# 商务智能在现代企业中的应用

汪林林[1] 焦慧敏[2]

(重庆邮电学院软件学院 重庆 400065)[1] (重庆邮电学院计算机学院 重庆 400065)[2]

**摘 要** 针对目前现代企业信息系统存在的问题,本文引入商务智能架构体系,提出了一种改进的商务智能体系结构,给出了一个完整的商务智能系统建设方案,并对其中的数据仓库的主题分析的选择、ETCL过程实现以及数据仓库的查询优化进行了研究和实现。实际应用中的结果表明,提高了企业对现有的信息数据的利用效率,提高了企业决策分析的能力。

**关键词** 商务智能,数据仓库,数据集市,OLAP,ETCL,数据挖掘

## Application of Bussiness Intelligence in Modern Enterprise

WANG Lin-Lin[1] JIAO Hui-Min[2]

(Computer College, Chongqing University of Posts and Telecommunications, Chongqing 400065)[1]

(Software College, Chongqing University of Posts and Telecommunications, Chongqing 400065)[2]

**Abstract** Aiming at the problems of modern enterprise information system, this paper introduces business intelligence architecture, proposes one kind of improvement system structure of business intelligence, puts forward a comprehensive method of business intelligence system construction, researches and realizes the essential technology of the business intelligence system: subject analysis choice, ETCL's process realization and data warehouse query optimization. The result in the practical application indicates that the using efficiency of the existed information data has been heightened, and the ability of the enterprise decision analysis has been improved.

**Keywords** Business intelligence, Data warehouse, Data mart, OLAP(Online analytical processing), ETCL(extract, transform, cleaning, load), Data mining

## 1 引言

随着信息化的发展,现代企业信息化进程得到巨大发展和广泛应用。各种应用系统的广泛应用以及互联网的蓬勃发展,为计算机应用系统的运行积累了大量的历史数据。但在很多情况下,这些海量数据在原有的作业系统中是无法提炼并升华为有用的信息,提供给业务分析人员与管理决策者的。一方面,联机作业系统因为需要保留足够的详细数据以备查询而变得笨重不堪,系统资源的投资跟不上业务扩展的需求;另一方面,管理者和决策者只能根据固定的、定时的报表系统获得有限的经营与业务信息,无法适应激烈的市场竞争。现在,大多数企业并不缺少数据,而是苦恼于海量数据以及数据的不一致性。随着数据量的增加,数据变得越来越难以访问、管理。如何把已有的海量数据转换成更有价值的商用信息,以便用于决策支持呢?商务智能被广泛认为是最好的解决方案之一。

商务智能系统建设的目标就是要为企业提供一个统一的分析平台,充分利用原有系统中积累的宝贵数据,对其进行深层次的发掘,并从不同的角度分析企业的各种业务指标和构建起业务知识模型。本文在讨论了商务智能的基本概念后,以大连商品交易所为例,具体分析了商品交易所计算机监控系统的系统结构及建设方案,并对其中的关键技术进行了详细描述。

## 2 商务智能体系结构及系统实现

商务智能系统是建立在数据仓库、OLAP、数据挖掘等技术的基础之上,通过收集、整理和分析企业内外部的各种数据,加深企业对客户及市场的了解,并运用一定的工具对企业运营状况、客户需求、市场动态等做出合理的评价及预测,为企业管理层提供科学的决策依据。商务智能体系结构一般为:源数据层、数据转换层、数据仓库(数据集市)层、OLAP及数据挖掘层、用户展现层。

商务智能的核心采用了数据仓库技术、OLAP技术及数据挖掘技术。数据仓库是商务智能系统的基础,是面向主题的、集成的、稳定的和随时间不断变化的数据集合。联机在线分析处理(OLAP)技术的核心是"维",通过对多维数据的钻取、切片及旋转等分析动作,来完成决策支持和多维环境下的查询及报表。基于底层数据存储的不同,可以将OLAP分为MOLAP和ROLAP两种。当进行OLAP查询时,需将用户的多维分析动作解释成相应的结构化查询语言(SQL)语句来执行操作并返回查询结果。数据挖掘是指运用人工智能、机器学习、统计学等技术,对企业中的数据进行分析推理,找出隐含或未知的模式,提供给管理人员,提高其决策水平。基于OLAP的决策支持,是更多地依靠系统与专业技术人员的交互来完成对历史数据的统计和分析的,而数据挖掘则自动化程度较高,主要用于发现隐含在数据中的有用信息。

