

面向 Web 网页的区域用户行为实证研究

马卫东^{1),2)} 李幼平²⁾ 马建国³⁾ 周明天¹⁾

¹⁾(电子科技大学计算机学院 成都 610054)

²⁾(中国工程物理研究院电子工程研究所 四川 绵阳 621900)

³⁾(西南科技大学信息工程学院 四川 绵阳 621010)

摘 要 用户对 Web 网页的访问是由用户需求行为确定的一个随着时间演化的复杂双模式二分网络. 通过对网站聚类生成的二分网络的实证研究表明, 其入度分布呈现出典型的无标度特征和集聚现象, 幂指数介于 1.7 到 1.8 之间. 将这种双模式二分网络映射为两种含权单模式网络: 用户群体兴趣广义关联网络和网站资源广义关联网络, 从而深入研究用户群体行为的关联性和从用户行为角度网站资源的关联性. 实证分析其统计特性表明, 两者的边权分布是幂律的, 网络节点关联紧密且呈现簇聚特征. 用户行为的无标度特征和集聚特点对优化 Internet 网络拓扑结构, 改善其网络性能具有重要意义.

关键词 复杂网络; 二分网络; 无标度网络; 用户需求行为

中图法分类号 TP393

Empirical Study of Region User Behaviors for Web Pages

MA Wei-Dong^{1),2)} LI You-Ping²⁾ MA Jian-Guo³⁾ ZHOU Ming-Tian¹⁾

¹⁾(School of Computer Science and Engineering, University of Electronics Science and Technology of China, Chengdu 610054)

²⁾(Institute of Electronic Engineering, China Academy of Engineering Physics, Mianyang, Sichuan 621900)

³⁾(School of Information Engineering, Southwest University of Science and Technology, Mianyang, Sichuan 621010)

Abstract The Web-visited bipartite networks, called the user interested networks, display a natural bipartite structure: two kinds of nodes coexist with links only between nodes of different kinds. The Web-visited bipartite networks are constructed dynamically in time series by user requirement behaviors, and the characteristic of user requirement collective behaviors can be analyzed through the bipartite networks. The empirical study of the bipartite networks express that the visiting frequency and in-degree distribution are power-law, which the exponential is between 1.7 and 1.8, and the networks have a clustering characteristics. The bipartite networks can be projected to two kinds of affiliation unipartite networks dividedly from the Web nodes and user nodes, and produce the collective interesting affiliation networks and Web resource affiliation networks. The empirical results express that the edge weight distribution of the unipartite affiliation networks is power-law, and the nodes relations are tightness and clustering. The scale-free and clustering characteristics are important to optimize resource distribution and improve the topology structure and performances of Internet.

Keywords complex networks; bipartite networks; scale-free networks; user requirements behaviors

收稿日期: 2007-03-25; 最终修改稿收到日期: 2008-10-08. 本课题得到国家自然科学基金(60272014)和中国工程院信息学部 2006 年度咨询项目资助. 马卫东, 男, 1968 年生, 博士研究生, 副研究员, 主要研究兴趣为复杂网络、信息共享和普适计算技术. E-mail: bobmars@163.com; mwd@uestc.edu.cn. 李幼平, 男, 1935 年生, 研究员, 博士生导师, 中国工程院院士, 当前主要研究兴趣为信息共享工程. 马建国, 男, 1959 年生, 教授, 主要研究兴趣为网络计算和主动服务技术. 周明天, 男, 1939 年生, 教授, 博士生导师, 主要研究兴趣为网络计算和数据挖掘等.

1 引言

复杂网络作为自然界与人类社会事物之间相互关系的一种描述方法和数学抽象得到了广泛的研究^[1-4]. 如以数据业务为核心的 Internet, 按照链路、带宽、流量、节点入度/出度、传输迟延、路由器队列长度等等刻画参数生成了规模宏大的复杂链路网络; 提供资源服务的万维网, 按照网页、网站等资源链接分布也形成了复杂资源网络. 当前对复杂网络主要从模型和实证两个角度研究其性质和动力学规律. 对于实证研究来说, 全局性细致的观察与度量研究非常困难, 因此通常采用基于局域数据经验型测度和基于简化条件的生成模拟研究. 在模型方面, 则主要关注复杂网络的生成机理、分布规律和动力学特性. 国内外有很多研究小组都做出了有影响的工作: Internet 路由器网络^[5-7]、万维网网页(站点)链接网络^[8-12]、科学家合作网络^[13-15]、中国城市航空网^[16-17]等. 尽管这些网络发展包含许多随机因素, 实证研究表明其结构呈现出微观世界不能解释的宏观特性: 无标度拓扑^[5-12]、小世界现象^[1-2]等等. 无标度拓扑的幂律分布是大量高变化参量对最大化、边缘化及混合化的演化结果, 是用户需求行为中马太效应的体现.

