

基于图卷积网络的多源本体相似度计算方法

孙留倩, 魏玉良, 王佰玲

(哈尔滨工业大学(威海)计算机科学与技术学院, 山东 威海 264209)

摘要: 在信息时代, 数据量呈指数式增长, 而不同数据源存在难以统一表示的异构问题, 给数据共享、重用造成不便。语义网络的迅速发展, 使本体映射成为解决该问题的有效手段, 其核心是本体相似度计算, 提出了一种基于图卷积网络的计算方法。将本体建模为异构图网络, 再使用图卷积网络学习文本嵌入规则, 得到全局统一表示, 完成多源数据的融合。实验结果表明, 所提方法计算准确性高于其他传统方法, 有效地提高了多源数据融合的准确度。

关键词: 多源数据融合; 图卷积网络; 本体映射; 相似度计算

中图分类号: TP393

文献标识码: A

DOI: 10.11959/j.issn.2096-109x.2021071

Novel similarity calculation method of multisource ontology based on graph convolution network

SUN Liujian, WEI Yuliang, WANG Bailing

School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China

Abstract: In the information age, the amount of data is growing exponentially. However, different data sources are heterogeneous, which makes it inconvenient to share and multiplex data. With the rapid development of semantic network, ontology mapping is an effective method to solve this problem. The core of ontology mapping is ontology similarity calculation. Therefore, a calculation method based on graph convolution network was proposed. Firstly, ontologies are modeled as a heterogeneous graph network, then the graph convolution network was used to learn the text embedding rules, which made ontologies were defined in global unified representation. Lastly, multisource data fusion was completed. The experimental results show that the accuracy of the proposed method is higher than other methods, and the accuracy of multi-source data fusion was effectively improved.

Keywords: heterogeneous data fusion, graph convolution network, ontology mapping, similarity calculation

收稿日期: 2021-03-11; 修回日期: 2021-05-13

通信作者: 王佰玲, wbl@hit.edu.cn

基金项目: 国家重点研发计划(2018YFB2004200)

Foundation Item: The National Key R&D Program of China (2018YFB2004200)

论文引用格式: 孙留倩, 魏玉良, 王佰玲. 基于图卷积网络的多源本体相似度计算方法[J]. 网络与信息安全学报, 2021, 7(5): 149-155.

SUN L Q, WEI Y L, WANG B L. Novel similarity calculation method of multisource ontology based on graph convolution network[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 149-155.

1 引言

随着信息技术的发展和数据库技术的普及,越来越多的领域开始使用信息技术来管理数据信息,这在一定程度上提高了工作效率,但不同领域独立管理数据源,不同管理者对数据的描述存在差别,导致信息只能在内部交换,给数据共享和信息交流方面带来极大不便^[1]。由于数据量巨大且分散,为了建立多源数据的全局统一表示,提高数据利用率和集成度,数据融合工作成为亟待解决的问题^[2]。

语义网络和集成技术的不断发展,使本体映射成为多源数据融合的有效手段。本体映射的过程如图 1 所示,通过相似度计算方法得到与目标本体相似的源本体,形成映射关系,进而得到统一的全局表示,从而实现多源数据的有效融合。本体映射的核心是本体语义相似度计算,相似度结果的准确性决定着本体映射工作的科学性,因此,如何提高其准确性逐渐成为本体映射、数据融合等领域的研究热点,具有重要的研究意义和价值。

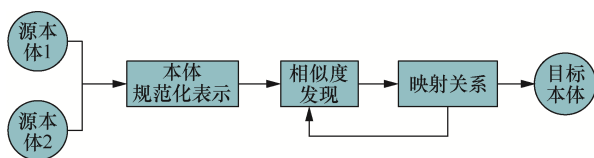


图 1 本体映射的过程
Figure 1 The flow chart of ontology mapping

本文在对传统本体语义相似度计算方法进行分析研究的基础上,提出一种基于图卷积网络(GCN, graph convolution network)进行实体相似度计算方法。该方法将要处理的本体建模为异构图网络,并使用 GCN 学习文本及文档的嵌入规则。该方法在不使用预先训练的文本和外部知识的情况下,相似度准确性优于目前的语义相似度计算方法。

