计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(4):303−308 [doi: 10.15888/j.cnki.csa.008424]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于句子的多属性融合相似度计算方法①

袁绍正, 周艳平

(青岛科技大学 信息科学技术学院, 青岛 266061)
通信作者: 袁绍正, E-mail: 860959136@qq.com

摘　要: 针对现有的句子相似度计算方法没有考虑句子中的关键词的多属性信息, 无法更好衡量句子相似度的问题, 综合考虑句子的结构和包含的属性, 提出一种基于句子的多属性融合相似度计算方法. 该方法通过提取句子的词频属性、词序属性、词性属性及句长属性, 采用层次分析法 (AHP) 计算出各属性的权重, 并验证权重值的合理性, 继而加权融合 4 种属性的相似度. 将本文提出的多属性融合相似度计算方法在构建的数据集上进行实验, 验证此方法的可靠性及可行性, 并以召回率、准确率以及归一化 F-度量值为标准和其他传统方法进行对比分析, 结果表明, 该方法不仅有着均衡的召回率和准确率, 且 F-度量值较高, 达到 83.57%.

关键词: 多属性; 权重; 句子相似度; 层次分析法 (AHP); F-度量值

引用格式: 袁绍正,周艳平.基于句子的多属性融合相似度计算方法.计算机系统应用,2022,31(4):303–308. http://www.c-s-a.org.cn/1003-3254/8424.html

Multi-attribute Fusion Similarity Calculation Method Based on Sentence

YUAN Shao-Zheng, ZHOU Yan-Ping

(College of Information Science and Technology, Qingdao University of Science &Technology, Qingdao 266061, China)

Abstract: The current sentence similarity calculation method does not consider the multi-attributes of the keywords in the sentence and cannot better measure the sentence similarity. Therefore, this study proposes a sentence similarity calculation method based on multi-attribute fusion, considering the sentence structure and the attributes contained. First, this method extracts the attributes of the sentence including the word frequency, word order, part of speech, and sentence length. Next, the analytic hierarchy process (AHP) is used to calculate the weight of each attribute and verify the rationality of the weight, and then the weighted fusion of the similarity of the four attributes is conducted. This proposed calculation method for multi-attribute sentence similarity is tested on the constructed dataset to verify its reliability and feasibility, and it is compared with other traditional methods in recall rates, accuracy rates, and normalized F-measure values. The results show that this method has balanced recall and accuracy rates and a high F-measure value of 83.57%.

Key words: multi-attribute; weight; sentence similarity; analytic hierarchy process (AHP); F-measure

计算句子相似度是自然语言处理领域研究的一个基础且重要的工作, 有着广泛的应用方向, 多用于智能问答、信息检索、语义分析和文本分类等场景.

目前对于句子相似度的研究停留在语义理解范围, 依托越来越庞大的数据库做大量的仿真, 做到让机器理解人类的语言, 但现有的句子相似度计算方法主要分为两大类: 基于统计的方法和基于深度学习的方法[1]. 典型的方法有莱文斯坦距离、BM25、TF-IDF、Word2Vec 余弦相似度、Jaccard 系数相似性计算等.

国内外各个学者对句子相似度的研究做了广泛的探索. Tian 等[2] 提出一种基于同义词表的改进 Word2Vec 句子相似度算法, 通过构建同义词表和融合词向量来

① 收稿时间: 2021-06-29; 修改时间: 2021-07-30; 采用时间: 2021-08-12; csa 在线出版时间: 2022-03-22

提高句子相似度计算的准确性; Wilson 等[3]提出一种使用组合语义方法来测量文档相似性的有效方法, 该方法结合了多个语义计算; 文献 [2,3] 的研究由于语义工具和应用逻辑的效率决定了应用程序的准确度和整体性能, 待进一步提升; Goz 等[4] 研究基于关键字的社交网络相似性的适用性; Ruan 等[5] 计算句子相似度是将 Word2Vec 方法和词嵌入相似度方法结合, 二者对于关键词词性信息稍欠考虑; 翟社平等[6]提出多特征的句子词形、词序及句长特征融合的相似度计算方法, 由于句子关键词存在一词多义, 只考虑了句子的字面特征, 将导致相似度匹配不准确.

