

基于深度学习的海量本体匹配模型构建研究

常万军, 张东方

(河南工学院 计算机科学与技术学院, 河南 新乡 453003)

摘要: 本体匹配是语义网中多源异构本体之间数据交流和操作的有效方法。在本体匹配算法设计中, 如何实现海量本体科学分类, 减少异质本体匹配过程中的时间、空间和资源消耗, 是研究人员面临的重大挑战。将自动编码器技术和深度学习技术引入海量本体匹配算法设计, 使用自动编码器对输入本体进行语义嵌入和模型训练, 以降低异质本体的复杂性, 提高本体分类的准确性, 为海量本体的语义匹配和数据挖掘提供重要支持。实验结果显示, 该算法在时间复杂度、运行精度方面的优势明显, 具有较高的应用价值。

关键词: 本体; 自动编码器; 语义嵌入; 深度学习

中图分类号: TP311.1

文献标志码: A

文章编号: 2096-7772-(2024)02-0021-06

0 引言

近年来, 语义网技术快速发展。各领域专家团队基于发展需要, 纷纷构建特点鲜明的语义本体框架, 如病历的语义标注^[1]、医学数据格式标准化^[2]、OGC 地理信息服务本体库^[3]和医学系统术语本体^[4]等, 本体类型越来越多, 本体规模日益增大。由于专家团队研究基础、研究目标、区域文化的差异, 以及技术保密、技术壁垒等因素, 各领域本体之间、同领域不同本体框架之间均存在着严重的异质问题, 影响了智能语义网信息的交互协作和发展。如何解决本体异构问题, 消除本体数据操作与交流障碍, 实现语义实体对应关系的准确匹配, 成为语义网技术发展的关键所在。

1 本体匹配技术研究

在本体匹配技术研究进程中, Khat & Benaissat 团队和 Lutke 团队提出将以参考本体为依托的自动映射方法引入本体匹配研究的思路, 构建 AGM 本体匹配模型^[5,6], 实现了快捷确认异构本体相关实体概念之间对应关系的目标, 但在处理医学、天文学等领域海量本体时存在着准确性不足和匹配效率太低的严

重问题。Monych 设计的 DESKMatcher 本体匹配模型^[7], 将人工智能深度学习技术引入本体匹配研究, 可以适度提高海量本体匹配的准确性, 但是匹配效率难以令人满意。Chang 等构建的 FAGMap-KG 本体匹配模型^[8]将分区法和深度学习方法结合, 对海量的输入本体进行分区匹配、分区合并, 有效提升了本体匹配效率, 却也因分区过程中的参数设置、语义精度丢失等造成匹配准确性下降。Chakraborty 构建的 OntoConnect 模型^[9]和王颖等设计的基于进化算法生物学本体匹配技术^[10], 提出将模糊神经网络的冗余裁剪用于本体匹配算法, 可以很好解决本体匹配的准确性问题, 但它的匹配效率仍然达不到预期目标。

基于以上研究, 本文构建基于深度学习的本体匹配模型 (Deep learning for Ontology-matching Model, Deep OM)。本模型利用参考本体将输入本体概念转换为深度学习的数值向量, 通过自动编码器进行样本训练, 得到高精度本体概念后输出。为确保输出本体的准确性, 本模型利用设定的余弦相似度阈值对输入、输出本体进行过滤, 保留符合标准的输出本体及对应关系, 得到本体匹配的最终结果。本模型可以免去海量本体匹配分区预处理的繁琐过程, 具有准确性高、运行时间短和内存资源占用少的优势。

收稿日期: 2024-01-10

基金项目: 河南省科技攻关项目 (232102220018)

第一作者简介: 常万军 (1979—), 男, 河南内乡人, 副教授, 硕士, 主要从事计算机应用技术研究。

2 基于深度学习的本体匹配模型构建

2.1 基于深度学习的本体匹配模型构建流程

构建 Deep OM 是依据分段自动处理理念进行的。

首先通过参考本体概念相似度计算,对输入本体概念进行语义嵌入;再通过以深度学习为核心的自动编码器技术实现对输入本体的高水平训练和高精度匹配。Deep OM 构建流程如图 1 所示。

