DOI:CNKI:11-2746/G2.20131030.1424.005

图书标注环境下用户的认知特征*

马费成　张　斌

摘　要　本文采集了豆瓣网上2,534本图书情报学科领域的图书在四个时间点的标签数据构建标签共现网络,利用标签分析和确定概念的序化和聚类(知识结构),揭示用户在图书标注环境下的认知特征。统计分析发现:标签共现网络的网络结构具有幂律特征,并服从稠密化幂律分布,适用于研究用户认知状态动态演化;标签共现网络为非均匀网络,度数较高的节点词汇处于核心位置,是最为基本的概念,而中介中心性较高的节点词汇反映概念在知识结构中的控制能力强,对标签共现影响较大,对网络的流动性至关重要,这些词汇的共同点在于相对概念比较宽泛,说明用户会优先选择易于理解的词汇进行标注和认知。图4。表4。参考文献25。

关键词　图书标注　标签共现网络　用户认知特征

分类号　G350

Cognitive Characteristics of Users in the Tagging Books Environment

Ma Feicheng & Zhang Bin

ABSTRACT　This paper collects 2,534 books in LIS and their tags data at four timing points at www.douban.com to construct the co-occurrence network of tags, analyze and determine the ordering and clustering of concepts (knowledge structure) by use of tags, and reveal the cognitive characteristics of users in the context of tagging book. This paper finds that, the structure of the co-occurrence network of tags is characterized by "power law" and obeys densification power law distribution, which is appropriate for studying the dynamic evolution of users' cognitive state; the co-occurrence network of tags is a inhomogeneous network, in which node terms with higher degree are at the core position as the most basic concepts, while node terms with higher betweenness centrality reflect the ability of the concepts in controlling knowledge structure, have greater influence on co-occurrence of tags, and are crucial to network mobility, which have relatively broad concept in common and explain that users would take priority to choose easily-understood terms to tag and recognize. 4 figs. 4 tabs. 25 refs.

KEY WORDS　Tagging books.　Co-occurrence network of tags.　Users' cognitive characteristics.

当代知识门类在高度分化的同时不断趋向综合,逐步演化成一个复杂的网络体系,将人类知识构建成一个时间上和空间上无限延展的关联系统。庞大而复杂的知识网络,督促我们去寻找更加有效的知识组织方法和途径,以使用户能够更加方便、快捷地获取所需的知识。

在Web 2.0环境下,以博客(Blog)、维基(Wiki)、分众分类(Folksonomy)为代表的应用服务让人们直观地感受到了一系列看似矛盾的哲学关系中却暗含着和谐的规律。如:看似混乱的标签,却蕴含着有序的幂律分布,稍作梳理后,不仅能跟语义网很好地结合,还能构造出知识元关系图;以往被人们认为权威的主题标引,也正遭遇多元化标引方法的挑战。由此引发了以下思考:以往图书情报学

* 本文系国家自然科学基金项目"知识网络的形成机制及演化规律研究"(编号:71173249)的研究成果之一。
通讯作者:张斌,Email: zb0205@126.com

者为了描述知识网络，普遍以引文网络为线索，揭示学科领域的结构和发展状态，这是从知识生产者角度出发，是一种高阶的知识网络；能否利用标签等 Web 2.0 服务，从用户角度出发，揭示知识接受者的认知状态，继而找到知识传播过程中的薄弱环节和问题。本文按照这一思路循序向前探索。

1　相关研究回顾

认知科学（cognitive science）是 20 世纪 70 年代末诞生并发展起来的，由哲学、心理学、语言学、人类学、计算机科学和神经科学等学科整合在一起的一门交叉学科，核心是认知心理学。图书情报学与认知科学在研究领域中有很多重合之处，特别是关于读者或者用户的研究，尤其关注用户的信息搜索行为。

1980 年，Brookes[1] 提出了布氏基本方程 $K[S]+\Delta I=K[S+\Delta S]$，用来表达知识接受者的知识结构 $K[S]$ 由于吸收了情报 ΔI 而变化为 $K[S+\Delta S]$。这里的知识是“由概念间关系连接起来的一种概念结构”，知识结构可能是客观的，也可能是主观的。它描述了影响个人知识结构的信息交流过程，因而被视为图书情报学认知范式的基础[2]。Belkin[3] 认为从认知角度来看，图书情报学的基本问题是“促进有需求的信息在生产者和接受者之间进行有效传播”，它的交流系统包含语言和认知两个层面（见图 1）。

