arXiv:2601.19513v1 [cs.IR] 27 Jan 2026

利用细粒度知识实体和多维文档嵌入提升学术论文推荐

Haixu Xi[1,2]，Heng Zhang[3]，Chengzhi Zhang[1*]

[1]南京理工大学信息管理系，南京，210094，中国。[2]江苏理工学院计算机工程学院，常州，213001，中国。[3]华中师范大学信息管理学院，武汉，430079，中国。

*通讯作者。电子邮箱：zhangcz@njust.edu.cn；贡献作者：xihaixu@jsut.edn.cn；zh heng@ccnu.edn.cn；

摘要

在学术文献爆炸式增长的时代，学者承担的文献综述负担日益加重。在研究过程中，主动推荐符合学者文献需求的学术论文，已成为提高研究效率、激发创新思维的关键途径之一。当前的学术论文推荐系统主要基于一般主题或领域相似性，侧重于提供宽泛的粗排建议。尽管这些系统能够有效识别相关文献，但在满足学者更具体、更细粒度的需求方面仍存在不足，例如查找采用特定研究方法的论文，或查找在同一主题下处理特定研究任务的论文。为满足学者在研究过程中多样化且具体的文献需求，本文提出了一种新颖的学术论文推荐方法。该方法融合新型细粒度知识实体、文档标题与摘要以及引文数据，对多维信息进行嵌入。随后，通过计算组合论文向量之间的相似性生成推荐结果。本文使用 STM-KG 数据集对所提出的推荐方法进行了评估。该数据集是一个知识图谱，包含从涵盖十个不同领域的论文中提取的科学概念。实验

结果表明，我们的方法优于基线模型，在前 50 个推荐结果中取得了 27.3% 的平均精确率。这一结果较现有方法提高了 6.7%。

关键词： 论文推荐、知识实体、知识图谱、语义嵌入

1 引言

全球学术成果的迅速增长极大地增加了已发表研究的数量，加重了文献综述的负担。高效且准确地找到符合研究人员兴趣和专业领域的论文，已成为一项核心挑战（Gusenbauer & Haddaway，2020年）。应对这一挑战有助于发现新的研究主题，并推动创新方法的发展。通过主动推送相关论文，学术推荐系统能够减少检索所需的精力；这一能力已受到学术界和产业界的广泛关注（Kreutz & Schenkel，2022年）。

现有方法通常侧重于全局内容分析（Zhu, Patra,& Yaseen，2021年）或基于历史交互的协同过滤（Murali,Vishnu, & Victor，2019年）。一些方法还会纳入作者间关系（Du 等人，2020年）或引文网络（Son& Kim，2018年），以评估论文或研究人员之间的相似度。然而，这些技术通常以较粗的粒度运行，且未能充分利用文本中存在的细粒度知识实体（Cao 等人，2021年）。因此，它们往往无法解释为何推荐某篇论文，或该论文与研究人员的特定兴趣有多么契合——例如，同一主题领域内的特定任务、问题或方法选择。

近期研究已开始关注细粒度信号（Takahashi & Katsurai，2022年）、特定知识实体（Ni, SumLiu, & Torzec，2020年）、知识图谱（Brack,Hoppe,& Ewerth，2021年；Cao 等人，2021年；Im`ene, Badia, & Nadir，2022年）。主流搜索引擎同样会将查询与知识库中的实体关联起来，以改进排序。尽管取得了这些进展，现有解决方案仍无法满足多样化的文献综述需求。研究人员经常希望找到针对相似任务、但采用不同方法或理论视角的论文，以全面了解某一领域。现有技术通常依赖知识图谱中静态的语义关系，且不会根据给定任务动态建模异构实体类型之间的语义组合，因此难以揭示多样化的研究视角和方法体系。

总之，当前推荐方法未能充分满足精排相关性与多样性的联合需求。为弥合这一差距，本文提出了一种推荐框架，用于建模异构细粒度知识实体之间动态且多样的语义组合。该方法将多维信号——标题、摘要、引文关系，以及细粒度实体及其关系——整合到文档嵌入中，并基于拼接的向量表示计算相似度，从而提供细粒度的基于内容的推荐结果

同时提升多样性。具体而言，本文构建了细粒度科学知识图谱（FG-SKG），用于捕捉详细的实体——研究任务、方法、材料和指标——及其关系。该图谱支持创建多维嵌入，将核心文本内容、引文上下文和自动提取的知识实体进行联合编码，从而使推荐结果与研究人员的细粒度信息需求保持一致。通过在相似度计算中采用受控组合和加权方案，进一步促进了多样性。

本研究的主要贡献有两点。(i) 提出了一种推荐方法，通过动态组合细粒度实体，使推荐结果契合研究人员具体且多样的文献需求，从而同时提高覆盖率和准确率。(ii) 在 $STM-KG$ 数据集（Brack、M¨uller、Hoppe 和 Ewerth，2021年）上系统评估了嵌入模型和拼接策略。所提出的方法实现了 27.3% 的 Top-50 精确率，较基线模型提高了 6.7 个百分点。所有数据集和源代码均可在以下网址公开获取：https://github.com/jsutxhx/SKG-AR。

2 文献综述

本节回顾以下领域的最新进展：(i) 基于知识单元的学术论文推荐，(ii) 面向学术文献的基于图的推荐，以及 (iii) 科学文档中的多维信息嵌入。

基于知识单元的学术论文推荐

现有系统主要通过识别相似学者或相似文档来推荐论文，采用内容的语义表示（Ma、Zhang、Zhang 和 Liu，2021年），或基于过往交互进行协同过滤（Sakib、Ahmad 和 Haruna，2020年）。为缓解数据稀疏性，许多方法会纳入图结构，例如社会关系和引文网络（Du 等人，2020年；Hao、Liu 和 Pan，2021年；Kanwal 和 Amjad，2024年）。然而，这类图衍生信号的客观性和中立性仍存在争议，并可能对推荐多样性和公平性产生不利影响（Ekstrand 等人，2018年）。与此同时，针对论文中知识层面内容的分析与挖掘受到的关注相对较少（Wu 等人，2020年），这限制了对细粒度、内容特定推荐的支持。

随着大规模知识库（例如，DBpedia[1]，YAGO[2] 和 Free- base[3]）的出现，主流搜索引擎开始将查询关联到实体，并对以实体为中心的结果进行排序。正如表 1所总结的那样，学术推荐研究同样探索了多种粒度的知识单元嵌入：话语级片段（例如引言、相关工作；（Santosa 等人，2024年））、句子级单元

[1]DBpedia（https://dbpedia.org/）旨在从维基百科文本中提取知识，并提供结构化内容。[2]YAGO（https://yago-knowledge.org/）是一个开源知识库，其内容由维基百科及其他来源自动提取而来。[3]Freebase API 现已由 Google 知识图谱 API 取代，且 Freebase.com 于 2016 年 5 月正式关闭。

表 1 基于知识单元的学术论文推荐代表性研究综述。

类别	作者	方法	主要发现
话语-层级	Santosa , Liu, Han, and Miyazaki (2024年)	S3par	S3PaR，一种序列式科学论文推荐——基于论文篇章结构的推荐系统，旨在支持学术写作，并提供上下文相关参考文献。
句子级别	Ma, Zhang, Zhang, and Gao (2025年)	多任务学习	基于多层感知机开发了一种模型，可同时进行形成引文推荐和论证引文句子的分类。
	Takahashi and Katsurai (2022年)	向量相似度	它通过计算来推荐学术论文.....之间向量相似度的差异研究目标句与方法论句子子。
短语-级别	Sarwar等人 al. (2021年)	多级时间逻辑基于学习的	使用关键词提取和余弦相似度计算，从参考文献中推荐相关文章，帮助理解某篇文章的内容。
	Mohamed, El-Kilany, and Mokhtar (2021年)	基于概念-的	通过提取语义概念来建模研究者兴趣这些概念来自他们的文章（由他们撰写、下载，或阅读的。）
	Luan (2018年)	科学术语基于	利用.....构建科学术语知识图谱半监督自然语言处理技术，并推导潜在关联，以推荐与研究人员相关的技术和理念。

（例如研究目标/方法句；（Takahashi & Katsurai，2022年））、修辞类型（目标/方法/结论；（Ma 等人，2025年））、短语级元素（关键词；（Sarwar 等人，2021年））、科学概念（Mohamed 等人，2021年）以及任务/方法实体（Luan，2018年）。相关研究也致力于推荐特定的精细知识，例如主题（Q. Liu、Du、Hong 和 Fan，2021年）、方法（Huang、Wan 和 Tang，2013）和数据集（Viswanathan、Gao、Wu、Liu 和 Neubig，2023年）。

这些研究通过利用以知识为中心的内容特征推荐论文，提升了精细粒度的相关性。然而，将知识单元孤立处理可能会忽视其语义结构和单元间关系，从而限制对文档内容的深入理解，并降低推荐效果。未来需要进一步对知识单元之间的结构和语义关系进行建模，以提供更加详细、准确且多样化的推荐结果。

2.2 基于图的学术论文推荐

基于图的推荐方法通过显式建模论文及相关实体之间的结构化关系，提升学术资源的匹配效果。如表2所总结，先前研究利用了作者间合作网络

(Hwang、Wei 和 Liao，2010年）、论文引文网络（Sakib 等人，2020年），以及二者的组合（Kanwal 和 Amjad，2024年）。然而，仅依赖结构可能引发有关客观性和公平性的担忧（Adomavicius 和 Tuzhilin，2015年）。为缓解这些问题，后续研究将学术档案（例如研究兴趣和领域）与文本语义相结合，构建表达能力更强的异构知识图谱（Du 等人，2020年；Gupta 和 Varma，2017年；Hao 等人，2021年）。

为满足细粒度内容感知推荐的需求，该领域已从整体内容建模转向细粒度知识建模。近期方法会提取语义单元——知识实体、科学概念或承载知识的句子——并构建融合语义信息与结构关系的科学知识单元图（Brack, Hoppe, & Ewerth，2021年；Cao 等人，2021年；Im`ene 等人，2022年；Ni 等人，2020年；Xi、Zhang 和 Zhang，2024年）。通过嵌入这些图并进行结构推理，推荐模型能够更全面地捕捉学者偏好、论文语义及潜在关联，从而提升结果的准确率和可解释性。

图表示（实体和关系嵌入）结合向量空间运算、链接预测和元路径推理，已被证明能够提高准确率，并在一定程度上提升多样性。尽管如此，当前的多维嵌入和推理技术仍难以满足多样化的文献需求，尤其是在需要动态组合知识时（例如，评估给定研究对象的某项任务当前所处的状态，或列举针对特定任务的替代方法）。因此，学术论文多维信息嵌入仍有待进一步研究。

2.3 学术论文的多维信息嵌入方法

学术论文编码了异构信号，包括作者身份与合作关系、论文之间的引文链接，以及细粒度知识实体（例如研究任务、方法、数据集和评估指标）及其语义关系。单一嵌入通常不足以捕捉这些多维语义。因此，多维信息嵌入被用于分别建模不同的语义单元，并将其整合为结构化且信息丰富的文档表示。

SPECTER 等具有代表性的方法（Cohan、Feldman、Beltagy、Downey 和 Weld，2020年）以及 SPECTER2（A. Singh、D'Arcy、Cohan、Downey 和 Feldman，2023年）利用引文监督和 Transformer编码器生成引文感知论文嵌入。SciNCL（Ostendorff、Rethmeier、Augenstein、Gipp 和 Rehm，2022年）进一步引入多视图对比学习，从元数据和内容相似度中构造负样本，以增强判别能力。其他方法则在结构层面或语义角色层面开展工作：HyperDoc2Vec（Han、Song、Zhao、Shi 和 Zhang，2018年）将文档划分为不同章节，以生成层次嵌入；而 COSAEMB（S. Singh 和 Singh，2024年）则从科学论述结构（例如方法、结果和讨论）中学习角色感知向量，从而提升可解释性和模块化程度。

表2 概要　　图神经网络学术论文推荐代表性研究　　。

类别	作者	方法	主要发现
关系网络为基础	Hwang等人 (2010年)	co-作者身份基于	构建学术界的合著网络，然后进行路径探索并针对特定任务进行筛选，以识别推荐论文。
	Sakib et al. (2020年)	基于引文的使用	公开上下文信息改进通过两级引文关系进行相似性检测等。
	Kanwal and Amjad (2024年)	RRMF	RRMF 通过整合论文推荐，引文网络和作者合作关系，利用结构模式并识别关键作者。
知识图谱-基于	Hao et al. (2021年)	兴趣-基于	一种方法利用作者不断变化的兴趣以及学术图结构来提升论文推荐的准确率和相关性。
	Du 等人(2020年)	HNPR	HNPR 使用带有引文信息的异构网络链接，作者合作　　，和研究领域　　，采用随机游走评估论文相关性。
	Gupta and Varma (2017年)	CCA	他们的方法融合了语义文本嵌入和引文图网络嵌入，并通过典型相关分析将其映射到同一空间（CCA）。
精排获得知识图谱基于	Xi 等（2024年）	FG-SKG & 多维嵌入	一种基于细粒度知识实体的多维文档嵌基于细粒度知识实体的表征及其关系。
	Brack, Hoppe, and Ewerth (2021年)	多重嵌入 & Con-概念	整合从.....中学习到的文档嵌入论文的文本内容和引文图，以及论文中提及的科学概念。
	Cao et al. (2021年)	GNN & Deep network	一种连接数据集和方法实体，并整合图结构以及通过图实现的文本特征表示神经网络和深度协同过滤网络，以实现适当的方法推-荐，针对特定数据集提供建议。

在本研究中，这些方向得到结合和扩展，形成了知识实体感知嵌入，能够显式编码任务、方法、材料和指标。文档级向量与引文级向量相结合，构成统一的混合表示，融合结构与语义，从而提升推荐的语义覆盖度和表示保真度。在以既有研究中的不足为出发点的基础上，本文探讨两个问题：（i）能否在知识实体层面，根据学者的精细化意图定制推荐结果？（ii）能否通过利用知识图谱和文档的多维嵌入，进一步提升论文推荐系统的性能？

3 方法论

所提出的推荐框架包含三个阶段。首先，构建细粒度科学知识图谱，并对其节点进行嵌入，以表示任务、方法、材料和指标。其次，通过拼接沿节点—关系路径的向量来构成文档表示，从而生成多维嵌入，在多个粒度层次上捕捉内容及其引文关系。第三，基于观察到的文献检索行为，对论文向量进行重新加权和拼接，以筛选候选项并对结果进行排序。该流程能够实现候选筛选和重排序，以满足学者多样化的信息需求，同时保留细粒度的语义特异性。

3.1 研究框架

图 1 展示了所提出方法的整体架构，该方法由三个组成部分构成：(i) 细粒度科学知识图谱（FG-SKG）的构建，(ii) 多维文档嵌入，以及 (iii) 基于向量的推荐。通过联合利用 FG-SKG 和多维嵌入，该系统旨在平衡精确率与多样性，从而生成既量身定制又多方面的推荐结果。

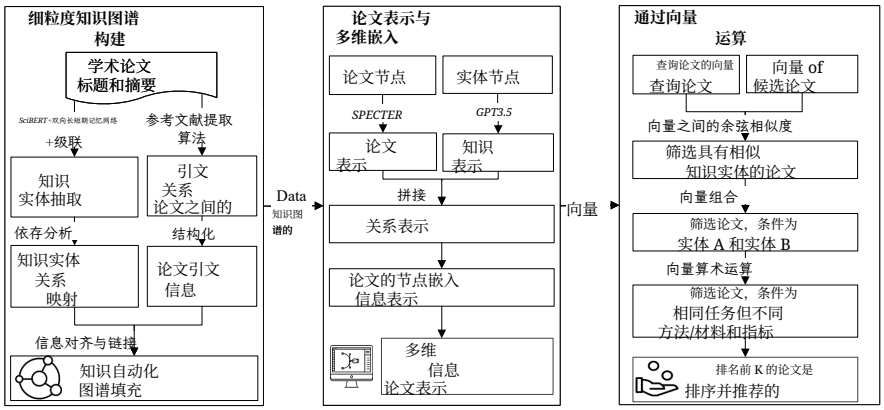

图 1 论文推荐研究框架。

FG-SKG 强调细致的关系结构，以提升相关性。从标题和摘要中提取四类知识实体——任务、方法、材料/数据集和指标，并映射它们之间的关系，涵盖语义链接和基于引文的关联，从而丰富上下文信号。生成的嵌入整合文本内容、引文信息和实体语义，为论文生成全面的向量表示。在推荐过程中，首先通过余弦相似度筛选候选项，然后利用学习得到的行为信息引导的权重进行排序，从而确保相关性，同时满足多样化的学术信息需求。

