第45卷第2期
2005年3月

大连理工大学学报
Journal of Dalian University of Technology

Vol. 45. No. 2
Mar. 2005

文章编号：1000-8608(2005)02-0291-07

基于语义理解的文本相似度算法

金　博[1,2]，史彦军[1,2]，滕弘飞*[1]

（1. 大连理工大学 机械工程学院，辽宁 大连 116024；
2. 大连理工大学 计算机科学与工程系，辽宁 大连 116024）

摘要：相似度的计算在信息检索及文档复制检测等领域具有广泛的应用前景. 研究了文本相似度的计算方法，在知网语义相似度的基础上，将基于语义理解的文本相似度计算推广到段落范围，进而可以将这种段落相似度推广到篇章相似度计算. 给出了文本（包括词语、句子、段落）相似度的计算公式及算法，用于计算两文本之间的相似度. 实例验证表明，该算法与现有典型的相似度计算方法相比，计算准确性得到提高.

关键词：知网；语义；文本相似度；复制检测；信息检索
中图分类号：TP391.1　　**文献标识码**：A

0　引　言

文本相似度计算在信息检索、数据挖掘、机器翻译、文档复制检测等领域有着广泛的应用. 目前，关于文本相似度的研究主要有篇章与篇章之间的相似度，如 Willett 研究的文档分类算法[1]；短语与篇章之间的相似度，如 Salton 等研究的信息索引方法[2]；短语与篇章中某一部分（如段落）的相似度，如 Callan 研究的文章段落检索等[3]. 所用到的文本相似度计算方法主要有以下几种：向量空间模型、广义向量空间模型、隐性语义索引模型、基于属性论的方法、基于海明距离的计算方法、基于数字正文的重构方法等，均是基于统计学的计算方法，需要大规模语料库的支持和长时间的训练过程，具有一定的局限性.

与基于统计学的相似度计算方法相比，基于语义理解的相似度计算方法不需要大规模语料库的支持，也不需要长时间的训练，具有准确率高的特点，相关的研究主要有使用 WordNet 进行相似度计算的方法，如 Agirre 等的词语消歧研究[4]；使用同义词词林进行相似度计算的方法，如车万翔等的句子相似度计算研究[5]；使用知网知识结构进行相似度计算的方法，如刘群、李素建的词语相似度及句子相似度的研究[6,7]. 目前，基于语义理解的相似度计算大多限于词语或句子范围. 本文以知网词语相似度计算为基础，将其应用范围扩大到段落，以用于抄袭识别或信息检索领域.

1　文本相似度的 3 个层次

本文所研究的文本包括词语、句子和段落等，关于篇章的相似度将另文讨论. 本文的文本相似度计算将包括以下几种关系：词语与词语、词语与句子、词语与段落、句子与句子、句子与段落和段落与段落等. 上述的各种相似度关系可分别用于不同的研究领域，如词语与词语之间的相似度计算可用于机器翻译领域；词语与句子、词语与段落、句子与段落之间的相似度计算可以用于信息检索领域；句子与句子之间的相似度计算可以用于自动问答领域；段落与段落之间的相似度计算可以用于复制检测或剽窃检测领域等. 所以文本相似度计算的研究适用范围较广，是信息处理技术中一项基础性的研究. 本文将文本相似度的计算划分为 3 个层次：词语层次，包括词语与词语、词语与句子、词语与段落之间的相似度计算；句子层次，包括句子与句子、句子与段落之间的相似度计算；段落层次，包括段落与段落之间的相似度计

收稿日期：2003-12-25；　**修回日期**：2005-01-24.
基金项目：国家自然科学基金资助项目(60073036，50275019)派生研究.
作者简介：金　博(1978-)，男，博士生；滕弘飞*(1936-)，男，教授，博士生导师.

算. 3个层次的相似度计算各有侧重,其共性表现为可替换性,即若两对象(词语、句子或段落)相似,则此两对象在各自的上下文关系中是可以相互替换的.

