

融合语义映射和 CNN 的知识图谱智能搜索方法

宋 钰¹
SONG Yu

摘 要

针对已有方法存在的语义特征提取准确性低、信息搜索性能不足等问题,文章提出融合语义映射和 CNN 的知识图谱智能搜索方法研究。构建语义映射关系,关联用户查询与知识图谱中的实体和关系,并通过语义相似度筛选出候选集合。基于改进 CNN 模型对候选实体与关系的语义特征进行深度挖掘与提取,并判断相关性类别,以此为基础,对候选集合进行进一步的筛选,生成初始的知识图谱信息搜索结果,通过综合评分方法对信息单元进行评分与降序排序,获取最终的知识图谱信息搜索结果,从而实现研究目标。消融实验结果显示:改进 CNN 模型候选集合特征提取准确性最大值达到了 97.5%;对比实验结果显示:所提方法应用后反馈第一个信息单元与查询的余弦相似度最大值达到了 0.97。

关键词

语义映射关系构建;知识图谱;特征提取与分类;改进 CNN;智能信息搜索

doi: 10.3969/j.issn.1672-9528.2025.12.040

0 引言

近年来,网络技术得到了飞速的发展,应用范围逐渐扩大,致使人们生产与生活中的信息呈现急剧增加的趋势,信息检索面临着海量数据与用户精准需求之间的巨大矛盾^[1]。而知识图谱是一种结构化语义知识库,其在信息检索领域占据着较为重要的地位,已有方法由于语义理解偏差大、搜索结果精度差等问题,无法满足现今信息检索领域的发展需求。现有算法路径搜索问题与序列决策问题的转换需要大量的训练数据和计算资源,而训练数据的获取难度较大,进而影响设计算法的整体性能。为了解决上述问题,提出融合语义映射和 CNN 的知识图谱智能搜索方法研究。

1 知识图谱智能搜索方法设计

1.1 语义映射关系构建

通过特定嵌入方法学习得到,可以捕捉实体和关系之间的语义信息,为后续语义相似度衡量提供一定的便利。其中,知识图谱实体为 $E = \{e_1, e_2, \dots, e_i, \dots, e_n\}$, n 代表实体的总数

量;知识图谱关系为 $R = \{r_1, r_2, \dots, r_j, \dots, r_m\}$, m 代表关系的总数量^[2]。

设计方法利用余弦相似度衡量查询与知识图谱元素的语义相似度,其取值范围为 $[-1, 1]$ 。常规情况下,余弦相似度数值越接近 1,则表示查询与知识图谱元素的语义相似度越高;反之,余弦相似度数值越接近 -1,则表示查询与知识图谱元素的语义相似度越低。查询与知识图谱实体、关系之间的余弦相似度计算公式为:

$$\begin{cases} \rho(q, e_i) = \frac{q \cdot e_i}{\|q\| \|e_i\|} \\ \rho(q, r_j) = \frac{q \cdot r_j}{\|q\| \|r_j\|} \end{cases} \quad (1)$$

式中: $\rho(q, e_i)$ 代表查询 q 与知识图谱实体 e_i 的余弦相似度; $q \cdot e_i$ 代表查询 q 与知识图谱实体 e_i 的点积; $\|q\|$ 代表查询 q 的模; $\|e_i\|$ 代表知识图谱实体 e_i 的模; $\rho(q, r_j)$ 代表查询 q 与知识图谱关系 r_j 的余弦相似度; $q \cdot r_j$ 代表查询 q 与知识图谱关系 r_j 的点积; $\|r_j\|$ 代表知识图谱关系 r_j 的模。

以式(1)计算的余弦相似度 $\rho(q, e_i)$ 与 $\rho(q, r_j)$ 为基础,联合预先设定的相似度阈值 θ (取值范围为 $[0, 1]$),筛选出与查询语义相关的实体及其关系,从而生成候选集合^[3]。具体来说,若 $\rho(q, e_i)$ 大于或者等于 θ ,将实体 e_i 加入候选实体集合 C_E ;若 $\rho(q, r_j)$ 大于或者等于 θ ,将关系 r_j 加入候选关系集合 C_R 。则候选集合表达式为:

$$C = C_E \cup C_R \quad (2)$$

式中: C 代表候选集合,其包含了与查询语义相关的所有实体和关系。

1. 广州工商学院 广东广州 510850

[基金项目] 2024 年度广东省本科高校教学质量与教学改革工程建设项目“数据库课程知识图谱的构建及应用研究”(粤教高函(2024)30号-1125); 2024—2025 学年广州工商学院校级质量工程项目“数据库课程知识图谱的构建及应用研究”(JXGG202401); 2023 年度广东省本科高校教学质量与教学改革工程项目“数据库课程群教研室”(粤教高函(2024)9号-308)成果。

