2020 年 第 20 卷 第 17 期
2020，20(17)：06950-06

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

引用格式：刘继明，于敏敏，袁 野. 基于句向量的文本相似度计算方法[J]. 科学技术与工程，2020，20(17)：6950-6955
Liu Jiming，Yu Minmin，Yuan Ye. Computing method of text similarity based on sentence vector[J]. Science Technology and Engineering，2020，20(17)：6950-6955

基于句向量的文本相似度计算方法

刘继明，于敏敏，袁 野
(重庆邮电大学经济管理学院电子商务与现代物流重点实验室，重庆 400065)

摘 要 为进一步提高文本相似度计算的准确性，提出基于句向量的文本相似函数(part of speech and order smooth inverse frequency，PO-SIF)，从词性和词序方面优化了平滑反频率(smooth inverse frequency，SIF)计算方法，SIF 算法的核心是通过加权和去除噪声得到句向量来计算句子相似度。在具体计算时，一方面通过增加词性消减因子调节 SIF 句向量计算权重参数，获得带有词性信息的句向量，另一方面通过将词序相似度与 SIF 句向量相似度算法进行线性加权优化句子相似度得分。实验结果表明，增加词性和词序的方法可以提升算法准确率。

关键词 平滑逆频率；句向量；词性；词序相似度
中图法分类号 TP391.2； **文献标志码** A

Computing Method of Text Similarity Based on Sentence Vector

LIU Ji-ming，YU Min-min，YUAN Ye
(The Key Lab of E-commerce and Modern Logistics，School of Economics and Management，Chongqing University of Posts and Telecommunications，Chongqing 400065，China)

[Abstract] In order to improve the accuracy of text similarity calculation，a text similarity function part of speech and order smooth inverse frequency (PO-SIF) was introduced based on sentence vector. The smooth inverse frequency (SIF) calculation method was optimized from aspects ofpart of speech and word order. The core of SIF algorithm was to get sentence vectors by weighting and removing noise to calculate sentence similarity. On one hand，the weight parameters of SIF sentence vectors were adjusted by adding part of speech subtraction factor to obtain sentence vectors with part of speech information. On the other hand，the similarity scores of sentences were optimized by linear weighting based on word order similarity and SIF sentence vector similarity algorithm. The results show that the method of adding part of speech and word order can improve the accuracy of the algorithm.

[Key words] smoothing inverse frequency；sentence vector；part of speech；word order similarity

为了提高问答系统的性能，对于用户输入的常问问题，通常采用面向常见问题集 FAQ(frequently asked questions)库的问答策略。基于 FAQ 库的限定域自动问答系统由于更具实用性而成为自然语言处理领域的研究热点，而问题之间的相似度计算是 FAQ 中最关键的技术。问答系统中的句子相似度是指两个问题的匹配符合程度，相似度用 0～1 之间的数值表示，数值越大两个句子越相似[1]。

目前常见的计算句子相似度的方法有基于统计的、基于语义的、基于句法结构的、基于编辑距离的以及词向量的方法[2]。基于统计的方法以词频-逆文档频率(term frequency-inverse document frequency，TF-IDF)算法及其改进算法为主，根据词频衡量该词在句子中的重要程度并以此赋予权重，通过计算向量距离表示句子之间相似度。该方法主要从词的表面特征进行匹配，计算复杂度低，速度快，但缺点是没有考虑词的语义本身，无法识别同义词或近义词[3]。基于语义信息的方法通过《同义词词林》《知网》、WordNet 等语义资源对通用词汇进行扩展，识别问句中的同义或近义信息，从而提高问句的匹配度。但该方法受限于知识源，不适用特定领域，且对人力成本和语言学知识要求高。基于句法结构的方法通过分析一个句子中的各个成分的依存关系得到有效配对，再计算有效配对之间的相似度得到句子之间相似度[4]。该方法引入句法特征，体现了句子内部的结构，但没有考虑语义在句子中的作用，现有的依存句法分析准确率并不高。基于编辑距离的方法以词串之间变化所需要的最少的编辑操作次数作为变量，衡量句子之间的相似度。虽然考虑了词语顺序，但是这种关系变化太过机械，无法体现出词语在语句中不同位置对语义的影响。基于词向量(word embedding)的方法[5]

