

模型参数联合求解的网络流量混沌预测

张志华, 盘炜生

(肇庆学院 教育技术与计算机中心, 广东 肇庆 526061)

摘 要: 流量预测是网络管理的基础, 为了提高网络流量的预测精度, 考虑模型参数之间的相互影响, 提出一种模型参数联合求解的网络流量混沌预测模型. 首先收集网络流量历史数据, 采用混沌理论对历史数据进行重构, 并确定模型参数范围, 然后运用遗传算法模拟自然界的“适者生存、优胜劣汰”机制对模型参数进行联合求解, 根据最优个体得到模型的最合理参数. 最后根据最合理参数对网络流量训练样本进行学习, 建立网络流量预测模型, 并采用仿真实验对模型的性能进行对比分析. 结果表明, 此模型通过混沌理论可以有效挖掘历史数据中的网络流量变化特点, 通过参数联合求解建立了精度高的网络流量预测模型, 为非线性网络流量预测提供了一种新的建模思路.

关键词: 网络流量; 混沌理论; 最小二乘支持向量机; 径向基核函数; 参数优化

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)02-0072-05

Network Traffic Chaotic Forecasting Based on Jointly Solving Forecasting Parameters of Model

ZHANG Zhi-hua, PAN Wei-sheng

(Educational Technology and Computer Center, Zhaoqing University, Zhaoqing 526061, China)

Abstract: Network traffic forecasting is the basis of network management, in order to improve the network traffic forecasting precision, considering the mutual influence between parameters of model, so this paper put forward a network traffic chaos forecasting model based on jointly solving forecasting parameters of model. Firstly, network traffic history data are collected and use chaotic theory reconstruct historical data, and determine the scope of the model parameters, and secondly, genetic algorithm is used to determine most reasonable parameters according to optimal individuals based on nature "the survival of the fittest, superior bad discard" mechanism, finally, the most reasonable parameters are used to establishing network traffic forecasting model by using training sample of network traffic, and the simulation experiments is used to analyze performance of the mode. Results show that the proposed model can accurately mine network traffic network flow change characteristics among historical data by chaos theory and get high forecasting accuracy of network traffic forecasting, it provides a modeling tool for nonlinear network traffic forecasting.

Key words: Network traffic; Chaos theory; Least squares support vector machine; radial basis kernel function; Parameter optimization

1 引言

随着 Internet 和通信技术的迅速发展, 网络上的业务种类不断增加, 网络拥塞的频率越来越高, 网络管理难度越来越大, 而网络流量是网络运营中的最关键数据, 因此对网络流量进行建模与分析显得

格外重要, 尤其在大数据背景下, 网络流量预测成为网络管理研究中的热点^[1-2].

在网络发展初期, 网络流量的预测没有引起人们的重视, 直到上世纪 90 年代, 人们对网络流量的预测才开始重视, 许多研究机构、学校、企业都投身于网络流量的建模与预测中, 取得不错的研究成果,

收稿日期: 2015-03-20; 修回日期: 2015-05-11

基金项目: 广东省科技项目(2012B061700063)

出现了许多类型的网络流量预测模型^[3,4]. 最原始的网络流量预测为用 Poisson 模型, 假设网络流量服从 Poisson 分布, 当用户数量较少时, 其预测精度相当高^[5], 而对于大规模的网络, Poisson 模型就无法准确刻画网络流量的变化特性. 随后有学者提出了基于 Markov 和 ARMA 的网络流量预测模型^[6-7], 它们假设网络流量是一种线性变化系统, 它们仍然无法准备描述网络流量的变化特性, 预测结果不理想. 随着研究的不断深入, 人们发现网络流量受到用户上网习惯、上网价格、节假日等因素的综合作用, 存在非平稳性、突变性, 因此出现了基于支持向量机、神经网络等网络流量预测模型^[8-9], 它们具有自学习能力, 要求先验适应少, 预测精度得到了相应提高. 近年来, 相关研究结果表明网络流量具有一定的混沌性, 有学者提出基于混沌理论的网络流量预测模型, 较好地刻画了网络流量的时变性、非平稳, 预测精度较高. 在网络流量混沌建模过程中, 需要确定预测模型的参数, 当前没有考虑模型各参数之间的相互影响, 而是独立地进行确定, 这样预测结果不稳定, 预测效果有待进一步提高^[10].

