doi:10.3969/j.issn.1001-4616.2023.03.015

多维属性视角下学者用户画像构建及合作学者推荐研究

王大阜,邓志文,贾志勇,王　静

(中国矿业大学图书馆 江苏 徐州 221116)

[**摘要**]　从多个维度属性构建学者用户画像,向目标学者推荐属性特征相似度高和易建立合作关系的合作学者,有助于增强学术交流合作、促进科研产出. 本文以学者的论文成果为数据源,设计并探讨了合作学者推荐系统模型. 首先采用 Louvain 社区发现算法划分学者社团,其次根据学者的基础属性、学术能力、研究兴趣、社交影响力四个维度构建用户画像,并从学者合作网络中获取合作关系强度、Katz 相似性指标,最后根据候选学者的融合推荐评分实现合作学者推荐. 通过构建学者用户画像,呈现学者全面的特征信息,赋予推荐结果可解释性. 实证表明本文所提出的推荐模型具有良好的推荐效果,为目标学者的合作学者遴选提供了决策依据.

[**关键词**]　用户画像,智慧图书馆,合作学者推荐,作者合作网络

[**中图分类号**]TP391.3　[**文献标志码**]A　[**文章编号**]1001-4616(2023)03-0112-11

Research on the Construction of Scholar User Portrait and the Recommendation of Cooperative Scholars from the Perspective of Multidimensional Attributes

Wang Dafu, Deng Zhiwen, Jia Zhiyong, Wang Jing

(Library, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Building user portraits of scholars from multiple dimension attributes and recommending collaborators who have high similarity in attribute characteristics and are easy to establish cooperative relationships to target scholars will help strengthen academic exchanges and cooperation, promote scientific research output. Taking the bibliographic data of scholars' papers as the data source, this paper designs and discusses the model of collaborative scholars' recommendation system. Firstly, Louvain algorithm is used for community discovery. Secondly, user portraits are constructed according to the four dimensions of scholars' basic attributes, academic ability, research interest, social influence, and indicators of cooperation relationship strength and Katz similarity indicator are obtained from the author's cooperation network. Finally, the recommendation of cooperative scholars can be realized according to the fusion recommendation score of candidate scholars. By constructing a scholar user portrait, the comprehensive characteristic information of the scholar is presented, giving the recommendation results interpretability. The empirical results show that the recommendation model proposed in this paper has a good recommendation effect, which provides a decision-making basis for the selection of cooperative scholars.

Key words: user persona, smart library, recommendation of cooperative scholars, author cooperative network

随着科学技术的快速发展和科学研究的不断深入,学者们为了共同解决重大的科研难题,经常跨学科、跨机构、跨地域地进行相互合作,有效促进了学术知识的流动和交叉渗透,并通过合著方式发表论文,逐渐形成稳定的科研合作关系. 相应地,学者间的合作关系网络可以建模为一种复杂网络类型,即科研合作网络[1]. 普赖斯指出"科研合作已成为当今科学发展的重要动力",换言之,科研合作网络的建立日益成

收稿日期:2022-11-25.

基金项目:江苏省高校哲学社会科学研究项目(2022SJYB1129)、国家社会科学基金项目(22BTQ023).

通讯作者:王大阜,硕士,馆员,研究方向:推荐系统、知识图谱. E-mail:wdf@ cumt.edu.cn

为影响科研成果产出能力的重要因素[1]. 大多数学者在科研起步阶段,通过师生、同事或朋友等熟人关系寻找科研合作对象,当某项研究需要进行跨学科甚至是跨地域合作的时候,就要凭借其历史合作学者、学术社交好友等社会关系来判断选择,但往往由于掌握的信息不全面、不准确,造成找到的合作学者未必合适. 科研大数据时代,面对海量的学术资源和学者,学者们难以就其感兴趣的学科领域进行相关文献或学者的择优选取,从而给学者们造成"信息迷航"的局面和困扰. 自智慧图书馆概念提出以来,围绕图书或论文等文献资源的个性化推荐服务,一直是学界和业界关注的研究热点,而关于推荐"人"(如权威学者或合作学者)的研究相对较少. 本文研究内容是以合作学者为推荐对象,为目标学者寻找潜在、适配的合作学者,这有利于增强学者间的学术合作,激发其科研活跃度和科研成果产出.

1 研究现状

当前,学者推荐的相关研究主要围绕权威学者和合作学者这两种推荐对象. 前者是推荐与用户的研究兴趣匹配度高的权威专家,李春英等[2]利用学术社交网络中影响力大的学者圈对学术社区检测,并实现社区内权威学者推荐服务. 熊回香等[3]基于学者的特征词和共被引关系,组合计算学者研究内容的相似性,实现个性化学者推荐. CHAKRABORTY 等[4]基于主题模型和聚类算法,通过 MOOC 平台为学生寻找与其兴趣领域相关的学者. 后者是根据科研合作网络中的连通关系推荐潜在的合作学者,这也是本文所要研究的主题,相关文献从以下几个视角展开研究:(1)社区发现及研究兴趣. 刘萍等[5]在社区划分的基础上,基于 LDA 构建的作者兴趣模型,实现相似研究兴趣的合作学者推荐. (2)学术能力、研究兴趣及社交属性. 熊回香等[6-7]从学者的学术能力挖掘候选推荐学者的知识覆盖度,结合历史合作关系挖掘合作质量实现合作学者推荐. 其还基于用户相似度和信任度对虚拟学术社区科学网中的学者进行推荐,提高学者推荐的质量. 董文慧等[8]挖掘学者的自然属性、兴趣属性、能力属性、社交属性(中心性)四个维度特征以构建用户画像,并基于学者偏好开展科研合作者推荐. 杨梦婷等[9]基于 LDA 主题模型和时间因子提取学者的动态科研兴趣特征,并对研究兴趣相似的学者进行聚类,综合学者的科研能力和社交属性两个维度构建学者推荐模型. 王妞妞等[10]从可合作性和易合作性两个维度挖掘科研合作者. JIN 等[11]基于文本相似度、社会相关度以及个人贡献度构建多维特征学者推荐模型 Mul-RSR,有效提高推荐准确性和可解释性. YUAN 等[12]从学术社交网络中提取学者属性特征和文本特征,通过图嵌入框架描述学者间复杂的异质学术网络,进而推荐相似学者. (3)研究兴趣及链路预测指标. 秦红武等[13]按照学者学术水平特征进行聚类,在同一类簇中利用链路预测算法中的 Katz 指标及兴趣指标对学者进行相似度计算及排序推荐. 汪俊等[14]通过综合分析科研社交网络中专家所具有的知识信息以及社会关系信息,构建链接预测 SVM 分类模型进行专家推荐. (4)网络表示学习. 刘云枫等[15]和张金柱等[16]分别采用 LINE、Word2vec 网络表示学习方法学习作者的上下文语境向量,进而发现作者间的关联性,提高推荐准确率.

