情报学报　2023 年 5 月　第 42 卷　第 5 期
Journal of the China Society for Scientific and Technical Information, May 2023, 42(5): 511-524

DOI: 10.3772/j.issn.1000-0135.2023.05.001

知识单元重组视角下的科学主题预测研究

梁继文[1,2]，杨建林[1,2]，王　伟[1,2]

（1. 南京大学信息管理学院，南京　210023；2. 江苏省数据工程与知识服务重点实验室，南京　210023）

摘　要　准确的科学主题预测能够明确学科未来的发展方向，为科研领域的发展规划和管理决策提供参考。本文着眼于新生科学主题的预测，基于知识单元重组视角，将主题-特征词的表征关系类比为科学概念-知识单元的表征关系，提出科学主题预测方法。首先，使用LDA（latent Dirichlet allocation）主题模型获取全局主题、特征词与概率矩阵，通过转置向量空间获得特征词向量；其次，运用ARIMA（autoregressive integrated moving average model）模型预测特征词的词频并计算向量调节系数，从而获得特征词预测向量，运用*t*-SNE（*t*-distributed stochastic neighbor embedding）算法将预测向量降维，并使用模糊*C*-均值算法将低维预测向量聚类生成预测主题，实现知识单元的重组；最后，筛选出由多个原始主题聚合而来、具有全新释义的预测主题，将其视为科学主题预测结果。本文以“知识管理-知识组织-知识服务”领域为例进行实证研究，预测出智库、数字人文等在已有领域研究中尚未出现的新词与相关主题，并通过特征词直接聚合与概念集成这两种主题映射模式，获得这些新生主题的基本内涵与相关研究内容。实证结果表明，本文提出的科学主题预测方法能够准确地预测出新生主题。

关键词　知识单元；科学概念；科学主题；主题预测；向量调节

Research on Scientific Topic Prediction from the Perspective of Knowledge Unit Reorganization

Liang Jiwen[1,2], Yang Jianlin[1,2] and Wang Wei[1,2]

(1. *School of Information Management, Nanjing University, Nanjing*　210023;
2. *Jiangsu Key Laboratory of Data Engineering & Knowledge Service, Nanjing*　210023)

Abstract: Accurate scientific topic prediction can clarify the future development direction of a given discipline and provide a reference for the development planning and management decision-making in the field of scientific research. This paper focuses on the prediction of new scientific topics based on the perspective of knowledge unit reorganization, compares the representation relationship between the topic and feature words to the representation relationship between scientific concepts and knowledge units, and proposes a scientific topic prediction method. First, the LDA (latent Dirichlet allocation) topic model is used to obtain the global topic, feature words, and probability matrix and obtains the feature word vector by transposing the vector space; second, the vector adjustment coefficients are calculated based on the feature word frequencies predicted by the ARIMA (autoregressive integrated moving average model) model to obtain the feature word prediction vectors, the *t*-SNE (*t*-distributed stochastic neighbor embedding) algorithm is applied to reduce the dimensionality of the prediction vectors, and then the low-dimensional prediction vectors are clustered by the fuzzy *C*-mean algorithm to generate prediction topics to realize the reorganization of knowledge units. Finally, the prediction topic with a new interpretation is selected from the aggregation of several original topics, and this is regarded as the scientific topic prediction re-

收稿日期：2022-05-11；**修回日期**：2022-10-28

基金项目：国家社会科学基金重点项目“大数据环境下领域知识加工与组织模式研究（20ATQ006）。

作者简介：梁继文，女，1995年生，博士研究生，主要研究领域为智能化信息处理；杨建林，男，1970年生，博士，教授，博士生导师，主要研究领域为信息检索、智能化信息处理、学术评价、情报学基础理论，E-mail：yangjl@nju.edu.cn；王伟，男，1994年生，博士，主要研究领域为知识组织、智能化信息处理。

sult. This paper takes the field of “knowledge management-knowledge organization-knowledge service” as an example for conducting empirical research. The results show that the proposed scientific topic prediction method in this paper can effectively predict new scientific topics from which the essential concepts and the corresponding research content of some words have not appeared at that time, such as digital humanities and knowledge payment.

Keywords: knowledge unit; scientific concepts; science topics; topic prediction; vector adjustment

0 引 言

科学预测是指以合适的理论方法为指导、以搜集分析对象的历史数据为前提，依据事物发展的客观规律揭示事物变迁过程的内在联系与发展趋势，预测未来的科学新概念或未知的科学现象[1-2]。对科学主题进行准确预测，可以把握领域研究的动向与趋势，有助于领域知识发现与知识管理，并能够提前规划学科发展部署，从而强化科技创新布局、优化科技资源配置，为科学管理与科学决策提供参考。

目前，图情学界对于主题挖掘与主题演化的研究已经较为成熟，但总体呈现“重演化，轻预测”“强指标，弱解释”的特点，多数研究侧重于使用定量指标分析已知主题的演化过程并预测未来发展趋势，但对于在过去尚未明确出现的主题的预测能力稍显欠缺，如何基于现有主题的发展趋势来预知未来可能出现的新生主题仍有待探索。通常情况下，某一时段突然涌现的新词表征着新生主题，预测未曾出现过的新词具有较大难度，但若从知识单元重组视角出发，将新词的内涵概念析毫剖厘，不难发现新概念的诞生其实有迹可循。学者们认为科学概念是由知识单元构成的，指出科学概念与知识单元具有可分解与可重组的特性，将科学概念-知识单元的转化过程视为科学创造的过程，即将科学概念分解为若干知识单元，再根据实际需求，融合人类认知，将获取的知识单元进行创新性的重组与凝聚，极有可能衍生全新的科学概念，最终由新科学概念析出新知识单元[3-5]。本文聚焦于新生科学主题的预测，从知识单元重组的视角出发，将主题-特征词间的表征关系类比为科学概念-知识单元间的表征关系，提出基于主题模型与特征词向量调节的科学主题预测方法。该方法通过主题建模提取原始主题、主题-特征词概率矩阵与特征词向量，使用时间序列方法预测特征词未来发展趋势，并设置基于词频的向量调节系数，将调节后的特征词向量视为预测向量，继而进行降维聚类获取预测主题，筛选出由多个不同原始主题聚合而来、释义发生显著变化的预测主题，将其视为新生主题并作为预测结果。最后，面向“知识管理-知识组织-知识服务”领域进行实证分析，验证本文方法的可行性与可靠性。

1 相关研究综述

1.1 主题挖掘与主题演化

主题挖掘侧重于从宏观层面对已有的领域研究进行整体分析，主题演化围绕强度演化与内容演化两个方面探寻领域研究主题状态的划分与关系的渐进。

在研究内容方面，学者们根据不同主题的特征，将科学主题分为新兴主题、热点主题与潜在主题等多种状态，设置文献数量、主题强度、主题新颖度与主题演化偏离度等特征测度指标作为强度演化标识[6-7]，并据此判断识别各类主题；部分研究融合相似度或距离等主题关联测度指标，结合内容分析来探寻主题间的融合、继承、分裂或消亡等演化关系，综合构建主题演化路径[8-10]。

在研究方法方面，现有研究大致可分三类：①基于计量的方法，通常使用领域内科技文献的外部特征进行量化分析，通过数理统计的方法刻画和评价当前领域现状[11]。②基于网络的聚类方法，通过计算关键词的共现关系构建共词矩阵，或基于文献的引用、同被引和耦合关系构建引文网络，部分研究使用UCINET、CiteSpace等网络分析软件提取文献外部特征、结合多种相似度度量方法进行多维聚类，从而识别关键主题、分析主题发展趋势，并提取主题演化规律[12-13]；部分研究使用社区划分算法等复杂网络方法，进行主题挖掘与结构分析[14]。相较之下，基于引文网络的主题分析能较好地反映领域内知识结构特征。③基于主题建模的方法，通过主题挖掘和揭示领域内的主题与研究内容，具有较强的可解释性[15]。研究常使用LDA（latent Dirichlet allocation）模型进行主题挖掘[16-18]，结合主题关联测度指标筛选并构建主题间的关联[19]。此外，随着表示学习的兴起，部分研究将不同粒度的文本表示

与主题建模相结合，先后衍生出LDA2vec模型[20]、主题词嵌入模型[21]等融合了语义信息的主题挖掘模型。

1.2　主题预测

主题预测以主题挖掘与演化研究为基础，强调对未来内容与发展方向进行预见。在以定性方法为主的预测研究中，学者们辅以可视化工具构建演化路径、识别演进轨迹，从内容的层面感知主题的演化特征与发展趋势，据此对未来可能出现的研究方向进行人为研判，具有较好的可解释性[22-23]。该类研究受演化路径构建效果和选取的主题关联阈值的影响较大；对于预测结果的分析需要融合专家智慧，或拥有较强的领域背景与全面的知识储备来支撑结论，强主观性的方式对于当前大数据环境下的主题迁徙特征感知的敏感度较低。

