

图书漂流网络模型实证研究

马杰良¹ 宋 艳² 潘贞贞² 韩 路²

(南京信息工程大学信息与控制学院 南京 210044)¹

(南京信息工程大学电子与信息工程学院 南京 210044)²

摘 要 通过收集整理图书漂流(bookcrossing)网站一个月内的图书漂流信息,建立图书与用户的数据库模型,并且构建两者间关系的二分图。从复杂网络的角度分析计算该网络的相关参数,如度分布、聚集系数、平均最短路径、节点项目度、项目大小、点强度及节点兴趣度,得到的图书漂流网络模型同时具有无标度特性和小世界网络的特性。

关键词 复杂网络,二分图,加权网,无标度,小世界

中图法分类号 TP393.02,TP391.9

文献标识码 A

DOI 10.11896/j.issn.1002-137X.2015.3.011

Empirical Research on Book-crossing Network Model

MA Jie-liang¹ SONG Yan² PAN Zhen-zhen² HAN Lu²

(College of Information and Control, Nanjing University of Information Science and Technology, Nanjing 210044, China)¹

(College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China)²

Abstract By collecting the data of bookcrossing website as data source in one month, this paper constructed the database model of books and the users, and then structured the bipartite graphs to describe the relationship between them. From the perspective of complex network, this paper analyzed and calculated the network related parameters, such as degree distribution, clustering coefficient, the average path length, node strength distribution, act degree distribution, act size distribution and node interest, and the conclusion is that the book-crossing network model has scale-free characteristics and small-world network characteristics in the same time.

Keywords Complex networks, Bipartite graphs, Weighted networks, Scale-free, Small-world

1 引言

随着时代与科技的发展,在满足了物质需求的同时,人们开始追求精神文明的建设。如何在众多图书中找到自己感兴趣的图书类型或者评价很高的图书,甚至找到与自己有相似阅读习惯的书友,成为了广大读者关心的问题。

图书漂流活动于20世纪60年代源于欧洲,该活动是指书友将自己不再阅读的书籍上特定的标签投放到公共场所,如公园的长凳上,无偿地提供给拾取的人阅读。拾取的人阅读之后,根据标签提示,再以相同的方式将该书投放到公共环境中去。互联网的出现加速了图书漂流活动的普及,2001年4月,在美国Kansas市附近的一个小村庄,罗恩·霍恩贝克(Ron Hornbaker)开设了一个图书漂流网站 www.bookcrossing.com,从此图书漂流活动得以迅速发展,虽不足4年,却已波及欧美。为了使每个图书的传递者都能了解到某一本书的漂流(传递)信息,获取图书的人需要尽义务到图书漂流网站上以该书标签上的唯一标识码“漂流号码”检索该书,撰写获取日志、趣闻或阅读笔记。这样,参与了该书漂流的书友就可以在网络上、电话里或聚会中相互沟通、交流。

复杂网络作为近年来研究热门的交叉学科,研究者们从不同领域对其进行理论及实证研究。实证方面,例如科技网

络中的航空网^[1]、万维网^[2]、科研引用网^[3]、生物网络中的蛋白质网^[4]以及新陈代谢网^[5]等,通过研究这些网络的拓扑结构、网络特性以及计算相关的统计数据,研究者们从不同角度理解网络所代表的现实意义。最近几年新兴的网络实例更贴近人们的生活,例如微博传播网络模型、城市公交网络模型及图书借阅网络模型等。在图书借阅方面,洪少春^[6]和唐金文^[7]分别建立了图书借阅系统的网络模型,并分析了度分布、聚集系数、平均路径,其结果表明各项统计参数值均服从指数分布 $N_{(k)} \propto \exp(-ck)$,表现出单标度性质,而不是常见的 $N_{(k)} \propto k^{-\gamma}$ 无标度性质;李楠楠等^[8]同样建立了图书借阅网络的二分图,发现项目大小、节点项目度以及节点强度服从指数分布的统计规律。

在以前的研究中,由于学校专业方向的差异,学生的借阅倾向也不尽相同。例如语言类学校中语言类图书的借阅数量和频次均大于计算机类图书,而理工院校的数学、工程类图书借阅量明显高于文学类图书。本文数据中的研究对象为全网用户,因此统计数据结果更具有普适性,所形成的网络模型也更具有随机性与传播性。