句子由多个词组成也包含多种属性, 句长度、词出现的频率和词在句中的词性以及其在句中的顺序对句子语义的影响度不同, 需综合考虑到句子深层和表层所有因素[7]. 因此本文以句子的语序结构、词性信息和形态结构等特点为核心要素, 构建研究领域本体库, 通过给各属性分配权重, 提出一种基于句子的多属性融合相似度计算方法, 以提升句子相似度计算的合理性.

1 相似度概念

1.1 句子相似度

文本相似度一般指文本在语义上的相似程度[8], 句子相似度指的是句子在语义上的相似程度, 用来评估句子之间符合程度. 如果两个句子之间符合程度高, 那两者一定有相似或相同的属性, 令$SIM(S_1,S_2)$作为两个句子S_1和S_2的相似度, 则其具有以下几个特点:

(1) $SIM(S_1,S_2)\in[0,1]\cap SIM(S_1,S_2)\in R$, 表示两个句子相似度的取值;

(2) $SIM(S_1,S_2)=0$, 表示句子之间没有任何相同的属性, 两个句子不相似;

(3) $SIM(S_1,S_2)=1$, 表示两个句子在形态结构、语序结构、语义信息等方面具有完全相同的属性;

(4) $SIM(S_1,S_2)=SIM(S_2,S_1)$, 表示两个句子相似且具有对称性.

1.2 余弦相似度

两个句子的相似度可以用向量余弦值的值来衡量, 称为余弦相似度[9]. 首先, 将两个句子数字化变成向量, 其次, 计算其夹角余弦$\cos(\theta)$, 衡量两个向量之间差异的大小. 余弦值接近 1, 夹角趋于 0, 表明两个向量越相似, 余弦值接近 0, 夹角趋于 90 度, 表明越不相似. 如图 1 所示, 向量 a 和向量 b 的余弦夹角小于向量 a 和向量 c 的余弦夹角, 表示 a 和 b 具有更高的相似度.

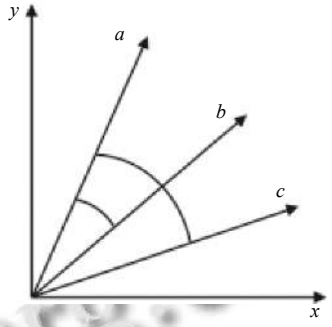

图 1 向量的余弦相似度

将句子 S_1 和 S_2 以向量v_i,v_i'表示:

$$S_1:(v_1,v_2,v_3,\cdots,v_n)$$

$$S_2:(v_1',v_2',v_3',\cdots,v_n')$$

则句子的余弦相似度计算公式为:

$$SIM(S_1,S_2)=\frac{\sum_{i=1}^{n}v_i\times v_i'}{\sqrt{\sum_{i=1}^{n}v_i^2\times\sum_{i=1}^{n}v_i^2}} \tag{1}$$

2 基于句子的多属性融合相似度计算方法

一个句子除包含的关键词外, 也不能忽略自身的一些属性, 比如词频、词序、词性和句长, 本文将 4 个属性进行加权融合得到句子相似度. 句子之间在词语形态上的相似度, 即出现共有关键词的频率为词频相似度; 句子之间共有关键词的相对位置关系的相似度为词序相似度; 句子之间共有关键词的词性的相似度为词性相似度; 两个句子的长度的关系为句长相似度.

传统方法对句子进行分词、去停等操作提取关键词进行表面特征的相似度比较, 这是不全面的, 中文自然语言不同于英文自然语言, 中文有着丰富且灵动的表达形式. 如词性方面, “退役士兵有什么需要? ”与“退役士兵需要有什么? ”, 此处的两个“需要”字面特征一致, 本质却不同, 名词和动词的词性不同导致句子所表达的意思有差别. 词性序列如表 1 所示.

表 1 句子关键词词性

原句与相似句	词性序列
原句: 退役士兵招聘有什么需要	/v/n/v/v/r/n
相似句: 退役士兵需要有什么招聘	/v/n/v/v/r/v

本文将使用哈尔滨工业大学开发的在线语言技术平台 (language technology platform, LTP)[10] 进行分词并得到所需的句子属性信息. LTP 处理后的结果如图 2.

今年	的	退役	大学生	士兵	有	什么	政策
nt	u	v	n	n	v	r	n

图 2 分词和词性标注的结果

2.1 词频相似度计算方法

改进基于向量空间模型 TF-IDF (term frequency-inverted document frequency) 的词频计算方法[11]. TF-IDF 简单结构没有考虑词语的语义信息, 无法处理一词多义与一义多词的情况.