在以量化方法为主的研究中，将预测指标与时间序列分析方法相融合是较为经典的研究模式[24-25]。部分研究基于多重主题特征测度指标筛选出热门或前沿主题，并提取区分度较强的指标作为预测指标（如热度、新颖度、迁徙度等），选用曲线拟合或自回归的方法进行指标的时间序列预测，研判领域未来发展方向[8]。同时，随着深度学习技术的发展，以LSTM（long short-term memory）为代表的神经网络也被应用于主题特征指标预测[26-27]。与定性方法相比，该类方法更具科学性，也能够体现出主题随时间变化的趋势与惯性。但仍存在以下问题：①预测分析通常以新兴、前沿等特定类别主题的识别为基础，但各类主题的含义、特征与甄别标准均由人为制定，仅预测限定类别的主题趋势，以此表征领域未来的发展态势难免会有失偏颇。②基于特征或指标来描述主题发展的趋势是一个由表及里的过程[28]，所选用的预测指标是否具有较强的代表性与科学性、能够与主题发展态势高度契合仍有待商榷，预测结果的可解释性稍显欠缺。

总体而言，与主题挖掘演化的研究体量相比，关于主题预测的研究相对偏少，主题预测似乎更多地被作为主题挖掘与演化研究结果的副产品出现。同时，现有的主题预测研究侧重于探索预测指标的选取、更多地关注指标的强弱变化趋势，是根据已有主题的发展惯性而做出的趋势预测，但无法发现新生科学主题。鉴于此，有必要基于领域内的全局主题，探索如何预测未来新生成的科学主题。

2　研究设计

2.1　总体思路与研究框架

本文将主题-特征词的表征关系类比为科学概念-知识单元的表征关系，具体如图1所示。在具体情景限定下，已知科学概念通常由多个知识单元的组合来表征其内涵。知识单元及其组合方式的变化隐含新生科学概念的发生和发展，随着新生科学概念的进一步完善，可能导致新知识单元的产生。这与科学概念和知识单元之间往复的解构与重构过程类似，领域研究主题和特征词之间也具有这样的表征关系。某一领域研究主题的文献集合经过情报组织与加工后，形成能够描述阐释该类主题基本含义的科学概念；选择提取特征并聚类，抽象出一组特征词构成特征词群，使用特征词群与词频共同表征该科学主题；在不断演化的过程中，科学主题既有原始主题的惯性延续，又有特征词群分解后融合人类认知、面向特定情景需求、经过重组与解读构成的新生主题。所以特征词的组合变化可以反映已有主题的演变和新生主题的涌现。

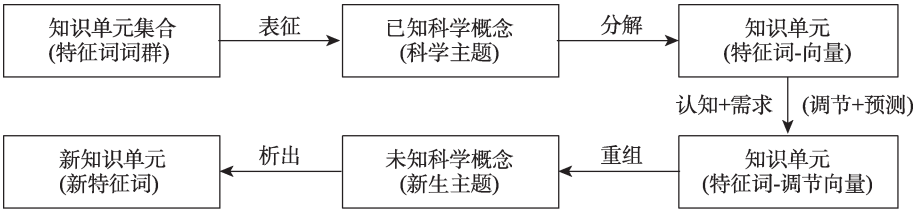

图1　“科学概念-知识单元”与“领域主题-特征词”的表征关系的转化

本文以图情档学科中的“知识管理-知识组织-知识服务”领域为例，从知识单元重组视角出发，构建研究框架如图2所示。①数据获取及预处理：详见3.1节；②全局主题提取：以LDA主题建模为基础，将原始主题分解为多个特征词，通过转置向量空间获取特征词向量；③词频预测与向量调节：将特征词词频作为基础预测指标，使用时间序列模型预测特征词未来词频，基于相对词频设置特征词

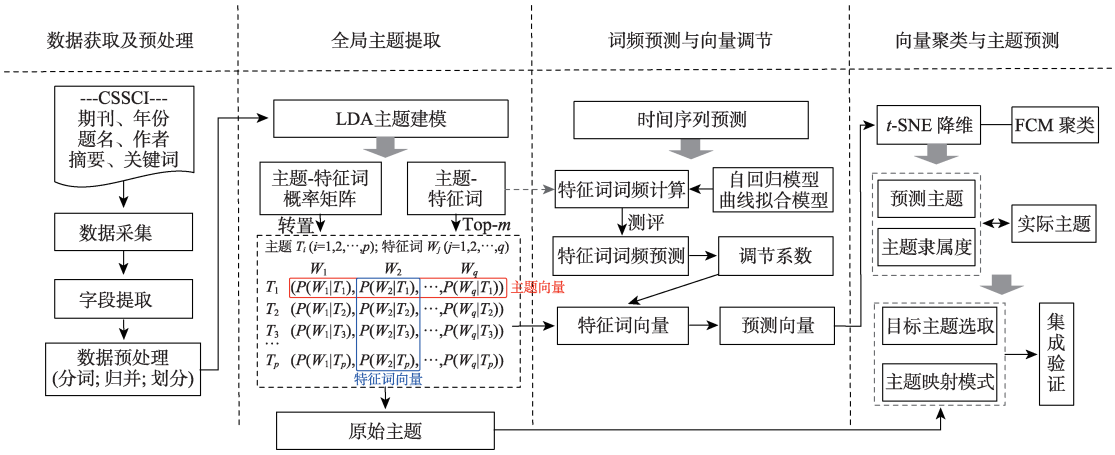

图2 科学主题预测研究框架

向量的调节系数，将过去时段的特征词向量转化为预测向量；④向量聚类与主题预测：通过对预测向量进行降维与聚类来将特征词重组，并通过目标主题选取、主题映射模式发掘与主题内容推理演绎来与实际主题进行对比，综合对领域主题预测进行实证分析。

图3展示了本文所采用的预测与验证模式。首先，为验证预测主题的准确性，根据文献的出版年份将数据集划分为训练集和测试集：去过时段 P（1998—n 年）的数据为训练集，未来时段 F（n+1—2021 年）的数据为测试集；其次，对训练集数据进行主题建模获取原始主题 T，通过词频预测与向量调节、向量聚类与主题预测后，获取预测主题 PT；最后，使用测试集数据进行主题建模，获取实际主题 RT，通过对比分析预测主题 PT 与实际主题 RT 的释义异同，用于验证本文所提出的主题预测方法的可行性与适用性。

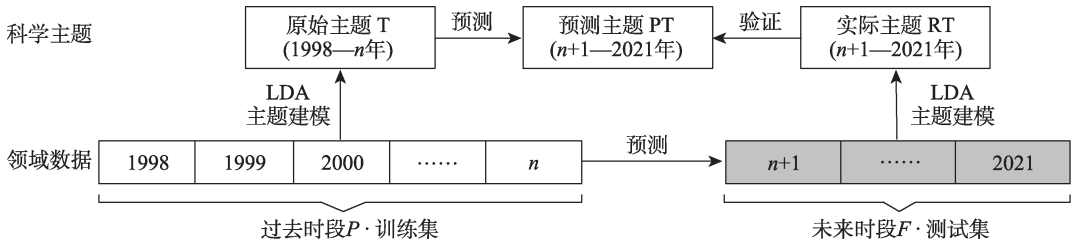

图3 科学主题预测及验证模式

2.2 全局主题提取

LDA 模型常用于挖掘领域文本的主题，每个主题均由具有代表性的多个特征词构成的集合来表征概括主题的内容，同时生成文档-主题和主题-特征词两个关联概率矩阵，特征词可以被转化为基于主题的数值向量，各维度上的数值表征属于相应主题的概率分布。本文使用 LDA 模型对训练集数据进行主题挖掘，通过计算困惑度（perplexity）辅助确定主题数量。可获取主题 T_i（i = 1,2,⋯,p）与相应的特征词 W_j（j = 1,2,⋯,q），以及生成的主题-特征词概率矩阵。主题可被视为存在于由 q 个特征词构成的 q 维向量空间中，各特征词向量是空间中主题向量的分量，将主题向量表示为 $\overrightarrow{T_i}=(P(W_1|T_i),P(W_2|T_i),\cdots,P(W_q|T_i))$，将主题-特征词矩阵中的概率进行归一化处理，权重即概率分布值。

