基于用户画像与关联规则的图书馆资源组合推荐算法

李蒙蒙
（温州商学院，浙江温州 325000）

摘要：图书馆资源推荐算法难以与用户的偏好相结合，导致其推荐精度较差。为提高推荐结果的准确性，基于用户画像与关联规则设计图书馆资源组合推荐算法。通过用户画像算法得到图书资源兴趣估计，在特征样本集合的基础上，计算主题权重的标准值，建立用户情景兴趣度表达式。基于关联规则建立图书馆资源聚合模型，计算文档内相同词汇出现的频率，计算不同书籍的相似度，并在书籍指标权重的基础上，得到关联规则下图书馆资源的聚合函数，以此设计资源组合推荐算法。实验结果表明，其最高精确率、召回率、*F*1 值分别为 0.92、0.73 和 0.69，该推荐算法的推荐精度较高。

关键词：用户画像；关联规则；图书馆资源；资源组合推荐算法；词汇相似度
DOI：10.3772/j.issn.1674-1544.2023.02.012
CSTR：15994.14.issn.1674-1544.2023.02.012
中图分类号：TN06
文献标识码：A

Library Resource Combination Recommendation Algorithm Based on User Portrait and Association Rules

LI Mengmeng
(Wenzhou Business School, Wenzhou 325000)

Abstract: Library resource recommendation algorithm is difficult to combine with users' preferences, resulting in poor recommendation accuracy. In order to improve the accuracy of recommendation results, library resource combination recommendation algorithm is designed based on user portrait and association rules. The interest estimation of book resources is obtained through the user portrait algorithm. Based on the feature sample set, the standard value of subject weight is calculated, and the expression of user scenario interest is established. The library resource aggregation model is established based on association rules, the frequency of the same vocabulary in the document, and the similarity of different books are calculated. Based on the book index weight, the aggregation function of library resources under association rules is obtained, so as to design the resource combination recommendation algorithm. The experimental results show that its highest accuracy, recall, and *F*1 values are 0.92, 0.73, and 0.69, respectively, indicating that the recommendation accuracy of this recommendation algorithm is superior to other algorithms.

Keywords: user portrait, association rules, library resources, resource combination recommendation algorithm, lexical similarity

作者简介：李蒙蒙（1993—），女，大学本科，温州商学院助理实验师，研究方向为资源管理。
收稿时间：2022 年 9 月 9 日。

0 引言

为保证学生可以随时随地访问图书馆中的数字教学资源，学校将图书馆中的书目全部存放在了计算机网络中[1-2]，并设置了一种用于自动推荐书目的算法。这种算法可以在一定程度上帮助学生更精确地寻找适合自己的数字资源，提高针对用户的服务质量，提高学生的自学效率。在现有的相关研究中，李宇佳等[3]以提高学术信息的精准性与专业性为根本目标，进一步构建了学术媒体的用户画像，并在新媒体技术的基础上，分析了用户需求的分层表达形式，通过理论框架建立了一个信息推荐的模型。于非[4]建立了一个以增强用户感知价值的逻辑框架，构建了一个资源推荐的生态服务圈，并在核心的情景要素中，以关联机理为核心，合理有序地保证了高校图书馆资源的动态平衡，系统地识别了高校数字图书馆的资源推荐序列。叶颖[5]以空间、资源、服务作为组合推荐的核心，大范围地构建了多元异构的数据，为之提供了新的分析角度，并逐步拓宽了资源服务的范围，使之成为更重要的拼图，在推荐结果中，这种以多元数据融合为核心的推荐方法可以针对更多种类的信息逐步推荐，满足更宽广的推荐需求。

综合以往研究方法，本文将设计一种基于用户画像与关联规则的图书馆资源组合推荐算法，进一步提高其推荐的精度。

1 图书馆资源组合推荐算法

1.1 基于用户画像的图书资源兴趣估计

本文以用户画像[6-7]为核心，在建立图书馆资源推荐指标的过程中，需要建立用户兴趣模型[8-9]，并描述用户的需求信息与其在图书馆中的兴趣度。在这种情景环境下，已经存在的用户向量空间中通常都会以n维向量的形式存在，可建立以下特征样本集合[10]：

