doi:10.3772/j.issn.2095-915x.2026.02.001

开放科学（资源服务）标识码（OSID）

基于知识图谱的知识产权竞争情报文本语义过滤模型

孙利娟[1,2] 韩宜书[2,3] 吴旭[2,3]

1. 国家图书馆信息技术部 北京 100081；
2. 北京邮电大学可信分布式计算与服务教育部重点实验室 北京 100876；
3. 北京邮电大学网络空间安全学院 北京 100876

摘要：[目的 / 意义] 为满足知识产权竞争情报实际工作中以某主题为核心过滤所有语义相关文本的任务需要，解决长短文本间语义过滤不精准的问题。[方法 / 过程] 本文构建了一个基于知识图谱的知识产权竞争情报文本语义过滤模型。通过知识产权竞争情报领域知识图谱对查询主题和候选文档分别进行实体链接，并通过知识图谱的嵌入表示方法将其分别表示为实体向量集合，最后通过晚交互的神经网络计算两者的相似度，依据相似度对候选文本排序完成过滤任务。[结果 / 结论] 本文在自主构建的知识产权领域竞争情报知识图谱和语义过滤任务数据集上进行了系统地实验，结果表明本模型的平均倒数排名可达到 9.41，且排名前十的文档平均召回率高达 84.25%，上述各项指标均优于对比算法。本文模型能够克服实际任务中长短文本的差异，充分利用知识图谱的语义信息，在知识产权领域竞争情报文本语义过滤任务中具备可行性和准确性，能够满足现实工作的需求。

关键词：知识产权竞争情报；知识图谱；语义过滤；实体链接；BERT

中图分类号：G35；TP391

Knowledge Graph-Based Text Semantic Filtering Model for Intellectual Property Competitive Intelligence

SUN Lijuan[1,2] HAN Yishu[2,3] WU Xu[2,3]

1. Information Technology Department, National Library of China, Beijing 100081, China;
2. Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing 100876, China;
3. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China

基金项目 信息产业部标准研究项目“网络电子身份标识 eID 移动应用接口测试方法”（2014-1020T-YD）。

作者简介 孙利娟（1993-），博士研究生，主要研究方向为多模态数据挖掘、知识产权智能情报分析；韩宜书（1998-），硕士研究生，主要研究方向为数据挖掘、网络内容智能监管；吴旭（1963-），通信作者，研究员，主要研究方向为智能情报与情报安全，E-mail：wux@bupt.edu.cn。

引用格式 孙利娟 , 韩宜书 , 吴旭 . 基于知识图谱的知识产权竞争情报文本语义过滤模型 [J]. 情报工程 , 2026, 12(2): 3-14.

doi:10.3772/j.issn.2095-915x.2026.02.001

Abstract: [Objective/Significance] To address the practical need for filtering semantically related texts around a specific topic in intellectual property competitive intelligence tasks, and to solve the issue of inaccurate semantic filtering between long and short texts. [Methods/Processes] This paper constructs a knowledge graph-based semantic filtering model for intellectual property competitive intelligence texts. The model performs entity linking on both the query topic and candidate documents using a domain-specific knowledge graph for intellectual property competitive intelligence, representing them as sets of entity vectors through knowledge graph embedding methods. A late-interaction neural network then calculates their similarity, and candidate texts are ranked and filtered based on similarity score. [Results/Conclusions] Systematic experiments were conducted on a self-constructed intellectual property competitive intelligence knowledge graph and a semantic filtering dataset. Experimental results show that the Mean Reciprocal Rank (MRR) of this model is 9.41, and the average recall rate of the top ten documents is as high as 84.25%. Both indicators surpass those of the contrasting algorithms. The model effectively overcomes the discrepancy between long and short texts in practical tasks by leveraging semantic information from knowledge graphs, demonstrating feasibility and accuracy in intellectual property competitive intelligence text filtering.
Keywords: Intellectual Property Competitive Intelligence; Knowledge Graph; Semantic Filtering; Entity Linking; BERT

引言

随着全球化的深入发展与技术进步的日新月异，知识产权已成为企业核心竞争力的重要组成部分。在当今激烈的市场竞争环境中，获取并有效利用知识产权信息对于企业制定发展战略至关重要。然而，由于知识产权信息源的多样性和复杂性，传统的信息检索和分析方法往往难以满足企业在快速变化的市场环境中的需求。在实际的知识产权竞争情报研究工作中存在以知识产权竞争情报主题为核心，研究其发展演化规律的重要任务。这一任务顺利开展的必要前提是从海量数据中精准过滤出以某确定主题为核心的关联信息，但网络知识产权竞争情报具有来源复杂、内容多源异构、组织结构松散、语义表达浅层等特点，给实际研究工作带来了巨大的挑战。为解决这一基础性关键问题，本文提出构建基于知识图谱的精准语义过滤模型，为后续演化规律研究提供高质量的情报数据。知识产权竞争情报文本语义过滤任务的难点在两大方面，一是传统过滤手段对于语义信息的利用有限。基于语言规则构建分析模型，操作简便、执行速度快、业务吞吐量大，但非常依赖一个全面、准确的过滤词库，词库维护成本高，而且只考虑关键词损失了很多语义信息。而浅层语义过滤主要以句子谓词为核心，并未真正对句子含义进行分析。二是任务中查询语句和候选文本通常篇幅差距较大，传统的文本匹配模型对长文本不友好。查询语句通常为 20 字以内短文本，甚至是简单的动宾短语，候选文本则多为篇章，或者结构内容完整的长句子。要在篇幅差异巨大的文本之间计算相似度完成匹配需要建模更科学合理的方法并实验验证其效果。本文提出构建基于知识图谱的知识产权竞争情报文本语义过滤模型，通过图谱结构化组织实体关系，在理解同义词、技术归属等深层语义的基础上，利用关系推理精准锚定长文本中的核心情报片段，旨在通过知识图谱将知识产权竞争情报文本信息进行更加有序、有机地组织，有效补充以往文本语义分析中缺乏的知识信息，同时向量化的表示学习能够在保持文本语义特征的基础上更直观地进