收稿日期：2019-09-07； 修订日期：2019-11-06
第一作者：刘继明(1964—)，男，汉族，博士，教授。研究方向：人工智能。E-mail：liu.jiming@jcubing.com。

投稿网址：www.stae.com.cn

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在大规模语料库中训练词向量,然后将问句转化为由词向量构成的句向量计算问句的相似度。通过训练特定领域的语料获得的词向量模型,能有效克服词典在特定领域作近义词查询时准确度不高的问题。

基于词向量的方法充分考虑语义特征并且解决了维度灾难问题,具有良好的性能。基于神经网络的 word2vec 在大规模语料条件下训练出的词向量包含了词本身的含义以及词与其他词之间的联系,是词向量计算的有力工具。在进行句子相似度计算时,往往需要将词向量转化为句向量进行相似度比较,平滑反频率(smooth inverse frequency,SIF)算法能够将词向量通过加权算法转化为句向量,其核心思想是通过加权和去除噪声得到句向量来计算句子相似度,相对于简单平均加权和 TF-IDF 加权具有绝对优越性[4]。然而 SIF 算法主要从语义和词频角度进行加权计算,未考虑词性与词序因素,因此,提出基于 word2vec 与 SIF 改进的文本相似度计算方法,提高相似度计算的准确度。

1　相关研究

1.1　词向量表示

传统的词表示方式为独热表示(one-hot representation),它将语料中所有非重复词语数量作为向量的维数,每一个词对应该向量的一个维度,对应位置的值为 1,其余位全为 0。假设词表中有 n 个词语,"集装箱"在词表中处于第 1 个,那么"集装箱"词向量表示为 [1 0 0 0…],"码头"在此表中处于第 2 个,其词向量表示为 [0 1 0 0 0…],即词表中的 n 个词都有唯一的离散向量表示。One-Hot 编码方法简单,但缺点是容易引起维数灾难,且词向量孤立,无法判别语义关系。1986 年 Hinton[7] 基于 Harris 的分布假说(distributional hypothesis)提出词的分布式表示(distributional representation),能够刻画语义之间的相似度并且用维度较低的稠密向量表示词语,即具有相似上下文的词,应该具有相似的语义。例如"铁路""公路"和"蛋糕"3 个词中"铁路"和"公路"相关性较大,这两个词对应的词向量距离就小。另外分布式表示通过使用较低维度的特征刻画词语,大大降低了计算的复杂性。

2013 年 Mikolov 等[8] 开发的 Word2vec 作为深度学习模型中的一种分布式表达(distributed representation),能够从大规模未经标注的语料中高效地生成词的向量形式,且可以较好地适用于中文处理。其基本思想是通过训练将每个词映射成高维实数向量,通过词之间的距离判断它们之间的语义相似度。word2vec 包含两种语言模型:CBOW(continuous bag-of-words model)和 Skip-Gram(continuous skip-gram model),如图 1 所示。CBOW 和 Skip-gram 模型都包括输入、投影和输出 3 层,其中输入与输出层表示词向量。不同的是,CBOW 的目标是根据上下文来预测当前词语的概率,Skip-gram 则是根据当前词语来预测上下文的概率。CBOW 模型能够获取更好的语法信息,因此主要采用 word2vec 的 CBOW 模型训练词向量。