为了提高网络流量的预测精度, 考虑模型参数之间的相互影响, 提出一种模型参数联合求解的网络流量混沌预测模型 (GA-LSSVM), 通过网络流量预测仿真对比实验对其性能进行测试和分析.

2 相关理论

2.1 混沌理论

设网络流量的历史样本为 $\{x_k, k = 1, 2, \dots, N\}$, 其中 N 为样本数量, 根据混沌理论可知, 历史样本可以重构成为

$$X_n(m, t) = (X_{1+(n-1)}, X_{1+(n-1)+t}, \dots, X_{1+(n-1)+(m-1)t}) \quad (1)$$

式中, t 是时间延迟; m 是嵌入维.

延迟时间 τ 与时间延迟 t 间的变化关系为

$$\tau = t \times T_s = t/F_s \quad (2)$$

式中, F_s 为采样频率^[11].

2.2 最小二乘支持向量机

最小二乘支持向量机 (LSSVM) 的建模效率要远高于支持量机, 其通过找到描述网络流量历史数据的变化函数进行建模预测, 即有

$$y_i = \mathbf{w}^T \cdot \varphi(x_i) + b \quad (3)$$

式中, $\mathbf{w}$ 和 b 为表示权向量和偏置量^[11].

通过松弛向量 (ξ_i) , 对 $\mathbf{w}$ 和 b 进行, 这样式 (3) 的求解变化为

$$\min \|\mathbf{w}\|^2 + \frac{1}{2}\gamma \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (4)$$

s. t.

$$y_i - \mathbf{w}^T \varphi(x_i) + b = e_i$$

式中, γ 为正则化参数.

直接对式 (4) 进行求解非常困难, 为了简化问题, 提高求解成功率, 引入拉格朗日乘子 (α_i) 得到

$$L(\mathbf{w}, b, \zeta, \alpha) = \frac{1}{2}\mathbf{w}^T \mathbf{w} + \frac{1}{2}\gamma \sum_{i=1}^n \xi_i^2 + \sum_{i=1}^n \alpha_i (\mathbf{w}^T \varphi(x_i) - b + \xi_i - y_i) \quad (5)$$

根据 $\frac{\partial L}{\partial \mathbf{w}} = 0, \frac{\partial L}{\partial b} = 0, \frac{\partial L}{\partial \xi_i} = 0, \frac{\partial L}{\partial \alpha_i} = 0$, 可以得到

$$\begin{cases} \mathbf{w} = \sum_{i=1}^n \alpha_i \varphi(x_i) \\ \sum_{i=1}^n \alpha_i = 0 \\ \mathbf{w} \varphi(x_i) + b + \xi_i - y_i = 0 \end{cases} \quad (6)$$

对于非线性预测问题求解, 需要引入核函数 $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ 得到 LSSVM 估计函数为

$$y_i = \sum_{i=1}^n \alpha_i K(x_i, x_j) + b \quad (7)$$

采用 RBF 函数建立 LSSVM 回归模型, 即有

$$y_i = \sum_{i=1}^n \alpha_i \exp(-\|x_i - x_j\|^2 / 2\sigma^2) + b \quad (8)$$

式中, σ 为 RBF 的参数^[12].

3 参数联合求解的网络流量预测模型

3.1 建模思路

在 GA-LSSVM 的建模过程中, 参数 γ 和 σ 共同决定网络流量预测结果的好坏, 当前通常独立确定 γ 和 σ , 实际 γ 和 σ 互相影响, 不能分开确定, 为此, 本文提出联合求解 γ 和 σ 的思想, 具体流程如图 1 所示.

3.2 具体设计

3.2.1 个体编码

个体编码是遗传算法与问题之间的映射关系, 对于 LSSVM 参数, 每一个个体与一组参数 (γ, σ) 相对应, 这样个体编码方案具体见图 2.