1.2 改进 CNN 特征提取与分类

在改进 CNN 模型应用之前, 对候选实体和关系表示进行优化处理, 为其高级语义特征挖掘及其提取提供便利。候选实体和关系综合表示向量为:

$$\begin{cases} \mathbf{e}_i = [\mathbf{e}_{ie}; \mathbf{e}_{ia}; \mathbf{e}_{is}] \\ \mathbf{r}_j = [\mathbf{r}_{js}; \mathbf{r}_{jc}; \mathbf{r}_{jp}] \end{cases} \quad (3)$$

式中: $\mathbf{e}_{ie}$ 代表候选实体嵌入向量; $\mathbf{e}_{ia}$ 代表候选实体属性信息向量; $\mathbf{e}_{is}$ 、 $\mathbf{r}_{js}$ 代表候选关系语义特征向量; $\mathbf{r}_{jc}$ 代表候选关系上下文语义信息向量; $\mathbf{r}_{jp}$ 代表候选关系路径信息向量。

改进 CNN 模型采用更深的网络结构, 增加卷积层和池化层的数量, 并引入注意力机制, 以提取更高级的语义特征^[4]。将优化处理后的候选实体 $\mathbf{e}_i$ 和关系 $\mathbf{r}_j$ 存储于候选集合 C 中, 输入到改进 CNN 模型中, 卷积层利用不同尺寸的卷积核在候选集合 C 上进行滑动卷积操作, 提取局部特征, 表达式为:

$$\mathbf{F}_{i,j} = \text{ReLU} \left\{ \sum_{a=0}^{k-1} \sum_{b=0}^{k-1} C(\mathbf{e}_i, \mathbf{r}_j) \cdot \kappa(a, b) + Q \right\} \quad (4)$$

式中: $\mathbf{F}_{i,j}$ 代表候选集合的局部特征向量; $\text{ReLU}\{\cdot\}$ 代表 ReLU 激活函数; k 代表卷积核尺寸大小; $C(\mathbf{e}_i, \mathbf{r}_j)$ 代表卷积核在候选集合上的滑动结果; $\kappa(a, b)$ 代表卷积核函数; Q 代表偏置项。

为了更好地整合不同卷积核提取的局部特征 $\mathbf{F}_{i,j}$, 改进 CNN 模型引入注意力机制, 计算每个特征向量的注意力权重, 表达式为:

$$\alpha_{i,j} = \text{softmax} \left(\frac{\mathbf{F}_{i,j} \cdot \mathbf{PK}^T}{\sqrt{d_k}} \right) \mathbf{V} \quad (5)$$

式中: $\alpha_{i,j}$ 代表局部特征向量的注意力权重; $\text{softmax}(\cdot)$ 代表 softmax 函数; $\mathbf{P}$ 、 $\mathbf{K}$ 与 $\mathbf{V}$ 代表查询矩阵、键矩阵与值矩阵; d_k 代表键矩阵的维度。

利用式 (5) 计算的注意力权重对不同局部特征向量进行加权求和, 获得融合后的候选集合特征向量, 表达式为:

$$\mathbf{f} = \sum_{i=1}^n \sum_{j=1}^m \alpha_{i,j} \cdot \mathbf{F}_{i,j} \quad (6)$$

式中: $\mathbf{f}$ 代表候选集合融合特征向量。

应用改进 CNN 模型中的全连接层将候选集合融合特征向量 $\mathbf{f}$ 转换为概率分布, 确定信息单元 (候选实体与关系) 与查询的相关性类别, 表达式为:

$$p_c = \frac{f_c(\mathbf{e}_i, \mathbf{r}_j)}{\sum_{c=1}^{c_{\text{total}}} f_c(\mathbf{e}_i, \mathbf{r}_j)} \quad (7)$$

