DOI:10.12154/j.qbzlgz.2022.06.002

基于多元数据融合的科学文献主题识别研究*

邱均平[1,2,3]　孙月瑞[2]　周贞云[1,2,3]
([1]杭州电子科技大学中国科教评价研究院　浙江　310018;
[2]杭州电子科技大学管理学院　浙江　310018;
[3]杭州电子科技大学数据科学与信息计量研究院　浙江　310018)

摘　要: **[目的/意义]**科学文献的主题识别研究是科研管理的重要内容之一,如何全面把握文献的多元数据、提升自动文献主题识别的效果是一个值得研究的问题。**[方法/过程]**文献的关键词、摘要是判断文献主题的重要依据,文章提出基于文献多元数据融合的主题识别模型,使用Word2vec模型、AP聚类及Node2vec模型表示出关键词层的主题向量,使用LDA模型表示出摘要层的主题向量,通过多视图聚类中的SGF方法进行数据融合并识别文献主题。**[结果/结论]**以不同规模的文献集为例,通过主题识别研究,验证该模型识别效果的准确性和可解释性优于典型LDA方法、Doc-LDA模型。

关键词: 科学文献　主题识别　数据融合　多视图聚类　多元数据

Research on the Topic Identification of Scientific Literature Based on Multivariate Data Fusion

Qiu Junping[1,2,3]　Sun Yuerui[2]　Zhou Zhenyun[1,2,3]
([1]Chinese Academy of Science and Education Evaluation, Hangzhou Dianzi University, Zhejiang, 310018;
[2]School of Management, Hangzhou Dianzi University, Zhejiang, 310018;
[3]Academy of Data Science and Informatics, Hangzhou Dianzi University, Zhejiang, 310018)

Abstract: **[Purpose/significance]** The research on topic identification of scientific literature is one of the important contents of scientific research management. How to comprehensively grasp the multivariate data of literature and effectively improve the accuracy of automatic literature topic identification is a problem worthy of research. **[Method/process]** Keywords and abstracts of documents are important basis for judging document topics. This paper proposes a topic identification model based on multi-data fusion of documents. Word2vec model, AP clustering and Node2vec model are used to represent the topic vector of the keyword layer. The topic vector of the abstract layer is represented by the LDA model, and the SGF method in the multi-view clustering method is used to perform data fusion and extract document topics. **[Result/conclusion]** Taking document sets of different scales as an example, through topic identification research, it is verified that the accuracy and interpretability of the recognition effect of the model are better than the typical LDA method and the Doc-LDA model.

Keywords: scientific literature　topic identification　data fusion　multi-view clustering　multivariate data

*本文系2019年国家社会科学基金重大项目"基于大数据的科教评价信息云平台构建和智能服务研究"(项目编号:19ZDA348)和2020年浙江省软科学研究计划重点项目"创新强省背景下浙江高校科技创新竞争力评价及提升研究"(项目编号:2020C25027)的研究成果之一。

1 引言

目前常用的文献主题识别方法是从文献的关键词和摘要这两个元数据入手。文献中总结出的关键词能准确地反映出文献的核心特征，是文献内容的高度提炼，直观地展现出作者的研究主题，因此关于学科领域主题识别的文献研究常常会通过文献资料里的关键词数据来讨论学科主题的变化情况。具体地，其多数从词义与词频的变化、词语共现的转变及新词的出现这些角度对文献进行研究。Bhattacharya等[1]以凝聚态物理研究领域的文献为例，用文章题目中的词语作为关键词，并以其为基础构建了共现网络，对网络使用聚类分析，进而有效识别出学科领域的研究前沿。Li[2]在论文中使用了突发词概念，对关键词与突发词进行了关联规则挖掘。霍朝光等[3]以国际深度学习领域文献为例，通过PageRank算法对研究热点进行比较、排序，并在关键词网络中使用Node2vec算法和t-SNE算法来发现集群，并进而揭示了学科领域的脉络情况。关于文献摘要这一信息单元，它是对文献内容的简要概括，清晰明了地介绍了研究工作的整体思路。相较于基于关键词的科学主题识别方法，基于摘要的主题识别研究多数利用主题概率模型对文章进行主题识别。主题模型的前身可以回溯到潜在语义索引算法（Latent Semantic Indexing，LSI）[4]，该算法认为文档内的词汇之间隐藏着语义结构。后来许多学者在此基础上不断提出新的模型，其中最著名的是Blei等在2003年提出的隐含狄利克雷分布模型（Latent Dirichlet Allocation，LDA）[5]。该模型自提出就获得了广泛应用，目前该模型及改进模型在科学主题研究工作中仍为常见。侯捷[6]在我国管理科学类基金项目的关键词及摘要文本中应用LDA模型来对文献主题进行识别，并将获得的主题对国家政策和基金项目做了关联分析。谭春辉等[7]利用LDA主题模型对国内外数据挖掘领域文献的研究主题进行抽取，并结合时间片构建主题的演化路径，从理论和应用两个角度来分析热点主题的变化情况。