**汪林林** 教授,硕士生导师,副博导,主要研究领域:数据库与数据挖掘、空间数据库与GIS、计算机网络、计算金融、电信增值业务、分布式计算;
**焦慧敏** 硕士研究生,研究方向:数据库、数据仓库、系统体系结构设计。

### 2.1 现有系统存在的问题

20世纪90年代以来，计算机监控系统得到了很大的发展和应用。随着网络管理系统、数据库管理系统、主机管理系统、杀毒软件以及防火墙等系统的相继投入使用，各管理系统积累了大量的历史数据。这些系统的建设为商品交易所的期货交易系统提供了有力的安全保障。但它们是在不同历史时期，根据不同业务需要，由不同供应商提供的，体系结构和管理实施等方面存在着较大的差异，各系统间的数据也相对分散和独立，难以共享，没有建立起统一的、用于分析处理的基础数据平台。

要把这些来自于不同的系统中的海量数据提炼并升华为有用的信息，及时提供给业务分析人员与管理决策者，使计算机系统的运行更加安全和稳定，并能在指定的条件下及时预警。而目前管理者和决策者只能根据固定的、定时的报表系统获得有限的信息，无法适应越来越庞大的计算机系统的稳定性、健壮性和安全性要求。

本系统收集分散的各种详细数据源，建立以各种主题为导向的数据仓库，并从中分析监控数据的内在规律，方便企业高层行政主管、业务主管对主机状况、网络状况、数据库状况、环境状况以及部分应用状况进行综合分析，及时掌握系统的各项运行指标等诸多有效信息，使管理人员的工作由事务型向思维型转变由事后处理型向事前预测型转化，为管理者做出科学的计划、判断和决策提供帮助。

### 2.2 系统设计

本系统采用B/S架构。服务器端提供的服务包括数据逻辑、数据服务、性能监控以及元数据存储、数据挖掘、数据聚集等。客户端执行的功能包括用户界面、查询分析、报表格式化以及数据访问等图1。

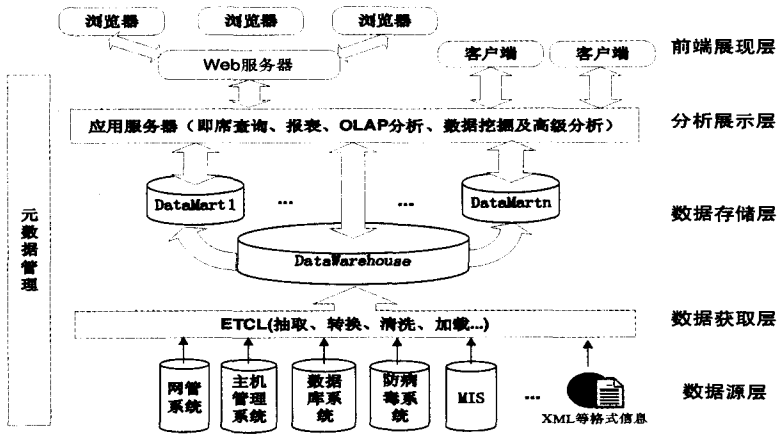


图1 系统结构图

1)数据源层：也可称作操作型数据层，是整个数据仓库的基础，提供了整个系统最原始的数据。通常为业务数据库和其他外部数据，业务数据包括：网管管理系统数据、主机相关运行指标的数据、数据库监控信息的数据、杀毒软件的数据、防火墙系统的数据、入侵检测系统的数据、基本信息维护的数据以及外部其他各式的数据。

