Internet 网络的访问直径分析

徐 野[1),2)] 赵 海[1),2)] 苏威积[1),2)] 张文波[2)] 张 昕[1)]

[1)](东北大学复杂网络研究中心 沈阳 110004)

[2)](东北大学嵌入式技术辽宁省重点实验室 沈阳 110004)

摘 要 结合复杂网络理论与 CAIDA 授权的关于 Internet 网络的真实海量数据,从复杂网络理论角度对真实的 Internet 数据进行分析与研究.首先借助物理学和生物学研究的方法,将 Internet 网络视为具有生命涨落特征的活体系统,形式化定义了 Internet 物理特征量——访问直径.然后根据目标复杂系统涨落演化特点,提出了 3 种基于 Logistic 模型的、以带衰减因子的正余弦函数组合模拟振荡涨落的数学模型.使用浮点型遗传算法分别进行拟合实验,并通过实验结果对上述 3 种模型进行优选.最终优选模型的拟合准确度为 97.87%,预测准确度为 97.47%,准确度高,符合 Internet 网络真实数据变化情况.文中使用模型对较远未来网络情况进行了预测,并得出结论:从现在开始至 2011 年 12 月,将是 Internet 网络高速发展时期,之后发展速度变缓,并于 2021 年 10 月左右趋于稳定,此时 Internet 网络访问直径为 10.2073 跳.最后,应用文中模型重点预测出了 2008 年 8 月北京奥运期间 Internet 网络访问直径为 10.7726 跳,并得出奥运期间 Internet 网络效率较高的结论.

关键词 复杂网络;访问直径;Internet 物理表征量;Logistic 模型;遗传算法;浮点遗传算法

中图法分类号 TP393

Analysis on Traveling Diameter of Internet

XU Ye[1),2)] ZHAO Hai[1),2)] SU Wei-Ji[1),2)] ZHANG Wen-Bo[2)] ZHANG Xin[1)]

[1)](*Complex Networks Research Center, Northeastern University, Shenyang* 110004)

[2)](*Laboratory of Embedded Technology, Northeastern University, Shenyang* 110004)

Abstract Based on the theory of complex networks and the giant data samples authorized by CAIDA, a research of real Internet samples is performed in view of complex networks theory. Firstly, Internet is regarded as an alive system with fluctuation property according to methods from Physics and Biology, and a formalized definition — Internet Traveling Diameter (ITD) — is put foreword. Secondly, according to features of real samples, three models simulating the development of ITD are put foreword. All three models include two parts, one part is Logistic function, which simulates the basic developing trace of ITD, and the other part is composed of sine and cosine functions, which simulates the fluctuations during the ITD's development. Then Experiments are performed by Float-point GA to train and learn the parameters of the three models, and the most optimized model is selected with a fitting accuracy of 97.87% and a forecast accuracy of 97.47%. Finally, a forecast of Internet development is performed and a conclusion is made that Internet would develop at top speed from now on to Dec. 2011, and after that the speed would slow down and become stable in Oct. 2021 when ITD would be 10.2073 hops. What's more, a forecast of Internet in Beijing Olympic Games in Aug. 2008 is performed and results are yielded that ITD would be 10.7726 hops and Internet validity would be in a high state at that time.

收稿日期:2005-05-20;**修改稿收到日期**:2006-02-15.**本课题得到国家"八六三"高技术研究发展计划项目基金**(863-317-01-04-99,2001AA415320)**资助.****徐 野**,男,1976 年生,博士研究生,主要研究方向为复杂网络、信息融合、嵌入式技术.E-mail:xuy.mail@gmail.com;xuy.mail@163.com.**赵 海**,男,1959 年生,教授,博士生导师,主要研究领域为复杂网络、信息融合、嵌入式技术.**苏威积**,男,1975 年生,博士研究生,主要研究方向为复杂网络、信息融合.**张文波**,男,1973 年生,博士研究生,主要研究方向为复杂网络、嵌入式技术.**张 昕**,男,1979 年生,博士研究生,主要研究方向为复杂网络、嵌入式技术.

Keywords　complex networks；traveling diameter；Internet physical property；Logistic model；genetic algorithms；float-point genetic algorithms

1　引　言

继对随机网络、小世界网络和无尺度网络的理论研究之后，复杂网络研究开始逐渐转向真实的复杂网络——Internet. 文献[1～4]的研究表明，Internet 具有小世界、无尺度和随机网络的特征. 目前复杂网络领域较常见的研究方法主要是国外某些统计物理学家使用统计和拟合的方法(拟合出幂率曲线求幂率值)，对演员网络、科学家网络等进行幂率和无尺度特性的研究；此外是借助几种经典传播模型，对复杂网络的传播特性的研究. 此外，一些文献上提到了对电力网、WWW、Internet 拓扑结构等内容的研究[5]. 其研究方法是首先从某角度定义反映目标网络特性的概念，然后根据数据或仿真实验结果对概念进行量化分析，从而得出关于目标网络某种特性的结论. 如一些文献对某些实体复杂网络定义了类似“网络直径”、“度”、“介数”、“流量熵”、“结构熵”等概念，然后根据数据量化这些概念，得出关于目标网络特性的结论.

