基于离散 Fourier 变换的通信网可靠性数据挖掘算法模型*)

周中定　孙青华　梁雄健
(北京邮电大学　北京100876)

A Data Mining Model Based on Discrete Fourier Transform(DFT)Theory for Reliability Data of Communication Networks

ZHOU Zhong-Ding　SUN Qing-Hua　LIANG Xiong-Jian
(Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract　In this paper, it presents a data mining model based on Discrete Fourier Transform (DFT)theory for reliability data of communication networks, it helps us analyze aberrance data information and provide a method to analysis and evaluate reliability data for management of communication network. It's effective and convenient by a practical application.

Keywords　DFT-discrete fourier transform, DM-data mining

1　引言

通信网络的可靠性是与网络设备、链路与网络拓扑结构紧密相关的。目前，通讯网络的可靠性研究主要是基于网络设备、链路与网络拓扑结构，对通信网络可靠性的研究大都是将通信网络抽象为节点和链路以及各种信息(业务流)的流图进行分析，以此为基础来考虑通信网络可靠性测度，并提出了各种分析方法，这些大都是属于可靠性分析与设计的范畴，通信网络可靠性设计则考虑网络拓扑设计和业务性能设计，文[1]、文[2]对通信网可靠性设计与分析的研究进行了概述与总结。通信网络可靠性测度归纳起来有三种：网络的抗毁性、网络的生存性和网络的有效性，通信网络可靠性分析的问题可以分为三类：二端可靠性问题、K 端可靠性问题和全端可靠性问题，分析方法主要采用解析法和模拟法。

由于通信网络结构的复杂化、网络操作系统和网络协议的多样化、物理传输媒介如传输距离等的限制、分组及通信带宽的限制等多方面因素，给网络安装和运行维护部门带来了新的挑战。另外，网络中断引起的直接或间接的经济损失也对网络的维护和管理提出了更高的需求。如何有效地维护网络，优化网络性能，在网络中断时迅速隔离故障，从而最大限度地减少网络中断时间就显得至关重要。为此，通信网络可靠性研究也越来越重要。随着通信网络与技术的发展，通信网络可靠性研究在观念和范围方面将不断演进。通信网络可靠性的研究范围正从设备和系统可靠性向外延伸，从技术范畴向管理范畴延伸，使通信网可靠性的研究形成一个综合性的系统工程研究。通信网络可靠性管理就是从系统的观点出发，对各项可靠性工程技术活动进行规划、组织、协调、控制和监督，以实现既定的通信网络可靠性目标[3]。通信网络可靠性管理主要分为3个阶段来管理即：分析设计阶段可靠性管理、建设实施阶段可靠性管理、运行维护阶段可靠性管理[4]。对于运行维护阶段可靠性管理，各通信运营商对网络与节点设备的运行与故障处理都有实时监测数据，目前，这些实时运行数据只是作为一个设备情况数据记录，没有有效的方法与数学模型对这些网络运行数据进行分析或数据挖掘，分析各种故障规律，从而提出相应的可靠性措施和对重大异常故障制定应急通信制度和措施。也就是说，在收集了大量的数据之后，如何将这些数据进行整理、分析，为企业进行科学决策提供支持，也是面临的一个主要问题。数据挖掘(DM)技术，能提供一条新的解决方法。所谓数据挖掘，就是从大量不完全的、有噪声的、模糊的、随机的数据中，提取人们事先不知道、但又有用的信息和知识。数据挖掘其实是知识发现(KDD)的核心部分，而知识发现是在积累了大量数据后，从中识别出有用的知识。

数据挖掘是 KDD 最关键的步骤，也是技术难点所在。研究 KDD 的人员中大部分都在研究数据挖掘技术，采用较多的技术有决策树、分类、聚类、粗糙集、关联规则、神经网络、遗传算法等。数据挖掘根据 KDD 的目标，选取相应算法的参数，分析数据，得到可能形成知识的模式模型。概括地说，数据挖掘中最常用的技术有：1)人工神经网络：仿照生理神经网络结构的非线形预测模型，通过学习进行模式识别。2)决策树：代表着决策集的树形结构。3)遗传算法：基于进化理论，并采用遗传结合、遗传变异以及自然选择等设计方法的优化技术。4)近邻算法：将数据集合中每一个记录进行分类的方法。5)规则推导：从统计意义上对数据中的"如果-那么"规则进行寻找和推导。

