● 李秀霞[1]，邵作运[2]
（1. 曲阜师范大学传媒学院，山东 日照 276826；2. 曲阜师范大学图书馆，山东 日照 276826）

基于离群主题词跨学科组合的学术创新机会发现研究*

摘 要：［目的/意义］学科领域的离群主题词可为创新机会发现提供新颖、稀缺的信息，离群主题词跨学科组合能催生新的知识，产生突破性学术创新机会。［方法/过程］以情报学和政治学为例，利用LDA提取不同学科文献的主题，以概率分布低的主题词为数据对象，利用Word2Vec和PCA技术将题名和摘要中包含文本语义的主题词表示为低维稠密向量，根据主题词在二维空间的分布发现学科内的离群主题词；利用余弦相似度计算不同学科离群主题词之间的语义相似度，将相似度高的不同学科的离群主题词组合视为具有创新潜能的组合。根据设计的需求度指标进一步筛选离群主题词组合，最终确定未来具有研究潜力的学术创新机会。［结果/结论］将主题提取与语义分析相结合，充分考虑了离群主题词的价值和语义环境；将离群主题词跨学科组合的语义相似度与需求度结合，能够兼顾学术创新的新颖性和有用性特征。研究表明，此研究方法能够有效发现学术创新机会，为科研指导、知识服务提供可靠参考。

关键词：离群主题词；语义相似度；需求度；创新机会发现

DOI：10.16353/j.cnki.1000-7490.2023.12.016

引用格式：李秀霞，邵作运．基于离群主题词跨学科组合的学术创新机会发现研究［J］．情报理论与实践，2023，46（12）：122-130.

Research on the Discovery of Academic Innovation Opportunity Based on the Interdisciplinary Combination of Outlier Topic Words

Abstract：［Purpose/significance］Outlier topic words play a crucial role in providing fresh insights for the exploration of innovative opportunities. By combining outlier topic words from different disciplines, the potential for generating groundbreaking academic innovations can be unlocked. ［Method/process］In this study, Information Science and Political Science are taken as examples. The first step involves applying LDA to extract topics from literature across various fields. Utilizing Word2Vec, topic words are then transformed into low-dimensional dense vectors that capture their semantic meanings in titles and abstracts. Subsequently, outlier topic words within each discipline are identified based on their distribution in a two-dimensional space. The semantic similarity between outlier topic words from different disciplines is calculated using cosine similarity, aiming to identify combinations of outlier topic words that exhibit high similarity. To gauge the demand for these combinations, an index is devised. ［Result/conclusion］By combining topic extraction with semantic analysis, the value and semantic context of outlier topic terms are fully considered. Integrating the semantic similarity of interdisciplinary combinations of outlier topic terms with their demand level allows for a balance between novelty and utility in academic innovation. Research indicates that the proposed method can effectively discover academic innovation opportunities and provide reliable references for research guidance and knowledge services.

Keywords：outlier topic words；semantic similarity；degree of demand；innovation opportunity discovery

0 引言

学术创新是科研发展的本质要求。在科学技术飞速发展的今天，学术创新已经成为衡量科研机构和个人研究能力、研究成果价值的一个重要因素。而学术创新的起点和根本是发现学术创新机会。学术创新机会的发现对科研人员的创新选题，开展突破性科技研发、科技范式转变、知识革新以及新技术预测等均具有重要战略意义，因而学术创新机会发现成为当前备受关注的话题。

创新机会发现研究已有20多年的发展历史，多数研究是基于文献计量分析、文献内容分析等，分析粒度较粗，缺乏语义特征分析，且研究或通过集群点共有特征发现创新机会，或通过弱链接关系发现创新点，忽视了离群点在创新机会发现中的潜在价值。马费成教授[1]曾指出情报学领域急需解决两个关键性问题：其一是信息知识的组

* 本文为国家社会科学基金一般项目“跨学科知识元迁移组合与学术创新机会发现研究”的成果，项目编号：22BTQ061。

织和表达需要从物理层面的文献单元转换到认知层面的知识单元；其二是信息知识的计量单位需要从语法层次向语义层次和语用层次发展。考虑到离群数据中往往隐藏着新颖、稀缺的信息，这些信息对创新机会发现具有极高的研究价值。因此，本文提出基于离群主题词跨学科组合的学术创新机会发现方法，将主题词视为学科领域的学术研究知识元，从知识元组合创新理论出发，通过学科间离群主题词的语义相似性分析，结合社会需求度分析发现学术创新机会。这种方法考虑了离群主题词的价值和语义环境，能够从内容上细粒度探究创新机会的背景和根源，而且兼顾了创新研究的新颖性和有用性特征。

1 研究现状

目前，学术创新机会的概念尚未有明确的界定，与其相近的概念是技术创新机会。早在 20 世纪 70 年代，S. Peter[2]就提出技术创新机会。20 世纪 90 年代，Zhu Donghua 等[3]将技术机会定义为：通过挖掘领域内已有技术的发展趋势及关系，推断该领域未来可能出现的技术形态或技术发展热点。李保明[4]将技术机会分为内涵的技术机会和外延的技术机会，前者是在现存技术规范或性能改进上发现技术机会，后者是在技术跨领域转移实现新功能上发现技术机会；伊惠芳等[5]认为技术创新的科学知识基础和技术知识基础的载体来源一般为论文或专利，技术创新机会的发现就是挖掘蕴藏在大量专利信息中的有用知识，提炼出潜在的创新机会。归纳学界对技术创新机会的理解，分析其概念内涵，根据技术与学术的同源性[6]，本文认为学术创新机会发现就是通过分析知识需求和现有相关文献，识别具有潜在价值的科学研究和应用机会的过程，是领域专家对创新可能的一种猜想。

