北京大学

本科生毕业论文

题目：基于 BERT 和 DGCN 的弱信号识别研究——以人工智能在金融领域的应用为例

Weak Signal Recognition Using BERT and DGCN: An AI Application Study in Finance

姓　　名：	黄赢泽
学　　号：	2100093001
院　　系：	信息管理系
专　　业：	大数据管理与应用
导师姓名：	周庆山　教授

二〇二五年　五月

摘要

随着人工智能在金融领域的快速发展，弱信号识别在预测未来趋势和潜在风险方面变得尤为重要。本文提出了一种基于 BERT（Bidirectional Encoder Representations from Transformers）和 DGCN（Directed Graph Convolutional Network）的弱信号识别框架，旨在通过融合文本深层语义理解和知识网络拓扑主题建模，并进行主题与术语的过滤，从中提取出相关领域的微弱信号。研究以 2018 年至 2024 年人工智能在金融领域的学术文献为数据源，通过构建引用网络和应用有向图卷积网络，结合 BERT 模型进行文本向量化，实现了对潜在研究主题的精准识别。进一步，通过弱函数筛选出潜在的弱信号主题，并基于 c-TF-IDF（class-based Term Frequency-Inverse Document Frequency）和文献被引量过滤出具有代表性的弱信号术语。研究发现，早期识别的弱信号中有 57%在后续发展中转化为行业重要趋势，验证了所提方法的有效性。最后，本文基于 2024 年的弱信号，预测未来人工智能在金融领域三大重点方向：绿色金融科技的崛起、普惠金融的新突破以及金融科技与政府合作的深化。

本文的研究不仅丰富了弱信号识别的技术体系，还为金融机构的技术布局、投资决策和风险管理提供了早期预警支持，同时也为监管机构和政策制定者提供了技术发展动态的监测依据。通过多学科交叉和技术融合，本文为不确定环境下的战略决策提供了有力支持，为金融科技领域的可持续发展提供了新的视角和方法论支持。研究结果表明，所提出的框架在识别潜在技术趋势和概念方面具有较高的准确性和前瞻性，能够有效捕捉金融科技领域的弱信号，为相关主体提供了宝贵的信息资源。未来研究将从数据来源、模型优化和应用场景等方面进行拓展和深化。在数据层面，构建多语言、多来源的异构数据集，整合非英语文献、专利、行业报告等数据，以捕捉更全面的弱信号。在模型优化上，探索轻量化模型和分布式训练框架，提升计算效率和实时处理能力。在应用层面，将弱信号识别方法应用于其他高风险、高动态性领域，开发可视化分析工具，帮助决策者更直观地理解术语之间的关联性和演化趋势。此外，还将关注弱信号识别的伦理和社会影响，探索合规性框架，确保其在促进创新的同时，符合数据隐私和知识产权保护的要求。通过这些努力，弱信号识别技术有望在更多领域发挥重要作用，为战略决策提供更加有力的支持。

关键词：弱信号识别，BERT，有向图卷积网络，金融科技，主题建模

ABSTRACT

With the rapid advancement of artificial intelligence in finance, the identification of weak signals has become increasingly critical for forecasting future trends and mitigating potential risks. This paper introduces a novel weak signal recognition framework that integrates BERT (Bidirectional Encoder Representations from Transformers) and DGCN (Directed Graph Convolutional Network). The framework leverages deep semantic understanding of text and knowledge network topology through topic modeling, enhanced by topic and term filtering, to extract weak signals from relevant domains.

Using academic literature on AI in finance from 2018 to 2024 as the primary data source, the study achieves precise identification of emerging research topics by constructing a citation network and applying a directed graph convolutional network alongside BERT-based text vectorization. Potential weak-signal topics are further refined using weak function filtering, while representative terms are selected based on c-TF-IDF (class-based Term Frequency-Inverse Document Frequency) and citation analysis. The results demonstrate that 57% of the early-detected weak signals evolved into significant industry trends, validating the efficacy of the proposed methodology. Based on the weak signals identified in 2024, the paper forecasts three pivotal future directions for AI in finance: the emergence of green fintech, transformative breakthroughs in financial inclusion, and deeper collaboration between fintech and government entities.

This research not only advances the technical framework for weak signal detection but also offers strategic insights for financial institutions in technology deployment, investment decisions, and risk management. Additionally, it provides regulators and policymakers with a robust monitoring tool to track technological developments. By integrating multidisciplinary approaches, the study enhances strategic decision-making in uncertain environments and contributes new perspectives for the sustainable growth of financial technology.The findings underscore the framework’s accuracy and foresight in identifying nascent technological trends, effectively capturing weak signals within the fintech domain to deliver actionable intelligence.

Future research will focus on expanding data sources by developing heterogeneous datasets that incorporate multilingual sources, non-English literature, patents, and industry reports to capture a more comprehensive spectrum of weak signals. Model optimization will explore lightweight architectures and distributed training frameworks to improve computational efficiency and real-time processing capabilities. Application diversification will extend weak signal recognition to other high-risk, dynamic fields while developing visualization tools to help decision-makers better understand term correlations and evolutionary

patterns.Furthermore, ethical and societal implications will be carefully considered, with the development of compliance frameworks to ensure adherence to data privacy and intellectual property standards while fostering innovation. Through these advancements, weak signal recognition technology is poised to play a transformative role across diverse sectors, offering robust support for strategic decision-making in an increasingly complex world.

KEYWORDS: Weak Signal Detection, BERT, DGCN, FinTech, Topic Modeling

目　录

图 目 录

表 目 录

第一章 引言

1.1 研究背景

进入 21 世纪以来，全球范围内的信息技术革命与数字化转型进程加速，数字经济已成为推动社会发展和产业升级的核心引擎。根据中国信通院发布的《全球数字经济白皮书（2022 年）》显示，2021 年测算的 47 个国家数字经济增加值规模为 38.1 万亿美元，同比名义增长 15.6%，占 GDP 比重为 45.0%。[1]在这一背景下，各行各业尤其是金融领域，面临着前所未有的技术重塑和模式创新。与此同时，大数据环境下的信息爆炸现象尤为显著，学术论文、专利文档、行业报告等文本数据量呈指数级增长，信息过载成为制约高效决策与前瞻判断的主要挑战。这种背景促使人们不仅需要挖掘已有数据中的显性知识，更亟需发掘隐藏在噪声中的微弱征兆，为未来趋势研判提供支撑。

在信息洪流之中，弱信号（Weak Signals）作为尚未形成显著趋势、但可能对未来产生重要影响的早期征兆，逐渐成为战略管理、风险预警和竞争情报等领域的重要研究对象。弱信号的概念最早由战略管理学者 Ansoff 提出，用以描述那些尚未形成明确趋势但可能对未来产生重大影响的早期征兆。[2]与传统意义上的“强信号”不同，弱信号往往具有三个典型特征：一是信息表达的模糊性，通常以碎片化形式分散在不同数据源中；二是影响的不确定性，其潜在价值或风险需要经过时间验证；三是识别的滞后性，常规监测手段难以实时捕捉。传统的弱信号识别依赖于专家经验或基于规则的系统，但随着数据规模扩大和环境复杂性上升，人工方法面临效率低、准确率低的问题。近年来，随着自然语言处理（Natural Language Processing, NLP）、机器学习（Machine Learning, ML）与图神经网络（Graph Neural Network, GNN）技术的兴起，弱信号识别方法从规则驱动逐步转向数据驱动与智能化。然而，现有研究仍存在诸多挑战：一方面，基于语义的文本挖掘方法往往忽视了信息之间的结构性关联；另一方面，基于图结构的模型又难以充分理解深层次语义，导致弱信号识别在精度和覆盖度之间难以取得有效平衡。

与此同时，人工智能（Artificial Intelligence, AI）技术在金融领域的应用正在不断深化，形成了以智能投顾、信用风控、智能交易、反欺诈等为代表的应用体系。从最初的基于规则的专家系统，到当前基于深度学习（Deep Learning）的复杂决策系统，AI 技术极大地推动了金融行业的数据分析、风险管理与客户服务水平的提升。例如，长短期记忆网络（Long Short-term Memory, LSTM）已广泛应用于股价预测，情感分析模型辅助市场情绪判断，图神经网络用于反洗钱和欺诈识别，极大提升了金融业务的智能化水平。然而，金融科技（FinTech）领域的技术创新也伴随着新的挑战，特别是在数据隐私保护、模型可解释性和合规性要求方面。以欧盟《人工智能法案》为例，其对金融场景中 AI 应用提出了高标准

[1] 中国信息通信研究院. 全球数字经济白皮书(2022 年)[R]. 北京: 中国信息通信研究院, 2022.
[2] Ansoff H I. Managing strategic surprise by response to weak signals[J]. California Management Review, 1975, 18(2): 21-33.

的透明性和问责性要求，限制了部分深度模型的直接应用。这种技术演化与监管要求交织的环境，使得金融领域对新兴趋势、潜在风险的早期识别能力提出了更高要求。

综上所述，数字经济环境下数据爆炸与信息超载加剧了未来趋势预测的困难，弱信号识别作为应对不确定性的重要工具，其方法体系在近年来虽有所突破，但仍存在语义理解与结构建模割裂的问题。同时，金融科技领域作为技术迭代最为迅速的行业之一，既需要应对快速变化带来的技术风险，又要满足合规性与伦理性的高标准要求。这种背景下，如何建立一种能够兼顾文本深层语义理解与知识网络结构建模的弱信号识别框架，以支持金融领域对潜在趋势与隐性风险的敏锐感知，成为亟待解决的关键课题。本研究基于此问题意识，提出融合预训练模型与神经网络的弱信号识别方法，旨在为业界提供一种高效、前瞻且符合合规性要求的趋势洞察工具。

1.2 研究目的与意义

基于前述背景，本文旨在针对金融科技领域中弱信号识别存在的语义与结构信息割裂问题，提出一种融合文本深层语义理解与知识网络拓扑建模的新型识别框架。具体而言，本文以 BERT 为基础，提取文献和专利数据中的语义特征；同时引入有向图卷积网络（DGCN），建模金融科技知识传播过程中的引用关系，以实现多维度信息的综合分析与弱信号的高效挖掘。

本研究拟实现以下具体目标：

(一) 设计并验证一种结合 BERT 与 DGCN 的弱信号识别方法，提升识别的准确性与前瞻性；

(二) 利用可量化的弱函数（Weakness Function），从主题紧密度、权重、自相关性三个维度筛选潜在弱信号主题；

(三) 应用上述方法对金融科技领域的数据进行实证研究，识别具有潜在影响力但尚未爆发的新兴技术趋势。

通过上述目标的实现，本文力求在理论探索与应用实践之间搭建桥梁，丰富弱信号识别的技术体系，并为金融领域趋势洞察提供可操作的解决方案。

本研究在理论与实践两个层面均具有重要意义。在理论方面，本文拓展了弱信号识别的研究范式，打破了以往主要依赖浅层文本特征或简单共现网络的局限，提出融合深度语义理解与有向知识网络拓扑建模的新方法，为弱信号识别提供了新的理论模型与分析框架。同时，本文深化了自然语言处理与图神经网络交叉应用的研究，通过整合 BERT 与 DGCN，探索了多源异构数据环境下语义与结构信息协同建模的新路径，为趋势预测与风险预警等领域开辟了方法论上的新方向。此外，本文结合金融科技领域的具体应用背景，丰富了该领域前瞻性技术识别与演化研究的理论基础。在实践层面，本文提出的方

法能够为金融机构的技术布局、投资决策和风险管理提供早期预警支持，帮助机构在高度动态与不确定的环境中提高敏锐感知能力，抢占战略先机；同时，研究成果也可为监管机构与政策制定者提供技术发展动态的监测依据，助力科学制定引导与监管策略；更进一步，本文注重模型可解释性与结构透明性，符合金融科技领域对合规性和伦理性的要求，有助于推动人工智能技术在金融行业的稳健应用与可持续发展。

1.3 研究方法

为了实现金融科技领域弱信号的高效识别与深度挖掘，本文设计了一套系统化的研究流程，综合运用了自然语言处理与图神经网络的方法。具体步骤如下：

(一) 数据来源：以开放文献数据库 OpenAlex 为主要数据源，选取金融科技领域的学术论文与专利摘要，确保数据的权威性、前沿性与时效性；

(二) 信息表征：采用预训练语言模型 BERT 对文本进行深层次语义向量化，充分捕捉文档中的语义特征；

(三) 结构建模：构建论文引用网络，并基于有向图卷积网络（DGCN）融合节点间的引用关系信息，实现文本语义与知识网络拓扑的双重建模；

(四) 主题建模：应用 UMAP（Uniform Manifold Approximation and Projection）降维与 K-means 聚类算法识别潜在研究主题；

(五) 主题过滤：结合自定义弱函数从主题紧密度、权重及自相关性三个维度筛选出潜在弱信号主题；

(六) 术语识别：基于类别的 TF-IDF（c-TF-IDF）与文献被引量过滤术语，完成对弱信号的识别。

1.4 研究创新点

本研究在方法与应用两个方面具有显著创新。首先，本文提出了基于 BERT 与 DGCN 融合的弱信号识别框架，突破了以往弱信号研究中仅依赖文本语义或单一网络结构建模的局限，实现了语义理解与知识传播关系的协同建模，提升了弱信号识别的准确性与系统性。其次，引入了弱函数用于主题筛选，综合考虑主题紧密中心度、主题权重与主题自相关性三个维度，以量化方式刻画主题的潜在弱性特征，有效避免了噪声信息干扰，提高了筛选的科学性与精确性。此外，本文还设计了从主题到术语的双重过滤机制，结合 c-TF-IDF 值和文献被引量筛选弱术语，不仅增强了弱信号识别结果的解释性，也提升了对潜在趋势的洞察能力。最后，本文将上述方法应用于金融科技领域的实证研究中，围绕智能投顾、金融反欺诈、高频交易等关键场景进行了系统验证，丰富了弱信号识别在金融产业应用中的案例积累，拓

展了相关研究的实践边界。总体而言，本文从理论建模、技术路径到应用验证，系统构建了面向金融科技领域的弱信号识别新框架，为后续相关研究和实际应用提供了可参考的方法论支持。

第二章 国内外文献综述

2.1 弱信号识别相关研究现状

2.1.1 国外研究

弱信号识别作为技术预见、竞争情报和风险管理的重要工具，其研究方法经历了从规则驱动到数据驱动的显著转变。根据技术手段和分析范式的差异，相关研究可划分为三个主要发展阶段，体现了从人工依赖向自动化、智能化方向的演进。第一阶段（2000-2010 年）以专家系统和规则引擎为主导，典型代表如 1975 年 Ansoff 提出的战略预警系统。该方法依赖领域专家设定的规则进行信号识别，虽然逻辑清晰且易于解释，但其性能受限于专家知识的覆盖范围和更新效率，难以适应动态变化的环境。第二阶段（2011-2014 年）标志着统计学习和语义分析技术的兴起，研究重点转向数据驱动的弱信号挖掘。Yoon 率先将关键词共现网络与时间序列分析结合，通过监测新闻文本中的词汇突发性增长来识别潜在趋势，但该方法难以捕捉词汇间的深层语义关联。[3]Thorleuchter 等进一步引入潜在语义分析（LSI），构建概念空间模型以识别表达形式不同但语义相近的弱信号，提升了主题演变的追踪能力。[4,5]

第三阶段（2015 年至 2021 年），机器学习技术的突破推动了弱信号识别的定量化发展，如 Kim 等提出的 NEST 模型整合全球专家网络与贝叶斯网络，为新兴技术趋势评估提供了结构化框架。[6]El Akrouchi 等开发的基于 LDA 的动态主题模型能够自动追踪技术术语在专利文献中的演变轨迹，其创新点在于引入了“技术成熟度曲线”的概念，通过计算主题热度变化率来预测技术拐点。[7]Gutsche[8]、Bouktaib 等人[9]、Mühlroth 和 Grottke[10]也通过 LDA 等语义聚类算法与其他技术的结合来检测网络数据中的弱信号。Edo-Osagie 等通过半监督分类挖掘 Twitter 数据，成功应用于公共卫生监测。[11]Yu 等提出的 FRAM-机器学习框架（如平衡随机森林）可有效识别工业系统中的事故预警信号。到了第四阶段（2022 年至

---

[3] Yoon J. Detecting weak signals for long-term business opportunities using text mining of Web news[J]. Expert Systems with Applications, 2012, 39(16): 12543-12550.
[4] Thorleuchter D, Van den Poel D. Weak signal identification with semantic Web mining[J]. Expert Systems with Applications, 2013, 40(12): 4978-4985.
[5] Thorleuchter D, Scheja T, Van den Poel D. Semantic weak signal tracing[J]. Expert Systems with Applications, 2014, 41(11): 5009-5016.
[6] Kim S, Kim Y E, Bae K J, et al. NEST: A quantitative model for detecting emerging trends using a global monitoring expert network and Bayesian network[J]. Futures, 2013, 52: 59-73.
[7] El Akrouchi M, Benbrahim H, Kassou I. End-to-End LDA-based automatic weak signal detection in Web news[J]. Knowledge-Based Systems, 2021, 212: 106650.
[8] Gutsche T. Automatic weak signal detection and forecasting[D]. Enschede: University of Twente, 2018.
[9] Adil B, Abdelhadi F. A framework for weak signal detection in competitive intelligence using semantic clustering algorithms[J]. International Journal of Advanced Computer Science and Applications, 2021, 12(12): 555-565.
[10] Mühlroth C, Grottke M. Artificial intelligence in innovation: how to spot emerging trends and technologies[J]. IEEE Transactions on Engineering Management, 2020, 69(2): 493-510.
[11] Edo-Osagie O, Smith G, Lake I, et al. Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance[J]. PLoS ONE, 2019, 14(7): e0210689.

