在线读书社区中的用户阅读偏好及社团发现

许益贴，刘红丽，胡海波*

(华东理工大学商学院 上海 徐汇区 200237)

【摘要】为了揭示在线读书社区中用户阅读的学科偏好及阅读兴趣的多样性，抓取了豆瓣读书社区的数据，利用用户共同阅读关系构建图书网络，结合复杂网络理论和机器学习方法对网络进行了研究。发现图书网络中，学科之间的双向权重近乎相等；阅读哲学、政治学等人文社科的用户跨学科阅读最为广泛，而阅读矿业工程、核科学与技术等工程科技学科的用户跨学科阅读最窄；二级学科网络具有3个明显的社团，对应人文社科、工程科技和基础科学三大领域，跨学科研究的适合度由高到低依次为基础科学、人文社科和工程科技。研究结果对于图书跨学科交叉推荐具有重要意义。

关 键 词 图书网络; 社团发现; 学科偏好; 在线读书社区; 文本分类; 用户行为

中图分类号 TP393; N949 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.06.020

User Reading Preference and Community Detection in an Online Reading Community

XU Yi-tie, LIU Hong-li, and HU Hai-bo*

(School of Business, East China University of Science and Technology Xuhui Shanghai 200237)

Abstract To reveal the subject preference of reading and the diversity of reading interest of users in online reading communities, this paper crawls the data of Douban reading, uses the common reading relationship to construct the book networks, and combines the complex network theory and machine learning methods to study the book networks. We find that in the book networks, the two-way weights between disciplines are nearly equal. Users who read philosophy, political science, and other humanities and social sciences have the most extensive interest in reading, while users who read engineering technology disciplines such as mining engineering and nuclear science and technology have the narrowest interest in reading. The network constructed by the secondary disciplines has three distinct communities, corresponding to the three major areas of humanities and social sciences, engineering technology, and basic sciences. For the suitability of interdisciplinary research, the basic sciences are the highest, the humanities and social sciences are the second, and the engineering technology is the lowest. Research results are of great significance to interdisciplinary cross recommendation of books.

Key words book network; community detection; discipline preference; online reading community; text classification; user behavior

随着互联网尤其是移动互联网的发展，以社会协作技术为特征的各种社交媒体网站大量涌现[1]。它们一类以在线交友为目的，如Facebook、人人网等[1]，一类以信息发布与传播为目的[2-5]，如Twitter、微博、微信等，还有一类以内容分享为目的，如优酷、豆瓣网等。它们以其潜在的研究价值吸引了来自不同学科学者们的关注[1-3, 6]。

阅读是人们获取信息的重要途径。作为从符号中获得意义的一种实践活动，阅读随着语言载体及媒介的变化而不断变化。以Web 2.0为特征的在线阅读社区，如豆瓣读书、LibraryThing、Goodreads等，充分利用大众参与，使用户能够在平台上留下大量阅读记录和评论信息，利用这些数据，学者们可以对用户的阅读偏好及用户之间、图书之间的相关性做分析。

早期对大众阅读行为的研究往往基于读者在图书馆的借阅记录，利用复杂网络理论[7-8]分析根据阅读行为构建的网络的统计特征。如文献[9]用图书及

收稿日期：2018 – 09 – 21；修回日期：2019 – 03 – 15
基金项目：国家自然科学基金(61473119, 61973121)；中央高校基本科研业务费(222201718006)
作者简介：许益贴(1993 –)，男，主要从事社交媒体方面的研究.
通信作者：胡海波，E-mail: sdhuzi@163.com

其借阅者构成的二分图对图书馆的图书借阅网进行了研究。文献[10]等对大学图书馆的图书借阅记录进行了分析，研究了两类网络，一是图书借阅形成的用户到图书的“图书借阅网络”，即二分图，二是“共同借阅网络”，即相同的图书被不同的读者所借阅，从而形成的读者间的知识分享社会网络。文献[11]基于研究生读者的借阅信息构建了读者间的共同借阅网络，研究了该网络的统计特性，并对借阅网络的社团结构进行了分析。文献[12]利用高校图书馆的学生借阅数据，构建了图书共现网络，该网络为无向加权图，节点表示书籍，连边表示共同借阅关系，利用复杂网络理论揭示了书籍间的内在联系。另有学者从其他角度对读者借阅行为进行了研究，如文献[13]利用图书管理系统中借阅图书的间隔时间数据，基于人类动力学理论[14]分析了借阅行为，发现在群体层间隔时间分布可用幂律近似刻画。

在数字化时代，学者研究书籍的方式发生了根本性的变化，文献[15]对数百万册的数字化图书所包含的人类文化进行了定量分析，研究了文化趋势或大众关注的热点随时间的变化。文献[16]利用美国亚马逊购书数据分析了用户的政治倾向，发现共同购买关系中90%以上的图书具有相同的政治倾向，自由派政治书籍的读者更偏好基础科学，如物理学、天文学和动物学，而保守派读者则更偏好应用和商业主题，如犯罪学、医学和地球物理学。