式中: p_c 代表信息单元融合特征向量 $\mathbf{f}_c(\mathbf{e}_i, \mathbf{r}_j)$ 与查询之间的

相关性属于第 c 个类别的概率; $f_c(\mathbf{e}_i, \mathbf{r}_j)$ 代表信息单元融合特征向量 $\mathbf{f}_c(\mathbf{e}_i, \mathbf{r}_j)$ 在第 c 个相关性类别上的分布概率; c_{total} 代表相关性类别总数量。

在公式 (7) 计算结果中寻找最大值 $\max\{p_c\}$, 其对应类别即为信息单元 (候选实体与关系) $C(\mathbf{e}_i, \mathbf{r}_j)$ 与查询之间的相关性类别, 记为 c^* ^[9]。

上述过程基于改进 CNN 模型完成了候选集合特征提取与分类, 能够更有效地表征信息单元 (候选实体与关系) 的高级语义特征, 从而提升知识图谱语义理解与匹配任务的性能。

1.3 知识图谱搜索结果生成与排序

以前输出的信息单元 (候选实体与关系) 与查询相关性类别 c^* 为依据, 对候选集合 C 进行进一步的匹配与筛选, 从而生成初始的知识图谱信息搜索结果, 并通过综合评分方法对候选集合中的信息单元进行评分与降序排序, 获取最终的知识图谱信息搜索结果, 将其反馈给用户, 使用户能够快速、准确地获取到所需信息。

设计方法设定相关性类别编号从 $1 \rightarrow c_{\text{total}}$, 对应的相关性逐渐升高。根据用户查询需求, 设置信息单元 (候选实体与关系) 与查询相关性阈值 ζ 。当相关性类别 c^* 小于或等于阈值 ζ , 对信息单元 (候选实体与关系) $C(\mathbf{e}_i, \mathbf{r}_j)$ 进行删除处理; 当相关性类别 c^* 大于阈值 ζ , 对信息单元 (候选实体与关系) $C(\mathbf{e}_i, \mathbf{r}_j)$ 进行保留处理, 从而完成了候选集合 C 的再次筛选, 记为 $\hat{C}$ 。

针对候选集合中的信息单元 $C(\mathbf{e}_i, \mathbf{r}_j)$, 综合考虑语义相似度、信息单元占比与查询上下文信息等因素, 对其进行综合评分, 计算公式为:

$$S[\hat{C}(\mathbf{e}_i, \mathbf{r}_j)] = \beta \times \rho[\hat{C}(\mathbf{e}_i, \mathbf{r}_j), q] + \chi \times \omega[\hat{C}(\mathbf{e}_i, \mathbf{r}_j)] + \delta \times W \quad (8)$$

式中: $S[\hat{C}(\mathbf{e}_i, \mathbf{r}_j)]$ 代表信息单元 $\hat{C}(\mathbf{e}_i, \mathbf{r}_j)$ 的综合评分。一般情况下, 综合评分越高, 表明信息单元越符合用户需求; β 、 χ 与 δ 代表权重系数; $\rho[\hat{C}(\mathbf{e}_i, \mathbf{r}_j), q]$ 代表信息单元 $\hat{C}(\mathbf{e}_i, \mathbf{r}_j)$ 与查询之间的语义相似度 (余弦相似度); $\omega[\hat{C}(\mathbf{e}_i, \mathbf{r}_j)]$ 代表信息单元 $\hat{C}(\mathbf{e}_i, \mathbf{r}_j)$ 的占比; W 代表查询上下文信息量化因子。

依据式 (8) 计算结果 $S[\hat{C}(\mathbf{e}_i, \mathbf{r}_j)]$ 对信息单元进行降序排列, 生成最终的知识图谱信息搜索结果——候选集合 $\tilde{C}$, 将其反馈给用户, 以此来满足用户的信息搜索需求^[10]。综上所述, 此研究在语义映射关系与改进 CNN 模型的联合应用下, 实现了知识图谱信息的智能搜索, 能够为用户信息搜索提供更加有效的方法支撑。