采用词语逆频率方式计算加权算法 TF-IWF (term frequency-inverse word frequency), 将句子S_1和S_2各自的词频向量映射到向量空间中, $S_1=(f_1,f_2,\cdots,f_n)$, $S_2=\left(f_1',f_2',\cdots,f_n'\right)$, 其中$f_i=tf_{w_i}\times iwf_{w_i}$, f_i为关键词w_i的词频-逆词频率, tf_{w_i}为关键字w_i在文本中出现的频率, 即 TF 值, 表示关键词w_i出现的次数与所有词汇量的比值, iwf_{w_i}为逆词频率, 即 IWF 值, 表示为所有词语的频数之和与关键词w_i出现的频数和的比值. 结合向量余弦相似度方法[12]:

$$WorSim(S_1,S_2)=\frac{\sum_{i=1}^{n}f_i\times f_i'}{\sqrt{\sum_{i=1}^{n}f_i^2\times\sum_{i=1}^{n}f_i^2}} \tag{2}$$

2.2 词序相似度计算方法

句子中共有关键词需考虑其相对位置关系, 词序相似度是共有关键词在两个句子中的位置相似度, 词位置顺序不同导致句子意思不同. 句子S_1="青岛籍退役士兵在北京服役政策", 句子S_2="北京籍退役士兵在青岛服役政策". 经词性和词频相似度计算, S_1和S_2相似度是 100%, 但实际意义差别较大, 采用逆序数与向量距离相似度度量方法[13,14] 融合计算词序相似度.

举例说明, S_1的中心词={'青岛', '退役士兵', '北京', '服役', '政策'}; S_2的中心词={'北京', '退役士兵', '青岛', '服役', '政策'}.

以句子S_1的序列为标准序列 (1, 2, 3, 4, 5).

首先以两个句子所含相同关键词的逆序数作为衡量因素, S_2的序列为 (3, 2, 1, 4, 5), $Ron(S_1,S_2,s)$代表句子S_2中词汇的逆序数, S 为相同关键词个数, 则采用逆序数衡量词序相似度公式为:

$$RevOrdSim(S_1,S_2)=\begin{cases}1-\dfrac{Ron(S_1,S_2,s)}{s-1}, & s>1\\ 0, & s=0\\ \dfrac{1}{2}, & s\leqslant 1\end{cases} \tag{3}$$

得出句子S_1和S_2的逆序数词序相似度为:

$$RevOrdSim(S_1,S_2)=\frac{1}{4}$$

然后以两个句子所含相同关键词的向量距离[14] 作为衡量因素, $distance(S_1,S_2)$代表句子S_1到S_2的向量距离, $maxDistance(S_1,S_2)$为$distance(S_1,S_2)$的最大值, 其计算公式为:

$$VecOrdSim(S_1,S_2)\begin{cases}1-\dfrac{distance(S_1,S_2)}{maxDistance}, & s>1\\ 0, & s=0\\ \dfrac{1}{2}, & s\leqslant 1\end{cases} \tag{4}$$

得出句子S_1和S_2的向量距离词序相似度为:

$$VecOrdSim(S_1,S_2)=\frac{2}{3}$$

逆序数$Rev(S_1,S_2)$与向量距方法$Vec(S_1,S_2)$融合计算词序相似度为:

$$OrdSim(S_1,S_2)=\frac{Rev(S_1,S_2)+Vec(S_1,S_2)}{2}=\frac{11}{24} \tag{5}$$

2.3 词性相似度计算方法

词性相似度 (nature similarity) 指两个句子中共有关键词的词性相似度, 此相似度计算方法用来完善一词多义的情况[15], 定义为具有相同词性的共有关键词数与两个句子总关键词数和之比. 计算公式为:

$$NatSim(S_1,S_2)=\frac{2\times Ncs(S_1,S_2)}{Com(S_1)+Com(S_2)} \tag{6}$$

使用 LTP 分词后并将得到的关键词词性进行比较, 式子中$Ncs(S_1,S_2)$是句子S_1和S_2相同词性的共有关键词数, $Com(S_1)$和$Com(S_2)$即句子S_1和S_2的总关键词数. 显而易见的, 如果得到的两个句子词性相同的关键词数越多, 那么两个句子词性相似度越高.

2.4 句长相似度计算方法

以词频、词序、词性为核心要素计算相似度时, 而要完整、准确的反映句子的信息也要考虑句长的存在.