随着网络用户数量的快速增长, 用户需求行为对网络流量分布造成的影响也越来越不可忽视. ISC(Internet Society of China)的最新统计数据表明, 中国流量排名第一的站点吸引了 31% 的网民点击, 达到每人每天 6.8 次的平均点击率. Internet 上关联着的链路带宽、分组流量、传输迟延、路由器队列长度等等许多刻画参量, 都受到用户需求和信息分布的影响. 因此 Internet 的动力学特性完全受到用户需求行为的影响.

本文通过实验分析手段研究区域用户对信息资源访问行为的统计特性, 即 Internet 用户对 Web 网页访问的实验统计特性. 由于网络用户和网站资源数量巨大, 基于区域的实证研究是一种可行的办法. Internet 用户需求行为确定了一个随着时间演化的用户对网页访问的复杂二分网络, 属于双模式节点的用户兴趣网络. 实证研究表明该网络的频度特性接近幂律分布, 其度分布服从幂律分布, 幂指数介于 1.7~1.8 之间. 将这种双模式二分网络映射为两种含权单模式网络: 面向用户的群体兴趣广义关联网络和面向网站的资源广义关联网络, 实证分析表明,

两者的边权分布是幂律的; 群体兴趣广义关联网络的节点关联紧密且呈现簇聚特征; 网站资源广义关联网络节点关联较为松散, 但也呈现出簇聚特征.

面向 Web 的基于区域用户行为的时域动态二分网络的统计特性研究国内外尚未见报道.

2 复杂网络模型、概念与属性

2.1 复杂网络概念与模型

当前对复杂网络的研究主要从随机图、层次结构、无标度拓扑、小世界现象等范畴出发.

在数学上网络 $G=(V, E)$ 是指由一个点集 $V(G)$ 和一个边集 $E(G)$ 组成的一个图. 网络中节点的度是其最简单也是最重要的刻画单位. 复杂网络的度分布函数 $P(k)$ 定义为: 随机选择一个节点, 其度数恰好为 k 的概率. 另一种表示度分布的方法是绘制累积度分布函数(cumulative degree distribution function), 它表示的是度不小于 k 的节点的概率分布, $P_k = \sum_{k'=k}^{\infty} P(k')$.

网络的平均度值定义为

$$\langle k \rangle = \sum_{x \in V} \deg(x) / N \quad (1)$$

规则网络所有节点度值相同, 度分布为 Delta 分布 $\delta(k-k_0)$.

随机网络定义为 N 个节点构成的图 G 按照给定概率 p 连接节点边所生成的 $G(N, p)$ 网络, 也称为 ER 网络. 随机网络是对 Delta 分布的扩展. 随机网络随着 p 从 0 增长到 1 形成了从空图到完全图的所有变化过程, 其几何性质需要对每一种可能构成做统计平均. 随机网络演化所生成的网络往往是高度民主的, 绝大多数节点的连接数目会大致相同. 在 $G(N, p)$ 网络中, 任意一个节点 v 与其它节点按照概率 p 相连, 相当于节点 v 对连接与不连接进行 $n=N-1$ 次重复实验, 其度分布服从二项分布:

$$P(|v| = k) = P_n(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (2)$$

设 $np_n = \lambda > 0$ 是常数, 则当 n 趋于无穷大时有

$$\lim_{n \rightarrow \infty} P\{|v_n| = k\} = \lambda^k e^{-\lambda} / k! \quad (3)$$

即当 n 充分大时, 随机网络 $G(N, p)$ 节点度值的分布方式符合 $np = \lambda$ 的 Poisson 分布, 其平均度值 $\langle k \rangle = pn \approx pN$. Poisson 分布对于节点度数 k 是按照指数快速下降的, 远离平均度值的节点几乎不可能存在.

复杂网络的度分布以及其它统计特性如网络直径、平均路径长度、簇系数和边权分布等概念和属性刻画了其基本拓扑特性和演化规律。

复杂网络 G 的直径 $D(G)$ 定义为网络任意一对节点 (x, y) 之间距离的最大值, 即

$$D(G) = \max_{x, y \in V} \{d(x, y)\} \quad (4)$$

平均路径长度 L 定义为网络所有节点对的距离的平均值, 即

$$L(G) = \frac{1}{n(n+1)} \sum_{x, y \in V, x \neq y} d(x, y) \quad (5)$$

其确定了网络中一对节点的典型分离程度。规则网络其平均最短距离为 $\langle d \rangle = (L \times N) / (2 \times C_N^2) \sim O(N)$ 。对实际网络来说, 其平均路径长度通常都不大。文献[18]给出的 Internet 网络访问直径分析和预测值大约 10~11 跳。文献[19]则给出万维网络直径的经验公式为 $\langle d \rangle = 0.35 + 2.06 \log(N)$ 。对于网络规模为 $N = 8 \times 10^8$, 其网络直径 $\langle d_{web} \rangle \approx 18.59$ 。

