存放位置的控制、数据的并行存储/管理、元数据管理技术、语言接口、数据的高效装入和高效索引技术、数据压缩技术、复合键码技术、变长数据、锁管理切换技术等。

1.2 数据仓库的体系结构及各自特点

1.2.1 单层体系结构及其特点

特点是:数据是"拷贝"到数据仓库上的,要通过决策支持工具直接访问业务数据。单层次模型有时称为虚拟数据仓库,仅在很少情况下使用。若业务数据是不必调和、存储在一个 RDBMS 中,对 DSS 的请求数量很少,并且应用程序设计能够应付时间需求的时候,它就是个可行的模型。

1.2.2 双层调和体系结构及其特点

特点是:适当的业务数据存放在经过调和、清洁的全局数据仓库中,决策支持工具在这样的仓库里访问数据,全局数据仓库包含细节记录,该模型很少使用。

1.2.3 双层衍生体系结构及其特点

特点是:业务数据经过过滤和概括后直接成为衍生数据,决策支持工具访问衍生的数据。数据集市和专用高速缓存之间的区别在于数据集市一般是关系数

收稿日期:2007-08-28
基金项目:安徽省自然科学基金资助项目(2006KJ052B)
作者简介:王 预(1965-),女,天津人,副教授,硕士生导师,研究方向为数据库与网络。

据集市,而专用高速缓存公用于专用数据分析的数据集市,它是目前最为常见的数据仓库模型,主要是因为便利和对IT的依赖性最小[2]。

1.2.4　三层体系结构及其特点

特点是:业务数据流失要经过调和,放进全局仓库,然后再过滤和概括,存入数据集市或专用高速缓存中。决策支持工具访问的是衍生数据。这种模型的倡导者强调指出,在建立数据仓库以前,数据必须先在全局仓库中经过调和,这样才能避免出现同一个问题得出不同答案的局面。

2　数据挖掘与数据仓库的关系

2.1　数据挖掘与数据仓库的联系

数据挖掘和数据仓库作为决策支持新技术,在近十年来迅速发展。数据仓库和数据挖掘二者既相互结合、共同发展,又相互影响、相互促进。二者的联系概括如下。

2.1.1　数据仓库为数据挖掘提供了更好更广泛的数据源

数据仓库中集成和存储着来自异质的信息源的数据,而这些信息源本身就可能是一个规模庞大的数据库。同时数据仓库存储了大量长时间的历史数据,这可以进行数据长期趋势的分析,为决策者的长期决策行为提供了支持。数据仓库中数据在时间轴上的纵深性是数据挖掘不能回避的又一个新难点。

2.1.2　数据仓库为数据挖掘提供了新的支持平台

数据仓库的发展不仅为数据挖掘开辟了新的空间,更对数据挖掘技术提出了更高的要求。作为数据挖掘对象,数据仓库技术的产生和发展为数据挖掘技术开辟了新的战场,提出了新要求和挑战。数据仓库的体系结构努力保证查询和分析的实时性。数据仓库一般设计成只读方式,数据仓库的更新由专门一套机制保证,数据仓库对查询的强大支持使数据挖掘效率更高[3]。

2.1.3　数据仓库为更好地使用数据挖掘工具提供了方便

数据仓库的建立,充分考虑数据挖掘的要求。用户可以通过数据仓库服务器得到所需的数据,形成开采中间数据库,利用数据挖掘方法进行开采,获得知识。数据仓库为数据挖掘集成了企业内各部门的全面的、综合的数据,数据挖掘要面对的是关系更复杂的企业全局模式的知识发现。而且,数据仓库机制大大降低了数据挖掘的障碍,一般进行数据挖掘要花大量的精力在数据准备阶段:数据仓库中的数据已经被充分收集起来,进行了整理、合并,并且有些还进行了初步的分析处理。这样,数据挖掘的注意力能够更集中于核心处理阶段。另外,数据仓库中对数据不同粒度的集成和综合,更有效地支持广多层次、多种知识的开采[4]。

