互联网传播行为的时序演化与预测

田鹤[1]，赵海[2]，王进法[2]，林川[2]
（1. 辽宁科技学院工程实践中心，辽宁 本溪 117004；2. 东北大学计算机科学与工程学院，辽宁 沈阳 110004）

摘 要：互联网的传播行为对研究网络拓扑结构和动态行为的关系具有重要作用。选取 CAIDA_Ark 项目下不同地区 4 个监测点的有效路径样本数据，统计网络访问时间与访问直径，发现它们的相关性极弱，网络访问时间呈多峰重尾分布。采用非线性时间序列分析方法对网络访问时间演化序列混沌辨析，结果表明其时序演化具有混沌特征。在此基础上，引入 Logistic 方程建立网络传播行为预测模型，并用粒子群优化算法对模型参数取优，用 4 个监测点的网络访问时间序列对模型进行实验，从准确性和可用性这 2 个方面对模型进行评价，结果表明，短期内该模型能够对网络传播行为做出准确预测，在一段时期内，可作为网络行为演化预测的工具。

关键词：互联网传播；网络访问时间；Logistic 模型；混沌特征；行为预测
中图分类号：TP393.6
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018096

Timing evolution and prediction of Internet transmission behavior

TIAN He[1], ZHAO Hai[2], WANG Jinfa[2], LIN Chuan[2]
1. Engineering Practice Center, Liaoning Institute of Science and Technology, Benxi 117004, China
2. School of Computer Science and Engineering, Northeastern University, Shenyang 110004, China

Abstract: The transmission behavior of Internet plays an importance role in the research on the relationship between network topology structure and dynamic behavior. Selecting effective path samples in four monitoring points which from different regions authorized by CAIDA_Ark project and statistics network traveling time and traveling diameter, their correlation is very weak, network traveling time is presented on multi-peak and heavy tail distribution. Using nonlinear time sequences analysis method to identify the Chaos characteristics of network traveling time evolution sequences. The results show that their timing evolution has Chaos characteristics. Based on this, the Logistic equation was lead to establish network transmission behavior prediction model, and particle swarm optimization (PSO) was used to optimize model parameters. The model by the network traveling time sequences of four monitoring points was experimented, evaluated it from accuracy and availability, the results show that the model can predict network transmission behavior accurately in the short term. It can be used as a tool for predicting the network behaviors' evolution in a period of time.

Key words: Internet transmission, network traveling time, Logistic model, Chaos characteristics, behavior prediction

1 引言

互联网作为一种典型的复杂网络，其宏观拓扑结构表现出明显的复杂网络特性[1-3]。从网络病毒传播以及网络拓扑结构的统计特征和演化等方面研究互联网的性能、结构和发展趋势已取得了丰硕的成果[4-6]。随着互联网的飞速发展，人们对网络的需求不断增加，使网络的应用得到广泛的扩展。互联网已经发展成为一个复杂的、非线性的系统。然而，网络的大规模扩张使网络安全性、资源调度与优化以及服务质量等方面也面临着巨大挑战，仅从网络宏观拓扑静态特征指标的度量、统计和建模不足以形象地描述网络的传播行为。为提高网络各方面的性能，人们需要对网络行为的特征规律有深刻的认

收稿日期：2017-08-08；**修回日期**：2018-04-02
基金项目：国家自然科学基金资助项目（No.60973022）
Foundation Item: The National Natural Science Foundation of China (No.60973022)

识，发现网络行为的内在机制是认识网络的必然过程。网络的传播行为包含了多种网络动态行为[7]，如链路访问、分组时延以及路由转发等，反映了网络拓扑结构对动态行为的影响。掌握网络传播行为的特征和规律有助于对网络的异常行为做出分析与评估，为防范网络攻击和病毒传播提供预警手段，在一定程度上可控制和预测网络动态行为的发生。

网络的传播行为特征规律可通过定义一些能够反映网络行为的特征指标来描述，然后实时监测网络，从中获取这些特征指标的样本数据，并对测量结果进行整理、统计、归纳和推断，透过指标的变化和性质对网络行为的各方面表现进行解释。研究初期，徐野等[8]定义并分析了访问直径网络物理特征量，利用其演化特征分析网络涨落现象。然而，互联网的规模呈指数级增长态势，导致仅从时间维度并不足以分析网络动态行为。结合非线性动力学和混沌理论对网络传播特征量进行统计，进而分析互联网传播行为特征，为互联网的演化分析开辟了新思路。隋岩等[9]从混沌学角度分析互联网群体传播特性，表明互联网群体传播就是一种混沌系统，具有非线性秩序性和自组织性等特征。Ye 等[10]基于 ARIMA 和 Holt-Winters，用多元时间序列方法建立长期预测模型，但只适用于网络静态预测，动态性较差。Chai 等[11]将时延坐标嵌入方法和混沌分析方法应用于神经网络构建预测模型，测试结果表明混沌方法可以显著提高预测能力。基于以上研究背景，本文以时间为主线，统计和筛选 CAIDA_Ark 项目下位于不同大洲的 4 个监测点的有效路径样本数据，对互联网传播行为进行统计和分析，利用非线性时间序列分析方法对网络访问时间序列的时序演化特征进行混沌辨识。在此基础上，引入 Logistic 方程建立以混沌网络访问时间序列为基础的网络传播行为预测模型，采用粒子群优化算法对模型参数取优。最后，分别将 4 个监测点的网络访问时间序列在预测模型上进行实验与验证，对预测模型的有效性和准确性做出评价。

2　数据与推演指标

CAIDA 是一个对互联网的网络结构和数据进行获取、测量、可视化以及分析的国际合作研究机构。2007 年 9 月，CAIDA 开展 Ark 探测项目计划，将原有的 Skitter 探测架构升级，采用 Scamper 技术、traceroute 主动探测方式和元组空间实现各监测点间的探测和通信。本文选取 CAIDA_Ark 项目下 4 个位于不同大洲的监测点 amw、san、bcn 和 mnl，利用 Scamper 技术对网络中随机抽取的目的 IP 地址发送 ICMP 探测数据分组，同时 traceroute 检查 ICMP 的 *Echo_request* 分组头部的 *TTL* 值的有效性，追踪路由的地址路径。其中，每一个监测点在同一探测周期内只能探测到一个 IP 地址。

探测源 *SRC* 向目的端 *DST* 发送探测数据分组，经过中转路由器 $R_1, R_2, \cdots, R_n$，则探测数据分组从 *SRC* 到 *DST* 所经过的路径表示为 $R=(SRC, R_1, R_2, \cdots, R_{n-1}, DST)$。探测数据分组的路由选择与转发都是由各中间路由器决定的，然而，由于监测点内的分组发送设置、中间路由器的个体差异性以及受 ICMP 分组接收率的限制等原因，网络中并不是所有的中转路由器都能对 ICMP 分组做出响应，所以探测数据分组在某些中转路由器处有可能不可达。虽然 *SRC* 未收到中转路由器的响应，但 Scamper 探测技术仍能够增加 *TTL* 值直至探测到 *DST*，而这时生成的 *R* 是不完整的。若所经过的中转路由器都能对探测数据分组做出响应并返回到 *SRC*，则所生成的 *R* 是一个完整的有效路径。IP 级拓扑是抽取网络 IP 接口和链路而成的，数据包含了网络拓扑最原始的内容，数据量非常庞大，忽略不可达路径，提取完整的有效路径上的样本数据。选取 2012-2015 年共 48 个月 IPv4 互联网 IP 级拓扑数据，以 3~4 天为一个探测周期，每个月选取一个探测周期的结果并提取有效路径样本，统计结果如表 1 所示。