行相似度的运算，规避基于规则方式存在的语义分析深度不足缺陷，提高基于深度学习的语义分析信息描述能力和模型的可解释性，为后续演化规律分析提供高价值的结构化数据输入，为企业的竞争情报分析提供强有力的技术支撑，有助于企业更好地把握市场动态，提升创新能力。

1 语义过滤相关工作

语义过滤技术分为浅层语义过滤和深度语义过滤。语义过滤技术旨在克服传统关键词匹配的语义局限，其发展呈现从浅层结构分析向深度语义理解演进的趋势，并需应对专业领域中的计算效率与适配性挑战。浅层方法虽能快速处理基础语义特征，却难以捕捉复杂语境与知识关联；深度方法虽显著提升语义理解能力，但在实际落地中常受限于计算开销、标注成本及跨篇幅文本匹配的鲁棒性。尤其在知识产权竞争情报场景下，多源异构数据、专业术语密集及长短文本交互等特性，进一步放大了现有方法的不足。为此，亟须探索融合领域知识、高效且深度的语义过滤新范式，以支撑复杂情报分析任务的需求。

1.1 浅层语义过滤技术及其局限性

浅层语义过滤技术旨在从文本中提取基础语义信息，基于句法树的语义角色标注方法[1]依赖于句法分析结果，识别句子中的谓词并标注时间、地点等核心论元及其语义角色。它能较好地捕捉句子的核心谓词 - 论元结构。基于特征向量的过滤技术[2]将文本中的结构化信息（如词、短语、句法关系等）转化为高维特征向量，利用向量空间模型进行相似度计算或分类，优势在于计算相对简便（可将结构关系转化为空间距离），并能保留词、短语级别的浅层语义信息。但在将复杂结构“扁平化”为向量的过程中，原有的层次化结构信息不可避免地会遭受损失。基于树核的过滤方法[3]直接对句法树进行操作，通过计算子树结构的相似性来衡量句子的语义相似度。其核心优势在于能够有效利用句子的层次结构信息，显著降低特征维度，节省存储空间。但其主要缺点在于子树匹配可能引入与核心语义无关的“噪声”信息，且每次计算都需要对复杂的树结构进行遍历和匹配分析，导致计算效率低下。

浅层语义过滤技术在自然语言处理发展史上扮演了重要角色，特别是在深度学习兴起之前，是理解文本基础语义的主要手段。它们在信息抽取、文本分类等特定任务上取得过成功应用。然而，其对句法分析的强依赖性、语义表示的浅层性、计算效率瓶颈、领域适应性差以及难以处理复杂语义的固有的局限性最终导致其被取代。当前，这些传统方法更多作为特定场景的补充手段、提供结构化监督信号。

1.2 深度语义过滤方法的分类与挑战

深度语义过滤方法分为三类：基于知识图谱的语义过滤、有监督的语义过滤、半监督或无监督的语义过滤[4]。其中知识图谱[5]以三元组形式积累事实，对给定的主题进行抽取可得到其形式化表示，而对于待过滤文本和给定查询均可以通过实体链接技术链接到实体，再从知识图谱中引入外部知识并完成形式化表示，后续通过表示学习的方法计算相似度或通过图结构路径搜索的方法比较子网都可以实现更深

doi:10.3772/j.issn.2095-915x.2026.02.001

层次的语义过滤[6]。有监督语义过滤是深度学习通用方式，通过高维度神经网络结构学习语义特征完成分类任务进而实现过滤工作，但有监督的方法往往需要大量人工标注的语料作为训练的数据集，专业性强，人工成本高。半监督或无监督语义过滤则不需要人工标注语料，因此相较有监督方法多了一个自建训练数据集的步骤，即通过知识图谱中的实体名和关系名，使用 EM 算法[7]不断迭代，选出高置信度文本和结果作为训练数据，之后模型的训练和调整与有监督方式基本相同。

1.3 知识图谱增强的深度语义过滤

以往对语义过滤的研究大多在公开数据集上进行，存在计算存储开销过大和过滤深度有限的问题，在知识产权领域竞争情报文本过滤的实际任务中可行性受限。此外，由于在竞争情报文本语义过滤任务中待匹配文本的篇幅差异较大，难以在长短文本间提取语义信息并进行相似度计算。因此，本文基于知识产权领域知识图谱引入语义信息以提高语义深度，通过实体链接[8]发掘关联信息，依靠表示学习完成向量化，最后通过晚交互的神经网络结构学习查询和候选文档间的语义信息并衡量相似度，按照相似度排序并确定阈值完成过滤任务。