综上,当前相关研究所采用的技术手段各有侧重点,主要手段是针对目标学者的相关属性,比如研究兴趣、学术能力、社交属性及链路预测指标等对合作学者进行相似度计算,其中研究兴趣是必要属性,其表征方法有特征词向量、主题模型等. 辅助手段是采用社区划分方法识别社群团体,或者采用聚类方法识别学者类群,由于推荐学者需要从目标学者相连通的合作网络中进行筛选,因此社区划分更为重要,同时也能够提升推荐模型的计算性能. 在现实世界中,选择理想的合作学者应当考虑兴趣、能力、关系三个要素,其中能力包括学术能力和社交能力,关系包括合作网络中已有的和未知的连通性. 基于此,本文研究综合两种手段,首先将学科领域中大规模的作者合作网络进行划分,并构建多种指标标签为每位学者构建用户画像. 进一步地,对多种指标进行融合评分,以向目标学者推荐兴趣、学术能力相当,合作关系易达的合作学者,从而有效增强学者之间的科研合作关系.

2 推荐系统模型设计

2.1 模型架构

合作学者推荐模型架构如图 1 所示,以 CNKI 期刊论文的题录数据作为数据源,首先根据论文合著关系构建学者之间的合作网络,其次对科研合作网络进行社区划分,然后提取学者的论文特征和网络特征,从而为每位学者构建用户画像. 标签主要包括基础属性(姓名与机构)、学术能力(学术水平与学术影响

力)、研究兴趣、社交影响力(中心性)、合作关系强度及 Katz 相似性指标,其中学术能力、研究兴趣从学者发表的论文中提取,代表学者的论文特征;中心性、学者之间的合作关系强度及 Katz 相似性指标从作者的科研合作网络中提取,代表学者的合作关系网络特征. 最后实现合作学者的遴选推荐,算法流程为:首先寻找与目标学者同在一个社区的学者集(即候选学者集),其次依据学术能力和研究兴趣进行相似度计算,排除学术能力和研究兴趣度相似度较低的候选学者,然后依据学术能力、研究兴趣、中心性、合作关系强度以及 Katz 相似性指标这五种学者特征进行加权融合计算推荐评分,特征权重通过层次分析法(AHP)确定,最终评分较高的 TOP-N 候选学者即为推荐合作学者集.

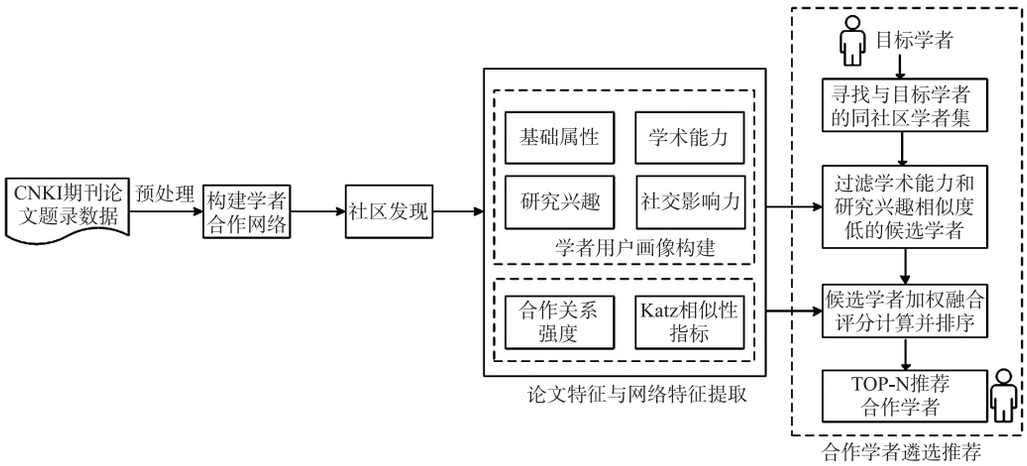

图 1　合作学者推荐模型架构

Fig. 1　Partner scholar recommendation model architecture

2.2　社区划分

社区划分是根据网络的属性特征,将网络节点划分到具有特殊含义的社区中的过程. 社区划分的算法包括模块度优化算法、分层聚类算法、图分割算法,其中模块度优化算法应用最为广泛,适用于大规模网络中的社区划分处理[17-18]. 模块度(Modularity)是衡量网络社区划分好坏的度量方法,其物理意义是社区内部边的比例减去在同样社区结构下任意边的比例期望值. 模块度越接近 1,表明社区结构越明显,社区划分效果越好. Louvain 算法是基于模块度最优化思想的启发式算法,其基本原理是通过迭代合并邻近节点或节点群来划分社区,每一次合并过程都伴随着模块度值的增加,若下一次合并的模块度增量不为正,那么此次合并后取得最优社区划分结果[17].

2.3　学术能力

文献计量学中,学者的学术能力是根据学者的科研产出的数量和质量进行综合评价,评价指标主要包括 G 指数、H 指数和 M 指数三种,其中 H 指数在学界得到最为广泛的认可和应用,H 指数由美国物理学家 Jorge E.Hirsch 提出,其计算方式是学者至少发表了 H 篇至少被引用 H 次的文献[19]. 然而以上三种评价指标没有考虑多作者合作情况下各个学者贡献度的区别以及各个期刊学术质量的区别,导致评价“有失公允”[20]. 为此,本文采用学术水平和学术影响力这两个指标对学者的学术能力进行评价度量.