另一方面,图书漂流网络以评级信息而不是常见的借阅时间为权值。借阅时间并不能完全代表读者对图书的兴趣度,因为读者借阅图书期间有可能因为忘记归还图书而产生

到稿日期:2014-04-29 返修日期:2014-07-11 本文受国家自然科学基金(61372128)资助。

马杰良(1964—),男,副教授,主要研究方向为复杂网络、离散系统, E-mail: njkjmjl@163.com; 宋 艳(1990—),女,硕士生,主要研究方向为复杂网络; 潘贞贞(1990—),女,硕士生,主要研究方向为复杂网络; 韩 路(1989—),女,硕士生,主要研究方向为复杂网络。

较长的借阅时间,但是评级信息则可以反映读者对该书的评价及兴趣度。

2 图书漂流网络模型的构建

2.1 复杂网络理论

从 18 世纪伟大的数学家欧拉(Euler)对著名的“Konigsberg 七桥问题”的研究开始,随后 20 世纪 50 年代两位匈牙利数学家 Erdos 和 Rényi 建立了随机图理论,其被公认为是在数学上开创了复杂理论的系统性研究。在 1998 年 Watts 和 Strogatz 提出小世界网络模型^[9],1999 年 Barabási 和 Albertt 提出无标度网络^[10]。在图论知识发展的基础上,根据复杂网络性质的不同,可以将网络分为有向网络和无向网络、加权网络和无权网络以及单元网和二元网等。图书漂流网络中,将图书和读者作为节点,两者间的阅读关系作为边,由于节点类型的不同,可以得出此网络为典型的二元网(bipartite network)。

2.2 二分图

二分图 $G=(V,E)$ 的节点分成两类节点 V_1 和 V_2 ,只有不同类的节点之间才有边相连, V_1 和 V_2 为相互独立的点集。二分图网络中描述的一种关系称为“合作”,其中一类节点是某种活动、事件或者组织的“参与者”,例如读者、科研人员等;另一类节点则是它们参与的活动、事件或者组织,称为“项目”,例如图书、科学论文等^[11,12]。若读者和图书之间存在阅读关系,则将两者用一条边相连,并以评级信息作为权值。

2.3 图书漂流二分图网络模型

从 book-crossing 网站得到的原始数据如图 1 所示。其中分为两类节点:“User-ID”和“ISBN”。当“User-ID”中任一节点与“ISBN”中节点之间产生联系,表示读者阅读过该图书并留下了对该书的评价,因此将两节点连线,形成一条边。同类结点间没有连线。“Book-Rating”表示用户对图书的评级信息,可以作为权值,形成加权网络。将数据进行处理后,读者人数为 95509,图书数为 320747,总阅读次数为 1047491。

	A
1	User-ID; "ISBN"; "Book-Rating"
2	276725; "034545104X"; "0"
3	276726; "0155061224"; "5"
4	276727; "0446520802"; "0"
5	276729; "052165615X"; "3"
6	276729; "0521795028"; "6"
7	276733; "2080674722"; "0"
8	276736; "3257224281"; "8"
9	276737; "0600570967"; "6"
10	276744; "038550120X"; "7"
11	276745; "342310538"; "10"
12	276746; "0425115801"; "0"
13	276746; "0449006522"; "0"
14	276746; "0553561618"; "0"
15	276746; "055356451X"; "0"
16	276746; "0786013990"; "0"
17	276746; "0786014512"; "0"
18	276747; "0060517794"; "9"