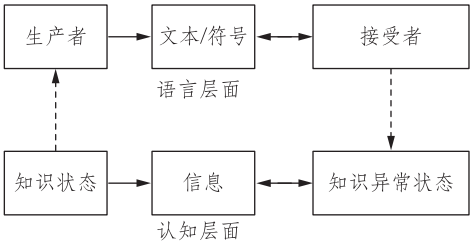

图 1　交流系统的两个层面

在语言层面上，文本是生产者有目的组织起来的符号集，试图去改变接受者的概念结构；在认知层面上，接受者由于认识到自身存在着某种知识异常状态（Anomalous States of knowledge, ASK），便将这种知识异常状态转换成某种可交流的结构，用于检索那些可以解决异常状态的文本或文本集，并加以解读。Ingwersen[4] 指出接受者会根据他的认知模型、实际的问题空间及不确定状态去感知和解释文本或符号，而这些来自于 ASK。Brookes、Belkin、Ingwersen 等人倡导的认知范式使得人们把注意力从集中在信息的实体和符号上转至理解主题或者接受者上[5]。

随着网络的普及，学术界关注网络环境下用户的认知特征以及对信息交流的影响，也促进了信息行为的认知方法从个体向整体转变，强调复杂的信息搜索情境下人类的群体行为[6]。情境是指认知行动者在信息行为过程中所处的特定环境和状态。Web 2.0 信息自组织与有序化的环境是网络信息生产和搜索的基础，社会标签的存在也使得计量和检索联系得更加密切，这将用户行为研究从认知性、个体性研究拓展到社会性、群体性研究，在更为广泛的社会情境下考察用户认知、行为、交流和对信息的利用[7]。有鉴于此，学者们对社会标注行为给予了重点关注，研究也从系统视角转向用户视角。

首先，对标注系统的结构以及标签共现的网络特征进行研究。社会化标注的主要组成元素有三个：用户、标签和资源。用户使用标签来描述资源，资源拥有唯一的身份标识，如 URL、ISBN。三个元素自身和相互间的关系构成了社会化标注系统的内在结构，三者之间的互动使这一结构不断变化，形成了社会性[8]。标签共现可以形成一类表达概念结构的加权无向网络[9]。网络会呈现出小世界现象和无标度特征，体现了词汇的相关性[10]。其次，依据用户特定的信息需求进行信息推送。易明等[11] 设计了一种基于标签的知识主题网络构建方法，先对单个用户拥有的知识进行聚类形成知识子类，再对所有用户的知识子类进行聚类形成不同的知识主题，进而形成知识主题网络，并用一个图书标注实验来分析其有效性。虽然没有给出网络的整体结构参数，但从中可以看出，用户对某个学科领域的图书进行标注能够大致反映该学科领域中知识接受者的认知状态。另外，标签不但反映了用

户对信息对象的看法,还可以增加对内容的描述,并能够提供更有效的信息表示。Yoon[12]从语言学角度在微观层面上对标签的组合关系进行研究,分三个主要概念类别进行分析,即内容主题(content-topic)、内容评论(content-comment)和使用情境(context of use),组合关系对信息搜索的意义在于可以结合由实际使用术语的情境产生的某个特定关联而将概念组合起来,特别是谓词关系超出了单纯的共现关系。

不难看出,学者们已经意识到,在认知方面,要对用户的知识状态和信息需求进行深入研究,需要与搜索行为进行整合,将系统研究和用户研究紧密结合起来,并强调情境的重要性。但在研究思路上还存在一些问题。第一,社会标签与受控的主题词表是两个不同但是又紧密关联的概念体系,在假设标签可以反映用户认知状态的前提下,从事此方面的研究就要关注网络环境对用户行为的改变,对信息交流的影响和对信息生产的促进作用。第二,现有研究局限在用户已经形成的标签中进行微观层面分析,但这也在一定程度上割裂了用户研究和系统研究。在图 1 中,交流系统不仅要关注语言和认知两个层面的问题,还需要考虑到信息从生产者到接受者这一过程,这也是发现潜在关联、提升信息服务水平的关键。

针对这些问题,本文利用图书标注环境这一特定的情境展开探索。图书是人类用来记录知识的工具之一,是人类交融感情、获取知识、传承经验的重要媒介,其内容比较系统、全面、成熟、可靠,能满足语言层面上的“文本”要素。在进入基于标签的社会化标注阶段后[13],用户不仅可以通过标签来收集、分类、聚合感兴趣的图书信息,还可以快速和方便地与他人分享这些信息,这些标签蕴含着用户的认知、行为、交流和对信息的利用。如果利用某种策略将生产者和接受者关联起来,利用标签揭示用户的思维认知,将有利于设计基于用户心理认知特点的各类信息服务系统的解决方案。

2 研究设计

在本研究中,为了探究图书标注环境下用户的认知特征,标签被用来分析和确定概念的序化和聚类特征。研究基于如下假设:标签中的一组特定术语(或概念)在语义关联上会比其他术语表现得更为强烈。研究目的在于确定由标签引出的语义核心能更好地反映信息内容和辅助信息搜索。这些语义核心被定义为概念实体(conceptual entities)及它们之间的关系,不妨称之为知识结构。