2 基于知识图谱的文本语义过滤模型

针对传统文本过滤模型语义深度有限，且难以克服用户简短查询与海量长文本候选集之间篇幅与信息密度差异的问题，本文创新性地引入自主构建的知识产权领域竞争情报知识图谱，以深度融合领域知识突破语义理解瓶颈。传统基于向量编码与余弦相似度的方案虽在通用场景有效，但在知识产权竞争情报场景面临显著挑战：用户查询多为高度浓缩的 20 字左右短文本，而候选文本（如专利全文、技术报告）结构复杂、篇幅冗长。直接在长短文本对间计算向量相似度，易因长文本噪声干扰导致核心语义关联被稀释。因此，本文通过知识图谱对文本中的机构等实体及其竞争、创新关系进行有序组织，补充纯文本语义分析缺失的专业逻辑；将查询短文本与长文本经实体链接对齐至图谱空间后，通过图谱嵌入的向量运算实现知识推理[9]，再结合晚交互神经网络动态捕捉语义交互特征；在知识增强的向量空间中计算距离，确保相似度结果既保留文本语义特征，又符合领域知识逻辑，最终实现对竞争情报文本的精准语义过滤与排序。

本文以实际网络知识产权竞争情报研究工作需求为导向，以跟踪某一主题竞争情报，分析其发展演化规律为目的开展研究，情报研究人员通常需要从海量的知识产权竞争情报文本数据中过滤出某确定主题为核心的关联信息，或与该主题近似或同义的信息，本文将这一任务定义为竞争情报文本语义过滤任务。在定义语义过滤任务之前，需要首先引入用户和查询的概念，用户即提出需求并使用服务的人，在实际应用场景中就是需要使用模型完成竞争情报文本过滤任务的人，即竞争情报研究工作者。查询（query）是用户在搜索引擎查找数据时输入的条件，迁移到应用场景中则是用户在使用模型时输入的语句，即用户设定的过滤条件，通常为需要保留的主题描述，知识产权竞争情

报文本语义过滤任务中限定查询语句 20 字以内的短文本，可以以“主谓宾”结构表述，查询的集合记作 Q，候选文本集 D 是所有待过滤文本的集合，根据知识产权领域竞争情报文本特征分析通常是篇章级文本，属于长文本范围。

文本语义过滤任务指的是依据用户输入的查询文本，将候选文本集 D 中与查询语句相似度高的文本保留下来，而相似度低的文本舍弃掉的过程，可以简单建模为一个前 k 项（top-k）的排序问题，通过合理设计阈值或根据用户需求保留前 n 条文本，即可视作完成语义过滤任务。本模型设计和实现重点在于数据集的构建、长短文本的向量化表征和评分函数的设计。数据集的构建方面，查询集 Q 依赖自主构建的知识产权领域知识图谱进行构造，而候选文本则借助搜索引擎的推荐功能进行爬虫采集，保留对应关系进行存储得到候选文本集 D。向量化表示依靠知识图谱的双曲嵌入 [10] 表示学习方法完成，通过晚交互 [11] 的神经网络提取向量特征，最后是评分函数的设计，本文采用了相似度计算函数 S，使得如果候选文本属于候选文本集 D，且文档与查询相似，那么相似度得分 S 越高，排序越靠前；反之如果候选文本与查询越不相关，则相似度得分 S 越低，排序越靠后。相似度得分 S 归一化后也可以视为相关的概率大小。

基于知识图谱的知识产权竞争情报文本语义过滤模型包括知识图谱构建、查询和候选文本的实体链接、查询和候选文本的向量化表示和向量相似度计算模块四大部分。首先构建知识产权竞争情报领域知识图谱，其次基于知识图谱信息由实体链接和实体识别模块直接提取查询语句（短文本）中实体，并通过向量匹配得到相关间接实体；其次是候选文本的实体链接模块通过实体链接技术获取待过滤文本（候选文本、长文本）的实体列表；第三，将查询语句实体列表和候选文本实体列表通过嵌入表示学习模型向量化，从而拼接得到详细向量表示。最后向量相似度计算模块通过晚交互神经网络计算长短文本向量表示的相似度确定文本相似度，并根据相似度对候选文本进行排序，结合领域知识确定过滤阈值。基于知识图谱的竞争情报文本语义过滤模型流程图如图 1 所示。

