数据挖掘技术在IT基础设施监控系统中的应用

宋应湃　汪林林
（重庆邮电大学计算机科学与技术学院　重庆 400065）

摘　要　本文介绍了数据挖掘技术在IT基础设施监控系统中的应用，着重阐述了采用业界广泛采纳的数据挖掘流程标准CRISP-DM，利用时序数据挖掘技术和一元线性回归预测技术从监控历史数据中发现有趣模式的过程，并针对该应用的特殊性，对相关技术的参数设置方法进行了改进，引入了一个新的指数——负载指数——以增加模型的精确性。

关键词　IT基础设施监控，数据挖掘，CRISP-DM，线性回归，时序数据

Application of Data Mining in IT Infrastructure Monitoring

SONG Ying-Pai　WANG Lin-Lin
(Computer Science and Technology Institute of CQUPT, Chongqing 400065)

Abstract　We explore a new application of data mining technology in IT infrastructure monitoring system. To find interesting patterns in historical data stored by monitoring system, including performance data and event data, we adopt the standard data mining process described in CRISP-DM and use two methods, including the time series data mining method and linear regression fit. Because of the specialties of this application, we improve the way of setting model parameter and induct a new index- load index-to improve the accuracy of model built by data mining method.

Keywords　IT Infrastructure monitoring, Data mining, CRISP-DM, Linear regression, Time series

1　引言

数据挖掘是信息技术自然演化的结果，它是从存放在数据库、数据仓库或其它信息库中的大量数据中挖掘有趣知识的过程。它与多学科领域技术的集成紧密联系，其中包括数据库技术、统计学、机器学习、高性能计算、模式识别、神经网络、数据可视化、信息检索，图像与信号处理以及空间数据分析。通过数据挖掘，可以从数据库中抽取并从多角度观察浏览有趣的知识、规律或高层信息。知识一经发现，便可应用于决策制定、过程控制、信息管理和查询处理等方面。因此，人们普遍认为，数据挖掘是数据库系统最为重要的发展前沿，是信息产业最具发展潜力的交叉学科之一。

本文即将介绍的是数据挖掘技术在IT基础设施监控系统中的应用。

2　数据挖掘在IT基础设施监控中的具体应用

国内某商品交易所是经国务院批准的三家期货交易所之一，承载其交易核心业务的IT基础设施，包括核心交易机群、场内交易网、远程交易网等，在交易期内，必须健康、高效地运作。为此，该所启动了IT基础设施监控系统，以建立一个统一的监控管理平台，实现对交易系统的网络设备、主机、数据库的实时监控，并实现所有监控数据的分析挖掘。

本文参考业界广泛采用的数据挖掘流程标准CRISP-DM（跨行业数据挖掘流程标准，CRoss-Industry Standard Process for Data Mining），实现数据挖掘在IT基础设施监控系统中的具体应用。CRISP-DM将数据挖掘视为一个迭代的、自适应的过程，由六个阶段组成，如图1所示。

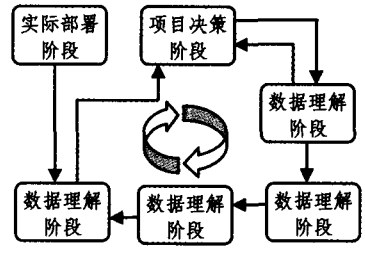

图1　CRISP-DM的数据挖掘流程

2.1　项目决策阶段

该阶段着重于从决策的角度理解项目的目标和需求，并将这些转化为数据挖掘的问题定义和实现这些目标的初步规划。

经项目初期的数据量评估，每一年存储的监控数据将达到200 G，除了用于日常的运维报告，这些数据中还包含了丰富的系统运行知识，对其进行数据挖掘是很有前景的。定义的数据挖掘目标如下：