图 1 图书漂流数据格式

将数据导入 Gephi 可视化数据软件,绘制网络模型。由于数据较大,只模拟部分网络。由图 2 可以看出,网络具有少量的高度节点和大量的低度节点。

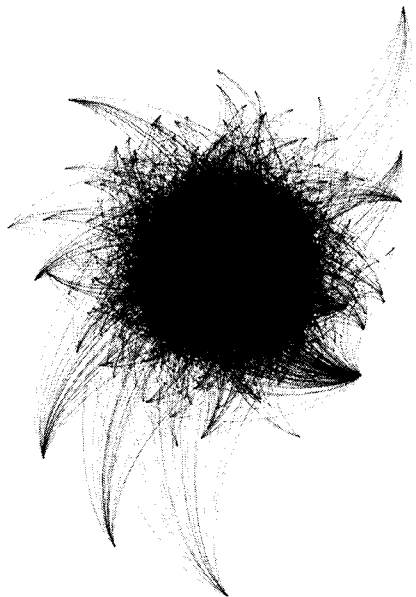


图 2 网络模型

2.4 图书漂流单元网模型

在建立读者-图书二分图网络模型的基础上,将网络转换为两个单元网络,即读者网和图书网。二分图网络模型中读者之间和图书之间没有连边,但当两名读者阅读过同一本书时,则在这两名读者间产生连边,由此将二分图投影成为读者单元网,在此基础上将读过相同图书的数量作为权值,生成加权网络。同理,当两本书被同一读者阅读过时,则将这两本书连接,形成加权的图书单元网络。图 3(a)为建立的二分图网络模型,节点 A 与节点 B 有两个共同的邻居节点,则在单元网中产生 A 与 B 的连边,边的权值为 2,得到的读者单元网如图 3(b)所示;同理也可得到图书单元网图,如图 3(c)所示。

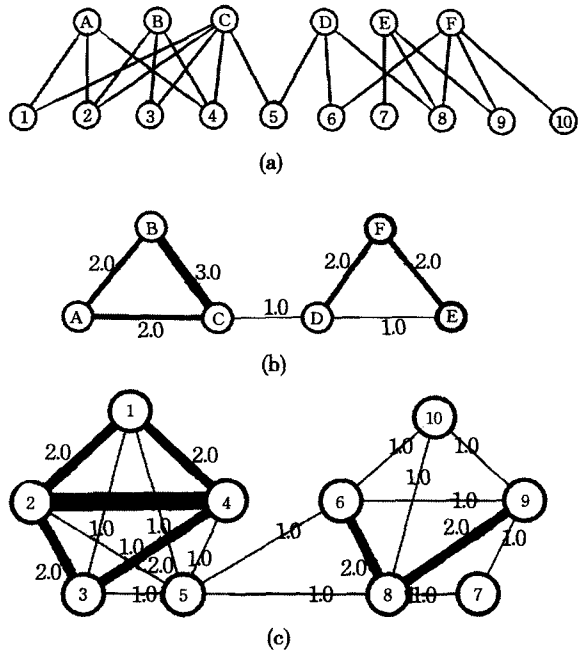


图 3 二分图网络模型及单元网

3 图书漂流网络模型仿真计算及分析

3.1 聚集系数

聚集系数反映网络中某个节点的其它两个邻居节点之间

彼此可能相连的程度,这种属性称为网络的聚类特性。聚集系数是考察网络的集团化程度的重要拓扑参数,可以反映网络的集团化程度。将图书漂流二分图网络模型投影到单元网络后,聚集系数可以考量读者(图书)网络中,读者(图书)之间连接的紧密程度。对于单元网络中每一个节点*i*,其邻居节点集合记为 N_i (某节点的邻居节点的个数等于该节点的度),这些邻居节点间最多可能存在 $N_i(N_i-1)/2$ 条边, N_i 中实际存在的边的数目为 E_i ,则聚集系数定义为 E_i 与 N_i 中所有可能存在的边数的比值,即节点 V_i 的聚集系数 $C_i = \frac{2E_i}{N_i(N_i-1)}$,整个网络的平均聚集系数 $C = \frac{1}{N} \sum C_i$ 。从几何特点来看,聚集系数也可定义为

$$C_i = \frac{\text{与点 } i \text{ 相连的三角形的数量}}{\text{与点 } i \text{ 相连的三元组组的量}}$$

与节点*i*相连的三元组指包括节点*i*的3个节点,并且至少存在从节点*i*到其他两个节点的两条边。当且仅当所有节点均为孤立节点,即没有任何连接边时, $C=0$;当且仅当网络是全局耦合的,即网络中任意两个节点都直接相连时, $C=1$ 。许多大规模的实际网络具有明显的聚类效应,例如在很多类型的网络(如社会关系网络)中,你的朋友的朋友同时也是你的朋友,的概率会随着网络规模的增加而趋向于某个非零常数,即当 $N \rightarrow \infty$ 时, $C=O(1)$ 。