在网络中, 节点 x 的邻接点 $\partial\{x\}$ 构成的连通子图表明了网络中“簇”的存在。簇系数 C 定义为网络中一个节点邻接点集所构成的子图中存在的连接数与其完全图边数的比例。设节点 k 的临界点集 $\partial\{k\}$ 构成的子图所包含的节点数为 n_k , 边数为 E_k , 则节点 k 的簇系数 $C(k)$ 为

$$C(k) = \frac{2E_k}{n_k(n_k - 1)} \quad (6)$$

整个网络的簇系数 C 是所有节点簇系数的平均值。显然, $C \leq 1$, 并且 $C = 1$ 当且仅当网络是完全图。

随机网络的簇系数 $C = p = \langle k \rangle / N$, 平均距离 $L \sim \ln(N) / \ln \langle k \rangle$ 。对于足够大的 N , 当 $p > (\ln N) / N$ 时, 几乎所有的节点都可能存在连接。在随机网络中连接数目比平均数高许多或低许多的节点, 都十分罕见, 很难呈现出簇聚现象, 且平均距离较小。与随机网络不同, 大多数实际网络都有构成一个个簇的趋势, 即大多数实际网络都不是随机生成的。

2.2 复杂网络的幂律分布

近年来的实证研究表明许多实际网络的度分布符合幂律分布, 即拥有 k 个连接的节点数目的概率分布服从

$$P(k) = ck^{-\tau} \quad (7)$$

其中 c 和 τ 均为正常数。由于幂律分布满足概率分布函数的“无标度条件”, 具有与规模无关的特点, 具有幂律分布的网络被称为无标度网络。幂律分布按照度值下降的速度比指数分布慢得多, 因此可能出现拥有少量高度值的节点(集散节点)存在。与随机

网络中连接的民主分布不同, 无标度网络是由集散节点所主控的系统。

Barabási 等提出的 BA 模型^[4]表明无标度网络可用网络的成长性和优先连接性来刻画。BA 模型的生成可用如下算法描述:

- 1. 设少量的初始节点数 m_0 ;
- 2. 加入 1 个新节点 v_{new} , 将之以概率 $\pi(v_i)$ 随机连接 $m \leq m_0$ 条边到已有系统 V 中 m 个不同节点上(成长性); 连接偏好概率 $\pi(v_i)$ 定义为

$$\pi(v_i) = \frac{|v_i|}{\sum_{j \in V} |v_j|} \quad (8)$$

- 3. 继续步 2 直到达到需求节点数 N 。

采用一种称之为平均场分析的方法, 可以得到 BA 模型的解析解^[20]。设 $k_i = |v_i|$, 则上述算法等价于微分动力方程:

$$\partial k_i / \partial t = m k_i / \sum_j k_j \quad (9)$$

节点 v_i 在时刻 t_i 加入系统, 在时刻 t , 所有节点的度数和 $\sum_j k_j = 2mt$, 于是 $\partial k_i / \partial t = m k_i / 2mt$, 由 $P(k) = \partial(P_i(t) < k) / \partial k$, 可解得 $P(k) = 2m^2 t / (m_0 + t) k^3$ 即拥有 k 度的节点的概率分布为

$$P(k) \sim 2m^2 k^{-3} \quad (10)$$

成长性和优先连接有助于解释集散节点的存在: 当新节点出现时, 它们更倾向于连接到已经有较多连接的节点, 随着时间的推移, 这些节点就拥有比其它节点更多的连接数目。

3 面向 Web 的用户兴趣二分网络

3.1 统计建模

为了揭示用户访问 Web 行为的时域统计分布特征, 我们对区域群体用户需求行为进行了实验统计分析。

选择一个确定的区域用户群和适当的时间段, 设所有用户的集合为 $U, |U| = N$; 所有网页的集合为 H ; 网页按照网站归属集聚成一系列网站的集合 W , 即 W 是 H 的一个划分, $|W| = M$ 。在该区域内确定的时间段, 用户对网页的访问表现为 U 对 H 按照时间 T 的映射过程 σ :

$$\sigma: (U \times T) \rightarrow H \quad (11)$$

是一个随着时间 T 瞬间衍变的二分网络。

为了体现数据的来源情况, 将这些被访问的网页按照网站来源归类, 可以形成用户对网站的访问频度映射:

$\sigma: (U \times T) \rightarrow W$ (12)