表 1　　有效路径样本数的统计结果

监测点	2012 年/条	2013 年/条	2014 年/条	2015 年/条	总计/条
amw	53 万	38 万	57 万	64 万	212 万
san	67 万	36 万	63 万	69 万	235 万
bcn	67 万	42 万	51 万	69 万	229 万
mnl	68 万	25 万	48 万	65 万	206 万

从表 1 可知，提取的有效路径样本数达 900 多万条。高冗余数据更有利于网络传播行为的研究分析。本文从时间维度上对网络传播行为进行统计，相关定义如下。

定义 1　访问时间[12]。在网络中，将监测点发送探测数据分组的时间与收到目的端返回响应时间之差定义为该条路径的一次访问时间，记为 $T_d(t)$。

定义 2　网络访问时间。大量探测数据分组从

任一源 IP 地址到任一目的 IP 地址所经过的有效路径的访问时间均值，记为

$$T(t)=\frac{1}{n}\sum_{d=1}^{n}T_{\mathrm{d}}(t) \tag{1}$$

其中，n 为数据样本总数。

定义 3 访问直径[8]。在一个完整的有效路径中，探测数据分组所经过的路由跳数。

定义 4 Pearson 相关系数。用来衡量定距变量间的线性关系，计算式为

$$r=\frac{\sum\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum\left(X_i-\overline{X}\right)^2}\sqrt{\sum\left(Y_i-\overline{Y}\right)^2}} \tag{2}$$

3 网络传播行为特征

3.1 网络访问时间的分布特征

首先，提取 4 个监测点 amw、san、bcn 和 mnl 探测得到的 2012-2015 年每月同一周期的有效路径样本数据。然后，分别对 4 个监测点的有效路径数据的访问时间做概率分布统计，结果如图 1 所示。

从图 1 可直观看出，4 个监测点在探测期间有效路径中网络访问时间分布至少有 2 个峰值，且它们的尾部几乎重合，呈多峰重尾分布[13]，并且在这 4 年内，同一个监测点的网络访问时间的分布具有较强的相似性，这是由于不同的监测点所处的不同地理位置影响了对目的端的访问。对于网络的动态传播行为，网络的访问时间直接影响有效路径上网络端到端的连接行为，例如，对互联网传输协议中重传超时时间（RTO）的设置，若 *RTO* 值过小则会加重网络不必要的负载，若 *RTO* 值过大则会浪费网络带宽。此外，各监测点的高冗余数据在探测有效路径中所呈现的网络访问时间的相似分布特征说明互联网具有自相似特性。

在网络动态传播演化的过程中，网络访问时间越短，数据分组在一定有效路径传输距离的时延越小，网络的传播性能和效率越高。观察图 1 可以进一步发现，网络访问时间大于 400 ms 的有效路径只占很小的比例，这说明虽然各监测点所处地理位置相距甚远，但它们的通信效率仍然很高，探测期间内整个网络的性能非常好。随着互联网的飞速发展，依靠互联网来实现跨国家、跨海洋以及跨大洲通信已不是问题。由于峰值附近的有效路径样本数据比例较大，具有代表性，本文截取各监测点的网络访问时间分布在峰值附近较密集的有效路径样本数据做统计分析。amw、san、bcn 和 mnl

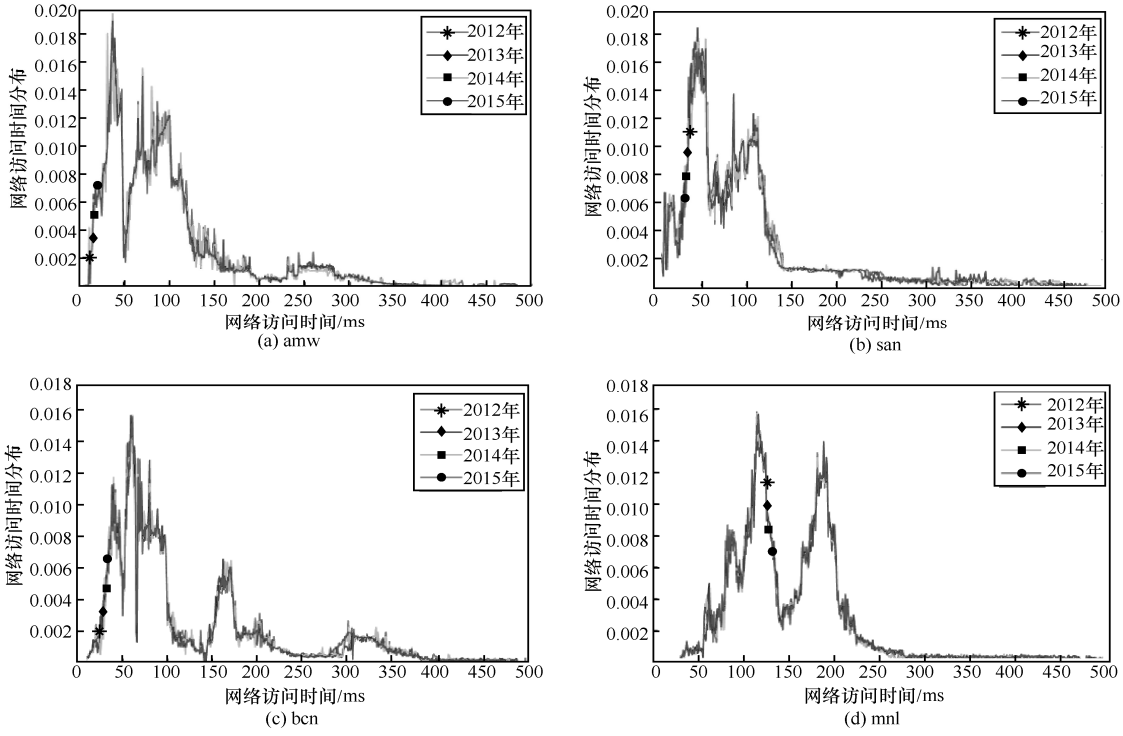


图 1 有效路径网络访问时间分布

这 4 个监测点的网络访问时间主要集中的区间如表 2 所示。

表 2　　截取的网络访问时间区间

监测点	区间 1	区间 2
amw	[10,50]	[60,120]
san	[30,60]	[90,130]
bcn	[15,55]	[55,115]
mnl	[110,140]	[160,220]

3.2　相关性分析

对于 IP 级拓扑，探测数据分组从探测源 *SRC* 到目的端 *DST* 以动态选路的方式得到的有效路径是 IP 级路径，其中，每个中转路由器 $R_1,R_2,\cdots,R_n$ 对应的 IP 地址为 $IP_1,IP_2,\cdots,IP_n$，中转路由器的个数即跳数。访问直径是网络拓扑传输效率的度量指标，反映了网络宏观拓扑结构对网络动态传播行为特征的影响。图 2 为 4 个监测点在探测有效路径中网络访问直径的累积分布。