今），深度学习和自动化技术的融合推动了弱信号识别的智能化发展。[12]预训练语言模型（如 BERT）的引入进一步提升了语义理解能力，Ebadi 等通过 BERT 的多层定量分析实现了高超声速技术领域弱信号的精准追踪。[13]弱信号识别研究经历了从人工规则到统计学习，再到深度学习的范式转换，标志着该领域逐渐向自动化算法和大规模数据处理的技术手段迈进。

2.1.2 国内研究

国内弱信号识别研究虽然起步较晚，但在应用创新方面展现出鲜明的特色。早期研究（2015 年前）主要聚焦于科技情报领域，致力于理论框架的构建和方法移植，如方微和邵波利用信息交流、情景分析、竞争对手学习与跟踪等，从理论角度阐述企业如何在实践中实现弱信号分析与风险识别。[14]邓胜利等的研究则更具实践指导意义，他们创新性地将波特的五力模型扩展为七力模型，并运用层次分析法和隶属度函数实现了弱信号的定量识别，显著提升了研究成果在企业战略决策中的应用价值。[15]这一阶段的研究虽然以理论探讨为主，但为后续的实证研究提供了重要的方法论指导。在指标评价方面，邓建军等提出的“发现—遴选—评价”（DSE）识别思路，通过结合情报大数据分析和科学知识图谱技术，为颠覆性技术识别提供了新视角。[16]

2016 年以后的研究以主流的机器学习和深度学习技术为主。韩盟等则基于范式转移理论、知识层次理论和意义建构理论，运用文献计量、文本挖掘、主题聚类、相似度计算和社会网络分析等技术，创建了基于三元计量特征的新兴技术弱信号识别方法，并以量子信息技术领域为实证对象进行方法检验和识别预测。[17]刘俊婉等的专利分析方法[18]、孙涛等的文本相似度-主题发现混合模型[19]、王莉晓等的多模型融合预测方法[20]，都体现了人工智能技术在弱信号识别中的创新应用。特别值得关注的是，将动态主题模型应用于比亚迪新能源汽车的顾客评论分析，通过新颖度、出现率、支持率和点击率等指标体系，成功预测了多个新兴消费趋势的爆发。刘宇飞等采用基于引用网络和文本信息的知识融合路径分析方法，结合图神经网络（GNN）模型，对论文数据进行聚类处理，分析了不同技术领域的融合路径和趋势。[21]林晗提出了一种基于引用网络结构和文本语义融合的热点识别方法，通过改进深度学习主题建模技术

---

[12] Yu M, Pasman H, Erraguntla M, et al. A framework to identify and respond to weak signals of disastrous process incidents based on FRAM and machine learning techniques[J]. Process Safety and Environmental Protection, 2022, 158: 98-114.
[13] Ebadi A, Auger A, Gauthier Y. Detecting emerging technologies and their evolution using deep learning and weak signal analysis[J]. Journal of Informetrics, 2022, 16(4): 101344.
[14] 方微, 邵波. 基于弱信号分析的企业风险识别[J]. 图书情报工作, 2009, 53(14): 80-83.
[15] 邓胜利,林艳青,王野.企业竞争弱信号的特征提取与定量识别研究[J].图书情报工作,2016,60(10):67-75.
[16] 邓建军,刘安蓉,曹晓阳,等.颠覆性技术早期识别方法框架研究－－基于科学端的视角[J].中国科学院院刊,2022,37(05):674-684.
[17] 韩盟,陈悦,王玉奇,等.新兴技术弱信号识别：理论模型与测度方法[J].科学学研究,2024,42(11):2262-2274.
[18] 刘俊婉,庞博,徐硕.基于弱信号的颠覆性技术早期识别研究[J].情报学报,2023,42(12):1395-1411.
[19] 孙涛,张秉坤,成磊峰,等.基于文本相似度和主题发现的弱信号识别方法[J].电脑知识与技术,2024,20(23):34-36.
[20] 王莉晓,陈伟,邱含琪.基于机器学习的颠覆性技术弱信号识别模型研究[J].数据分析与知识发现,2024,8(Z1):63-75.
[21] 刘宇飞,苗仲桢,黎凌峰,等.战略性新兴产业多领域知识融合路径研究——基于引用网络和文本信息的分析[J].中国工程科学,2020,22(02):120-129.

AVIGTM，结合元聚类融合策略，实现了对研究热点的精准识别和语义提取。[22]尹娟则通过结合动态主题建模和改进的 Autoformer 时序预测算法，实现了社交媒体数据中弱信号的自动提取和趋势预测。[23]不过，现有国内研究在理论基础构建和国际影响力方面仍有提升空间，多数成果尚未形成系统化的方法论体系。除此之外，在跨学科融合和实际应用方面仍具有较大发展前景，应推动形成“数据-算法-应用”的完整创新链条，使弱信号识别技术真正成为支撑战略决策的有力工具。

整体而言，现有的弱信号识别方法主要存在两个方面的不足：一是过度依赖专家经验，导致识别过程主观性强、效率低下；二是分析方法单一，往往仅关注文本语义或元数据分析中的某一方面，难以全面捕捉弱信号特征。针对这些问题，本研究创新性地提出了一种融合 BERT 语义表征和 DGCN 网络结构分析的多模态识别方法，通过构建量化弱函数实现客观筛选，既克服了传统方法的主观性局限，又实现了多维度特征的有效融合，以此提升弱信号识别的准确性和时效性，为相关研究提供了更科学的技术路径。

2.2 人工智能相关领域应用现状

2.2.1 金融领域

近年来，金融科技（FinTech）的快速发展正在深刻重塑传统金融行业的格局。人工智能技术在金融领域的应用已从局部试点转向全链条渗透，形成了覆盖基础设施、中台决策到终端服务的完整技术生态。这一变革不仅重构了传统金融业务模式，更催生了以数据驱动为核心的新型业态。从早期的规则引擎到如今的深度学习，金融业的智能化转型呈现出明显的技术迭代路径，而这一过程中，弱信号的捕捉与识别往往成为技术突破的关键。

在风险管理这一核心领域，人工智能的应用已从单一的信用评分扩展到包含破产预测、反欺诈、合规监控等多元场景。研究表明，神经网络（ANN）和支持向量机（SVM）等算法在破产预测中的表现显著优于传统统计模型。例如，Kwak 等人应用多准则线性规划（MCLP）以及支持向量机来预测韩国企业的破产，两种方法的预测准确率都在 87%左右。[24]类似地，Khandani 等利用机器学习算法在预测消费者信用卡违约和拖欠方面表现出色，其预测能力显著优于传统的信用评分模型。[25]值得注意的是，技术

[22] 林晗.引用网络结构和文本语义融合的热点识别方法研究[D].军事科学院,2023.
[23] 尹娟.基于深度学习的社交媒体网络舆情弱信号识别与预警研究[D].重庆理工大学,2024.
[24] Kwak W ,Shi Y ,Kou G .Bankruptcy prediction for Korean firms after the 1997 financial crisis: using a multiple criteria linear programming data mining approach[J].Review of Quantitative Finance and Accounting,2012,38(4):441-453.
[25] Khandani E A ,Kim J A ,Lo W A .Consumer credit-risk models via machine-learning algorithms[J].Journal of Banking and Finance,2010,34(11):2767-2787.

迁移的趋势日益明显——以摩根大通（JPMorgan）COiN 平台为例，其通过自然语言处理与深度学习技术解析商业贷款合同，将原本需要 36 万小时人工审阅的工作压缩至秒级完成[26]，这种效率的跃升正是人工智能替代规则引擎的典型案例。

投资管理领域的技术革新则更加激进。股票价格预测已从传统的技术指标分析转向基于 LSTM 和新闻情感分析的混合建模。Ghosh 等构建的 LSTM 模型在标普 500 指数预测中实现了 0.64%的交易前每日回报率[27]；而 Manela 和 Moreira 则证明，新闻文本对股票市场波动率的预测效力甚至超过部分宏观经济指标。[28]机构层面的实践更具颠覆性：BlackRock 的 Aladdin 系统[29]和 Bridgewater[30]高达 2 亿美金的纯 AI 基金通过融合多源异构数据（包括卫星图像、供应链日志等非结构化数据），能够实时识别市场波动并执行毫秒级交易。Sentient Technologies 开发的 AI 交易系统能够自主识别市场模式并实时调整策略，其表现已超越传统量化模型。[31]这种能力使得算法交易逐渐从执行工具演变为决策主体。数据显示，全球算法交易市场表现出显著增长，从 2022 年的 160.2 亿美元扩大到 2023 年的 180.6 亿美元，复合年增长率（CAGR）为 12.7%。[32]

反洗钱与欺诈检测是另一个技术密集度极高的场景。2020 年，一般基于传统规则的检测系统误报率高达 95%左右，而根据麦肯锡报告，通过引入 AI 手段，可以减少 20%-30%的误报率，能够极大地减少人员投入。[33]Jullum 等采用随机森林算法为挪威银行构建的反洗钱模型，在公平的性能度量上优于银行当时的方法。[34]更值得关注的是技术融合带来的范式创新——Mastercard 的 Decision Intelligence 技术通过建立用户行为基线，结合实时交易流分析，使欺诈识别的准确率欺诈检测率平均提高 20%，在某些情况下甚至可提高 300%。[35]这种从“事后侦测”到“事前预防”的转变，体现了 AI 对金融安全防护体系的重构能力。

在零售金融端，AI 的渗透同样深刻。Bank of America 的聊天机器人 Erica 通过自然语言理解技术，

---

[26] Superior Data Science. J.P Morgan - COiN - a case study of AI in finance[EB/OL]. (n.d.)[2023-12-01]. https://superiordatascience.com/jp-morgan-coin-a-case-study-of-ai-in-finance/.
[27] Pushpendu G ,Ariel N ,Keshari J S .Forecasting directional movements of stock prices for intraday trading using LSTM and random forests[J].Finance Research Letters,2022,46(PA):
[28] Manela A ,Moreira A .News implied volatility and disaster concerns[J].Journal of Financial Economics,2017,123(1):137-162.
[29] Filipsson F. Case study: AI-based predictive analytics for investments at BlackRock (Aladdin)[EB/OL]. (2025-02-12)[2023-12-01]. https://redresscompliance.com/case-study-ai-based-predictive-analytics-for-investments-at-blackrock-aladdin/.
[30] Basak S. Bridgewater launches $2 billion fund[EB/OL]. (2024-07-01)[2023-12-01]. https://finance.yahoo.com/news/bridgewater-launches-2-billion-fund-150000992.html.
[31] Roetzer P. Sentient Technologies applied its AI platform to marketing[EB/OL]. (2017-10-19)[2023-12-01]. https://www.marketingaiinstitute.com/blog/sentient-technologies-spotlight.
[32] Global Algorithmic Trading Market. Algorithmic trading global market report[EB/OL]. (2023-09-11)[2023-12-01]. https://finance.yahoo.com/news/algorithmic-trading-global-market-report-092300576.html.
[33] 支付界. 反洗钱任重道远：行业平均误报率仍 95%[EB/OL]. (2020-03-24)[2023-12-01]. https://www.sohu.com/a/382744499_208700.
[34] Jullum M, Løland A, Huseby R B, et al. Detecting money laundering transactions with machine learning[J]. Journal of Money Laundering Control, 2020, 23(1): 173-186.
[35] Mastercard. Mastercard supercharges consumer protection with gen AI[EB/OL]. (2024-02-01)[2023-12-01]. https://www.mastercard.com/news/press/2024/february/mastercard-supercharges-consumer-protection-with-gen-ai/.

单单在 2024 年就能处理超过 2.5 亿常规客户咨询，节省客服成本的同时也提高服务效率，98%以上的用户能够仅在平均 44 秒内获得他们想要的答案。而蚂蚁集团的“智能风控引擎”则展示了技术下沉的价值——融合生成式 AI 技术，大幅提升风控智能化水平，能够在风险管理复杂度综合降低 50%的前提下，将风险对抗时效从天级别降低到了小时级别。[36]目前已服务金融、出行、电商等十多个行业的 1000 余家企业。这些案例共同揭示了一个规律：金融 AI 的价值实现往往依赖于“技术适配性”，即如何针对特定业务场景优化通用算法。例如，Huang 等开发的 FinBERT 通过注入金融知识图谱，使情感分析准确率比其他方法提升了至少 12%，这种领域专用模型的崛起正是弱信号转化为主流应用的重要路径。[37]

然而，金融 AI 的发展仍面临监管与伦理的双重挑战。欧盟《人工智能法案》对高风险金融 AI 系统提出了“决策可解释性”的硬性要求，这直接制约了深度学习在信贷审批等场景的部署。不过，一些国家却对金融科技持有包容态度，如新加坡金融管理局与金融机构合作，开发了多个反洗钱解决方案，利用先进的数据分析和机器学习技术来提高反洗钱监测的效率和准确性。这些解决方案包括客户身份验证、交易监测和风险评估等方面。以上技术演进路径表明，未来金融 AI 的创新将更多出现在“合规性”与“性能”的平衡点上，而早期识别这类技术交叉产生的弱信号，对于把握行业变革先机具有战略意义。

2.2.2 其他领域

医疗 AI 的发展呈现出从辅助诊断向全流程智能化的演进路径。影像识别技术已实现从二维到三维、从静态到动态的跨越，西门子医疗的 AI-Rad Companion 系统能够自动标注 CT 影像中的可疑病灶，简化病灶测量以节省医生的时间。在药物研发领域，DeepMind 的 AlphaFold 系统破解了蛋白质折叠难题，有望显著缩短药物研发的整体周期。值得关注的是，医疗 AI 正在突破单一应用场景的限制，向诊疗全流程延伸。例如，Mayo Clinic 开发的临床决策支持系统整合了电子病历、基因检测和实时监护数据，能够为复杂病例提供个性化治疗方案建议。

自动驾驶技术的商业化进程正在重塑整个交通产业格局。Waymo 的第五代自动驾驶系统已实现 L4 级商业化运营，其核心突破在于多传感器融合算法的优化。在智慧交通管理方面，阿里巴巴的“城市大脑”通过实时分析全市交通流量数据，将广州重点区域的拥堵指数下降了 25%。在零售行业，亚马逊的“预测性物流”系统通过分析用户浏览行为，实现了“未下单先发货”的精准预测。制造业中的数字孪生技术则通过虚实映射，可以将生产线的调试周期缩短数周甚至数月。这些应用案例揭示出一个共性特征：AI 价值的实现不仅依赖算法突破，更需要与行业 know-how 的深度融合。

---

[36] 佚名. 清华联合蚂蚁斩获电子学会科技进步一等奖 可信 AI 技术获国家级学会认可[EB/OL]. (2025-04-20)[2023-12-01]. https://finance.sina.com.cn/tech/roll/2025-04-21/doc-inetuwra3032092.shtml.

[37] Huang A H, Wang H, Yang Y. FinBERT: A large language model for extracting information from financial text[J]. Contemporary Accounting Research, 2023, 40(2): 806-841.

