DOI：10.13998/j.cnki.issn1002-1248.22-0662

新冠领域溯源类论文筛选及全文实体标注研究

徐　硕[1]，张萌萌[2]，柳力元[2]，王聪聪[1]，孙　睿[2]，李怡琳[2]，徐金楠[2]，安　欣[2*]

（1. 北京工业大学 经济与管理学院，北京 100124；2. 北京林业大学 经济管理学院，北京 100083）

摘　要：［目的／意义］新冠病毒出现以来，国内外与新冠病毒研究相关的论文迅猛增长。整理国内外 COVID-19 相关学术论文，创建针对新冠溯源类论文的数据集和细粒度的实体数据集能为新冠病毒的起源和传播机理等相关研究提供坚实的数据支撑。［方法／过程］提出基于主动学习模型的论文筛选方法，从海量论文中高效精准地定位与新冠溯源相关的论文。同时，设计了一种新冠领域 18 类实体的标注方案，不仅包含生物领域通有的基因、蛋白质和化合物等实体，还涵盖新冠领域特有的冠状病毒、野生动物等实体。［结果／结论］构建了一个新冠溯源类论文数据集，共包含 885 篇文章；基于提出的实体标注方案，标注全文本论文 99 篇，构建了一个细粒度的实体数据集，包含 39 118 个实体，是目前新冠领域规模最大、最全面的实体标注数据集。

关键词：新冠病毒；数据收集；SARS-CoV-2 起源；文档筛选；实体标注

中图分类号：G255.51　　**文献标识码：**A　　**文章编号：**1002-1248（2023）01-0087-12

引用本文：徐硕，张萌萌，柳力元，等. 新冠领域溯源类论文筛选及全文实体标注研究[J]. 农业图书情报学报，2023，35（1）：87-98.

1 引　言

2019 年 12 月，新型冠状病毒（SARS-CoV-2）导致 COVID-19 暴发，目前仍然在许多国家和地区肆虐。为打好新冠肺炎疫情防控全球阻击战，相关科研工作者积极投入研究工作[1]。以 *The New England Journal of Medicine*（NEJM）、*Lancet*、*The Journal of the American Medical Association*（JAMA）和 *The British Medical Journal*（The BMJ）为代表的各大期刊纷纷开设新型冠状病毒专题，开放绿色通道，加速同行评议，提前网络发布，采取各种手段为新冠病毒相关科研成果的快速发表提供帮助。

新冠相关研究涉及病毒起源、传播、治疗及社会

收稿日期：2022-09-07

基金项目：国家自然科学基金项目“基于全文本的微观实体扩散机制研究”（72004012）；北京工业大学 2022 年度“研究生思政教育进科研团队——抗疫专项探索项目”

作者简介：徐硕（1979-），男，博士，教授，博士生导师，研究方向为科学计量学、科技情报分析和数据挖掘等。张萌萌（1997-），女，硕士，研究方向为科技情报分析。柳力元（1998-），女，硕士，研究方向为知识扩散。王聪聪（1999-），女，硕士，研究方向为科学计量学。孙睿（1996-），女，硕士研究生，研究方向为知识扩散。李怡琳（1998-），女，硕士，研究方向为科学计量学。徐金楠（1998-），女，硕士，研究方向为科学计量学

*通信作者：安欣（1980-），女，博士，教授，硕士生导师，研究方向为科学计量学和数据挖掘等。Email：anxin@bjfu.edu.cn

DOI：10.13998/j.cnki.issn1002-1248.22-0662

影响等多个领域，其中寻找 SARS-CoV-2 的起源最为关键。但解决新冠溯源问题是艰难且复杂的。创建专门的新冠溯源类论文数据集能为科研人员提供数据基础，方便后续研究的开展。但目前规模最大的数据集 CORD-19[2]并未对论文进行分类，Lit-COVID[3]只是将论文分为一般信息、机理、传播、诊断、治疗、预防、案例报告和预测等 8 个大类，缺乏针对新冠溯源研究的论文数据集。其次，现有数据集主要涉及题目、摘要、作者等信息，并未对全文本内容进行深入分析和挖掘。尤其是全文本中提到的领域实体信息，这不利于新冠病毒的起源和传播过程及模式等相关研究的顺利开展。

基于此，本文提出基于主动学习的论文筛选方法，从海量论文中高效精准的筛选与新冠溯源相关的论文，构建新冠溯源类论文数据集；其次，本文制定实体标注方案，对新冠领域的全文本内容进行实体标注，创建新冠领域的细粒度的实体数据集，为新冠病毒的产生和传播机理等相关研究提供数据支撑。

2 相关研究

2.1 新冠相关数据集

根据数据集格式和应用，可将新冠相关的数据集分为 3 类：时间序列数据集、社会媒体数据集和知识库数据集。时间序列数据集可用于预测对人类社会的影响，其结构相对简单，通常包含时间、确诊病例数、死亡统计、人口流动量等字段。如 XU 等[4]从国家及省级卫生报告和在线报告中收集实时病例数据，构建涵盖地理编码数据（旅行史、症状和日期信息）的数据集。约翰·霍普金斯大学系统科学与工程中心发布约翰·霍普金斯流行病学数据集，包含病例报告和时间序列汇总表[5]。社会媒体数据集多用于行为决策、情感等社会科学研究，其信息类型丰富，有文本、图像和视频等多种形式。如密苏里大学收集并整理 1 亿多条推文构建 SARS-CoV-2 推特数据集[6]，数据集有推文创建时间、内容和用户 ID 等 6 个字段。杨崇洛等[7]利用中国各地发布的疫情新闻等结构化数据，采用多种神经网络模型拼接的方法抽取新冠相关的实体及实体关系，该数据集包含 9 类实体及 11 类实体关系，其实体包含患者、地点、机构和联系方式等信息。

表 1 列举了 COVID-19 相关的论文和知识库类数据集。CORD-19[2]包含冠状病毒的历史和最新科学研究论文，包含论文题目、摘要、期刊、年份等字段信息。Lit-Covid[3]除了基本的论文信息外，按照不同的研究主题和地理位置对新冠论文进行分类，并且每天更新。除了文本数据集外，知识图谱的形式能让科研人员清晰分析个体间的联系和区别。DOMINGO-FERNÁNDEZ 等[8]基于 COVID-19 相关论文构建知识图谱，知识图谱包含 SARS-CoV-2 病毒蛋白、潜在药物靶点等信息。ZHANG 等[9]基于 SARS-CoV-2 的生物过程、药物靶点、基因和蛋白质等构建知识图谱。