目前多数的学科领域文献的主题识别研究是对文献题录中的单要素进行区别分析而后汇总，缺乏对要素间关联性的考虑，难以全面利用文献的多元数据，其存在由于文献各单元关系研究不充分而产生判断失误等问题，降低了文献主题识别的准确性。当前对多个文献对象内部及对象之间的多种关系进行融合分析的研究较少。许海云等[8]利用科学文献中引文、主题词和作者之间的共现关系，通过PathSelClus算法来发现文献之间在研究主题上的关联强度并进行聚类，其效果好于单一的共词聚类，但由于PathSelClus算法十分依赖专家的知识及其判断，该方法未能实现很好的自动主题识别效果。在已有的其他关系融合方法里，单纯的线性融合策略占多数，并没有产生合理的面向学科领域文献主题识别的多元数据融合方法[9]。

随着文献数据源的不断丰富、复杂网络及机器学习研究的深入发展，关于同时利用关键词和摘要这两个信息单元进行科学主题发现的研究有待继续深化。本文尝试构建合理且可实现的多元数据特征抽取及特征关系融合的主题识别模型，其重点是将文献中的关键词及摘要两层文本数据分别进行主题空间的向量化表示，利用能挖掘出特征间关联的多视图聚类方法进行数据融合并获取文献的主题类别。为了验证模型的有效性，以不同规模的文献集为例进行自动主题识别验证研究。

2 模型框架

使用计算机技术对多维度观测、收集到的信息资源进行加工处理并综合分析，能做出比利用单一信息资源进行研究时更准确、可靠的决策方案。本文提出的基于文献多元数据融合的主题识别模型致力于对文献实体的多元数据进行加工和综合评价，在主题判断上利用起多种信息源之间的关联性，避免受到单一信息源中噪声数据的干扰。该模型在构建上可分为两部分，分别是以文本挖掘方法为核心的元数据主题向量表示和通过多视图聚类达到多元数据融合的文本主题发现技术，如图1所示。

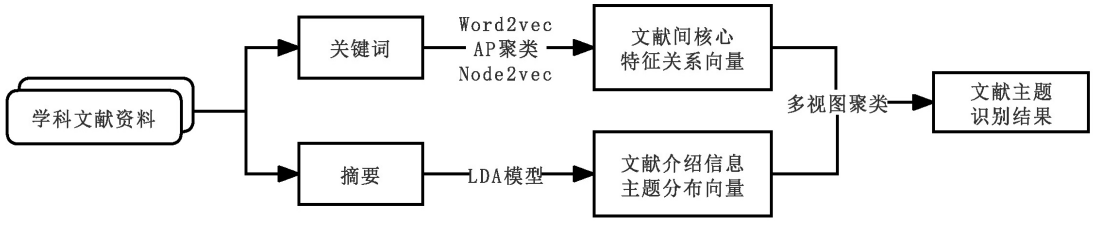

图1 模型框架图

2.1 多元数据主题向量表示

2.1.1 关键词层的向量表示

文献中人为归纳出的关键词集是文献核心内容的重要体现。目前对关键词集分析可以从结构和内容两个维度展开。近年来，随着自然语言处理和机器学习的发展，针对在关键词分析中存在一义多词、低频词易忽略等缺点而导致不能有效挖掘出科学文献领域主题的问题，许多学者给出了从内容维度入手的新方法。巴志超等[10]利用浅层神经网络语言模型Word2vec来构造出关键词的词向量形式，并进一步构建出关键词的语义网络。陈翔等[11]使用分段线性表示法和Word2vec模型来创建关键词的动态语义网络，并且以该网络为基础通过社区发现算法来发现其社区信息，将网络中的社区作为研究主题。通过关键词之间的语义网络构建方法可以提供关键词之间的语义关联信息，但如何跟原有的关键词共现网络相结合是目前关键词分析研究的难点。为了有效表示出文献在关键词层次上的主题倾向与其他文献的相近程度，本文首先利用Word2vec模型表示出关键词的词向量形式，利用AP聚类对原有关键词进行聚类并分组，然后使用分组编号与原有文献的关键词集相替换，作为文献的主题特征编码。利用该特征编码使文献与其他文献相连，相连权重等于文献之间主题特征编码的重叠个数，最终构建出文献在关键词层上的关系网络。利用网络表示学习算法Node2vec模型将文献节点的网络信息映射到低维、稠密的向量空间中，其整个过程如图2所示。