虽然关于图书阅读的实证研究已得到学者们的广泛关注，但现有研究往往未能从学科的角度研究读者或用户的阅读行为。图书自身内容决定了所隶属的学科，从学科类别来划分图书从而形成不同的社团能够揭示用户的跨学科阅读特征。此外，由于大量电子图书的出现以及社会化阅读的兴起，用户能够及时依照个人兴趣进行图书选阅，通常，用户不会仅局限于阅读某一学科的图书，但目前涉及跨学科阅读的研究仍然很少。

本文利用中国最活跃的读书社区——豆瓣网“豆瓣读书”中的数据建立刻画图书间用户共同阅读关系的有向无权网络，通过ISBN(国际标准书号)进行跨库查询并结合文本分类方法对网络中每一本图书进行学科类别标注，利用复杂网络理论分析网络特征，从学科角度研究用户阅读学科偏好及其多样性，并利用社团发现算法进一步分析图书网络中图书所隶属学科之间的相关性。

1 数据收集与预处理

在豆瓣读书版块(https://book.douban.com/)中，每本图书的主页均会列出喜欢该书的用户也喜欢其他哪些书，这些书之间存在共同阅读关系，在群体层面揭示了用户的跨学科阅读及其分布。利用爬虫程序从豆瓣读书版块获取数据，过程如下：创建空的待爬取队列和已爬取队列，随机选择一本图书作为种子节点加入到待爬取队列，根据图书主页上的信息“喜欢该本书的人也喜欢”获取相应图书集合，再对该图书集合中的每本书进行判断，是否已经存在于待爬取队列和已爬取队列。若两个队列中均未含有该图书，则将该本书的URL添加进待爬取队列，将已爬过的图书URL从待爬取队列中删除，添加到已爬取队列。再选取待爬取队列队首的图书URL获得相关信息，如此循环，直至待爬取队列为空。

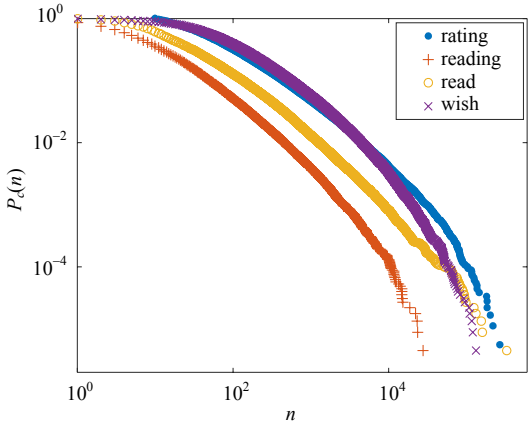


图1 图书被关注程度的互补累积分布

表1 幂律及对数正态分布的拟合参数

分布		评分人数	在读人数	已读人数	想读人数
幂律	$\hat{n}_{\min}$	518	118	507	8 677
	$\hat{\alpha}$	1.98	2.12	2.15	2.78
对数正态	$\hat{\mu}_{\log}$	−0.82	−1.72	−9.76	6.11
	$\hat{\sigma}_{\log}$	3.03	2.70	3.92	1.50

利用上述方法，获取了225 210本图书数据，图书属性包括ISBN、简介、评分、评分人数、在读人数、读过人数、想读人数、短评和书评。在豆瓣读书中，图书被用户关注的程度可用其评分人数(rating)、在读人数(reading)、已读人数(read)和想读人数(wish)进行描述，它们的互补累积概率分布如图1所示。4种指标均为长尾分布，表明图书被关注的程度存在异质性，绝大多数图书很少有人关注，但少数图书拥有庞大的读者群。4种分布的Gini系数[17]分别为0.840 8、0.843 8、0.854 2和0.803 7，表现出强异质性。对4种分布的幂律拟合均未通过阈值p = 0.1的Kolmogorov-Smirnov测试，因而并不满足幂律分布，实际上4种分布的幂律拟合与对数正态拟合的

对数似然比[18]分别为–61，–37，–5，–6，表明对数正态分布比幂律具有更好的拟合效果。本文对4种分布分别进行了幂律及对数正态分布的拟合(取相同的$\hat{n}_{\min}$)，得到参数如表1所示。

2　图书网络及图书隶属学科分类

2.1　图书网络结构特征

设*A*为一图书，*B*为喜欢*A*的用户喜欢的其他图书中的一本，则从*A*到*B*存在一有向边，通过“喜欢该本书的人也喜欢”可构建有向图书网络。豆瓣图书网络包含150 513个节点，1 438 390条边。该网络的平均路径长度*L*=10.28，其值跟网络规模的对数在同一数量级，聚类系数*C*=0.25，远大于同等规模随机图的聚类系数，可见该图书网络具有小世界特性。