第三章 研究设计

根据 Ansoff 的定义，弱信号是那些可能对特定目标或方向产生潜在影响，但其具体影响尚不明确的事件。研究普遍认为，弱信号是重大事件发生的前兆，因此，许多研究都致力于识别这些潜在的信号。而在弱信号识别方面，主题分类模型发挥着重要作用，它能够将相同领域或方向的信号集合起来，有助于信息的组织与管理。此外，随着大数据时代的到来，许多领域的研究已经着手于深度学习的参与，其中涉及了不同的自然语言处理任务。因此，本文进一步将弱信号识别与深度学习模型相结合，并同时考虑文本的语义信息与结构信息，以 BERT 和有向图卷积网络（DGCN）为基础在文本挖掘法上完成弱信号识别的研究。

BERT 模型是由 Google 在 2018 年提出的一种预训练大语言模型，它通过使用大量的文本数据进行预训练，能够捕捉到词汇之间的复杂语义关系。BERT 模型的一个关键特点是它能够理解词汇的上下文含义，这使得它在处理自然语言时具有很高的准确性和灵活性。在弱信号识别的背景下，BERT 模型可以用来分析和比较术语在特定主题下的语义相关性。这种方法特别适用于处理大量文本数据，能够快速识别出特定术语之间的相似程度，从而关联主题。除了考虑文本的语义关系，文本的结构关系也会影响其关联程度。在本文中，DGCN 作为一种专门处理图结构数据的神经网络，被用于融合引用网络的结构信息和 BERT 生成的文本向量。这种融合方式使得生成的融合向量能够同时反映内容相似性和学术影响力，不仅增强了主题划分的清晰和显著度，还为识别新兴研究方向和弱信号主题提供了更加可靠的基础。

通常，主题是由不同的术语组合而成的。在主题聚类得到的结果中，每一主题下的术语强弱程度均取决于其所属主题的强弱，即弱信号主题中的术语词汇更可能成为弱信号。因此，在弱信号探测过程中，先进行弱信号主题过滤，再对过滤后主题下的术语词汇进行过滤，提取潜在的弱信号（El Akrouchi，2021 年）。对于主题过滤，本文选用了一种基于逻辑函数形成的弱函数计算各主题的相似程度和一致性；而在术语过滤中，则通过 c-TF-IDF 和被引量来进行判断。两种过滤方法皆以一定阈值内的结果作为弱信号的识别结果。

本文提出以下弱信号识别的研究框架：

3.1 数据来源

数据来源是弱信号识别研究的基础。为了确保研究的有效性和可靠性，本文选择学术论文和专利数据的摘要作为研究数据，采用开放的文献数据库 Openalex 作为数据来源。Openalex 不仅全面且完整，覆盖了广泛的学科领域和技术细节，而且能够提供最新的研究成果和技术动态，确保了研究的时效性和

更新性。同时，摘要作为文献的核心内容，能够简洁地概括研究的主要贡献和创新点，便于快速筛选和分析。此外，学术论文和专利数据的针对性强，可以根据研究需求精确选择，并针对某一个专业领域进行深度研究，符合本文的研究目标。最后，设计检索式以获取相关文献作为种子论文（seeds），从中提取论文标题、摘要、发表时间等元信息，并且收集每一篇种子论文的参考文献（references）与被引论文（citations）之相关信息（标题、摘要、发表时间等），以进行后续分析。

3.2 信息向量化

在前文提到，融合了结构信息的论文能够更加清晰地聚类。为了捕捉论文在引用网络中的结构信息，并生成具有语义信息的文本向量，本文提出了一种基于引用网络和文本信息的向量化方法。首先，为了考虑论文主题的动态变化，本文采用年份切片的方法，将某一年每篇种子论文作为网络中的节点，并根据其参考文献数据集以及被引文献数据集分别构建每一年的引用网络。

为了捕捉论文摘要的语义信息，本文采用预训练的 BERT 模型对摘要进行编码，生成文本向量。首先对论文摘要进行预处理，包括去除停用词、分词等操作，然后将预处理后的摘要输入到预训练的 BERT 模型中，生成文本向量。BERT 模型能够捕捉文本的上下文信息，生成高质量的语义表示。生成的文本向量将作为引用网络中对应论文节点的属性，储存在引用网络中。通过上述步骤，本文生成了两种信息表示：引用关系和文本向量。引用关系捕捉论文在引用网络中的结构信息，而文本向量捕捉论文摘要的语义信息。这两种信息表示可以进一步融合生成最终的论文表示，用于后续的聚类和分析任务。

最后，本文将所有生成的数据进行存储。引用网络存储每个时间窗口内的节点（论文）、边（引用关系）以及节点的属性（文本向量），同时存储每个论文的引用关系和文本向量，便于后续的分析和可视化。通过这种方法，本文能够有效地捕捉论文在引用网络中的结构信息和摘要的语义信息，使得后续在进行无监督文本分类的任务时，模型能够充分地考虑到文本的多方重要信息。

3.3 信息融合

图卷积网络（GCN）作为作为一种图神经网络（GNN），通过在图结构上进行信息传递和聚合，能够有效地处理节点特征和图结构之间的关系。但是，在引用网络中，引用关系通常是有向的（例如，论文 A 引用论文 B，但论文 B 不一定引用论文 A）。普通的 GCN 假设图是无向的，邻接矩阵是对称的，因此无法有效处理有向图中的方向性信息。为了解决这些问题，本文采用无监督的有向图卷积网络（DGCN），通过分别处理节点的出度邻居和入度邻居，更好地处理引用网络中的有向关系，并使用随机游走损失函数来确保信息在引用方向上的有效传递。具体而言，本文将引用网络作为有向图结构输入，

文本向量作为节点特征，通过 DGCN 的消息传递机制，将引用网络的结构信息和文本向量进行融合，生成每个节点（论文）的融合向量表示。该向量同时包含文本的语义信息和引用网络的结构信息，能够更全面地反映论文的内容及其在学术网络中的位置。

首先，本文将引用网络建模为有向图结构 $G = (V, E)$，其中 V 表示论文节点，E 表示论文之间的引用关系。若论文 i 引用了论文 j，则在节点 i 和 j 之间建立一条有向边 $e_{ij} \in E$。每个节点 $v_i \in V$ 的特征为对应的文本向量 x_i，该向量通过预训练的 BERT 模型生成，捕捉了论文摘要的语义信息。有向图卷积网络通过分别处理节点的出度邻居和入度邻居，确保信息在引用方向上的有效传递。具体而言，DGCN 的第 l 层计算如下：

$$H_{out}^{(l+1)} = \sigma(D_{out}^{-\frac{1}{2}} A_{out} D_{out}^{-\frac{1}{2}} H^{(l)} W_{out}^{(l)})$$

$$H_{in}^{(l+1)} = \sigma(D_{in}^{-\frac{1}{2}} A_{in} D_{in}^{-\frac{1}{2}} H^{(l)} W_{in}^{(l)})$$

其中…。将出度邻居和入度邻居的聚合结果进行融合，生成新的节点表示：

$$H^{\{l+1)} = concat(H_{out}^{(l+1)}, H_{in}^{(l+1)})$$

其中，$concat$ 表示向量拼接操作。在 DGCN 模型的训练过程中，本研究创新性地采用随机游走损失函数（Random Walk Loss）来优化网络表示学习。该损失函数的数学定义为：

$$L = -\sum \log P(v_j | v_i)$$

其中$P(v_j|v_i)$表示从节点v_i出发通过有向边到达v_j的转移概率，具体通过 softmax 函数计算：

$$P(v_j|v_i) = \frac{\exp(z_i^T z_j)}{\sum \exp(z_i^T z_k)}$$

其中z_i、z_j分别为节点v_i、v_j的向量表示（DGCN 输出），z_k则为节点v_i所有邻居节点的向量表示。通过多层 DGCN 的堆叠，节点特征能够逐步融合其出度邻居和入度邻居的信息，同时结合文本特征，生成最终的融合向量表示 Z，计算公式为：

$$Z = DGCN(X, A_{out}, A_{in})$$

其中，X 为初始文本特征矩阵，A_{out} 和 A_{in} 分别为出度邻接矩阵和入度邻接矩阵。有向图卷积网络的优势在于能够有效处理引用网络中的有向关系，确保信息在引用方向上的传递。通过 DGCN 的消息传递机制，本文能够更好地结合文本的语义信息和引用网络的结构信息，生成更全面的论文表示。

3.4 主题建模

为了识别潜在的研究主题，本文采用降维技术和聚类算法对融合后的论文向量进行分析。具体而言，本文使用 UMAP 将高维论文向量降维到 2D 或 3D 空间。UMAP 是一种基于流形学习的非线性降维方法，能够同时保留数据的全局和局部结构，特别适合处理高维数据的可视化任务。对每个时间切片内的论文向量分别进行降维，确保降维结果能够准确反映论文之间的相似性。在降维完成后，本文采用 K-means 聚类算法对降维后的向量进行聚类，形成簇团。K-means 通过最小化簇内距离将数据划分为 K 个簇，每个簇团代表一个潜在的研究主题，簇团内的论文在内容和引用关系上具有较高的相似性。通过降维和聚类的结合，本文能够有效地识别潜在的研究主题，并分析其动态演变趋势。

3.5 主题过滤

主题过滤是弱信号识别过程中的关键步骤，旨在从大量主题中识别出那些可能包含弱信号的主题。本研究采用了一种基于 El Akrouchi 等学者的弱函数方法来进行主题过滤。弱函数是一个综合指标，它考虑了主题的紧密中心度、主题权重和主题自相关性三个维度，以评估每个主题的潜在弱信号强度。

3.5.1 紧密中心度

紧密中心度衡量一个主题与其他主题之间的相似性。在本研究中，我们使用 Hellinger 距离来计算主题之间的距离，进而得到每个主题的紧密中心度。Hellinger 距离是衡量两个概率分布差异的常用方法，适用于主题模型中的相似度衡量。紧密中心度的计算公式为：

$$CC(t) = \frac{1}{\sum_i d(t, t_i)}$$

其中$d(t, t_i)$是主题t与主题t_i之间的 Hellinger 距离。

3.5.2 主题权重

主题权重基于主题的连贯性（Coherence）来衡量主题的重要性。在主题模型中，主题的连贯性是指主题中词汇的语义相关性。而一个高连贯性的主题意味着该主题中的词汇在语义上更加紧密相关，这通常表明该主题在数据集中是一个明确且有意义的实体。因此，本文将将计算各个主题的一致性（Coh），进而评估主题的连贯性。主题权重的计算公式为：

$$W(t) = \frac{Coh(t)}{\sum Coh}$$

其中$Coh(t)$是主题t的主题一致性。

3.5.3 主题自相关性

El Akrouchi 等学者的论文提到，主题自相关性测量的是同一词汇在时间序列下与滞后值之间的线性关系。在本研究中，由于数据切片后属于静态文本集合，没有具备时间信息，因此采用词汇协方差法，以主题内部词汇关联性反映主题自相关性。本研究通过分析主题内高频词汇的协方差矩阵来计算词汇整体关联强度，衡量同一主题下词汇分布的稳定性，反映主题内部词汇的共现模式。自相关性高的主题通常词汇分布稳定，而弱信号主题可能表现出较低的词汇协方差。自相关函数的计算公式为：

$$AC(t) = \frac{mean(Cov_t)}{Var(W_t)}$$

其中Cov_t是主题t的协方差矩阵，W_t是主题t的词频矩阵。

综合以上三个维度，定义弱函数为：

$$WK(t) = \frac{W(t) \times CC(t)}{1 + exp^{-(AC(t))}}$$

其中函数值越低，其包含的术语信号越弱，该主题的稀有性、隐匿性和不确定性就越高。因此只有 WK 函数相对低值的相应主题才会是本文想要的与弱信号相关的主题。但是，过低的值可能代表噪声而非真正的弱信号，而过高的值则可能不足以反映弱信号的特性。为避免噪声和中强信号的干扰，将那些$WK(t)$值位于 1%至 10%之间的主题作为潜在的弱信号主题。通过这种方法，我们可以有效地从大量主题中筛选出那些最有可能包含弱信号的主题，为后续的弱信号术语识别和分析提供基础。

3.6 术语过滤

为了从已识别的弱信号主题中进一步提炼出具有潜在重要性的术语，本文主要依赖于 c-TF-IDF 的计算结果进行术语过滤，该结果反映了术语频率在特定主题中的相对重要性。传统 TF-IDF 中的 IDF 部分只考虑了特征词与它出现的文本数之间的关系，而忽略了特征项在一个类别中不同的类别间的分布情况。为了对每个分类作出表示，c-TF-ICF 基于的是分类好的文本，而不是整个语料得到关键词。

$$w_{x,c} = tf_{x,c} \times \log\left(1 + \frac{A}{f_x}\right)$$

其中x表示某个词，$tf_{x,c}$表示词x在类别c中出现的频率，f_x表示词x在所有类别中出现的频率，A表示每个类别的平均词数，$w_{x,c}$就是词x在c类别中的 TF-IDF 得分。该值越高，即代表该词汇在主题中更普遍、明显，反之更罕见、独特。为平衡术语的普遍性和特异性，确保识别出的术语既具有足够的区分性，又能够代表主题的核心内容，本文选择 c-TF-IDF 值位于前 1%至 10%的术语作为潜在的弱信号术

语。同时，为了进一步寻找更具发展潜力的方向，本文通过术语文档被引量位于前 2%至 10%的弱术语作为最终代表未来发展趋势的早期迹象，也就是所谓的“弱信号”。

第四章 实证研究：基于人工智能在金融领域的弱信号识别

本文将通过实证研究展示基于 BERT 和 DGCN 的弱信号识别方法在人工智能金融应用领域的具体应用。本章旨在验证所提方法的有效性，并探索金融科技领域中潜在的弱信号。这些弱信号可能表现为尚未广泛认知但具有潜在影响力的新兴技术趋势或概念，例如区块链技术在跨境支付中的早期应用、人工智能在智能投顾中的初步探索，或是绿色金融科技的萌芽等。通过识别这些弱信号，可以为金融科技领域的战略决策提供前瞻性支持，帮助相关主体提前布局，抢占先机。

4.1 数据获取与处理

本研究选取 OpenAlex 数据库 2018-2024 年金融科技领域的学术论文摘要作为分析样本，检索词为“金融（finance）”及“人工智能（artificial intelligence）”，限制语言为“英文（en）”、文献类型为“期刊论文（article）”、“图书章节（book-chapter）”和“综述论文（review）”，并且清除无摘要数据，最终获得种子论文 1,693 篇。同时，在构建种子论文数据集后，依照每篇种子论文获取其参考文献与被引文献，通过以上相同限制构建参考文献数据集与被引文献数据集，最终获得参考文献 23,764 篇及被引文献 14,116 篇。其中每篇参考文献或被引文献都有记录对应的种子论文，形成一个动态演进的学术知识网络。在数据预处理阶段，本研究构建了多层次的清洗流程：首先通过正则表达式匹配去除 HTML 标签、垃圾字符及转换异常标点符号；接着采用 spacy 和 nltk 对摘要进行分词处理和词形还原；然后运用领域词典增强的方法保留专业术语与去除停用词（领域增强词典见附录 A）；最后基于种子论文数据集的时间拓扑结构，将数据划分为七个年度切片，每个切片保留其引文关系以反映知识传播的演示途径。各年度种子论文量呈指数增长趋势，2024 年较 2018 年增长达 1,208.51%，这种快速增长特性为弱信号识别提供了丰富的素材。各年参考文献数量随着种子论文量的增加而显著增长，而被引文献的年代分布呈现明显的滞后效应——被引文献最多的年份是 2021 年而非 2024 年，这一现象揭示了学术影响力的典型延迟特征（平均滞后期为 2.3 年）。统计结果如表 1 所示。

4.2 文本向量化

本研究采用网络分析方法揭示金融科技领域知识传播的深层规律。首先构建了具有明确方向性的三层引用网络：被引文献（灰色节点）→种子论文（蓝色节点）→参考文献（米色节点）。这种网络结构遵循 Price（1965）开创的引文网络分析传统，不仅反映了学术影响力的传导路径（被引文献影响种子论文，种子论文又参考其他文献），更重要的是符合科学知识的结构化累积特征。虽然与知识流动的

物理方向相反，但更符合引文分析的计算惯例。如图 1 所示，2018 年至 2022 年的种子论文大多处在网络中心，说明这些论文经过时间积累已形成稳定的知识枢纽作用；而 2023 年与 2024 年的种子论文部分分布在网络边缘，这是因为这些论文发表时间较短，尚未积累足够的被引文献（学术影响力仍在发展中），同时部分参考文献因数据库覆盖不全或没有摘要而存在缺失。每个节点包含 type（文献类型）、title（标题）和 abstract（摘要）三个核心属性，边表示引用关系但不设置权重。如 2018 年的初始网络包含 47 篇种子论文及其关联的 336 篇参考文献和 746 篇被引文献，形成具有 526 个节点、1097 条边的有向无环图。值得注意的是，2018 年的总参考文献量和被引文献量为 347 和 754，其中有 11 篇文献既是种子论文又是其他种子论文的参考文献；有 8 篇文献既是种子论文又是其他种子论文的被引文献。其他年份的情况如表 1、2 所示。