$$(K,M)=\{(k_1,m_1),\cdots,(k_n,m_m)\} \quad (1)$$

式中，(K,M)表示用户模型中特征向量的权重；(k_n,m_m)则是其在n维空间中的表现形式。当情景元组中的兴趣模型通过特征集合来表示时，可以得到：

$$H_m=\{(f_1,p_1,n_1,d_1),\cdots,(f_n,p_n,n_n,d_n)\} \quad (2)$$

式中，H_m表示用户主体的兴趣特征集合；f_n表示n维兴趣主题的特征向量；p_n表示n维向量空间中的主题权重；n_n表示n维空间的情景要素；d_n表示主题文档中的实际特征项。此时可以通过初始化的方式，获取主题权重的标准值：

$$g_i=\sum_{i=1}^{n}I(page_i) \quad (3)$$

式中，g_i表示主题权重标准值；$I(page_i)$表示用户对该主题文档的兴趣度，主题文档一般通过*page*来表示。在估计多维用户模型的同时，还可以建立用户模型：

$$Q(h,m,n)=\frac{f_{user}}{\sum_{i=1}^{n}f_{context_i}} \quad (4)$$

式中，$Q(h,m,n)$表示用户模型在相同文化水平因素、时间因素、地点因素下的信息评分；f_{user}表示用户的兴趣度；$f_{context_i}$表示该主题文档所能带来的兴趣标准值。这时，可以建立用户情景兴趣度的表达式：

$$Q_{context}=\frac{\prod_{i=1}^{n}(context_h\times context_m\times context_n)}{Q(h,m,n)} \quad (5)$$

式中，$Q_{context}$表示用户兴趣度矩阵的评分；$context_h$、$context_m$和$context_n$分别表示文化水平、时间、地点3类因素的评分标准。此时就可以得到用户画像的图书资源兴趣估计，并基于该用户信息建立图书资源在用户处的信息集合。

1.2 基于关联规则建立图书馆资源聚合模型

在关联规则中，通常需要将海量数据以某种形式挖掘出来，并建立一定的联系，作为规则的集合。这可以通过图书馆资源中某词汇的权重值来衡量[19]。当词汇权重值不断增大的同时，可以得到一个文档中出现的两个同样词汇的出现频率为：

$$w_{ij}=\frac{t_i\times D_i}{\sqrt{t_j}} \quad (6)$$

式中，w_{ij}表示某词汇在图书资源集合i与集合j中的关联频率；t_i和t_j分别表示该词汇在图书资源数据库i和数据库j中出现的次数；D_i表示食物数据库对该频率的支持度。在所有孤立点中，需要寻找关联规则的隐藏规律，此时通过支持度与映射关系将子节点连接起来，就能够得到事务聚合的结果。此时可以依据图书模型得到两个书籍之间的距离：

$$X_{AB}=\sqrt{\frac{d_A-d_F}{d_B-d_F}} \tag{7}$$

式中，X_{AB}表示书籍A与书籍B之间的欧式距离；d_A表示书籍A的虚拟节点深度；d_B表示书籍B的虚拟节点深度；d_F表示分类节点x在中图法分类中的深度。将其作为共同祖先，用于图书模型的构建时，可以得到书籍A与书籍B之间相似度的计算：

$$\sin(A,B)=\frac{\sum_{i=1}^{n}\left(h_{ap}\times h_{bp}\right)}{\sqrt{\sum_{i=1}^{n}h^2{}_{ap}}\times\sqrt{\sum_{i=1}^{n}h^2{}_{bp}}} \tag{8}$$

式中，$\sin(A,B)$表示书籍A与书籍B之间的相似度指标；h_{ap}和h_{bp}分别表示两种书籍的特征参数。此时可以通过标准化处理的方式，建立不同类型书籍之间的指标权重：

$$A_{ij}'=\frac{\left(\frac{a_{ij}-a_{\min}}{a_{\max}-a_{\min}}\right)\times\frac{\alpha_{ij}}{\alpha_{ij}'}}{\left(\frac{a_{\max}-a_{ij}}{a_{\max}-a_{\min}}\right)} \tag{9}$$

式中，A_{ij}'表示经过标准化处理后某书籍的指标权重；a_{ij}表示书籍在该指标中所占的比重；$a_{\max}$和$a_{\min}$分别表示指标的最大和最小信息熵；α_{ij}表示调节系数，α_{ij}'则表示调节系数的预测值。基于该指标权重的参数，可以直接得到关联规则下图书馆资源的聚合函数：

$$f(x)=-\log\left[\frac{freq(h_c)}{N_m}\right] \tag{10}$$

式中，$f(x)$表示图书馆资源的聚合函数；h_c表示所有概念分支的谜底信息；N_m表示同一概念树在相同语义上的系数。结合该函数，可以建立一个图书馆资源的聚合模型。