文[5]利用 Hamming 人工神经网络对通信网络可靠性指标体系进行了数据挖掘及综合评价[5]，本文提出利用离散 Fourier 变换(DFT)对通信网实时监测数据进行数据挖掘的算法模型，从而对通信网络可靠性性能进行实时分析及数据挖掘，例如：网络流量(网络中各种协议的流量)、交换设备故障率、交换设备的温度、机房室温、用户服务质量申告率等通信网络可靠性运行与维护数据指标。当这些数据信息突变或畸变时，在时域中有时是很难判断与分析的，因此，我们利用

*)国家自然科学基金资助项目(NO. 79870069)。**周中定**　博士研究生，研究方向为通信网络与优化、数据挖掘、神经网络、信息管理与信息系统等。**孙青华**　博士后研究生，研究方向为通信网络与优化、数据挖掘、神经网络、信息管理与信息系统等。**梁雄健**　教授，博士生导师，研究方向为通信网络与优化、数据挖掘、神经网络、电信网络管理、电子商务、信息管理与信息系统等。

离散 Fourier 变换将这些离散数据信息进行 Fourier 变换，从而能够更准确地对这些数据进行判断与分析，即对通信网络可靠性运行与维护数据指标进行数据挖掘，从而得到有效的数据知识即知识发现(KDD)。

2 离散 Fourier 变换(DFT)数据挖掘算法模型

对于通信网管，每天都要对网络设备运行可靠性实时数据进行记录和分析，如何对这些网管数据进行有效分析和进一步的数据挖掘，目前缺乏更有效的方法和数据挖掘工具。这些网络设备运行可靠性实时数据是实时采集的，是一个具有 N 个采样的离散的序列 $x(k)$：$\{x_0, x_1, x_2, x_3, \cdots, x_k, \cdots, x_{N-1}\}$。

对于网络设备运行可靠性实时网管数据离散序列 x(k)，为了便于分析数据是否有突变或畸变的可能性，本文讨论了对这些离散数据信息的分析及数据挖掘模型与算法。

1. 数据归一化

为了便于分析与比较，因而在进行 Fourier 变换前，先将网络设备运行可靠性性能序列数据特征值统一进行规范化处理，即对特征值进行归一化，当然，由于网络设备运行网管离散序列数据的类型有效益型指标、成本型指标、固定型指标、区间型指标等，其特征值数据归一化方法也有不同的方法，下面分别予以说明。

设网络设备运行某可靠性性能序列数据特征值 x_i 区间为 $x_i \in [a,b]$，则：

1)对于效益型指标，即指标特征值越大越好型指标有：

$$x(k)=x_i-a/b-a \quad (1)$$

2)对于成本型指标，即指标特征值越小越好型有：

$$x(k)=a-x_i/b-a \quad (2)$$

3)对于固定型指标，即指标特征值既不能太大，又不能太小，而以稳定在某个固定值 c 为最佳的一类指标有：

$$x(k)=\begin{cases}1, x_i=c(\text{固定值})\\ 1-\dfrac{|x_i-c|}{\max|x_i-c|}, x_i\neq c(\text{固定值})\end{cases}$$

4)对于区间型指标，即指标特征值以落入某个区间内为最佳的类指标有：

$$x(k)=\begin{cases}1-\dfrac{x_i-q_2}{\max(q_1-a, b-q_2)}, x_i>q_2\\ 1-\dfrac{q_1-x_i}{\max(q_1-a, b-q_2)}, x_i<q_1\end{cases}$$

其中，$[q_1,q_2]$为指标特征值的最佳稳定区间。

2. 离散 Fourier 变换数据挖掘数学算法模型

对于一个连续函数 $f(t)$，其 Fourier 变换 $F(f)$定义为：

$$F(f)=\int_{-\infty}^{+\infty} f(t)e^{-j2\pi ft}dt$$

其 Fourier 逆变换为：

$$f(t)=\int_{-\infty}^{+\infty} F(f)e^{j2\pi ft}df$$

对于具有有限长 N 的离散数据信息序列 $x(k)$，其 Fourier 变换 $X(n)$定义为：

$$X(n)=\frac{1}{N}\sum_{k=0}^{N-1} x(k)e^{-jk2\pi n/N}, \text{ 其中 } n=0,1,2,\cdots,N-1。$$

其 Fourier 逆变换为：

$$x(n)=\sum_{k=0}^{N-1} X(k)e^{jk2\pi n/N}, \text{ 其中 } n=0,1,2,\cdots,N-1$$

图1所示表示数据经 DFT 变换后的过程与数据结果。从图1可以直观看出数据畸变(突变)位置非常明显直观，从而容易直观判决通信网络设备运行可靠性性能序列数据畸变的可能性，对通信网络维护和数据分析、数据挖掘是方便、有效的。