所有的统计数据可以用 $N \times M$ 阶矩阵 $B = (b_{ij})$ 表示. 其中 b_{ij} 是在时间段 T 内记录的第 i 个用户访问第 j 个网站的统计次数. 于是可得用户 i 的访问量为 $e_i = \sum_{j=1}^M b_{ij}$, 网站 j 的访问量为 $f_j = \sum_{i=1}^N b_{ij}$. 不失一般性, 对所有网站, 按照网站访问量进行不升序排列(频度相同的网站随机排序). 同时, 定义访问累积频度占比为 F_j , 即访问前 j 个网站的频度之和除以总访问量的比. 由于许多网站区域用户群不访问, 不妨设 M 为网站访问频度不为 0 的网站个数. 则有

$f_j > 0, j = 1, 2, \dots, M; f_{M+k} = 0, k = 1, 2, 3, \dots$

$$F_j = \sum_{j=1}^j \frac{f_j}{F_\infty}, j = 1, 2, \dots, M$$
 (13)

访问频度为 k 的网站个数为

$$d_k = \{j; f_j = k\}, \sum_{k=1}^\infty d_k = M$$
 (14)

则用户需求网络的入度分布为

$$P(k) = d_k / M$$
 (15)

3.2 实验数据分析

我们对校园网路由器出口用户访问日志进行了统计分析. 其中分别选择了 2006 年 5 月、7 月各一周的 HTTP 协议数据. 前者总共产生了 1050154 条网络访问记录, 访问了 32578 个网站, 共涉及 8407 个用户(所有源 IP 地址相同的主机认为是同一个用户). 后者总共产生了 1641879 条网络访问记录, 访问了 31581 个网站, 共涉及 23178 个用户(见表 1).

表 1 校园网访问情况(5 月、7 月各一周数据)

数据来源	链接总数	被访问网站数	用户数	前 800 个网站链接数	所占比例/%
5 月	1050154	32578	8407	766744	73.01
7 月	1641879	31581	23178	1356439	82.62

两组数据按照 Web 序号和访问频度由低到高排序, 网站访问频度测量结果参见图 1.

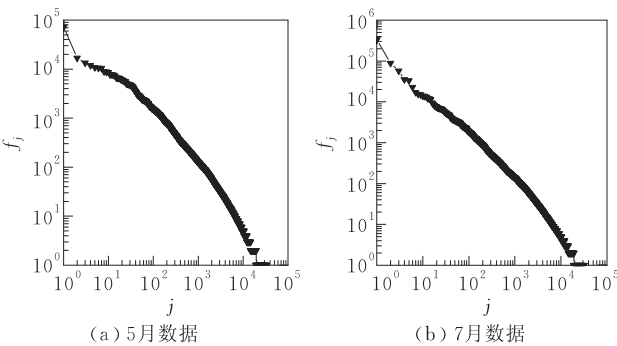


图 1 用户访问网站频度统计

由图 1 可以看出, 两组用户需求访问频度在双对数坐标系上均显示出近似幂律分布的特征. 因此具有最高访问量的网站虽然比例很小, 但是访问占的比例却很大(参见图 2). 据 CNNIC 公布的统计数据, 截止到 2006 年 6 月 30 日, 我国网民人数达到了 1.23 亿, 网站总数达 788400 个, 上网计算机总数约 5450 万台. 我们考察具有最高访问量的网站总数的千分之一(约 800 个网站)的链接访问量情况, 发现其大约占有总体网站访问量的 70%~80%.

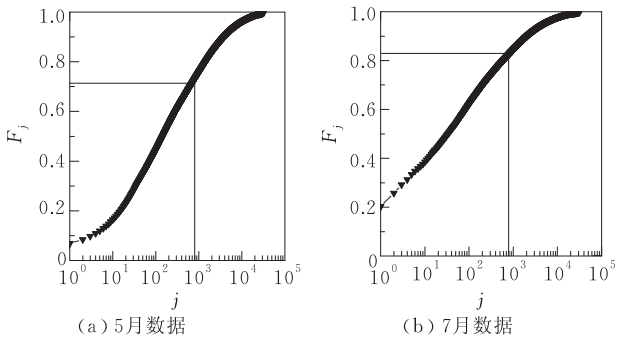


图 2 用户访问网站频度累积度结果

用户对网站的需求行为形成了一个随着时间 T 瞬间衍变的二分网络, 其网站访问度分布和累积度分布函数的统计结果分别参见图 3 和图 4.