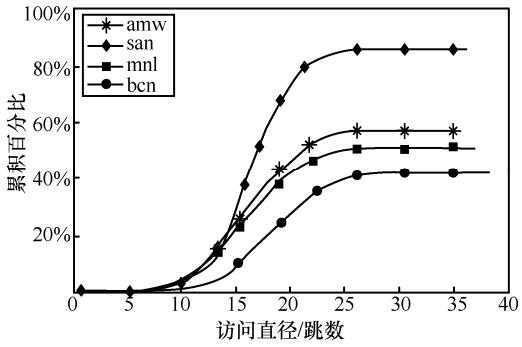
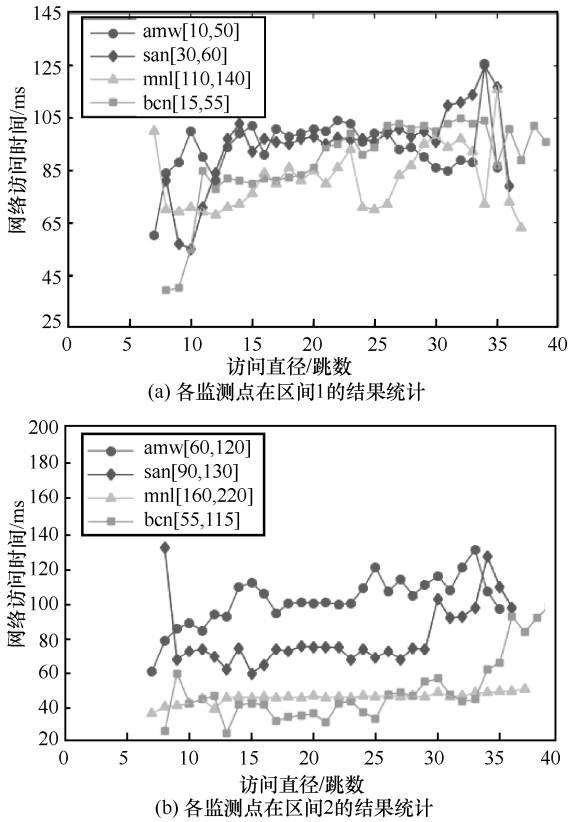


图 2　有效路径中网络访问直径的累积分布

从图 2 可以看到，只有不到 10%的有效路径的访问直径在 15 跳以下，说明数据分组从有效路径的源 IP 地址到目的 IP 地址传输一般要经过较多的中转路由器。分别来看，amw 监测点探测的有效路径访问直径有 90%超过 12 跳，bcn 监测点是 15 跳，而 san 和 mnl 监测点是 13 跳。选取 4 个监测点探测的具有代表性的访问直径区间内有效路径样本数据，amw、bcn、san 和 mnl 监测点的有效访问直径区间分别为 7~35 跳、8~39 跳、8~36 跳和 7~37 跳，相应的平均访问直径分别为 14 跳、16 跳、14 跳和 15 跳。统计不同访问直径的有效路径的网络访问时间，在 4 个监测点的有效路径的网络访问时间分布的峰值范围内（如表 2 所列的每个监测点截取的 2 个区间），结果如图 3 所示。



(a) 各监测点在区间1的结果统计

(b) 各监测点在区间2的结果统计

图 3　访问直径对网络访问时间的影响

从整体趋势上看，区间 1 和区间 2 中有效路径的访问直径与网络访问时间的变化趋势是振荡上升的，也就是说，随着访问直径的增大，网络访问时间也随之增大，说明网络拓扑中有效路径的访问直径越大，数据分组经过的中转路由器越多，所需的网络访问时间就越长。进一步观察图 3，对于区间 1，如图 3(a)所示，san 和 mnl 监测点的网络访问时间开始时急剧下降，然后大幅上升，而 amw 和 bcn 监测点则相反；amw、bcn 和 san 监测点的访问直径在 13~33 跳时，网络访问时间变化的振荡幅度较小，而 mnl 监测点的网络访问时间的平缓变化区间相对较小，主要集中在 9~23 跳。对于区间 2，如图 3(b)所示，4 个监测点的网络访问时间随着访问直径的增大并没有大幅的振荡，并且也没有出现相对平缓的变化。相较于其他 3 个监测点，mnl 监测点的网络访问时间随着访问直径的增大而小幅增大，尽管出现微小的波动，但总体趋势是平缓的。

为了深入分析访问直径与网络访问时间的关系，截取 4 个监测点的 2 个峰值范围内有效路径的网络访问时间与访问直径进行量化统计，如表 3 所示。

表 3　　有效路径的网络访问时间与访问直径的统计结果分析

监测点的访问时间区间	访问时间的均值/ms	访问时间的标准差	访问时间的中位数/ms	访问直径的均值/跳	访问直径的标准差	访问直径的中位数	Pearson 相关系数
amw[10,50]	30.592 2	9.314 9	29.941 2	14.498	3.569 8	14	0.176
amw[60,120]	88.530 2	15.489 9	88.011 2	15.773	3.854	16	0.132
san[30,60]	44.829 7	8.544 2	45.300 7	15.492	3.607 7	15	0.201
san[90,130]	102.622 8	10.071 9	102.329 1	16.788	3.525 6	17	0.068
mnl[110,140]	127.397 3	9.680 3	126.675	16.394	3.444 7	16	0.091
mnl[160,220]	182.339 2	14.914 1	181.505	15.483	3.442 6	15	0.098
bcn[15,55]	35.677	8.062 1	35.311 4	15.228	2.705 2	15	0.164
bcn[55,115]	81.792	15.138 2	78.991 4	17.448	3.343 7	17	0.092

由表 3 可得，4 个监测点提取的网络访问时间区间内访问时间的均值与中位数相差不大，访问直径的均值与中位数也很相近，说明区间内均为有效样本。从 Pearson 相关系数值看到，最大值只有 0.201，最小值是 0.068，所以可以认为访问直径与访问时间是不相关的。进一步分析 4 个监测点的访问直径与访问时间 Pearson 相关系数随着时间的演化趋势，如图 4 所示。以月为单元，在 2012-2015 年共 48 个月 Pearson 相关系数值的演化范围集中在 0.05~0.25，期间并没有呈现增大或减小的变化趋势，而是随着时间振荡演化，出现的最大值也不超过 0.35，说明访问直径与访问时间之间是一种极弱的关系，可以视为是不相关的。而图 3 所示的网络访问时间随着访问直径增大的可能的原因是对整体样本数据进行统计时，由于数据的高冗余性，有效路径的某一访问直径下个体样本数据中较大的访问时间样本数据没有体现出来，对总体样本数据的统计掩盖了个体之间的真实关系。

