计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.181202351

基于 SA-BP 算法的本体概念语义相似度综合计算

许飞翔[1]　叶　霞[1]　李琳琳[1]　曹军博[1]　王　馨[2]
1 火箭军工程大学作战保障学院　西安 710025
2 马里兰大学信息系统学院　马里兰州 巴尔的摩 21250
(894409949@qq.com)

摘　要　不同作战部队在指挥信息系统测试评估中建立的指标存在异构问题,导致在信息交互和测试数据共享上存在较大困难。实现指标本体概念的映射和集成,建立一个统一的全局指标本体树可以有效地解决该问题,其中本体概念相似度计算的准确性至关重要。针对现有本体概念相似度计算模型中存在的精度不高的问题,提出了基于模拟退火改进 BP(Back Propagation)神经网络(Simulated Annealing Back Propagation,SA-BP)算法的相似度综合计算模型。首先,对经典的基于语义距离、信息内容和概念属性的相似度计算模型进行改进,同时提出了基于概念子节点重合度的相似度计算模型;然后,采用 SA-BP 算法进行相似度综合计算,避免现有方法中人为确定权重的主观性和简单线性加权的不准确性问题;最后,从某作战部队不同单位建立的各异的指挥信息系统评估指标的本体概念中提取样本数据,对相似度综合计算模型进行训练测试。实验数据表明,相比于 PSO-BP 计算模型和主成分分析确定权值的线性加权计算模型,基于 SA-BP 算法的相似度综合计算模型的计算结果与专家评价结果的 Pearson 相关系数分别提升了 0.0695 和0.1351,达到了极强相关的一致性。实验数据充分说明,模拟退火算法改进的 BP 神经网络在训练后可以较好地收敛,在综合计算本体概念相似度时更加准确,从而有效地解决了本体概念集成的关键问题。

关键词:本体集成;语义相似度计算;BP 神经网络;模拟退火算法;子节点重合度
中图法分类号　TP391

Comprehensive Calculation of Semantic Similarity of Ontology Concept Based on SA-BP Algorithm

XU Fei-xiang[1],YE Xia[1],LI Lin-lin[1],CAO Jun-bo[1] and WANG Xin[2]
1 Academy of Combat Support,Rocket Force University of Engineering,Xi'an 710025,China
2 Information Systems Department,University of Maryland,Baltimore,Maryland 21250,USA

Abstract　There are heterogeneous problems in indicators that are established by different combat forces when evaluating and testing command information systems,which leads to great difficulties in information interaction and data sharing. In order to achieve mapping and integration of indicator's ontology-concept,building a unified global indicator ontology tree is an effective solution. In this case,the accuracy of similarity calculation for ontology-concept becomes crucial. Aiming at the problem of low accuracy in the existing ontology-concept similarity calculation model,a comprehensive simi-larity calculation model based on BP neural network algorithm which is improved by Simulated Annealing (SA-BP),was proposed. This paper first improved the classical similarity calculation models based on semantic distance,information content and conceptual attribute. Besides,a similarity calculation model in view of concept's sub-node coincidence was proposed in order to avoid the subjectivity of artificially determined weights and the inaccuracy of simple linear weighting in existing models. At last,a training test on the comprehensive similarity calculation model was performed,while the sample data were extracted from ontology-concept of variable evaluation indicators that come from command information systems established by different departments of combat forces. Experimental data show that compared with PSO-BP calculation model and principal-component linear weighted calculation model,the comprehensive similarity calculation model based on SA-BP algorithm achieves strong correlation,since its results and its Pearson correlation coefficient of the results evaluated by experts are increased by 0.0695 and 0.1351 respectively. The experimental results verify that,after training,SA-BP algorithm can converge better and achieve higher accurate when calculating ontology-concept similarity. Hence,

到稿日期:2018-12-18　返修日期:2019-04-28　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61702525)
This work was supported by the National Natural Science Foundation of China (61702525).
通信作者:叶霞(yex_qing@163.com)

key issues of integration for ontology-concept can be effectively solved.