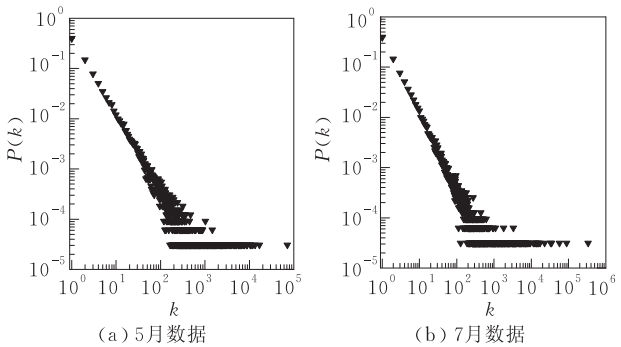


图 3 用户访问网站入度分布

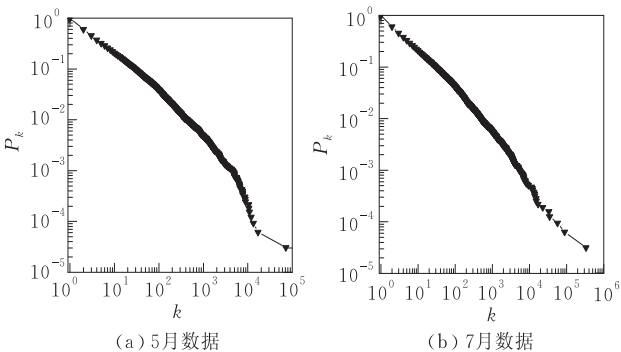


图 4 用户访问网站累积度分布

从用户对信息需求的时域统计行为可以看出, 用户对网站访问频度呈现出幂律分布. 以该时间段网站

被访问频度构成的度分布,也呈现幂律分布特性.由于数据采样的有限性,采用截断幂函数定义度分布:

$$P(k)=\begin{cases}ck^{-\gamma}, & 1\leq k\leq F\\0, & k>F\end{cases}\quad (16)$$

这里 $F=f_1$ 为网络最高度值. 常量 c 为

$$c=\frac{\gamma-1}{1-F^{1-\gamma}}\quad (17)$$

累积度分布函数(度不小于 k 的节点的概率)为

$$P_k=\sum_{k'=k}^FP(k')\approx\int_{x=k}^FP(x)\mathrm{d}x=\frac{k^{1-\gamma}-F^{1-\gamma}}{1-F^{1-\gamma}}\quad (18)$$

设网站访问量 $k\geq k_m$ 的网站数为 m , 则

$$m=M\cdot\int_{k_m}^FP(k)\mathrm{d}k\quad (19)$$

前 m 个网站的访问总量 V_m :

$$V_m=M\cdot\int_{k_m}^FkP(k)\mathrm{d}k\quad (20)$$

期望值 E (平均访问量)为

$$E=\int_1^FkP(k)\mathrm{d}k=\frac{\gamma-1}{1-F^{1-\gamma}}\cdot\frac{1-F^{2-\gamma}}{\gamma-2}\quad (21)$$

前 m 个网站访问量占总访问量的比例 $q(m)$ 为

$$\begin{aligned}q(m)&=\frac{\int_{k_m}^FkP(k)\mathrm{d}k}{\int_1^FkP(k)\mathrm{d}k}\\&=\frac{\left[\frac{m}{M}(1-F^{-\gamma+1})+F^{-\gamma+1}\right]^{\frac{\gamma-2}{\gamma-1}}-F^{-\gamma+2}}{1-F^{-\gamma+2}}\quad (22)\end{aligned}$$

对于截断幂函数来说, $1<\gamma<+\infty$ 构成一个 $q(m)$ 曲线族(图 5), 其中 $M=32578, F=71960, \gamma$ 为幂指数, 图 5 中还包含一组实测数据曲线. 从图中可以看出, 具有高度值的一个小子集就具有较大的贡献率. 所有 $q(m)$ 曲线都是单调递增的函数. 其中, 对于 $\gamma=2$ 来说, 在半对数坐标系上的 $q(m)$ 是一条直线.

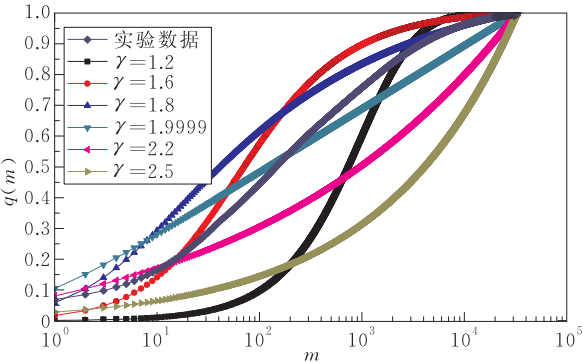


图 5 幂律分布小子集贡献率曲线

根据式(18),由实验采样数据 k 和 P_k , 可以计算用户需求入度幂律分布的 γ 值. 图 6 给出了由 700 个样本点计算得到的 γ 值的计算统计结果. 对 5 月和 7 月数据的统计平均值分别为 1.7503 和 1.7334. 即实验数据的幂律分布的参数方程为

