基于人工免疫检测的商业智能系统 *

徐 锐[1], 马文丽[1,2], 郑文岭[1,2]
(1. 上海大学 电子生物技术研究中心, 上海 200072; 2. 广州南方医科大学 基因工程研究所, 广州 510515)

摘 要: 设计商业智能系统,分别构建 ETL 模块、数据仓库和 OLAP 系统。对于 Cube 中经常出现的异常数据问题,提出使用人工免疫系统进行检测,将 Cube 查询的 KPI 历史数据进行二进制编码作为自我集合,并用阴性选择算法产生检测器。设计基于人工免疫检测的新商业智能系统,测试表明,改进的系统可以有效地检测异常数据的存在,从而保障了最终用户端使用数据的准确性和整个商业智能系统的可靠性。

关键词: 商业智能; 联机分析处理; 数据仓库; 人工免疫; 阴性选择
中图分类号: TP301 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0209-03
doi:10.3969/j.issn.1001-3695.2010.01.061

Business intelligence system based on artifical immune detection

XU Rui[1], MA Wen-li[1,2], ZHENG Wen-ling[1,2]
(1. *Bioelectrical Research Center, Shanghai University, Shanghai* 200072, *China*; 2. *Institute of Genetic Engineering, Southern Medical University, Guangzhou* 510515, *China*)

Abstract: Designed the business intelligence system, and built ETL module, data warehouse and OLAP system. Aiming at the abnormal data in the Cube, introduced the artificial immune system to detect by binary decoding the historical KPI data from Cube as the self set and generating the detectors based on negative selection. Designed the new business intelligence system using the artifical immune detection. The testing results show that the improved system can detect abnormity successfully, which ensures the accuracy of the data that the end customer access and the reliability of the BI system.

Key words: business intelligence(BI); on line analytical processing(OLAP); data warehouse; artifical immunity; negative selection

网际网络盛起前,握有最多信息的便是赢家。迈入信息爆炸的网络时代后,原先善于掌握信息的赢家却纷纷淹没在信息洪流中,于是 Bill Gates 在《数字神经系统》一书中大力急呼企业获取及利用信息的方式将决定企业的竞争优势。对于现代企业而言,数据可以被视为重要的资产,但是又必须要能利用这些数据,也就是把这些庞大的数据转换为有用的信息,才能产生真正的价值。

商业智能(BI)是一种综合运用了数据仓库、联机分析处理(OLAP)和数据挖掘(data mining)技术来处理和分析数据的崭新技术,它能够将数据转变为信息和知识[1]。BI 的应用为商业分析和报表制作提供了极大的便利,但同时 BI 正处在一个发展阶段,其中很多技术还尚未完全成熟,开发和使用流程中也存在一些风险因素。其中最大的风险出现在数据载入的稳定性和数据呈现的准确性上。对此引入人工免疫检测机制对 Cube 数据进行异常检测,以提高 OLAP 客户应用数据的准确性和降低商业风险。

1 商业智能系统设计

本文根据实际应用的某公司部门需求为基础,设计编制 BI 系统。

1.1 ETL 模块和数据仓库

根据具体需求建立 ETL 模块,用于数据的载入、清理和抽取。数据库软件采用 SQL Server 2005;根据分析,数据的原始来源包括工程师使用的工具 Clarify、DealtrackManagement 等,工程师使用的网站为 Decent、Webcat 等。在这些数据源中,数据事实表包括 case 的基本信息、sales 的基本信息、engineer 的基本信息、customer 的基本信息等,维度表包括 time 信息、region 信息、unitcode 信息、casestatus 信息等。由于数据源分别来自不同的部门,同时数据可得到的方式也不同,包括 SQL Server 数据表、Excel 表格、Access 文件、邮件、网站报表等,针对每种数据呈现模式分别开发不同的数据载入工具。对于 SQL Server 数据表和 Access 文件,通常采用开发 SQL Server Integration Service(SSIS)包并使用 SQL Server agent 自动运行的模式。对于规范的 Excel 表格同样也可以采用这种模式;对于不规范的 Excel 表格、邮件和网站报表,则需要开发一些抓取数据并提供异常分析的小工具,其中有用 C#编写的邮件定时采集工具,用 C#和 JS 写的网站报表抓取工具等。使用以上这些工具将来自各个数据源的数据导入到本地数据库 XDatawareHouse 中,数据的更新采取每日差量导入的方法,根据每个表时间字段 lastmodifytime 来实现更新数据的判断。

数据成功载入后,对事实表做一些相应的清理和抽取工作,然后针对每个事实表建立清理抽取后的 view。这样用于存放基本数据的数据仓库就建立好了。

收稿日期: 2009-03-30; **修回日期**: 2009-05-11 **基金项目**: 国家自然科学基金资助项目(39880032);广州市重大科技资助项目(199-2005-001);广东省自然科学基金资助项目(5004737)

作者简介:徐锐(1983-),男,山东人,博士,主要研究方向为计算机免疫学、网络安全、数据挖掘等(x120r021@163.com);马文丽(1964-),女,江西人,教授,博导,主要研究方向为电子生物学、数据挖掘;郑文岭(1962-),男,江西人,教授,博导,主要研究方向为电子生物学、数据挖掘.