Keywords Ontology integration, Semantic similarity calculation, BP neural network, Simulated annealing algorithm, Sub-node coincidence

1 引言

指挥信息系统在作战部队中的广泛运用，产生了大量的测试与评估数据，由于不同单位建立的指标体系存在差异，相同的概念经常被使用不同的指标本体进行描述，导致测试与评估数据在共享和集成上存在较大困难。随着语义网络和本体集成技术的不断发展，本体映射[1]成为解决此问题的有效手段。通过映射完成相同语义指标概念的合并，得到统一的指标体系，从而实现指挥信息系统测试评估数据的有效集成。其中，本体概念语义相似度的计算是本体映射技术的关键，计算结果的准确性直接影响着本体映射的科学性。因此，如何提高本体概念语义相似度计算的准确性逐渐成为本体映射、信息检索、语义网络和数据集成等领域研究的热点。

2 相关研究

目前本体概念语义相似度计算的方法主要有 4 种，分别是基于语义距离、基于信息内容、基于属性和混合式的相似度计算方法[2]。

在基于语义距离的相似度计算方法研究中，Rada 等[3]认为本体概念间的语义相似度与其在本体分类树中的距离有关，距离越大则相似度越小，进而提出 Shortest Path 法；Wu 等[4]提出基于本体概念对与其最近公共父节点的位置关系的计算方法；Leacock 等[5]提出了考虑本体树深度的相似度计算方法。

在基于信息内容的计算方法的研究中，Goble 等[6]提出考虑共享父节点所包含的信息内容来计算概念对之间相似度的计算方法；Resnik[7]提出考虑本体概念对的公共父节点中信息量最大的节点信息内容的改进方法；Lin[8]提出概念词属于同一本体时需要考虑自身信息内容的计算方法。

在基于属性的计算方法研究中，Tversky[9]提出只利用本体属性集合中的信息的计算方法；Wan 等[10]提出通过两个概念在本体树中的注释重合程度进行计算的方法；张忠平等[11]提出属性名称、数据类型和属性值的属性相似度计算模型。

在混合式的计算方法研究中，Li 等[12]提出同时考虑概念对的最短路径、最近公共父节点的深度以及局部密度等信息的计算方法；张沪寅等[13]提出综合考虑概念对路径重合度、节点密度、节点深度和概念属性等因素的 PRSSC 方法，权值由人为确定；郑志蕴等[14]提出自适应相似度综合加权计算的 ACWA 方法，综合考虑概念对的信息量、距离和属性的因素，使用主成分分析法计算权值，然后进行线性加权计算；Gao 等[15]提出基于树结构的本体概念相似度计算模型，考虑不同概念的出现次数，引入专家权重，一定程度地提高了概念相似度计算的准确性；韩学仁等[16]提出基于 PSO-BP 算法的地理本体概念语义相似度度量，用粒子群算法改进 BP 神经网络进行本体概念的综合相似度计算。

除了本体概念语义相似度的 4 种主要计算方法，基于语义网的词语相似度计算也是使用较为广泛的方法。郭小华等[17]提出在路径和深度的基础上，通过边权重改善 WordNet 结构中的层次不均匀性，利用余弦函数修正相似度计算结果的非线性误差，具有一定的借鉴意义。Fan 等[18]提出基于 HowNet 的词相似度计算的三大步骤，提升了词语相似度计算的准确性。受语义网的启发，基于知识图谱的概念相似度计算得到发展，李阳等[19]通过领域知识图谱计算实体间的相似度，运用了相似实体推荐及知识推理，相似度计算效果较好。

总体来看，基于语义网和知识图谱的语义相似度计算方法在各领域的表现不尽相同，本文涉及作战指挥等术语知识，目前没有较为权威的领域内语义网络和知识图谱作为参考，无法准确计算相关本体概念的相似度，因此不予以运用。而近年来本体概念语义相似度计算方法中仅考虑单一要素的方法研究不断减少，对混合式的相似度计算方法的研究逐渐成为热门，但大多数模型的因子权重过于依赖专家意见和经验数据，同时采用简单的线性加权法，计算准确性较差。神经网络通过对样本的训练可以准确地模拟出复杂计算模型中各因子之间的关系，有效适用于本体概念语义相似度的综合计算模型。但是，常用的 BP 神经网络存在着收敛速度慢、振荡、容易陷入局部最优点等问题，文献[16]中的 PSO 算法存在容易产生早熟收敛、局部寻优能力较差等缺陷，尤其是在处理复杂的多峰搜索问题时，将导致改进后模型计算的准确性不够理想。而启发式算法中的模拟退火算法在高温时的全局搜索能力和低温时的局部寻优能力可以较好地与 BP 神经网络相结合，进一步提高了模型计算的准确性。因此，本文提出基于模拟退火算法改进的 BP 神经网络综合计算模型来计算本体概念对的相似度。

3 模拟退火算法改进的 BP 神经网络算法

3.1 BP 神经网络

BP 神经网络是以 Rumelhart 等[20]为首的科学家小组于 1986 年提出的，是一种按误差逆向传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。该模型使足够多的样本集经过误差逆向传播算法的训练，得到稳定的具有最优权值阈值的神经网络模型。