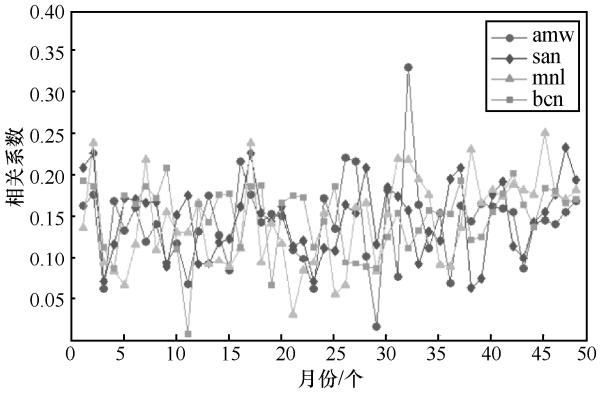


图 4　访问直径与访问时间 *Pearson* 相关系数演化趋势

然而，纵向观察表 3，对于同一监测点的不同访问时间区间，访问时间的均值和中位数相差很大，但访问直径的均值和中位数很相近，也就是说，在访问直径相差不大的情况下，访问时间却相差很大。数据分组从网络中某一源 IP 地址到任一目的 IP 地址传播过程中，由于链路吞吐量的差异以及数据传输过程中的分组丢失、链路消耗和时延等原因，造成网络拓扑中某一特定的访问直径下，大量数据分组的访问时间是不同的。在实际的网络传播过程中，访问直径对数据分组访问时间的影响并不大。

4　网络访问时间的演化

4.1　网络访问时间演化序列

无论是网络本身内部拓扑结构还是其外在状态表现，网络都时刻处于动态变化中，因此网络的行为也是随着时间不断变化的。选取 amw、san、bcn 和 mnl 监测点探测得到的 2012-2015 年共 48 个月的有效路径样本数据，根据定义 2，结合图 1 的有效路径中网络访问时间分布，对 4 个监测点的有效路径中网络访问时间演化序列的结果做统计，如图 5 所示。

从图 5 可以看到，网络访问时间整体的演化趋势是缓慢下降的，表明随着时间的推移，网络的有效性能不断增强，网络的传播效率不断提高。然而，这种下降趋势不会一直出现，数据分组在网络中传播会受到多方面因素的影响，例如，互联网服务提供商因为一些特殊情况对局部地区的网络结构进行调整，尤其是骨干网上的网络结构调整，引起网络绕路问题等原因使网络访问时间不会无限减小。

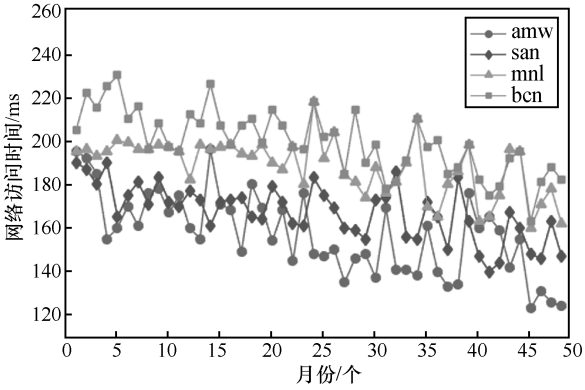


图 5　网络访问时间的演化

聚焦各趋势线上的波动点，4 个监测点在探测期间网络访问时间是振荡变化的，甚至在某些月份的振荡幅度较大，说明互联网在平稳演化的过程中时常伴有突变的发生，导致这一现象的原因可从主、客观因素这 2 个方面来分析：就互联网本身而言，在时间和空间的有限结合中，世界上每天都会有大量的节点接入互联网，同时也会有许多节点由于各种原因而消亡，互联网不断破坏自身系统的平衡，却又平稳地选择生成新的拓扑结构，互联网的演化使其内部各组织之间以及与其他外界社会环境之间不断地相互作用和影响，为了能够长期稳定地生存，演化必须进行物质、能量和信息代谢，而代谢活动势必会使自身的拓扑结构产生重组（自复制）与变异（突变）以适应变化的环境，在某一个较短的时间内不精确的自复制或是有误差的数据传输使参数大幅动荡，但是为了维持自身拓扑结构的稳定，动荡持续时间并不会太长；另一方面，在互联网演化过程中时刻存在着互联网异常事件，如 DDoS 攻击、僵尸网络、病毒传播和网络群体事件等，这些异常事件将直接带来网络中流量增加，引起局部路由上的吞吐量突然增加，进而使网络时延增加，网络访问时间会突然增加。另外，由于 CAIDA 多点探测方式的限制、路由配置问题以及随机噪声的干扰，会使网络通信路径和传输发生异常，也会使网络访问时间的演化产生振荡。由此来看，由于网络自身拓扑结构和通信链路的变化以及本地网络环境的影响，网络访问时间的演化并没有一些明显的特定规律，但从图 1 可知，有效路径中网络访问时间的分布特征又表现出一种规律性，说明互联网宏观拓扑结构带有一种序，具有自相似性，这种自相似性是由于互联网拓扑演化时不断进行自复制行为而产生的，尽管网络演化时常出现突变，但从总体来看，互联网仍是一个稳定的确定系统。从另一角度来说，新事物的产生通常由许多差异引起，突变是新信息的主要来源，所以，大幅异常的波动点并不是一种坏现象，它们是网络进化的动力，也是改造网络的有利时机。

4.2　混沌特征

根据非线性动力学理论，一些看似无规则的随机行为实际上是一个真实的非线性确定系统内在随机性的表现，这种现象可用混沌运动来解释，即在非线性确定系统中不需要附加任何随机因素就能发生类随机行为[14]。因此，互联网访问时间的时序演化符合混沌运动特征，同时，互联网拓扑有序层次化的自相似性说明互联网具有分形特征。

以混沌理论来分析非线性时间序列的基础是相空间重构[15]，即把低维的时间序列重构成一个高维的相空间。重构的关键是确定 2 个参数，即时延 τ 和嵌入维数 m。

对于网络访问时间的 n 个一维时间序列 $x_1,\cdots,x_n$，采用时延坐标法[16]，重构相空间为

$$\begin{aligned}&\left[\boldsymbol{Y}(t_1),\boldsymbol{Y}(t_2),\cdots,\boldsymbol{Y}\left(t_{n-(m-1)\tau}\right)\right]\\&=\begin{bmatrix}x_1 & x_2 & \cdots & x_{n-(m-1)\tau}\\ x_{1+\tau} & x_{2+\tau} & \cdots & x_{n-(m-2)\tau}\\ \vdots & \vdots & \cdots & \vdots\\ x_{1+(m-1)\tau} & x_{2+(m-1)\tau} & \cdots & x_n\end{bmatrix}\end{aligned}\tag{3}$$

其中，重构相空间矢量长度为 $N=n-(m-1)\tau$。

1) 时延 τ

时延 τ 的选取应使重构相空间中各矢量相互独立，统计网络访问时间跨度为 k 的自相关系数，计算式为

$$R_k=\frac{\sum_{i=1}^{n-k}\left(x_i-\overline{X}\right)\left(x_{i+k}-\overline{X}\right)}{\sum_{i=1}^{n-k}\left(x_i-\overline{X}\right)^2}\tag{4}$$