3.2 细粒度知识图谱构建

本研究通过整合研究论文、其核心知识实体、实体间关系以及论文间引文网络，构建细粒度科学知识图谱（$FG-SKGs$）。流程如下：首先，从论文标题和摘要中识别四类知识实体。接下来，提取并标注这些实体之间的论文内关系。然后，通过分析参考文献推导论文间引文链接。最后，将论文、实体及其关系融合为一个结构化、相互关联的图谱，从而连贯、全面地呈现研究领域。

3.2.1 学术论文中细粒度知识实体的自动识别

我们首先定义与科技论文相关的细粒度知识实体类型，随后开发一种自动化实体识别模型，其性能可与最先进水平（SOTA）的方法相媲美。

（1）细粒度知识实体分类

以往研究引入了多种实体类型来组织科学信息（Brack, Hoppe, & Ewerth, 2021；Hou, Jochim, Gleize, Bonin, & Ganguly, 2021；Luan, He, Ostendorf, & Hajishirzi, 2018；Zhang, Zhang, & Wang, 2024），包括任务、方法、工具、过程、材料、数据/数据集、指标、得分、其他科学术语和通用类别。随着标签集不断扩展，识别准确率通常会下降，从而在推荐场景中形成覆盖率与可靠性之间的权衡（Brack, M¨uller, et al., 2021；Dess´ı, Osborne, Reforgiato Recupero, Buscaldi, & Motta, 2022；Ni et al., 2020）。在上述类别中，过程和工具用于描述特定于方法的细节，而得分表示指标下的具体结果；在文献探索的早期阶段，这些要素通常并非主要关注点。材料、数据和数据集均指支持研究的资源，并且在不同学科之间可能存在较大差异；为确保跨领域一致性，本文将它们统一归入材料类别。最后，其他科学术语和通用类别缺乏足够的特异性，无法捕捉细致的用户偏好。

为兼顾特异性与稳健性，提炼出四种主实体类型：任务、方法、材料和指标（表3）。这种分类保留了与学术搜索意图最相关的概念，同时控制了标签复杂度，从而提升推荐结果的相关性和可解释性。

（2）细粒度知识实体识别模型

准确识别细粒度实体对于表示学者的详细信息需求至关重要。为此，采用 $SciBERT$ + 双向长短期记忆网络 + 级联架构（Zhang 等人，2024年）。该模型（图 2）沿用 Zhang 等人（2024年）的数据增强方法，以增加样本数量和多样性；通过 $SciBERT$提升词元级语义，借助 双向长短期记忆网络捕获上下文依赖关系，并通过级联学习完成实体边界检测和类型分配。采用 条件随机场 层来约束有效的标签转移并提升序列一致性。

T表 3 学术论文中的知识实体分类 。

Our 实体	实体存在的	实体描述	示例
Task	Task	描述需要观察的具体事物或现象、分析和研究的对象，概述需要解决的问题以及需要完成的任务，即 *what is being researched*。	缓解学者的信息过载是移动阅读时代的关键问题。句法步骤与多特征文本相结合，用于自动化结构式科学论文的摘要生成或许是不错的选择。
方法	方法，工具，过程	探讨理论模-els，具体技术，and 程序所需 to con-开展研究任务，具有专有名称，即有专有名称，即 *how the research is conducted*。	This 论文 提出了 a Moves-Features摘要方法该方法融合了多种文本特征以权重的形式纳入TextRank算法的迭代计算过程，并引入MM 处理，以处理预先选定的摘要集合。
材料	材料，Data and 数据集	支持研究所需的资源研究活动，包括物理材料和数字语料库和数据集，即 *what is used to conduct research*。	实验研究 on the SUMPUBMED 数据集表明 com-与基线模型相比，该方法效率很高。
指标	指标	用于.....的参数或指标比较、测量和评估的参数或指标研究成果的有效性，即 *how the results of the research are evaluated*。	这种方法在提升不同句 摘要的 ROUGE 的方面存在差异。

注： 任务: 方法: 材料: 指标:

具体而言，该流程由三个组件构成：（i）*SciBERT* 编码；（ii）双向长短期记忆网络 上下文化；以及（iii）级联解码（实体标注器 +实体类型标注器）。输入首先被编码为词元向量，得到隐藏序列 H。在级联解码器中，两个二元分类器分别识别实体的起始位置和结束位置；随后，条件随机场 约束预测的标签序列，以确保有效的*BIO*/IOB模式，并生成实体跨度。接下来，实体类型分类器为每个跨度分配四个目标类别之一（任务 /方法/材料 /指标）。与此同时，词元级表示 H 经由双向长短期记忆网络 进一步进行上下文化处理，生成H'，并据此得到跨度表示，再将其输入类型分类器。该设计在保持较高边界精确率和稳健类型判别能力的同时，提高了多标签分类效率。

3.2.2 学术论文中细粒度知识实体的自动关系抽取

我们首先定义与科技论文相关的细粒度知识实体之间的关系类型。基于这些定义，我们采用类型到关系的映射策略，自动推断实体对之间的关系类型。这一抽取流程的有效性取决于底层实体类型识别模型的准确率。

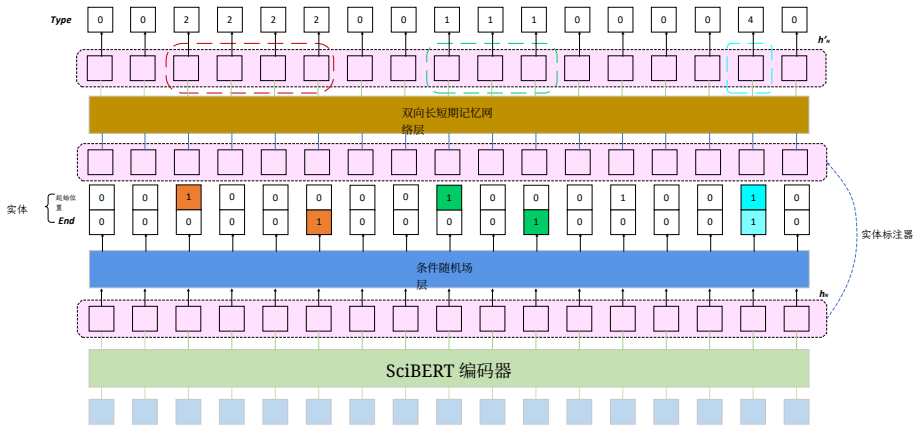

图 2 深度学　　习模型用于细粒度知识实体识别　　。

（1）细粒度知识实体关系分类

近期研究确定了知识实体之间广泛的关系类型，以支持知识图谱构建（Brack, Hoppe, & Ewerth, 2021年；Dess´ı 等，2022年；Luan 等，2018年；Mondal, Hou, & Jochim, 2021年）。其中包括比较、合取、评估对象、在……上评估、由……评估、部分关系、特征关系、用途关系、下位词关系、共指、相关等，共计 179 种不同的语义关系。然而，在文献综述过程中，学者通常关注与特定实体相关的论文，而非探究所有可能的关系细微差别。因此，出于检索目的，可以将比较、合取、部分关系、下位词关系和特征关系合并为更宽泛的“相关”关系。与此同时，用途关系经过细化，以反映特定类型的交互：由……实现（任务 ↔方法）、被……使用（材料 ↔ 任务）以及由……评估（任务 ↔ 指标）。这些区分体现了四种实体类型的语义角色，并提升了表征清晰度。

为降低抽取复杂度而不牺牲覆盖率，本文不对显式共指链接建模；相反，通过表面形式匹配和语义匹配，将提及与外部知识库（例如 维基数据）进行匹配，以统一不同提及。因此，本文提出一组精简的四种核心关系——由……实现、被……使用、由……评估和相关——（表4）。为进一步捕捉细微的同类型交互，本文引入了层次化细化（表4）：任务细分为对象和问题，其中相关关系连接任务–对象和任务–问题；方法细分为过程，其中相关关系连接方法–过程。

（2）细粒度知识实体关系抽取

传统流程在识别实体后再检测关系，这可能导致错误传播并降低整体性能（Q. Li 和 Ji, 2014年；Shang、Huang 和 Mao,

表 4 学术论文中知识实体之间的关系 。

主实体	Sub-实体	实体描述	实体关系
Task	对象	直接观察到的具体事物或现象，分析， 和研究， 界定研究任务的范围，即研究对象。	任务 *related* 对象
	问题	研究任务中的子任务或子问题。	任务 *related* 问题
方法	过程	改变或生成有关.....的信息的操作研究方法中研究对象的状态。	方法 *related* 过程 任务 *achievedBy* 方法
材料	—	—	材料 *usedBy* 任务
指标	—	—	任务 *evaluatedBy* 指标

2022年）。鉴于摘要中的许多关系很大程度上由类型决定（表3），关系抽取被表述为：（i）判断两个实体是否相关的二元决策，随后进行（ii）类型到关系的映射。

由于摘要中通常包含多个同类型实体，因此采用基于依存句法、以精确率为导向的启发式配对策略，以避免静态映射产生虚假配对。每种可接受的类型对都关联一小组依存路径模板（例如，主语 + 直接宾语、名词性修饰语、关系从句修饰语；表5）。只有当实体对 (a) 符合其中一种模板，且 (b) 在局部语义上连贯时，才会实例化关系；否则不生成边。该设计优先保留可安全注入知识图谱并用于下游推荐的高置信度边。

对标注数据的初步评估表明，$Precision > 85\%$、$Recall \approx 60\%$ 和 $F1 \approx 0.70$，说明该方法在标注资源有限的情况下具有实用性。小规模人工审查进一步证实，与模板匹配的关系大多正确；因此，在知识图谱构建过程中为这些边赋予较高权重，以减轻噪声边带来的误差传播，并降低推荐系统对偶发遗漏的敏感性。

表 5 用于实体关系抽取的依存路径模板。

实体组合	关系	典型路径/关系 依存关系	精确率	召回率	$F_1 - score$
任务—方法	由.....实现	nsubj+dobj，nmod	0.85	0.65	0.73
任务—材料	被.....使用	nsubj+dobj，nmod	0.88	0.62	0.73
任务—指标	由.....评估	nmod	0.90	0.58	0.70
任务—对象/问题	相关	名词修饰语、从句修饰语	0.87	0.63	0.73
方法—过程	相关	名词性主语+直接宾语、名词性修饰、从属从句修饰	0.83	0.60	0.70

随后，我们在§4.3.2（表12）中按由.....实现 的关系类型量化对缺失边的敏感性，结果表明，平均准确率均值/归一化折损累计增益的准确率最受被.....使用 对多样性的贡献更大（列表内多样性/覆盖率）。”

需要指出的是，关系归纳组件旨在用于大规模科学文献的组织与推荐，强调可扩展性和低标注成本；此阶段不会训练完全监督的关系抽取器。由于基于依存关系的召回受模板覆盖率限制，未来工作将探索联合建模以及半监督/弱监督关系抽取，在保持高精确率的同时提高覆盖率。

3.2.3 学术论文知识图谱的自动填充

我们首先概述与科技论文相关的细粒度科学知识图谱（FG-SKG）的结构，然后介绍一种图谱自动填充流程。FG-SKG 的有效性在很大程度上取决于实体识别的准确率，因为识别实体并提取其关系是必要的前置步骤。

（1）FG-SKG 的结构

FG-SKG 旨在系统地表示学术论文的核心要素——摘要、细粒度知识实体，以及论文内部和论文之间的关系。该图谱采用异构信息网络形式，以便在统一的多关系空间中灵活编码多种节点和边类型。这种设计支持推荐任务中的高效检索与遍历，同时能够捕捉学术内容固有的复杂度（例如引文链接和实体间语义）。

形式化地，知识图谱定义为：$G = (P, E, L_{PP}, L_{PE}, L_{EE})$，其中 P表示论文，E 表示知识实体。实体分为四种主要类型：$E = T \cup M \cup D \cup R$，其中包括 T（任务）、M（方法）、D（材料/数据集）和 R（指标）。对于论文 $p \in P$，其实体集合为 $E_p = (T_p, M_p, D_p, R_p) \subseteq E$。边编码三类关系：（i）$L_{PP} \subseteq P \times P$，表示论文之间的引文关系；（ii）$L_{PE} \subseteq P \times E$，表示论文-实体链接；（iii）$L_{EE} \subseteq E \times E$，表示实体间关系，包括表 4 中定义的四种基本语义链接（例如 achievedBy、usedBy、evaluatedBy、related）。每种节点/边类型都同时存储内容属性（例如摘要、实体描述）和结构属性（例如引文邻域、类型化语义连接），从而形成对研究主题、方法体系、资源和评估实践的动态互联表示。

（2）细粒度科学知识图谱的自动填充

为构建知识图谱，我们采用多阶段流程来识别、分类并链接论文及其组成实体。该过程包括：（i）实体识别和文档内关系归纳，（ii）引文链接，以及（iii）图谱实例化。

第 3.2.1 节所述的模型首先应用于论文标题和摘要，以检测细粒度实体——任务 *(T)*、方法 *(M)*、材料 *(D)*和指标 *(R)*——并根据预定义模式对其进行分类。随后，将类型关系——由……实现、被……使用、由……评估和相关——推导为L_{EE}的组成元素。提取出的实体被存储为节点集合，从而形成每篇论文细粒度内容的结构化表示。

与此同时，借助引文解析工具和辅助脚本建立论文间引用（L_{PP}）和论文-实体链接（L_{PE}），以丰富知识图谱中的论文间及论文与实体之间的连接关系。

FG-SKG 可物化为以下任一形式：(a) 支持异构网络的图数据库，或 (b) 单个 JSON 文件。节点和边的形式化定义如下：

$$\text{Papers} = [\{\text{paper_id}, \text{title}, \text{abstract}\}, \dots]$$

$$\text{Entities} = [\{\text{entity_id}, \text{name}, \text{domains}, \text{top_type}\}, \dots]$$

$$\text{data} = \begin{Bmatrix} \text{nodes} : [\{\text{id}, \text{category}\}, \dots], \\ \text{links} : [\{\text{source}, \text{target}, \text{relation}\}, \dots], \\ \text{categories} : [\text{name}, \dots] \end{Bmatrix}$$

此处，节点包括论文和实体，每个节点都标有类别（论文、任务、方法、材料、指标）。链接用于编码类型化关系，包括 L_{PP}（引文）、L_{PE}（论文-实体关联）和 L_{EE}（实体间关系），并指定源节点、目标节点和关系类型。类别列举了节点类型，以便高效筛选和执行查询操作（例如，以任务或方法为重点的检索）。

这种结构化设计构建出高效且可扩展的 FG-SKG，作为推荐流程中候选生成和关系感知重排序的基础，利用相互关联的实体和多跳关系路径，支持精准且多样化的学术推荐。

3.3 论文的嵌入表示

论文的多维信息包括内容（标题/摘要和精细粒度的知识实体）以及引文上下文。为满足多样化的学术需求，我们将这些信号嵌入为向量，并根据节点关系将其组合起来，从而形成多个粒度下可比较的表示。

3.3.1 图节点嵌入

嵌入的目标是学习低维向量，使相互引用、文本内容相似或包含相似精细实体的论文在嵌入空间中彼此接近。本文考虑两种节点类型：论文和知识实体。

对于论文节点，采用基于 Transformer 的文档编码器 *SPECTER* (Cohan 等人, 2020年)，SPECTER 联合建模标题和摘要的语义以及引文关系，并在学术嵌入任务中展现出优异的性能。将预处理后的标题和摘要输入 SPECTER，以获得论文向量$\mathbf{s}_P$。

对于实体节点，使用基于*GPT* 的编码器对简短的定义性或描述性上下文进行建模。由于实体语义主要是序列式的（而非图结构化的），且通常缺少引文链接，因此*GPT*的序列建模非常适用。直接将*SPECTER* 应用于实体会导致结构不匹配

并且需要从头开始训练新的实体级语料库。因此，使用GPT对实体描述进行编码，以获得语义向量。对任务和方法下的子实体进行一致分组，以保持类型一致性。

论文中的实体嵌入通过求和池化进行聚合，这是一种简单且可解释性强、具有良好可扩展性和泛化能力的集合表示技术（Zaheer等人，2017年）。令 $\mathbf{c}_p^t$、$\mathbf{c}_p^m$和 $\mathbf{c}_p^d$ 分别表示 任务、方法和 材料/指标实体的聚合向量，并令 $GPT(\cdot)$ 为实体编码器：

$$\mathbf{c}_p^t = \sum_{e_i^t \in T_p} GPT(e_i^t) \tag{1}$$

$$\mathbf{c}_p^m = \sum_{e_i^m \in M_p} GPT(e_i^m) \tag{2}$$

$$\mathbf{c}_p^d = \sum_{e_i^d \in D_p} GPT(e_i^d) \tag{3}$$

3.3.2 通过节点关系进行向量组合

实际上，研究人员会以不同粒度进行查询（例如仅查询任务；任务 +方法；任务 + 材料/指标；或全部查询）。为支持候选项选择和排序之前的此类使用场景，论文表示通过将论文向量与一个或多个聚合实体向量拼接而成：