1.3　设计资源组合推荐算法

结合实际的书籍历史借阅记录，以及在关联规则下的图书馆资源聚合模型，可以得到书籍推荐的算法流程，见图1。

如图1所示，可以通过获取个人基础信息的方式来启动资源组合推荐算法，采用用户喜爱的历史标签来决定用户的兴趣偏好。此时可以得到矩阵分解模型的损失函数：

$$d(x)=\sum_{i=1}^{n}\left(a_{ij}-k_{ij}\right)^2 \tag{11}$$

式中，$d(x)$表示矩阵分解模型损失函数；a_{ij}表示模型预测参数；k_{ij}表示预测结果的评分值。通过损失函数，可以判断书籍预测的结果，若预测成功，则可以直接得到书籍名单，即得到图书馆资源的组合推荐算法。

2　实验研究

2.1　实验数据集选取与环境设置

本文实验的数据来源于某大学的图书馆，包括其系统内的书目名称、作者、出版社、所属类别等信息。依据现有的书目数据对图书馆中的资源文本进行分析与处理，并随机抽取其中的360篇书目报告作为模型的初始项，得到了6种主体类型的文档模型。其实例归类如图2所示。

从图2可以看到，构建用户画像文档结构，实现资源的实时推荐，并收集整理数据信息，从5000名用户的借阅数据中，分析关键情景要素，包括用户的借阅记录、评分、书籍的特征参数。书籍的特征参数计算可以通过对书籍信息进行分类、关键字提取等方式得到，以此提取出用户兴趣模型并进行解析，每个用户对于不同主题的兴趣程度由0到1的浮点数表示。每本书都被描述为由不同主题构成的向量，每个维度表示该主题在书籍中出现的频率，向量长度为100。采用公式（8）计算书籍A和B之间的相似度，统计该主题在书籍中出现的次数，并计算每个主题在书籍中出现的频率，同时分析书籍A和B在主题上