2.3.1　学术水平

论文发表的数量和质量是体现学者学术水平的重要指标,本研究根据学者发表论文数量、期刊级别、署名次序三个因素综合评估学者的学术水平.

(1)期刊级别. 论文的质量与期刊影响因子(impact factor,IF)密切相关,IF 直观反应期刊整体的论文质量,也表征同一期刊中每篇论文的通用质量. 期刊级别分为 A、B、C 和 D 四个,分别表示领域内的权威期刊、重点期刊、核心期刊和普通期刊. 根据 CNKI 2021 年图书情报类期刊的 IF 值分布,将期刊划分为四个等级区间,每个等级取平均影响因子进行量化.

$$Class=\begin{cases}3.883 & A\text{ 级别},IF\in[3.03,7.343]\\ 2.464 & B\text{ 级别},IF\in[1.683,2.934]\\ 0.984 & C\text{ 级别},IF\in[0.811,1.41]\\ 0.552 & D\text{ 级别},IF\in[0.331,0.771]\end{cases}\tag{1}$$

(2)学者贡献度. 在合著论文中,学者不同署名次序表征其对论文的贡献度,通常署名次序靠前的作者贡献度较大. 另外,通讯作者的贡献度通常与第一作者相近,但鉴于论文的题录数据中没有标注通讯作者,因此本文未计算通讯作者的贡献度. 假定某学者的发表论文集合为 P,某篇论文 p 的合著学者数为 n,学者署名次序为 k,$1 \leqslant k \leqslant n$,则该学者在论文 p 中的贡献度 *Contribution* 的计算公式为:

$$Contribution = \frac{n - k + 1}{\sum_{k=1}^{n} k} \tag{2}$$

综合学者发表论文的期刊等级与贡献度两个因素,最终该学者的学术能力指标 Level 的计算公式为:

$$Level = \sum_{p \in |P|} Class \times Contribution \tag{3}$$

2.3.2 学术影响力

在同一研究领域,某篇论文被引次数越高,表明该论文更受学者们的青睐和认可,其学术影响力较高,其中可能包括历久不衰的经典文献. 综合学者发表论文的期刊等级、贡献度及被引用频次三个因素,学者影响力指标 *Impact* 计算公式如下,式中 *Citation* 表示论文的被引用频次.

$$Impact = \sum_{p \in |P|} Class \times Contribution \times Citation \tag{4}$$

假定学者的学术水平特征向量表示为 $Ability = (Level, Impact)$,鉴于表征学者能力的学术水平和学术影响力会存在明显的量纲差异,在计算学者能力相似度之前,采用 *Min-Max* 离差标准化法进行归一化处理,最终学者 a 与 b 的学术能力相似度为:

$$Aiblity_{Sim(a,b)} = \mathrm{Cos}(Ability_a, Ability_b) \tag{5}$$

2.4 研究兴趣

学者的研究兴趣是一种从宏观到微观的层级关系,例如某学者的研究兴趣为:智慧图书馆→信息技术,前者表示研究领域,后者表示研究主题. 本文选取的智慧图书馆领域的期刊文献,研究兴趣可简化为研究主题. 聚类和主题模型是学者识别学科领域研究主题的常用技术方法,据有关学者研究表明,对于学术文献文本而言,全局语义信息有助于挖掘出细粒度的主题,因此主题模型的主题挖掘效果优于聚类[21]. 潜在狄利克雷分布(latent dirichlet allocation,LDA)主题模型通过语义分析技术,对上下文理解后挖掘出隐含的抽象主题[22]. LDA 模型由文档、主题、词组成的三层贝叶斯概率分布生成,其基本思想是:一篇文档隐含了多个主题,一个主题由多个词语构成,通过迭代模拟文档生成过程,进而识别文档潜在的主题信息.

语料库中每篇文档 d 生成的过程如图 2 所示,步骤是:(1)从超参数 α 采样生成“文档-主题”概率分布 θ,对于文档 d 的每个词从 θ 参数的 Multinomial 分布中采样生成主题 z;(3)从超参数 β 采样生成“主题-词”概率分布 ψ,从 ψ 参数的 Multinomial 分布中采样生成词 w;(4)上述三个步骤重复 N 次,产生文档 d.

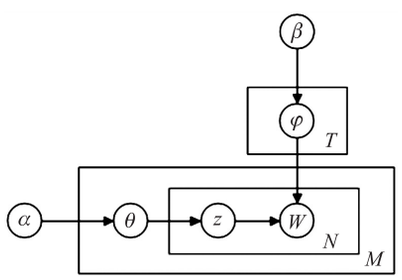

图 2 LDA 主题模型结构图

Fig. 2 LDA theme model structure diagram

假定文档集包含的研究主题集合为 T,$T = \{topic_1, topic_2, topic_3 \cdots topic_n\}$,某学者的相关论文集的文档——主题分布矩阵可表示为:

$$Matrix(D) = \begin{bmatrix} P_{11} & P_{12} & P_{13} & \cdots & P_{1n} \\ P_{21} & P_{22} & P_{23} & \cdots & P_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{m1} & P_{m2} & P_{m3} & \cdots & P_{mn} \end{bmatrix} \tag{6}$$

其中,m 表示文档数,矩阵中的每一行表示每篇文档的主题概率向量,记作 $\boldsymbol{V}(P_i)$,$\boldsymbol{V}(P_i) = (P_{i1}, P_{i2}, P_{i3}, \cdots, P_{in})$. 学者的研究主题特征向量记作 $\boldsymbol{V}(T)$,其计算公式为:

$$\boldsymbol{V}(T)=\frac{\sum_{i=1}^{m}\boldsymbol{V}(P_i)}{m} \tag{7}$$

学者 a 和 b 的研究主题的余弦相似度为:

$$Interest\_\mathrm{Sim}(a,b)=\mathrm{Cos}(\boldsymbol{V}(T)_a,\boldsymbol{V}(T)_b) \tag{8}$$

2.5 科研合作网络

科研合作网络的概念起源于文献计量学,是科学知识图谱领域重要的研究对象之一. 其基本思想是基于某个学科或研究领域,根据科技文献中的合著关系构建科研合作网络,用于揭示网络内部结构和个体对整个群体的影响[23]. 科研合作网络可以从微观、中观、宏观三个视角分别建立作者、机构及国家的合作网络[24],合作网络中的节点代表作者、机构、国家三种类型实体,节点的大小通常代表发文量,连线(边)代表作者、机构或国家间的合作关系,连线的厚度代表关系权重,通常用合作频次表示.

