文章编号：2096-1472(2020)-10-05-03
DOI:10.19644/j.cnki.issn2096-1472.2020.10.002

基于强类别特征的文本相似度计算及其性能评估

刘 辉

(上海理工大学信息化办公室，上海 200093)
✉liu_hui@usst.edu.cn

摘 要：本文基于强类别特征识别算法，研究一种文本语义相似度的计算算法并对其性能进行评估。为实现该功能并形成一种通用算法，本文设计了一种基于语义识别码的语义函数库作为比较对象，使用两次模糊神经元深度卷积机器学习算法模块，并在两次机器学习之间使用一次基于傅立叶变换的频域特征提取的刚性算法，最终在该算法模块前后使用外置的数据模糊算法和解模糊算法，实现了一个较复杂的机器学习通用算法。而该算法也是本文的一次技术创新。通过基于志愿者主观评价的性能评估，发现该系统重点实现了汉语言的文本语义相似度评价，且实现了81.78%的人工判断准确率对比结果，且只有5.52%的志愿者认为系统判断结果与人工判断结果完全不一致。

关键词：强类别特征算法；机器学习；文本相似度；语义识别；性能评估
中图分类号：TP309 **文献标识码**：A

Text Similarity Calculation and Performance Evaluation based on Strong Category Features

LIU Hui

(*Information Office*, *University of Shanghai for Science and Technology*, *Shanghai* 200093, *China*)
✉liu_hui@usst.edu.cn

Abstract: This paper studies the algorithm of text semantic similarity calculation and its performance evaluation, based on the recognition algorithm of strong category features. In order to realize this function and form a general algorithm, this paper designs a semantic function library based on the semantic identification code as the comparison object, uses two fuzzy neuron deep convolution machine learning algorithm modules. Between two machine learning modules, one frequency domain feature extraction rigid algorithm is used based on Fourier transform. Finally, a more complex general algorithm of machine learning is realized by using external data before and after the algorithm module. This algorithm is also a technical innovation. Through the subjective performance evaluation of volunteers, it is found that the system realizes the semantic similarity evaluation of Chinese text, and achieves 81.78% of the compared manual judgment accuracy rate, and only 5.52% of the volunteers think that the results of the system are completely inconsistent with the results of manual judgment.

Keywords: strong class feature algorithm; machine learning; text similarity; semantic recognition; performance evaluation

1 引言(Introduction)

如果单纯比较文本的BIG码串或者ASIC码串，几乎不可能获得文本语义上的相似度，比如“今天是晴天。”和“It is sunny today.”两串文本之间，如果不使用深度机器学习，很难实现对其语义的比较[1−3]。再比如“今天是晴天。”和“冬日里阳光和煦。”之间，更无法使用传统方式对其进行语义相似度的比较。而如果单纯使用任何一种神经元网络架构对上述字符串之间进行比较，也很难实现足够精确的文本语义相似度的比较结果[4−6]。所以，近年来基于语义函数库和频域特征的前置机器学习比较算法提取文本语义特征串，结合后置机器学习文本语义特征串的比较算法，在当前文本相似度比较领域得到了较广泛的应用。

2 语义函数库的搭建模式(Building model of the semantic function library)

早期无法使用语义函数库对相关语义比较过程进行大数据支持，是因为函数库的数据结构难以得到有效且高效的设置。因为汉语言中的名词、代词、动词、副词、形容词、介词等，均有相对独立且几乎没有重合度的语义评价指标[7−9]。特别是名词用作动词、名词用作形容词、虚介词等复杂语法环境下，即便使用机器判断其真实的词性词义都是一个复杂

的计算量。部分研究中使用多级模糊比较的方式实现对语义函数库的搭建，即使用词性比较模块先划分输入词的词性，再根据其上下文和二级库实现对其语义语境的判断。

本文重点研究语义的直接模糊实现，即该语义函数库的输出目标并非针对人机界面的直接判断输出而是采用一个语义深度码指标，面向后续机器学习模块进行语义识别，比如图1所示。