传统向量空间转置后，可以获取词在文本空间中的表示[29]。相应地，对由特征词构成的向量空间进行转置，可得到特征词在主题空间中的向量表示 $\overrightarrow{W_j}=(P(W_j|T_1),P(W_j|T_2),\cdots,P(W_j|T_p))$，将转置后的词-主题概率矩阵进行归一化处理，用于获取融合了主题信息的特征词向量。鉴于特征词数量过多，为避免向量空间高维稀疏，本文将特征词概率值降序排列，选用各主题中权重最大的前 m 个特征词，取并集构建特征词集合。

2.3 词频预测与向量调节

2.3.1 词频计算与预测

学科领域研究主题的状态具有延续性，特征词

的出现趋势具有惯性，特征词频次是反映主题状态最为直接、有效的外部特征。同人工设定的主题状态预测指标相比，未经加工的词频指标更为客观、准确，具有较强的科学性与普适性，在预测研究中可以最大限度地减少误差，因此，本文使用时间序列方法预测特征词的未来词频，用于计算后续向量调节系数。

为全面精准地感知词频的演化趋势，本文将时间窗口设定为1年进行多步预测。在预测序列数据时，选用直接预测与递归预测两种方式。其中，直接预测是指基于过去时段P（1998—n年）中特征词集合的逐年词频，直接预测其在未来时段F（n+1—2021年）的逐年词频；递归预测是指通过时段P（1998—n年）中逐年词频预测第n+1年的数据，然后将第n+1年的预测值递归至原始数据中，再使用（1998—n+1年）年的数据预测第n+2年的数据，依此类推。模型方面，多项式曲线拟合与自回归是进行时序预测常用的两类方法。其中，多项式曲线拟合模型包含线性、二次等多种形式；自回归方法中的ARIMA（autoregressive integrated moving average model）常用于非平稳时序数据，与本文的特征词词频预测需求相契合。

为减少预测误差，本文同时选用多项式曲线拟合模型与ARIMA模型进行词频预测，采取直接预测与递归预测两种预测方式进行对比预测效果，使用平均绝对误差MAE（mean absolute error）作为误差检验指标，用于确定最终的预测模型与预测方式，计算方式为

$$\mathrm{MAE}=\frac{1}{n}\sum_{t=1}^{n}|a(t)-f(t)| \tag{1}$$

其中，a表示实际值；f表示预测值。MAE值越小，预测效果越好。

2.3.2　向量调节系数测度

LDA模型以文档与特征词构成的词频矩阵为基础，得到主题与特征词间的概率分布，因此，通过向量空间转置得到的特征词向量与各特征词词频间存在映射关系。词的逐年词频即绝对词频，但词频的真实价值受多重因素影响，将词的逐年频次归一化结果作为相对词频，能够有效弱化绝对词频的数值差异，更好地反映词本身的变化情况[30]。所以本文在衡量特征词向量的过去与未来变化时，使用特征词相对词频进行映射。鉴于研究秉承全局视角，仅有过去与未来两个整体时段，将现有的逐年加权相对词频调整为总体相对词频。

将同一特征词W_j的向量按时间发展分为原始向量与预测向量两种状态。预测向量$\overrightarrow{W'_j}$是在原始向量$\overrightarrow{W_j}$的基础上受趋势惯性影响所产生的一种调节与转化，可以使用数乘向量的形式近似地表示，即

$$\overrightarrow{W'_j}=\delta_j\cdot\overrightarrow{W_j} \tag{2}$$

其中，δ_j为向量的调节系数。结合特征词向量与词频间存在的映射关系，δ_j等于未来的相对词频与过去的相对词频之比，计算方式为

$$\delta_j=(\mathrm{tf}'_j/\mathrm{TF}')/(\mathrm{tf}_j/\mathrm{TF}) \tag{3}$$

其中，tf_j表示W_j在过去时段P中的词频；TF表示特征词集合总词频；tf'_j表示W_j在未来时段F中的预测词频；TF′表示特征词集合的总预测词频。综上，基于词频预测计算相对词频与向量的调节系数，进而获取各特征词预测向量。

2.4　向量降维与聚类

词与主题共同构建的向量空间维度过高，为避免维数灾难，需要在向量聚类前进行降维处理。t-SNE（t-distributed stochastic neighbor embedding）算法由SNE算法改进而来[31]，基于局部的流形学习可以更好地表达高维目标数据的复杂非线性关系，降维效果明显且呈现较好的聚集性，主题轮廓鲜明[32]，因此，本文选用t-SNE算法进行特征词向量降维。

聚类分析方法包含层次聚类、密度聚类和划分聚类等。其中，层次聚类适用于探寻目标数据间存在的上下位、并列、重叠等层级关系；密度聚类适用于密度分布均匀的目标数据；划分聚类分析是进行科学主题识别时常使用的方法，如使用k-means进行关键词聚类来识别学科主题[33]，但k-means算法非此即彼的划分状态与实际情况存在偏差。模糊C-均值聚类算法（fuzzy C-means，FCM）由k-means算法改进而来，其与k-means硬聚类相比更为灵活。FCM算法是基于对象数据自身的属性特征构造模糊矩阵，每个对象不再仅属于某一特定的簇，而是以特定概率对应多个簇，这符合本文中词与主题所呈现的一对多的映射关系，因此，本文选用FCM算法进行词向量的聚类与主题提取，将目标向量集合分为多个模糊簇，目标向量分属各簇的隶属度总和为1。

3　“知识管理-知识组织-知识服务”领域研究主题预测

3.1　数据获取及预处理

“知识管理-知识组织-知识服务”是图书情报学

界中重要的研究领域。知识管理聚焦于管理知识的流程架构，知识组织侧重于对知识进行序化，知识服务则面向用户需求提供知识或对策，三者联系紧密，相辅相成[34]。本文以“知识管理-知识组织-知识服务”领域为例，使用CSSCI（Chinese Social Sciences Citation Index）中的期刊文献作为研究数据源，分别以知识管理、知识组织与知识服务作为主题词进行检索。检索时间范围限定为1998—2021年，将获取的文献题录信息作为目标数据，并进行数据清洗与预处理，具体包含去重、字段选取、无关文本剔除、分词与同义词合并。本文选用目标数据中的题名、关键词与摘要字段作为实验数据，同时保留文献的作者、期刊与时间信息作为解释性数据。剔除与实验目标无关的数据，如征稿启事、领域要闻等，并将摘要等目标字段为空的文献信息剔除。与词相比，词组更能体现出研究内容，因此，本文将关键词作为补充词典用于提升领域文本分词的准确性，使用jieba对目标文本做分词处理，构建领域同义词、缩略词词典，并进行跨语言文本替换，统一词义相同但形式不同的词，如“非遗”与“非物质文化遗产”、“blog”与“博客”。

最终，获取10870篇文献作为实验数据，文献数量分布如图4所示。领域文献数量呈先增长后下降的趋势，并于2009年达到峰值，为856篇，累积文献数量增长速度逐渐放缓。从整体上看，我国“知识管理-知识组织-知识服务”领域的前期研究取得了较多成果，发展至今已步入成熟的缓慢增长阶段。

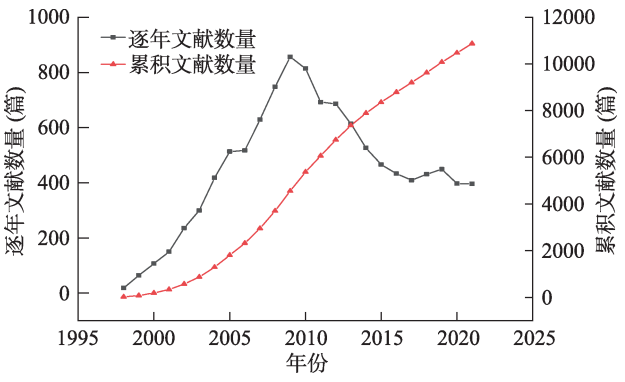

图4 领域文献数量年度分布

使用时间序列模型进行预测时，需要足够长的历时数据来充分反映变化趋势与惯性，2015年是国内“知识管理-知识组织-知识服务”领域中一个重要的时间节点，在研究内容、应用技术与发展方向上做出了调整改变[35]，因此，将2015年作为时间划分节点。设定1998—2015年为过去时段P，对应数据作为训练集；2016—2021年为未来时段F，对应数据作为测试集，旨在更好地验证主题预测的有效性。