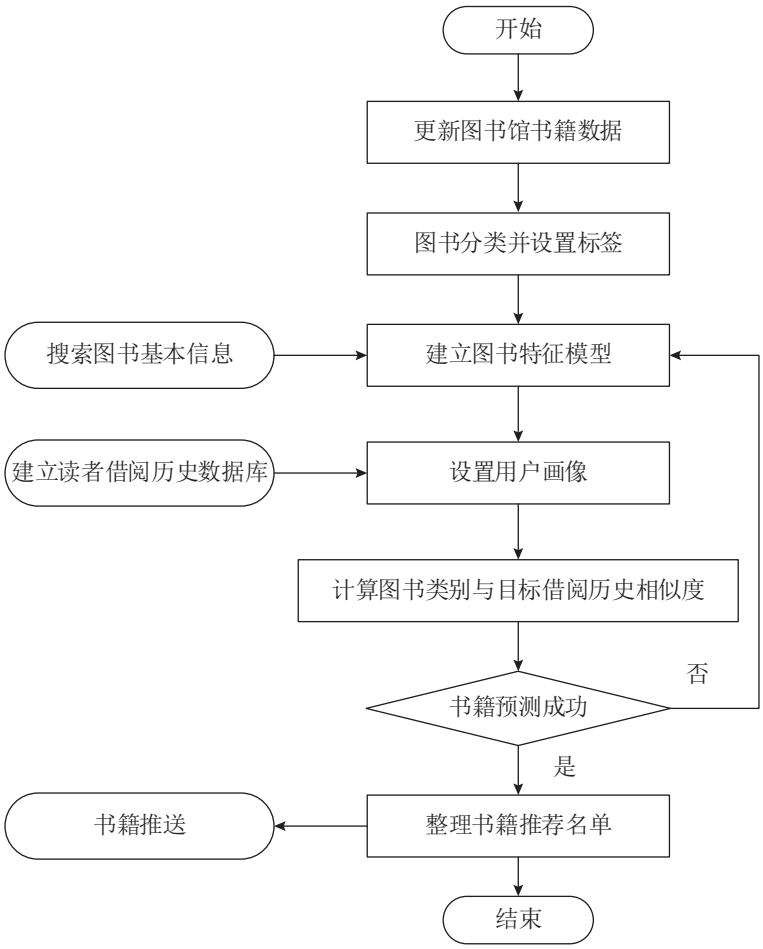

图 1　算法流程

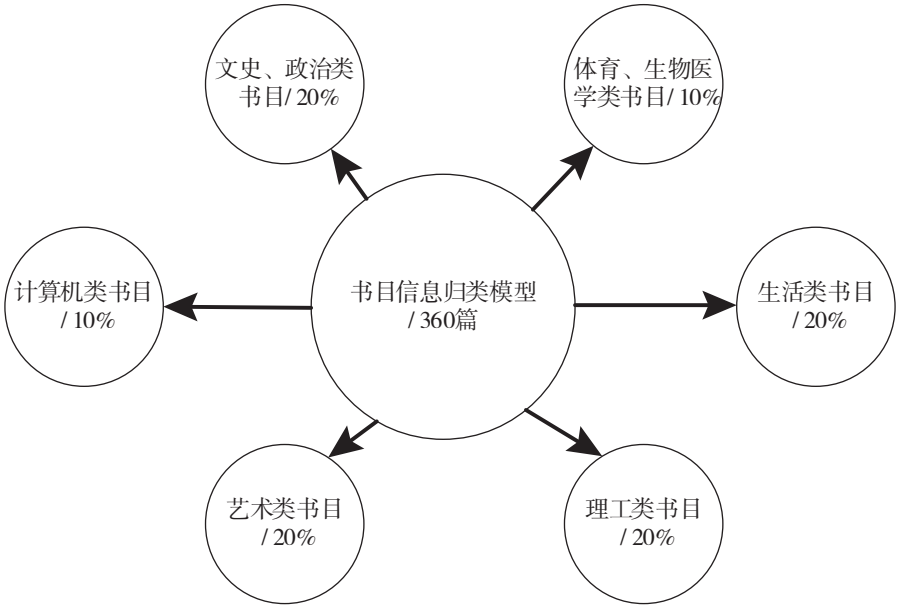

图 2　书目归类模型

的频率值，进行双向设置，即：

$cosine_{AB} = \frac{\sum\limits_{i=1}^{n} a_i b_i}{\sqrt{\sum\limits_{i=1}^{n} a_i^2} \sqrt{\sum\limits_{i=1}^{n} b_i^2}}$

根据书籍中各自所含的不同主题频率值，将每个主题的频率值作为向量的一个维度，构建出一个100维的主题向量作为书籍的特征参数。抽取向量节点，进行数据的预处理。

2.2　模型训练

在训练资源推荐模型的过程中，存在两个影响推荐效果的参数，分别为迭代次数、超参数。这两个参数的变化，会直接导致训练模型的结果发生变化，模型的好坏可以通过均方根误差直接体现出来，该系数的计算公式为：

$$M_{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}\left|f_i-\hat{f}_i\right|}{|H_n|}} \tag{12}$$

式中，M_{RMSE}表示模型训练的均方根误差，M_{RMSE}越小，表明模型训练的效果越好；f_i表示模型的实际评分结果，$\hat{f}_i$表示预测评分结果；H_n表示评分总数量。结合该公式，可以在大量的图书馆资源中，获取科研资源的数据。因此，在本文实验模型的训练中，分别设置迭代次数为50、100、150、200、250、300，设置超参数分别为0.01、0.05、0.1、0.5、1、5，并分别对其均方根误差进行测试，得到的结果如图3所示。

从图3可以看到，随着迭代次数的逐渐增加，均方根误差会逐渐减小，直至迭代次数达到200次后，均方根误差达到最小值，为0.94。在超参数由小变大的过程中，均方根误差呈现出先减小后增加的趋势，当超参数为0.1时，均方根误差达到最小值，为0.90。因此，在本文实验中，可以选择迭代次数200、超参数0.1为模型的参数。

在实验的预处理阶段，聚类个数的大小K值也能够改变数据集本身的推荐效果，因此可以分析不同K值变化对数据集推荐精度的影响。设定K值为0～60，得到的均方根误差如图4所示。

从图4可看到，在聚类数量由0变化为60的过程中，均方根误差呈现出不规则的“U”型，在聚类数量为40时，达到均方根误差的最小值，约为0.92。由此可见，在实验中，可以将数据集的聚类数量设置为40，以提高算法训练的效率。

