·人工智能及识别技术· 文章编号：1000-3428(2017)09-0234-06 文献标志码：A 中图分类号：TP391

基于修正偏移量的句子相似度算法

裴飞龙，闵华松

（武汉科技大学 冶金自动化与检测技术教育部工程研究中心，武汉 430081）

摘 要： 在计算词语语义相似度的基础上，提出一种以句子中心词为基准衡量词语组合相对位置偏移量的句子相似度计算方法。根据词语词性和语法规则确定句子中心词并剔除全局性限定词，以句子中心词为基准对词语相对位置进行标定并计算词语组合的相对位置偏移量，综合句子长度差异性信息、浅层次结构信息和语义信息计算句子相似度。实验结果表明，与基于统计与基于偏移量的句子相似度算法相比，该算法具有较高的相似度计算准确率、召回率及综合指标 F 值。

关键词： 句子相似度；中心词；相对位置；修正偏移量；全局性限定词

中文引用格式： 裴飞龙，闵华松. 基于修正偏移量的句子相似度算法[J]. 计算机工程，2017，43(9)：234-239.

英文引用格式： PEI Feilong，MIN Huasong. Sentence Similarity Algorithm Based on Corrected Offset[J]. Computer Engineering，2017，43(9)：234-239.

Sentence Similarity Algorithm Based on Corrected Offset

PEI Feilong，MIN Huasong

（Engineering Research Center of Metallurgical Automation and Measurement Technology of Ministry of Education，Wuhan University of Science and Technology，Wuhan 430081，China）

【Abstract】 On the basis of calculating the semantic similarity of words，a sentence similarity calculation method is proposed to measure the relative position offset of word combination according to the sentence center word. Firstly，central word of the sentence is determined by part of speech and grammatical rules，and the global qualifiers of the sentence are eliminated. Secondly，the relative position of words is marked according to the center word，and the relative position offset of the combination of two words is computed. At last，the sentence length difference information，shallow structure information and semantic information are integrated to calculate sentence similarity. The experimental results show that the algorithm can effectively improve the accury，recall rate，and F value of the comprehensive index compared with the sentence similarity algorithm based on statistics and the one based on offset.

【Key words】 sentence similarity；center word；relative position；corrected offset；global qualifier

DOI： 10.3969/j.issn.1000-3428.2017.09.041

0 概述

句子相似度计算在自然语言处理领域是一项应用非常广泛的技术。在信息检索过程中，句子相似度反映的是文本与用户查询在意义上的符合程度，从相关的文档中检索出与用户查询相似度高的文档。相似度计算的准确度影响着检索出的信息是否符合用户的需求，并且在机器翻译、文本聚类、自动摘要以及自动问答系统中句子相似度的计算都具有重要作用。

由于句子具有数据稀疏性和时序性等特征，使得某些句子相似度算法的准确率并不高，难以区分句子信息的差异程度[1]。目前，常用的句子相似度算法有 3 类[2-4]：基于统计的相似度算法，基于语义的相似度算法和基于句法结构的相似度算法。然而，这 3 类算法都有各自的缺点。基于统计的相似度算法没有考虑文本的结构和语义信息，对于句子级的短文本，特征词重复出现的概率很低，会造成数据稀疏的现象。基于句法结构的相似度算法依赖句法分析器，目前句子各成分依存关系分析准确率不高，使得句子相似度计算准确率低。基于语义的相似度算法虽然挖掘了文本的蕴含信息，但没有考虑结构信息。