2 相关工作

基于本体的相似度计算方法的研究,大体上可以分为 5 类:基于语义距离的相似度计算方法;基于信息内容的相似度计算方法;基于概念属性的相似度计算方法;混合式语义相似度计算方法

^[3];基于深度学习的相似度计算方法^[4]。

2.1 传统相似度计算方法

2.1.1 基于语义距离的相似度计算方法

Rada^[5]提出一种基于语义距离的相似度计算方法,它的基本思想是预设本体结构树中的权重大小相等,根据概念节点词在本体结构树中的位置语义距离来衡量相似度,语义距离越大说明相似度越低,语义距离越小说明相似度越大,该方法计算公式如下:

$$\text{sim}(i, j) = \frac{1}{\text{dis}(i, j)} \quad (1)$$

其中, $\text{sim}(i, j)$ 表示概念节点 i 和 j 之间的相似度, $\text{dis}(i, j)$ 表示概念节点 i 和 j 之间的语义距离。该方法的优点是计算复杂度不高,计算速度快,不足之处在于该方法有个前提是本体结构树中的每条边的权重大小是相等的,没有考虑到边的类型、位置信息等因素的影响。

2.1.2 基于信息内容的相似度计算方法

基于信息内容的相似度计算方法的主要思想是求概念节点之间的熵,熵值越大,代表节点之间共享信息越多,相似度也就越大。在一棵本体结构树中,任意一个子概念节点都可以包含其祖父节点的全部信息内容。故 Sun^[6]提出任意两个概念节点的相似度可以通过计算最邻近祖父概念节点的信息量和出现频率来求得,本体中任意两个概念节点之间的语义相似度计算公式如下:

$$\text{Sim}(c_i, c_j) = \frac{2\text{IC}(\text{Lcan}(c_i, c_j))}{\text{IC}(c_i) + \text{IC}(c_j)} \quad (2)$$

其中, $\text{Lcan}(c_i, c_j)$ 表示本体结构树中概念节点 c_i, c_j 的最近邻公共祖先节点, $\text{IC}(c_i)$ 、 $\text{IC}(c_j)$ 分别表示概念节点 c_i, c_j 的信息量。

算法的不足之处在于当两个概念次节点属于同一个本体结构时,不仅需要计算被比较概念次节点的共享信息,还要计算两个概念本身的信息熵之和。

2.1.3 基于概念属性的相似度计算方法

基于概念属性的相似度计算方法的主要思想是根据概念节点的属性来说明概念节点的特征, Tversky^[7]提出可以通过计算两个概念节点公共属性对的数量来计算相似度,具体公式如下。

$$\text{Sim}(c_i, c_j) = Xf(c_i \cap c_j) - Yf(c_i - c_j) - Zf(c_j - c_i) \quad (3)$$

其中, $f(c_i \cap c_j)$ 表示概念节点 c_i, c_j 具有的公共属性集, $f(c_i - c_j)$ 表示概念节点 c_i 具备而 c_j 不具备的属性集, $f(c_j - c_i)$ 表示概念节点 c_j 具备而 c_i 不具备的属性集。该算法的缺点是仅考虑了概念节点的属性, 而完全忽略了概念节点之间位置、内容信息。

2.1.4 混合式语义相似度计算方法

单一的相似度计算方法导致相似度计算结果线性程度低, 故混合式语义相似度计算方法成为了研究的一个方向。Zheng^[8]针对相似度进行人工加权计算时效率低的问题, 设计了一种自适应的主动加权相似度计算方案。Lu^[9]提出一种综合相似度计算模型, 构建了一种新的语义相似度度量方法的组合框架和参数调整方案, 有效组合了3种相似度度量方法来确定高质量的本体匹配结果。该方法考虑因素比较全面, 计算结果比较准确且稳定, 不足之处在于考虑的因素过多, 导致计算复杂度大大增加, 降低了计算效率。