对于知识结构,不同领域有不同的解释。图书情报学中将知识结构理解为人类知识的分类,而心理学中将知识结构定义为在人脑中形成的概念体系,并用概念地图[14]进行表示。Brookes 也没有明确指出知识结构该如何表示,只是用认知地图(cognitive map)去描述它。更为普遍的处理方式是采用心智模型(mental model)的提法。虽然这些概念间存在细微差别,但通常可以认为它们表达了相近的内容,在一定条件下可以通用。

研究采用连续实验方法,即在一个时间段内,通过执行连续的任务,观测用户在知识结构及其行为上发生的变化,属于纵向变化比较分析。让具有同等背景知识的用户执行图书标注任务,通过观测任务执行的行为特征和效果来了解他们目前的知识状态。由于大多数的标签是名词形式[15],这意味着将它们放在一起代表着一系列离散且有联系的概念。共现关系并不能非常明显地指出标签之间的具体语义关系,但却是一种简单有效的表达组合,而且能方便测量两个概念节点的相似程度。使用共词网络方法去探讨知识结构,虽然没有概念地图中丰富的命题,但却能将意群(概念簇)最大化地呈现出来。当然,在方法论上还要求将用户认知微观层面的解释纳入到标签共现关系的分析中来。

2.1 研究对象的选取

研究选择图书情报学科领域的图书作为研究对象。吴丹等[16]对 iSchools 的研究兴趣进行过广泛统计,认为表 1 中的 50 个关键词可以较好地代表图书情报学科领域的研究内容,并展开对图书[17]、学术博客[18]、学术论文[19]的社会标签规范性研究,社会标签与受控的主题词重合率不高这一

结论为数据规范处理提供了一定的参考。图书情报学科规模适中，利用这些关键词在读书或书评网站上检索相关书目，并获得相应标注信息，由此避免收集综合学科带来的概念复杂性。

表 1　图书情报学科关键词

	英文关键词	中文关键词		英文关键词	中文关键词
1	Cataloging/classification	编目	26	Information security	信息安全
2	Digital library	数字图书馆	27	Information architecture	信息构建
3	Information organization	信息组织	28	User interaction	用户交互
4	Metadata	元数据	29	Human-computer interaction	人机交互
5	Information retrieval	信息检索	30	Information consultation	信息咨询
6	Knowledge management	知识管理	31	Social web	社会网络
7	Information management	信息管理	32	Information policy	信息政策
8	Information system	信息系统	33	Information resources development	信息资源建设
9	Reference service	参考咨询服务	34	Information planning	信息规划
10	Information literacy	信息素养	35	Enterprise information	企业信息
11	Library management	图书馆管理	36	Public information	公共信息
12	Library	图书馆	37	Information industry	信息产业
13	Information service	信息服务	38	Information ecosystem	信息生态
14	Intellectual property	知识产权	39	Government information	政府信息
15	Intelligence information	竞争情报	40	Information culture	信息文化
16	Archive	档案	41	Information platform	信息平台
17	Publishing	出版	42	Information portal	信息门户
18	Information	情报	43	Search engine	搜索引擎
19	Data mining	数据挖掘	44	Information economics	信息经济
20	Information visualization	信息可视化	45	Medical information	医学信息
21	Electronic commerce	电子商务	46	Museum	博物馆
22	E-government	电子政务	47	Reference book	工具书
23	Cultural heritage	文化遗产	48	Web information	网络信息
24	Ancient book	古籍	49	E-journal	电子出版物
25	Catalog	目录学	50	Information evaluation	信息评价

2.2　数据获取与处理

豆瓣网(www.douban.com)是一家 Web 2.0 网站，可以自由发表有关书籍、电影、音乐的评论，注册用户已经超过 5,000 万，并以受过高等教育的青年大学生为主，符合对数据源的要求。数据采集步骤如下：① 用关键词在豆瓣读书上检索书目，采集被标注过的书目；② 由于部分图书可同时被若干关键词检索到，根据 URL(或 ISBN)进行去重处理；③ 剔除跟关键词或学科背景明显偏离的书目。最终得到 2,534 本图书及其标注信息。

在豆瓣网中，图书标注系统并不会将所有的历史标签信息显示出来，而是显示按标注次数降序排

列的前 8 位。这意味着,由标签构成的共词网络的演化必然包括标签节点的产生和消亡、共现关系的产生和消亡,以及网络结构的变化。从逻辑上讲,在时间维度上,用户的认知是不断发展的,伴随着各种各样的修正。