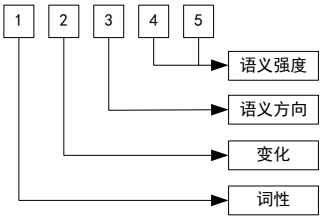

图1 语义识别码定义图

Fig.1 The definition diagram of semantic identification code

如此，在语义识别库中，将每个固定词转化为一个5位的语义识别码，该识别码在实际刚性比较过程中并未能提供数据支持，但足以在机器学习中提供异构化自然文本数据的同构化支持过程。

在语义识别库中，可能存在一个固定词对应多个语义识别码的情况，比如“观察”一词，可能对应名词的弱语义强度选项，也可能对应动词的强语义强度选项，且其也可以用作名词转动词或者动词转名词的应用。这就需要在卷积神经网络支持下进行根据上下文的语义筛选机器学习判断。该判断模式将在下文重点分析[10]。

实际语义函数库的设计过程，并不需要对现代汉语词典中的每个词均进行语义函数特征的设计，只需要对2000—3000个常用词的语义语势进行囊括，即可实现对大部分汉语言词语的语义语势提供对比数据支持，即在该语义函数库中，约包含8000—12000个比较关联函数。

3 强类别特征比较的整体算法模式分析(Analysis of the whole algorithm pattern of strong category feature comparison)

如图2所示，系统中输入两个待比较的字符串，字符串长度不限。在强卷积和流输入模式下，在语义函数库的支持下，使用一个模糊神经元卷积网络实现对其语义串的生成过程。使用傅立叶变换函数为核心基函数的频域特征分析模块，对该算法生成的语义串进行二次分析，各得到一个时域特征串。然后两列特征串经过一个模糊神经元卷积网络得到一个比较值Double结果，最后使用一个解模糊模块对其进行解模糊计算，使用一个普通格式化输出模块对其比较结果进行输出。这一整合算法共使用了两个模糊神经元卷积网络对两列字符串进行了语义比较，可以最大程度减少每个神经元网络的算力需求，以提升系统效率。

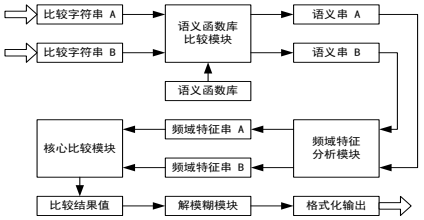

图2 强类别特征比较整体算法数据流图

Fig.2 Data flow graph of the algorithm for strong category feature comparison

3.1 语义函数库比较模块设计

语义函数库采用双环卷积的流数据比较模式运行，即针对语义函数库的每记录输入，分别对比较字符串进行遍历，获得对应结果并进行输出。该模块属于典型的模糊神经元深度卷积网络算法，其核心控制变量为语义函数库的指针变量，次要控制变量为两列比较字符串的指针变量。输出变量为针对两列比较字符串指针的语义串数据。详见图3。

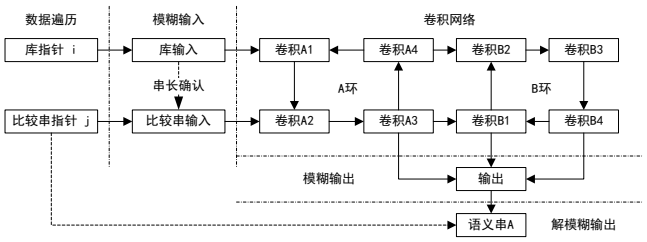

图3 语义函数库比较模块设计图

Fig.3 Comparison module design of the semantic function library

在图3功能设计中，两个比较字符串独立实现比较，即实现一个针对比较函数库的遍历指针i，针对每个i，对比较串中根据比较函数库中的目标字符串长度，使用一个指针j对比较字符串进行逐字符遍历，形成一个比较串指针。即对于库输入和比较串输入来说，其字符串长度相等。本文限定其每个比较字符串的长度不超过4字符即8字节。古该模糊神经元网络的唯一两个输入量，长度均不超过8字节的Bit数据。但因为系统需要充分考虑上下文影响，所以应对该输入数据进行深度卷积，所以该模块使用了双环卷积的方法，其中A环和B环都是四个模块，每个模块按照3、7、13、5、1的隐藏层结构进行节点设计，且每个卷积模块的节点均按照高阶多项式回归的方式进行节点设计，其节点函数可写做：