两个句子长度的差的绝对值, 可以反映一定程度上的句子相似度, 其与绝对值的大小成反比, 值越小, 说明此种程度上的相似度越大. 假设句子S_1长度表示

为$Len(S_1)$, S_2长度表示为$Len(S_2)$, 句长相似度表示为$LenSim(S_1,S_2)$, 则其计算公式如下:

$$LenSim(S_1,S_2)=1-\frac{abs(Len(S_1)-Len(S_2))}{Len(S_1)+Len(S_2)} \tag{7}$$

式中, $abs()$ 为绝对值函数.

2.5 句子的多属性融合相似度计算方法

综合词频、词序、词性、句长 4 种属性, 融合 4 种相似度, 其计算流程如图 3 所示.

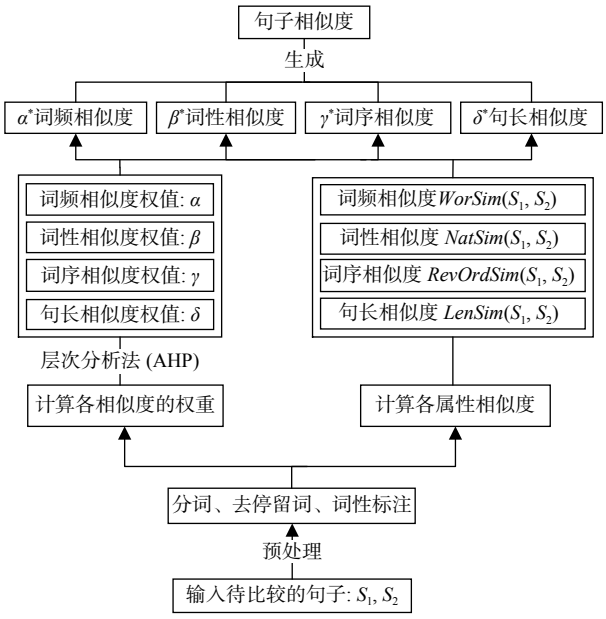

图 3　方法计算流程

首先输入句子S_1, S_2, 经过分词处理、去停留词、词性标注等预处理操作, 最终得出句子相似度公式为:

$$\begin{aligned}&StrSim(S_1,S_2)=\\&\alpha*WorSim(S_1,S_2)+\beta*OrdSim(S_1,S_2)\\&+\gamma*NatSim(S_1,S_2)+\delta*LenSim(S_1,S_2)\end{aligned} \tag{8}$$

式中, α、β、γ、δ分别是词频、词序、词性和句长相似度的权重值, 其中 $0\leqslant\alpha\leqslant1$, $0\leqslant\beta\leqslant1$, $0\leqslant\gamma\leqslant1$, $0\leqslant\delta\leqslant1$, 且满足$|\alpha+\beta+\gamma+\delta|=1$.

本文采用层次分析法 (AHP)[16] 通常被用到处理复杂的决策问题, 准备采取决策的问题分成 3 个层次, 基于该方法计算出的各相似度的权重. 步骤如下:

(1) 将问题条理化、层次化, 根据词频、词序、词性和句长相似度建立层次结构模型.

(2) 根据经验赋予各相似度重要程度并构造判断矩阵, 词频与词序相似度重要程度高且一致, 相较而言, 词性和句长相似度重要程度低且一致, 并使用 1~9 及其倒数作为标度来确定a_{ij}的值, 如表 2 所示.

显然, 表 2 中的元素满足:

$$a_{ij}>0;\ a_{ij}=\frac{1}{a_{ij}};\ a_{ij}=1$$

根据层次分析法规则可得比较矩阵, 如表 3 所示.

表 2　重要程度

i比j强的重要程度	相等	稍强	强	很强	绝对强
a_{ij}	1	3	5	7	9

表 3　比较矩阵

相似度	词频相似度	词序相似度	词性相似度	句长相似度
词频相似度	1	1	5	5
词序相似度	1	1	5	5
词性相似度	1/5	1/5	1	1
句长相似度	1/5	1/5	1	1

(3) 层次单排序并进行一致性检验, 根据表 3 可得判断矩阵 A.

$$A=\begin{bmatrix}1&1&5&5\\1&1&5&5\\\frac{1}{5}&\frac{1}{5}&1&1\\\frac{1}{5}&\frac{1}{5}&1&1\end{bmatrix}$$

计算该判断矩阵的最大特征值$\lambda_{\max}=3$, 其对应的特征向量为 [0.6934, 0.6934, 0.1387, 0.1387], 求出一致性指标 CI (consistency index):

$$CI=\frac{\lambda_{\max}-n}{n-1} \tag{9}$$

其中, n 为矩阵的维度, 得出 CI=0, 表示完全一致. CI 的值越小, 表示越一致, CI 的值越大, 表示越不一致.