在词语层次中,相似度用于衡量文本中词语的可替换程度,这里的词语相似度不等同于词语的相关度. 例如"军人"和"武器"两个词,其相似度非常低,而相关度却很高. 可以这样认为,词语相似度反映的是词语之间的聚合特点,而词语相关度反映的是词语之间的组合特点. 本文所研究的词语是处于一定的上下文环境中,而且词语是通过句子结构分析(如分词处理)得到的,词语作为句子组件作用的特性很明显,所以按照词语的可替换程度(即组合特点)度量词语相似程度的方法是合理的. 与词语作为句子组件的特性类似,在句子层次,句子可以看做是段落的组件;在段落层次,段落可以看做是篇章的组件,所以句子和段落层次的相似度可以由词语层次的相似度推广得出.

2 文本相似度计算方法

本文按照文本相似度的3个层次,首先通过语义分析计算出词语相似度,接着通过分词及对句子结构进行分析计算出句子相似度,最后按照句子与段落之间的关系得到段落相似度的计算方法.

2.1 词语相似度

本文研究的词语相似度部分的计算借鉴文献[6]的词语相似度研究. 文献[6]主要介绍了知网的知识结构及其知识描述语言的语法,并提出利用知网进行词语相似度的计算方法,给出了词语相似度的计算公式,但由于其用到的一些经验公式未充分地利用知网的知识结构,如义原相似度计算公式等,计算结果不尽准确. 本文对此进行了改进.

知网是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[8],其中包含丰富的词汇语义知识和世界知识. 在知网中,词汇语义的描述被定义为义项(概念). 每一个词可以表达为几个义项. 义项又是由一种知识表示语言来描述的,这种知识表示语言所用的词汇称做义原. 如图1所示,知网与一般的语义词典(如同义词词林或WordNet)不同,其语义树并不涵盖所有词语,而将描述词汇语义的义原用树状结构组织起来,义原根据义原之间的属性关系分为多棵义原树,树与树之间又存在一定的关系,形成知网所具有的网状知识结构. 相比词汇的规模,义原的数量很少,只有1 500多个,但其组合起来可以表达数以万计的词语. 这样,将词语的相似度计算转化为义原的相似度计算可以提高计算效率,有利于知识库的扩展. 义原的相似度可以根据义原树结构的相对位置关系求出.

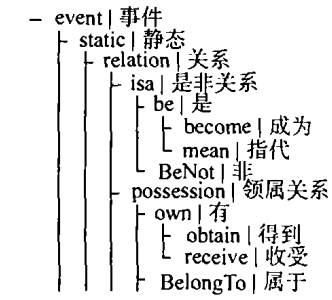

图1 义原层次结构

Fig. 1 Level structure of sememe

用知网计算词语相似度时首先要找出与词语对应的义项表达式,即知网的知识描述语言,表1为知网字典中部分词语的知识表达式. 比如要计算"剽窃"与"抄袭"两词的相似度,首先要找出两词的义项表达式,由表1可得,"剽窃"只有1个义项,而"抄袭"一词在知网中有3个义项. 显而易见,在计算时应将标号为114646的"剽窃"义项条目与标号为011942的"抄袭"义项条目相比较,而不是与标号为011940或011941的"抄袭"义项条目相比较. 由此即可根据知网本身的知识结构实现词语的消歧处理.

表1 知网知识描述语言实例

Tab. 1 Description of knowledge specimen in HowNet

标号	词语	义项
061554	男人	human| 人,family| 家,male| 男
005241	必须	{modality| 语气}
114646	剽窃	steal| 偷,* copy| 抄写
011940	抄袭	attack| 攻打,military| 军
011941	抄袭	imitate| 模仿
011942	抄袭	steal| 偷,* copy| 抄写

在算法表达上,设定两个词语 w_1 和 w_2,如果 w_1 有 n 个义项 $s_{11},s_{12},\cdots,s_{1n}$,$w_2$ 有 m 个义项 $s_{21},s_{22},\cdots,s_{2m}$,规定 w_1 与 w_2 之间的相似度是各个义项的相似度最大值,即

$$\mathrm{sim}W(w_1,w_2)=\max_{i=1,\cdots,n,j=1,\cdots,m}\mathrm{sim}WS(s_{1i},s_{2j}) \tag{1}$$

其中 $\mathrm{sim}W(w_1,w_2)$ 表示两词语的相似度，$\mathrm{sim}WS(s_1,s_2)$ 表示两义项的相似度. 这样，就把两个词语之间的相似度问题归结为两个词语的义项之间的相似度问题.