总体表示（包含所有实体的论文）：

$$\mathbf{p}_g = [\mathbf{c}_p^t, \mathbf{c}_p^m, \mathbf{c}_p^d, \mathbf{s}_p] \tag{4}$$

双实体组合：

$$\mathbf{p}_{tm} = \left[\mathbf{c}_p^t, \mathbf{c}_p^m, \mathbf{s}_p\right], \quad \mathbf{p}_{td} = \left[\mathbf{c}_p^t, \mathbf{c}_p^d, \mathbf{s}_p\right]. \tag{5}$$

单实体组合：

$$\mathbf{p}_t = \left[\mathbf{c}_p^t, \mathbf{s}_p\right], \quad \mathbf{p}_m = \left[\mathbf{c}_p^m, \mathbf{s}_p\right], \quad \mathbf{p}_d = \left[\mathbf{c}_p^d, \mathbf{s}_p\right]. \tag{6}$$

这些组合为下游推荐提供了多粒度、可比较的表示。

3.4 基于向量运算的论文推荐

学者在文献检索过程中往往有着多样的信息需求（Bawden，2007）。除了寻找主题相似的论文，许多人还会有意查阅采用不同方法或材料来处理相同任务的研究，以拓宽视角并促进创新（J.Liu, Mitsui, Belkin, & Shah，2019年；Talley等人，2011）。构建多维嵌入后，候选筛选与排序被建模为基于相似度的检索，并采用学习得到的组合权重。

（1）向量运算的设计

给定一篇查询论文 $q \in P$，分别沿精细维度（任务、方法、材料/指标）以及整体内容计算语义相似度。令cos_{task}、cos_{method}和 $cos_{met\ metric}$ 表示各维度相似度，cos_{enti}表示整体相似度。目标向量 $\mathbf{p}$ 由四种表示的凸组合构成：整体 $\mathbf{p}_g$，以及单实体组合 $\mathbf{p}_t$、$\mathbf{p}_m$和$\mathbf{p}_d$：

$$\mathbf{p} = \sum_{i\in\{g,t,m,d\}} w_i\,\mathbf{p}_i, \qquad w_i \geq 0,\ \sum_i w_i = 1. \tag{7}$$

$$\cos(\mathbf{q},\mathbf{p}) = \frac{\mathbf{q}^\top \mathbf{p}}{\|\mathbf{q}\|\,\|\mathbf{p}\|}. \tag{8}$$

此处，$p_g = [c_p^t, c_p^m, c_p^d, s_p]$；$p_t$，$p_m$，$p_d$ 分别对应公式（6）中的单实体组合。在任务相似子集 $CP_{task} \subset C_P$中，多个信号（方法相同/不同；材料/指标相同/不同）通过另一种学习得到的凸组合进行聚合：

$$\cos(\mathbf{q},\mathbf{p}) = \sum_{j=1}^{4} \alpha_j\, s_j(\mathbf{q},\mathbf{p}), \qquad \alpha_j \geq 0,\ \sum_j \alpha_j = 1, \tag{9}$$

其中，$\{s_j\}_{j=1}^4$ 依次对应 $cos_{task\ method}$ $(\mathbf{p} = \mathbf{p}_{tm})$、$cos_{task} - cos_{method}$、$cos_{task} - cos_{mat/metric}$ 和 $cos_{task\ mat/metric}$ $(\mathbf{p} = \mathbf{p}_{td})$。候选论文按以下方式排序：

$$\mathrm{rank}(\mathbf{q},\mathbf{p}) = \sum_{i\in\{g,t,m,d\}} w_i \cos(\mathbf{q},\mathbf{p}_i) \quad \text{or} \quad \sum_{j=1}^{4} \alpha_j\, s_j(\mathbf{q},\mathbf{p}). \tag{10}$$

并返回 $top-N$ 个项目。

（2）组合权重的学习与优化

公式（7）和（9）中的权重在开发集划分上进行学习，唯一优化目标是经准确率调优（例如 平均准确率均值）（$\lambda = 0$），随后固定用于测试。我们采用两阶段过程：（i）以原始启发式配置初始化的粗粒度网格搜索（步长 0.05）；（ii）进行坐标上升并细化网格（步长 0.01），将更新投影回单纯形，并在连续五次迭代未能改进后提前停止。敏感性分析通过每次扰动一个权重 $\pm$10%，以及 $\pm$20%（随后重新归一化）进行，同时保持所有其他设置不变。在 STM-KG（域内）上，每种配置使用三个随机种子运行；报告均值和 $\pm$标准差，并通过配对 t检验评估显著性。

如图 3所示，学习得到的权重使 MAP@50 和 nDCG@50 均获得稳定且显著的提升，而最优值附近的局部扰动会导致性能变化 ≤ 1.5%，这表明模型具有鲁棒性。降低任务权重会导致最大幅度的性能下降，证实了任务语义在学术相似性中的主导作用；对方法和材料/指标权重进行适度扰动，影响则较小，这表明它们具有互补作用。STM-KG（域内）的学习得到的权重如下：

$$w^\star = (0.46, 0.28, 0.18, 0.08) \quad \text{and} \quad \alpha^\star = (0.40, 0.30, 0.20, 0.10),$$

条目顺序与原始表述一致。在独立测试集上也复现了这一提升。

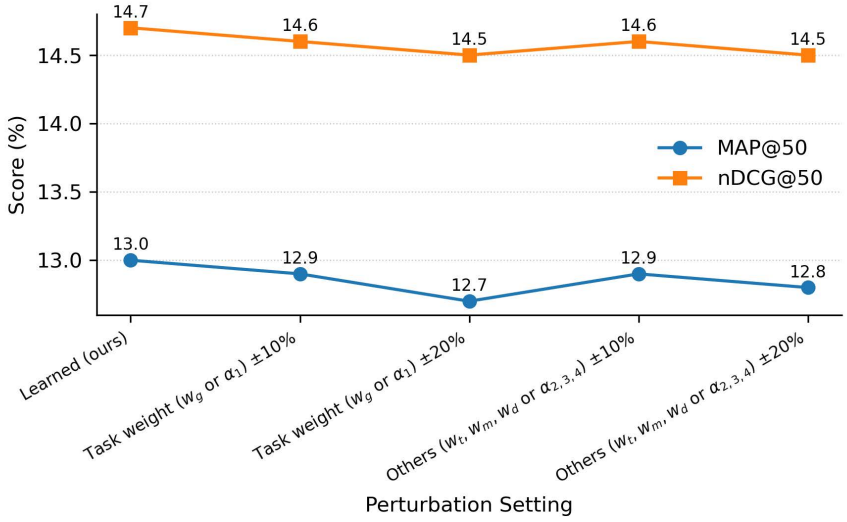

图 3 STM-KG（域内）上对权重扰动的敏感性。

3.5 论文推荐的算法细节

为提升可复现性和清晰度，我们总结了从 FG-SKG 构建到排序的完整流程。这些步骤对应于 §3.2–§3.4 和公式 (1)–(10)：节点嵌入（§3.3.1）、多向量组合（§3.3.2）、候选生成与任务感知优化（§3.4），以及结合敏感性分析的权重学习。

3.5.1 我们的方法的伪代码

附录 B 展示了本文提出的推荐流程的伪代码。表 6 列出了伪代码中使用的符号及其定义。对于大小为 $|\mathbf{CP}| = K$、向量维度为 d 的候选池，公式 (8) 中的余弦评分耗时 $O(K \cdot d)$。公式 (9) 中的任务感知优化还会增加 $O(K \cdot d)$。公式 (4)–(6) 中的拼接对每篇论文执行一次（$O(d)$）。关系抽取（§3.2）在线下执行，不会影响测试时复杂度。

3.5.2 与先前工作的差异

（1）与单向量文本编码器相比。以往系统通常用单个主题级向量表示每篇论文，这限制了控制能力和可解释性。我们的方法组合了多个实体感知向量（任务/方法/材料/指标），并学习

T表 6 第3.3–3.4节中使用的符号和定义（续）　　续）。

符号	定义
P, E	论文集合；FG-SKG 中的实体集合
T_p, M_p, D_p	论文 p 的任务、方法、材料/指标实体
$\mathbf{s}_p$	论文 p 的 SPECTER 嵌入（标题+摘要）
GPT（e）	实体 e的 GPT 嵌入
$\mathbf{c}_p^t, \mathbf{c}_p^m, \mathbf{c}_p^d$	p 的求和池化实体向量（任务/方法/材料）
$\mathbf{p}^g$, $\mathbf{p}_t$, $\mathbf{p}_m$, $\mathbf{p}_d$	拼接论文向量（公式 4–6）
$\mathbf{q}$, $\mathbf{p}$	查询论文向量和目标论文向量
w_i	在 $\{g, t, m, d\}$ 上的非负权重，含 $\sum_i w_i = 1$ （公式 7）
α_j	四种任务子集信号上的非负权重，含 $\sum_j \alpha_j = 1$ （公式 9）
CP, CP_{task}	候选池；任务相似子集
余弦相似度$(\cdot, \cdot)$	余弦相似度（公式 8）
排序$(\mathbf{q}, \mathbf{p})$	最终排序得分（公式 10）

凸权重（公式 7 和 10），从而能够精细控制哪些方面影响相似度和推荐理由。

(2) 相较于纯粹基于引文的推荐系统。基于引文的方法可能过度强调流行度，并带来公平性或多样性问题。我们融合文本语义、实体语义和引文上下文，并通过精确率优先的关系归纳（§3.2）注入高置信度边，从而提升排序稳定性，并增强系统对缺失链接的鲁棒性（见 §4.4.2）。

(3) 相较于缺乏动态组合的知识单元方法。现有的单元方法通常将各单元孤立处理，或依赖静态角色。我们引入任务感知优化（公式 9），根据条件重新组合信号（相同/不同的方法/材料），以满足常见的学术意图（“相同任务，不同方法/材料”）。

我们的方法优于上述方法，主要原因如下：(i) 学习得到的凸权重相较于均匀加权和单向量基线带来了具有统计显著性的提升，而敏感性分析显示，在±10%–±20%的扰动下，波动幅度为 ≤1.5%；(ii) 以精确率为导向的关系归纳避免了噪声边导致的误差传播，带来了稳定的净收益；(iii) 多向量组合提供了可解释、可调节且具备多样性感知能力的推荐结果，能够满足用户的细粒度需求。

4 结果分析

为验证基于细粒度知识实体的论文推荐方法的有效性，本研究从$STM\text{-}KG$ (Brack, M¨uller等人，2021年）数据集中抽取细粒度知识实体，以构建细粒度知识图谱。随后，实施了基于多维文档嵌入组合的学术论文推荐。之后，将知识实体抽取和学术论文推荐子任务的实验结果与相关基准进行了比较和分析。

4.1 数据集

本文借鉴了Brack等人的系列研究，使用STM-corpus数据集验证知识实体任务，并使用STM-KG数据集验证论文推荐任务。*STM-corpus* (Brack、D'Souza、Hoppe、Auer和Ewerth，2020年）包含110篇论文摘要，分别属于10个不同的科学领域（农业、天文学、生物学、化学、计算机科学、地球科学、工程学、材料科学、数学、医学），每个领域有11篇摘要。*STM-KG* (Brack, M¨uller等人，2021年）是一个基于上述10个学科的55,485篇学术论文摘要构建的知识图谱，通过自动抽取科学概念形成。该图谱涵盖论文标题、摘要和引文网络（包括15,395条引文链接，其中包含2,200条跨学科链接），并标注了细粒度知识实体及其关系。这种结构与我们的推荐模型中"多维语义组合和实体增强表示"的建模目标高度一致。该数据集用于基于知识图谱的引文推荐子任务。其标注质量良好且具备复用基础，为我们评估不同实体嵌入策略和推荐性能提供了有力支持。尽管*STM-KG* 不如一些通用数据集（如 *OpenAlex* 和*AMiner*）知名，但其在结构完整性和实体语义信息方面的优势，使其非常适合用于验证我们提出的推荐方法。

由于本研究使用的知识实体类型与该图谱中的科学概念类型并不完全一致，因此选取包含完整标题和摘要信息的*STM*语料库数据集中的论文作为实体识别语料库。完成句子切分后，使用*BRAT* 标注工具。在*STM*语料库(Brack等人，2020年)的110篇论文摘要标注基础上，新增了对象、任务、问题等实体的标注，并重新训练了实体自动识别模型。我们采用*BIO* 标注格式识别实体边界，并制定了统一的标注规范（见附录 A），以明确各类实体的定义标准、上下文判别策略，以及嵌套或重叠冲突的处理方法。标注工作由3名具有自然语言处理或计算机科学背景的研究生完成。所有标注人员在正式标注前均接受了培训，并通过预标注测试，以确保理解一致。每个样本均由两名标注人员独立完成，存在冲突的样本则由一名经验丰富的研究人员进行复核。我们从标注集随机抽取300个样本，计算实体层面的科恩κ系数。平均一致性得分为0.82，表明标注结果具有较高的一致性。实体标注结果的统计数据见表7。

为避免同期引文链接造成训练集与测试集相互交织，我们在编码器预训练和评估时间窗口之间按时间划分。具体而言，所有评估论文及其入向／出向引文的时间均晚于编码器的预训练截止时间；编码器保持冻结，且我们的流程不会使用评估期间的任何引文边来拟合嵌入参数。此协议杜绝了评估时的引文关系泄漏到表示学习中的可能性。

随后，我们从剩余的 *STM-* 知识图谱 中选取至少有 4 条引文的论文作为查询论文，并构建了一个包含 720 篇论文和 4069 条引文链接的推荐评估集。使用第 3.2.1 节训练的细粒度知识实体识别模型，以及第 3.2.2 节的实体关系映射方法，为这些论文构建知识实体图后，我们使用各种嵌入向量表示每篇查询论文，并通过向量运算从数据集中的其他所有论文中检索排名前 K 的论文，以供推荐。

表7 *STM-corpus* 数据集的实体标注和数据增强结果。

类别			Task	对象	问题 方法 过程			材料 指标	
数量标注	of	实体	811	698	287	258	1,301	2,099	1,658
数量论元	of	实体	1,072	923	379	341	1,720	2,774	2,191
句子数量			总计：2,487						
数量增强	句子		总计：3,287；训练集：2,629；验证集：329；测试集：329						

4.2 评估指标

本研究的评估指标包括知识实体抽取和论文推荐两个子任务。对于知识实体抽取，精确率、召回率和 $F_1-score$ 是常用的指标。我们采用片段级严格匹配来评估知识实体识别，要求预测实体的起始位置和结束位置必须与标注实体完全一致，边界必须完全相同，且实体类别也必须相同，才能被视为“命中”；否则，无论部分重叠程度如何，都视为“未命中”。这一判断设置与 CoNLL-2003（Tjong Kim Sang & De Meulder, 2003）等主流命名实体识别评估标准一致，有助于准确反映模型的边界判断能力。

论文推荐的目的是推荐符合学者特定研究需求的多样化论文。在本研究中，我们通过衡量查询论文的参考文献与相似度高于平均相似度的推荐结果之间的重叠程度，来评估推荐方法的准确率。重叠程度越高，表明推荐论文越符合学者的需求。此外，我们还通过比较查询论文的参考文献与推荐结果之间的重叠程度来评估推荐结果的多样性。重叠程度越低，说明该推荐方法能够为学者针对特定研究任务提供更加多样化的论文。

这两个重叠值均使用 K点平均准确率均值（$MAP@K$）指标计算。 $MAP@K$是所有查询论文的前 N 个推荐结果的平均准确率（$AP@N$）的均值。如公式（11）所示， $AP@N$

用于衡量针对特定查询论文推荐的前 1 至 N 个结果的精确率，即这些结果与目标论文的匹配准确度。

$$AP@N(p) = \frac{\sum_{k=1}^{N} \text{Precision@}k(p) \cdot \text{rel}(k)}{\#\text{ relevant documents for } p} \tag{11}$$

$Precision@k$ 表示前 k 个推荐结果中被引用的比例。如果位置 k 上的推荐结果被引用，则 $rel(k)$ 的取值为 1，否则取值为 0。

除了考虑是否推荐了相关文献外，$nDCG@K$还重点关注相关文献在推荐列表中的排序，强调高度相关的文献应尽可能靠前出现。如公式（12）所示

$$nDCG@K(p) = \frac{\sum_{i=1}^{K} \frac{2^{r_i}-1}{\log_2(i+1)}}{IDCG@K(p)} \tag{12}$$