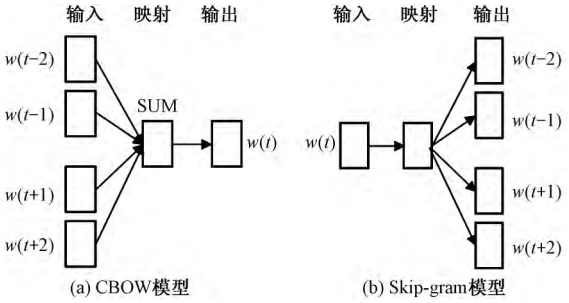

图 1　CBOW 模型与 Skip-gram 模型

Fig. 1　CBOW model & Skip-gram model

1.2　句向量表示

关于词向量(word embedding)的研究相对较早,目前分布式表示方式已经相对成熟并得到广泛应用[9]。但从词向量向句向量(sentence embedding)乃至段落向量(paragraph embedding)等的扩展却仍处于探索阶段。词向量可以通过均值模型或加权模型转化为句向量。其中,最简单直接的方法是将句子中所有词向量的平均值作为句向量,但这种方法的缺点是忽略了词频对句子的影响,认为句子中的所有词对于表达句子含义同样重要。加权模型考虑到词频的影响,其效果要明显优于均值模型。TF-IDF 加权平均词向量以 TF-IDF 为权重,对所有词向量加权得到句向量[10];SIF 改进了 TF-IDF 加权平均词向量方法,它以 SIF 为权重对所有词的词向量加权平均,最后从中减掉第一主成分(principal component)的投影得到句向量[4]。

SIF 模型以简洁的思想和优异的性能成为句向量领域最新的强大基线。其句向量表示为

$$\boldsymbol{v}_S \leftarrow \frac{1}{|S|}\sum_{w\in S}\frac{a}{a+p(w)}\boldsymbol{v}_w \tag{1}$$

式(1)中:$\boldsymbol{v}_S$ 表示去除主成分前的句向量;S 表示句子中词的总数;$\boldsymbol{v}_w$ 表示词向量;$a/[a+p(w)]$ 表示句子 S 中词 w 的权重,其中 a 为平滑系数,取 0.001;$p(w)$ 为(估计的)词频。式(1)表明频率越低的词权重越大,其在句子中的重要性也越大。

$$\boldsymbol{v}_S \leftarrow \boldsymbol{v}_S - \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}_S \tag{2}$$

式(2)中最终的句子 S 的句向量 $\boldsymbol{v}_S$ 为式(1)中的句

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

向量减去句子中的共有信息(主成分)。其中 $uu^{\mathrm{T}}v_S$ 表示 v_S 的最大主成分向量; u 为去"所有构成 v_s 的矩阵"通过奇异值分解的特征矩阵; u^{T} 为 u 弘的转置。即减去所有句子的共有信息,使得句子确保留下来的句子向量更能够表示本身意思。

1.3 相似度算法

两个句子的相似度可用欧式距离、余弦距离、杰卡德系数等公式计算[11],其中余弦距离应用较为广泛。余弦相似度指利用向量空间中两个向量夹角的余弦值表示两个个体间差异的大小程度,余弦值在 0~1 之间越接近 0,两个向量越相似。

2 算法设计

目前,对 SIF 算法计算出的句向量进行相似度计算能够准确地匹配到 FAQ 库中相似问句,取得较好的效果。SIF 算法兼顾词频与语义信息,但缺点是未考虑词性与词序因素。因此,综合考虑词性、词序、句法结构、词频、语义信息,从词性和词序方面提出基于句向量的句子相似度计算方法。具体步骤如下。

步骤 1: 将预处理后的文本 s_n 和 s_m 采用词集合 s_n' 和 s_m' 表示。

步骤 2: 采用 word2vec 获得每个词的词向量,采用词性消减因子为每个词赋词性调节权重因子 w_{p}; 最终权重为 w_{p} 与 SIF 原始权重之积。

步骤 3: 利用调节后的权重将每个文本的词向量相加,获得句向量。

步骤 4: 利用余弦距离公式计算用户输入问句与 FAQ 问题库中每个问句的相似度并获取前 5% 相似句集合 T。

步骤 5: 采用词序的方法计算 T 中的句子与测试句子的词序相似度 $\mathrm{sim}_{\mathrm{o}}$ 并与 SIF 方法进行线性加权,得到最终相似度得分。基于 SIF 问句相似度算法的基本流如图 2 所示

