arXiv:2302.03341v1 [cs.DL] 7 Feb 2023

元数据对科学文献标注的影响：跨领域跨模型研究

Yu Zhang、Bowen Jin、Qi Zhu、Yu Meng、Jiawei Han 伊利诺伊大学厄巴纳-香槟分校 {yuz9, bowenj4,qiz3, yumeng5,hanj}@illinois.edu

摘要

随着万维网上的科学出版物呈指数级增长，迫切需要为每篇论文标注细粒度主题，以便研究人员追踪自己感兴趣的研究领域，而不至于淹没在浩瀚的文献之中。科学文献标注并非单纯的多标签文本分类任务，因为万维网上的论文通常附有元数据信息，例如发表场所、作者和参考文献，这些信息可以作为额外信号，用于推断相关标签。尽管已有研究在学术论文分类中利用元数据，但其关注范围通常局限于一两个科学领域（例如计算机科学和生物医学），且只采用一种特定模型。在这项工作中，我们系统地研究了元数据对19个字段中科学文献标注的影响。我们选取了三种具有代表性的多标签分类器（即词袋模型、基于序列的模型和预训练语言模型），并探究将元数据作为额外特征输入分类器后，它们在科学文献标注任务中的表现变化。我们发现，元数据的影响在所有领域中呈现出一些普遍规律（例如，在几乎所有情况下，发表场所都能持续提升论文标注效果）；此外，在计算机科学和生物医学以外的领域中，也存在一些以往研究尚未探讨的独特规律。

CCS 概念

• **信息系统** → **数字图书馆与档案**；**数据挖掘**；**万维网**。

关键词

科学文献标注；元数据；文本分类

ACM 参考文献格式：

Yu Zhang、Bowen Jin、Qi Zhu、Yu Meng、Jiawei Han。2023。《元数据对科学文献标注的影响：跨字段、跨模型研究》。载于《*2023* 年 *ACM Web* 会议论文集（*WWW '23*）》，*2023*年*5*月*1*日至*5*日，美国得克萨斯州奥斯汀。ACM，纽约州纽约，11 页。https://doi.org/10. 1145/3543507.3583354

1 引言

各种学术服务平台，如谷歌学术、AMiner [40], Microsoft Academic [38], Semantic Scholar [1], 和 PubMed [26],在万维网上广受关注。此类平台的一项主要目标是帮助研究人员查询和跟踪学术信息与资源。与此同时，科学出版物的数量正呈指数级增长，每 12 年翻一番 [12]，到 2019 [42]时达到 240,000,000 篇。在这个信息爆炸的时代，为每篇科学论文准确标注相关主题，比以往任何时候都更加重要，这样研究人员就能跟踪自己感兴趣的研究领域，而不至于被浩如烟海的文献淹没。图 1 展示了科学文献标注任务的一个示例：该任务旨在根据论文《*Web*中的图结构》，预测“万维网”“网络图”和“链接农场”等标签。

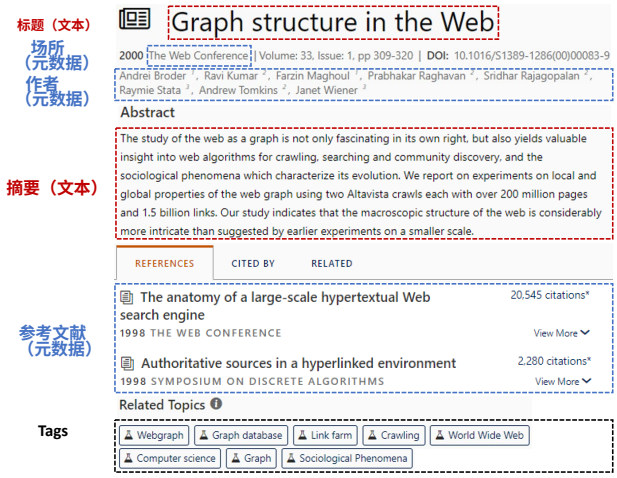


图 1：一篇包含来自 Microsoft Academic 的元数据的科学论文 [38]。科学文献标注的目标是预测其相关主题。

先前研究 [47, 59] 指出，科学文献标注不只是多标签文本分类任务，因为学术论文附带元数据，使其比纯文本序列更加复杂。这些元数据信息，包括发表场所、作者和参考文献，能够有力地指示每篇论文的相关主题。例如，在图 1 中，发表场所“*WWW*”表明该论文与“计算机科学”和“万维网”相关，而作者和参考文献则可能进一步指示“网络图”和“链接农场”等细粒度标签。

尽管现有关于科学文献标注的研究 [45, 47,57, 59, 60] 提出将元数据特征融入标注器，但仍存在两大主要局限。首先，研究范围仅限于一两个科学领域（例如计算机科学和生物医学）。由于缺乏这些领域中用于细粒度论文标引的基准数据集，元数据在其他领域（例如艺术、经济学、数学、物理学等）中的影响一直难以通过实证检验。其次，已有研究提出的

†代码和数据集可在以下地址获取：https://github.com/yuzhimanhua/MAPLE 以及 https://doi.org/10.5281/zenodo.7611544，分别对应代码和数据集。

在不收取费用的情况下，允许为个人或课堂使用而制作本作品全部或部分内容的数字或纸质副本，前提是这些副本不得用于营利或商业利益，也不得为此进行分发，且副本首页须保留本声明和完整引文。必须尊重本作品中由作者以外的其他人持有版权的组成部分。注明出处后可以摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，则需事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。
WWW '23，*2023*年*5*月*1*日至*5*日，美国得克萨斯州奥斯汀© 2023 版权所有归所有者/作者所有。出版权授权给 ACM。ACM ISBN 978-1-4503-9416-1/23/04...$15.00
https://doi.org/10.1145/3543507.3583354

元数据感知型标注器主要从头开始训练循环神经网络 [5]或 Transformer[41] 架构。目前尚不清楚，所提出的元数据用法能否推广到其他方法，例如词袋分类器（在关注效率的大规模多标签文本分类研究中，这类分类器仍得到广泛研究 [23]）以及预训练语言模型。

贡献。 为解决以往研究的上述两项局限，本文针对元数据对科学文献标注的影响，开展了一项系统的跨领域、跨模型研究。首先，我们基于微软学术图谱 [38]构建了大规模科学文献标注基准 Maple（Metadata-Aware Paper colLEction，元数据感知型论文集）。Maple 覆盖 19 个科学领域，包含超过 1190 万篇论文。每个领域的候选标签数量介于 ~700 和 ~64,000 之间。随后，我们考察了三种主要的多标签分类方法：词袋分类器 [2, 16, 19, 31–34, 39, 49, 50, 54],基于序列的分类器 [21, 44, 45, 47, 53, 59], 和预训练语言模型 [4, 17, 48, 52, 56, 60]。我们从这三类方法中各选取一种代表性模型，即 Parabel [33], Transformer [44],和 OAG-BERT [24],，这些模型均可通过简单修改，将文本和元数据共同作为分类输入。基于 Maple，我们使用选定的三种分类器，探索发表场所、作者和参考文献对 19 个领域论文标注的影响。我们得出以下主要观察结果：

• 元数据的影响因不同领域和分类器而有显著差异。总体而言，发表场所在几乎所有情况下都持续发挥积极作用，而作者和参考文献的作用则高度依赖于领域和分类器。例如，作者信息显然有助于哲学领域的论文标注，却会对地质学领域产生负面影响；在数学领域使用 Parabel 时，参考文献会带来帮助，但采用 Transformer 时，参考文献在同一领域反而会产生不利影响。

• 属于同一高层级科学领域的两个领域，其元数据影响往往相似。例如，生物学与医学都属于生命科学（根据 [35, 51]），发表场所、作者和参考文献在这两个领域中的影响大体一致。这一发现表明，在一个领域使用元数据的经验可以推广到相近领域。

• 我们还研究了元数据在预测不同粒度级别标签时的影响。在许多领域，发表场所不仅能提升粗粒度标签的预测表现，也能提升非常细粒度标签的预测表现；这在计算机科学领域可能并不常见。主要原因是，这些领域中的某些发表场所能够指示非常细粒度的标签（例如，历史领域中的“《罗马考古学杂志》”，哲学领域中的“*Mediaeval Studies*”）。

总而言之，本研究的贡献如下：(1) 我们构建了大规模基准 Maple，用于涵盖19个字段的科学文献标注，其字段覆盖范围远超以往论文标注研究所使用的数据集 [45, 47, 59]。(2) 我们全面评估了纳入元数据特征后，不同类型的多标签分类器在科学文献标注中的表现。(3) 我们的实证发现揭示了元数据影响在所有字段中普遍存在的一些模式，以及在计算机科学和生物医学以外的字段中独有的一些模式，而以往研究尚未探讨这些模式。我们的系统性研究旨在为从业者提供见解，帮助他们构建能够惠及所有字段的科学文献标注器。

表1：Maple中涵盖19个字段的20个数据集的统计信息。计算机科学领域包含2个数据集，其中一个收集自顶级会议，另一个收集自顶级期刊。

字段	论文来源	论文数	标签数	出版场所数	作者数	参考文献数
Art	期刊	58,373	1,990	98	54,802	115,343
哲学	期刊	59,296	3,758	98	36,619	198,010
地理学	期刊	73,883	3,285	98	157,423	884,632
商业	期刊	84,858	2,392	97	100,525	685,034
社会学	期刊	90,208	1,935	98	85,793	842,561
历史	期刊	113,147	2,689	99	84,529	284,739
政治《科学》	期刊	115,291	4,990	98	93,393	480,136
环境《科学》	期刊	123,945	694	100	265,728	1,217,268
经济学	期刊	178,670	5,205	97	135,247	1,042,253
工程学	期刊	270,006	10,683	100	430,046	1,867,276
心理学	期刊	372,954	7,641	100	460,123	2,313,701
计算机《科学》	会议	263,393	13,613	75	331,582	1,084,440
	期刊	410,603	15,540	96	634,506	2,751,996
地质学	期刊	431,834	7,883	100	471,216	1,753,762
数学	期刊	490,551	14,271	98	404,066	2,150,584
材料《科学》	期刊	1,337,731	6,802	99	1,904,549	5,457,773
物理学	期刊	1,369,983	16,664	91	1,392,070	3,641,761
生物学	期刊	1,588,778	64,267	100	2,730,547	7,086,131
化学	期刊	1,849,956	35,538	100	2,721,253	8,637,438
医学	期刊	2,646,105	36,619	100	4,345,385	7,405,779

2 预备知识

文本与元数据。 我们将论文 p的文本信息表示为单个文本序列 $\mathrm{T}_p = w_1 w_2 ... w_{|T_p|}$，该序列由论文的标题和摘要拼接而成。论文 p的元数据表示为集合 $\mathrm{M}^p = \{m_1, m_2, ...m_{|\mathcal{M}_p|}\}$，其中包含其场所、作者和参考文献。

问题定义。 科学文献标注任务可以表述为一个大规模多标签分类问题，其中所有候选标签（例如“有机化学”“偶联反应”“铃木反应”）构成一个大型标签空间 L（例如包含 10^3– 10^5 个标签）。给定一篇包含文本和元数据信息的科学论文，任务是从 L 中找出一组与该论文相关的标签。形式化地说，该问题定义如下。