$$P_5(k)=0.7503\times k^{-1.7503}, P_7(k)=0.7334\times k^{-1.7334}\quad (23)$$

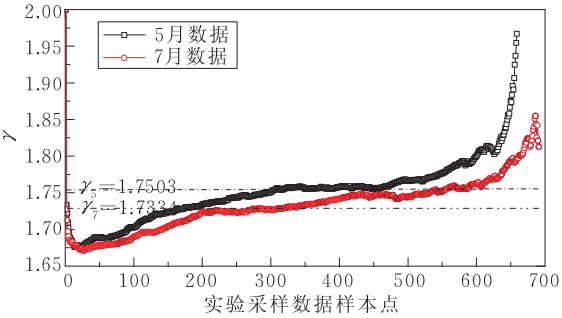


图 6 用户访问入度的 γ 分布情况

用户需求行为的幂律特征揭示了这样一个本质: 高度值的网站虽然在整个网站资源中是个小子集, 但往往对网络的行为有着深刻的影响.

4 单模式关联网络的统计特性

用户兴趣二分网络属于双模式网络, 即网络是由两类不同属性的节点构成(用户节点和 Web 节点), 且网络中并不考虑同类节点的关系. 这种网络类似于广义合作网的构成方式^[13-15], 将其投影到其中一类节点上可以构成单模式网络, 从而能够揭示同类节点之间的相互关系及其统计特性^[1].

图 7 给出了二分网络 a 按照上层节点投影到下层单模网络 b 的过程. 由二分网络的对称性, 也可以按照下层网络 b 投影到上层单模网络 a 的方式进行.

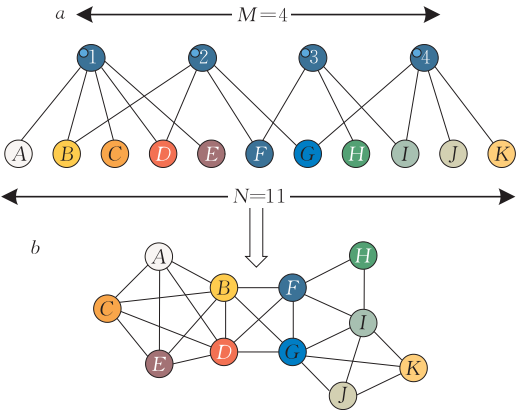


图 7 二分图的单模投影过程

二分网络构成 $N \times M$ 阶矩阵 $\mathbf{B} = (b_{ij})$ ，则单模式网络可由式(24)和式(25)计算得到：

$$\mathbf{U}_{N \times N} = \mathbf{B}_{N \times M} \times \mathbf{B}_{M \times N}^T \tag{24}$$

$$\mathbf{W}_{M \times M} = \mathbf{B}_{M \times N}^T \times \mathbf{B}_{N \times M} \tag{25}$$

$\mathbf{U}_{N \times N}$ 即为用户群体广义兴趣关联网络， $\mathbf{W}_{M \times M}$ 即为网站资源广义关联网络。矩阵 $\mathbf{U}_{N \times N}$ 和 $\mathbf{W}_{M \times M}$ 中的元素值，分别定义为用户群体广义兴趣关联网络和网站资源广义关联网络的边权值。

用户群体广义兴趣关联网络节点之间的连接表明的是两个节点具有相同的兴趣模式，其权值越高，表明其兴趣关联程度越高。网站资源广义关联网络节点之间的连接表明的是两个节点被相同的用户群所访问，其边权值越高，表明其资源被相同的用户群所关心的程度越高。

由于用户需求网络的数据量巨大，按周来分析

数据极为困难。我们对 5 月某天 24 小时的数据采样进行单模映射分析，得到了两个网络的节点度分布和边权分布的统计情况。实验数据构成的二分网络的源节点数为 1382 个，目的节点数为 6863 个，连接边数为 18545 条。则网站资源广义关联网络为 $\mathbf{W}_{6863 \times 6863}$ ，其边数为 8394411；群体兴趣广义关联网络为 $\mathbf{U}_{1382 \times 1382}$ ，其边数为 881765。

图 8 给出了网站资源广义关联网络(左图)和群体兴趣广义关联网络(右图)的节点相互关系平面分布图。可以看出网站和用户群体之间有明显的簇聚现象存在。用户群体兴趣广义关联网络有一个相对集中的小方块区域，通过分析发现其为校外访问校园网 Web 服务器的用户群。由于我们的路由器统计是局部的，这部分用户群对其它网站的访问我们无法知道，因而形成图 8 的分布结果。