针对以上 3 类算法的缺点，国内外学者就多特征融合的句子相似度算法做了很多研究[5-9]。文献[10]提出一种结合词项语义信息的 TF-IDF 方法

基金项目： 国家自然科学基金(61175094，61673304)。

作者简介： 裴飞龙(1990—)，男，硕士研究生，主研方向为嵌入式系统及应用、智能机器人；闵华松，教授、博士生导师。

收稿日期： 2016-08-04 **修回日期：** 2016-09-05 **E-mail：** 576467670@qq.com

的文本相似度度量方法,在 TF-IDF 模型的基础上,借助词典分析词项之间的语义相似度进而起到降维作用。由于句子词语数据的稀疏性,因此使得句子相似度计算准确率不高。文献[11]对自然语言进行句法分析生成语义树,在此基础上衡量 2 个语义树的相似度,最终可对文本进行分类,但是实验依赖句法分析器。文献[12]提出一种基于空间向量法的改进句子相似度算法,其在空间向量法的基础上考虑了词序信息,在一定程度上提高了算法准确率,但没有考虑语义信息。文献[13]对字符串的相似度度量方法提出改进,通过计算向量空间模型中向量的汉明距离来描述字符串的相邻程度,以向量的曼哈顿距离作为衡量字符词序的指标。该方法可应用到句子相似度的计算中,但向量的生成过程以句首为基准可能会导致偏移量的累计增加。

考虑到目前句法分析器技术还不够成熟,多特征融合的方法复杂度高,句子存在数据稀疏的特点,本文提出一种以句子中心词为基准衡量词语组合相对位置偏移量的句子相似度计算方法。

1　基于《知网》的词语语义相似度计算

句子相似度计算依赖词语语义相似度。《知网》[14]是目前计算词语语义相似度广泛用到的词典。《知网》中一个词语被表达成各种各样的概念,而概念被描述成义原,因此,当计算 2 个词语之间的相似度时就需要计算义原相似度、概念相似度。

1.1　义原相似度计算

在《知网》中义原是不可分割的最小单位。文献[15]中介绍了 3 种义原相似度的计算方法并作出了比较:

1)义原层次体系中节点的路径长度为:

$$SimP(p_1,p_2)=\frac{\alpha}{dis(p_1,p_2)+\alpha} \tag{1}$$

其中,p_1 和 p_2 表示义原;$dis(p_1,p_2)$ 表示两义原 p_1 和 p_2 在义原层次中的路径长度;α 表示相似度为0.5时的路径长度。

2)在计算方法 1)的基础上考虑义原的深度:

$$SimP(p_1,p_2)=\frac{\alpha\times\min(depth(p_1),depth(p_2))}{dis(p_1,p_2)+\alpha\times\min(depth(p_1),depth(p_2))} \tag{2}$$

其中,$depth(p)$表示义原 p 在层次体系中的深度。

3)根据义原层次中两节点包含的共有信息有:

$$SimP(p_1,p_2)=\frac{2\times\log_a(S(p_3))}{\log_a(S(p_1))+\log_a(S(p_2))} \tag{3}$$

其中,p_3 表示义原 p_1 和 p_2 的最近公共节点;$S(p)$表示该节点的子节点个数与层次树中的总节点个数比。

根据文献[15]的分析对比可知,式(3)对义原相似度的计算更精确,因此,本文选取式(3)计算义原相似度。

1.2　概念相似度计算

义原相似度是概念相似度计算的基础。在汉语中,词语分为虚词和实词 2 类,实词一般有实际的意义,本文只考虑实词的概念相似度计算。根据《知网》对实词的描述,实词的概念相似度 $SimS(\boldsymbol{S}_1,\boldsymbol{S}_2)$是将 4 个义原相似度进行加权相加。4 种义原相似度为:第一独立义原相似度 $SimP_1(p_1,p_2)$,其他独立义原相似度 $SimP_2(p_1,p_2)$,关系义原相似度 $SimP_3(p_1,p_2)$以及符号义原相似度 $SimP_4(p_1,p_2)$。

$$SimS(\boldsymbol{S}_1,\boldsymbol{S}_2)=\sum_{i=1}^{4}\beta_i SimP_i(p_1,p_2) \tag{4}$$

其中,$\beta_1,\beta_2,\beta_3,\beta_4$ 应满足以下条件:$\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1>\beta_2>\beta_3>\beta_4$。