(4) 使用 Satty 模拟 1 000 次得到的 RI 表计算一致性比率, RI 表如表 4 所示.

表 4　RI 表

阶数	3	4	5	6	7	8	9
RI	0.58	0.90	1.12	1.24	1.32	1.41	1.45

$CR=\frac{CI}{RI}$, 判断当$CR<0.1$时, 认为判断矩阵的一致性满意, 此时权值可用判断矩阵的特征向量; 若$CR\geqslant0.1$, 认为判断矩阵的一致性不合适, 应考虑修正判断矩阵使一致性比率$CR<0.1$. 计算得出 CR 值为 0.

根据$|\alpha+\beta+\gamma+\delta|=1$, 得出基于词频和词序属性的相似度权值为 0.417, 基于词性和句长属性的相似度权

值为 0.083.

3 实验及分析

本文算法实验中, 开发环境为 Windows 10 X64, 开发工具为 VSCode1.54.1, 开发语言为 Python 3.6.4. 采用哈尔滨工业大学开发的在线语言技术平台 (LTP) 进行关键词分词并得出所需的句子的属性.

为验证方法的效果, 对本文提出的基于句子的多属性融合相似度计算方法和 Jarcarrd、文献 [14] 方法, 设计对比试验, 以召回率 (Recall)、准确率 (Precision)、F-度量值 (*F*-Measure)[17] 对比不同算法的性能, F 度量值综合涵盖召回率与准确率两个指标, 值越接近于 100%, 说明准确率和召回率越均衡, 方法的效果越好, 相反, 如果 F 度量值越接近于 0, 说明两个指标的均衡性越差, 方法效果欠缺.

(1) 召回率 (Recall) 衡量相似度匹配的查全率.

$$召回率 = \frac{正确检测到的相似句子数}{实际存在的相似句子数} \tag{10}$$

(2) 准确率 (Precision) 衡量相似度匹配的查准率.

$$准确率 = \frac{正确检测到的相似句子数}{所有检测到的相似句子数} \tag{11}$$

(3) *F*-度量值 (*F*-Measure) 是召回率与准确率的指标归一化平均值, 用于反映整体的指标.

$$F-度量值 = 2 \times \frac{准确率 \times 召回率}{准确率 + 召回率} \tag{12}$$

实验步骤如下:

实验所需数据为随机从国家与地方退役军人事务局等网站爬取的, 经过数据处理建立的问答库, 从中选取 300 条问答对作为初始数据集S. 随机从S中选取 50 条作为初始标准集, 余下 250 条作为初始噪声集, w 为标准集的问句, $w \in S$. 依次使用w作为百度知道的查询条件, 利用 Python 正则表达式对查询返回的网页进行标签分析处理, 提取出网页中前 5 个标题, 问句w会有 1–5 个相似问句, 人工处理筛选出标题和问句相似度高的句子, 得到完善好的包含 223 个元素的扩充标准集, 将扩充标准集和初始标准集混合成为 273 个元素的标准测试集, 相同的将初始噪声集处理得到包含 1 190 个元素的扩充噪声集, 并和初始噪声集混合得到 1 440 个元素的噪声测试集, 最后将二者混合得到测试集. 依次从标准测试集的 273 个句子中抽出一个问句 P, 然后将其与测试集的问句的相似度计算出进行逆序排列, 如果前 5 个句子包括了扩充标准集中问句 P 所以对应的 1–5 个句子, 则表明句子相似度计算达到预期. 实验结果如表 5 所示.