定义 *2.1*。 （问题定义）给定训练语料库 D 和标签空间 L，其中每篇论文 $p \in \mathrm{D}$ 都关联着其文本 T^p、元数据信息 M^p以及相关标签 $\mathrm{L}_p \subseteq \mathrm{L}$，我们的目标是学习一个多标签文本分类器 f_{class}，该分类器能够将未标注论文 $p' \notin \mathrm{D}$ 映射到其相关标签 $\mathrm{L}_{p'} \subseteq \mathrm{L}$。

3 数据集构建

以往关于科学文献标注的研究 [9, 31, 45, 47, 52, 59]主要使用计算机科学和生物医学领域的论文来评估其提出的模型，同时较少关注其他科学领域。为弥补这一不足，我们构建了 Maple，这是一个用于评估科学文献标注的多领域基准。Maple 基于微软学术图谱（MAG） [38], 中的数据构建，该图谱已广泛应用于科学文本挖掘 [6, 51, 59]。MAG 涵盖 19 个学术领域，详见表 1。对于每个领域，我们执行以下步骤来构建数据集。

出版场所选择。 MAG 根据 *h*-index [13]维护着各个领域排名前100的期刊列表。在构建 Maple 时，我们重点关注发表在这些顶级期刊上的论文。需要注意的是，MAG 也将一些预印本服务（例如“*arxiv*”、“*bioRxiv*”、“*SSRN*”、“*NBER*工作论文”） [43]视为顶级期刊，但我们将其排除在考虑范围之外

因此，在表1中，构建的一些数据集包含的发表场所不足100个。在19个字段中，计算机科学（计算机科学）有着独特的出版文化：计算机科学论文通常首先或仅在会议上发表，而非期刊。因此，对于计算机科学领域，除了期刊论文集外，我们还根据CSRankings[1]构建了另一个包含来自75个顶级会议的论文的数据集。尽管如此，我们总共将为这19个领域构建20个数据集。

标签空间构建。对于每个字段，我们都需要一组用于论文标注的候选标签。MAG具有一个采用有向无环图（DAG）结构的标签分类体系 $\mathrm{L}_{\mathrm{MAG}}$ [37]。该分类体系有6个层级，包含超过10^5个标签，其中19个字段中的每一个都是第0层标签（即粒度最粗的标签）。给定一个字段 F，我们从分类体系 $\mathrm{L}_{\mathrm{MAG}}$ 中提取其后代标签，作为该字段的标签空间 L_F，但将 F 本身排除在 L_F 之外，因为在分类中，预测根标签毫无意义。由于一个标签在有向无环图结构的分类体系中可能有多个父节点，因此它可能出现在两个或多个字段的标签空间中。例如，标签“解剖学”既是“生物学”也是“医学”的子标签，因此在 Maple 中，它既是生物学数据集的候选标签，也是医学数据集的候选标签。

论文筛选。MAG 中的每篇论文 p 都标注有其相关标签 $\mathrm{L}_p \subseteq \mathrm{L}_{\mathrm{MAG}}$ [37][2]。要纳入 Maple（给定一个字段F），论文 p 需要满足以下两个条件：（1） p 发表于 F的一个选定场所；（2） p标注有 F，且至少标注有 F的一个后代标签（即 $F \in \mathrm{L}p$和 $|\mathrm{L}_p \cap \mathrm{L}_F| \geq 1$）。研究某个字段 F中的科学文献标注任务时，我们只关注与该字段相关的标签。因此， p的真实标签定义为 $\mathrm{L}p|F = \mathrm{L}_p \cap \mathrm{L}_F$。请注意，一篇论文可能属于多个字段；在考虑不同字段时，其真实标签也会不同。例如，如果一篇论文标注有“医学”“多重用药”“计算机科学”和“图嵌入”，且“多重用药”是医学字段中的候选标签，“图嵌入”是计算机科学字段中的候选标签，那么该论文会同时出现在 Maple 的医学和计算机科学数据集中。然而，在医学字段中进行标注时，该论文的真实标签是“多重用药”；在计算机科学字段中进行标注时，该论文的真实标签是“图嵌入”。

文本与元数据提取。对于每篇入选论文 p，我们从 MAG 中提取其标题、摘要、场所、作者和参考文献。标题和摘要拼接后构成文本信息 $\mathrm{T}p$。场所、作者和参考文献构成该论文的元数据特征 $\mathrm{M}p$ 。

训练-验证-测试划分。Maple 包含 1981 年 1 月 1 日至 2020 年 12 月 31 日期间发表的学术论文。我们使用 1981 年至 2015 年的论文进行训练和验证，使用 2016 年至 2020 年的论文进行测试。

构建的20个数据集在 Maple 中的统计信息见表1。更多详细信息请参见附录A.1。下文中，为便于讨论，在不引起歧义的情况下，我们将“字段”和“数据集”这两个术语互换使用。（换言之，我们将“计算机科学（会议）”和“计算机科学（期刊）”视为不同字段。）

[1]https://csrankings.org/[2]论文标签 $\mathcal{L}p$ 来自 Microsoft Academic 提出的一套系统 [37] 的预测结果。这些标签经人工核验，准确度较高 [37]，并已用于支持重要发现 [18, 51]。此外，我们还在三个带有医学主题词表标签 [7], 的数据集上进行了实验，这些数据集由生物医学专家整理。关于另外三个数据集的讨论，请参见附录A.5。

4 模型

过去十年间，大规模多标签文本分类（LMTC）得到了广泛研究。人们提出了多种方法，可用于科学文献标注。根据文本在分类器中作为特征的使用方式，现有的 LMTC 方法可分为三大类：（1）词袋分类器 [2, 16, 19, 31–34, 39, 49, 50, 54] 将每篇文档视为词元多重集，同时忽略词语位置和顺序。这类模型通常采用树、嵌入和线性层。（2）基于序列的分类器 [21, 44, 45, 47, 53, 59] 将每篇文档视为词元序列，并从头开始训练卷积神经网络、循环神经网络或 Transformer 架构，以构建多标签分类器。（3）预训练语言模型分类器 [4, 17, 48, 52, 56, 60] 旨在将从网络规模语料库（如维基百科和 PubMed）中学到的知识迁移到 LMTC 任务中，从而补充训练语料库中的文本信息。由于本文旨在研究元数据的影响，我们从这三类中各选取一个易于扩展以纳入元数据信息的 LMTC 模型。下面介绍这三个模型——Parabel [33]（词袋）、Transformer [44]（基于序列）和 OAG-BERT [24]（预训练语言模型）——以及它们如何将文本和元数据特征作为输入用于分类。

4.1 词袋分类器：Parabel [33]

4.1.1 仅使用文本。一般而言，词袋分类器将每个文档 p表示为一个 $|\mathrm{V}_{\mathcal{D}}|$维向量 $\boldsymbol{x}_p$，其中 $\mathrm{V}_{\mathcal{D}}$是训练语料库 D 的词汇表。给定一个词 $w \in \mathrm{V}_{\mathcal{D}}$，其在$\boldsymbol{x}_p$ 中的对应条目使用 TF-IDF 分数定义：

$$x_{p,w} = \mathrm{tf}(w, p) \cdot \mathrm{idf}(w, \mathcal{D}). \tag{1}$$

其中，$\mathrm{tf}(w, p)$ 表示 w在文档 p中的词频；$\mathrm{idf}(w, \mathrm{D}) = \log \frac{|\mathcal{D}|}{|\{p' \in \mathcal{D}(\mathcal{D}: w \in \mathcal{T}p'\}|}$表示 w的逆文档频率。

每个文档的相关标签由一个 $|\mathrm{L}|$维向量 $\boldsymbol{y}_p$表示，其中$y_{p,l} = 1$ 如果 l是与 p相关的标签（即，$l \in \mathrm{L}p$），而$y_{p,l} = 0$ 则表示其他情况。

为了执行多标签文本分类，Parabel 学习由多棵标签树组成的集成模型，其中每棵树都通过递归地将标签划分为两个大小均衡的组来构建，直到每个节点包含的标签少于某个指定数量。该划分过程基于标签表示 $\boldsymbol{v}_l$，通过球面二均值聚类实现；$\boldsymbol{v}_l$是一个单位向量，其方向为包含标签 l的训练点均值的方向。在完成标签空间划分后，Parabel 学习一个层次化判别分类器 $\mathrm{Pr}(\boldsymbol{y}_p|\boldsymbol{x}_p)$。具体而言，在每个非叶节点处，学习一个分布，以确定应遍历哪些子节点；在每个叶节点处，学习一个分布，以预测相关标签集合。有关更多技术细节，可参考 [33],，但由于这些细节与元数据的使用没有直接关系，本文不在此展开。

4.1.2 使用文本 + 元数据。将词袋分类器推广为使用元数据特征是直接可行的。令 $\mathrm{U}_{\mathcal{D}}$ 表示出现在 D 中的元数据实例集合。给定一个元数据实例 $m \in \mathrm{U}_{\mathcal{D}}$和一篇论文 p，我们定义 $\mathrm{tf}(m, p)$ 和 $\mathrm{idf}(m, \mathrm{D})$ 如下：

$$\mathrm{tf}(m, p) = \mathbf{1}(m \in \mathcal{M}_p), \quad \mathrm{idf}(m, \mathcal{D}) = \log \frac{|\mathcal{D}|}{|\{p' \in \mathcal{D} : m \in \mathcal{M}_{p'}\}|}, \tag{2}$$

其中 $\mathbf{1}(\cdot)$ 是指示函数。可以发现，公式（2）中的定义与 $\text{tf}(w, p)$ 和 $\text{idf}(w, \text{D})$ 的定义完全一致，只是元数据实例不会在文档中重复出现。

现在，我们可以为每篇论文构建一个“元数据袋”表示 $\tilde{\boldsymbol{x}}_p$。$\tilde{\boldsymbol{x}}_p$ 是一个 $|\mathbb{U}_\mathcal{D}|$ 维向量，其中

$$\tilde{x}_{p,m} = \text{tf}(m, p) \cdot \text{idf}(m, \mathcal{D}). \tag{3}$$

最后，我们将每篇论文表示为一个 $p(|\mathbb{U}_\mathcal{D}|+|\mathbb{V}_\mathcal{D}|)$ 维向量，即词袋表示与“元数据袋”表示的拼接 $\boldsymbol{x}_p \,||\, \tilde{\boldsymbol{x}}_p$。该向量被输入 Parabel 进行训练和预测。

4.2 基于序列的分类器：Transformer [44]

4.2.1 使用纯文本。为将 Transformer [41]应用于大规模多标签文本分类，我们遵循 [44], 提出的架构，该架构在文档文本 T_p 前添加多个“[CLS]”标记，作为输入序列。形式上，给定 $\text{T}_p = w_1 w_2 ... w_{|\mathcal{T}_p|}$，论文 p 的输入序列为

$$\mathcal{I}_p = [\text{CLS}_1]\ [\text{CLS}_2]\ ...\ [\text{CLS}_C]\ w_1\ w_2\ ...\ w_{|\mathcal{T}_p|} \tag{4}$$

这里的动机是，当标签空间较大时（例如包含 10_4 个标签），单个“[CLS]”标记的输出表示（例如，一个包含数百个维度的向量）可能不足以携带预测相关标签所需的信息。因此，需要多个“[CLS]”标记，从不同角度探查文本的语义。

