文章编号:1000 - 582X(2004)12 - 0116 - 04

面向商业 OLAP 的并行数据抽取接口设计*

冯 永,吴开贵,熊忠阳,吴中福
(重庆大学 计算机学院,重庆 400030)

摘 要:商业竞争日趋激烈的今天,单纯的联机事务处理系统已经不能满足管理者的决策支持要求,联机分析处理(OLAP)技术的出现具有重要意义。研究了目前联机分析处理的关键技术,数据仓库的经典解决方案,数据预处理的相关方法,提出了一种面向商业 OLAP 的并行数据抽取接口设计方案,并对设计过程中的数据清理、数据集成和变换、数据归约等数据预处理技术作了重点介绍和应用研究。最后结合实际应用阐明了提出的数据抽取方案对于实现商业 OLAP 功能的有效性和实用性。

关键词:OLAP;数据抽取;数据清理;数据集成和变换;数据归约
中图分类号:TP391 **文献标识码**:A

目前大多数企业都先后拥有了自己的信息管理系统,基本满足了企业的联机事物处理要求。近年来,伴随商业竞争的日趋激烈和信息产业的快速发展,单纯的联机事物处理已经很难满足企业的发展需要,企业决策者迫切需要的是具有联机分析处理(OLAP, On-line Analytical Processing)功能的决策支持系统。OLAP 专门用于支持复杂的分析操作,具有汇总、合并和聚集功能,以及从不同角度观测信息的能力,侧重于对高层管理人员的决策支持[1]。

在信息化社会中,企业为了自身的发展和运作,收集了大量的信息和数据,用于分析自身的业务运行状态。而决策所需要的是综合性的、规范的、关键的信息和数据,很少使用或者不使用大量的细节信息和数据。据统计,源数据库存放的细节数据与决策时需要的有效数据之比非常悬殊,大约是 1 000 : 1[2],因此决策者们希望设计一种数据抽取系统,建立专为决策者使用的专业数据库,存放有效数据,从而改进决策者对数据的利用率。

针对上述应用要求,提出了一种面向商业 OLAP 的并行数据抽取接口设计方案。首先对企业的各种数据源进行数据清理,然后整合各种异构数据源为统一的数据存储模式,最后为企业决策者建立了用于 OLAP 的专业数据库。整个数据抽取系统基于企业服务器组成的并行网络,采用 Master - Slave 计算模式,因而具有良好的可伸缩性和可扩展性。

1 面向商业 OLAP 的系统设计

一般的大型商务公司下属若干家分公司。总公司拥有数台数据库服务器,通过局域网互连,各分公司分别拥有自己的客户机,这些客户机通过专用网络与总公司的局域网连接,总公司进行 OLAP 的数据应来自各分公司提供的数据。

但是,分公司存放的数据可能是大量细节性数据和异构数据源数据(如文本等)[3],这就使得数据的存储模式不统一,不符合 OLAP 的要求。因此首要问题就是需要设计一个数据抽取接口,首先对各分公司的数据进行清理工作,然后再把清理后的分公司数据进行集成和变换,整合成为与总公司数据库统一的数据存储模式,最终形成总公司一级进行 OLAP 的专业数据库。

基于上述分析,提出了整个系统的结构设计,如图 1 所示。系统共有 3 层,分公司层是各分公司的数据库和一些异构数据源,其中存放了大量的细节数据和不规范数据;数据抽取层包括数据抽取接口和感应器,数

* 收稿日期:2004 - 09 - 02
基金项目:重庆市科技计划项目应用基础项目(7968)
作者简介:冯永(1977 -),男,山东平度人,重庆大学博士研究生,主要从事数据挖掘及并行处理研究。

据抽取接口的主要作用是对分公司级的数据进行预处理，形成规范的总公司级的统一数据存储模式。感应器设定了一个数据量阀值，当各分公司的变动数据量超过这个阀值时，数据抽取接口开始工作；总公司层包括专门为决策分析而设计的专业数据库和决策支持查询应用。