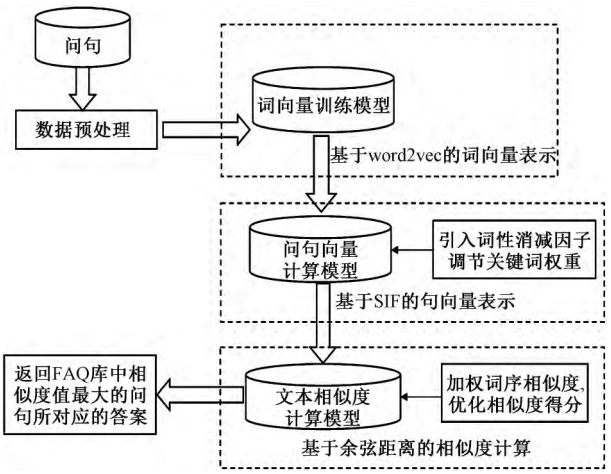

图 2 基于 SIF 问句相似度算法的基本流程

Fig. 2 Basic flow based on SIF question similarity algorithm

2.1 引入词性消减因子调节关键词权重系数

除词频外,词性也是判断词语重要性的重要指标。汉语词语中实词有 6 类: 动词、名词、形容词、量词、数词和代词,虚词也有 6 类: 介词、副词、助词、叹词、语气词和连词。一般情况下,动词、名词和形容词是句子的核心词,应该给予更高的权重,其中动词在句子中的作用尤为重要。同时,查询的意图往往依赖于动词的位置,动词越接近句子的末尾,越能代表整个句子的意思,其重要程度也相对越高[12],因此还应该考虑动词所处句子的位置。例如"我/可以/申请/退货/吗","申请"和"退货"都是动词,"退货"位于"申请"之后,更能反映整个句子的意思。为了表示词语位置与权重关系,按逆序法分析词语和句子。若句子中的词为动词、名词、形容词,则词性调节权重因子为

$$w_{\mathrm{p}} = \{\mathrm{score}_{\min}, \mathrm{score} - \mathrm{decay}\},\ 0 < w_{\mathrm{p}} \leqslant 1 \tag{3}$$

式(3)既考虑了不同词性的重要程度,也考虑了不同位置的词的重要性,即越靠近句尾的词语更能说明句子的意图。其中, w_{p} 为词性调节权重因子,其大小取决于词性及词在句中的位置; $\mathrm{score}_{\min}$ 表示每类词给定的下限值,它保证了所有词性的词的最低权重大于 0,不同词性的下限值设定为: 1 > 动词 > 名词 > 形容词 > 其他 > 0; decay 为词性消减权重,表示句子中的关键词随着单词接近句首而变小的幅度。以动词为例,"我/可以/申请/退货/吗",假定 decay = 0.2,那么按照逆序法,"退货"一词为第一个动词,假定其权重为 1,那么"申请"一词的词性权重为 1 - 0.2 = 0.8。句向量的权重系数 $aw_{\mathrm{p}}/[a + p(w)]$,句向量表示为

$$v_S \leftarrow \frac{1}{|S|}\sum_{w \in S}\frac{aw_{\mathrm{p}}}{a + p(w)}v_w \tag{4}$$

2.2 融合词序相似度优化相似度值

人类很容易处理词序信息,然而对于计算机来说,在理解自然语言的计算方法中加入词序信息是一个很大的挑战。词序是指词在词组或句子里的先后次序,它包含了句子所传达的语义和语法信息。在汉语中,即使包含完全相同词语的两个句子,若词语位置发生微调,他们所表达的意思也会发生改变。以下面两句为例。

S_1: 那些北京来的客人很热情。

S_2: 那些客人来北京很热情。

S_1 和 S_2 所包含的词相同,但由于词序存在差

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

异,两句话所表达的意思明显不同。由于 SIF 加权算法仅仅将词频作为句向量加权的依据,不考虑句子中词语顺序信息,因此通过 SIF 加权后得到的 S_1 和 S_2 的句向量完全相同。这种情况下,计算两个句子之间相似度时考虑词序问题会使结果更加精确。