1.3　词语语义相似度计算

《知网》中一个词被表达成各种概念。假设词语 W_1 有 n 个概念 $S_{11},S_{12},\cdots,S_{1n}$,词语 W_2 有 m 个概念 $S_{21},S_{22},\cdots,S_{2m}$,本文将词语 W_1,W_2 相似度定义为两词语概念相似度的最大值:

$$Sim(W_1,W_2)=\max Sim_S(S_{1i},S_{2j}) \tag{5}$$

其中,$i=1,2,\cdots,n$;$j=1,2,\cdots,m$。

2　基于修正偏移量的句子相似度计算

词语语义相似度是句子相似度计算的基础,只考虑词语语义相似度会使得句子相似度的计算不精确,因此,加入词语组合相对位置偏移量的计算可提高句子相似度计算的准确率[16]。先选取相似度最大的词语组合,再考虑词语组合中以句子中心词为基准的 2 个词语相对位置偏移量,相对位置偏移量计算的基础是词语相对位置的确定。

2.1　词语组合的选择

词语组合选取的原则是寻找 2 个句子经分词、词性标注、虚词剔除等预处理后的集合中语义相似度最大的词语组合。为了便于寻找词语组合,将待比较的句子 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 预处理后表示成集合:$\boldsymbol{S}_1=\{W_{11},W_{12},\cdots,W_{1m}\}$,$\boldsymbol{S}_2=\{W_{21},W_{22},\cdots,W_{2n}\}$,并生成特征矩阵如下:

$$\boldsymbol{S}_1\times\boldsymbol{S}_2^{\mathrm{T}}=\begin{bmatrix} W_{11}W_{21} & W_{12}W_{21} & \cdots & W_{1m}W_{21} \\ W_{11}W_{22} & W_{12}W_{22} & \cdots & W_{1m}W_{22} \\ \vdots & \vdots & & \vdots \\ W_{11}W_{2n} & W_{12}W_{2n} & \cdots & W_{1m}W_{2n} \end{bmatrix}$$

其中,$W_{1i}W_{2j}=Sim(W_{1i},W_{2j})$。

相似度最大的词语组合选取方法为:遍历特征矩阵,选取出相似度最大的词语组合,再将其所在的行和列从特征矩阵中删除并生成新的特征矩阵,按照此规则继续取词语组合,直到特征矩阵中没有元素为止,最终获得相似度最大词语组合的集合:

$$\max C=\{Sim(W_{1i},W_{2j})_{\max 1},Sim(W_{1i},W_{2j})_{\max 2},\cdots,Sim(W_{1i},W_{2j})_{\max k}\}$$

其中，$i \leqslant m; j \leqslant n; k = m \leqslant n ? m : n$。

以“我喜欢你”和“你喜欢我”为例计算相似度时，相似度最大的词语组合的集合为：$\max C = \{Sim(我,我) = 1, Sim(喜欢,喜欢) = 1, Sim(你,你) = 1\}$。如果以集合中词语组合的值进行加权平均，那么其相似度为 1，很明显不符合人的直观判断。因为这里宾语和主语倒置，所以在计算句子相似度时，只考虑词语的语义相似度不够全面，还要考虑词语组合相对位置偏移量，偏移量计算的基础是词语相对位置的确定。

2.2 词语相对位置的确定

目前词语相对位置的确定一般以句首为基准，将第一个词位置设为 0，后面的词语位置以步长为 1 递增，再将词语组合中两词语位置差值的绝对值作为此词语组合的相对位置偏移量。这样会由于插入一个词语导致其他词语相对位置不准确，因此影响词语组合偏移量计算的准确性。为了修正偏移量，本文提出一种以句子中心词为基准确定词语相对位置的方法，将句子中心词位置设为 0，中心词前面的词语相对位置以步长为 1 递减，中心词后面的词语相对位置以步长为 1 递增。

本文根据词语词性和语法规则进行中心词的选取，分词和词性标注工具是中科院的 ICTCLAS 系统，其可以在二级词性标注时辨别出不及物动词、其他动词、时间词以及地点词，为本文的计算方法提供可行性条件。本文涉及该系统中的词语种类和相对应的词性标注符号见表 1。