表 1 部分 COVID-19 知识库数据集详细信息

Table 1 Detailed information of part of the COVID-19 knowledge datasets

数据集	类型	规模/篇	主要内容
CORD-19[2]	论文信息	1 056 660	COVID-19 相关
Lit-COVID[3]	论文信息	308 046	COVID-19 相关
COVID-19kg1[8]	知识图谱	41 609	流行病学相关
COVID-19kg2[9]	知识图谱	339 638	COVID-19 相关

综上，目前已有多种形式的 COVID-19 相关数据集，但随着新冠相关研究的不断深入，仍然缺乏溯源类论文数据集，同时目前主要针对论文的题目和摘要进行分析，而全文本数据包含的丰富信息尚未被挖掘。

2.2 文档筛选方法

文档筛选是指尽可能查找与给定主题相关的所有相关文档，在信息检索领域被称为总召回问题（The Total Recall Problem）[10]。具体来说，这个问题可被描述如下：给定一组候选文档（其中只包括小部分相关文档），每个候选文档都可以被核查以确定其是否为相关文档，最终目标是核查尽可能少的候选文档但达到非常高的召回率。

自 COUNSELL 的开创性研究工作[11]以来，论文中已经提出了许多方法。它们在许多领域中都有很好的

应用，包括循证医学[12]和软件工程[13]中的系统评价，法律诉讼中的电子证据筛选[14]等。此外，TREC[15]和CLEF eHealth[16]竞赛类任务进一步推动了文档筛选的发展。总体来说，论文中涉及信息检索和机器学习两个主要研究分支。

在信息检索领域，相关研究可进一步分为3类：相关反馈[17]、查询扩展[18]和排名学习[19]。前两种方法侧重于转换或改进原始查询。主要区别在于，相关反馈致力于收集代表用户需求的信息并自动创建新检索式，而查询扩展则使用同义词或语义相关术语重新构造给定的检索式，以匹配其他相关文档。排名学习方法是对所有文档进行排序，以便尽可能多地将相关文档排在不相关文档之前。

事实上，文档筛选也可以被视为一个二元分类问题（相关或非相关）。从理论上来说，任何用于文本分类的监督机器学习模型都可以直接用于文档筛选，如朴素贝叶斯[20]、支持向量机[22]、随机森林[22]等。然而，由于相关和非相关实例的严重不平衡、标注耗时且工作量巨大，许多监督模型的性能欠佳。

近年来，基于主动学习的分类方法被引入到文档筛选中[23]。该方法的主要思想是，如果允许监督模型从候选文档中选择学习的实例，则监督模型可以在标注数据较少的情况下达到不错的分类性能。大量研究表明[15,16]，基于主动学习的文档筛选方法在许多实际应用中都有不错的性能表现。新冠溯源问题的复杂性，获得大量的标注样本的财务和时间成本巨大，本文采用主动学习模型筛选与新冠溯源相关的论文，能在减少财务和时间成本的同时，获得更好的模型性能。

2.3　领域实体标注

传统的实体标注多以新闻数据为主，随着科技论文资源的迅速增长，从大量论文中提取领域实体信息变得越来越重要。高质量实体语料库的构建是许多文本挖掘任务的基础，且语料库的规模和质量对算法的性能和表现至关重要。目前，语料库的建设大多是面向特定领域，比如司法领域[24]、国防科技领域[25]、语言学领域[26]和医学领域[27]等。尤其在生物医学领域，随着BioCreative 和 BioNLP 会议等评测生物医学语料库任务的发布，学者们构建了多个用于解决特定任务的语料库。

根据实体类别可将语料库分为多种类型（表2），有标注基因及基因家族的 GNormPlus[28]、NLM-gene[29]语料库，标注化合物及药物的 CHEMDNER[30,31]、BC5CDR[32]和 NLM-Chem[33]语料库，标注生物体的 Species[34]和 Cancer Genetics（CG）[35]语料库。也有部分语料库标注了多个类型的实体，如 CRAFT[36]语料库，涉及基因 / 蛋白质、物种、细胞和化学物质等 9 类实体。其次，语料库的标注规模和范围也各有差异，大部分语料库仅标注了题目和摘要部分。NLM-Chem 语料库是近期鲜有的基于全文本的实体语料库，共标注了 150 篇全文本论文的化合物相关实体，实体提及数量达 38 342 个，去重后的实体有近 5 000 个。

目前，新冠领域论文量呈爆炸式增长，但尚没有基于全文本的实体数据集。因此，本文针对新冠领域

表 2　部分生物医学领域语料库详细信息

Table 2　Detailed information of part of the biomedical corpus

语料库名称	标注规模/篇	标注范围	实体类别	标记数量/个	实体数量/个
CRAFT	97	full-text	基因/蛋白质、物种、细胞和化合物等 9 类	~970 000	~14 000
Species	800	abstracts	生物物种	3 708	718
Cancer Genetics（CG）	600	abstracts	解剖学、分子实体和生物体	21 683	/
GNormPlus	694	abstract、text snippets	基因、基因家族和蛋白质结构域	10 640	1 996
CHEMDNER	10 000	abstracts	化合物和药物	84 355	19 805
BC5CDR	1 500	title、abstracts	化合物和疾病	31 901	13 343
NLM-gene	550	abstracts	基因	15 950	5 500
NLM-Chem	150	full-text	化合物	38 342	4 867

DOI：10.13998/j.cnki.issn1002-1248.22-0662

构建一个高质量的实体数据集，以便推动相关研究工作的顺利开展。

综上，本文将基于主动学习方法从海量论文中筛选与新冠溯源相关的论文，该方法仅需要较少的标注样本便能达到更好的分类效果，可大大缓解传统分类方法对大量标注数据的要求。另外，以往生物领域所标注的领域实体多为化合物、基因 / 蛋白质、生物体等类型，且标注范围局限在论文题目和摘要中，本文结合新冠领域，添加冠状病毒、野生动物等实体类型，且基于全文本进行实体标注，更能充分概括该领域论文的研究内容。