1.1 学术创新机会发现方法

1）定性分析法。定性分析法是依靠专家意见发现创新机会，包括专家访谈法、德尔菲法、技术路线图等。主要是通过专家访谈获取专家见解和意见，深入了解专家的知识、经验和见解，发现潜在的创新机会；或由专家团队进行预测和决策，发现和评估未来的创新机会；或通过邀请专家参与制定和评估技术路线图，发现知识空缺和新的创新机会。定性分析法主要依赖专家经验学识进行判断，存在程序复杂、社会成本高、专家意见差异性不易处理等问题。

2）定量分析法。根据对创新机会内涵揭示程度的不同，定量分析法可分为以下 3 类：①基于文献计量发现创新机会。利用计量分析技术，如引用频次分析[7]、引文网络[8]、关联关系[9]等识别前沿研究领域，揭示某一领域或某一主题的研究趋势和模式，发现当前研究的空白和缺口，并预见未来可能出现的新的研究主题和领域。这类方法主要依赖文献的外在关系，不能深度分析创新机会的内涵。②基于文献关键词发现创新机会。通过挖掘和分析已有文献的研究内容，了解当前的研究动态、技术趋势和市场需求，以发现潜在的技术机会或科技创新点。常用的方法是将词簇分析（如关键字分析、共词分析）与因子分析、网络分析、链接预测结合，如依据领域中的关键词构建共词网络，结合链接预测算法预测学术研究机会的可能性[10]；或通过文本挖掘获取代表技术特征的关键字，基于关键字的因子分析构建技术词典，利用技术字典识别技术创新机会[11]。这类方法识别出的创新机会由关键词组合特征来表征，不利于揭示创新机会的细节。③基于文献内容的功能信息发现创新机会。学术创新是为了解决现实问题，实现生活、生产所需的功能是创新研究的目标之一。因此，有学者提出基于功能的技术机会发现框架，根据技术或产品之间的语义功能相似性推导出潜在的技术机会[12]。这种方法关注文本的功能实现，但忽视了信息背景，在创新机会的细节解读中存在一定的局限性。④基于核心词的语义分析发现创新机会。这种方法通过语义分析获取文本中核心词的上下文信息，帮助理解文本的实际含义，获得的信息更准确，表达创新机会的意义更明确。Feng Sijie [13]结合文本挖掘和 Word2Vec 语义分析，通过分析创新要素的内在联系，发现专利技术机会；韩晓彤等[14]通过文本语义相似性关联网络，基于 Louvain 算法进行主题聚类，识别科学研究推动的技术机会。目前，此类方法已成为当前创新机会发现的研究热点，但在跨学科学术创新机会发现中的应用还相对匮乏。

1.2 跨学科创新机会发现研究

学术创新主要来源于知识要素的游离和重组[15]，以创造性的方式整合不同领域的知识尤其有助于新技术的出现，可以促进突破性创新[16]。在跨学科创新机会发现研究中，最重要的方法是通过构建新颖性指标识别跨学科相关知识。其核心思想是根据核心词的新颖性判断跨学科研究的创新点[17]。主题模型作为文本挖掘中的重要技术模型，在创新知识识别研究中也有广泛应用。商宪丽[18]从集成数据集文献中提取交叉学科的潜在主题，并通过交叉学科主题的共现关系发现知识创新过程中知识组合。主题模型法能够帮助发现学科间潜在的知识关联和交叉创新机会，而新颖性指标则提供了对创新机会定量评估的标准，两者结合能够实现更全面的创新机会分析，帮助决策者更好地识别和选择具有潜力的创新方向。如陈虹枢等[19]运用主题模型、词嵌入算法、复杂网络分析等构建动态主题网络，结合突破性创新的新颖性和学科交叉性特征识别突

破性创新主题。此外，李长玲研究团队提出了两种创新机会识别方法：一种是从跨学科的弱连接关系中识别具有高合作潜力的知识组合[20]；另一种是从非相关知识中识别潜在的跨学科合作研究主题[21-22]。

1.3 基于离群点检测的创新机会发现

离群点（Outlier）又称为异常点，是在数据分布中远离其他数据点的点。离群点检测是识别不符合预期模式或数据集的项目、事件、观察结果[23]。分析这些离群点的异常特征，一方面可以降低错误决策的风险，帮助识别和预防不良影响；另一方面可以发现有潜在意义的信息，帮助领域从业者快速定位特殊信息，制定高质量决策。目前，离群点检测在医疗知识识别[24]、气象分析[25]、新兴主题发现[26]、异常情报分析[27]、异常用户预测[28]、网络异常数据挖掘[29]等方面均有应用。

在学术研究领域，离群点可能代表着一些突破性的研究成果或模型，具有创新的研究方法或观点。已有研究表明，非典型的知识组合对科学发展具有更强的整合能力[30]，更有助于解决新出现的技术问题[31]。在早期的研究中，人们主要使用离群文献和离群关键词来发现创新机会。离群文献和离群关键词可能代表着新的研究方向、研究问题、研究方法、理论观点。因此，通过分析文献引用网络中离群程度较高的专利文献，可以发现代表技术创新机会的内容[32]；基于技术关键词的特征向量，结合向量距离计算和离群点检测算法，能够识别离群技术关键词所指向的新领域、新问题、新方法或新观点[33]。随着数据挖掘和自然语言处理技术的发展，主题提取技术逐渐应用于创新机会的发现。应用主题模型提取技术主题，结合离群检测算法评估离群主题，可以有效识别技术机会[34]。此外，词嵌入技术也被应用于创新机会发现的研究中，J. Lee等[35]结合词嵌入技术和离群点检测算法获取离群关键字，根据离群关键词组合发现技术机会，极大地提高了技术机会发现的效果。