关于词序相似度,一种是通过找出两个句子中既是相同词语又在相同位置的词语来计算词序相似度[13]。另一种使用较多的是基于向量距离的词序算法[14]。基于这两种计算词序的思想,提出一种新的词序相似度算法,其核心是 S_2 中与 S_1 重叠的词的个数越多并且 S_2 达到 S_1 顺序所需交换的次数越少,则两句话的词序相似度越大。公式表示为

$$\mathrm{sim}_o(S_1,S_2)=2\frac{\mathrm{common}(S_1,S_2)}{\mathrm{len}(S_1)}\frac{b}{b+T}$$

$$0\leqslant \mathrm{sim}_o(S_1,S_2)\leqslant 1 \tag{5}$$

式(5)中:$\mathrm{common}(S_1,S_2)$ 表示 S_2 中与 S_1 重叠的词的个数;$\mathrm{len}(S_1)$ 表示 S_1 中词语的个数;T 表示 S_2 中与 S_1 重叠词语按照 S_1 顺序排列所需要的交换次数,$\frac{b}{b+T}$ 确保当 T 为 0 时,此项仍有意义。具体做法如下。

首先,写出待比较句子的特征顺序向量。将句子 $S_1=(w_1,w_2,w_3,\cdots,w_n)$ 的向量特征项顺序作为标准顺序,其对应的特征顺序向量为 $v_1=(v_1,v_2,v_3,\cdots,v_n)=(1,2,3,\cdots,n)$,那么 $S_2=(w_1,w_2,w_3,\cdots,w_m)$ 所对应的顺序特征向量 v_2 是向量空间 S_2 按照特征项在 S_1 中生成的,没有相同特征的位置用 0 表示[15]。以上述例句为例,将文本 S_1 的向量特征顺序作为标准顺序 $S_1=$(那些,北京,来,的,客人,很,热情),其对应的特征顺序向量 $v_1=(1,2,3,4,5,6,7)$,那么 $S_2=$(那些,客人,来,北京,很,热情)对应的特征顺序向量 $v_2=(1,5,3,2,6,7)$。

然后,对 S_2 中不是正常顺序的索引进行惩罚。当 S_2 中第 n 个索引比第 $n-1$ 个索引值大时,表示该词在 S_2 中出现的顺序与 S_1 中保持一致,v_2 转化为标准特征顺序向量只需要 0 次交换,语序相似度 $\frac{b}{b+T}=1$。当 S_2 中第 n 个索引比第 $n-1$ 个索引值小时,对词序因素进行惩罚,T 越大,词序相似度越小。

3　实验

3.1　实验数据及性能指标

3.1.1　实验数据

词向量训练常用质量较好的中文语料库为维基百科的中文语料库,它具有质量高、领域广泛而且开放的优点,因此本文采用 2017 年 10 月发布的维基百科中文数据作为 word2vec 的训练集。过滤掉标点符号和其他无关符号等数据清洗之后,通过张华平等[16]研究的中文分词工具 NLPR 汉语分词系统(又名 ICTCLAS)进行分词,形成提供学习训练的文本数据集,然后使用基于 python 语言的自然语言处理库 Gensim 中 word2vec 模块,采用 CBOW 模型,维度为 200、窗口大小为 10 等参数对该文本数据集进行学习训练得到词向量模型文件。

FAQ 库语料来源由 3 部分组成共 1 200 个问答对。①厦门科技局提供的真实问答语料。②根据科技政策有关的等相关办事指南及法律条文人工编写的问答对。③网络爬取相关语料。随机抽取 300 条语料进行人工改写相似问句用于测试。

3.1.2　评价指标

在智能问答系统方面,中国研究起步相对较晚,评价标准相对也不是很完善。召回率(recall)、准确率(precision)和 F 测度值是检测问答系统中常用的 3 个参数[17]。召回率是匹配到的相关问题数目与 FAQ 库中所有相关问题数目的比率,衡量问答系统的查全率;准确率是匹配到的相关问题数目与匹配到的问题总数的比率,衡量问答系统的查准率;F 测度值是把召回率和准确率的函数两个参数结合起来的综合评价指标。