2.1.4　数据挖掘为数据仓库提供了更好的决策支持

企业领导的决策要求系统能够提供更高层次的决策辅助信息,而基于数据仓库的数据挖掘能更好地满足高层战略决策的要求[5]。数据挖掘对数据仓库中的数据进行模式抽取和发现知识,从数据仓库中揭示出对企业有潜在价值的规律知识,形成知识发现,为知识管理提供了内容,在知识管理中起到中流砥柱的作用。这些正是数据仓库所不能提供的。

2.1.5　数据挖掘对数据仓库的数据组织提出了更高的要求

数据仓库作为数据挖掘的对象,要为数据挖掘提供更多、更好的数据。其数据的设计、组织都要考虑到数据挖掘的要求。

2.1.6　数据挖掘还为数据仓库提供广泛的技术支持

数据挖掘的可视化技术、统计分析技术等都为数据挖掘提供了强有力的技术支持。

总之,数据仓库在纵向和横向都为数据挖掘提供了更广阔的活动空间。数据仓库完成数据的收集、集成、存储、管理等工作,数据挖掘面对的是经初步加工的数据,使得数据挖掘能更专注于知识的发现。又由于数据仓库所具有的新特点,对数据挖掘技术提出了更高的要求。另一方面,数据挖掘为数据仓库提供了更好的决策支持,同时促进了数据仓库技术的发展。可以说,数据挖掘和数据仓库技术要充分发挥潜力,就必须结合起来。

2.2　数据仓库与数据挖掘的区别

数据仓库是一种存储技术,它的数据存储量是一般数据库的100倍,它包含大量的历史数据、当前的详细数据以及综合数据。它能为不同用户的不同决策需要提供所需的数据和信息。数据挖掘是从人工智能机器学习中发展起来的,它研究各种方法和技术,从大量的数据中挖掘出有用的信息和知识。

3　数据仓库的安全性问题

3.1　数据仓库的数据更新

数据仓库中,数据量的增长速度比任何人预计得都要快,数据的质量对数据仓库的性能有很大影响,因此要经常对它们进行更新。对数据仓库中数据的更新不仅是指对原有数据的刷新,还包括定期向数据仓库追加数据。数据仓库中的数据来源于大量的源数据库,要对这些数据进行更新,一种方法是直接读取源数

据库,对那些发生变化的数据加以更新;另一方法是仅当数据变化时才对数据更新。更新数据方法如下。

3.1.1 直接读取

直接读取老的传统的数据库是大多数的组织在考虑刷新数据仓库时最容易想到的一种方法。而且在一定环境下进行某些处理时,刷新只能直接读取旧的传统文件来实现。例如,当数据必须从多个不同的传统数据源收集,从而组成一个整体放入数据仓库中时,直接读取传统数据可能是进行数据仓库刷新的惟一选择;当一个事务处理同时为多个传统文件更新时,直接读取传统数据也是刷新惟一的方法。

3.1.2 捕捉数据

刷新数据仓库的一个更吸引人的方法是在传统环境中捕捉正在被修改的数据。当传统环境中数据发生变化时,通过捕捉它,就不需要当刷新数据仓库时对传统环境中的表全部扫描。另外,因为数据是在改变时被捕捉的,所以不需要传统 DBMS 联机来进行长时间的顺序扫描。相反,可以在脱机时处理捕捉到的数据。在传统操作型环境中,当数据改变时,有下面两种基本技术来捕获这种数据。

a.数据复制:这种方法要求将要捕获的数据在修改之前标识出来。那么,改变发生时数据就能被捕获。

b."变化数据捕获"(CDC):指将发生了的变化从在联机更新时生成的日志或日志磁带中提取出来,是一种对数据仓库环境高效的刷新方法。CDC 使用日志磁带来捕获和确定联机处理时的变化,CDC 需要读取日志或日志磁带。

捕捉变换数据的常用途径有[6]:

(1)时标方法:如果数据含有时标,对新插入或更新的数据记录,在记录中增加更新时的时标,只需根据时标即可判断哪些数据是上次追加后变化的。但许多数据仓库中的数据并不包含有时标。

(2)DELTA 文件:它是由应用生成的,记录了应用所改变的所有内容。利用 DELTA 文件的效率比较高,它避免了扫描整个数据库。

(3)前后映像文件的方法:在上次抽取数据库中的数据到数据仓库之后及本次将抽取数据库中的数据之前,对数据库分别做一次快照,然后比较上次之后本次之前的两幅快照的不同,从而确定实现数据仓库追加的数据。这种方法需占用大量的资源,可能较大地影响系统性能,因此并无多大实际意义。