2)数据获取层：也可称作数据转换层，主要是把数据源层的数据通过ETCL过程转换到数据仓库中，包括数据的抽取(Extract)、转换(Transform)、清洗(Cleaning)和加载(Load)4个部分，这一层在一定程度上决定着数据仓库中数据的质量。ETCL指的是数据抽取、转换、清洗、加载的过程，其目的是从数据源中抽取该系统所需要的数据，对其进行检验和清洗，并根据数据仓库设计要求对数据进行重新组织和加工，且装载到数据仓库的目标数据库中。

3)数据存取层：该层是按主题进行分析和对相关的数据进行挖掘的数据源，包括每一个按主题进行分类的数据集市和专门用于数据挖掘的数据挖掘库。

4)数据分析服务层：该层是数据存储和前端分析工具的桥梁，它包括OLAP分析引擎、安全控制机制等，能按照用户的要求设计、生成具有多维分析功能的分析主题，予以组织，以便进行多角度、多层次的分析，并发现趋势。它们响应前端用户的分析请求，将多维数据传送给前端的分析工具显示。

5)前端展现层：在用户眼中，用户界面的显示才是最重要的，本系统采用Hyperion公司的BI软件来设计基于Web的数据展现和图形展现，并提供给用户多种查询方式，能根据用户要求钻取到相关层，获得相关的明细数据。用户可以通过B/S和C/S两种方式获取展示的数据。展现层中主要使用一些展现方法，如二维笛卡尔坐标系、三维笛卡尔坐标系、二维墙面坐标系、三维墙面坐标系和一些特殊效果处理，如光照、雾化、融合、纹理等，还有可视化工具，如图表、曲线、决策树、规则图、直方图、饼图、高低区域图等。展示样式见图2、图3。

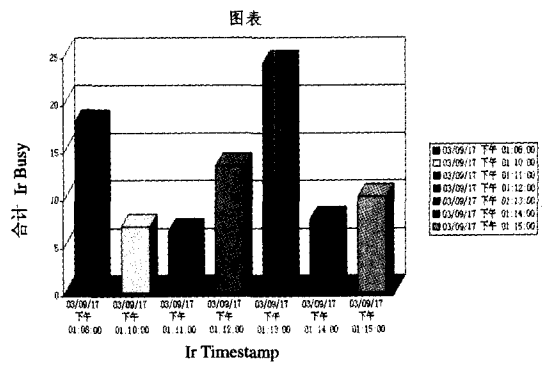


图2 展示样式一

和传统的系统结构相比，本系统采用了基于主题的数据集市，而不是针对部门的数据集市。针对不同的主题，用户可以进行不同的查询和分析。对于主题本身的选择，采取传统方法和“基于特征值的数据仓库主题搜索方法”相结合；并单独成立一个数据集市，作为数据挖掘库，这样对数据挖掘来讲，可以减少冗余数据，大大提高数据挖掘的速度和效率；对于客户端的用户，可以在本地对数据进行离线统计分析，减少了网络的负荷，提高了运行速度和效率。

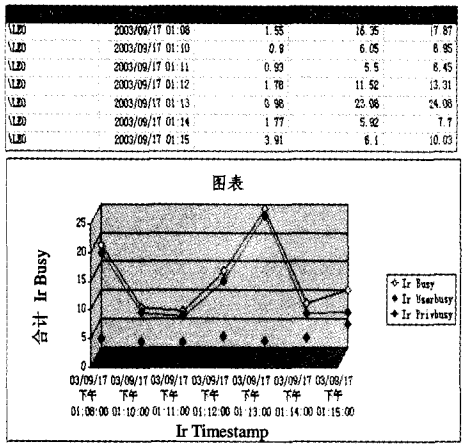


图 3 展示样式二

## 3 系统实现中的关键技术研究

### 3.1 主题分析和数据建模

由于数据仓库在规模大小、集成程序和体系结构方面都与传统的数据库不同，因此传统的数据结构无法满足其要求，必须寻找一种新的数据结构来描述数据模型。在该系统的数据分析中，会涉及到大量的 OLAP 操作，包括切片、切块、旋转和钻取等。通过对现有系统大量数据的分析，并结合数据仓库环境的特点，我们选定星型模式来描述数据模型。星型模式通常由一个包含主题的事实表和多个包含事实的非规范化的维表组成。通过事实表将各种不同的维表连接起来，维表中的对象通过事实表与另一维表中的对象相关，这样建立各个维表对象之间的联系。