由于标签是一种个性化的表达,部分用户倾向使用自己易理解或者有特殊含义的词汇,而这种词汇一般不为大众所广泛接受,表现为大量标注次数为 1 的标签(豆瓣"标签"后括号中的数字表示有几个用户用这个标签来进行标注),因此予以剔除,保留标注次数 2 次及其以上的标签。需要特别说明的是,在传统的共词分析中,研究者需要考虑诸如同义词、反义词、近义词、缩写、中英文互译、简繁统一等问题,在本研究中并不对标签数据进行任何修改。原因主要有:一是用户的语言表达代表着用户的认知状态;二是少量不规范用语并不会影响共词网络的整体结构。

分别在 2011. 12. 31、2012. 08. 31、2013. 01. 02、2013. 07. 31 这四个时间点抓取目标图书的标注信息,获得所需的标签及其共现数据并构建各自的共词矩阵,形成网络 T_1、T_2、T_3、T_4。网络为加权无向图,自编程序对其处理,得到相应的 Pajek 数据格式文件(*. net)。

3 分析与结果

3.1 数据集的统计特征

对上述四个网络的规模及其拓扑结构进行基本统计,如表 2 所示。T_4 相比 T_1 而言,标签节点数量增加了 23. 2%,边的数量(共现关系)增加了 32. 6%。显然,随着时间的推移,网络中节点和连边的数量保持着增长趋势,网络结构关系变得复杂起来。

表 2 标签网络的拓扑结构

网络参数	T_1	T_2	T_3	T_4
标签节点数量 N	2, 196	2, 441	2, 531	2, 706
边的总数 A	9, 777	11, 349	11, 885	12, 964
平均度 $\langle k \rangle$	8. 904	9. 299	9. 392	9. 582
网络密度 ρ	4.057×10^{-3}	3.811×10^{-3}	3.712×10^{-3}	3.542×10^{-3}
平均聚类系数	0. 857	0. 858	0. 857	0. 859
平均路径长度	3. 301	3. 264	3. 261	3. 253
幂指数 γ	2. 310	2. 280	2. 320	2. 320
k_{min}	10	10	11	11

这四个网络的平均度 $\langle k \rangle \approx 9$,密度 $\rho \approx 4 \times 10^{-3}$,说明网络整体稀疏。而网络的平均聚类系数约为 0. 86,平均路径长度约为 3. 3,说明相邻标签的簇内部的网络连接完备程度较高,词汇间容易连接,标注行为易于传递,标签之间的相关性较大,符合"小世界网络"的特征。基于给定的实证数据,作出网络节点的度分布和累积度分布图形(见图 2),将加权网络转变为无权网络,采用极大似然法[20] 直接估计幂指数,T_4 的幂指数为 $\gamma = 2.32 \in (2,3]$,使得幂律成立的度的最小值 $k_{min} = 11$ 。也就是说,网络为非均匀网络(Inhomogeneous network),那些度数相对很高的节点是网络的"集线器"。

因此,无权网络中度数较高的节点是分析用户认知状态的基础和重点。此外,词频较高的节点、加权网络中强度较高的节点也可以反映用户会优先从这些概念出发对图书进行认知;而中介中心性较高的节点反映概念在知识结构中的控制能力。表 3 列出了排名前 20 位的标签,可以发现四者之间有很大的重叠。这些词汇的共同点在于相对概念比较宽泛,说明用户会优先选择易于理解的词进行标注和认知,它们构成了用户知识结构的概念维度。

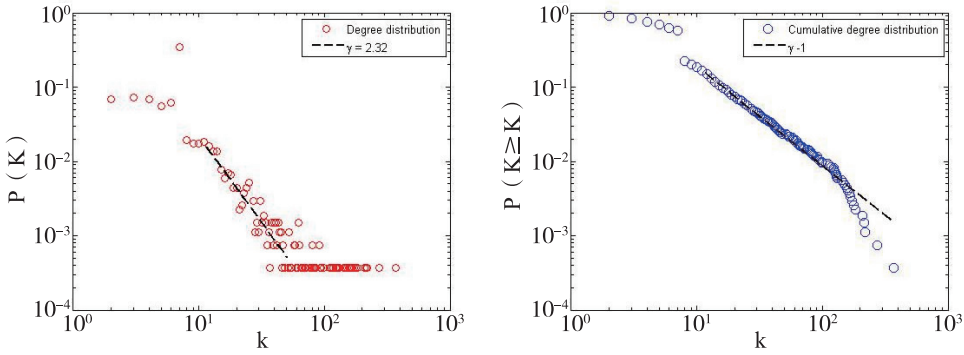

图 2　双对数坐标系中 T_4 的度分布和累积度分布

表 3　T_4 中词频、度、强度和中介中心性排名前 20 的标签节点

序号	词频	度	强度	中介中心性
1	历史	历史	历史	历史
2	博物馆	中国	小说	中国
3	文献学	小说	计算机	计算机
4	计算机	文化	日本	管理
5	出版	计算机	文献学	教材
6	小说	管理	中国	日本
7	中国	艺术	文化	文化
8	管理	日本	管理	小说
9	数据挖掘	文献学	博物馆	出版
10	日本	出版	艺术	设计
11	目录学	文学	数据挖掘	博物馆
12	艺术	博物馆	互联网	艺术
13	版本学	教材	搜索引擎	图书馆
14	文化	图书馆	出版	美国
15	教材	互联网	目录学	互联网
16	图书馆学	外国文学	版本学	经典
17	搜索引擎	设计	漫画	文学
18	图书馆	数据挖掘	古籍	图书馆学
19	电子商务	古籍	文学	杂志
20	古籍	国学	教材	文献学