本文研究方法与此稍有不同. 一是我们根据 CAIDA 机构授权取得了关于 Internet 网络的真实海量数据；二是我们借助物理学和生物学研究的方法，将 Internet 网络视为具有生命涨落特征的活体系统，从定义其物理特征量(如 Internet 网络密度、温度、直径、分割度等)出发，基于 CAIDA 海量数据所反映的目标网络的某些特性，主要采用拟合与回归的方法，得出其物理特征属性的动态演化模型. 本文即是 Internet 网络物理特征量研究的一部分，将基于海量 Internet IP 层数据，使用统计的方法，定义 Internet 网络的访问直径并对其进行初步研究.

1.1　访问直径的定义

我们首先定义“访问直径”的概念. 一些文献上也提到与“访问直径”类似的定义，如“网络直径”，它从图论角度出发，将“网络直径”定义为：网络拓扑中任意两点间最短距离中的最大值. 但复杂网络(尤其是真实的 Internet 网络)的拓扑结构中结点数和边数大大增多，使用传统的图论算法(如最短路径算法)的复杂度也将急剧增加，成为强 NP 难题.

为更真实地表现 Internet 网络，本文从统计的角度出发，基于表现 Internet 网络客观属性的路由跳数，给出如下定义.

定义 1.　在 Internet 网络中，如果数据包从源 IP 地址到目的 IP 地址所经过的路由跳数称为一次访问的直径，那么，大量数据包从 Internet 中任一源 IP 地址到任一目的 IP 地址所经过的路由跳数的统计均值称为 Internet 网络访问直径，简称访问直径. 设一个数据包路由转发的路由跳数为 J_i，样本总量为 n，跳数 J_i 出现的频数为 F_i，统计频率为 p_i，则访问直径 D 为

$$D=\sum_{i=1}^{n}J_ip_i=\frac{1}{n}\sum_{i=1}^{n}J_iF_i \tag{1}$$

作为真实存在的网络，其数据包在网络上转发，直观地讲，如果数据包所经的路由跳数越少，那么，数据包因路由器故障而引发的丢包率就越小，数据包到达目的 IP 地址的时间可能就越短(当然还取决于链路状态)，因此，数据包路由跳数在一定程度上反映了 Internet 网络转发数据包的有效性，也就是说，如果数据包能被更少的路由转发而到达目的地，那么网络也就更有效. 因此对基于路由跳数的 Internet 网络访问直径研究是具有实际意义的.

1.2　数据样本

本文数据样本选自 CAIDA①，主要采用了来自北美洲美国的圣地亚哥(riseling)、欧洲荷兰的阿姆斯特丹(k-peer)和亚洲的日本东京(apan-jp)3 个结点②的五年间③(从 1999 年 7 月～2004 年 6 月)近 7500 万条数据. 此外，为更好地说明访问直径的特性，本文亦引用了另外几个结点的数据.

① CAIDA(The Cooperative Association for Internet Data Analysis)是一个对全球范围 Internet 结构及数据进行研究的国际合作机构. 研究的主要内容包括 Internet 网络的产生、发展及演化趋势以及 Internet 网络行为、动力、网络传播特征和 Internet 宏观拓扑结构的变化规律.
CAIDA 在世界范围内的参与者共有 30 余家，主要分布在北美洲、欧洲的许多国家中的研究院所、军事机构及高等学府中. 亚洲仅有三家，其中两家在日本东京，另外一家在中国东北大学复杂网络研究中心.

② 之所以选取 CAIDA 在北美洲、欧洲和亚洲 3 个结点，主要是因为分别从三大洲抽取的数据能更客观地表现 Internet 网络. 而每洲仅取 1 个结点，是在保证了数据冗余度的基础上简化了统计分析的复杂程度.
没有选取亚洲中国结点的原因是，中国结点成立的时间较短，数据时间跨度不长.

③ 每月抽取当月 15 日一天的数据. 3 个结点分别存在缺失数据的情况，其中 riseling 缺少 2001 年 7 月～9 月的数据，k-peer 缺少 1999 年 7 月～2001 年 5 月的数据和 2004 年 1 月的数据，apan-jp 缺少 1999 年 7 月、2000 年 6 月和 2003 年 10 月的数据. 本文中，一般不对缺失数据做补充，在计算需要情况下，使用另两月同时段数据的均值代替.

本文根据数据包能否到达目标 IP 地址将样本分成两类：一类是可达的，由 C 表示，即数据包所经由的路由链路是完整的；另一类是不可达的，由 I 表示，即在 Internet 传输中，被路由器舍弃的那些数据包，这类数据包返回源地址时所载的信息是无效的，路由链路是不完整的. 因此，在本文研究中，将忽略路径不可达的数据样本，只考虑路径可达的数据样本，并将之称为可达样本. 且表 1 统计分析(可达样本约占总样本的 54.98%，共计超过 4100 万条)表明，即使忽略不可达样本，仍可保证数据的冗余性.

表 1 可达样本统计分析结果

节点	C	I	$C/(C+I)\times 100(\%)$	总计
riseling	16448329	13669728	54.6	30118057
k-peer	9479028	9555955	49.8	19034983
apan-jp	15172408	10423192	59.3	25595600
总计	41099765		54.98	74748640

1.3 访问直径的直观特性

为了更客观地描述 Internet 网络访问直径的直观特性，本文将分别从空间维和时间维对其进行统计分析.

