doi:10.16652/j.issn.1004-373x.2016.11.037

基于数据挖掘的商务智能系统的设计与实现

李　菲
（广西财经学院，广西 南宁　530003）

摘　要：信息系统逐渐活跃在各大企业的管理系统中，帮助企业解决日常事务，但由于其相互独立，集成度不够高，给企业产生了海量的历史数据旦无法充分利用。针对上述现象，通过对企业的经营和业务活动进行分析判断，使分散在企业各个系统中的信息有机集成，并且结合恰当的分析模型和算法，利用现有的企业信息库为企业的发展和市场竞争提供有效的企业参考，提高企业的竞争力；通过分析企业采取和最终实施商务智能的系统全过程，重点介绍企业基于报表系统的领导决策系统的构建过程，为其他企业提供了宝贵的借鉴案例。

关键词：数据挖掘；报表系统；商务智能；数据仓库；联机分析处理
中图分类号：TN915.09-34；TM417　　**文献标识码**：A　　**文章编号**：1004-373X(2016)11-0152-04

Design and implementation of business intelligence system based on data mining

LI Fei
(Guangxi University of Finance and Economics, Nanning 530003, China)

Abstract: The information system is gradually active in the management system of various large enterprises, and helps the enterprises to solve the daily affairs. However the information system is mutually independent, and its integration is low, so the enterprises can't make full use of the massive historical data. In view of the above phenomenon, the enterprise's management and business activities are analyzed and judged to integrate the information dispersed in various systems of the enterprises. In combination with the appropriate analysis model and algorithm, the existing enterprise information database is used to provide the effective enterprise reference for enterprise development and market competition, and improve the enterprise competitiveness. The whole process to finally realize the enterprise, bussiness inteligence system is analyzed. The constructure process of enterprise's leader decision-making system based on report system is introduced emphatically, which provides the valueble reference case for other enterprises.

Keywords: data mining; report system; business intelligence; data warehouse; OLAP

面对瞬息万变的市场，有许多问题需要企业的生产经营者去调查、分析、研究[1]。为了应对激烈变化的市场环境，企业开始充分利用信息技术提高其竞争力，各种信息系统，如CRM，ERP，EIS开始在企业中得到广泛的应用。虽然这些系统能够满足企业日常事务性工作的需要，但是各信息系统之间相互独立，关联性并不强，各系统对与企业多年经营积攒下来被束之高阁的海量数据的处理，以及及时、准确的商务分析力不从心，无法为企业的管理和决策给出指导性建议。

1　系统需求分析

1.1　现状分析

研究企业已经通过使用Office Automatic System、用友NC财务核算系统以及U9报表填报系统等系统，基本上实现了办公的自动化、信息化[2]。而研究企业是一个有着20多家项目公司的大型企业，项目公司分布广泛，管理层级多，大部分核心业务数据仍停留在手工采集、汇总、分析的阶段，同时多年来运营积累下来的历史数据量较大，以各种形态散落各处，无集中管理，数据梳理较难，无法将数据变成信息或知识，无法对未来的经营预测、战略决策提供支持。本系统利用数据统计管理平台，以报表形式收集下属项目公司的填报数据，实现总部与下属公司之间的办公自动化和信息化；BI系统通过对收集到的数据进行分析加工，最终利用BI分析工具加以展现。生产经营管理系统利用数据统计管理平台收集填报报表数据，实现各业务部门报表数据收集统计业务，将数据加工后在领导决策分析平台（BI分析工具）加以展现分析[3]。

1.2　数据统计管理平台

数据统计管理平台采用的是IUFO报表系统，实现

收稿日期：2015-11-16
基金项目：2015年广西财经学院信息与统计学院学科建设课题(2015XK33)

各部门及下属项目公司日常填报数据管理,将管理人员从大量的数据收集、整理工作中解放出来,有效地提高了工作效率[4]。

1.3　领导决策分析平台

BI系统实现对收集到的数据进行有效的统计分析,对公司决策工作起到辅助决策作用[5]。整个系统的功能模块划分如图1所示。

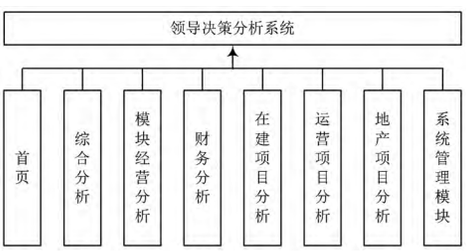

图1　BI系统功能分析图

决策分析系统由8部分组成。首页主要分析的内容是公司总体的新签合同额、完成投资额、利润总额的计划执行情况,同时了解下属项目区域分布情况,查看不同区域项目重点指标完成情况;综合分析模块是对公司总体情况进行计划和分析的模块;之后的板块经营分析、财务分析、在建项目分析、运营项目分析、地产项目分析都是针对R企业某一特定经营项目的分析;系统管理模块是专门为系统管理员单独设定的一个模块,管理员可利用这一模块对整个系统进行维护和更新[6-7]。