随着时间的推移，标签持续增加，信息不断交流，网络的平均度增长，使得标签网络适于作为研究用户认知状态动态演化的对象。使用幂函数对边数与节点数的关系进行拟合，两者的关系服从 $A(t) \sim N(t)^{1.3529}$，$R^2 = 0.9994$，服从稠密化幂律（Densification power law）分布，幂指数为 1.3529，介于 1 到 2 之间，这意味着，网络会随着时间的演化变得越来越稠密，但仍然是稀疏的。同时，网络的

平均路径长度随时间的演化呈现出缓慢下降趋势，表现出直径收缩(Shrinking diameters)现象。这些与 Leskovec 等[21]的研究结论相符。

3.2 标签的聚类及可视化

研究使用可视化工具 Gephi(版本 0.8.2)去探索发现用户的知识结构。第一步是聚类，它不仅有助于分析基于概念实体相关性的内容主题，而且还涉及分析过程中产生的概念范畴的演化。Gephi 将 T_4 的知识结构划分为 15 个模块(见表 4)，以每个模块中度数最高的节点加以命名，代表了接受者的认知；同时，将每个模块包含的表 1 中的关键词也列出，它们代表了生产者的认知。

表 4 T_4 的模块划分

模块	名称/接受者	占比(%)	图书/生产者
5	中国	13.63	—
7	文化	12.46	文化遗产、博物馆
1	计算机	11.66	信息检索、数据挖掘、企业信息、搜索引擎、网络信息
6	管理	11.06	知识管理、信息系统、知识产权、竞争情报、情报、电子商务、信息咨询、信息经济
0	历史	9.61	档案
2	文献学	8.60	古籍、目录学、工具书
12	心理学	6.92	社会网络
8	出版	6.59	出版、信息资源建设
3	图书馆	5.67	编目、数字图书馆、信息组织、元数据、信息管理、图书馆、信息可视化、图书馆管理
4	日本	3.70	—
9	设计	2.73	信息服务、信息构建、人机交互
10	绘本	2.45	—
13	考古	2.25	—
14	生活	1.53	—
11	信息安全	1.13	信息安全

为了进一步明晰用户的知识结构，将不含生产者认知的模块过滤掉。图 3 为过滤后 T_4 的知识结构可视化，使用的是 Yifan Hu 布局。从图中可以明显发现度数较高的节点词汇处于核心地位，是最为基本的概念；而中介中心性较高的节点词汇对标签共现影响较大，对网络的流动性至关重要。

在本例中，在“历史”周围聚集着其他具有重要地位的节点，如“古籍”、“博物馆”、“图书馆”、“文化”、“艺术”、“国学”。历史可以是指对过去事实的记载，已过去的事实、经历和事迹的痕迹；还可以指自然界和社会的发展进程、沿革、来历。其语义核心就是“过去的”，这在它周围聚集的节点定义里都有涉及，比如“博物馆”。一般博物馆的工作对象是历史流传至今的实物和自然标本，有文物、文化遗产、人工制品、物品和物证等几种说法，这些具有重要的科学、历史和艺术价值[22]。知识接受者在认知过程中，这些概念会非常自然地结合在一起，这里可以大胆假设受控的主题词表及其定义是影响用户知识结构的关键。