$$Y=\sum_{m}\sum_{n=6}^{0}A_nX_m^n \tag{1}$$

其输入模块输入1个8字节bit变量，输出一个4字节Double变量，隐藏层按照3、7、3的隐藏层结构进行节点设计，采用线性函数对其节点进行设计。其节点函数可写做：

$$Y=\sum\left(A\cdot X_i+B\right) \tag{2}$$

其输出模块整合$A1$、$B1$、$B4$三个卷积模块的输出量，均为Double变量，该模块的统计学意义是将该三组输入数据充分二值化，故采用二值化函数对其进行管理。其隐藏层应达到足够的深度，故采用五层隐藏层设计，按照5、17、31、13、3的隐藏层结构进行节点设计。其节点函数可写做：

$$Y=\sum\frac{1}{A\cdot e^{X_i}+B} \tag{3}$$

在语义串的输出模块中，根据实时输出的比较结果，当其结果接近1.000时，读取特征语义识别码与比较串指针生成该比较串指针位置的语义识别码序列，而当其结果接近0.000时，该比较串指针对应的语义识别码定义为0。当比较串指针对应的语义识别码已经存在定义时，则使用算数平均法，给出均值结果。即使用该算法生成的机器学习语义串，无法从语义函数库中反查其语义，但足以供后续的三个模块生成机器学习结果。

综合该模块的子模块设计，可以得到表1。

表1 语义函数库比较模块设计参数总表

Tab.1 The parameters for comparison module of semantic function library

子模块	输入节点	输出节点	隐藏层	总节点数	节点函数
库输入比较串输入	1×64bit	1×Double	3、7、3	15	$Y=\sum(A\cdot X_i+B)$
卷积A1	2×Double	1×Double		32	
卷积A2	2×Double	1×Double		32	
卷积A3	1×Double	1×Double		31	
卷积A4	1×Double	1×Double	3、7、13、5、1	31	$Y=\sum_m\sum_{n=6}^{0}A_nX_m^n$
卷积B1	2×Double	1×Double		32	
卷积B2	2×Double	1×Double		32	
卷积B3	1×Double	1×Double		31	
卷积B4	1×Double	1×Double		31	
输出	3×Double	1×Double	5、17、31、13、3	73	$Y=\sum\frac{1}{A\cdot e^{X_i}+B}$

3.2 频域特征分析模块设计

3.1中生成的机器学习语义串的本质是一个时域函数，即其标定了在输入字符串字符顺序指针序列上的语义识别码信息。但该串仍存在一定的时域上的特异性。所以，频域特征分析模块的统计学意义是将该时域上的特异性进行削弱，从而得到一个频域特征数据。该模块需要进行一次基于时域数据的频域特征提取计算过程。而该过程通过一次傅立叶变换即可实现。

首先根据语义串的指针t，得到其语义识别码$f(t)$，对$f(t)$进行基于傅立叶变换的频域特征提取：

$$f(\omega)=\int_{-\infty}^{+\infty}f(t)\cdot e^{-i\omega t}dt \tag{4}$$

得到特征函数$f(\omega)$后，根据指针t的总长度，将$f(\omega)$进行划分，并提取其$f(\omega)$结果，即可形成其频域特征串。

该过程属于刚性计算过程，并未牵扯到任何机器学习算法，即本文是在两个模糊神经元网络模块之间，进行一个基于刚性算法的数据治理过程。

3.3 核心比较模块设计

两列频域特征串，即频域特征串A与频域特征串B，输入到核心比较模块中，该比较模块也是一个模糊神经元网络卷积算法模块。详见图4。

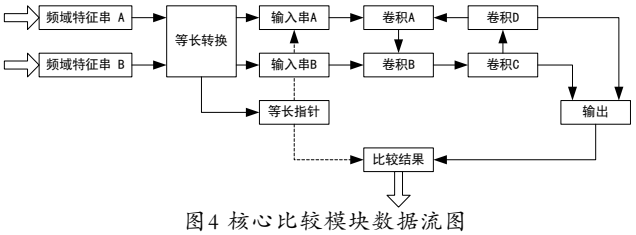

图4 核心比较模块数据流图

Fig.4 The data flow diagram of core comparison module

该模块的模糊化过程核心算法是判断两个频域特征串的长度，使用差值法将频域特征串进行等长转化。然后根据等长转化后的特征串指针作为控制变量，构成两个输入串，然后形成一个4模块(A、B、C、D)的卷积模块，其中卷积A和卷积B的统计学意义是整合输入串数据(Long型变量)到卷积循环中，卷积C和卷积D的统计学意义是为输出模块各提供一个Double数据。最终在等长指针的条件下，对所有比较结果提供一个算数平均值结果。该结果即是两组待比较字符串的模糊比较结果。