2　系统详细设计

2.1　报表服务设计

根据研究企业的业务需求,报表服务设计选择的是固定式报表服务,应能支撑查询分析所需的报表功能[8]。报表服务具有以下功能和要求:

(1) 应完全针对国内报表的需求设计和开发。做到美观、直观、简单和实用。

(2) 除了支持常规BI报表的功能以外,还应有独创的报表功能。

(3) 可视化操作,设计和预览应在同一个页面上进行,以拖拽的方式设计报表,做到真正意义上的"所见即所得"应用。

(4) 支持"业务视图"。能够保证从业务角度访问、使用和分析数据,将数据库和数据的复杂性隐藏在业务视图之后。

(5) 支持报表模板功能,可以自定义模板,也可以将报表转化成模板。

(6) 支持"自由钻取"。自由钻取是数据钻取(Drill-down)功能的扩展。通过自由钻取可以将数据、文档、图片、视频等一切电子形式的内容相互关联。

(7) 具有更强的交互功能。通过动态参数、提示筛选、上下文自动计算等功能,增强了报表的交互能力。

(8) 拥有丰富的图表功能。常用的统计分析图表被固化在报表中,可以随时随地生成直观的图表。

(9) 支持OLE,可与Microsoft Office等软件相互嵌套。

(10) 具有丰富的格式和样式设置。格式和样式(包括报警)的设置参考了Excel的实现方式,从而满足绝大多数用户的使用习惯。

(11) 支持自定义函数。预定义报表除了包括数据库的内建函数和扩展的函数之外,还支持用户自定义函数功能。用户可以根据需要创建自己的计算函数。

(12) 严格的数据安全控制。支持行级的数据访问权限,并可以通过数据库视图和业务视图两条途径实现数据访问权限的设置。

2.2　数据仓库设计

(1) 数据仓库(高层模型)概念模型

概念模型设计要完成的工作有界定系统边界、确定主要的主题域和内容。

界定系统边界。研究企业是一家大型投资企业,所包含的业务分类也很多,决策分析时,对不同业务的审查需要不同的数据,例如,对运营项目审查时,就要对下列信息进行分析:项目年度经营计划表、吞吐量情况表、项目基本情况表、采购分析表。所以把系统边界定为研究企业所经营的范围内的各分公司的经营信息:分公司基本信息、分公司经营计划、分公司经营情况、分公司财务数据等[9]。

确定主要的主题域和内容。系统边界界定之后,根据各分公司经营项目的不同,将其分为四个模块:板块经营分析、在建项目分析、运营项目分析、地产项目分析。按用户要求,本文将财务数据和综合分析单独存放,作为额外两个主题域:财务分析和综合分析。

(2) 数据仓库(中层模型)逻辑模型

中间层逻辑模型是对高层数据概念模型的细分,在高层数据模型中所标识的主题域都需要与一个逻辑模型相对应。通过中层逻辑模型的设计,可以向用户提供一个比概念模型更详细的设计结果。在这一步主要进行的设计有:丰富和分析主题域;粒度的确定;数据分割策略的确定;关系模式的定义;记录系统的定义。

本文将整个时间粒度划分为日、周、月、季度和年,其中月、季度和年是所有报表都有的粒度,而周和日粒度是某些报表特有的粒度。系统主要按照业务板块进行讨论与开发,所以,在数据分割这一块也按照板块来

划分数据存储单元，即按照板块分析、在建项目分析、运营项目分析、地产项目分析和财务报表分析划分。之后，再根据不同板块的业务需求，以各版块的数据报表来划分。

（3）数据仓库（底层模型）物理模型

在综合考虑了研究企业服务器的存储空间利用率、购置的成本、存取的速度以及维护的代价后，本文采用了目前比较通用的容错结构廉价冗余磁盘阵列（Redundant Array of Inexpensive Disk，RAID5）。RAID5通过某种算法决定某组数据块的校验块的存放位置，正是这样的结构，保证当某个磁盘故障时不会丢失数据，而且其读取速度较快，但是写入速度由于校验过程会受到轻微的影响。

2.3 数据的组织结构

根据之前数据仓库的设计，可以根据事实表与维度表的外键关系设计出OLAP过程中的重要组成部分——数据立方体。由于所有的事实表都有四个维度表：时间维度表、指标维度表、项目维度表、单位维度表，由于决策一般用到一级指标，但是二到四级指标也都是最终分析转换为一级指标的依据，这里使用四级指标，所以，立方体的维度为四级指标，月份和项目公司，如图2所示。