2.2 基于深度学习的相似度计算方法

随着语义网络、知识图谱等新兴技术的发展, 基于语义网的词语相似度计算方法开始广泛应用。传统的语义相似度计算方法并没有考虑语义和词语进行有效组合, 后来 WordNet 计算方法被提出, 该方法主要通过 WordNet 义原进行分类, 利用义原计算概念之间的相似度。Guo^[10]对其进行了一定的优化。Li^[11]等运用相似实体推荐及知识推理来计算实体间的相似度, 实验效果较好。Xu^[12]针对现有本体概念相似度计算模型中存在的精度不高问题, 提出了基于模拟退火改进 BP (SA-BP, simulated annealing back propagation) 神经网络算法的相似度综合计算模型, 利用 BP 网

络可以对复杂相似度计算模型的算术因子进行模拟, 但一般的神经网络模型存在收敛速度过慢的问题, 导致最后计算结果会陷入局部最优解。另外, 大多数相似度计算模型的权重对领域专家和历史数据依赖性较大, 存在主观性、滞后性, 且对不同本体适用性较差, 不适合拓展。以上所提及不得相似度求解方法的优缺点如表1所示。

也有些研究尝试用 GCN 进行文本分类、异常检测^[13]相关方面的工作, 它们将一份文档或者一句话视为图的一个节点^[14], 或用不常用的引用关系构建图网络^[15], 而检索未发现将 GCN 应用到本体相似度计算方面的研究。本文将单词和文档都视为节点, 构建出大型的异质图网络, 不需要利用文档间的关系, 有效地提高了相似度计算的准确性。因此, 本文提出一种基于 GCN 的计算模型来计算本体之间的语义相似度。

3 基于图卷积网络多源本体相似度算法

3.1 图卷积网络

GCN 模型是通过对相邻节点的特征进行卷积来对图进行操作的, 最早是由 Kipf^[16]提出来的, 相较于卷积神经网络, GCN 模型计算效率更高, 在 GCN 模型中, 输出的特征与节点本身及邻近节点有密切关系, 这体现了 GCN 模型是在图的基础上进行特征的学习并输出的。

GCN 模型的结构包含三层, 分别是输入层、隐藏层、输出层。输入层的输入主要有网络节点的特征矩阵和邻接矩阵, 特征矩阵描述了网络中各个节点的特征之间的区别; 邻接矩阵是为了方便网络节点之间的信息传播。隐藏层的作用是线性划分不同类型的数据, 其中, 经常使用 ReLu 作为激活函数, 引入 dropout 是为了防止过拟合。输出层的作用是将隐藏层学习的抽象特征转化为

表1 不同相似度求解方法的优缺点对比
Table 1 Comparison of advantages and disadvantages of different similarity methods

方法	优点	缺点
基于语义距离的相似度计算方法	计算复杂度较低	考虑因素单一, 使结果准确性不高且不稳定
基于信息内容的相似度计算方法	结果较准确	针对完备性不高的语料库, 算法不具备鲁棒性
基于概念属性的相似度计算方法	结果较准确	针对不同本体, 算法不具备普适性
混合式语义相似度计算方法	结果较客观准确	计算复杂度高
基于深度学习的相似度计算方法	结果准确	收敛性差

预测值输出^[17]。

GCN 模型可以通过层之间的相互叠加来实现不同空间的信息传播及特征提取^[18], 提取的特征表示为

$$h_v^l = \sigma(W^l \sum_{u \in N(v)} \frac{1}{\sqrt{\deg(v)} \cdot \sqrt{\deg(u)}} h_u^{l-1}) \quad (4)$$