该模块的子模块设计思路与语义函数库比较模块类似，其中两个输入串按照线性回归函数进行节点管理，隐藏层结构为3、7、3，节点函数如函数(2)，四个卷积模块按照高阶多项式回归函数进行节点管理，隐藏层结构为3、7、13、5、1，节点函数如函数(1)，一个输出模块按照二值化回归函数进行接地单管理，隐藏层结构为3、7、3，节点函数如函数(3)。所以，该模块的实际设计参数汇总表如表2。

表2 核心比较模块设计参数总表

Tab.2 The parameters of core comparison module

子模块	输入节点	输出节点	隐藏层	总节点数	节点函数
输入串A 输入串B	1×Long	1×Double	3、7、3	15	$Y=\sum(A\cdot X_i+B)$
卷积A	2×Double	1×Double		32	
卷积B	2×Double	1×Double	3、7、13、5、1	32	$Y=\sum_m\sum_{n=6}^{0}A_nX_m^n$
卷积C	1×Double	1×Double		31	
卷积D	1×Double	1×Double		31	
输出	2×Double	1×Double	3、7、3	16	$Y=\sum\frac{1}{A\cdot e^{X_i}+B}$

3.4 解模糊及格式化输出模块设计

根据前文分析，该算法的最终输出结果，是经过深度代数平均计算的二值化结果均值结果，所以，最终数据的二值化特征并不显著。即该模型的最终落点基本集中在[0,1]区间上，也有部分结果超出了该区间。即该输出结果是一个深度模糊化的输出结果。

在模糊化过程中，可以限定两个阈值，即输出结果大于某值M时，此两段文本的相似度处于高置信区，输出结果小于某值N时，此两端文本的相似度处于低置信区，但仍有较大可能处于[N,M]区间中，此时系统给出一个弱相似结果。即本文算法最终的格式化输出结果中，包含三种判断结果的输出可能，即两端文本的语义强相似、弱相似、不相似，而强相似和不相似结果的输出频率，应确保在80%以上，才可以实现该算法的实际应用场景适应性。

4 算法性能评估(The performance evaluation of algorithm)

因为文本语义的相似性评价结果的本质是用户的主观评价结果，所以，在进行评估的过程中，选择100位志愿者，均为有一定文学批评功底的汉语言文学、国际汉语、汉语言教育专业的本科以上在校生，选取50对文本段进行比较，以发现系统对该50对文本段的评价结果与志愿者的人工判读结果的一致性。志愿者将对系统给出的判断结果给出非常一致(10分)基本一致(6分)不一致(3分)完全不一致(0分)的主观评价，以判断系统的文本语义相似性判断的准确率。最终评估结果中，100位志愿者在对应的5000次评价比较过程中，给出非常一致评价2763次，占55.26%，给出基本一致评价1326次，占26.52%，给出不一致评价635次，占12.70%，给出完全不一致评价276次，占5.52%。该系统的综合判断准确率(非常一致与基本一致的综合占比)为81.78%，综合主观得分为74.98分

(下转第4页)