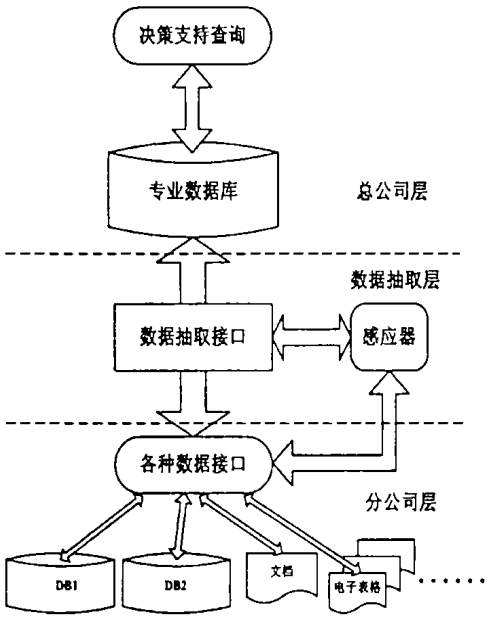

图 1　整个系统的结构设计图

2　数据抽取接口设计中的数据预处理

在整个接口设计中，数据预处理是一个非常重要的部分。现实世界中的应用数据存在着许多不一致数据、不完整数据和噪声数据，产生这种现象的原因是多方面的，主要有滥用缩写词、数据输入错误、数据中的内嵌控制信息、不同的惯用语、重复记录、丢失值、拼写变化、不同的计量单位、过时的编码等 9 个方面，因此预处理是必须的。数据预处理包括数据清理、数据集成和变换、数据规约。

2.1　数据清理

数据清理主要包括 2 个方面的工作：填补丢失的数据、清除噪声数据。

当分析数据时，可能会发现许多元组的一些属性没有记录值，这将对数据分析的结果产生不利影响。在抽取接口的设计中采用了填入同一类元组在该属性上的平均值的方法来填入空缺值。此方法通过分类属性，计算每种属性的平均值，将平均值填入对应属性的空缺值中，使分析结果更加逼近真实情况。例如，公司的客户编码从 1 到 k，$v(i)$ 表示客户 i 的销售额，其中 $1 \leqslant i \leqslant k$，客户销售额的平均值记为 v_{ave}。若客户 2、16、30 的销售额为空，则应该把 vave 填入 $v(2)$、$v(16)$、$v(30)$。

噪声指没有包含在任何分组当中的数据对象，应该被丢弃，否则会影响分析结果的准确性。噪声数据可以通过聚类分析检测到。其主要思想就是将一组数据按照某种相似性划分为若干组，遗留在所有小组之外的零散数据将被作为一种噪声数据而剔除。在抽取接口的设计中采用了 V. Barnett 和 T. Lewis[4] 提出的基于统计的孤立点检测来发现并去除噪声数据。该方法要求数据集的分布参数（如平均值和方差），但计算复杂度为线性，适合商业型数据。

2.2　数据集成和变换

数据集成将多个数据源中的数据结合起来存放在一个一致的数据存储中。数据源可能包括多个数据库、文本文档、电子表格。

对于大多数商务公司，可能会存在多个不同的数据库，其它的数据源一般都是文本文档和电子表格。对于不同的数据库之间可以利用数据库中的元数据（关于数据的数据），使多个数据库中的实体匹配。对于文本文档和电子表格，可以通过编制数据管道将其导入数据库。

数据集成剩下的问题就是冗余问题了。这里对于冗余数据采用相关分析的方法来检测数据之间的相关性，然后把冗余的数据删除。对于给定的两个属性，相关分析可以度量一个属性在多大程度上蕴涵另一个属性。属性 A 和属性 B 之间的相关性可以用下面的公式度量：