当 R_k 下降至初始值的 $1-\dfrac{1}{\mathrm{e}}$ 时，即最佳时延 τ。

2) 嵌入维数 m

首先，定义相空间矢量间的关联积分为

$$C_m(r)=\frac{\sum_{i,j=1}^{N}H\left(r-R(i,j)\right)}{C_N^2}\tag{5}$$

其中，$H(X)$ 是 Heaviside 函数，当 $X<0$ 时，

$H(X)=0$；当 $X\geqslant 0$ 时，$H(X)=1$。$R(i,j)$ 是相空间矢量 $\boldsymbol{Y}(t_i)$ 和 $\boldsymbol{Y}(t_j)$ 间的距离，于是 $C_m(r)$ 则表示相空间矢量间距离小于 r 的比例。根据重建复杂系统动力学原理，当 r 足够小而 N 足够大时，$C_m(r)$ 与 $r^{D(m)}$ 成正比，即 $C_m(r)=Ar^{D(m)}$，$D(m)$ 就是混沌吸引子的关联维数，其值等于 $\ln C_m(r)$ 与 $\ln r$ 的斜率，则有 $\ln C_m(r)=D(m)\ln r+\text{const}(\text{常数})$。若 $D(m)$ 随着 m 的增大而保持收敛，则系统是混沌的，此时，$D(m)$ 为饱和关联维，m 为最小嵌入维。

利用混沌理论来分析非线性访问时间序列，首先，确定时延 τ。网络访问时间演化序列的 R_k 随 k 的变化如图 6 所示。由图 6 可知，最佳时延 $\tau=3$。然后，利用分形维的 G-P 算法[17]，绘制最佳时延 $\tau=3$ 下，嵌入维数 m 为 3~12 时 $\ln C_m(r)-\ln r$ 曲线，如图 7 所示。

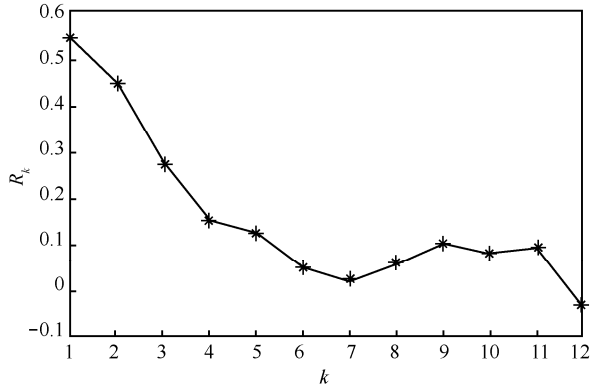


图 6 网络访问时间演化序列的 R_k 随 k 的变化

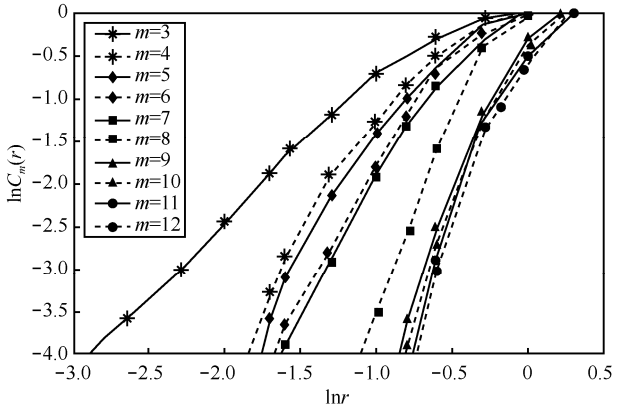


图 7 网络访问时间演化序列的关联积分

从图 7 可以看到，随着 m 的增大，$\ln C_m(r)$ 随 $\ln r$ 变化的曲线斜率逐渐收敛。当 $m>4$ 时，各条曲线斜率趋于平行。当 $m>9$ 时，各条曲线几乎重叠，说明 $D(m)$ 趋于收敛，由此得出重构相空间的最小嵌入维数 $m=9$。进一步对 $m=9$ 时 $\ln C_m(r)-\ln r$ 曲线做线性回归分析，得到饱和关联维 $D(m=9,\tau=3)=2.830\,4$，是一个分数维，说明网络访问时间演化序列具有混沌特征。

5 网络传播行为的预测模型

依据网络访问时间演化序列的混沌特性，在网络时序演化中引入 Logistic 方程[18]并加以改进，基于网络访问时间演化序列建立预测模型，并进行仿真分析与验证。

5.1 模型的建立

第 4 节通过对网络访问时间演化序列分析得到，在探测期间网络访问时间演化趋势是缓慢振荡下降的，其振荡的幅度相对并不大，由于自身拓扑结构的突变或节点间通信异常使其呈现出一种准周期性振荡衰减趋势。但是随着互联网的迅猛发展，网络业务不断增多，给网络通信传输带来了一定的压力，网络访问时间并不会一直衰减，且一定存在某一下限。因此，采用 Logistic 方程描述网络访问时间演化行为是可行的。建模过程如下。

步骤 1 将网络访问时间演化序列代入 Logistic 模型的非线性微分方程，有

$$\frac{\mathrm{d}T}{\mathrm{d}t}=rT\left[1-\frac{T}{K}\right] \tag{6}$$

其中，r 为网络访问时间变化率，T 为 t 时刻（以月为单位）的网络访问时间。

步骤 2 对式(6)进行积分，得

$$T=\frac{K}{1+m\mathrm{e}^{-rt}} \tag{7}$$

由式(7)可得，当 $r>0$ 时，T 随着 t 的增大而单调递增。

步骤 3 对 Logistic 方程变换，使其符合网络访问时间演化序列的振荡衰减特征，变换式为

$$T=d-\frac{K}{1+m\mathrm{e}^{-rt}} \tag{8}$$

步骤 4 引入带正余弦的指数线性组合作为振荡衰减因子，同时引入校正系数 p 确保模型能反映实际演化特征，最终得到 Logistic 方程的改进模型为

$$T=d-\frac{K}{1+m\mathrm{e}^{-\left\{p+r_1\mathrm{e}^{\sin\left[\pi\left(\frac{t}{u_1}+v_1\right)\right]}+r_2\mathrm{e}^{\cos\left[\pi\left(\frac{t}{u_2}+v_2\right)\right]}\right\}t}} \tag{9}$$

其中，r_1和r_2为振幅，v_1和v_2为初始幅角，u_1和u_2为振荡半周期，T为输出值。模型方程简化表示为$T=f\left(d,K,m,p,r_1,u_1,v_1,r_2,u_2,v_2,t\right)$。

5.2　参数的确定

模型参数的选择对预测模型的准确性有很大影响。本文采用粒子群优化算法（PSO）[19]根据探测期间网络访问时间演化序列对 Logistic 模型取优。算法流程如下。