式(4)中, 设定一个图 $G=(V, E)$, V 表示图中的节点集合, E 表示图中的边集合。 h_v^l 代表节点 v 在第 l 层的输出特征, σ 代表非线性激活函数, W^l 代表学习的参数矩阵, $N(v)$ 代表节点 v 和其最邻近节点的集合, $\deg(\cdot)$ 代表某个节点的度, h_u^{l-1} 代表节点 u 在第 $l-1$ 层的隐藏特征。

3.2 基于图卷积网络多源本体相似度计算方法

基于 GCN 的本体相似度计算过程如图 2 所示, 算法具体步骤如下。

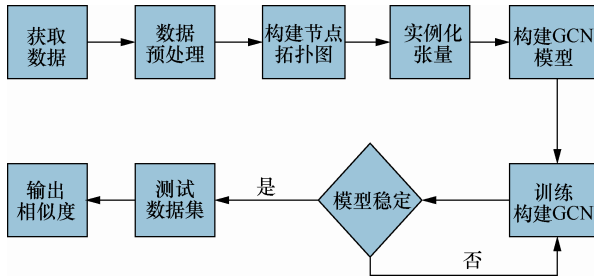


图 2 基于 GCN 的本体相似度计算过程

Figure 2 The process diagram of similarity calculation method multisource ontology based on graph convolution network

输入 来源于数据库 1 的数据集 A , 字段为 $\{A_1, A_2, A_3, \dots, A_n\}$; 来源于数据库 2 的数据集 B , 字段为 $\{B_1, B_2, B_3, \dots, B_m\}$ 。

输出 来源于数据库 1 的本体 x 与来源于数据库 2 的本体 y 的相似度。

步骤 1 数据预处理

获取数据集 A 和数据集 B 之后, 将获取的数据保存在 mysql 中。为了提高计算的精确度, 在计算前对数据进行分词、停用词过滤、词干提取、保留相似属性字段、删除不同属性字段等预处理操作。得到数据集 $A'\{C_1, C_2, C_3, \dots\}$, 数据集 $B'\{C_1, C_2, C_3, \dots\}$ 。两个数据集再经 protégé 处理后分别转化为本体集 $X\{x_1, x_2, x_3, \dots, x_n\}$ 、 $Y\{y_1, y_2, y_3, \dots, y_m\}$, 以 RDF 格式存储。

步骤 2 构建拓扑图

利用步骤 2 得到的本体集构建拓扑图, 拓扑

图的节点数就是本体的数量, 用单位矩阵 I 来表示特征矩阵 X , 向量形式采用 one-hot 稀疏矩阵, 这样的形式扩充了本体的特征, 使特征的距离计算更加合理; 邻接矩阵 A 使用点互信息 PMI 表示, 具体公式如下。

$$A_{x,y} = \begin{cases} 0, & x \neq y \\ \text{PMI}(x,y), & \text{其他} \\ 1, & x = y \end{cases} \quad (5)$$

点互信息的计算公式如下。

$$\text{PMI}(x,y) = \log \frac{p(x,y)}{p(x) \cdot p(y)} \quad (6)$$

$$p(x,y) = \frac{\#W(x,y)}{\#W} \quad (7)$$

$$p(x) = \frac{\#W(x)}{\#W} \quad (8)$$

$\#W(x,y)$ 表示节点 x 和节点 y 特征向量相同的个数, $\#W(x)$ 表示节点 x 的特征向量在所有本体中出现的个数。

由以上步骤所得的每个节点的特征矩阵及邻接矩阵可以构建出两个无向网络拓扑图 $G_1=(V_1, E_1)$ 、 $G_2=(V_2, E_2)$ 。

步骤 3 实例化张量

对构建好的拓扑图节点进行实例化张量, 需要实例化的张量包括特征矩阵 X 、邻接矩阵 A 、节点的度 degree 、节点的标签 label , 无向图 G_1 实例化后的张量可以表示为 $[X_1, A_1, \text{degree}_1, \text{label}_1]$, 无向图 G_2 实例化后的张量可以表示为 $[X_2, A_2, \text{degree}_2, \text{label}_2]$ 。