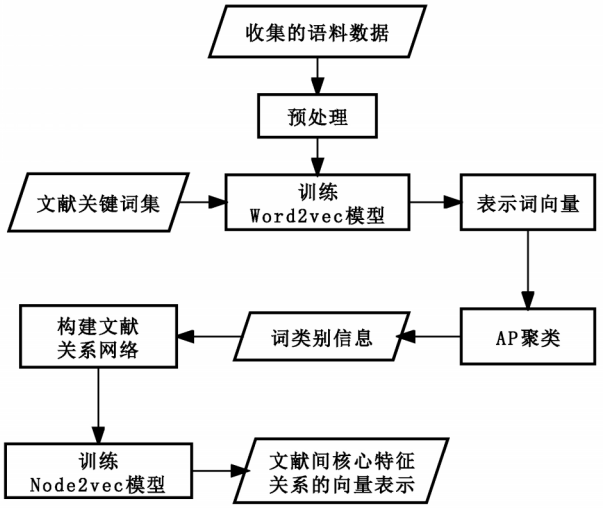

图2　关键词层向量生成图

由Mikolov等[12]提出的Word2vec模型，其词向量表示形式名为“Distributed Representation”，能有效避免原有“One-hot Representation”词向量形式中的维度高、词间关系难以表示等问题。Word2vec模型可以用语料库进行训练，能够以词向量形式表示出语料库中的词语，并且词之间的相近程度可以通过词在向量空间上的距离来度量。Word2vec模型有以输入词来预测上下文为核心的Skip-Gram和以输入上下文来预测当前词为核心的CBOW两种训练模式。本文以Skip-Gram作为Word2vec模型的训练模式，利用摘要及外部文本资料作为语料库，生成关键词的词向量形式。

由Frey等[13]提出的AP(Affinity Propagation)聚类算法是将所有样本当作潜在的聚类中心看待，其中每个样本对每一个样本具有吸引值(Responsibility)和归属值(Availability)两种属性。在其算法迭代过程中，不断更新样本对每一个样本的属性值，直至达到最大迭代次数或聚类中心决策稳定。样本对每一个样本的吸引值和归属值相加后得到的最大的、相对应的样本对象作为该样本的聚类中心。AP聚类不需要设置初始的聚类数目，其初始的聚类过程依托于样本之间的相似度，适合多维度的数据集。

样本的相似度量方法有欧氏距离法、余弦系数法等，本文采用余弦系数法进行关键词聚类，样本特征即为词向量。利用AP聚类对关键词进行聚类，生成$f\in N^+$种类别，并对不同类别附上标签$l\in\{1,2,\cdots,f\}$。对于文献集V，单篇文献$v\in V$，其对应关键词集为$C_v=\{c_1,c_2,\cdots,c_n\}$，$n$为单篇文献的关键词个数。利用类别标签1对关键词集进行替换，形成主题特征编码$L_v=[l_1,l_2,\cdots,l_n]$，其为多重集。构建以文献为节点的关系网络$G=(V,E)$，定义$E\subseteq V\times V$表示文献节点之间的连边，$e=(u,v)$、$e\in E$代表文献节点u、v附有权重$w_{uv}\geq 0$的连边。文献节点u、v之间的特征编码重叠个数作为文献连边的权重w_{uv}，即$w_{uv}=|L_u\cap L_v|$，以其视为文献节点u、v之间在关键词层面的主题重叠强度。

Node2vec是借鉴词向量表示(即Word2vec)的方法思想提出的一种节点网络表示算法。使用该算法可以用来减小原来由网络结构复杂而导致节点间结构关系难以表示的问题，其核心内容是通过添加潜在变量来学习节点周围的连接关系，将网络节点的结构关系通过低维向量表示出来[14-15]。Node2vec算法是本文实现文献间关系网络特征表示的关键。将已生成的文献网络G输入Node2vec算法中，表示出文献在关键词元数据上的特征数据，即对应文献v有该层的向量表示$m_v\in M$，$m\in R^d$ $(d<<|V|)$。