Transformer 对输入序列 I_p进行编码后，每个标记 $w \in \text{I}_p$ 都会有一个输出表示 $\boldsymbol{h}_w$。我们将所有“[CLS]”标记的表示拼接起来，作为论文嵌入：

$$\boldsymbol{h}_p = \boldsymbol{h}_{[\text{CLS}_1]} \,||\, \boldsymbol{h}_{[\text{CLS}_2]} \,||\, ... \,||\, \boldsymbol{h}_{[\text{CLS}_C]}. \tag{5}$$

$\boldsymbol{h}_p$ 进一步输入全连接层，以进行多标签分类：

$$\hat{\boldsymbol{y}}_p = \text{Sigmoid}(\boldsymbol{W}^\top \boldsymbol{h}_p + \boldsymbol{b}). \tag{6}$$

其中，$\hat{\boldsymbol{y}}_p$ 是一个 $|\mathbb{L}|$ 维向量，$\hat{y}_{p,l}$ 表示论文 p 与标签 l 相关的预测概率。训练分类器的目标是最小化以下二元交叉熵（BCE）：

$$-\sum_{l\in\mathcal{L}} (y_{p,l} \log \hat{y}_{p,l} + (1 - y_{p,l}) \log(1 - \hat{y}_{p,l})). \tag{7}$$

4.2.2 使用文本 + 元数据。Transformer 中的全连接注意力机制为整合元数据提供了简便途径。具体而言，遵循 [59],，我们可以直接将元数据标记插入输入序列。给定论文文本 $\text{T}_p = w_1 ... w_{|\mathcal{T}_p|}$和元数据 $\text{M}_p = \{m_1, ..., m_{|\mathcal{M}_p|}\}$，输入序列为

$$\widetilde{\mathcal{I}}_p = [\text{CLS}_1]\ ...\ [\text{CLS}_C]\ m_1\ ...\ m_{|\mathcal{M}_p|}\ [\text{SEP}]\ w_1\ ...\ w_{|\mathcal{T}_p|} \tag{8}$$

与卷积神经网络 [21] 或循环神经网络 [45, 53] 分类器不同，在 Transformer 中，元数据实例的顺序并不重要，因为每个在编码过程中，（元数据，词）或（元数据，元数据）对可以通过全连接注意力机制相互作用。在我们的实验中，列出 $\widetilde{\text{I}}_p$ 中的作者时，我们遵循作者顺序（即第 1位 作者、第 2位 作者，.....）；列出 I_p 中的参考文献时，我们采用随机顺序，因为 MAG 中未存储引文顺序和行内上下文。

接下来，Transformer 的编码过程将每个元数据实例视为一个标记。例如，场所“*WWW*”被表示为一个标记“[Venue_1135342153]”，而不是包含多个标记的文本名称“网络会议”。

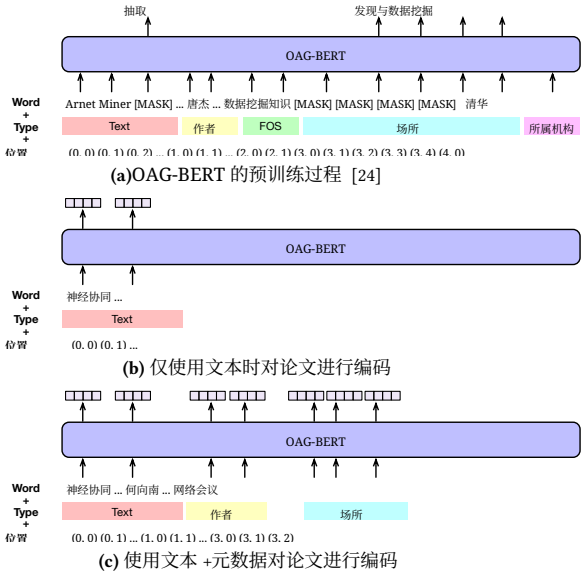


图2：预训练过程及我们对 OAG-BERT [24] 的使用。

将 $\widetilde{\text{I}}_p$ 输入 Transformer 后，元数据感知分类器的其余设计完全遵循公式（5）—（7）。

4.3 预训练语言模型分类器：OAG-BERT [24]

以预训练语言模型（PLM）为基础构建的分类器，主要利用 BERT [10],、BioBERT [20],、SciBERT [3],、XLNet [46], 或 RoBERTa [25] 来生成文档表示。然而，这些预训练语言模型在预训练期间大多侧重于文本信息，并非专为处理元数据而设计。为弥补这一不足，我们采用 OAG-BERT [24],，这是一种实体增强型学术语言模型，可以联合编码科学文本与场所/作者信息。

OAG-BERT 的预训练过程简要展示于图2(a)中。它将文本与元数据信息置于同一序列中进行掩码语言建模。此外，它还提出了三种处理元数据实体的策略：(1) 异构实体类型嵌入使模型能够识别不同的元数据类型；(2) 跨度感知实体掩码会在较长实体中选取一个连续跨度（例如场所“知识发现与数据挖掘”）；(3) 二维位置嵌入联合建模实体间与实体内的词元顺序。模型首先在500万篇全文论文上训练，然后使用来自开放学术图谱 [55]的元数据，在1.2亿篇论文标题/摘要上继续训练。

4.3.1 仅使用文本。当仅考虑论文文本 T_p 进行分类时，我们首先使用 OAG-BERT 对文本序列进行编码（如图2(b)所示）。

$$\boldsymbol{H}_p = \text{OAG-BERT}(\mathcal{T}_p). \tag{9}$$

这里，$\boldsymbol{H}_p = [\boldsymbol{h}_{p,1}, ..., \boldsymbol{h}_{p,N}]$ 包含所有标记的输出向量。然后，我们采用均值池化来获得论文表示 $\boldsymbol{x}_p$。

$$\boldsymbol{x}_p = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}_{p,i}. \tag{10}$$

直接使用二元交叉熵损失（即式（7））对预训练语言模型进行微调非常耗时，尤其是在标签空间较大时。考虑到效率和模型简洁性，我们固定论文表示$\boldsymbol{x}_p$，以训练一个 Parabel分类器$f_{\text{Parabel}}(\boldsymbol{x}_p) = \Pr(\boldsymbol{y}_p|\boldsymbol{x}_p)$。

在预测阶段，给定一篇论文 $p_\prime$，训练好的分类器会预测 $p_\prime$ 与每个标签 l相关的概率：

$$\hat{\boldsymbol{y}}_{p'} = f_{\text{Parabel}}(\boldsymbol{x}_{p'}). \tag{11}$$

随后，我们可以根据$\hat{}y_{p',l}$对标签进行排序，以预测最相关标签的列表。然而，在实际操作中，我们发现这种策略无法取得具有竞争力的预测准确率。这一发现与相关研究 [4, 31] 中的设计一致，即在进行分类时，应将离散的词汇特征与连续表示结合起来。为添加词汇特征，受 [60],启发，我们提出了一种简单的启发式方法来重新排序标签：给定一篇论文$p_\prime$，名称出现在论文文本 Tp' 中的所有标签，其排序都将高于未出现在 Tp'中的标签。这种启发式方法等价于根据修改后的分数$\hat{}z_{p',l} = \hat{}y_{p',l} + \mathbf{1}(t_l \subseteq \text{T}p')$对标签进行排序，其中 t_l表示标签 l的文本名称。

4.3.2 使用文本 $_+$元数据。 当有元数据信息时，我们使用 OAG-BERT 对论文文本和元数据进行联合编码（如图 2(c) 所示）。

$$\widetilde{\boldsymbol{H}}_p = \text{OAG-BERT}(\mathcal{T}_p, \mathcal{M}_p). \tag{12}$$

其余步骤与仅文本模型完全相同。值得注意的是，由于 OAG-BERT 预训练时不涉及参考文献，因此这里我们只能使用发表场所和作者作为元数据。此外，与将每个发表场所/作者视为一个标记的 Transformer分类器不同，OAG-BERT 在编码时会根据其词汇表对发表场所/作者的文本名称进行分词。

5 实验

5.1 设置

数据集与对比方法。 我们已在第 3 节介绍了这 20个数据集。对于三种分类器中的每一种，我们分别测试仅使用文本、文本+发表场所、文本+作者和文本+参考文献时的性能，以单独考察每种元数据类型的影响。（请注意，OAG-BERT 无法将参考文献作为元数据特征，因此不支持文本+参考文献这一变体。）有关详细的超参数设置，请参阅附录 A.2。

评估指标。 遵循以往关于多标签文本分类的研究 [21, 44, 59], 我们采用 P@k和 NDCG@k作为评估指标，其中 $k = 1$，3，5。给定一篇论文 p，令 rank(i) 表示根据各分类器预测结果中第 i高标签的索引，则

$$\text{P@}k = \frac{1}{k}\sum_{i=1}^{k} y_{p,\text{rank}(i)}.$$
$$\text{DCG@}k = \sum_{i=1}^{k}\frac{y_{p,\text{rank}(i)}}{\log(i+1)},\quad \text{NDCG@}k = \frac{\text{DCG@}k}{\sum_{i=1}^{\min(k,\|\boldsymbol{y}_p\|_0)}\frac{1}{\log(i+1)}}. \tag{13}$$

5.2 总体分析

表 2 展示了三种分类器在 20 个数据集上的 P@k 和 NDCG@k 得分。每项实验运行 5 次，并报告平均得分。我们进行双尾 t 检验，以检查元数据的影响是否具有统计显著性。具体而言，给定一个字段和一种分类器，如果使用某种类型的元数据后，得分相比仅使用文本显著提高（p 值为 < 0.05），我们会将该得分标记为表[illegible]的蓝色；如果得分显著下降（p 值为 < 0.05），则会将该得分标记为红色。

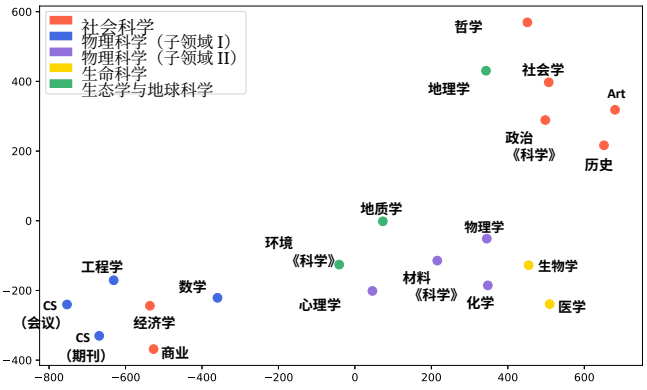


图 3：我们根据场所、作者和参考文献信息对三个分类器的影响，用一个24维向量表示每个字段。随后，我们应用 t-SNE [27]将这些字段可视化到二维空间中。颜色方案突出显示了若干高层次科学领域，这些领域遵循 [35, 51] 检测到的科学主要聚类，并表明每个领域内的元数据影响相似。

从表 2 中，我们可以观察到：（1）元数据的影响在不同字段之间存在显著差异。例如，当使用 Parabel 作为分类器时，作者信息对计算机科学与工程学具有显著帮助；但采用同一分类器时，在生物学和医学领域却会产生不利影响。当使用 Parabel 时，参考文献在商业与经济学领域中发挥作用；但如果使用 Transformer 作为标注器，在相同领域中则会产生不利影响。（2）在几乎所有情况下，发表场所都持续有利于科学文献标注。具体而言，使用 Parabel 时，全部20个字段都显著受益于发表场所[3]；使用 Transformer 时，有18个字段受益；使用 OAG-BERT 时，也有18个字段受益。在其余4个发表场所信息无益的情况下，它也不会产生负面影响。（3）当 Parabel 和 OAG-BERT 分别作为标注器时，作者信息在大多数领域（分别为15个和17个）中有所帮助，但采用 Transformer 时几乎没有帮助。参考文献发挥作用的情况更少。三类元数据的总体性能体现在20个数据集上的 P@k 和 NDCG@k 分数的宏平均中，相关结果见表 3。作者和参考文献在 Transformer分类器中不起作用的原因在于，它们的嵌入需要从头开始训练，缺乏良好的初始化。相比之下，仅利用发表场所只会引入少量额外参数。因此，如果希望利用作者和参考文献信息，一些嵌入预训练技术 [59] 可能会有所帮助。

5.3 不同领域中的影响

本节考察元数据在不同领域中的影响。给定一个领域 F，为描述元数据在 F 中的影响，我们按照以下方式构造一个 24 维向量：研究涉及三种分类器（即 Parabel、Transformer 和 OAG-BERT）以及三种元数据类型（即场所、作者和参考文献）

[3]如果在使用某种分类器时，纳入某一类型的元数据后，五项指标中至少有一项显著改善（p 值为 < 0.05），我们就称该领域显著受益于这一类型的元数据。

表 2：P@k 和 NDC G@k 在20个数据集上的得分。 显著优于仅使用文本（p 值 $<$ 0. 05）。 显著差于仅使用文本时的 n（p 值 $<$ 0.05）。“–”：OAG-BERT 无法将参考文献作为元数据信号 .