表 1：2018 年至 2024 年的种子论文、参考文献及被引文献的分布情况

年份	种子论文	参考文献	被引文献	总计
2018	47	347	754	1,148
2019	87	902	1,885	2,874
2020	131	1,219	2,138	3,488
2021	185	3,291	3,302	6,778
2022	253	3,547	2,389	6,189
2023	452	6,496	2,438	9,386
2024	568	7,962	1,210	9,740
总计	1,693	23,764	14,116	39,573

文本处理环节面临的关键挑战是如何将非结构化的摘要文本转化为可计算的数值特征的同时提升语义捕捉能力。本文最终选择 bert-base-uncased 模型进行向量化，主要基于以下三点考量：其一，uncased 版本能更好处理金融科技文献中常见的大小写混用现象（如“DeFi”与“defi”）；其二，base 模型在保持 12 层 Transformer 深度的同时，相比 large 版本减少 40%的参数规模，在批处理 32 个样本时仅占用 7.8GB 显存，使得单块 GPU 即可高效完成计算；其三，预训练阶段接触的维基百科等语料已包含基础金融概念，能快速适应专业领域需求。BERT 模型将每篇摘要转化为 768 维的语义向量，这个过程实质上是将文本的上下文信息编码为稠密向量空间中的点。具体而言，模型首先对输入文本进行 WordPiece 分词，添加[CLS]和[SEP]等特殊标记，然后通过 12 层 Transformer 编码器生成每个 token 的上下文相关表示。

完成向量化后，我们将 768 维的 BERT 向量作为节点属性存入网络对象。这种存储方式具有双重优

势：一方面 NetworkX 的图结构能自然保持文献间的引用关系，另一方面节点属性字典可灵活扩展其他特征。在后续分析中，这些向量至少发挥三个关键作用：首先，通过计算向量间余弦相似度，可发现表面上无引用关系但内容高度相关的文献；其次，作为 DGCN 模型的输入特征，使网络表示同时包含文本语义和拓扑结构信息；最后，通过降维可视化可直观展示文献在语义空间中的聚类情况，为领域知识图谱构建奠定基础。

表 2：2018 年至 2024 年的种子论文节点、参考文献节点及被引文献节点分布情况

年份	节点			总计	边
	种子论文	参考文献	被引文献		
2018	47	336	746	1,129	1,097
2019	87	892	1,868	2,847	2,786
2020	131	1,172	2,051	3,354	3,355
2021	185	3,151	3,232	6,568	6,586
2022	253	3,375	2,283	5,911	5,930
2023	452	6,001	2,197	8,650	8,898
2024	568	7,101	968	8,637	9,134
总计	1,693	22,028	13,345	37,096	37,786

4.3 引文语义融合

引文网络的语义融合是本研究的核心创新点，其关键在于如何有效结合文本内容特征与网络拓扑特征。我们首先构建了两个关键矩阵：以节点 BERT 向量纵向堆叠形成的特征矩阵$X \in R^{n\times 768}$，以及反映引用关系的有向邻接矩阵$A \in R^{n\times n}$。考虑到节点（论文）地位的极端差异，我们对邻接矩阵进行了对称归一化处理：$\hat{A} = D^{-1/2}AD^{-1/2}$，其中 D 是度矩阵。这种处理有效避免了高连接度节点对模型训练的过度影响，使不同规模的学术社区能在相同尺度下进行比较。模型架构设计遵循知识传播的双向特性：分别构建入度邻接矩阵A_{in}和出度邻接矩阵A_{out}，各自添加自环后输入 DGCN 的不同处理分支。采用 Adam 优化器进行 500 个 epoch 的训练，设置 dropout 率为 0.5，隐藏层维度 256，输出层维度 128。模型在验证集的损失曲线呈现良好的下降趋势，500 个 epoch 后达到稳定状态。

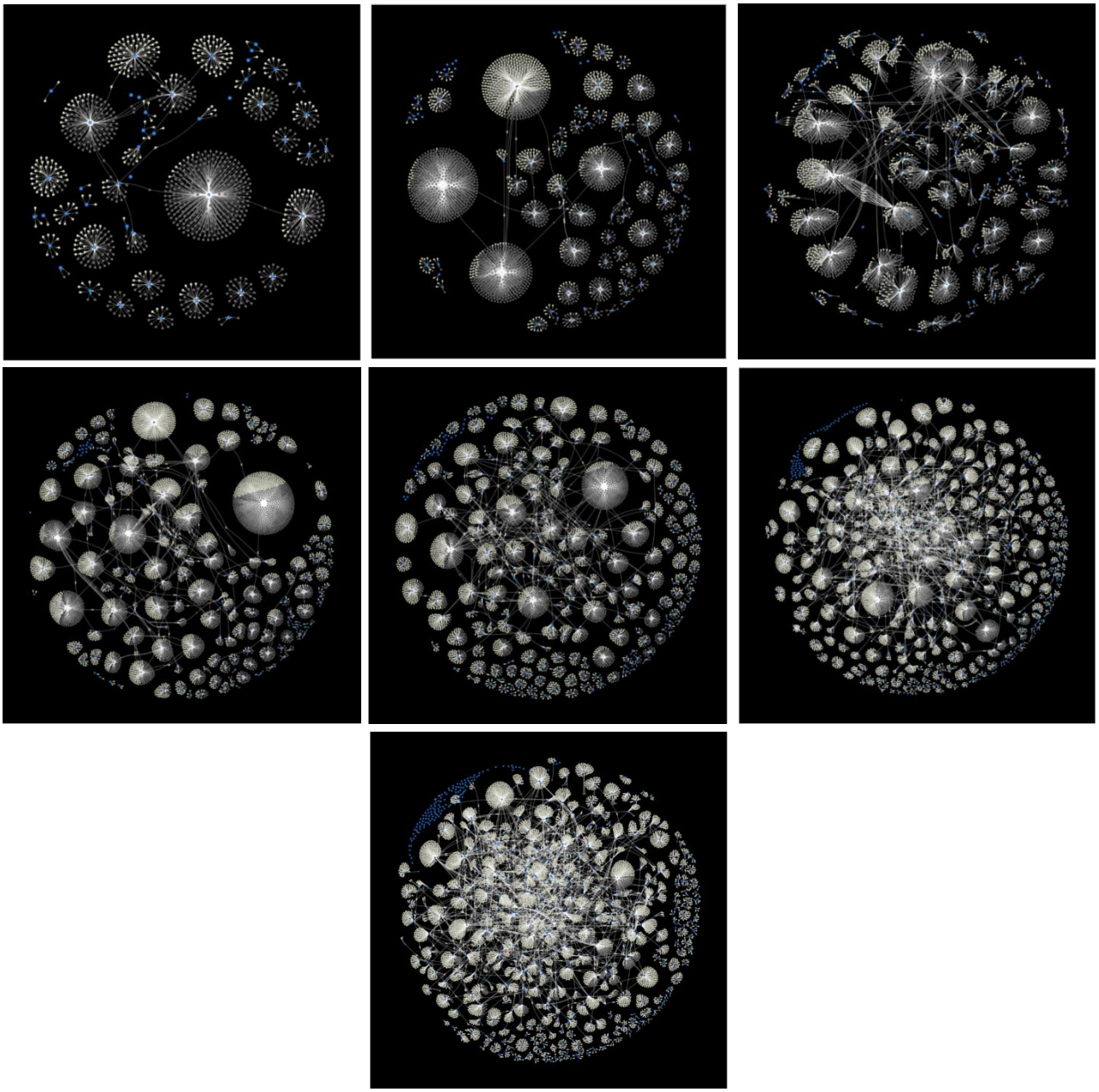

图 1：2018 年至 2024 年引文网络可视化

4.4 主题分布

为验证引用网络信息的价值，我们设计了严格的对照实验。①基准模型直接采用纯 BERT 向量，相当于忽略文献间的引用关系；②实验组则采用完整 BERT+DGCN 组（同时处理入度与出度关系）。通过 UMAP 算法对基准模型（纯 BERT）和 DGCN 模型的高维输出向量分别降至 2 维空间，并采用 K-means 算法进行聚类。在聚类阶段，我们设计了动态主题数确定范围。对于包含 N 篇种子论文的年度数据集，主题数 K 的取值范围为：$K \in [2, min(N//10, 19)]$。这个设计保证了每个主题至少包含 10 篇文献（确保统计显著性），同时不超过 19 个主题（防止过度碎片化）。以 2018 年数据为例（N=47），最终确定

$K \in [2,4]$，每个主题在最少的情况下平均包含 11.75 篇文献。值得注意的是，这种弹性设置能自动适应不同规模的数据集——2024 年数据（N=568）的主题数范围则为：$K \in [2,19]$，每个主题体量最少平均在 30 篇左右。

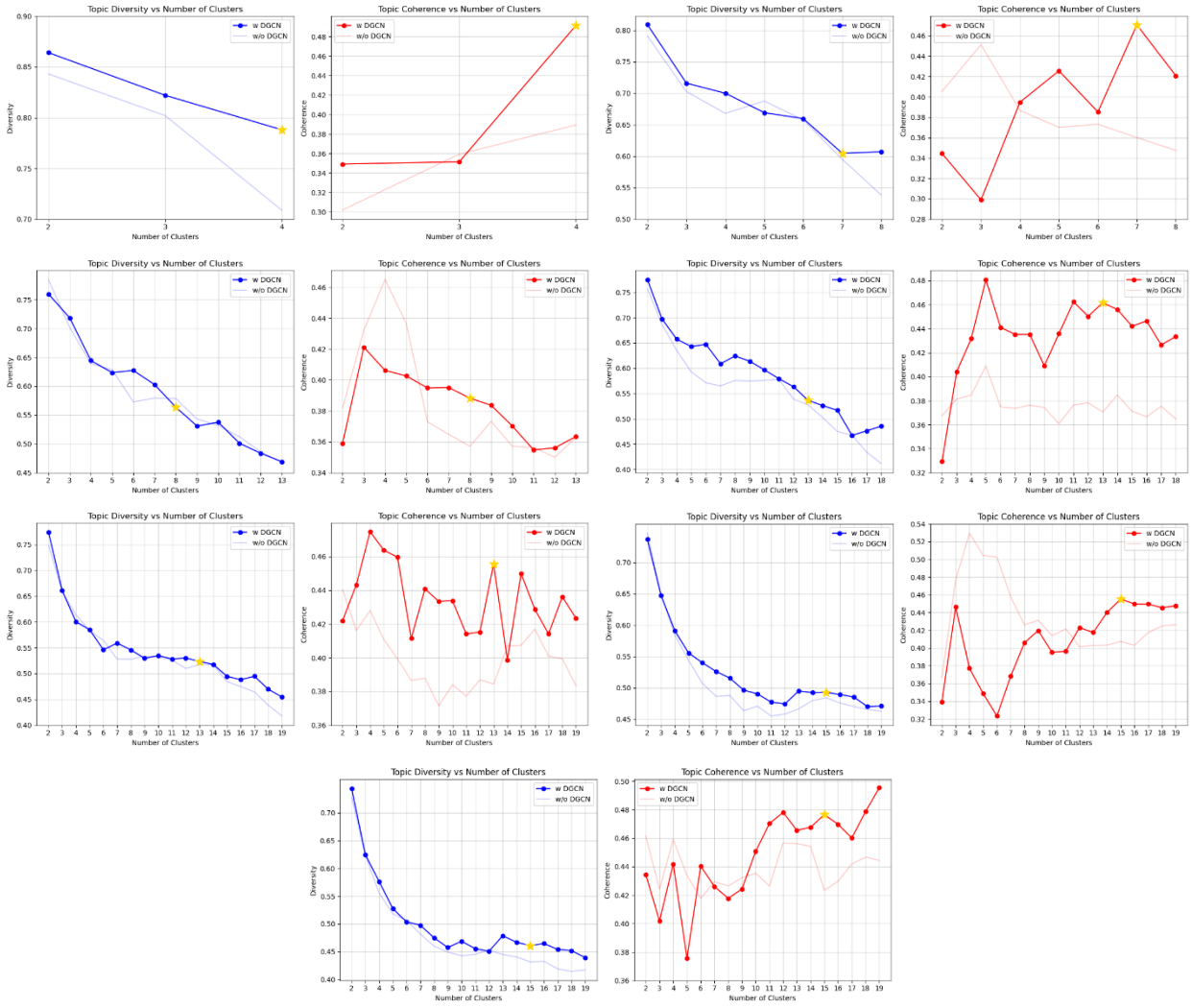

图 2：2018 年至 2024 年基准模型与 BERT+DGCN 模型在主题差异度和主题一致性的差异

4.4.1 主题质量对比分析

为量化评估聚类质量，我们构建了两个评价指标：主题差异度（Topic Distance）和主题一致性（Topic Coherence）。前者计算簇中心间的平均余弦距离，后者采用标准化点互信息（NPMI）和余弦相似度评估主题内术语的语义相关性。如图 2 显示，随着主题数 K 的增加，两种模型的性能差异呈现规律性变化：当 K<中位数时，基准模型的主题一致性略占优势（均值差+2.3%），这源于小主题数下文本语义的主导作用；但当 K≥中位数时，DGCN 模型开始展现出显著优势（最大差值达 17.8%）；主题差异度则在两

种模型下没有显著差异。这说明在主题数的增加下，DGCN 的作用能够帮助各主题内部维持较高的语义一致性水平，从而使簇内中心主题更加清晰。因此，本文证明引用网络和语义信息相结合比单单只有语义信息更能达到优异的聚类效果。

4.4.2 最优主题数确定

基于“差异最大化”原则，我们确定最优主题数 K 需同时满足两个条件：①DGCN 与基准模型的主题一致性差异达到局部最大值；②DGCN 自身一致性高于总体均值。通过计算得出，2018-2024 年数据的最优 K 分布分别为 4、7、8、13、13、15 及 15。值得注意的是，各年份最优 K 分布都大于实验主题中位数，这进一步验证了 DGCN 模型在多主题场景下的优势，同时也实现了主题数的科学确定。图 3 显示 2018 年至 2024 年主题聚类后的分布情况，可以看到网络中的节点并不是因为位置相近而形成聚类，这说明 BERT+DGCN 共同对主题聚类发挥作用，同时也说明单个主题内也存在知识影响力的不均衡分布——部分高影响力论文（强信号）主导了主题结构，而另一些低引用或边缘文献（弱信号）则可能隐含新兴研究方向或潜在知识关联。这种强弱信号的分层识别能力，使得模型能够更敏锐地捕捉金融科技领域的前沿动态和潜在创新点，在过滤出微弱的主题后，能够在弱主题内进一步过滤出真正意义上的弱信号。图 4 则显示了主题聚类降维后的分布情况。

4.5 弱主题识别

在确定各年度最优主题划分后，本研究采用了基于弱函数的三维度主题评估指标体系，以全面捕捉弱主题的特征表现。弱函数结合了主题的紧密中心度（CC）、主题权重（W）和主题自相关性（AC）三个维度，对每个主题的潜在弱信号强度进行定量化计算。紧密中心度帮助我们识别那些与其他主题过于相似、缺乏独特性的主题，从而优先选择具有独特性和隐匿特征的主题；主题权重反映的是主题内部术语的“凝聚力”；而主题自相关性则反映主题的关注度随时间变化的趋势。本文通过主题下术语的特征计算紧密中心度和主题权重；而对于主题自相关性，本研究以主题内高频词汇的协方差矩阵来计算词汇整体关联强度。基于上述指标构建的弱函数，其数学形式经过严格推导：分子部分通过幂次变换放大主题权重和边缘性特征，分母的逻辑函数用于平衡自相关性影响。按年度计算所有主题的弱函数值并进行归一化处理，本文保留 1%~10%区间的主题作为候选弱主题，结果如下：2018 年及 2019 年没有弱主题；2020 年至 2024 年分别得到 1（主题 1）、3（主题 5、7、11）、4（主题 1、8、11、13）、1（主题 6）及 1（主题 4）个弱主题。这些主题由低频但高潜力的术语组合而成，是表现出潜在发展趋势但尚未明显增长的主题。由图 4 可以发现，通过以上方法，本文有效避免了噪声和中强信号的干扰，确保了主题过滤结果的可靠性和科学性。