由于义项都由义原表示，义项相似度计算又归结为义原相似度的计算. 所有的义原根据上下位关系构成了一个树状的义原层次体系(见图1)，可以通过计算语义距离的办法计算义原相似度. 文献[6]根据经验公式得到这两个义原之间的相似度：

$$\mathrm{sim}WP(p_1,p_2)=\alpha/(d+\alpha) \tag{2}$$

其中 p_1 和 p_2 表示两个义原(primitive)；d 是 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数，这里作为两义原的义原距离；α 是一个可调节的参数，其含义是相似度为0.5时的路径长度，这里根据义原树的深度取 $\alpha=1.6$.

在研究中发现，义原之间的相互关系并不仅仅与义原之间的距离相关，义原之间相对位置关系对义原的相似度也有着较大的影响，如图1中义原“有”和义原“得到”之间的义原距离与义原“事件”和义原“静态”之间的义原距离相同，但明显可以看出，这两对义原的相似度关系并不完全相同，参照 Agirre 等在用 WordNet 知识结构进行相似度计算时使用的方法[4](加入了概念树深度及概念树密度的影响因子)，作者对式(2)进行了改进. 但由于知网知识结构与 WordNet 不同，知网知识结构密度不均匀的问题并不明显，本文仅引入义原相对位置的影响因子 h_1h_2.

$$\mathrm{sim}WP(p_1,p_2)=\alpha h_1h_2/(d+\alpha) \tag{3}$$

其中 h_1、h_2 为两义原 p_1、p_2 在义原树中的深度.

另由表1可见，知网的知识表达是比较复杂的，这种知识描述语言归纳为以下几条[6]：

(1) 知网收录的词语主要归为两类，一类是实词，一类是虚词.

(2) 虚词的描述比较简单，用“{句法义原}”或“{关系义原}”进行描述.

(3) 实词的描述比较复杂，由一系列用逗号隔开的“语义描述式”组成，这些“语义描述式”又有以下3种形式：

ⓐ独立义原描述式. 用“基本义原”，或者“(具体词)”进行描述.

ⓑ关系义原描述式. 用“关系义原 = 基本义原”或者“关系义原 = (具体词)”或者“(关系义原 = 具体词)”来描述.

ⓒ符号义原描述式. 用“关系符号 基本义原”或者“关系符号(具体词)”加以描述.

(4) 在实词的描述中，第1个描述式总是1个基本义原，这也是对该实词最重要的一个描述式，这个基本义原描述了该实词的最基本的语义特征，称做第一独立义原描述式.

实现时考虑到在汉语中实词才是表达文章意义的关键词汇，所以在相似度算法中省略了虚词部分的相似度计算，这样可以在保证计算准确性的前提下提高计算效率.

对于实词概念的语义表达式，参照知网对实词的描述，将实词相似度计算分成以下4个部分[6].

(1) 第一独立义原描述式：两个义项的这一部分的相似度记为 $\mathrm{sim}WP_1(p_1,p_2)$，其可由式(3)直接算出.

(2) 其他独立义原描述式：语义表达式中除第一独立义原以外的所有其他独立义原称做其他独立义原. 两个义项的这一部分的相似度记为 $\mathrm{sim}WP_2(p_1,p_2)$，其值是独立义原相似度最大组合序列的加权平均值；

$$\mathrm{sim}WP_2(p_1,p_2)=\sum_{k=1,\cdots,p}\lambda_i\max_{i=1,\cdots,m,j=1,\cdots,n}\mathrm{sim}WP(p_i,p_j) \tag{4}$$