字段	输入	Parabel [33]					Transformer [44]					OAG-BERT [24]				
		P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
Art	Text	0.7203	0.4829	0.3392	0.7585	0.7921	0.6995	0.4396	0.3008	0.6956	0.7183	0.6982	0.4411	0.3030	0.7024	0.7253
	+场所	0.7235	0.4861	0.3417	0.7630	0.7969	0.7109	0.4456	0.3038	0.7065	0.7279	0.7032	0.4476	0.3066	0.7116	0.7335
	+作者	0.7247	0.4827	0.3393	0.7598	0.7936	0.7060	0.4447	0.3035	0.7024	0.7238	0.7032	0.4454	0.3062	0.7085	0.7318
	+参考文献	0.7186	0.4818	0.3389	0.7570	0.7913	0.7052	0.4432	0.3034	0.6997	0.7222	–	–	–	–	–
哲学	Text	0.8124	0.6033	0.4617	0.7578	0.7434	0.7493	0.5103	0.3712	0.6524	0.6219	0.7652	0.5548	0.4157	0.6980	0.6738
	+场所	0.8164	0.6322	0.4909	0.7834	0.7738	0.7889	0.5725	0.4285	0.7159	0.6897	0.7679	0.5621	0.4271	0.7045	0.6852
	+作者	0.8128	0.6050	0.4637	0.7593	0.7455	0.7544	0.5151	0.3743	0.6577	0.6262	0.7673	0.5594	0.4206	0.7030	0.6799
	+参考文献	0.8023	0.5968	0.4581	0.7478	0.7344	0.7442	0.5072	0.3698	0.6463	0.6160	–	–	–	–	–
地理学	Text	0.6233	0.4114	0.3015	0.6331	0.6640	0.5239	0.3140	0.2237	0.4887	0.5036	0.6220	0.3997	0.2843	0.6158	0.6341
	+场所	0.6283	0.4147	0.3036	0.6384	0.6694	0.5457	0.3251	0.2304	0.5083	0.5223	0.6244	0.4036	0.2879	0.6213	0.6406
	+作者	0.6215	0.4113	0.3019	0.6321	0.6637	0.5241	0.3141	0.2239	0.4888	0.5038	0.6237	0.4046	0.2885	0.6219	0.6414
	+参考文献	0.6135	0.4104	0.3021	0.6277	0.6603	0.4960	0.3030	0.2181	0.4665	0.4835	–	–	–	–	–
商业	Text	0.6601	0.4775	0.3666	0.6192	0.6351	0.6375	0.4344	0.3238	0.5736	0.5780	0.6687	0.4859	0.3648	0.6328	0.6408
	+场所	0.6667	0.4811	0.3690	0.6247	0.6402	0.6550	0.4435	0.3293	0.5878	0.5914	0.6693	0.4899	0.3688	0.6371	0.6462
	+作者	0.6626	0.4782	0.3673	0.6208	0.6367	0.6265	0.4255	0.3161	0.5619	0.5649	0.6701	0.4886	0.3679	0.6362	0.6455
	+参考文献	0.6587	0.4798	0.3694	0.6219	0.6389	0.5872	0.3985	0.2947	0.5237	0.5242	–	–	–	–	–
社会学	Text	0.6173	0.4207	0.3057	0.6219	0.6509	0.5851	0.3611	0.2514	0.5489	0.5614	0.6177	0.4004	0.2799	0.5998	0.6144
	+场所	0.6227	0.4257	0.3080	0.6297	0.6573	0.6026	0.3704	0.2576	0.5644	0.5768	0.6195	0.4046	0.2834	0.6048	0.6202
	+作者	0.6182	0.4206	0.3054	0.6220	0.6505	0.5970	0.3625	0.2517	0.5548	0.5664	0.6191	0.4027	0.2820	0.6026	0.6179
	+参考文献	0.6112	0.4210	0.3068	0.6200	0.6500	0.5805	0.3573	0.2491	0.5395	0.5516	–	–	–	–	–
历史	Text	0.7164	0.4574	0.3173	0.7480	0.7799	0.6941	0.4210	0.2869	0.6928	0.7174	0.6591	0.4122	0.2822	0.6776	0.7029
	+场所	0.7199	0.4615	0.3200	0.7537	0.7856	0.7008	0.4264	0.2897	0.7010	0.7245	0.6602	0.4147	0.2841	0.6809	0.7064
	+作者	0.7144	0.4571	0.3171	0.7469	0.7789	0.6988	0.4233	0.2881	0.6959	0.7198	0.6627	0.4161	0.2853	0.6832	0.7089
	+参考文献	0.7142	0.4565	0.3169	0.7463	0.7786	0.6999	0.4248	0.2892	0.6990	0.7233	–	–	–	–	–
政治学	Text	0.7723	0.5153	0.3767	0.7252	0.7291	0.7624	0.4863	0.3429	0.6892	0.6797	0.7327	0.4805	0.3393	0.6721	0.6624
	+场所	0.7776	0.5216	0.3817	0.7327	0.7374	0.7723	0.4948	0.3494	0.7004	0.6910	0.7320	0.4843	0.3438	0.6756	0.6672
	+作者	0.7737	0.5173	0.3779	0.7275	0.7313	0.7644	0.4882	0.3439	0.6913	0.6810	0.7356	0.4850	0.3436	0.6771	0.6680
	+参考文献	0.7633	0.5107	0.3750	0.7182	0.7237	0.7549	0.4812	0.3392	0.6811	0.6712	–	–	–	–	–
环境《科学》	Text	0.6484	0.4254	0.3087	0.6990	0.7299	0.6447	0.4003	0.2856	0.6706	0.6964	0.7185	0.4465	0.3156	0.7449	0.7663
	+场所	0.6488	0.4269	0.3094	0.7005	0.7313	0.6469	0.4023	0.2867	0.6730	0.6981	0.7196	0.4468	0.3157	0.7454	0.7669
	+作者	0.6484	0.4267	0.3095	0.7002	0.7313	0.6333	0.3945	0.2816	0.6590	0.6845	0.7189	0.4474	0.3166	0.7461	0.7683
	+参考文献	0.6486	0.4278	0.3112	0.7016	0.7340	0.6126	0.3843	0.2752	0.6388	0.6648	–	–	–	–	–
经济学	Text	0.7288	0.5842	0.4793	0.6483	0.6301	0.7375	0.5785	0.4635	0.6448	0.6167	0.6861	0.5651	0.4612	0.6229	0.6037
	+场所	0.7339	0.5885	0.4830	0.6534	0.6353	0.7460	0.5848	0.4690	0.6521	0.6243	0.6868	0.5677	0.4645	0.6251	0.6069
	+作者	0.7288	0.5856	0.4809	0.6496	0.6315	0.7224	0.5607	0.4477	0.6258	0.5967	0.6875	0.5676	0.4655	0.6254	0.6080
	+参考文献	0.7306	0.5866	0.4826	0.6511	0.6340	0.6867	0.5303	0.4203	0.5921	0.5609	–	–	–	–	–
计算机《科学》（会议）	Text	0.7158	0.5594	0.4597	0.6222	0.5909	0.6980	0.5287	0.4196	0.5924	0.5498	0.6516	0.5299	0.4349	0.5835	0.5541
	+场所	0.7196	0.5638	0.4638	0.6268	0.5958	0.7079	0.5400	0.4300	0.6040	0.5618	0.6508	0.5308	0.4364	0.5832	0.5541
	+作者	0.7164	0.5608	0.4610	0.6234	0.5923	0.6770	0.5122	0.4068	0.5727	0.5308	0.6519	0.5315	0.4375	0.5847	0.5561
	+参考文献	0.7151	0.5614	0.4618	0.6236	0.5927	0.6377	0.4777	0.3774	0.5353	0.4940	–	–	–	–	–
工程学	Text	0.7881	0.6383	0.5274	0.7064	0.6776	0.7801	0.6154	0.4961	0.6852	0.6461	0.6780	0.5668	0.4757	0.6234	0.6058
	+场所	0.7904	0.6399	0.5285	0.7087	0.6798	0.7818	0.6168	0.4980	0.6869	0.6486	0.6819	0.5699	0.4787	0.6270	0.6101
	+作者	0.7885	0.6397	0.5286	0.7078	0.6790	0.7679	0.5995	0.4820	0.6686	0.6286	0.6784	0.5686	0.4788	0.6251	0.6091
	+参考文献	0.7879	0.6398	0.5288	0.7079	0.6793	0.7261	0.5579	0.4453	0.6242	0.5832	–	–	–	–	–
心理学	Text	0.7784	0.6272	0.5133	0.7073	0.6889	0.8035	0.6434	0.5172	0.7274	0.7007	0.7503	0.6148	0.4986	0.6909	0.6686
	+场所	0.7871	0.6323	0.5173	0.7143	0.6958	0.8108	0.6481	0.5214	0.7339	0.7078	0.7536	0.6200	0.5047	0.6964	0.6758
	+作者	0.7775	0.6283	0.5154	0.7078	0.6903	0.7918	0.6300	0.5062	0.7126	0.6858	0.7489	0.6165	0.5015	0.6921	0.6712
	+参考文献	0.7821	0.6334	0.5209	0.7136	0.6973	0.7776	0.6205	0.4979	0.7017	0.6749	–	–	–	–	–
计算机《科学》（期刊）	Text	0.7075	0.5611	0.4645	0.6209	0.5862	0.7125	0.5477	0.4415	0.6110	0.5663	0.6356	0.5266	0.4416	0.5760	0.5481
	+场所	0.7098	0.5639	0.4673	0.6238	0.5893	0.7146	0.5501	0.4439	0.6134	0.5689	0.6359	0.5296	0.4465	0.5789	0.5529
	+作者	0.7094	0.5629	0.4662	0.6228	0.5883	0.6937	0.5278	0.4241	0.5897	0.5449	0.6362	0.5299	0.4467	0.5792	0.5533
	+参考文献	0.7089	0.5636	0.4675	0.6234	0.5894	0.6534	0.4920	0.3936	0.5503	0.5062	–	–	–	–	–
地质学	Text	0.8618	0.7327	0.6213	0.8052	0.7765	0.8752	0.7412	0.6254	0.8147	0.7825	0.7832	0.6622	0.5605	0.7276	0.7004
	+场所	0.8634	0.7347	0.6234	0.8071	0.7787	0.8779	0.7455	0.6299	0.8190	0.7875	0.7851	0.6637	0.5631	0.7295	0.7031
	+作者	0.8601	0.7319	0.6215	0.8036	0.7756	0.8641	0.7310	0.6171	0.8018	0.7693	0.7814	0.6600	0.5606	0.7251	0.6990
	+参考文献	0.8432	0.7197	0.6138	0.7886	0.7622	0.8485	0.7162	0.6031	0.7848	0.7503	–	–	–	–	–
数学	Text	0.7406	0.6191	0.5326	0.6584	0.6192	0.7715	0.6368	0.5403	0.6796	0.6335	0.6174	0.5315	0.4641	0.5612	0.5334
	+场所	0.7421	0.6210	0.5341	0.6603	0.6211	0.7756	0.6421	0.5455	0.6848	0.6390	0.6187	0.5351	0.4689	0.5643	0.5376
	+作者	0.7408	0.6214	0.5346	0.6603	0.6212	0.7642	0.6294	0.5330	0.6718	0.6252	0.6175	0.5347	0.4687	0.5637	0.5372
	+参考文献	0.7412	0.6237	0.5383	0.6623	0.6242	0.7324	0.5974	0.4999	0.6390	0.5895	–	–	–	–	–
材料《科学》	Text	0.8193	0.6638	0.5436	0.7746	0.7720	0.8670	0.6994	0.5696	0.8203	0.8157	0.7490	0.5935	0.4808	0.6932	0.6845
	+场所	0.8233	0.6667	0.5459	0.7786	0.7761	0.8692	0.7009	0.5707	0.8228	0.8180	0.7496	0.5943	0.4823	0.6946	0.6867
	+作者	0.8203	0.6659	0.5459	0.7766	0.7745	0.8558	0.6879	0.5597	0.8069	0.8016	0.7469	0.5928	0.4812	0.6921	0.6843
	+参考文献	0.8206	0.6654	0.5462	0.7762	0.7747	0.8510	0.6840	0.5567	0.8021	0.7970	–	–	–	–	–
物理学	Text	0.8367	0.7228	0.6219	0.7597	0.7050	0.8739	0.7552	0.6490	0.7941	0.7368	0.7461	0.6399	0.5585	0.6730	0.6300
	+场所	0.8452	0.7269	0.6248	0.7650	0.7095	0.8787	0.7593	0.6520	0.7985	0.7407	0.7491	0.6419	0.5609	0.6754	0.6328
	+作者	0.8372	0.7235	0.6236	0.7602	0.7061	0.8665	0.7472	0.6406	0.7857	0.7276	0.7415	0.6372	0.5566	0.6698	0.6270
	+参考文献	0.8340	0.7197	0.6213	0.7563	0.7029	0.8682	0.7486	0.6423	0.7873	0.7294	–	–	–	–	–
生物学	Text	0.8446	0.7503	0.6609	0.7804	0.7292	0.8764	0.7743	0.6798	0.8069	0.7524	0.7448	0.6517	0.5772	0.6804	0.6371
	+场所	0.8474	0.7523	0.6626	0.7829	0.7317	0.8791	0.7766	0.6813	0.8094	0.7544	0.7484	0.6571	0.5840	0.6856	0.6436
	+作者	0.8433	0.7495	0.6610	0.7793	0.7286	0.8652	0.7574	0.6600	0.7909	0.7332	0.7427	0.6528	0.5806	0.6808	0.6393
	+参考文献	0.8456	0.7505	0.6618	0.7807	0.7300	0.8722	0.7668	0.6697	0.8000	0.7432	–	–	–	–	–
化学	Text	0.8448	0.7208	0.6211	0.7720	0.7310	0.8866	0.7658	0.6587	0.8187	0.7749	0.7480	0.6311	0.5334	0.6768	0.6312
	+场所	0.8485	0.7229	0.6229	0.7756	0.7347	0.8896	0.7668	0.6590	0.8214	0.7774	0.7503	0.6361	0.5413	0.6819	0.6390
	+作者	0.8444	0.7220	0.6232	0.7727	0.7324	0.8811	0.7577	0.6499	0.8105	0.7654	0.7480	0.6346	0.5403	0.6794	0.6366
	+参考文献	0.8425	0.7182	0.6204	0.7694	0.7295	0.8810	0.7568	0.6484	0.8102	0.7649	–	–	–	–	–
医学	Text	0.7894	0.6491	0.5325	0.7335	0.7227	0.8421	0.6935	0.5669	0.7849	0.7711	0.6935	0.5744	0.4789	0.6458	0.6413
	+场所	0.7936	0.6528	0.5360	0.7379	0.7275	0.8436	0.6951	0.5685	0.7867	0.7729	0.7012	0.5818	0.4884	0.6534	0.6516
	+作者	0.7809	0.6428	0.5297	0.7246	0.7148	0.8291	0.6800	0.5565	0.7688	0.7549	0.6873	0.5727	0.4804	0.6424	0.6397
	+参考文献	0.7913	0.6520	0.5362	0.7364	0.7264	0.8394	0.6902	0.5633	0.7817	0.7672	–	–	–	–	–