2 实例测试与结果分析

选取基于可搜索加密的知识图谱搜索方法与基于收益型

造技术的知识图谱搜索方法作为参考方法，简称为方法 1 与方法 2，联合本文方法共同进行知识图谱搜索实验，分析应用方法的性能优劣。

2.1 实验数据集选取与介绍

依据本文方法应用性能测试需求，选取适当的公开数据集作为实验数据集，具体如表 1 所示。

表 1 实验数据集示意图

数据集	知识图谱数据集	查询数据集
名称	DBpedia	LC-QuAD 2.0
规模	68 GB	52 GB
涵盖领域	人物、地点、组织、作品等	建筑、人物、地点、风景、作品等
数据存储格式	结构化格式	结构化格式与非结构化格式
语言支持	多种语言版本	多种语言版本
更新模式	持续更新	持续更新
查询类型	—	单跳查询、多跳查询、聚合查询等

DBpedia 与 LC-QuAD 2.0 在知识图谱研究与应用中具有广泛的影响力与认可度，能够满足后续实验的需求。

2.2 实验参数设置

本文方法应用了改进 CNN 模型，为了保障本文方法实验数据的精准性：输入层数量为 1；卷积层数量为 5；卷积核大小为 1×1、3×3 与 5×5（组合应用）；池化层为 4；池化窗口大小为 2×2（最大池化窗口大小）与 3×3（平均池化窗口大小）。

2.3 消融实验

为了验证本文方法中改进 CNN 模型的应用性能，设计消融实验。参与消融实验的模型如表 2 所示。

表 2 消融实验参与模型示意图

模型	模型 1	模型 2	模型 3	模型 4
输入层数量	1	1	1	1
卷积层数量	1	2	3	5
卷积核大小	1×1	1×1 与 3×3	1×1、3×3 与 5×5	1×1、3×3 与 5×5
池化层数量	1	1	2	4
池化窗口大小	2×2	3×3	2×2 与 3×3	2×2 与 3×3
全连接层数量	1	1	1	1
输出层数量	1	1	1	1
注意力机制	×	×	×	√

应用表 2 所示模型 1~4 进行候选集合特征提取实验，获取候选集合特征提取准确性，如图 1 所示。模型 1 应用后候

选集合特征提取准确性最低，主要是因为其整体结构较为单薄，致使其特征提取性能较差；相较于模型 1，模型 2 与模型 3 应用后候选集合特征提取准确性得到了一定的提高，主要是因为其整体结构层次更深，再加之卷积核与池化窗口的混合应用，有效地提升了模型的特征提取性能；模型 4（本文方法应用模型）应用后候选集合特征提取准确性最高，其不但增加了模型结构层次的深度，还引入了注意力机制，最大程度提升了模型特征提取性能，使得模型 4 候选集合特征提取准确性最大值达到了 97.5%。

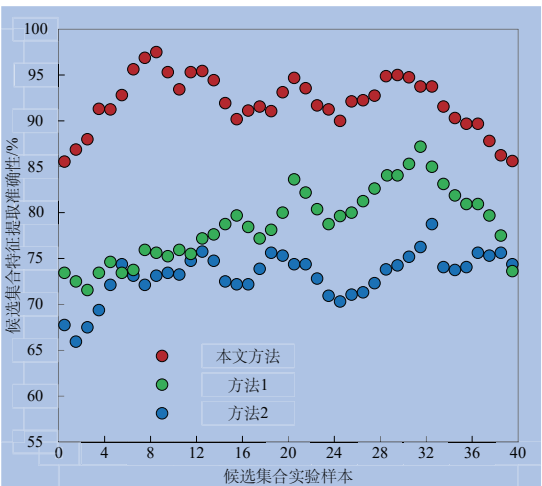


图 1 候选集合特征提取准确性示意图

2.4 对比实验

应用本文方法、方法 1 与方法 2 进行知识图谱搜索实验，根据反馈给用户的第一个信息单元（实体与关系）与查询之间的余弦相似度，分析应用方法的信息搜索性能，如图 2 所示。