输入　网络访问时间演化序列

输出　最优适应度个体$T=f(d,K,m,p,\ r_1,u_1,v_1,r_2,u_2,v_2,t)$

步骤 1　设置初始参数最大迭代次数、群体规模M和加速度c等，并确定各参数的取值范围。根据问题的复杂程度和需求设置群体规模和算法的终止条件。

步骤 2　定义适应度函数。标准的 PSO 算法中，适应度函数$f\left(X\right)$是一个最小优化目标距离，即$f\left(X\right)=\min\limits_{i,j\in\{1,2,\cdots,M\}}\left|x_j-x_i\right|$，$x_i$为第$i$个粒子的空间位置，那么个体经历的最优位置所对应的适应度为$f_{\text{best}}\left(x_i\right)$；所有粒子经历的最优位置所对应的适应度为$f_{\text{best}}$。本模型以评价标准的角度出发，建立模型的输出值$T^*\left(i\right)$与实际值$T\left(i\right)$的累积误差作为适应度函数，为

$$S\left(i\right)=\frac{1}{n}\sum_{i=1}^{n}\left|T\left(i\right)-T^*\left(i\right)\right| \tag{10}$$

其中，n为以月为单元统计的时间跨度。$S\left(i\right)$值越小，模型的输出值与真实数据总体误差越小，预测模型就能够准确地模拟真实数据。

步骤 3　在参数的取值范围内随机生成初始群体，计算每个个体的适应度$S\left(i\right)$。若$S\left(i\right)<f_{\text{best}}\left(x_i\right)$，则个体所处于局部最好位置；若$S\left(i\right)<f_{\text{best}}$，则个体处于全局最好位置。

步骤 4　判断适应值是否超过最大迭代次数或预设值。若不满足则继续进行步骤 3 的计算和判断；若满足则结束，输出结果。

5.3　实验与分析

以月为单元，选择 4 个监测点的 2012-2015 年共 48 个月的网络访问时间演化序列。分别将 4 个监测点的前 40 个月作为预测模型的输入值进行模拟演化，然后，对比后 8 个月的数据输出值，以此来评价预测模型的准确度。

首先，设置群体规模$M=50$，最大迭代次数为 1 000，加速度为 2。然后，根据第 3 节和第 4 节对网络访问时间序列的分布特征与演化特征的分析，确定模型中各参数的取值范围。接着，对每个监测点在预测模型中进行反复多次实验以取得最优解，如图 8 所示。

实际上，由于系统的复杂性、算法的缺陷以及外界随机噪声等因素的影响，PSO 输出会带有一定的误差，绝对最优解是不存在的。理想状态并不存在，寻找完备空间也没有必要，只要能在一个非完备空间中找到相对最优解就是合理的。从图 8 可以看到，对 4 个监测点的数据进行实验，当迭代次数超过 800 以后，群体较难产生更优的个体，说明模型参数算法的收敛性很好，此时参数的选择较为合理。将 4 个监测点 PSO 输出的参数优化值代入预测模型，为

$$T_{\text{amw}}=214.592\,7-\frac{60.747\,9}{1+64.444\,8\text{e}^{-\left\{3.369\,5+0.115\,7\text{e}^{\sin\left[\pi\left(\frac{t}{0.245\,8}+1.062\,9\right)\right]}+0.350\,4\text{e}^{\cos\left[\pi\left(\frac{t}{0.543\,8}+0.228\,9\right)\right]}\right\}t}} \tag{11}$$

$$T_{\text{san}}=235.283\,2-\frac{61.4211}{1+65.398\,7\text{e}^{-\left\{3.378\,8+0.091\,8\text{e}^{\sin\left[\pi\left(\frac{t}{0.248\,7}+1.277\,3\right)\right]}+0.263\,4\text{e}^{\cos\left[\pi\left(\frac{t}{0.565\,1}+1.246\,9\right)\right]}\right\}t}} \tag{12}$$

$$T_{bcn}=228.737\,9-\frac{52.844\,3}{1+84.489\,8\text{e}^{-\left\{2.988\,3+0.131\,9\text{e}^{\sin\left[\pi\left(\frac{t}{0.494\,3}+1.088\,3\right)\right]}+0.286\,3\text{e}^{\cos\left[\pi\left(\frac{t}{0.129\,1}+1.874\,7\right)\right]}\right\}t}} \tag{13}$$

$$T_{mnl}=250.211\,3-\frac{76.454\,2}{1+91.417\,8\text{e}^{-\left\{3.992\,5+0.033\,9\text{e}^{\sin\left[\pi\left(t+1.062\,9\right)\right]}+0.350\,4\text{e}^{\cos\left[\pi\left(\frac{t}{0.543\,8}+0.228\,9\right)\right]}\right\}t}} \tag{14}$$

5.4　模型评价

为了评价模型的预测准确性，引入相对平均误差作为模型评价指标，计算式为

$$\overline{\varepsilon_r}=\frac{\frac{1}{n}\sum\limits_{i=1}^{n}\left|T\left(i\right)-T^*\left(i\right)\right|}{\frac{1}{n}\sum\limits_{i=1}^{n}T^*\left(i\right)} \tag{15}$$

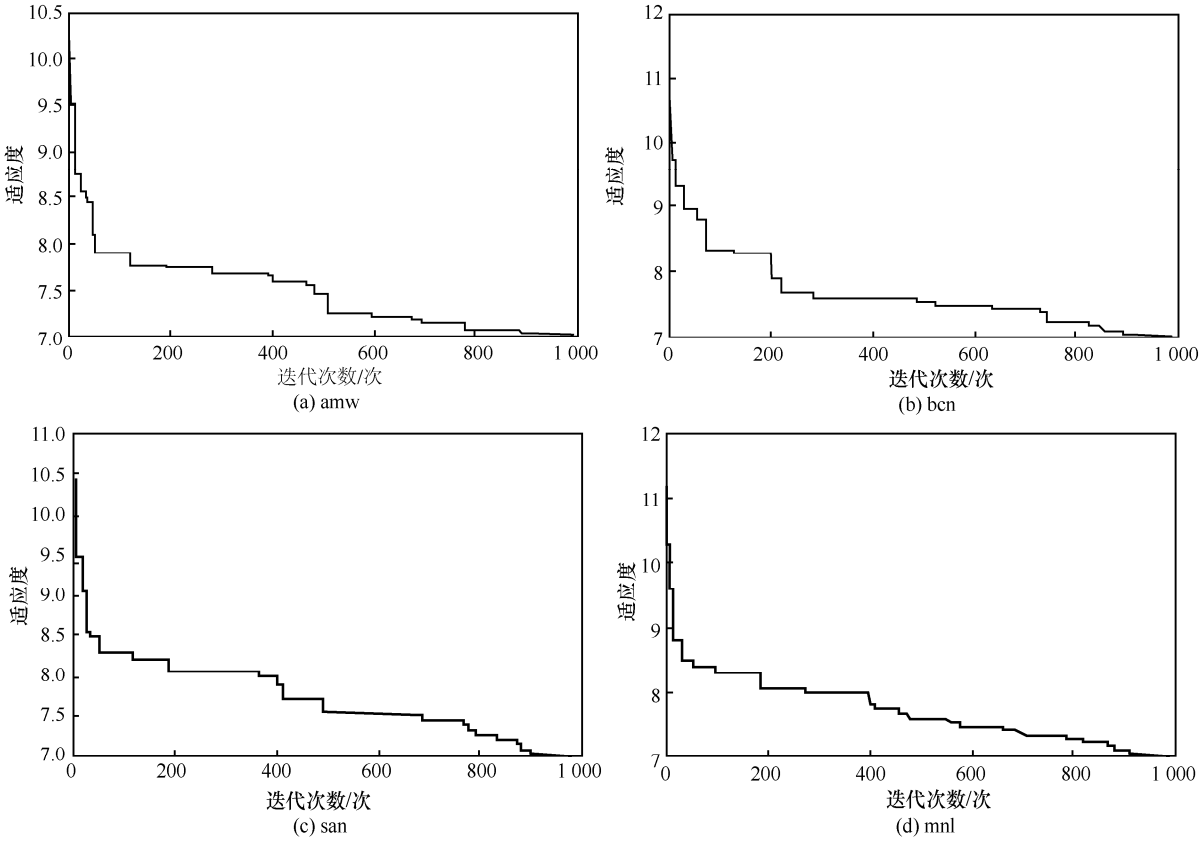


图 8　PSO 收敛过程

其中，$\bar{\varepsilon}=\frac{1}{n}\sum_{i=1}^{n}\left|T(i)-T^{*}(i)\right|$ 为绝对平均误差。对于一个模型来说，如果其拟合准确度大于 95%，并且预测准确度大于 80%，这样的模型是可以接受的。