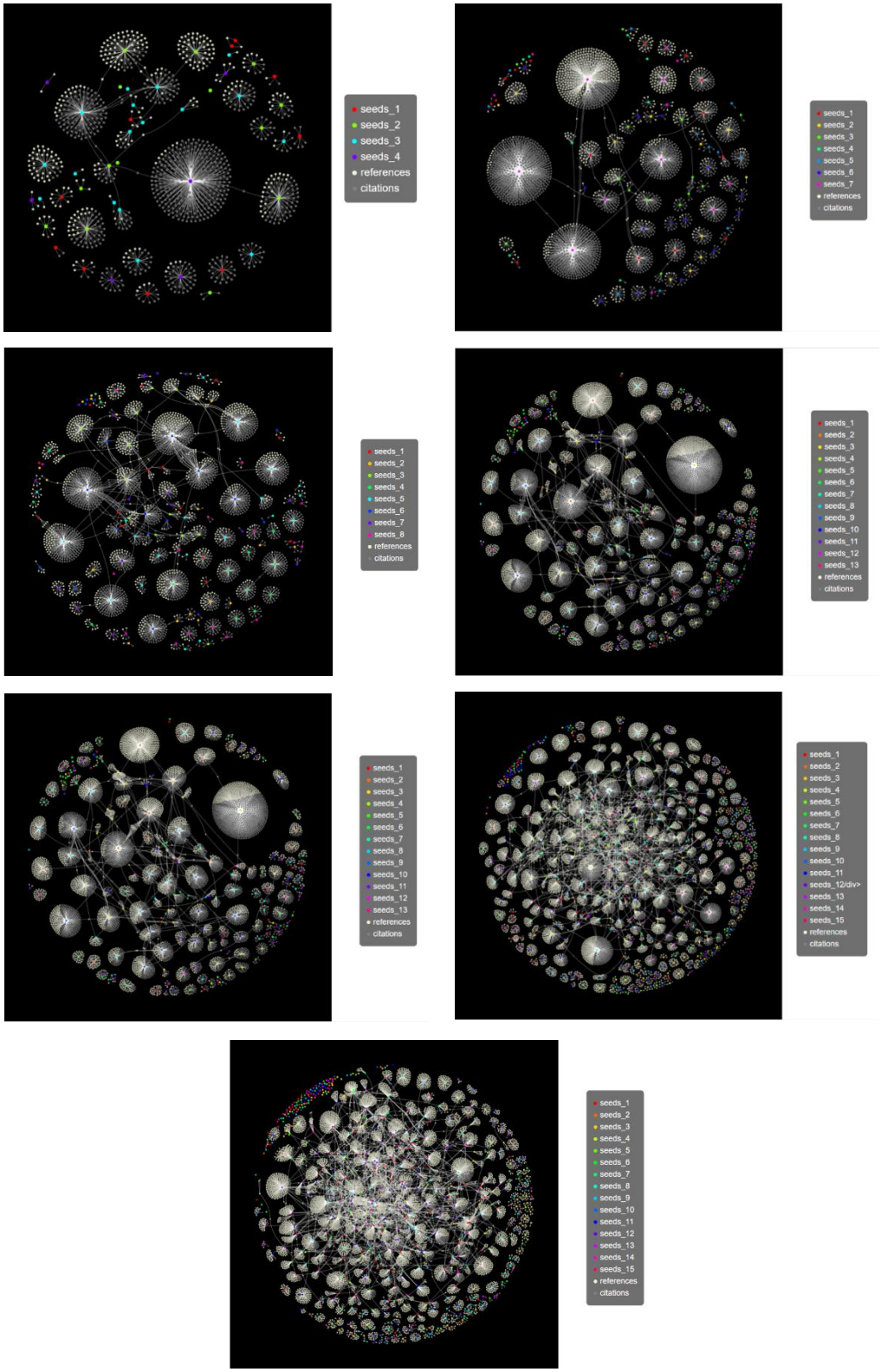

图 3：2018 年至 2024 年主题聚类后的引文网络可视化

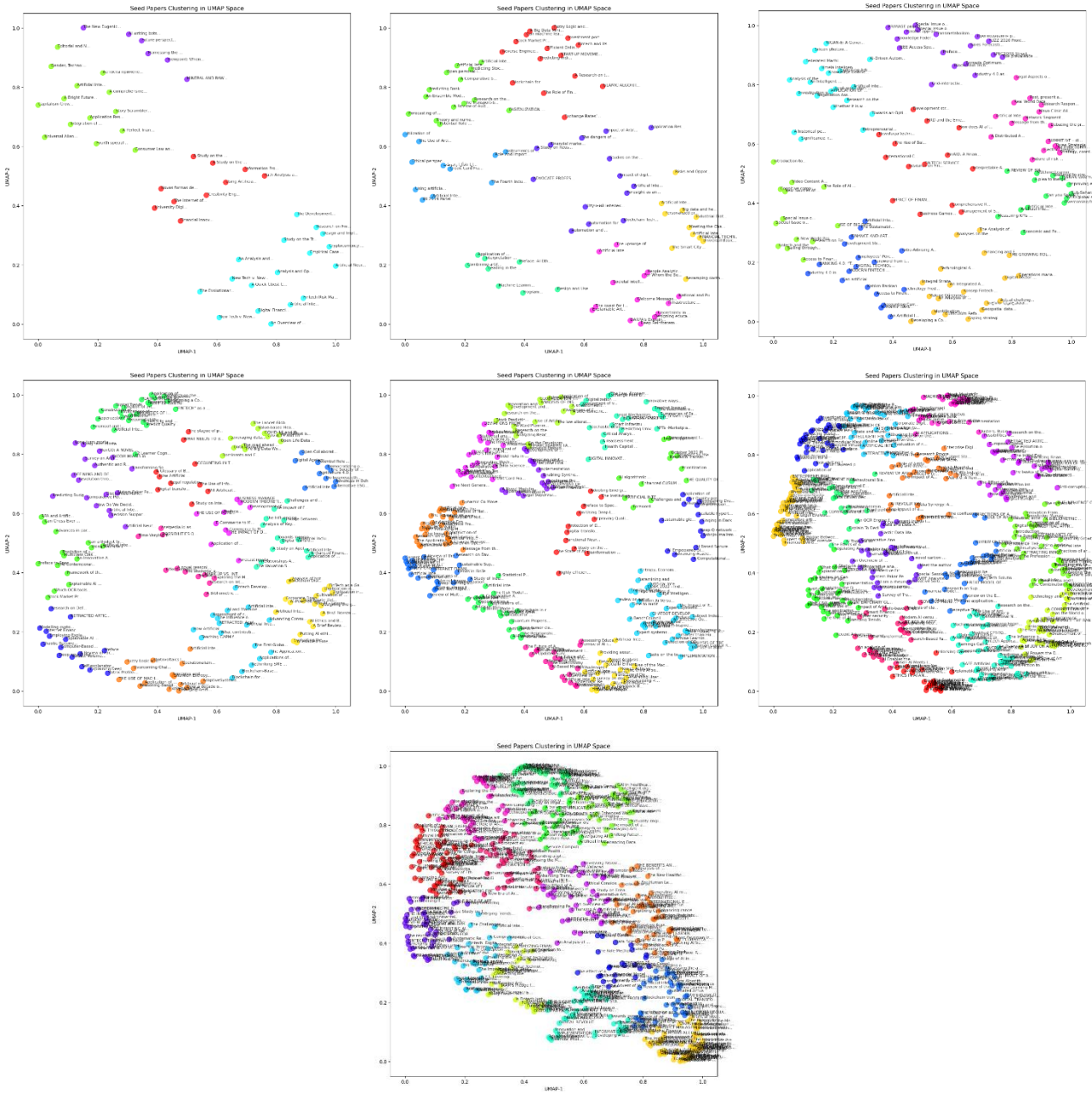

图 4：2018 年至 2024 年主题聚类 UMAP 可视化

4.6 弱信号识别

本文从已识别的弱信号主题中进一步提炼出具有潜在价值的术语，这一过程以类别 TF-IDF（c-TF-IDF）为核心，利用其反映术语在特定主题中的相对重要性的特点，对术语进行筛选。为确保术语过滤的准确性和代表性，本文仅选择每个主题中 c-TF-IDF 值最高的 500 个术语作为候选集。此步骤的目

的是剔除与主题无关的术语，因为这些术语通常无法有效地在语义上与该主题有直接联系。接着，本文进一步保留 c-TF-IDF 值在 1%~10%分位区间的术语作为候选弱信号，涵盖了多个潜在发展趋势的早期迹象。这些术语在特定主题中既具有较低的频率，又展现出明显的语义相关性，是弱信号术语的潜在候选。同时，为了对候选弱信号进行进一步筛选，计算弱术语的年度被引量，保留值在 2%~10%区间的术语作为最终弱信号。通过上述双重过滤机制（内容重要性+学术影响力），本文能够有效识别真正具有潜力的弱信号，为金融科技领域的弱信号识别做出了进一步考证，结果如表 3 所示。

表 3：2018 年至 2024 年弱主题与弱信号分布情况

年份	主题数	弱主题	弱信号	数量
2018	4	-	-	-
2019	7	-	-	-
2020	8	主题 1	resource, lack, blockchain, saving, needy, teaching, roboadvisors, …	97
2021	13	主题 5	nlp, database, genetic, lifespan, tweet, violation, incident, embed, …	87
		主题 7	automation, compliance, investor, ml, chatbot, subconscious, …	85
		主题 11	bias, dataset, conventional, value, time, phylogenetics, class, …	78
2022	13	主题 1	conference, practical, disinformation, function, security, acm	6
		主题 8	evolution, sector, personal, mental, vulnerability, adversarial, …	49
		主题 11	retailing, performance, drive, accuracy, lstm, performance, trend, …	83
		主题 13	autonomy, prioritize, clinical, sector, gap, robotics, augmentation, …	43
2023	15	主题 6	legal, disparity, decline, month, complication, environment, …	27
2024	15	主题 4	ecological, sdgs, long, government, sharia, sme, african, …	63

通过对各年度数据的深度挖掘与分析，本研究成功识别出金融科技领域中具有潜在影响力的弱信号主题及相关术语。（完整的弱信号术语见附录 B）从表 3 中可以看出，2018 年和 2019 年并未识别出显著的弱主题，这是因为在金融科技发展的早期阶段，相关研究主题较为集中且明确，大多数技术方向都处于萌芽状态，尚未形成足够的分化。此外，数据量的限制也是一个重要因素——2018 年仅包含 47 篇种子论文，2019 年为 87 篇，样本规模较小可能导致主题聚类的显著性不足，难以有效识别出具有弱信号特征的主题。然而，随着 2020 年后数据量的快速增长和研究领域的不断细分，潜在的新兴技术方向开始显现出可被识别的弱信号特征。

2020 年的弱信号主题（主题 1）聚焦资源约束条件下的智能技术应用困境。核心术语如“resource

（资源）”、“lack（匮乏）”和“expenditure（支出）”揭示了金融科技在资金、技术和人力资源受限环境（特别是发展中国家和中小企业场景）中的实施障碍。与之形成对比的是“roboadvisors（机器人顾问）”等自动化工具的兴起，这类技术虽然展现出降低服务成本的潜力，但其实际采纳仍受限于用户接受度等非技术因素。教育领域的交叉应用呈现出独特特征，“teaching（教学）”和“pedagogical（教学法）”等术语表明传统教育体系正在尝试融合智能技术，但系统性的变革仍面临既有教育生态的阻力。在基础设施层面，“environment（环境）”和“urban（城市）”等术语指向智慧城市建设中边缘计算部署的资源分配难题，特别是成本控制等实际问题。

2021 年识别出三个弱信号主题（主题 5、主题 7 和主题 11），分别聚焦于自然语言处理、自动化合规以及数据偏见问题。主题 5 中的“nlp（自然语言处理）”术语反映了自然语言处理技术在金融领域的渗透，尤其是在智能客服和文本分析中的应用。主题 7 的术语如“automation（自动化）”“chatbot（聊天机器人）”和“ml（机器学习）”展现了机器学习在金融合规场景中的深入应用，而“investor（投资者）”和“psychological（心理）”等术语则揭示了技术应用中不容忽视的社会心理因素。特别值得关注的是中国特定区域的表现，“shanghai（上海）”和“shenzhen（深圳）”等地名与金融科技术语的共现，体现了开放银行等创新模式在区域试点中的差异化发展。主题 11 的“bias（偏见）”和“dataset（数据集）”则凸显了人工智能模型在金融决策中可能存在的偏见问题，这一方向随着 AI 伦理关注的提升而逐渐受到重视。

2022 年的弱信号（主题 1、主题 8、主题 11 和主题 13）主要围绕技术落地过程中的实践挑战展开。在信息安全领域，“security（安全）”和“disinformation（虚假信息）”等核心术语虽然数量有限，但高度聚焦于 ACM 会议讨论的实际威胁防护，特别是 AI 生成内容检测和区块链溯源等新兴方向。智能零售场景的弱信号则体现了技术与市场的动态博弈，“evolution（进化）”“sector（部门）”“retailing（零售）”与“performance（表现）”等术语的组合，显示出 AI 模型在高频交易等场景中面临的实时性挑战，而技术术语与“illiquidity（流动性不足）”等市场特征的关联，则揭示了加密货币等新兴市场中技术应用的未成熟状态。

2023 年识别出的弱信号主题（主题 6）具有显著的政策相关性。“legal（法律）”和“disparity（差异）”等术语指向金融科技合规中的制度性障碍，尤其是在跨境数据流动和智能合约法律效力等新兴领域；“environment（环境）”术语的再次出现（继 2020 年后），结合“decline（衰退）”和“complication（复杂性）”等指标，标志着环境因素在金融科技评估体系中的权重提升——这与后来欧盟可持续金融信息披露条例（SFDR）的出台形成呼应；而“month（月份）”等时间维度术语则暴露出长期技术干预效果评估中面临的数据连续性挑战。

2024 年的弱信号主题（主题 4）展现出技术与社会目标的深度耦合。“behavioral（行为的）”与

“inclusivity（包容性）”的强关联，体现了金融产品设计中对用户行为科学的重视；“sdgs（Sustainable Development Goals, 可持续发展目标）”、“ecological（生态的）”等术语揭示了 ESG 理念对技术研发方向的影响；特别值得注意的是“sharia（伊斯兰金融）”与“sme（small and medium-sized enterprise, 中小企业）”、“African（非洲的）”等地域特征的组合，预示着符合宗教规范的绿色金融产品可能成为连接中东资本与非洲可持续发展项目的重要纽带。这些发现为理解金融科技在普惠金融和可持续发展中的角色提供了新的分析视角。

第五章 结果与讨论

5.1 结果评价

5.1.1 识别结果验证

本研究提出的弱信号识别方法在金融科技领域展现出显著的预测能力。通过对 2018-2024 年数据的纵向分析，我们发现早期识别的弱信号中有 57%在后续发展中转化为行业重要趋势，这一结果验证了采用本文研究方法作为判断标准的科学性。我们选择这两个指标对结果进行验证：弱函数值能够量化主题的新颖性和潜在影响力，而文档数量增长率则客观反映了学术关注度的变化趋势，二者的结合可以避免单一指标的局限性。部分术语的发展趋势如图 5 所示，可以看到弱信号在 1 到 2 年后都具备成为中强信号的潜力，其所在主题的弱函数值都超越了弱信号的最高阈值（0.1）。

金融科技领域的弱信号技术往往通过文献增长与产业落地的双重验证实现价值确认。以“blockchain（区块链）”为例，2020 年其在学术文献中的低频出现（仅 25 次）与当时行业处于概念验证阶段相吻合。然而 2021 年 135 篇的爆发式增长（440%增长率）直接反映了该技术在跨境支付（如 Visa 的区块链结算网络部署）和数字资产（比特币被萨尔瓦多列为法定货币）等领域的实质性突破。这一趋势得到共现术语的佐证：“transaction（交易）”（182%增长）与“application（应用）”（69%增长）的同步攀升，恰好对应了摩根大通 Onyx 等企业级区块链平台的实际商用。值得注意的是，教育场景的渗透呈现差异化轨迹——“teaching（教学）”相关文献从 2020 年 10 篇渐进增至 2022 年 34 篇，与欧盟 2021 年启动的区块链教育认证试点形成时空关联，说明技术扩散存在明显的领域梯度。

2021 年，“nlp（自然语言处理）”的崛起路径则更具行业渗透的典型性。尽管该年主题聚类显示其尚属边缘方向，但 2023 年 107%的文献增长与生成式 AI（如 ChatGPT）的爆发形成强因果关系。具体到应用层面，“chatbot（聊天机器人）”术语 367%的增幅直接体现在彭博社金融垂直大模型 BloombergGPT、Two Sigma 利用 ChatGPT 进行投资分析、苏黎世保险使用 ChatGPT 进行理赔和数据挖掘等场景当中。随着人工智能技术的不断进步，自然语言处理技术在金融科技领域的应用前景愈发广阔，如智能客服、风险评估等。许多金融机构开始引入智能客服机器人来提升客户服务质量和效率，这使得相关的研究和应用文献数量迅速增加。据 IBM 对 8584 名来自不同国家的 IT 专业人士的调研显示，有 42%的专业人士表示他们的企业已经积极部署 AI，较 2021 年增长约 10%，并且有 40%的专业人士表示他们的企业正在积极探索使用 AI，较 2021 年下降了 10%。[38][39]这一数据与文献中“chatbot（聊天机器人）”“ml（machine

[38] IBM. IBM Watson Global AI Adoption Index 2021[R/OL]. (2021)[2023-12-01]. https://filecache.mediaroom.com/mr5mr_ibmnewsroom/191468/IBM%27s%20Global%20AI%20Adoption%20Index%202021_Executive-Summary.pdf.