表 1 词语种类和词性标注符号

词语种类	词性标注符号
地点名词	ns
时间词	t
不及物动词	vi
趋向动词	vf
副动词	vd
名词	n
处所词	s
人称代词	rr
副词	d

2.2.1 中心词选取

本文提出一种以句子中心词为基准衡量词语相对位置的方法，词语相对位置的确定依靠中心词。本文从汉语特点进行考虑，根据词语词性和语法规则对中心词进行选取，选取原则如下：

1）动词是一个句子的核心词，在句法分析生成语义树时，以动词作为根节点延伸到句子的其他成分并且作为句子的核心词，不应该考虑其相对位置偏移量，只考虑其语义信息。

2）动词有不及物动词和其他动词，不及物动词相比于其他动词，其本身意义完整，其他动词必须跟宾语才是完整的实义动词。

3）根据汉语语义后置的原则，当一类动词的数量较多时，最后一个动词的语义信息最丰富；当句子没有动词时，最后一个词为句子的主体对象。

根据以上描述，本文选取中心词的优先级为：不及物动词 > 其他动词 > 最后一个词，句子中心词选取流程如图 1 所示。

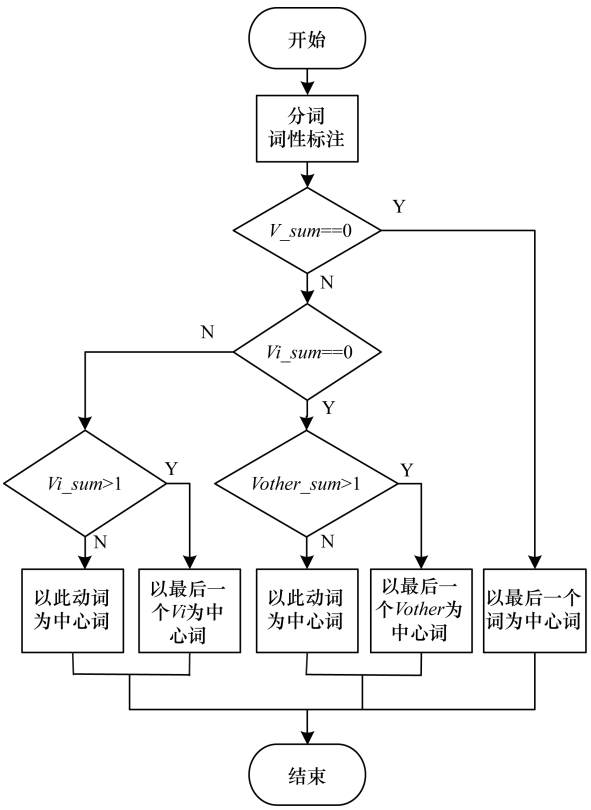

图 1 句子中心词选取流程

在图 1 中，*V_sum* 表示动词数量，*Vi_sum* 表示不及物动词数量，*Vother_sum* 表示其他动词数量。

2.2.2 词语相对位置标定

在中心词选取后可以对词语相对位置进行标定。标定方法如下：将中心词的位置设为 0，中心词前面的词语相对位置以步长为 1 递减，中心词后面的词语相对位置以步长为 1 递增。

以“我喜欢你的手表”和“我非常喜欢你的手表”为例确定词语相对位置时，这 2 句话中的中心词都是动词“喜欢”。如果以句首为基准确定词语相对位置，由于“非常”是对“喜欢”的修饰，因此导致动词以及动词后面的宾语都发生了一个位置的偏移，这里副词“非常”是对动词“喜欢”的修饰。从语法角度来看，不应对后面的句子成分产生影响，而当以动词为基准时，这里只对主语“我”产生一个偏移，更科学。以动词为基准衡量词语相对位置的实例分析见图 2。