$IDCG@K(p)$ 表示理想排序下的 DCG 值，用于归一化。$rel(i)$ 表示推荐结果中第 $i-th$ 个位置的相关性等级，通常取 0（不相关）或 1（相关）。

我们引入了两个多样性指标，主要将其应用于我们的方法，使用学习得到的权重组合不同的向量分量：

- 列表内多样性（ILD）：用于衡量推荐论文的平均成对不相似度。较高的 ILD 得分表示列表中的差异性更大。如公式（13）所示

$$ILD@K \;=\; \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} \Big(1 - \cos\big(f(r_i), f(r_j)\big)\Big) \tag{13}$$

$f(\cdot)$表示用于多样性评估的嵌入函数。

- 项目覆盖率：在所有测试查询的 $top-K$ 推荐结果中，不重复论文所占的比例。该统计量用于衡量推荐系统在语料库层面的广度和独特性。如公式（14）所示

$$\text{Coverage@}K \;=\; \frac{\big|\{\, b(r) \,|\, r \in R_K \,\}\big|}{\big|\mathcal{B}_{\text{cand}}\big|} \tag{14}$$

$\mathcal{B}_{\text{cand}}$ 表示候选池中该查询所涉及实体对应的语义桶集合，$b(r) \in B$ 将论文映射到其所属的语义桶，$R_K = \{r_1, \ldots, r_K\}$ 表示查询论文的前K 个推荐列表。

4.3 知识实体与关系抽取的实验结果

4.3.1 实验设置

为避免文献长度差异造成的训练偏差，并确保每个子集都包含所有实体，本文随机划分标注

数据集，并在句子级别按照 8:1:1 的比例划分为训练集、验证集和测试集。我们比较了使用双向长短期记忆网络+条件随机场 （称为方法 a）、$SciBERT$ （称为方法 b）、$SciBERT$+条件随机场 （称为方法 c）以及本文提出的方法进行知识实体识别的性能。

为确保模型评估结果的稳定性和代表性，我们在其余 90% 的句子数据上进行了 9 折交叉验证，同时保持原始测试集（10%）不变。我们将这部分数据等分为 9 份，每一折依次用作验证集，其余部分用于训练，最终在固定的测试集上进行评估。

为进一步验证所提出的方法与基线模型之间的性能差异是否显著，我们对 9 折交叉验证中 $SciBERT$、$SciBERT$+双向长短期记忆网络+级联 模型的 $F_1-score$ 进行了统计检验。

为评估大语言模型（LLM）在学术知识实体识别中的有效性，我们将监督模型与基于当前主流大语言模型 $GPT3.5$ 的零样本和少样本实体抽取方法进行了比较。我们通过 $OpenAI$应用程序接口，将 ***STM***语料库 中的标题和摘要作为输入传递给 $GPT\text{-}3.5$ 。每段文本均单独处理，并与金标准标注进行比较。我们测试了两种 大语言模型 提示策略：

(1) 零样本：要求模型仅使用自然语言指令（不使用任何标注示例）从摘要中提取实体。

(2) 少样本（3 个示例）：在提示词前添加三个标注示例，以引导模型进行上下文学习。

为评估从依存句法模板中提取的高精确率关系对推荐性能的贡献，以及它们在关系缺失情况下的稳健性，我们在 $STM-KG$数据集上进行了两组受控实验。第一组实验侧重于评估贡献：在保持推荐模型参数一致的前提下，我们比较了仅包含知识实体的 $Entity-only$ 模型与添加了高置信度关系的 $Entity+Rel$ 模型；这些关系经过依存路径模板匹配和小规模人工抽查验证。第二组实验侧重于评估关系缺失情况下的稳健性：以$Entity+Rel$模型为基线，我们随机保留一定比例的关系边：$r\in\{100\%，75\%，50\%，25\%\}$，同时保持所有其他训练和评估设置不变，以模拟召回率进一步不足的情况。两组实验均在域内独立运行三次。我们报告MAP@50 和 $nDCG$@50 的均值和标准差，并通过配对t检验评估显著性。

如表8所示，我们仅汇总了信息抽取组件的超参数；关系归纳基于规则，无需训练。

4.3.2 实验结果与分析

（1）实体识别的比较分析

的精确率、召回率和 $F_1-score$ 不同模型下知识实体识别的9。本文提出的 $SciBERT$ + 双向长短期记忆网络 + 级联模型表现最佳，微观 $F_1-score$ 达到 85.10。这一优异表现可归因于多个因素。首先，与方法

表8 知识实体识别的关键超参数

模块	方法	批次	LR	隐藏维度	训练轮数	最大序列长度
实体识别	方法a	64	0.001	128	—	—
	方法 b	512	2e-5	—	—	512
	方法 c	512	2e-5	—	4	512
关系抽取	使用依存路径模板（启发式），因此没有可训练的超参数。					

知识实体识别，运行时间：Python 3.11.5。

a 相比，本研究采用了$SciBERT$ 预训练语言模型，该模型专为科学论文设计，能够对句子级语义特征进行高度区分性的表示。此外，本研究采用数据增强技术扩充了原先由 110 篇论文组成的标注语料库，增加了训练样本的多样性和数量，从而增强了模型的鲁棒性。此外，级联方法实现了实体识别和分类任务的级联学习，减少了级联错误。与方法 b 和方法 c 相比，这种方法不仅减少了级联错误，还提供了更大的训练数据集。值得注意的是，实验结果显示，在$SciBERT$ 层之后添加条件随机场层，对知识实体抽取产生了轻微的负面影响。对抽取结果的详细分析表明，条件随机场学习到的实体顺序干扰了$SciBERT$ 模型捕捉到的上下文关系。虽然$SciBERT$模型能够准确嵌入词语语义，但也能捕捉上下文关系，因此无需额外的双向长短期记忆网络层即可获得令人满意的知识实体抽取结果。

从不同模型在各实体类型上的 $F_1-score$ 来看，所提出的 $SciB\text{-}ERT$+双向长短期记忆网络+级联 在 任务/方法上仍然表现最佳，而 $SciBERT$ 和$SciBERT$+条件随机场 模型分别在 材料 和 指- 标 实体的 $F_1-score$ 上略占优势。$SciBERT$ 能够捕捉任务句中的目标/问题导向语义以及方法句中的领域特定短语；双向长短期记忆网络能够平滑长分句和依存关系；而级联 阶段则先检测实体，再对其进行分类，从而消除实体类型混淆并减少误差传播。与材料相关的实体短语可能较为稀疏，且拼写独特；纯$SciBERT$的领域特定预训练子词统计已能很好地适应这些短语，而加入双向长短期记忆网络/级联 可能会对复合短语施加过度正则化或引入偏差，从而产生轻微偏差。由于条件随机场具有更强的标签转移建模能力和边界一致性，因此在识别较短的短语标签（例如指标 实体）方面表现更好。条件随机场的标签转移约束减少了数字/符号及相邻单位周围的边界翻转，从而使边界更加清晰，并减少 BIO错误。

如表9所示，结果表明$GPT3.5$ 模型表现出较强的泛化能力，尤其是在低样本环境中。然而，其整体表现仍逊于我们的监督模型。此外，$GPT3.5$ 的输出有时会出现结构不一致和实体边界重叠的问题，从而限制了其在知识图谱构建和论文推荐等下游任务中的可靠性。本文提出的微调后的监督模型

论文在细粒度学术环境中展现出出色的准确率、类别区分能力和输出一致性。

如表10所示，主模型*(SciBERT+*双向长短期记忆网络*+*级联）在不同测试子集上的表现稳定，$F_1 - score$ 范围为 -0.50 至 $+0.39$，波动很小（$\sigma = 0.28$，完整范围为 0.89），验证了当前数据划分的合理性以及该方法的鲁棒性。配对样本t检验的 p 值为 0.004，Wilcoxon符号秩检验的 p 值为 0.0039。两项检验均表明，性能差异具有统计显著性（$p < 0.05$），这说明所提出的模型在细粒度实体识别任务中具有稳定、可靠的性能优势。

为进一步评估我们的模型在其他数据集上的实体识别性能，我们收集了两个相关的开放数据集，即$SciERC$ 和TDM，以将我们模型的评估结果与现有模型进行比较。表11 显示，我们的模型在 $F_1-score$ 上取得了最高的 69.77 分，在TDM 数据集上表现出色，并且在$SciERC$ 数据集上也表现良好。结果表明，我们的方法在科学实体识别方面具有出色的泛化能力。

(2) 特定关系类型的遗漏分析

在一个经过标注的样本上，我们估计每种关系类型的遗漏率和典型错误来源；同时列出具有代表性的失败案例，以便复现。

- 由……实现 （任务–方法）——遗漏率为 $\approx 6\%$。原因在于长距离依存关系（任务和方法分布在不同分句中）、名词化或省略的方法提及，以及破坏依存链的并列结构。例如：“我们通过一种新颖的 <任务> 来处理 <方法>…”，其中方法被分散在多个句子中，因而无法匹配出有效路径；
- 被……使用 （材料–任务）——遗漏率为 $\approx 10\%$。原因在于介词结构或同位语表达（“使用/借助/基于”），以及掩盖具体数据集/材料的通用占位语（“一个标准基准”）。例如：“我们在该 <数据集> 上评估<任务>……”，被解析为修饰语而非链接；
- 由……评估 （任务–指标）——遗漏率为 $\approx 3\%$。原因在于隐含的指标提及（“宏平均精确率为0.91”中缺少显式的“F1”标记），以及符号/数字相邻导致的BIO边界偏移。例如：“取得宏平均精确率和召回率……”，未被整合为统一的指标实体；
- 相关 （类型内）——遗漏率为 $\approx 21\%$。原因在于，层次化变体或同义变体分布在不同句子中；跨句共指未被具体化为显式边（我们转而依赖外部知识库匹配）。

大多数遗漏出现在宽泛的相关 类别中（通常表现为分散的措辞和较长距离的依存关系），其次是被……使用 （资源使用的词汇表达多样）。由于具有明显的谓词线索，由……实现 相对而言捕获得较为充分；由……评估则受益于稳定的评估动词和模板。关系归纳采用高精确率的依存模板；众所周知，召回率受到模板覆盖率的限制。

（3）实体关系贡献分析

在高精度知识实体关系的贡献方面，如图 4 所示，$Entity + Rel$ 在 STM-KG（域内）上取得了稳定且显著提升的推荐性能，相较于 $Entity - only$：两者

MAP@50 和 $nDCG$@50 分别提高了约 $+2.1$ 和 $+2.4$ 个绝对百分点（配对t检验，$p < 0.01$）。这表明，高置信度关系为知识实体提供了文本之外的结构约束，从而有效提升排序质量。

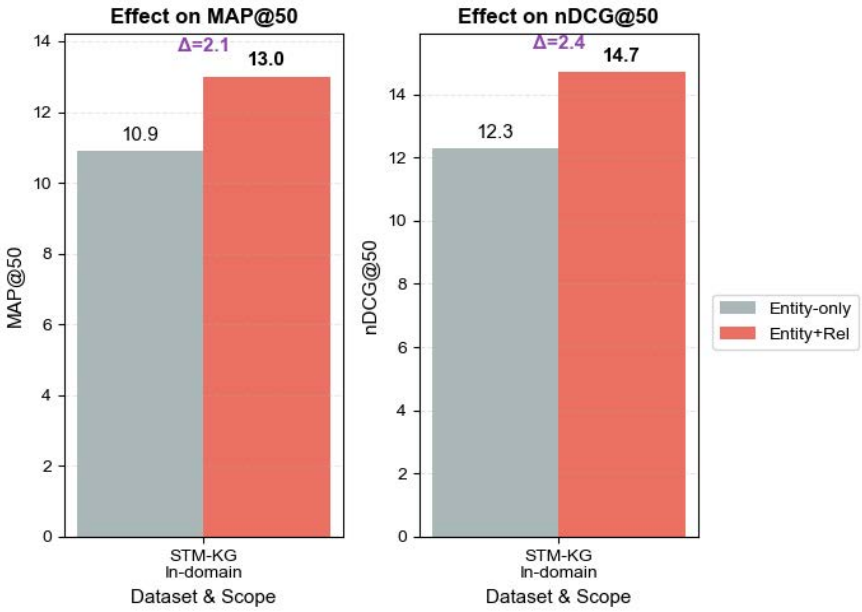

图 4 STM-KG 中高精度关系的贡献（域内）。

关于对缺失关系的鲁棒性，如图所示5，在其他设置均保持不变的情况下，可用关系边的保留比例逐步从 100% 降至 75%、50% 和 25%。两项指标均呈现近乎线性且均匀的下降：保留率为 50% 时，MAP@50 / $nDCG$@50 相较于仅实体仍保持显著的正向增益（约为 $+0.5$ 至 $+1.1$，$p < 0.01$）；即使保留率仅为 25%，性能仍与仅实体相当（约为 $+0.1$ 至 $+0.3$，$p \approx 0.05$）。这些结果表明，使用依存模板提取的关系切实可用且有效，其正确边的质量足以带来净收益。此外，该系统的性能下降幅度可控，并且在关系缺失时仍能持续发挥结构贡献，这验证了我们优先保证高精度、再逐步扩大召回率这一设计权衡的合理性。

（4）关系类型特定影响对推荐的影响

表12 显示，移除由……实现后，准确率（平均准确率均值、归一化折损累计增益）下降幅度最大，但多样性的下降相对温和，这表明“任务—方法”关系主要有助于提供精确的匹配信号。相反，移除被……使用后，列表内多样性和覆盖率显著下降，表明资源渠道是结果多样性的主要来源。关于由……评估和

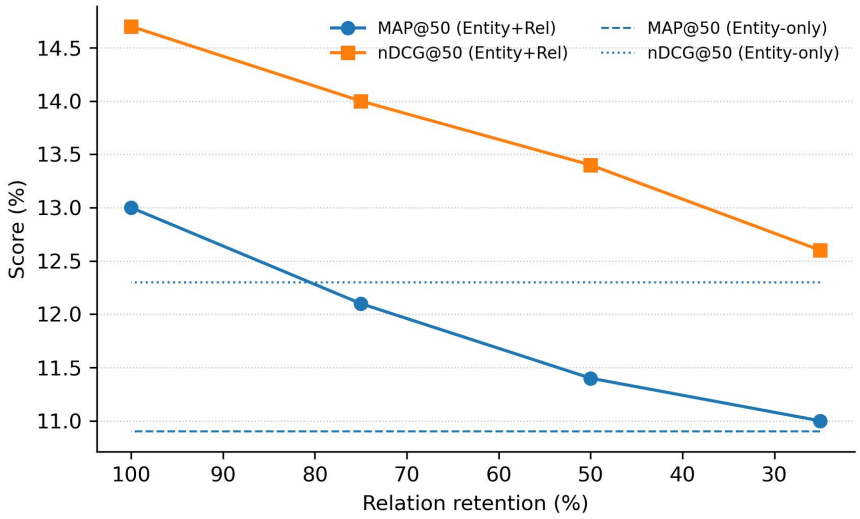

图 5 STM-KG（域内）对关系缺失的鲁棒性。

相关的影响中等或较小，更多体现在细粒度重排以及覆盖边界的轻微收缩上。

尽管本文采用的这种关系抽取策略在实际应用中具有较好的实用性，但在处理复杂的上下文敏感关系时仍存在一定局限，且尚未与现有的监督学习关系抽取模型进行直接比较。未来工作计划探索联合建模或半监督关系抽取方法，以进一步提升关系抽取效果，并在此基础上与主流关系抽取模型开展系统的比较分析。

4.4 学术论文推荐的实验结果

4.4.1 实验设置

我们提出的方法使用 *SPECTER* 嵌入来表示论文标题、摘要和引文信息，使用 *GPT3.5* 嵌入来表示从论文标题和摘要中提取的知识实体，并使用学习得到的权重组合不同的向量分量，以实现论文推荐。本实验设置主要包括在 *STM-KG* 和 Aminer[4] v12 数据集上的模型比较分析，以及在 *STM-KG* 数据集上的消融研究。

（1）模型的比较分析

在基线比较实验中，我们将未使用学习得到的权重来组合不同向量分量的方法，与两种相似方法进行比较，分别是基线1（Takahashi& Katsurai，2022年）、基线2（Brack, Hoppe, & Ewerth,

[4]https://www.aminer.org/citation

2021年）以及三种最先进的模型，分别是 SMIGNN（Niu, Xing,Jia,Xin,&Xing, 2024年）、TIDRec（Xiao et al., 2025年）、大语言模型增强（K. Liu,Zhang, Pan, Gao,& Wang, 2025年），在 STM-KG 和 *Aminer v12* 数据集上，对域内和跨域论文推荐场景进行比较。本实验旨在验证我们的方法在论文推荐准确率方面的优越表现，并评估各方法在跨域应用时的稳健性和泛化能力。