2.3　算法性能评价

本文设计了基于用户画像与关联规则的图书馆资源组合推荐算法。为测试该算法的有效性与优越性，可以将其与现有的几种算法对比。其评价指标的计算公式为：

$$\begin{cases} P_{precision}=\dfrac{\sum_{u\in U}|K_u\cap T_u|}{\sum_{u\in U}|K_u|} \\ P_{recall}=\dfrac{\sum_{u\in U}|K_u\cap T_u|}{\sum_{u\in U}|T_u|} \\ P_{F_1}=2\times\dfrac{P_{precision}\times P_{recall}}{P_{precision}+P_{recall}} \end{cases} \tag{13}$$

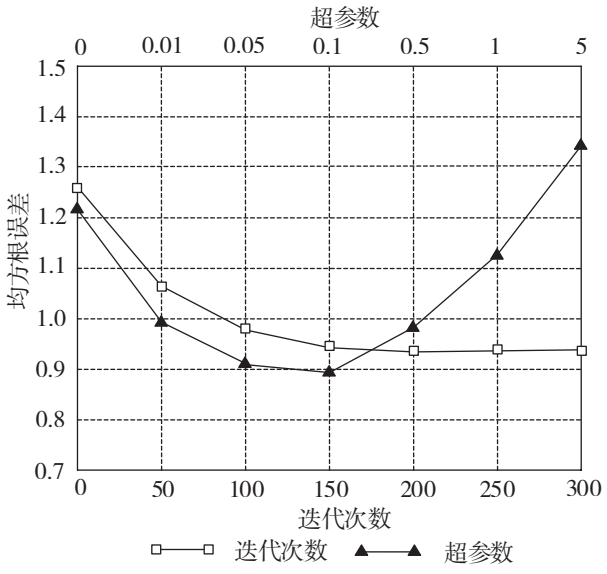

图3　模型参数选取

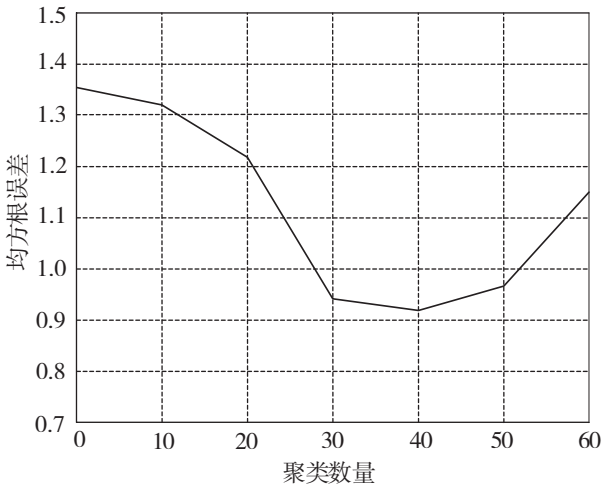

图4　聚类数量对均方根误差的影响

式中，$P_{precision}$表示资源推荐的精确率，P_{recall}表示资源推荐的召回率，精确率与召回率越高，证明推荐算法的效果越好；P_{F_1}是对精确率与召回率的算法平衡指标；K_u表示推荐系统中用户所收藏的书目信息；T_u表示被用户着重标记的书目信息。在该推荐指标的影响下，分别测试不同数值图书推荐数量下推荐精度的效果，如图 5 所示。

由图 5 数据可知，随着推荐数量的增加，3 种推荐精度均呈现出先增加后减小的趋势。其中精确率的推荐精度指标数值在同等推荐数量下为最大值，当推荐数量为 20 时到达最高点。在召回率指标中，推荐数量为 15 时精度指标达到最大值，在*F*1 值的计算中，推荐数量为 15 时到达最大值。由此可见，可以设置图书推荐列表的数量为 15，作为图书馆资源组合推荐的最优选项。在该参数下，可以得到不同算法间性能的对比结果，如图 6 所示。

从图 6 可以看到，在本文设计的用户画像与关联规则推荐算法中，精确率可以达到 0.92，召回率为 0.73，*F*1 值最大为 0.69。与这个推荐算法相比，用户动态画像、情景要素适配、多源数据融合 3 种推荐算法的精度指标均小于本文算法，其中用户动态画像算法大于多源数据融合算法，情景要素匹配算法的精度最差。