实例 1	以句首为基准	词语/词性标注	我/rr	喜欢/vi	你/rr	手表/n		我/rr	喜欢/vi	你/rr	手表/n	
		相对位置 *pos*	0	1	2	3		−1	0	1	2	
实例 2	以中心词为基准	词语/词性标注	我/rr	非常/d	喜欢/vi	你/rr	手表/n	我/rr	非常/d	喜欢/vi	你/rr	手表/n
		相对位置 *pos*	0	1	2	3	4	−2	−1	0	1	2

图 2　以动词为基准衡量词语相对位置的实例分析

以“农村经济”和“经济”为例确定词语相对位置时,如果以句首为基准确定词语相对位置,第一句的“农村”和第二句的“经济”在一个位置,那么显然不符合中文语法习惯。在汉语中,一般是修饰在前,主体在后,第一句中“农村”是对“经济”的修饰限定,这 2 个短语中没有动词,根据本文方法以最后一个词为基准考虑相对位置,符合中文语法习惯。以句尾词语为基准衡量相对位置的实例分析见图 3。

实例 1	以句首为基准	词语/词性标注	农村/n	经济/n	农村/n	经济/n
		相对位置 *pos*	0	1	−1	0
实例 2	以中心词为基准	词语/词性标注	经济/n		经济/n	
		相对位置 *pos*	0		0	

图 3　以句尾词语为基准衡量词语相对位置的实例分析

为了便于理解,本文用以上简单句型分析词语相对位置的确定过程。针对其他复杂句型,词语相对位置确定都以中心词为基准,中心词的位置设为 0,词语若在中心词左侧,则词语相对位置以步长为1递减;若在右侧,则词语相对位置以步长为 1 递增。

2.2.3　全局性限定词剔除

在对词语相对位置标定的过程中,不是所有词语都需要标定,一个句子中一般含有对整个句子进行修饰、限定的词语,这类词语放在句中任意位置不影响句子相似度,因此,不需要考虑词语组合相对位置偏移量。比如“我明天去上海吃饭”和“明天我去上海吃饭”,这两句话中的“明天”是对整个句子进行时间限定,放在不同位置不影响句子相似度。因此,在标定相对位置时先剔除句子中时间的限定词,只考虑其包含的语义信息,不进行相对位置偏移量的计算,不仅时间词对整个句子起限定作用,而且地点词在标定相对位置时也应该先剔除。

以“明天我们去吃饭”和“我们明天去吃饭”为例,这两句话的中心词都为不及物动词“吃饭”,如果以句首为基准确定词语的相对位置,那么相似度计算不为 1,明显不对;以本文方法确定词语的相对位置,先将 2 个句子中的时间词“明天”剔除出来考虑其语义相似度,再考虑剩下词语组合的语义相似度和相对位置移偏量,更符合人的经验判断。剔除限定词的词语相对位置标定的实例分析见图 4。

实例 1	以句首为基准	词语/词性标注	明天/t	我们/rr	去/vf	吃饭/vi	明天/t	我们/rr	去/vf	吃饭/vi
		相对位置 *pos*	0	1	2	3	/	−2	−1	0
实例 2	以中心词为基准	词语/词性标注	我们/rr	明天/t	去/vf	吃饭/vi	我们/rr	明天/t	去/vf	吃饭/vi
		相对位置 *pos*	0	1	2	3	−2	/	−1	0

图 4　剔除限定词词语相对位置标定的实例分析

2.2.4　词语组合相对位置偏移量计算

词语组合相对位置偏移量计算的基础是词语相对位置的确定。以句子中心词对词语相对位置进行标定后,就可以计算偏移量。本文为了减小单个词语组合偏移量对整个句子相似度产生的影响,将词语组合中 2 个词语相对位置差值的绝对值作为偏移量,偏移量的大小只对该词语组合的相似度产生影响,而不影响其他词语组合相似度。

2.2.5　改进的句子相似度计算方法

在计算句子相似度时,加入词语组合相对位置偏移量可以提高句子相似度计算的准确率,本文提出一种以句子中心词为基准衡量词语组合相对位置偏移量的句子相似度计算方法,并将句子 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 的相似度计算定义为式(6)。分子是相似度最大的