- 基线1：利用研究目标与方法句子之间语义相似度差异的方法；

- 基线2：根据从摘要中提取的科学概念构建论文嵌入，并使用*SPECTER* 嵌入表示论文标题、摘要和引文信息。

- SMIGNN：引入节点注意力机制和意图注意力机制，从写作-引文图和作者图中提取信息，以获取作者的研究兴趣。

- TIDRec：使用融合双图知识的主模型捕捉作者的全局研究兴趣，并使用单图知识辅助模型探索局部研究兴趣。采用三图交互蒸馏校正全局研究兴趣，最终向作者推荐高度相关的论文。

- 大语言模型增强：采用大语言模型改进论文表示的大语言模型增强引文网络模型。

（2）消融研究实验

为评估向量组合权重、节点文本嵌入模型、不同类型的知识实体以及引文信息对我们模型的影响，我们开展了一组消融研究。

•我们研究了不同词嵌入模型对域内论文推荐准确率的影响。具体而言，我们应用了三种嵌入模型——FastText（Bojanowski、Grave、Joulin 和 Mikolov，2017年）、*SciBERT*和*GPT-3.5* （Radford，2018年）——以获得文档内容的语义表示。我们计算相应文本单元（标题、摘要或实体）中词嵌入的平均值，并使用这些文档级向量进行基于相似度的推荐。

- 我们研究了不同知识实体类型和引文信息对域内论文推荐准确率的影响。具体而言，在移除材料/指标、方法和任务实体以及引文信息后，我们比较了使用不同类型文档信息嵌入进行论文推荐的性能。如果忽略引文信息，则使用*GPT3.5* 模型将论文标题和摘要表示为文档信息嵌入中的一个项目。

- 我们研究加权组合策略对域内和跨域论文推荐效果的影响。具体而言，在使用*GPT-3.5* 和*SPECTER* 进行语义表示的相同设置下，我们比较了采用和不采用学习得到的权重来组合不同向量分量的方法。这使我们能够揭示引入加权

方案对域内论文推荐和跨领域跨学科论文推荐的影响。

- 我们在时间去污染协议下研究正交的纯文本基线和消融实验：(i) BM25（纯文本）；(ii) SciBERT（纯文本）；(iii) 我们的方法（纯文本）：将引文信息编码器替换为纯文本编码器，同时保留我们流程的其余部分；(iv) 我们的方法（完整）：同时使用文本和引文。此外，我们通过移除共引和书目耦合特征，对引文侧结构进行消融。所有设置共享相同的候选池、训练方案和评估指标（$MAP@K$、$nDCG@K$ 以及实体感知相关性度量）。我们使用 3 个随机种子运行实验，并报告均值和标准差；显著性通过配对 t 检验进行评估。

- 为评估高被引论文是否会放大引文特征的作用，我们按照查询集的引文三分位数（低/中/高）报告分层结果，同时还在平衡子集上进行评估；该子集对每篇论文的引文数量设置上限（以上限截断于第 80 百分位数）。

由于论文推荐阶段是一个无需训练的向量检索与加权重排序流程。为确保可复现性，我们还列出了权重搜索设置（目标函数、搜索计划和敏感性分析），以及文档和实体嵌入所用的向量维度，如表 13所示。

4.4.2 实验结果与分析 ysis

我们在指定数据集上按照实验设置开展了各项实验，记录实验结果，并分析产生这些结果的可能原因。

(1) 近期论文推荐方法的性能比较。表 14 和图 6 展示了我们的方法和基线方法在跨域和域内论文推荐中的表现，包括两项评估指标（平均准确率均值和归一化折损累计增益）以及两个数据集（STM-KG 和 AMiner 12）的结果。对于两项评估指标，使用不同 top_k 值（即 top_{10}、top_{20}和 top_{50}）的推荐列表，用于考察这些模型的性能。我们的方法在所有数据集上均显著优于基线方法，且具有统计显著性。对这些表格和图表的分析得出以下见解和分析：

从整体推荐性能来看，首先，我们提出的方法在大多数排名靠前的相关指标上取得了最高得分，尤其是在$MAP@50$ 和$nDCG@50$方面。这表明，该方法在检索更广泛的语义相关论文方面具有优势。这一性能得益于我们模型的混合表示策略，该策略结合了：

- $GPT\text{-}3.5$ 跨领域嵌入用于捕捉跨学科语境语言知识；

- 精细化知识实体向量用于明确编码研究任务、方法和指标。

- 细粒度实体的文本语义、角色及异构关系贯穿召回和排序的全过程，并借助可跨领域迁移的结构不变模式。这些设计使系统即使缺乏通过大语言模型预训练过程获得的文本间深层语义，也能保持语义连贯性，表现可与近期基于大语言模型的方法相媲美。

其次，LLM增强方法在四项评估指标中均排名第二，且结果始终优于 SMIGNN 和 TIDRec。这得益于其大语言模型增强推荐策略。与此前在 STM-KG 数据集上的实验相比，该方法在 Aminer数据集上展现出更强的推荐性能，尤其是在 MAP@50（30.6）和 nDCG@50（34.7）方面。然而，若直接用作推荐系统或仅用于轻量增强，重排序器主要针对固定候选池优化文本级语义相似度，因此容易受到论文数据集中常见的长尾术语和别名现象影响，也无法有效利用图约束。因此，大语言模型的表现往往不及强大的基于图的方法或图文混合语义方法（例如我们的方法）。近期有研究表明，融入结构交互知识可以达到甚至超越这些性能（C. Li、Deng、Hu、Kan 和 Li，2025年）。因此，这或许是论文推荐领域值得进一步研究的方向。

第三，作为一种通用型图神经网络，SMIGNN能够捕捉多种意图和高阶邻域，整体上优于仅依赖文本语义（如基线1）或未能充分利用图结构（如基线2）的基线方法。然而，在学术论文推荐中，它缺乏显式的关系蒸馏和领域语义建模，例如作者之间的合作关系以及论文之间的引文关系。因此，其表现落后于专为学术网络设计的三图蒸馏模型 TIDRec 和 LLM语义增强方法。

TIDRec 显式整合多种关系图并进行蒸馏校正，从而在论文上下文中有效对齐论文内容、引文结构和学者兴趣。在域内论文推荐中，其表现显著优于通用型图神经网络（如 SMIGNN）。然而，由于它的优势主要依赖于训练领域内关系分布的一致性，因此在跨域时优势会减弱。因此，尽管该模型在跨域论文推荐中的表现优于 SMIGNN，但它与大语言模型增强方法和我们的方法之间的差距却逐渐扩大。

基线2通过利用论文间的引文信息和细粒度知识语义，增强了论文内容的语义表示。与仅依赖知识句子语义表示的基线1相比，这提高了论文推荐的准确率。

从跨学科论文推荐的表现来看，首先，我们发现，应用于跨域推荐时，所有六种方法在推荐列表中的性能都会下降，且top_k 值相同。不同模型的性能下降程度各异，并且随着推荐论文数量增加，性能下降幅度趋于减小。这一趋势在平均准确率均值和归一化折损累计增益得分中均十分明显，尤其是在较小的 K 值下。例如，基线1的 nDCG@20 从 17.6（域内）降至 15.2（跨域），而我们的方法从 28.1 降至 26.9，表明其排序质量更加稳定。这一趋势说明，域内推荐通常比跨域推荐更准确，这可能是因为不同学科的句子、科学概念和实体词之间存在语义差异。然而，当前K个推荐

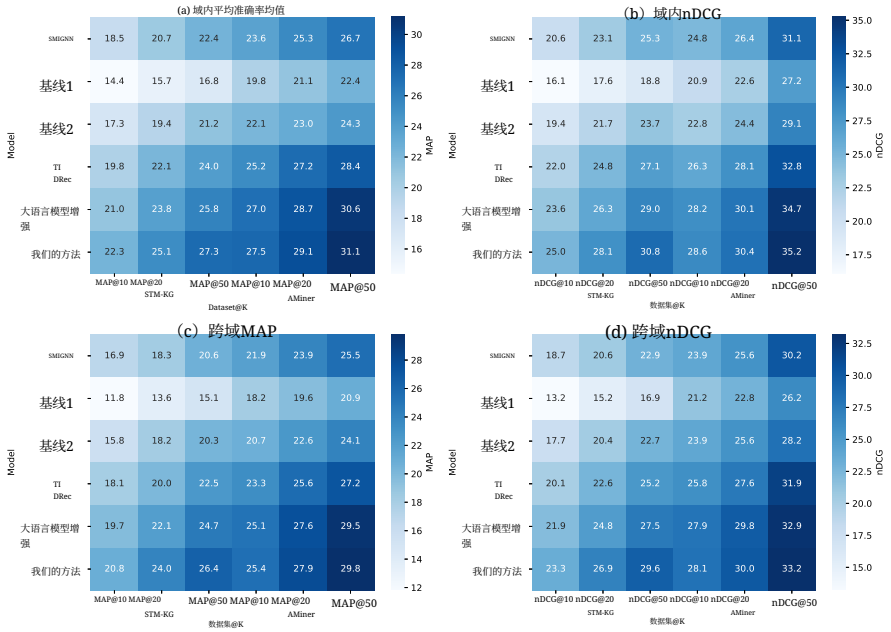

图 6 基线模型和我们的方法的性能。

列表扩大时，覆盖率的提升有助于抵消这些语义差异，从而在一定程度上恢复性能。

其次，在六种方法中，基线1的性能下降幅度最大，而我们提出的方法在跨领域场景中的退化程度最小。这表明，不同文档信息类型的语义嵌入受学科差异的影响并不相同。尽管基线2和我们的方法都依赖结构化概念或实体嵌入，而这些嵌入可能因领域而异，但引文信息的引入减轻了语义分歧带来的负面影响。因此，与未利用此类结构信号的基线1相比，它们的性能更为稳定。

（2）采用不同词嵌入模型的域内论文推荐性能

表15 和图7 展示了我们的方法、基线1和基线2在域内论文推荐方面的性能。这些方法均主要依赖文本语义嵌入，并采用了不同的词嵌入模型。表中列出了它们在 STM-KG 数据集上的两项评估指标（平均准确率均值和归一化折损累计增益）的结果。针对这两项评估指标，我们使用不同的 top_k 值（即 top_{10}， top_{20}和 top_{50}）来考察这些模型的性能。不同词嵌入模型对我们的方法性能有显著影响。使用领域专用词汇的预训练模型嵌入后，训练语料库规模对性能的影响小于引文信息的影响。通过分析这些表格和图表，可以得出以下见解和分析：

首先，就准确率而言，我们的方法对嵌入模型的选择较为敏感，其中$GPT\text{-}3.5$始终表现最佳，凸显了大规模语言模型的优势。将$FastText$ 升级为$SciB\text{-}ERT$ 可显著提升性能，而从$SciBERT$ 升级到$GPT\text{-}3.5$带来的提升则相当有限。这表明，领域专用词汇和训练语料库规模这两个变量能够提升嵌入质量。像$SciBERT$ 和$GPT\text{-}3.5$这样的预训练语言模型能够更准确地捕捉语义相似度，因此在基于嵌入的方法中使用时，可获得更好的推荐结果。

其次，使用$FastText$时，我们基于精细粒度知识实体嵌入的方法，其性能低于基线2（使用$SciBERT$生成科学概念嵌入）。有趣的是，使用$FastText$ 的基线2表现也优于基线1，即使基线1使用$SciBERT$ 或$GPT\text{-}3.5$也是如此。这些结果表明，在缺少引文信息的情况下，嵌入模型的选择对推荐性能的影响大于被嵌入的文档内容类型。然而，一旦纳入引文信息，嵌入模型选择的相对影响便会减弱，因为引文结构提供了互补的文档间关联信号，从而增强语义相似度匹配。

第三，在排序方面，$GPT\text{-}3.5$ 在 $nDCG$中也始终领先，得分为 $nDCG@10 = 25.0$、 $nDCG@20 = 28.1$和 $nDCG@50 = 30.8$，比SciBERT高出1.1至2.5分。GPT-3.5表现更优，是因为它能够表征从大规模跨领域语料库中学习到的复杂语义关系，从而实现更好的高排名论文匹配。这些结果证实，语义嵌入模型的质量会显著影响检索准确率和排序效果。

（3）使用不同维度的文档信息嵌入进行域内论文推荐的性能

图8 展示了我们的方法使用不同类型的文档信息嵌入时的域内论文推荐性能，包括在 STM-KG 数据集上使用两种评估指标（平均准确率均值和归一化折损累计增益）得到的结果。对于这两种评估指标，我们使用了具有不同 top_k 取值（即 top_{10}、top_{20}和 top_{50}）的推荐列表来探究这些模型的性能。我们发现，不同类型的知识实体以及引文信息对我们的方法的推荐性能有不同程度的影响。对这些表格和图表的分析得出了以下见解和分析：

首先，从我们的方法中依次移除从论文标题和摘要中提取的细粒度知识实体（即任务、方法和材料/指标实体）的语义嵌入后，推荐性能逐渐下降。单独考察各类实体时，它们对推荐性能的贡献依次为：任务 > 方法 >材料/指标。然而，与仅使用标题和摘要的语义嵌入进行推荐的性能相比，整体性能下降幅度较小。这表明，与标题和摘要相比，孤立的知识实体嵌入无法为有效推荐提供足够的上下文，且



图 7 我们的方法采用不同词嵌入模型时在 STM-KG 上的域内论文推荐性能比较。

研究人员优先关注与其研究任务相关的论文，而对材料和指标相关性相对较低的论文也会予以关注。

其次，请参见表 14，将标题、摘要和细粒度知识实体结合起来生成语义嵌入，可以提高推荐准确率。与仅基于标题和摘要的嵌入，以及使用摘要中的背景和目标陈述（即基线1）的嵌入相比，这种方法效果更佳。尽管仅使用完整摘要嵌入凭借其丰富的上下文也能取得相对出色的性能，但进一步嵌入更具针对性且语义更密集的研究目标陈述，能够提升推荐结果的相关性。值得注意的是，纳入实体嵌入后，研究意图和方法论的表示会更加细粒度，从而增强论文之间的语义对齐。

第三，将引文信息纳入文档嵌入可提升我们的方法的推荐性能。以表 14为例，具体而言，基于 SPECTER 的模型——该模型整合了摘要和引文图——表现优于基线1。在此基础上，基线2通过结合 *SPECTER* 嵌入与科学概念向量进一步提升了性能。我们提出的方法将 *SPECTER* 与结构化的细粒度知识实体相整合，取得了最高性能，相比基线2提高了5.0–6.1个百分点。这表明，基于引文的结构通过捕捉潜在主题关系，对语义相似度形成补充；同时，实体的覆盖率和分类方式也会影响嵌入质量。与缺乏上下文、目的和任务相关实体覆盖，且过程类型与方法类型之间存在重叠的基线2不同，我们的分类涵盖了关键研究维度——任务（研究什么）、方法（如何研究）

、材料，以及指标（工具和评估）——从而对论文的核心内容提供了更全面的表示。

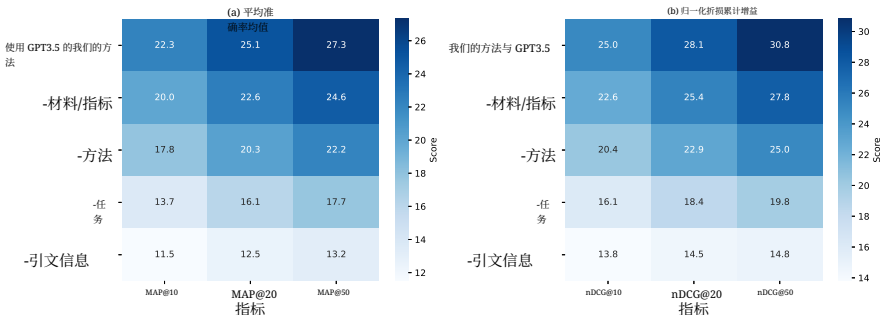

图8 我们的方法采用不同维度的文档信息嵌入时，域内论文推荐性能的对比。

我们还使用 nDCG@10、nDCG@20 和 nDCG@50 评估排序质量，如图所示。8。结果表明，摘要 + 引文 + 知识实体嵌入配置在所有 nDCG 指标中得分最高：nDCG@10= 25.0、nDCG@20 = 28.1 和 nDCG@50 =30.8，分别比基线2（摘要 +引文）高出 5.6、6.4 和 7.1 分。这表明，所提出的方法不仅提升了整体相关性，还将最相关的论文排到了前列，更符合用户在学术检索和引文行为中的预期。