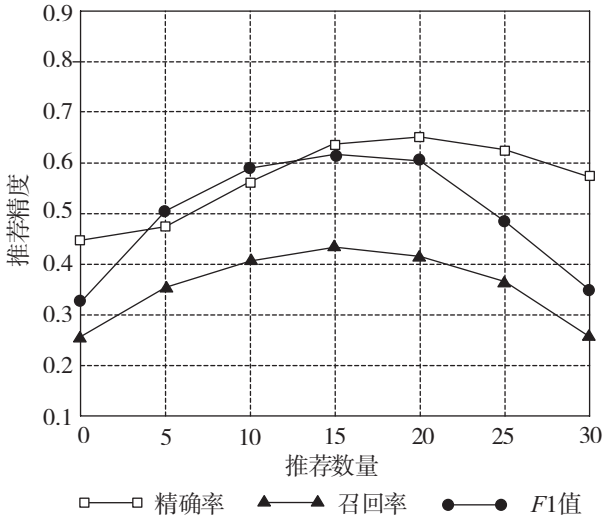

图 5　不同图书推荐数量对推荐精度的影响

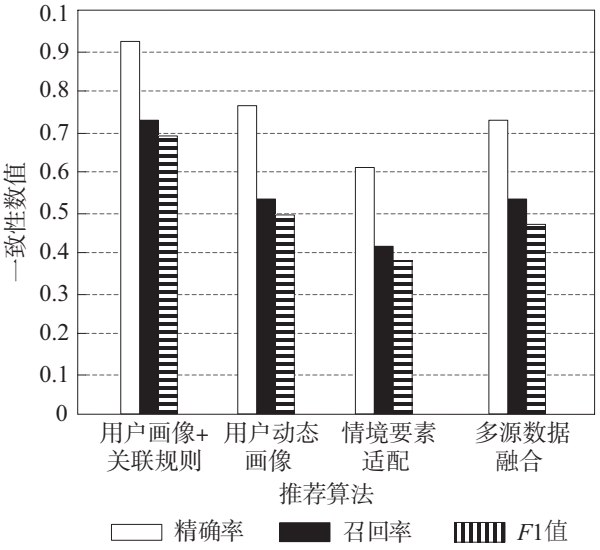

图 6　不同算法推荐结果对比结果

3　结语

本文基于用户画像与关联规则，设计了一种图书馆资源组合推荐的算法。经实验证明：①当推荐数量为 20 时精确率到达最高点，推荐数量为 15 时召回率达到最大值，推荐数量为 15 时*F*1 值达到最大值。②精确率为 0.92，召回率为 0.73，*F*1 值最大为 0.69。

本文研究认为，在保证书籍与历史借阅项目具备较好的相似度的同时，还要提高书籍借阅推荐的精确度。在未来的研究中，可以进一步细化学生群体的阅读需求，结合学生的专业与课程进行组合推荐，并建立学生反馈机制，让读者反馈推荐结果，优化推荐质量。

参考文献

[1] 张炜，敦文杰.国家图书馆影音视听资源智慧化服务的实践与思考[J].图书馆，2022(7): 37−43.

[2] 刘宇航. 基于微服务的智慧图书馆信息系统的设计[J]. 微型电脑应用，2022(8): 201−204.

[3] 李宇佳，王益成.基于用户动态画像的学术新媒体信息精准推荐模型研究[J].情报科学，2022, 40(1): 88−93, 101.

[4] 于非.基于情境要素适配的高校数字图书馆资源情境化推荐路径研究[J].图书馆工作与研究，2021(6): 74−

81.
[5] 叶颖.多源数据融合视角下的智慧图书馆个性化推荐方法[J].图书馆论坛, 2022, 42(3): 154−161.
[6] 张海涛, 栾宇, 周红磊.用户画像: 向知识迈进[J].图书情报知识, 2020(5): 131−134.
[7] 刘海鸥, 李凯, 何旭涛, 等.面向信息茧房的用户画像多样化标签推荐[J].图书馆, 2022(3): 83−89.
[8] 刘森晶, 马雪梅.基于用户兴趣模型的数字多媒体信息智能推送方法[J].自动化技术与应用, 2021, 40(8): 52−56.
[9] 张彬, 徐建民, 吴姣.大数据环境下基于知识图谱的用户兴趣扩展模型研究[J].现代情报, 2021, 41(8): 36−44.
[10] 史蕴豪, 许华, 郑万泽, 等.基于集成学习与特征降维的小样本调制识别方法[J].系统工程与电子技术, 2021, 43(4): 1099−1109.

