arXiv:2305.04055v1 [cs.DL] 6 May 2023

科学技术本体：新兴主题分类体系*

MahenderKumar[1], *, RubyRani[1], Mirko Bottarelli[1], Gregory Epiphaniou[1] 和 Carsten Maple[1]

[1]安全网络系统研究组，*WMG*，华威大学，考文垂，英国，*CV4 7AL*

摘要

本体在语义网技术中发挥着关键作用，它以结构化、标准化的方式表示知识，使机器能够理解数据的含义。尽管已有多种分类体系和本体问世，但它们通常只针对单一领域，而且其中一些需要耗费大量时间和人工投入。此外，这些体系还需要覆盖更多非常规主题，从而更全面、整体地呈现知识格局和跨学科协作。因此，我们需要一种覆盖科学与技术的本体，通过连接不同字段和领域中可能相关或存在共性的主题，促进多学科研究。为解决这些问题，我们提出了一种自动构建的科学技术本体（S&TO），涵盖科学与技术不同领域中的非常规主题。所提出的 S&TO 有助于发现新的研究领域，并促进跨学科协作。该本体通过对从 Semantic Scholar 收集的 2021年10月 至 2022年8月 期间的 393,991 篇科学文章数据集应用 BERTopic 构建，涵盖四个科学字段。目前，S&TO 包含 5,153 个主题和 13,155 个语义关系。通过对更新的数据集运行 BERTopic，可以更新 S&TO 模型。

关键词

科学技术本体、非常规主题、BERTopic、科学知识图谱

1. 引言

本体是一种宝贵的工具，可利用领域内的一组概念、关系和规则，表示和组织特定主题或一组主题的知识 [1, 2]。本体有许多应用，包括数据标注和可视化 [3], 预测新的研究领域 [4], 以及学术数据发现 [5]。不同领域创建的一些主题本体包括 ACM 计算分类系统 [1]、物理学与天文学分类方案（PACS）[2]，该方案于2016年由物理学主题标题（PhySH）取代 [3], 数学

文本知识图谱生成国际研讨会（*TEXT2KG*），与扩展语义……同期举行 *tic*
网络会议（*ESWC 2023*）日期：*2023*年*5*月*28*日至*6*月*1*日，地点：希腊赫索尼索斯

*mahender.kumar@warwick.ac.uk (M. Kumar); ruby.rani@warwick.ac.uk (R. Rani); mirko.bottarelli@warwick.ac.uk (M. Bottarelli); gregory.epiphaniou@warwick.ac.uk (G. Epiphaniou); CM@warwick.ac.uk (C. Maple) 0000-0001-7082-0475 (M. Kumar)

©2023 年本文作者版权所有。允许依据知识共享署名 4.0 国际许可协议（CC BY 4.0）使用。CEUR 研讨会论文集（CEUR-WS.org）

[1]ACM 计算分类系统：http://www.acm.org/publications/class-2012。[2]物理学与天文学分类方案：https://publishing.aip.org/publishing/pacs [3]PhySH - 物理学主题词：https://physh.aps.org/about。

学科分类（MSC）[4]，这是数学领域使用的分类体系，以及医学主题词（MeSH）[5]。创建这些大规模分类体系是一项复杂且成本高昂的工作，通常需要多位领域专家的专业知识，因此既耗时又耗费资源。因此，这些分类体系往往难以更新和维护，随着新信息和新发现不断涌现，很快就会过时。由此，这些分类体系的实用性和价值受到显著限制。本体生成领域最引人注目的进展之一，是开发出一种名为计算机科学本体（CSO）的大规模自动化本体 [6]。CSO 本体在计算机科学领域的研究主题表示方面实现了重大突破，提供了一个结构化且全面的框架，用于组织和整合知识，但其范围仅限于计算机科学概念。

研究挑战：理解非常规主题相关的动态变化颇具挑战，因为这些主题能够更全面、整体地呈现知识格局与跨学科合作。为这类非常规主题构建本体，需要从多个领域识别并收集关键概念及其关系。此外，非常规主题可能需要多学科研究，因此必须整合来自众多领域的信息。克服这些挑战后，研究人员和学者便有机会研究科学与技术领域的新兴发展方向，并促进不同领域之间的跨学科合作。

贡献。本文介绍了构建 S&TO本体的初步工作，该本体可自动生成非常规科学与技术主题的分类体系。S&TO本体通过对从 Semantic Scholar 收集的 2021年10月至 2022年8月期间的 393,991 篇科学文章数据集应用 BERTopic 构建，涵盖四个科学领域：计算机科学、物理学、化学和工程学。目前，S&TO 包含 5,153 个主题和 13,155 个语义关系。与现有本体不同，S&TO本体能够为知识表示与发现带来诸多益处，促进跨学科研究并支持动态更新。