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A \sigma_B}$$

其中，n 是元组个数，$\bar{A}$、$\bar{B}$ 分别是 A、B 的平均值，σ_A 和 σ_B 分别是 A 和 B 的标准差。如果公式值大于 0，则 A 和 B 是正相关的。该值越大，表明 A 或 B 偏离均值越多，即两个属性的相关性就越大。因此，一个很大的值表明 A（或 B）可以作为冗余数据被去掉。

进行数据变换的主要目的是使数据规范化，转换成适合 OLAP 的形式。因为 OLAP 用到的数据基本是数值型的，而有的数据值非常大，不利于决策分析，把这些数据投影到一个相对较小的区间，在相对较小的数值区间上分析会比较容易。

主要的方法有：Z - Score 规范化、小数定标规范化、最小 - 最大规范化。

Z - Score 规范化基于属性值的平均值的标准差来进行规范化，当存在噪声数据时，该方法比较有效；小数定标规范化通过移动属性的小数点位置进行规范化，该方法易破坏原始数据之间的关系；最小 - 最大规

范化对原始数据进行线性变换。假定 v_{max} 和 v_{min} 分别为属性 A 的最大和最小值,属性 A 的值 v 被规范成 v',由下面的公式计算:

$$v' = \frac{v - v_{min}}{v_{max} - v_{min}}(v_{n\_max} - v_{n-min}) + v_{n\_min}$$

其中$[v_{n_min}, v_{n_max}]$是新的值区间。这种方法保持了原始数据值之间的关系,由于在数据清理阶段已经处理了噪声数据,因此实际设计时采用了这种方法。

2.3 数据规约

数据归约技术用来降低分析数据量。通过数据规约后,得到的数据集要小得多,而且分析结果也几乎相同。由于针对的是大量的商业数值型数据,因此设计采用了数据立方体聚集[5]的方法来进行数据规约。这种方法可以对数据按不同维度进行汇总计算,形成多个高层次的数据立方体,从而减低了数据量,提高了 OLAP 的效率。

3 数据抽取接口的具体实现

3.1 并行网络结构

总公司下属有若干家分公司,需要抽取的数据量非常庞大,因此在实际设计时,采用了并行网络结构。如图 2 所示,并行网络结构采用主从模式,这类模型由一个主进程(master)和若干个从进程(slave)组成。主进程运行于一台主服务器上,负责生成从进程、初始化网络、计算任务分配、收集结果。而从进程运行于各从服务器上,执行相同程序,处理不同数据(处理的数据为整个数据集的子集,且互不交叉)。各从进程之间互不通信,以减少通信开销。

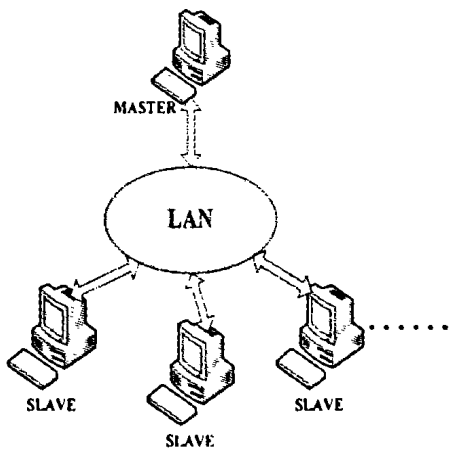

图 2 并行网络结构

3.2 具体实现步骤

Master 过程:

1) Detect(UpdateLog),监视器探测各分公司的数据变动日志;

2) Sum(UpdateNumber),累加各分公司的变动数据量,记为 SUN;

3) 如果 SUN 大于设定的阀值 M,则进行数据抽取,否则返回;

4) Select(UpdateData),当 SUN > M 时,从各分公司提取变动数据;

5) Separate(Slave, dataset),按照属性将变动数据集分配到各个 Slave;

6) Gather(n, e,σ,Slave),接收各 Slave 发送的 n、e、σ,并进行汇总;