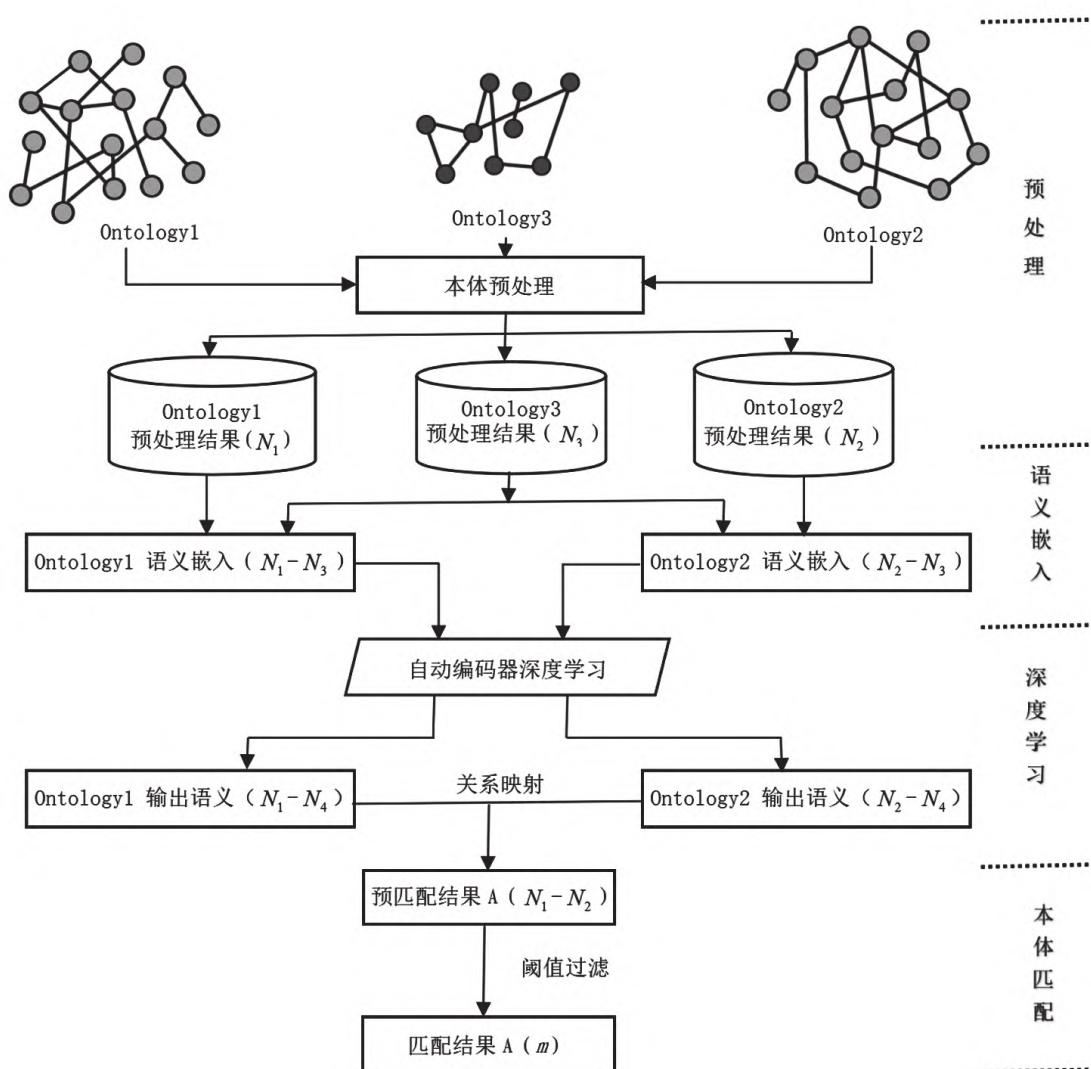


图 1 Deep OM 构建流程

图 1 展示了以 Ontology1 和 Ontology2 为输入本体,以 Ontology3 为参考本体,通过 Deep OM 实现本体匹配的过程。在该过程中,得到了具有 m 个对应关系的本体匹配结果,用 A 表示, $A = \{a_1, a_2, a_3, \dots, a_m\}$ 。 A 中的元素 a_i 是四维向量组, $a_i = \{id_i, C, C', \dots, val_i\}$ ($i = 1 \sim m$), id_i 为通信标识符, C 为 Ontology1 的概念集合, C' 为 Ontology2 的概念集合, val_i 为 Ontology1 和 Ontology2 在 Deep OM 中的相似度值,其置信区间为 $[0,1]$ 。