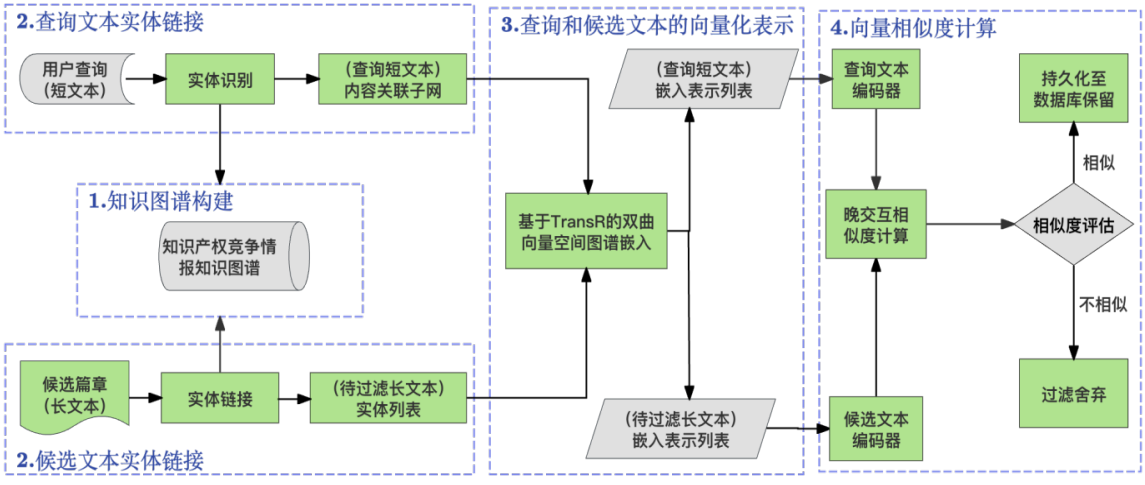

图 1　基于知识图谱的知识产权竞争情报文本语义滤模型流程

doi:10.3772/j.issn.2095-915x.2026.02.001

2.1 知识图谱的构建

本文对于知识产权领域竞争情报文本知识图谱的构建采用了自顶向下的设计和自底向上的抽取结合的方式进行。通过调研和自主归纳方式针对知识产权领域竞争情报文本信息特征和图谱实际需求完成图谱结构的顶层设计，其次通过广泛采集竞争情报文本并从大量文本数据中通过实体识别关系、抽取技术批量抽取知识，完成图谱内容的补充。具体包括知识产权领域竞争情报数据收集和预处理、情报数据的实体识别和关系抽取、知识图谱的存储等步骤[12]。本文从知识产权相关官方网站、社交媒体、专业论坛、新闻报道等网络平台采集知识产权领域竞争情报文本，比如国家知识产权公共服务网、知识产权一站式服务平台、地方性知识产权公共服务平台、万商云集、全国企业信用信息公示系统等，结合已有先验领域知识作为数据来源。对数据预处理后通过词关系的命名实体识别模型进行实体识别和关系抽取，在此基础上引入少量人工校验和标注，完成知识图谱构建。知识图谱包括知识产权领域以名词性实体为核心的知识图谱和以动词性事件为核心的事理概念库两大部分。

网络知识产权领域的竞争情报内容涵盖专利、商标、版权等多种知识产权类型，信息更新速度快，可及时反映市场和技术的变化。信息量庞大，数据之间存在着复杂的关联性，需要运用知识图谱等技术揭示隐含的关系。有些重要信息具有隐蔽性，需要采用网络爬虫、数据挖掘等技术手段进行深入挖掘。针对以上特征本文考虑分别从竞争情报主体和竞争情报事件两方面进行知识图谱结构的设计，知识图谱构建流程如图 2 所示。

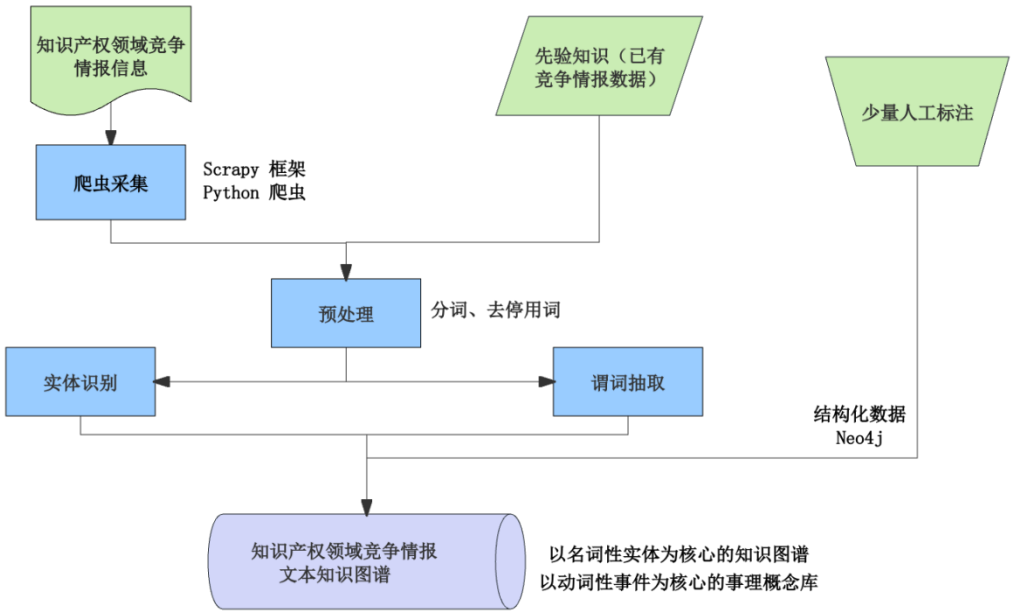

图 2 知识产权领域竞争情报文本知识图谱的构建流程

综合考虑竞争情报涉及的人物、发生的地点，从概念和实例的角度进行分析，设计了人物实体四类：情报生产者、情报消费者、情报管理者和竞争对手，机构实体两类：内部机构、外部机构，两者之间存在互补关系，事件实体一类，共计 7 类实体。并根据 7 类实体设计了

11 类关联，具体表示如下：人物 - 发明 - 专利、人物 - 工作于 - 研究机构、研究机构 - 持有 - 专利、人物 - 担任 - 职务、机构 - 注册 - 商标、技术领域 - 关联 - 专利、事件 - 涉及 - 人物、人物 - 持有 - 版权、机构 - 持有 - 商标、实验室 - 产生 - 专利和机构 - 使用 - 技术等。根据领域知识特征和我们构建图谱关注的信息，充分考虑了不同网络平台数据特征，设计了对采集到的情报文本数据的存储结构。存储包括信息来源、作者、内容、发布事件在内的各种信息，具体的字段设计如表 1 所示。