3.2 领域主题识别与预测

使用LDA模型对训练集数据进行主题建模，为保证建模效果过滤掉文档中词频小于3的词。在计算困惑度时，考虑到主题数量较多会出现过拟合的情况，将主题个数的阈值设置为[5,50]，步长为1，迭代次数为200，计算困惑度。当主题数为16与35时，分别对应了困惑度的最小值与次小值。辅以pyLDAvis主题可视化工具，观察主题的重叠与分布情况，最终确定主题数量为35。为避免概率矩阵的维度灾难、保障主题的完整性与可解释性，本文提取每个科学主题中概率分布较大的前20个词来表征主题内容。由此获取了表征“知识管理-知识组织-知识服务”领域1998—2015年的研究主题与主题-特征词概率矩阵，对矩阵进行转置与归一化处理后生成特征词向量。

基于上文生成的主题-特征词概率矩阵提取特征词集合，获取特征词672个，且多为词组结构。其中有494个特征词的词长大于2，占比73.51%；有446个特征词的词长大于3，占比66.37%。首先，计算特征词集合在1998—2021年的逐年词频，构建维度为672×24的特征词-年份词频矩阵。其次，使用训练集数据预测未来P时段的逐年词频，分别运用Python语言自编程序与SPSS软件进行多项式曲线拟合与ARIMA建模。其中，多项式曲线拟合的最大阶数由对比R^2确定，ARIMA模型的差分阶数由专家建模器自动计算选取。两种模型与两种预测方式的年平均预测误差值如图5所示，其中选用多

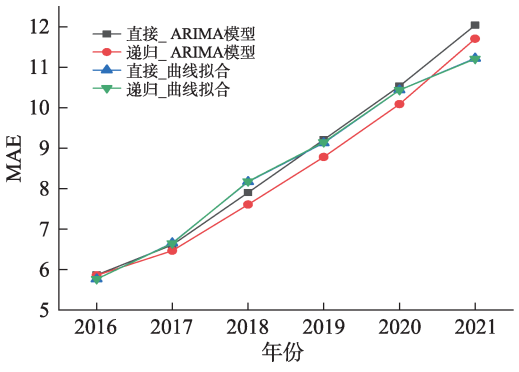

图5 曲线拟合与自回归模型MAE值比较

项式曲线拟合方法进行直接预测与递归预测的效果相差无几，而选用 ARIMA 模型和递归方式时的 MAE 值在多数时间低于其他方法，预测效果最好，因此使用 ARIMA 模型进行递归词频预测。最后，根据预测词频计算各特征词向量的修正系数 δ，并获取预测向量。

特征词向量经由调节系数转化为预测向量后，对生成的预测向量进行 *t*-SNE 降维与 FCM 聚类。FCM 的聚类效果受人为设定的模糊加权指数与聚类数量影响，有研究表明[1.5,2.5]是模糊加权指数取值的最佳区间，2 是可以取得较好聚类效果的理想取值[36]。因此，将模糊加权指数设定为 2。在确定最优聚类数量时，为降低主观性判断的影响，引入评价聚类效果的量化指标模糊划分系数 fpc（fuzzy partition coefficient）[37]，当 fpc 值趋于 1 时，对应的模糊聚类数的划分效果最佳。为防止过拟合，本文将聚类中心数阈值设置为[5,50]，对应的 fpc 值如图 6 所示。当模糊聚类数为 34 时，fpc 值最趋近于 1。因此，将预测向量聚类数量确定为 34，使用 FCM 聚类算法进行预测主题聚类，并计算各特征词与主题对应的隶属度。

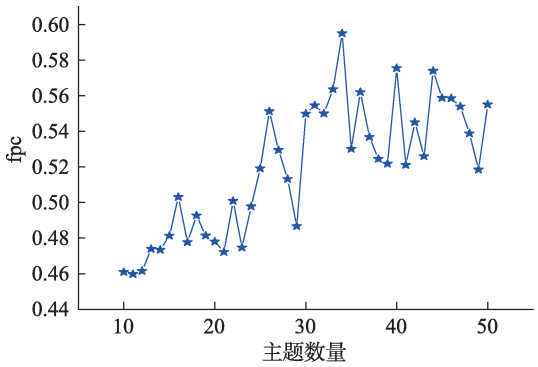

图 6　fpc 数值-主题数量关系

3.3　主题预测结果验证

3.3.1　预测主题选取与映射模式分析

本文的预测主题 PT 是由原始主题 T 对应的特征词群拆分与重组而来，而领域内研究是长期且持续的，并非所有主题均会在短期内发生突变。对比预测主题与原始主题，将发生显著变化的预测主题视为新生科学主题，因此，首先需要探寻哪些预测主题由多个含义不同的原始主题聚合而来。选用预测主题与原始主题的特征词群重叠程度来衡量预测主题的变化程度，重叠是指词汇完全匹配或词义相同，重叠度是重叠词数与主题内总词数的比值。依据实验结果设定主题特征词重叠度的阈值为 0.7，将重叠度大于 0.7 的预测主题定义为变化不显著的预测主题；重叠度小于 0.7 的预测主题视为由多个原始主题拆分后聚合构成，定义为变化显著的预测主题。此外，存在部分预测主题仅包含少量特征词，与原始主题重叠程度较高且意义不明。因此，将特征词对应各预测主题隶属度进行降序排列，保留隶属度最大的前两个主题，对特征词较少的预测主题进行特征词扩展，并纳入后续对比验证。通过上述环节来发掘变化显著的预测主题。

为验证预测主题的准确性，需要与实际主题进行对比分析。使用 LDA 模型基于测试集数据进行主题建模，计算困惑度并结合可视化主题分布确定实际主题的数量为 30，最终获取由 573 个特征词集合构建的 30 个实际主题。原始主题与预测主题的特征词集合相同，而实际主题的特征词集合与前者存在交叉关系，其由部分原始主题中的特征词与部分未曾出现过的新词共建而成。

观察分析实验结果中变化显著的预测主题与原始主题、实际主题间存在的映射关系，可以提炼出如图 7 所示的两种模式。模式一是特征词直接聚合，主要关注原始主题、预测主题与实际主题三者中共现的词——将分别来源于不同原始主题 T1、T2 的特征词 W_a、W_b 通过主题预测流程聚合到同一个主题 PT1 中，且与实际主题 RT1 中出现的特征词相符。模式二是特征词概念集成，在提取由不同原始主题聚合到同一预测主题中特征词的基础上，探寻实际主题中的新词与原始旧词的词义关联——分别来源于原始主题 T1、T2 且直接被聚合到预测主题 PT1 中的特征词 W_a、W_b 与实际主题 RT1 中的特征词 W_c 不同，且 W_c 未曾出现在原始特征词集合中，是新特征词。尝试通过将原始主题的释义进行拆分重组来解析预测主题。为更深刻地理解由原始主题聚合而来的预测主题的含义，构建融合情景/背景、对象/主体、方法/途径、本质/原理 4 个维度的主题内容解构框架，对预测主题 PT1 的内容进行分解，同时剖析 W_c 的概念，可发现其与 W_a、W_b 集成并融合特定情境后的含义相差无几。预测主题与实际主题对比验证将围绕这两种模式展开。

3.3.2　主题预测结果与实际主题对比分析

通过 3.3.1 节预测主题的选取，可得到部分变化显著的候选预测主题，进行定性分析后筛选出预测主题 PT-5、PT-20、PT-3 与 PT-7，将其与原始主题关联后进行具体解读，并与实际主题进行对比分析。