1.4 研究现状述评

目前，创新机会发现的研究成果较为丰富，创新机会的表现形式呈多样化。分析已有研究发现，创新机会主要表现为粗粒度的文献组合、细粒度的关键词组合和主题组合。细粒度的知识组合不仅具有知识浓缩性，而且知识表达更具有实质性意义，是表征创新机会简单且有效的方法。跨学科创新机会发现和基于离群点的创新机会发现呈现出相同的发展趋势，即从知识的继承关系（引用关系）和共现关系逐渐转向知识要素（关键词、主题词）的关联关系，越来越重视知识要素的语义分析。

跨学科知识组合能够将不同学科的知识结合起来，产生知识价值的增值，催生新的知识、发现突破性学术创新机会[36]；离群词组合可以突破原有的思维模式，从新的角度看待问题，是一种探索新思维、发现创新机会的重要途径[35]。相对跨学科创新机会发现研究，基于离群点组合的创新机会发现研究相对较少，在国内通过离群点检测发现学术创新机会的研究更少，尤其缺乏通过离群点的跨学科组合来发现学术创新机会的研究。因此，为了充分挖掘学科领域离群主题词的潜在价值，进一步丰富学术创新机会发现的方法，本文提出一种离群主题词跨学科组合的学术创新机会发现方法。通过挖掘不同学科离群主题词的特征信息，分析离群主题词跨学科组合的新颖性和有用性，发现有价值的新的学术创新机会，助力学术研究的发展。

2 离群主题词跨学科组合的学术创新机会发现研究模型

本研究的主要方法是通过LDA模型从不同学科文献的关键词中提取主题词，并利用不同学科文献的关键词训练Word2Vec的skip-gram模型，将提取到的主题词向量化；通过PCA降维后将不同学科主题词的词向量在2维空间呈现，据此发现不同学科的离群主题词；通过语义相似度模型计算不同学科间离群主题词之间的语义相似性，结合需求度计算，筛选学术创新机会。综合上述方法，深入挖掘和理解不同学科之间新颖知识的关联，为学术研究提供新的视角和方向。

构建研究模型如图1所示。研究总体分为以下步骤：学科划分、关键词获取、主题词提取及向量化、离群主题词识别、相似度和需求度计算及学术创新机会发现。

步骤1：收集相关文献数据，并按学科分类号对其进行学科分类。

1）以跨学科知识吸收能力强的期刊论文为研究样本，导出期刊文献及其中文参考文献的题录数据；

2）根据参考文献的学科分类号将参考文献数据集划分为不同的子集，不同子集代表不同的学科领域知识。

步骤2：获取学科领域的关键词。

1）提取不同学科领域论文的文献题名、摘要；

2）使用文本预处理技术清洗文本数据，如删除标点符号、进行分词、停用词处理等，建立学科关键词语料库。

步骤3：主题词提取和主题词向量化。

1）以学科领域关键词为词典，利用LDA模型从步骤2构建的语料库中提取学科领域主题，获取不同学科的主题词；

2）利用不同学科文献的关键词训练Word2Vec的skip-gram模型，将提取到的主题词向量化。

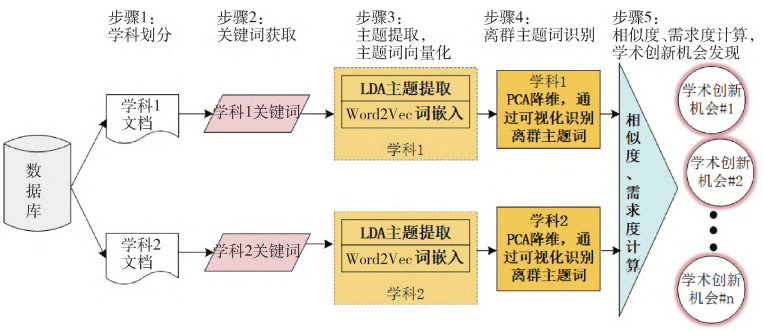

图 1　研究模型

Fig. 1　The research framework

步骤 4：离群主题词识别。

1）利用 PCA 技术将高维主题词向量降维至 2 维，在二维平面呈现主题词；

2）根据主题词的散点分布发现不同学科领域的离群主题词。

步骤 5：相似度分析和需求度分析。

1）利用相似度算法发现学科间相似度高的离群主题词，得到具有创新潜能的离群主题词组合；

2）设计需求度指标，计算离群主题词跨学科组合的社会需求度；

3）分析学术创新机会。

3　实证研究

3.1　数据收集及预处理

情报学是一门受信息技术驱动并广泛服务于社会的交叉学科[37]，有效组织和整合其他学科知识是情报学研究快速增长、不断创新的重要途径。因此，本文选取情报学为目标学科，根据情报学期刊论文参考文献所在期刊的学科分类号确定辅助学科。

选择情报学领域跨学科知识吸收能力较强的 3 种期刊《情报学报》《情报理论与实践》《数据分析与知识发现》为研究样本，选取时间窗口为 2020—2022 年。通过维普期刊的文献导出功能直接获取这 3 种期刊在上述时间窗口刊载的文献及其中文参考文献的题录数据，检索获取这 3 种情报学期刊论文 2084 篇，参考文献 18144 篇。剔除通知、会议纪要、书评等，最终得到情报学论文 2021 篇，本文以此作为知识输入的目标文献。国内情报学的参考文献分类号主要集中在 G 类、TP 类、D 类。除 G 类外，TP 类、D 类是我国情报学知识输入的主要学科。目前，分析 G 类与 TP 类之间知识交流以及其间交叉知识结构的研究已很多[38-39]。考虑到研究结果的实践价值，本文选取政治学（学科分类号为 D0-D8）学科作为向情报学（学科分类号为 G35）输入知识的辅助学科。根据中图分类号，从 18144 篇参考文献中筛选出学科分类号为 D0-D8 的参考文献，并删除中图分类号标识有误的参考文献（如标识为 D0-D8 的参考文献，对应的期刊却是 G 类期刊），最终得到政治学参考文献 549 篇。本文以此作为向情报学输入知识的政治学领域的文献。最后，提取情报学和政治学上述文献的提名和摘要，利用 Python 环境下的 Jieba 库进行分词、利用哈工大停用词表去除停用词，建立两类学科文献的单词序列语料库，作为学科主题提取和语义分析的数据基础。