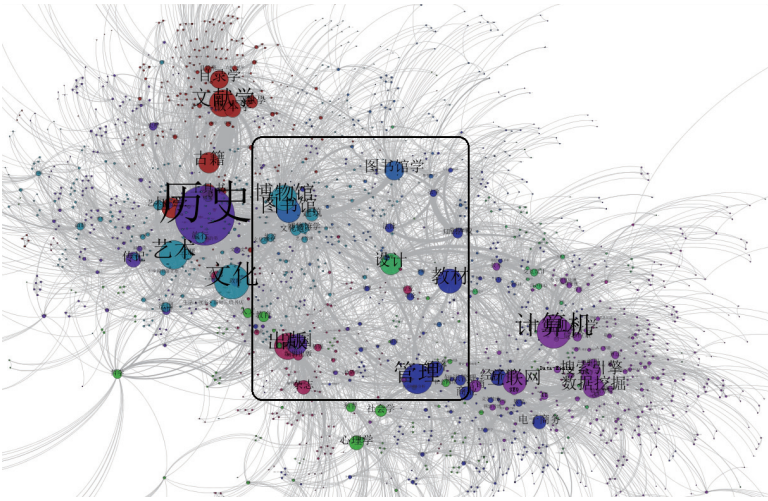

图 3　T_4 的知识结构可视化

值得注意的是：① 图书情报学科领域视为核心的“图书馆”、“情报”和“档案”三个关键词不属于同一个模块，分属于模块 3、模块 6 和模块 0，这说明在图书标注环境下用户的认知中，对这三个概念的理解相对独立，这与专家知识中将它们视为不同的二级学科是一致的；② 从整个图谱展现的结构上看，图书情报学科属于交叉学科，连接了计算机科学、管理学、历史学等，而文献学相对独立。

图 4 表示了信息在生产者和接受者之间进行有效传播后，用户知识结构的变化。在图中，“信息检索”、“数据挖掘”、“企业信息”、“搜索引擎”和“网络信息”共同组成了内容主题，而这些在知识结构中表现出了不同的空间性特征，几何空间的距离表征用户认知心理上的接近程度，如：“计算机”—“信息检索”，“搜索引擎”—“数据挖掘”。此外，图 3 中各个模块的空间分布还反映了各自概念范畴的边界。

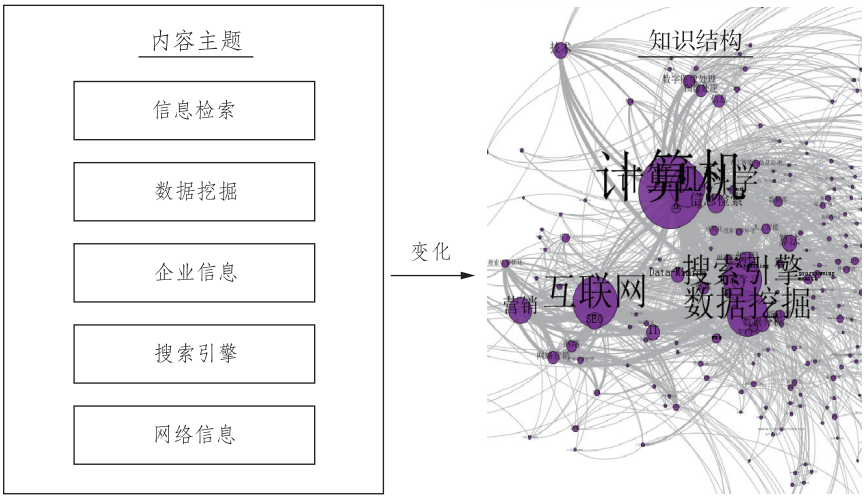

图 4　内容主题向知识结构的转变

4 讨论

总体而言,上述研究结果表明标签共现网络及其可视化在一定程度上可以反映出用户的认知状态。下面将对这些结果反映在标签上的概念和关系进行理论和实践意义上的讨论。

认知心理学认为,概念是人脑反映客观事物本质属性与共同特征的思维形式,是人对一个类别的心理表征,是分类过程的产物。标签表示图书内容中的概念实体,簇间关系表明不同内容主题间的关系,簇内关系则表明了更为一般的概念实体间的关系。图书标注环境下,用户通过加工和运用标签(符号代码)来表征并解释外界事物以及发生在自己身上的事情,属于基于符号取向的知识表征,虽然是不规范的,但却能体现群体智慧。

网络结构被广泛用于认知科学的研究中。理论上,标签共现网络属于语义网络(semantic network),其知识表征是有层次性的,可以经过一定的算法处理构建出标签之间的层次性[23-24]。通常情况下,标签的选择遵循源文本中出现的词汇,但也包括用户自己定义的词汇,其簇内关系比较一般,如:同义、反义、上下位,而且依赖于用户标注图书所处的情境。这种情境通常是指使用标签的目的和范围,比如:标注行为表示参与者的信息共享。标签在概念层次上的选择似乎与传统理论保持一致。概念的内涵和外延是由定义给出的,概念定义和语义关系又是基于经验主义、理性主义、诠释学、实用主义、批判理论等认知理论的[25]。这意味着在日常交流中,人们从概念层次中的基本分类里去选择词汇是普遍现象。在图书标注环境下,用户选择标签也是如此,这些可以从图3中的簇间关系加以证实。