前 k 个词语组合的相似度值和相对应的偏移量乘积之和，k 为 2 个句子中词语数较少的词语数值。分母为 2 个句子中词语数较多的词语数值，其比值从一定程度上体现出 2 个句子长度的差异性；$pos(W)$ 是浅层次句子结构信息的表达；$Sim(W_{1i},W_{2j})$ 计算词语的语义相似度。该公式融合了句子的句长差异性信息、浅层次结构信息以及语义信息，避免了经验取值以及算法复杂度高的问题。

$$Sim(\boldsymbol{S}_1,\boldsymbol{S}_2)=\frac{\sum_{g=1}^{k}\left(\begin{array}{l}Sim(W_{1i},W_{2j})\\ \max_g\left(1-\dfrac{|pos(W_{1i})-pos(W_{2j})|}{\max(len(\boldsymbol{S}_1),len(\boldsymbol{S}_2))}\right)\end{array}\right)}{\max(len(\boldsymbol{S}_1),len(\boldsymbol{S}_2))} \tag{6}$$

其中，$k=m\leqslant n$？$m:n$；$i\leqslant m$；$j\leqslant n$；$len(S)$ 表示句子 S 的词语个数；$pos(W)$ 表示词语 W 基于句子中心词的相对位置；$Sim(W_{1i},W_{2j})$ 表示 W_{1i} 和 W_{2j} 基于《知网》的词语语义相似度。

2.3 改进的句子相似度算法描述

本文提出算法可以有效修正词语组合相对位置偏移量，式(6)综合了句子的长度差异性信息、浅层次结构信息以及语义信息。改进的句子相似度算法具体描述如下：

输入 待计算相似度的 2 个句子 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$

输出 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 的相似度

步骤 1 将待计算相似度的 2 个句子 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 进行分词、词性标注、剔除停用词和虚词等预处理后，获得词语集合 $\boldsymbol{S}_1=\{W_{11},W_{12},\cdots,W_{1m}\}$，$\boldsymbol{S}_2=\{W_{21},W_{22},\cdots,W_{2n}\}$。

步骤 2 根据步骤 1 中获得的 2 个词语集合，利用《知网》计算两集合中两两词语间的语义相似度形成特征矩阵。

$$\boldsymbol{S}_1\times\boldsymbol{S}_2^{\mathrm{T}}=\begin{bmatrix}W_{11}W_{21} & W_{12}W_{21} & \cdots & W_{1m}W_{21}\\ W_{11}W_{22} & W_{12}W_{22} & \cdots & W_{1m}W_{22}\\ \vdots & \vdots & & \vdots\\ W_{11}W_{2n} & W_{11}W_{22} & \cdots & W_{1m}W_{2n}\end{bmatrix}$$

其中，$W_{1i}W_{2j}=Sim(W_{1i},W_{2j})$。

步骤 3 遍历特征矩阵，取出相似度值最大的词语组合，删除其所在的行和列的所有元素并得到新的特征矩阵，反复此过程，直至特征矩阵中没有元素。

步骤 4 根据词语词性和语法规则确定句子中心词，剔除步骤 1 集合中的地点、时间词语，标定句中剩余词的相对位置。

步骤 5 根据步骤 3 中遍历特征矩阵，获取相似度最大词语组合的集合和步骤 4 中标定的词语相对位置，计算词语组合中两词语的相对位置偏移量。

步骤 6 根据改进的句子相似度计算式(6)计算两句子的相似度。

3 实验与评价

为验证本文算法的优越性，实现实验所需要的 3 种算法。算法 1 为文献[12]提出的基于向量词序的句子相似度算法；算法 2 以句首为基准，衡量词语相对位置偏移量，以式(6)作为句子相似度计算公式；算法 3 为本文算法。将算法 1 和算法 2 作为对比算法，利用典型句子对对 3 种算法作出分析并以人工聚类的句子库作为测试集进行验证，评价指标为准确率、召回率以及综合指标 F 值。