2.1.2 摘要层的向量表示

科学文献中的摘要单元是文献内容的简要概括，

相较于原有如TF-IDF的主题词统计模型，LDA模型是利用联合概率分布（即生成文档、主题和词三层次贝叶斯概率）来构建的。关鹏等[16]通过LDA模型分别对关键词、摘要及关键词+摘要三种结构的语料进行主题判断，以国内风能领域的科学文献资料为例，实验结果表明摘要、关键词+摘要作为语料是合理的选择。张文伟等[17]以纳米材料领域数据为例，利用LDA与BTM概率主题模型来抽取科学主题，结果表明，以摘要为语料时，LDA模型在主题颗粒度方面优于BTM模型。本文为了表示出文献摘要在主题方面的分布特征，将摘要文本视为文档，利用LDA模型进行主题建模，生成的摘要主题分布特征数据作为文献摘要层的向量表示，即对应文献v有该层的向量表示$z_i \in Z, z \in R^d$ ($d << |V|$)。

2.2 多元数据融合

多视图聚类（Multi-view Clustering）是从多视图数据入手来生成样本类别。其中，多视图数据来自同一实体在不同视角下的描述信息。随着大数据时代的来临，数据的收集和存储变得越发容易，使用多视图数据的情形变得十分普遍。多视图聚类可以挖掘出各视图数据之间的一致性、互补性信息，提升仅利用单一视图数据进行聚类的效果。目前关于多视图聚类算法可以分为四类，分别为图学习算法、后融合算法、协同训练算法以及子空间学习算法。其中，关于图学习算法的一种核心观点是指通过多视图数据构建多份相似度矩阵，利用这些矩阵来求出一份高质量的关系图，最后在该图上通过聚类算法得到样本类别[18]。Liang等[19]提出了一种基于图学习的、可行的多视图聚类算法（即Consistent Graph Learning算法），该算法将多个视角下的多个相似度矩阵分解为一致性和不一致性两部分，利用一致性和不一致性之间的关系构造目标函数并学习出一份统一的相似图（即融合图）。在得到该图后，为得到最终的多视图聚类结果，对融合图使用谱聚类。该多视图聚类算法在多种数据集中同其他多视图聚类算法相比，展现了出色的性能。

文献的关键词、摘要都属于作者展现文献主题的重要表象，两方面结合分析可以对学科领域的研究主题有更为准确的判断。本文通过文献中的关键词和摘要得到多元特征向量，将其作为多视图数据，使用Consistent Graph Learning算法中的SGF（Similarity Graph Fusion）方法来获取科学文献的主题类别。具体地，对关键词、摘要两层次向量表示M、Z进行处理，通过平方欧氏距离式（1）生成距离矩阵$D^{(1)}$、$D^{(2)}$。

$$d_{ij}=\| x_i-x_j \|_2^2 \tag{1}$$

式（1）中，x_i、x_j代表文献i、$j \in V$在同一层次下的向量表示。在每个距离矩阵中，将文献之间的距离通过高斯核函数RBF（Radial Basis Function）表示出文献之间的相似度，进而生成对应的初始相似度矩阵H，见式（2）。

$$h_{ij}=e^{(-\frac{\| x_i-x_j \|_2^2}{2\sigma^2})}=e^{(-\frac{d_{ij}}{2mean(D)})} \tag{2}$$

在初始相似度矩阵$H^{(i)}$ ($i=1,\cdots,t$)上，使用KNN（K-Nearest Neighbor）算法遍历所有的文献集V，每个文献都标记出最近的$k \in N^+$个点，即文献v对应的近邻文献集为$P_v^{(i)}=\{p_1,\cdots,p_k\}, p \in V$。进一步表示出多重集$Q_v=P_v^{(i)}=\{p_1,\cdots,p_k\}, p \in V$。进一步表示出多重集$Q_v=P_v^{(1)}+P_v^{(2)}+\cdots+P_v^{(t)}$，将其作为文献$v$最终的近邻文献集。通过所有近邻文献集$Q$表示出近邻矩阵$O$，其中$o_{ij}$为$Q_i$中元素$j$的重复度与$Q_j$中元素$i$的重复度之和的一半（当某一近邻文献集中缺少某一元素时，默认该元素在该集合的重复度为0）。通过式（3）求出经过处理后的相似度矩阵$W^{(1)}$、$W^{(2)}$。