2.5.1 中心性

社会网络分析是研究社会结构的一种全新的科学范式,中心性是其较为重要研究范畴,用于反映行动者在社会网络中所处的核心地位与影响力[25]. 中心性的度量指标有三个:度中心性(Degree Centrality)、中介中心性(Betweenness Centrality)和接近中心性(Closeness Centrality),在学者合作网络中分别代表了学者的合作活跃度、控制他人合作的能力以及与他人建立合作的能力. 此外,中心性可分为绝对中心性和相对中心性,绝对中心性是指一个节点的中心性,而相对中心性是其标准化形式,其计算方法为绝对中心性与图中节点的最大可能的中心性之比[25]. 假定学者的社交影响力向量表示为 Centrality = (Degree, Betweenness, Closeness),则学者 a 与 b 的社交影响力的余弦相似度为:

$$Social\_\mathrm{Sim}(a,b)=\mathrm{Cos}(Centrality_a,Centrality_b) \tag{9}$$

2.5.2 合作关系强度

作者合作网络是根据作者的历史合作关系而生成的,该合作关系可以用方形共现矩阵来表示,元素代表作者在合著论文中的共现频次,主对角线上的元素表示每位作者出现的频次[1]. 学者共现频次的高低反映了两作者的亲疏关系,若两作者的共现频次越高,代表其关系越亲近,反之,代表其关系越疏远. 同时,这种亲疏关系也能够代表作者的信任程度,关系越亲近的作者的信任度越高,越容易在往后建立科研合作关系[26]. 为抑制高频次对相关性分析的影响,通过 Ochiai、Dice、Jaccard 等系数算法对共现矩阵进行处理,从而转换成标准化的相关矩阵,矩阵数值区间为[0,1]. 其中最常采用的是 Ochiia 系数算法,其计算公式如下:

$$Ochiia=\frac{c_{ij}}{\sqrt{c_i\times c_j}} \tag{10}$$

式中,c_{ij}表示作者 i 和作者 j 共同出现的频次,c_i 表示作者 i 出现的频次,c_j 表示作者 j 出现的频次.

2.5.3 链路预测

链路预测是通过复杂网络中已知的节点关系预测未来节点产生连接的可能性[27]. 在科研合作网络中,链路预测被广泛应用于学者推荐,其中基于网络结构信息的相似性链路预测方法最为常用,例如基于邻居节点的相似性指标主要包括 CN(共同邻居)、AA(节点赋权)、RA(资源分配)、PA(择优连接)等,其中后三者是 CN 指标的扩展,以惩罚或奖励大度的共同邻居节点为原则. 由于基于邻居节点的相似性分数区分度不高,预测精度受限,随后有学者提出基于路径的相似性指标,主要包括 LP(局部路径)、Katz 和 LHN-Ⅱ. 张斌等[28-29]针对 CSSCI 中的七门学科进行合作网络链路预测,使用不同指标对链路预测效果进行评价,实验表明,基于路径的指标的总体表现要优于基于共同邻居的相关指标,就图书情报学科而言,Katz 效果最好.

Katz 指标的基本思想是考虑网络中所有节点之间各种长度的路径,较长(高阶)路径赋予的权重低于较短路径的权重,以此模拟学者合作关系的传递随路径长度逐渐衰减的过程. Katz 指标定义为:

$$S_{xy}=\sum_{l=1}^{\infty}\beta^l\cdot|path_{x,y}^{\langle l\rangle}|=\sum_{l=1}^{\infty}\beta^l\cdot(A^l)_{x,y} \tag{11}$$

式中,β 表示衰减因子,是控制路径权重的可调参数,通常设定为 0.005,$|path_{x,y}^{\langle l\rangle}|$ 表示节点 v_x 和 v_y 之间路

径长度为 l 的路径数量. 若上述公式收敛,β 的取值应小于邻接矩阵最大特征值的倒数,则该公式可转化为式(12),其中 $\boldsymbol{A}$ 为网络的邻接矩阵,$\boldsymbol{I}$ 为单位矩阵.

$$S=(I-\beta\times\boldsymbol{A})^{-1}-\boldsymbol{I} \tag{12}$$

2.6 学者用户画像

用户画像(Persona)于 1998 年由交互之父 Alan Cooper 提出,他指出用户画像是“基于用户真实数据的虚拟代表”. 用户画像是在用户真实数据的基础上用来勾勒用户基本属性,描述用户兴趣特征的重要技术手段,通过全面细致地刻画用户的信息全貌,进而为用户提供精准的个性化推荐服务[30-31]. 本文从学者的基础属性、学术能力、研究兴趣、社交影响力四个维度构建学者用户画像标签,各维度标签包括的指标为:基础属性标签(姓名、机构),学术能力标签(发文量、篇均被引次数、篇均贡献度、篇均期刊级别、学术水平、学术影响力),研究兴趣标签(研究主题),社交影响力标签(度中心性、中介中心性、连接中心性).

3 实证分析

3.1 数据来源及预处理

进入 CNKI 期刊高级检索页面,输入检索条件主题 OR 篇关摘=“智慧图书馆”OR 主题 OR 篇关摘=“智慧化图书馆”OR 主题 OR 篇关摘=“图书馆 * 智慧服务”,期刊来源类别为北大核心和 CSSCI,时间范围为 2011 年 1 月—2022 年 2 月. 剔除选题指南、名人专访、著录错误等题录数据,最终有效期刊论文为 927 篇. 本文将 927 篇期刊论文的摘要文本集作为训练 LDA 模型的语料库,预先采用哈尔滨工业大学停用词表,并增加了“智慧”“图书馆”“智慧图书馆”等本位词以及“目的”“意义”“方法”等无意义的词汇.

