二部图在用户-网站中的实证研究①

曹 易，张 宁

(上海理工大学 管理学院，上海 200093)

摘 要：通过分析用户浏览网站的访问日志，建立用户-网站的二部图。其次分别通过对用户和网站进行投影，构建出用户网和网站网。然后通过计算节点间的相似度来确定边的权值。最后计算了用户网和网站网进行了度分布、平均最短路径、平均群聚系数、点强度等拓扑参数以及时间间隔分布等人类动力学特性。证实了该网络是无标度网络，且具有“小世界”效应特性。

关键词：二部图；拓扑参数；无标度网络；“小世界”网络

Demonstrative Research of User-Website Network Based on Bipartite Network

CAO Yi, ZHANG Ning

(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The Bipartite network of user-website network is built by analysing the internet browsing history. Secondly, it builds two networks of users and websites by mapping the users and websites respectively. The value of edge is the similarity of two nodes. By calculating the topological parameters, including degree distribution, average shortest path length, average clustering coefficient and node strength, the conclusion is that both the network of users and network of websites are scale-free networks with small world effect .

Key words: bipartite network; topological parameters; scale-free networks; small world effect

1 引言

近年来，复杂网络理论得到了广泛的发展，我们真实世界中有很多系统可以用网络来描述，例如Internet、社会关系网络、学术合作网络以及生物中的食物链网络等等[1-4]。这些网络与我们的生活实践息息相关，研究它们不仅可以促进新科学分支的发展而且可能引起人类生活的重要变革[5]。

社会网络分为“单模式网络”和“双模式网络”以及更多模式的网络，主要是后者网络中存在不同类型的节点。“双模式网络”中一种称为“隶属网络”，其中一类节点是参与某项活动、事件或者组织的中“参与者”；而另一类就是它们参与的活动、事件或者组织（我们称为“项目”），例如演员合作网[6]。但是实际中往往只关心同一类节点之间的相互关系，这样就可以把二部图向其中一类节点投影，组成某一类节点相互关系的单模式网络。为了克服简单图不能描述节点间相互作用的关系的强弱，给边赋上权值，构成加权网络。周涛[7]等人在 2007 年提出了平均资源分配法来得到边的权值。

本文以某大学网络中心的访问日志，根据用户访问网页的浏览记录，将其表示成用户-网站的二部图模型。在二部图中引用节点相似性关系，映射成用户网和网站网两个单模式加权网络，研究度分布等拓扑参数，以及用户访问网站的时间间隔分布等人类动力学特性，发现用户网和网站网内部之间蕴藏的关系。

2 用户-网站的二部图模型

2.1 数据处理

本文根据上海某高校网络中心服务器的网页浏览日志，该日志记录了大学校园网 2933 个 IP 用户的近 6

① 基金项目:国家自然科学基金(70971089);上海市重点学科建设项目(S30501)
收稿时间:2011-10-07;收到修改稿时间:2011-11-18

© 中国科学院软件研究所 http://www.c-s-a.org.cn

个月的访问信息，一条完整的记录的形式为：

编号	时间	用户 IP	网站 IP	网站	网页	类别

对数据进行预处理，提取其中的“用户 IP”和“网站网址”对应的记录，本文取其中一天的数据，共有 2583 个用户以及 20522 个网站。

2.2 二部图模型建立

用户-网站的二部图模型是一个包括两类节点集合的网络，分别包括用户 IP 节点(以下简称用户节点)集合以及网站网址节点(以下简称网站节点)集合，如果某个用户节点访问某个网站节点，则将该用户节点与访问的网站节点用一条边相连，如图 1(b)所示。

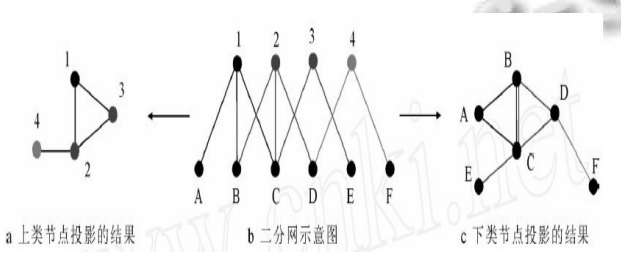


图 1 二部图投影

2.3 二部图投影

图 2 二部图投影后的单模式网示意图

根据以上建立的用户-网站二部图，用户设定为参与者，网站设定为项目，向参与者和项目两类节点分别进行投影，结果如图 2 所示。一般我们把投影方法分为无加权投影和加权投影。无加权投影方法简单，如果投影前的两个节点至少有一个公共节点，则投影后该两个节点相连，如图 1(a)(b)所示。但是该投影方式不能反映出节点间的作用强度，还容易导致信息丢失。在加权网络中，一般是用节点间的多重边数目作为权值。但是该方法具有一定的局限性，当两个参与者刚开始合作时，每合作一次就使得它们之间的影响大，但是当合作到了一定程度时，这种影响效果就呈现“饱和”效应。周涛等人基于上述缺点，提出了资源平均分配法的加权二部图投影。首先假设每个参与者有一定数量的资源，然后平均分配给每个参与的项目，最后每个项目又把自己得到的资源平均分给每个参与者。