典型的 BP 神经网络具有 3 层结构，分别为输入层、隐含层和输出层，如图 1 所示。设 BP 神经网络的网络结构包括 n 个输入层节点、p 个隐含层节点、q 个输出层节点；输入向量为 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$，隐含层输入向量为 $\boldsymbol{hi}=(hi_1,hi_2,\cdots,hi_p)$，隐含层输出向量为 $\boldsymbol{ho}=(ho_1,ho_2,\cdots,ho_p)$，输出层输入向量为 $\boldsymbol{yi}=(yi_1,yi_2,\cdots,yi_q)$，输出层输出向量为 $\boldsymbol{yo}=(yo_1,yo_2,\cdots,yo_q)$，期望输出向量为 $\boldsymbol{do}=(do_1,do_2,\cdots,do_q)$。输入层与隐含层的连接权值为 W_{ih}，隐含层与输出层的连接权值为 W_{ho}，隐含层各神经元的阈值为 b_h，输出层各神经元的阈值为 b_o。激活函数一般选择 Sigmoid 函数（又称 S 函数），即

$f(x)=\frac{1}{1+e^{-x}}$，误差函数为 $e=\frac{1}{2}\sum_{o=1}^{q}(do(k)-yo(k))^2$。具体的 BP 神经网络的训练学习可以分为信号的前向传播和误差的反向传播两个过程。训练过程中，向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 输入到神经网络后，通过 W_{ih} 的加权计算得到 $\boldsymbol{hi}=(hi_1,hi_2,\cdots,hi_p)$，再通过 Sigmoid 函数以及阈值 b_h 计算得到 $\boldsymbol{ho}=(ho_1,ho_2,\cdots,ho_p)$，同理经过 W_{ho}、Sigmoid 函数和阈值 b_o 的计算最终得到输出向量 $\boldsymbol{yi}=(yi_1,yi_2,\cdots,yi_q)$。根据误差函数计算输出与期望之间的误差，通过误差对 W_{ho} 的偏导数来修正 W_{ho}，再将误差传播到隐含层节点上，通过误差对 W_{ih} 的偏导数来修正 W_{ih}。然后从样本中抽取另一个输入向量进行同样的训练，直至误差达到预设精度或者达到迭代次数的要求，从而完成模型的训练。

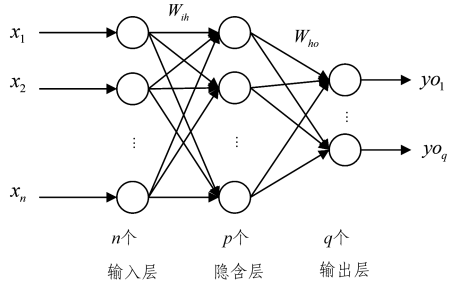

图 1 神经网络示意图

Fig. 1 Schematic diagram of neural network

3.2 模拟退火算法

模拟退火（Simulated Annealing，SA）算法的思想最早由 Metropolis 等[21]于 1953 年提出。1983 年，Kirkpatrick 等[22]成功将模拟退火的思想应用到组合优化领域。该算法来源于高温固体退火的原理，将固体的温度加热到足够高，使得分子呈随机排列状态，再让其缓慢冷却。其主要分为 3 个过程：1）加温过程，固体内部粒子随温度的升高变为无序状态，消除系统原先可能存在的非均匀状态；2）等温过程，对于与外界环境交换热量而温度不变的封闭系统，系统状态的自发变化总是朝着自由能减少的方向进行的，当自由能达到最小时，系统达到平衡态；3）冷却过程，粒子的热运动减弱并趋于有序，系统能量逐渐降低，最后在常温时达到基态，得到低能量的晶体结构。

根据 Metropolis 准则[23]，粒子在温度 T 时出现能量 ΔE 时降温的概率为：

$$p(\mathrm{d}E)=\exp(-\frac{\Delta E}{kT}) \tag{1}$$

其中，exp 表示自然指数，E 为温度 T 时的内能，ΔE 为其改变量，k 为 Boltzmann 常数。设目标函数为 $f(\boldsymbol{x})$，初始温度为 T_0，温度下限为 $T_{\min}$，当前温度为 t，当前的可行解为 $\boldsymbol{x}$，根据扰动模型得到的新解为 $\boldsymbol{x}'=\boldsymbol{x}+\Delta\boldsymbol{x}$，则对应的能量差定义为 $\Delta f=f(\boldsymbol{x}')-f(\boldsymbol{x})$。若 $\Delta f<0$ 则接受 $\boldsymbol{x}'$ 作为新的当前解，否则以概率 $p(\Delta f)=\exp(-\frac{\Delta f}{kt})$ 接受 $\boldsymbol{x}'$ 作为新的当前解。满足迭代次数后缓慢降温，重置迭代次数，最终当 $t<T_{\min}$ 时，停止迭代，输出最优解 $\boldsymbol{x}'$。