3.1 实验测试集及评价指标

本文利用人工聚类的句子库作为测试语料，共有 20 类，每类有 60～70 个句子，涉及不同领域。随机抽取一类中的任意一句作为实验句，其他句子作为干扰句，并按照相似度从大到小进行排序，在此基础上计算准确率(P)、召回率(R)以及综合指标 F 值(F)。

$$P=\frac{\text{系统检索到的该类句子数}}{\text{系统检索到的句子总数}}$$

$$R=\frac{\text{系统检索到的该类句子数}}{\text{该类句子总数}}$$

$$F=\frac{2\times P\times R}{P+R}$$

3.2 实验结果分析

为证明本文算法相比其他 2 种算法的优越性，选取典型句子对并对算法 1、算法 2 以及本文算法进行相似度计算并作出分析，计算结果见表 2。

表 2 3 种算法典型句子对相似度对比

编号	句子对	相似度		
		算法 1	算法 2	本文算法
1	我喜欢你 我爱你	0.54	1.00	1.00
2	我明天去上海读书 明天我去上海上学	0.57	0.92	1.00
3	我喜欢你手上美丽的手表 我非常喜欢你手上美丽的手表	0.74	0.76	0.84
4	实体经济 农村实体经济	0.48	0.44	0.67

从表 2 可以看出，编号 1 和编号 2 的句子对以算法 1 作为计算方法时，其相似度值低于 0.9，明显与事实不符，因为句子对中含有同义词“喜欢”和“爱”、“读书”和“上学”，而算法 1 没考虑语义信息，将这些同义词按照不同词语处理是不合适的；编号 2 和编号 3 句子对的相似度按照经验判断非常高，本文算法相比算法 2 更能体现该信息。主要是因为以句子中心词衡量词语相对位置可以修正词语组合的相对位置偏移量，进而提高句子相似度计算的准确率，所以本文算法具有一定的实用性。

为进一步验证本文算法的优越性，根据 3.1 节描述的公式，利用人工聚类句子库作为测试集，按照

实验所需的 3 种算法计算出准确率、召回率以及综合指标 F 值,计算结果见表 3。

表 3　3 种算法评价指标对比

算法	准确率	召回率	综合指标 F 值
算法 1	0.571 2	0.626 5	0.597 6
算法 2	0.672 3	0.615 7	0.642 8
本文算法	0.756 5	0.664 3	0.707 4

综合指标作为评估算法效果的标准,F 值越大说明算法越有效。从表 3 中可以看出,综合指标最大的是本文算法,其次是算法 2,最小的是算法 1。主要是因为算法 1 没有考虑词语的语义信息,而中文中一词多义和同义词的现象很多,所以其综合指标相对较小。本文算法的综合指标 F 值大于算法 2,说明以句子中心词为基准确定词语的相对位置可以减小偏移量的累计,并且偏移量的修正对于提高综合指标 F 值有一定作用。

4　结束语

为提高算法的综合指标 F 值,本文对词语组合相对位置偏移量进行修正,提出一种以句子中心词为基准衡量词语组合相对位置偏移量的句子相似度计算方法。在人工构造的实验测试集下验证了相似度算法的有效性。由于句子中心词能够左右句子的表达含义,正确分配句子中心词的权重对于提高相似度计算的准确率具有重要意义,因此下一步将对中心词的权重分配进行研究。

参考文献

[1] 黄九鸣,吴泉源,刘春阳,等. 短文本信息流的无监督会话抽取技术[J]. 软件学报,2012,23(4):735-747.

[2] 李景玉,张仰森,陈若愚. 面向用户查询意图的句子相似度分层计算[J]. 计算机科学,2015,42(1):227-231.

[3] 李　茹,王智强,李双红,等. 基于框架语义分析的汉语句子相似度计算[J]. 计算机研究与发展,2013,50(8):1728-1736.

[4] 郑　诚,李　清,刘福君. 改进的 VSM 算法及其在 FAQ 中的应用[J]. 计算机工程,2012,38(17):201-204.