$$w_{ij}=h_{ij}o_{ij} \tag{3}$$

将相似度矩阵W看作为由一致性矩阵A和不一致性矩阵E组成，即式（4）。

$$W^{(i)}=A^{(i)}+E^{(i)} \tag{4}$$

为了生成公共相似矩阵（即融合图）并确保其能学习到各个视图的一致性信息、使视图之间的不一致性表现稀疏，制定出优化目标函数式（5）。

$$\min_{\alpha,S,A^{(1)},\cdots,A(t)} \sum_{i=1}^{t} \| \alpha_i A^{(i)}-S \|_F^2 + \sum_{i,j=1}^{t} b_{ij}\alpha_i\alpha_j Tr((W^{(i)}-A^{(i)})\cdot(W^{(j)}-A^{(j)})^T) \tag{5}$$

$$b_{ij}=\begin{cases}\gamma, i \neq j\\ \beta, i=j\end{cases}$$

$$\text{s.t.} \| \alpha \|_1=1, \alpha \geqslant 0, S \geqslant 0, W^{(i)} \geqslant A^{(i)} \geqslant 0, i=1,\ldots,t$$

式（5）中，相似度矩阵$W^{(i)}$($i=1,\ldots,t$)的一致性矩阵$A^{(i)}$、缩放系数α_i及公共相似矩阵S是优化目标函数中的变量，γ和β为超参数。对目标函数式（5）使用交替迭代优化算法进行优化。具体地，固定$(A^{(1)},\ldots,A^{(t)})$、更新α和S，然后固定α和S、更新$(A^{(1)},\ldots,A^{(t)})$，以此重复。当达到最大迭代次数或目标值收敛时，迭代停止。通过$W^{(1)}$、$W^{(2)}$，获得迭代结果中的公共相似矩阵S，将其作为关键词、摘要这两个层次的融合图，并通过谱聚类得到聚类类别，将其作为文献集的主题划分结果。

3 方法验证

3.1 小型文献集

以近年来分别涉及情报学、管理学领域的两种文

献资料开展方法验证。文献资料均来源自CSSCI在2016年至2020年内收录的学术文献信息。从情报学文献资料中根据是否围绕信息传播、数据管理、评价研究、专利分析、用户分析这五种文献主题来挑选文献及其数据，生成含文献数1127篇的小型文献集A。同时，从管理学文献资料中根据是否围绕经济预测、能源问题、决策优化、员工关系、线路调度这五种文献主题来挑选文献及其数据，生成含文献数479篇的小型文献集B。利用基于文献多元数据融合的主题识别模型对这两个文献集进行主题识别研究，来判断模型的有效性。

将小型文献集作为已知类别的测试集，将其带入本文模型中进行处理。在关键词层的向量表示中，关于Word2vec模型生成关键词词向量部分，先收集所选文献集中的关键词集合，并将其加入jieba工具的词典中。将维基百科等外部资料和文献摘要等内部资料作为最终语料库，使其通过jieba工具进行分句、分词、去停用词等处理并输入到词向量维度为250、上下文窗口参数为10的Word2vec模型中。对未生成词向量的关键词，如“GREP模型”，用其上位词“模型”进行代替。对含词向量的关键词进行AP聚类。在文献集A的研究中的部分聚类结果如表1所示。

表1　部分关键词编号表

类别	关键词
1	网络谣言信息公开运行机制……
2	TOPSIS因子分析聚类分析……
3	社会影响力社会网络分析法服务质量评价……
4	学术期刊专利申请电子书……
5	中国台湾地区澳大利亚……

进一步地，以文献为节点构建出关键词层上的关系网络，并将其放到节点向量维度为20、节点游走次数为80、游走长度为20、游走方式为广度优先采样策略的Node2vec模型中，求出文献在关键词层的向量表示。另一方面，将处理后的摘要文本数据放入主题数为5的LDA模型中，求出文献在摘要层的向量表示。

为了探明SGF方法的效果，本文采取两种手段生成主题类别。一种是对关键词、摘要两个层次的向量分别构建出相似度矩阵a、b，并分别对其进行谱聚类。另一种是利用γ为10、β为10^{-5}的SGF方法同时对关键词、摘要两个层次的向量进行处理，生成公共相似矩阵c，对其进行谱聚类。谱聚类过程均采用经典算法Jordan算法进行计算[20]。本次初始近邻个数均为15。

将上述聚类结果及典型LDA方法（取文档在主题分布中的最大值索引作为主题类别）、Doc-LDA模型[21]生成的类别结果分别跟人工标签进行标准化互信息[22]（Normalized Mutual Information，NMI）计算。标准化互信息能够评价模型给出的样本类别结果与实际类别之间的差距情况，其范围在0到1之间，并且数值越高、方法效果越好。本文使用标准化互信息作为评价方法好坏的指标，关于文献集A、B的指标结果如图3所示。