知识产权领域竞争情报文本的采集主要依赖爬虫技术实现，本文使用 scrapy 框架编写 Python 爬虫，通过分析采集源对应的网络平台网页结构，通过 Xpath、CSS 选择器或模拟登录的方式获取竞争情报文本数据。使用 jieba 分词工具对语料的标题和正文内容进行分词、去停用词和主题无关词汇，对剩余词汇利用同义词林等方式进行语义增强。通过词关系的实体识别技术统一处理扁平实体、重叠实体和非连续实体三种实体识别任务，通过规则模板匹配事理关系，后引入了部分人工标注完成图谱构建，其中人工标注是为避免误差积累影响后续知识图谱补充模型准确性，知识图谱示例如图 3 所示。

表 1　知识产权领域竞争情报文本数据字段表

字段名	类型	注释
bid	varchar(50)	board id 代表板块 id
aid	varchar(50)	数据贴主键 id
group_id	varchar(50)	数据组 aid
reply_id	varchar(50)	数据组回复信息 aid
title	longtext	文本标题
Image	longtext	图像地址
content	longtext	正文内容
uid	varchar(50)	发表该文本数据的用户 id
post_time	timestamp	该文本的发布时间
source	varchar(50)	文本数据的信息源 id

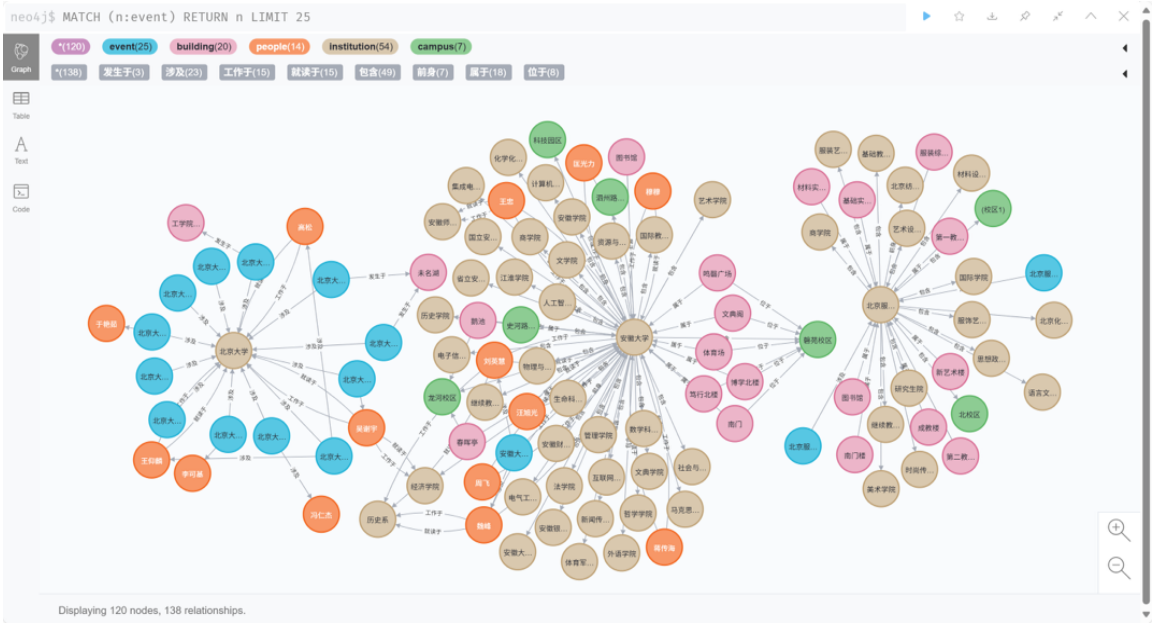

图 3　图数据库中知识图谱可视化示例

doi:10.3772/j.issn.2095-915x.2026.02.001

本文从预定的 13 个网络站点共采集得到 1 万余条知识产权领域网络竞争情报文本数据，通过词间关系实体识别模型 [13] 完成实体识别任务知识图谱实体库部分包括机构（instruction）、人物（people）、事件（event）三大类 7 小类共计 4912 个实体，事件发生于机构、事件涉及人物等 11 类关系；24 类属性。图 2 中是在图数据库 Neo4j 中使用 cypher 语句查询到的可视化图结构。蓝色节点（node）为机构，橘色节点是人物，棕色节点为实体机构等，关系由有向箭头标识，鼠标悬停时可显示实体属性。

2.2 查询文本和候选文本实体链接

为以实体列表的形式表示查询语句和候选文本，首先由实体链接和相关实体发现模块直接提取查询语句（短文本）中实体，并通过向量匹配得到相关间接实体；其次是候选文本的实体链接模块通过实体链接技术获取待过滤文本（候选文本、长文本）的实体列表。由于查询语句直接从知识图谱实体列表中组合得到，不存在实体消歧的问题，因此只考虑相关实体的发现问题。“主谓宾”结构的查询语句是从事件实体的属性中抽取到的文本描述，可以直接对应到具体的事件实体上，因此选择与该实体直接相连的一跳邻居作为相关实体。对于候选文本需要对文本进行实体链接，由于本文知识产权领域竞争情报文本知识图谱构建过程中实现过从数据源文本中识别命名实体的工作，此处再次依靠这个程序对候选文本完成实体识别，得到文本内所有的实体提及，并将实体及链接到知识图谱中得到候选文本的实体列表。