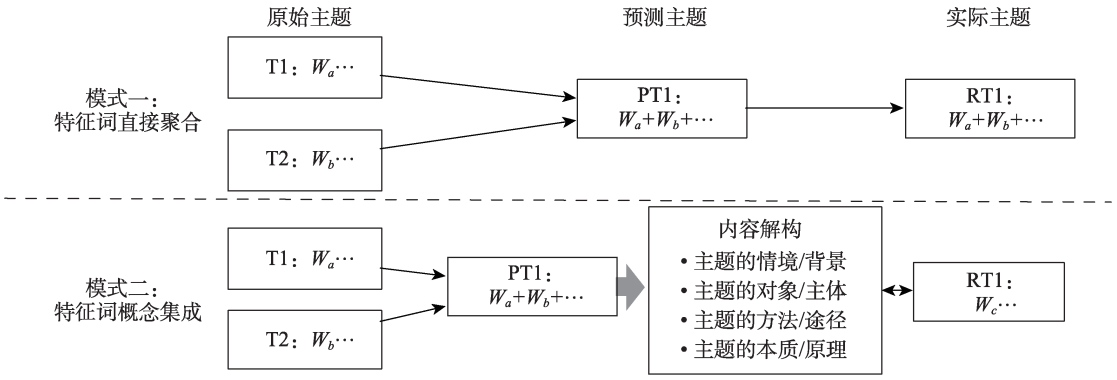

图 7 原始主题-预测主题-实际主题的映射模式

（1）预测主题 PT-5——智库相关主题预测

在主题映射模式一中，隶属于原始主题 T-11 的“科技成果转化”与隶属于 T-14 的“知识生产”被一同聚合到预测主题 PT-5 中，其与实际主题 RT-9 相符，即 T-11（科技成果转化）+T-14（知识生产）= 部分 PT-5=部分 RT-9。在主题映射模式二中，预测主题 PT-5 的特征词与原始主题特征词重叠情况如表 1 所示，其由 T-1（领域科技知识组织与管理）、T-17（嵌入式知识服务模式下图书馆的知识转移）、T-14（新环境下图书馆知识生产与服务）、T-21（传统高校图书馆的知识管理）与 T-11（知识共享知识转移与知识服务）等原始主题的特征词群拆分聚类生成。参考上述原始主题自身含义，着重关注重叠特征词的释义，按照模式二中主题内容分解的 4 个维度厘清 PT-5 的主题内容，可知，预测主题 PT-5 的情境/背景是“大数据时代、竞争环境、领域知识”等，对象/主体是“图书馆与高校”，方法/途径是“知识生产、知识分享与虚拟参考咨询”，本质/原理是“知识管理与知识服务”。由此可以将预测主题 PT-5 理解为“在大数据时代与竞争情境下，面向某一特定领域，通过知识生产、知识分享等途径提供有指向性的新型知识服务”。结合时代背景与领域发展，发现 PT-5 的主题释义与智库的内涵较为相符，尤其是高校智库与图书馆智库：聚焦于某一专业领域、利用领域背景知识从专业客观的角度辅助决策[38]。模式二得出的结论与模式一的结果互为佐证，实际主题 RT-9 的主题释义正是“图书馆与高校智库服务”。

表 1 预测主题 PT-5 与原始主题的重叠特征词

预测主题	对应原始主题	重叠特征词（部分）
PT-5	T-1	知识分享|知识创造模型|知识市场
	T-17	图书馆知识转移|竞争环境|环境因素
	T-14	知识生产|虚拟参考咨询|信息共享
	T-21	高校图书馆知识管理|知识构建
	T-11/T-26 等	科技成果转化|大数据时代

综合参考模式一与模式二，可将预测主题 PT-5 与原始主题、实际主题的关系进行如图 8 所示的可视化。在图 8 中，小圆圈表示特征词，左侧原始主题中的实线圆圈表示被聚合到预测主题中的特征词，虚线圆圈表示未被聚合到预测主题中的特征词，在此处用于为实线圆圈特征词提供相关语境信息；右侧实际主题中的白色圆圈表示来源于原始特征词集合中的旧词，灰色圆圈则表示过去时段中未曾出现过的新词。与预测主题 PT-5 对应的是实际主题 RT-9（图书馆智库|行业知识服务|知识生产）与 RT-19（高校智库|新型智库|智库服务），分别表示图书馆智库与高校智库的知识服务创新。

2015 年，《关于加强中国特色新型智库建设的意见》等规划实施后，智库相关研究逐渐兴起，深耕于智库研究的期刊《智库理论与实践》于 2016 年正式创刊。由此认为，智库相关研究起始于 2016 年，属于新生主题，将智库相关的特征词视为新词。当囊括资源建设、参考咨询与学科服务等业务的图书馆与高校等机构处于大数据时代浪潮中时，辅以新兴技术、顺应领域发展趋势、进行知识生产与知识服务革新后，即可衍生转化为专职知识生产组织的专业智库，并作为沟通产学研多方主体的桥梁，为科技成果转化等知识活动赋能。综上，认为主题 PT-5 的预测结果准确可信。

（2）预测主题 PT-20——数字人文相关主题预测

在实际主题中，预测主题 PT-20 没有与之对应的多个共现特征词，不符合主题映射模式一。在模式二中，预测主题 PT-20 与原始主题重叠情况如表 2 所示，其由 T-10（智能化知识组织与知识检索）、

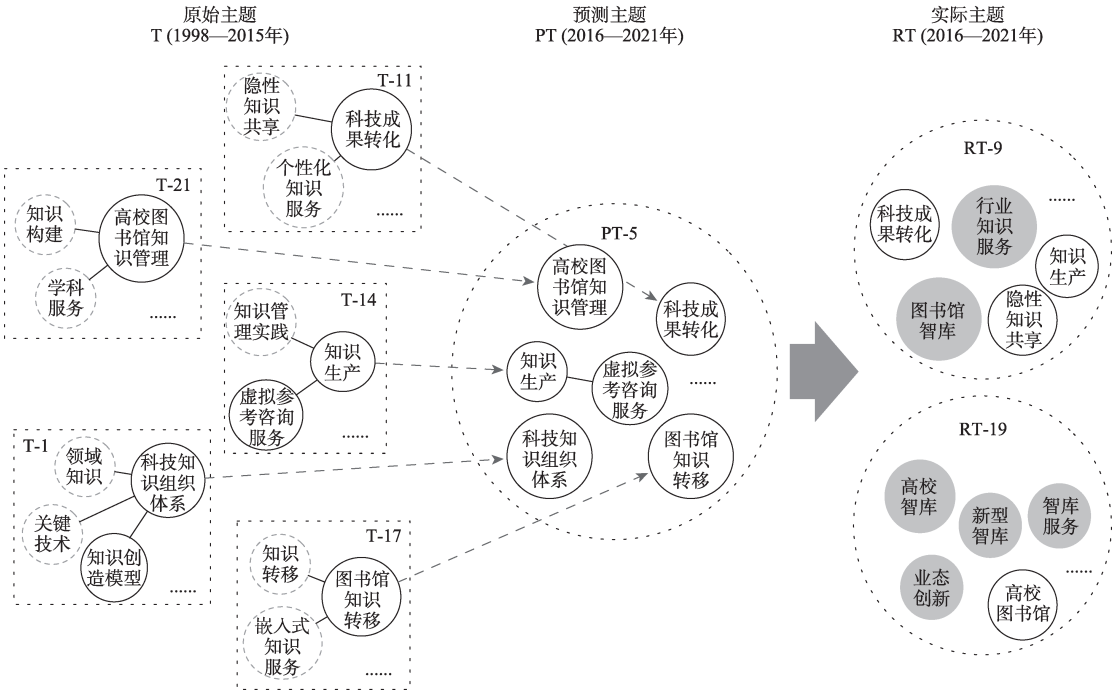

图 8　预测主题 PT-5 的对比验证

T-5（网络环境下学科服务创新）、T-33（图书馆智慧转型）等原始主题的特征词群拆分而来。按照模式二中 4 个维度分解 PT20 的主题内容，其情境/背景是“全媒体、竞争优势”等，对象/主体是“图书馆、开放知识”，方法/途径是“知识检索、知识表示、知识本体、语义关联”，本质/原理是“智慧服务、学科服务”。将预测主题 PT-20 理解为“在全媒体环境下图书馆通过语义关联、知识表示、知识本体等知识组织技术，面向开放知识提供的知识检索与展示等新型智慧服务”。预测主题 PT-20 的释义与数字人文类研究内容较为相符。

表 2　预测主题 PT-20 与原始主题的重叠特征词

预测主题	对应原始主题	重叠特征词（部分）
PT-20	T-10	知识检索|开放知识|知识表示|知识本体|智慧服务
	T-5	语义关联|学科服务|全媒体|高校图书馆知识服务
	T-33/T-1 等	智慧图书馆|领域知识|知识聚合

文中的数字人文含义与普遍认知中的“将网络信息技术与传统人文学科相融合的交叉领域研究”略有不同。图情领域中常见的数字人文研究通常围绕古籍知识建模与可视化或古诗词领域智能化抽取等多方面展开[39-40]，在表面上似乎侧重于人文语料的开发与数字方法的应用这两处。但从“知识管理-知识组织-知识服务”的领域视角来看，最初国内数字人文的概念被引入并应用到实践中时，是以数字图书馆的智慧转型为驱动的，多数研究以知识组织与知识服务为最终落脚点，研究本质仍是基于海量的开放数据与特色资源，借助新兴的知识表示、知识组织与可视化技术，提供全媒体形式的新型智慧服务[41]。将预测主题 PT-20 与原始主题、实际主题的关系进行如图 9 所示的可视化，与 PT-20 对应的实际主题是 RT-28，代表数字人文与知识组织相关的研究。国内的数字人文研究与有关项目兴起于 2015 年后，与之相关的数字人文、古籍等词属新词范畴。虽无法预估出能体现人文特质的古籍之类的人文对象，但能够通过原始主题分解重组的方式预测出与数字人文内涵相符的新生主题。因此，认为主题 PT-20 的预测结果准确、可信。