步骤 4 构建 GCN

为了防止构建的 GCN 模型过拟合, 特引入 dropout 层, 构建两层 GCN 模型, 非线性激活函数使用 LeakyReLU, 损失函数采用 SoftMax 函数, 优化器采用 ADAM, 如式(9)、式(10)所示, GCN 模型如图 3 所示。

$$H = \text{ReLU}(\tilde{A}XW_0) \quad (9)$$

$$\text{Output} = \text{SoftMax}(\tilde{A}HW_1) \quad (10)$$

步骤 5 训练数据集

将无向图 G_1 中的一部分本体概念节点 V_1 和无向图 G_2 中的一部分本体概念节点 V_2 作为训练数据, 将概念节点通过人工评价得到的相似度以

及分别基于语义距离、信息内容、概念属性、混合式语义、SA-BP 算法得到的相似度作为模型的输入,通过 GCN 模型的学习计算出针对本训练集最合适的特征矩阵 $\mathbf{X}_{\text{best}}$ 和邻接矩阵 $\mathbf{A}_{\text{best}}$, 以此得到最稳定的计算模型。

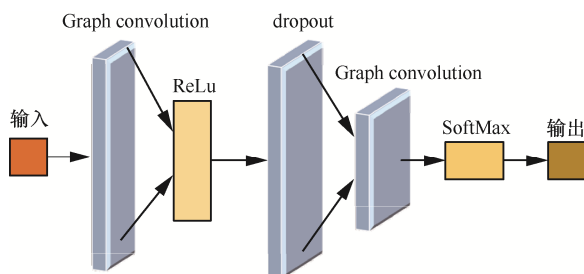


图3 GCN 模型
Figure 3 GCN model

步骤6 测试数据集

得到稳定的计算模型后,将之应用于测试数据集中,在本体集 X 、 Y 中选取需要融合的本体概念节点 x 、 y 作为测试集,分别计算测试数据集基于语义距离、信息内容、概念属性、混合式语义、SA-BP 算法的相似度,将计算结果代入稳定的 GCN 模型,进行正向计算,输出的结果即本体概念节点 x 、 y 的相似度计算结果。

4 实验结果与分析

4.1 算法运行环境

集成开发环境为 PyCharm 2020.3, 编码语言为 Python 3.0, 使用的框架为 TensorFlow。

4.2 实验数据集

本文的数据获取分别来自“*The Movie Database (TMDb)*”和“*豆瓣网电影排行榜*”经过对数据的去重、去噪、删除不同属性、保留相同属性的处理,共计获得演员数量 1 982 位;电影数量 91 369 部;20 类电影类型;人物与电影的关系 7 119 287 对;电影与类型的关系 196 354 对。

对以上数据进行概念节点的构建,构建出的电影本体概念节点的属性结构为{电影名称, 参演演员, 电影评分, 电影发行日期, 电影类型}。根据概念节点的属性构建出网络拓扑如图 4 所示。

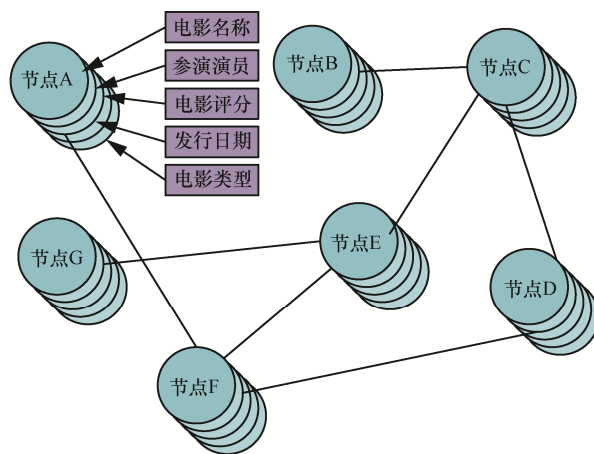


图4 网络拓扑
Figure 4 Network topology

4.3 实验结果分析

从 TMDb 数据库和豆瓣数据库构建的概念节点中,如图 4 所示。随机选取 1 000 个本体概念节点,组成 1 000 组概念节点对,其中 700 组用作于训练数据集,剩下的 300 组作为测试数据集。

