

Internet 网络访问直径的短期及长期预测

徐 野¹ 赵 海² 张文波¹

(沈阳理工大学信息科学与工程学院 沈阳 110168)¹ (东北大学信息科学与工程学院 沈阳 110004)²

摘 要 本文首先形式化定义表征 Internet 的参量—访问直径,然后根据海量数据变化特点,提出了基于 Logistic 模型的、以正余弦函数模拟振荡变化的数学模型,使用浮点型遗传算法确定模型形式。由于 Internet 访问直径演化复杂,传统一维模型在长期预测中已不适用。因此在长期预测方面,首先计算得到 Internet 访问直径的关联维数,然后根据关联维数及奇异吸引子相近空间混沌轨道运动特性,提出了基于三维常微分方程组数学模型形式。

关键词 复杂网络,访问直径,Logistic 模型,浮点遗传算法,关联维数,奇异吸引子

Short-term and Long-term Forecast of Internet Traveling Diameter

XU Ye¹ ZHAO Hai² ZHANG Wen-bo¹

(College of Information Science and Engineering, Shenyang Ligong University, Shenyang 110168, China)¹

(School of Information Science and Engineering, Northeastern University, Shenyang 110004, China)²

Abstract Based on giant samples, a property of Internet-traveling diameter was firstly defined, and its mathematical model which was composed by a Logistic Model and sine and cosine function groups simulating oscillations during the growth of traveling diameter was then determined by a Float-point GA. Since growth of traveling diameter is complex, the model mentioned above is not suitable while being used for a long-term forecast. To solve this problem, Correlation Dimension of traveling diameter was calculated, and a model comprising three-dimension function groups was put forward based on the value of Correlation Dimension and properties of track near the strange attractor in Chaos system. This model was proved to be comparatively suitable for long-term forecast.

Keywords Complex networks, Traveling diameter, Logistic model, Float-point GA, Correlation dimension, Strange attractor

1 引言

Internet 已经成为复杂网络研究的热点。文献[1-5]的研究使用无尺度、幂律等参量描述 Internet 的复杂性。与此相似,本文提出 Internet 网络访问直径参量描述其特性,并对该参量展开研究。

首先提出基于 Logistic 方程的数学模型对访问直径未来发展做短期预测。但由于属于传统的一维(一个自变量)模型,仅能表现目标系统的一维物理过程(常见为时间),无法全面地描述目标系统的所有物理过程,因此当时间增长后,模式系统的预测准确率将会迅速下降^[12]。但限于可计算性,拟合模型不能以无穷多维映射目标系统的无穷多物理过程,因此需要确定目标系统的具体维数。据此,本文将从分形角度确定长期预测模型形式。

1.1 定义访问直径

定义 1 Internet 中,如果数据包从源 IP 到目的 IP 所经路由跳数为一次访问的直径,那么大量数据包从任意源 IP 到任意目的所经路由跳数的统计均值称为 Internet 网络访问直径。设数据包跳数为 J_i ,频数为 F_i ,统计频率为 p_i ,样本总量

为 n ,则访问直径

$$D = \sum_{i=1}^n J_i p_i = \frac{1}{n} \sum_{i=1}^n J_i F_i \quad (1)$$

1.2 样本量

复杂网络是一种宏观网络,需要通过海量级数据样本进行有效的研究。本文样本选自 CAIDA¹⁾,采用了北美洲美国圣地亚哥(riseling)、欧洲荷兰阿姆斯特丹(k-peer)和亚洲日本东京(apan-jp)三个结点²⁾五年间³⁾(99.7~04.6)7500 万条数据。

1.3 可达样本量

本文忽略被路由器舍弃样本,针对可达样本进行研究,其样本量参见表 1。

表 1 中可达样本约占总量 55%,共约 41000 万条,保证了数据的冗余性。

表 1 可达样本统计分析结果

节点	可达(C)	总量	C%
1	16448329	30118057	54.6%
2	9479028	19034983	49.8%
3	15172408	25595600	59.3%