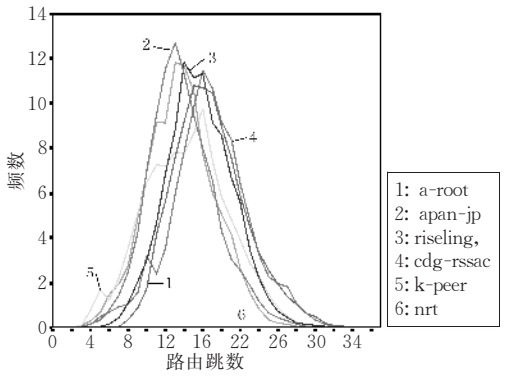
空间维上，本文在已有 3 个结点的基础上，分别从三大洲多选一个结点(包括美国弗吉尼亚的 a-root、法国巴黎的 cdg-rssac 和亚洲日本的 nrt)，从这 6 个结点抽取 2003 年 12 月 15 日一天数据共 1198204 条，进行统计分析，分析结果见图 1(a).

在时间维上，本文分别对 riseling，k-peer 和 apan-jp 进行从 2003 年～2004 年的时间维统计分析，分析结果相近. 不失一般性，本文对 riseling 的分析结果制图，见图 1(b).

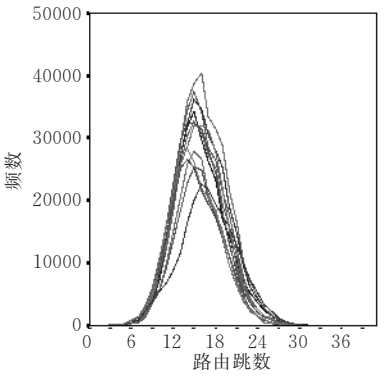
从图 1 可以看到，曲线波峰集中地收敛于[10,18]之间，表明高冗余的数据表现出统一的特征；此外，跳数 J 的全幅跨度约在 2～32 跳之间，由于下文中访问直径演化模型需要的是跳数上下区间，要考虑误差因素，因此本文不妨将其设大一点：(1,36)，以确保区间的完整性. 对于访问直径 D，根据式(1)，D 不可能超出上述区间，因此有

$$D\in(1,36) \tag{2}$$

本文从三大洲选取 6 个结点，从 riseling 选取长达一年的数据，保证了对访问直径分析的客观性. 同时，1198204 条空间维数据和 3270289 条时间维数据，又充分保证了分析的可信度. 因此，本文认为图 1 的分析结果可接受.



(a) 从北美洲、欧洲、亚洲选取具有代表性的 6 个结点、从中分别抽取 2003 年 12 月 15 日当天数据共 1198204 条，对 Internet 网络访问直径进行空间统计分析的结果



(b) 从 riseling 抽取 2003 年 7 月～2004 年 6 月的 3270289 条数据、对 Internet 网络访问直径进行基于时间维统计分析的结果(每月一条曲线)

图 1 Internet 网络访问直径的空间及时间统计分析结果(图中横坐标是可达数据包经由路由跳数，即式(1)中的 J，纵坐标是不同访问直径的数据包的频数，即式(1)中的 F)

2 模 型

2.1 模型选择

根据访问直径随时间变化的特性选择模型.

首先按照式(1)对 riseling，k-peer 和 apan-jp 3 个结点的数据计算访问直径 D，然后对 D 按时间变化作图，如图 2 所示.

从图 2 中可以看到：3 个结点的访问直径 D 随时间 t 增加，明显地呈现出一致缩减的变化趋势. 尽管 3 个结点在时间 t 的初始一段区间内 D 差值较大，但在 $t>50$ 后差值明显收缩，在 t 临近 60 处 3 条曲线收敛于一个 D 差值极小的区间.

因此可以判断，3 个结点访问直径 D 随时间呈现出一致的变化规律，本文做出 3 个结点 D 的均值与时间 t 变化图，结果如图 3 所示.

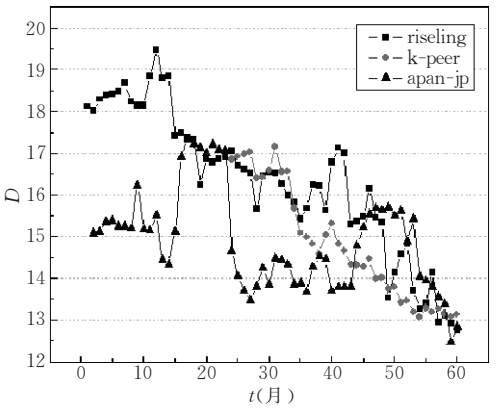


图 2　riseling，k-peer 和 apan-jp 3 个结点访问直径 D 随时间 t(1999 年 7 月～2004 年 6 月共 60 个月)变化情况(图中 3 个结点存在某些缺失数据，一般不对缺失数据做补充，如 k-peer 结点缺 1999 年 7 月～2001 年 5 月的数据；在计算需要情况下，缺失数据使用另两月同时段数据算术均值代替，如 riseling 结点 2001 年 7 月～9 月的数据)