可以看出，利用单相似图进行谱聚类后的得分较差，还有改进的空间，而利用融合相似图的聚类得分均高于其他单相似图的聚类得分，这表明SGF方法能够在一定程度上提升原来依据单视图数据进行聚类的效果。另一方面，利用融合相似图的聚类得分也高于典型LDA方法及Doc-LDA模型的得分，即与人工主题类别划分结果最为接近，表明本文模型的主题识别性能出色。

3.2　大型文献集

对原有的情报学文献集进行扩充，生成含文献数10523篇且无主题标签的大型文献集。利用上述所提到的各种方法对该文献集进行同样地处理，其中关于文献集主题个数设置问题，以摘要层上的LDA模型中的困惑度[5]确定。困惑度越低意味着LDA模型主题识别表现越好。对摘要层上的LDA模型主题设置数（从0开始）每增加5个时求一次困惑度。随着LDA模型的主题设置数的增多，困惑度变化呈现下落趋势且下降速度先快后慢，在主题设置数为60时，困惑度曲线逐渐平缓，因此以60作为大型文献集的主题数。

为了有效评价在大型文献集中上述方法在主题识别效果上的差异，参考林江豪等[23-24]研究中的评估方法，以主题可解释性和类内准确度这两方面进行评

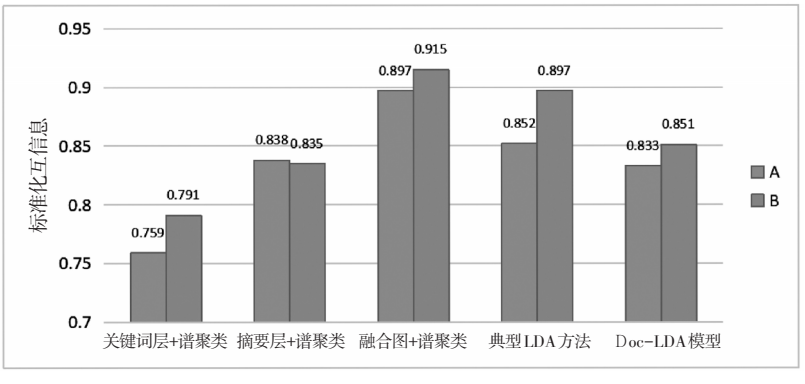

图3　标准化互信息对比图

估。具体地，在上述方法的主题类别结果中，均随机抽取不同主题类别文献集5份，并均对其进行编号（从1,2,...,5）。邀请4名情报学领域的学者进行评估工作，要求评估者通过某一编号类别内文献的题录信息，用1-3个词概括出该类别所涉及的主题内容，同时给出概括过程中生成主题词的难易等级，等级范围为1-5（5个等级，分值越高越容易）。其部分概括词如表2所示（表中小括号内为某一评估者对类内文献集所给出的概括词）。

从表2中可以看出，在某一方法中的某一类别内，评估者之间所给出的主题概括词较为一致，因此以某一方法内给出的所有难易等级的平均值代表该方法在生成主题类别上的可解释性程度是合理的。另一方面，设置平均准确率指标来评估某一方法的类内准确度。即在某一方法的主题类别结果中，在$i\in(1,2,...,y)$编号类别下识别出明显不能与对应主题相匹配的文献数量c_i，并记录i类别的文献总数t_i，通过式（6）求出平均准确率P。

$$P=\frac{\sum_{i=1}^{y}(t_i-c_i)}{y} \quad (6)$$

计算出的平均难易等级及平均准确率结果如图4、图5所示。

由图4、图5可以看出，基于多元数据融合的主题识别模型（融合图+谱聚类）在平均难易等级及平均准确率上优于其他方法，说明该模型相较于其他方法在主题识别上的可解释性高、同一主题类别内的文献关联度强，适合科学文献的主题识别工作。

4 结论与不足

多元数据融合指利用数据融合方法有效加工、整合不同属性但存在关联性的数据，进而利用更加丰富的信息来发现更准确、更深层次的实体关系[25]。为了避免原有利用单一要素判断科学文献主题的方法弊端，本文提出了基于多元数据融合的主题识别模型，其重点分别是构建了基于关键词层的文献间核心特征关系向量、基于摘要层的文献介绍信息主题分布向量和通过多视图聚类方法来处理这些特征数据、生成主题类别。为了比较

