DOI:10. 20033/j. 1003-7241. (2025)12-0138-04

基于深度学习的多源信息融合知识图谱智能化构建技术

刘 敏
(湖南科技职业学院软件学院,湖南 长沙 410004)

摘 要:知识图谱作为一种强大的知识表示和推理工具,已广泛应用于智能问答、语义搜索和决策支持等领域,为实现不同来源数据的整合以及深度推理,研究基于深度学习的多源信息融合知识图谱智能化构建技术。采用深度学习中全局向量模型,确定不同来源文本信息中实体及其关系;针对不同来源文本信息内词汇,设定具有相关性实体与关系的整合方式;在获取不同实体属性对应的关系后,基于自动嵌入技术实现知识图谱智能构建。实验结果表明,该研究方法可以实现标准化拟合,能够实现测试对象的知识图谱智能化构建,具有应用价值。
关键词:深度学习;知识图谱;多源信息融合;图谱智能化;自动嵌入
中图分类号: TP311. 5 **文献标志码**: A **文章编号**: 1003-7241(2025)12-0138-04

Intelligent construction technology of multi-source information fusion knowledge map based on deep learning

LIU Min
(Software College of Hunan Vocational College of Science and Technology, Changsha 410004, China)

Abstract: As a powerful tool for knowledge representation and reasoning, knowledge map is widely used in intelligent question answering, semantic search and decision support. In order to realize the integration of data from different sources and deep reasoning, the intelligent construction technology of multi-source information fusion knowledge map based on deep learning is studied. The global vector model in deep learning is used to determine the entities and their relationships in text information from different sources. Aim at that words in the text information from different source, the integration mode of related entities and relationships is set. After obtaining the corresponding relations of different entity attributes, the intelligent construction of knowledge map is realized based on automatic embedding technology. The experimental results show that the research method can realize standardized fitting and intelligent construction of the knowledge map of the test object, which has application value.
Keywords: deep learning; knowledge graph; multi source information fusion; intelligent map; automatic embedding

0 引言

随着信息技术的迅猛发展,知识图谱作为一种结构化的知识表示方法,逐渐在各个领域展现出其独特的价值。知识图谱的构建技术作为支撑知识图谱应用的基础,不仅为数据组织与管理提供了新的手段,而且在智能问答、语义搜索、推荐系统等领域展现出广阔的应用前景。它能够将多源、异构的数据融合成一个统一的语义网络,为用户提供更为便捷、精准的信息获取与决策支持。目前,知识图谱构建技术的研究已成为学术界和工业界关注的热点。然而,传统的知识图谱构建方法往往依赖于人工参与,效率低下且难以应对大规模数据的处理。因此,国内外学者开展了对知识图谱智能化构建技术,并取得了较好的成果,文献[1]提出了基于关键词的知识图谱构建技术,解决企业海量信息中的数据关系,实现对企业数据的深层次探索和有效应用,能够实现相关企业数据的知识图谱构建。但向量模型在处理复杂语义关系时,存在映射错误的情况,对于较为复杂的企业数据,可能不能及时表达与追溯。文献[2]针对危险化学品风险信息,设计知识图谱的构建技术,该技术应用了标签传递模型与注意力机制,实现对不同类型编码字符的分类与图谱构建。但标签传递模型在对风险信息特征筛选和提取时,需要质量较高的数据,在清洗和处理时具有一定难度,一旦数据质量较差,会影响知识图谱构建的精度。文献[3]提出了一种基于长短期记忆网络的知识图谱构建技术。该技术利用专家知识构建本体,通过双向长短期记忆网络抽取实体和关系,构建了一个具有全面性的飞机电源系统故障诊断知识图谱,在飞机相关故障搜索中具有应用价值。但由于长短期记忆网络在处理自然语言时,对于较为复杂的语句和词汇,不能明确划分关系边界,可能会导致抽取与识别的准确性。文献[4]选择教育领域中的在线协作学习交互文本作为来源领域,提出跨领域的知识图谱构建技术。该技术以个性化推荐方式,对海量交互文本的实体与关系进行识别和抽取,生成具有个性化推荐的知识图谱。但个性化

