

Internet 访问时间的混沌性解析^{*})

李超 赵海 张昕 袁韶谦

(东北大学信息科学与工程学院 沈阳 110004)

摘要 以 CAIDA 组织授权的海量数据为样本空间,利用相空间重构技术以及 G-P 算法,对时间维的 Internet 访问时间进行相空间重构,得到其混沌吸引子饱和关联维数为 2.8308,证实 Internet 访问时间的演化过程具有混沌特征。在此基础上,根据关联维数及奇怪吸引子在相空间混沌轨道运动的特性,提出了三维微分方程组的 Internet 访问时间长期预测模型。

关键词 Internet 访问时间,时间序列,相空间重构,饱和关联维数,混沌吸引子

Chaotic Analysis of Internet Traveling Time

LI Chao ZHAO Hai ZHANG Xin YUAN Shao-qian

(School of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

Abstract As the giant data samples authorized by CAIDA to be the sample space, the saturated correlative dimension of chaotic attractor of Internet traveling time with phase space reconstruction and G-P algorithm are calculated about 2.8308. It approves that the evolvement of Internet traveling time exists characteristic of chaos. Based on the saturated correlative dimension and strange attractor movement near the path of the phase space, a model of three dimension differential equation of Internet traveling time is put forward, which is suitable for the long-term forecast.

Keywords Internet traveling time, Time series, Phase space reconstruction, Saturated correlative dimension, Chaotic attractor

1 引言

与传统的人工系统相比,Internet 更像一个生态系统^[1]。在 Internet 复杂表象下,自发地涌现出某种秩序以及产生这种秩序的机制。由于 Internet 缺少中央控制,其网络特征量,如访问直径^[2]、网络分割比、访问时间的变化都具有较强的随意性,导致其看起来似乎无任何规律。但是非线性动力学研究表明,一些貌似随机的过程是非线性确定系统内在随机性的表现^[3]。混沌是非线性耗散系统中产生的一种貌似随机的不规则现象。因此,尝试从混沌动力学的角度研究 Internet 网络特征量不失为一个可行的方法。

由于从 Internet 上获得一组在时间维度上的观测值十分方便,因此对于如何用一维时间序列观测值来准确描述 Internet 是一个很实际的问题。本文根据 CAIDA¹ 提供的海量样本数据,对基于时间维度的 Internet 访问时间进行相空间重构,并建立相应的状态微分方程,以此对 Internet 的混沌特征进行初步分析。

2 Internet 访问时间的相空间重构

2.1 相空间重构

混沌是自然界中一种普遍现象,它具有对初始值敏感、不可预测、周期点稠密等特点^[3]。其中,混沌系统的辨识是研究混沌的首要基础。但是对于大多数非线性系统而言,描述其状态变化的微分方程是未知的,实际中往往只能测得该系统某一特征量的时间序列。

从时间序列研究混沌性始于 Parkcard 提出的相空间重构理论。由于系统任一变量的时间演化行为均包含了系统所有变量长期演化信息,因此可通过决定系统长期演化的任一单变量时间序列来研究系统的混沌特征。对该时间序列进行重构相空间就可以在拓扑等价意义上与原系统保持一致。

对系统特征量的一维时间序列 $x_1, x_2, \dots, x_n$, 采用延迟坐标法,选定嵌入维数 m ,延迟时间 τ ,重构相空间如式(1):

$$[Y(t_1), Y(t_2) \dots Y(t_{n-(m-1)\tau})] = \begin{pmatrix} x_1 & x_2 & \dots & x_{n-(m-1)\tau} \\ x_{1+\tau} & x_{2+\tau} & \dots & x_{n-(m-2)\tau} \\ \vdots & \dots & \dots & \vdots \\ x_{1+(m-1)\tau} & x_{2+(m-1)\tau} & \dots & x_n \end{pmatrix} \quad (1)$$

式(1)中, n 表示原始样本空间容量,重构相空间中共有 $N=n-(m-1)\tau$ 个矢量点,且每个矢量有 m 个分量。只要延迟时间 τ 和嵌入维数 m 选取的合适,就可以在相空间中恢复出系统的混沌吸引子。