根据公式,由 Gephi 可视化软件可以计算得到图书漂流网络模型读者及图书单元网的聚集系数。图书网络中的平均聚集系数说明了网络中两本被同一个读者阅读过的图书再次被另一个读者都借阅的倾向。图书单元网中的平均聚集系数为 0.934,读者单元网的平均聚集系数为 0.905,结果表明网络具有良好的聚集性,并不是完全随机的,而是在某种程度上具有类似于“物以类聚,人以群分”的特性。

3.2 平均路径

网络中两个节点*i*和*j*之间的距离 d_{ij} 定义为连接着两个节点的最短路径上的边数。网络中任意两个节点之间的距离的最大值称为网络的直径(diameter),记为*D*,即

$$D = \max_{i,j} d_{ij}$$

网络的平均路径长度*L*定义为任意两个节点之间的距离的平均值,即

$$L = \frac{1}{\frac{1}{2}N(N+1)} \sum_{i>j} d_{ij}$$

其中,*N*为网络节点数。网络的平均路径长度也称为网络的特征路径长度。

同样由 Gephi 软件计算得到图书漂流网络模型中读者单元网平均路径长度为 2.247,网络直径为 5;图书单元网平均路径长度为 2.157,网络直径为 5。这说明了例如读者单元网中,在平均情况下,经过不到 3 个人就可将任意两名读者联系起来。

网络统计特征如表 1 所列。

表 1 网络统计特征		
	读者网络	图书网络
聚集系数	0.905	0.934
平均路径	2.247	2.157
网络直径	5	5

平均路径长度和聚集系数是网络拓扑结构中的两个重要参数,在图书漂流网络模型中,网络同时具有较短的平均路径

长度和较高的聚集系数,因此表现出了小世界特性,符合小世界网络的基本特征。

3.3 节点项目度和项目大小

节点项目度表示二分图中参与者参与的项目数,即一位读者阅读过的图书数量。

$$h_i = \sum a_{ij}$$

项目大小表示二分图一个项目中包含的参与者数,即一本图书被阅读的次数。

$$T_i = \sum a_{ij}$$

其中, a_{ij} 为网络模型的二分图邻接矩阵的矩阵元,即邻接矩阵为 $A_{(G)}$,当读者与图书之间存在阅读关系时, $a_{ij} = 1$;否则 $a_{ij} = 0$ 。图 4 和图 5 分别示出了节点项目度分布和项目大小度分布。本文中的度分布并没有采用累积度分布的形式,而是采用更为直观的度分布形式。根据图中的显示,近似地认为度分布符合幂律分布: $N(x) \approx x^{-\alpha}$ 。节点项目度曲线拟合的度分布指数为 2.052,其中节点度为 1 的读者节点有 53768 个,占 56.3%;阅读过 2—10 本书的读者有 30808 人,占总读者数的 32.26%;阅读过 11—20 本书的读者有 4510 人,占 4.72%。经计算得到 $\langle h \rangle = 10.03$,即读者在一个月內平均阅读的图书数量为 10.31 本。项目大小曲线拟合的度分布指数为 2.09,其中节点度为 1 的图书节点包含 188016 个,占图书总量的 58.61%,即超过半数的图书只被阅读过 1 次;被阅读过 2—10 次的图书有 118250 本,占 36.87%;被阅读过 11—20 次的图书为 8199 本,占 2.56%。经计算得到 $\langle T \rangle = 3.21$,即每本书在一个月內被阅读的平均次数为 3.21。

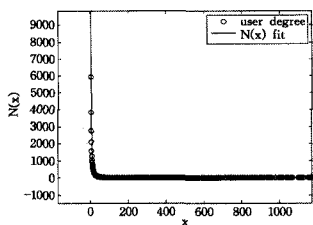


图 4 节点项目度的分布图

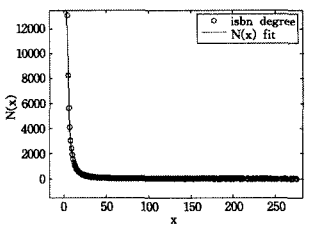


图 5 项目大小的分布图

图 6 和图 7 为对数坐标下的度分布,图中后半区域存在“肥尾现象”,这是由于网络中多数的节点有比较少的连接,仅有少数的节点有大量的连接,因此网络不存在特征度数,表现出了无标度特性。无标度网络在演化的过程中,新的节点更倾向于与那些拥有较高连接度的节点相连,这种优先连接的特性说明读者挑选一本书来阅读并不是完全随机的,可能会根据之前的用户在阅读该书时留下的评级信息或读书报告而选择是否阅读该书。