徐 野 博士,副教授,主要研究方向为智能计算与模式识别、复杂网络。

¹⁾ CAIDA(The Cooperative Association for Internet Data Analysis)是合作性国际研究机构,其参与者共 30 余家,主要分布在北美洲、欧洲国家的研究院及高校。亚洲有三家,两家在日本东京,另一家在中国东北大学。

²⁾ 选取 CAIDA 在北美洲、欧洲和亚洲三个结点的原因是它们能客观地表现 Internet 网络。而每洲仅取一个结点是简化统计分析的复杂度。没有选取亚洲中国结点的原因是中国结点成立时间短,数据时间跨度不长。

³⁾ 每月抽取当月 15 日一天的数据。三个结点分别存在数据缺失,其中 riseling 缺 01.7-9 数据和 kpeer 缺 99.7-01.5 数据和 04.1 数据,apan-jp 缺少 99.7、00.6 和 03.10 数据。缺失数据由另两结点同时段数据均值代替。

1.4 访问直径的直观特性

空间维上,在已有三个结点基础上,分别从三大洲多选一个结点(美国弗吉尼亚 a-root、法国巴黎 cdg-rssac 和亚洲日本

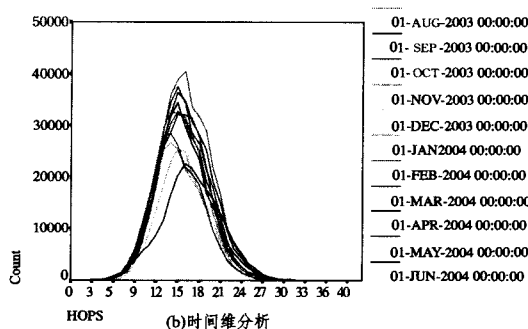
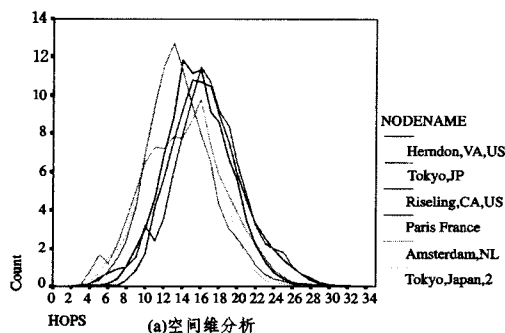


图1 访问直径空间与时间维分析

从图1看到,曲线波峰集中地收敛于 $[10, 18]$;跳数全幅跨度约在2到32跳之间,考虑误差因素,将其设大一点(1, 36),以确保区间的完整性。故

$$D \in (1, 36) \quad (2)$$

1.5 访问直径随时间变化的特性

三个节点 D 按时间变化情况参见图2。

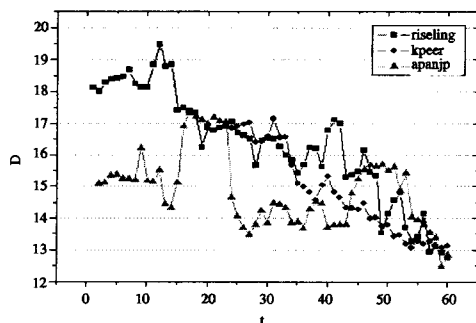


图2 三个结点访问直径 D 随时间 t (月, 99.7~04.6) 的变化

从图2看到,三个结点的访问直径 D 随时间 t 增加呈现一致缩减。三个结点在初始区间差值较大,在 $t > 50$ 后差值明显收缩,在 t 近60处收敛于一个差值极小区间。

对三个节点 D 均值与时间 t 变化做图,如图3所示。

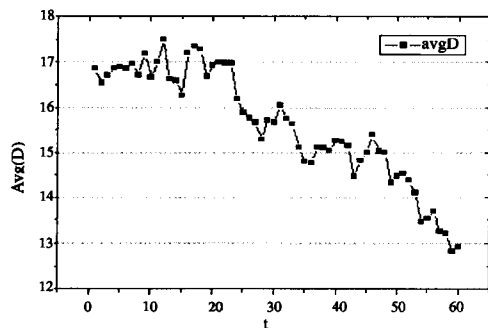


图3 访问直径均值变化

从图3看到, D 均值随时间变化规律与 Logistic 方程研究的随时间生长变化规律相似,可用来拟合图3数据^[6]。

2 短期预测

2.1 改进模型

的 nrt),抽取 03.12.15 数据 1198204 条,进行统计分析,分析结果见图1(a)。时间维分析结果见图1(b)。

Logistic 模型^[7,8,11]由 Verhulst 于 1837 年提出,用于研究有限资源条件下的人口发展模型,将其积分式应用于本文数据,得到

$$D = \frac{k}{1 + \frac{k}{D_0 - 1} e^{-\alpha}} = \frac{k}{1 + m e^{-\alpha}} \quad (3)$$