2.2 Internet 访问时间的时序序列样本

Internet 访问时间表示在对 Internet 探测的过程中,从源监测节点发送探测数据包开始到接收目的端发送回响应的时间之差。由于 Internet 访问时间可以作为衡量整个网络延迟的指标,可以在一定程度上反映源端到目的端网络的动力学特性^[5],因此本文以 Internet 访问时间对 Internet 的混沌特征进行研究。

通过 CAIDA skitter^[6] 项目授权的海量样本数据,对 Internet 访问时间分布做直观统计分析。由于网络的复杂性

^{*}) 基金项目: 国家自然科学基金资助项目(69873007); 国家级火炬计划项目(2002EB010154)。李超 博士生; 赵海 教授, 博士生导师。

1. CAIDA (The Cooperative Association for Internet Data Analysis), 是一个对全球范围内 Internet 结构及数据进行研究的国际合作机构, 其主要目的是在全球范围合作基础上最大限度地对 Internet 数据进行获取、分析、研究和共享。东北大学嵌入式技术实验室已经成为 CAIDA 在中国授权建立的第一监测节点。

(如网络繁忙或目的端设置防火墙等),可能导致 skitter 探测网络目的端时不可达^[7],因此只选取目的端可达的有效数据作为样本进行分析。从时间维度上选取 CAIDA 的 riesling 监测节点于 2003 年 4 月到 2007 年 3 月共计 48 个月²对 Internet 进行探测的有效样本共计近 1700 万条,为不失一般性,本文仅对每隔 6 个月的 Internet 访问时间分布作图,结果如图 1 所示。

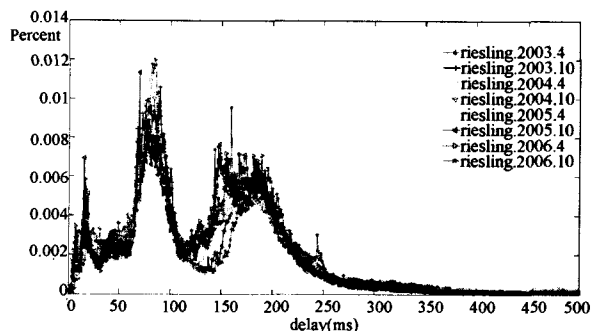


图 1 时间维度上 Internet 的访问时间分布

从图 1 可以看出, Internet 访问时间在时间维度上呈现出了较为统一的多峰重尾分布。在充分保证较高冗余度的条件下,采样得到的数据其分布具有较强的规律性,说明了 skitter 探测样本数据的有效性,可以从时间维度上对 Internet 访问时间进行分析。

对每个月的 Internet 访问时间取统计平均以反映整个网络的延迟,其变化如图 2 所示。从图 2 可以看到, Internet 平均访问时间主要集中在 [160ms, 250ms] 范围内,且在 200ms 左右 Internet 平均访问时间呈现出一定的振荡性。本文根据该时间序列进行相空间重构,以此恢复出 Internet 访问时间的混沌吸引子。

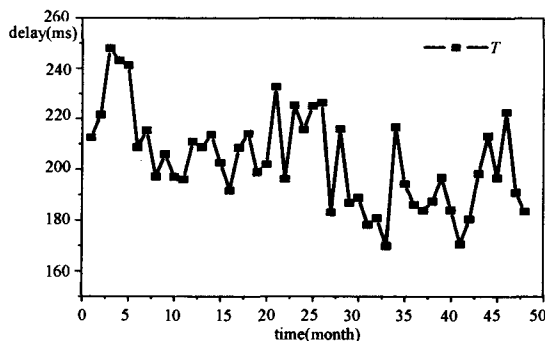


图 2 Internet 访问时间随时间(2003 年 4 月到 2007 年 3 月共 48 个月)变化的情况

3 数值分析

3.1 滞后时延 τ 的确定

根据式(1),在重构 Internet 访问时间的相空间之前,首先需要确定滞后延迟时间 τ ,以保证相空间中的矢量点 $Y(t_i), Y(t_j)$ 是不相关的^[8]。对于 Internet 平均访问时间的的时间序列 $x_1, x_2, \dots, x_n$,其时间跨度为 k 的自相关函数如式(2):

$$R_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{X})(x_{i+k} - \bar{X})}{\sum_{i=1}^{n-k} (x_i - \bar{X})^2} \quad (2)$$