表3中列举的词汇在用户的知识结构中起着主导作用。理论上讲,这些词汇假定是指向内容主题的,往往涉及一个名词或者名词短语。更为直观的是,它们作为一般性概念出现,可以很容易进行联想(association)。对于联想来说,标签共现网络是按照标签的相似程度(语义距离)进行结构组织的,距离的远近(连线的长短)表示了联系的紧密程度,显然,“计算机”—“信息检索”比“计算机”—“算法”更具典型性,也更容易发生联想。不同概念间的语义联想可以更好地确定特殊含义,这也变相解释了以关键词为导向的信息组织和信息检索是非常有用的。然而,如何区分联想类型以及将联想构建成命题需要搜集更多的数据,并结合人类联想记忆模型(HAM)进行研究。

在图书标注环境下,标签实际上是用户增加的有关图书内容背景和应用的表述,这也表明了它具有增值功能。标注行为假设用户自由创造和分配的词汇是建立在用户和图书的互动基础之上,从图1所示的交流系统来看,这也间接代表了接受者和生产者之间的互动。用户会依据自身的认知空间对图书进行定位和理解,这通常包括了用户之间的知识互动关系,以及表4中接受者和生产者的认知对应关系。图书和标签之间的关系暗示了一种标签使用维度,即帮助其他有着相同兴趣、相似思考与表达方式的用户进行搜索。这意味着,标签共现关系可以为用户和标签间的关系(即信息需求)提供建议。在信息搜索状态下,用户的信息需求实际上是一个不确定的状态。用户表达的需求指向有关他们知识异常状态(ASK)的一个概念实体,但是用户并不能很好地定义这些概念实体以及它们之间的关系。因此,用户可以通过标签共现关系去识别对他们特别有用的信息,而标签共现网络的拓扑性质也可以用于显示搜索结果,并帮助用户识别热点概念和预测演化趋势。

5 结语

本文试图将网络结构分析和认知心理学有关理论整合到用户标注行为研究中,重点考察用户的认知特征。标签的序化和聚类可以看成一种关于用户生成内容(UGC)的非常有价值的描述,并能有效引导其他用户收集信息。它的目的在于刻画用户标注图书时的认知状态,并找出主导用户知识结构的概念主体。在网络结构上,标签共现网络具有幂律特征,并服从稠密化幂律分布,它们刻画了用

户认知的宏观演化特征。在微观上,重点分析了标签共现关系,按照语义距离的远近抽取了一些概念样本,并结合定义进行了一些认知心理分析。

需要说明的是,实证部分是通过固定目标图书,在不同时间点观察用户群体的标注行为,进而研究认知特征。这一过程虽然相对理想化,但在一个较大尺度上得出的结论是具有可信性的。当然,实际的情况是生产者在不停生产知识,表现为目标图书呈现出增长状态,因此,后续研究需要考虑从动态增长的视角来观察用户的认知特征。

参考文献

[1] Brookes B C. The foundations of information science: Part I. Philosophical aspects [J]. Journal of Information Science, 1980, 2(3-4): 125-133.

[2] Belkin N J. The cognitive viewpoint in information science [J]. Journal of Information Science, 1990, 16(1): 11-15.

[3] Belkin N J. Information concepts for information science [J]. Journal of Documentation, 1978, 34(1): 55-85.

[4] Ingwersen P. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory [J]. Journal of Documentation, 1996, 52(1): 3-50.

[5] 陆伟,万维雅. 基于认知观点的信息检索交互模型 [J]. 中国图书馆学报, 2005, 31(2): 54-57. (Lu Wei, Wan Weiya. An interactive model of information retrieval based on cognitive viewpoint [J]. Journal of Library Science in China, 2005, 31(2): 54-57.)

[6] 迪莉娅. 西方信息行为认知方法研究 [J]. 中国图书馆学报, 2011, 37(2): 97-104. (Di Liya. Studies on cognitive approach to information behavior in western countries [J]. Journal of Library Science in China, 2011, 37(2): 97-104.)

[7] Ding Y, Jacob E K, Zhang Z, et al. Perspectives on social tagging [J]. Journal of the American Society for Information Science and Technology, 2009, 60(12): 2388-2401.

[8] Marlow C, Naaman M, Boyd D, et al. Tagging paper, taxonomy, flickr, academic article, to read [C]// Proceedings of the Seventeenth Conference on Hypertext and hypermedia. ACM, 2006: 31-39.

[9] Cattuto C, Schmitz C, Baldassarri A, et al. Network properties of folksonomies [J]. AI Communications, 2007, 20(4): 245-262.

[10] 李亚婷,马费成. 基于标签共现的社会网络分析研究 [J]. 情报杂志,2012,31(7):103-109. (Li Yating, Ma Feicheng. Social network analysis based on tags co-occurrence [J]. Journal of Intelligence, 2012, 31(7): 103-109.)

[11] 易明,曹高辉,毛进,等. 基于 Tag 的知识主题网络构建与 Web 知识推送研究 [J]. 中国图书馆学报, 2011, 37(4): 4-12. (Yi Ming, Cao Gaohui, Mao Jin, et al. Knowledge topic network construction and Web knowledge push based on tag [J]. Journal of Library Science in China, 2011, 37(4): 4-12.)