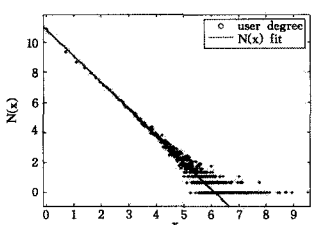


图 6 对数坐标下节点项目度的分布图

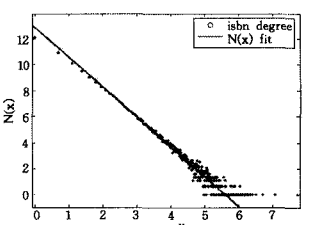


图 7 对数坐标下项目大小的分布图

3.4 节点强度

在图书漂流网络模型中,将读者对图书的评价作为权值,建立阅读关系的加权二分图。加权网络中,节点强度表示为某一节点的邻边权值的总和,用*S*表示。若 V_i 与其邻节点相

连的边 $e_{i,j}$ 的权值为 $w_{i,j}$ (本例中用读者对图书的评级信息表示,范围在 0 到 10 之间)。该点的点强度为 $S_i = \sum_{j \in N_i} w_{i,j}$, 其中, N_i 为节点 i 的邻节点集合。

点强度表示读者给予图书的评级信息或图书收到的读者给其的评级信息的总和,点强度越大,表示该读者(图书)在网络中的竞争力越强。点强度同时考虑了节点的近邻数和该节点与其近邻节点之间的权重,是该节点局域信息的综合体现。当某图书节点的度很大时,并不代表该图书很受读者欢迎,只有当其节点度和其与邻节点间权值均很大时,才说明该书具有阅读价值或为畅销书籍。例如有 3 本图书,其中一本图书被 8 名读者评价,且评分均为 1,得到其点强度为 8;第二本图书收到 5 名读者的评价,但评分均为 8,则其点强度为 40;第三本图同样得到 5 名读者的评价,评分均为 5 分,其点强度为 25,此时我们可以看出,第二本图书相比其他两本书更具有竞争力。图 8 和图 9 分别示出了对数坐标下的图书节点和读者节点,同样存在“肥尾现象”,近似符合幂律分布。

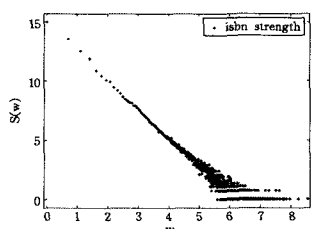


图 8 图书点强度分布

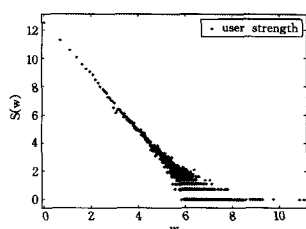


图 9 读者点强度分布

3.5 节点兴趣度

节点兴趣度表示节点在其邻节点中引起的兴趣程度,得到的节点兴趣度的统计参量可以用于后期热门图书和活跃用户的推荐。图书漂流网络数据模型中包含大量权值为零的连边,即评级信息(Rating)为 0。这种情况属于图书阅读时的特殊情况,可能存在两种原因:一是读者并不喜欢该图书,给予了图书 0 分的评价;另一种为读者并没有阅读该书,网站将评级信息默认为 0。考虑到这种情况,本文将每本图书评级信息中的非零值与该图书总的评价次数的比值作为节点兴趣度。