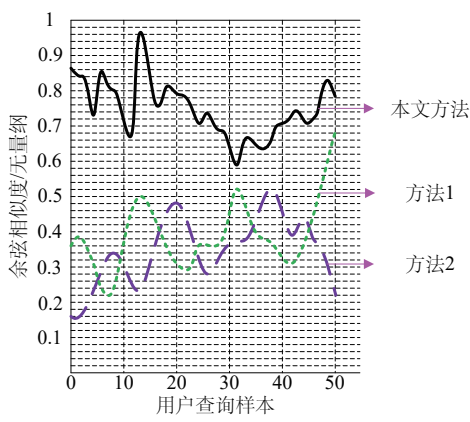


图 2 余弦相似度示意图

常规情况下，知识图谱反馈第一个信息单元与查询的余弦相似度越高，表明应用方法反馈的信息搜索结果越符合用户的需求，则其信息搜索性能越高。如图 2 所示，相较于方法 1 与方法 2，本文方法应用后反馈第一个信息单元与查询的余弦相似度显著较高，最大值达到了 0.97。

基于高性能磁存储与蓝光近线存储结合的数据中心存储系统设计

刘梦莹¹ 贾学志¹ 王天宇¹
LIU Mengying JIA Xuezhi WANG Tianyu

摘要

面对数据中心存储成本持续攀升与高性能资源利用率不足的双重挑战,文章创新性地提出一种基于高性能磁存储与蓝光近线存储深度融合的分级存储架构。该方案摒弃复杂数据迁移算法,聚焦于业务导向的静态分级模型与系统级架构设计,将高时效性、低容量的活跃业务数据部署于磁存储层,而将低访问频率、大容量的归档与备份数据迁移至蓝光存储层。通过物理隔离、业务驱动的分级策略,此架构在保障核心业务性能的前提下,显著降低总体存储成本(TCO),提升资源利用效率,为企业数据中心提供了一种高性价比、易于落地的存储优化方案。实际应用分析表明,所提架构可降低长期存储成本达60%,同时保持关键业务响应时间毫秒级延迟。

关键词

磁存储; 蓝光存储; 成本优化; 数据生命周期

doi: 10.3969/j.issn.1672-9528.2025.12.041

0 引言

随着大数据、人工智能技术的不断发展,企业数据总量呈指数级增长,数据中心面临着前所未有的存储压力。一方面,为满足业务实时性需求而配置的高性能存储(如全闪存阵列、分布式全对称磁存储)资源利用率普遍偏低,造成巨额投资浪费;另一方面,海量的历史数据、备份数据、归

档数据长期占用昂贵的在线存储空间,导致存储总拥有成本(TCO)居高不下。尤其对于预算有限、技术力量相对薄弱的企业,如何在保障核心业务性能的同时,有效控制存储成本,已成为亟待解决的痛点问题^[1]。

当前主流的分级存储方案多依赖复杂的自动数据迁移算法,根据数据热度动态调整存储位置。然而,此类方案在企业实际部署中面临诸多挑战:算法设计与调优复杂度过高;频繁数据迁移消耗额外CPU内存计算与数据I/O资源;

1. 长光卫星技术股份有限公司 吉林长春 130102

3 结语

知识图谱的兴起与应用,为信息检索带来了革命性的变化,其主要以实体为中心,可以有效地将分散的信息整合成有机的整体,并能够清晰地区分语义意思,让用户快速获取所需信息。与此同时,知识图谱具备强大的语义关联能力,可以在理解语义字面意思的基础上,挖掘背后的语义关系,大大提高了信息搜索的准确性和效率。但随着知识图谱规模的扩大,信息搜索难度也随之增加,已有方法无法取得较好的信息搜索效果,故提出融合语义映射和CNN的知识图谱智能搜索方法研究。实验结果显示:设计方法显著提升了候选集合特征提取准确性与知识图谱信息搜索性能,为相关研究提供一定的帮助。

参考文献:

- [1] 诸云强,代小亮,杨杰,等.地球科学知识图谱一站式共享服务系统[J].高校地质学报,2023,29(3):325-336.
- [2] 何杰,王佳蓉,王恒恒.基于节点语义相似度的本体映射方法[J].吉林大学学报(理学版),2024,62(2):399-409.
- [3] 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.
- [4] 唐文春,栗强强,闫会峰.改进知识图谱的城市交通拥堵智能预测仿真[J].计算机仿真,2024,41(12):229-233.

【作者简介】

宋钰(1984—),女,湖南郴州人,硕士,副教授,研究方向:数据挖掘与数据库技术、知识图谱。

(收稿日期:2025-06-30 修回日期:2025-12-08)