实验共采用了 6 种算法对 300 组测试样本均分为 6 组进行相应的相似度计算,最后将所得结果进行分析对比,6 种算法分别为基于语义距离的相似度计算方法、基于信息内容的相似度计算方法、基于概念属性的相似度计算方法、基于混合语义式的相似度计算方法、基于 SA-BP 算法的相似度计算方法和基于 GCN 的相似度计算方法。实验结果如表 2 所示。

以人工评价作为参考标准,分别计算以上几种算法得出的相似度与人工评价相似度的差值的最大值、最小值、标准差并计算其准确度,得到的结果如表 3 所示。为了更直观进行数据展示,将实验数据绘制成柱状图,如图 5 所示。

为了评估各种算法得出的相似度与人工评价得出的结果的相关性,本文采用皮尔逊相关系数作为参考指标,它在统计学中用于表征两个变量之间的相关性,其值介于 0~1。值越大代表两者相关性越大,定义如下。

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2} \sqrt{E(Y^2) - (E(Y))^2}} \quad (11)$$

其中, $E()$ 表示某变量的期望。当皮尔逊的值介于 0~0.4, 表示两者极弱相关或者不相关;当皮尔逊的值介于 0.4~0.6, 表示两者弱相关;当皮尔

表2 不同计算方法求测试样本的相似度结果对比

Table 2 Comparison of similarity results of different calculation methods for test samples

计算相似度方法	代表性文献	第一组	第二组	第三组	第四组	第五组	第六组
人工评价	—	0.735 5	0.596 7	0.638 6	0.266 1	0.128 3	0.908 9
基于语义距离的相似度计算方法	文献[5]	0.547 8	0.671 2	0.601 9	0.362 1	0.562 4	0.928 7
基于信息内容的相似度计算方法	文献[6]	0.991 2	0.172 2	0.351 3	0.266 1	0.133 9	0.987 2
基于概念属性的相似度计算方法	文献[7]	0.802 8	0.821 1	0.211 2	0.503 2	0.003 7	0.740 3
混合式语义相似度计算方法	文献[8,9]	0.357 2	0.308 8	0.708 9	0.100 2	0.067 1	0.653 2
基于 SA-BP 算法的相似度计算方法	文献[12]	0.498 4	0.201 9	0.339 8	0.099 8	0.089 2	0.798 7
基于 GCN 的相似度计算方法	—	0.692 1	0.387 1	0.511 4	0.112 7	0.138 2	0.887 3

表3 不同方法相似度计算结果误差及皮尔逊相关系数

Table 3 Error of similarity calculation results and Pearson correlation coefficient of different methods

计算相似度方法	误差最大值	误差平均值	标准差	皮尔逊相关系数
基于语义距离的相似度计算方法	0.434 1	0.141 4	0.141 5	0.715 6
基于信息内容的相似度计算方法	0.424 5	0.175 2	0.158 1	0.781 5
基于概念属性的相似度计算方法	0.427 4	0.208 2	0.113 7	0.680 0
混合式语义相似度计算方法	0.378 3	0.203 2	0.115 3	0.834 9
基于 SA-BP 算法的相似度计算方法	0.394 8	0.198 8	0.123 0	0.898 3
基于 GCN 的相似度计算方法	0.209 7	0.094 2	0.074 0	0.960 2