此外,本文使用较为流行的结巴分词工具对文本进行分词处理,将文本切分为多个词条(Token),为 LDA 模型训练和文本向量化表示做前期准备. 众多分词工具对于专业性术语词汇,不能精准识别,这些术语词汇需要自定义词典文件并进行加载. 术语词汇分为两种:通用术语和学科领域术语. 一般性的术语词汇,包括“5G”、“大数据”、“物联网”、“人工智能”等,图书馆学领域的术语词汇,包括“知识服务”、“空间再造”、“转知成慧”、“阅读推广”等. 为了使得分词和语义表征更加精准,笔者通过 BICOMB 工具从题录数据中提取全部关键词,进而构建包含 1 724 个词的自定义词典.

3.2 社区网络发现

笔者通过社会网络软件 Gephi 绘制作者合作网络图谱,如图 3 所示.

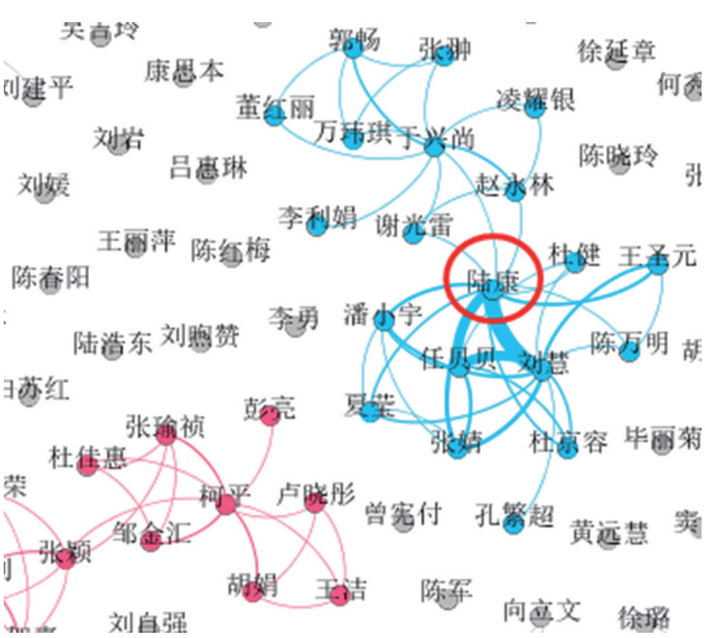

图 3 作者合作网络图谱

Fig. 3 The cooperative network graph

经 Gephi 统计,网络中共计有学者 1 309 名,边 1 373 条,总连接数 1 586 条. Gephi 的模块度(Modularity)统计功能采用 Louvain 算法划分社区,共识别社区 276 个,模块度为 0.978,表明社区结构非常显著,相关指标统计值如表 1 所示,其中学者陆康(图 3 中红色圆圈标识)的合作团体的模块度类(Modularity Class)为 251.

表 1　作者合作网络的指标统计

Table 1　Index statistics of author cooperation network

平均度	图密度	模块度	平均聚类系数	平均路径长度
2.67	0.003	0.978	0.901	1.79

以学者陆康为例,使用 UCINET 软件绘制其合作网络图谱,如图 4 所示,节点代表某个学者,边的厚度代表学者间的关联强度. 由图可见,该合作网络规模较大,由 20 个学者节点构成,是整个网络的第二大连通网络(最大连通网络是以杨新涯为核心的合作网络,共有节点 37 个). 其中与陆康有过直接合作关系的学者有 12 名(度中心性=12),刘慧合作次数高达 16 次,任贝贝高达 10 次,另外通过学者于兴尚和刘慧可间接建立合作关系的学者有 7 名.

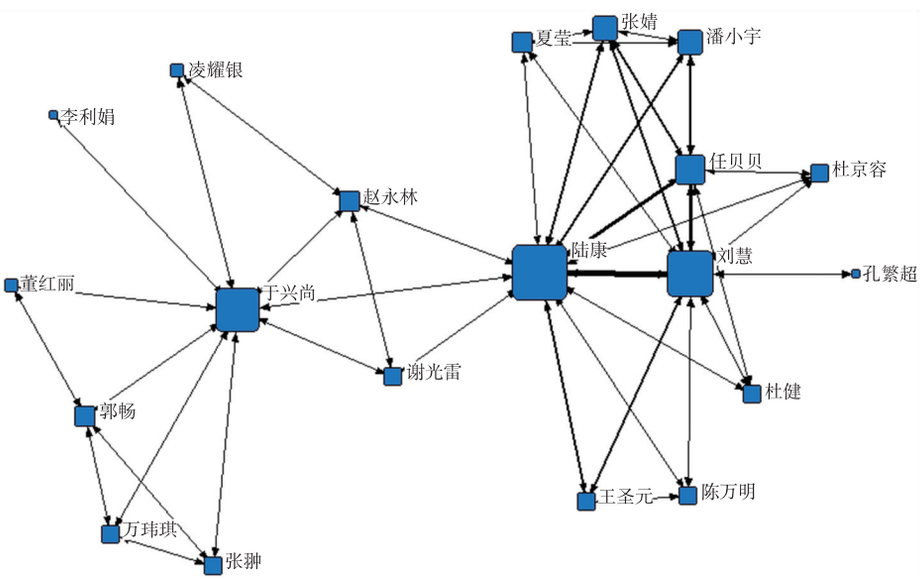

图 4　学者陆康的合作网络图谱

Fig. 4　The cooperative network graph of scholar Lu Kang

3.3　基础信息

笔者编写 Python 脚本对学者的学术能力指标进行统计,由于篇幅原因,依据普赖斯定律识别具有重要学术地位的高产学者(发文量≥5),其中 TOP-10 的高产学者列表如表 2 所示. 由表可见,学者王世伟发文量虽然排名第五,但是其论文均是独著发表,而且被引用次数高达 1524,其学术能力和学术影响力指标均排名第一,遥遥领先于其他学者. 学者曾子明被引次数和篇均贡献度较高,其学术能力排名第三,学术影响力排名第二. 学者邵波发文量排名第一,但是其论文的篇均贡献度低于学者王世伟、曾子明,其学术能力排名第二,影响力排名第三.