3.2　学科文献主题词提取

基于自建的情报学、政治学两个学科的词典，利用 LDA 主题模型分别提取两类学科的研究主题。根据 pyLDAvis 呈现的主题可视化结果（各主题之间互不交叉，相互独立，表明主题提取效果较好）确定两类学科最佳的主题数目，最后提取到情报学、政治学各 5 个主题。每个主题内的主题词按概率分布大小排序，高概率的主题词对主题的贡献较大，由于本研究意在获取学科中的离群主题词，考虑到同一个主题词会出现在不同的主题中，因此截取每个主题内的低概率主题词，除去重复的主题词，共获取情报学、政治学两学科各 50 个主题词，作为跨学科学术创新机会发现的主题词。

3.3　学科文献主题词向量化

LDA 模型能够识别主题—类别之间的语义关系，反映了文本的全局隐性信息。然而，LDA 模型是典型的词袋模型，假设词语之间是独立的，忽略了主题词上下文间的局部语义信息。而 Word2Vec 模型能够学习单词的上下文信息，捕捉主题词之间的语义关系，从而更好地理解句子和文本的语义，可有效解决 LDA 主题特征语义信息不足的问题。因此，将两者结合起来，能够增强文本语义的表达，使提取的主题信息更丰富，更接近于特定的文本内容。在本文中，分别将学科领域对应的文献摘要和标题生成的词序列输入 Word2Vec 的 Skip-Gram 模型进行训练。训练参数设置如下：n_dim = 100、window = 5，alpha = 0.025、epochs = 5、sg = 1、min_count = 5。使用训练过的模型将截取的两学科对应的主题词表示为密集词向量，以表征主题词的上下文语义关系。

3.4　学科离群主题词识别

目前，离群点探测方法可归纳为以下 4 种：统计方法、可视化方法、分类方法和聚类方法[40]。其中，最常用的方法是聚类法。聚类方法是根据距离或密度设计聚类模型，将数据集中的样本划分为不同的聚类。在聚类中，正常数据通常归属于某个密度较大、距离较近的聚类，而

距离超过设定的阈值或密度低于设定的阈值的数据点则被视为离群点。这种方法适用于处理大规模、高维数据，其缺点是离群点的识别结果受聚类算法选择和聚类数目设定的影响较大。可视化方法是通过可视化手段将数据映射为易于辨识的图形，从而辅助我们发现与其他数据明显不同的离群点。这种方法的突出优势在于通过观察散点分布，以直观地发现复杂数据中的离群点。但可视化方法对于高维数据的分析存在一定的困难。由于本研究中所使用的数据规模不大，因此选择可视化检测方法。

利用 PCA 将上述 100 维数据降维到 2 维，并在平面上对其进行可视化呈现，如图 2 所示。由于本文所使用的数据量较少，数据点在平面上的位置分布容易识别，每个学科都只有一个核心类团，因此直接根据主题词的散点分布（不需要通过聚类分析）就能发现不同学科的离群主题词，即图 2 中带有方块的点对应的主题词。

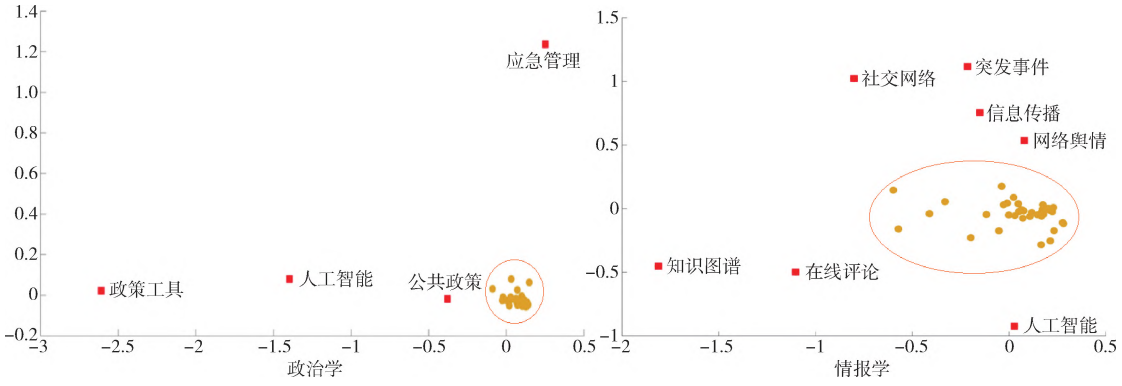

图 2 学科主题词二维语义向量的散点图分布

Fig. 2 The scatter plot distribution of two-dimensional semantic vectors for subject keywords

根据图 2 中不同学科主题词的分布，统计不同学科的离群点。政治学领域的离群点是：政策工具、应急管理、人工智能、公共政策等。由于人工智能是来自计算机科学领域的研究内容，目前已成为情报学领域的主要技术方法之一。本文分析的是政治法律学向情报学的知识输入，人工智能虽然属于政治学学科的离群点，但这里不予考虑。情报学领域对应的离群点有知识图谱、社交网络、突发事件、信息传播、在线评论、网络舆情、人工智能等。