2.3 查询和候选文本的向量化表示

得到查询和候选文本的实体列表后，需要分别将查询和候选文本进行编码向量化表示。这部分主要包括查询编码器和候选文本编码器的构建和相似度评分函数的设计。查询语句和候选文本共享同一个编码器，通过基于 TransR[14] 改进的双曲对比嵌入模型得到向量表示，在尽可能保存知识产权领域竞争情报文本知识图谱语义特征包括层次结构和逻辑关系等信息的基础上，使用较少的空间对知识图谱进行准确表示。双曲空间嵌入能够保留知识图谱层次结构，通过庞加莱圆盘模型建模双曲空间。设置不同的双曲空间的曲率，可以决定空间的“弯曲”程度，曲率更大时双曲空间更接近树状，曲率小时则更为平坦更接近欧式空间。模型参考了 TransR 算法的重要假设：不同关系会考量实体不同方面的特征，将三元组按照不同关系进行分类，将实体投影到代表不同关系的双曲空间来进行实体的向量化表示，如图 4 所示。

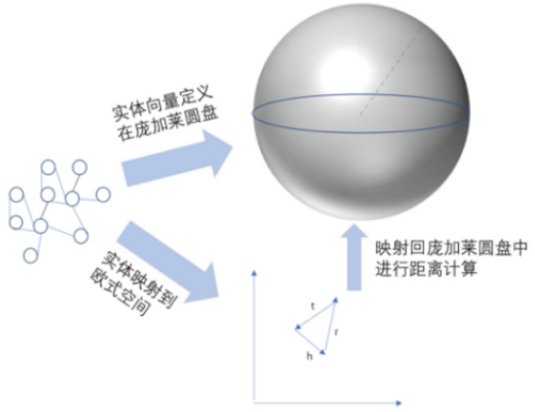

图 4 双曲对比嵌入流程

针对每个关系 r 学习一个特定关系绝对曲率 c_r，d 维的庞加莱圆盘模型建模双曲空间，负曲率为 $-c(c>0)$，缺主语表示为 $P^{d,c}$，R^d 是 d 维

欧氏空间，$\|\cdot\|$ 表示 L_2 范式。

$$P^{d,c}=\left\{x\in R^d:\|x\|^2<\frac{1}{c}\right\} \quad (1)$$

通过函数完成点在双曲空间和欧氏空间的转换，将点投影至对应的切线空间。对于任意一个点 $x\in P^{d,c}$，其切线空间 T_x^c 是一个 $P^{d,c}$ 中从 x 出发包含所有可能方向路径的 d 维向量空间。通过 *exp* 函数可以将 T_x^c 中的点投影到庞加莱圆盘 $P^{d,c}$ 中，而 *log* 函数则是将点从双曲空间中映射到切线空间中，两者互为反函数，如公式（2）（3）所示：

$$\exp_0^c(v)=\tan h\left(\sqrt{c}\|v\|\right)\frac{v}{\sqrt{c}\|v\|} \quad (2)$$

$$\log_0^c(y)=\arctan h\left(\sqrt{c}\|y\|\right)\frac{y}{\sqrt{c}\|y\|} \quad (3)$$

由于庞加莱圆盘模型中球内两点相加可能得到球外一点，所以双曲空间中的向量加法定义不明确，此处使用莫比乌斯加法 [10] 在双曲空间内模拟欧几里得加法。因此双曲空间 $P^{d,c}$ 内点 x,y 的显式距离为公式（4）：

$$d^c(x,y)=\frac{2}{\sqrt{c}}\arctan h\left(\sqrt{c}\left\|-x\oplus c_y\right\|\right) \quad (4)$$

2.4 向量相似度计算

将查询文本和候选文本的实体表示都替换为实体向量列表后需要分别将查询和候选文本进行编码向量化表示。这部分主要依靠晚交互范式排序模型实现，通过晚交互简化末层网络结构，提升计算效率。

查询文本和候选文本共享同一个编码器，通过 2.3 查询和候选文本的向量化表示模块构建的嵌入模型得到实体的向量表示，将查询文本和候选文本都变成向量列表，通过在查询和候选文本中分别添加一个特殊的标记[Q]和[D]，来区分输入序列类型，如（5）（6）式。

$$E_q:=Norm(CNN(HEC("[Q]q_0q_1\cdots q_m"\#\#\cdots\#))) \quad (5)$$

$$E_d:=Filter(Norm(CNN\left(HEC("[D]d_0d_1\cdots d_n"\right))) \quad (6)$$

其中“#”表示 [mask] 标记。CNN 是做维度变换，用来对 BERT 输出降维，Normalize 为了之后计算余弦相似度而做的正则化处理。交互计算查询实体列表中每个实体向量与候选文本中的每个实体向量的最大相似度，累加得到 score。由于之前进行过 Normalize 只需计算内积就得到余弦相似度，如公式（7）所示。

$$S_{q,d}:=\sum_{i\in[|E_q|]}\max_{j\in[|E_d|]}E_{qi}\cdot E_{dj}^T \quad (7)$$

最后，使用 Learning2Rank 中的 pairwise Loss，使得相关候选文本（正样本）的相似度得分高于不相关的文本（负样本），最后将相似度得分由高到低进行排序。