3.3 模拟退火算法改进的 BP 神经网络算法

模拟退火算法具有全局搜索能力和局部寻优能力，将其引入到 BP 神经网络的权值和阈值更新过程中，使用随机扰动模型取代梯度下降的搜索模型，以解决 BP 神经网络容易出现局部最优、收敛慢和振荡等问题，从而得到更精确的计算值。文献[24]采用模拟退火算法改进 BP 神经网络，利用 k-交换法产生新解，并将其应用于二分类和三分类问题，取得了较好的效果。本文对其产生新解的扰动模型进行改进，并将改进的神经网络模型应用到求取误差函数最小值的问题中。具体地，取 BP 神经网络中的误差函数 $e=\frac{1}{2}\sum_{o=1}^{q}(do(k)-yo(k))^2$ 作为目标函数；定义解向量 $\boldsymbol{x}$ 为 BP 神经网络中连接权值为 W_{ih}，W_{ho} 以及阈值 b_h，b_o 的综合向量；定义随机扰动产生新解的模型为 $\boldsymbol{x}'=\boldsymbol{x}+\alpha\times\frac{t}{M}\times\mathrm{rand}(-1,1)$，其中 α 为扰动幅度参数，$\frac{t}{M}$ 表示当前温度与最大迭代次数的比值，rand(−1，1)表示在−1 到 1 之间产生随机数的函数，从而实现了高温时的粗粒度搜索和低温时的细粒度寻优。使用模拟退火改进 BP 神经网络算法的具体步骤如下：

Step 1 初始化初始温度 T（足够大）、温度下限 $T_{\min}$（足够小），定义每个 t 状态下的迭代次数为 M，根据 BP 神经网络的训练产生初始连接权值，并与阈值合并成初始解向量 $\boldsymbol{x}$。

Step 2 对 $m=1,2,\cdots,M$ 做 Step 3 到 Step 6 的迭代。

Step 3 根据随机扰动模型产生新解 $\boldsymbol{x}'$。

Step 4 以新解 $\boldsymbol{x}'$ 中的连接权值与阈值作为 BP 神经网络的新的连接权值与阈值，重新计算误差函数 e'，计算误差增量 $\Delta e=e'-e$。

Step 5 若 $\Delta e<0$，则接受 $\boldsymbol{x}'$ 作为新的当前解，否则以概率 $p(\Delta e)=\exp(-\frac{\Delta e}{kt})$ 接受 $\boldsymbol{x}'$ 作为新的当前解。

Step 6 满足迭代次数后 t 缓慢降温，重置迭代次数，跳转至 Step 2。

Step 7 当 $t<T_{\min}$ 时，算法终止，输出最优解 $\boldsymbol{x}'$。

算法终止时最终确定的解向量 $\boldsymbol{x}'$ 中的连接权值及阈值即为 BP 神经网络计算模型中针对当前训练样本的权值阈值的最优解。改进后的算法表示为 SA-BP 算法。

4 基于 SA-BP 算法的语义相似度综合计算

4.1 基于语义距离的相似度计算

基于本体概念语义距离的相似度计算的经典方法是通过计算两个概念之间的几何距离来计算语义距离的。其中几何距离是通过概念对到最近公共父节点的路径和来计算的，计算模型为：

$$Sim^{sd}(c_1,c_2)=\frac{1}{Dis(c_1,c_2)+1} \tag{2}$$

其中，$Dis(c_1,c_2)=mp(c_1,RCPN(c_1,c_2)+mp(c_2,RCPN(c_1,c_2))mp(c_1,RCPN(c_1,c_2))$ 表示概念 c_1 到 c_1 与 c_2 的最近公共父节点的最短路径，$mp(c_2,RCPN(c_1,c_2))$ 表示概念 c_2 到 c_1 与 c_2 的最近公共父节点的最短路径。