Neural Information Processing Systems. Cambridge, MA: MITPress, 2014(06): 2672-2680.
[8] Salimans T, Goodfellow I. Zaremba W, et al. Improved techniques for training gans[J]. arXivpreprint, 2016(01): 2226-2234.
[9] LING M, WU QX. Deep unsupervised learning for image super resolution with generative adversarial network[J]. Signal Process: Image, 2018(08): 88-100.
[10] CAESAR H, UIJLINGS J, FERRARI V. Region-based semantic segmentation with end-to-endtraining[C]. Proceedings of the 14th European Conference on Computer Vision 2016, Amsterdam, The Netherlands, 2016(01): 381-397.
[11] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern. Recognition, LasVegas, USA, 2016, 39(6): 770-778.
[12] Luo Huilan, Lu Fei, Kong Fansheng. Image semantic segmentation based on region and depth residual network[J]. Journal of electronics and information, 2019, 41 (11): 2777-2786.
[13] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[14] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Comp, 2006, 18(7): 1527-1554.
[15] Aliya. Aierken, Halidan. Abudureyimu, et al. A Uyghur span classification method based on deep belief networks[J]. Computer Engineering & Science, 2016, 10(10): 2134-2139.
[16] Shi Ke, Lu Yang, Liu Guangliang, Bi Xiang, et al. Deep belief network training method based on multi hidden layer Gibbs sampling[J]. Acta automatica Sinica, 2019, 45(05): 975-984.
[17] 赵文迪,陈德旺,卓永强,等.深度神经模糊系统算法及其回归应用[J/OL].自动化学报,2020,46(8):1557-1570.

作者简介：

孙运文(1996-)，男，硕士生.研究领域：医学信息学.
徐秀林(1957-)，女，博士，教授.研究领域：康复设备与骨科器械，医学信息学.

(上接第7页)

(满分100分)。

在5000次评价中，系统共给出强相似评价1031次，占10.62%，弱相似评价391次，占7.82%，不相似评价3578次，占71.56%，其中强相似评价与不相似评价之和为4609次，占92.18%，满足了本文设计需求(详见3.4)。

5 结论(Conclusion)

该系统重点实现了汉语言的文本语义相似度评价，且实现了81.78%的人工判断准确率对比结果，且只有5.52%的志愿者认为系统判断结果与人工判断结果完全不一致。因为当前基于机器学习的自然语言语义判断仍属于尖端课题，实现该判断准确率较相关文献针对单一判断目标的判断准确率仍有一定的先进性。该系统是一种通用语义判断算法，通过对语义函数库的进一步完善，以及对两组判断神经元网络机器学习模块的更深度训练，该判断准确率还可以进一步提升。

参考文献(References)

[1] 刘思华,曾传禄.“能”和“会”的情态语义比较[J].沈阳大学学报(社会科学版),2020,22(01):95-100;105.
[2] 王友良.比较语义关系形容词的强语势表达探究[J].焦作大学学报,2019,33(04):7-11.
[3]祝晶.俄汉比较范畴的语义类型及其表达手段[J].中国俄语教学,2020,39(01):34-43.
[4] 颜冰,张辉.框架语义视角下中美贸易战话语的历时比较分析[J].外国语文,2020,36(01):1-8.
[5] 马慧芳,刘文,李志欣,等.融合耦合距离区分度和强类别特征的短文本相似度计算方法[J].电子学报,2019,047(006):1331-1336.
[6] 王伟,朱立明,章强,等.基于相似性分析和阈值自校正的烟箱缺条智能检测方法[J].烟草科技,2019,52(01):97-103.
[7] 宋呈祥,陈秀宏,牛强.文本分类中基于CHI改进的特征选择方法[J].传感器与微系统,2019,38(02):37-40.
[8] 何春辉.一种基于文本相似度的网页新闻标题自动抽取算法[J].湖南城市学院学报(自然科学版),2019,28(01):61-64.
[9] Liu W, Ma H, Tuo T, et al. Co-occurrence distanceand discrimination based similarity measure on short Text[J]. Computer Engineering and Science, 2018, 40(7): 1281-1286.
[10] Liu Weiru, Giunchiglia, Fausto, et al. International Conference on Knowledge Science, Engineering and Management[C]. Australia: Springer, 2018(08): 67-75.

作者简介：

刘 辉(1984-)，男，硕士，初级工程师.研究领域：信息研究，网络安全.

(上接第25页)

[7] 张建民.一种改进的K-means聚类算法[J].微计算机信息,2010(9):233-234.
[8] 黄晓辉,王成,熊李艳,等.一种集成簇内和簇间距离的加权k-means聚类方法[J].计算机学报,2019,59(42):1-15.

作者简介：

罗军锋(1976-)，男，硕士，高级工程师.研究领域：数据挖掘，区块链.
洪丹丹(1981-)，女，硕士，高级工程师.研究领域：主数据工程，高校信息化.