2.2　图书隶属学科分类

豆瓣网中的图书区别于线下图书馆记录的图书，缺乏中图分类号，因此无法通过其对图书进行学科分类，可根据图书的ISBN从中国高等教育文献保障系统(CALIS，http://opac.calis.edu.cn/opac/simpleSearch.do)获取其学科分类。部分图书冷门或太新，图书的ISBN缺失，此外有些图书在CALIS没有记录，这些图书也无法通过ISBN对其进行学科分类。对这类图书，采用机器学习方法对其进行隶属学科的归类。

2.2.1　基于CALIS的图书分类

对拥有ISBN的图书，向CALIS数据库递交其ISBN，便可获得图书的隶属学科分类。具体流程为从已爬取的图书数据中获取所有图书的ISBN并形成队列，从队列头部取出一个ISBN利用爬虫程序向CALIS数据库自动递交该号并进行页面跳转，跳转后的网页会出现图书隶属的学科分类，并可获得分类相对应的编号。在CALIS的分类体系中，图书可归为12个一级学科，85个二级学科，具体的分类见附录(https://doi.org/10.6084/m9.figshare.6728984)。

2.2.2　基于机器学习的图书分类

通过CALIS数据库对图书隶属学科进行标注形成具有标签的数据集，再利用此数据集结合机器学习中有监督的多类别分类算法，对缺失ISBN的图书进行学科分类。考虑到二级学科含有85个类别，数量较多无法保障分类结果的准确性，故本次机器学习的分类基于一级学科，即12大类。

在文本分类过程中，特征构建及分类器的选择可对结果准确性产生重要影响。本文首先考虑利用TF-IDF对文本进行特征构建[19-20]，之后用支持向量机(SVM)分类器[21]对文本进行隶属学科分类，考虑到SVM是目前最常用、效果最好的分类器之一，且可以很好地解决本文样本较小情况下的机器学习问题，故第一个模型选用TF-IDF+SVM组合。

此外，图书还存在学科类别不均衡及多语种问题。在获取的图书中，有部分为英语原版及其他不同语言的图书，为了防止出现外文图书数量较少从而导致文本训练集数据量不足的问题，本文在TF-IDF+SVM的基础上，引进基于词向量的FastText文本分类算法。FastText为Facebook在2016年开源的文本分类项目[22]，可利用类别不均衡分布的优势来加速运算过程，其不仅较好解决了图书学科类别不均衡问题，还支持多语言表达。故本文测试了3个模型，模型1为中文样本集+TF-IDF+SVM，2为中文样本集+FastText，3为全样本集+FastText。

值得注意的是，文本分类也经常利用主题模型，如概率潜在语义分析(PLSA)[23-24]和隐狄利克雷分布(LDA)[25]，但它们均为无监督算法，不适用于本文有监督的图书分类。此外，文本分类作为自然语言处理领域[26]的重要问题，目前亦有相应的深度学习[27-28]算法，如文本卷积神经网络(TextCNN)[29]、文本循环神经网络(TextRNN)[30]、文本循环卷积神经网络(TextRCNN)[31]、分层注意网络(HAN)[32]等，但这些算法一般应用于大规模文本或图像分类以及语音识别等问题，文本分类往往并不需要太深的网络结构，且本文用于分类的文本量较少，故无需应用深度学习。实际上，本文利用的FastText算法，也是一种浅层神经网络方法，虽非深度学习，但可快速进行文本分类，适用于本文的应用场景。

考虑到分类算法的运算效率和分类结果的准确性，本文随机选取已标注好学科类别的10万本图书作为全样本集，按照上述方法流程进行分类任务，表2为各个图书类别的精确率*P*、召回率*R*和F-measure值。

从表2可见除了学科09的F-measure较低，分类结果不够准确外，3种模型对于大部分学科的分类结果都比较准确。本文又计算了3种模型的准确率(accuracy)，作为最终选定模型的评价指标，3种模型的准确率分别为0.791 8，0.770 1和0.758 9。

对比模型1和2，可见对于中文图书的分类任务，模型1优于2，考虑到模型3在全样本数据集中的适用性，本文最终的图书学科分类流程为从数据库中筛选出未标注学科类别的图书，通过ISBN查询到该图书简介，得到类别标签+图书简介的样本集并以此作

为最终的分类文本。在每次分类前，先判断图书语言，根据语言的不同选择不同的分类模型，若为中文图书采用TF-IDF+SVM，非中文图书则采用FastText算法，以此保证分类的准确率。结合CALIS和机器学习分类方法，本文最终为150 513本图书标注了类别。