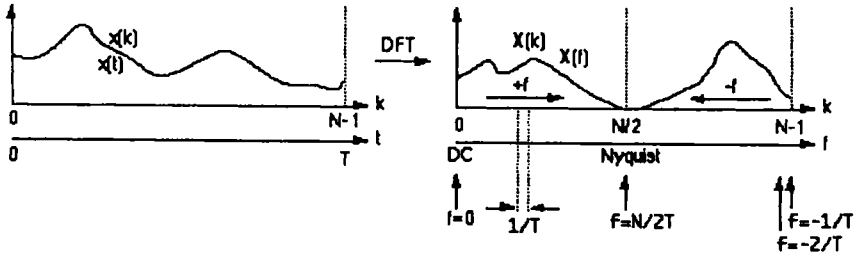

图1 离散 Fourier 变换(DFT)

3 应用实例

网络管理人员每个月的固定网络维护工作包括使用规程分析仪进行基准数据采集，并进行网络流量基准分析。规程分析仪一般具备以下功能：协议解码分析、在线监测、统计、专家系统、事后分析、模拟。协议解码分析是规程分析仪的重要功能之一，通过对网络数据流的实时的或事后的解码分析，可以了解到网络上跑的有什么协议(包括一些上层应用协议)，数据帧的源地址和目的地址、数据帧的格式等。对计算机局域通信网来说，规程分析仪一般能提供对 Ethernet MAC、TCP/IP、Novell IPX、AppleTalk、DECnet 等协议的解码。在线监测指的是将规程分析仪连接到网络上，实时地捕获网络上的数据流，并进行协议解码分析再进行统计。一些规程分析仪提供过滤器和事件触发功能，以便分析仪有选择地捕获数据。统计功能将捕获的数据、解码分析后的数据和经专家系统分析后的数据进行分类统计，给出统计结果，比如有多少超长帧、多少错帧、各种协议占网络流量的比例等。专家系统是根据已有的网络使用经验给出的网络告警或某一网络现象的可能原因，为网络管理人员提供一些参考。事后分析有别于在线检测，是事后对捕获的数据文件进行分析。模拟功能主要用来进行网络性能测试。如以不同的网络负载发送数据帧，或作为流量发生器来检测网络性能。

某单位的网络结构如图2所示。以以太网为主干，上面挂着若干个小网段，不同的小网段运行不同的协议，其中工程部网络运行 TCP/IP，市场部网络运行 AppleTalk，财务部网络运行 Novell IPX。只有经过路由的信息才能运行在主干上，不同网段上的流量只在本段内。文件服务器的备份工作在每天凌晨网络流量最低时进行。

某一个月，分析图表显示，主干网络流量与上个月相比增加了10%以上，由原来的平均20%～30%，增加到超过40%。

由于主干网络为以太网，40%以上的利用率严重影响到网络的性能，是什么原因造成网络利用率的增加呢?网络管理人员首先使用规程分析仪对网络流量进行测量，如图3所示。

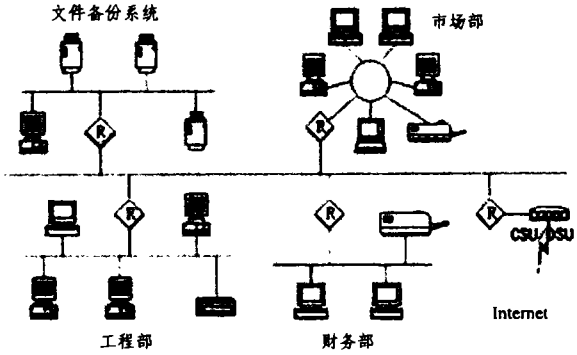

图2 某单位的计算机局域网络拓扑结构图

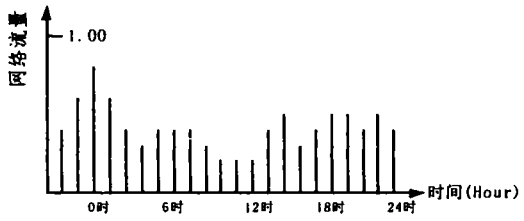

图3 平均流量/小时数据图(归一化)