Wu 等[4]提出的改进方法的主要思想是，在本体树中概念的层次越低，概念之间的相似度就越小。在具体计算时，其

考虑了两个概念最近公共父节点在本体树中的深度。Leacock 等[5]提出的改进方法除了考虑概念对之间的最短路径外，还考虑了其所处本体树的最大深度。综合以上研究，本文给出基于本体概念语义距离的相似度计算公式：

$$Sim^{sl}(c_1,c_2)=\frac{dp(RCPN(c_1,c_2),c_1)+dp(RCPN(c_1,c_2),c_2)}{(Dis(c_1,c_2)+1)\times(\max(dp(c_1))+\max(dp(c_2)))} \tag{3}$$

其中，$dp(RCPN(c_1,c_2),c_1)$表示概念对 c_1 和 c_2 的最近公共父节点在概念 c_1 所在本体树中的深度，$\max(dp(c_1))$表示概念 c_1 的本体树的最大深度。

4.2 基于信息内容的相似度计算

基于本体概念信息内容的相似度计算需要考虑概念对最近公共父节点的信息量，同时根据 Lin 等[8]的研究的改进，还需要考虑概念对本身的信息量。他们提出的计算方法为：

$$Sim^{ic}(c_1,c_2)=\frac{2\times IC(RCPN(c_1,c_2))}{IC(c_1)+IC(c_2)} \tag{4}$$

其中，$RCPN(c_1,c_2)$表示概念对 c_1 和 c_2 在本体树中的最近邻公共父节点，$IC(c_1)$和 $IC(c_2)$分别表示概念 c_1 和 c_2 包含的信息量。信息量的计算公式为：

$$IC(c)=-\log\frac{N(c)}{N} \tag{5}$$

其中，$N(c)$为概念 c 在训练样本中出现的次数，N 为训练样本的总数。

本文认为本体概念间的语义关系应作为信息内容相似度的一部分。概念对主要存在同义关系、IS-a 关系、Part-whole 关系或者 Other 关系。相比较而言，存在同义关系的概念对的信息内容相似度应明显高于后几种关系。定义贡献度为：

$$sr(c_1,c_2)=\begin{cases}1, & \text{同义关系}\\ 0.6, & \text{Is-a 关系}\\ 0.3, & \text{Part-whole 关系}\\ 0.1, & \text{Other 关系}\end{cases} \tag{6}$$

综上，基于本体概念信息内容的相似度计算公式为：

$$Sim^{ic}(c_1,c_2)=\frac{2\times IC(RCPN(c_1,c_2))\times sr(c_1,c_2)}{IC(c_1)+IC(c_2)} \tag{7}$$

4.3 基于概念属性的相似度计算

基于本体概念属性的相似度计算主要考虑能够表明概念特征的属性共有程度，Tversky[9]提出的算法就是通过计算概念对所共有的属性个数来计算属性相似度，公共属性个数越多，相似度就越大。张忠平等[11]提出属性信息包含属性名称、属性数据类型和属性值 3 个要素，综合计算属性 3 个要素的相似度更为科学。本文直接引用此计算模型进行概念属性的相似度计算。其中，属性名称和数据类型都是字符串形式，利用字符串相似度计算方法进行计算，如汉明距离计算法、余弦距离计算法。模型中设概念 c_1 和 c_2 的属性分别为 a 和 b，则属性 a 和 b 的相似度计算公式为：

$$Sim^{ab}(a,b)=\omega_1\times sim(a_{\text{name}},b_{\text{name}})+\omega_2\times sim(a_{\text{datatype}},b_{\text{datatype}})+\omega_3\times sim(a_{\text{value}},b_{\text{value}}) \tag{8}$$

其中，$\omega_1,\omega_2,\omega_3$为权重参数且 $\omega_1+\omega_2+\omega_3=1$，$sim(a_{\text{name}},b_{\text{name}})$表示属性名称相似度值，$sim(a_{\text{datatype}},b_{\text{datatype}})$表示属性数据类型相似度值，$sim(a_{\text{value}},b_{\text{value}})$表示属性值相似度值。当概念 c_1 和 c_2 共计算出 m 个 $Sim^{ab}(c_1,c_2)$时，设每个属性对的权重为 ω_k，则 c_1 和 c_2 基于本体概念属性的相似度计算公式为：

$$Sim^{pr}(c_1,c_2)=\frac{\sum_{k=1}^{m}w_{ab}^{k}Sim^{ab}(a_k,b_k)}{\sum_{k=1}^{m}w_{ab}^{k}} \tag{9}$$