其中 $D_0 = D(t=0)$, $m = \frac{k}{D_0 - 1}$ 。 $D(\geq 0)$ 是 t 时刻访问直径 D 值, $t(>0)$ 为时间(月), $k(>0)$ 是访问直径极限值。 $r(>0)$ 是与环境条件和种群物种特性有关的参数,这里表示访问直径 D 的增长速率。

(1)递增改进为递减

标准 Logistic 方程是单调递增函数,根据图4(a-d)变化,可得到递减 Logistic 方程,如式(4)所示。

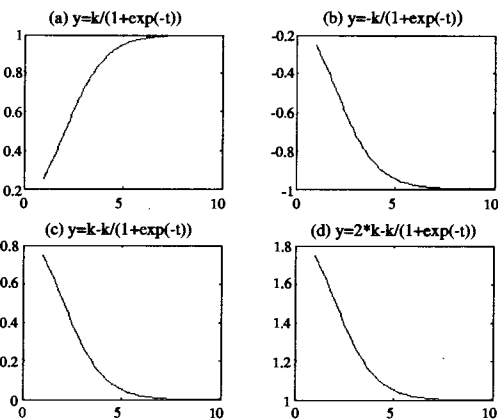


图4 以 Logistic 方程 $y = k/(1+e^{-t})$ ($k=1$) 为例,对其形式进行变换后做图。(a)是标准图形;(b)是 $y = -k/(1+e^{-t})$ 图形;(c)是 $y = k-k/(1+e^{-t})$ 图形;(d)是 $y = 2k-k/(1+e^{-t})$ 图形

$$D = d - \frac{k}{1 + m e^{-\alpha}} \quad (4)$$

其中 d 为沿 y 轴的平移量。

(2)模拟准周期振荡

图3曲线具有准周期性振荡,可由基于正弦、余弦的函数或组合描述^[6]。改进模型表示为

$$D = d - \frac{k}{1 + m e^{-(v + p_1 e^{-g_1 t} \sin[\pi(\frac{t}{h_1} + u_1)] + p_2 e^{-g_2 t} \cos[\pi(\frac{t}{h_2} + u_2)])t}} \quad (5)$$

其中 $D(\geq 0)$ 是 t 时刻访问直径 D 值, d 是平移量, v 为校正系

数, p_1, p_2 为相对振荡幅值, h_1, h_2 为振荡半周期长度(月), u_1, u_2 为振荡初始幅角, g_1, g_2 为振荡衰减系数。式(4)中参数 r 已经被式(5)中 v 和 p_1, p_2 吸收。

2.2 拟合

本文使用浮点型遗传算法^[6,9,10]作为拟合方法。

(1) 遗传基因编码

遗传基因 x 由待拟合的参数向量表示:

$$x = (d, k, m, v, p_1, p_2, g_1, g_2, h_1, h_2, u_1, u_2) \quad (6)$$

(2) 随机生成初始群体

设群体规模为 N (本文 $N=100$), 随机生成初始群体。

(3) 适应度函数

确定待定参数应使预测模型的访问直径 $D(t)$ 与实际值 $D^*(t)$ 拟合最好, 即从 99.7~04.6 共 60 个月的 $D(t)$ 与 $D^*(t)$ 的差值累计最小。由此, 适应度函数

$$f(x) = \sum_{i=1}^{60} |D(t_i) - D^*(t_i)| \quad (7)$$

遗传个体分值取误差累计 $f(x)$ 的倒数

$$score(x) = \frac{1}{f(x)} \quad (8)$$

(4) 选择

选择分数高的前 m 个 (m 为随机数, 且 $0 < m < 1$) 个体, 复制两份, 删除分数最低的后 m 个个体, 群体总数不变。

(5) 杂交

杂交操作为

$$\begin{aligned} v'_i &= v_i(1-\alpha) + \beta v_j \\ v'_j &= v_j(1-\alpha) + \beta v_i \end{aligned} \quad (9)$$