表 3：P@k 和 NDCG@k 在 20个数据集上的宏平均。蓝色、红色和"- 2 相同。

	输入	Parabel [33]					Transformer [44]					OAG-BERT [24]				
		P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
宏平均	Text	0.7513	0.5811	0.4678	0.7076	0.6977	0.7510	0.5673	0.4507	0.6896	0.6712	0.6983	0.5354	0.4275	0.6549	0.6429
	+场所	0.7554	0.5858	0.4717	0.7130	0.7033	0.7599	0.5753	0.4572	0.6995	0.6812	0.7004	0.5391	0.4318	0.6588	0.6480
	+作者	0.7512	0.5817	0.4687	0.7079	0.6983	0.7442	0.5594	0.4433	0.6809	0.6617	0.6984	0.5374	0.4305	0.6569	0.6461
	+参考文献	0.7487	0.5809	0.4689	0.7065	0.6977	0.7277	0.5469	0.4328	0.6652	0.6459	–	–	–	–	–

在我们的实验中，因此共有 $3 \times 3 - 1 = 8$ 种（分类器，元数据）组合，因为（OAG-BERT，参考文献）不适用。对于 8 种（分类器，元数据）组合中的每一种，我们通过比较仅使用文本的分类器与同时使用文本和元数据的分类器，计算 P@1、P@3 和 P@5 的相对性能变化。例如，对于（Parabel，场所）组合，我们计算以下 3 个值：

$$\frac{\text{P@}k\,(\text{Parabel, Text + Venue}) - \text{P@}k\,(\text{Parabel, Text})}{\text{P@}k\,(\text{Parabel, Text})},\ k = 1, 3, 5. \quad (14)$$

总共会有 $3 \times 8 = 24$ 个值，可以构成一个 24 维向量 $\boldsymbol{u}_F$。我们计算所有 20 个字段的 $\boldsymbol{u}_F$，然后应用 t-SNE [27]，将这些向量可视化在二维空间中。

可视化结果如图 3 所示，我们根据 [35, 51] 检测到的高层级科学领域/子领域为各个字段着色。具体而言，Rosvall 和 Bergstrom [35] 基于一个大规模论文引文网络，将科学领域划分为四个主要领域——社会科学、自然科学、生命科学以及生态学与地球科学。自然科学这一类别又进一步划分为两个子领域，其中一个与数学和计算机科学有关，另一个与物理学和化学有关；Yin等人[51] 在研究各字段中科学论文的公众使用情况时，也观察到了对这 19 个字段的类似划分。

在图 3 中，我们发现，颜色相同的字段通常嵌入得更近，这表明属于同一领域/子领域的字段受到元数据的影响相似。当我们需要将在某个领域使用元数据的经验推广到其他领域时，这一发现非常有用

从一个字段推广到相似字段。例如，已知在医学领域使用 Parabel 时，参考文献有益，而作者有害。根据我们的发现，在生物学领域使用 Parabel 时也可以推断出相同的规律，因为医学和生物学都属于生命科学。根据表 2，这一推断是正确的。

与此同时，我们观察到两个例外：（1）社会科学领域分为两个簇，其中一个包含艺术、历史、哲学、社会学和政治学，另一个包含经济学和商业。商业与经济学领域中的元数据效应，与数学和计算机科学领域中的效应更为相似。这可能是因为大量经济学和商业论文依赖对海量数据进行定量分析来得出结论，这与大多数艺术、历史和哲学论文采用的范式不同。（2）地理学与社会科学的距离比与地质学和环境科学的距离更近。这可能是因为地理学是一个跨学科领域。自然地理学子领域与地球科学的关系更为密切，而人文地理学子领域则更接近社会科学。总的来说，我们仍然观察到，在高层级的领域／子领域内，元数据效应存在共性。

5.4 不同标签粒度下的效果

现在，我们考察元数据对不同粒度级别标签预测的影响。MAG 分类体系 [37] 共有 6 层（从粒度最粗的第 0 层到粒度最细的

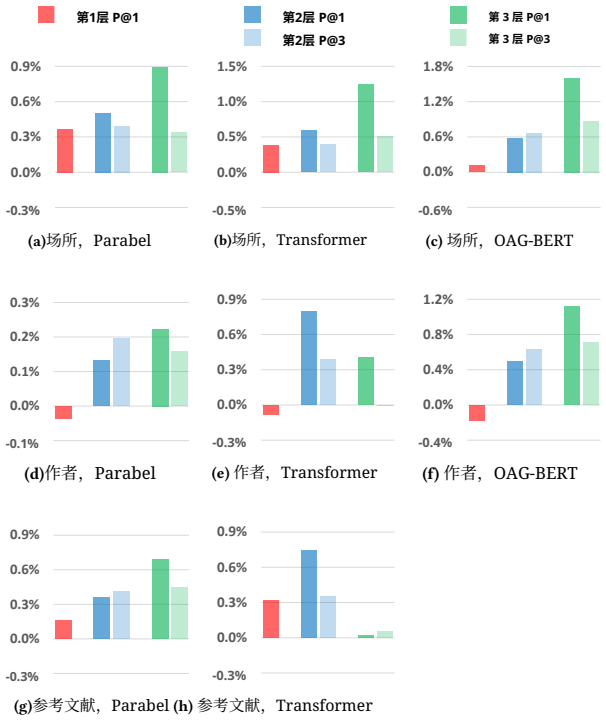


图4：元数据对第-j 层P@k 得分的影响（对显著受益于利用该类型元数据的字段进行分类器平均）。

表4：在第1层 P@1、第2层 P@1 和第 3 层 P@1 方面，最能从利用出版场所信息中获益的字段。"T"：仅文本。"+V"：文本+场所。"Δ"：绝对性能提升。