基金项目:湖南省自然科学基金项目(2021JJ60049)
收稿日期:2024-05-25

推荐过程中,会过于依赖用户的历史行为与兴趣数据,当数据更新不及时时,会导致个性化推荐方式抽取的实体与关系准确度降低,增加知识图谱构建的难度。

深度学习作为一种强大的机器学习方法,在特征提取、模式识别等方面具有显著优势。而多源信息融合技术的研发,主要是为了有效整合不同来源、不同形式的信息。基于此,为提高知识图谱构建的质量和效率,此次基于深度学习设计一种多源信息融合知识图谱智能化构建技术,为人工智能领域的发展作出贡献。

1 基于深度学习确定文本信息实体与关系

在知识图谱构建中,需要对该构建领域内的信息进行实体识别和关系抽取等任务,采用深度学习,尤其是深度学习网络模型,可以直接对原始文本信息中的特征提取,以此确定信息中的实体及其关系。

选择深度学习中全局向量模型,该模型中含有开放域和语料生成库,可以直接对文本信息内的实体进行标记,实现实体与关系输出[5-8]。并且,全局向量模型存在映射矩阵,可以完成文本信息中实体的多次复制,形成一个具有循环学习与提取功能的知识图谱结构,表示为

$$\begin{cases} M_{\mathrm{X}} = [a_s, \{a_{s1}, a_{s2}, \cdots, a_{sk}\}] \\ [\boldsymbol{f}] = [f_1, f_2, \cdots, f_k] \end{cases} \tag{1}$$

$$\begin{cases} [a_s]^{g_d} \to [h] = [d_s] \\ [s-s1, s-s2, \cdots, s-sk] \to [\boldsymbol{f}] = [f_{s1}, f_{s2}, \cdots, f_{sk}] \end{cases} \tag{2}$$

$$[a_s] = [d_s] \to [f_{s1}, f_{s2}, \cdots, f_{sk}] = [d_s, f_{s1}, f_{s2}, \cdots, f_{sk}] \tag{3}$$

式中,M_{X} 为用于识别的全局向量模型;a 为 M_{X} 中句子;a_s 为第 s 个词语;$\{a_{s1}, a_{s2}, \cdots, a_{sk}\}$ 为与 a_s 有关系实体,其中,k 为实体量;$[\boldsymbol{f}]$ 为位置嵌入矩阵;$[f_1, f_2, \cdots, f_k]$ 分别为位置嵌入,$[s-s1, s-s2, \cdots, s-sk]$ 为 a_s 与 $\{a_{s1}, a_{s2}, \cdots, a_{sk}\}$ 的相对距离;$[\boldsymbol{f}_{s1}, \boldsymbol{f}_{s2}, \cdots, \boldsymbol{f}_{sk}]$ 为对 $[s-s1, s-s2, \cdots, s-sk]$ 的映射向量;g_d 为维度;$[h]$ 为单词嵌入矩阵;$\boldsymbol{d}_s$ 为 a_s 在 g_d 内的映射向量。

综合式(1)、(2)呈现出最终的实体与关系转换形式,可以对知识图谱构建中实体与关系进行初步确定。完成映射后的单词,以对应的嵌入位置,可以被连接到对应向量中,基于此,采用信息融合方式表示词汇实体与关系,在知识图谱中的具体表达。

2 多源信息融合方式整合文本信息表达

对知识图谱的构建,其所需的文本信息来源较为广泛,且文本信息的形式也具有多样化,包含有结构化数据库、半结构化网页以及非结构化的文本等[9-12]。采用多源信息融合技术,对不同来源文本信息进行整合,确保不同来源的文本信息能够相互补充和验证,保证知识图谱中信息的一致性、完整性和准确性。融合过程为

$$S(z_{\mathrm{x}}, z_{\mathrm{c}}) = M_{\mathrm{X}} \times \frac{C(z_{\mathrm{x}}, z_{\mathrm{c}}) - \alpha}{C(z_{\mathrm{x}}) \times C(z_{\mathrm{c}})} \tag{4}$$