从 FAQ 数据集中进行问句匹配时,把文档分成 4 组:A 组为系统匹配到的相关问句;B 组为系统匹配到的不相关问句;C 组为相关但系统没有匹配到的问句;D 组为不相关且系统未匹配到的问句。召回率与准确率如图 3 所示。

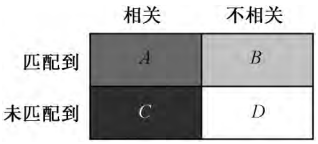

图 3　召回率与准确率

Fig. 3　Recall rate and precision rate diagram

$$\text{召回率 } R=\frac{\text{匹配到的相关问题数目}}{\text{FAQ 库中所有相关问题数目}}=\frac{A}{A+C};$$

$$\text{准确率 } P=\frac{\text{匹配到的相关问题数目}}{\text{匹配到的问题总数}}=\frac{A}{A+B};$$

$$F\text{ 测度值 }=\frac{2\times\text{召回率}\times\text{准确率}}{\text{召回率}+\text{准确率}}=\frac{2RP}{R+P}。$$

3.2　实验结果及分析

基于 word2vec 词向量和 SIF 句向量模型提出一种融合词性和词序的句子相似度算法,为验证方法的有效性,分别采用基于 SIF 的方法、结合词性消减因子调节 SIF 权重的方法、结合词序的方法以及融合词性和词序的方法对条测试问句进行相似度测试,测试问题个数为 300 个。实验结果如表 1 和图 4 所示。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表1 4种相似度算法的实验对比

Table 1 Experimental comparison of four similarity algorithms

相似度方法	准确率 P/%	召回率 R/%	平均调和值 F/%	平均用时/s
基于SIF方法	83.40	76.50	79.80	0.73
SIF+词性	87.30	79.10	83.00	0.79
SIF+词序	86.20	78.20	82.01	0.88
本文方法	87.9	79.80	83.70	0.9

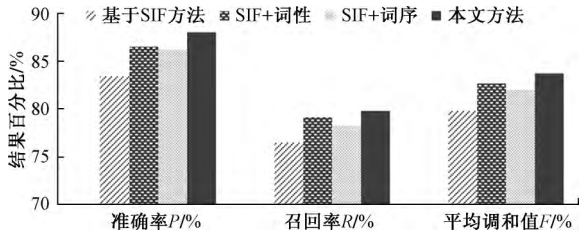

图4 4种句子相似度计算方法结果比较

Fig. 4 Comparison of results of four sentence similarity calculation methods

从以上试验结果可以看出,通过使用基于SIF的改进算法,其所得的结果在准确率、查全率和查准率方面均要高于基准SIF算法。同时,由于加入了词性和词序因素,算法的复杂性也相对提高,平均用时增加了0.13 s,在正常范围内。这说明不同词的词性对词的重要程度有一定影响,词序相似度在句子相似度计算中也发挥了作用,使计算结果更加准确。

4 结论

提出了一种基于word2vec词向量和SIF句向量模型的句子相似度计算方法,利用词性及词性消减因子调节SIF计算句子向量的参数,并使用词序相似度进行加权计算,使句子相似度的度量结果更为准确。在实验所用的语料中显示,本文方法相比其他传统的句子相似度算法能够在相似度测试中获得更好的效果,更加接近于人的相似排序。但算法对汉语中一些句法结构的考虑还有欠缺,下一步将结合中文句法结构对算法进行改进,继续提升句子相似度计算的准确率。

参 考 文 献

1 董自涛,包佃清,马小虎. 智能问答系统中问句相似度计算方法[J]. 武汉理工大学学报(信息与管理工程版),2010,32(1):31-34.
Dong Zitao, Bao Dianqing, Ma Xiaohu. Method for calculating similarity of question sentences in intelligent question answering system [J]. Journal of Wuhan University of Technology (Information and Management Engineering Edition), 2010, 32(1): 31-34.