为进一步区分细粒度内容维度的贡献，我们在固定编码器的同时评估了三种实体类型设置：仅任务（T ）、任务+方法（T+M ），以及任务+方法+材料/指标（T+M+D）。如表16所示，MAP@K 和 nDCG@K 的性能从 T → T+M → T+M+D 单调提升（$K \in [10, 20, 50]$）。这表明，实体维度越丰富，所有 K 值下的 MAP 和 nDCG 就越呈单调增长；这也说明，与过程和资源相关的语义（例如方法/材料）对任务相似度具有互补作用。这些结果支持了我们的设计目标：对齐学者域内信息需求的多维语义建模。任务实体锚定主题意图；方法和材料/指标则引入了过程层面和资源层级的区别，这些区别往往对学术检索中的论文相似度起决定性作用。纳入这些信息能够产生更具区分性的表示，并提升前K项排序质量。

结合上述消融实验，我们发现，基于语义向量相似度的论文推荐性能受三个主要因素影响：

- 嵌入内容反映论文核心研究维度的程度，以及
- 语义嵌入模型的能力，以及
- 引文信息的纳入。

其中，各类细粒度实体的总贡献对论文推荐性能提升的影响最大，其次是嵌入模型的选择，最后是引文信息嵌入。

（4）采用向量相似度组合与权重的论文推荐模型性能

图 9 展示了我们的方法、基线1和基线2在跨域和域内论文推荐中的性能。这些方法主要依赖文本语义嵌入，并结合了向量相似度组合与加权策略，其中包括两项评估指标（平均准确率均值和归一化折损累计增益）的结果。对于这两项评估指标，使用具有不同 top_k 值（即 top_{10}、 top_{20}和top_{50}）的推荐列表来探究这些模型的性能。在我们的方法结合相似度加权向量组合后，推荐性能评估指标均有所下降，且域内推荐中的下降幅度更高。对这些表格和图表进行分析后，可以得到以下见解和分析：

首先，在应用相似度加权向量组合后，所有方法的MAP@N得分都显著下降，我们的方法（跨域）的 $nDCG$@20从26.9降至15.1。尽管有所下降，我们的方法仍然优于基线方法，其nDCG@$50=$ 16.6仍高于其他方法。这表明，推荐论文列表与查询论文参考文献列表之间的重叠程度有所降低。从另一个角度来看，这意味着推荐结果开始偏离已知引文，从而可能为研究任务、方法或解决方案提供独特的见解。其基本逻辑是，当某一实体类型（例如任务）具有很高的相似度，而其他实体类型的相似度较低时，加权会降低整体相似度值。因此，排名靠前的推荐论文具有非对称相关性：突出某一组成部分（例如任务），但在其他方面有所偏离，从而为同一研究任务引入多样性和独特视角。

其次，在加权设置下，我们提出的方法始终取得最高的 MAP@N 得分，且与基线1和基线2相比，性能下降幅度更小。这表明，我们的方法与学者的实际引文行为更加吻合，更倾向于推荐在多个维度上具有较高语义相似度的论文。然而，这也意味着我们的方法在推荐多样化论文方面相对效果较差，因为它更倾向于相似性而非差异性。

这一观察引出了一个重要的研究问题：推荐性能应优先考虑准确率（以引用重叠衡量），还是多样性（体现为与被引作品的差异）？ 此外，这里观察到的趋势与先前实验的结果一致：

- 域内推荐性能仍高于跨域推荐；
- MAP@N得分随推荐论文数量的增加而提高。

我们严格按照 §3.4 中的定义评估向量运算，不引入外部聚合方案。具体而言，我们比较：

- 单向量：通过拼接仅使用 $p_g = [c_p^t, c_p^m, c_p^d, s_p]$（所有实体）。
- 等权凸组合：公式 (7) 中（p_g， p_t， p_m， p_d）采用相等权重

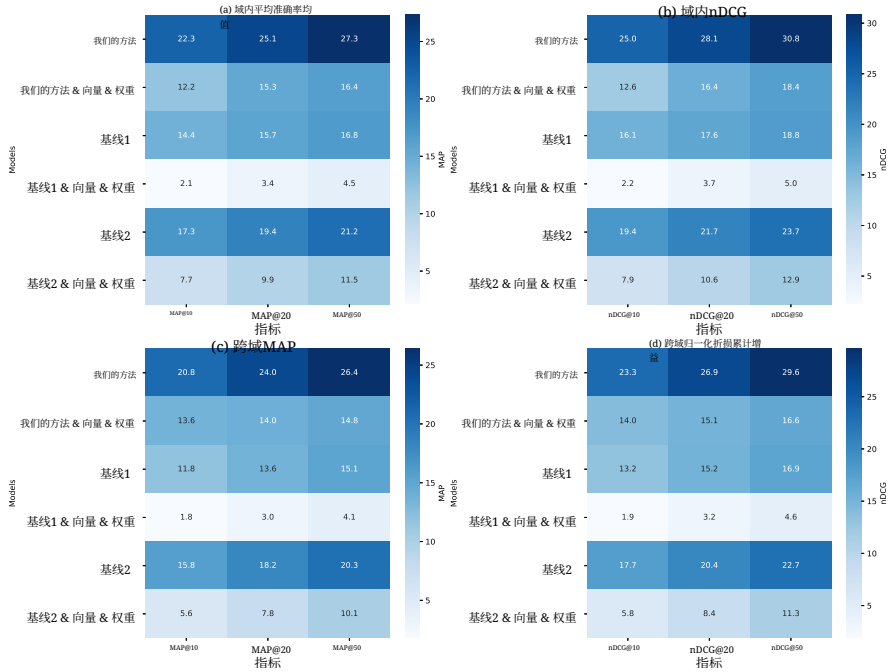

图9 采用向量相似度组合和权重的论文推荐模型性能比较。

• 精度调优权重：在开发集划分上学习公式 (7) 中的 w（MAP@50目标，$\lambda = 0$），然后在测试时固定。
• 任务子集重排序：在 $CP_{\text{任务}}$内，学习式（9）中的 $\boldsymbol{\alpha}$，以聚合 $\{s_j\}_{j=1}^{4}$（相同/不同方法；相同/不同材料/指标），并应用于$\mathbf{w}$之上。

表中结果 17 表明，在 STM-KG（域内）上，采用 p_g 的单向量表示取得了最优的 平均准确率均值/归一化折损累计增益 性能。随着聚合策略从均匀加权（$w = \frac{1}{4}\mathbf{1}$）、学习得到的权重（式（7），$\lambda = 0$）过渡到学习得到的权重 $+\,\boldsymbol{\alpha}$ 重排序（式（9））， 列表内多样性 和 覆盖率 均呈现持续提升，而准确率相应下降——这体现了典型的精度–多样性权衡。总体而言， p_g 非常适合面向精度的检索场景；在较大的K 值（例如@50）下，均匀加权和重排序配置能够带来更高的结果多样性和覆盖率，因此对探索性检索场景（包括方法扩展和跨视角比较）更为有利。

（5）与基于大语言模型的推荐系统的互补性

大语言模型（大语言模型）领域近期最具突破性的进展，例如$GPT\text{-}3.5$ 和 $GPT\text{-}4$，使模型能够基于论文标题、摘要或检索词等非结构化输入，高效完成学术内容的创建、检索与推荐。尽管大语言模型在一般推荐或问答任务中表现出色，但将其用于论文推荐时仍存在一些问题

，涉及透明性、真实性和控制能力。我们提出的方法在以下方面对大语言模型进行了补充：

• 结构化语义表示：我们的方法并非仅依赖端到端文本生成，而是采用细粒度科学实体抽取（任务、方法、指标），并构建能够保留语义结构的向量表示。这使可控的推荐策略（例如“相同任务，新方法”）和可解释检索成为可能。

• 可解释性与定制化：基于实体的向量拼接使用户能够将推荐结果追溯到具体的语义组成部分（例如任务相似度、方法差异）。相比之下，大语言模型通常像黑箱一样运作，除非投入大量快速工程设计，否则无法支持这种细粒度定制。

• 效率与部署能力：我们的方法所需计算量更少，无需 API 访问或快速调优，并支持批量计算和离线索引，因此更适合大规模或机构部署。

尽管如此，我们认为大语言模型在未来工作中能够增强我们的过程。例如：

• 大语言模型可以协助实现更准确的实体消歧和关系类型归纳；
• 大语言模型可以生成用户意图嵌入，或辅助进行零样本查询扩展；
• 大语言模型生成的反馈可用于微调多面向向量拼接中的权重策略。

总之，我们的方法构成了一个结构化且可解释的基础，可从与大语言模型功能的集成中受益，而非与之竞争的替代方案。在未来工作中，我们将研究一种融合知识嵌入表示与生成技术的混合推荐框架。

（6）STM-KG 和域内条件下的纯文本对比

表 18 显示，***BM25***（纯文本）< ***SciBERT***（纯文本）<我们的方法（纯文本编码器） < 我们的方法（完整）。我们的方法（完整）显著优于强大的纯文本基线 $SciBERT$，在 $MAP@50/nDCG@50$ 和实体感知相关性（$p < 0.01$ 三项指标均如此），这表明仅依赖文本表示不足以实现最佳性能。保留我们的流程并将其替换为纯文本编码器后，性能仍优于$SciBERT$，说明细粒度实体信号和异构结构本身有助于提升性能。进一步的消融分析表明，移除引文侧结构特征后，性能较我们的方法（完整）略有下降，但仍优于***SciBERT***和$BM25$，且实体感知指标的排序与各项指标的排序一致。这些结果表明，我们的方法优势并非源于潜在的引文泄漏，而是来自实体、结构和文本的协同效应；引文信息有积极贡献，但并非唯一因素。

（7）针对流行度偏差的分层与均衡评估

表19 显示，绝对得分会随着引文数量增加而提高——这符合预期——但 Our（完整）相较于$SciBERT$ 的相对改进在所有层级中都持续存在，并且

在引文数量最高的三分位组中最小，这表明存在温和但并非决定性的流行度效应。平衡子集上的结果与分层分析一致，表明我们的结论并不取决于引文数量较高的项目。

4.4.3 学术论文推荐的多样性感知评估

我们评估了一种多样性感知设置：从经过准确率调优的得分 $\lambda = 0$ 和共享候选池出发，应用重排序目标—— $J_{\text{div}} =$ 平均准确率均值@50+λ·列表内多样性@50，并结合 $\lambda > 0$ ，以明确提升列表内多样性（列表内多样性）和覆盖率。我们引入了两个衡量推荐多样性的指标，对比并分析了基线1、基线2以及本文提出的推荐方法在是否使用向量相似度组合及权重时，推荐准确率和多样性得分的变化，并进一步探讨了推荐准确率与多样性之间的权衡及其各自的主要应用场景。两个指标均基于文档嵌入之间的余弦距离计算。对于 ILD，我们使用任务 + 方法 + 材料/指标向量拼接来表示每篇论文。

（1）科学论文推荐中的多样性定义

在学术论文推荐的语境下，多样性指的是：在准确匹配学者查询的推荐论文列表中，各篇论文彼此之间的差异程度，以及它们在主题等多个维度上的覆盖范围，尤其是在方法论、材料和指标等方面。我们将多样性操作性地定义为：推荐与查询论文具有相同研究任务、但在方法体系或材料/指标方面存在显著差异的论文，从而为广为人知的主题提供不同的观点或独特的方法。

这种方法适用于早期学科探索、跨学科文献发现或方法扩展等场景。在这些场景中，目标并非只是检索已知内容（即平均准确率均值 和归一化折损累计增益 的重合度较高），而是拓展学术视野。

（2）实验结果与分析

我们进一步研究向量相似度与权重组合如何影响推荐准确率和多样性。在使用 *GPT-3.5*嵌入知识实体的设置下，我们比较基线1、基线2（主要使用文本嵌入并结合带有学习权重的向量）以及我们提出的方法在域内论文推荐场景中的表现，旨在揭示这些因素影响准确率和多样性的规律。

如图 10所示，采用向量相似度组合和加权会导致三种方法的推荐准确率均有所下降，这与使用向量相似度组合和权重的论文推荐模型性能比较中的观察结果一致。与此同时，我们观察到与多样性相关的指标均有所改善。对于我们提出的方法，ILD 得分从 32.4 提高到 47.2，且 $Coverage$ 指标从 37.2 提高到 50.1。这些提升幅度大于另外两种基线。与单向量和未加权多向量基线相比，多样性感知设置显著

提升了列表内多样性和覆盖率。平均准确率均值/归一化折损累计增益略有下降，反映了显式优化多样性组件所带来的预期权衡。

结合我们的方法与不同词嵌入模型进行域内论文推荐性能比较，以及跨域和域内论文推荐性能比较的结果，我们将这一优势归因于我们的方法借助 $GPT\text{-}3.5$ 嵌入实现了更优的知识实体语义表示。这使模型能够更好地区分和匹配具有不同或相似任务、方法和材料/指标的论文，从而在准确率与多样性之间取得更好的平衡。

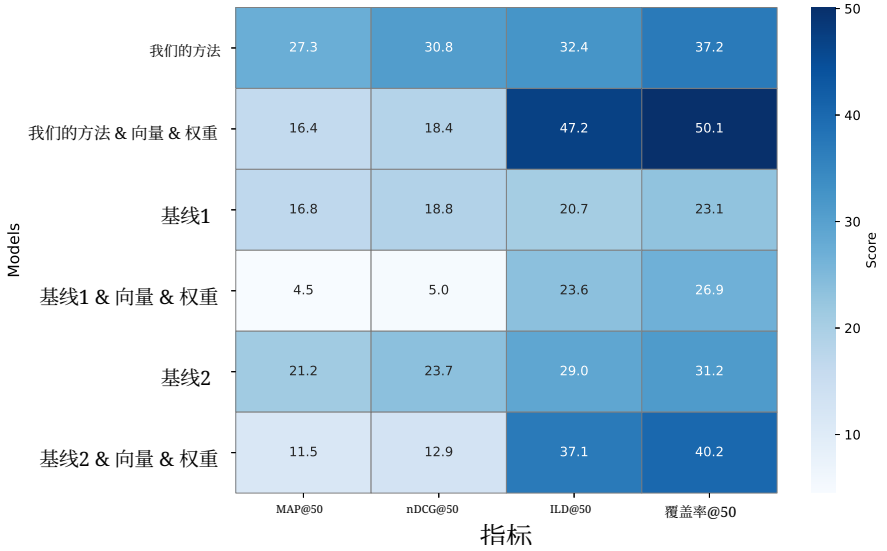

图10 采用向量相似度组合及权重、diversity-aware 重排序的论文推荐模型的准确率与多样性比较（$\lambda > 0$）。

结果表明，我们的加权组合技术提升了 ILD和 $Coverage$，使推荐列表更加多样化，但代价是准确率略有下降（MAP）。这体现了准确率与多样性之间的典型权衡。

（3）应用场景：准确率与多样性

我们还强调了使推荐策略与用户场景相匹配的重要性。对于预测引文或维持既定研究方向，建议采用较高的 MAP 和 $nDCG$ 值。对于概念生成、跨学科研究或早期评估，多样性感知排名可能更有益。

本研究表明，采用多样性感知的向量加权能够生成更多样化的论文推荐结果，从而更好地满足探索性研究目标。未来研究可探索自适应重排序算法，根据用户偏好或任务类别权衡准确率与新颖性。

4.4.4 我们的方法在学术论文推荐上的实验总结

本文验证了，整合多种类型的文档信息——例如标题、摘要、引文图和细粒度知识实体——能够持续提升基线论文推荐方法的性能。

融入大语言模型（尤其是*GPT-3.5*）生成的语义嵌入，可进一步提升推荐准确率，尤其是在标题、摘要和知识实体等文档组成部分上应用时。

我们的分析表明，尽管*SPECTER* 在域内知识映射方面表现略好，但从跨学科知识图谱中获得的知识实体嵌入，对跨域推荐性能的提升贡献更大。

此外，所提出的*SPECTER* + 知识实体嵌入 + 向量加权方法在加权设置下取得了最佳*MAP@K* 得分；尤其在*MAP@50*上，其表现始终优于所有其他向量加权基线。

在准确率调优设置（$\lambda = 0$）下，所提出的*SPECTER* +知识实体嵌入 + 向量加权方法取得了最佳或在统计学上不劣于单向量和无权重多向量基线的$MAP@K/nDCG@K$表现。相比之下，图10 报告了多样性感知重排序（$\lambda > 0$的结果，其中显式提升多样性显著改善了列表内多样性/覆盖率，同时使$MAP/nDCG$小幅、可控地下降。

我们认为，这一改进得益于加权相似度组合的灵活性。它使推荐系统不仅能够呈现与被引用的参考文献相匹配的论文，还能提供新颖且多样的观点。在准确率与新颖性之间取得平衡，使系统能够更好地满足学者各异的信息需求。

总之，我们的实验表明，将细粒度知识图谱作为辅助语义来源加以整合，能够显著提升论文推荐性能。无论采用何种嵌入策略，这一提升均适用，包括使用完整的摘要表示，或使用任务、方法和材料/指标实体的结构化嵌入。

5 结论与未来工作

本文提出了一种新的学术论文推荐方法，旨在解决现有方法难以契合学者的特定文献需求、且推荐结果广度不足的问题。本研究通过从论文标题和摘要中提取细粒度知识实体，并捕捉实体之间的关系及引文链接，构建了细粒度科学知识图谱。为此，作者结合*SciBERT*、双向长短期记忆网络和级联模型，提取细粒度实体并映射其关系。研究使用Transform模型，将摘要和引文数据整合为有意义的嵌入，同时由*GPT-3.5* 支持知识实体的语义嵌入。推荐过程依赖于

多维向量运算，通过知识实体关系表示论文。系统运用向量过滤和排序技术，推荐契合学者特定需求的论文。实验结果表明，命中准确率和多样性均有所提升，推荐结果也更相关、更多样。这种方法不仅契合学者的文献需求，还拓宽了他们的研究视角。未来的工作包括进一步细化实体之间的关系，以捕捉复杂关联，并改进图表示学习。本研究旨在平衡对推荐结果的客观评估与学者的主观意见，从而提升推荐结果的准确率和多样性。

致谢。本研究得到中国国家社会科学基金青年项目“基于学术论文全文识别算法实体的语义关系与演化路径”（项目编号：24CTQ027）的资助。本文在JCDL2024会议论文的基础上，进一步拓展了相关研究现状、研究方法和实验讨论。

声明

作者声明，就本文的研究、作者身份和/或发表而言，不存在任何潜在利益冲突。

参考文献

Adomavicius, G. 与 Tuzhilin, A.（2015）。情境感知推荐系统。F. Ricci、L. Rokach 与 B. Shapira（编），推荐系统手册（第191–226页）。Springer US。

Bawden, D.（2007）。转向：在上下文中整合信息寻求与信息检索。《文献学杂志》，*63*（2），290–292。

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.（2017年）。通过子词信息丰富词向量。《计算语言学协会会刊》，*5*，135–146，https://doi.org/10.1162/tacl a 00051

Brack, A., D’Souza, J., Hoppe, A., Auer, S., Ewerth, R. (2020). 从研究文章中提取领域无关的科学概念。欧洲信息检索会议 (pp. 251–266). Brack, A.,Hoppe, A., Ewerth, R. (2021). 通过知识图谱为研究论文提供引文推荐。数字图书馆理论与实践国际会议 (pp. 165–174).