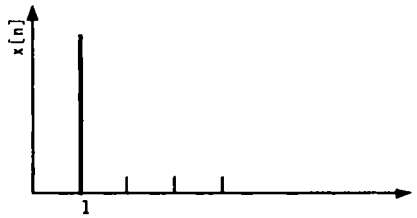

图4 离散 Fourier 变换(DFT)后

然后，经过本文提出的基于离散 Fourier 变换的通信网可靠性数据挖掘算法模型对网络流量进行离散 Fourier 变换，得出如图4所示的图形数据，直观地表现了网络流量在夜晚12时左右产生了突变。再用规程分析仪进一步查看 TCP/IP 协议帧，结果发现在夜晚12时左右 TCP/IP 协议帧大大增加，再测 TCP/IP 协议帧中上层应用 FTP、Telnet、DNS、SNMP 等所占的比率，结果发现，FTP 数据帧增加。问题出在使用 TCP/IP 协议的工程部。再使用规程分析仪所具有的 Top Talkers 测量功能。Top Talkers 测量功能能够显示同一时期发送和接收数据量最大的信息点。从以往 Top Talkers 表格显示的图形可以看出，占据 Top Talkers 前几位、处于高 Fps 群的是夜间进行数据备份的几个信息点，而本月的 Top Talkers 表格显示，处于高 Fps 群的信息点增加了。通过 Top Talkers 表格，网络管理人员很快找到了引起网络流量发生变化的信息点。通过走访工程部，网络管理人员了解到工程部新增加了一个工作组，该工作组与远程小组有大量的共享文件传输，导致网络拥塞。网管人员查看24小时网络利用率图形后，指定了一个利用率较低的时间段供该小组进行每天一次的远程集中传输。就这样，网管人员使用本文模型和规程分析仪在网络故障发生前消除了网络的隐患。

结束语 本文提出了基于离散 Fourier 变换理论的数据挖掘算法模型，利用该数据挖掘模型对通信网络可靠性运行与维护数据指标进行分析，因此，这一算法模型的提出是具有现实意义与价值的，为通信网络管理者提供了一个分析评价通信网络可靠性的理论依据和实用方法，促进了通信网络可靠性管理工作的完善和提高，推动通信网络可靠性管理理论的研究与完善。随着数据挖掘技术的发展与成熟、完善，对于数据信息突变时的分析与挖掘越来越有着重要的意义。

参 考 文 献

1 熊蔚明，刘有恒. 关于通信网可靠性的研究进展. 通信学报，1990，11(4)：43～49
2 刘有恒. 通信网可靠性研究的新进展. 通信学报，1990，11(4)：43～49
3 梁雄健. 国家自然科学基金申请报告-通信网可靠性的评价理论、方法和管理(NO. 79870069). 1999
4 梁雄健. 电信网可靠性及突发故障管理. 电信软科学研究，2000，10(10)：5～13
5 Zhou Zhongding, Liang Xiongjian. A Synthetic Evaluation Methodology Based on Neural Networks For Reliability Indexes of Communication Networks. In: Proc. of ICII'2001. Beijing, China
6 Oppenheim A V, Schafer R W. Discrete-Time Signal Processing. Prentice-Hall, Inc. 1998

(上接第81页)

结束语 在考察了遗传算法后，本文提出了基于多父体杂交和按排序的优劣对较差个体进行非均匀变异的改进策略的新算法。通过对5个算例测试并与数学计算软件 MATLAB 计算的结果相比，新算法计算结果得到最优解或优于用 MATLAB 得到的最好解；其余2个算例优于文[6～8]中计算的结果。计算结果表明：新算法具有通用，精确，稳健，高效等特点，是求解函数优化问题的一种较好的算法。

参 考 文 献

1 Michalewicz Z. Genetic Algorithm + Data Structures = Evolutionary Programms[M]. Berlin: Springer-Verlag, 1992
2 Michalewicz Z, Schoenauer M. Evolutionary algorithms for constrained parameter optimization problems [J]. Evolutionary computation, 1996, 4(1): 1～32
3 刘勇，康立山，陈毓屏著. 非数值并行算法——遗传算法[M]. 北京：科学出版社，1995
4 蒲俊，吉家锋，伊良忠编著. MATLAB6.0数学手册[M]. 上海：浦东电子出版社，2002. 176～200
5 杜藏，骆源编著. 科学计算语言 MATLAB 简明教程[M]. 天津：南开大学出版社，1998. 20～33
6 [美]z. 米凯利维茨著. 演化程序——遗传算法和数据编码[M]. 周家驹，何险峰译. 北京：科学出版社，2000. 24～33
7 [日]玄光男，程伟著. 遗传算法与程序设计[M]. 汪定伟，唐加福，黄敏译. 北京：科学出版社，2000. 5～11，100～200
8 何新贵，梁久祯. 利用目标函数梯度的遗伟算法[J]. 软件学报，2001，12(7)：981～986