3 数据及研究框架

3.1 CORD–19 数据集

CORD-19 数据集[2]主要考虑了 7 种数据源：WHO、PubMed、Medline、Elsevier、bioRxiv、medRxiv 和 arXiv，其中后 3 种为预印本论文库。所采用的检索式为："COVID" OR "COVID-19" OR "Coronavirus" OR "Corona virus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome" OR "2019 novel coronavirus" OR "2019 novel CoV" OR "coronavirus 2019" OR "new human coronavirus" OR "SARS-CoV-2"。截至 2021 年 12 月 13 日，该数据集共包括 864 554 篇科技论文，是目前新冠病毒研究领域最全面的数据集，故本文后续研究主要以 CORD-19 数据集为基础。

表 3 为每种数据源论文分布情况，由于部分论文同时被两种或两种以上数据源收录，所以所有数据源论文的占比总和大于 100%。从表 3 可以看出，在 7 个主要数据源中，来自 WHO 数据源的论文最多，其次是 Medline 和 PMC，其余 4 个数据源的论文占比均小于 10%。

3.2 研究框架

如图 1 所示，本文在 CORD-19 数据集的基础上，基于论文的题目和摘要等信息，提出基于主动学习的溯源类论文筛选方法，并根据筛选结果构建新冠溯源论文数据集。另外，基于论文的全文本内容，确定新冠领域的实体类别，并建立标注指南，由标注团队完成两轮的标注，最后根据标注结果构建新冠实体数据集。

4 新冠溯源类论文资源构建

4.1 数据提取

鉴于新冠病毒 SARS-CoV-2 溯源的困难性和复杂性，新冠病毒溯源论文非常稀少，这使得正负样本数据严重失衡。为了减轻领域专家的工作量，本文按以下步骤提前准备了一个具有类别信息的种子数据集。

（1）种子数据集的收集。种子数据集构建的总体思路是先确定小部分种子文章，然后根据这些种子文章的前向和后向引用进行扩展。“世卫组织召集的 SARS-CoV-2 全球溯源研究：中国部分”[37]和关于 SARS-CoV-2 起源的综述类文章[38-42]提供了有价值的线索。本文基于这 6 篇种子文章以及它们的前向和后向引用文章，利用以下 3 条规则从 282 种期刊共收集 470 篇文章：①文章发表时间为 2019 年 12 月及以后；②文章的主题与 COVID-19 相关；③仅限于经过同行

表 3 每种数据源论文分布情况

Table 3 Literature distribution of each data source

数据源	论文量/篇	占比/%	数据源	论文量/篇	占比/%
PMC	304 230	35.19	ArXiv	11 499	1.33
Medline	364 912	42.21	MedRxiv	17 436	2.02
BioRxiv	7 355	0.85	Elsevier	71 313	8.25
WHO	472 502	54.65			

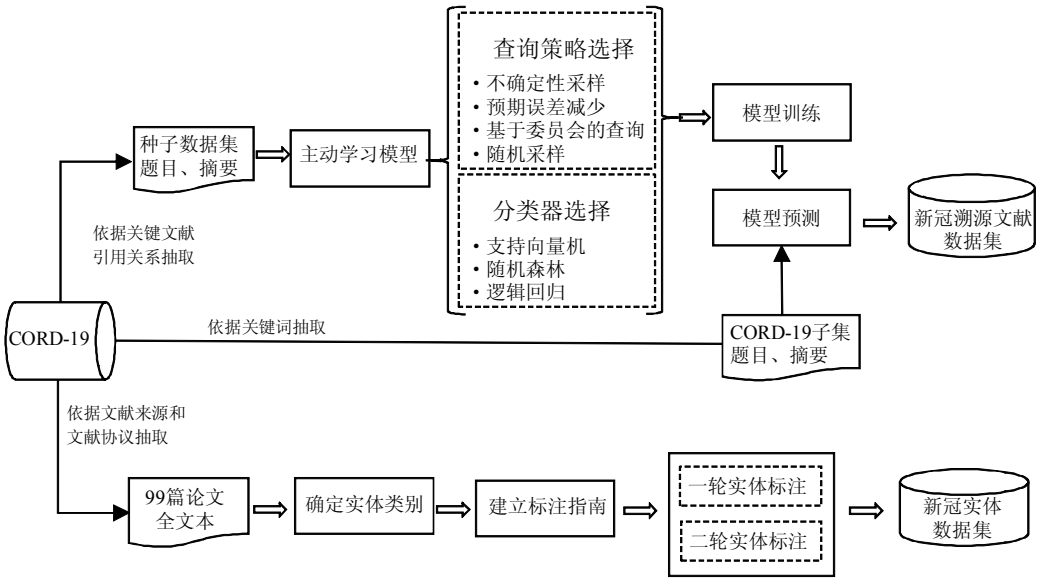


图 1　研究框架图

Fig.1　Research framework for selecting the papers on the origins of COVID-19 and annotating entities based on full texts

评议的文章。

（2）种子数据集的标注。构建好种子数据集后，需对每篇文章“是否与 SARS-CoV-2 起源相关”进行标注，以便进行主动学习。与 SARS-CoV-2 起源相关的文章标为“1”，否则标为“0”。该工作由两名生物学知识背景的研究生完成，整个标注过程主要分为 2 步。为了获得高质量的标注结果，先由两位标注人员分别标注相同的 50 篇文章，采用 multi- κ 指标[43]计算其一致性得分为 80.2%。经过讨论发现两名标注人员在研究中间宿主（狗、猫、水貂等）的文章上意见相左。经过深入讨论，我们认为这些文章与 SARS-CoV-2 起源密切相关，统一标注为“1”。第二步，其余文章随机分为大小相同的两组，分别由两名标注人员独立进行标注。最终，种子数据集包括 170 篇相关文章（正实例）和 300 篇非相关文章（负实例）。

（3）种子数据集的文档表示。CORD-19 每次更新时会同步发布 768 维的文档嵌入表示，这种嵌入表示是利用 SPECTER（Scientific Paper Embedding Using Citation-Informed TransforERs）方法在标题、摘要和引用网络的基础上得到。因此，本文通过 DOI 映射获得种子数据集的文档表示。