表 2　高产学者学术能力指标统计

Table 2　Statistics of academic ability indicators for high-yield scholars

序号	作者	发文量	总被引频次	篇均被引频次	篇均期刊等级	篇均贡献度	学术能力	学术影响力
1	邵波	26	739	28.42	B	0.359	23.67	693.30
2	陆康	21	107	5.10	B	0.468	18.03	113.40
3	杨新涯	19	315	15.58	B	0.403	21.45	336.52
4	刘慧	18	124	6.89	B	0.413	14.91	64.91
5	王世伟	16	1 524	95.25	B	1.000	43.56	4 550.54
6	许正兴	12	48	4.00	B	0.944	22.93	105.69
7	曾子明	12	445	37.08	A	0.625	23.33	844.75
8	杨文建	10	81	8.10	B	0.600	14.14	118.42
9	任贝贝	10	48	4.80	B	0.160	2.79	13.68
10	单轸	9	198	22.00	B	0.611	13.08	297.73

笔者采用 Python 版本的 Gensim 软件包训练 LDA 模型,以提取论文摘要文本中隐含的研究主题,相关参数设置为:超参数 α 设为 0.05,超参数 β 设为 0.01,iterations(迭代次数)设为 200,dictionary(字典)过滤词频小于 5 的词. 最终通过困惑度(Perplexity)指标[22]评估和确定最优主题数为 10. 各个研究主题及相关词(词频排名前 20)的统计结果如表 3 所示.

采用 Gensim 软件包的 LDA 模型计算得出高产学者发表论文所属的 10 个主题概率,选取最大概率值作为论文所涉及的研究主题,结果如表 4 所示,表中的中心性指标由 UCINET 软件计算得出. 通过公式(7)进一步计算出每位学者的研究兴趣向量,结果如表 5 所示.

表 3　研究主题及相关词的统计结果

Table 3　Statistical results of research topics and related words

序号	研究主题	相关词
1	个性化服务	用户 数据 模型 技术 环境 框架 设计 系统 感知 大数据 智慧服务 信息 资源 情境 特征 维度 情境感知 行为 融合 影响
2	智慧图书馆构建要素(馆员、技术、服务)	技术 模式 概念 智慧馆员 建设 智慧服务 内涵 要素 互联网 结合 特征 系统 阐述 理念 论述 智能化 图书馆智慧服务 个性化服务 时代 工作
3	智能技术	技术 物联网 融合 建设 高校图书馆 智能化 微服务 模式 结合 管理 5g 技术 云计算 5g 工作 未来 平台 领域 rfid 技术 建立 模块
4	未来发展趋势	建设 智慧图书馆建设 国内 我国 技术 文献 现状 未来 标准 参考 图书馆空间 系统 总结 智慧空间 机器人 理论研究 案例 调研 概念 建议
5	十四五规划	公共图书馆 环境 文化 建设 社会 创新 十四五 规划 转型 智慧城市 理念 促进 需要 空间 时期 成为 技术 城市 打造 方向
6	服务模式转型	人工智能 技术 数字图书馆 互联网 文献 中国 管理 传统 转型 模式 智慧服务 国内外 图书馆服务 服务内容 服务平台 融合 馆藏 智慧化 资源 重要
7	智慧服务	高校图书馆 建设 技术 空间 智慧服务 提升 高校智慧图书馆 馆员 资源 图书馆智慧服务 创新 学科服务 功能 推动 重要优化 机制 体系 数字孪生 双一流
8	数据安全	资源 数据 价值 信息 重要 管理 区块链技术 技术 读者 阅读 组织 有效 机制 数字资源 保护 提升 功能 融合 系统 提高
9	热点分析	领域 文献 学科 主题 图书馆学 阅读推广 研究热点 我国 发现 社会 内容 未来 区块链 大数据 数据 研究主题 热点 数据库 中国 情报
10	知识服务	读者 知识 知识服务 需求 用户 信息 图书 用户画像 个性化 提高 满足 平台 推荐 图书馆知识服务 情景 大数据 学习 创新 使用 图书情报领域

表 4　高产学者相关指标统计结果

Table 4　Statistical results of relevant indicators of high-yield scholars

序号	作者	机构名	研究主题	绝对度中心性	度、中介和接近相对中心性(数值 * 0.01)
1	邵波	南京大学图书馆	智能技术	16	0.158,0.018,0.078
2	陆康	南京晓庄学院图书馆	数据安全、个性化服务、构建要素	12	0.215,0.012,0.078
3	杨新涯	重庆大学图书馆	十四五规划、个性化服务、智慧服务	28	0.201,0.049,0.079
4	刘慧	南京晓庄学院图书馆	数据安全、个性化服务、构建要素	10	0.005,0.003,0.078
5	王世伟	上海社会科学院信息研究所	智慧图书馆构建要素、未来发展趋势	0	0.000,0.000,0.000
6	许正兴	南京交通职业技术学院图书馆	智慧服务	2	0.010,0.000,0.077
7	曾子明	武汉大学信息资源研究中心	个性化服务	8	0.062,0.003,0.077
8	杨文建	重庆第二师范学院图书馆	智慧服务	1	0.038,0.000,0.077
9	任贝贝	上海市网络技术综合应用研究所	数据安全、个性化服务、构建要素	6	0.134,0.000,0.078
10	单轸	南京大学信息管理学院	智能技术、个性化服务	3	0.053,0.000,0.078