（3）预测主题 PT-3——知识付费相关主题预测

在实际主题中，预测主题 PT-3 存在与之对应的多个共现特征词，但均来源于同一原始主题，不符合主题映射模式一。在模式二中，预测主题 PT-3 与原始主题特征词重叠情况如表 3 所示，其由 T-28（数字出版与期刊知识服务）、T-25（资源整合与知识服务）、T-5（网络环境下学科服务创新）、T-22（知识经济背景下的知识管理）、T-32（面向知识社会的知识信息服务）等原始主题的特征词群拆分而来。将预测主题 PT-3 按照模式二中内容解构的 4 个维度进行分解可知，该主题的情境/背景是“网络

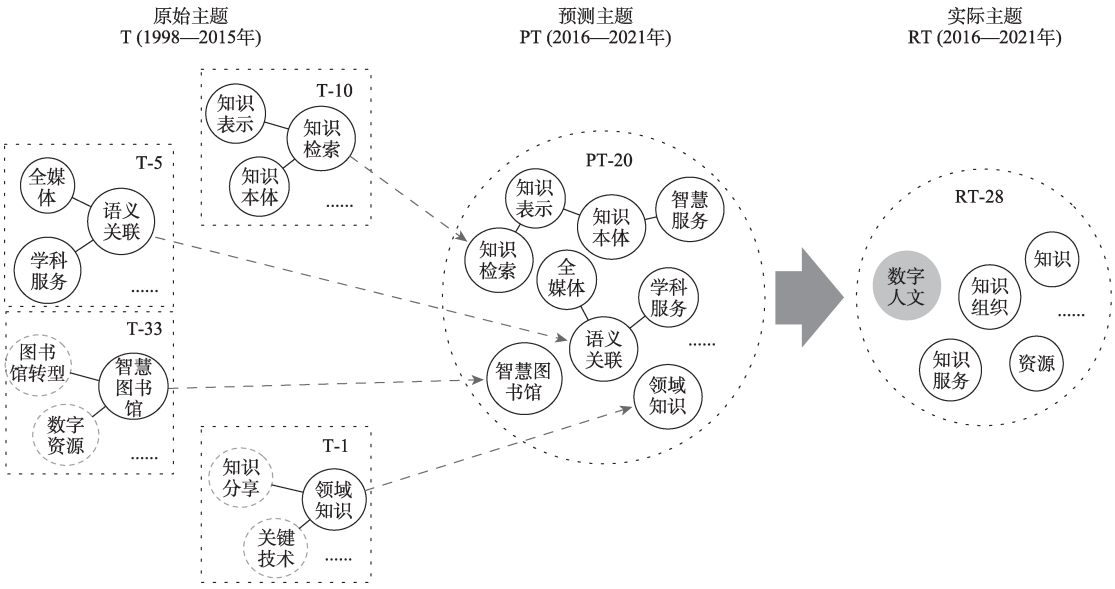

图9 预测主题PT-20的对比验证

环境、云环境、泛在知识环境、知识战略”，对象/主体是“数字出版、期刊”，方法/途径是“信息资源整合、知识服务平台、知识服务模式”，本质/原理是“知识服务”。因而，将预测主题PT-3理解为“数字出版行业在云服务环境与知识经济兴起的大背景下，以知识战略为指导，融合新兴技术，通过知识服务平台与知识服务模式等方面的革新转向新型出版知识服务”，这与知识付费的内涵不谋而合。

表3 预测主题PT-3与原始主题的重叠特征词

预测主题	对应原始主题	重叠特征词(部分)
PT-3	T-28	期刊|数字出版|知识服务
	T-25	信息资源整合|知识服务平台|知识网格
	T-5	知识服务模式|泛在知识环境|网络环境
	T-22	云环境|知识融合
	T-32	知识战略|公共知识

可将预测主题PT-3与原始主题、实际主题的关系进行如图10所示的可视化，与PT-3对应的实际主题是RT-21与RT-24，均与知识付费相关，RT-24侧重于数字出版（期刊）向知识付费的行业调整，RT-21侧重于知识付费商业模式的运营。2016年是知识付费元年，知识付费平台的兴起及服务模式引发了诸多学者的关注。知识付费虽是新词，但其雏形古已有之，传统出版与咨询行业自身便是一种古老的知识付费，本质上均是对知识内容进行整合加工并通过知识与价值的互换提供优质知识服务[42]，只是置身于大数据浪潮与知识经济社会中的知识服务行为被冠以与领域更为适配、更符合时代特征的全新称谓。知识付费的研究内容涉及影响因素、平台运营模式以及网络生态。本文仅考虑在“知识管理-知识组织-知识服务”领域下的知识付费，在广义上可指基于多种媒体技术融合，集知识的生产与传播、转移与扩散于一体的新型经济模式，旨在满足用户知识需求[43]；若将概念主体限定至数字出版，可指数字出版行业与期刊知识服务在新环境下面临革新时衍生出的新型商业模式与转型途径[42]。综上，认为主题PT-3的预测结果准确、可信。

（4）预测主题PT-7——学科服务相关主题预测

在实际主题中，存在与PT-7含义相近的多个共现特征词，如学科（知识）服务与数字图书馆等特征词被聚合到一起，T-12（学科知识服务）+T-34（数字图书馆）=RT-8（学科服务+数字图书馆），符合主题映射模式一。在主题映射模式二中，预测主题PT-7与原始主题特征词的重叠情况如表4所示，其由T-15、T-12、T-8、T-5与T-34等原始主题的特征词群拆分而来。其中T-15对应文档较多，是涵盖图书馆与企业知识管理与知识服务的宏观主题；T-12表示信息化时代学科知识服务；T-8表示学科领域热点挖掘；T-5表示网络环境下的学科服务。将预测主题PT-7分解至4个维度：该主题的情境/背景是“信息化”，对象/主体是“高校图书馆、学科馆员、学科服务”，方法/途径是“数据挖掘、知识图谱”，本质/原理是“知识服务、知识创新”等。因而，可以将预测主题PT-7理解为“在大数据与信息化时代背景下，学科馆员以用户需求为导向，参考

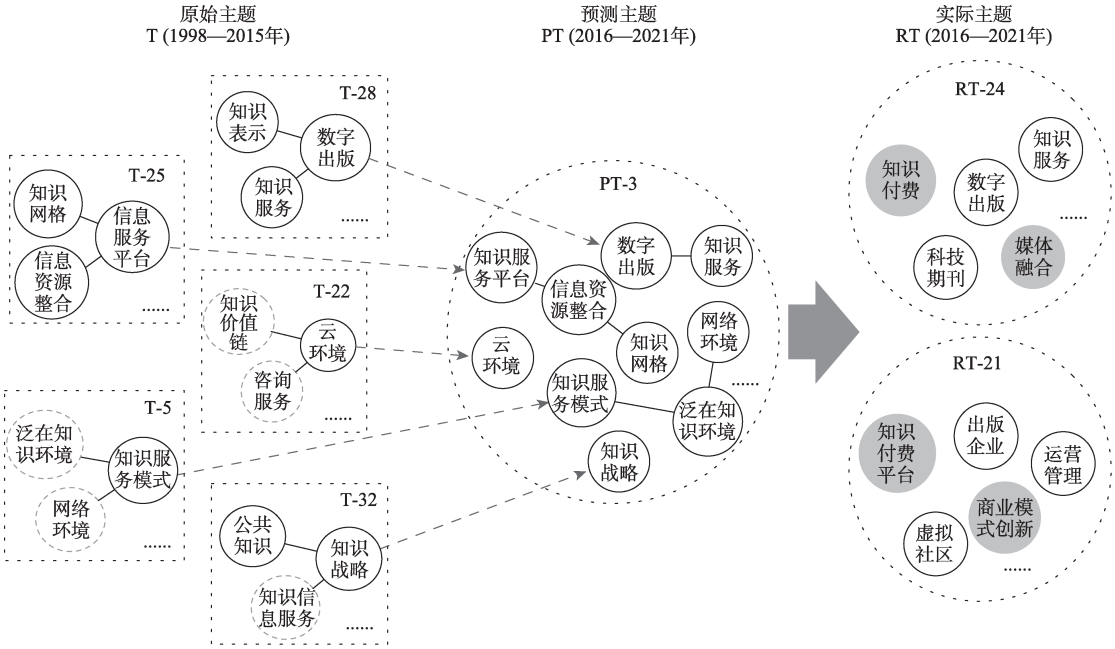

图 10　预测主题 PT-3 的对比验证

领域主题演化与识别的研究范式，使用数据挖掘等新兴技术，用于提升知识转化与知识创新效率，实现学科知识服务的智慧转型”。

表 4　预测主题 PT-7 与原始主题的重叠特征词

预测主题	对应原始主题	重叠特征词(部分)
PT-7	T-15	知识服务|知识转移|知识共享|隐性知识|知识创新
	T-12	学科知识服务|学科馆员|用户需求|数据挖掘
	T-8	学科|研究热点|知识图谱|文献
	T-5	高校图书馆|用户