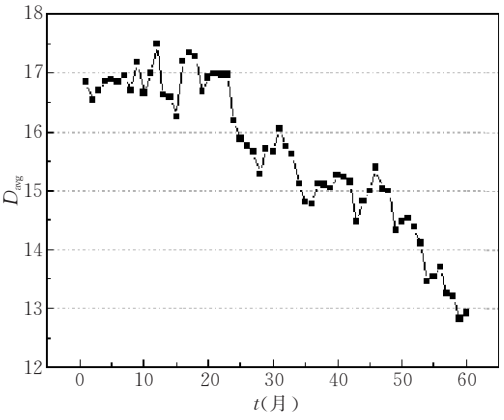


图 3　riseling，k-peer 和 apan-jp 3 个结点访问直径 D 的均值 D_{avg} 随时间 t(1999 年 7 月～2004 年 6 月共 60 个月)变化情况

从图 3 中可以明显地看出 D 均值随时间的变化规律，与 Logistic 方程研究的随时间生长变化规律相似，根据文献[6]，本文将选择基于 Logistic 方程的改进模型对图 3 中数据进行拟合，以得到访问直径 D 关于时间 t 的数学模型.

2.2　改进 Logistic 模型

Logistic 方程[7,8,11]是由 Verhulst 提出的研究在有限资源的条件下的人口发展模型，将其应用于本文数据，得到式(3)所示的非线性微分方程

$$\frac{\mathrm{d}D}{\mathrm{d}t}=rD\left(1-\frac{D}{k}\right) \tag{3}$$

其中 $D(\geq 0)$ 是 t 时刻访问直径 D 值，$t(>0)$ 为时间(月)，$k(>0)$ 是访问直径极限值量，$r(>0)$ 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. 本文对式(3)积分，得

$$D=\frac{k}{1+\dfrac{k}{D_0-1}\mathrm{e}^{-rt}}=\frac{k}{1+m\,\mathrm{e}^{-rt}} \tag{4}$$

其中 $D_0=D(t=0)$，$m=\dfrac{k}{D_0-1}$.

(1) 变换 1

标准的 Logistic 方程是单调递增函数，以式

$$y=\frac{k}{1+\mathrm{e}^{-t}},\ m=1,\ r=1,\ k=1 \tag{5}$$

为例，做其曲线如图 4(a)所示. 将式(5)沿 x 轴反转，得单调递减函数，如图 4(b)所示；再沿 y 轴向上平移 k，得单调递减且无限趋近 0 的函数，如图 4(c)所示；最后沿 y 轴再向上平移 k，得到单调递减且无限趋近 k 的函数，如图 4(d)所示.

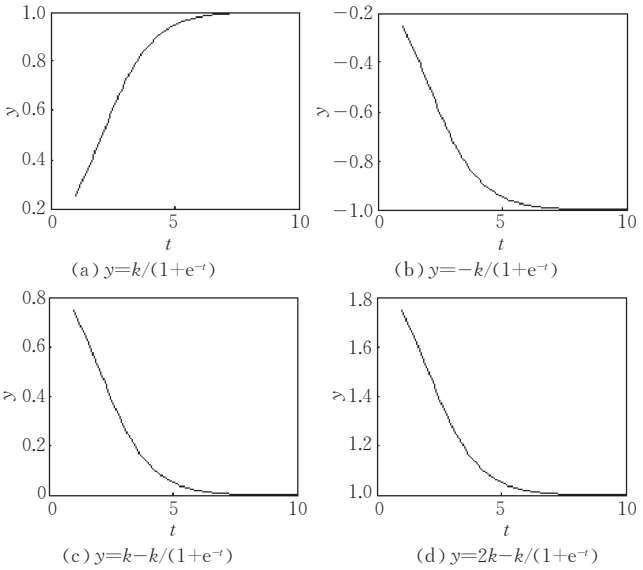


图 4　以 Logistic 方程 $y=k/(1+\mathrm{e}^{-t})$ 取 $k=1$ 为例，对其形式进行变换后形成的图形

图 4(d)所示模型是图 4(b)沿 y 轴向上平移 $2k$ 后得到的结果，符合图 3 访问直径 D 单调递减且趋近某一常数 k 的要求. 但图 4(d)所示模型并不严格要求平移量为 k 的整数倍，所以使用 d 表示平移量，因此对式(4)进一步改进，得到 Logistic 方程

$$D=d-\frac{k}{1+m\,\mathrm{e}^{-rt}} \tag{6}$$

其中 d 为沿 y 轴的平移量.

(2) 变换 2

图 3 所示曲线与 Logistic 的另一个不同点是，Logistic 曲线是光滑的，而图 3 曲线是在振荡中衰减的. 仔细观察图 3，曲线振荡具有准周期特性，这种振荡可以由基于正、余弦的函数或其组合描述，因此需要为经过一次改进的 Logistic 模型增加模拟振荡的正余弦模型[6]，实现另一次改进.

如果图 3 中曲线只是一种单周期振荡,那么单独的正弦函数足以拟合;否则,就需要组合正、余弦函数.但是,本文只能大略观察出曲线振荡具有准周期性,而无法确定曲线振荡到底具有什么样的周期性,因此本文在文献[6]方法的基础上,选用了 3 种振荡拟合模型——正弦函数、正、余弦相乘和正、余弦相加.选择上述 3 种模型的意义:(1)简单性,即模型不过于复杂而难于计算;(2)充分性,即在满足简单性原则的基础上,尽可能保证拟合出充分复杂的振荡曲线,正、余弦函数的相加和相乘即是如此.