(3) 关系义原描述式：语义表达式中所有用关系表示的义原称做关系义原. 两个义项的这一部分的相似度记为 $\mathrm{sim}WP_3(p_1,p_2)$，其值是关系相同的义原组合的最大值；

$$\mathrm{sim}WP_3(p_1,p_2)=\max_{i=1,\cdots,m,j=1,\cdots,n}\mathrm{sim}WP(p_i,p_j) \tag{5}$$

(4) 符号义原描述式：语义表达式中所有的用符号表示的义原称做符号义原. 两个义项的这一部分的相似度记为 $\mathrm{sim}WP_4(p_1,p_2)$，由于符号义原与关系义原在知网中表示方式相同，其公式与关系义原描述式类似.

$$\mathrm{sim}WP_4(p_1,p_2)=\max_{i=1,\cdots,m,j=1,\cdots,n}\mathrm{sim}WP(p_i,p_j) \tag{6}$$

于是，两个义项语义表达式的整体相似度记为[6]

$$\mathrm{sim}WS(s_1,s_2)=\sum_{i=1}^{4}\beta_i\mathrm{sim}WP_i(p_1,p_2) \tag{7}$$

其中 $\beta_i(1\leqslant i\leqslant 4)$ 是可调节的参数，且有 $\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$. 后者反映了 $\mathrm{sim}WP_1$ 到 $\mathrm{sim}WP_4$ 对于总体相似度所起到的作

用依次递减. 由于第一独立义原描述式反映了一个概念最主要的特征,应该将其权值定义得比较大,一般应在0.5以上.

在使用式(7)的算例验证中发现,如果$\mathrm{sim}WP_1$非常小,但$\mathrm{sim}WP_3$或者$\mathrm{sim}WP_4$比较大,将导致整体的相似度仍然较大的不合理现象. 因此对式(7)进行了修改,给出如下公式[6]:

$$\mathrm{sim}WS(s_1,s_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}\mathrm{sim}WP_j(p_1,p_2) \tag{8}$$

其意义在于,主要部分的相似度值对于次要部分的相似度值起到制约作用,也就是说,如果主要部分相似度相对比较低,那么次要部分的相似度对于整体相似度所起到的作用也要降低.

但式(8)中将除第一独立义原相似度之外的其他义原相似度也进行了相互制约,这里存在不当之处. 根据知网的描述,除第一独立义原相似度之外的其他义原相似度是相对独立的,所以本文提出新的计算公式:

$$\mathrm{sim}WS(s_1,s_2)=\beta_1\mathrm{sim}WP_1(p_1,p_2)+\sum_{i=2}^{4}\beta_1\mathrm{sim}WP_1(p_1,p_2)\times\beta_i\mathrm{sim}WP_i(p_1,p_2) \tag{9}$$

至此,可以算出两实词之间的相似度值,下面进一步介绍句子相似度的计算.

2.2 句子相似度

当前基于语义理解的相似度研究还大多停留在词语范围,主要是由于句子相似度较词语相似度的计算更为复杂,其不仅包括语义关系的辨别,还包括句子结构的辨别等问题. 本文通过对句子结构的分析及前文的词语相似度的研究,将相似度的研究推广到句子范围.

由于汉语自身的特点,词与词之间没有明显的分割符号,句子相似度计算的第一步是进行句子分词处理. 这里使用的是中国科学院计算技术研究所研制的基于多层隐马模型的汉语词法分析系统ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System)[9],该系统的功能有中文分词、词性标注、未登录词识别等.

如前所述,实词具有实在的意义,本文只关心句子中的实词,所以在分词后所得到的字符串中,只将句子中的实词取出. 实词包括名词(N)、动词(V)、形容词(A)、数词(M)、量词(Q)和代词(R)等[10].