首先，分别计算 4 个监测点的预测模型的拟合值和预测值，并与其实际值对比，如图 9 所示。

从整体演化趋势来看，模型计算的拟合值和预测值的趋势走向与实际数据所表现出的网络访问时间演化序列的变化态势大致相同，且随着时间的推移，其都是呈局部缓慢振荡下降的。在探测时间内，开始时拟合效果并不好，直观上来看，amw、bcn、san 和 mnl 监测点模型计算值与实际值分别在 14 个月、9 个月、11 个月和 10 个月前差距较大，这是因为初期粒子群体需要一段时间的适应过程。之后除了个别异常波动点以外，拟合值与实际值的重合性相对较好。而对于后 8 个月的预测值却有不同表现，在振荡幅度较小的时间点上，预测值和实际值差距不大，但从实际数据来看，短短 8 个月的序列演化仍会出现突变点，尤其 bcn 监测点后期各月波动性最大，这种情况下的预测准确度会受到影响，但是可以通过振荡的幅度和频率来判断未来的演化趋势，在实际应用中可以将这部分时间做记录，后续演化时需对它们进行密切关注以及重点研究。分别计算 4 个监测点的预测模型的拟合相对平均误差和预测相对平均误差，然后用 1 分别减去相应的相对平均误差值，得出拟合准确度和预测准确度，结果如表 4 所示。

表 4　　模型评价指标

监测点	拟合准确度	预测准确度
amw	95.02%	92.89%
bcn	96.84%	92.65%
san	96.31%	93.58%
mnl	96.78%	94.09%

由表 4 可知，模型拟合准确度和预测准确度都满足评价判定标准。这说明预测模型的构建合理，应用该模型能够对网络传播行为的演化做出准确的预测。

时间序列分析是一种广泛应用的数据分析方法，它研究的是代表某一现象的一串随时间变化而又相关联的动态数据，从而描述和探索该现象随时间发展变化的规律性。时间序列分析利用的手段可

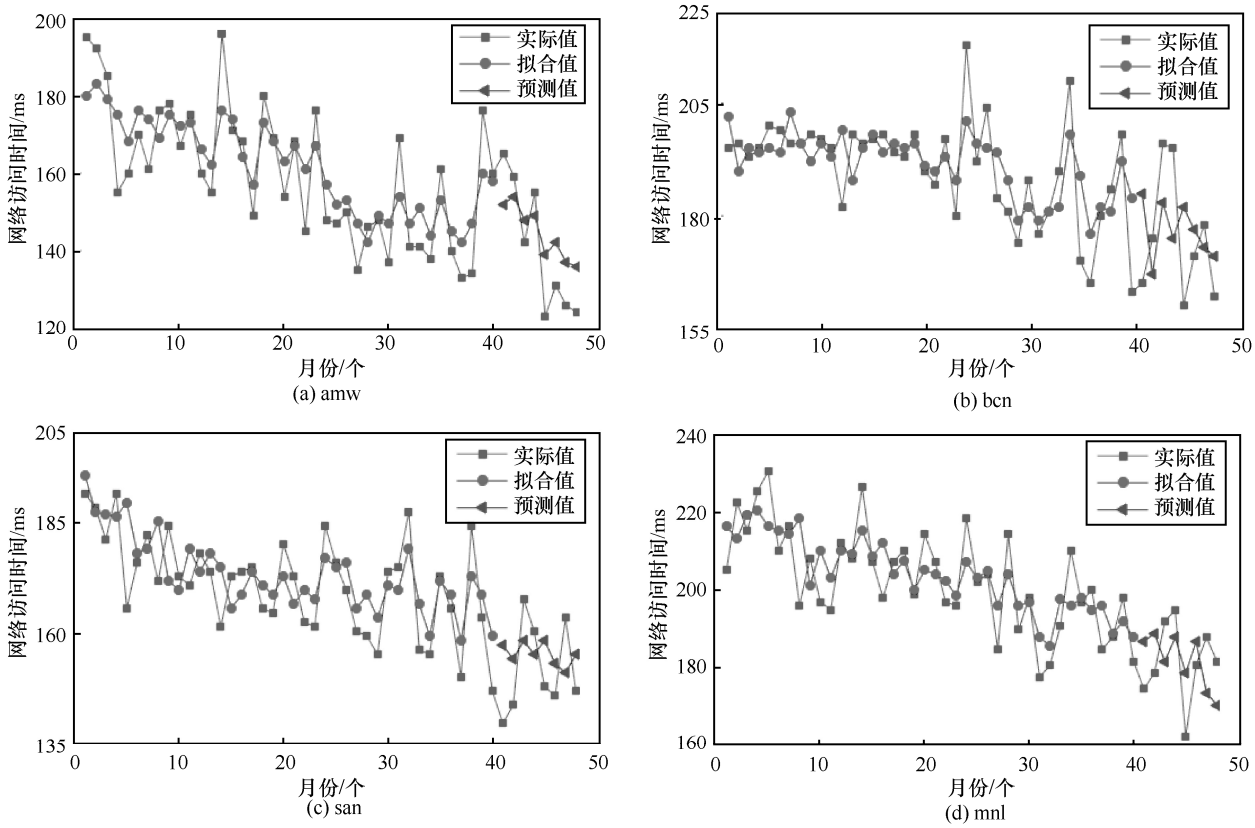


图 9　实际值、拟合值和预测值的对比

以是直观简便的数据图法、指标法、模型法等。而模型法相对来说更具体也更深入，能更本质地了解数据的内在结构和复杂特征，以达到控制与预测的目的。传统的一维拟合模型仅能表现目标系统的一维物理过程，在宏观拓扑结构下，网络的传播行为能够表征网络拓扑结构对网络动态行为的影响。互联网是基于时间和空间运行的抽象体，时间序列的演化是网络特征表现最直接的载体。网络访问时间序列的混沌特性说明了网络访问时间的自相似和稳定性，进而通过预测模型建立访问时间预测算法，可得到当前和未来一段网络环境的访问时间，这为面临恶意网络传播时有效的内容劫持提供约束条件，并为网络劫持提供时间约束条件，促进构建强大的、安全的网络空间。