3.5 两学科离群主题词的语义相似度计算

相似度计算是寻找相似点最简单、最有效的方法[41]。主题词语义相似度反映了两个主题词内容的相似程度。不同学科领域离群主题词的语义相似度越高，其组合创新的可能性就越大。本文选用余弦距离来计算学科间离群主题词的语义相似度。余弦相似度计算公式为：

$$\text{similarity}(V_X, V_Y) = \frac{V_X \times V_Y}{\|V_X\| \times \|V_Y\|} \quad (1)$$

式中，V_X、V_Y 表示两个学科对应的某个离群主题词的词向量。利用公式（1）计算两学科离群主题词之间的语义相似度，计算结果见表 1（由于主题词向量有正、负值，因此计算结果出现了正、负数据）。

表 1 两学科离群主题词的语义相似度

Tab. 1 The similarity of outlier topic words in two disciplines

情报学	政治学		
	应急管理	公共政策	政策工具
突发事件	0.902	-0.04	-0.08
在线评论	0.029	0.239	0.152
知识图谱	0.132	0	-0.04
网络舆情	0.457	-0.023	-0.066
信息传播	0.719	-0.143	-0.194
社交网络	0.506	-0.095	-0.161
人工智能	-0.013	0.946	0.889

两个学科离群主题词的相似性反映了学术创新的新颖性和跨学科知识组合的可能性。

离群主题词产生于不同的机制，而非随机偏差，反映了学科领域真实的研究内容，这种离群主题词有着区别于其他集群主题词的思想、观点或方法，能够提供全新的研究视角和解决问题的方式。离群主题词与核心类团的主题词距离越远，这个主题词对应的概念、观点、方法等越有可能具有超乎常规的思维模式。因此，离群主题词组合在内容上具有较高的新颖性。

当两个学科的离群主题词语义相似性较高时，意味着这些跨学科的离群主题词在某种程度上共享一些非传统、非常规的思维模式或研究观点，这种相似性可促使将一个学科中的离群点思想迁移到另一个学科，提高了学术创新的可能性。因此，两个学科的离群主题词语义相似性越高，将两学科之间的新奇思想和方法结合起来产生创新机会的可能性越高。

3.6 两学科离群主题词组合的需求度计算

离群主题词是学科领域内新奇的信息，离群主题词跨学科组合则反映了学术创新内容的新颖性。创新具有新颖性和有用性两个核心特征[42]，因此学术成果的创新价值应由创新成果的有用性来检验[43]。同时，还需要通过合适的评估指标对上述离群主题词组合进行进一步的筛选。

有用性作为学术研究评判的第一准则，由科学研究的目的和科学研究活动的性质决定。创新机会的有用性表现为：创新机会能否满足大众需求，是否具有可预见的应用价值。由于学术创新机会是对未来研究选题的指导，本文设计由社会关注度和当前学术研究热度两个子指标组成的需求度评估指标，用以反映学术创新机会的有用性。相关的学术创新机会需求度指标见表2。

表2　学术创新机会需求度指标

Tab. 2　The demand index of academic innovation opportunity

子指标	指标内容	测度方法
社会关注度	在社交媒体上出现越频繁，社会关注度越高	用社交媒体上对离群主题词跨学科组合的发布数、对应的转发数、点赞数、评论数来测度
研究热度	相关研究成果越多，研究热度越高	用相关研究成果的发文量来度量

1）离群主题词跨学科组合的社会关注度。关注度 R 的计算以微博官网（“微博—随时随地发现新鲜事”）数据为基础。微博是一种基于社交网络的用户信息交流平台，通过这个平台，学者们即时发布研究成果、想法和经验，进行学术交流和讨论，研究机构在微博上发布学术会议、论坛和研讨会。与小木虫、知乎、学术圈等更专业的社交媒体相比，微博的用户基数最大，可以获取更多样化的信息和观点，正好与政治学问题的复杂多样性相契合；微博的信息发布和传播速度极快，热点信息能够在第一时间得到广泛传播和讨论，微博上的信息实时性较强；微博平台信息互动性强，其评论、转发、点赞等功能可以使得用户方便地与其他用户进行互动，形成强大的社区效应。因此，微博在新词和热点话题发现、舆情分析、舆论引导、用户分析等中发挥着极其重要的作用。本文通过 Python 编程爬取两学科离群主题词组合在微博上的发布数、对应的转发数、点赞数、评论数4类数据，数据获取时间为2023-05-23 上午 10：35。根据对问题关注程度的差异性，主观确定上述4类数据的权重系数分别为0.5、0.1、0.3、0.1。由此得到离群主题词跨学科组合的关注度 R 计算公式为：

$$R = 0.5 \times \text{发布数} + 0.1 \times \text{转发数} + 0.3 \times \text{评论数} + 0.1 \times \text{点赞数} \quad (2)$$

2）离群主题词跨学科组合的研究热度。研究热度是指针对某研究对象、研究问题、研究方法等引起研究人员的关注数，具体表现为参与研究的人员数量、研究成果的多少等，可用发文量、被引量、下载量、转发量等来表示。鉴于上述数据存在重复性，为方便数据获取，本文用相关研究成果的发文量 N 来表示研究热度。

研究热度与创新机会之间的关系是动态变化的。就某一研究主题，在研究初期，研究热度可能会引发大量的创新机会；随着时间的推移，如果大家仍在追求研究同样的主题，创新机会则会降低。一项研究成果通常在发表3～5年后达到引用高峰，引用高峰期也是研究热度最高的时期。在这个时期，这项研究成果的重要性和影响力达到最高点，引起人们的广泛关注和引用。然而，这个引用高峰期也是研究创新性降低的拐点，因为在这个阶段之后，对研究成果再创新和改进的机会将相应减少。因此，在分析创新机会时，本文将发文量 N 限定为：从离群主题词跨学科组合的最早发文时间 t 延后4年，即第（$t+4$）年开始至今对应的发文量。由于本文以情报学为目标学科来分析创新机会，在检索文献时只选择了文献分类目录中的“信息科技”类，以确保所研究的文献具有与情报学相关的特点和内容。