为了给决策分析者提供不同的分析角度，每个主题的数据都采用数据集市的方式存贮。在主题的确立上采用两种方式来实现：

(1)根据业务需求的划分，我们将数据仓库系统按要求划分为以下几个主题：主机主题、网络主题、数据库主题、防病毒主题、防火墙主题、环境主题等。

(2)在其他主题的建立上，本系统使用了一种“基于特征值的数据仓库主题搜索方法”，根据数据仓库主题的特征属性来确定其他的主题。此方法能够自动地从关系数据库中寻找具有这种数据仓库主题特征的表，为数据仓库设计的主题选择提供辅助参考。定义给出了数据库中表 R 的主题相关属性值，其中外键特征值 $f$、数值特征值 $n$、时间特征值 $d$ 和视图特征值 $t$ 都是静态的，元组特征值是动态的。取静态特征值的平均数与动态特征值的乘积作为表的主题特征值，定义表 $R$ 的主题特征向量为 $Md=t\times(f+n+d+v)/4$。$Md$ 越大，越适合作为主题。

### 3.2 ETCL 中的实现方法

(1)抽取方式：根据本系统的特性，ETCL 数据加载采用时戳方式对数据进行抽取，在原有系统的业务表中统一添加时间字段作为时戳(如表中已有相应的时间字段，可以不必添加)。每当数据源系统中更新修改业务数据时，同时修改时间戳字段值。对关系型数据库中的表，可以使用触发器的技术来实现。对于其他的数据，通过编写抽取器代码的方式来实现。当进行 ETCL 加载时，通过系统时间与时戳字段的比较来决定进行何种数据抽取。ETCL 系统设计清晰、源数据抽取相对简单、速度快。在抽取策略上，制定一个行之有效的数据抽取时间和方式(比如 00:00～4:00)，为系统不断补充合格的数据，实现数据的递增加载。

(2)在 ETCL 中，数据清理是一项很重要的工作，而重复记录消除是数据清理研究中一个很重要的方面，它的目的是检测并消除那些冗余的、可能对后来的 OLAP 和数据挖掘造成影响的数据。已有研究都是通过设定一个相似度阈值来判断两条记录是否为重复记录。过大的阈值将导致返回率下降，过小的阈值将导致误检率上升。本系统使用了一种双阈值的重复记录消除方法，利用数据仓库环境下数据库表之间的外键联系做进一步判断，可以有效地提高判断质量，减小误检率。

FKSim(R1,R2)=sim (R1,R2 )，如果 sim(R1,R2)>=UP 或者 sim(R1,R2)<=LOW

FKSim(R1,R2)=(1−$d$ ) * sim(R1,R2)+$d$ * sim(FK(R1),FK(R2 )),LOW<sim(R1,R2 )<UP

集合 s1 和 s2 的相似度 sim(s1,s2) $d$ 为贡献因子，它的意思是在原始相似度处于 LOW 和 UP 之间的时候，应该用另一个相似度来衡量 r1、r2 的相似程度，则 FK(R)表示所有外键参照 R 中记录的记录集。

(3)缺失数据的处理：除了原始数据本身有缺失数据外，前面在抽取数据时也会产生新的缺失数据。如果缺失数据的前后时间间隔不大，我们采用线性插值的方法将其补上。例如，如果我们已知 $n$ 时刻、$n+i$ 时刻的 cpu 的占用率为：$T_n$，$T_{n+i}$，而缺少中间的数据，则中间时刻 $n+j$ 的取值为：