其中 α, β 为小于 1 的随机数。

(6) 变异

对每个参数以 0.3 的概率进行变异操作。以参数 v 变异为例, 有

$$\begin{aligned} v_i &= v_i(1+\alpha), \text{ if } \gamma \geq 0.5 \\ v_i &= v_i(1-\alpha), \text{ if } \gamma < 0.5 \end{aligned} \quad (10)$$

其中 α, γ 为小于 1 的随机数。

(7) 终止

设定计算轮数为 50000, 以保证得出优秀模型参数。

(8) 结果

根据遗传算法得模型为

$$D = 16.949545 - \frac{6.8262}{1 + 58.241775 \times e^{-(A+B)t}} \quad (11)$$

其中 $t > 0, t \in N, A, B$ 为

$$A = 0.007844 + 0.001218e^{-0.001191t} \times \sin\left[\pi\left(\frac{t}{0.000866} + 0.014911\right)\right]$$

$$B = 0.082215e^{-0.004712t} \times \cos\left[\pi\left(\frac{t}{0.001} + 0.003242\right)\right]$$

2.3 评价

定义 2 设访问直径原始样本为 $D^*(t_i)$, 模型计算值为 $D(t_i)$ ($i \geq 1, i \in N$), 相对平均误差为

$$\bar{\epsilon}_r = \frac{\bar{\epsilon}}{D^*} = \frac{\sum_{i=1}^n |D(t_i) - D^*(t_i)|}{\frac{1}{n} \sum_{i=1}^n D^*(t_i)} = \frac{\sum_{i=1}^n |D(t_i) - D^*(t_i)|}{\sum_{i=1}^n D^*(t_i)} \quad (12)$$

其中 n 为按月计算的时间跨度。

根据拟合模型的原始数据 (99.7~04.6), 使用式(12)计

算误差为

$$\bar{\epsilon}_r = \frac{\bar{\epsilon}}{D^*} = \frac{1.275278}{60} = 0.021255$$

模型拟合准确度为 $1 - 0.0212 = 97.87\%$, 可接受。

2.4 短期预测

使用 rising 结点 04.07~05.05 数据验证拟合模型的预测精度。据式(12)计算, 得

$$\bar{\epsilon}_r = \frac{\bar{\epsilon}}{D^*} = \frac{0.277985}{11} = 0.025271$$

故拟合模型在 04.7~05.5 预测精度为 $1 - 0.0252 = 97.47\%$, 与拟合准确度相比略降。将该模型应用到长期预测中, 虽无法验证, 但可想像其准确率必然会随时间增长而迅速下降。

在引言中提到, 这是由于拟合模型维数不能映射目标系统多个物理过程, 需要对拟合模型形式做进一步改进。

3 长期预测

3.1 混沌系统的长期预测方法

洛伦兹在研究天气长期预测中^[17]得出无法对混沌系统做长期预测的结论。然而, 混沌系统在奇异吸引子上的运动是遍历和混和的, 即在限定时间范围内, 系统的混沌轨道将在以奇异吸引子为中心的限定空间范围内运动^[12,13]。因此, 在限定的时间和空间范围内, 混沌系统的状态在奇异吸引子上的平均值不是条件敏感的, 是可预测的, 且比传统短期预测方法的预测时间长, 因此称为长期预测。

通过奇异吸引子进行预测的系统必须是混沌系统, 下面验证 Internet 混沌性。

3.2 混沌性

可以通过计算目标系统关联维数来判断是否为混沌系统。根据从目标系统提取的时间序列构造 m 维相空间, 如果随 m 增大, 计算得到关联维数 D_2 达到饱和值(收敛), 则目标系统存在吸引子, 必为混沌系统。而如果随 m 增大, D_2 不收敛而增大到无穷, 则不是混沌系统^[12]。计算关联维数算法参见表 2^[12,14-16]。