逊的值介于 0.6~0.8, 表示中等程度相关; 当皮尔逊的值介于 0.8~0.9, 表示两者强相关; 当皮尔逊的值介于 0.9~1.0, 表示两者极强相关。

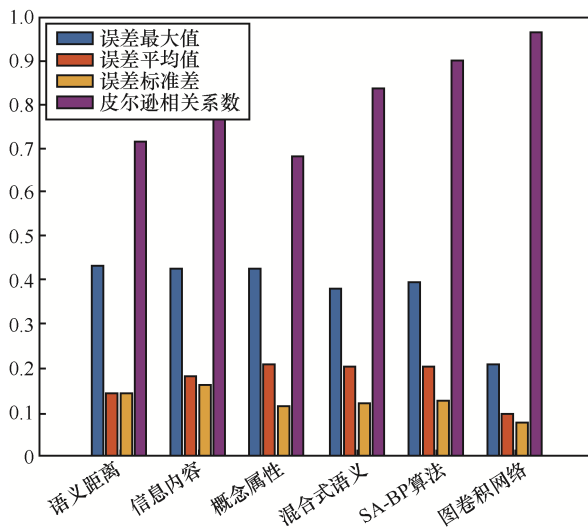


图5 6种算法结果对比

Figure 5 Comparison of six algorithms

从以上结果可以看出, 本文提出的基于 GCN 的本体相似度计算方法得出的结果在皮尔逊系数上是最大的, 达到了 0.960 2, 并且误差最大值、

误差平均值、误差标准差是最小的。这表明利用本文提出的相似度计算方法的结果收敛性好, 稳定性强, 准确率高。

为了验证不同体量的多源数据集对算法准确率的影响, 将数据大小分别为 1 000 条、10 000 条、100 000 条的 3 组多源数据集作为单独任务应用进行计算, 计算得出的皮尔逊相关系数均在 0.94 以上, 说明面对不同体量的数据, 本文方法依然具有很高的准确率, 性能良好。

5 结束语

本文分析了几种计算本体相似度方法存在的缺陷, 并针对相应的问题, 结合 GCN, 提出一种基于 GCN 的本体相似度计算方法, 实验结果表明, 本文提出的方法较传统的计算本体相似度的方法准确率更高。下一步工作是对 GCN 模型的参数进行优化以提高模型的准确性及稳定性。

参考文献:

- [1] 杨泉. 基于遗传算法的词语语义相似度计算研究[J]. 计算机技术与发展, 2021, 31(2): 8-13.
YANG Q. Research on word semantic similarity calculation based