1）定位系统性能瓶颈。系统的整体效能可能存在瓶颈，经数据挖掘，应能定位系统瓶颈所在，及时补充匮乏的系统资源，比如扩大内存容量、添加磁盘阵列、升级网络设备；

2）探寻系统负载规律。发现潜在的系统运行负载规律，供系统维护人员参考；

3）评估系统生命周期。随着交易会员的增加和交易品种的丰富，系统的负载必然日趋增大，当其潜能被开发殆尽，系统的更新换代不可避免，准确的评估系统寿命是非常有价值的；

4）优化人力资源配置。系统维护任务分为三大部分：网

宋应湃　硕士研究生，主要研究方向：数据挖掘；汪林林　教授，主要研究领域：数据库、GIS、计算机网络。

络、主机和数据库，而系统维护人员的配置是由技术难度和故障频率决定的，可以通过数据挖掘，发现系统故障发生规律，优化人力资源配置。

2.2 数据理解阶段

该阶段从汇总数据开始，然后，逐步深入，包括熟悉数据、鉴别数据质量问题并探察数据的深层次含义，或者搜索数据子集以形成对隐含知识的假想。

通过项目初期的理解，监控数据包括两大部分，一部分是网络设备、主机和数据库的运行指标，包括网络设备丢包率、主机 CPU 利用率、数据库数据字典命中率等 400 余项，其数据流如图 2 所示，这部分数据可以用作定位系统性能瓶颈和探寻系统负载规律。

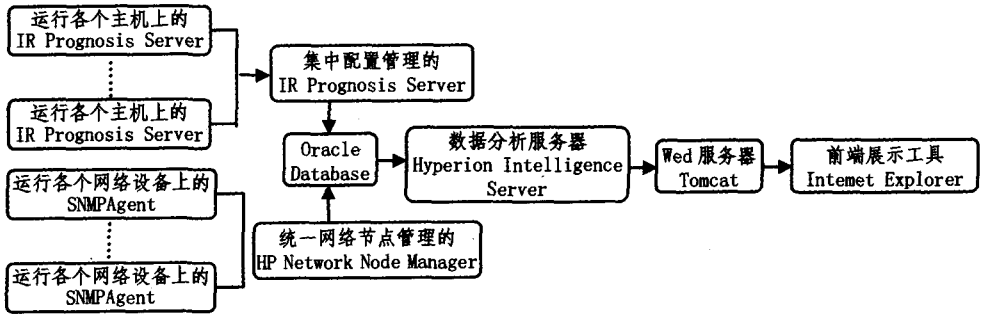

图 2 运行指标的数据流

一部分是网络设备、主机和数据库的事件数据，包括网络设备端口宕掉、主机可用内存数量低于 10%、数据库关键进程僵死等。网络设备的事件以及主机和数据库的事件分别由 SNMP Agent 和 IR Prognosis 发送 SNMP Trap 到 HP Network Manager，再由 HP Network Node Manager 完成数据写库的操作。这些数据中就包含了系统故障信息，可以用作评估系统生命周期和优化人力资源配置。

2.3 数据准备阶段

该阶段涵盖了所有从原始数据中构造最终数据集（这些数据将被用于构造数据模型）的活动。数据准备的任务可能以非预定的顺序执行多次，这包括表格、记录和属性的选取，以及数据的转化和清理。

通过对原始数据的分析，可以发现数据中存在不一致，比如，IR Prognosis 在发送的 SNMP Trap 中，将事件的严重程度分为四级：信息（Information）、警告（Warning）、错误（Error）和危急（Critical）；而 HP Network Node Manager 将事件分为五级：正常（Normal）、警告（Warning）、初级错误（Minor）、主要错误（Major）和危急错误（Critical）。必须进行数据清理和转换，以消除数据的不一致。

该阶段还应该针对项目决策阶段所指定的数据挖掘目标，进行数据的选择。通过与技术人员的交流，决定建立两个模型：

模型Ⅰ 选定网络设备的 CPU 利用率、网络设别的缓存利用率、主机的 CPU 利用率、主机的内存利用率和主机的逻辑卷繁忙事件百分比为达成数据挖掘目标 1 和目标 2 的数据集，建立系统负载模型；

模型Ⅱ 选定每周故障发生次数为达成数据挖掘目标 3 和目标 4 的数据集，建立系统故障规律模型。

2.4 模型建立阶段

该阶段选择和使用各种各样的建模技术，并将它们的参数调校到最佳值。一般说来，同一数据挖掘问题类型会有若干建模技术，其中的一些建模技术对数据格式有特殊的要求，因此，有时必须从该阶段返回到数据准备阶段。

2.4.1 负载模型

该模型包含的数据为网络设备的 CPU 利用率、网络设备的缓存利用率、主机的 CPU 利用率、主机的内存利用率和主机的逻辑卷繁忙事件百分比，这五个指标的原始采样间隔为 10 分钟，都带有时间戳，是典型的时序数据。