3 实验结果和分析

3.1 数据集构建

本文自主构建了高度专业化的知识产权竞争情报语义过滤数据集，该数据集由查询集和候选集共同组成，其中查询集 Q 源自知识产权竞争情报知识图谱——353 条“主谓宾”查询直接抽取真实事件属性，416 条“主谓”查询通过系统化组合 32 个专业谓词，如“某新型材料专利侵权诉讼”等，两类查询以 0.85 比例平衡真实性与泛化需求。候选集 D 通过百度新闻搜索引擎动态获取，每条查询采集 TOP10 结果涵盖企业公告、裁判文书等轻情报源，经人工校对核实后构建 7690 条高质量文本对（同查询

doi:10.3772/j.issn.2095-915x.2026.02.001

TOP10 为正样本，异查询文档为负样本）。该数据集虽规模受限于领域图谱实体，但其源于真实业务场景的构建逻辑、专业术语覆盖度如“专利侵权诉讼”的多形态表达，以及高纯度样本筛选机制，使实验结果在知识产权领域竞争情报语义过滤任务中具备较强验证力和说服力。在竞争情报文本语义过滤任务中数据集包括查询语句集 Q，候选文本集 D，且需要保留每条查询语句 q_i 与词条查询语句搜索得到的候选文本子集 D'_i 的对应关系。

$$Q:\{q_1, q_2, \cdots, q_i\} \quad (8)$$

$$D=D'_1 \cup D'_2 \cup D'_2 \cup \cdots \cup D'_i \quad (9)$$

$$D'_i:=\{d_{i,1}, d_{i,2}, \cdots, d_{i,j}\} \quad (10)$$

对于查询集 Q 依赖本文知识产权领域的竞争情报文本知识图谱进行构造。知识图谱的事件实体是知识产权领域发生过的实际竞争情报事件，此部分直接从实体的属性信息中抽取“主谓宾”格式的事件描述充当查询语句，共获得“主谓宾”结构查询语句 353 条。其次对于抽取后的事件描述中谓语进行统计，将事理概念库中所有谓词概念中未在词表中出现的谓词保留，还剩谓词 32 个，并分别从人物实体的概念层、机构实体、事件实体三大类中随机选择名词与这些谓词结合组成“主谓”结构查询语句，选择三类实体的数量按实体数量的比例关系定为（1，4，8）个，例如概念谓词“专利侵权诉讼”，从人物实体概念层内随机到实体概念“专利代理人”，可以得到查询语句“专利代理人专利侵权诉讼”；从机构实体中随机选取，可得“xx 学校专利侵权诉讼”；从事件实体中随机组合，可得到“某新型材料专利侵权诉讼”的查询语句，共得到“主谓”查询语句 416（32×13）条，共计 769 条查询语句，“主谓宾”格式和“主谓”格式查询语句数量比值为 0.85。

候选文本集 D 的构建则借助了搜索引擎的推荐排序功能进行搜集，由于本论文主要研究中文知识产权领域文本的语义过滤，因此选择百度搜索引擎采集候选文本，覆盖企业公告、法律文书等公开情报源，正样本为同查询 TOP10 文本，负样本来自异查询文档，经人工校对核实后形成 7690 条高质量文本对。通过 scrapy 框架针对百度新闻页面搜索功能定制爬虫，分别采集每个查询语句的前 10 条文本数据，构成候选文本集合 D。最后考虑训练效果，分别构建正样本和负例，其中通过查询语句 q_i 索引得到的前 10 篇文档 $d_{i,1}\cdots d_{i,10}$ 作为该查询的正样本，而其他查询 q_j 获得的文档可以作为查询 q_i 的负样本，其中 $i \neq j$。

3.2 比较方法

由于本文定义的竞争情报文本语义过滤任务为计算长短文本间的相似度，故选择了 BM25[15]、doc2query[16]、ColBERT 模型与本文模型进行对比实验，以验证其有效性。选择上述模型进行对比实验，是因为 BM25 和 doc-2query 都是经典的信息检索算法，分别是基于特征提取和基于文档扩展的方法，非常具有代表性。而本文模型是基于 ColBERT 方法改进的，与其对比可以验证改进的有效性。 BM25 属于检索算法，用来评价搜索词和文档之间的相关性，基于概率检索实现，是 TD-IDF 的优化版本，分别计算单词权重，单词和文档的相关度，以及单词和 query（关键词）的相关性最终得到相关度得分，并按分值排序取前十个结果。

doc2query 用于信息检索任务，原理简单巧妙，以文档扩展（document expansion）为核心，直接让相关的查询语句和文档“长得差不多”。即对于每个候选文档（document），利用 NLP 的方法根据生成可能相关的查询（query）文本，然后把候选文档和预测的查询文本进行拼接。比如候选文档是 D，预测的查询语句是 Q′，那么把 D 扩展成 DQ′，再将 DQ′ 与实际的查询 Q 进行相似度计算。ColBERT 是基于上下文化的后期交互的排序模型，是一种细粒度的基于 token 级的交互模式，用于估计查询 q 和文档 d 之间的相关性。查询 q 和文档 d 是共享同一个 transformer 编码器的，最后做交互取代直接计算求解查询 q 和文档 d 的分数。查询中的每个 token 和文档中的每个 token 求点积，然后取余弦相似度的最大值之和。