4.4 基于子节点重合度的相似度计算

在实际计算过程中，除了本体概念的语义距离、信息内容以及概念属性对本体概念的相似度有贡献外，被比较概念对的子节点的重合程度也体现了概念对的相似度信息。子节点共有重合程度高的概念对的相似度显然高于子节点重合程度低的概念对。因此，本文给出基于本体概念子节点的相似度计算公式：

$$Sim^{ln}(c_1,c_2)=\frac{2\times LN(c_1\cap c_2)+1}{LN(c_1\cup c_2)+1} \tag{10}$$

其中，$LN(c_1\cup c_2)$表示概念对 c_1 和 c_2 所包含的子节点集合中元素的个数，$LN(c_1\cap c_2)$表示概念 c_1 和 c_2 共有的子节点集合中元素的个数。

4.5 基于 SA-BP 算法的语义相似度综合计算模型

为了更加科学地计算本体概念语义相似度，解决现有算法中人为确定权重的主观性问题和线性加权法可能存在的不准确问题，在基于语义距离、信息内容、概念属性和子节点重合度的相似度计算模型的计算结果的基础上，采用 SA-BP 算法模型进行本体概念语义相似度的综合计算。其综合计算模型如图 2 所示。

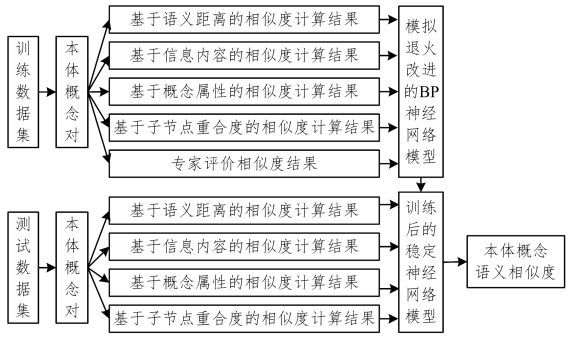

图 2 语义相似度的综合计算模型

Fig. 2 Comprehensive calculation model of semantic similarity

通过 SA-BP 算法模型对训练数据集进行训练，训练数据来自专家对样本的评价数据。将训练数据中本体概念对的基于语义距离、信息内容、概念属性和子节点重合度的相似度计算结果和专家评价的对应的相似度值输入到神经网络模型中，通过模型的学习计算出针对训练数据集的稳定最优的连接权值和阈值。之后将稳定的模型运用到测试数据集，测试数据集即本领域海量的没有经过专家评价的本体概念对，将测试数据基于语义距离、信息内容、概念属性和子节点重合度的相似度的计算结果输入到模型中，完成一次正向过程，便可以得到较为科学的本体概念语义相似度的综合计算结果。

5 实验结果与分析

5.1 实验数据集与参数选定

从不同单位建立的指挥信息系统测试评估指标中挑选

150 组本体概念对，来进行基于 SA-BP 算法的语义相似度综合计算模型的实验。随机选取其中 140 组本体概念对作为训练样本，用于训练 SA-BP 算法模型，将剩下的 10 组作为测试样本，用于检验模型的效果。

本文的测试仿真环境为 Python3.7，实验模型中相关参数的选定如下：BP 神经网络采用典型的三层结构，输入层节点个数为 4，隐含层节点个数根据经验以及测试效果选择为 4，输出节点个数为 1，设初始温度为 $T=100$，温度下限为 $T_{\min}=0.01$，每个温度迭代次数为 $M=100$，随机扰动幅度参数为 $\alpha=5$。

5.2 实验结果与分析

实验采用了 5 种算法对测试样本进行相似度计算，并将相关的实验结果进行比较分析。5 种算法分别为：基于模拟退火改进 BP 神经网络的综合相似度计算算法、基于 PSO-BP 神经网络的综合相似度计算算法、基于主成分分析获取权重的线性加权算法、基于概念属性的相似度计算算法以及基于 HowNet 的相似度计算算法。实验结果如表 1 所列。

表 1 测试样本中本体概念对语义相似度的综合计算结果

Table 1 Comprehensive calculation results of ontology concept semantic similarity in test samples