但由于混沌系统的初始敏感性、初始场的不准确性以及复杂系统内部随机性使计算极易出现较大的误差，随着时间的推移，误差会不断累积增大。另一方面，由于数据采样过程中可能带有噪声等因素的影响，使混沌序列的时序演化不断地振荡，长时间跨度下很难做出精准的预测。但混沌时间序列是由确定性非线性系统产生的，其内部存在确定性规律，因此，短期内预测模型可以对网络传播行为演化做出较好的预测。实际上，这样一个预测模型对于互联网来说是十分实用的，因为互联网自身的发展迅速，并且带有很多未知，长期预测并没有意义，浪费了时间和资源。如果在某段时间范围内能够准确地预测网络行为的演化特征和规律，在预测能力消失之前对网络进行适当的技术改善，那么这个预测模型也是圆满完成了任务。

6　结束语

为保证网络的正常服务、提升网络性能和应用拓展，在提取 CAIDA_Ark 项目下 4 个监测点的有效路径样本数据的基础上，本文从大时间尺度上对网络访问时间的分布和演化进行统计和描述。对 4 个监测点在探测期间内的有效路径中网络访问时间和访问直径进行分析，得出网络访问时间呈多峰重尾分布，具有自相似性；访问直径与网络访问时间具有极弱相关性，可认为不相关，说明在网络传播过程中，访问直径对数据分组的访问时间的影响并不大。因此在路由算法设计上，不仅要关注路由链路的长度还要注重路由链接的性能，从而改善了

路由算法，提高网络的通信效率。

统计 4 个监测点的网络访问时间演化序列，以非线性时间序列分析方法对其时序演化特征进行混沌辨识，得出网络访问时间演化序列具有混沌特征。借此可分析最大的网络时延点，通过在该点放置内容缓存服务器，降低访问时间和访问直径，提高网络内容传播效率，为内容分发网络中的内容缓存服务器部署提供指导建议。最后，引入 Logistic 模型并适当改进，采用粒子群算法（PSO）对模型参数取优，建立以网络访问时间演化序列为基础的网络传播预测模型。通过对 4 个监测点数据在模型中的实验分析，验证了模型的有效性，该模型短期内能够对网络传播行为做出准确的预测，可为下一代互联网建设提供指导性意见。

参考文献：

[1] YOOK S H, JEONG H, BARABASI A L. Modeling the Internet's large-scale topology[J]. Proceedings of the National Academy of Sciences, 2002, 99(21): 13382-13386.

[2] PASTOR-SATORRAS R, VESPIGNANI A. Evolution and structure of the Internet: a statistical physics approach [M]. Cambridge: Cambridge University Press, 2007.

[3] KROGFOSS B, WELDON M, SOFMAN L. Internet architecture evolution and the complex economies of content peering[J]. Bell Labs Technical Journal, 2012, 17(1): 163-184.

[4] DING W, YAN Z, DENG R H. A survey on future Internet security architectures[J]. IEEE Access, 2016, 4: 4374-4393.

[5] 除久强，王进法，张君，等. 基于度相关性病毒传播模型及其分析[J]. 中国科学: 信息科学, 2014, 44(6): 793-810.
XU J Q, WANG J F, ZHANG J, et al. Virus spreading model based on degree correlation and its analysis[J]. Scientia Sinica Informations 2014, 44(6): 793-810.

[6] ZHANG S, ZHAO H. Community identification in networks with unbalanced structure[J]. Physical Review E, 2012, 337(6092): 337-341.

[7] WANG P, AKYILDIZ I F. Improving network connectivity in the presence of heavy-tailed interference[J]. IEEE Transactions on Wireless Communications, 2014, 13(10): 5427-5439.

[8] 徐野，赵海，苏威积，等. Internet 网络的访问直径分析[J]. 计算机学报, 2006, 5(29): 690-698.
XU Y, ZHAO H, SU W J, et al. Analysis on traveling diameter of Internet[J]. Chinese Journal of Computers, 2006, 5(29): 690-698.

[9] 隋岩，曹飞. 从混沌理论认识互联网群体传播特性[J]. 学术界, 2013(2): 86-94.
SU Y, CAO F. Study on features of Internet group communication by chaos theory[J]. Condensed Paper, 2013(2): 86-94.

[10] YE Z, MISTRY S, BOUGUETTAYA A, et al. Long-term QoS-aware cloud service composition using multivariate time series analysis[J]. IEEE Transactions on Services Computing, 2016, 9(3): 1.

[11] CHAI S H, LIM J S. Forecasting business cycle with chaotic time series based on neural network with weighted fuzzy membership functions[J]. Chaos Solitons & Fractals, 2016, 90: 118-126.

[12] 林川，赵海，毕远国，等. 互联网网络时延特征研究[J]. 通信学报, 2015, 36(3): 163-174.
LIN C, ZHAO H, BI Y G, et al. Research on network delay of Internet[J]. Journal on Communications, 2015, 36(3): 163-174.

[13] HUFFAKER B, FOMENKOV M, PLUMMER D J, et al. Distance metrics in the Internet[J]. IEEE International Telecommunications Symposium, 2002: 200-202.

[14] 王光义,袁方. 级联混沌及其动力学特性研究[J].物理学报, 2013, 62(2): 020506.
WANG G Y, YUAN F. Cascade Chaos and its dynamic characteristics[J]. Acta Physica Sinica, 2013, 62(2): 020506.

[15] 张春涛，马千里，彭宏. 基于信息熵优化相空间重构参数的混沌时间序列预测[J]. 物理学报, 2010, 59(11): 7623-7629.
ZHANG C T, MA Q L, PENG H. Chaotic time series prediction based on information entropy optimized parameters of phase space reconstruction[J]. Acta Physica Sinica, 2010, 59(11): 7623-7629.

[16] 侯站. 基于预测的相空间重构技术研究[D]. 郑州: 郑州大学, 2010.
HOU Z. Research on phase space reconstruction based on prediction[D]. Zhengzhou: Zhengzhou University, 2010.

[17] GRASSBERGER P, PROCACCIA I. Measuring the strangeness of strange attractors[J]. Physica D, 1983, 9(1, 2): 189-208.

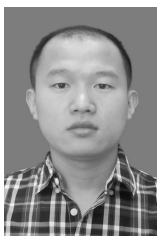
[18] JOVIC B, UNSWORTH C P. Fast synchronization of chaotic maps for secure chaotic communications[J]. Electronics Letters, 2010, 46(1): 49-50.

[19] DENNEDY J, EBERHART R C. Particle swarm optimization[C]//The IEEE International Conference on Neural Networks. 1995: 1942-1948.

[作者简介]

田鹤（1985-），女，辽宁沈阳人，辽宁科技学院讲师，主要研究方向为计算机网络、复杂网络。

赵海（1959-），男，辽宁沈阳人，博士，东北大学教授、博士生导师，主要研究方向为复杂网络、嵌入式系统、普适计算等。



王进法（1988-），男，山东德州人，东北大学博士生，主要研究方向为互联网拓扑分析、数据融合。

林川（1988-），男，辽宁凤城人，东北大学博士生，主要研究方向为复杂网络、软件定义网络。