将预测主题 PT-7 与原始主题、实际主题的关系进行可视化，如图 11 所示，与之对应的是实际主题 RT-8（人工智能技术与数据驱动下的学科服务）。其中，数据驱动与人工智能等词不属于已有特征词集合。2015 年后，数据驱动逐渐被应用于知识管理与知识服务领域，是指使用计算机技术采集海量数据，并在数据基础上通过组织、整合等流程生成自动化知识；人工智能始于 20 世纪 50 年代，直至 2015 年后人工智能方法的效度与准度才取得较大进展，并逐渐被应用于知识组织与知识管理的实践中。因此，可将数据驱动与人工智能视为领域内新词。图书馆服务随时代变迁几经更迭，由资源导向转向用户导向，由文献服务转向知识服务[44]。相应地，在当前人工智能背景下提出的智慧图书馆理念同样对学科服务提出了由信息化转向知识化、智慧化的全新要求[45]，需要依托数字图书馆的海量数字资源，在已有学科服务的基础上以用户知识需求为导向、融入领域热点主题分析方法与知识组织技术，用于了解领域演化历程、掌握领域研究趋势，从而提供主动化嵌入式的学科参考咨询服务、学科知识服务乃至学科情报服务，用于辅助领域教育与科研。由此认为，主题 PT-7 的预测结果较为准确。

3.3.3　讨论与总结

在面向“知识管理-知识组织-知识服务”领域的实证分析中，使用本文所提出的主题预测方法可以成功研判面向图书馆服务与出版行业的创新与转型。大数据时代的到来，以及人工智能的兴起与知识经济的发展，为“知识管理-知识组织-知识服务”领域研究提供了海量的数据、智能的技术环境与全新的研究理念，研究内容在宏观上朝向智能化与智慧化演进。由于领域内早期知识管理理论方面的研究已趋于成熟，本文预测出的新生主题侧重于知识组织方式与知识服务模式的实践应用创新，这与部分面向该领域的计量实证分析结论相符[46]。

与现有的其他科学主题预测方法相比，本文聚焦于新生科学主题的预测，重点关注未来可能会出现的、变化显著的新生主题。其中，预测主题 PT-5（图书馆与高校新型智库服务）、PT-20（图书馆数字人文服务）、PT-3（知识付费与数字出版行业转型）中出现了与智库、数字人文和知识付费这些在

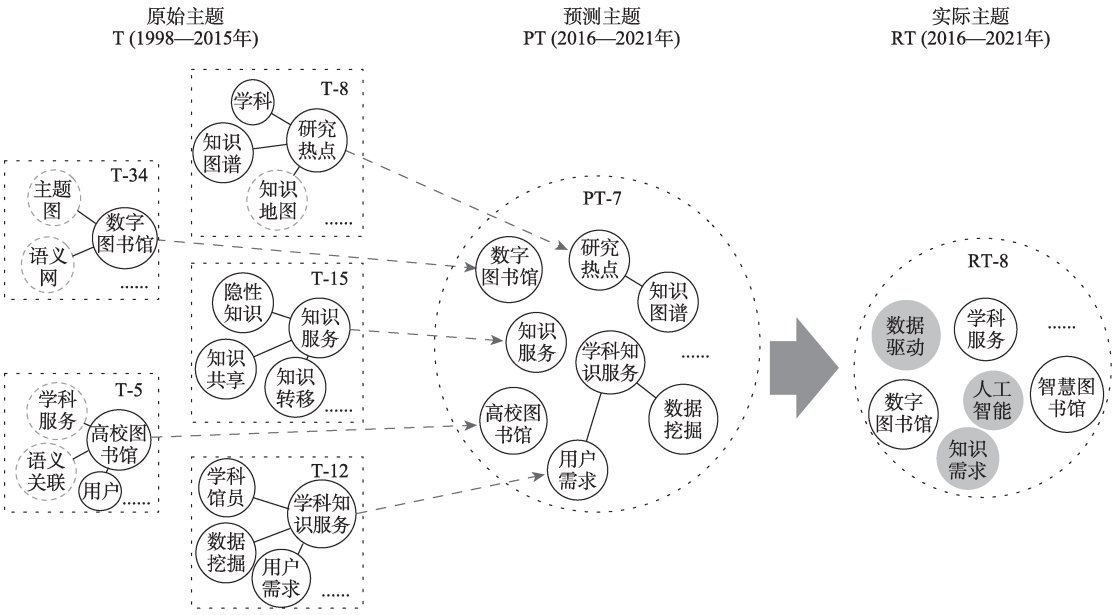

图 11 预测主题 PT-7 的对比验证

过去时段未曾出现的、频次极低的新词的内涵较为相符的主题释义，即通过将原始主题解构与重组，生成了新的领域概念。在实际研究中，学者们融合主观认知、客观情景与实际需求后，为这些新科学概念赋予全新的称谓。PT-7 则不同于上述预测主题，与之映射的原始主题和实际主题均围绕学科服务展开，这是因为该类主题是基于原始主题进行的理念与技术的革新，是属于实践模式的转变，因此，未衍生出领域中的新概念与新名词。

部分预测主题并无显著变化，如预测主题 PT-15 与原始主题 T-19 相互关联，T-19（供应链|客户知识管理|协同知识管理|出版企业）代表与“供应链与企业知识管理”相关的研究，预测主题 PT-15 与原始主题 T-19 重叠度为 0.83，可认为预测主题 PT-15 释义基本等同于原始主题 T-19，即该主题研究内容在 F 时段不会发生较大变革。与之对应的是实际主题 RT-17（科技出版|供应链|中小企业|产业价值链|知识价值链），其代表了与“企业价值链管理”相关的研究，而供应链以价值链的连接为基础，知识价值链是知识链与价值链耦合的产物，其本质均是企业的知识转移与知识管理。因此，认为 RT-17 与 PT-15 相符。这验证了预测主题 PT-15 的准确性，但 PT-15 由原始主题 T-19 平稳演化而来，并未发生显著变化。观察该类主题可知，通常情况下，由侧重于本质与原理、缺少表征应用与实践的特征词构成的原始主题内部常联结紧密，在拆分与聚类时较少会发生改变；而聚类后生成的预测主题特征词群中，若缺少能够催生变革的背景、情景或具体的方法和途径，主题同样不会发生显著变化。

此外需要注意：①本文是在限定“知识管理-知识组织-知识服务”领域的前提下，开展的特征词概念剖析以及原始主题-预测主题-实际主题的对比分析，相同的名词在不同领域中具有不同角度的解释。例如，知识分享在企业知识管理、社会学研究与图书情报学研究中扮演多重角色，知识付费在电子商务、行为学研究等领域下的侧重存异，只有限定了概念适配的领域才能更好地将预测主题与实际主题进行映射与关联。②使用本文提出的科学主题预测方法所获取的结果是基于已有研究进展推理演绎出的新内容，而不是直接预测出未来的某种对象或载体。例如，上文中的主题 PT-20 可以预测出数字人文研究的基本方法与模式，但无法预测出具体研究中应用的古籍、唐诗或家谱文本，需要依据现实中的资源优势、研究者的思维偏好与具体情景需求而定。

4 结 语

本文基于知识单元重组视角，将主题-特征词的表征关系类比为科学概念-知识单元的表征关系，提出了面向新生科学主题的预测方法，构建了相应的预测流程。以“知识管理-知识组织-知识服务”领域为例，使用 1998—2015 年的领域文献有效预测出了 2016—2021 年将出现的数字人文、知识付费等新词的内涵，分析了在图书馆智慧服务、出版行业

转型等实践方面的新生研究主题，总结了“知识管理-知识组织-知识服务”领域在大数据、人工智能与知识经济蓬勃发展背景下的发展方向，同时探讨了部分预测主题变化不显著的内在成因，综合验证了提出的科学主题预测方法流程的可行性与可靠性。

本文尚存在部分局限：仅选用领域科技文献数据进行预测。实际上除领域文献自身发展具有的惯性外，时代背景、国家战略需求、指导性政策文件以及领域意见领袖等均会对研究主题产生影响，如在全球疫情肆虐的大环境下涌现的以突发公共卫生事件、应急管理与危机管理为主题的研究成果。在后续研究中将尝试引入影响领域发展的多重因素，融合多源数据来进一步提升科学主题预测的精准程度。

参 考 文 献

[1] 陈玉祥, 朱桂龙, 陈德棉 . 科学发展预测的概念和功能[J]. 预测, 1994, 13(1): 57-61.