本体概念对相似度		SA-BP 算法	PSO-BP 算法	主成分分析法	基于概念属性	基于 HowNet
Sim(战场数据化能力，战场信息化能力)	0.9132	0.9081	0.8563	0.8034	0.7613	0.8431
Sim(信息更新速度，数据处理能力)	0.6397	0.4573	0.4134	0.5024	0.4732	0.4025
Sim(覆盖范围敏感度，云计算精度)	0.1886	0.3045	0.3303	0.3728	0.3816	0.0556
Sim(FDMA，频分多址)	1.0000	0.9493	0.8730	0.8141	0.5796	0.0000
Sim(传输误码率，I/O 性能)	0.2181	0.3686	0.2983	0.3617	0.3362	0.0437
Sim(组网方式，网络可扩展性)	0.7527	0.8744	0.5173	0.4978	0.4034	0.3134
Sim(丢包率，存储容量)	0.2633	0.3645	0.3801	0.4895	0.4272	0.2113
Sim(网络故障率，节点连通距离)	0.5993	0.7208	0.3503	0.3291	0.4694	0.1325
Sim(平均故障修复时间，容灾备份能力)	0.5449	0.6928	0.6936	0.4017	0.5243	0.1798
Sim(通信容量，存储容量)	0.6743	0.6210	0.4998	0.5938	0.5691	0.4973

以专家评价的相似度数值为标准，将 5 种算法的计算结果与专家评价数据进行比较，计算误差值。具体误差结果包括误差的最大值、平均值和标准差，如图 3 所示。

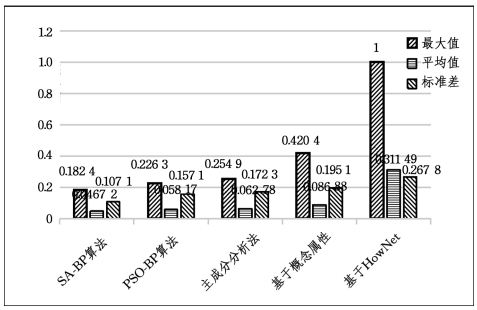

图 3 5 种算法的计算结果误差

Fig. 3 Errors in calculation results of five algorithms

在与专家评价结果进行比较的过程中，引入 Pearson 相关系数[25]来评价这 4 种算法针对测试样本所计算出的相似度和专家评价数据的一致性。Pearson 相关系数的定义为：

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E((X-EX)(Y-EY))}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{11}$$

其中，$\operatorname{cov}(X,Y)$ 表示协方差，E 表示数学期望或均值，D 表示方差，$\sqrt{D(X)}$ 表示 X 的标准差。Pearson 相关系数所对应的相关性强度，即一致性程度如表 2 所列。

表 2 Pearson 相关系数强度

Table 2 Intensity of Pearson correlation coefficient

Pearson 相关系数	相关强度
0.9～1.0	极强相关
0.8～0.9	强相关
0.6～0.8	中等程度相关
0.4～0.6	弱相关
0.0～0.4	极弱相关或不相关

通过对实验结果的计算，5 种算法的相似度计算结果与专家评价数据的 Pearson 相关系数如表 3 所列。

表 3 不同算法的 Pearson 相关系数计算结果

Table 3 Calculation results of Pearson correlation coefficients for different algorithms

算法	SA-BP 算法	PSO-BP 算法	主成分分析法	基于概念属性	基于 HowNet
Pearson 相关系数	0.9184	0.8489	0.7833	0.6847	0.3716

从图 3 和表 3 中可以看出，基于 SA-BP 算法的相似度综合计算结果与专家评价结果之间的 Pearson 相关系数最大，同时误差的最大值、平均值、标准差都为最小。实验结果表明，基于语义网的相似度计算由于缺少领域内的相关知识，计算结果误差较大，效果不尽人意。单因素的基于概念属性的相似度计算结果也存在较大的误差，效果不佳。相比而言，综合计算的算法准确性更高，基于神经网络的相似度综合计算算法相比文献[14]中的基于主成分分析和线性加权的算法更加科学和准确；相比于文献[16]中的 PSO-BP 算法，本文的 SA-BP 算法在样本中的寻优和收敛效果更好，与专家评价结果表现出了更好的一致性。

结束语 本文研究了基于 SA-BP 算法的本体概念语义相似度综合计算，在基于语义距离、信息内容、概念属性的相似度计算模型的基础上，提出了基于子节点重合度的相似度计算模型；而后将相似度计算结果输入神经网络模型进行综合相似度计算，引入模拟退火算法，对神经网络的梯度下降的搜索方式进行改进，优化了 BP 神经网络的寻优效果。实验结果表明，基于 SA-BP 算法的本体概念综合语义相似度计算模型具有较好的收敛效果和较高的准确性。下一步的工作主要是增强模型对不同样本的适应性，完成大量的指挥信息系统测试评估指标的映射和集成。