$$T_{n+j}=T_n+\frac{T_{n+i}-T_n}{i}\cdot j \quad 0<j<i$$

### 3.3 数据仓库中的查询优化

由于数据仓库中数据主要是批量增加数据，数据的修改量很小，维护索引的开销就很小。因此，在数据仓库中可以根据需要添加足够多的索引，以此来提高查询的性能。

在查询的过程中，事先生成一些实物化视图。这样，查询时就可以直接使用存储在实物化视图中的数据，而不需要实时计算。这是一种利用空间的代价来换取时间效率的方法。

基于数据仓库星型模式的特点，本系统在进行数据仓库的查询时引用了一种新的多表排序连接算法，这种算法直接应用于数据仓库，可以提高数据库的查询性能。算法过程如下：

```
Algorithm MulSortJoin (F, D1, D2,…,Dm)
  Input:事实表 F(ID1,…,IDm,A1,…,Ap)
        维表 D1(ID1,B11,…,B1q),D2(ID2,B21,…,B2r) ,…,
        Dm(IDm,Bm1,…,Bms)
  Output: F 和 D1,D2,…,Dm 的连表 T。
{
FOR (int i=1; i<=m; i++)
{
   将维表 Di 的所有记录读入内存;
   按属性值 Di [IDi]排序关系得到排序表 DDi;
}
FOR (int j=1;j<=F. count;j++) //F. count 为事实表的记录数
{
   r=事实表中的第 i 条记录
   FOR (int i=1;i<=m;i++)
   {
     di= 根据 r [IDi]查找排序表 DDi 得到的记录; //r [IDi]为事
     实表的对应于维表的外键的值
   }
   利用 r,d1,d2,…,dm 形成完整的连接结果元组 t;
   T = T ∪{t};
  }
  RETURN(T);
}
```

这个算法所产生的 I/O 开销为 BF+BD1+BD2+…+BDm+U 块磁盘存取。其中 BF 为事实表 F 的磁盘块数，

(下转第 162 页)

表 2 几种算法测试结果

| 算法 | | Oliver30 | | | att48 | | |
|---|---|---|---|---|---|---|---|
| | | 最好解 | 平均值 | 最差解 | 最好解 | 平均值 | 最差解 |
| 模拟退火算法 | | 438.5223 | 424.6918 | 479.8312 | 34958 | 35176 | 40536 |
| 基本遗传算法 | | 483.4572 | 467.6844 | 502.5742 | 38541 | 38732 | 42458 |
| ant-density system | | 425.6490 | 429.1071 | 433.1215 | 33786 | 35688 | 36559 |
| ant-quantity system | | 424.6727 | 429.7738 | 435.2420 | 33902 | 35714 | 36111 |
| ant-cycle system | | 423.7406 | 429.7032 | 432.4568 | 33780 | 35595 | 36534 |
| $m_1:m_2:m_3$ | 1:1:1 | 423.7406 | 429.2475 | 432.4168 | 33896 | 35792 | 36357 |
| | 2:1:1 | 423.7406 | 427.9214 | 431.4568 | 33522 | 35561 | 36100 |
| | 3:2:1 | 423.7406 | 429.3649 | 432.7944 | 33657 | 35599 | 36297 |

**结束语** 利用蚁群算法信息更新特性，提出的多样信息素的蚁群算法可以显著提高计算效率，具有较大的实用价值。尽管国内外研究蚁群算法的比较多，但还有许多问题值得研究，如算法的参数选择只能通过仿真实验，无法给出理论指导。但从当前的应用效果来看，这种模仿自然生物的新型系统寻优思想无疑具有十分光明的前景，更多深入细致的工作还有待于进一步展开。

## 参 考 文 献


1 Colorni A, Dorigo M, Maniezzo V. An investigation of some properties of an ant algorithm [A]. In: Proc. of the Parallel Problem Solving from Nature Conference (PPSN'92) [C]. Brussels, Belgium: Elsevier Publishing, 1992. 509～520

2 吴庆洪，张纪会，徐心和. 具有变异特征的蚁群算法[J]. 计算机研究与发展，1999，36(10)：1240～1245

3 马良，项培军. 蚂蚁算法在组合优化中的应用[J]. 管理科学学报，2001，4 (2)：32～33

4 Gunes M, Sorges U, Bouazizi I. ARA the ant colony based routing algorithm for MANETs [A]. In: Proceedings International Conference on Parallel Processing Workshops [C]. Uuncouver, B C, Canada, 2002. 79～85