(4)日志文件:最可取的技术是利用日志文件,因为它是数据库的固有机制,因此不会影响联机事务处理(OLTA)的性能;同时它还有 DELTA 文件的优越性质,提取数据只局限于日志文件,而不用扫描整个数据库。

3.2 数据仓库的管理

一般而言,数据仓库管理包括管理和更新元数据、数据的更新、备份和恢复数据、数据的存档、存储管理、网络管理以及安全管理等。

3.3 数据仓库中的安全问题

3.3.1 数据仓库的安全水平及控制

当企业对越来越多的用户开放其数据仓库中的数据时,提供数据访问和维护数据安全必然成为一对矛盾,尤其是当其数据仓库中有多种不同的数据库平台时,保持适当的安全水平就是一个极为重要的问题。安全水平可以分为三个层次:通过操作系统注册访问系统、应用程序水平的安全以及数据库访问。其中通过操作系统注册访问系统这一层次的安全是保证经过授权的用户能够访问计算机系统,这一般是借助于用户标志符和口令实现,即通常说的用户鉴定或登录,它为确定个人责任提供了依据。建立了用户标志符之后,就可以通过系统安全软件包授予系统特权。在某些环境中,这些系统特权也可以授予数据库管理员,没有数据库责任的管理人员也可以在数据库环境中拥有权力。使用这种方法,安全管理人员可以确定用户责任、口令模式、口令期限。另外,安全系统还可以控制无效尝试口令的次数,在注册时报告自上次成功注册后不成功尝试的次数等。

3.3.2 安全后门问题及其解决

初次安装安全系统时,开发商一般会提供一个默认的用户标志符和口令以进行首次注册。例如,DB2 for OS/2 提供的用户标志符和口令是 USERID 和 PASSWORD,这个标志符拥有在系统内可得到的任何权力:访问所有数据、执行所有管理功能、管理系统安全以及分配新口令等。解决这一问题的方法较简单,须重新建立一个用户标志符,删掉该默认值,或者改变默认标志符的口令[7]。

4 数据仓库未来的研究方向

数据仓库是数据管理技术和市场上一个方兴未艾的领域,有着良好的发展前景,数据仓库技术已经逐渐发展完善,并且随着电子商务的全面展开,数据仓库技术的一个分支:客户关系管理也与网上交易、供应链管理一起构成一个全面的整体电子商务解决方案。数据仓库技术的发展包括数据抽取、数据管理、数据表现和方法论等方面;在市场上看,对于提供数据仓库产品和解决方案的厂商来说,严酷的市场竞争是永恒的主题;从用户的角度看,数据管理在传统领域,如金融、保险、

(下转第 149 页)

(2) 代理签名的不可伪造性:因为 k_i 是每个 p_i 自己秘密选取的,因此除了 p_i 外,任何人都不能伪造 p_i 的有效部分代理签名;

(3) 代理签名的可跟踪性:由于委托过程是将 p_i 的身份 ID_i 与 p_0 绑定在一起的,因此 p_0 可以根据某一有效的代理确定代理签名者的身份,实现对代理签名者的事后监督功能;

(4) 代理签名者权力的限制:委托过程和签名过程中,p_0 和 p_i 以及签名合成者 DC 之间的交互,有效保证了授权证书 C_p 和时间证书 T_p 的安全性,限制了代理签名者 p_i 的权力;

(5) 参与者 p_i 之间的欺诈检测过程:对于参与者 p_i,根据上述方案,如果 $w_i^e \equiv H(\sigma)(\mathrm{mod}m)$,则 p_i 为出示了真正子密钥的合法参与者,否则为内部欺诈者或者外部欺诈者,如果欺诈者 p_i 改 σ 为 σ^*,则他必须计算 w_i^*,使得

$$w_i^{*e} \equiv H(\sigma)(\mathrm{mod}m)$$

才能通过验证,但是他不知道 d,所以他能够计算出 w_i^* 等价于攻破 RSA 公钥密码体制;如果欺诈者 p_i 改 w_i 为 w_i^*,则他必须计算 σ^*,使得