表 2 关联维数算法

输入 list // Internet 访问直径时间序列 (99.07~04.06)
// 初始化
tao = 3; length = list.length;
// 根据不同 m, 计算关联维数
loop when m=4,8,10,12 ... until 关联维收敛 do
// 根据 m, 由时间序列形成 length-(m-1)*tao 维的相空间
vecgrou P= zeros(m, length-(m-1)*tao);
vecgrou P= getValue(list);
// 计算相空间中两两向量之间距离
rij = calcRIJ(vecgroup(:,i), vecgroup(:,j));
// 确定 r(矩阵), 确定 r 的循环步数为 15
r = [maxRij, (maxRij-minRij)/15, minRij]
// 计算关联积分 cr(矩阵)
cr = calcCR(r, rij);
// 对 ln(r) 和 ln(cr) 做图, 计算关联维数
plot(log(r), log(cr));
// 当关联维数 D_2 收敛, 求斜率得到 D_2
calcD2();
// 结束循环
end of loop

文献[12]建议 τ 取值 4~8, 但由于 Internet 访问直径时间序列较小 (60 个点), 因此取 3, 保证了由时间序列形成的相空间中向量间的独立性。

根据算法得到计算结果参见图 5。

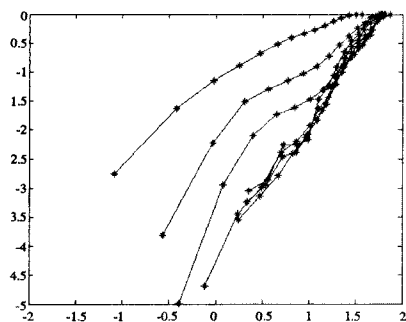


图5 关联维数计算结果

图5表明,当 $m=11,13,15,16$ 时,即图5中从左至右第3条曲线开始,各曲线直线部分平行性很好,几乎已经重叠,可以判定收敛,故判断目标系统存在吸引子,为混沌系统。测量其中直线部分斜率,得到Internet访问直径关联维数:

$$D_2(m_c=16, \tau=3)=2.2444 \quad (13)$$

3.3 模型变换

对关联维数为 D_2 的混沌系统,必须至少使用 $\lceil D_2 \rceil$ 维拟合模型,表征影响目标系统中 $\lceil D_2 \rceil$ 个重要物理过程,才能实现较准确的预测^[12]。因此据式(13),Internet网络访问直径长期预测的拟合模型至少为 $\lceil D_2 \rceil=3$ 维,其形式为

$$\begin{cases} x=f_1(x,y,z,x^{(n)},y^{(n)},z^{(n)}) \\ y=f_2(x,y,z,x^{(n)},y^{(n)},z^{(n)}) \\ z=f_3(x,y,z,x^{(n)},y^{(n)},z^{(n)}) \end{cases} \quad (14)$$

式中, $x \sim z$ 为表征目标系统重要物理过程的变量, $x^{(n)} \sim z^{(n)}$ 为该变量的 n 阶导数。

式(14)是表征奇异吸引子在限定时间空间内混沌轨道运动形式的数学模型形式。由于混沌轨道运动形式复杂,而Internet访问直径时间序列跨度较小(5年),因此尚不足以确定式中具体参数,这将是下一步工作。

结束语 本文首先形式化定义Internet访问直径参量,然后以CAIDA rising, k-peer和apan-jp三个结点5年(99.07~04.06)7500万条样本为基础,对Internet访问直径的短期及长期预测行为进行了初步分析。

