第16卷 第5期
2006年5月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.5
May 2006

基于数据仓库环境下的数据可用性研究

柳向斌,张志勇,黄 涛
(河南科技大学 电信学院,河南 洛阳 471003)

摘 要:随着数据仓库技术的广泛应用,如何存储数据并保证数据长期可用性已成为近年来的研究重点。文中阐述了在数据仓库环境下存在的数据长期可用性问题,并分析了已有解决方案的优劣,从而给出了一种较为完善的解决方案——模拟策略的核心思想和具体实现过程,以及相关的关键技术,指出了保障数据可用性的未来研究方向。该策略通过模拟历史的硬件平台环境,较好地解决了数据仓库中大量历史数据的长期可用性问题。

关键词:数据仓库; 数据存储 ;数据可用性; 模拟策略
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2006)05-0016-02

Research of Data Usability Based on Data Warehouse

LIU Xiang-bin, ZHANG Zhi-yong, HUANG Tao
(School of Electronic and Info. Eng., Henan Univ. of Sci. and Techn., Luoyang 471003, China)

Abstract: With the application of data warehouse technology abroad, how to save data and assure the readability of data long-term has become an important research direction recently. The paper expounds data usability problems in data warehouse, analyses the merits and shortages of some methods, and introduces an ideal scheme——Emulation Strategy, including its primary thinking, detailed realization process and related key technologies, future research direction of data usability is also presented. The strategy better resolves the question of a mass of history data usability in data warehouse through simulating former hardware platform.

Key words: data warehouse; data preservation; data usability; emulation strategy

0 引 言

在全球信息化进程中,数据库系统领域中的数据仓库技术作为一项前沿技术正在被广泛的应用。鉴于市场竞争日趋激烈,大型企业、公司、服务行业正在从基于MIS/LAN技术上的传统经营管理模式走向建立面向本单位(部门)的DSS系统,为中高层经营管理者提供决策支持。其中构建本单位企业级数据仓库将成为重点,随之而来的是如何存储这些大量的数据信息,以保证它们的长期可用性。

1 数据仓库与存储方式

1.1 数据仓库

"数据仓库之父"W. H. Inmon给出了数据仓库的经典定义:数据仓库是面向主题的、集成的、不可更新的且随时间不断变化的数据集合,用来支持管理人员的决策[1]。数据仓库作为一个数据仓储中心(Data Repository),有别于传统的数据库(如RDB)。首先,前者存储了大量的、历史的数据,用于分析型系统,提供日常性或战略性的决策支持服务;而后者存储的数据类型多为近期或当前的数据信息,主要用于操作型系统,进行日常检索、更新等服务。其次,在存储数据量和时间跨度上也有明显的差别,由于数据仓库中的数据信息的使用目的决定了它的存储时间应是长期的,一般为510年。至于数据量,企业级的数据仓库中的数据量也远远超过了普通的RDB。

1.2 目前的数据存储方案

目前,数据仓库中数据信息的存储技术是根据数据不同的特性,选择不同的存储介质来存放的。常用的存储介质主要有:磁带、磁盘、微缩胶片、CD等等。由于数据使用频度的差异,对于访问频繁的数据,将其存放于磁盘;访问频度低的数据,则存储在磁带或其他介质中;对于大量的历史数据也可以存放于CD中。这种存储方式对于一定时间阶段的数据存储来说是完全可以胜任的,但如果时间跨度变大,将会产生新的问题:能否保证这些数据信息在若干年之后仍然具有可读性。

2 数据存储问题

2.1 数据仓库中的数据信息

数据仓库在DSS系统中用来存储大量的、异质的、多源的数据信息,使其作为决策分析人员的分析对象。如果若干年之后,这些数据失去了可用性,则无法满足系统的需求。因此,保证数据可用性问题将逐渐成为一个大型数

收稿日期:2005-08-08
基金项目:教育部科学技术重点项目资助(03081)
作者简介:柳向斌(1971-),男,北京人,硕士,研究方向为控制理论与控制工程、信息系统。

据仓库环境下的瓶颈问题,这一点也正被业界所关注。

大多数的数据仓库和基于WEB的数据仓库管理员被一些部门,如市场部门,作为战术上的急需所驱动。很少的市场部门关心长于三年的历史数据,因为产品和市场的变化很快,以至于促使他们仅考虑市场客户,而遗弃了那些不再满足他们需要的历史数据。

现在有下列海量的数据信息需要长期保存[2]:

1)详细的销售记录,目的是为了法律、金融和税收。

2)长期跟踪的调查数据,它们具有战略价值。

3)所有政府法规的记录。

4)在某些方面必须保存一百年的医学记录。

5)病人的诊断记录和处理结果。

6)有毒废物的处理,燃料配送,安检记录。

7)所有对某人、在某时具有历史价值的其他数据。

面对上述这些数据信息,必须承认在五年、十年、甚至五十年之后仍需要检索这些数据,而是否能够再利用这些数据将成为一个重要的问题。随着数据仓库技术的深入发展,它和WEB技术相融合之后,将来在数据仓库中所存储的数据也不仅仅局限于文本信息,将会包括更多的图片、图像、声音等多媒体数字化信息。并且伴随着全球信息化进程,数据也正在剧增。这些都将给存储数据问题带来严峻的挑战。

2.2 已有的存储方案

上述这些需要长期存储的海量数据,在若干年后,由于硬件以及软件系统的过时、存储介质的转变、数据格式的差异,都将给读取这些历史数据造成很大的困难。在保证数据可读,而不丢失原来各种属性的过程中,最初建立历史数据的软件系统将起到决定性的作用。

在这个问题上,业界已有多种技术方案,这里将阐述两种方案。

(1)迁移技术(Migrate Technology)。它是指在一定的时间周期内,在某种介质未被淘汰之前,将所存储在其中的数据信息尽可能不丢失地迁移到新介质中。

(2)制定标准。为了保证数据长期可用,不再受到软硬件环境的限制,制定统一的标准来存储数据信息。这两项技术都不可能最终解决上述所提出的问题,但在问题的局部处理中具有重要的理论价值。

3 模拟策略

3.1 模拟策略的思想及所需技术

模拟策略的核心思想是通过在未来的,不可预测的系统上模拟过时的硬件平台,使得最初的软件能够在此平台上运行,从而完成读取任务[3,4]。

在模拟策略的实现过程中需要下列几个方面的技术:

(1)关于模拟机(Emulator)。模拟机用来在将来的计算机系统上模拟原始的硬件平台。因此,模拟机规格详细说明及解释程序将是最终模拟成功的必要和关键技术。

(2)封装技术(Encapsulate)。在模拟策略中,为保证数字信息将来可读,需要在原始数据信息中封装其他的数据信息。

(3)存储技术。只有在保证元数据的如实存储,才能确保模拟策略最终实现,所存储的信息将来找到原数据文件。

3.2 数据封装

关于数据封装,主要包括以下三种数据。

(1)需要读出的数据信息本身和最初的应用软件以及原OS软件。作为封装数据的核心,数据信息用二进制位流的形式存放于一个或多个文件中,这里不以编码的形式存放数字化信息。为了保证将来可用性,原应用软件也以可执行的位流存储。

(2)模拟机的规格说明书。这些信息包括必要的足够的数据使得模拟机能在将来的未知的机器上被创建。它本身并不是可执行程序,但必须能够解释原硬件平台的各种属性,这样原应用软件可以在模拟机上运行。

(3)元数据、注释信息等等各种附加的信息。这些附加信息用于为用户解释如何使用封装文件。这部分信息必须保证将来的可用性,它关系到封装文件的打开和使用。

3.3 保存信息的具体实现过程

具体实现过程主要分为以下三个步骤:a.创建注释信息;b.封装;c.模拟。注释信息用于提供数据文件的软硬件环境的上下文,即它的应用程序和OS软件。封装是上文所谈到的三种信息。模拟是在将来需要读取原文件信息时,打开封装,创建具体的模拟机,模拟原硬件平台,使其在未来机器上运行,最终原应用软件在模拟状态下运行,用于读取所需的数字化信息。

整个的读取数字化信息的过程如下所述:

在具体的读取过程上,分为三个分支来解释说明。第一个分支是模拟机,第二是现有系统(硬件和软件)分支,最后一个分支是现有存储介质。

关于第一分支,首先通过模拟机规格说明书解释程序对模拟机说明书本身解释执行生成原始的硬件平台模拟机。第二分支是将现有的硬件和OS软件相结合,使得OS运行于现有的硬件平台。这样,原始的模拟机便可以在现有的运行了OS软件的机器上运行原始的OS软件,然后再和最初的应用软件相结合。对于第三分支,需要读取的数据信息必须已经存储在现有的存储介质上,其中的实现可以通过迁移技术。然后现有的存储介质和驱动器结合使得介质在物理上可读取,在现有的OS软件上再通过驱动程序的驱动,来保证介质在逻辑上可读取。最后,需要读取的数据信息物理和逻辑上已经可读,加之最初用来写入数据信息的应用软件也已在现有的软硬件环境下得到了运行,这样原数据信息便可以被读取出来了。

3.4 关于模拟策略的补充说明

模拟策略是对硬件平台的模拟,为何选择模拟硬件,

(下转第20页)

④保留最优分类器的结构,利用整个的数据集进行参数学习(条件概率表)。

⑤输出这个新的分类器。

数据集规模不大的情况下,选择用 LOO 验证法。LOO 充分利用训练样本,是最严格、最精确的一种评估方法之一,但是它适用于小规模数据集,否则增加机器运行时间。数据集规模很大的情况下,选择保留(Holdout)方法来测试分类器的性能。保留方法中一般采用数据集的 2/3 作为训练集,数据集的 1/3 作为测试集。