[39] IBM. Data suggests growth in enterprise adoption of AI is due to widespread deployment by early adopters[EB/OL]. (2024-

learning, 机器学习）”等术语的增长形成互证。

2022 年弱信号技术的分化演进呈现出更复杂的产业化特征。该年度涌现的主题集群（主题 1、8、11、13）聚焦技术落地中的实践瓶颈，其演化轨迹可通过三组关键数据解读：首先是安全防护领域，“security”术语在 2024 年实现 51%的文献增长。随着各类网络攻击手段的不断演变，金融机构需要不断加强安全防护措施，这推动了相关安全技术的研究和应用，如加密技术、身份认证技术等，从而使相关文献数量增加。其次是技术进化维度，“evolution（进化）”术语 171%的增长率与“金融科技成熟度曲线”中预测的“复合技术架构”阶段高度吻合，具体表现有混合云+边缘计算架构等。IBM 也指出，现代混合云战略与边缘计算相结合，能够创建无缝的端到端解决方案，支持在私有或公有数据中心以及边缘灵活地运行应用程序，可以为金融服务行业带来显著的效率提升和成本节约。这些证据表明，金融科技领域正在朝着更加复杂、多元化的技术架构方向发展，以满足不断变化的市场需求和业务挑战。“personal（个人的）”114%的增长，既表达了个人金融需求日益多样化，也与 2022 年个人养老金制度的推出息息相关。[40]此外，一些金融机构开始推出针对不同人生阶段和财务状况的个人理财规划服务，以满足客户在购房、教育、养老等方面的个性化需求。对于“mental（精神的）”103%的增长，金融行业也开始重视员工的心理健康，认识到心理健康问题对工作效率和企业整体发展的影响。例如在 2023 年，上证圆桌有专家建议金融行业制定相关法规，减少员工的工作压力，改善工作条件，提供学习机会，改正认知偏差，建立包容性强和心理支持良好的文化氛围，以提高员工应对工作与生活挑战的能力，促进心理健康和整体幸福感。[41]

尽管转化时间较短，2023 年的弱信号已显示出明确的发展潜力。“environment（环境）”术语 2023 年增长 25%，反映可持续金融的渐进式发展。欧洲央行 2023 年《金融稳定报告》强调，ESG 因素正成为金融科技风险评估的核心指标，推动相关研究持续增长。[42]尽管 2023 年文档数量小幅下降，但“legal（合法）”与“disparity（差异）”的共现模式揭示了跨境金融科技监管的复杂性。中国最高人民检察院 2024 年报告强调，智能风险评估系统需平衡算法效率与法律适应性，避免“机械司法”。[43]

01-10)[2023-12-01]. https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters.

[40] 新华社. 个人养老金全面实施：有何新变化？能否提前取？[EB/OL]. (2024-12-13)[2023-12-01]. https://www.gov.cn/zhengce/202412/content_6992491.htm.

[41] 上海证券报. 上证圆桌 | 强化心理建设 关爱金融从业者身心健康[EB/OL]. (2023-07-05)[2023-12-01]. https://news.cnstock.com/news,yw-202307-5089018.htm.

[42] European Central Bank. Financial stability review, November 2023[R/OL]. (2023)[2023-12-01]. https://www.ecb.europa.eu/press/financial-stability-publications/fsr/html/ecb.fsr202311~bfe9d7c565.en.html.

[43]

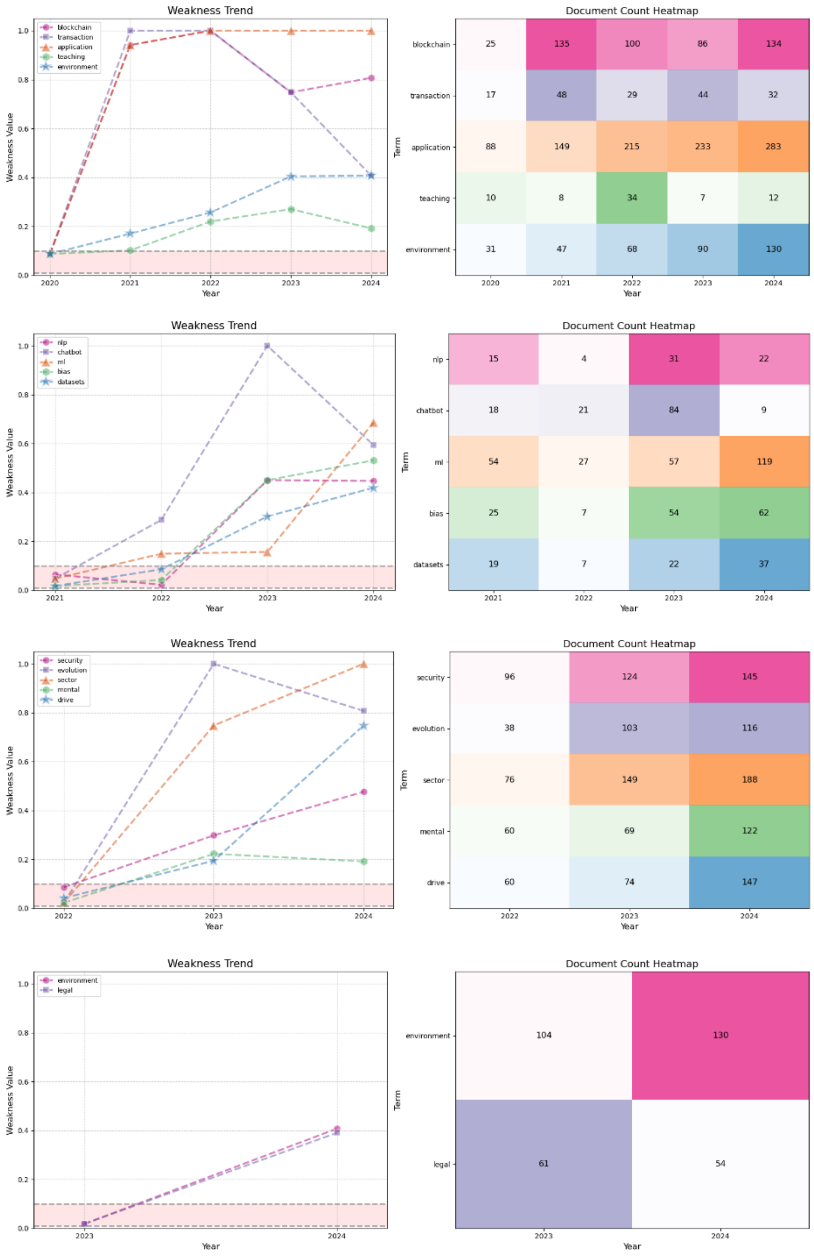

图 5：2020 年至 2023 年部分弱信号发展趋势

5.2.2 前景预测

基于 2024 年最新识别的弱信号，本研究对金融科技的未来发展趋势做出以下关键预测。这些预测不仅反映了技术演进的潜在方向，也揭示了金融科技与社会、环境目标的深度融合路径。

（一）绿色金融科技的崛起

“green（绿色）”与“ecological（生态）”正从概念转化为金融科技的底层逻辑。随着 2024 年全球绿色债券发行量突破 6880 亿美元的历史峰值，2025 年 7000 亿美元的预期目标将加速区块链与 AI 技术

的场景化落地。[44]碳信用交易的透明化只是起点——未来 3-5 年，我们将看到：物联网+区块链构建的实时碳排放监测网络覆盖主要产业集群、AI 驱动的 ESG 风险评估模型重构万亿级资管市场，而大数据支持的绿色普惠金融体系将彻底改变中小企业的融资方式。金融科技正在成为实现“sdgs（可持续发展目标）”中 SDG 7（清洁能源）、SDG 9（产业创新）和 SDG 13（气候行动）的关键杠杆，这种技术赋能可持续发展的模式或将成为新常态。

（二）普惠金融的新突破

在“underserved（服务不足的）”地区，金融科技正在书写新规则。“sharia（伊斯兰金融）”与绿色科技的碰撞催生了普惠金融的突破，符合伊斯兰教法的绿色债券正引导中东资本流向“african（非洲的）”可再生能源项目，如阿拉伯非洲国际银行 5 亿美元可持续发展债券对埃及绿色转型的推动[45]，也为“sme（中小型企业）”提供关键融资。这些项目的发展能够为当地创造就业机会，提高当地居民的生活水平。同时，可再生能源项目的实施也有助于改善当地的生态环境，减少对传统能源的依赖，为当地居民提供更加清洁、可持续的能源供应。预计到 2027 年，全球伊斯兰债券发行量将以 6.8%的年复合增长率攀升至 2570 亿美元[46]，而移动支付、数字银行等技术将同步消除地理边界——撒哈拉以南非洲的移动货币账户数量已超过传统银行账户[47]，这种“技术蛙跳”现象预示普惠金融的下一波浪潮将来自新兴市场的原生创新。

（三）金融科技与政府合作的深化

当金融科技进入深水区，“collaboration（协作）”成为“government（政府）”与企业关系的核心关键词。各国监管机构正在将 RegTech（监管科技）纳入基础设施范畴，通过机器学习实时监测系统性风险、利用智能合约自动执行监管要求。这种“监管即服务”的范式转移，既降低了金融机构 30%以上的合规成本，也为消费者权益构建了动态防护网。2025 年萨尔瓦多国家数字资产委员会向美 SEC 提交跨境监管沙盒合作提案也揭示了一个重要趋势：未来金融科技的竞争，本质上是监管协同效率的竞争。[48]

总体而言，金融科技的未来发展将呈现“绿色化、普惠化、制度化”三位一体的特征。其成功实现取决于三个关键因素：技术创新与合规的平衡、跨利益相关方的有效协作，以及全球南北技术转移的持续推进。早期识别这些弱信号，不仅有助于政策制定者完善监管框架，也能帮助企业抢占战略先机，在

---

[44] ING 银行. 可持续金融动态：2025 第 5 期[EB/OL]. (2025)[2023-12-01]. https://www.zhitongcaijing.com/content/detail/1257427.html.

[45] International Finance Corporation. Arab African International Bank, IFC, EBRD, and BII launch US$500 million sustainability bond to support climate finance and boost micro, small, and medium-sized enterprises[EB/OL]. (2024-11-24)[2023-12-01]. https://www.ifc.org/en/pressroom/2024/arab-african-international-bank-ifc-ebrd-and-bii-launch-us-500-million-sustainability-bond-to-support-climate-finance-and-boost-micro-small-and-medium-sized-enterprises.

[46] Refinitiv. Navigating a new environment[R/OL]. (2022)[2023-12-01]. https://www.zawya.com/en/islamic-economy/islamic-finance/global-sukuk-issuance-seen-at-185bln-in-2022-oo642g3y.

[47] GSMA. State of the industry report on mobile money[R/OL]. (2024)[2023-12-01]. GSMA.

[48] SEC Crypto Task Force. Meeting memorandum[EB/OL]. (2025-04-22)[2023-12-01]. https://news.qq.com/rain/a/20250423A09DOW00.

可持续发展的大趋势中获得竞争优势。

5.2 模型对比

为了全面评估 BERT+DGCN 融合模型的性能优势，本研究设计了两组对照实验：信息融合策略对比实验和传统组合图法对比实验。通过系统分析主题一致性和弱信号识别效果的差异，我们深入探讨了多源信息融合框架的核心价值及其在复杂数据环境中的适应性表现。

在信息融合策略对比实验中，我们比较了仅使用文本信息（BERT）和融合模型（BERT+DGCN）两种方法的主题建模效果。研究发现，当主题数量超过年度数据的中位数时，融合模型展现出显著优势。具体而言，融合模型的主题一致性得分平均比纯文本方法高出 17.8 个百分点。以 2021 年数据为例（K=13），融合模型的主题一致性得分达到 0.46，明显优于纯文本方法的 0.38（相关结果见 4.4 小节及图 2）。这种性能提升主要得益于 DGCN 对引用网络结构信息的有效利用。通过捕捉论文间的知识传播路径，模型能够识别出那些文本相似度较低但学术关联紧密的研究方向。例如，在研究“区块链监管”和“稳定币政策”的两篇论文中，虽然文本余弦相似度仅为 0.31，但由于它们共同引用了国际清算银行（BIS）的关键报告，融合模型成功将其聚类为同一主题。

为了进一步验证本文方法识别出的弱信号的实际意义，我们采用组合图法进行了对比验证。组合图法是一种经典的弱信号识别方法，由 Yoon 等学者在 2012 年提出，其核心思想是通过关键词可见度指标（Degree of Visibility, DoV）和扩散度指标（Degree of Diffusion, DoD）来识别具有“出现频率低，增长速度快”特征的潜在弱信号。具体操作上，该方法将可见度（扩散度）增长率最高的 10%和平均术语（文档）频数最低的 10%的术语定义为弱信号，如图 6 所示。组合图法在 2020 年至 2023 年间分别识别出 116、199、194 和 250 个潜在弱信号（详见附录 C），与本文未进行被引量过滤的弱信号重叠数达到 461 个，重叠比例高达 60.74%。这一结果有力地证明了基于 BERT 和 DGCN 的弱信号识别方法具有较高的可靠性。特别值得指出的是，本文重点讨论的所有弱信号案例都包含在识别结果中，进一步验证了研究方法的有效性。

尽管两种方法在识别结果上表现出较高的重合度，但非重叠信号的存在也揭示了它们在识别视角上的本质差异。融合模型主要依赖于深度语义分析和引用网络建模，能够捕捉概念间的深层次关联；而组合图法则更注重时间序列特性，擅长发现突发性增长信号。这种互补性特征为未来研究指明了改进方向：通过整合两种方法的优势，有望构建出识别准确率更高、覆盖面更广的弱信号检测体系，从而为战略决策提供更具前瞻性的信息支持。