本文根据节点间是否有另一类节点的公共邻居来判断是否连边，用节点之间的相似度来确定权值的大小。节点 i 和节点 j 之间的相似度可用公式 1 来计算

$$\delta(i,j)=\frac{|N(i)\cap N(j)|}{|N(i)\cup N(j)|} \tag{1}$$

其中 $N(i)$ 和 $N(j)$ 分别表示节点 i 和节点 j 相邻的节点数，$\delta(i,j)$ 表示 i 和 j 两节点间的相似度。在用户-网络二部图中，如果两个用户访问的网站有很多是相同的，或者两个网站同时被用户选择，则说明这两个用户或者网站具有一定的相似性。但是考虑到节点 i、j 它们的度大小相差很大，所以我们可以用下面公式 2 和公式 3 来计算它们之间的相似度。

$$\delta(i,j)=\frac{|N(i)\cap N(j)|}{\min(N(i),N(j))} \tag{2}$$

$$\delta(i,j)=\frac{|N(i)\cap N(j)|}{\max(N(i),N(j))} \tag{3}$$

其中 $\min(N(i),N(j))$ 和 $\max(N(i),N(j))$ 分别表示 $N(i)$ 和 $N(j)$ 中的较大者和较小者。本文用公式 2 表示的度较大的节点不利指标(HDI)来计算网络的权值。

3 网络中的统计属性以及实证结果

3.1 网络中的统计属性

复杂网络中节点的度是指与该节点相连的其他节点数，也就是该节点具有的边数。节点的度分布 $p(k)$ 指的是在网络中随机取一个节点，该节点的度为 k 的概率。对于无标度网络来说，$p(k)\approx k^{-\alpha}$，α 称为度分布指数[8]。

一个连通图中两个节点之间有一条最短路径，那

© 中国科学院软件研究所 http://www.c-s-a.org.cn

么任意两节点之间的平均距离就是平均最短路径长度，它可以证明网络的小世界效应。可用公式 4 来计算，其中 l 为平均最短路径，d_{ij} 是顶点 i 和 j 之间的最短路径。

$$l=\frac{1}{\frac{1}{2}n(n+1)}\sum_{i\geq j}d_{ij} \tag{4}$$

平均群聚 C 系数描述了网络中的传递性，用社会网络语言来讲，你朋友的朋友也可能是你的朋友，可以用公式 5 和公式 6 来计算

$$C_i=\frac{\text{包含}i\text{的三角形的个数}}{\text{以}i\text{为中心的三元组的个数}} \tag{5}$$

$$C=\frac{1}{n}\sum_{i=1}^{n}C_i \tag{6}$$

点强度表示加权网络中某点的邻边权的总和，如果用 $\delta(i,j)$ 来表示节点 i 和 j 的边权值，那么点强度(s)就可以表示为：$s_i=\sum_{j\in\Gamma}\delta(i,j)$，其中 Γ 表示节点 i 的邻点集。点强度分布就是指网络中强度为 s 的节点的概率 P(s)随 s 的变化规律，显然点强度包含的信息比度分布更多[9]。

时间间隔分布是人类动力学中一个很重要的特性，探索非泊松行为特征的动力学机制。它往往服从以下幂律分布：

$$p(k)\sim k^{-r} \tag{7}$$

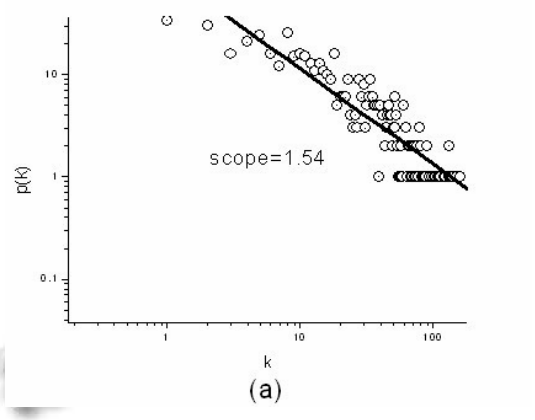
其中，r 是一个常量，通常被称为幂指数或者标度参数。

3.2 实证结果

根据用户访问网络日志，建立用户-网站的二部图模型，然后分别对用户和网站进行投影，就得到了用户网和网站网这两个单模式网络。

图 3 中(a)(b)分别是用户网和网站网的度分布，它们都是服从 $p(k)\approx k^{-\alpha}$ 的幂律分布，其中用户网中指数 $\alpha_1=1.54$，而网站网中指数 $\alpha_2=2.12$，都是属于无标度网络。

根据本文的数据，计算出用户网的平均最短路径 $l_1=1.95$，网站网平均最短路径 $l_2=2.47$。说明用户之间平均只需要 1.95 步就能与网络中任何其他用户联系起来，而网站则需要 2.47 步，这就说明了用户网和网站网都是具有小世界效应的。