3.2.2 个体适应度函数

在遗传算法, 个体的好坏由适应度函数的值来决定, 其决定着个体的进化方向, 本文采用平均相对预测误差计算个体适应度值, 其计算公式为

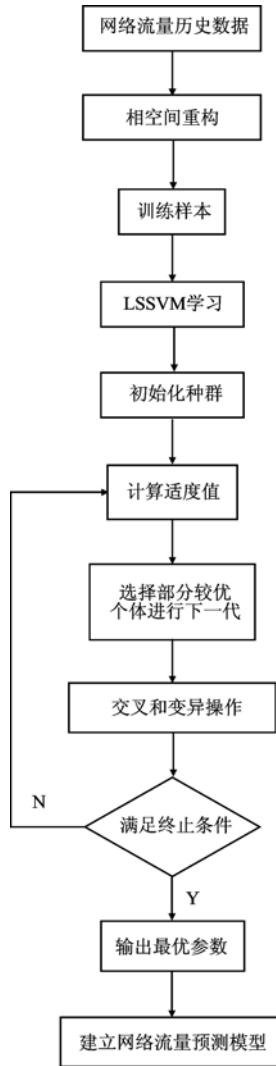


图1 GA-LSSVM 的网络流量建模流程

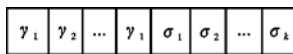


图2 个体编码

$$f(\gamma, \sigma) = 1 - \frac{1}{m} \sum_{i=1}^m \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (9)$$

式中, m 表示样本数.

3.2.3 网络流量的预测步骤

(1) 收集网络流量历史数据, 并根据采用混沌理论对历史数据重构, 挖掘数据之间的联系.

(2) 设置 LSSVM 参数 (γ, σ) 的取值范围, 以及遗传算法相关参数.

(3) 随机产生多个个体, 它们组成了初始种群, 每一个个体由 (γ, σ) 组成, 进化代数 $t = 0$.

(4) 对一个个体进行解码得到 LSSVM 的参数 (γ, σ) , 对训练样本进行学习得到个体的适应度函

数值.

(5) 根据选择概率选择一部分适应度函数值较高的个体直接进入下一代种群.

(6) 根据交叉概率随机部分个体进行交叉操作, 产生一些新个体, 根据新个体与其父个体之间的适应度函数值大小, 取适应度函数值大者进入下一代种群.

(7) 根据变异概率随机部分个体进行变异操作, 产生一些新个体, 根据新个体与其父个体之间的适应度函数值大小, 取适应度函数值大者进入下一代种群.

(8) 进化代数增加 $t = t + 1$.

(9) 如果进化代数超过最大进化代数, 参数搜索过程终止, 否则转步骤(4) 继续重复执行.

(10) 根据最优个体得到最适合的 (γ, σ) , 采用最适合的 (γ, σ) 建立网络流量预测模型.

4 仿真实验

4.1 数据来源

为了分析模型参数联合求解的网络流量预测模型(GA-LSSVM)性能, 选择肇庆学院教育技术与计算机中心的网络流量数据进行仿真实验, 共收集到 800 历史数据点, 选择最后 100 个数据对模型泛化能力进行测试, 所有数据见图 3.

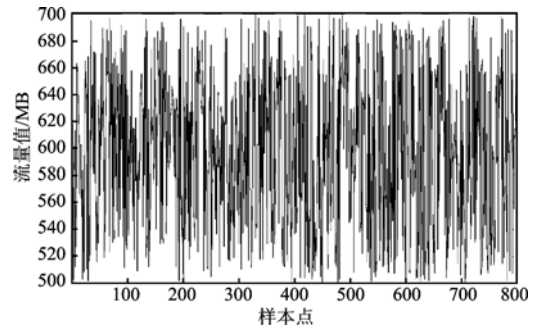


图3 网络流量数据

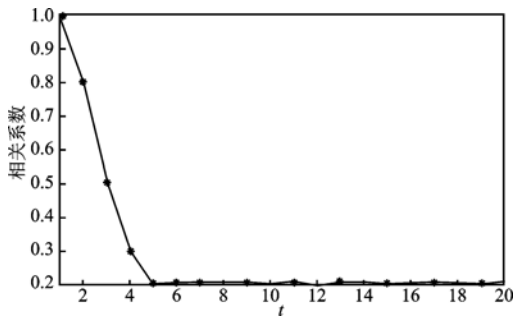
为了使预测结果更加具有说服力, 选择灰色模型(GM)、BP 神经网络(BPNN)、ARIMA 以及模型参数分开优化的最小二乘支持向量机(LSSVM)进行对照实验, 采用 MPAE 和 RMSE 对预测效果进行衡量, 具体为^[13]

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{\hat{y}(i) - y(i)}{y(i)} \right| \times 100\% \quad (10)$$