将 3 种模型代入式(6)得到经过第二次改进的 Logistic 方程,分别表示为式(7)~(9):

$$D=d-\frac{k}{1+m\,\mathrm{e}^{-(v+p\mathrm{e}^{-gt}\sin[\pi(\frac{t}{h}+u)])t}} \tag{7}$$

$$D=d-\frac{k}{1+m\,\mathrm{e}^{-(v+p\mathrm{e}^{-gt}\sin[\pi(\frac{t}{h_1}+u_1)]\cos[\pi(\frac{t}{h_2}+u_2)])t}} \tag{8}$$

$$D=d-\frac{k}{1+m\,\mathrm{e}^{-(v+p\mathrm{e}^{-gt}\sin[\pi(\frac{t}{h_1}+u_1)]+p_2\mathrm{e}^{-g_2t}\cos[\pi(\frac{t}{h_2}+u_2)])t}} \tag{9}$$

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 为相对振荡幅值,$h(h_1,h_2)$为振荡半周期长度(月),$u(u_1,u_2)$为振荡初始幅角,g 为振荡衰减系数.注意,式(6)中的参数 r 已经被式(7)~(9)中的 v 和 $p(p_2)$吸收.

对上述 3 种振荡模型,不仅要从三者之间选优,还要用实际数据检验模型预测的准确度.因为很可能 3 种模型都不能很好地拟合本文数据,而需要更复杂的正余弦组合.这些猜测都需要通过实验验证,我们将在实验结果与分析部分讨论.

2.3 拟合算法

本文使用浮点型遗传算法[6,9,10]作为拟合方法,算法如下.

输入:(list) /* Internet 访问直径时间序列(1999 年 7 月~2004 年 6 月) */

1. /* 初始化 */

1.1 $x=(d,k,m,p,g,h,u,v)$; /* 初始化遗传基因个体,基因体被表示为由待拟合的参数组成的矢量 */

1.2 *initGroup*(); /* 随机生成初始群体,群体大小为 N,本文取 $N=100$ */

1.3 *defineFunc*(*list*); /* 定义适应度函数:预测模型的访问直径 $D(t)$与实际值 $D^*(t)$的拟合最好,即从 1999 年 7 月~2004 年 6 月共 60 个月的 $D(t)$与 $D^*(t)$累计差的绝对值最小 */

2. repeat

2.1 /* 选择:选择遗传基因 */

doSelect(*list*); /* 从基因群体中选择最优的个体进入下一代循环操作;根据适应度函数计算结果确定评优的标准,适应度最强的基因个体可以被优选到下一代遗传计算 */

2.2 /* 杂交:对选择后留在群体中的遗传基因,进行杂交操作 */

doCrossover(*list*);

2.3 /* 变异:对杂交后的基因以较低的概率进行变异 */

doMutation(*list*);

2.4 until(个体适应度超过预定值 ‖ 遗传代数大于 50000)

2.5 /* 算法结束,取最优基因个体 x,确定模型形式 */

2.4 拟合结果

本文在对模型(7)~(9)拟合之外,为作比较,对原始 Logistic 模型(6)也使用相同算法进行了拟合,拟合结果见表 2.

表 2 4 个模型的拟合结果

模型		拟合结果		
		组 1	组 2	组 3
(6)	d	17.217682	17.305021	17.249234
	k	8.530387	10.000000	7.568359
	m	27.957824	27.785511	21.270079
	r	0.054708	0.049706	0.054579
	分数①	0.032529	0.032174	0.030638
	轮数②	16735	19951	15311
(7)	d	17.070376	17.443896	17.253545
	k	7.289685	10.175583	9.074355
	m	26.082255	25.948223	19.199425
	p	0.001831	0.003536	0.000928
	g	0.002402	0.033085	0.016955
	h	0.001000	0.168535	0.003114
	u	0.044763	0.076170	0.001675
	v	0.052811	0.052056	0.041737
	分数	0.046393	0.045637	0.044138
	轮数	29608	27486	45913
(8)	d	16.945643	17.074561	17.274702
	k	5.823017	7.862857	9.894616
	m	46.732679	47.204944	38.551788
	p	0.002553	0.002130	0.002047
	g	0.004033	0.000832	0.029910
	h_1	0.006077	0.002267	0.000619
	u_1	0.004484	0.002008	0.003771
	h_2	0.004258	0.009444	0.001924
	u_2	0.005146	0.001255	0.001605
	v	0.075158	0.064965	0.056829
	分数	0.050375	0.050142	0.050319
	轮数	20762	34565	30185
(9)	d	17.874506	17.424018	16.949545
	k	12.924366	11.961684	6.826200
	m	15.680552	27.153925	58.241775
	p	0.003642	0.002431	0.001218
	g	0.015851	0.003184	0.001191
	p_2	0.002560	0.000789	0.082215
	g_2	0.053497	0.004985	0.004712
	h_1	0.001746	0.004994	0.000866
	u_1	0.006999	0.001820	0.014911
	h_2	0.057158	0.043464	0.001000
	u_2	0.052713	0.001312	0.003242
	v	0.036473	0.045700	0.007844
	分数	0.050667	0.052162	0.051604
	轮数	25901	32311	38068

① 分数指根据适应度函数计算(取适应度函数计算值的倒数)得到的最高优度评分,是算法在 50000 轮遗传计算中的最高分数.

② 轮数指模型收敛到最高分数时遗传计算的轮数.