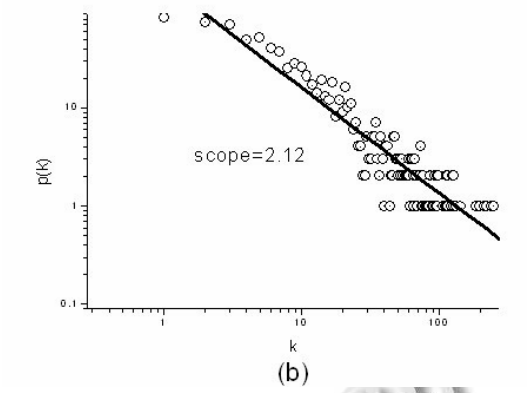


图 3 (a)用户网的度分布 (b)网站网的度分布

在本文数据中，用户网的平均群聚系数 $C_1=0.752$，而网站网中平均群聚系数 $C_2=0.427$。用户网中的平均群聚系数还是蛮大的，说明该网络中有相当数量的近于完美的集群。而网站网中的平均群聚系数稍小，因此表示该网络密度比用户网小，但是还是具有一定数量的近于完美的集群。

表 1　点强度最大的 5 个用户节点

用户节点	点强度	度
202.120.209.36	128.6	506
122.65.58.125	115.9	468
223.167.10.177	103.2	324
221.184.28.154	97.3	289
202.121.209.50	89.7	348

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 2 点强度最大的 5 个网站节点

网站节点	点强度	度
www.baidu.com	2414.1	5357
www.kaixin001.com	2057.8	3952
www.google.com.hk	1852.4	4485
www.usst.edu.cn	1697.9	3851
bbs.usst.edu.cn	1507.6	3375

表 1 和表 2 分别是用户网和网站网中节点的点强度最大的 5 个节点的信息。从表 2 可以看出，百度和谷歌等搜索引擎还是被访问的热点，具有很大的点强度和度，是一类重要的节点。网络中点强度分布和度分布是类似的，具有很大的相关性。

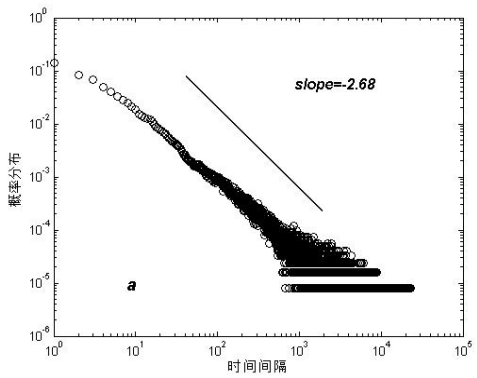


图 4 时间间隔分布

从图 4 可知，用户访问网站的时间间隔分布服从幂律分布，其中幂指数 r=2.68。从时间间隔方面来说，随着生活节奏的加快，互联网为人类的生活和学习带来极大的便利，在日常生活、工作中人们常常需要频繁地访问互联网，所以对于群体用户来说，大多数的时间间隔都很短，只有晚上休息的时候才会有比较长的时间间隔，这种时间间隔的极度不均匀性造成了幂律分布的形成。

4 结论与展望

二部图模型在复杂网络中具有很重要的地位，本文分析用户浏览网页日志，建立用户-网站的二部图模型。然后通过分别对用户和网站进行投影，得到两个单模式网络：用户网和网站网。通过计算其度分布、平均最短路径、平均群聚系数和点强度等拓扑参数，这些统计性质很好地反映出网络所具有的特点。经实证分析，二部图投影得到的两个网络都是无标度的，具有“小世界”效应特性。在以后的工作中，我们将研究用户-网站网络当中相关的动力学模型。

参考文献

1 Adamic LA. Huberman BA. Power-law distribution of the world wide web. Science, 2000,287(5461):2115.

2 Barabasi AL, Jeong H, Ravasz E, Neda Z, Schuberts A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A, 2002,311(3/4):590–614.

3 Cohen JE, Briand F, Newman CM. Community food webs: Data and theory. Springer, New York,1990.

4 Barabasi AL, Albert R, Jeong H. Scale-free characteristics of random newworks: The topology of the World Wide Web. Physica A, 2000,281(1/4):69–77.

5 吕金虎.复杂网络的同步：理论、方法、应用与展望.力学进展,2008,38(6):713–722.

6 官山,何大韧,朱陈平.跨越多个实际学科领域的合作网络与合作-竞争网络.力学进展,2008,38(6):827–834.

7 Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation. Phys Rev. E, 2007,76(4): 046115.

8 王林,戴冠中.复杂网络的度分布研究.西北工业大学学报, 2006,24(4):405–408.

9 李涛.关键词合作网络及实证研究.计算机工程,2010,36 (24):267–271.

© 中国科学院软件研究所 http://www.c-s-a.org.cn