1.2　Cube 设计

在数据仓库 XDatawareHouse 的基础上，本文使用.NET 2005 开发平台设计 OLAP。首先在工程中建立数据库视图和所需维度关联，如图 1 所示。

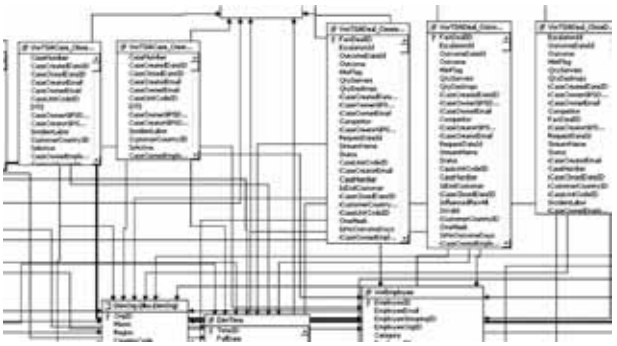

图 1　OLAP 数据库视图和表关联

然后设计为客户提供数据查询服务的 Cube。根据 Key Performance Indicator(KPI)定义，Cube 中分别需要计算如表 1 所示的指标。

表 1　KPI 定义

KPI 名称	KPI 定义
case volume#	case 数量
case closed%	完成 case 占总 case 数量的百分比
MPI#	平均处理每个 case 花费的时间
sales revenue#	销售金额
customer entitled#	注册用户数量

本文使用 Calculations 功能编写这些 KPI，如图 2 所示。

图 2　KPI 在 Cube 中的设计

完成编译后，最后在数据库部署 Cube。部署成功的 Cube 可以通过 Office Excel 2007 中集成的 Pivot Table 功能或者 MDX 查询等多种方式访问，最终用户可以从时间、地域、工程师、客户等多个角度查询 KPI，并十分容易地根据这些数据制作报表或者统计图。

2　人工免疫算法对 BI 系统的改进

对于实际应用的 BI 系统，因为数据来源复杂，数据源的个数通常也很多，如果某个数据源因异常问题而导致某天的数据出现错误，则可能最终导致某个 KPI 或者整个呈现数据的错误。与此同时，前端展示工具的使用人群往往是高级决策人员和报表制作人员，他们并不关心技术实现的细节和问题，一旦数据多次出现错误，往往造成的是对 BI 系统质疑和缺乏信任，这无疑是致命的。所以 Cube 中的数据纠错和异常检测问题在 BI 系统开发完善中亟待解决。

对此引入人工免疫检测机制对 Cube 数据进行异常检测。人工免疫系统是模仿自然免疫系统机制的一种智能方法，是一类基于生物免疫系统的功能、原理、基本特征以及相关理论免疫学说而建立的用于解决各种复杂问题的计算系统。人工免疫检测通过检测器对某个模式的匹配判别，可以有效判断该模式属于自我集合还是非我集合，从而实现对异常情况的检测。

2.1　阴性选择算法

人工免疫检测的核心算法是阴性选择算法。经典的阴性选择算法把整个空间集用二进制字符串来表达，并将所有模式分为自我和非我两部分。同时阴性选择算法采用 r 连续位匹配规则，即任意两种存在模式 m_1 和 m_2，当且仅当它们在 r 或多于 r 个连续位置上有相同字符时(同为 1 或 0)，称为它们在 r 连续位匹配规则下匹配，记为 $M(m_1,m_2)$。

阴性选择算法的核心是生成检测器集的过程，它的步骤为：对于确定的一个自我模式集合，从整个模式集合中随机挑选存在模式，如果该模式与任何一个自我模式集合中的模式都不匹配，则它成为一个成熟的检测器，否则被淘汰。检测器集制备完毕后，开始进行非我模式检测。如果外来的存在模式与检测器集中任意检测器匹配，则该模式被诊断为异常模式清除。阴性选择算法的流程如图 3 所示。