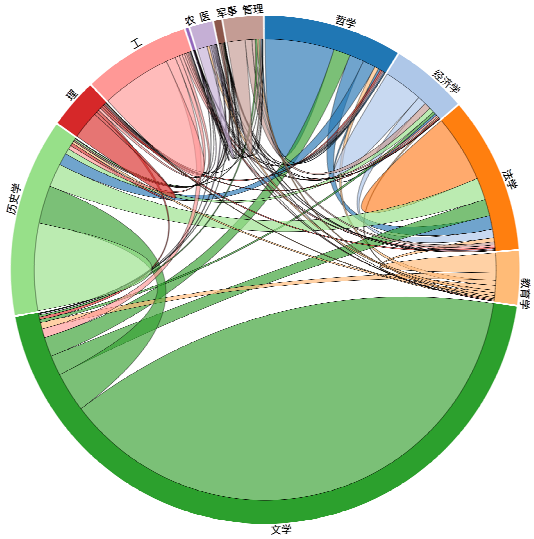
表2 文本分类结果

学科	模型1			模型2			模型3		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
01	0.761 8	0.745 5	0.753 562	0.722 9	0.746 0	0.734 3	0.745 0	0.760 2	0.752 5
02	0.801 8	0.686 1	0.739 452	0.660 8	0.779 1	0.715 1	0.642 4	0.741 0	0.688 2
03	0.673 0	0.616 1	0.643 294	0.551 8	0.671 4	0.605 8	0.570 5	0.677 4	0.619 3
04	0.708 2	0.637 8	0.671 159	0.571 4	0.711 1	0.633 7	0.577 9	0.690 5	0.629 2
05	0.849 6	0.910 6	0.879 043	0.910 3	0.837 6	0.872 4	0.894 3	0.814 4	0.852 5
06	0.752 2	0.753 8	0.752 999	0.726 5	0.705 8	0.716 0	0.702 4	0.685 6	0.693 9
07	0.787 4	0.767 0	0.777 066	0.708 7	0.768 4	0.737 4	0.730 4	0.758 2	0.744 0
08	0.836 8	0.823 4	0.830 046	0.790 2	0.851 8	0.819 8	0.778 1	0.842 0	0.808 8
09	0.750 0	0.166 7	0.272 772	0.666 7	0.400 0	0.500 0	0.733 3	0.392 9	0.511 6
10	0.755 6	0.531 2	0.623 834	0.796 9	0.622 0	0.698 6	0.582 1	0.639 3	0.609 4
11	0.727 3	0.510 6	0.599 983	0.702 1	0.452 1	0.550 0	0.666 7	0.500 0	0.571 4
12	0.708 5	0.771 2	0.738 522	0.756 5	0.745 5	0.750 9	0.687 7	0.743 0	0.714 3

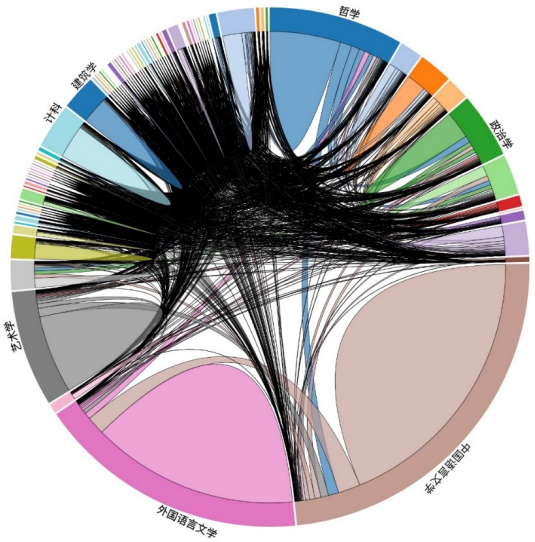
3 用户阅读学科偏好分析

为研究豆瓣网用户跨学科阅读偏好，基于图书网络及节点隶属学科分类，构建图书隶属学科之间的关联，该关联可用有向加权图来表示。节点为学科，用学科编号、学科名称、隶属学科的图书总数3个属性来刻画；用图书之间的联系来构建学科之间的关系，两个学科之间关系的权重由一个学科的图书指向另一个学科图书的数量来决定。图2分别从一级和二级学科的角度构建学科网络，形成学科交互弦图，学科间连线的宽度代表权重，连线在某一学科处的宽度正比于该学科的链出权重。

从图2a可知，隶属文学的图书占据大部分，其次是历史学、法学、哲学和工学等。偏好历史学和法学的用户跨学科阅读的程度较高，与其他学科交互比例较大，而阅读工学图书的用户跨学科阅读程度较低，集中于单学科的阅读。从图书数据中筛选出具有二级学科分类标签的数据形成二级学科网络，以此得到二级学科交互弦图，如图2b所示，并对出入度较大的学科进行了标注。由于豆瓣用户阅读文学学科的图书较多且大部分为外国语言文学和中国语言文学，从而导致二级学科划分下这两类的度值较大。阅读学科领域不同的用户，他们阅读的广泛性也有所不同。阅读外国语言文学的用户大多专注于本二级学科内图书的阅读，对其他学科的图书兴趣不大。中国语言文学、计算机科学与技术、建筑学等二级学科同样如此。而政治学、哲学等学科则有所不同，喜欢阅读该类学科图书的用户兴趣相对更为广泛。