在短期预测方面,提出了基于Logistic模型的改进模型。在长期预测方面,计算了Internet访问直径的关联维数,提出

了三维微分方程组的长期预测模型形式。但限于时间跨度小,尚不能确定参数,这是下一步要开展的工作。

参考文献

- [1] Floy S, Paxson V. Difficulties in simulating the Internet. IEEE/ACM Trans. on Networking, 2001, 9(4): 392-403
- [2] Watts D, Strogatz S. Collective dynamics of 'small-world' networks. Nature, 1998, 393(6684): 440-442
- [3] Aiello W, Chung F, Lu L Y. A random graph model for massive graphs // Proc. of the ACM STOC 2000. Portland: ACM Press, 2000: 171-180
- [4] 张宇, 张宏莉, 方滨兴. Internet 拓扑建模综述. 软件学报, 2004, 15(8): 1221-1226
- [5] 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. 系统仿真学报, 2002, 14(11): 1450-1455
- [6] 尹朝庆, 尹皓. 人工智能与专家系统. 北京: 中国水利水电出版社, 2002: 286-295
- [7] 杨志杰, 徐中儒. 用动态逻辑斯蒂预测法研究黑龙江省乡村人口增长. 黑龙江八一农垦大学学报, 1997, 6(9): 23-28
- [8] 吴淑玲. 利用 Logistic 模型预测我国数字图书馆的发展趋势. 情报方法, 2004(4)
- [9] 汪剑鸣, 许镇琳. 浮点遗传算法中一种新的杂交算子. 控制理论与应用, 2002, 19(6)
- [10] Rudolph G. Convergence properties of canonical genetic algorithms[J]. IEEE Trans. on Neural Networks, 1994, 5(1): 96-101
- [11] 张荷观. 二个新的种群增长方程. 生物数学学报, 1995, 10(4): 78-82
- [12] 黄润生. 混沌及其应用. 武汉: 武汉大学出版社, 2000: 191-239
- [13] 张琪昌, 王洪礼, 竺致文. 分岔与混沌. 天津: 天津大学出版社, 2005: 256-281
- [14] Falconer K J. Fractal Geometry- Mathematical Foundations and Applications. Northeastern University Press, 1991: 41-75
- [15] 徐平. 大坝观测数据的吸引子分析. 大坝观测与土工测试, 1992(10)
- [16] 康卓, 黄竞伟, 李艳, 等. 复杂系统数据挖掘的多尺度混合算法. 软件学报, 2003, 14(7): 1229-1237
- [17] Lorenz E N. Deterministic nonperiodic flow. J. Atmos Sci, 1963, 20(2): 325-334
- [8] Zorzi M, Rao R R. Geographic Random Forwarding (GeRaF) for Ad Hoc and Sensor Networks: Multihop Performance. IEEE Transactions on Mobile Computing, 2003, (2): 337-348
- [9] Zorzi M, Rao R R. Energy and Latency Performance of Geographic Random Forwarding for Ad Hoc and Sensor Networks. IEEE Wireless Communications and Networking (WCNC 2003), New Orleans, Louisiana, USA, Mar. 2003
- [10] Gao Q, Blow K J, Holding D J, et al. Radio Range Adjustment for Energy Efficient Wireless Sensor Networks. Ad Hoc Networks, 2006(4): 75-82
- [11] Kuruvila J, Nayak A, Stojmenovic I. Hop Count Optimal Position-based Packet Routing Algorithms for Ad hoc Wireless Networks with a Realistic Physical Layer. IEEE Journal on Selected Areas in Communications, 2005, 23(6): 1267-1275
- [12] Guo Z W, LI Z W, YU L. Lifetime prolonging algorithms for underwater sensor networks. China Ocean Engineering, 2006, 20(2): 325-334
- [13] Chen B, Jamieson K, Balakrishnan H, et al. Span: an energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks // Proc. Seventh Ann. Int'l Conf. Mobile Computing and Networking. July 2001
- [14] Schurgers C, Tsiatsis V, Ganeriwal S, et al. Optimizing Sensor Networks in the Energy-Latency-Density Design Space. IEEE Trans. Mobile Computing, 2002, 1(1): 70-80
- [15] Riedy E J, Szewczyk R. Power and Control in Networked Sensors. http://today.cs.berkeley.edu/tos/
- [16] Heinzelman W. Application-Specific Protocol Architectures for Wireless Networks (D). Ph. D. thesis. Massachusetts Institute of Technology, 2000
- [17] Deng J, Han Y S, Heinzelman W B, et al. Balanced-energy Sleep Scheduling Scheme for High Density Cluster-based Sensor Networks. Computer Communications, 2005, 28: 1631-1642

(上接第44页)