7) Compute(n, e,σ),根据上步汇总得到的参数,进行属性间的相关度计算,属性间的相关度记为 r;

8) Broadcast(Slave, r),向所有 Slave 广播 r;

9) Congregate(sta_subdataset),接收各 Slave 发送的规范数据子集,进行合并,并进行数据立方体聚集,形成适合 OLAP 的统一数据存储模式。

Slave 过程:

1) Save(subdataset, tempbuffer),接收 Master 分配的数据子集,并存储到临时数据工作区 tempbuffer 中;

2) DataCleaning(subdataset),对数据子集进行数据清理;

3) Integration(subdataset),对清理后的数据子集进行数据集成,并计算属性的元组数目 n、元组均值 e、元组标准差 σ;

4) Send(n, e,σ,Master),把属性的 n, e,σ 发送给 Master;

5) Delete(redundance, r),根据接收到的属性间的相关度,删除冗余属性;

6) Standardization(subdataset),对上步操作得到的数据子集进行数据变换,得到规范的数据子集,记为 sta_subdataset;

7) Send(sta_subdataset,Master),把 sta_subdataset 发送给 Master。

3.3 接口的特点

1) 采用增量式更新的方法,减少了资源的消耗,使效率明显改善;

2) 数据清理的两步工作,以及数据集成都是基于数据属性进行的,因此设计中按照属性分配数据集到各 Slave,进行并行处理时就不会破坏数据的完整性;

3)进程间的通信都是 Master 与 Slave 间的通信,Slave 间互不通信,而且所有的通信都是异步通信,这大大降低了通信开销。

4　数据抽取接口的实际应用

药品销售预测是对影响药品销售的各种因素进行综合分析，及时预知药品的销售趋势和走向，对于指导进货、促销策划、地域分布分析和客户分析等决策提供必要的支持，是一个有实际价值和广阔前景的 OLAP 应用。因此，应用并行数据抽取接口到医药公司，进行药品销售预测，具有实际应用价值。

医药公司下属 11 家分公司，各分公司长期进行销售业务，累积了大量数据。数据来源有异构数据库数据、文本文档、电子表格。总公司拥有 6 台 HP6000 服务器，现将 6 台服务器互连，利用可移植异构编程环境 PVM[6] 组成并行计算网络。表 1 列出了并行数据抽取接口各部分的执行时间，总的执行时间，各台服务器的 CPU 平均占用率，以及采用串行结构抽取所需花费的总时间，CPU 占用率，最后计算了加速比。

表 1　并行数据抽取接口的执行效率

变动数据量阀值/万个数据元组	划分数据集时间/s	数据清理时间/s	计算集成时间/s	计算相关度时间/s	广播通信时间/s	删除冗余时间/s
100	21	43	34	11	6	10
200	40	79	66	19	11	18
300	59	121	100	27	17	25
400	78	167	132	41	22	39
500	101	211	164	52	29	47

数据变换时间/s	立方体计算时间/s	并行抽取总时间/s	并行 CPU 占用率/%	串行执行总时间/s	串行 CPU 占用率/%	加速比
14	9	148	11	607	45	4.1
27	16	276	19	1 187	71	4.3
40	23	412	27	1 936	96	4.7
53	34	566	36	2 887	100	5.1
66	42	712	51	3 916	100	5.5

从表 1 的实验结果可以看出并行数据抽取接口在进行大规模数据处理时，运行时间分配合理，工作效率高，各台服务器除了完成抽取，还可以兼顾其它计算任务，这有利于大型商务公司频繁的联机事务处理。当抽取数据量增加时，抽取接口的加速比成上升趋势，表明抽取接口具有良好的可伸缩性。若采用串行结构，当数据量上升到 400 万个元组以上，CPU 占用效率达到 100%，完成抽取任务十分困难。

5　结　论

1）抽取接口的设计基于并行计算网络，整个系统的性价比高，可扩展性好，适合目前的商务公司进行 OLAP；

2）并行执行过程中采用异步通信模式，降低了通信开销；

3）数据抽取是增量式进行的，减少了数据量，从而降低了计算复杂度；

4）主进程分配数据集到子进程时，按照属性分配，保证并行计算结束后数据对象的完整性，不会破坏最终结果的准确性；

5）实际应用表明抽取接口的工作效率高，可伸缩性好，适合商业 OLAP 的要求。

参考文献：

[1] 王珊. 数据仓库技术与联机分析处理[M]. 北京：科学出版社，1998.