	L1 精确率@1			L2 精确率@1			L3 精确率@1		
Parabel [33]									
物理学	T	0.7178	哲学	T	0.7846	哲学	T	0.6118	
	+V	0.7284		+V	0.8170		+V	0.7209	
	Δ	1.06%		Δ	3.24%		Δ	10.91%	
Transformer [44]									
物理学	T	0.7393	哲学	T	0.6728	哲学	T	0.4105	
	+V	0.7495		+V	0.7504		+V	0.6071	
	Δ	1.02%		Δ	7.76%		Δ	19.66%	
OAG-BERT [24]									
地理学	T	0.6626	Art	T	0.6747	历史	T	0.3455	
	+V	0.6679		+V	0.6879		+V	0.4130	
	Δ	0.53%		Δ	1.32%		Δ	6.75%	

第5层）。第0层标签是指我们从标签空间中排除的19个字段。在全部20个数据集中，分别有85.2%、86.5%、70.4%、35.0%和13.2%的论文具有真实标注的第1层、第2层、第3层、第4层和第5层标签。由于与第4层和第5层标签相关的论文比例较低，我们只研究元数据对预测第1层、第2层和第3层标签的影响。

对于一种分类器和一种类型的元数据，我们考察使用该分类器时能从元数据中显著受益的所有字段。在这些字段中，我们计算第j层P@k分数（(j，k) =(1，1)，(2，1)，(2，3)，(3，1)，(3，3)）。第j层P@k的定义与P@k非常相似，不同之处在于，第j层P@k考察的是第j层中置信度最高的k个标签，而非整个标签空间中的标签。计算第j层P@k时，我们关注至少有一个真实标注的第j层标签的论文。随后，我们计算纳入所考察的元数据类型后，第j层P@k分数的绝对性能变化。结果如图4所示。

从图4中，我们发现：（1）平均而言，发表场所不仅有助于科学文献的粗粒度标签标注，也有助于细粒度标签标注。从计算机科学家的角度来看，这一结论可能有些违反直觉，因为计算机科学出版场所（例如“*WWW*”）很难指示非常细粒度的标签（例如“链接农场”）。事实上，在计算机科学（会议）数据集中，发表场所对第3层 P@1 的贡献十分有限（例如，使用 Parabel 时为 0.19%，使用 Transformer 时为 -0.11%）。然而，在计算机科学以外的字段中，一些发表场所确实携带非常细粒度的信号。例如，历史与哲学领域中分别有两个发表场所：“《罗马考古学杂志》”和“*Mediaeval Studies*”。这两个发表场所可能强烈暗示论文与“古典考古学”和“中世纪主义”相关，而这两个标签分别属于第2层和第3层标签。在表4中，我们列出了在利用出版场所信息后，就第1层 P@1 、第 2 层 P@1 和第 3 层 P@1 而言获益最大的字段；我们确实观察到，就第3层 P@1 而言，哲学和历史是受益最大的领域。（2）与发表场所不同，作者有利于细粒度标注，却不利于粗粒度预测。这一观察结果在所有三种分类器中都一致成立。

6 相关工作

极端多标签文本分类。与第 4 节中的讨论一致，我们将极端多标签分类的相关研究分为三大类。(1) 词袋分类器 [2, 16, 19, 31–34, 39, 49, 50, 54]以稀疏 tf–idf 特征作为输入。为提高模型效率，一对其余方法（如 DiSMEC [2] 和 PPDSparse [49] ）通过模型权重截断来探索并行化和模型规模缩减。另一方面，基于树的方法在较大的标签空间上采用各种划分技术。例如，Parabel [33] 使用二均值聚类将标签划分为平衡的树结构；Bonsai [19]通过允许多路且不平衡的划分来改进 Parabel；XR-Linear [54] 通过结合各种困难负采样方案来改进 Parabel；AnnexML [39] 通过基于图的近邻索引划分标签。(2) 基于序列的分类器 [21, 44, 45, 47, 53, 59] 采用卷积神经网络（例如 XML-CNN [21]）、循环神经网络（例如 MeshProbeNet [45] 和 AttentionXML [53]）以及 Transformer（例如 BertXML [44]）等深度神经网络架构，学习输入文本序列的语义表示以进行分类。还有一些研究聚合浅层词嵌入和/或应用多层感知机层来获取文档嵌入，例如 Slice [15], 、DeepXML [8], 、DECAF [29], 、GalaXC [36], 和 ECLARE [30]。(3)预训练语言模型分类器 [4, 17, 48, 52, 56, 60] 提出将预训练语言模型从网络规模语料库中学到的知识迁移到分类任务中。例如，X-Transformer [4] 将 BERT [10], 、XLNet [46], 或 RoBERTa [25] 的输出与稀疏 tf–idf 特征结合；LightXML [17] 采用预训练语言模型作为文本编码器，并使用同一个预训练语言模型进行标签短名单筛选和重新排序；XR- Transformer [56] 进一步提出快速多分辨率预训练语言模型微调。尽管这些模型在极端多标签分类中取得了成功，但它们主要关注对纯文本序列进行分类，对文档元数据的利用较少。相比之下，我们的工作提出了几种简单直接的方法，利用元数据信号增强文本分类器。

科学文献标注。将学术论文分类是文本挖掘 [6, 28, 58] 和图挖掘 [11, 14]研究中常见的评估任务。然而，大多数研究只考虑粗粒度的论文分类（例如，标签空间中有 5–20 个类别），其结果细分程度不足，无法满足用户对细粒度分类的需求。为了使用细粒度医学主题词表对 PubMed 上的论文进行标注，MeSH 索引任务已得到广泛研究 [9, 22, 31, 45, 52]。然而，这些模型仅将文本或文本+场所用作输入，尚未探索作者和参考文献的影响。近期，Zhang 等人 [59, 60] 和 Ye 等人 [47] 利用元数据为 MAG 中的论文标注研究领域。不过，这些研究仅限于计算机科学和生物医学领域。相比之下，我们的研究针对 19 个领域和三种主要类型的分类器开展了系统性研究。

7 结论与未来工作

在这项工作中，我们使用三种分类器，考察了元数据对19个领域科学文献标注的影响。我们的结果为致力于构建精准科学文献标注器的从业者提供了以下见解：第一，尽管以往的元数据感知方法通常会直接使用所有可用类型的元数据，但我们发现，并非所有元数据始终都有益。应根据分类器的类型、所属领域以及预测标签的粒度级别，选择有用的元数据特征。第二，尽管文本序列建模的最先进模型一直由基于Transformer的模型主导，但我们证明，在大规模细粒度论文标注的许多情况下，简单的词袋分类器表现相当，甚至更好，而且在利用不同类型的元数据方面可能更有效。第三，尽管不同元数据类型在不同领域中的影响各异，但每种元数据类型带来的收益或损失在相似领域之间相当一致。

本研究分别探讨每种元数据类型，以避免混杂因素。然而，不同类型的元数据（例如场所和作者）可能会相互作用，并为分类提供额外线索。未来，我们有兴趣研究多种元数据类型的综合效应。

致谢

我们感谢匿名审稿人提出宝贵且富有见地的意见。研究部分由 IBM-Illinois Discovery Accelerator Institute、美国 DARPA KAIROS 计划 No. FA8750-19-2-1004 和 INCAS计划 No. HR001121C0165、美国国家科学基金会 IIS-19-56151、IIS-17-41317 和 IIS 17-04532，以及 Molecule Maker Lab Institute 提供支持；该研究所是 NSF 根据资助编号 2019897 支持的 AI Research Institutes 计划的一部分。研究还由 NSF 根据资助编号 2118329 支持的 Institute for Geospatial Understanding t hrough an Integrative Discovery Environment (I-GUIDE) 提供支持。本文所表达的任何意见、研究结果、结论或推荐均属于作者所有，并不一定代表美国国防高级研究计划局或美国政府明示或暗示的观点。

A 附录

附录A.1 数据集

作为表1的补充，表5汇总了 Maple 中更多的数据集统计信息，包括每篇论文的作者、参考文献和标签平均数量，以及训练、验证和测试论文的数量。

表5：Maple中20个数据集的更多统计数据。

字段	#作者/论文	#参考文献/论文	#标签/论文	#训练论文	#验证论文	#测试论文
Art	1.314	2.813	2.425	39,901	9,975	8,497
哲学	1.117	8.121	3.715	40,696	10,174	8,426
地理学	3.625	31.793	2.231	44,710	11,178	17,995
商业	2.346	37.635	3.556	49,481	12,370	23,007
社会学	1.592	24.474	2.318	60,815	15,204	14,189
历史	1.218	4.405	1.938	80,830	20,208	12,109
政治《科学》	1.496	10.198	3.107	73,650	18,413	23,228
环境《科学》	4.301	34.156	2.138	70,479	17,620	35,846
经济学	1.966	26.608	4.961	119,309	29,827	29,534
工程学	3.055	19.703	4.987	179,804	44,951	45,251
心理学	3.805	41.201	5.032	245,288	61,322	66,344
计算机科学（会议）	3.389	17.417	6.089	136,322	34,081	92,990
计算机科学（期刊）	3.296	21.968	5.976	214,616	53,654	142,333
地质学	3.552	36.510	5.957	284,022	71,006	76,806
数学	2.151	19.068	6.433	338,454	84,613	67,484
材料《科学》	4.829	26.838	4.428	783,289	195,822	358,620
物理学	9.841	23.509	7.271	905,613	226,403	237,967
生物学	5.611	36.137	8.306	1,125,605	281,401	181,772
化学	4.293	28.382	6.288	1,252,531	313,133	284,292
医学	5.647	12.888	5.376	1,620,165	405,041	620,899

A.2 超参数

在20个数据集中的每一个数据集中，我们都移除了出现在少于5篇论文中的元数据实例。使用仅文本、文本+场所、文本+作者和文本+参考文献作为输入时，我们采用相同的超参数配置。下文将介绍各分类器的代码来源和超参数。

A.2.1 Parabel。[4] 我们移除了出现在少于5篇论文中的词。所有参数均采用默认设置。具体而言，线程数为$T = 1$；树的数量为 $t = 3$；叶节点中的最大标签数为 $m = 100$；预测时的束搜索宽度为 $B = 10$。

A.2.2 Transformer。[5] Transformer层数为1；注意力头数量为2； [CLS] 标记数量为 $C = 10$；嵌入维度为100；最大序列长度为500；批次大小为256。我们采用GloVe.6B.100d作为初始化词嵌入。对于（训练轮数，预热轮数），艺术、哲学和地理学采用（100，20），商业、社会学、历史和政治学采用（80，16），环境科学、经济学、工程学和计算机科学采用（60，12），心理学、地质学和数学采用（40，8），材料科学、物理学和生物学采用（20，4），化学和医学采用（15，3）。其他超参数均采用默认设置。

A.2.3 OAG-BERT。[6] 我们使用“oagbert-v2”作为预训练语言模型。经过预训练语言模型编码后，我们固定论文嵌入来训练 Parabel 分类器。Parabel 的所有参数均采用默认设置。

[4]http://manikvarma.org/code/Parabel/download.html [5] https://github.com/XunGuangxu/CorNet（我们使用了该 GitHub 仓库中的 BertXML 分类器。尽管该模型名为 BertXML，但它从头开始训练 Transformer 架构，并未使用 BERT 初始化。）[6]https://github.com/THUDM/OAG-BERT

A.3 关于元数据在不同标签粒度下影响的更多内容

作为表4的补充，表6展示了在第1层 P@1、第2层 P@1 和第 3 层 P@1 方面，从利用作者或参考文献信息中获益最多的字段。我们发现，在许多情况下，作者和参考文献对艺术、哲学和历史字段都有帮助。可能的原因是，这些字段中的每篇论文所含的作者和参考文献数量都较少（根据表5中的统计数据）。因此，作者列表或参考文献列表可能包含较少的混淆信号，且更能指示主题。