输入本体 Ontology1、Ontology2 的匹配过程可细分为预处理、语义嵌入、深度学习和本体匹配等 4 个阶段。

2.2 预处理阶段

该阶段的主要目的是为本体匹配准备本体数据,包括本体信息提取和本体预匹配两项任务。

本体信息提取是预处理阶段的基础任务。本项任务主要目的是从输入本体 Ontology1 和 Ontology2 中提取可以描述该本体概念的代表性词汇,即概念要

素。一般来说, 从本体中提取的每一个概念要素与其他概念要素之间都存在一对多的数据关系, 具有典型树形结构的祖先、后代或兄弟类三元数据关联关系特征。

本体预匹配的主要目标是实现对本体概念要素的标注与降噪处理。本体概念要素标注包括对每个概念要素标签名称的标注和各要素之间所具有的祖先、后代或兄弟类三元数据关联关系的标注。本体概念要素标注之后, 需要对其进行降噪操作, 将大写形式的本体概念词汇以小写形式呈现, 将与匹配支持没有任何关联的无用本体概念词汇、长度为1的非数字词汇、标点符号以及特殊字符做删除处理, 以减少对本体关联关系判断的干扰。

2.3 语义嵌入阶段

2.3.1 选择参考本体

依据本体匹配准确性需要, 在语义嵌入阶段引入与 Ontology1 和 Ontology2 语义均接近的小规模独立本体 Ontology3 作为参考本体。通过参考本体 Ontology3 对降噪后的输入本体 Ontology1 和 Ontology2 进行语义嵌入, 将它们转换为深度学习模型可理解的向量数值表示形式。

Ontology1 的概念集用集合 $C = \{C_{i,i=1 \sim N_1}\}$ 表示, Ontology2 的概念集用集合 $C' = \{C'_{i,i=1 \sim N_2}\}$ 表示, Ontology3 的概念集用集合 $C'' = \{C''_{i,i=1 \sim N_3}\}$ 表示, N_1 、 N_2 、 N_3 分别为三个本体对应的概念长度。

2.3.2 计算语义相似性

语义相似性是体现输入本体和参考本体之间相似度的关键指标, 也是进行语义嵌入的重要前提。将 Ontology1 的概念集 C 和 Ontology2 的概念集 C' 组合形成新的输入本体概念集 C_1 , 将 Ontology3 的概念集 C'' 用等价替代概念集 C_2 来标示。对于组合输入本体概念集 C_1 和等价参考本体概念集 C_2 , 其语义相似性需要从概念相似度、结构相似度、术语相似度三个方面分别计算后, 再进行综合分析得到。

(1) 概念相似度。概念相似度是反映上下文中对应概念相似性及语法相似性的数据, 可通过公式 (1) 得到。

$$TerSim(C_1, C_2) = \frac{2 * |Label_1 \cap Label_2| + D_1 + D_2}{|Label_1| + |Label_2|} \quad (1)$$

式中: $Label_1$ 、 $Label_2$ 分别是本体概念集 C_1 、 C_2 的预处理标签, D_1 是向量 $(Label_1 \rightarrow Label_2)$ 与向量 $(Label_2 \rightarrow Label_1)$ 的相似度值, D_2 是向量 $(Label_2 \rightarrow Label_1)$ 与 $(Label_1 \rightarrow Label_2)$ 的相似度值。

(2) 结构相似度。结构相似度反映的是本体概

念元素的祖先、后代或兄弟之间的相似结果, 可通过公式 (2) 得到。

$$StrSim(C_1, C_2) = \frac{2 * |RC_1 \cap RC_2| + D_1 + D_2}{|RC_1| + |RC_2|} \quad (2)$$