（4）提取待筛选论文。CORD-19 数据集涵盖了多种流行病相关学术文章，而本节的目的在于筛选新冠溯源论文，因此为加快筛选速度，根据以下两个步骤从 CORD-19 数据集提取出与新冠病毒相关的文章：①提取题目或摘要包含“COVID-19”“2019-nCoV”“SARS-CoV-2”或“coronavirus2019”关键词的文章；②剔除种子数据集中的文章。最终，待筛选论文有 371 664 篇，并将其命名为“CORD-19 子集”。

4.2　筛选步骤

由于带有类别信息的种子数据集已提前准备好，因此实验是通过模拟标注过程进行。也就是说，通过查询策略所选择的文档标签被假定为事先未知，并且需要在主动学习过程中由领域专家进行标注。如图 2 所示，筛选过程的输入包括初始标注集、未标注数据集、查询策略、基分类器和 CORD-19 子集，其中初始标注集、未标注数据集共同由种子数据集组成。筛选主要由以下 3 部分组成。

（1）首先用具有标注信息的训练集初始化分类器；

（2）用查询策略从未标注数据集中选择一个实例，交由领域专家进行人工标注，然后把该实例从未标注数据集中删除并添加到训练集，用更新后的重新训练分类器。重复此步骤，直到分类器达到最佳性能；

DOI：10.13998/j.cnki.issn1002-1248.22-0662

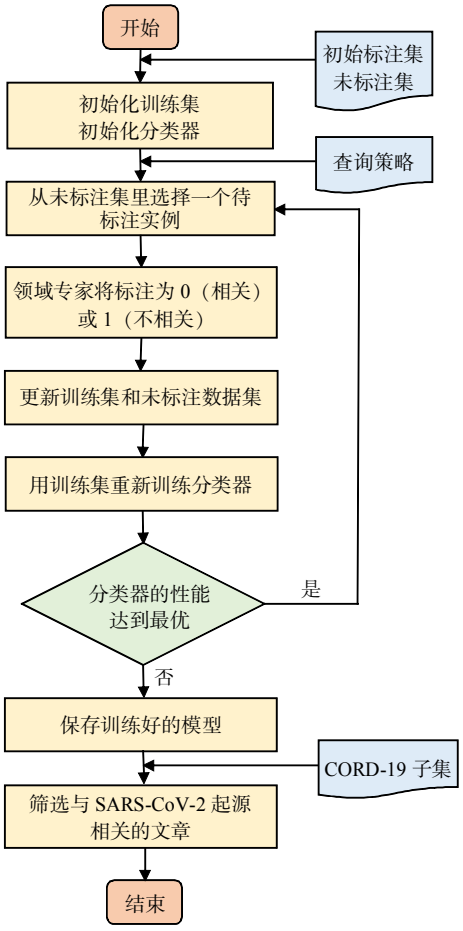


图 2　筛选过程流程图

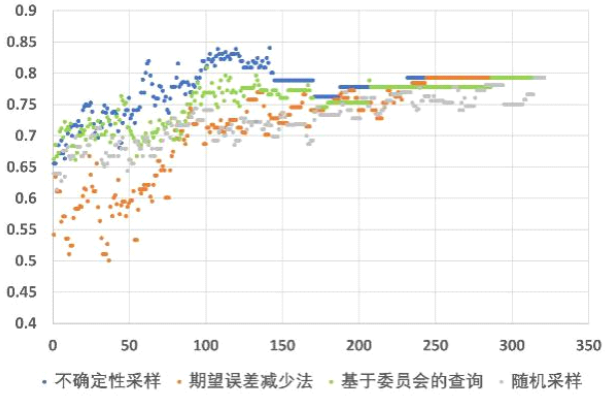
Fig.2　Screening process flow chart

（3）最后，利用训练好的分类器从 CORD-19 子集中筛选与 SARS-CoV-2 起源相关的文章。

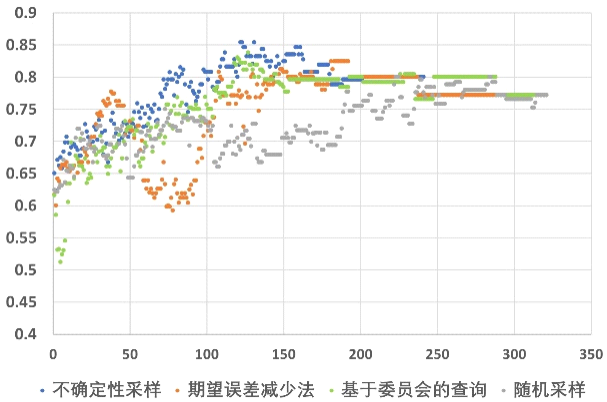
4.3　实验过程及结果

本文考虑了 3 种基分类器：支持向量机（SVM）、逻辑回归（LR）和随机森林（RF），同时考虑了 4 种查询策略：不确定性采样（Uncertain Sampling）、预期误差减少（Expected Error Reduction）、基于委员会的查询（Query by Committee）和随机采样（Random Sampling）。对于每种基分类器，本文优化了查询策略，如图 3 所示。可以看出，不确定性采样查询策略性能表现最好，因此该查询策略后续被用于筛选 SARS-CoV-2 起源相关的学术文章。

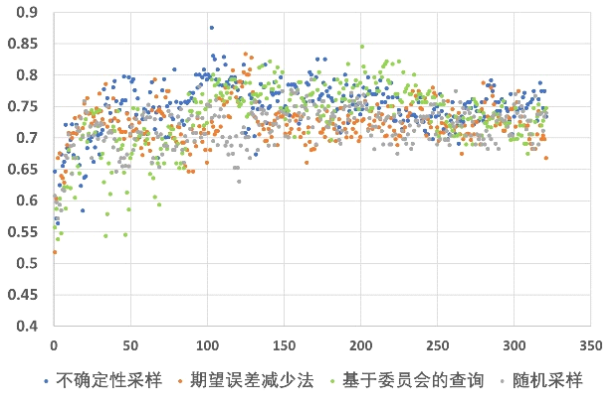
SVM、LR 和 RF 基分类器可为每个待筛选文章分配一个后验概率，基于后验概率，本文人工逐个检查