节点兴趣度分为图书兴趣度和用户兴趣度。图书兴趣度为该书在被浏览多次之后,有多少读者对它留下了评价,比值越大,表明该书引起的读者兴趣度越高;比值越小,则读者兴趣度越小。用户兴趣度为该读者在浏览了多本图书后,为多少本书留下了评价,比值越大,说明读者在网站中的活跃度越高;反之则越小。

$$I_i = \frac{R_i}{R_{all}}$$

图 10 示出了分图模型的图书兴趣度分布。从中可以看出,兴趣度为 0 的图书所占比例最大,表示读者浏览过这些图书后并未给出评分及评价,因此该类图书很有可能为并不畅销的图书;兴趣度为 100 的图书则表示其大部分很受读者欢迎。除去这两种情况,一般图书的兴趣度集中在 50 左右,即在收到图书之后,有 50% 的读者给予了该书评价。图 11 示出了读者兴趣度分布,结果表明用户兴趣度为 100 的用户占很大的比例,说明网站中存在大量的活跃用户群体,对于收到的图书均进行了阅读并评价。第二个峰值出现在用户兴趣度为 50 时,说明一般情况下用户给予了所收到图书的 50% 的图书评价。

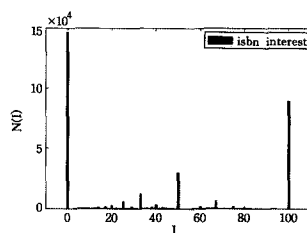


图 10 图书兴趣度

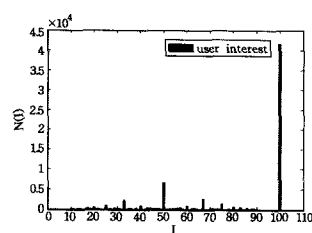


图 11 用户兴趣度

结束语 通过整理图书漂流网站一个月内的数据信息,建立图书漂流网络模型,仿真得到了网络模型的聚集系数、平均路径、项目大小分布、节点项目度分布、节点强度分布,结果表明它们均服从幂律分布,并且存在“肥尾”现象。读者在收到传递过来的图书之后,可以根据自己的兴趣阅读该图书,并给出评价。如果图书附带的评级信息越高或者别人留下的阅读笔记更有趣,则读者越有可能阅读该图书,从而使得该图书的评级信息也越高。BA 模型中,这种“富者更富”的倾向性就会导致图书的项目度分布呈现幂律分布。同时,网络具有较短的平均路径长度和较高的聚集系数,也表现出了小世界特性。

将图书漂流二分图映射成为单模式的读者单元网和图书单元网,随后可以在两个网络中进行相应的社团划分,应用于网站的推荐系统,例如向用户推荐与之有相同阅读兴趣的读者以及评价较高的书籍;还可以根据文中统计得到的节点兴趣度,选出活跃用户群体和热门图书组群,进行图书和用户的推荐。

参考文献

- [1] Guimera R, Mossa S, Turttschi A, et al. The worldwide air transportation network: anomalous centrality, community structure, and cities global roles[J]. Proc Natl Acad Sci USA, 2005, 102(22): 7794-7799
- [2] Albert R, Jeong H, Barabási A-L. Diameter of the world wide Web[J]. Nature, 1999, 401: 130-131
- [3] Redner S. How popular is your paper? an empirical study of the citation distribution[J]. The European Physical Journal B, 1998(4): 131-134
- [4] Jeong H, Mason S, Barabási A-L, et al. Lethality and centrality in protein networks[J]. Nature, 2001, 411: 41-42
- [5] Jeong H, Tombor B, Albert R, et al. The large-scale organization of metabolic networks[J]. Nature, 2000, 407: 651-654
- [6] 洪少春. 基于图书借阅关系的复杂网络的度分布和集聚程度[J]. 海南师范大学学报: 自然科学版, 2007, 20(3): 229-233
- [7] 唐金文. 图书借阅系统网络模型[J]. 计算机应用, 2010, 30(2): 10-12, 46
- [8] 李楠楠, 张宁. 图书馆借阅网的二分图研究[J]. 复杂系统与复杂性科学, 2009, 6(2): 33-39
- [9] Watts D J, Strogatz S H. Collective dynamics of small-world networks[J]. Nature, 1998, 393: 440-442
- [10] Barabási A-L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286: 509-512
- [11] Ramasco J J, Dorogovtsev S N, Pastor-Satorras R. Self-organization of collaboration networks[J]. Physical Review E, 2004(70): 36-106
- [12] Zhang P P, Chen K, et al. Model and empirical study on some collaboration networks[J]. Physica A, 2006, 360(2): 599-616