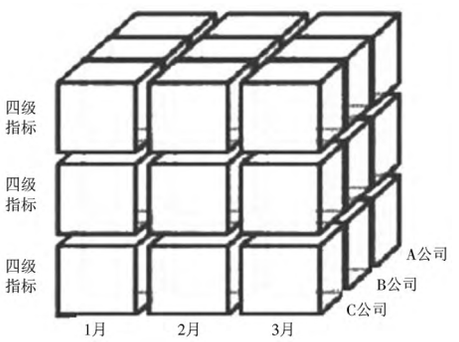

图2 数据立方体示意图

3 系统开发与实现

3.1 数据仓库

系统的数据准备过程是从IUFO数据报表系统后台数据库中抽取数据，进行一定的处理（将不同时期的数据进行统一定义、通过以单位等）存入中间表中。系统可以从各个数据库表中抽取数据，经过特定的Query查询语句，把相对应的数据存入到相对应的维度表中，在数据统计管理平台后台数据库中，DW_Rport中存储了各报表的关键信息，而DW_Rport_Item中存储的是相应报表的具体数据项。需要把原始报表存入临时表temp_report_ distinc中，之后进行初始化操作。然后，把数据格式、名称、时间格式等进行标准化处理，为转换阶段做好准备并且删除掉重复的数据。处理完数据之后，系统中的维度表已全部生成，数据也全部导入到临时表temp_report_distinc中。

3.2 联机分析处理

（1）Data Cube优化

由于本系统就单个项目公司来说，所要填报的报表就有上百张，况且研究企业的项目公司多达几十家，所以总共的报表就有上千张，要在这么多报表中进行联机分析，必须对既有的分析框架结构进行优化，以提高总体的分析效率，减少报表之间的重复查询。

① 排序、分组和散列。将所有维度表进行排序、散列和分组，使维度表有顺序的集成，以便于相关的元组数据重新排列和聚集。

② 把中间结果同时缓存和聚集。用之前计算的较低层次的结果计算较高层次的数据，而不是从最底层的数据开始计算，减少了输入输出次数，增加了效率。

③ 当存在多个维度时，从最小的维度开始聚集。比如说项目中报表的经营计划表，分为月、季度、年三个类型，若要计算经营计划表，那么最有效、最直接的方法就是从月度开始聚集。

（2）部分物化视图的实现

在进行部分物化之前，从系统的主要数据类型确定如何建立物化视图，分析平台的数据都是由数据统计管理平台得来，而数据管理平台的数据不外乎两种，数值型数据和文字型数据，所以，选择数据泛化的方向为面向属性的泛化。

（3）数据立方体的计算

采用多维数组作为基本数据结构，计算整个数据立方体，将数据分成三块，按时间维度，项目维度和指标维度聚集，并把按各维度所分的块再细化，时间维度块又分成月、季度和年维度分别聚集；项目维度块分成企业下的各项目公司分别聚集，并将这些小维度块排序，使这些维度块的被访问次数降到最低，以减少内存的占用和输入输出开销，最后将时间维度和项目维度聚集完的总聚集块作为指标维度聚集的基础，从而提高OLAP的分析速率。

3.3 数据挖掘

C4.5算法用信息增益率选择属性，在构造过程中进行枝剪，增强了对不完整数据处理的能力以及对连续属性的离散化处理，基于这些优点，本文选择C4.5作为数据挖掘的算法。C4.5用信息增益率函数作为属性选择的标准，定义如下：

假设 L 是一个包含1个数据样本的集合，类别有

m 个值,对于 m 个不同类别 $D_i, i \in \{1,2,\cdots,m\}$。假设 l_i 为类别 D_i 上的样本数量,则需要的信息量为:

$$\mathrm{Inf}(l_1,l_2,\cdots,l_m)=-\sum_{i=1}^{m} p_i \log_2 p_i \quad (1)$$

再假设一个属性 B,有 n 个不同的值 $\{b_1,b_2,\cdots,b_n\}$, 属性 B 可以将集合 L 划分为 n 个子集 $\{l_1,l_2,\cdots,l_n\}$,其中 l_j 包含了 L 集合中属性 B 取 b_j 的数据样本,若属性 B 被选择为测试属性, l_{ij} 为子集 l_j 中属于 D_i 类别的样本个数,则利用 B 属性划分当前样本集合所要的信息熵为:

$$E(B)=\sum_{j=1}^{n}\frac{l_{1j}+l_{2j}+\cdots+l_{mj}}{L}\mathrm{Inf}(l_{1j}+l_{2j}+\cdots+l_{mj}) \quad (2)$$