（a）支持向量机（SVM）



（b）逻辑回归（LR）



（c）随机森林（RF）

图 3　基于（a）支持向量机，（b）逻辑回归和（c）随机森林基础分类器的主动学习方法的性能比较（F1 值）

Fig.3　Performance comparison of active learning methods in term of F1 score on the basis of base classifiers: (a) support vector machine, (b) logistic regression and (c) random forest

了每个基本分类器输出的前 1 000 篇文章，如图 4 所示。在前 1 000 篇文章中，SVM、LR 和 RF 三个基分

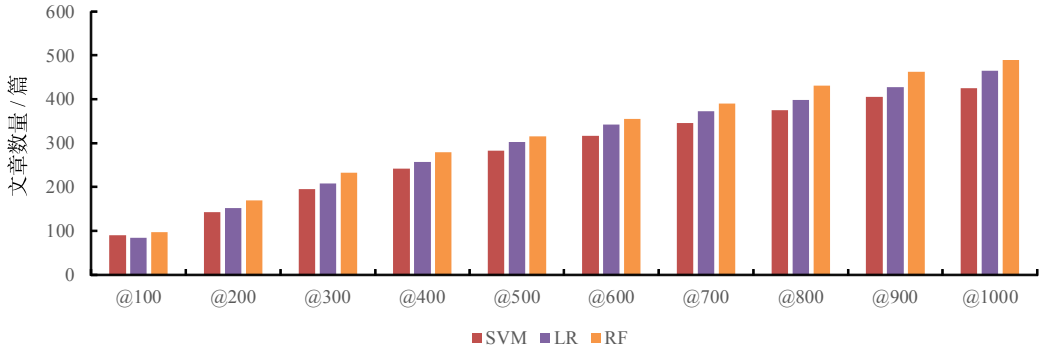


图 4　SARS-CoV-2 起源相关文章筛选性能评估

Fig.4　Evaluation of screening performance of articles related to the origin of SARS-CoV-2

类器分别能够正确筛选 425 篇、465 篇和 489 篇文章。由于 3 个基分类器筛选出的文章有很大比例的重叠，最终共筛选出来 715 篇与 SARS-CoV-2 溯源相关的文章。加上训练集中的 170 篇标注为“1”的文章，最终一共得到 885 篇与 SARS-CoV-2 溯源相关的文章。

5　新冠领域全文实体标注

5.1　实体类别

本文按照论文来源和授权访问协议对 CORD-19 数据集的论文进行分组，等比例随机抽取 99 篇论文进行后续标注。通过参阅大量的论文和前期的研究工作（见 2.3 节），除了关注基础生物实体（基因、疾病、化学、突变等），还重视与 COVID-19（如冠状病毒）相关的实体。基于此，本文确定 18 类需要标注的微观实体，实体定义参照 UMLS 语义网络及维基百科，如表 4 所示。

5.2　标注指南及标注过程

参考 NLM-Chem 标注指南[33]和 bc5 CDR 标注指标[31]，本文制定了新冠领域实体标注指南，该指南主要包括 3 部分：（a）实体类别及定义；（b）实体标注的一般规则；（c）不应该被标注为实体的情况。

为了加快标注进程，并尽可能减少标注错误的出现，本文采用了可视化标注工具 BRAT [45]进行实体标注工作。标注团队由一名管理员和六名标注人员组成，实体标注由两轮组成。第一轮是 6 个标注人员独立标注相同的 99 篇文章，标注过程会参考 UMLS 语义网络上的实体分类以及谷歌、百度、必应等网站。第一轮完成后，每个文档的所有标注结果都由管理员合并，以识别相同的标注和不同的标注，并使用 multi-κ 一致性指标计算了标注结果的一致性分数[37]。如图 5 所示，有 70%左右的文章一致性分数聚集在 0.4~0.7 之间。其中，一致性分数在 0.5~0.6 区间的文章最多（30 篇），其次是 0.6~0.7（21 篇）和 0.4~0.5（20 篇）。需要注意的是，如果一篇文章有标注人员标了部分实体，但有标注人员未标注任何实体时，会导致一致性分数小于 0。在 99 篇文章中，有 5 篇文章的一致性分数小于 0。

第二轮是针对实体标注有差异的文章，由管理员将这些文章单独列出，6 位标注人员讨论存在差异的部分，直到达成共识，至此完成标注。

5.3　标注结果

从表 5 的标注结果可以看出，基因蛋白类实体占比达 25.35%，数量最多，非冠状病毒类和化合物占比超过 10%，疾病、实验技术等实体分别占 8.48%和 7.04%，其余实体占比均小于 5%。

99 篇文档共标注了 39 118 个实体，平均每篇文章约有 395 个实体。依据每个实体的出现次数对实体进行排序，得到了每类实体中排名前 5 的实体。在表 6 中展示了实体分布超 5%的实体类别（除引用类，即化合物类、基因 / 蛋白 / 酶类、非冠状病毒类、疾病类和实验技术类），每个类别给出了排名前 5 的实体实例。这些实体从多个角度来了解体现了目前针对新冠病毒的学术研究