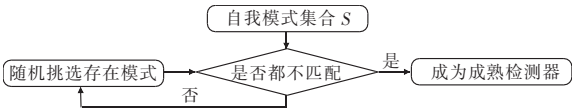

图 3　阴性选择算法流程图

2.2　BI 系统中异常数据检测

引入人工免疫检测理念和阴性选择方法，同样采用二进制字符串进行待讨论问题集合的描述。首先建立自我模式集合，该集合是指可能出现的合理 KPI 组合情况。比如在一地区某个财月内[Case Volume#]值为 100，则[Sales Revenue#]应该在 300 万左右，MPI 在 200 左右是合理的数值。如果出现较大偏差，如 MPI 在某个月低于了 150，则有 50% ~70% 的可能性是 Cube 内的数据出现了问题，如果经确认后数据没有问题，则决策者可以根据这一变化引起注意，作出相应的决策调整。基于这一方法分别对 KPI 中的[Case Volume#][MPI#]和[Sales Revenue#]的取值范围进行二进制编码，并取合理的组合模式作为自我模式集合，如表 2 所示。

表 2　KPI 二进制编码

KPI 名称	KPI 取值范围	KPI 取值范围编码
case volume#	0 ~ 99	00
	100 ~ 149	01
	149 ~ 200	10
MPI#	0 ~ 150	00
	151 ~ 200	01
	201 ~ 250	10
	251 ~ 300	11
sales revenue#	0 ~ 200	000
	201 ~ 300	001
	301 ~ 400	010
	401 ~ 500	011
	501 ~ 600	100
	601 ~ 700	101
	701 ~ 800	110
	801 ~ 900	111

3　系统设计和测试

首先设计基于人工免疫检测的 BI 系统，在该系统中对历史 KPI 数据根据经验编码得到各种正常模式组合，并生成检测器集；然后引入临时 Cube 模块，每个周期 ETL 工作完成后首

先重新部署这个临时 Cube,再通过 MDX 语句对临时 Cube 的更新 KPI 数据进行查询,经编码后得到待检测模式。使用人工免疫检测模块对这些模式进行异常检测,如果发现异常,则通知管理人员进行错误排查,如果没有数据异常,则将临时 Cube 同步到客户端使用的 Cube 中去。系统设计如图 4 所示。

根据表 2 的编码,取{0101001,0110001,0101010,0110010,1011010,1011011,1011100,1011101}为自我模式集合,其他为异常模式。选择匹配参数 $r=5$,通过基于特征值的阴性选择算法生成异常模式检测器 80 个,通过对全体模式集合的检测测试发现,共检测到包括检测器模式在内的异常模式 116,并有 4 个漏检模式被 MHC 检测窗发现。

测试和实际应用证明,在需要检查多种 KPI 组合和检查多个财月以及财年数据准确性时,该方法可以便捷有效地提供异常检测,图 5 是 BI 系统运行 12 个财月后检测情况图。

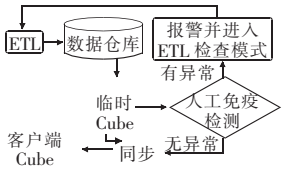

图 4　基于人工免疫检测的 BI 系统设计

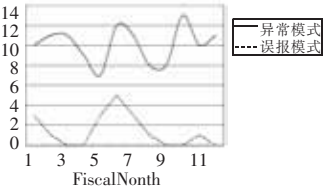

图 5　12 个财月异常模式检测统计图

因为每个财年商业情况不同,所以刚开始异常检测的误报模式还比较多,随着本财年数据的积累和对自我模式的补充,误报模式慢慢趋近于零,检测系统检测成功率大大提高。与此同时,使用人工免疫检测方法与传统的两日数据对比方法相比,异常数据的检测更为准确和高效,并真正实现了异常检测的自动化。

4　结束语

本文在设计商业智能系统基础上,将人工免疫检测引入商业智能系统的异常数据检测。通过对 KPI 的有效组合进行二进制编码作为自我集,并使用阴性选择算法生成检测器,实现了对 Cube 中异常数据的检测。对基于人工免疫检测的新 BI 系统测试表明,改进的商业智能系统可以有效地检测异常数据的存在,从而保障了最终用户端查询数据的准确性和可靠性。

参考文献:

[1] 陈京民. 数据仓库原理、设计与应用[M]. 北京:中国水利水电出版社,2004:91-112.

[2] ABDULEZER L. Going beyond spreadsheets:how visual modeling can enhance decision analysis[R]. New York:Evolving Technologies Corporation,2006.

[3] ELSON R J. Data warehouse strategy[D]. Saratoga, Florida:Dissertation of University of Saratoga,2001.