3）需求度评估指标。离群主题词跨学科组合的社会关注度越高，研究热度越低，需求度越高。由此，设计需求度评估指标为：

$$N_{\text{time-quantity}} = \ln \frac{10 \times R}{N + 1} \quad (3)$$

为避免研究热度对应的发文量 N 为0的情况出现，公式（3）的分母取（$N+1$）。根据公式（2）和公式（3）计算离群主题词跨学科组合的需求度，结果见表3。

表3　离群主题词跨学科组合的学术创新机会需求度评估结果

Tab. 3　The evaluation results of the demand for academic innovation opportunity in interdisciplinary combinations of outlier topic words

组合	发布数	转发数	评论数	点赞数	关注度（R）	研究热度（N）	需求度
在线评论—政策工具	99	341	194	1749	365.1	0	8.203
人工智能—公共政策	297	1453	357	1599	792	15	6.205
在线评论—应急管理	13	126	112	330	108.3	2	5.889
在线评论—公共政策	22	38	82	152	57.8	2	5.261
人工智能—政策工具	220	204	124	1999	364.3	22	5.065
信息传播—应急管理	155	470	237	2022	460.8	152	3.405
网络舆情—应急管理	239	235	181	491	245.6	141	2.850
社交网络—应急管理	16	41	92	272	71.9	63	2.419
知识图谱—应急管理	23	12	8	28	15.7	56	1.013
突发事件—应急管理	391	565	367	989	495.8	1867	0.976

注：结果按需求度由大到小排序。

取需求度的中值（3.11）为阈值，将大于这个阈值的离群主题词组合视为具有高优先级的创新组合，即表3中的前6组离群主题词组合："在线评论—政策工具""在线评论—应急管理""在线评论—公共政策""人工智能—公共政策""人工智能—政策工具""信息传播—应急管理"。而"网络舆情—应急管理""社交网络—应急管理""知识图谱—应急管理""突发事件—应急管理"等组合的创新机会相对较小，属于低优先级的创新组合。

3.7 学术创新机会发现

根据表3识别出的高优先级的创新组合，结合现有研究，为情报学解决政治学问题提供相应的学术创新机会，见表4。

表4 基于离群主题词跨学科组合的情报学领域的学术创新机会

Tab.4 The academic innovation opportunities in the field of Information Science based on interdisciplinary combination of outlier topic words

序号	创新组合	具体学术创新机会（每对组合提供3个示例）
1	在线评论—政策工具	政策工具是政策达成目标的手段，政策工具的制定、实施及实施效果的评估都应遵循民意。面向服务大众，基于在线评论给出以下创新研究机会：（1）分析影响政策工具选择和正常实施的因素；（2）建设适应特定环境和人群的动态政策工具；（3）建设政策工具绩效评估的指标体系，以提高政策工具的管理效能，为政策顺利执行提供民意基础
2	人工智能—公共政策	基于人工智能技术快速分析多源数据，为决策者制定有利于可持续发展、公平合理的公共政策提供支持。基于人工智能技术的相关研究机会有：（1）通过多源信息的分析，制定、改进公共政策，提升公共政策的社会公平性；（2）发现不同地域、不同时期公共政策的主题及其相互联系，解释公共政策的特征、变迁与演变；（3）分析影响公共政策制定的因素，提高公共政策的效用
3	在线评论—应急管理	基于在线评论建立针对性的应对机制，采取针对性的措施，降低突发事件带来的负面影响，保障公众生命健康安全和利益、提高应急管理机构的信誉度。具体研究机会有：（1）分析人们对突发事件感知的影响因素，提高应急管理的效果；（2）预测公众在紧急情况下的行为表现和各类需求；（3）分析应急管理机构的信誉度，评估危机事件造成的影响，为应急管理提供参考
4	在线评论—公共政策	公共政策集中反映了社会利益，因而公共政策必须反映公众的意愿，保证公众对公共政策的满意度。基于在线评论的相关研究机会有：（1）分析大众参与公共政策制定的热情、公民在线参与公共政策的过程，提高人们的参政议政热情；（2）分析公共政策的本质、归属及价值取向；（3）检测、评价公共政策的目标能力、执行能力
5	人工智能—政策工具	基于人工智能技术模拟不同的情景，实现政策工具管理效能的最大化。相关研究机会有：（1）借助人工智能技术以优化求解的方式构建有效指标，预判、识别风险态势，判断政策工具的应用场景，形成政策智能工具；（2）针对不同问题，借助人工智能技术实现政策工具的优化组合；（3）借助人工智能技术，分析政策工具的多样性和动态性特征
6	信息传播—应急管理	社交媒体的使用增加了公众对快速接收信息、与传播者互动的渴望，面向应急管理。相关研究机会有：（1）分析社交媒体中信息的真实性和可信度；帮助用户正确理解和解读信息；（2）提供开放高效的信息查询和交流平台，减少谣言传播的负面效应，维护信息秩序和公共安全；（3）分析应急管理信息传播的及时性、一致性、差异性，为事前预警、事发应对、事中处置和善后管理提供信息保障

在表4中，政治学中的政策工具、应急管理、公共政策与情报学中离群主题词组合创新的机会较大，情报学中的在线评论、人工智能与政治学中的离群主题词实现组合创新的潜力较大。基于在线评论，利用人工智能技术解决政治学领域的政策工具、应急管理、公共政策等问题，已成为情报学领域未来学术创新的主要机会。当前，在线评论已成为网络环境下获取大众意愿、需求、情感等的重要数据资源，利用人工智能技术（如机器学习、深度学习、网络爬虫、图像处理、自然语言处理、知识图谱等），通过学习和分析大量的在线数据，快速、高效地挖掘在线评论中的公众诉求，针对社会经济、文化教育、医疗卫生、交通运输、信息服务等问题，实现对大众意愿的数据化、智能化、中立化管理与分析，能够为政治学领域提供更有价值的情报产品；精准发现公共政策、政策工具问题，及时进行应急管理，维护政策工具的权威性、公共政策的共有性价值。