表2 部分主题概括词表

	编号1	编号2	编号3
关键词层+谱聚类	(大数据)(大数据、情报研究)(情报研究)(大数据)	(专利研究)(专利分析)(专利)(专利分析)	(数据开放)(高校图书馆、开放数据)(政府数据开放)(图书馆)
摘要层+谱聚类	(高校图书馆)(高校图书馆、服务)(高校图书馆)(图书馆)	(知识服务)(知识、图书馆)(知识管理)(知识)	(网络研究)(网络舆情、知识网络)(网络分析)(网络)
融合图+谱聚类	(评价研究)(评价方法、期刊评价)(科研评价)(科技评价)	(图书馆服务)(高校图书馆)(公共图书馆)(图书馆)	(知识管理)(知识、协同)、(知识管理)(知识网络)
典型LDA方法	(突发事件)(应急决策)(突发事件)(公共安全)	(图书馆)(图书馆、文化、信息情报)(图书情报)(文化传播)	(网络舆情)(情感分析、网络舆情)(网络舆情)(网络舆情)
Doc-LDA模型	(社交网络)(网络舆情、应对机制)(文本挖掘)(算法)	(电子政务服务)(开放式创新、知识)(网络舆情)(舆情)	(信息生态)(图书馆、素养教育)(知识管理)(知识)

出该模型跟现有主题识别技术的差别情况，在两种规模的文献集中，经过对比验证，该模型可以挖掘出学科领域文献在主题上的分布信息，并且在科学文献主题识别的准确性和可解释性上优于其他方法。

该模型的提出能给科研人员在面对海量的科研成

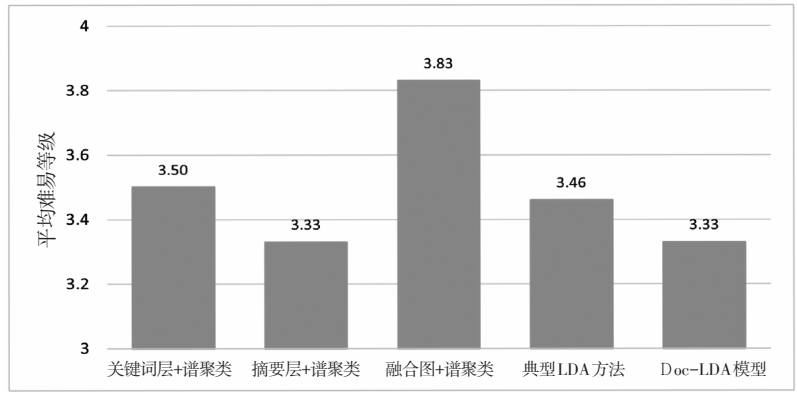

图4 平均难易等级对比图

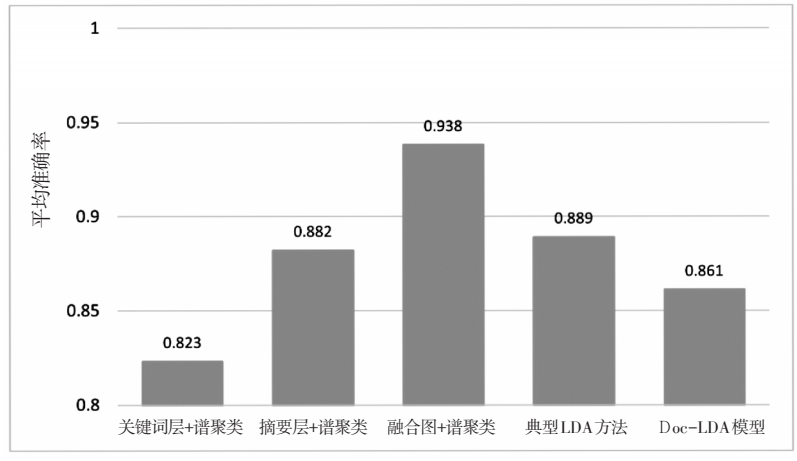

图5 平均准确率对比图

果、如何清晰且准确地把握学科领域中的前沿主题时提供新思路。同时，在另一方面，对科研成果的准确区分有利于解决现有科研评价总以数量为准的弊端。本文模型对期刊论文按研究内容进行有效区分，能对科研机构的评价工作内容向层次性、相对性转变提供了有利工具。

此次研究也存在着可以改进的地方。科学文献除了关键词、摘要两种元数据外，还存在作者、引文等其他元数据，这些数据也同样含有文献之间的关联性，如何将这些关联性用到文献主题识别上还有待继续研究。学科领域的研究主题在不同时间段内的分布特征是不同的，如何利用本文模型进行学科领域文献的演化分析、探明学科领域的发展趋势和研究前沿是未来值得深化的问题。

参考文献

[1] Bhattacharya S ,Basu P K . Mapping a research area at the micro level using co-word analysis[J]. Scientometrics, 1998, 43(3):359-372.