5 Lumer E, Faieta B. Diversity and adaptation in populations of clustering ants [A]. In: Proc. of the 3 Conf on Simulation of Adaptive Behavior [C]. MIT Press, 1994. 499～508

6 Parpinelli R S, Lopes H S, Freitas. Data mining with an Ant Colony optimization algorithm [J]. IEEE Transactions on Evolutionary Computation, 2002, 6(4): 321～332

7 Dorigo M, Maniezzo V, Colorni A. Ant system: optimization by a colony of cooperating agents [J]. IEEE Trans on System, Man and Cybernatics, 1996, 26(1): 28～41

8 Stutzle T, Hhoos H. The MAX-MIN ant system and local search for the traveling salesman problem [A]. In: Proceedings of the IEEE Interna tional Conference on Evolutionary Computation (ICEC'97) [C]. Indianapolis, USA, 1997. 309～314

9 康立山，谢云，尤矢勇，等. 模拟退火算法[M]. 北京：科学出版社，1994. 130～151

10 高尚. 基于 MATLAB 遗传算法优化工具箱的优化计算[J]. 微型电脑应用，2002，18(8)：52～54


(上接第 130 页)

BD1，BD2，…，BD$m$ 分别为维表 D1，D2，…，$D_m$ 的磁盘块数，U 为结果 T 的磁盘块数。因为事实表的记录数远远大于维表，所以 BF>>BD1，BF>>BD2，…，BF>>BD$m$，总的磁盘存取块数近似等于 BF+U。

对于传统的多表连接处理方法，如果两两表连接的次序为 F×D1×D2×…×D$m$，则因为 BF×D1×D2×… ×D$i$>BF，最终连接运算所要求的磁盘块存取数将大于(2$m$−1)BF+U，其中读操作为 $m$ 次，写操作为 $m$−1 次。显然，当 $m$>1 时，多表排序连接算法要比普通的多连接算法磁盘存取块数少很多。

**总结** 在当今信息时代，企业往往被淹没在来源于多个渠道的庞大、丰富的海量数据中。只有及时地将数据有机地组合在一起，及时地将信息转化为知识和智能，才能更好地指导企业进行商业决策和行动。商务智能的作用就在于此，它帮助管理者做出科学的计划、判断、决策，避免主观、片面等因素引起的重大失误。本文主要从技术的角度对商务智能的基本概念、技术架构和涉及到的技术进行了探讨，并对具体实现进行了深入研究。大连商品交易所计算机系统监控项目已经投入到应用中，对整个期货交易系统起到保驾护航的作用，有效地增强了企业的敏捷度和竞争力。通过集成化的商务智能系统的构建，企业高效地利用现有的信息基础设施中的数据，大大提高了用户对数据查询分析和决策分析的能力。商务智能是一个跨学科的方向，还存在很多没有解决的问题，比如在数据仓库中的可视化的数据挖掘系统等问题，需要进一步进行研究。

## 参 考 文 献


1 Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM Sigmod Record, 1997, 26(1): 65～74

2 Inmon W H. Building the Data Warehouse [M]. John Wiley &Sons Inc, 1996

3 O'Neil P, Quass D. Improved query performance with variant indexes. ACM Sigmod Record, 1997, 26 (2) : 38～ 49

4 Wang L, Li Y J, Wijesekera D, et al. Precisely answering multi-dimensional range queries without privacy breaches. Proceedings of the Eighth European Symposium on Research in Computer Security (ESORICS'03), 2003

5 Gao Aiqiang, Li Qingzhong. The Data Manipulation and Querying Optimization. In SCDDWS, Computer Engineering and Application, 2002

6 王珊. 数据仓库技术与联机分析处理. 北京：科学出版社，1999

7 陈晓云，郭朝珍. 数据析取分类研究与设计. 计算机应用，2001，21(8)：1～2

8 王裕明，吴忠. 商务智能中元数据管理模型研究. 计算机应用与软件，2005(8)

9 王卫平，徐宏发. 基于 Web Services 的商务智能的网络研究. 计算机系统应用，2005(7)