相较于前5个组合，"信息传播—应急管理"的需求度不高，但仍可作为下一步研究创新的机会。在社交媒体环境下，突发事件具有信息传播速度快、受众面广的特点，因此情报人员应充分发挥自身信息分析的技术特长，准确分析信息传播的特点、机制，有效辨识信息的真伪，在应急管理中真正起到舆论监督和引导作用。

表4仅从宏观上给出有限的几个学术创新机会，学者们可以就当前经济、教育、交通、医疗、环境等某一具体的社会问题设计合适的研究课题，如基于在线评论研究古迹、文物保护管理政策；基于人工智能技术的疫情防控管理、洪涝灾害应急管理；分析人工智能在公共政策制定中的伦理问题；基于人工智能技术分析疫情信息传播的及时性、一致性、差异性；基于人工智能技术对当前政府服务产生的影响和冲击，研究公共政策的应对策略等。

4 结束语

离群主题词识别旨在寻找学科领域内的研究冷点，将离群主题词作为学术创新的重要信息，能够发现跨学科学术创新机会。通过情报学和政治学的跨学科学术创新机会发现的实证研究，验证了这个观点的实效性。本文将 LDA 模型与 Word2Vec 结合，能够同时获取文档的主题信息以及主题词之间的上下文语义关系，提取的主题信息更接近真实的文本内容。将 Word2Vec 和 PCA 结合识别离群主题词，不仅体现了离群主题词与学科领域内其他主题词内容上的语义差异，而且根据主题词在二维空间上分布的位置可直观呈现离群主题词。给出的学术创新机会需求度评估指标从社会关注度、学术研究热度两个维度可以分析离群主题词跨学科组合的有用性，可为学术创新机会评估研究提供参考。相似度计算将两学科中语义相似度高的离群主题词组合作为内容上具有新颖性的学术创新机会，并结合需求度评估，筛选出具有较高学术创新机会的两学科的离群主题词组合："在线评论—政策工具""在线评论—应急管理""在线评论—公共政策""人工智能—公共政策""人工智能—政策工具""信息传播—应急管理"；根据上述组合，结合现实研究，设计出相应的跨学科学术创新机会，研究结果有助于科研人员开展差异化的学术研究，同时也能为科研管理部门进行科研管理、开展学术评价提供指导。

本研究尚需完善的环节在于：①离群主题词识别。本研究按照主题词对所属主题的贡献度排序选择了贡献度较低的主题词。然而，主题词贡献度越低，与相应主题的关联度越低，因此选取的主题词数量不能太多，这就导致具有潜在创新价值的主题词被遗漏，使发现的学术创新机会不充分。未来将考虑以低频关键词为数据基础，从低频词的语义网络中获取学科领域的离群词，以解决离群主题词识别不全的问题。另外，由于本文实证研究所用的数据量较小，每个学科的主题词只有一个核心类团，离群点与类团中心的距离较远，因此选取可视化方法识别离群主题词。可视化方法避免了主题词之间距离的精准计算，但缺乏严谨性。当数据量较大时，主题词数量增多，会形成多个类团，并且每个类团周围都可能有自己的离群主题词。因此，为适应多类群的数据对象，有必要对降维后的主题词进行聚类分析。②学术创新机会有用性评估。本文仅通过需求度评估筛选到的离群主题词跨学科组合，而学术成果的研究潜力是对其未来潜在研究价值的一种预判，其大小受多种因素影响，如人们当前的知识结构、认知水平、研究方向的发展趋势、技术条件等。因此，后续研究需要构建更完备的评估体系，以利克特量表（Likert Scale）形式通过专家打分全面评估学术创新机会的有用性。□

参考文献

[1] 马费成. 情报学的进展与深化［J］. 情报学报，1996，15（5）：337-343.
[2] PETER S. Technological innovation opportunities［J］. Computers and People，1974，23（5）：33-33.
[3] ZHU Donghua，PORTER A L. Automated extraction and visualization of information for technological intelligence and forecasting［J］. Technology Forecasting and Social Change，2002，69（5）：495-506.
[4] 李保明. 技术机会与技术创新的决策［J］. 科学管理研究，1990（5）：61-62.
[5] 伊惠芳，刘细文，龙艺璇. 技术创新全视角下技术机会发现研究进展［J］. 图书情报工作，2021，65（7）：132-142.
[6] 赵可云. 新媒体干预农村留守儿童学习社会化研究［M］. 北京：中国社会科学出版社，2020：201-209.
[7] PORTER A L，DETAMPEL M J. Technology opportunities analysis［J］. Technological Forecasting and Social Chang，1995，49（3）：237-255.
[8] YOON B，MAGEE C L. Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction［J］. Technological Forecasting and Social Change，2018，132（1）：105-117.
[9] 石幸，战洪飞，余军合，等. 基于关联规则和技术功效矩阵的企业技术创新机会发现和辅助决策方法［J］. 宁波大学学报（理工版），2021，34（4）：35-42.
[10] 任海英，于立婷，黄鲁成. 基于链接预测的科学研究机会发现方法研究［J］. 情报杂志，2016，35（10）：53-58，37.
[11] YOON B，PARK Y. A Systematic approach for identifying technology opportunities：keyword-based morphology analysis［J］. Technological Forecasting & Social Change，2005，72（2）：145-160.
[12] YOON J，PARK H，SEO W，et al. Technology opportunity discovery（TOD）from existing technologies and products：a function-based TOD framework［J］. Technological Forecasting & Social Change，2015，100（4）：153-167.
[13] FENG Sijie. The proximity of ideas：an analysis of patent text using machine learning［J］. PLoS One，2020，15（7）：e0234880.
[14] 韩晓彤，朱东华，汪雪锋. 科学推动下技术机会发现方法研究［J］. 图书情报工作，2022，66（10）：19-32.
[15] 赵红州，蒋国华. 科学计量学的历史和现状［J］. 科学学研究，1984（4）：26-37.
[16] KAPLAN S，VAKILI K. The double-edged sword of recombination in breakthrough innovation［J］. Strategic Management