[4] HAN Jia-wei, HUANG Yue, CERCONE N, *et al*. Intelligent query answering by knowledge discovery techniques[J]. IEEE Trans on Knowledge and Data Engineering,1996,8(3):373-390.

[5] LIAUTAUD B, HAMMOND M. E-business intelligence:turning information into knowledge into Profit[M]. [S. l.]:McGraw-Hill Trade, 2000.

[6] FREEMAN O. Competitor intelligence: information or intelligence [J]. Business Information Review,1999,16(2).

[7] Cognos enterprise business intelligence for e-business[R]. [S. l.]: Ontario User Group,2000.

[8] HAN J W, CAMBER M. Data mining concepts and techniques[M]. San Francisco:Morgan Kaufmann Publishers,2001:225-244.

[9] INMON W H. Building the data warehouse[M]. New York:John Wiley,1996:45-71.

[10] TANG Zhao-hui, JAMIE M. 数据挖掘原理与应用——SQL Server 2005 数据库[M]. 北京:清华大学出版社,2007:72-93.

[11] TONY B. SQL Server 2000 数据仓库与 Analysis Services [M]. 北京: 中国电力出版社, 2003 : 11216.

[12] QUINN K R. Data visualization:gaining perspective[R]. [S. l.]:Information Builders,2006.

[13] INMON W H. Building the data warehouse[M]. New York:John Wiley,1997.

[14] BERRY M J A, LINOFF G S. Data mining techniques: for marketing, sales, and customer relationship management[M]. New York:John Wiley,1997.

[15] 王茁,顾洁. 三位一体的商业智能——管理、技术与应用[M]. 北京:电子工业出版社,2004.

[16] FORREST S, PEREI SON A S, ALLEN L, *et al*. Self-nonself discrimination in a computer[C]//Proc of IEEE Symposium on Security and Privacy. 1994:202-212.

[17] FORREST S, HOFMEYR S A. Immunology as information processing [C]//Proc of Design Principles for the Immune System and Other Distributed Autonomous Systems. Oxford: Oxford University Press, 2000:361-387.

[18] ESPONDA F, FORREST S, HELMAN P. A formal framework for positive and negative detection scheme[J]. IEEE Trans on Systems, Man, and Cybernetics,2004,34(1):357-373.

(上接第 208 页)

c)采用框架和规则相结合的知识表示方法,提高了推理效率,使得该系统在实践中更加有效。

d)利用不确定推理方法对评估事实进行推理,符合等级测评的实际情况;并且,本文对搜索策略和冲突消解策略进行了优化,提高了推理效率,并使推理结果更加准确可信。

e)利用粗糙集理论从信息安全事件中自动提取推理规则,使得推理规则能随着安全状态的变化而自动更新,从而使适合等级测评专家系统的推理结果更加准确有效。

参考文献:

[1] 杨兴,朱大奇,桑庆兵. 专家系统研究现状与展望[J]. 计算机应用研究,2007,24(5):2-5.

[2] 齐俊鹏,孙四明,王化鹏. 信息安全风险评估专家系统技术研究[J]. 计算机仿真,2008,25(9):128-129.

[3] 戴钎,王力生. 基于故障树和规则匹配的故障诊断专家系统[J]. 计算机应用,2005,23(9):2034-2037.

[4] JOSEPH C G, GARY D R. Expert system principles and programming [M]. 4th ed. [S. l.]:Thomson, 2005.

[5] 蔡自兴,约翰· 德尔金,龚涛. 高级专家系统:原理、设计及应用[M]. 北京:科学出版社,2005.

[6] 安跃文,李贤玉,王晖,等. 基于故障树的导弹发动机失效诊断专家系统设计与实现[J]. 弹箭与制导学报,2006,26(2):218-220.

[7] 苏羽,赵海,苏威积,等. 基于模糊专家系统的评估诊断方法[J]. 东北大学学报:自然科学版,2004,25(7):654-656.

[8] 龙志强,吕治国,常文森,等. 基于模糊故障树的磁浮列车悬浮系统故障诊断[J]. 控制与决策,2004,19(2):139-142.

[9] 韩祯祥,张琦,文福拴. 粗糙集理论及其应用综述[J]. 控制理论与应用,1999,16(2):153-157.

[10] 魏大宽. 不完备模糊目标信息系统粗糙集模型与知识约简[J]. 计算机工程,2006,32(8):48-51.

[11] WEI D K. Rough set model and precision reduction in incomplete and fuzzy decision information systems [J]. Computer Engineering, 2006,32(8):48-51.

[12] 陈淑珍. 基于粗糙集的几种约简算法分析[J]. 武汉工业学院学报, 2005,24(3):118-120.