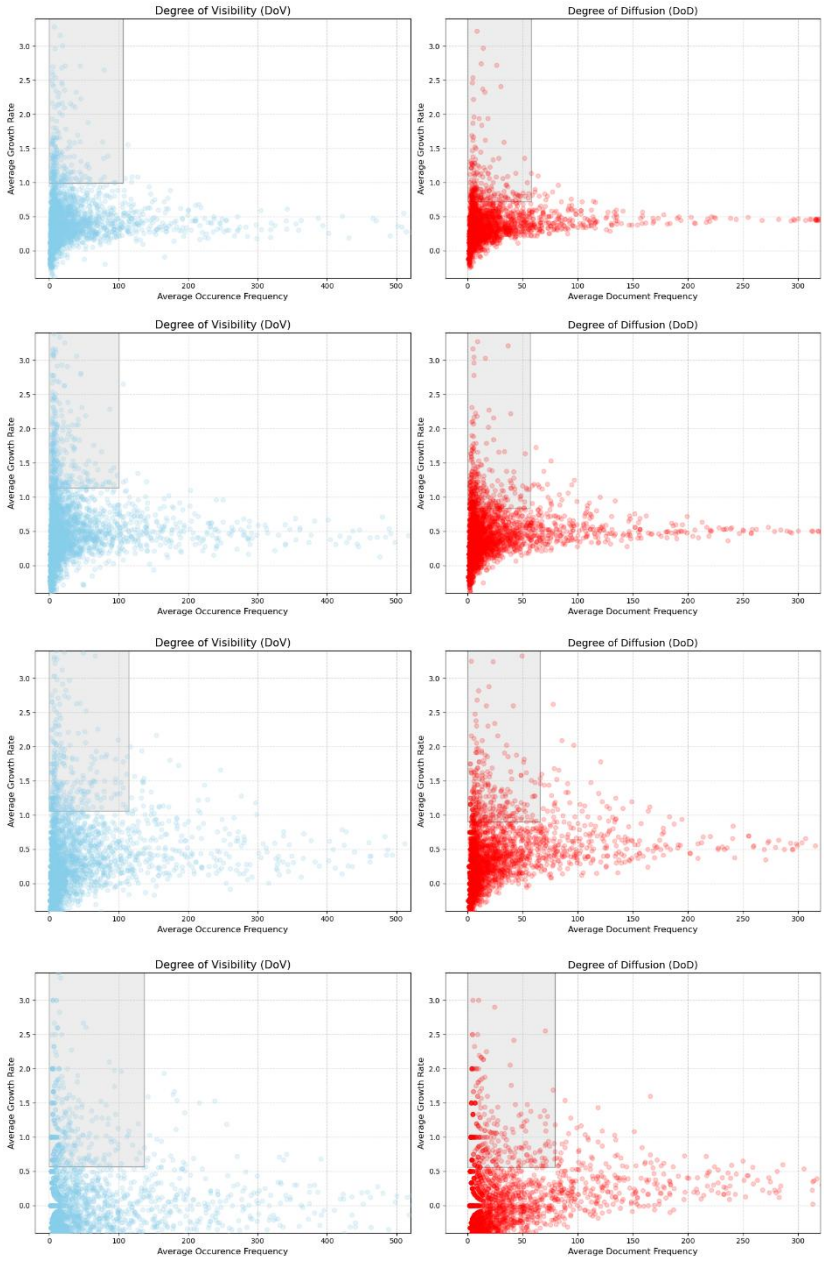

图 6：2020 年至 2023 年组合图法的弱信号识别情况

第六章 总结

6.1 研究发现与建议

本研究基于 BERT+DGCN 融合模型对 2018-2024 年金融科技领域的学术文献进行弱信号识别，成功捕捉到多个具有潜在影响力的新兴技术方向。研究发现，弱信号的演化具有明显的阶段性特征，不同年份的弱信号反映了金融科技发展的不同趋势和挑战。在 2018 至 2019 年，弱信号尚未显著显现。由于金融科技仍处于早期发展阶段，研究主题较为集中，数据量有限，因此未能识别出显著的弱信号主题。这一现象表明，在技术萌芽期，弱信号可能尚未形成足够的学术关注度，或仍以隐性方式存在于行业实践中。

在 2020 年，弱信号主要集中在区块链技术的早期探索阶段。尽管当时相关文献数量较少，但术语如“blockchain”和“transaction”已显示出潜在的技术突破方向。这些信号在 2020 年以后迎来爆发式增长，文献数量增幅高达 440%，与跨境支付和数字资产的实际应用落地形成直接关联。例如，摩根大通的 Onyx 平台和萨尔瓦多的法定数字货币实践，均印证了早期弱信号的前瞻性价值。2021 年，弱信号的主题逐渐向自然语言处理和智能投顾领域扩展。术语如“nlp”和“chatbot”在聚类分析中被识别为边缘方向，但 2023 年后随着生成式 AI（如 ChatGPT）的爆发，相关文献增长率达到 107%。金融机构如彭博社和苏黎世保险迅速将 NLP 技术应用于投资分析和客户服务，进一步验证了弱信号对技术商业化路径的预测能力。

2022 年的弱信号则呈现出更加多元化的特征，尤其是安全防护和技术架构优化方向。术语如“security”和“evolution”的文献增长分别达到 51%和 171%，反映了金融科技在快速发展中对风险管理和技术升级的迫切需求。例如，混合云与边缘计算的结合成为金融机构提升效率的重要路径，而“personal”和“mental”等术语的增长则体现了金融行业对个性化服务和员工心理健康问题的关注。2023 年至 2024 年，弱信号进一步向绿色金融和普惠金融领域延伸。术语如“green”和“underserved”的文献增长虽处于早期阶段，但已显示出明确的政策和技术联动趋势。例如，全球绿色债券发行量的快速增长与区块链技术的碳排放监测应用，预示了金融科技在可持续发展中的潜在作用。同时，伊斯兰金融与绿色科技的融合为非洲等地区的普惠金融提供了新思路，展现了弱信号在全球化背景下的广泛影响力。

基于以上发现，本研究为金融科技领域的相关主体提出以下建议：企业层面，应建立常态化的弱信号监测机制，尤其关注技术交叉领域的边缘创新，以便在技术爆发前抢占先机；投资机构可以结合弱信号识别与行业动态分析，优先布局具有高增长潜力的早期技术项目，从而降低投资盲目性；监管部门则需重视弱信号反映的合规与伦理问题，提前制定适应性政策框架，以平衡创新与风险。

6.2 研究局限

尽管本研究在弱信号识别方法上取得了一定突破，但仍存在一些局限性。首先，数据来源主要依赖于 OpenAlex 数据库的英文文献，虽然该数据库覆盖范围广泛，但可能无法完全代表非英语国家或地区的金融科技发展动态。例如，中国在移动支付和区块链应用方面的实践具有鲜明特色，但这些成果可能更多以中文形式发表，未被纳入本研究的数据集。这种语言和地域的局限性可能导致部分弱信号的遗漏或偏差。

其次，在模型设计上，虽然 BERT+DGCN 框架能够有效融合语义和结构信息，但对计算资源的要求较高。尤其是在处理大规模引用网络时，DGCN 的消息传递机制会显著增加训练时间和显存占用。尽管本研究通过批次处理和降维技术缓解了这一问题，但对于实时性要求较高的应用场景如市场舆情监测，现有方法的效率仍有提升空间。此外，弱信号术语的过滤依赖于 c-TF-IDF 和被引量指标，这些指标虽然能够反映术语的重要性，但可能无法完全捕捉术语在跨领域或跨学科中的潜在价值。例如，某些术语在金融科技领域出现频率较低，但在其他领域（如医疗或教育）已有广泛应用，其跨界融合的可能性未被充分评估。

最后，实证研究的时间跨度（2018 年至 2024 年）虽然能够覆盖金融科技快速发展的关键阶段，但对于长周期技术趋势的预测能力有限。弱信号的验证需要更长时间的追踪，而本研究仅能基于现有数据对部分术语的演化进行初步分析。未来需要通过更长期的纵向研究，进一步检验弱信号识别方法的稳定性和泛化能力。

6.3 研究展望

未来研究可以从以下几个方面进一步拓展和深化弱信号识别的方法与应用。首先，在数据层面，可以构建多语言、多来源的异构数据集，例如整合中文、西班牙语等非英语文献，以及专利、行业报告、社交媒体等非学术数据。这种多元化的数据来源有助于捕捉更全面的弱信号，尤其是在金融科技这种全球化特征明显的领域。同时，可以探索动态图神经网络（Dynamic GNN）的应用，以更好地建模主题和术语随时间演化的轨迹，从而提升弱信号识别的时效性和准确性。其次，在模型优化上，可以尝试将弱信号识别与生成式 AI 相结合。未来可以利用生成式模型对弱信号术语进行语义扩展或情景推演，模拟其潜在的发展路径和影响范围。此外，可以引入强化学习机制，通过反馈循环动态调整弱函数的参数，使模型能够自适应不同领域或不同阶段的数据特性。对于计算效率问题，可以探索轻量化模型（如知识蒸馏技术）或分布式训练框架，以降低资源消耗并提升实时处理能力。

在应用层面，未来研究可以进一步拓展弱信号识别的场景，将本方法应用于其他高风险、高动态性

领域如医疗健康、气候变化或地缘政治等，验证其普适性。同时，可以开发弱信号的可视化分析工具，帮助决策者更直观地理解术语之间的关联性和演化趋势。例如，通过交互式知识图谱展示弱信号主题的扩散路径，或利用时间轴工具对比不同术语的热度变化。这类工具可以成为战略规划和风险管理的有效辅助手段。而弱信号识别的伦理和社会影响也值得关注，避免算法偏见在弱信号筛选中的干扰，或者平衡商业机密与学术开放性的矛盾。未来研究可以探索弱信号识别的合规性框架，确保其在促进创新的同时，符合数据隐私和知识产权保护的要求。总之，弱信号识别作为连接学术研究与产业实践的重要桥梁，其方法论和应用场景仍有广阔的探索空间。通过多学科交叉和技术融合，有望为不确定环境下的战略决策提供更加有力的支持。

参考文献

中文参考文献

[1] ING 银行. 可持续金融动态：2025 第 5 期[EB/OL]. (2025)[2023-12-01]. https://www.zhitongcaijing.com/content/detail/1257427.html.

[2] 邓建军,刘安蓉,曹晓阳,等.颠覆性技术早期识别方法框架研究－－基于科学端的视角[J].中国科学院院刊,2022,37(05):674-684.

[3] 邓胜利,林艳青,王野.企业竞争弱信号的特征提取与定量识别研究[J].图书情报工作,2016,60(10):67-75.

[4] 方微, 邵波. 基于弱信号分析的企业风险识别[J]. 图书情报工作, 2009, 53(14): 80-83.

[5] 韩盟,陈悦,王玉奇,等.新兴技术弱信号识别：理论模型与测度方法[J].科学学研究,2024,42(11):2262-2274.

[6] 林晗.引用网络结构和文本语义融合的热点识别方法研究[D].军事科学院,2023.

[7] 刘俊婉,庞博,徐硕.基于弱信号的颠覆性技术早期识别研究[J].情报学报,2023,42(12):1395-1411.

[8] 刘宇飞,苗仲桢,黎凌峰,等.战略性新兴产业多领域知识融合路径研究——基于引用网络和文本信息的分析[J].中国工程科学,2020,22(02):120-129.

[9] 上海证券报. 上证圆桌 | 强化心理建设 关爱金融从业者身心健康[EB/OL]. (2023-07-05)[2023-12-01]. https://news.cnstock.com/news,yw-202307-5089018.htm.

[10] 孙涛,张秉坤,成磊峰,等.基于文本相似度和主题发现的弱信号识别方法[J].电脑知识与技术,2024,20(23):34-36.

[11] 王莉晓,陈伟,邱含琪.基于机器学习的颠覆性技术弱信号识别模型研究[J].数据分析与知识发现,2024,8(Z1):63-75.

[12] 新华社. 个人养老金全面实施：有何新变化？能否提前取？[EB/OL]. (2024-12-13)[2023-12-01]. https://www.gov.cn/zhengce/202412/content_6992491.htm.

[13] 佚名. 清华联合蚂蚁斩获电子学会科技进步一等奖 可信 AI 技术获国家级学会认可[EB/OL]. (2025-04-20)[2023-12-01]. https://finance.sina.com.cn/tech/roll/2025-04-21/doc-inetuwra3032092.shtml.

[14] 尹娟.基于深度学习的社交媒体网络舆情弱信号识别与预警研究[D].重庆理工大学,2024.

[15] 支付界. 反洗钱任重道远：行业平均误报率仍 95%[EB/OL]. (2020-03-24)[2023-12-01]. https://www.sohu.com/a/382744499_208700.

[16] 中国信息通信研究院. 全球数字经济白皮书(2022 年)[R]. 北京: 中国信息通信研究院, 2022.

[17] 中华人民共和国最高人民检察院. "三个善于"：揭示司法办案客观规律与本质要求[EB/OL]. (2024-05-23)[2023-12-01]. https://www.spp.gov.cn/spp/llyj/202405/t20240523_654818.shtml.

英文参考文献

[1] Adil B, Abdelhadi F. A framework for weak signal detection in competitive intelligence using semantic clustering algorithms[J]. International Journal of Advanced Computer Science and Applications, 2021, 12(12): 555-565.

[2] Ansoff H I. Managing strategic surprise by response to weak signals[J]. California Management Review, 1975, 18(2): 21-33.

[3] Basak S. Bridgewater launches $2 billion fund[EB/OL]. (2024-07-01)[2023-12-01]. https://finance.yahoo.com/news/bridgewater-launches-2-billion-fund-150000992.html.

[4] Ebadi A, Auger A, Gauthier Y. Detecting emerging technologies and their evolution using deep learning and weak signal analysis[J]. Journal of Informetrics, 2022, 16(4): 101344.

[5] Edo-Osagie O, Smith G, Lake I, et al. Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance[J]. PLoS ONE, 2019, 14(7): e0210689.

[6] El Akrouchi M, Benbrahim H, Kassou I. End-to-End LDA-based automatic weak signal detection in Web news[J]. Knowledge-Based Systems, 2021, 212: 106650.

[7] European Central Bank. Financial stability review, November 2023[R/OL]. (2023)[2023-12-01]. https://www.ecb.europa.eu/press/financial-stability-publications/fsr/html/ecb.fsr202311~bfe9d7c565.en.html.

[8] Filipsson F. Case study: AI-based predictive analytics for investments at BlackRock (Aladdin)[EB/OL]. (2025-02-12)[2023-12-01]. https://redresscompliance.com/case-study-ai-based-predictive-analytics-for-investments-at-blackrock-aladdin/.

[9] Global Algorithmic Trading Market. Algorithmic trading global market report[EB/OL]. (2023-09-11)[2023-12-01]. https://finance.yahoo.com/news/algorithmic-trading-global-market-report-092300576.html.

[10] GSMA. State of the industry report on mobile money[R/OL]. (2024)[2023-12-01]. GSMA.

[11] Gutsche T. Automatic weak signal detection and forecasting[D]. Enschede: University of Twente, 2018.

[12] Huang A H, Wang H, Yang Y. FinBERT: A large language model for extracting information from financial text[J]. Contemporary Accounting Research, 2023, 40(2): 806-841.

[13] IBM. Data suggests growth in enterprise adoption of AI is due to widespread deployment by early adopters[EB/OL]. (2024-01-10)[2023-12-01]. https://newsroom.ibm.com/2024-01-10-Data-Suggests-

Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters.

[14] IBM. IBM Watson Global AI Adoption Index 2021[R/OL]. (2021)[2023-12-01]. https://filecache.mediaroom.com/mr5mr_ibmnewsroom/191468/IBM%27s%20Global%20AI%20Adoption%20Index%202021_Executive-Summary.pdf.

[15] International Finance Corporation. Arab African International Bank, IFC, EBRD, and BII launch US$500 million sustainability bond to support climate finance and boost micro, small, and medium-sized enterprises[EB/OL]. (2024-11-24)[2023-12-01]. https://www.ifc.org/en/pressroom/2024/arab-african-international-bank-ifc-ebrd-and-bii-launch-us-500-million-sustainability-bond-to-support-climate-finance-and-boost-micro-small-and-medium-sized-enterprises.

[16] Jullum M, Løland A, Huseby R B, et al. Detecting money laundering transactions with machine learning[J]. Journal of Money Laundering Control, 2020, 23(1): 173-186.

[17] Khandani E A ,Kim J A ,Lo W A .Consumer credit-risk models via machine-learning algorithms[J].Journal of Banking and Finance,2010,34(11):2767-2787.

[18] Kim S, Kim Y E, Bae K J, et al. NEST: A quantitative model for detecting emerging trends using a global monitoring expert network and Bayesian network[J]. Futures, 2013, 52: 59-73.

[19] Kwak W ,Shi Y ,Kou G .Bankruptcy prediction for Korean firms after the 1997 financial crisis: using a multiple criteria linear programming data mining approach[J].Review of Quantitative Finance and Accounting,2012,38(4):441-453.

[20] Manela A ,Moreira A .News implied volatility and disaster concerns[J].Journal of Financial Economics,2017,123(1):137-162.

[21] Mastercard. Mastercard supercharges consumer protection with gen AI[EB/OL]. (2024-02-01)[2023-12-01]. https://www.mastercard.com/news/press/2024/february/mastercard-supercharges-consumer-protection-with-gen-ai/.

[22] Mühlroth C, Grottke M. Artificial intelligence in innovation: how to spot emerging trends and technologies[J]. IEEE Transactions on Engineering Management, 2020, 69(2): 493-510.

[23] Pushpendu G ,Ariel N ,Keshari J S .Forecasting directional movements of stock prices for intraday trading using LSTM and random forests[J].Finance Research Letters,2022,46(PA):

[24] Refinitiv. Navigating a new environment[R/OL]. (2022)[2023-12-01]. https://www.zawya.com/en/islamic-economy/islamic-finance/global-sukuk-issuance-seen-at-185bln-in-2022-oo642g3y.

[25] Roetzer P. Sentient Technologies applied its AI platform to marketing[EB/OL]. (2017-10-19)[2023-12-01]. https://www.marketingaiinstitute.com/blog/sentient-technologies-spotlight.

[26] SEC Crypto Task Force. Meeting memorandum[EB/OL]. (2025-04-22)[2023-12-01]. https://news.qq.com/rain/a/20250423A09DOW00.