式中: RC_1 、 RC_2 分别是 C_1 、 C_2 的关联概念, 它们之间是树形结构祖先、后代或兄弟类三元数据关联关系; D_1 是向量 $(RC_1 \rightarrow RC_2)$ 与 $(RC_2 \rightarrow RC_1)$ 的相似度值; D_2 是向量 $(RC_2 \rightarrow RC_1)$ 与 $(RC_1 \rightarrow RC_2)$ 的相似度值。

结构相似度均值是反映本体结构相似度的更为客观的重要指标, 可通过计算本体概念元素的祖先、后代和兄弟元素之间的相似结果平均值得到, 用 $Strval$ 标示。

(3) 术语相似度。术语相似度是反映本体概念与对应实例之间的相似性结果, 可通过公式 (3) 得到。

$$ExtSim(C_1, C_2) = \frac{|Inst_1 \cap Inst_2|}{|Inst_1| + |Inst_2|} \quad (3)$$

式中: $Inst_1$ 、 $Inst_2$ 分别是 C_1 、 C_2 对应的实例集, 其相似度值可通过 Jaccard 函数得到。

(4) 综合语义相似度。在计算概念相似度、结构相似度、术语相似度的基础上, 可按照设定的阈值条件, 得到两个本体概念集的相似度值。

对 $ExtSim(C_1, C_2)$ 、 $TerSim(C_1, C_2)$ 和 $Strval$, 若 $Strval$ 和 $ExtSim(C_1, C_2)$ 均大于设定阈值, 则以 $Strval$ 、 $TerSim(C_1, C_2)$ 和 $ExtSim(C_1, C_2)$ 的平均值作为综合相似度值; 若只有 $Strval$ 大于设定阈值, 则以 $Strval$ 和 $TerSim(C_1, C_2)$ 的平均值作为综合相似度值; 若只有 $ExtSim(C_1, C_2)$ 大于设定阈值, 则以 $ExtSim(C_1, C_2)$ 和 $TerSim(C_1, C_2)$ 的平均值作为综合相似度值; 若 $ExtSim(C_1, C_2)$ 、 $TerSim(C_1, C_2)$ 和 $Strval$ 均小于设定阈值, 则以 $TerSim(C_1, C_2)$ 作为综合相似度值。

2.3.3 语义嵌入实现

以参考本体概念元素 C'' 与输入本体概念元素 C 和 C' 之间的相似性计算为前提, 以 N_3 为长度标准, 实现 Ontology1 和 Ontology2 的本体概念语义嵌入。具体步骤如下:

(1) 将 Ontology1、Ontology2、Ontology3 的概念元素, 即 $C = \{C_{i,i=1 \sim N_1}\}$ 、 $C' = \{C'_{i,i=1 \sim N_2}\}$ 、 $C'' = \{C''_{i,i=1 \sim N_3}\}$ 输入到模型;

(2) 计算数据集 C 中元素 C_i 与数据集 C'' 中元素 $C''_{j,j=1 \sim N_3}$ 的相似度值, 存入向量数组 $vec_i[j]$ 中构成空间向量集合 V , 完成对 C 的语义标注, 如公式 (4) 所示; 计算数据集 C' 中元素 C'_i 与数据集 C'' 中元素

$C'_{j,j=1 \sim N_3}$ 的相似度值,存入向量数组 $vec'_j[j]$ 中构成空间向量集合 V' ,完成对 C' 的语义标注,如公式(5)所示。

$$V = \begin{pmatrix} vec_1 = [V_{1-1}, V_{1-2}, V_{1-3}, \dots, V_{1-N_3}] \\ vec_2 = [V_{2-1}, V_{2-2}, V_{2-3}, \dots, V_{2-N_3}] \\ vec_3 = [V_{3-1}, V_{3-2}, V_{3-3}, \dots, V_{3-N_3}] \\ \dots \\ \dots \\ vec_{N_1} = [V_{N_1-1}, V_{N_1-2}, V_{N_1-3}, \dots, V_{N_1-N_3}] \end{pmatrix} \quad (4)$$