2 赵 谦,荆 琪,李爱萍,等. 一种基于语义与句法结构的短文本相似度计算方法[J]. 计算机工程与科学,2018,40(7):1287-1294.
Zhao Qian, Jing Qi, Li Aiping, et al. A short text similarity calculation method based on semantic and syntactic structure [J]. Computer Engineering and Science, 2018, 40(7): 1287-1294.

3 黄承慧,印 鉴,侯 昉. 一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J]. 计算机学报,2011,34(5):856-864.
Huang Chenghui, Yin Jian, Hou Fang. A text similarity measurement method combining term semantic information and TF-IDF method [J]. Chinese Journal of Computers, 2011, 34(5): 856-864.

4 李 琳,李 辉. 一种基于概念向量空间的文本相似度计算方法[J]. 数据分析与知识发现,2018,2(5):48-58.
Li Lin, Li Hui. A method of text similarity calculation based on concept vector space [J]. Data Analysis and Knowledge Discovery, 2018, 2(5): 48-58.

5 高明霞,李经纬. 基于word2vec词模型的中文短文本分类方法[J]. 山东大学学报(工学版),2019,49(2):34-41.
Gao Mingxia, Li Jingwei. Chinese short text classification method based on word2vec word model [J]. Journal of Shandong University (Engineering Science Edition), 2019, 49(2): 34-41.

6 Arora S, Liang Y, Ma T. A simple but tough-to-beat baseline for sentence embeddings [C] // Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017: 1-16

7 Hinton G E. Learning distributed representations of concepts [C] // Proceedings of the Eighth Annual Conference of the Cognitive Science Society. Amherst: University of Pittsburgh, 1986: 1-12.

8 Mikolov T, Sutskever L, Chen K. et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

9 Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information [J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

10 段旭磊,张仰森,孙祎卓. 微博文本的句向量表示及相似度计算方法研究[J]. 计算机工程,2017,43(5):143-148.
Duan Xulei, Zhang Yangsen, Sun Yizhuo. Research on sentence vector representation and similarity calculation method of microblog text [J]. Computer Engineering, 2017, 43(5): 143-148.

11 王春柳,杨永辉,邓 霏,等. 文本相似度计算方法研究综述[J]. 情报科学,2019,37(3):158-168.
Wang Chunliu, Yang Yonghui, Deng Fei, et al. A review of text similarity calculation methods [J]. Information Science, 2019, 37(3): 158-168.

12 Bi Y, Deng K, Cheng J X. A keyword-based method for measuring sentence similarity [C] // Proceedings of the 2017 ACM on Web Science. New York: ACM, 2017: 379-380.

13 谭咏梅,王敏达,牛少彰. 使用有序词语移动距离特征进行中文文本蕴含识别[J]. 北京邮电大学学报,2017,40(5):123-128.
Tan Yongmei, Wang Minda, Niu Shaozhang. Chinese text implication identification using ordered word moving distance feature [J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40(5): 123-128.

14 程志强,闵华松. 一种基于向量词序的句子相似度算法研究[J]. 计算机仿真,2014,31(7):419-424.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Cheng Zhiqiang, Min Huasong. Research on sentence similarity algorithm based on vector word order [J]. Computer Simulation, 2014, 31(7): 419-424.

15 李 峰,侯加英,曾荣仁,等. 融合词向量的多特征句子相似度计算方法研究[J]. 计算机科学与探索,2017,11(4):608-618.
Li Feng, Hou Jiaying, Zeng Rongren, et al. Research on multi-character sentence similarity calculation method of fusion word vector [J]. Journal of Computer Science and Technology, 2017, 11(4): 608-618.

16 张华平,刘 群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报,2004,27(1):85-91.
Zhang Huaping, Liu Qun. Research on automatic recognition of Chinese names based on role tagging [J]. Journal of Computer Science, 2004, 27(1): 85-91.

17 杨 晨,张 鹏. 基于词向量相似度的食品安全问答系统设计与实现[J]. 软件导刊,2019,18(8):16-20.
Yang Chen, Zhang Peng. Design and implementation of food safety question answering system based on word vector similarity [J]. Software Guide, 2019, 18(8): 16-20.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net