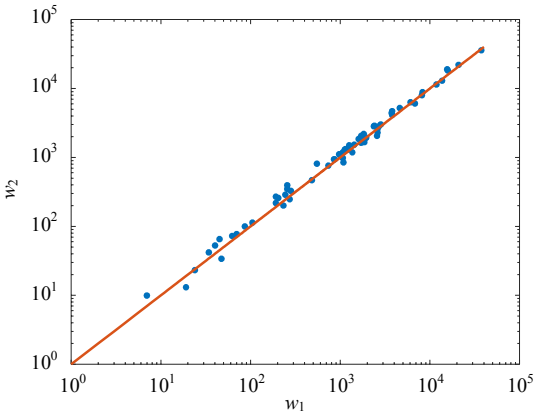


a. 一级学科

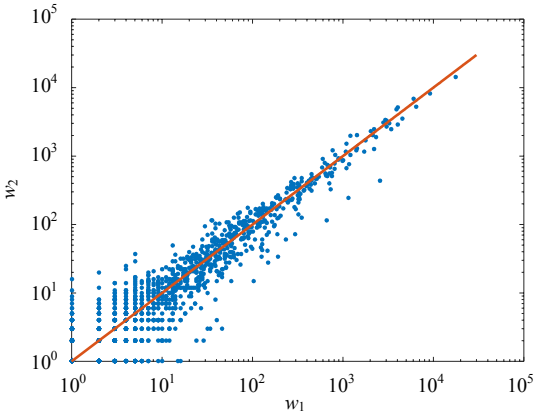


b. 二级学科

图2 学科交互弦图



a. 一级学科



b. 二级学科

图3　学科之间权重的相关性，直线为对角线$w_2 = w_1$

对于A、B两个学科，从A指向B与从B指向A的有向边的权重之间存在相关性，图3给出了一级学科间和二级学科间的相关关系，横坐标为学科编号较小的学科指向编号较大学科的权重，纵坐标则为编号较大学科指向较小学科的权重。可见，一、二级学科均表现出明显的正相关性，其Pearson相关系数分别为0.994 6(p<0.001)和0.981 5(p<0.001)，且学科之间的双向权重近乎相等，即A指向B的有向边的权重近似等于B指向A的权重。

从图2可知，阅读某学科某本书的用户往往也倾向于阅读该学科的其他图书，为了定量刻画这种偏好性，设p_{ij}为学科i的图书指向学科j的图书的数量相对于学科i图书总出度的比例，$\sum_j p_{ij} = 1$，p_j为学科j的图书占图书总数的比例，$\sum_j p_j = 1$，则$d_{ij} = p_{ij} - p_j$可刻画学科间的偏好性：$d_{ij} > 0$表示相对于随机阅读，阅读i学科图书的用户更倾向于阅读j学科的图书，$d_{ij} = 0$表示阅读i学科图书的用户对j学科的图书没有偏好性，d_{ij} <0则表示阅读i学科图书的用户倾向于不阅读j学科的图书。图4给出了一级学科间的偏好关系，图中数字为d_{ij}。可见每个学科的读者都偏好于同学科的图书；文学类图书占比较高，除了文学自身，其他学科图书的读者均倾向于不阅读文学图书。此外，只有文学和工学图书读者仅偏好于本学科图书，其他学科读者均对与之相关的邻近学科有所偏好。

根据学科交互关系，可以分析用户跨学科阅读的多样性。在图书网络中，用图书之间的联系构建学科之间关系后，阅读不同学科图书的用户其跨学科阅读多样性可用信息熵[33]表示：$H = -\sum_i (T_i / N)\log(T_i / N)$，其中$N$为图书总量，$T_i$代表阅读某一学科的用户阅读的学科$i$的图书数量，且$\sum_i T_i = N$。表3分别给出了信息熵最大和最小的前10个二级学科。