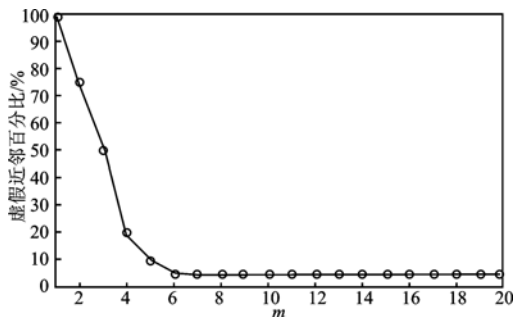
$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}(i) - y(i))^2} \quad (11)$$

4.2 网络流量的混沌处理

网络流量具有一定的混沌性,即看似没有规律,实际隐含有一定的变化规律,该变化规律需求通过确定最合理的延迟时间(τ)和嵌入维数(m)来反映,根据自相关法估计最合理的 τ ,结果如图4(a)所示,从图4(a)我们可以发现,当 $\tau=5$ 时,数据之间相关系数出现了第一个最小值,这就可以确定最合理 $\tau=5$;根据假近邻法估计最合理的 m ,结果如图4(b)所示,从图4(b)我们可以发现,当 $m=6$ 时,假近邻数的百分比不再出现较大波动,变化比较平稳,此时 $m=6$ 为最合理的嵌入维数,根据 $\tau=5, m=6$ 对网络流量的历史数据进行相空间重构,得到有规律的LSSVM建模样本。



(a) τ 与相关系数之间的变化关系

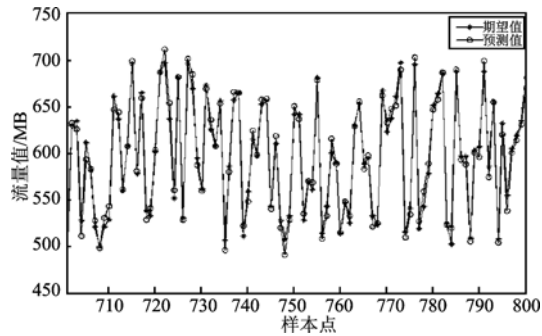


(b) 嵌入维数和虚假近邻数的关系

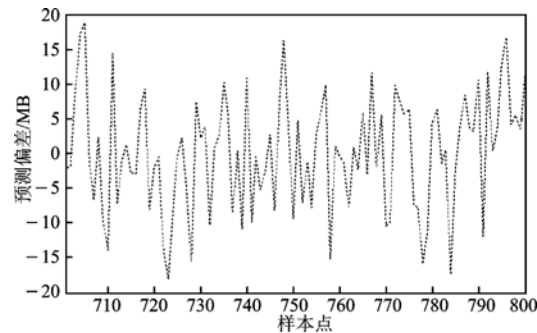
图4 网络流量的混沌分析

4.3 单步预测性能

GA-LSSVM 重构的网络流量训练集进行学习,得到最优参数 $\gamma=89.242, \sigma=1.248$, 根据 $\gamma=89.242, \sigma=1.248$ 对 100 个测试样本进行预测,检验 GA-LSSVM 的泛化能力,得到结果如图 5 所示. 从图 5(a) 的网络流量预测结果可以知道,GA-LSSVM 的预测结果十分理想,可以准确刻画网络流量的非平稳性、时变性变化趋势,同时从图 5(b) 可以看出,GA-LSSVM 的预测偏差变化范围相当的小,可以很好地拟合网络流量的期望值,结果证明了 GA-LSSVM 的有效性。



(a) 预测值与实际值的拟合曲线



(b) 预测值与实际值的偏差变化曲线

图5 GA-LSSVM 的单步预测性能

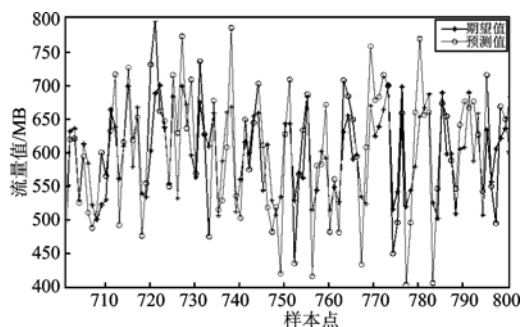
4.4 多步预测性能分析

网络流量数据之间具有长时间相关性,单步预测结果的应用价值有限,由于网络管理员需要了解网络流量将来一段时间变化趋势,因此对 GA-LSSVM 的多步预测性能进行分析,本文进行提前 5 步的网络流量预测,结果如图 5 所示. 从图 6 可以很好了解到,5 步预测精度没有单步预测精度高,而且预测偏差增加幅度大,但 GA-LSSVM 仍然可以描述了网络流量将来的变化趋势,预测偏差处于控制范围内($<10\%$),再一次证明了本文提出模型参数联合优化的思路是可行性。