式中,z_{x}、z_{c} 分别表示文本信息中随机挑选的信息词汇;$C(z_{\mathrm{x}}, z_{\mathrm{c}})$ 为在不同来源文本信息中出现的次数;$C(z_{\mathrm{x}})$、$C(z_{\mathrm{c}})$ 分别为 z_{x}、z_{c} 出现的次数;α 为折扣系数;$S(z_{\mathrm{x}}, z_{\mathrm{c}})$ 为连续词汇出现的频率。以随机文本信息词汇进行整合,存在有

$$\begin{cases} S(z_{\mathrm{x}}, z_{\mathrm{c}}) < \alpha, z_{\mathrm{x}}, z_{\mathrm{c}} \underset{\neq}{\to} z_{\mathrm{xc}} \\ S(z_{\mathrm{x}}, z_{\mathrm{c}}) \geqslant \alpha, z_{\mathrm{x}}, z_{\mathrm{c}} \to z_{\mathrm{xc}} \end{cases} \tag{5}$$

其中,当 z_{x}、z_{c} 连续出现时,若词汇出现的频率为 $S(z_{\mathrm{x}}, z_{\mathrm{c}}) < \alpha$ 时,不能构成双词汇短语;反之,若 $S(z_{\mathrm{x}}, z_{\mathrm{c}}) \geqslant \alpha$,当可以构成双词汇概念时,则表明具有整合意义,可以进一步将其与其他词汇进行整合。基于此,不断地对文本信息中词汇进行整合,实现所有类型词汇的划分,即

$$D_{\mathrm{mn}} = \frac{b_{\mathrm{p,o}}}{\sum_{\mathrm{u}} b_{\mathrm{u,o}}} \tag{6}$$

$$D_{\mathrm{qwe}} = \log \frac{|w|}{|E|} \tag{7}$$

$$D_{\mathrm{mn}} \Leftrightarrow D_{\mathrm{qwe}} = D_{\mathrm{mn}} \times D_{\mathrm{qwe}} \tag{8}$$

式中,i_{p} 为选定的融合依托词汇;D_{mn} 为以多源融合方式处理时,i_{p} 出现在文本中的频率;D_{qwe} 为逆文本频率;$b_{\mathrm{p,o}}$ 为其在不同本文来源中出现的次数;$\sum_{\mathrm{u}} b_{\mathrm{u,o}}$ 为文本中所有词出现次数的和;$|w|$ 为语料库中文本总数;$|E|$ 为包含 i_{p} 的文本数目;$D_{\mathrm{mn}} \Leftrightarrow D_{\mathrm{qwe}}$ 为多源信息融合指标。

结合式(6)、式(7)得出,多源信息融合方式下整合知识图谱中的文本信息,是以关键词作为依托,将其作为信息融合判断指标,实现在多个来源文本信息中的搜索和整合。通过对不同来源文本信息的整合方式设计,再以文本信息中实体和关系的确定方式,以自动嵌入技术构建知识图谱。

3 实体自动嵌入技术智能构建知识图谱

设计知识图谱智能化构建技术,旨在提高构建的效率和准确性,综合对文本信息中的实体与关系判断方式的设计,以及不同来源文本信息的整合方式设计,以实体自动嵌入技术,实现不同领域知识图谱构建[13-15]。不同领域中文本信息涵盖的实体与关系具有差异性,为实现智能化知识图谱的构建过程,将不同领域文本的实体与关系,以参考概念进行描述,如表 1 所示。