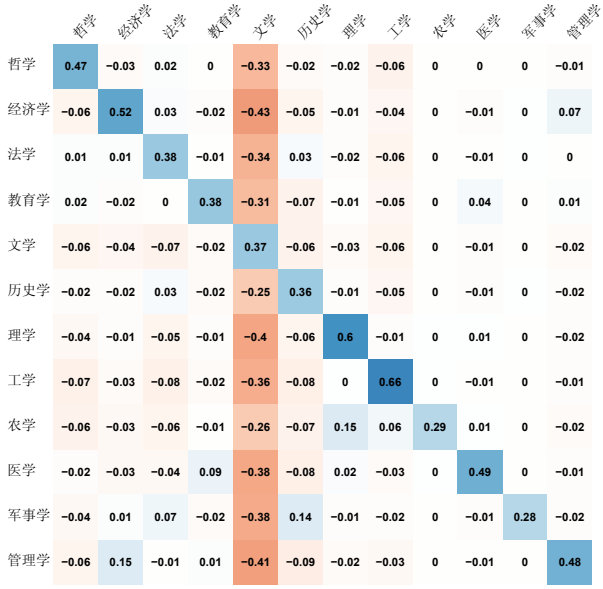


图4　一级学科间的偏好关系

表3　不同二级学科用户的信息熵

学科	信息熵	学科	信息熵
哲学	1.555 6	矿业工程	0.000 6
政治学	1.153 1	核科学与技术	0.000 6
艺术学	1.081 6	军制学	0.000 6
法学	0.915 0	石油与天然气工程	0.000 6
心理学	0.748 4	兽医学	0.000 8
历史学	0.635 7	口腔医学	0.001 1
工商管理	0.631 3	农业资源利用	0.001 5
理论经济学	0.595 7	战术学	0.001 5
应用经济学	0.580 8	测绘科学与技术	0.001 8
计算机科学与技术	0.566 5	农业工程	0.002 2

由表3可知，阅读哲学、政治学、艺术学等人文社科的用户信息熵最大，表明阅读这几个学科的用户跨学科阅读最为广泛。相反，信息熵最小的学科为矿业工程、核科学与技术、军制学等，它们多属

于工程科技学科，阅读这些学科类别的用户跨学科阅读最窄，他们往往专注阅读某一学科的图书。

4 社团发现

4.1 图书网络中的社团发现

图书网络中存在社团结构[34-35]，研究社团结构对于深入理解网络拓扑特性、学科关系和用户阅读偏好具有重要意义。学者们基于不同的优化目标和使用场景提出了不同的社团发现算法，包括Girvan-Newman算法[36]、Fast Greedy算法[37]、Louvain算法[38]、Walktrap算法[39]和Infomap算法[40]等。由于一些算法只适用于无向有权网络，而本文构建的图书网络为有向无权网络，且规模较大，考虑到计算复杂度和使用情况，最终选用Infomap算法来研究图书网络的社团结构。在图书网络中共发现6,557个社团，平均每个社团有15本图书，社团数量较多且大部分社团规模较小，意味着在该网络中，大部分图书之间的联系都较为松散。本文选取了规模最大的前20个社团(共包含6 624本图书)进行分析，如表4所示。

由表4可知排名前20的社团均包含文学学科，这主要是因为图书类别不均衡，豆瓣读书中隶属于文学学科的图书较多。此外隶属文学与隶属管理学的学科编号为1205(图书馆、情报与档案管理)的图书分在同一社团的情况出现了4次，表明它们之间交互较多，从用户的角度来看，文学与管理学中的图书馆、情报与档案管理之间的距离更近。

表4 图书数排名前20的社团

社团编号	图书数量	图书隶属学科编号	社团编号	图书数量	图书隶属学科编号
1	676	0503, 1205, 0504, 0501, 0502	11	265	0501, 0502, 0503, 0504
2	605	0502, 0503, 0504, 0501, 1205	12	255	0501, 0502, 0504
3	503	0502, 0503, 0504, 0822, 0501, 0101	13	250	0501, 0502, 0504
4	487	0502, 0503, 0504, 0501, 0812	14	249	0501, 0502, 0503, 0504
5	427	0502, 0503, 0504, 0501, 0812	15	246	0502, 0503, 0504, 0601, 0501
6	329	0501, 0502, 0503, 0504	16	242	0501, 0502, 0503, 0504
7	310	1205, 0501, 0502, 0504	17	240	0501, 0502, 0503, 0504
8	289	0501, 0502, 0503, 0504	18	236	0501, 0502, 0504
9	278	0501, 0502, 0504	19	235	0502, 0503, 0504, 0501, 0101
10	268	0501, 0502, 0503, 0504	20	234	0502, 0503, 0504, 0501, 1205

4.2 学科网络中的社团发现

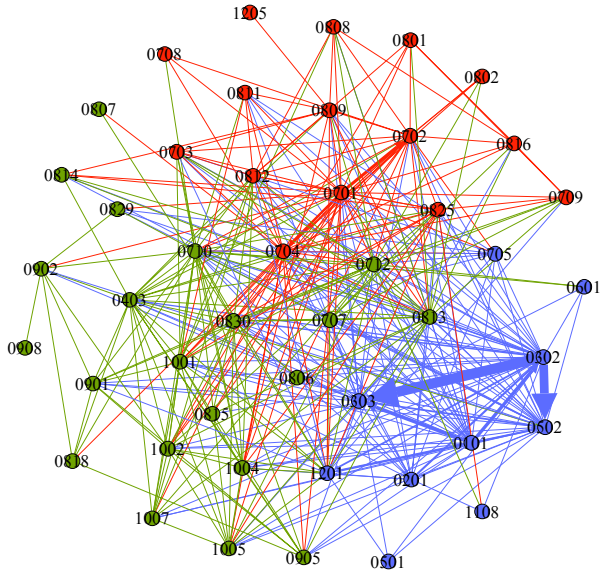


图5 二级学科网络社团划分

将隶属相同二级学科的图书进行合并后，图书之间的联系转换为二级学科之间的联系，图书的出入度转换为各学科的出入度及各学科间的权重，最终可得到二级学科之间的交互网络。该网络为有向加权网络，共有46个节点，391条边，对该网络进行社团划分，结果如图5所示。它可划分为3个社团(用不同颜色表示)，权重用边的宽度表示，图中未显示自环。表5给出了3个社团的网络结构参数及社团特征。