R_k 表示第 k 阶自相关函数,根据上式得出 Internet 平均访问时间的自相关函数,其随 k 变化的情况如图 3。

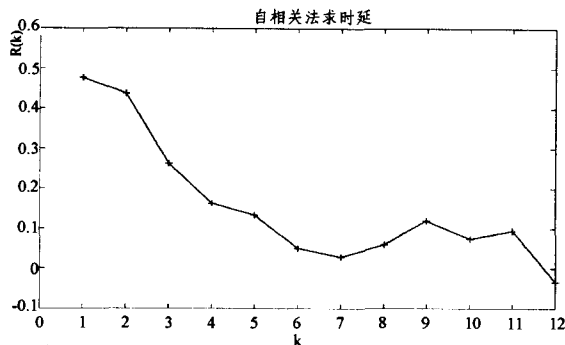


图 3 Internet 访问时间的自相关函数 R_k 随 k 变化的情况

一般认为,自相关函数下降到初始值的 $1 - \frac{1}{e}$ 时,所得到的时间就是重构相空间的最佳延迟时间 τ ^[8]。根据图 3,得出 Internet 平均访问时间的自相关函数下降到初始值的 $1 - \frac{1}{e}$ 时,最佳延迟时间 $\tau = 3$ 。因此可以说明,第 1 个月的平均访问时间对第 3 个月的平均访问时间没有影响,也就是说到第 3 个月的时候,第 1 个月的网络访问时间信息已经基本消失,可以保证相空间中矢量点的相关性为最小。

3.2 饱和关联维数

根据重建复杂系统动力学原理,定义重构相空间的关联积分如下^[9]:

$$C_m(r) = \frac{\sum_{i,j=1}^N \theta(r - R(i,j))}{C_N^m} \quad (3)$$

其中, θ 为 Heaviside 函数:

$$\theta(z) = \begin{cases} 1 & (z \geq 0) \\ 0 & (z < 0) \end{cases} \quad (4)$$

$R(i,j)$ 表示相空间矢量点 $Y(t_i), Y(t_j)$ 之间距离,因此,关联积分表示相空间矢量点之间距离小于 r 的比例。只要 N 足够大, r 足够小时,关联积分 $C_m(r)$ 就满足如下关系:

$$C_m(r) = A \times r^{D(m)} \quad (5)$$

对两边取对数,得到:

$$\ln C_m(r) = \ln A + D(m) \ln r \quad (6)$$

如果 $D(m)$ 随着 m 不断增大而保持收敛,则系统是混沌的,否则,所考虑的系统是一个随机时间序列。此时所对应的 m 为最小嵌入维数, $D(m)$ 为饱和关联维数。

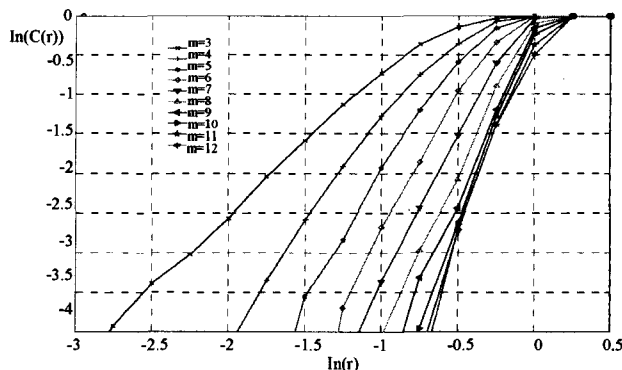


图 4 Internet 访问时间的关联积分图

(下转第 130 页)

兵 杭州 开始 十分 喜欢 应该 英雄 乘务 东西 码头 空中小姐 高兴 明白 乘务队 希望 几乎 简直 民航 情况 确实 已经 变化 城市 吃饭 而且 九溪 军舰 全部 王眉 影响 报纸 充满 房间 关系 航班 建设 精神 蓝色 泪水 妈妈 重新 乘务员 第二天 开玩笑 老百姓 一架飞机 容易 虽然 写信 信任 衣服 意思