$$H(\sigma^*) \equiv w_i^{*e}(\mathrm{mod}m)$$

在单向函数 H 具有足够的安全性时,这也是难以实现的。因此,本方案能有效阻止参与者之间的欺诈,更加有效地克服了伪造签名攻击。

(6) 抗合谋攻击:假设 $t-1$ 个签名代理人与 p_1,…,p_{t-1} 与 p_j 联合,以让 p_j 冒充 p_i 签名,但是由于离散对数问题的难解性,他们无法求得 k_j 而获得 x_j,因此不能伪造代理签名。假设 p_1,…,p_{t-1} 与外部攻击者 p' 联合进行内外合谋攻击,但由于 p' 并未得到 p_0 的签名权力委托,故而不能通过合成者 DC 的认可,也就无法完成签名。因此,本方案能抵抗合谋攻击。

3　结束语

基于离散对数问题和 RSA 公钥密码体制以及单向函数的安全性,文中提出的新方案除了保持所引文献的优点外,引入了授权文书和时间证书等概念,把代理签名者限制在特定时间和特定次数下代表授权人进行代理签名,有效限制了代理人的签名权力;验证片段的引入,有效阻止了参与者之间的欺诈,克服了合谋攻击和伪造签名攻击。分析表明,新方案满足了数字签名中安全、高效的要求,是一个安全可行的方案,具有较好的理论和应用价值。

参考文献:

[1] Mambo M, Usuda K, Okamoto E. Proxy signatures for delegating signing operation[C]//Proc 3rd ACM Conference on Computer and Communications Security. [s. l.]: ACM Press, 1996:48 - 57.

[2] Mambo M, Usuda K, Okamoto E. Proxy signatures: delegation of the power to sign messages[J]. IEICE Trans Fundam, 1996, E79 - A (9):1338 - 1354.

[3] Zhang K. Threshold proxy signature schemes[C]//Information Security Wority Workshop. Japan: [s. n.], 1997:191 - 197.

[4] Kim S J, Park S J, Won D H. Proxy Signatures[C]//revisited. ICICS' 97, LNCS 1334. [s. l.]: Springer - Verlag, 1997: 223 - 232.

[5] Yang C Y, Tzeng S F, Hwang M S. On the efficiency of nonrepudiable threshold proxy signatures with known signers[J]. The Journal of Systems and Softwares, 2003, 22 (9): 1 - 8.

[6] XU Ying, Liu Huan - ping. An improvement of nonrepudiable (t,n) threshold proxy signature scheme with know signers [J]. Natural Sciences Journal of Harbin Normal University, 2006, 22(2):45 - 47.

[7] Stinson D R. 密码学原理与实践[M]. 第 2 版. 冯登国译. 北京:电子工业出版社, 2003.

(上接第 146 页)

电信等行业有其特定应用,如信用分析、风险分析、欺诈检测等,是数据仓库的主要市场。在未来大规模定制经济环境下,数据仓库将成为企业获得竞争优势的关键武器。数据仓库的发展趋势主要表现在三个方面:对非结构化数据的处理,实现共享数据、对信息进行打包处理[8]。

参考文献:

[1] 史忠植. 知识工程[M]. 北京:清华大学出版社, 1988.

[2] 王　珊, 罗　力. 从数据库到数据仓库[R]. 北京:中国人民大学数据与知识工程研究所, 1997:26 - 28.

[3] 韩客松, 王永成. 文本挖掘数据挖掘和知识管理[J]. 情报学报, 2001(1):45 - 47.

[4] 陈文伟, 邓　苏. 经验数据发现技术[J]. 计算机世界. 1995(8):28 - 30.

[5] 陈文伟. 决策支持系统及其开发[M]. 第 2 版. 北京:清华大学出版社. 2000.

[6] 李德毅. 从数据库中发现知识的策略和方法[J]. 计算机 HI 界报, 1995(3):22 - 24.

[7] 陈文伟, 邓　苏. 可视化机器学习研究][J]. 国防科技大学学报, 1995(3):10 - 12.

[8] 张维群. 数据挖掘研究和应用的现状和前景[J]. 统计与信息论坛, 2004(1):23 - 25.