定义 1 时序数据是指那些随时间变化的序列值或状态，比如 CPU 利用率和程序端口的状态。

因此，可以采用常用的时序数据挖掘技术，对这些数据进行处理：

1）为了减少偶然出现的波动对模型准确度的负面影响，决定对这些数据进行平滑处理，同时，为了抵消平滑效果，以保留负载曲线趋势变化，决定对这些指标采用 3 阶加权移动平均，加权分别为（1，3，1）：

$$\frac{x_1\times1+x_2\times3+x_3\times1}{1+3+1},\frac{x_2\times1+x_3\times3+x_4\times1}{1+3+1},$$

$$\frac{x_3\times1+x_4\times3+x_5\times1}{1+3+1},\cdots\cdots$$

2）为了定位系统瓶颈，应将五项指标放在一起进行比较，但是，它们的绝对数量没有可比性，比如，交易主机 CPU 利用率的典型值为 40%，而路由器的 CPU 利用率如果达到 30%，则应该向维护人员报警。为了准确反映负载数据的相对大小，仿效季节指数（Seasonal Index），引入负载指数（Load Index），即，从系统维护人员处获取某一指标的常规值，并用此值去除对经过 3 阶加权移动平均处理后的数据，比如，主机 CPU 的利用率在交易期通常为 40%，则主机 CPU 利用率的负载指数就是 0.4；

3）为了获取系统总体负载的趋势变化，对经过 3 阶加权移动平均处理后五项指标进行加权加和后，得出系统的总体负载曲线，权值同样经系统维护人员分析后得出。

负载模型最终由五项负载指标曲线图和系统总体负载趋势图构成。

2.4.2 故障模型

如前所述，HP Network Manager 将时间分为五个等级，最高的三个等级的事件均可以视为系统故障，将每周的故障次数分级统计后，对每类事件次数加权，由高到低依次为（1，1/2，1/3），得到的数据是一组以周为计时单位的离散的故障数目点。

维护人员将每周的故障数作为衡量系统健康程度的主要标准，当该值达到一定量后，就可将系统视为不可维护的，其寿命也就结束了，因此，可以用一元线性回归方程去拟合每周的故障数目点，并使用得到的一元线性回归方程来预测系统的寿命。一元线性回归方程的概念如下：

定义 2 已知一批离散的数据 (x_i,y_i)，$i=1,2,\cdots,n$，寻求一个函数，$f(x_i)=a_0+a_1\times x_i$，$i=1,2,\cdots,n$ 使 $f(x_i)$ 与 y_i 的误差平方和 $\sum_{i=1}^{n}(f(x_i)-y_i)^2$ 达到最小，则称 $f(x_i)$ 为该组

数据的一元线性回归方程。

故障模型最终由网络和主机两部分的每周故障数目曲线图以及系统总体每周故障数目的一元线性回归方程构成。

2.5 模型评价阶段

从数据分析的角度来说，进入此阶段，表明一个或者多个高质量的模型已经建立。在进入模型的最终部署之前，为了保证模型真正体现了商业/研究目标，应对模型进行全面而细致的评估，并且审查构造模型的每一个具体步骤。其中，一个重要的目的就是确保所有的目标和需求都得到了足够的重视。在该阶段的末尾，应该对数据挖掘结果的使用完成决策。

在该项目中，通过负载模型，数据挖掘人员得出以下结论：

1)网络设备的负载，包括其 CPU 利用率和缓存利用率，普遍高于主机类的所有负载指标，因此，应酌情对网络设备进行升级；

2)系统整体负载呈缓慢上升趋势，与系统维护人员的预期相符。

通过故障模型，数据挖掘人员得出以下结论：

1)网络部分的故障数量明显多于主机部分的故障数量，因此，在系统维护人员数量配置上，应向网络维护部分倾斜；

2)利用一元线性回归方程，预测系统还可以健康运行 3～4 年，与系统维护人员的预期相符。

通过系统维护人员的验证，表明上述结论具有相当的准确性和较高的参考价值。同时，也提出了修改意见，比如，增加负载模型的指标类型，以加大趋势的全面性；将故障模型的一元线性回归该为高阶多项式回归，以提高预测的准确性。