3　选择模型

(1)淘汰模型(6).

从表 2 看到,模型(6)轮数明显小于其它 3 个模型,考虑到(6)的参数也明显少于其它模型,其收敛速度最快是合理的. 但是,由于模型(6)是平滑的 Logistic 曲线,无法表示图 3 所示原始数据的波动振荡,因此其优度评分明显低于其它模型,将被淘汰.

(2)淘汰模型(7).

表 2 中对模型的评分是基于遗传算法适应度函数(从 1999 年 7 月～2004 年 6 月共 60 个月的 $D(t)$ 与 $D^*(t)$ 累计差的绝对值最小)进行计算. 为做出明确判决,使用式(10)求平方和来放大评分结果.

$$f(d,k,m,p,g,h,u,v)=\sum_{i=1}^{60}[D(t_i)-D^*(t_i)]^2 \tag{10}$$

式(10)计算结果如图 5 所示.

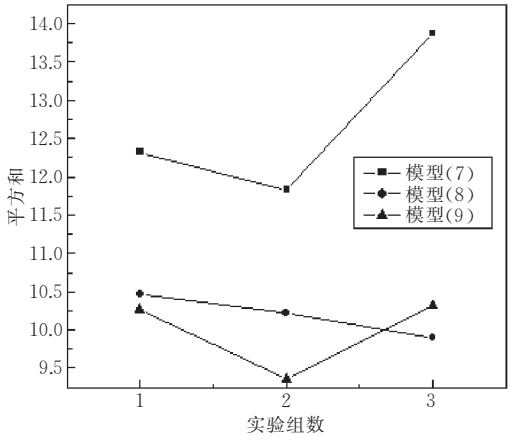


图 5　模型(7),(8),(9)三组实验根据式(10)计算结果

从图 5 可见,根据式(10)对评分结果放大之后,明显看出模型(7)与模型(8),(9)的差别,因此支持淘汰模型(7)的判断.

(3)从模型(8),(9)中选择.

定义 2.　设本文中访问直径原始数据值为 $D^*(t_i)$,模型计算值为 $D(t_i)(i\geq 1, i\in N)$,则平均误差为

$$\bar{\varepsilon}=\frac{\sum_{i=1}^{n}|D(t_i)-D^*(t_i)|}{n} \tag{11}$$

其中,n 为按月计算的时间跨度.

平均误差表示了误差的绝对大小,根据式(11)分别计算 6 组实验结果的平均误差,可以判断 6 个模型的优劣. 经过计算表明:实验组 6(即模型 9 的实验组 3)误差低于其它实验结果. 因此确定本文的优选模型为

$$D=16.949545-\frac{6.8262}{1+58.241775\times e^{-(A+B)t}} \tag{12}$$

其中,$t>0$, $t\in N$, A, B 为

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

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

4　评　价

4.1　优选模型的量化分析

根据客观实践,如果一个模型拟合准确度超过 95%,而预测准确度大于 80%,那么这个模型是可接受的. 因此,针对 3 种模型,尽管分别仅做三组实验,实验空间并不完备,但某些情况下(如资源受限或空间不可穷尽),找出完备的空间和最优解是不可能的,也是不需要的,而可能需要的是非完备空间中的一个次优或局部最优的解,只要该解能够满足上述客观判定. 因此,本文优选的模型可能只是次优解,但只要能够在拟合的准确度和预测的准确度两方面通过进一步验证,便是可接受的.

式(11)的 $\bar{\varepsilon}$ 表示的是误差的绝对大小,不能表示误差的相对大小,在评价模型时是不完备的. 例如,对于 10^3 和 10^1 数量级数据,$\bar{\varepsilon}=1$ 的意义是完全不同的. 因此引入相对平均误差 $\bar{\varepsilon}_r$ 概念.

定义 3.　设本文中访问直径原始数据值为 $D^*(t_i)$,模型计算值为 $D(t_i)(i\geq 1, i\in N)$,则相对平均误差 $\bar{\varepsilon}_r$ 为

$$\bar{\varepsilon}_r=\frac{\bar{\varepsilon}}{\overline{D^*}}=\frac{\frac{\sum_{i=1}^{n}|D(t_i)-D^*(t_i)|}{n}}{\frac{1}{n}\sum_{i=1}^{n}D^*(t_i)}=\frac{\sum_{i=1}^{n}|D(t_i)-D^*(t_i)|}{\sum_{i=1}^{n}D^*(t_i)} \tag{13}$$

其中,n 为按月计算的时间跨度.

(1)计算选定模型的拟合准确度.

根据拟合模型的原始数据(1999 年 7 月～2004 年 6 月),使用式(13)计算的结果为

$$\bar{\varepsilon}_r=\frac{\bar{\varepsilon}}{\overline{D^*}}=\frac{1.275278}{60}=0.021255,$$

故选定模型(式(12))的拟合准确度为

$$1-0.021255=97.8745\%,$$

可以接受.

(2)计算选定模型的预测准确度.

根据用于预测的原始数据(2004 年 7 月～2005 年 5 月),使用式(13)计算结果为

$$\bar{\varepsilon}_r=\frac{\bar{\varepsilon}}{\overline{D^*}}=\frac{0.277985}{11}=0.025271,$$

故选定模型(式(12))的预测准确度为

$$1-0.025271=97.4729\%,$$

可以接受.