根据表 1 中内容所示,在实体中可以划分为四个基础属性,对应不同的属性存在有具体的描述,而每一个属性处于关系类型中,具有差异性,可以表现为多种关系。为此,为实现智能化知识图谱构建,在整合文本信息时,需要快速且准确地筛选出词汇,以相似关系作为自动嵌入原则,计算公式为

$$S(A_1,A_2)=\beta_1\times S_{les}(A_1,A_2)+\beta_2\times S_{con}(A_1,A_2) \quad (9)$$

$$\begin{cases} S_{les}(A_1,A_2)=\dfrac{|\ \mathrm{T}(\cdot)\cap \mathrm{T}(A_2)\ |}{|\ \mathrm{T}(A_1)\cup \mathrm{T}(A_2)\ |} \\ S_{con}(A_1,A_2)=\dfrac{S_{cos}[\boldsymbol{E}_{\mathrm{T}}(A_1),\boldsymbol{E}_{\mathrm{T}}(A_2)]}{2} \end{cases} \quad (10)$$

式中，A_1、A_2 为候选的实体属性；S_{les} 为实体在词法上的相似度；$\boldsymbol{E}_{\mathrm{T}}(A_1)$、$\boldsymbol{E}_{\mathrm{T}}(A_2)$ 为 A_1、A_2 的转换向量；S_{con} 为实体在词义上的相似度；S_{cos} 为在 $\boldsymbol{E}_{\mathrm{T}}(A_1)$、$\boldsymbol{E}_{\mathrm{T}}(A_2)$ 的余弦相似度；$\mathrm{T}(\cdot)$ 为提取实体对应概念词汇集合的函数；β_1、β_2 为关系相关量参数，具体的相关参数需根据文本来源领域确定；$S(A_1,A_2)$ 为 A_1、A_2 的划分关系的相关值。

表 1　参考概念下实体与关系描述

类型	属性	描述
实体	模块	{模块名称,描述}
	类别	{类别名称,描述}
	方式	{方法名称,描述,相关参数,变量,方法类型}
	领域	{领域概念,描述,相关参数,变量,领域类型}
关系	包含	{模块,模块}、{模块,类别}、{模块,方式}、{类别,方式}、{类别,领域}
	继承	{类别,类别}
	提及	{模块,领域}、{类别,领域}、{方式,领域}、{领域,领域}
	重载	{方法类型,方法类型}

通过深度学习和多源信息融合技术，在判断文本中实体与其关系的相关性基础上，可以直接对待选的不同实体属性进行关系判断，在其具有关系时，可以直接对应为具体的关系，完成对相关词汇的关系自动嵌入，可以实现知识图谱的自动化构建和更新。

4　实验分析

为验证所设计技术的有效性，可以实现在不同领域内多源信息融合的知识图谱智能化构建，采用对比测试的方式进行论证，分别选择文献[3]面向天域感知领域的知识图谱构建技术、文献[4]基于在线协作学习交互文本的跨领域知识图谱构建技术作为对照组，与所设计技术同时进行知识图谱的构建测试，验证不同技术的应用效果。

4.1　选择构建对象

多源信息融合知识图谱的构建验证，既要满足信息对象的多类型要求，又要满足信息的多来源要求，基于此，以不同范围的高中数据学科信息作为文本来源，通过所选择的三组构建技术，实现对高中数学领域知识的知识图谱构建，验证不同构建技术的应用效果。此次选择的文本信息来源如下。

1）主要来源：高中数学教材以及电子教材；

2）辅助来源：网络教育平台提供的高中数学知识。

根据文本信息来源可知，此次对测试领域内的知识图谱构建，其知识性要求较高，因此，在设计知识图谱时，必须具有较高的覆盖性，即对高中数学知识相关的实体部分，均需要所有涉及，并形成一个核心体系。以现阶段普通高中数学课程为标准，其应具有的核心本体应涵盖 4 个类别，具体如表 2 所示。

表 2　高中数学知识图谱核心要素

一类	二类
预备知识	集合
	逻辑用语
	相等关系
	不等关系
函数	函数概念
	函数性质
	集合
	幂函数
	对数函数
	指数函数
	三角函数
	函数应用
	一元函数导数
	数列
概率与统计	统计
	概率
	计数原理
几何与代数	立体几何
	复数
	平面向量
	平面解析几何
	空间向量

根据表 2 中内容所示，按照一类和二类标准对高中数学核心要素划分，在对应一类标准中，可以划分出多个二类要素。在实际中，二类要素也会涵盖三类要素，但为节省实验环境，总体呈现为 4 个一类要素、21 个二类要素。