上面的结果中,带下划线的字串是小说中的人名,带波浪线的是地名,带方框的字串“一架飞机”是错误的。与以往的方法相比,本文方法抽取的频繁字串数量和质量都有一定的提高。

结束语 本文吸取了数据挖掘中 Apriori 关联规则挖掘的思想,提出了逐层扫描算法和有效支持数统计方式。与以往算法相比,该算法具有更快的速度和更高的准确性,可广泛应用于自动标引、分类、信息检索、摘要等各种信息处理领域。

参考文献

- [1] 王 永成. 中文信息处理技术及其基础[M]. 上海: 上海交通大学出版社, 1991
- [2] Tan Hong-ye. Research on Method of Automatic Recognition of Chinese Place Name based on Transformation[J]. Journal of Software, 2001, 12(11): 1608-1613
- [3] Nie Jian-Yun. Unknown Word Detection and Segmentation of

Chinese using Statistical and Heuristic Knowledge. Communications of COLIPS, 5(1&2): 47-57

- [4] Ling G C, Asahara M, Matsumoto Y. Chinese Unknown Word Identification Using Character-based Tagging and Chunking. In Companion Volume to the Proceedings of ACL 2003, Interactive. Poster/Demo Sessions, 197-200
- [5] 崔世起, 刘群, 孟遥, 等. 基于大规模语料库的新词检测[J]. 计算机研究与发展, 2006, 43(5): 927-932
- [6] 黄萱菁, 吴立德, 王文欣, 等. 基于机器学习的无需人工编制词典的切词系统[J]. 模式识别与人工智能, 1996, 9(4): 297-303
- [7] 罗盛芬, 孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报, 2003, 17(3): 9-14
- [8] 刘挺, 吴岩, 王开铸. 串频统计和词形匹配相结合的汉语自动分词系统[J]. 中文信息学报, 1998, 12(01): 17-25
- [9] 任禾, 曾隽芳. 一种基于信息熵的中文高频词抽取算法[J]. 中文信息学报, 2006, 20(5): 40-90
- [10] 韩 客松, 王永成, 陈桂林. 无词典高频字串快速提取和统计算法研究[J]. 中文信息学报, 2001, 15(02): 23-30
- [11] 姜韶华, 党延忠. 基于长度递减与串频统计的文本切分算法[J]. 情报学报, 2006, 25(1): 74-79
- [12] 金翔宇, 孙正兴, 张福炎. 一种非受限中文文档自动抽词方法[J]. 中文信息学报, 2001, 15(6): 33-39
- [13] 马颖华, 王永成, 苏贵洋. 一种在汉语文本中抽取重复字串的快速算法[J]. 电子学报, 2002, 12(12): 2177-2179
- [14] 刘华. 一种快速获取领域新词语的新方法[J]. 情报学报, 2006, 20(5): 17-23
- [15] 彭波. 搜索引擎的混合索引技术[J]. 计算机工程与应用, 2004, 22: 16-18

(上接第 86 页)

在确定了最佳延迟时间后,可以根据式(1)将 Internet 平均访问时间的时间序列构造造成 m 维的相空间,通过 G-P 算法来确定 Internet 访问时间的关联维数。在滞后时延 $\tau=3$ 下, m 由 3 逐渐递增至 12 时 $\ln(C_m(r))-\ln(r)$ 变化的曲线,如图 4 所示。

从图 4 看到,随着嵌入维数 m 增大, $\ln(C(r))$ 随 $\ln(r)$ 变化的曲线斜率呈逐渐收敛趋势,并且图 4 中从左到右的第 4 条曲线开始,各曲线的线性部分平行性较好,且 m 增大到 9 之后的各曲线几乎重叠,即 $D(m)$ 呈收敛趋势,故认为 $m=9$ 是 Internet 访问时间重构空间的最小嵌入维数。