组织结构。本文其余部分安排如下：第 2 节介绍用于构建所提出的 S&TO 的数据集和方法。第 3 节给出所提出的 S&TO。第 4 节讨论实验结果。第 5 节介绍 S&TO 的应用和用例，第 6 节讨论当前版本的局限性。最后，第 7 节给出结论。

2. 数据与方法

2.1. Semantic Scholar

Semantic Scholar 收录了来自各个领域的大量学术出版物，包括医学、农业、地球科学、生物医学文献和计算机科学。我们使用 RESTful Semantic Scholar 学术图谱（S2AG）API 检索了这些文章中的一个样本 [7]。该

[4]2010 数学学科分类：https://mathscinet.ams.org/msc/msc2010.html。[5]MeSH - 医学主题词：https://www.nlm.nih.gov/mesh。

API 可按需向用户提供有关作者、论文、标题、引文、发表 venue 等信息。我们通过 S2AG API 从 Semantic Scholar 获取了 393,991 篇科学与技术领域的文章。该 API 提供了可靠的数据源，允许用户直接链接到 semanticscholar.org 上的相关页面，因此成为检索学术论文信息的一种便捷且易于使用的方式。

2.2. 方法



图1：数据流

下载数据集后，我们使用 BERTopic 方法从文章中提取主题——一些涉及多学科的文章需要作为离群值纳入。为了纳入这些非常规文章并降低离群值比例，我们在主题聚类过程中调整了 BERTopic 的参数 [8]。表3列出了用于分类体系生成的关键 BERTopic 参数。

如图1所示，我们提出的主题建模工作流包含五个重要步骤：句子嵌入、降维、聚类、主题质量评估和主题表示。尤其是句子嵌入阶段，需将文本输入转换为能够捕捉文本内在语义的数值向量。随后对这些向量进行降维，以降低其维度并提升聚类过程的有效性。聚类是将相似向量组合起来，从而生成与相关材料对应的连贯簇的过程。向量化可确保提取的主题具有高质量，而主题表示则为每个主题生成易于理解的摘要。总之，我们结合 BERTopic 的优势与审慎的参数优化，提供了一种稳健有效的方法，能够从海量异质数据集中提取有意义的非常规主题，并确保获得最佳结果。

1. **句子嵌入**：在分析输入文章之前，我们首先将其转换为数值表示。为此，我们使用了句子转换器，这是 BERTopic 使用的默认嵌入模型。该模型可以确定不同文档之间的语义相似度。默认情况下，BERTopic 提供了许多预训练模型，其中我们尝试了以下两个：“`all-MiniLM-L6-v2`“ 和 “`paraphrase-MiniLM-L12-v2`“。虽然有多种句子嵌入模型可供选择，但在本研究中，我们选用了“`paraphrase-MiniLM-L12-v2`“模型。该模型能够有效平衡性能与速度，非常符合我们的需求。因此，借助句子转换器与 BERTopic 的结合，我们可以有效地将文本数据转换为数值形式，并从大量且多样化的数据集中获得相关洞见。

表 1BERTopic 参数

参数	值
嵌入参数	
嵌入_模型	句子模型_
UMAP 参数	
n_邻居	2
n_成分	5
低_memory	true
HDBSCAN 参数	
5min_簇 _size	10
度量	欧几里得
簇 _选择 _方法	eom
预测_数据	true
最小_样本	1
CountVectoriser 参数	
词汇表	词汇表
Min_df	10
类别 TF-IDF 参数	
前_个_词	30
详细输出	True
最小值_主题_大小	20
向量化器 _模型	CountVectorizer
低_memory	True
计算_概率	假
多样性	0.4

表 2用于减少离群值的 BERTopic 聚类模型选择

特征	HDBSCAN	K-means	HDBSCAN 与概率
主题质量	Good	Less	Good
离群值	High	No	Low
遗漏风险	Less	少量	最少
非常规主题			

2. **降维**：聚类可能很复杂，因为嵌入通常具有高维度。为解决这一问题，嵌入的维度通常会降低到更实用的水平。我们采用 UMAP（统一流形逼近与投影）技术，在一个流形中表示局部和全局高维特征。

低维域 [9]。“n_neighbors”和“n_components”是 UMAP 方法中的两个重要参数。这些参数对生成的簇的大小有相当大的影响。特别是，这些因素取值较大时，会形成更多重要的簇。通过仔细调整这些参数，我们获得了最佳聚类结果，并从输入数据中提取出重要主题。