表 5　高产学者研究兴趣向量

Table 5　Research interest vector of high-yield scholars

序号	主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8	主题 9	主题 10
1	0.069	0.025	0.076	0.379	0.065	0.143	0.083	0.042	0.043	0.074
2	0.104	0.064	0.018	0.097	0.048	0.052	0.091	0.477	0.012	0.037
3	0.070	0.060	0.195	0.067	0.129	0.096	0.065	0.181	0.049	0.089
4	0.110	0.094	0.007	0.098	0.053	0.055	0.119	0.395	0.014	0.053
5	0.057	0.232	0.010	0.035	0.157	0.221	0.013	0.184	0.064	0.026
6	0.015	0.482	0.064	0.006	0.056	0.058	0.069	0.163	0.023	0.064
7	0.455	0.033	0.045	0.012	0.030	0.038	0.089	0.041	0.014	0.243
8	0.176	0.087	0.091	0.039	0.085	0.105	0.401	0.011	0.003	0.003
9	0.087	0.049	0.003	0.166	0.049	0.091	0.083	0.457	0.011	0.003
10	0.127	0.125	0.057	0.100	0.075	0.095	0.113	0.217	0.026	0.066

3.4　学者用户画像

以学者陆康为例,结合前文统计的与其相关的基础信息,刻画出该学者的用户画像,如图 5 所示.

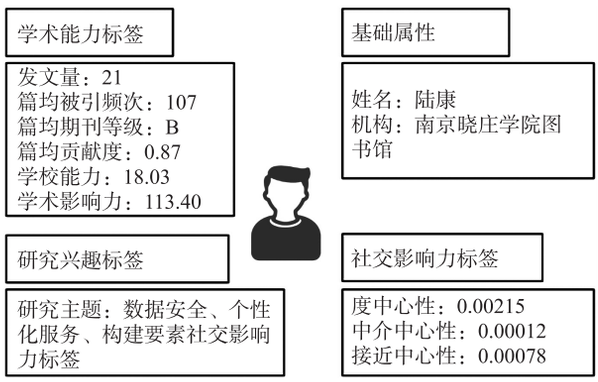

图 5　学者陆康的用户画像

Fig. 5　User Portrait of Scholar Lu Kang

3.5　学者推荐结果

本文从学者的发文题录数据中提取学者的五种特征(学术能力、研究兴趣、中心性、合作关系强度以及 Katz 指标)进行融合推荐评分测量,采用信息检索领域中用于数据融合评分的 CombMNZ 算法作为整合策略[26],计算候选学者的推荐评分并进行排序,以确定待推荐学者的优先级. 由于五种特征的评分数值区间为[0,1],不存在量纲差异,因此未进行标准化处理. CombMNZ 计算公式如下:

$$Score_{CombMNZ}=\sum_{i=1}^{m} w_i \times Score_i \tag{13}$$

式中,m 表示评分项数量(本研究为 5,即五种特征),w_i 表示评分项 i 的权重,关于权重的确定,笔者依据中国矿业大学图书馆 9 位专家学者的问卷评分构造判断矩阵,采用层次分析法(AHP)计算权重并通过一致性检测,赋予的权重值分别为 0.264 70、0.409 00、0.008 16、0.183 80、0.006 08,$Score_i$ 表示候选学者在评分项 i 上得到的评分,对应于候选合作学者与目标学者之间各个特征的相似度.

以学者陆康为例,将其候选学者按照融合推荐评分进行降序排名,结果如表 6 所示. 于兴尚、谢光雷、赵永林等排名后 10 位学者,由于其研究兴趣与目标学者不重合,且融合评分较低,因而不适宜推荐给目标学者. 鉴于此,学者陆康的 TOP-9 推荐合作学者集合为{刘慧、任贝贝、杜京容、张靖、潘小宇、王圣元、杜健、夏莹、孔繁超}. 学者刘慧的研究兴趣涉及“数据安全”“个性化服务”“构建要素”,与目标学者陆康的研究兴趣匹配度高,学术能力较为相当,并且其与目标学者是同事关系,均在南京晓庄学院图书馆就职,经历过 16 次科研合作,其与目标学者进行科研合作的可能性高. 任贝贝与目标学者虽不是同事关系,但经历过 10 次跨机构合作,也较容易与目标学者建立合作关系. 学者孔繁超与目标学者未曾有过历史合作,但融合推荐评分为 0.545,因此其也在推荐合作学者集合中.

表 6　陆康的合作学者推荐列表

Table 6　Recommended list of Lu Kang's collaborating scholars

序号	候选学者	学术能力	研究兴趣	中心性	相关性	Katz 指标	融合推荐评分
1	刘慧	0.986	0.988	0.999	0.801	0.085	0.821
2	任贝贝	0.997	0.984	0.983	0.690	0.055	0.802
3	杜京容	0.981	0.960	0.646	0.309	0.012	0.714
4	张婧	0.979	0.895	0.503	0.436	0.023	0.710
5	潘小宇	0.999	0.884	0.782	0.378	0.017	0.702
6	王圣元	0.864	0.945	0.679	0.378	0.016	0.690
7	杜健	0.915	0.931	0.501	0.218	0.006	0.667
8	夏莹	0.626	0.845	0.646	0.309	0.011	0.573
9	孔繁超	0.849	0.775	0.400	0.000	0.000	0.545
10	于兴尚	0.999	0.379	0.811	0.098	0.005	0.444
11	谢光雷	0.780	0.438	0.501	0.218	0.005	0.430
12	赵永林	0.970	0.334	0.602	0.154	0.005	0.427
13	郭畅	0.993	0.332	0.602	0.000	0.000	0.404
14	陈万明	0.918	0.194	0.501	0.218	0.006	0.367
15	董红丽	0.961	0.238	0.457	0.000	0.000	0.355
16	李利娟	0.982	0.158	0.400	0.000	0.000	0.328
17	万玮琪	0.780	0.235	0.501	0.000	0.000	0.307
18	张翀	0.780	0.235	0.501	0.000	0.000	0.307
19	凌耀银	0.847	0.189	0.457	0.000	0.000	0.305