$$V' = \begin{pmatrix} vec'_1 = [V'_{1-1}, V'_{1-2}, V'_{1-3}, \dots, V'_{1-N_3}] \\ vec'_2 = [V'_{2-1}, V'_{2-2}, V'_{2-3}, \dots, V'_{2-N_3}] \\ vec'_3 = [V'_{3-1}, V'_{3-2}, V'_{3-3}, \dots, V'_{3-N_3}] \\ \dots \\ \dots \\ vec'_{N_2} = [V'_{N_2-1}, V'_{N_2-2}, V'_{N_2-3}, \dots, V'_{N_2-N_3}] \end{pmatrix} \quad (5)$$

(3) 输出语义标注后的空间向量 V 和 V' ,作为下一阶段深度学习的高精度输入数据。

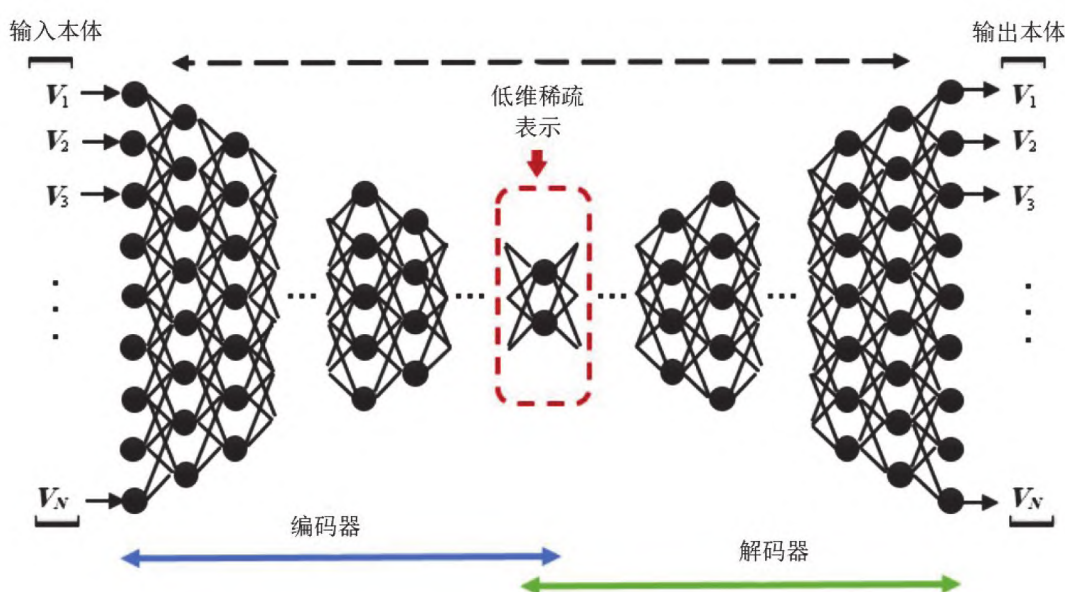


图2 自动编码器构建模型

2.4 深度学习阶段

利用多层神经网络的深度学习技术,对语义标注的本体概念以无监督的方式训练产生自动编码器模型。本模型包括编码器和解码器两个组件,编码器通过深度学习方法提取本体概念关键特征、获得本体概念的紧凑表达,将输入本体映射至低维稀疏的表示形式;解码器则依据紧凑的表达和关键特征反向创建原始输入,形成与输入本体具有相同维度的数据输出。自动编码器的构建模型如图2所示。

2.5 本体匹配阶段

本体匹配阶段,首先需要建立 Ontology1 和 Ontology2 对应概念的映射关系,然后利用余弦相似

函数 $\text{Cos}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}$ 计算具有映射关系向量

之间的相似度值。对相似度值超过设定阈值(阈值建

议取值 0.95)的重要映射关系予以保留,组成具有 m 个对应关系的本体匹配结果 A ;对相似度值达不到设定阈值的映射关系则做删除处理,以此提升两个本体概念元素映射的精准度,确保本体匹配的可靠性。