[12] Yoon K. Conceptual syntagmatic associations in user tagging [J]. Journal of the American Society for Information Science and Technology, 2012, 63(5): 923-935.

[13] 庄晓喆,吴丹. 图书标注系统的比较分析 [J]. 图书情报工作, 2012, 56(5): 73-77. (Zhuang Xiaozhe, Wu Dan. Comparison and analysis of book tagging systems [J]. Library and Information Service, 2012, 56(5): 73-77.)

[14] 马费成,郝金星. 概念地图在知识表示和知识评价中的应用(I)——概念地图的基本内涵 [J]. 中国图书馆学报, 2006, 32(3): 5-9, 49. (Ma Feicheng, Hao Jinxing. Applications of concept maps in knowledge representation and knowledge evaluation (I) [J]. Journal of Library Science in China, 2006, 32(3): 5-9, 49.)

[15] Spiteri L F. The structure and form of folksonomy tags: The road to the public library catalog [J]. Information Technology and Libraries, 2007, 26(3): 13-25.

[16] Wu D, He D Q, Jiang J P, et al. The state of iSchools: An analysis of academic research and graduate education [J]. Journal of Information Science, 2012, 38(1): 15-36.

[17] 吴丹,林若楠,冯倩然,等. 社会标签的规范性研究——图书标注 [J]. 图书馆论坛, 2012, 32(1): 1-7, 56. (Wu Dan, Lin Ruonan, Feng Qianran, et al. The standardization of social tagging: A study on tagging books [J]. Library Trib-

动态

德国数字图书馆提供开放数据服务

德国数字图书馆(Deutsche Digitale Bibliothek, DDB)于2013年11月4日正式发布应用程序接口(API),向公众提供开放数据服务。图书馆以 API 的方式提供馆藏数字资源的开放共享在全球尚属首次。

德国数字图书馆的元数据是以不保留任何权利的知识共享(CCO)形式授权发布,任何用户都可利用德国数字图书馆的应用程序接口运行自己编写的计算机程序,以不受任何限制的方式使用这些元数据。用户可以借此利用公共数字图书馆的资源,在网站或者手机上开发新的互联网信息服务,包括充分利用"德国数字文化遗产"等资源。德国数字图书馆计划为此组织起一个创新应用大赛,以激励开发者和公众充分利用图书馆的数字化网络化资源。

资料来源

1. Deutsche Digitale Bibliothek releases API. https://www.deutsche-digitale-bibliothek.de/content/news/2013-11-04-001.
2. API der Deutschen Digitalen Bibliothek. https://api.deutsche-digitale-bibliothek.de/doku/display/ADD/API+der+Deutschen+Digitalen+Bibliothek.
3. API Terms and Conditions. https://www.deutsche-digitale-bibliothek.de/content/terms/api/.

(顾立平　整理)

une, 2012, 32(1): 1-7, 56.)

[18] 吴丹,杨艳,马曦. 社会标签的规范性研究——学术博客标注[J]. 情报资料工作, 2011(6): 11-15. (Wu Dan, Yang Yan, Ma Xi. The normative studies of social label: Academic blog labeling [J]. Information and Documentation Services, 2011(6): 11-15.)

[19] 吴丹,王艳妮. 社会标签的规范性研究——学术论文标注[J]. 图书馆, 2012(1): 85-88. (Wu Dan, Wang Yanni. The standardization of social tagging: A study on tagging academic articles [J]. Library, 2012(1): 85-88.)

[20] Clauset A, Shalizi C R, Newman M E J. Power-law distributions in empirical data [J]. SIAM Review, 2009, 51(4): 661-703.

[21] Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: Densification and shrinking diameters [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 1-41.

[22] 宋向光. 博物馆定义与当代博物馆的发展[J]. 中国博物馆, 2003(4): 1-6. (Song Xiangguang. Definition of museum and its contemporary development [J]. Chinese Museum, 2003(4): 1-6.)

[23] Heymann P, Garcia-Molina H. Collaborative creation of communal hierarchical taxonomies in social tagging systems [R]. Stanford InfoLab Technical Report, 2006.

[24] Begelman G, Keller P, Smadja F. Automated tag clustering: Improving search and exploration in the tag space[C]// Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, 2006: 15-33.

[25] Stock W G. Concepts and semantic relations in information science [J]. Journal of the American Society for Information Science and Technology, 2010, 61(10): 1951-1969.

马费成　武汉大学信息资源研究中心教授,博士生导师。通讯地址:武汉市珞珈山。邮编:430072。

张　斌　武汉大学信息资源研究中心博士研究生。通讯地址同上。

(收稿日期:2013-04-02;修回日期:2013-08-15)