对 $m=9$ 时 $\ln(C(r))-\ln(r)$ 曲线进行线性回归,得到 Internet 平均访问时间的饱和关联维数为:

$$D(m=9, \tau=3) = 2.8308 \quad (7)$$

饱和关联维数存在且为分数,故认为 Internet 访问时间的演化过程具有混沌特征,系统表现出对初始条件敏感的混沌振荡。

4 Internet 访问时间的预测模型

一般认为,某条路径的访问时间受多方面因素的影响,其中主要包括路由节点处理数据包的能力,路由节点所承受的负载,链路带宽以及数据包大小等^[10]。文献[3]指出,对关联维数为 D 的混沌系统,必须至少使用 $\lceil D \rceil$ 维模型才可以表征影响目标系统 D 个重要的物理过程。

根据 Internet 访问时间的关联维数为 2.8308,说明至少需要使用 3 个物理特征量来表征 Internet 访问时间的混沌过程。在由多因素引起的 Internet 访问时间变化中,关联维数的确定为建立 Internet 访问时间的理论模型提供了变量个数的标准,而不必将各种有相关联系的因素都考虑在模型方程之中。因此,对 Internet 访问时间的模型建立如下:

$$\begin{cases} x_1 = f_1(x_1, x_2, x_3, x_1^{(n)}, x_2^{(n)}, x_3^{(n)}) \\ x_2 = f_2(x_1, x_2, x_3, x_1^{(n)}, x_2^{(n)}, x_3^{(n)}) \\ x_3 = f_3(x_1, x_2, x_3, x_1^{(n)}, x_2^{(n)}, x_3^{(n)}) \end{cases} \quad (8)$$

式(8)表征了 Internet 访问时间在混沌轨道运动上的数学模型,其中 $x_1 \sim x_3$ 为表征 Internet 访问时间的重要物理量, $x_1^{(n)}$ 为该变量的 n 阶导数,该三维状态微分方程可以作为 Internet 访问时间的预测模型。

由于混沌轨道运动形式十分复杂,考虑到目前 Internet 访问时间的时间序列跨度较小,该样本数据集还不能确切地描述式(8)中方程的具体形式。确定 Internet 访问时间状态微分方程中的具体参数将是本文的下一步工作。

结束语 以 CAIDA skitter 采样得到的 Internet 访问为样本空间,对其进行延迟坐标相空间重构,得到 Internet 访问时间的最佳滞后时间为 3,以此说明了信息损失的时间尺度范围。通过 G-P 算法,得到 Internet 访问时间的相空间最小嵌入维数为 9,饱和关联维数为 2.8308,关联维数存在且为分数,说明了 Internet 访问时间的演化过程具有混沌特征,并以此建立了 Internet 访问时间三维微分方程在混沌运动轨道上的数学模型。该工作对于以混沌动力学来分析 Internet 网络特征量并在此基础上进一步预测及控制网络性能开辟了一条新路。

参考文献

- [1] 张家才, 周登勇. 从开放的复杂巨系统来看 Internet 中的大范围模式[J]. 系统仿真学报, 2002, 14(11): 1450-1454
- [2] 徐野, 赵海. Internet 网络的访问直径分析[J]. 计算机学报, 2006, 29(5): 690-698
- [3] 黄润生. 混沌及其应用[M]. 武汉: 武汉大学出版社, 2000: 183-246
- [4] Takens F. Lecture notes in mathematics, 1981, 898: 366
- [5] Kamrani E, Hamid H R. Modeling Internet delay dynamics for teleoperation//Proceedings of the 2005 IEEE Conference on Control Applications, 2005: 1528-1533
- [6] CAIDA skitter Project. <http://www.caida.org>
- [7] 张宏莉, 方滨兴. Internet 测量与分析综述. 软件学报, 2003, 14(1): 110-116
- [8] 吕金虎, 陆君安. 混沌时间序列分析及其应用[M]. 武汉: 武汉大学出版社, 2002: 60-63
- [9] 李后强, 汪富泉. 分形理论及其在分子科学中的应用. 北京: 科学出版社, 1993: 190-195
- [10] Han K H, Kim S K Y J, Kim Y J. Internet control architecture for Internet-based personal robot. Autonomous Robots, 2001, 10: 135-147