- on genetic algorithm[J]. Computer Technology and Development, 2021, 31(2): 8-13.
- [2] 丁悦航, 于洪涛, 黄瑞阳, 李英乐. 本体摘要技术综述[J]. 网络与信息安全学报, 2018, 4(10): 12-21.
DING Y H., YU H T, HUANG R Y. Ontology summarization technology survey[J]. Chinese Journal of Network and Information Security, 2018, 4(10): 12-21.
- [3] DAO W. Distance learning techniques for ontology similarity measuring and ontology mapping[J]. Cluster Computing, 2017, 20(2) : 959-968.
- [4] 周爱武, 翟增辉, 刘慧婷. 基于模拟退火算法改进的 BP 神经网络算法[J]. 微电子学与计算机, 2016, 33(4): 144-147.
ZHOU A W, ZHAI Z H, LIU H T, et al. Improved BP neural network based on simulated annealing[J]. Microelectronics and Computer, 2016, 33(4): 144-147.
- [5] RADA R, MILI H, BICKNELL E. Development and application of a metric on semantic nets[J]. IEEE, Transactions on Systems, Man and Cybernetics, 1989, 19(1): 17-30.
- [6] 孙丽莉, 张小刚. 基于 WordNet 的概念语义相似度的计算方法[J]. 统计与决策, 2017(23): 79-82.
SUN L L, ZHANG X G. A novel concept semantic similarity calculation method based on WordNet[J]. Statistics & Decision, 2017(23): 79-82.
- [7] TVERSKY A. Feature of similarity[J]. Psychological Review, 1997, 84(4): 222-226.
- [8] 郑志蕴, 阮春阳, 李伦, 等. 本体语义相似度自适应综合加权算法研究[J]. 计算机科学, 2016, 43(10): 242-247.
ZHENG Z Y, RUAN C Y, LI L, et al. Adaptive ontology semantic similarity comprehensive weighted algorithm[J]. Computer Science, 2016, 43(10): 242-247.
- [9] 卢家伟, 薛醒思, 肖祖宇, 等. 一种基于混合语义相似度度量方法的本体元匹配技术[J]. 宝鸡文理学院学报(自然科学版), 2020, 40(2): 59-63.
LU J W, XUE X S, XIAO Z G, et al. An ontology meta-matching technique based on the hybrid semantic similarity measure[J]. Journal of Baoji University of Arts and Sciences(Natural Science Edition), 2020, 40(2): 242-247.
- [10] 郭小华, 彭琦, 邓涵, 等. 基于边权重的 WordNet 词语相似度计算[J]. 计算机工程与应用, 2018, 54(1): 172-178.
GUO X H, PENG Q, DENG H, et al. Edge weight-based word similarity computation in WordNet[J]. Computer Engineering and Applications, 2018, 54(1): 172-178.
- [11] LI Y, GAO D Q. Research on entity similarity computation in knowledge map[J]. Chinese Journal of Information Science, 2017, 31(1): 145-151.
- [12] 许飞翔, 叶霞, 李琳琳, 等. 基于 SA-BP 算法的本体概念语义相似度综合计算[J]. 计算机科学, 2020, 47(1): 199-204.
XU F X, YE X, LI L L, et al. Comprehensive calculation of semantic similarity of ontology concept based on SA-BP[J]. Computer Science, 2020, 47(1): 199-204.
- [13] 曲强, 于洪涛, 黄瑞阳. 基于图卷积网络的社交网络 Spammer 检测技术[J]. 网络与信息安全学报, 2018, 4(5): 39-46
QU Q YU H T, HUANG R Y. Spammer detection technology of social network based on graph convolution network[J]. Chinese Journal of Network and Information Security[J]. 2018, 4(5): 39-46 .
- [14] PENG H, LI J. Large-scale hierarchical text classification with recursively regularized deep graph[J]. International World Wide Web Conference, 2018: 1063-1072.
- [15] ROUSSEAU F, KIAGIAS E. Text categorization as a graph classification problem[J]. ACL, 2015, 1: 1702-1712.
- [16] KIPF T N. Semi-supervised classification with graph convolutional networks[J]. ICLR, 2017.
- [17] 姚佳奇, 徐正国, 燕继坤, 等. GCN-PU: 基于图卷积网络的 PU 文本分类算法[J]. 计算机工程与应用, 2020: 1-8
YAO J Q, XU Z G, YE J K, et al. GCN-PU text classification algorithm based on graph convolutional network[J]. Computer Engineering and Applications. 2020: 1-8.
- [18] 李慧, 胡吉霞. 一种基于图卷积自编码模型的多维度学科知识网络融合方法[J]. 图书情报工作, 2020, 64(18): 114-125.
LI H, HU J X. Multi-dimensional subject knowledge network fusion method based on graph convolution self-encoding model[J]. Library and Information Service, 2020, 64(18): 114-125.

[作者简介]



孙留倩 (1997-), 女, 山东菏泽人, 哈尔滨工业大学(威海)硕士生, 主要研究方向为多源数据融合、网络大数据安全。



魏玉良 (1989-), 男, 山东寿光人, 博士, 哈尔滨工业大学(威海)助理研究员, 主要研究方向为自然语言处理、知识图谱、工业互联网安全。



王佰玲 (1978-), 男, 黑龙江哈尔滨人, 哈尔滨工业大学(威海)教授、博士生导师, 主要研究方向为工业互联网安全、信息对抗、信息安全、信息搜索、金融安全。