2.6 实际部署阶段

根据需求，该阶段可能只是简单的生成报告，也可能是复杂到在整个组织内实施一个可重复的数据挖掘流程，最重要的是，如何将建立的模型应用到实际环境中去。

该项目的实际部署阶段，就是交付给系统维护人员和高层决策人员的一份对指标数据和事件数据的数据挖掘报告，该报告的框架如下：

1)描述了数据挖掘的基础概念和将其技术应用到 IT 基础设施监控系统中的可行性；

2)介绍了数据挖掘流程标准 CRISP-DM 以及采用该标准的优势；

3)阐述了按照标准的数据挖掘流程，将具体的数据挖掘技术应用到 IT 基础设施监控系统中的过程；

4)分析了上述过程中可以改进的方法和步骤；

5)探讨了将其它数据挖掘方法应用到该系统中的可行性。

结论 数据挖掘是跨学科的前沿领域，能够从各个学科汲取营养，博采众长，为我所用；它也是一个通用的方法学，在将抽象的方法和具体的应用相结合的过程中发生矛盾，解决矛盾，不断进步。将数据挖掘技术应用到 IT 基础设计监控系统中，就是一个鲜活的例子。本文采用了数据挖掘中的时序数据挖掘技术和统计学中的一元线性回归技术，论述从大量的监控历史数据中发现有趣模式的过程，再次证明了数据挖掘的魅力所在。

参 考 文 献

1 Han J, Kamber M. Data Mining: Concepts and Techniques. Beijing: High Education Press, 2001

2 Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. Washington: John Wiley & Sons, 2003

3 Daniel T. Larose. Discovering knowledge in data: an introduction to data mining. Washington: John Wiley & Sons, 2005

4 Hand D, Mannila H, Smyth P. Principles of Data Mining. Cambridge: The MIT Press, 2001

5 Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Silicon Valley: Morgan Kaufmann, 2005

6 罗可，林睦纲，郗东妹. 数据挖掘中分类算法综述. 计算机工程，2005,31(1)：3～5,11

7 吴志勇，吴跃. 数据挖掘在电信业中的应用研究. 计算机应用，2005,25(S1)：213～214

8 王曙燕，周明全，耿国华. 医学图像的关联规则挖掘方法研究. 计算机应用，2005，25(6)：1408～1409

9 刘蓉，陈晓红. 利用数据挖掘技术提高电信企业管理决策水平. 计算机应用与软件，2005，22(12)：66～68

(上接第 199 页)

效果较差，劣于 CVI,S-Link，而对球状数据效果优于 S-Link，但劣于 CVI。实验表明，本文的方法有较小的错误率，优于 K-means,S-Link 方法。

表 3 本文方法和其它聚类方法错误率之比较

有效性函数	数据集 1	数据集 2	数据集 3	Iris
CVI	0.028	0.010	0.000	0.020
K-means	0.200	0.136	0.000	0.100
S-Link	0.072	0.092	0.002	0.300

结束语 本文提出了一种元组间新的相似度计算方式，对计算中可能出现异常点的情况做了平均化的处理，并提出了一种新的有效性函数。实验表明，该方法对有重叠交叉的簇、方形的簇有较好的计算效果，能正确发现簇的个数。本文的方法也能处理多维的数据集。由于有效性函数增量 ΔCVI 计算简单，算法复杂性较低，效率较高。本文的方法与 k-means,S-Link 聚类方法比较，具有较小的错误率。

参 考 文 献

1 Jain A, Murty M, Flynn P. Data clustering: a review. ACM Computing Surveys, 1999,31(3):264～323

2 Ka L,ufman, Rousseeuw P. Finding Groups in Data: An Introduction To Cluster analysis, New York: Wiley, 1990

3 Salvador S, Chan P. Determining the number of Clusters/segments in Hierarchical Clustering/Segmentation Algorithms. Tools with Artificial Intelligence, ICTAI 2004. In: 16th IEEE International Conference, 2004. 576～584

4 Maulik U, Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 20002, 2004(12): 1650～1654

5 Kim M, Ramakrishna R S. A New Clustering Algorithm Based on Cluster Validity Indices. In: Proc. 7th International Conference, DS 2004, Padova, Italy, October 2004. 2～5

6 Halkidi M, Vazirgiannis M. Quality scheme assessment in the clustering process. In: European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD). Lecture Notes in Artificial Intelligence, 2000, 1910: 265～276

7 Schwarz G. Estimation the dimension of a model. Annals of Statistics, 1978, 6(1): 461～464

8 Bezdek J. Mathematical taxonomy with fuzzy sets. Journal of Mathematical Biology, 1974,1:57～71

9 http://www.ics.uci.edu/～mlearn/MLRepository.html

10 Newman M E J. Fast algorithm for detecting community structure in networks. Phys. Rev. E 69,066133,2004

11 卜东波. 聚类/分类理论研究及其文本挖掘中的应用:[中科院计算所博士论文]. 2000.10