参 考 文 献

[1] SCHADD F C, ROOS N. Word-Sense Disambiguation for Ontology Mapping: Concept Disambiguation using Virtual Documents

and Information Retrieval Techniques[J]. Journal on Data Semantics, 2015, 4(3): 167-186.
[2] GAO W, FARAHANI M R, ASLAM A, et al. Distance learning techniques for ontology similarity measuring and ontology mapping[J]. Cluster Computing, 2017, 20(2): 959-968.
[3] RADA R, MILI H, BICKNELL E, et al. Development and application of a metric on semantic nets[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2002, 19(1): 17-30.
[4] WU Z, PALMER M. Verb Semantics and Lexical Selection [C]// Proceedings of 32nd Annual Meeting on Association for Computational Linguistics. LasCruces, New Mexico, 1994: 133-138.
[5] LEACOCK C, CHODOROW M. Combining Local Context and WordNet Similarity for Word Sense Identification[M]. WordNet: An Electronic Lexical Database, 1998.
[6] GOBLE A, STEVENS J R, BRASS C A, et al. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation[M]. Oil Bhales of the World: Pergamon Press, 2003.
[7] RESNIK P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language [J]. Journal of Artificial Intelligence Research, 2011, 11(1): 95-130.
[8] LIN D. An Information-Theoretic Definition of Similarity[C]// International Conference on Machine Learning. New Brunswick, NJ, 1998.
[9] TVERSKY A. Features of Similarity[J]. Readings in Cognitive Science, 1988, 84(4): 290-302.
[10] WAN S, ANGRYK R A. Measuring semantic similarity using WordNet-based Context Vectors[C]// 2007 IEEE International Conference on Systems, Man and Cybernetics. Montreal, 2007: 908-913.
[11] ZHANG Z P, TIAN S X, LIU H Q. A Comprehensive Method for Calculating Ontology Similarity[J]. Computer Science, 2008, 35(12): 142-145.
[12] LI Y, ZA B. An approach for measuring semantic similarity between words using multiple information sources [J]. IEEE Transactions on Knowledge & Data Engineering, 2003, 15(4): 871-882.
[13] ZHANG H Y, WEN C Y, LIU D B, et al. Improved ontology-based semantic similarity calculation[J]. Computer Engineering and Design, 2015, 36(8): 2206-2210.
[14] ZHENG Z Y, RUAN C Y, LI L, et al. Research on Adaptive Synthetic Weighting Algorithm for Ontology Semantic Similarity[J]. Computer Science, 2016, 43(10): 242-247.
[15] GAO X R, XU Y Z. Research on Improved Model for Concept Similarity Computation in Domain Ontology and Application [C]// 2017 International Conference on Robots & Intelligent System (ICRIS). Huai'an, 2017: 257-261.
[16] HAN X R, WANG Q S, GUO Y, et al. Semantic similarity measure of geographic ontology based on PSO-BP algorithm[J]. Computer Engineering and Applications, 2017, 53(8): 32-37.
[17] GUO X H, PENG Q, DENG H, et al. WordNet word similarity calculation based on edge weight [J]. Computer Engineering and Application, 2018, 54(1): 172-178.
[18] FAN M, ZHANG Y, LI J. Word similarity computation based on HowNet[C]// International Conference on Fuzzy Systems & Knowledge Discovery. IEEE, 2016.
[19] LI Y, GAO D Q. Research on Entity Similarity Computation in Knowledge Map [J]. Chinese Journal of Information Science, 2017, 31(1): 145-151, 159.
[20] WANG S, NA Z, LEI W, et al. Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method[J]. Renewable Energy, 2016, 94(1): 629-636.
[21] METROPOLIS N, ROSENBLUTH A W, ROSENBLUTH M N, et al. Equation of State Calculations by Fast Computing Machines[J]. The Journal of Chemical Physics, 2004, 1087(1953): 21.
[22] KIRKPATRICK S, VECCHI M P. Optimization by simulated annealing[M]. Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications, 1987.
[23] MAMANO N, HAYES W B. SANA: Simulated Annealing far outperforms many other search algorithms for biological network alignment[J]. Bioinformatics, 2017, 33(14): 1-9.
[24] ZHOU A W, ZHAI Z H, LIU H T. An improved BP neural network algorithm based on simulated annealing algorithm [J]. Microelectronics and Computer, 2016, 33(4): 144-147.
[25] DE WINTER J C, GOSLING S D, POTTER J. Comparing the Pearson and Spearman Correlation Coefficients Across Distributions and Sample Sizes: A Tutorial Using Simulations and Empirical Data[J]. Psychological Methods, 2016, 21(3): 273.

XU Fei-xiang, born in 1995, postgraduate. His main research interests include ontology integration and semantic network.

YE Xia, born in 1977, Ph.D, associate professor. Her main research interests include database technology and computer network.