式(2)中所有属性中 B 取 b_j 值的样本数总和除以 L 集合的样本总数。则属性 B 对当前集合的信息增益 Gain(B)为:

$$\mathrm{Gain}(B)=\mathrm{Inf}(l_{1j}+l_{2j}+\cdots+l_{mj})-E(B) \quad (3)$$

C4.5的信息增益定义为:

$$\mathrm{Gainratio}(B)=\frac{\mathrm{Gain}(B)}{\mathrm{Spliti}(B)} \quad (4)$$

其中Gain(B)如式(3)定义。

$$\mathrm{Spliti}(B)=-\sum_{i}^{n}\frac{p_i}{m}\log_2\left\langle\frac{p_i}{m}\right\rangle,\ m=p_1+\cdots+p_i+\cdots+p_n \quad (5)$$

C4.5算法产生的分类规则易于理解,而且准确率很高,误差比较小,整体预测效果比较理想,它所要求的样本数量比较庞大,可以避免因数量收取过少而产生的偏差,同时能够充分利用数据仓库庞大的数据资源,适应性比较强,基本满足了决策者对属性选择的要求。

3.4　系统界面

(1) 数据统计管理平台界面

将企业原有的数据填报平台NC系统与IUFO数据统计平台系统相对接,展示在一个界面上是本系统实现的一个特点,充分考虑到建设原则中的兼容性这一原则。

(2) 首页

将之前设计好的表格改成折线图,这样的设计,使登陆者一目了然,迅速获得企业经营状态。

(3) 综合分析模块

综合分析模块把企业决策者重视的指标,即新签合同额、完成投资额、利润总额、归属母公司净利润,包括年度、季度和月度的指标显示于图上,以便于公司高层能够先从整体上掌握整个企业的营业状况。

(4) 运营项目分析

运营项目下又分为多个部分,包括完成投资额、建安工程费、交母公司现金总额、融资额和回收款,这些项目又分为年度、季度和月度,可以根据决策者的需要点击标签,切换时间维度的分析。

(5) 地产项目分析

地产项目分析模块也分为多个部分,包括投资额、新开工面积、销售面积、销售额、回款额以及营业收入,这些项目也分为年度、季度和月度,可以根据决策者的需要点击标签,切换时间维度的分析。

4　结　论

通过对研究企业经营现状的分析,在明确了企业实际业务需求的基础上,集中围绕着BI的核心技术(数据仓库),联机分析处理和数据挖掘进行设计,开发过程中对企业的业务逻辑进行了梳理,并且合理地设计了系统各功能模块的划分及其模块的主要内容,并且根据企业高层领导的决策需求,设计了合理的数据集合,抽取以及展现的过程,最终实现了领导决策分析平台和数据统计管理平台。数据管理平台的使用,使得数据来源集中化,数据格式规范化,同时,其精确的数值也为领导决策平台提供新的可靠保障;领导决策平台的使用实现了企业的无纸化办公,更加能让用户不限时间地点的掌握企业的运营状况,能够对企业运营当中突发的事件作出快速的回应,从而大大地提高了工作的效率。

参　考　文　献

[1] 张美虎,陈网凤.基于ERP的商务智能系统的设计[J].牡丹江教育学院学报,2008(6):107-108.

[2] 程蕾,程鹏.知识管理与商务智能整合思考[J].情报科学,2012,30(7):1084-1087.

[3] 张哲,戎立.商务智能技术及其应用[J].中国商界,2010(9):318.

[4] 牟少霞.基于智能终端的移动电子商务商业模式研究[D].济南:山东师范大学,2014.

[5] 夏国恩,金炜东,张葛祥.商务智能在中国的现状和发展研究[J].科技进步与对策,2006(1):173-177.

[6] 刘泽.我国企业应用商务智能的现状、挑战与对策研究[J].科技管理研究,2012(2):34-37.

[7] 郭星明,何勇.商务智能本体云架构设计[J].电信科学,2011(11):35-40.

[8] 谭有福,岑贤生,王咏梅,等.下一代商务智能的研究与设计[J].广西轻工,2011(8):122-123.

[9] 屠建平,杨雪.基于电子商务平台的供应链融资模式绩效评价研究[J].管理世界,2013(7):182-183.

[10] 陈小燕.机器学习算法在数据挖掘中的应用[J].现代电子技术,2015,38(20):11-14.

作者简介:李　菲(1981—),女,壮族,广西防城人,硕士,讲师。研究方向为电子商务、计算机应用。