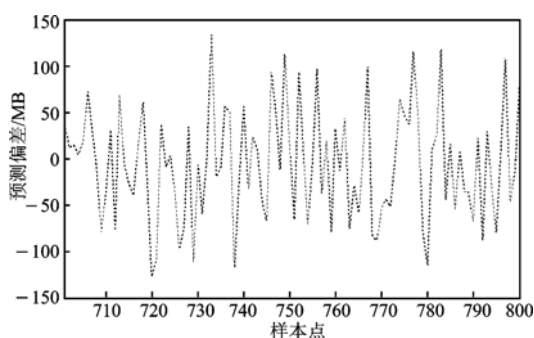
GA-LSSVM 与其他模型单步和多步预测结果的 MAPE、RMSE 见表 1. 从表 1 对比结果可知,GA-LSSVM 的网络流量预测结果的评价指标均最优,对比结果验证了 GA-LSSVM 的优越性。

表1 GA-LSSVM 与其他模型的性能对比

模型	单步		4 步	
	MAPE/%	RMSE	MAPE/%	RMSE
GM	1.773	13.500	16.973	90.288
BPNN	1.545	11.748	16.005	86.976
ARIMA	1.880	14.339	18.904	95.244
LSSVM	1.452	10.292	11.423	73.590
GA-LSSVM	1.148	8.305	8.664	62.653



(a) 预测值与实际值的拟合曲线



(b) 预测值与实际值的偏差变化曲线

图6 GA-LSSVM的多步预测性能

4.5 训练时间对比

随着网络的应用范围越来越大,网络规模更加庞大,网络流量的建模效率越来越关键,因此对不同模型的训练时间进行仿真分析,训练时间变化曲线如图7所示。从图7可以看出,GA-LSSVM的训练时间相对于较少,预测步长越大,建模效率高的优势更加显著,主要因为对LSSVM参数进行联合优化,大幅度减少了支持向量数量,网络流量建模速度加快,GA-LSSVM的适用范围更广,实际应用价值更高。

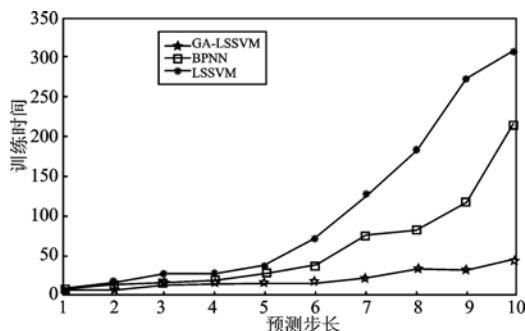


图7 不同模型的训练时间比较

5 结束语

针对当前LSSVM参数独立优化产生的不匹配问题,以提高网络流量预测准确性为目标,本文提出一种新的网络流量建模思路,并通过仿真实验得到

了以下结论:

(1)采用混沌理论确定网络流量历史数据的嵌入维和延迟时间,可以找到隐含其中的变化特点,有利于网络流量建模与预测;

(2)利用遗传算法强大搜索能力对LSSVM的参数进行联合求解,可自动、快速找到最合理的参数(γ 、 σ),建立了预测精度高的网络流量预测模型;

(3)相对于其他网络流量预测模型,本文模型的各项评价指标更优,验证了本文建模思路的有效性和优越性。

为了更好地刻画网络流量的变化特点,需要考虑嵌入维和延迟时间与LSSVM的参数(γ 、 σ)之间关系,以及如何对它们进行优化,这是下一步我们要研究的方向。

参考文献:

- [1] 邱婧,夏靖波,吴吉祥. 网络流量预测模型研究进展[J]. 计算机工程与设计, 2012, 33(3): 865-869.
- [2] 姜明,吴春明,张曼,等. 网络流量预测中的时间序列模型比较研究[J]. 电子学报, 2009, 37(11): 2353-2358.
- [3] 高波,张钦宇,梁永生. 基于EMD及ARIMA的自相似网络流量预测[J]. 通信学报, 2011, 32(4): 47-56.
- [4] 张晓天,刘静娴. 改进的基于小波变换和FARIMA模型的网络流量预测算法[J]. 通信学报, 2011, 32(4): 153-165.
- [5] 姚奇富,李翠凤,马华林,等. 灰色系统理论和马尔柯夫链相结合的网络流量预测方法[J]. 浙江大学学报:理学版, 2007, 34(4): 396-400.
- [6] 刘杰,黄亚楼. 基于BP神经网络的非线性网络流量预测[J]. 计算机应用, 2007, 27(7): 1770-1772.
- [7] 赵振江. 基于PSO-BP神经网络的网络流量预测与研究[J]. 计算机应用与软件, 2009, 26(1): 218-221.
- [8] 魏永涛,汪晋宽,王翠荣,等. 基于小波变换与组合模型的网络流量预测算法[J]. 东北大学学报:自然科学版, 2011, 32(10): 1382-139.
- [9] 罗赞,夏靖波,王焕彬. 混沌-支持向量机回归在流量预测中的应用研究[J]. 计算机科学, 2009, 36(7): 244-246.
- [10] 张颖路. 基于遗传算法优化支持向量机的网络流量预测[J]. 计算机科学, 2008, 35(5): 177-179.
- [11] 王升辉,信学报,裴正定. 结合多重分形的网络流量非线性预测[J]. 通信学报, 2007, 28(2): 45-57.
- [12] 张文金,许爱军. 混沌理论和LSSVM相结合的网络流量预测[J]. 计算机工程与应用, 2013, 49(15): 101-104.

(下转第82页)

计采样算法估算图直径,简化混合式算法中对每个遍历层的条件判断,而无需已知图的结构. 实验证明相对于GPU边并行算法,本文提出的采样混合式算法针对不同结构的图均有较理想的加速比. 针对采样混合式算法的伸缩性测试结果证明当节点个数高于8 192时,采样混合式算法性能高于GPU边并行算法,并随着节点个数的增多,差距加大.

参考文献:

- [1] Floyd R W. Algorithm 97: shortest path[J]. Communications of the ACM, 1962, 5(6): 345-348.
- [2] Watts D J, Strogatz S H. Collective dynamics of small-world networks [J]. nature, 1998, 393 (6684): 440-442.
- [3] Barabási A L, Albert R. Emergence of scaling in random networks[J]. science, 1999, 286(5439): 509-512.
- [4] Johnson D B. Efficient algorithms for shortest paths in sparse networks [J]. Journal of the ACM (JACM), 1977, 24(1): 1-13.
- [5] Cormen T H, Leiserson C E, Rivest R L, et al. Introduction to algorithms [M]. Cambridge: MIT press, 2001:531-539.
- [6] Jia Y. Large graph simplification, clustering and visualization[D]. University of Illinois at Urbana-Champaign, 2010:50-54.
- [7] Saad Y. SPARSKIT: A basic tool kit for sparse matrix computations[J]. 1990,1(3):45-47.
- [8] Merrill D, Garland M, Grimshaw A. High performance and scalable gpu graph traversal[J]. Department of Computer Science, University of Virginia, 2011, 1 (8): 117-128.
- [9] Dice D, Hendler D, Mirsky I. Lightweight contention management for efficient compare-and-swap operations [M]. Berlin Heidelberg:Springer,2013: 595-606.
- [10] Bader D A, Meyerhenke H, Sanders P, et al. Graph partitioning and graph clustering[C]//10th DIMACS Implementation Challenge Workshop. Atlanta, GA: IEEE, 2012.
- [11] Palmer C R, Steffan J G. Generating network topologies that obey power laws [C]//Global Telecommunications Conference [s.l.]: IEEE, 2000: 434-438.
- [12] Leskovec J, Adamic L, Adamic B. The Dynamics of Viral Marketing[J]. ACM Transactions on the Web (ACM TWEB), 2007,1(1):468-476.

作者简介:

李寅女,(1987-),硕士研究生,研究方向为高性能计算. E-mail:liying@huawei.com.

邓仰东男,(1972-),副教授,NVIDIA合作教授、研究方向为高性能计算机体系结构、电子设计自动化和嵌入式系统.

(上接第76页)

- [13] 喻皓,陈志峰. 基于小波分析和Hopfield的网络流量预测[J]. 计算机应用与软件, 2013,30(6):246-249.

作者简介:

张志华男,(1978-),硕士. 研究方向为计算机网络管理、网络安全. E-mail:zhang@ZQU.edu.cn.

盘炜生男,(1964-),高级实验师. 研究方向为计算机应用、信息安全.