（上接第74页）

将通用的科普驱动力指标体系应用于研究特定地方的问题也存在一定的局限性，如明光市作为一个县级市，不具备图书报刊、影像制品等传播媒介的生产与运营能力，而当地在科普志愿服务活动等方面的突出优势却难以通过当前的指标体系充分体现。为了减少这些因素带来的影响，本文在分析明光市科普驱动力情况的基础上，注重将宏观状况与地区特点相结合，深入展开探讨与分析。今后可进一步拓展多源数据进行地区科普的动力研究，并通过对多类型区域、多层级地区的比较分析提出更为丰富的结论与建议。

参考文献

[1] 张锋, 杜发春, 石顺科. 边境民族地区科普的本土化思考(1) [J]. 科技导报, 2012, 30(14): 81.
[2] 王贵彦, 陈曦, 张永升, 等. 中国农村不同地区科普模式比较分析[J]. 中国农学通报, 2010, 26(15): 432−436.
[3] 朱洪启.新时代农村科普初探[J].科技传播, 2019, 11(12): 137−138.
[4] 王德海, 刘玉花, 高翠玲, 等. 国外农村科普运作模式经验简介[J]. 世界农业, 2006(8): 8−11.
[5] 于洁, 佟贺丰, 黄东流, 等. 基于三阶段DEA的我国区域科普投入产出效率研究[J]. 科技管理研究, 2018, 38(6): 40−47.
[6] 陈套, 罗晓乐. 我国区域科普能力测度及其与科技竞争力匹配度研究[J]. 科普研究, 2015, 10(5): 31−37.
[7] 任嵘嵘, 郑念, 赵萌. 我国地区科普能力评价: 基于熵权法—GEM[J]. 技术经济, 2013, 32(2): 59−64.
[8] RAZA G, SINGH S, KUMAR P V S. Public understanding of science: glimpses of the past and roads ahead [M]//SCHIELE B, CLAESSENS M, SHI S. Science communication in the world: practices, theories and trends. Dordrecht: Springer Netherlands, 2012: 139−150.
[9] CEES M K. An example of a science communication evaluation study: Discovery07, a Dutch science part[J]. Journal of science communication, 2008, 29(4): 1−8.
[10] BERTALANFFY L V, VON B, RAPOPORT A. General systems theory[J]. Taehan kanho the Korean nurse, 1989, 28(3): 36.
[11] THELEN E, SMITH L. Dynamic systems theories[M]// DAMON W,LERNER R M. Handbook of child psychology: theoretical models of human development. New York: John Wiley & Sons, Inc, 1998: 563−634.
[12] 赫尔. 行为的原理: 行为理论导论[M]. 北京: 中国传媒大学出版社, 2016: 16−31.
[13] 徐建平. 组织行为学[M]. 北京: 中国人民大学出版社, 2014: 113−115.
[14] 赵建平. 玉龙雪山景区旅游发展驱动力变化研究[D]: 昆明: 云南师范大学, 2021.
[15] 卓丽洪, 李群, 王宾, 等. 中国地区科普驱动力指标体系构建与评价[J]. 中国科技论坛, 2016 (8): 95−101.
[16] 陈昭锋. 我国区域科普能力建设的趋势[J]. 科技与经济, 2007 (2): 53−56.
[17] 潇湘晨报.最佳志愿服务项目“抱团科普”志愿服务项目事迹材料[EB/OL].(2021−11−25)[2022−02−10]. https://baijiahao.baidu.com/s?id=1717407106947735264&wfr=spider&for=pc.
[18] 明光市三界镇. 三界镇科协: 小板凳科普接地气、近群众[EB/OL]. [2021−12−22]. https://www.mingguang.gov.cn/zwdt/xwlb/xzzc/278301535.html.
[19] 央视新闻客户端. 科学技术部发布2020年度全国科普统计数据 [EB/OL]. [2021−11−27]. http://www.chinanews.com.cn/gn/2021/11−23/9614406.shtml.
[20] 彭华. 旅游发展驱动机制及动力模型探析[J]. 旅游学刊, 1999 (6): 39−44.