Brack, A., M¨uller, D.U., Hoppe, A., Ewerth, R. (2021). 多领域研究论文中的共指消解。 欧洲信息检索会议 (pp. 79–97). Cao, X., Shi, Y., Yu, H., Wang, J., Wang, X., Yan, Z., Chen, Z. (2021). Dekr：用于机器学习方法推荐的描述增强型知识图谱。 第*44*届*ACM SIGIR*信息检索研究与发展国际会议论文集 (pp. 203–212). Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S. (2020). SPECTER：利用引文信息增强的Transformer进行文档级表示学习。第*58*届计算语言学协会年会论文集 (pp.2270–2282).Dess´ı, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E. (2022). CS-KG：一个面向计算机科学研究实体与论断的大规模知识图谱。 国际语义网会议 (pp. 678–696). Du, N., Guo, J., Wu, C.Q., Hou, A., Zhao, Z., Gan, D. (2020). 基于异构信息网络的学术论文推荐。 *2020*年*IEEE/ACS*第*17*届计算机系统与应用国际会议（*AICCSA*） (pp. 1–6). Eberts, M., & Ulges, A. (2020). 采用Transformer预训练的基于跨度的实体与关系联合抽取。 *ECAI 2020* (pp. 2006–2013). IOS Press. Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D., Pera, M.S. (2018). 所有酷孩子都如何融入其中？：推荐系统评估与效果中的受欢迎程度和人口统计偏差。 公平性、问责制与透明度会议 (pp. 172–186). Gupta, S., & Varma, V. (2017). 利用文本和图的分布式表示进行科学文章推荐。 第*26*届万维网国际会议伴随会议论文集 (pp. 1267–1268). Gusenbauer, M., & Haddaway, N.R. (2020). 哪些学术检索系统适用于系统综述或元分析？评估Google Scholar、PubMed及其他26种资源的检索质量。 *Research Synthesis Methods*，*11*(2), 181–217，https://doi.org/10.1002/jrsm.1378Han, J., Song, Y., Zhao, W.X., Shi, S., Zhang, H. (2018). HyperDoc2Vec：超文本文档的分布式表示。 第 *56* 届计算语言学协会年会论文集（第 *1* 卷：长篇论文） (p. 2384)。

Hao, L., Liu, S., Pan, L.（2021）。基于作者—论文兴趣和图结构的论文推荐。*2021*年*IEEE*第*24*届计算机支持的协同工作与设计国际会议（*CSCWD*）（第256–261页）。Hou, Y., Jochim, C., Gleize, M., Bonin, F., Ganguly, D.（2019）。识别科学排行榜构建所需的任务、数据集、评估指标和数值评分。计算语言学协会第*57*届年会论文集（第5203–5213页）。Hou, Y., Jochim, C., Gleize, M., Bonin, F., Ganguly, D.（2021）。Tdmsci：用于科学文献中任务、数据集和指标实体标注的专用语料库。计算语言学协会欧洲分会第*16*届会议论文集：主卷（第707–714页）。Huang, S., Wan, X., Tang, X.（2013）。Amrec：一种智能方法推荐系统。第*17*届*AAAI*人工智能领域最新进展会议论文集（第47–49页）。Hwang, S.-Y., Wei, C.-P., Liao, Y.-F.（2010年）。合著网络与学术文献推荐。《电子商务研究与应用》，*9*（4），323–334，Im`ene, S., Badia, K., Nadir, M.（2022）。基于知识图谱的相关实体推荐方法。《智慧城市智能能源效率的人工智能与启发式方法：案例研究：蒂帕萨，阿尔及利亚》（第488–496页）。Kanwal, T., & Amjad, T.（2024）。基于引文网络多种特征的研究论文推荐系统。《科学计量学》，*129*（9），5493–5531，https://doi.org/10.1007/s11192-024-04567-8Kreutz, C.K., & Schenkel, R.（2022）。科学论文推荐系统：近期出版物的文献综述。《国际数字图书馆期刊》，*23*（4），335–369，https://doi.org/10.1007/s00799-022-00316-0Li, C., Deng, Y., Hu, H., Kan, M.-Y., Li, H.（2025）。Chatcrs：将外部知识和目标引导融入基于大语言模型的对话式推荐系统。计算语言学协会研究成果：*NAACL 2025*（第295–312页）。Li, Q., & Ji, H.（2014）。增量联合抽取实体提及和关系。计算语言学协会第*52*届年会论文集（第*1*卷：长篇论文）（第402–412页）。

Liu, J., Mitsui, M., Belkin, N.J., Shah, C.（2019年）。任务、信息寻求意图与用户行为：迈向对网络搜索的多级理解。*2019*年人类信息交互与检索会议论文集（第123–132页）。Liu, K., Zhang, Y., Pan, R., Gao, T., Wang, H.（2025）。大语言模型增强的大规模引文网络学术文献推荐。*arXiv*预印本 *arXiv:2503.01189*，Liu, Q., Du, Q., Hong, Y., Fan, W.（2021）。开放式创新平台中的创意推荐：一种设计科学方法。中国互联网经济研究中心（*CCIE*）研究论文，Luan, Y.（2018年）。从科学文献中进行信息抽取以推荐方法。*arXiv*预印本 *arXiv:1901.00401*，Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.（2018）。面向科学知识图谱构建的实体、关系和共指多任务识别。*2018*年自然语言处理中的经验方法会议论文集（第3219–3232页）。Ma, S., Zhang, C., Zhang, H., Gao, Z.（2025）。基于用户查询论证分区的引文推荐。《信息计量学期刊》，*19*（1），101607, https://doi.org/10.1016/j.joi.2024.101607Ma, S., Zhang, H., Zhang, C., Liu, X.（2021）。结合时间偏好的时序引文推荐。《科学计量学》，*126*，2991–3010, https://doi.org/10.1007/s11192-021-03912-0Mohamed, D., El-Kilany, A., Mokhtar, H.M.（2021）。使用基于概念的表示进行学术论文推荐。智能系统与应用：*2020*年智能系统会议（*IntelliSys*）论文集，第*2*卷（第733–744页）。Mondal, I., Hou, Y., Jochim, C.（2021）。自然语言处理知识图谱的端到端构建。计算语言学协会研究成果：*ACL-IJCNLP 2021*（第1885–1895页）。Murali, M.V., Vishnu, T., Victor, N.（2019）。一种基于协同过滤的推荐系统，用于推荐任何研究领域的新趋势。*2019*年第五届先进计算与通信系统国际会议（*ICACCS*）（第550–553页）。

Ni, C.-C., Sum Liu, K., Torzec, N. (2020年). 使用维基百科和 Yahoo!知识图谱进行实体推荐的分层图嵌入。 *2020*年网络会议伴随论文集 (第 811–818 页). Niu, Y., Xing, X., Jia, Z., Xin, M., Xing, J. (2024年). SMIGNN：基于图神经网络、采用多意图知识蒸馏的社交推荐。《超级计算机杂志》， *80*(5), , Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., Rehm, G. (2022年). 基于引用嵌入的科学文档表示邻域对比学习。 *2022*年自然语言处理中的经验方法会议论文集 (第 11670–11688 页). Radford, A. (2018年). 通过生成式预训练提升语言理解能力。计算语言学协会第 *56* 届年会论文集（第 *1* 卷：长篇论文） (第 175–186 页). Sakib, N., Ahmad, R.B., Haruna, K. (2020年). 一种利用引文上下文进行科学论文推荐的协作方法。 *IEEE Access*， *8*， 51246-51255, https://doi.org/10.1109/ACCESS.2020.2980904Santosa, N.C., Liu, X., Han, H., Miyazaki, J. (2024年). S3par：基于章节的科学论文序列推荐，用于论文写作辅助。《知识型系统》， *303*， 112437, Sarwar, T.B., Noor, N.M., Miah, M.S.U., Rashid, M., Al Farid, F., Husen, M.N. (2021年). 研究文章推荐：一种结合无监督关键词短语提取与词汇相似度计算的多级时间顺序学习方法。 *IEEE Access*， *9*， 160797–160811, Shang, Y.-M., Huang, H., Mao, X. (2022年). OneRel：一步中通过单个模块联合抽取实体与关系。 *AAAI*人工智能会议论文集 (第 36 卷，第 11285–11293 页). Singh, A., D'Arcy, M., Cohan, A., Downey, D., Feldman, S. (2023年). SciRepEval：面向科学文档表示的多格式基准。 *2023*年自然语言处理中的经验方法会议（ *EMNLP 2023*）(第 5548–5566 页). Singh, S., & Singh, M. (2024年). COSAEMB：面向科学文章的对比式章节感知方面嵌入。 第四届学术文档处理研讨会论文集（*SDP 2024*）(第 283–292 页)。

Son, J., & Kim, S.B. (2018年). 利用多层同步引文网络的学术论文推荐系统。 决策支持系统， *105*， 24–33， https://doi.org/10.1016/j.dss.2017.11.005Takahashi, T., & Katsurai, M. (2022年). Solutiontailor：基于细粒度摘要分析的科学论文推荐。 欧洲信息检索会议 (第 316–320 页). Talley, E.M., Newman, D., Mimno, D., Herr, B.W., Wallach, H.M., Burns, G.A., . . . McCallum, A. (2011). 利用机器学习类别和图形聚类构建的美国国立卫生研究院资助项目数据库。 《自然·方法》， *8*(6), 443–444, Tjong Kim Sang, E.F., & De Meulder, F. (2003). CoNLL-2003 共享任务简介：语言无关的命名实体识别。 *HLT-NAACL 2003* 第七届自然语言学习会议论文集，第 *4* 卷 (第 142–147 页). Viswanathan, V., Gao, L., Wu, T., Liu, P., Neubig, G. (2023年). Datafinder：根据自然语言描述推荐科学数据集。 计算语言学协会第 *61* 届年会。 Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y. (2020年). 图神经网络综合综述。 《*IEEE* 神经网络与学习系统汇刊》， *32*(1), 4–24, https://doi.org/10.1109/TNNLS.2020.2978386Xi, H., Zhang, C., Zhang, H. (2024年). 利用细粒度知识实体和多维文档嵌入增强学术论文推荐。第 *24* 届 *ACM/IEEE* 数字图书馆联合会议论文集 (第 1–3 页). Xiao, X., Liu, Y., Huang, J., Jin, D., Shen, Z., Zhang, C. (2025年). TIDRec：一种用于论文推荐的新型三图交互式蒸馏方法：X. xiao 等。 《知识与信息系统》， 1–33, Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J. (2017年). 深度集合。 神经信息处理系统进展， *30*， , Zaratiana, U., Holat, P., Tomeh, N., Charnois, T. (2022年). 用于科学命名实体识别的层次化 Transformer 模型。 *arXiv* 预印本 *arXiv:2203.14710*， ,

Zhang, H., Zhang, C., Wang, Y. (2024年)。揭示自然语言处理的技术发展：以科学实体为中心的视角。*InformationProcessing & Management*，*61*(1), 103574, https://doi.org/10.1016/j.ipm.2023.103574

Zhong, Z., & Chen, D. (2021年)。一种出奇简单的实体与关系抽取方法。*2021*年北美计算语言学协会分会会议：人类语言技术，*NAACL-HLT 2021*年 (第 50–61 页)。

Zhu, J., Patra, B.G., Yaseen, A. (2021年)。使用公共数据集的学术论文推荐系统。*AMIA* 转化科学峰会论文集，*2021*年，672，https://doi.org/10.1007/s13197-020-04656-5

附录 A：实体识别标注指南

A.1 概述

本标注任务旨在从科学论文的标题和摘要中识别七类细粒度知识实体，包括：任务、对象、问题、方法、过程、材料和指标。所有实体均采用标准的 BIO（开始-内部-外部）标注格式进行句子级标注。标注结果将用于构建细粒度知识图谱，并支持下游推荐任务。

A.2 实体类别

1. 任务

- **定义：** 研究的总体目标或任务，通常回答“要做什么”这一问题。
- **典型位置：** 标题中部、摘要开头或目的句。
- **关键词线索：**分类、预测、生成、检测、分析、建模、估计等。
- **示例：**
  - 我们提出了一种用于机器阅读[redacted]
  - 本文聚焦于产品评论的[redacted]
- **标注为：**B-TASK, I-TASK

2. 对象

- **定义：**在研究中被操作、处理或分析的具体对象或实体，限定了研究任务的范围，也就是研究“谁”。

- **典型位置：**紧接在任务词或方法词之后。
- **关键词线索：**文本、图像、基因、问题、文章、蛋白质、句子、分子等。
- **示例：**

  – 我们提出了一种新的模型，用于对生物医学文档进行机器阅读理解。– 该系统将客，

- **标注为：**B-OBJECT, I-OBJECT

3. 问题

- **定义：**研究中需要解决的核心问题、困难或痛点，强调“存在什么困难”。

•**典型位置：**摘要开头或研究动机的描述中。

- **关键词线索：**挑战、局限、瓶颈、问题、缺点、弊端等。
- **示例：**

  – 神经句法分析中的一个关键挑战是领域
  – 现有方法的抗噪性较低。

- **标注为：**B-PROBLEM, I-PROBLEM

4. 方法

- **定义：** 用于完成研究任务的方法、模型、算法、技术和架构。
- **典型位置：** 位于句子中部或结束位置，通常与 propose/use / apply/develop 等动词共现。
- **关键词线索：**模型、框架、架构、算法、方法、技术等。
- **示例：**

  – 我们设计了一种新
  – 我们的方法采用图卷积网

- **标注为：** B-METHOD, I-METHOD

5. 过程

•**定义：** 研究方法中改变或生成研究对象状态信息的行为。

- **典型位置：** 介绍研究方法时，或作为模型模拟步骤时。

•**关键词线索：**光合作用、蛋白质折叠、扩散、信号转导、发酵等。

- **示例：**

– 我们模拟使用图神[redacted].
– 我们的研究对热带森林[redacted]

- **标注为：**B-PROCESS, I-PROCESS

6. 材料

- **定义：**实验、评估、训练或数据分析中使用的外部资源，包括数据集、工具、软件和设备等。
- **典型位置：**实验描述中，句子末尾或括号内。
- **关键词线索：**数据集、语料库、系统、模拟器、应用程序接口、软件、设备、平台等。
- **示例：**

– 实验在 CIFAR-10 数据集上进行。–[redacted]
PyTorch 实现了模型。[redacted]