表6：在第 1 层 P@1、第 2 层 P@1 和第 3 层 P@1 方面，最能从利用作者或参考文献信息中受益的字段。“T”：仅文本。“+A”：文本+作者。“+R”：文本+参考文献。“Δ”：绝对性能提升。

	L1 精确率@1			L2 精确率@1			L3 精确率@1	
Parabel [33]								
数学	T	0.5960	政治《科学》	T	0.7928	哲学	T	0.6118
	+A	0.5971		+A	0.7969		+A	0.6160
	Δ	0.11%		Δ	0.41%		Δ	0.42%
心理学	T	0.6992	商业	T	0.6704	商业	T	0.6124
	+R	0.7086		+R	0.6801		+R	0.6265
	Δ	0.94%		Δ	0.97%		Δ	1.41%
Transformer [44]								
历史	T	0.5471	哲学	T	0.6728	Art	T	0.4859
	+A	0.5489		+A	0.6900		+A	0.4910
	Δ	0.18%		Δ	1.72%		Δ	0.51%
历史	T	0.5471	Art	T	0.7859	历史	T	0.4857
	+R	0.5537		+R	0.7953		+R	0.4890
	Δ	0.66%		Δ	0.94%		Δ	0.33%
OAG-BERT [24]								
Art	T	0.6241	Art	T	0.6747	历史	T	0.3455
	+A	0.6254		+A	0.6875		+A	0.3819
	Δ	0.13%		Δ	1.28%		Δ	3.64%

A.4 元数据对模型效率的影响

下面报告元数据对模型效率的影响。表7展示了在20个数据集中分别加入发表场所、作者和参考文献后，三种分类器训练时间的平均相对增幅。所有模型均在 Intel Xeon E5-2680 v2 @ 2.80GHz 上运行，并使用一块 NVIDIA GeForce GTX 1080 Ti GPU（如果需要 GPU）。可以观察到，添加参考文献作为特征后，训练时间显著增加。当 Parabel 作为分类器时，增幅源于用于表示论文的词袋向量大幅变长；当 Transformer 作为分类器时，增幅则是由于增加了大量参数（即参考文献嵌入），导致模型收敛速度变慢。

表7：加入一种类型的元数据后，20个数据集中训练时间的平均相对增幅。

	Parabel [33]	Transformer [44]	OAG-BERT [24]
+场所	+0.15%	+0.13%	+0.11%
+作者	+0.72%	+1.86%	+0.11%
+参考文献	+22.68%	+10.75%	–

图5展示了三种分类器在20个数据集上的训练时间。报告的训练时间是20次运行的平均值（即，5次运行 ×{仅文本, 文本+场所, 文本+作者, 文本+参考文献}），适用于测试 Parabel 和 Transformer；若测试 OAG-BERT，则为15次运行的平均值（即，5次运行 ×{仅文本, 文本+场所, 文本+作者}）。在这三种分类器中，Parabel 在20个数据集上始终效率最高；在

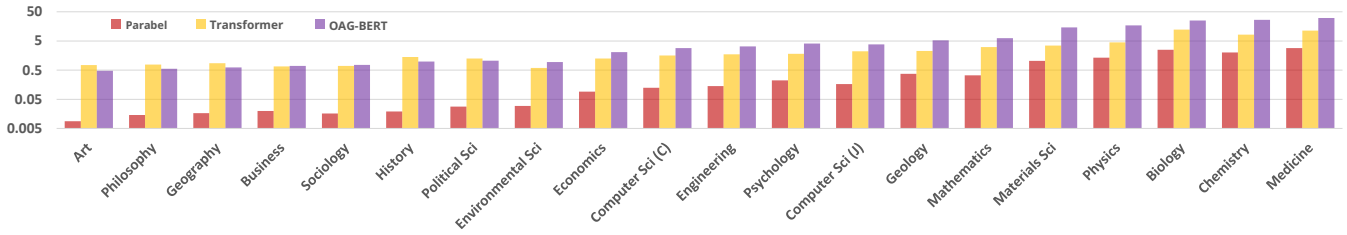


图5：三个分类器在20个数据集上的训练时间（小时）。

表8：三个带有 MeSH 标签的数据集的统计信息。

字段	论文来源	论文数	标签数	出版场所数	#作者	#参考文献	#作者/论文	#参考文献/论文	标签数/论文	#训练论文	#验证论文	#测试论文
生物学-MeSH	期刊	1,379,393	25,039	100	2,486,814	6,876,739	5.686	40.043	13.870	985,364	246,341	147,688
化学-MeSH	期刊	762,129	21,585	87	1,498,358	5,928,908	4.741	34.344	10.984	511,814	127,954	122,361
医学-MeSH	期刊	1,536,660	25,188	100	2,791,165	7,190,021	5.254	20.931	11.819	1,020,969	255,242	260,449

表 9：P@k 和 NDCG@k 在另外三个数据集上的得分。 蓝色，红色和“ ”表 2 相同。

字段	输入	Parabel [33]					Transformer [44]					OAG-BERT [24]				
		P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
生物学-MeSH	Text	0.8924	0.7964	0.7088	0.8194	0.7576	0.9112	0.8098	0.7193	0.8341	0.7698	0.7506	0.6185	0.5460	0.6490	0.5945
	+场所	0.8934	0.7976	0.7101	0.8206	0.7587	0.9119	0.8105	0.7194	0.8348	0.7701	0.7520	0.6200	0.5473	0.6504	0.5958
	+作者	0.8932	0.7985	0.7112	0.8212	0.7596	0.9090	0.8028	0.7093	0.8281	0.7614	0.7504	0.6184	0.5464	0.6488	0.5946
	+参考文献	0.8976	0.8058	0.7198	0.8279	0.7674	0.9079	0.7988	0.7034	0.8248	0.7565	–	–	–	–	–
化学-MeSH	Text	0.8447	0.7340	0.6407	0.7603	0.6947	0.8445	0.7113	0.6082	0.7427	0.6684	0.6971	0.5804	0.5099	0.6071	0.5557
	+场所	0.8453	0.7354	0.6421	0.7615	0.6960	0.8465	0.7151	0.6121	0.7460	0.6720	0.6995	0.5826	0.5132	0.6093	0.5587
	+作者	0.8450	0.7350	0.6419	0.7611	0.6957	0.8439	0.7105	0.6066	0.7419	0.6670	0.6981	0.5819	0.5126	0.6086	0.5580
	+参考文献	0.8490	0.7409	0.6491	0.7667	0.7023	0.8248	0.6802	0.5743	0.7139	0.6366	–	–	–	–	–
医学-MeSH	Text	0.9673	0.8410	0.7454	0.8712	0.8075	0.9683	0.8567	0.7583	0.8842	0.8191	0.7343	0.6229	0.5629	0.6491	0.6089
	+场所	0.9673	0.8422	0.7472	0.8722	0.8090	0.9688	0.8585	0.7606	0.8856	0.8210	0.7370	0.6235	0.5607	0.6503	0.6081
	+作者	0.9671	0.8342	0.7424	0.8656	0.8039	0.9687	0.8495	0.7496	0.8786	0.8117	0.7347	0.6185	0.5549	0.6459	0.6027
	+参考文献	0.9666	0.8382	0.7463	0.8687	0.8073	0.9662	0.8470	0.7458	0.8762	0.8084	–	–	–	–	–

小型数据集上，Transformer 耗时比 OAG-BERT 更长，但在中型和大型数据集上情况则相反。

A.5 带有医学主题词表标签的其他数据集

为进一步增强我们的研究发现，我们构建了三个以医学主题词表（MeSH）术语 [7] 作为标签的数据集，这些术语由美国国家医学图书馆的专家精心整理。具体而言，我们从 Maple 中选取生物学、化学和医学这三个数据集，并获取每篇论文的真实医学主题词标签[7]。移除没有医学主题词标签的论文后，我们得到三个新数据集：生物学-MeSH、化学-MeSH 和医学-MeSH。其统计数据见表8。

我们在这三个新数据集上运行 Parabel、Transformer 和 OAG-BERT。运行 Transformer 时，三个数据集的（训练轮数，预热轮数）均设为（20，4）。运行 OAG-BERT 时，我们需要将出现在论文文本中的标签排在未出现的标签之前。请注意，一个医学主题词标签可能有多个名称（即一个规范名称以及零个、一个或多个入口词；参见医学主题词标签“COVID-19”[8]）。给定一个医学主题词标签，如果其任一名称出现在论文文本中，我们就视为该标签出现在论文中。其他所有超参数均与附录 A.2 相同。

Parabel、Transformer 和 OAG-BERT 在三个新数据集上的 P@k 和 NDCG@k 得分见表 9。总体而言，我们仍然发现，在几乎所有情况下，发表场所都有助于科学文献标注，而作者和参考文献的影响则取决于分类器的类型和字段。

另外三个数据集也可在我们的Maple基准：https://doi.org/10.5281/zenodo.7611544。

[7]https://learn.microsoft.com/en-us/academic-services/graph/reference-data-schema#paper-mesh [8]https://meshb.nlm.nih.gov/record/ui?ui=D000086382

参考文献

[1] Waleed Ammar、Dirk Groeneveld、Chandra Bhagavatula、Iz Beltagy、Miles Craw- ford、Doug Downey、Jason Dunkelberger、Ahmed Elgohary、Sergey Feldman、Vu Ha 等。2018。Semantic Scholar 文献图谱的构建。载于 *NAACL-HLT'18*。84–91。[2] Rohit Babbar 和 Bernhard Schölkopf。2017。Dismec：用于极端多标签分类的分布式稀疏机器。载于 *WSDM'17*。721–729。[3] Iz Beltagy、Kyle Lo 和 Arman Cohan。2019。SciBERT：一种面向科学文本的预训练语言模型。载于 *EMNLP'19*。3615–3620。[4] Wei-Cheng Chang、Hsiang-Fu Yu、Kai Zhong、Yiming Yang 和 Inderjit S Dhillon。2020。驾驭预训练 Transformer，实现极端多标签文本分类。载于 *KDD'20*。3163–3171。[5] Junyoung Chung、Caglar Gulcehre、KyungHyun Cho 和 Yoshua Bengio。2014。门控循环神经网络在序列建模中的实证评估。*arXiv:1412.3555* (2014)。[6] Arman Cohan、Sergey Feldman、Iz Beltagy、Doug Downey 和 Daniel S Weld。2020。SPECTER：利用引文感知 Transformer 进行文档级表示学习。载于 *ACL'20*。2270–2282。[7] Margaret H Coletti 和 Howard L Bleich。2001。用于检索生物医学文献的医学主题词。*JAMIA* 8, 4 (2001), 317–323。[8] Kunal Dahiya、Deepak Saini、Anshul Mittal、Ankush Shaw、Kushal Dave、Akshay Soni、Himanshu Jain、Sumeet Agarwal 和 Manik Varma。2021。Deepxml：一种应用于短文本的深度极端多标签学习框架。载于 *WSDM'21*。31–39。[9] Suyang Dai、Ronghui You、Zhiyong Lu、Xiaodi Huang、Hiroshi Mamitsuka 和 Shanfeng Zhu。2020。FullMeSH：利用全文改进大规模医学主题词表索引。*Bioinformatics* 36, 5 (2020), 1533–1541。[10] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向 Transformer 预训练。载于