3. **聚类**：BERTopic 在降维过程完成后，将输入数据划分为嵌入相似的多个簇。聚类技术的准确性直接影响生成主题的质量。K-means [10], 基于层次密度的含噪声空间聚类（HDBSCAN） [11], 和凝聚聚类[12] 是 BERTopic 提供的聚类技术。表2总结了这些聚类算法的优缺点，重点强调它们生成高质量主题、处理离群点比例以及降低遗漏非常规主题风险的能力。HDBSCAN 是一种基于密度的聚类算法，用于在数据集中查找密度各异的簇。它通过根据数据点的密度构建簇的层级树来实现这一目标。首先，它识别密度最高的点，并围绕这些点形成一个簇。随后，它逐步将密度较低的点添加到该簇中，直到达到自然截断点，表明该簇的结束。根据我们的研究结果，将预测_数据参数设置为“True”的 HDBSCAN 是最佳选择。我们的方法有效地平衡了上述各项因素，使我们能够从大型复杂数据集中获取有意义且有价值的见解。

4. **向量化**：CountVectorizer 技术将文本文档转换为短语频率向量。然而，该方法存在明显的缺点，例如无法考虑某个特定主题在文章中的相对相关性。为了解决这一问题，我们采用了 C-TF-IDF（基于类别的 TF-IDF），这是一种经典 TF-IDF（词频-逆文档频率）方法的变体，它根据词项与特定文档类别的相关性为其分配权重。通过调整 BERTopic 中的参数，我们可以利用结合 C-TF-IDF 的 CountVectorizer 优化聚类过程，从而微调模型的性能。该方法使我们能够生成质量更高的主题，这些主题更具吸引力，也与输入数据更加相关。

5. **主题表示**：BERTopic 可以调整 TF-IDF，使其在簇级别而非文档级别发挥作用，从而从词袋矩阵中获得主题的具体表示。这种修改后的 TF-IDF 称为 c-TF-IDF。对于类别 c 中的词 x，c-TF-IDF 值为：

$$w_{x,c} = |tf_{x,c}| \times log(1 + A/f_x) \tag{1}$$

其中，$tf_{x,\ c}$ 表示单词 x 在类 c 中出现的频率，f_x 表示单词 x 在所有类中的出现频率，A 表示每类的平均单词数。

3. 科学与技术本体生成

为了创建主题网络（也称为知识图谱），需要使用 Semantic Scholar 提供的元数据。主题本体的构建涉及以下成分的定义：

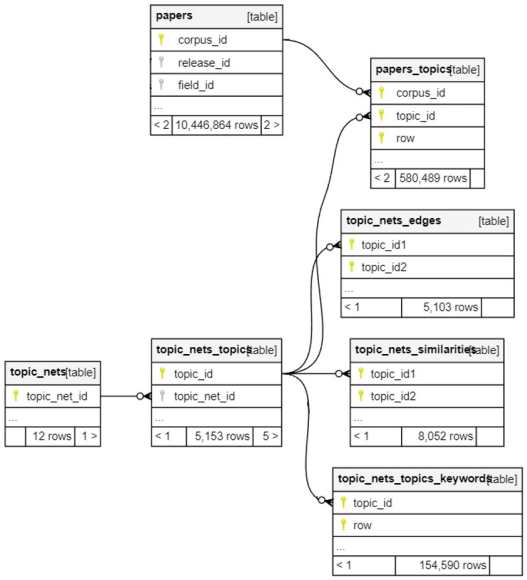


图 2： 主题提取和主题网络创建的架构

- 主题：主题本体中的概念（例如体育、艺术、政治）。
- 谓词：定义本体概念之间所建立语义链接的关系类型。在主题本体中可以定义许多谓词：层级化谓词（例如 superTopicOf）和非层级化谓词（例如，属于、促成）。

- 关系：根据谓词及其所链接的元素集合，可以区分不同的关系。关系可用于刻画图中的路径，并表示为三元组（T1, P, T2），其中 T1 和 T2 表示主题，P 表示连接 T1 和 T2 的谓词。

3.1. 主题

KeyBERT 工具用于处理相关论文，以提取代表每篇文档中核心概念和主题的关键词，从而生成 BERTopic 的词汇表。随后，这些关键词被输入 BERTopic，由其生成完整的主题集合，以捕捉数据集中的总体主题。提取出的主题被保存到数据库的“topic_nets_topics”和“topic_nets_topics keywords”中（见图2）。每个主题的权重表示其作为主要关联主题的论文数量，体现了该主题在数据集中的相关性。