3 实验验证

3.1 实验数据选择

本实验选择 OAEI 本体概念集作为输入本体,以 AEO 为参考本体,以 BioWordVec 本体作为术语相似度检测的外部实例本体^[11],对文中的构建模型进行实验和测试。

OAEI 是国际通用共享的解剖学轨迹本体概念集,包含以 OWL 语言表示的 3304 个人类本体和 2744 个鼠类本体。AEO 是国际通用解剖学本体概念集,包含 250 个以 OWL 表示的解剖结构本体,具有注释清晰、知识量储备丰富、操作便捷等优势。

BioWordVec 本体是由医学控制词汇和未标注医学词汇联合组成的开放医学数据集，包含 232 万个不同单词。

由于参考本体 AEO 的长度为 250，在语义嵌入过程中 OAEI 概念集中的每个概念都会生产一个包含 250 个数字的向量表示形式。因此，本文将训练的自动编码器模型的结构设定为 [250-200-150-100-150-200-250] 的形式，作为自动编码器模型功能构建规则，以支撑和服务自动编码器中本体数据的输入、输出。

3.2 实验性能评估

以常用的精准率 p 、查全率 r 和综合度量值 F 三个参数指标对本体匹配模型的性能进行评估。若用 A 表示匹配的结果本体，用 R 表示参考本体，在实验测试中三个参数指标的含义及计算方法为：

(1) 精准率 p 表示正确匹配的本体与匹配的结果本体 A 的比例，可通过公式 (6) 得到。

$$p = \frac{|A \cap R|}{|A|} \quad (6)$$

(2) 查全率 r 表示正确匹配的本体与参考本体 R 的比例，可通过公式 (7) 得到。

$$r = \frac{|A \cap R|}{|R|} \quad (7)$$

(3) 综合度量值 F 是精准率 p 和查全率 r 的加权调和平均值，是对精准率 p 和查全率 r 之间客观存在的反比例反差关系的修正指标，可通过公式 (8) 得到。

$$F = 2 * \frac{p * r}{p + r} \quad (8)$$

3.3 实验结果

以既定测试数据为对象，将本模型与文献 [5][6] 中的 AGM 模型、文献 [7] 中的 DESKMatcher 模型、文献 [8] 中的 FAGMap-KG 模型和文献 [9] 中的 OntoConnect 模型进行对照测试，测试结果对照如表 1 所示。

表 1 本体模型性能测试指标结果对照表

测试模型	性能测试指标			运行时间 秒
	精准率	查全率	综合度量值	
AGM	0.152	0.195	0.171	628
DESKMatcher	0.472	0.623	0.537	391
FAGMap-KG	0.976	0.662	0.782	165
OntoConnect	0.986	0.648	0.797	248
Deep OM	0.996	0.665	0.797	149

由表 1 可知：Deep OM 的匹配精准率、查全率、综合度量值均是 5 个对照模型中得分最高的，说明本

模型进行本体匹配的可靠性更有保证，匹配效果令人满意；虽然 OntoConnect 模型在综合度量值测试上与 Deep OM 的得分相当，但在运行时间方面，Deep OM 展示出优越的计算性能和运行效率。

4 结束语

本文设计的 Deep OM 运用深度学习技术训练产生自动编码器，将海量本体以更准确、更少维度的形式进行表达，进而实现海量本体的精准匹配。通过对大型 OAEI 本体和对等本体的匹配测试表明，Deep OM 的匹配准确性更高、时间消耗更少，在海量异构本体匹配领域具有很好的研究与应用价值。