[2] 陈德棉, 潘皖印, 毛家杰 . 科学预测和技术预测的方法研究[J]. 科学学研究, 1997(4): 56-62.

[3] 赵红洲, 蒋国华 . 知识单元与指数规律[J]. 科学学与科学技术管理, 1984, 5(9): 39-41.

[4] Swanson D R. Undiscovered public knowledge[J]. The Library Quarterly, 1986, 56(2): 103-118.

[5] 王子舟, 王碧滢 . 知识的基本组分——文献单元和知识单元[J]. 中国图书馆学报, 2003, 29(1): 5-11.

[6] Rotolo D, Hicks D, Martin B R. What is an emerging technology? [J]. Research Policy, 2015, 44(10): 1827-1843.

[7] 白敬毅, 颜端武, 陈琼 . 基于主题模型和曲线拟合的新兴主题趋势预测研究[J]. 情报理论与实践, 2020, 43(7): 130-136, 193.

[8] Li Y T, Chen Y, Wang Q Y. Evolution and diffusion of information literacy topics[J]. Scientometrics, 2021, 126(5): 4195-4224.

[9] 王康, 陈悦, 苏成, 等 . 多维视角下科学主题演化分析框架[J]. 情报学报, 2021, 40(3): 297-307.

[10] Huang L, Chen X, Zhang Y, et al. Identification of topic evolution: network analytics with piecewise linear representation and word embedding[J]. Scientometrics, 2022, 127(9): 5353-5383.

[11] Mryglod O, Holovatch Y, Kenna R, et al. Quantifying the evolution of a scientific topic: reaction of the academic community to the Chornobyl disaster[J]. Scientometrics, 2016, 106(3): 1151-1166.

[12] 马费成, 陈潇俊, 刘向 . 基于科学知识图谱分析的知识演化研究—以生物医学为例[J]. 情报科学, 2012, 30(1): 1-7, 15.

[13] Li P, Yang G L, Wang C Q. Visual topical analysis of library and information science[J]. Scientometrics, 2019, 121(3): 1753-1791.

[14] 王伟, 杨建林 . 基于引文网络重叠社团发现的图书情报领域学科主题结构分析[J]. 情报学报, 2020, 39(10): 1021-1033.

[15] 王曰芬, 傅柱, 陈必坤 . 基于LDA主题模型的科学文献主题识别: 全局和学科两个视角的对比分析[J]. 情报理论与实践, 2016, 39(7): 121-126, 101.

[16] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[17] 李湘东, 张娇, 袁满 . 基于LDA模型的科技期刊主题演化研究[J]. 情报杂志, 2014, 33(7): 115-121.

[18] 赵新琴, 吴鹏 . 基于TDT技术的新冠肺炎疫情文献主题演化研究[J]. 科技情报研究, 2022, 4(2): 49-60.

[19] Figuerola C G, Marco F J G, Pinto M. Mapping the evolution of library and information science (1978-2014) using topic modeling on LISA[J]. Scientometrics, 2017, 112(3): 1507-1535.

[20] Wang Z B, Ma L, Zhang Y Q. A hybrid document feature extraction method using latent Dirichlet allocation and word2vec[C]// Proceedings of the 2016 IEEE First International Conference on Data Science in Cyberspace. Piscataway: IEEE, 2016: 98-103.

[21] Liu Y, Liu Z Y, Chua T S, et al. Topical word embeddings[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 2418-2424.

[22] 陈茫, 张庆普 . 我国知识服务研究的演进历程知识图谱与研究态势探讨[J]. 情报资料工作, 2018(2): 80-91.

[23] 隗玲, 许海云, 胡正银, 等 . 学科主题演化路径的多模式识别与预测——一个情报学学科主题演化案例[J]. 图书情报工作, 2016, 60(13): 71-81.

[24] Chen W, Lin C R, Li C Y, et al. Tracing the evolution of 3-D printing technology in China using LDA-based patent abstract mining[J]. IEEE Transactions on Engineering Management, 2022, 69(4): 1135-1145.

[25] 岳丽欣, 周晓英, 陈旖旎 . 基于ARIMA模型的信息构建研究主题趋势预测研究[J]. 图书情报知识, 2019(5): 54-63, 72.

[26] 朱光, 刘蕾, 李凤景 . 基于LDA和LSTM模型的研究主题关联与预测研究——以隐私研究为例[J]. 现代情报, 2020, 40(8): 38-50.

[27] 霍朝光, 霍帆帆, 董克 . 基于LSTM神经网络的学科主题热度预测模型[J]. 图书情报知识, 2021(2): 25-34.

[28] 董克 . 预见学科之美: 学科主题预测研究[J]. 图书情报知识, 2021(2): 封二 .

[29] 赵一鸣, 张进, 黎苑楚 . 基于多维尺度模型的潜在主题可视化研究[J]. 情报学报, 2014, 33(1): 45-54.

[30] 奉国和, 孔泳欣, 肖洁琼 . 基于加权关键词的领域热点与趋势分析新方法[J]. 图书情报工作, 2018, 62(18): 102-109.

[31] der Maaten L, Hinton G E. Visualizing data using *t*-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

[32] 陈挺, 李国鹏, 王小梅 . 基于t-SNE降维的科学基金资助项目可视化方法研究[J]. 数据分析与知识发现, 2018, 2(8): 1-9.

[33] Chang I C, Yu T K, Chang Y J, et al. Applying text mining, clustering analysis, and latent Dirichlet allocation techniques for topic classification of environmental education journals[J]. Sustain-

ability, 2021, 13(19): Article No.10856.

[34] 朱晓峰, 葛锐, 盛天祺. 四十年来我国情报学研究的学术变迁与学理支撑[J]. 科技情报研究, 2022, 4(2): 1-14.

[35] 朱晓峰, 蒋旭牧, 张卫. 领域知识组织研究的历史演化与未来展望[J]. 情报资料工作, 2021, 42(5): 23-31.

[36] Pal N R, Bezdek J C. On cluster validity for the fuzzy c-means model[J]. IEEE Transactions on Fuzzy Systems, 1995, 3(3): 370-379.

[37] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques[J]. Journal of Intelligent Information Systems, 2001, 17(2): 107-145.

[38] 初景利, 栾瑞英, 孔媛. 国外高水平高校智库运行机制特征剖析[J]. 图书馆论坛, 2018, 38(4): 8-16.

[39] 张琪, 王东波, 黄水清, 等. 史书多维知识重组与可视化研究——以《史记》为对象[J]. 情报学报, 2022, 41(2): 130-141.

[40] 张卫, 王昊, 邓三鸿, 等. 面向数字人文的古诗文本情感术语抽取与应用研究[J]. 中国图书馆学报, 2021, 47(4): 113-131.

[41] 夏翠娟, 张磊. 关联数据在家谱数字人文服务中的应用[J]. 图书馆杂志, 2016, 35(10): 26-34.

[42] 梁徐静. 数字出版与知识付费[M]. 广州: 中山大学出版社, 2020: 168-170.

[43] 郭宇, 郭勇, 刘文晴, 等. 国内互联网知识付费研究现状与发展趋势[J]. 图书情报工作, 2021, 65(24): 100-108.

[44] 初景利, 张冬荣. 第二代学科馆员与学科化服务[J]. 图书情报工作, 2008, 52(2): 6-10, 68.

[45] 张晓林. 颠覆性变革与后图书馆时代——推动知识服务的供给侧结构性改革[J]. 中国图书馆学报, 2018, 44(1): 4-16.

[46] 石玉玲, 陈万明. 我国知识管理研究现状、热点与趋势[J]. 新世纪图书馆, 2020(4): 85-91.

（责任编辑 冯家琪）