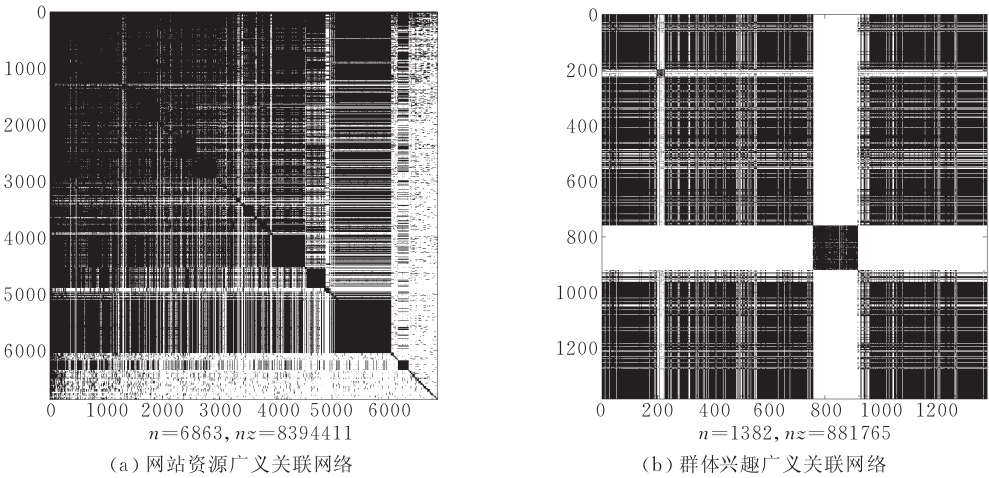


图 8 广义关联网络的节点相互关系平面分布图

从图 8 可以看出，网站资源广义关联网络和群体兴趣广义关联网络的簇集聚现象非常明显，经计算其平均簇系数分别为 0.885979 和 0.976503。

图 9 给出了网站资源广义关联网络和群体兴趣广义关联网络的度分布示意图。

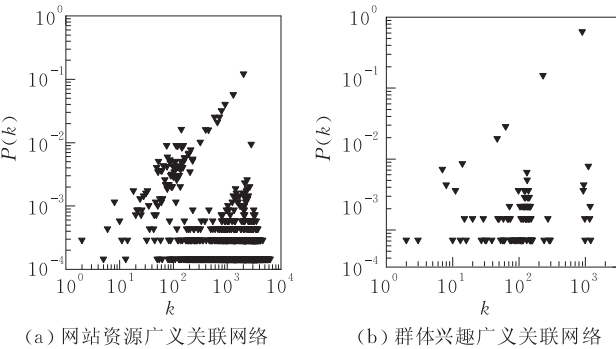


图 9 网站资源和群体兴趣广义关联网络度分布

趣广义关联网络的边权分布图。可以明显地看出，二者的边权分布服从幂律分布。

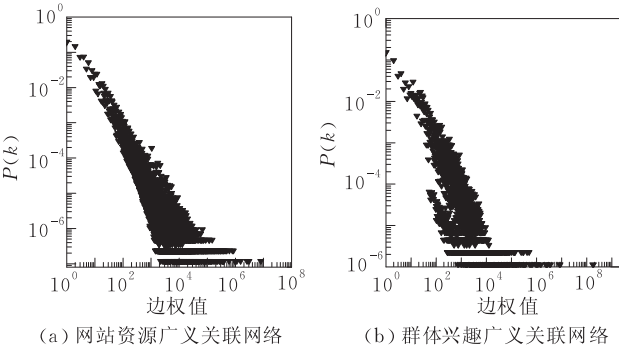


图 10 网站资源和群体兴趣广义关联网络边权分布

对 24 小时的数据分析还表明了用户群体广义兴趣关联网络的用户节点关联程度很高，节点连接率高达 75%，而网站资源广义关联网络的节点连接率仅为 12.5%。

5 结 论

实证研究表明,用户兴趣网络的入度分布呈现出无标度特征和集聚现象.将这种双模式二分网络映射为两种含权单模式网络:面向用户的群体兴趣广义关联网络和面向网站的资源广义关联网络,实证分析其统计特性表明,两者的边权分布是幂律的;群体兴趣广义关联网络的节点关联紧密且呈现簇聚特征;网站资源广义关联网络节点关联较为松散,但也呈现出簇聚特征.

本文揭示了区域用户行为具有一致性统计特征,这些特征有利于我们深入思考 Internet 结构的演化和优化过程,更好地建设信息共享基础设施.用户需求行为实际上是在真实物理网络的大量用户对网络信息资源进行操作的随机过程.从时域上来看,用户需求行为的幂律分布对 Internet 的主要特性:数据分组、业务量密度和流量的波动产生深刻的影响.由于基于 HTTP 协议的 Web 服务是当前 Internet 业务的重要组成部分,用户群体需求行为在时间上的幂律分布特性,其行为导致了对网络流量时空分布的不均匀性.