由表5可知，蓝色节点所代表的社团主要是人文社科类图书，包括哲学、历史学、政治学等；绿色节点大部分为工程科技领域的图书，包括海洋科学、系统科学、冶金工程等，体育学也被分在了该类；红色节点主要为基础科学领域的图书，包括数学、物理学、化学等，还有电气、机械等领域的图书。此外，基础科学类别的社团C网络直径及平均路径长度较短，社团内成员的联系较为紧密，学科交互较多。而工程科技类别的社团B网络直径和平均路径长度较长，社团内成员联系比社团C更松散，人文社科类别的社团A内的成员交互介于社团C和B之间。可见，就跨学科研究的适合度而言，基础科学最高，人文社科次之，工程科技再之。

社团发现的结果不仅可以揭示图书之间的关系，而且由于构建的网络均基于用户行为，其结果同样对图书推荐具有重要意义。以往的推荐系统往往较少考虑用户跨学科阅读的多样性，本文的研究表明，可以从学科角度，更好地为偏好不同类型图书的用户进行多学科交叉推荐。

表5　二级学科网络社团特性

社团	网络直径	聚类系数	平均路径长度	学科编号	社团特征
蓝色社团(社团A)	3	0.36	1.27	0101，0201，0302，0303，0501，0502，0601，0705，1108，1201	哲学，经济学，政治学，社会学，中外文学，历史学，地理学，军事学，管理科学
绿色社团(社团B)	4	0.34	1.52	0403，0707，0710，0712，0806，0807，0813，0814，0815，0818，0829，0830，0901，0902，0905，0908，1001，1002，1004，1005，1007	体育学，海洋科学，生物学，系统科学，冶金工程，动力工程，建筑学，土木工程，水利工程，环境科学，作物园艺，医学
红色社团(社团C)	2	0.44	1.15	0701，0702，0703，0704，0708，0709，0801，0802，0808，0809，0811，0812，0816，0825，1205	数学，物理学，化学，天文学，地质学，力学，机械工程，电气工程，计算机科学与技术，航空宇航科学与技术，图书馆情报学

5　结 束 语

本文基于豆瓣阅读的数据，利用用户共同阅读关系构建图书网络，结合复杂网络理论和机器学习对图书网络进行分析，揭示了用户阅读的学科偏好、跨学科阅读的多样性以及学科之间的联系强度。研究发现，图书被用户关注的程度存在异质性，阅读哲学、政治学、艺术学等人文社科的用户跨学科阅读最为广泛，而阅读矿业工程、核科学与技术、军制学等工程科技学科的用户跨学科阅读最窄。二级学科网络具有3个明显的社团，对应于人文社科、工程科技和基础科学三大领域，基础科学最适合做跨学科研究，人文社科次之，工程科技再之。

在前互联网时代，大众的阅读记录往往难以保存，近年来以互联网为基础的新媒体的出现极大的改变了大众的阅读方式，可搜集的数据也不再局限于图书馆的借阅记录。作为中国最活跃的读书社区，豆瓣阅读记录了大量用户的阅读信息，进而可以对用户的阅读行为及偏好进行深入研究，本文在这方面做了有益的探索，但仍存在不足，如数据量较少、数据采样可能有偏及分析层面较为宏观等。将来的工作希望结合用户自身属性及阅读行为的时间序列信息对跨学科阅读进行进一步研究，如跨学科阅读的稳定性及影响用户阅读偏好的因素等。

参 考 文 献

[1] 胡海波, 王科, 徐玲, 等. 基于复杂网络理论的在线社会网络分析[J]. 复杂系统与复杂性科学, 2008, 5(2): 1-14.
HU Hai-bo, WANG Ke, XU Ling, et al. Analysis of online social networks based on complex network theory[J]. Complex Systems and Complexity Science, 2008, 5(2): 1-14.

[2] 李栋, 徐志明, 李生, 等. 在线社会网络中信息扩散[J]. 计算机学报, 2014, 37(1): 189-206.
LI Dong, XU Zhi-ming, LI Sheng, et al. A survey on information diffusion in online social networks[J]. Chinese Journal of Computers, 2014, 37(1): 189-206.

[3] ZHANG Z K, LIU C, ZHAN X X, et al. Dynamics of information diffusion and its applications on complex networks[J]. Physics Reports, 2016, 651: 1-34.

[4] 罗春海, 刘红丽, 胡海波. 微博网络中用户主题兴趣相关性及主题信息扩散研究[J]. 电子科技大学学报, 2017, 46(2): 458-468.
LUO Chun-hai, LIU Hong-li, HU Hai-bo. Research on correlation of users' topic interests and topic information diffusion in microblog networks[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(2): 458-468.