Journal，2015，36（10）：1435-1457.
[17] 杜德慧，李长玲，相富钟，等．基于引文关键词的跨学科相关知识发现方法探讨［J］．情报杂志，2020，39（9）：189-194.
[18] 商宪丽．基于潜在主题的交叉学科知识组合与知识传播研究［D］．武汉：华中师范大学，2017.
[19] 陈虹枢，宋亚慧，金茜茜，等．动态主题网络视角下的突破性创新主题识别：以区块链领域为例［J］．图书情报工作，2022，66（10）：45-58.
[20] 牌艳欣，李长玲，徐璐．弱引文关系视角下跨学科相关知识组合识别方法探讨——以情报学为例［J］．图书情报工作，2020，64（21）：111-119.
[21] 李长玲，刘小慧，刘运梅，等．基于开放式非相关知识发现的潜在跨学科合作研究主题识别——以情报学与计算机科学为例［J］．情报理论与实践，2018，41（2）：100-104.
[22] 刘小慧，李长玲，崔斌，等．基于闭合式非相关知识发现的潜在跨学科合作研究主题识别——以情报学与计算机科学为例［J］．情报理论与实践，2017，40（9）：71-76.
[23] HAN J W，KAMBER M，PEI J. 数据挖掘概念与技术［M］．3版．范明，孟小峰，译．北京：机械工业出版社，2012：351-352.
[24] HANAUER D A，SAEED M，ZHENG Kai，et al. Applying MetaMap to Medline for identifying novel associations in a large clinical dataset：a feasibility analysis［J］．Journal of the American Medical Informatics Association Jamia，2014，21（5）：925-937.
[25] 周志光，谢琬滢，郑微桦，等．降水时空关联特征可视分析［J］．中国图象图形学报，2021，26（3）：619-632.
[26] 何文静．基于异常检测的学科领域新兴主题识别研究［D］．武汉：武汉大学，2017.
[27] 李勇男．基于聚类的反恐情报异常数据分析方法研究［J］．现代情报，2019，39（10）：32-37，65.
[28] 邓胜利，夏苏迪，汪奋奋．基于图注意力网络的社交媒体异常用户预测研究［J］．情报理论与实践，2022，45（3）：94-102.
[29] 周燕，肖莉．基于改进关联聚类算法的网络异常数据挖掘［J］．计算机工程与设计，2023，44（1）：108-115.
[30] UZZI B，MUKHERJEE S，STRINGER M，et al. Atypical combinations and scientific impact［J］．Science，2013，342（6157）：468-472.
[31] WANG Jian，VEUGELERS R，STEPHAN P. Bias against novelty in science：a cautionary tale for users of bibliometric indicators［J］．Research Policy，2017，46（8）：1416-1436.
[32] 罗素平，寇翠翠，金金，等．基于离群专利的颠覆性技术预测——以中药专利为例［J］．情报理论与实践，2019，42（7）：165-170.
[33] 关杏彬．基于离群专利的技术机会分析研究［D］．广州：华南理工大学，2019.
[34] 宋凯，冉从敬．基于主题挖掘与专利评估的技术机会识别研究——以智慧农业为例［J］．图书情报工作，2023，67（3）：61-71.
[35] LEE J，PARK S. Technology opportunity analysis based on machine learning［J］．Axioms，2022，11（12）：708.
[36] 孙震，冷伏海．一种基于知识元迁移的ESI研究前沿知识演进分析方法［J］．情报学报，2021，40（10）：1027-1042.
[37] 王芳．情报学理论：哲学基础与应用发展［M］．北京：科学技术文献出版社，2021：9.
[38] 刘婷，李长玲，刘运梅，等．基于参考文献分类号的图书情报学跨学科知识输入特点分析［J］．情报科学，2018，36（10）：99-104.
[39] 吴陈昊，杨建林．国内情报学研究中的知识输入状况定量分析［J］．情报理论与实践，2020，43（4）：66-73.
[40] 陶盈春，张红丽，徐健．异常值探测在大数据分析中的应用研究［J］．情报科学，2018，36（3）：75-80.
[41] KANUNGO T，MOUNT D M，NETANYAHU N S，et al. An efficient K-means clustering algorithm：analysis and implementation［J］．IEEE Transactions on Pattern Analysis & Machine Intelligence，2002，24（7）：881-892.
[42] LEE Y N，WALSH J P，WANG Jian. Creativity in scientific teams：unpacking novelty and impact［J］．Research Policy，2015，44（3）：684-697.
[43] CUMMING B S. Innovation overview and future challenges［J］．European Journal of Innovation Management，1998（1）：21-29.

作者简介：李秀霞（通信作者，Email：lixiuxia@qfnu.edu.cn），女，1971年生，硕士，教授，硕士生导师。研究方向：学术评价，知识发现。**邵作运**，男，1972年生，硕士，副研究馆员，硕士生导师。研究方向：文本分析与知识发现。

作者贡献声明：李秀霞，确定内容与研究方向，撰写论文。**邵作运**，制定论文与研究框架，数据处理。

录用日期：2023－07－24