建立论文与主题之间关系的方法有两种：

1. 概率：首先，每篇论文的 BERTopic/HDBSCAN 概率都会作为条目保存到“paperstopics”表中。这些概率表示每篇文档与提取出的各个主题的相关程度。2. 嵌入相似度：其次，利用“主主题 ID”字段（即 BERTopic 为每篇论文识别出的主要主题），通过 SQL

触发器将其直接关联到论文。这样便能高效查询和分析语料库中的主题及其相关论文。

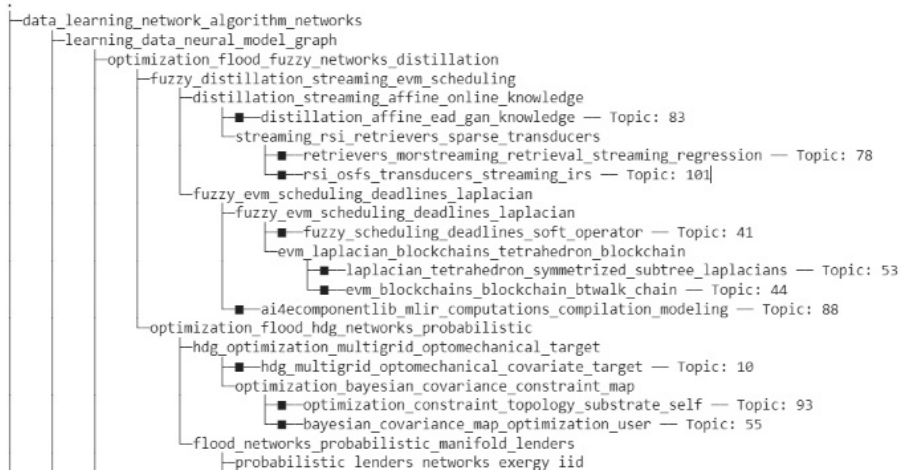


图3： 主题层级的一个实例

3.2. 关系

下一步是利用主题之间的关系创建主题网络。目前，S&TO本体基于从Semantic Scholar收集的393,991篇科学论文构建，时间范围为2021年10月至2022年8月。它涵盖四个科学领域：计算机科学、物理学、化学和工程学。S&TO本体遵循数据模型简单知识组织系统[6]，并包括以下语义关系：

- “relatedIdentical“，它是 SKOS相关的子度量，表示在评估研究主题时，两个主题可视为相同。主题之间的相似度在 SQL 存储过程中以余弦相似度计算，该过程为 create_topic_nets。如果相似度阈值高于 0.9，则建立主题之间的关系。

- “superTopicOf“：它是SKOS较窄的一个子度量，表示在图中，一个主题是另一个主题的上级领域。例如，"stream- inmg_rsi_retrieval_streaming_regression" 是主题ID为78和101的主题的上级主题，如图 3所示。

- “CommonArticles”：Itextracts 提取出现在两个主题中的共同文章。两个主题之间的关联通过分配给这些主题的共同文章所对应的概率分布之和进行评估。

- “nSimilarTopics“：它会返回与输入关键词相似度最高的 x 个主题。例如，表格中列出了与关键词“motor”相关的相似度最高的 5 个主题3。

[6]简单知识组织系统（SKOS）- http://www.w3.org/2004/02/skos。

表3与关键词“motor”最相似的5个主题

主题编号	Name
33826	33826_电动的_主轴_静压_节流器
50435	50435_多电机_齿轮传动系_摇臂_卡车
62709	62709_衬套_摩托车_电动的_珩磨
49677	49677_电动的_主轴_纳米催化剂_拖曳
46619	46619_动力传动系统_动力传动系统_土方工程_2025

4. 实验结果与讨论

在文献中， [13], 本体通过四种方法进行评估：基于金标准 [14],基于语料库 [15], 基于应用 [16], 以及基于结构的方法 [17]。基于金标准的方法将所构建的本体与先前开发的参考本体进行比较。基于语料库的方法将所构建的本体与涵盖特定领域的文本语料库内容进行比较。基于应用的方法会根据应用在各个用例中的表现来考量应用并进行评估。基于结构的方法则量化大小和本体复杂度等基于结构的属性。

选择最佳评估方法并阐明评估已开发本体的理由是必要的。在本研究中，科学与技术本体尚处于开发早期，未来工作中将继续扩展。因此，基于应用的方法并非合适的评估方法，因为该本体目前仍在开发中，不适用于应用目的。所提出的本体基于 Semantic Scholar 数据集的一个子集构建。因此，最佳参考本体应为 Semantic Scholar。然而，由于无法获取 Semantic Scholar，将其作为黄金标准本体并不切实际。