3.3 实验结果

在知识产权领域竞争情报文本过滤任务中，当前缺乏公开可用的标准化数据集。为契合知识产权竞争情报过滤的实际需求，本文采用自建数据集重点测试了 MRR 与 Recall@10 两项核心指标。MRR 反映排序准确性，在专利无效检索、技术方案比对等场景中，模型需将高相关度文本优先排序。Recall@10 体现信息完备性，知识产权分析要求尽可能不遗漏潜在相关文本（如在先专利、竞品技术文档）。实验结果显示，本文所提出的模型 MRR 值达到了 9.41，Recall@10 达到了 84.25，均超越了其他三个对比模型，表现最优。证明了本文所提方法在复杂语义环境下的文本匹配能力，为知识产权领域奠定了技术可行性基础。

从表 2 的数据可以看出，BM25 依赖统计方法进行过滤，由于缺乏深度语义信息利用，在 MRR 指标上表现较弱，但在 Recall@10 指标上的结果接近 Doc2query，Doc2query 通过生成式过滤机制增强了语义信息处理，其 MRR 值比 BM2 提升近 50%，语义理解能力有所进步；ColBERT 作为基于 BERT 预训练的上下文后期交互排序模型，凭借预训练优势也具备较优的实验结果。

本文模型通过优化语义处理机制，在竞争情报复杂语义过滤场景中实现了更高的检索精准度与召回效率，有效验证了其在实际竞争情报处理任务中的可行性与技术领先性，为后续在知识产权领域的落地应用提供了技术支撑。

表 2　对比实验结果

模型	MRR	Recall@10
BM25	4.10	69.72
Doc2query	6.13	70.03
CoLBERT	9.22	79.58
本文模型	9.41	84.25

4　结论

针对知识产权领域竞争情报文本具有信息密度高、语义关系复杂且长短文本间语义差异显著的特性，传统文本匹配方法在处理时面临语义信息挖掘不足及跨长度文本相似度计算敏感度偏低的问题，本文深度剖析该领域竞争情报文本特征，构建领域知识图谱以引入深层语义信息，设计并实现了适用于知识产权场景的竞争情报文本语义过滤模型。首先采用实体列表对查询语句与候选文本进行结构化表征，精

doi:10.3772/j.issn.2095-915x.2026.02.001

准捕捉领域特有的实体关联关系；继而借助双曲对比嵌入模型完成语义向量化编码，有效处理高维语义空间中的层次化关联；最终通过晚交互模式设计评分函数，实现对跨长度文本语义匹配度的精细化计算。基于自建的知识产权领域数据集，通过与 BM25、Doc2Query、ColBERT 对比实验，本文模型在 MMR 和 Recall@10 指标上均展现出显著优势，获得最优表现。

参 考 文 献

[1] 岳世峰 , 林政 , 王伟平 , 等 . 智能回复系统研究综述 [J]. 信息安全学报 , 2020, 5(1): 20-34.

[2] 阮冰 . 基于特征向量的文本信息过滤算法研究 [J]. 现代电子技术 , 2010, 33(4): 145-147, 151.

[3] 庄成龙 , 钱龙华 , 周国栋 . 基于树核函数的实体语义关系抽取方法研究 [J]. 中文信息学报 , 2009, 23(1): 3-8, 34.

[4] 潘怀奇 , 覃正优 . 基于知识图谱元路径的协同过滤算法 [J]. 现代计算机 , 2021, 27(27): 17-23.

[5] 徐增林 , 盛泳潘 , 贺丽荣 , 等 . 知识图谱技术综述 [J]. 电子科技大学学报 , 2016, 45(4): 589-606, 345.

[6] 王根生 , 潘方正 . 融合语义相似度的协同过滤推荐算法 [J]. 中国科学技术大学学报 , 2019, 49(10): 835-841.

[7] 周俊生 , 黄书剑 , 陈家骏 , 等 . 一种基于图划分的无监督汉语指代消解算法 [J]. 中文信息学报 , 2007(2): 77-82.

[8] LI J Y, JI DH, LIU J, et al. Unified named entity recognition as word-word relation classification[C]// Proceedings of AAAI Conference on 350 Artificial Intelligence, 2021: 10965-10973.

[9] 刘峤 , 李杨 , 段宏 , 等 . 知识图谱构建技术综述 [J]. 计算机研究与发展 , 2016, 53(3): 582-600.

[10] CHAMI I, WOLF A, JUAN D C, et al. Low-dimensional hyperbolic knowledge graph embeddings[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6901-6914.

[11] KHATTAB O, ZAHARIA M, et al. ColBERT: Efficient and effective passage search via contextualized 355 late interaction over BERT[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 39-48.

[12] 夏立新 , 陈健瑶 , 余华娟 . 基于事理图谱的多维特征网络舆情事件可视化摘要生成研究 [J]. 情报理论与实践 , 2020, 43(10): 157-164.

[13] LAN G, LI Y, HU M, et al. Knowledge graph integrated graph neural networks for Chinese medical text 340 classification[C]// Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021: 682-687.

[14] LIN Y K, LIU Z Y, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of AAAI Conference on Artificial Intelligence, 2015: 2181-2187.

[15] 陈乐乐 , 黄松 , 孙金磊 , 等 . 基于 BM25 算法的问题报告质量检测方法 [J]. 清华大学学报 (自然科学版), 2020, 60(10): 829-836.

[16] LI Z H, YANG N, WANG L, et al. Learning diverse document representations with deep query interactions for dense retrieval[EB/OL]. (2022-08-08) [2025-05-01]. https://arxiv.org/abs/2208.04232.

（责任编辑：高雄）