[2] 王亚芬. 智能数据抽取技术在决策支持系统中的应用研究[J]. 情报学报，1996，15(2)：89－94.

[3] HAN J, KAMBER M. Data Mining: Concepts and Techniques [M]. Beijing: High Education Press, Morgan Kaufman Publishers, 2001.

[4] BARNETT V, LEWIS T. Outliers in Statistical Data[M]. New York: John Wiley & Sons, 1994.

[5] SARAWAGI S, STONEBRAKER M. Efficient Organization of Large Multidimensional Arrays[A]. In Proc. 1994 Int. Conf. Data Engineering(ICDE94)[C]. Houston, TX USA: [s.n.], 1994. 328－336.

[6] MIGLIARDI M, SUNDERAM V. The HARNESS PVM-Proxy: gluing PVM Applications to Distributed Object Environments and Applications[A]. (HCW 2000) Proceedings 9th[C]. Cancun Mexico: Heterogeneous Computing Workshop, 2000. 309－322.

（下转第 123 页）

in Convex Metric Spaces[J]. J Appl Math Mech, 2002, (9):1 001 - 1 008.

[10] TAKAHASHI. A Convexity in Metric Space and Nonexpansive Mappings[J]. J I Kodai Math Sem Kep, 1970,22: 142 - 149.

[11] ZHANG S S. Convergence Problem of Ishikawa Type Iterative Sequence With Errors for ϕ-quasi-contractive Mappings [J]. Applied Mathematics and Mechanics, 2000, 21 (11):1 - 10.

Convergence of Ishikawa Type Iterative Sequence With Errors for Asymptotically Quasi-nonexpansive Mappings in Convex Metric Spases

TIAN You-xian

(Institute of Computer Science & Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Liu Qihou, in 2001 and in 2002, extended the results of Petryshgh and Williamson, Ghosh and Debnath respectively in 1973 and in 1977, proved some sufficient and necessary condition for Ishikawa iterative sequence and for Ishikawa iterative sequence with error member of asymptotically quasi-nonenpansive mappings to converge to fixed poind in Banach space and in uniform convex Banach space. In convex metric spaces, the Ishikawa iteration process with errors is defined for asymptotically quasi-nonexpansive mappings. Some sufficient and nevessary conditions proved for the iterative schene converges to the fixed point of the asymptotically quasi-nonexpansire mappings. These results qeneralize and unify many important known results in recent literature.

Key words: convex metric space; asymptotically nonexpansive mappings; asymptotically quasi-nonexpansive mapings; ishikawa iteration process with errors; fixed point

(编辑　吕赛英)

(上接第 119 页)

Frame of Parallel Data Extraction Interface for Commercial OLAP

FENG Yong, WU Kai-gui, XIONG Zhong-yang, WU Zhong-fu

(College of Computer Science, Chongqing University, Chongqing 400030, China)

Abstract: The OLTP system can not afford the demand of decision support due to the drastic commerce competition, so the appearance of OLAP technology is importanct. The key technical of OLAP, the classical resolvent of data warehouse and the methods of data preprocessing are studied. A frame of parallel data extraction interface for commercial OLAP is proposed. Data preprocessing technology (data cleaning, data integration and transform, data reduction) connected with the frame is practically studied and applied. The eficiency and the practicality of the frame are illustrated by practical application.

Key words: OLAP; data extraction; data cleaning; data integration and transform; data reduction

(编辑　张　苹)