参考文献：

- [1] YAN S K, WONG K C. Elucidating high-dimensional cancer hallmark annotation via enriched ontology[J]. Journal of Biomedical Informatics, 2017(73):84-94.
- [2] PING P P, HERMIAKOB H, POLSON J S, et al. Biomedical informatics on the cloud: a treasure hunt for advancing cardiovascular medicine[J]. Circulation Research, 2018, 122(9):1290-1301.
- [3] STRANG J F, MEAGHER H, KENWORTHY L, et al. Initial clinical guidelines for Co-occurring autism spectrum disorder and gender dysphoria or incongruence in adolescents[J]. Journal of Clinical Child & Adolescent Psychology, 2018, 47(1):105-115.
- [4] SANZ X, PAREJA L, RIUS A. Definition of a SNOMED CT pathology subset and microglossary, based on 1.17 million biological samples from the Catalan Pathology Registry[J]. Journal of Biomedical Informatics, 2018(78): 167-176.
- [5] KHIAT A, BENAÏSSA M. A new instance-based approach for ontology alignment[J]. International Journal on Semantic Web and Information Systems, 2015, 11(3):25-43.
- [6] LÜTKE A. AnyGraphMatcher Submission to the OAEI Knowledge Graph Challenge[J]. OM@ISWC, 2019, 2536:86-93.
- [7] MONYCH M, PORTISCH J, HLADIK Met, et al. DESKMatcher[J]. CEUR, 2020, 2788:181-186.
- [8] CHANG F, CHEN G, ZHANG S. FCAMap-KG Results for OAEI[J]. OM@ISWC, 2019, 2536:138-145.
- [9] CHAKRABORTY J, YAMAN B, VIRGILI L, et al. OntoConnect: Results for OAEI[J]. Ontology Matching, 2020:204-205.
- [10] 王颖, 薛醒思, 卢家伟, 等. 基于进化算法的生物医学本体匹配技术 [J]. 重庆大学学报, 2023, 46(6):130-135.
- [11] ZHANG Y, CHEN Q, YANG Z. BioWordVec: improving biomedical word embeddings with subword information and Medical Subject Headings [J]. Scientific Data, 2019, 6(1):1-9.

Research on Constructing a Mass Ontology-matching Model Based on Deep Learning

CHANG Wanjun, ZHANG Dongfang

(School of Computer Science & Technology, Henan Institute of Technology, Xinxiang 453003,China)

Abstract: Ontology-matching is an effective method for data communication and operation of multi-source heterogeneous ontologies in semantic network. It is a major challenge in the design of Ontology-matching algorithm for researchers to achieve scientific classification of mass ontology and reduce the time, space and consumption during the process of heterogeneous Ontology-matching. This paper presents autoencoder and Deep Learning technology of mass Ontology-matching algorithm design which provides important support for mass ontology semantic matching and data mining to deal with this problem. It consists of semantic embedding and model training for input ontology by using autoencoder, which can reduce the complexity of heterogeneous ontology and improve the accuracy of heterogeneous ontology classification. It is clearly showed that the algorithm has obvious advantages in time complexity and operation accuracy through experiments of multiple Ontology-matching systems, which has a high value of research and application.

Key words: ontology; autoencoder; semantic embedding; deep learning

(责任编辑 王磊)

(上接第 16 页)

Study on the Limit and Mechanism of Carbon Fiber/Aluminum Alloy Laminates for Automotive with Elastic Mold Plasticizing Forming

LI Qian, LIU Yifan, WANG Yaoyao, LIU Qing

(School of Vehicle And Traffic Engineering, Henan Institute of Technology, Xinxiang 453003, China)

Abstract: In order to solve the problems of poor deformation coordination of heterogeneous materials and low plastic forming limit of heterogeneous materials in the traditional forming process of automotive carbon fiber/aluminum alloy laminates, a method of elastic mold plastic forming for carbon fiber metal laminates is proposed. Combined with theoretical analysis, finite element simulation, experiments and other means, the deformation behavior and forming limit of carbon fiber metal laminates in the plastic forming process of elastic mold were studied. The experimental results show that the elastic mold can significantly improve the forming performance of carbon fiber metal laminates, and as the hardness of the elastic mold increases, both the forming limit displacement and the load-bearing limit of the laminates are improved. When the hardness of the elastic mold reaches HS90A, compared with the traditional non-elastic mold forming, the forming limit displacement of the laminates can be increased by 30.6% and the load-bearing limit can be increased by 21.8%.

Key words: carbon fiber / aluminum alloy laminates; elastic mold; plastic forming

(责任编辑 吕春红)