DOI：10.13998/j.cnki.issn1002-1248.22-0662

表 4 实体类型及实体定义表

Table 4 Entity types and entity definitions

ID	实体类型	定义	例子
1	冠状病毒	冠状病毒科的成员，可引起多种脊椎动物的呼吸道或胃肠道疾病	SARS-CoV-2, SARS-CoV, MERS-CoV
2	非冠状病毒	冠状病毒以外的病毒的成员	rotavirus, norovirus, adenoviruses
3	家畜	为家庭使用或盈利而饲养的家养农场动物	cattle, sheep, pig
4	野生动物	被认为是野生的或不适合家庭使用的动物	bat, pangolin, monkey
5	实验动物	用于或打算用于研究、测试或教学的动物	C57BL/6 mice
6	宠物	人类饲养的用于陪伴和享受的动物，而不是家畜	cat, dog
7	人	可能被细菌、病毒或其他生物感染的人。此类别不包括角色（患者、医生）和职称（教授、主任）	people, children
8	其他动物	家畜、野生动物、实验动物和宠物以外的动物	parasites
9	细菌	一种小型的、典型的单细胞原核微生物。它们是单细胞原核微生物，通常具有刚性细胞壁，通过细胞分裂进行繁殖，并表现出 3 种主要形式：圆形或球形、杆状或杆菌状以及螺旋状	streptococcus, enterococcus
10	身体器官	位于特定区域，或结合并执行生物体的一种或多种专门功能的细胞和组织的集合。其范围大到结构，小到复杂器官的组成部分。与组织相比，这些结构相对有明确的位置	heart, right ventricle, mitral valve,
11	基因/蛋白/酶等	该类别包括基因、蛋白质、酶、基因组和蛋白质组	CSHGs, ACE2
12	化合物	完全由有机或无机化学式定义的物质，包括其他化学物质的混合物	DMSO, amoxicillin, gentamicin
13	身体基质	细胞外物质或细胞和细胞外物质的混合物，由身体产生、排泄或增加	Serum, urine, sputum
14	材料	构成物理对象的有形物质	Silver
15	实验技术	用于确定样本的成分、数量或浓度的方法或技术	qRT-PCR
16	疾病或症状	一种改变或干扰生物体正常过程、状态或活动的现象。它通常以宿主的一个或多个系统、部分或器官的功能异常为特征。这里包括描述疾病的一系列症状	Aseptic meningitis, encephalitis
17	引用	信息来源或参考论文的简短说明	[1] (Wang et al., 2020)
18	参考	论文中的图表	Fig.1, Table 1

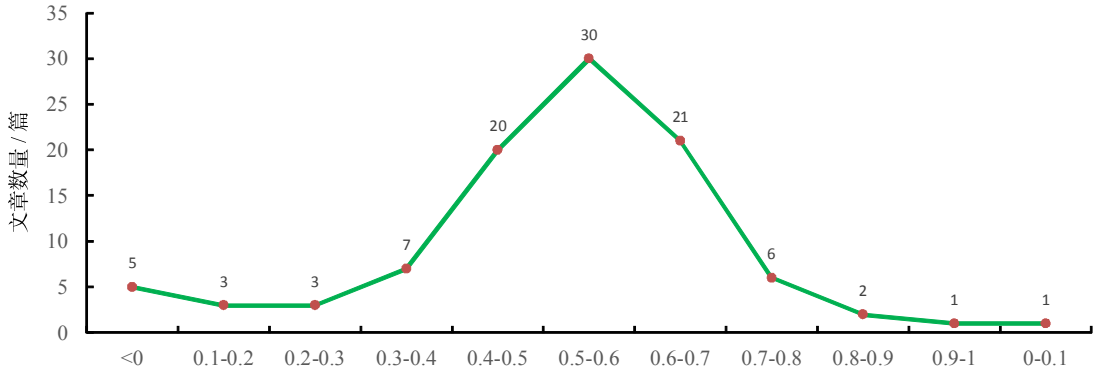


图 5 99 篇文章的一致性分数

Fig.5 A consensus score of 99 articles

情况，对比仅从文章的题目或摘要中得到的信息，实体中更丰富的信息扩展了我们对新冠相关研究的了解，对后续的与新冠主题相关的研究有较大的帮助。

6 结 语

COVID-19 的爆发扰乱了人们的日常生活和工作，

表 5　精标实体分布

Table 5　The distribution of annotated entities

序号	实体类别	实体数量/个	百分比/%	序号	实体类别	实体数量/个
1	基因/蛋白/酶等	9 917	25.35	11	细菌	665
2	引用	5 957	15.23	12	身体基质	599
3	非冠状病毒	4 128	10.55	13	冠状病毒	554
4	化合物	4 040	10.33	14	野生动物	543
5	疾病或症状	3 319	8.48	15	家畜	446
6	实验技术	2 754	7.04	16	其他动物	425
7	参考	1 856	4.74	17	材料	340
8	身体器官	1 594	4.07	18	宠物	36
9	实验动物	1 184	3.03			
10	人	761	1.95			

表 6　前 5 类实体类别中排名前 5 的实体分布

Table 6　Top 5 entities for each of Top 5 entity types

实体类别	实体	全称
化合物	PA	Patchouli alcohol（广藿香醇）
	Poly(I:C)	聚肌胞苷酸
	L-DOPA	左旋多巴
	PBS	phosphate buffer saline（磷酸缓冲盐溶液）
	OM	Oblongifolin M（长叶菌素 M）
基因/蛋白/酶等	parkin	Parkin 蛋白
	ACE2	Angiotensin-converting enzyme 2（血管紧张素转化酶 2）
	API5	apoptosis inhibitor protein 5（凋亡抑制蛋白 5）
	GPNMB	glycoprotein non-metastatic melanoma protein B（非转移性黑色素瘤糖蛋白 B）
	HACE1	HECT domain and Ankyrin repeat Containing E3 ubiquitin-protein ligase 1（E3 泛素连接酶）
非冠状病毒	TMV	Tobacco mosaic virus（烟草花叶病毒）
	IAV	Influenza A virus（甲型流感病毒）
	EBOV	Ebola virus（埃博拉病毒）
	HIV	Human Immunodeficiency Virus（艾滋病病毒）
	HIV-1	Human Immunodeficiency Virus-1（艾滋病病毒 1 型）
疾病或症状	pneumonia	肺炎
	influenza	流感
	GMH	Germinal Matrix Hemorrhage（生发性基质出血）
	Oxidative stress	氧化应激
	infectious diseases	传染性疾病
实验技术	PCR	polymerase chain reaction（聚合酶链反应）
	Western blotting	蛋白质印迹法
	unpaired Student's t test	非配对学生 t 检验
	TWIRLS	Topic-wise inference engine of massive biomedical literatures（海量生物医学论文的主题推理引擎）
	qPCR	Quantitative polymerase chain reaction（荧光定量 PCR 法）

DOI：10.13998/j.cnki.issn1002-1248.22-0662

为此，全球相关科研工作者积极投入研究工作，引起新冠领域的论文呈爆炸式增长。为加快新冠相关研究工作，国内外学者已经构建了多个新冠相关数据集，这些数据集大致可分为时间序列数据集、社会媒体数据集和知识库数据集3类。然而，新冠领域的研究涉及病毒起源、传播、治疗及社会影响等多个方面，目前尚未有针对新冠溯源的论文数据集，也没有基于全文本的领域实体数据集。为加快新冠相关研究的开展，本文基于主动学习方法筛选新冠溯源类同行评议的文章885篇，同时，提出一种新冠领域实体标注方案，涉及冠状病毒、非冠状病毒、野生动物等18类实体，完成了99篇全文本论文的标注工作，共计39 118个实体。

参考论文：

[1] MEI-HSIU-CHING H, LIU JOHN-S. The swift knowledge development path of COVID-19 research: The first 150 days[J]. Scientometrics, 2021, 126(3): 2391-2399.