- **标注为：**B-MATERIAL, I-MATERIAL

7. 指标

- **定义：**用于评估模型性能或比较方法的指标。
- **典型位置：**实验结果描述或结束位置。
- **关键词线索：** $BLEU$， $ROUGE$， $F1$， $accuracy$， $precision$， $recall$， $RMSE$，等等。
- **示例：**

– 我们使用 $BLEU$ 和 $ROUGE$[redacted]评估[redacted]
– 在测试集[redacted]高准确率为 94.5%。

- **标注为：**B-METRIC, I-METRIC

A.3 边界判定规则

- 遇到复合实体（例如“图像分类任务”）时，应标注整个短语；嵌套实体：仅标注最外层实体，不进行嵌套：
- 术语被打断时（例如在方法名称中插入修饰语），可以跨越短暂间隔进行标注，例如：“一种新颖的基于卷积神经网络的方法” →整体标注为 METHOD；
- 以下类型不予标注：作者姓名、年份、谓语动词、模糊词语（例如，如果未指定技术，“方法”不予标注）：
- 形容词性代词（this、its、these、such）和限定词（the、a）不应包含在该跨度内。

附录 B：我们提出的论文推荐方法的伪代码。

注：
• 权重学习：采用单纯形投影的粗网格坐标上升；在测试集上固定；敏感性为 $\pm\{10\%, 20\%\}$。
• 复杂度：对 $|CP| = K$ 进行余弦相似度计算，向量维度为 d 时，复杂度为 $O(K \cdot d)$；细化增加$O(K \cdot d)$；拼接的复杂度为 $O(d)$。

表9 不同模型下细粒度知识实体识别的各类别结果。

模型	实体类型	精确率	召回率	F_1-得分
双向长短期记忆网络 + 条件随机场	Task	78.50	80.20	79.34
	方法	80.60	81.10	80.85
	材料	78.10	79.40	78.74
	指标	80.3	82.00	81.14
	Micro-average	*79.36*	*80.85*	*80.10*
SciBERT	Task	83.10	84.50	83.79
	方法	84.00	85.30	84.65
	材料	84.50	83.00	**83.74**
	指标	83.90	85.00	84.45
	Micro-average	*83.78*	*84.77*	*84.27*
SciBERT+条件随机场	Task	82.30	85.10	83.68
	方法	82.70	86.40	84.51
	材料	83.00	84.20	83.60
	指标	84.90	85.70	**85.30**
	Micro-average	*82.12*	*85.83*	*83.93*
SciBERT+双向长短期记忆网络+级联（所提出的方法）	Task	85.90	85.60	**85.75**
	方法	86.40	86.00	**86.20**
	材料	84.20	82.70	83.44
	指标	83.70	84.30	84.00
	Micro-average	*84.84*	*85.36*	***85.10***
9 折交叉验证（平均值）	Task	85.10	84.80	84.95
	方法	85.50	84.90	85.20
	材料	84.00	83.20	83.60
	指标	84.40	84.10	84.25
	Micro-average	*85.04*	*84.80*	*84.92*
GPT-3.5 零样本	Task	70.40	72.60	71.48
	方法	72.00	73.90	72.94
	材料	70.10	73.00	71.52
	指标	72.50	74.80	73.63
	Micro-average	*71.25*	*73.58*	*72.39*
GPT-3.5 少样本（3 个示例）	Task	76.10	78.90	77.47
	方法	77.80	79.90	78.84
	材料	75.50	79.20	77.31
	指标	77.30	79.30	78.29
	Micro-average	*76.92*	*79.34*	*78.11*

表10 SciBERT 与我们的方法的 9 折交叉验证结果。

Fold	SciBERT F_1	(SciBERT+双向长短期记忆网络+级联) F_1
第1折	84.49	84.79
第2折	84.23	85.00
第3折	83.81	85.05
折 4	84.31	84.42
折 5	84.56	84.81
折 6	84.20	85.19
折 7	84.50	84.62
折 8	84.37	85.31
折 9	84.73	85.08

表11 在*SciERC* 和*TDM* 数据集上的实体识别模型评估。

数据集	作者	模型	精确率	召回率	F_1-得分
SciERC	(Zhong& Chen,2021年)	PURE	-	-	68.90
	(Eberts & Ulges,2020年)	SpERT	70.87	69.79	70.33
	(Zaratiana, Holat, Tomeh , & Charnois , 2022年)	层次化 Transformer	67.99	74.11	**70.91**
	Our	SciBERT+双向长短期记忆网络+级联	66.95	71.49	69.14
TDM	（Hou、Jochim, Gleize, Bonin, & Ganguly, 2019年）	SCIIE	67.17	58.27	62.40
	（Zaratiana 等，2022年）	层次化 Transformer	65.56	70.21	67.81
	Our	SciBERT+双向长短期记忆网络+级联	68.84	70.73	**69.77**

表 12 移除每种关系类型对推荐指标的影响（STM-KG，域内； Δ 与完整图相比）。负值表示下降。3 次运行的均值±标准差；采用配对 t-检验判断显著性。

移除的关系类型	ΔMAP@50	ΔnDCG@50	ΔILD@50	Δ覆盖率@50
由……实现（任务–方法）	-1.6	-1.4	-1.2	-1.0
被……使用（材料–任务）	-1.0	-0.9	-3.8	-3.4
由……评估（任务–指标）	-0.6	-0.5	-1.5	-1.2
相关（类型内）	-0.4	-0.3	-0.8	-0.6

注：每行仅移除一种关系类型，同时保留其他类型。数据划分、候选池和评估均遵循 §§4.1–4.4，以确保可比性。

表 13 权重学习／重排序（不可训练）和向量维度设置。

Item	**设置**
向量维度	SPECTER 论文嵌入：**768**；GPT-3.5 实体嵌入：**1536**。拼接后的向量遵循定义 Concatenated vectors follow definitions $p_g = [c_t, c_m, c_d, s_p]$, $p_t = [c_t, s_p]$, $p_m = [c_m, s_p]$, $p_d = [c_d, s_p]$。
目标函数（针对准确率调优）	开发集划分上的 MAP@50，$\lambda = 0$；学习得到的权重在测试时排序中保持固定。
目标函数（多样性感知，可选）	$J_{\text{div}} = \text{MAP@50} + \lambda \cdot \text{ILD@50}$（仅用于第 4.4.3 节中的权衡分析）。
搜索空间与约束	单纯形：$w_i \geq 0$，$\sum_i w_i = 1$；$\alpha_j \geq 0$，$\sum_j \alpha_j = 1$。
搜索计划	粗网格：步长 0.05；坐标上升 + 细网格：步长 0.01；每一步投影回单纯形；在连续 5 次迭代未取得改进后提前停止（开发集 MAP@50）。
敏感性分析	对 ±10%和 ±20%进行单维扰动（重新归一化），使用 3 个随机种子，配对 t-测试集。
实现说明	余弦相似度；向量的 ℓ_2-范数；采用时间切分以避免泄漏。

Table 14 Performance comparison across baseline on two datasets, and the outcomes of the best performing model are shown in bold.

Dataset	Scope	Metric	Top_k	Baseline 1	Baseline 2	SMIGNN	TIDRec	LLM-enhanced	Our method
STM-KG	In-domain	MAP	10	14.4	17.3	18.5	19.8	21.0	22.3
			20	15.7	19.4	20.7	22.1	23.8	25.1
			50	16.8	21.2	22.4	24.0	25.8	27.3
		nDCG	10	16.1	19.4	20.6	22.0	23.6	25.0
			20	17.6	21.7	23.1	24.8	26.5	28.1
			50	18.8	23.7	25.3	27.1	29.0	30.8
	Cross-domain	MAP	10	11.8	15.8	16.9	18.1	19.7	20.8
			20	13.6	18.2	18.3	20.0	22.1	24.0
			50	15.1	20.3	20.6	22.5	24.7	26.4
		nDCG	10	13.2	17.7	18.7	20.1	21.9	23.3
			20	15.2	20.4	20.6	22.6	24.8	26.9
			50	16.9	22.7	22.9	25.2	27.5	29.6
Aminer v12	In-domain	MAP	10	19.8	22.1	23.6	25.2	27.0	27.5
			20	21.1	23.0	25.3	27.2	28.7	29.1
			50	22.4	24.3	26.7	28.4	30.6	31.1
		nDCG	10	20.9	22.8	24.8	26.3	28.2	28.6
			20	22.6	24.4	26.4	28.1	30.1	30.4
			50	27.2	29.1	31.1	32.8	34.7	35.2
	Cross-domain	MAP	10	18.2	20.7	21.9	23.3	25.1	25.4
			20	19.6	22.6	23.9	25.6	27.6	27.9
			50	20.9	24.1	25.5	27.2	29.5	29.8
		nDCG	10	21.2	23.9	23.9	25.8	27.9	28.1
			20	22.8	25.6	25.6	27.6	29.8	30.0
			50	26.2	28.2	30.2	31.9	32.9	33.2

表 15 不同词嵌入模型的性能比较，表现最佳模型的结果以粗体显示。

指标（Top_k）	MAP			nDCG		
	10	20	50	10	20	50
基线1使用 FastText	12.6	13.6	14.4	13.0	14.6	16.3
基线1使用 SciBERT	14.1	15.3	16.2	14.5	16.4	18.3
基线1使用 GPT3.5	14.4	15.7	16.8	16.1	17.6	18.8
基线2使用 FastText	15.4	16.4	18.3	15.9	17.6	20.7
基线2搭配 SciBERT	17.0	19.0	20.6	17.5	20.4	23.3
基线2搭配 GPT3.5	17.3	19.4	21.2	19.4	21.7	23.7
我们的方法搭配 FastText	16.1	18.5	20.1	16.6	19.9	22.7
我们的方法搭配 SciBERT	21.8	24.2	26.3	22.5	26.0	29.7
我们的方法搭配 GPT3.5	22.3	25.1	27.3	25.0	28.1	30.8

表16 实体类型配置的统一消融实验（第4.4.2(3)节），STM-KG（域内）。所有运行均使用相同的候选池和评估方法。每列最优结果以**粗体**标示。

实体类型	**向量操作**	**MAP@K ↑**			**nDCG@K ↑**		
		@10	@20	@50	@10	@20	@50
标题+摘要+引文	单向量（文档 s_p）	13.7	16.1	17.7	16.1	18.4	19.8
T	单向量 p_t	17.8	20.3	22.2	20.4	22.9	25.0
T+M	拼接 $[p_t, p_m]$	20.0	22.6	24.6	22.6	25.4	27.8
T+M+D	$p_g = [c_t, c_m, c_d, s_p]$	22.3	25.1	**27.3**	25.0	28.1	**30.8**

注：$p_t = [c_t, s_p]$，$p_m = [c_m, s_p]$，$p_g = [c_t, c_m, c_d, s_p]$。

表 17 §3.4 中定义的向量相似度组合／加权策略（第 4.4.2(4) 节），STM-KG（域内）。所有运行使用相同的候选池和评估。每列最佳结果以**粗体**显示。

组合类型 向量操作		**平均准确率均值/列表内多样性@K**			**归一化折损累计增益/覆盖率@K**		
		@10	@20	@50	@10	@20	@50
T+M+D	$p_g = [c_p^t, c_p^m, c_p^d, s_p]$	22.3/28.9	25.1/30.7	27.3/32.4	25.0/29.8	28.1/33.5	30.8/37.2
T+M+D	未加权：$w = \frac{1}{4}\mathbf{1}$	19.3/30.4	21.6/32.1	24.1/35.6	21.3/31.2	24.0/26.0	26.6/40.2
T+M+D	*learned* w (λ=0)	15.2/37.1	16.9/38.4	18.7/39.6	16.3/37.9	18.3/41.0	20.1/44.1
T+M+D, CP_{task}	*learned* $w + \boldsymbol{\alpha}$	12.2/**38.6**	15.3/**43.1**	16.4/**47.2**	12.6/**40.4**	16.4/**45.7**18.4/**50.1**	

注释：(1 7 $p = \sum_{i \in \{g,t,m,d\}} w_i p_i$ $w_i \geq 0$ $\sum w_i = 1$ 2 9 $\{s_j\}_{j=1}^{4}$ CP_{task} $\alpha_j \geq$) 0

式 ( ) 与 , 形成；( ) 式 ( ) 在内部聚合，使用 , $\sum \alpha_j = 1$。

表18时间去污染协议下的仅文本对比（STM-KG，域内）。3 次运行的均值±标准差；与 SciBERT 进行配对 t-检验。

设置	MAP@50	nDCG@50	实体感知相关性
BM25（仅文本）	22.6±0.1	23.0±0.1	0.412±0.006
SciBERT（仅文本）	24.7±0.1	25.3±0.1	0.441±0.005
我们的方法（仅文本）	26.3±0.1	29.7±0.1	0.489±0.005
我们的方法（完整）	**27.3**±0.1	**30.6**±0.1	**0.507**±0.004
p-value vs. SciBERT	<0.01	<0.01	<0.01
Ablation: remove citation-side structure (no co-citation / bib.-coupling)			
我们的方法（完整） → 无引用图	26.6±0.1	29.9±0.1	0.498±0.004

表19 用于评估流行度偏差的分层均衡评估（STM-KG，MAP@50）。3次运行的均值±标准差。

子集	SciBERT（仅文本）	我们的（完整）	Δ
低引用三分位组	23.6±0.1	**26.2**±0.1	+2.6
中引用三分位组	24.9±0.1	**27.4**±0.1	+2.5
高引用三分位组	25.6±0.1	**28.3**±0.1	+2.7
平衡子集（顶部编码）	24.5±0.1	**27.0**±0.1	+2.5

算法 1 实体感知多向量论文推荐

1: **输入：** 2: q – 查询论文3: P – 论文集4: $G = (P,\ E,\ L_{PP},\ L_{PE},\ L_{EE})$ – FG-SKG：论文、实体、引文、论文-实体

y

链接、实体-实体链接 5：K – 候选池大小（前-K）6：N – 返回的推荐结果数量（前-N）7：8：// A. 节点嵌入与实体聚合 [§3.3.1，公式 (1)(3)]9：对每篇论文 $p \in P$ 执行 10：$s_p \leftarrow$SPECTER（ p的标题/摘要）11：// 使用 GPT 对实体文本进行编码，并通过求和池化进行聚合 12：$c_p^t \leftarrow \sum_{e \in \mathcal{T}_p}$GPT($e$) 13：$c_p^m \leftarrow \sum_{e \in \mathcal{M}_p}$ GPT(e) 14：$c_p^d \leftarrow \sum_{e \in \mathcal{D}_p}$GPT($e$) 15：结束 16：17：// B. 多向量组合 [§3.3.2，公式 (4)(6)]18：对每篇论文 $p \in P$ 执行19 $p_g \leftarrow [c_p^t, c_p^m, c_p^d, s_p]$ ：： $p_t \leftarrow [c_p^t, s_p]$ 21：$p_m \leftarrow [c_p^m, s_p]$22：$p_d \leftarrow [c_p^d, s_p]$23：结束 24：25：// C. 候选生成（粗排） [§3.4，公式 (7)(8)]26：对每篇论文 $p \in P$ 执行 27：$p^* \leftarrow \sum_{i \in \{g,t,m,d\}} w_i \cdot p_i$ // $w_i \geq 0$、$\sum_i w_i = 1$ （在开发集上学习）28：得分 $\leftarrow \cos(q, p^*) = \frac{q^\top p^*}{\|q\| \cdot \|p^*\|}$ 29：结束 30：按得分排列的前-K 篇论文 31：32：// D. 任务感知优化（精排） [§3.4，公式 (9)]33： $CP_{\text{task}} \leftarrow$ 的任务相似子集 34：对每篇论文 $p \in CP_{\text{task}}$ 执行 35：$p_{tm} \leftarrow [c_p^t, c_p^m, s_p]$36：$p_{td} \leftarrow [c_p^t, c_p^d, s_p]$37：// 四个信号：任务±方法/材料 38：$\{s_j$（$q$，$p$）计算信号（$q$，$p_{tm}$，$p_{td}$）39：$\widehat{\ }\cos(q,\ p) \leftarrow \sum_{j=1..4} \alpha_j \cdot s_j$（$q$，$p$）// $\alpha_j \geq 0$、$\sum_j \alpha_j = 1$ （在开发集上学习） 40：结束 41：42：// E. 排序并返回 [§3.4，公式 (10)]43：对每篇论文 $p \in CP$ 执行 44：排序(q，p) $\leftarrow \sum_{i \in \{g,t,m,d\}} w_i \cdot \cos(q,\ p_i)$ 或 $\sum_{j=1..4} \alpha_j \cdot s_j$（$q$，$p$）45：结束 ^{55}46：返回按排序($q$，$p$)排列的前-$N$ 篇论文