基于结构的评估通过多项度量执行，包括知识覆盖度和流行度度量（即属性和类的编号），以及结构度量（即最大深度、最小深度和深度方差）。采用这些度量的依据是：具有高深度和广度方差的密集本体更有可能产生有意义的语义内容。结构指标与本体中自适应建模知识的语义准确性相关 [19]。

在所提出的 S&TO 本体背景下，我们通过考察其分类结构，对一些结构指标进行了量化。S&TO 本体包含 5,153 个主题和 13,155 个语义关系（其中 8,052 个主题基于余弦相似度，5,103 个主题基于概率分布）。图 4 展示了一个涵盖主题及语义关系的知识图谱示例。我们使用 neo4j 作为图数据库来存储最终的本体 [18]。绿色连线表示通过余弦相似度评估的语义关系，黄色连线则表示基于分配给各主题的论文概率分布确定的关系。S&TO 在主题聚类中涵盖了最多的文章，仅将 15.47% 的文章归为离群值，从而能够提取属于非常规文章的主题。

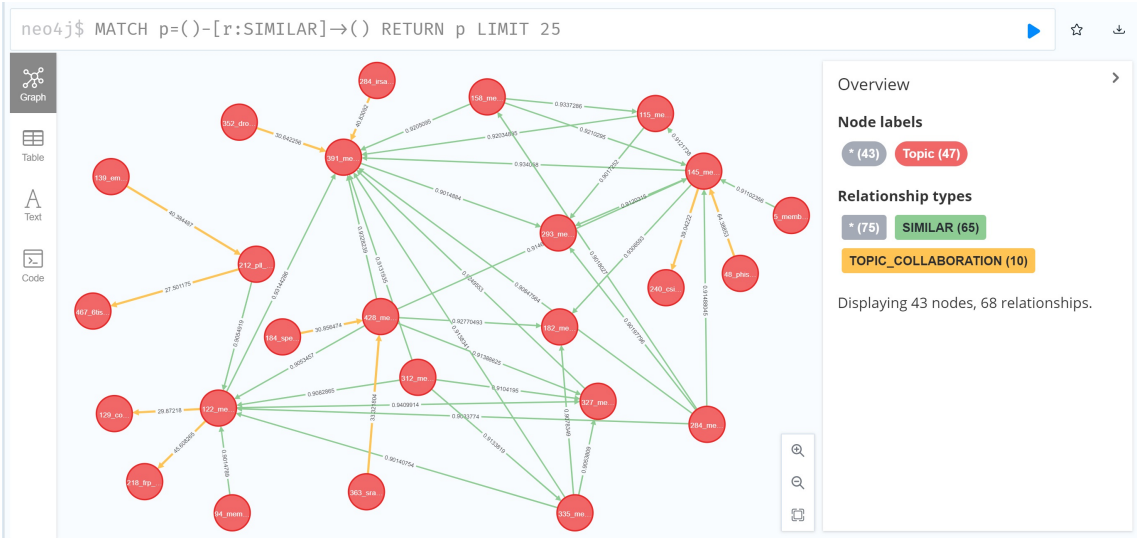


图 4： Neo4j Browser 中的一个知识图谱实例 [18]

表4表格“主题_网络”的结构

字段	描述
主题_net_id	主键[7],网络唯一标识符
创建时间 _on	表示网络创建的时间
状态	表示网络的状态（新建、完成等）
年份_月份	表示该网络创建的月份

4.1. 主题详情

本节讨论主题网络的结构、主题网络中的主题，以及与网络中主题相关的关键词。

“主题_网络”表（表4）提供了有关主题建模网络发展和状态的信息。每个网络都会被分配一个唯一标识符，称为“主题_网络_id”，即其主键。“创建_时间”参数指定了网络建立的日期和时间。“状态”字段提供网络当前状态的信息，该字段可能具有多种取值。此字段有助于跟踪处理进度，并确保所有网络均已正确生成和评估。此外，“年_月份”参数提供了网络创建的月份。由于该功能使研究人员能够了解主题及其关联如何随时间变化，因此对于跟踪语料库中主题的时间演变十分有益。

“主题_网络_主题”表（表5）提供了与主题网络相关主题的基本信息。每个主题都有一个唯一标识符，即“主题_ID”，也就是主键。每个主题通过一个唯一标识符与相应的网络关联，即“主题_网络”，使用“主题_网络_ID”作为外键。主题的描述性标签

表5表 'topic_nets_topics' 的结构

字段	描述
topic_id	主键，主题的唯一标识符
主题_net_id	外键[8] 指向主题_网络表的编号 整数 主题编号
标签	主题标签
主题_权重	与此主题相关的论文数量
嵌入	用于计算余弦相似度的主题嵌入
相似_主题	与相似主题相关的主题 ID 数组