[2] LUCY L, LO K, CHANDRASEKHAR Y, et al. CORD-19: The covid-19 open research dataset[J]. ArXiv: 2004.10706v4, 2020.

[3] CHEN Q, ALLOT A, LU Z. LitCovid: An open database of COVID-19 literature[J]. Nucleic acids res, 2021, 49(d1): D1534-D1540.

[4] XU B, GUTIERREZ B, MEKARU S, et al. Epidemiological data from the covid-19 outbreak, real-time case information[J]. Scientific data, 2021, 7(1): 1-6.

[5] DONG E, DU H, GARDNER L. An interactive web-based dashboard to track covid-19 in realtime[J]. The lancet infectious diseases, 2020, 20(5): 533-534.

[6] KABIR M, MADRIA S. Coronavis: A real-time covid-19 tweets analyzer[J]. ArXiv: 2004.10706v4, 2020.

[7] 杨崇洛, 生龙, 魏忠诚, 等. 新冠文本实体关系抽取及数据集构建方法研究[J/OL]. 计算机工程与应用: 1-9[2023-02-06]. http://kns.cnki.net/kcms/detail/11.2127.tp.20220622.1100.010.html.
YANG C L, SHENG L, WEI Z C, et al. Research on COVID-19 text entity relation extraction and dataset construction methods[J/OL]. Computer engineering and applications: 1-9 [2023-02-06]. http://kns.cnki.net/kcms/detail/11.2127.tp.20220622.1100.010.html.

[8] DOMINGO-FERNáNDEZ D, BAKSI S, SCHULTZ B, et al. Covid-19 knowledge graph: A computable, multi-modal, cause-and-effect knowledge model of covid-19 pathophysiology[J]. Bioinformatics, 2021, 37(9): 1332-1334.

[9] ZHANG R, HRISTOVSKI D, SCHUTTE D, et al. Drug repurposing for COVID-19 via knowledge graph completion[J]. Journal of biomedical informatics, 2021, 115: 103696.

[10] GROSSMAN M, CORMACK G, ROEGIEST A. TREC 2016 total recall track overview[C]. Proceedings of the twenty-fifth text retrieval conference, 2016: 15-18.

[11] COUNSELL C. Formulating questions and locating primary studies for inclusion in systematic reviews[J]. Annals of internal medicine, 1997, 127(5): 380-387.

[12] CARVALLO A, PARRA D, LOBEL H, et al. Automatic document screening of medical literature using word and text embeddings in an active learning setting[J]. Scientometrics, 2020, 125(3): 3047-3084.

[13] HASSLER E, HALE D, HALE J. A comparison of automated training-by-example selection algorithms for evidence based software engineering[J]. Information and software technology, 2018(98): 59-73.

[14] CORMACK G, GROSSMAN M. Evaluation of Machine-Learning protocols for technology-Assisted review in electronic discovery[C]. Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval, 2014: 153-162.

[15] ROEGIEST A, CORMACK G, GROSSMAN M, et al. TREC 2015 total recall track overview[C]. Proceedings of the 24th text REtrieval-conference (TREC 2015), 2015.

[16] KANOULAS E, LI D, AZZOPARDI L, et al. CLEF 2018 technologically assisted reviews in empirical medicine overview[C]. CEUR workshop proceedings, 2018: 10-14.

[17] DONOSO-GUZMáN I, PARRA D. An interactive relevance feedback interface for evidence-based health care[C]. The 23rd international conference on intelligent user interfaces, 2014: 103-114.

[18] WENG L, LI Z, CAI R, et al. Query by document via a decomposition-based two-level retrieval approach[C]. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval, 2011: 505-514.

[19] LEE G, SUN A. Seed-driven document ranking for systematic reviews in evidence-based medicine[C]. The 41st international

ACM SIGIR conference on research & development in information retrieval, 2018: 455–464.

[20] XU S. Bayesian na ve bayes classifiers to text classification[J]. Journal of information science, 2018, 44(1): 48–59.

[21] XU S, AN X, QIAO X, et al. Multi–task least–squares support vector machines[J]. Multimedia tools and applications, 2014, 71(2): 699–715.

[22] AN X, SUN X, XU S, et al. Important citations identification by exploiting generatiïe model into discriminative model[J]. Journal of information science, 2023, 49(1): 107–121.

[23] SETTLES B. Active learning literature survey[R]. University of Wisconsin–Madison, Madison, USA: Computer sciences technical report 1648, 2010.

[24] 沙九, 冯冲, 周鹭琴, 等. 面向司法领域的高质量开源藏汉平行语料库构建[J]. 中文信息学报, 2021, 35(11): 51–59.
SHA J, FENG C, ZHOU J Q, et al. Constraction of high–quality and open source Tibetan–Chinese parallel corpus judicial domain[J]. Journal of Chinese information processing, 2021, 35(11): 51–59.

[25] 冯鸾鸾, 李军辉, 李培峰, 等. 面向国防科技领域的技术和术语语料库构建方法[J]. 中文信息学报, 2020, 34(8): 41–50.
FENG L L, LI J H, LI P F, et al. Constructing a technology and terminology corpus oriented national defense science[J]. Journal of Chinese information processing, 2020, 34(8): 41–50.

[26] 刘妍, 熊德意. 面向小语种机器翻译的平行语料库构建方法[J]. 计算机科学, 2022, 49(1): 41–46.
LIU Y, XIONG D Y. Construction method of parallel corpus for minority language machine translation[J]. Computer science, 2022, 49(1): 41–46.

[27] 杨锦锋, 关毅, 何彬, 等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报, 2016, 27(11): 2725–2746.
YANG J F, GUAN Y, HE B, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records[J]. Journal of software, 2016, 27(11): 2725–2746.