4 结论

本文以智慧图书馆研究领域为例,从多个维度对科研学者进行全面的用户画像,描述其学术能力、学术影响力、社交影响力等特征,在此基础上,为其推荐适配的合作学者遴选集合,并取得良好的推荐效果. 但同时存在一定的局限性:(1)本文选用的学者的科研成果仅包括期刊论文,题录数据的不完整会导致学者的研究兴趣和合作关系信息的遗漏;(2)研究兴趣具有一种层级结构关系,本文采用粒度略粗的研究主题作为学者的研究兴趣,未能细化到以关键词表征研究方向;(3)未考虑时间上下文因素,学者兴趣会随着时间发生动态变化,容易造成推荐过去兴趣相似的"过时"学者. 同时,学者之间的合作密切程度会随着时间有所淡化,容易造成推荐过去关系密切的"过时"学者. 未来笔者将从以下几个方面进行改进优化:(1)融合多源数据,引入专利、会议、专著等多种类型科研成果,提取增补的学者研究兴趣和合作关系,同时融合基金项目、学术社交网络等合作关系和朋友关系,从而扩大合作学者的候选范围. (2)细化学者的研究兴趣粒度,增加研究方向细粒度兴趣特征,从而满足分级推荐合作学者的需求. (3)引入时间衰退因子,对学者的研究兴趣和合作关系加权,使推荐的学者在研究兴趣和合作关系方面都较为适配.

[参考文献]

[1] 李杰,陈超美. Citespace:科技文本挖掘及可视化[M]. 北京:首都经济贸易大学出版社,2016.

[2] 李春英,汤庸,肖政宏,等. 学术社交网络中的权威学者推荐模型[J]. 计算机应用,2020,40(9):2594-2599.

[3] 熊回香,李晓敏,杜瑾. 基于学术关键词与共被引的学者推荐研究[J]. 情报学报,2021,40(7):725-733.

[4] CHAKRABORTY J,THOPUGUNTA G,BANSAL S. Data extraction and integration for scholar recommendation system[C]// IEEE International Conference on Semantic Computing,Los Alamitos,CA,2018:397-402.

[5] 刘萍,郑凯伦,邹德安. 基于 LDA 模型的科研合作推荐研究[J]. 情报理论与实践,2015,38(9):79-85.

[6] 熊回香,杨雪萍,蒋武轩,等. 基于学术能力及合作关系网络的学者推荐研究[J]. 情报科学,2019,37(5):71-78.

[7] 熊回香,顾佳云,代沁泉,等. 基于用户相似度与信任度的虚拟学术社区中学者推荐研究[J]. 情报科学,2022,40(2):74-81.

[8] 董文慧,熊回香,杜瑾,等. 基于学者画像的科研合作者推荐研究[J]. 数据分析与知识发现,2022,6(10):20-34.

[9] 杨梦婷,熊回香,肖兵,等. 基于动态特征的学者推荐研究[J]. 情报理论与实践,2022,45(4):120-127.

[10] 王妞妞,熊回香,刘梦豪,等. 基于多维决策属性的科研合作者推荐研究[J]. 情报科学,2022,40(7):93-101.

[11] JIN H Y,ZHANG P C,DONG H,et al. Personalized scholar recommendation based on multi-demensional features[J]. Applied sciences,2021,11(18):8664.

[12] YUAN C,HE Y,LIN R,et al. Graph embedding for scholar recommendation in academic social networks[J]. Frountiers in physics,2021,768006.

[13] 秦红武,赵猛,马秀琴,等. 基于学术水平聚类的科研合作者推荐模型[J]. 计算机工程与应用,2022,58(21):172-181.

[14] 汪俊,岳峰,王刚,等. 科研社交网络中基于链接预测的专家推荐研究[J]. 情报杂志,2015,34(6):151-157.

[15] 刘云枫,孙平,葛志远. 基于网络表示学习的作者合作推荐模型[J]. 情报科学,2020,38(2):75-80.

[16] 张金柱,于文倩,刘菁婕,等. 基于网络表示学习的科研合作预测研究[J]. 情报学报,2018,37(2):132-139.

[17] 褚叶祺,丁佳骏. 基于 Louvain 算法的作者合著网络社区划分研究[J]. 高技术通讯,2021,31(3):257-262.

[18] 武森,卢丹,冯小东,等. 基于大规模复杂网络社区发现的科研合著网络分析[J]. 中国科技论文,2014,9(4):474-478.

[19] 邱均平. 文献计量学(第二版)[M]. 北京:科学出版社,2019.

[20] 熊回香,叶佳鑫,丁玲,等. 基于改进的 h 指数的学者评价研究[J]. 情报学报,2019,38(10):1022-1029.

[21] 曲靖野,陈震,胡铁楠. 共词分析与 LDA 模型分析在文本主题挖掘中的比较研究[J]. 信息资源管理学报,2018,36(2):18-23.

[22] 赵凯,王鸿源. LDA 最优主题数选取方法研究:以 CNKI 文献为例[J]. 管理决策,2020,39(16):175-179.

[23] 姜鑫. 社会网络分析方法在图书情报领域的应用研究[M]. 北京:知识产权出版社,2015.

[24] 李纲,李岚凤,毛进,等. 作者合著网络中研究兴趣相似性实证研究[J].图书情报工作,2015,59(2):75-81.

[25] 刘军. 整体网分析讲义——UCINET 软件实用指南[M]. 上海:上海人民出版社,2009.

[26] 熊回香,杨雪萍,蒋武轩,等. 科研社交网站中基于相似性趣的学者推荐研究[J]. 情报科学,2017,35(9):3-11.
[27] 余传明,龚雨田,赵晓莉,等. 基于多特征融合的金融领域科研合作推荐研究[J]. 数据分析与知识发现,2017,1(8):39-47.
[28] 张斌,马费成. 科学知识网络中的链路预测研究综述[J]. 中国图书馆学报,2015,41(3):99-113.
[29] 张斌,李亚婷,戴怡清. 学科合作网络的链路挖掘与应用分析[J]. 情报理论与实践,2018,41(9):108-113.
[30] 高广尚. 用户画像构建方法研究综述[J]. 数据分析与知识发现,2019,3(3):25-35.
[31] 刘海鸥,孙晶晶,苏妍嫄,等. 国内外用户画像研究[J]. 情报理论与实践,2018,41(11):155-160.

[责任编辑:杜忆忱]