[5] 裴　婧,包　宏. 汉语句子相似度计算在 FAQ 中的应用[J]. 计算机工程,2009,35(17):46-48.

[6] 程传鹏,吴志刚. 一种基于知网的句子相似度计算方法[J]. 计算机工程与科学,2012,34(2):172-175.

[7] 赵　臻,吴　宁,宋盼盼. 基于多特征融合的句子语义相似度计算[J]. 计算机工程,2012,38(1):171-173.

[8] 冯　凯,王小华,谌志群. 基于动态规划的汉语句子相似度算法[J]. 计算机工程,2013,39(2):220-224.

[9] 姜　华,韩安琪,王美佳,等. 基于改进编辑距离的字符串相似度求解算法[J]. 计算机工程,2014,40(1):222-227.

[10] 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.

[11] GALITSKY B. Machine Learning of Syntactic Parse Trees for Search and Classification of Text [J]. Engineering Applications of Artificial Intelligence,2013,26(3):1072-1091.

[12] 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.

[13] 肖　雨,崔荣一,怀丽波. 一种融合位置信息的字符串相似度度量方法[J]. 计算机应用研究,2015,32(11):3287-3290.

[14] 知网[EB/OL]. (2012-07-15). http://www.keenage.com/.

[15] 刘青磊,顾小丰. 基于《知网》的词语相似度算法研究[J]. 中文信息学报,2010,24(6):31-36.

[16] 白培发,王成良,徐　玲. 一种融合词语位置特征的 Lucene 相似度评分算法[J]. 计算机工程与应用,2014,50(2):129-132.

编辑　陆燕菲

(上接第 233 页)

[3] ZHANG Zhongfu, CHEN Xiangen, LI Jingwen. On Adjacent Vertex Distinguish Total Coloring of Graphs [J]. China Science:Series A,2005,48(3):289-299.

[4] ZHANG Zhongfu, QIU Pengxiang, XU Baogen. Vertex Distinguishing Total Coloring of Graphs [J]. ARS Combinatoria,2008,87(2):33-45.

[5] 张忠辅. 图的 Smarandachely 邻点可区别边染色[J]. 兰州交通大学学报,2009,1(1):1-13.

[6] 朱恩强,王治文,张忠辅. 若干倍图的 Smarandachely 邻点可区别边染色[J]. 山东大学学报(理学版),2009,44(12):25-29.

[7] 刘信强,刘旺发. 图的 Smarandachely 邻点无圈边色数的一个上界[J]. 系统科学与数学,2013,33(5):550-554.

[8] 李敬文,张　欣,王治文,等. 路图的 Smarandachely 全染色算法[J]. 计算机应用研究,2011,28(3):848-850.

[9] 陈祥恩,王治文,赵飞虎,等. 若干强积图及合成图的邻点可区别一般边染色[J]. 山东大学学报(理学版):2013,48(6):18-22.

[10] 李瑾姝,刘　静,焦李成,等. 一种基于多智能体进化的广义图染色算法[J]. 软件学报,2009,20(2):315-326.

[11] 赵焕平,刘　平,李敬文. 完全图的点可区别强全染色算法[J]. 计算机工程,2012,38(17):32-34.

[12] 陈祥恩,王治文,赵飞虎,等. 几类弱积图的邻点可区别一般边染色[J]. 兰州大学学报(自然科学版),2012,48(1):97-99,103.

[13] 伍大清,邵　明,李　俊,等. 基于奖惩机制的协同多目标优化算法[J]. 计算机工程,2015,41(10):186-191,198.

[14] 戚玉涛,刘　芳,常伟远,等. 求解多目标问题的 Memetic 免疫优化算法[J]. 软件学报,2013,24(7):1529-1544.

[15] 尚荣华,胡朝旭,焦李成,等. 多目标优化算法在多分类中的应用研究[J]. 电子学报,2012,40(11):2264-2269.

编辑　顾逸斐