事实上,信息网络呈现出的无标度拓扑是人类用户需求行为参与网络发展与演化的必然结果.对于这些现实的或虚拟的网络拓扑结构,无标度拓扑构成的这些事实促使我们开始重新思考下面这些问题:当前的 Internet 的网络结构是否合理?信息传输的路由机制能否更加有效的改善?资源分布与服务是否需要因为幂律分布而进行优化?不同类型的信息资源分别需要什么样的信息传输机制?如何能够更加充分与高效地利用有限的带宽资源等等.

致 谢 本文在研究过程中,得到北京应用物理与计算数学研究所水鸿寿研究员、陈谋松研究员等人的真诚帮助,在此表示感谢!

参 考 文 献

[1] Strogatz S H. Exploring complex networks. *Nature*, 2001, 410(3): 268-276

[2] Watts D J, Strogatz S H. Collective dynamics of small-world networks. *Nature*, 1998, 393(6): 440-442

[3] Albert R, Barabási A-L. Statistical mechanics of complex networks. *Review of Modern Physics*, 2002, 74(1): 47-97

[4] Barabási A-L, Albert R. Emergence of scaling in random networks. *Science*, 1999, 286(10): 509-512

[5] Yook S H, Jeong H, Barabási A-L. Modeling the Internet's large-scale topology. *PNAS*, 2002, 99(10): 13382-13386

[6] Govindan R, Tangmunarunkit H. Heuristics for Internet map discovery//*Proceedings of the IEEE INFOCOM 2000*. Tel Aviv, Israel, 2000: 1371-1380

[7] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology//*Proceedings of the ACM SIGCOMM*, Cambridge, Massachusetts, 1999: 251-262

[8] Tadic B. Dynamics of directed graphs: The world-wide Web. *Physica A*, 2001, 293(4): 273-284

[9] Adamic L A, Huberman B A. Power-law distribution of the world-wide Web. *Science*, 2000, 287(3): 2115

[10] Huberman B A, Adamic L A. Growth dynamics of the world wide Web. *Nature*, 1999, 401(9): 131

[11] Broader A Z, Kumar S R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J L. Graph structure in the Web. *Computer Networks*, 2000, 33(6): 309-320

[12] Barabási A-L, Albert R, Jeong H. Scale-free characteristics of random networks: The topology of the world-wide Web. *Physica A*, 2000, 281: 69-77

[13] Newman M E J. The structure of scientific collaboration networks. *PNAS*, 2001, 98(2): 404-409

[14] Newman M E J. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 2001, 64(1): 016131

[15] Newman M E J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 2001, 64(1): 016132

[16] Li W, Cai X. Statistical analysis of airport network of China. *Physical Review E*, 2004, 69(4): 046106

[17] Liu Hong-Kun, Zhou Tao. Empirical study of Chinese city airline network. *Acta Physica Sinica*, 2007, 56(1): 106-112 (in Chinese)

(刘宏鲲,周涛.中国城市航空网络的实证研究与分析. *物理学报*, 2007, 56(1): 106-112)

[18] Xu Ye, Zhao Hai, Su Wei-Ji, Zhang Wen-Bo, Zhang Xin. Analysis on traveling diameter of Internet. *Chinese Journal of Computers*, 2006, 29(5): 690-698(in Chinese)

(徐野,赵海,苏威积,张文波,张昕. Internet 网络的访问直径分析. *计算机学报*, 2006, 29(5): 690-698)

[19] Albert R, Jeong H, Barabási A-L. The diameter of World Wide Web. *Nature*, 1999, 401(9): 130-131

[20] Barabasi A-L, Albert R, Jeong H. Mean-field theory for scale-free random networks. *Physica A*, 1999, 272(1): 173-187



LI You-Ping, born in 1935, professor, Ph. D. supervisor, member of China Academy of Engineering. His main

Background

The work of this paper is supported by the National Natural Science Foundation of China under grant No. 60272014 and China Academy of Engineering' consultation project in 2006. The key problems of the projects are how to improve the information infrastructure of network information environment, construct information share engineering based on Internet and data broadcasting system.

This paper presents the statistics characteristics of the Web-visited bipartite networks, and projected to two kinds of unipartite networks; the collective interesting affiliation net-

research interests recently focus on information share system architecture.

MA Jian-Guo, born in 1959, professor. His main research interests include network computing and active service technology.

ZHOU Ming-Tian, born in 1939, professor, Ph. D. supervisor. His main research interests include distributed computing, middleware and data mining.

works and the Web resource affiliation networks. The networks are dynamically produced in time series that are mainly influence by user requirement collective behaviors. The statistics characteristics of the Web-visited bipartite networks and its affiliation networks in a region network are not reported before the paper. The study results of the scale-free and clustering characteristics for the Web-visited bipartite networks are important to optimize the topology structure of Internet and improve the network performance and resource distribution.