NAACL-HLT'19。4171–4186。[11] Yuxiao Dong、Nitesh V Chawla 和 Ananthram Swami。2017。metapath2vec：面向异构网络的可扩展表示学习。载于 KDD'17。135–144。[12] Yuxiao Dong、Hao Ma、Zhihong Shen 和 Kuansan Wang。2017。科学的百年历程：科学合作、引文与创新的全球化。载于 KDD'17。1437–1446。[13] Jorge E Hirsch。2005。一种量化个人科学研究产出的指标。PNAS 102, 46 (2005), 16569–16572。[14] Ziniu Hu、Yuxiao Dong、Kuansan Wang 和 Yizhou Sun。2020。异构图Transformer。载于 WWW'20。2704–2710。[15] Himanshu Jain、Venkatesh Balasubramanian、Bhanu Chunduri 和 Manik Varma。2019。Slice：在一亿个标签上训练的可扩展线性极端分类器，用于相关搜索。载于 WSDM'19。528–536。[16] Himanshu Jain、Yashoteja Prabhu 和 Manik Varma。2016。用于推荐、标注、排序及其他缺失标签应用的极端多标签损失函数。载于 KDD'16。935–944。[17] Ting Jiang、Deqing Wang、Leilei Sun、Huayi Yang、Zhengyang Zhao 和 Fuzhen Zhuang。2021。Lightxml：采用动态负采样的 Transformer，实现高性能极端多标签文本分类。载于 AAAI'21。7987–7994。[18] Ching Jin、Yifang Ma 和 Brian Uzzi。2021。科学奖项与科学主题的非凡增长。Nature Communications 12, 1 (2021), 5619。[19] Sujay Khandagale、Han Xiao 和 Rohit Babbar。2020。Bonsai：用于极端多标签分类的多样化浅层树。Machine Learning 109, 11 (2020), 2099–2119。[20] Jinhyuk Lee、Wonjin Yoon、Sungdong Kim、Donghyeon Kim、Sunkyu Kim、Chan Ho So 和 Jaewoo Kang。2020。BioBERT：一种用于生物医学文本挖掘的预训练生物医学语言表示模型。Bioinformatics 36, 4 (2020), 1234–1240。[21] Jingzhou Liu、Wei-Cheng Chang、Yuexin Wu 和 Yiming Yang。2017。极端多标签文本分类的深度学习。载于 SIGIR' 17。115–124。[22] Ke Liu、Shengwen Peng、Junqiu Wu、Chengxiang Zhai、Hiroshi Mamitsuka 和 Shanfeng Zhu。2015。MeSHLabeler：通过整合多种证据提高大规模医学主题词表索引的准确性。Bioinformatics 31, 12 (2015), i339–i347。[23] Weiwei Liu、Haobo Wang、Xiaobo Shen 和 Ivor W Tsang。2021。多标签学习的新兴趋势。IEEE TPAMI 44, 11 (2021), 7955–7974。[24] Xiao Liu、Da Yin、Ji ngnan Zheng、Xingjian Zhang、Peng Zhang、Hongxia Yang、Yuxiao Dong 和 唐杰。2022。OAG-BERT：迈向学术知识服务的统一骨干语言模型。载于 KDD'22。3418–3428。[25] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019。Roberta：一种稳健优化的 BERT 预训练方法。arXiv 预印本 arXiv:1907.11692 (2019)。[26] Zhiyong Lu。2011。PubMed 及其他工具：用于检索生物医学文献的万维网工具综述。Database 2011 (2011)。[27] Laurens van der Maaten 和 Geoffrey Hinton。2008。使用 t-SNE 可视化数据。JMLR 9 (2008), 2579–2605。[28] Dheeraj Mekala、Xinyang Zhang 和 Jingbo Shang。2020。META：元数据赋能的文本分类弱监督。载于 EMNLP'20。8351–8361。[29] Anshul Mittal、Kunal Dahiya、Sheshansh Agrawal、Deepak Saini、Sumeet Agarwal、Purushottam Kar 和 Manik Varma。2021。Decaf：融合标签特征的深度极端分类。载于 WSDM'21。49–57。[30] Anshul Mittal、Noveen Sachdeva、Sheshansh Agrawal、Sumeet Agarwal、Purushottam Kar 和 Manik Varma。2021。ECLARE：结合标签图相关性的极端分类。载于 WWW'21。3721–3732。[31] Shengwen Peng、Ronghui You、Hongning Wang、Chengxiang Zhai、Hiroshi Mamitsuka 和 Shanfeng Zhu。2016。DeepMeSH：用于改进大规模医学主题词表索引的深度语义表示。Bioinformatics 32, 12 (2016), i70–i79。[32] Yashoteja Prabhu、Anil Kag、Shilpa Gopinath、Kunal Dahiya、Shrutendra Harsola、Rahul Agrawal 和 Manik Varma。2018。利用标签特征进行极端多标签学习，以实现热启动标注、排序和推荐。载于 WSDM'18。441–449。[33] Yashoteja Prabhu、Anil Kag、Shrutendra Harsola、Rahul Agrawal 和 Manik Varma。2018。Parabel：用于极端分类的分区标签树及其在动态搜索广告中的应用。载于 WWW'18。993–1002。[34] Yashoteja Prabhu 和 Manik Varma。2014。Fastxml：一种用于极端多标签学习的快速、准确且稳定的树分类器。载于 KDD'14。263–272。[35] Martin Rosvall 和 Carl T Bergstrom。2011。网络随机游走的多层压缩揭示了大型综合系统中的层次组织。PLOS ONE 6, 4 (2011), e18209。[36] Deepak Saini、Arnav Kumar Jain、Kushal Dave、Jian Jiao、Amit Singh、Ruofei Zhang 和 Manik Varma。2021。GalaXC：用于极端分类的逐标签注意力图神经网络。载于 WWW'21。3733–3744。[37] Zhihong Shen、Hao Ma 和 Kuansan Wang。2018。用于科学知识探索的网络规模系统。载于 ACL'18 系统演示。87–92。[38] Arnab Sinha、Zhihong Shen、Yang Song、Hao Ma、Darrin Eide、Bo-June Hsu 和 Kuansan Wang。2015。微软学术服务（MAS）概述及其应用。载于 WWW'15 Companion。243–246。[39] Yukihiro Tagami。2017。AnnexML：用于极端多标签分类的近似最近邻搜索。载于 KDD'17。455–464。[40] 唐杰、Jing Zhang、Limin Yao、Juanzi Li、Li Zhang 和 Zhong Su。2008。Arnet-miner：学术社交网络的抽取与挖掘。载于 KDD'08。990–998。[41] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。注意力就是你所需要的一切。载于 NIPS'17。5998–6008。[42] Kuansan Wang、Zhihong Shen、Chiyuan Huang、Chieh-Han Wu、Yuxiao Dong 和 Anshul Kanakia。2020。微软学术图谱：专家不足以胜任之时。《定量科学研究》1，1（2020），396–413。[43] Boya Xie、Zhihong Shen 和 Kuansan Wang。2021。预印本是科学的未来吗？在线预印本服务的三十年历程。arXiv 预印本 arXiv:2102.09066（2021）。[44] Guangxu Xun、Kishlay Jha、Jianhui Sun 和 Aidong Zhang。2020。用于极端多标签文本分类的相关网络。载于 KDD'20。1074–1082。[45] Guangxu Xun、Kishlay Jha、Ye Yuan、Yaqing Wang 和 Aidong Zhang。2019。MeSHProbeNet：用于 MeSH 索引的自注意力探测网络。《Bioinformatics》35，19（2019），3794–3802。[46] Zhilin Yang、Zihang Dai、Yiming Yang、Jaime Carbonell、Russ R Salakhutdinov 和 Quoc V Le。2019。XLNet：用于语言理解的广义自回归预训练。NeurIPS'19。[47] C henchen Ye、Linhai Zhang、Yulan He、Deyu Zhou 和 Jie Wu。2021。超越文本：利用异构图整合元数据和标签结构，以进行多标签文档分类。载于 EMNLP' 21。3162–3171。[48] Hui Ye、Zhiyu Chen、Da-Han Wang 和 Brian Davison。2020。用于极端多标签文本分类的预训练广义自回归模型与自适应概率标签簇。载于 ICML'20。10809–10819。[49] Ian EH Yen、Xiangru Huang、Wei Dai、Pradeep Ravikumar、Inderjit S Dhillon 和 Eric Xing。2017。PPDSparse：一种用于极端分类的并行原始-对偶稀疏方法。载于 KDD'17。545–553。[50] Ian En-Hsu Yen、Xiangru Huang、Pradeep Ravikumar、Kai Zhong 和 Inderjit S Dhillon。2016。Pd-sparse：一种用于极端多类别与多标签分类的原始与对偶稀疏方法。载于 ICML'16。3069–3077。[51] Yian Yin、Yuxiao Dong、Kuansan Wang、Dashun Wang 和 Benjamin F Jones。2022。科学的公共使用与公共资助。《自然·人类行为》6，10（2022），1344–1350。[52] Ronghui You、Yuxuan Liu、Hiroshi Mamitsuka 和 Shanfeng Zhu。2021。BERTMeSH：利用全文进行大规模高性能 MeSH 索引的深度上下文表征学习。《Bioinformatics》37，5（2021），684–692。[53] Ronghui You、Zihan Zhang、Ziye Wang、Suyang Dai、Hiroshi Mamitsuka 和 Shanfeng Zhu。2019。AttentionXML：用于高性能极端多标签文本分类的基于标签树的注意力感知深度模型。NeurIPS'19（2019），5820–5830。[54] Hsiang-Fu Yu、Kai Zhong、Jiong Zhang、Wei-Cheng Chang 和 Inderjit S Dhillon。2022。23，98（2022），1–32。PECOS：面向海量相关输出空间的预测。《JMLR》。[55] Fanjin Zhang、Xiao Liu、唐杰、Yuxiao Dong、Peiran Yao、Jie Zhang、Xiaotao Gu、Yan Wang、Bin Shao 等。2019。OAG：构建大规模异构实体图的探索。载于 KDD'19。2585–2595。[56] Jiong Zhang、Wei-Cheng Chang、Hsiang-Fu Yu 和 Inderjit S Dhillon。2021。用于极端多标签文本分类的快速多分辨率 Transformer 微调。载于 NeurIPS'21。7267–7280。[57] Yu Zhang、Xiusi Chen、Yu Meng 和 Jiawei Han。2021。弱监督下的层次化元数据感知文档分类。载于 WSDM'21。770–778。[58] Yu Zhang、Shweta Garg、Yu Meng、Xiusi Chen 和 Jiawei Han。2022。MotifClass：利用高阶元数据信息进行弱监督文本分类。载于 WSDM'22。1357–1367。[59] Yu Zhang、Zhihong Shen、Yuxiao Dong、Kuansan Wang 和 Jiawei Han。2021。MATCH：大型层次结构中的元数据感知文本分类。载于 WWW'21。3246–3257。[60] Yu Zhang、Zhihong Shen、Chieh-Han Wu、Boya Xie、Junheng Hao、Ye-Yi Wang、Kuansan Wang 和 Jiawei Han。2022。用于零样本多标签文本分类的元数据诱导的对比学习。载于 WWW'22。3162–3173。