利用保留方法测试小数据集的分类器性能会降低分类器的分类精度;而利用 LOO 方法测试大数据集的分类器,会增大机器的运行时间。因此不同的数据规模下选择合理的检验方法既保证分类器分类精度,又可以减少机器运行时间。

4 实验结果

标准数据集是从 UCI[5] 上下载的,选取了 8 个数据集。有关数据集的概况见表 1。所选择的数据集都是小规模的,所以本实验只针对评估方法是 LOO 情况进行。在实验中不同的数据集是在相同的环境下运行的。

表 1 算法的实验结果

Dataset	Attributes	Classes	Instances	HCS	SP	HCS&SP
Vehicle	18	4	846	70.1	70.3	70.3
Post-op	9	3	90	72.7	72.1	72.7
Australia	14	2	690	84.7	85.3	85.3
Hepatitis	19	2	270	84.8	84.3	84.8
Vote	16	2	435	95.6	95.7	95.7
Heart	13	2	270	78.7	76.1	78.7
Soybean-Large	35	19	562	88.6	88.4	88.6
Pima	8	2	768	78.0	78.2	78.2

表 1 列出了实验结果,实验结果基本和文献[4]中的数据差不多,第七列是改进算法的实验结果,每次输出的分类器确实是 HCS 和 SP 中性能最好的,验证了所建立的分类器的有效性和正确性。同时也发现当属性变量之间的依赖关系不是很复杂的情况下,分类器的分类精度会很高。例如数据集 Vote,说明文中讨论的分类器在对 Vote 操作时几乎把影响分类准确率的弧都加上了,丢失有价值的弧的个数相应减少。当属性间的关系很复杂时,由于分类器属性之间构成的是有向森林,有可能丢失一些对分类有价值的弧,如 Vehicle。再一次表明了属性结点之间的条件依赖关系不是很复杂的情况下,HCS 和 SP 算法的分类准确性非常高,改进的算法的分类准确性更高。

5 结束语

通过对 HCS 算法和 SP 算法的分析,又提出了一个改进的算法,明显提高了分类器的分类精度。扩展朴素贝叶斯分类器既有朴素贝叶斯分类器的简单性,又有很好的分类能力,应努力把它应用到更多的现实领域中。当属性结点之间的相关性很复杂的情况,即有简单性又有很好的分类性能的分类器待于进一步的研究。

参考文献:

[1] 林士敏,田凤占,陆玉昌.用于数据采掘的贝叶斯分类器研究[J].计算机科学,2000,27(10):73-76.

[2] Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers[J]. Machine Learning,1997(29):131-163.

[3] Cheng J, Greiner R. Comparing Bayesian network classifiers [A]. Proceedings of the fifteenth conference on uncertainty in artificial intelligence[C]. San Francisco: Morgan Kaufmann, 1999.101-108.

[4] Eamonn J, Pazzani J. Learning augmented bayesian classifiers: a comparison of distribution-based and classification-based approaches[A]. Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics[C]. Lauderdale:[s.n.],1999.225-230.

[5] Blake C, Keogh E, Merz C. UCI repository of machine learning database[EB/OL]. http://www.ics.uci.edu/mlearn/ML Repository.html,1998.

(上接第 27 页)

就可以确定输出层的神经元数目为 4。但是通过其后的实践发现,当采用(0,0,0,0)这样的目标输出向量时,BP 网络无法收敛,那是因为采用的激发函数永远不可能达到 0 或 1,而只能是接近。所以还要对其重新编码,最后确定其编码为:0(0.1,0.1,0.1,0.1),1 (0.1,0.1,0.1,0.9),…,9 (0.9,0.1,0.1,0.9)。

3 实验结果

本系统采用了 100 个训练样本用来测试算法的性能,这 100 幅图像包括 Arial, Batang, Gautami 字体,单个数字的平均识别率达到 94%,网络训练时间大约 3~6 秒,当然如果采用更多的训练样本,则识别率将会进一步提高。

参考文献:

[1] 张宏林,蔡 锐. Visual C++数字图像识别技术及工程实践[M].北京:人民邮电出版社,2003.422-442.

[2] 冈萨雷斯.数字图像处理(第 2 版)[M].北京:科学出版社,2003.60-127.

[3] Hagan M T. Neural Network Design[M].北京:机械工业出版社,2002.16-64.

[4] Bruck J, Sanz J. A Study on Neural Networks[J]. Int J Intelligent System,1988(3):4-15.

[5] 徐丽娜.神经网络控制[M].哈尔滨:哈尔滨工业大学出版社,1999.7-16.

[6] 袁曾任.人工神经网络及其应用[M].北京:清华大学出版社,1999.16-29.