在测试前通过专家和爬虫技术两种标注方式，分别对高中数学教材、电子教材和教育平台中的相关文本进行标注，共计有 800 组数学教案。统计完毕后，将处理好的文本数据构建为集合，按照 75%、15%、10%的比例划分为训练集合、验证集合和测试集合，详情如表 3 所示。

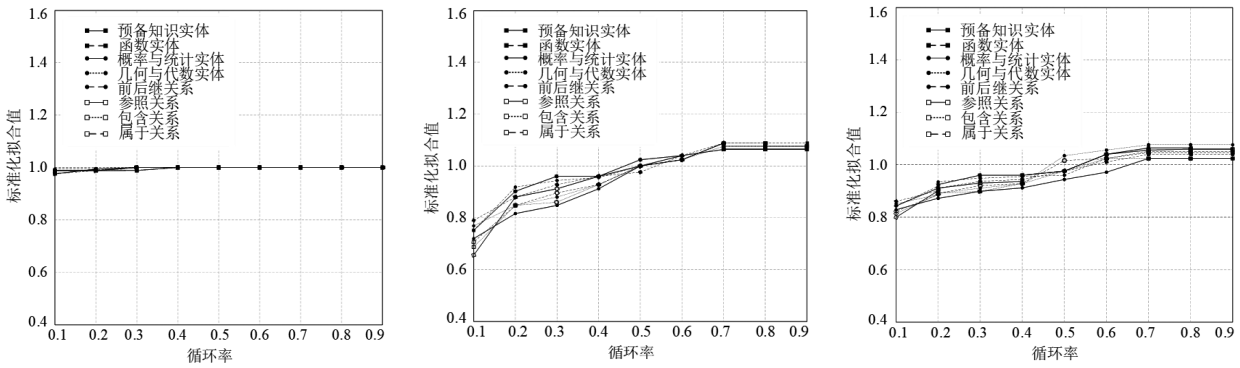
表 3　数据集详细信息

类型	实体数
标注句子数	12 250
标注教案数	600
实体总数	5 547
关系总数	7 148
预备知识实体	1 212
函数实体	1 535
概率与统计实体	2 142
几何与代数实体	658
前后继关系	752
参照关系	682
包含关系	3 102
属于关系	2 612

根据表 3 中内容所示，对于三个类型要素来讲，其存在多种关系，而对于不同领域内知识图谱的构建，其主要步骤即为知识实体的识别与关系抽取。因此，在验证所设计技术与其他两组技术时，分别将知识实体的识别与知识关系的抽取作为验证指标，完成后续的测试内容。

4.2 验证构建效果

在测试前对实验环境进行配置,具体情况为:在硬件方面,选择 CPU 为 INTEL-XEON-E5-296-V2;在软件方面,选择编程语言为 PYTHON3.8,学习框架为 TENSORFLOW。由于各组构建技术均以模型为主,为此,以标准化拟合指标作为评价结果。该指标最佳值为 1,表示可以覆盖全部的实体对象,以及对应关系的抽取,若超过 1 或小于 1,说明知识图谱构建的拟合效果较差,当大于 1 时为过度拟合状态,小于 1 时为拟合不足情况。分别验证各组方法,结果如图 1 所示。



(a) 新技术　(b) 面向天域感知领域的知识图谱构建技术　(c) 基于在线协作学习交互文本的跨领域知识图谱构建技术

图 1 不同技术下知识图谱构建要素拟合结果

根据图 1 可知,以循环率作为划分标准,不同技术下构建的高中数学知识图谱,其图谱要素拟合结果具有差异性。其中,面向天域感知领域的知识图谱构建技术、基于在线协作学习交互文本的跨领域知识图谱构建技术在循环率为 0.5 时出现了过度拟合问题,而在 0.3、0.4 时,又存在拟合度不足的情况,说明在对高中数学知识图谱构建时,容易产生实体的识别缺失,或者是对应关系属性的错误判断。