4.2 Internet 网络访问直径的预测分析

数学模型的意义在于预测,上文通过对 2004 年 7 月～2005 年 5 月时间段的数据对访问直径进行了预测,验证了模型的预测准确性. 现在将对较远未来(350 个月)进行大范围预测,预测结果如图 6 所示.

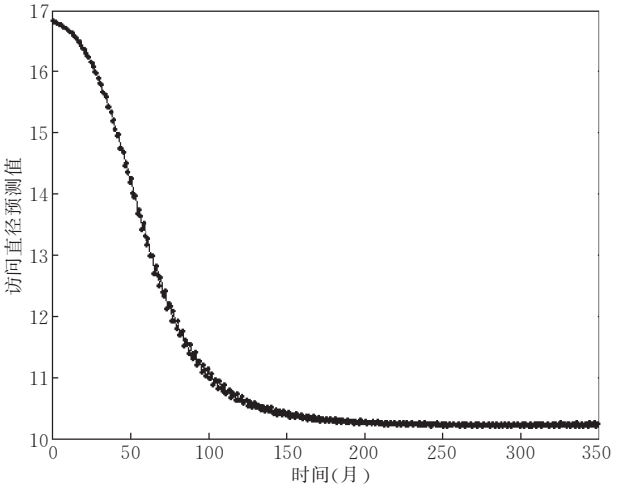


图 6 1999 年 7 月～2028 年 8 月(350 个月)模型预测数据(图中横坐标为时间,从 1999 年 7 月开始为 1,以后递增,到 2028 年 8 月为 350)

(1)跳数衰减的原因

无论从图 2 所示的原始数据,还是从图 6 所示的模型预测结果,都可以看到 Internet 网络访问直径随时间的衰减变化,这也表明 Internet 网络中数据包转发所经路由跳数的统计值在衰减.

实际上,Internet 网络是一个不断发展、逐渐完善的网络,近几年来各国家不断修建光纤骨干网络和许多关键路由结点之间的直达链路,是 Internet 网络路由跳数衰减的主要原因. 例如,假设某数据包从路由 A 至路由 C,在 A 与 C 之间无直接链路时,需要绕路 B,即 $A\rightarrow B\rightarrow C$; A 与 C 之间铺建直接链路后路由路径为 $A\rightarrow C$,节省了路由跳数,这也意味着提高了网络效率.

(2)跳数衰减的极值(稳定值)

首先可以从理论上进行分析. 假设理想情况下,Internet 成为全连通网络,即任意两个结点之间都存在着直接链路,那么根据本文定义,整个 Internet 网络的访问直径为 1. 因为在这种理想 Internet 网络下,数据包经过了 0 次路由,这时已经达到数据包在 Internet 网络上转发的极限,因此访问直径不可能比这再低了,1 是衰减的极值. 但全连通的 Internet 网络只是理想存在的,根据常识,这种网络不可能存在,真实 Internet 网络路由跳数衰减的极值必定是大于 1 的某个值.

从图 6 中可以看到,本文预测模型在时间大于 150 个月时趋近平稳收敛,由于曲线是在振荡和涨落中收敛的,本文取某一近似跳数衰减的值为 10.2073,时间为 268 月(从 1999 年 7 月为 1 月开始计算),即 Internet 网络访问直径将在近 2021 年 10 月趋于平稳,其统计值为 10.2073.

模型计算值为 10.2073 跳,考虑到这是整个 Internet 访问直径的统计值,是可以接受的.

(3)高速衰减区域

图 6 中从第 1 月至约第 150 月(即从 1999 年 7 月～2011 年 12 月,约 12 年)访问直径通过从 17 跳衰减至 10 跳左右. 由此可以预测未来几年将是 Internet 网络的高速发展期.

(4)2008 年奥运期间 Internet 预测

由图 6 可预测,2008 年 8 月(110 月)北京奥运期间 Internet 网络访问直径为 10.7726 跳,网络可用性较高.

4.3 长期预测

根据客观实践,如果使用本文基于 1999～2004 年数据的数学模型对几十年后的 Internet 网络变化进行长期预测,准确度将迅速降低.

文献[12,13]指出,按照传统的方法(如本文拟合模型法)对复杂的目标系统(如本文 Internet)进行长期预测,将会遇到巨大困难. 其原因是模式系统在映射目标复杂系统的过程中,其所有动力模式都只能突出目标系统的某些物理过程,而忽视了其它过程,从而无法全面地描述目标系统的所有物理过程. 因此,当时间增长以后,模式系统的预测准确率将会迅速下降.

传统的一维(一个自变量,如本模型中参数 t)拟合模型仅能表现目标系统的一维物理过程(常见

为时间)，在长期预测中，已不再适用. 但限于可计算性，拟合模型又不能以无穷多维映射目标系统的无穷多物理过程，因此，需要在既满足可计算性，又满足提高长期预测准确率的基础上，确定拟合目标系统的具体模型维数. 根据文献[13]，这需要从混沌学与分形学角度对目标系统做混沌属性的测试，然后计算其关联维数，确定其奇异吸引子的存在结构，从而才能确定预测模型的数学形式. 这也将是作者今后研究的重要部分.