[27] Superior Data Science. J.P Morgan - COiN - a case study of AI in finance[EB/OL]. (n.d.)[2023-12-01]. https://superiordatascience.com/jp-morgan-coin-a-case-study-of-ai-in-finance/.

[28] Thorleuchter D, Scheja T, Van den Poel D. Semantic weak signal tracing[J]. Expert Systems with Applications, 2014, 41(11): 5009-5016.

[29] Thorleuchter D, Van den Poel D. Weak signal identification with semantic Web mining[J]. Expert Systems with Applications, 2013, 40(12): 4978-4985.

[30] Yoon J. Detecting weak signals for long-term business opportunities using text mining of Web news[J]. Expert Systems with Applications, 2012, 39(16): 12543-12550.

[31] Yu M, Pasman H, Erraguntla M, et al. A framework to identify and respond to weak signals of disastrous process incidents based on FRAM and machine learning techniques[J]. Process Safety and Environmental Protection, 2022, 158: 98-114.

附录 A：金融科技领域术语增强词典

术语
ai, artificial, intelligence, machine, learning, ml, deep, learning, dl, neural, network, nn, transformer, attention, gpt, llm, bert, gpt-3, gpt-4, nlp, natural, language, processing, computer, vision, cv, generative, gan, reinforcement, learning, rl, supervised, unsupervised, semi-supervised, finance, financial, fintech, banking, investment, stock, market, trading, portfolio, risk, management, credit, loan, blockchain, crypto, cryptocurrency, bitcoin, ethereum, defi, algorithmic, forex, quantitative, hedge, fund, robo-advisor, fraud, detection, regulation, compliance, insurtech, payments, wealth, management, asset, lending, mortgage, insurance, audit, accounting, financial, inclusion, microfinance, crowdfunding, p2p, peer-to-peer, gcn, graph, convolutional, network, lstm, cnn, rnn, transformer, attention, svm, random, forest, xgboost, clustering, regression, classification, optimization, bayesian, forecasting, prediction, anomaly, detection

附录 B：2020 年至 2024 年的弱信号术语

年份	弱主题	弱信号
2020	主题 1	maneuvering, achieve, resource, opportunity, control, mixed, lack, aim, also, participant, provide, make, boundary, domain, domestically, become, verge, ever, internationally, due, hefty, firebase, logged, fetch, fulfils, implement, expenditure, help, skilled, correspond, anti, among, sound, needy, firstly, plaid, context, collapse, vigorous, pressure, saving, improve, trading, dialogflow, investor, pure, roboadvisors, demand, application, practical, individual, every, reliable, handle, cause, easy, unique, server, answer, big, detail, blockchain, important, early, collaborate, urban, mobile, environment, cloud, housing, preparation, switch, contrary, aforementioned, teaching, focus, alliance, tendency, accelerate, traditional, small, things, apply, pedagogical, productionand, dlt, sympathy, lower, insurance, tool, apis, accordingly, currently, transaction, rejection, plant, therefore
2021	主题 5	forward, logistic, logistics, internal, unified, point, education, several, improved, proprietary, comparative, nlp, although, toolsets, integrate, cost, experimental, play, efficiency, entire, allow, nonlinear, autoregressive, statement, forecast, fully, optimize, prolong, hadamard, rmqcs, astronomy, polynomial, quantum, type, vectorization, distributed, allocation, sampling, speedup, runtime, final, fault, cluster, dram, applications, complexity, controller, hpc, extension, criminal, recall, ensemble, theoretical, charging, embed, version, database, genetic, lifespan, tweet, violation, incident, achieve, technologies, contextual, scheduling, mobile, valuable, vital, finally, find, accord, review, processor, request, international, qos, optimal, discover, claim, adaptive, speed, zhou, communication, daily, signaltovoice, realize
	主题 7	increase, landscape, nation, recognition, automation, hence, ml, authority, advanced, compliance, integration, complete, bring, investor, possible, employ, due, consider, evasion, play, know, inefficient, sanction, administration, recently, perceive, subconscious, enrich, quantify, ceos, scrutiny, always, attitude, become, towards, psychological, compared, mediating, faceless, upper, tedious, penalty, manpower,

		taxpayer, offence, echelon, taking, shanghai, pronounce, profitability, heterogeneity, explore, trait, shenzhen, nonborrowed, alleviate, selling, analyze, literacy, centric, platform, highly, internally, banks, immerging, open, convenience, thrive, happen, apis, however, moreover, digitizing, consumer, motorized, survive, eliminate, identical, purchase, detect, chatbot, tough, conventional, evaluation, measure
	主题 11	neuroscience, choice, appropriate, interdependent, conventional, value, time, phylogenetics, class, intertwining, group, current, find, reliable, best, neural, expectation, significance, functioning, discipline, brain, genetics, background, results, allocation, cost, frequently, without, expert, choose, datasets, good, conclusions, estimate, median, ii, fairness, successful, institutional, accurate, government, clinical, however, tématerületi, filed, mid, nvkp_, source, nkfia, pension, minimization, bioimaging, fraction, ejection, therapy, cardiac, únkp, primary, decrease, apply, alone, adjust, bias, variable, less, kiválósági, itm, higher, ms, respective, male, therapeutic, predictor, university, auc, classification, variance, nkfih
2022	主题 1	conference, practical, disinformation, function, security, acm
	主题 8	theory, object, much, personal, center, cloud, face, malware, solution, life, large, sector, implement, tool, implementation, api, analysis, successful, context, trend, scope, plan, scale, vulnerability, adversarial, modern, instantiation, shop, mathematical, bot, novel, benchmark, web, write, evolution, hour, inspire, volume, hard, website, second, follow, mental, last, record, professional, others, integer, four
	主题 11	finally, safety, association, module, university, achieve, drive, construct, college, least, well, advance, internal, retailing, put, promising, advantage, attacker, explore, intelligent, exist, practical, mechanism, recent, order, measure, cause, task, crucial, performance, attnblstm, attnbilstm, defective, densely, dlcnn, round, recurrent, warn, many, actual, improvement, prediction, phenomenon, competitiveness, accuracy, lstm, simultaneously,

		long, blstm, layer, closedform, replica, pave, definitive, others, trend, corollary, test, conundrum, regulate, inability, adoption, crunch, period, relies, apply, subsidiary, clearcut, breakneck, promise, among, illiquidity, forward, centric, halflife, clever, objective, prescriptive, novel, tuning, catalog, enhancing, market
	主题 13	example, practice, procedure, autonomy, already, frame, prioritize, motivate, practical, bias, clinical, create, sdgs, uncertainty, advanced, fairness, sector, deployment, tension, alongside, anatomical, emerge, gap, neural, major, contract, detailed, aspect, operation, robotics, augmentation, technological, policing, politics, extensive, crucial, case, water, good, rapidly, limited, ultimately, group
2023	主题 6	associate, setting, huge, teacher, predetermine, impairment, extensive, complication, demonstrate, lumineticscore, visit, disparity, clearance, month, bfs, environment, pedagogical, mems, define, decline, aivi, participants, pedagogically, convergence, dialogue, carer, legal
2024	主题 4	effort, improve, behavioral, inclusivity, increase, emphasize, behavior, promising, collective, author, hybrid, ecological, advanced, leverage, examine, indicator, condition, platform, feasibility, interface, reality, programming, identification, successful, namely, reduce, practice, sharia, narrow, expand, inclusive, emerge, green, organization, sme, among, academia, break, establish, way, sdgs, importance, theme, underserved, cities, urban, shed, think, collaboration, material, portfolio, government, focus, african, promise, primary, price, hub, nations, ensure, delineate, hinder, long

附录 C：组合图法弱信号术语

年份	弱信号
2020	sympathy, next, facial, alliance, vigorous, revolutionize, server, death, dream, aim, laboratory, th, require, corporation, expenditure, reliable, move, duration, insignificant, image, frequency, flexibility, belt, predictive, antagonism, effective, household, asset, saving, control, oversight, boundary, encryption, literacy, small, domain, emerge, housing, ever, resource, lack, dlt, help, favorable, stocks, damage, macro, south, verge, tendency, accelerate, statement, strive, actionable, optical, collapse, make, preparation, fetch, primarily, taxonomy, exploration, elucidate, adoption, unfold, api, china, firstly, latter, risk, traditional, legislation, among, collaborate, complex, surrogate, less, lost, technological, apparent, focus, cryptocurrency, likely, xai, sound, streamline, algorithmic, year, essay, one, approach, technologist, span, productivity, scoring, alibaba, cooperation, essentially, dominant, area, million, intelligent, mobile, institution, connection, teaching, productive, relation, oligopoly, chat, underscore, due, contradiction, trading, opinion, aside
2021	median, bayes, et, finding, gru, toolsets, superior, medicine, discipline, larose, explainable, criminal, moreover, sec, finally, possibility, tough, fit, website, player, explanations, digitize, carry, chinese, incorporation, positively, induce, silicon, impose, empirical, smes, different, nature, place, launder, variance, min, sons, pronounce, psychology, personalize, transformer, sovereign, take, advanced, cluster, ecological, lime, corresponding, degradation, search, upper, nano, refinement, pillar, reap, ml, describe, manpower, explore, temperature, ifc, logistic, base, sector, collect, give, hierarchy, relationship, machine, fraction, correction, congo, application, section, expert, topic, automation, delineate, language, shap, specifically, lastly, description, precision, offence, functioning, intertwining, among, senior, discover, transitioning, minimization, inherent, advantage, graph, technology, prescriptive, administration, calculate, cost, compliance, invert, reference, rank, earnings, taxonomy, recently, comparative, lender, phishing, neural, trade, physician, explainability, compared, ibm, predicting, residential, persistent, urban, future, advancement, elucidate, seek, biobank, apis, conclusions, acquire, regulatory, flow, secondary, shape, abstract, survive, replicate, modeling, activity, human, stage, datos, nvkp_, prior, al, open, encompass, importantly, hu, popularization, share, mediating, credibility, propose, distinguish, screen, university, polish, predicción, encryption, hazard,

reason, repay, sport, transparency, recommend, communicate, sexual, part, dt, fail, alleviate, financial, reduce, weight, credit, shanghai, selling, corporate, showcase, significance, state, insurance, embrace, store, request, point, charging, mention, database, decline, class, referential, enrich, extract, foster, review, genetics, transportation, include

2022 span, part, prescription, open, important, include, desire, hospital, awareness, control, nlp, argue, participate, complex, redundant, abstract, mainly, embrace, like, opportunity, crucial, low, emerald, subsidiary, closedform, kari, offline, transport, inventory, medical, first, scarcity, aim, delve, article, level, promising, future, pave, opinion, banking, xai, adapt, exist, alpha, reflective, understand, give, conundrum, readiness, precede, attnbilstm, thoracic, david, hour, zero, kharkiv, bias, protection, crunch, centric, colonoscopy, transition, focal, malicious, definitive, put, diagnostics, state, face, process, produce, vital, trustee, connect, discrimination, app, last, shock, personnel, advice, transparent, optimization, fourth, outside, transparency, profit, al, international, adversarial, role, flow, tangible, malware, elucidate, content, contribution, societal, cycle, advantage, web, ultimately, inspire, paas, exemplify, prescriptive, context, replacement, theorize, teach, dlcnn, provide, critical, write, match, disposition, wgo, radically, recently, nonetheless, alaimo, breakthrough, three, solution, numerous, life, information, beacon, sdgs, prevail, consumption, beyond, equity, digitalize, highly, translate, today, thinking, robustness, interests, amount, show, quantum, stanford, oviedo, theorist, confidence, fulfil, narrow, criterion, writer, responsible, expose, vocabulary, diffusion, ensembling, lesson, take, consider, innovation, loop, pursuit, proficiency, investment, recommend, end, advanced, natural, productdominant, author, available, allow, find, frame, marketing, planning, healthy, test, choose, good, model, interface, fluctuation, rasdaman, deficiency, mitigation, respect, grant, comprise, session, trajectory, typically, worth, trend

2023 author, huge, curation, different, literature, multifaceted, role, simultaneously, volatility, operation, emphasize, engine, complication, focus, aid, extensively, usefulness, convergence, concurrently, day, term, dialog, ct, categorize, modality, collection, process, aforementioned, embark, responsible, robust, survey, hypothesis, seek, addition, respond, reliance, implication, significantly, finance, execution, disparity, aspect, full, option, methodology, model, inform, shapley, superiority, uncover,

safeguard, augment, radiology, intersection, language, fair, deployment, priorities, offer, blood, shed, myriad, democratization, illustrate, impairment, literacy, auditing, provider, promise, retail, invasion, pedagogically, closing, important, rule, focal, finding, compatibility, focusing, balanced, profoundly, esg, personalize, intricacy, autonomously, intensify, advantage, automation, subsequently, hesitant, penetrate, necessitate, realm, question, journey, population, cybersecurity, furnish, lead, interpretable, accessibility, clearance, guarantee, exploration, shap, financing, unstructured, daily, streamline, refine, technology, identify, analyze, inference, emerge, across, encryption, xgboost, teacher, company, mohammad, geoeconomic, transcend, product, reskilling, incorporation, reduce, navigate, competitiveness, sensitive, disease, venture, illuminate, intellectual, range, extensive, historical, tiny, maximize, understand, facet, interact, life, certain, geopolitical, outcome, integrate, detailed, may, multiple, augmentation, reminder, text, impact, factor, spot, planning, cohesion, challenge, scf, play, separately, industry, significant, topic, actionable, interdependent, decision, netherlands, investigate, attribute, quality, friendly, dalle, tip, professional, stability, harmonious, clustering, profound, trustworthy, base, aivi, societal, destination, ukraine, similarity, examples, assist, environment, shape, protection, productivity, tailor, medication, additionally, cohesive, detect, advocate, investing, consideration, facilitate, potentially, reputable, trading, practice, efficacy, capability, represent, clinicians, foster, designer, artificially, consultation, response, require, paper, whilst, transformative, conversation, bfs, type, general, associate, importance, displacement, utilize, privacy, meticulous, emotional, domain, field, expand, must, balance, internal, pivotal, descriptive, meticulously, diverse, condition, filter, utilization, indirectly, adaptive, predetermine, risk, hurdle, search

致谢

行文至此，意味着我的本科生涯即将画上句点。回首这段求学时光，心中充满感激与不舍，正是众多师长、同窗和亲友的支持与陪伴，让我不仅仅是顺利完成学业，更是人格涵养的洗礼与锤炼。

感谢我的父母，是你们含辛茹苦把我养大，教我人生道理，如果没有你们就没有现在的我。我不善言辞，但永远把你们放在第一位。

感谢我的三位姐姐，从小就对我疼爱有加，不管是金钱还是精神，你们都很愿意支持我。祝愿你们都能过上自己理想中的生活。

感谢我的伴侣张丝敏，这一年多的时间里我们留下了许多美好的回忆，也都珍惜彼此的存在。如今即将共同迈入人生的另一阶段，还请多多关照。

感谢周庆山教授，在您的指导下我才能够完成这篇毕业论文。感谢北京大学，为我提供如此优质的教育资源。

感谢我的两位室友蔡昇杰、黄子铭，以及番外室友邝训贤、余勇乐，这一年来你们带来了很多欢笑，让我的大学生活充满活力，也磨练了我的厨艺。也感谢我的前室友李均皓、张垐安、詹宇谦，和你们同住屋檐下让我感到很安心、舒适。感谢 hdl 小分队、清流、所有本硕博已毕业未毕业延毕休学退学的朋友，没有你们我的大学生活一定很枯燥。

感谢信管、国信排球队，是你们让我对排球的热爱更上一层楼。特别感谢陈雨航老师，你对球队的用心与付出我们有目共睹。感谢国际球友、五四球友、比赛队友，使我有机会磨练自己的技术。感谢所有陪我打过羽毛球、网球、台球、皮克球、乒乓、爬山、篮球、足球的朋友，让我经历了那么多有趣的体验。

感谢所有亲戚、老师，感谢实习单位的同事，感谢拉茶的小伙伴们，感谢世界上所有的猫和花，感谢中关新园，感谢 hdl 的 69 折，感谢麦当劳，感谢史迪仔，感谢车管所让我考到驾照，感谢我的小电驴，感谢所有考试，感谢所有熬过的夜，感谢网课，感谢博雅计划，感谢海淀，感谢北京，感谢中国。

感谢我的本科生涯，让我不断寻找人生的价值与意义，虽然有过怀疑、失望，但也有憧憬、理想。

愿你我实现未来期待，愿你我一切安好顺遂。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：

日期：　　年　　月　　日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；

论文作者签名：

导师签名：

日期：　　年　　月　　日