表 5 句子相似度对比实验 (%)

实验	召回率	准确率	*F*-度量值
Jaccard方法	33.04	93.74	48.86
文献[14]方法	92.63	69.17	79.20
本文方法	85.24	89.09	87.12

Jaccard 方法和文献 [14] 方法与本文方法的实验结果从表 5 可以看出, Jaccard 方法具有较高准确率和较低召回率, 但该方法以句子的字面量为特征, 所以有一定限制在一词多义层面的相似度计算方面, 句子中包含的关键词相似, 但却忽略了关键词词性的不同. 文献 [14] 方法解决了一义多词问题, 比较而言, 其召回率比 Jaccard 方法高 64.3%, 准确率却下降了 26.2%, 显而易见, 两种方法均没有达到均衡稳定的效果. 本文方法相较于文献 [14] 方法准确率提高约 20%, 且 F-度量值更接近于 100%, 明显优于使用 Jaccard 方法和文献 [14] 方法.

4 结论与展望

本文提出的基于句子的多属性融合相似度计算方法, 综合考虑了句子的结构和包含的属性, 以词频、词序、词性和句长 4 种相似度加权融合计算, 对提高句子相似度计算的准确率有利, 且不会大范围出现召回率的降低, 其可靠性及可行性优于传统方法, 召回率、准确率不仅均衡且归一化 *F* 度量值较高, 达到 87.12%, 拥有综合优势. 接下来, 将该方法应用于智能问答系统的句子匹配, 可适用普遍存在的句子语法情况, 进一步研究时, 将继续优化此方法的复杂度及问答效率.

参考文献

1 李慧. 词语相似度算法研究综述. 现代情报, 2015, 35(4): 172–177.

2 Tian HN, Guo X. Research on improved sentence similarity calculation method based on Word2Vec and synonym table in interactive machine translation. 2021 5th International Conference on Robotics and Automation Sciences (ICRAS). Wuhan: IEEE, 2021. 255–261.

3 Wilson PK, Jeba JR. An efficient methodology for measuring sentence similarity using combinational semantics. 2021 7th

International Conference on Advanced Computing and Communication Systems (ICACCS). Coimbatore: IEEE, 2021. 1872–1876.
4 Goz F, Kabasakal O, Mutlu A. Experimental analysis of keyword-based social network similarity approach for document classification. 2020 28th Signal Processing and Communications Applications Conference (SIU). Gaziantep: IEEE, 2020. 1–4.
5 Ruan HP, Li Y, Wang QL, *et al*. A research on sentence similarity for question answering system based on multi-feature fusion. 2016 IEEE/WIC/ACM International Conference on Web Intelligence. Omaha: IEEE, 2017. 507–510.
6 翟社平, 李兆兆, 段宏宇, 等. 多特征融合的句子语义相似度计算方法. 计算机工程与设计, 2019, 40(10): 2867–2873, 2884.
7 吴全娥, 熊海灵. 一种综合多特征的句子相似度计算方法. 计算机系统应用, 2010, 19(11): 110–114.
8 王寒茹, 张仰森. 文本相似度计算研究进展综述. 北京信息科技大学学报, 2019, 34(1): 68–74.
9 翟永杰, 吴童桐. 基于语义空间信息映射加强的零样本学习方法. 计算机应用与软件, 2020, 37(12): 113–118, 196.
10 郎君, 刘挺, 张会鹏, 等. LTP: 语言技术平台. 中国中文信息学会. 第三届学生计算语言学研讨会论文集. 沈阳: 中国中文信息学会, 沈阳航空工业学院, 2006. 64–68.
11 Pang SC, Yao JM, Liu T, *et al*. A text similarity measurement based on semantic fingerprint of characteristic phrases. Chinese Journal of Electronics, 2020, 29(2): 233–241.
12 张俊飞. 改进 TF-IDF 结合余弦定理计算中文语句相似度. 现代计算机, 2017, (32): 20–23, 27.
13 Abuobieda A, Osman AH. An adaptive normalized Google distance similarity measure for extractive text summarization. 2020 2nd International Conference on Computer and Information Sciences (ICCIS). 2020. 1–4.
14 周艳平, 李金鹏, 蔡素. 基于同义词词林的句子语义相似度方法及其在问答系统中的应用. 计算机应用与软件, 2019, 36(8): 65–68, 81.
15 吴浩, 艾山·吾买尔, 卡哈尔江·阿比的热西提, 等. 融合词性特征的中文句子相似度计算方法. 计算机工程与设计, 2020, 41(1): 150–155.
16 刘万里, 刘卫锋, 常娟. AHP 中互反判断矩阵的区间权重确定方法. 统计与决策, 2021, 37(6): 33–37.
17 刘辉. 基于强类别特征的文本相似度计算及其性能评估. 软件工程, 2020, 23(10): 5–7, 4.