表 6 表 'topic_nets_topics_keywords' 的结构

字段	描述
主题_标识	主键，主题的唯一标识符
编号	主题编号
row	按分数递增排序的自增编号
关键词	关键词
分数	与该主题的关键词相关联的分数

根据相关论文中最常见的术语设定。主题具有“主题_权重”，表示与之相关的论文数量，体现其在语料库中的重要性和相关性。它还存储“嵌入”，用于计算余弦相似度，以及一个名为“相似_主题”的数组，其中包含与搜索主题相关的主题 ID。

“主题_网络_主题_关键词”表（表6）提供了与主题相关联的关键词的重要信息，这些关键词由称为“主题_id”的唯一标识符表示，该标识符是主键。此表存储的字段包括：“number”表示主题编号，“row”是自增编号，“keyword”表示关键词名称。此外，“score”表示与该主题的关键词相关联的权重。

“论文_主题”表（表7）展示了学术论文与其所涉及主题之间的关系。“语料库_id”表示论文提取自哪个语料库的唯一标识符。“主题_id”表示论文所涉及主题的唯一标识符，便于追踪该主题下的论文。“probability”表示分配给该主题的论文权重。该权重表明论文涉及该主题的程度。概率值通常会进行归一化，缩放至 0 到 1 之间。

4.2. 关系详情

本节讨论主题关系的结构，例如有关主题之间的边和相似性的信息。

“主题_网络_主题_边”存储了网络“主题_网络”中各个“主题_网络_主题”之间边的信息（表 8）。边连接由其唯一标识符表示的两个主题（即“主题_id1”和“主题_id2”）。“边_权重”是

表 7“论文_主题”表的结构

字段	描述
语料库_标识	主键的一部分，并且是 papers 表的外键
主题_标识	主题_网络_主题表的主键和外键的一部分
row	自增编号，按概率升序排列
概率	论文被分配至该主题的概率

表 8“主题网络主题边”表的结构

字段	描述
主题_id1	主键的一部分，也是指向 topic_nets_topics 表的外键
主题_id2	主键的一部分，也是指向 topic_nets_topics 表的外键
边_权重	共享这两个主题的论文的概率之和
列_ _的字符串	按调和平均数计算的字段权重和归一化权重

表 9“主题_网络_主题_相似度”表的结构

字段	描述
主题_id1	主键的一部分，也是指向 topic_nets_topics 表的外键
主题_id1	主键的一部分，也是指向 topic_nets_topics 表的外键
相似度	两个主题嵌入之间的余弦相似度

两个主题之间共享论文的可能性之和（表7中的“可能性”字段）。协作强度“协作强度_的_协作”表示通过调和平均数计算并根据主题权重归一化后的权重，如式（2）所示。

$$str_of_col = HarmonicMean\left(\frac{edge_weight\ \ (id1,id2)}{topic_weight(id1)}\ \frac{(id1)}{edge_weight(id1,id2)}\right) \tag{2}$$

“主题_网络_主题_相似度”存储了基于相似度的主题之间边的信息（表9）。

5. 优势与用例

所提出的包含非常规主题的 S&TO 具有以下优势和用例。

5.1. 知识扩展

S&TO 可通过纳入以往未曾考虑的主题，拓展知识表示的范围，突破现有本体的局限。提供一种更全面、更整体的

知识格局视角，能够带来新的见解与发现。在医学研究中，可以将整体疗法或正念练习等非常规主题纳入本体，从而更全面地呈现更广泛的健康与福祉格局 [20]。

5.2. 跨学科合作

其次，通过连接可能具有共性或彼此相关的不同领域主题，非常规主题本体能够促进跨学科合作。这有助于发现新的研究领域并推动跨学科协作，从而为复杂问题带来新颖的解决方案。例如，融合计算机科学与心理学问题的本体，可以让这两个领域的学者更容易在人机交互或情感计算方面开展协作 [21]。

5.3. 可扩展性与适应性

非常规主题本体的优势在于易于更新和调整，能够反映最新的发展动态和主题，从而形成一个动态且灵活的知识表示系统。这一能力在技术和医疗保健等快速发展的领域尤为重要，因为新的主题和概念会定期涌现。例如，包含与人工智能[22] 或区块链 [23] 等新兴技术相关主题的本体，可以轻松更新，以纳入新的概念和趋势。

6. 局限性

当前提出的 S&TO 本体版本存在以下局限。

6.1. 有限数据集

当前版本的 S&TO本体基于 Semantic Scholar 数据集构建，涵盖了 2021年10月至2022年8月的 393,991 篇科学与技术文章。不过，它也可以基于更多数据集构建。