[2] Li M . An exploration to visualise the emerging trends of technology foresight based on an improved technique of co-word analysis and relevant literature data of WOS[J]. Technology Analysis and Strategic Management, 2017, 29(6):655-671.

[3] 霍朝光, 魏瑞斌, 张 斌. 基于PageRank和Node2vec的研究热点与集群发现——以国际深度学习研究领域为例[J]. 情报杂志, 2020, 39(8):7.

[4] Papadimitriou C H , Raghavan P , Tamaki H , et al. Latent semantic indexing: a probabilistic analysis[J]. Journal of Computer and System Sciences, 1998, 61(2):217-235.

[5] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[6] 侯 捷. 基于文本挖掘的管理科学学科研究热点及前沿发现与分析[D]. 北京:北京工业大学,2019.

[7] 谭春辉, 熊梦媛. 基于LDA模型的国内外数据挖掘研究热点主题演化对比分析[J]. 情报科学, 2021, 39(4):12.

[8] 许海云, 武华维, 罗瑞,等. 基于多元关系融合的科技文本主题识别方法研究[J]. 中国图书馆学报, 2019, 45(1):13.

[9] 武华维, 罗 瑞, 许海云,等. 科学技术关联视角下的创新演化路径识别研究述评[J]. 情报理论与实践, 2018, 41(8):7.

[10] 巴志超, 杨子江, 朱世伟,等. 基于关键词语义网络的领域主题演化分析方法研究[J]. 情报理论与实践, 2016, 39(3):6.

[11] 陈 翔, 黄 璐, 倪兴兴,等. 基于动态语义网络分析的主题演化路径识别研究[J]. 情报学报, 2021, 40(5):13.

[12] Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[EB/OL].(2013-09-07)[2021-09-26].https://arxiv.org/pdf/1301.3781.pdf.

[13] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[14] Grover A, Leskovec J. Node2vec: scalable feature learning for networks[C].Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco,USA,2016: 855-864.

[15] Jaeger S ,Fulle S , Turk S . Mol2vec: Unsupervised machine learning approach with chemical intuition[J]. Journal of Chemical Information & Modeling, 2018, 58(1):27-35.

[16] 关 鹏, 王曰芬, 傅 柱. 不同语料下基于IDA主题模型的科学文献主题抽取效果分析[J]. 图书情报工作, 2016, 60(2):10.

[17] 张文伟, 赵 辉. LDA与BTM概率主题模型抽取科学主题效果比较研究[J]. 情报工程, 2020, 6(2):12.

[18] Kang Z , Pan H , Hoi S , et al. Robust graph learning from noisy data[J]. IEEE Transactions on Cybernetics, 2019, 50(5): 1833-1843.

[19] Liang Y , Huang D , Wang C D . Consistency meets inconsistency: a unified graph learning framework for multi-view clustering[C]. IEEE International Conference on Data Mining (ICDM), Beijing, China, 2019.

[20] Luxburg U V . A tutorial on spectral clustering[J]. Statistics and Computing, 2004, 17(4):395-416.

[21] 张卫卫, 胡亚琦, 翟广宇,等. 基于LDA模型和Doc2vec的学术摘要聚类方法[J]. 计算机工程与应用, 2020, 56(6):6.

[22] Li Y ,Nie F , Huang H , et al. Large-scale multi-view spectral clustering via bipartite graph[C]. Twenty-ninth Aaai Conference on Artificial Intelligence, Austin, Texas, USA, 2015.

[23] 林江豪, 周咏梅, 阳爱民,等. 结合词向量和聚类算法的新闻评论话题演进分析[J]. 计算机工程与科学, 2016, 38(11):7.

[24] 王小红, 浦江淮, Colin Allen . 探索主题模型可解释性问题[J/OL]. 中国社会科学报, [2020-11-03]. http://sscp.cssn.cn/xkpd/kxyrw/202011/t20201103_5210759.html.

[25] Hai-YunXu, Zeng-HuiYue, Chao Wang, et al. Multi-source data fusion study in scientometrics[J]. Scientometrics, 2017, 111(2):773-792.

[作者简介]邱均平，男，1947年生，杭州电子科技大学中国科教评价研究院博士生导师，资深教授。

孙月瑞，男，1997年生，杭州电子科技大学管理学院硕士研究生在读。

周贞云，男，1979年生，杭州电子科技大学中国科教评价研究院博士研究生，副教授（通讯作者）。

收稿日期：2022-04-11