[5] 陆豪放, 张千明, 周莹, 等. 微博中的信息传播: 媒体效应与社交影响[J]. 电子科技大学学报, 2014, 43(2): 167-173.
LU Hao-fang, ZHANG Qian-ming, ZHOU Ying, et al. Information spreading in microblogging systems: Media effect versus social impact[J]. Journal of University of Electronic Science and Technology of China, 2014, 43(2): 167-173.

[6] 许小可, 胡海波, 张伦, 等. 社交网络上的计算传播学[M]. 北京: 高等教育出版社, 2015.
XU Xiao-ke, HU Hai-bo, ZHANG Lun, et al. Computational communication on social networks[M]. Beijing: Higher Education Press, 2015.

[7] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012.
WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Introduction to network science[M]. Beijing: Higher Education Press, 2012.

[8] BARABÁSI A L. Network science[M]. Cambridge, UK: Cambridge University Press, 2016.

[9] 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.
LI Nan-nan, ZHANG Ning. The study of the bipartite graph about the library lending network[J]. Complex Systems and Complexity Science, 2009, 6(2): 33-39.

[10] 燕飞, 张铭, 孙韬, 等. 基于网络特征的用户图书借阅行为分析——以北京大学图书馆为例[J]. 情报学报, 2011, 30(8): 875-882.
YAN Fei, ZHANG Ming, SUN Tao, et al. Network based users' book-loan behavior analysis: A case study of Peking university library[J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(8): 875-882.
[11] 张柯, 赵金龙, 胡小丽. 基于复杂网络理论的高校图书馆借阅网络研究[J]. 大学图书情报学刊, 2014, 32(1): 75-77.
ZHANG Ke，ZHAO Jin-long，HU Xiao-li. Research on book-borrowing network of university library based on the complex network theory[J]. Journal of Academic Library and Information Science, 2014, 32(1): 75-77.
[12] 陈晓威, 孙建军. 基于图书借阅网络的各类书籍关系研究[J]. 图书情报工作, 2017, 61(11): 21-28.
CHEN Xiao-wei, SUN Jian-jun. The relationships among books based on the book-borrowing network[J]. Library and Information Service, 2017, 61(11): 21-28.
[13] 王福生, 杨洪勇. 图书管理系统中的借阅行为分析[J]. 复杂系统与复杂性科学, 2012, 9(1): 55-58.
WANG Fu-sheng, YANG Hong-yong. Books-borrowing behavior in library management system[J]. Complex Systems and Complexity Science, 2012, 9(1): 55-58.
[14] BARABÁSI A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435: 207-211.
[15] MICHEL J B, SHEN Y K, AIDEN A P, et al. Quantitative analysis of culture using millions of digitized books[J]. Science, 2011, 331(6014): 176-182.
[16] SHI F, SHI Y, DOKSHIN F A, et al. Millions of online book co-purchases reveal partisan differences in the consumption of science[J]. Nature Human Behaviour, 2017, 1(4): 0079.
[17] HU H B, WANG X F. Unified index to quantifying heterogeneity of complex networks[J]. Physica A, 2008, 387(14): 3769-3780.
[18] CLAUSET A, SHALIZI C R, NEWMAN M E J. Power-law distributions in empirical data[J]. SIAM Review, 2009, 51: 661-703.
[19] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.
[20] MANNING C D, RAGHAVAN P, SCHÜTZE H. Introduction to information retrieval[M]. New York, USA: Cambridge University Press, 2008.
[21] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.
[22] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[EB/OL]. (2016-08-09) [2017-07-12]. https://arxiv.org/abs/1607.01759.
[23] HOFMANN T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999: 50-57.
[24] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1-2): 177-196.
[25] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
[26] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.
[27] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521: 436-444.
[28] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge, MA: The MIT Press, 2016.
[29] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1746-1751.
[30] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. California: IJCAI, 2016: 2873-2879.
[31] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 2267-2273.
[32] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2016: 1480-1489.
[33] WENG L, MENCZER F. Topicality and impact in social media: Diverse messages, focused messengers[J]. PLoS ONE, 2015, 10(2): e0118410.
[34] FORTUNATO S. Community detection in graphs[J]. Physics Reports, 2010, 486(3-5): 75-174.
[35] JAVED M A, YOUNIS M S, LATIF S, et al. Community detection in networks: A multidisciplinary review[J]. Journal of Network and Computer Applications, 2018, 108: 87-111.
[36] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826.
[37] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks[J]. Phys Rev E, 2004, 70: 066111.
[38] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics, 2008(10): P10008.
[39] PONS P, LATAPY M. Computing communities in large networks using random walks[EB/OL]. (2005-12-12) [2017-07-23]. https://arxiv.org/abs/physics/0512106.
[40] ROSVALL M, BERGSTROM C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118-1123.

编 辑 蒋 晓