5　小结与展望

本文创新点在于：

(1)结合了复杂网络理论与 CAIDA 授权的关于 Internet 网络的真实海量数据，从复杂网络理论角度对真实的 Internet 数据进行分析与研究，与传统的网络工程(如网络协议、流量研究等)方法不同，它为揭示 Internet 客观属性提供了更多的方法.

(2)借助物理学和生物学研究的方法，将 Internet 网络视为具有生命涨落特征的活体系统，定义了 Internet 物理特征量，如 Internet 网络访问直径、温度、密度、分割度等. 本文的研究即为上述研究中的 Internet 网络的访问直径分析. 借助物理学和生物学方法定义 Internet 网络的物理特征量的研究方式，目前是较新的.

(3)基于 CAIDA 海量数据所反映的目标网络的某些特性，主要采用拟合与回归的方法，得出其物理特征属性的动态演化模型. 这是将老方法(传统的拟合方法)应用在新问题(复杂网络实例的物理特征量研究)上的一次尝试和创新.

对今后的展望：

(1)根据上述创新点的介绍，本文的研究方向、方法与研究结论在复杂网络研究领域可以说是比较新的. 但是，对 Internet 网络的物理特征量的研究还仅仅是个开始，根据现在的研究结果，尚不能肯定从物理特征量角度对 Internet 网络的研究的正确性；而且，根据当前研究结论对 Internet 网络物理特征量(如访问直径)预测的准确性仍需要未来几年 Internet 网络真实变化情况加以验证. 因此，本文及其相关研究仅是结合复杂网络和真实特例(Internet)进行研究的一次初步尝试，还需要更进一步的工作.

(2)根据混沌系统相关理论，本文一维预测模型在预测长期(几十年甚至上百年)数据时将不再准确，根据相关文献，对此问题的解决方法需要从混沌和分形学入手，确定目标系统的奇异吸引子结构及分维数(关联维数)，然后才能确定目标复杂系统的预测模型形式. 这也将是本文研究的下一步重要工作.

参　考　文　献

1　Floyd S., Paxson V.. Difficulties in simulating the Internet. IEEE/ACM Transactions 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. In: Proceedings of the ACM STOC 2000, Portland, 2000, 171～180

4　Zhang Yu, Zhang Hong-Li, Fang Bin-Xing. A survey on Internet topology modeling. Journal of Software, 2004, 15(8): 1221～1226(in Chinese)
(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　Zhang Jia-Cai, Zhou Deng-Yong. View the large-scale modes of the Internet as an open complex giant system. Journal of System Simulation, 2002, 14(11): 1450～1455(in Chinese)
(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　Yin Chao-Qing, Yin Hao. Artificial Intelligence and Expert System. Beijing: China Water Publication, 2002(in Chinese)
(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)

7　Yang Zhi-Jie, Xu Zhong-Ru. Forecast of the population growth in the country of Heilongjiang by the forecast method of dynamic logistic. Journal of Agriculture University of Heilongjiang, 1997, 9(2): 23～28(in Chinese)
(杨志杰，徐中儒. 用动态逻辑斯谛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，1997，9(2)：23～28)

8　Wu Shu-Ling. Forecast of development of China Numerical Library by logistic model. Journal of Information, 2004, 23(4): 56～57(in Chinese)
(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，2004，23(4)：56～57)

9　Wang Jian-Ming, Xu Zhen-Lin. New crossover operator in float-point genetic algorithms. Control Theory and Applications, 2002, 19(6): 977～980(in Chinese)
(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)：977～980)

10　Rudolph G.. Covergence properties of canonical genetic algorithms. IEEE Transactions on Neural Networks, 1994, 5(1): 96～101

11　Zhang He-Guan. Two new population growth equation. Journal of Bimathematics, 1995, 10(4): 78～82(in Chinese)
(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 Lorenz E. N.. Deterministic nonperiodic flow. Journal of Atoms Science, 1963, 20: 130～141

13 Huang R. S.. Chaos and Its Application. Wuhan: Wuhan University Press, 2000(in Chinese)
(黄润生. 混沌及其应用. 武汉:武汉大学出版社, 2000)

XU Ye, born in 1976, Ph. D. candidate. His current research interests include complex networks, information fusion and embedded technology.

ZHAO Hai, born in 1959, professor, Ph. D. supervisor. His current research interests include complex networks, information fusion and embedded technology.

SU Wei-Ji, born in 1975, Ph. D. candidate. His current research interests include complex networks and information fusion.

ZHANG Wen-Bo, born in 1973, Ph. D. candidate. His current research interests include complex networks and embedded technology.

ZHANG Xin, born in 1979, Ph. D. candidate. His current research interests include complex networks and embedded technology.

Background

Over past few years, the research on complex network has been developed rapidly and extended any science fields, such as biology, physics and social science. It boils down to two reasons: (1) With the improvement of computing capability, people can research various realistic networks including multimillion nodes. It could not be realized in the past; (2) People need to recognize various networks urgently to find out instructional macroscopical-rules. The research discovered that many practical networks have some characteristics of complex networks, and Internet is a typical one.

Supported by CAIDA (The Cooperative Association for Internet Data Analysis), the authors introduce a new method, in view of physical properties, to start research on Internet, and yielded two forecast models which respectively suit short-term forecast and long-term forecast of Internet. The work is sponsored by the National High Technology Research and Development Program (863 Program) of China under grant No. 863-317-01-04-99, 2001AA415320.