6.2. 主题标注

由于 S&TO本体采用了 BERTopic（一种无监督主题计算库），因此本体会受到未标注主题所带来问题的影响。由于缺乏标注数据，确定非常规主题的重要性和相关性，并将其与噪声或不相关主题区分开来，可能颇具挑战性。在处理包含各种主题的大规模复杂数据集时，这一问题尤为棘手。

6.3. 领域覆盖面

S&TO 可以通过涵盖以下四个领域来捕捉各种研究主题和领域：计算机科学、物理学、化学和工程学。然而，科学与技术领域内的许多其他学科和子领域尚未纳入 S&TO。例如，生物学，

环境科学和神经科学都是重要的研究领域，可将它们整合到本体中，以创建一个更全面的多学科框架来理解科学研究。扩大 S&TO 的覆盖范围，将其他领域纳入其中，可能会带来多项益处。

6.4. 主题质量

虽然 S&TO 在组织和分类科学与技术主题方面做出了重大努力，但本体中所包含主题的质量仍有提升空间。

7. 结论与未来工作

本文介绍了科学与技术本体（S&TO），这是一种涵盖所有科学研究主题的自动化本体。我们利用 BERTopic 对从 Semantic Scholar 获取的 2021年10月至2022年8月期间的 393,991 篇科学文章进行分析，构建了一个涵盖四个不同科学领域的本体。S&TO 可以利用近期数据集通过 BERTopic 进行更新，从而为知识表示提供动态且灵活的基础。S&TO本体有望拓展知识表示的范围并促进跨学科合作，使其成为科学家和技术专家的重要资源。

S&TO本体不断演进，需要持续改进，以满足不断扩展的知识格局的需求。目前，S&TO正在发展中，我们正采用主题标注技术，通过为不同主题赋予有意义的“标签”，来改进其组织方式并加深对它们的理解。这使用户能够轻松浏览本体并获取有价值的见解。此外，我们计划探索更多用于主题建模和聚类的方法与算法，以提高主题质量。这将增强本体描述知识格局的准确性和效能。此外，我们计划将本体扩展到更大规模的数据集，以纳入更多非常规类别和主题，从而改进并丰富知识库。

参考文献

[1] H. Saif、Y. He、H. Alani，Twitter语义情感分析，载于：语义网——ISWC 2012：第11届国际语义网会议，波士顿，马萨诸塞州，美国，2012年11月11日至15日，论文集，第一部分11，Springer，2012年，第508–524页。[2] F. Osborne、A. Salatino、A. Birukou、E. Motta，利用智能主题挖掘器自动分类Springer Nature会议论文集，载于：语义网——ISWC 2016：第15届国际语义网会议，日本神户，2016年10月17日至21日，论文集，第二部分15，Springer，2016年，第383–399页。[3] M. Dudáš、S. Lohmann、V. Svátek、D. Pavlov，本体可视化方法与工具：最新进展综述，《知识工程评论》33（2018）e10。[4] A. A. Salatino、F. Osborne、E. Motta，Augur：预测新研究主题的涌现，载于：第18届ACM/IEEE数字图书馆联合会议论文集，2018年，第303–312页。