所设计技术在构建高中数学知识图谱时,对应每一个划分标准,均可以实现标准化的拟合效果,并且,所设计的构建技术,可以在较少的循环次数下,完成对文本信息对象的知识图谱构建,同时具有全面覆盖和快速组建的效果。主要是所设计技术同时融合了多源信息技术和深度学习技术,可以直接对实体知识特征进行分析和提取,实现标准化的知识图谱构建,具有应用价值。

5 结束语

全文针对不同来源的信息文本,设计了基于深度学习的多源信息融合知识图谱智能化构建技术,该技术综合了深度学习算法、多源融合方式以及自动化识别技术,可以在海量文本信息中实现对信息实体关系的排列组合,构建同时具备准确性和完整性的知识图谱,且经过实验论证,所研究技术具有应用价值。但由于时间有限,在分析和论证过程中仅能以某个类型的多源信息数据作为对象,仅针对这一领域内的文本信息进行知识图谱的构建,具有一定的不足之处,后续会针对多个领域进行测试,为实现不同领域内文本信息的知识图谱构建,以及信息挖掘提供技术支持。

参考文献

[1] SELLAMI S, ZAROUR N E. Keyword-based faceted search interface for knowledge graph construction and exploration[J]. International Journal of Web Information Systems, 2022, 18(5):453-486.

[2] CHEN G, HU Q, LU Q, et al. A hazardous chemical knowledge base construction method based on knowledge graph[J]. International Journal of Reasoning-based Intelligent Systems, 2022, 14(4): 184-193.

[3] 潘越, 刘云汉, 于荣欢, 等. 面向天域感知领域的知识图谱构建技术研究[J]. 中国电子科学研究院学报, 2023, 18(8):707-716.

[4] 郑兰琴, 范云超, 牛佳玉. 基于在线协作学习交互文本的跨领域知识图谱构建技术[J]. 电化教育研究, 2022, 43(12):70-77.

[5] 许雪晨, 田侃. 智能投顾领域的知识图谱构建与应用研究[J]. 学习与探索, 2023(5):122-133.

[6] 郑庆华, 师斌, 董博. 面向智慧税务的大数据知识工程技术及应用[J]. 中国工程科学, 2023, 25(2):221- 231.

[7] 张雅晴, 单中原, 赵俊峰, 等. 基于智能映射推荐的知识图谱实例构建与演化方法[J]. 计算机科学, 2023, 50(6):142-150.

[8] 王书鸿, 郑少明, 刘中硕, 等. 面向某地区电网继电保护装置缺陷知识图谱构建的实体关系抽取[J]. 电网技术, 2023, 47(5):1874-1885.

[9] 武月佳, 周建涛. DL$^+$:一种增强型双层知识图谱推理框架[J]. 计算机科学, 2023(12):302-313.

[10] 史慧洋, 魏靖烜, 蔡兴业, 等. 威胁情报提取与知识图谱构建技术研究[J]. 西安电子科技大学学报, 2023, 50(4):65-75.

[11] 梁柱, 杜赛赛, 沈国栋, 等. 基于多源数据融合及知识图谱的桥梁智慧康养平台设计[J]. 中国建设信息化, 2023(4):64-67.

[12] 胡杰, 李源洁, 耿號, 等. 基于深度学习的汽车故障知识图谱构建[J]. 汽车工程, 2023, 45(1):52-60.

[13] 刘言东. 基于多源信息融合的电子政务信息风险评价研究[J]. 自动化技术与应用, 2023, 42(5):92-95.

[14] 张佳玲, 刘谦, 陈奕, 等. 多源数据融合与驱动背景下枸杞科学协作与热点前沿知识图谱构建及可视化分析[J]. 中草药, 2023(24):8165-8179.

[15] 罗顺财, 李庆印, 魏福禄, 等. 基于知识图谱的多源信息融合事故处理系统[J]. 山东理工大学学报(自然科学版), 2024, 38(3):28-34.

作者简介:刘　敏(1982—),女,硕士,副教授,研究方向:知识图谱、推荐系统。