[28] WEI C, KAO H, LU Z. GNormPlus: An integrative approach for tagging genes, gene families, and protein domains[J]. BioMed research international, 2015: 918710.

[29] ISLAMAJ R, WEI C, CISSEL D, et al. NLM–Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi–species gene recognition[J]. Journal of biomedical informatics, 2021, 118: 103779.

[30] KRALLINGER M, RABAL O, LEITNER F, et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles[J]. Journal of cheminformatics, 2015, 7(1): S2.

[31] XU S, AN X, ZHU L, et al. A CRF–based system for recognizing chemical entity mentions (CEMs) in biomedical literature[J]. Journal of cheminformatics, 2015, 7(1): S11.

[32] LI J, SUN Y, JOHNSON R, et al. BioCreative v CDR task corpus: A resource for chemical disease relation extraction[J]. Database (oxford), 2016: Baw068.

[33] ISLAMAJ R, LEAMAN R, KIM S, et al. NLM–Chem, a new resource for chemical entity recognition in PubMed full text literature[J]. Scientific data, 2021, 8(1): 91.

[34] PAFILIS E, FRANKILD S, FANINI L, et al. The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text[J]. PLoS one, 2013, 8(6): E65390.

[35] PYYSALO S, OHTA T, ANANIADOU S. Overview of the cancer genetics (cg) task of bionlp shared task 2013[C]. Proceedings of the BioNLP shared task 2013 workshop, 2013: 58–66.

[36] BADA M, ECKERT M, EVANS D, et al. Concept annotation in the CRAFT corpus[J]. BMC bioinformatics, 2012, 13: 161.

[37] Joint WHO–China study team. WHO–convened global study of origins of SARS–CoV–2: China part [EB/OL].[2022–03–10]. https://www.who.int/publications/i/item/who–convened–global–study–of–origins–of–sars–cov–2–china–part.

[38] DOMINGO J. What we know and what we need to know about the origin of SARS–CoV–2[J]. Environmental research, 2021, 200: 111785.

[39] HOLMES E, GOLDSTEIN S, RASMUSSEN A, et al. The origins of SARS–CoV–2: A critical review[J]. Cell, 2021, 184(19): 4848–4856.

[40] VAN HELDEN J, BUTLER C, ACHAZ G, et al. An appeal for an objective, open, and transparent scientific debate about the origin of SARS–CoV–2[J]. Lancet, 2021, 398(10309): 1402–1404.

[41] KARLSSON E, DUONG V. The continuing search for the origins of SARS–CoV–2[J]. Cell, 2021, 184(17): 4373–4374.

[42] LEITNER T, KUMAR S. Where did SARS–CoV–2 come from[J].

DOI：10.13998/j.cnki.issn1002-1248.22-0662

Molecular biology and evolution, 2020, 37(9): 2463-2464.

[43] DAVIES M, FLEISS J. Measuring agreement for multinomial data[J]. Biometrics, 1982, 38(4): 1047-1051.

[44] COHAN A, FELDMAN S, BELTAGY I, et al. SPECTER: Document-level representation learning using citation-informed transformers[C]. Proceedings of the 58th annual meeting of the association for computational linguistics, 2020: 2270-2282.

[45] PONTUS S, SAMPO P, GORAN T, et al. BRAT: A web-based tool for NLP-assisted text annotation[C]. Proceedings of the demonstrations at the 13th conference of the European chapter of the association for computational linguistics, 2012: 102-107.

[46] WANG X, SONG X, LI B, et al. Comprehensive named entity recognition on CORD-19 with distant or weak supervision[J]. ArXiv: 2003.12218v5, 2020.

Selection of Papers on the Origins of COVID-19 and Entity Annotation Based on Full Texts

XU Shuo[1], ZHANG Mengmeng[2], LIU Liyuan[2], WANG Congcong[1], SUN Rui[2], LI Yilin[2], XU Jinnan[2], AN Xin[2*]

(1. School of Economics and Management, Beijing University of Technology, Beijing 100124;

2. School of Economics and Management, Beijing Forestry University, Beijing 100083)

Abstract: [Purpose/Significance] Since the outbreak of COVID-19, there has been a rapid increase in the number of studies related to COVID-19 at home and abroad. Review of relevant literature on COVID-19 provides data resources for related research on the emergence and transmission mechanism of SARS-CoV-2. However, the current COVID-19 related dataset is a collection of the literature, without classifying the data for each subfield, and the coarse-grained information such as the title and author fails to provide an in-depth understanding of the progress of COVID-19 research. Therefore, this paper created a dataset for the COVID-19 sub-domain and a fine-grained entity dataset. [Method/Process] Firstly, this paper proposed a literature screening method based on active learning model, which can obtain more valuable marker samples with less labor cost, so that the classifier has better generalization performance. We considered three base classifiers: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF), while considering four query strategies: uncertainty sampling, expected error reduction, committee-based query, and random sampling. Taking the origin of SARS-CoV-2, one of the sub-fields related to SARS-CoV-2, as an example, articles related to the origin of SARS-CoV-2 were efficiently and accurately located from the literature. At the same time, this paper designed a labeling scheme covering 18 types of entities, including not only genes, proteins, compounds and other entities that are universal in the biological field, but also corona viruses and wild animals that are unique to the field of SARS-CoV-2. In this paper, visual annotation tool BRAT was used for entity annotation. The tagging team consisted of an administrator and six annotators, and the entity tagging consisted of two rounds. What's more, multi-k consistency index was used to calculate the consistency score of annotation results. [Results/Conclusions] The results of the active learning model show that the uncertain sampling query strategy has the best performance. SVM, LR and RF based on uncertain sampling can correctly screen 425, 465 and 489 articles, respectively. After the removal of overlapping articles, a dataset related to the origin of SARS-CoV-2 was constructed, containing a total of 885 articles. Secondly, based on the proposed entity labeling scheme, 6 annotators completed 99 papers. Based on the results of fine marking, this paper constructed an entity dataset containing 39,118 entities, which is the largest and most comprehensive entity corpus in the field of COVID-19.

Keywords: SARS-CoV-2; data collection; origins of SARS-CoV-2; document screening; entity annotation