[5] S. Fathalla、S. Vahdati、S. Auer、C. Lange，Semsur：用于研究发现语义表示的核心本体，《Procedia计算机科学》137（2018）151–162。[6] A. A. Salatino、T. Thanapalasingam、A. Mannocci、F. Osborne、E. Motta，计算机科学本体：大规模研究领域分类体系，载于：语义网——ISWC 2018：第17届国际语义网会议，美国加利福尼亚州蒙特雷，2018年10月8日至12日，论文集，第二部分17，Springer，2018年，第187–205页。[7] A. D. Wade，Semantic Scholar学术图（s2ag），载于：2022年Web大会伴随论文集，2022年，第739–739页。[8] M. Grootendorst，BERTopic：采用基于类别的tf-idf流程的神经主题建模，arXiv预印本arXiv:2203.05794（2022）。[9] L. McInnes、J. Healy、J. Melville，UMAP：用于降维的统一流形近似与投影，arXiv预印本arXiv:1802.03426（2018）。[10] J. A. Hartigan、M. A. Wong，算法AS 136：一种K-means聚类算法，《英国皇家统计学会杂志：C辑（应用统计）》28（1979）100–108。[11] L. McInnes、J. Healy、S. Astels，HDBSCAN：基于层级密度的聚类，《开源软件杂志》2（2017）205。[12] D. Müllner，现代层级凝聚聚类算法，arXiv预印本arXiv:1109.2378（2011）。[13] M. Fernández、C. Overbeeke、M. Sabou、E. Motta，什么造就了优质本体？细粒度知识复用案例研究，载于：语义网：第四届亚洲会议ASWC 2009，中国上海，2009年12月6日至9日，论文集4，Springer，2009年，第61–75页。[14] A. Maedche、S. Staab，衡量本体之间的相似性，载于：知识工程与知识管理：本体与语义网：第13届国际会议EKAW 2002，西班牙锡圭恩萨，2002年10月1日至4日，论文集13，Springer，2002年，第251–263页。[15] C. Brewster、H. Alani、S. Das mahapatra、Y. Wilks，数据驱动的本体评估（2004）。[16] M. Sabou、J. Gracia、S. Angeletou、M. d’Aquin、E. Motta，评估语义网：一种基于任务的方法，载于：语义网：第六届国际语义网会议、第二届亚洲语义网会议，ISWC 2007+ ASWC 2007，韩国釜山，2007年11月11日至15日，论文集，Springer，2007年，第423–437页。[17] P. Buitelaar、T. Eigner、T. D. OntoSelect，支持本体选择的动态本体库，载于：国际语义网会议演示环节论文集，日本广岛，11月（2004）。[18] D. Fernandes、J. Bernardino，图数据库比较：Allegrograph、arangodb、infinitegraph、neo4j和orientdb，载于：Data，2018年，第373–380页。[19] D. Sánchez、M. Batet、S. Martínez、J. Domingo-Ferrer，语义方差：一种直观的本体准确性评估指标，《人工智能工程应用》39（2015）89–99。[20] J. Howard，人工智能：对未来工作的影响，《美国工业医学杂志》62（2019）917–926。[21] W. Xu，迈向以人为本的人工智能：来自人机交互的视角，《Interactions》26（2019）42–46。[22] C. Zhang、Y. Lu，人工智能研究：现状与未来展望，

《工业信息集成期刊》23 (2021) 100224。[23] A. A. Monrat, O. Schelén, K. Andersson,《从应用、挑战和机遇的角度对区块链进行综述》，IEEE Access 7 (2019) 117134–117151。

A. BERTopic 参数

此处，我们总结了 S&TO 开发过程中设置的参数。以下参数已设定：

• “n_neighbors”指估算流形所需的邻近数据点数量。较大的样本点嵌入能呈现更全局的结构视角，而较低的值则会产生更狭窄的视角。为取得良好平衡，我们将 n=2 设为估算结果。

• “n_components”指降维后保留的成分数量。此值会直接影响聚类性能，因此有必要将其设为最优值。默认情况下，该值设为 5，以尽可能降低维度，同时最大限度地保留生成嵌入中的信息。

• “low_memory“：由于我们使用的是超大数据集，需要大量内存，因此将其设置为 TRUE。
• “min_cluster_size“：簇生成的数量很大程度上取决于簇的大小。必须调整最小值。经过多次实验，发现将簇大小设为 50 是最优选择。高值会生成数量较少但大小可观的簇，而低值则会生成微簇。

• “metric“：度量（例如 HDBSCAN）用于计算距离。这里我们选择了欧几里得度量，因为在降低维度后，我们得到的是低维数据，不需要进行太多优化。不过，如果你增大 UMAP 中的“n_components”，建议研究适用于高维数据的度量。

• “prediction_data“：请务必始终将此值设置为 True，因为之后预测新的数据点时需要用到它。如果你不希望预测任何未见过的数据点，可以将其设置为 False。

•“min_samples“：此项会自动设置为“min_cluster_size“，并控制生成的离群点数量。将此值设置得显著低于“min_cluster_size“，可能有助于减少噪声量。请注意，离群点通常是预期会出现的，而强行让输出中没有离群点可能无法准确呈现数据。

• “top_n_words”指每个主题提取的词语数量。实际上，我们将该值保持在 30 以下，最好介于 10 到 20 之间。原因是，代表某个主题的词语越多，其相关性可能越低。在这种情况下，排名靠前的词语最能代表该主题，也是关注的重点。

• “min_topic_size“ 指定了主题的最小大小。大小值越低，生成的主题就越多。如果该值设置得过高，则可能不会生成任何主题。我们将该值设置得过低，结果得到了许多微簇。

• “calculate_probabilities“ 为每个文档给出所有主题的概率。这可能会减慢大量文档的主题提取速度。

• “low_memory“ 设置为 true 可确保计算时占用较少的内存。虽然这会减慢计算速度，但能让 UMAP 在内存较少的机器上运行。
• “diversity”报告的主题多样性范围为 0 到 1，其中 0 表示没有多样性，1 表示多样性很高。主题多样性越高，簇越小，主题的一致性就越低。在我们的案例中，假设多样性为 0.4 或更高。