设句子s分词后包括k个词:

$$s=\{w_1,w_2,\cdots,w_k\} \tag{10}$$

而句子s可根据实词的种类分为6个集合:

$$s=\{N,V,A,M,Q,R\} \tag{11}$$

则根据分词后的词语属性将实词取出并归入相应的集合中:

$$s=\{N(w_i,\cdots),V(w_j,\cdots),A(w_k,\cdots),M(w_l,\cdots),Q(w_m,\cdots),R(w_n,\cdots)\} \tag{12}$$

由于汉语句子结构相当复杂,准确地分析句子结构在目前是困难的,这里采用了一种简单的方法进行处理,即根据实词属性分别计算两句子中各个词性集合中实词的词语相似度. 例如有两个句子:"在基于实例的机器翻译中,词语相似度的计算有着重要的作用."和"我们认为词语相似度的计算对于基于实例的机器翻译来说是十分重要的."根据对句子相似度的定义可以认定两句是相似的,因为其讨论的是同一个问题,即"词语相似度计算是重要的",并将其所要表达的内容完整地进行了描述,在上下文关系中两个句子可以相互替换. 但两句的句子结构却有较大的不同,第一句的"词语相似度的计算"是主语,而在第二句中"词语相似度的计算"则作为宾语,所以将句子结构进行准确的划分意义不大. 名词和代词一般作为主语和宾语,动词一般作为谓语,形容词、数词和量词一般作为定语或状语[11],考虑到句子倒装等问题,根据词性集合计算句子相似度是合适的. 如上述两个例句,名词集合完全相同,在名词集合中计算出的相似度为1,此结果与事实相符.

设待计算的两句子s_1、s_2的实词集合分别为$\{N_1,V_1,A_1,M_1,Q_1,R_1\}$、$\{N_2,V_2,A_2,M_2,Q_2,R_2\}$,以名词集合$N$为例讨论句子相似度的计算过程.

$$N_1=(w_{11},w_{12},\cdots,w_{1m}) \tag{13}$$

$$N_2=(w_{21},w_{22},\cdots,w_{2n}) \tag{14}$$

设N_{12}为句子s_1、s_2相似度的特征矩阵,

$$N_{12}=N_1\times N_2^{\mathrm{T}}=\begin{bmatrix} w_{11}w_{21} & \cdots & w_{1m}w_{21} \\ \vdots & & \vdots \\ w_{11}w_{2n} & \cdots & w_{1m}w_{2n} \end{bmatrix} \tag{15}$$

其中

$$w_{1i}w_{2j}=\mathrm{sim}W(w_{1i},w_{2j}) \tag{16}$$

计算时首先遍历相似度特征矩阵,取出相似度最大的词语组合,再将其所属行和列从相似度特征矩阵中删除,继续选取余下矩阵中相似度最

大的组合，直到矩阵中元素为零，此时可得到词语最大组合序列：

$$\max L = \{\mathrm{sim}W_{\max_1}, \mathrm{sim}W_{\max_2}, \cdots, \mathrm{sim}W_{\max_k}\} \tag{17}$$

句子 s_1、s_2 名词集合相似度

$$\mathrm{sim}S_1(s_1, s_2) = \frac{1}{k}\sum_{i=1}^{k}\mathrm{sim}W_{\max_i} \tag{18}$$

同理，可以得到动词、形容词等其他集合的相似度. 整句相似度 $\mathrm{sim}S(s_1, s_2)$ 由各集合相似度加权平均得到，如下式：

$$\mathrm{sim}S(s_1, s_2) = \sum_{i=1}^{6}\beta_i \mathrm{sim}S_i(s_1, s_2) \tag{19}$$

其中 β_i 为权系数，其值的选取根据语言学知识及实验得到.

2.3　段落相似度

根据上述给出的词语及句子相似度计算方法就可以进行段落相似度的计算. 参照句子相似度的计算过程，段落文本相似度的计算过程描述如下：

设两段落文本为 t_1、t_2，而段落文本又是由多个句子组成的，

$$t_1 = \{s_{11}, s_{12}, \cdots, s_{1m}\} \tag{20}$$

$$t_2 = \{s_{21}, s_{22}, \cdots, s_{2n}\} \tag{21}$$

则段落相似度是由句子相似度计算出的. 其算法与句子相似度的计算方法相类似，根据式(17)推导出最大相似度句子组合为 $\max L\_S$：

$$\max L\_S = \{\mathrm{sim}S_{\max_1}, \mathrm{sim}S_{\max_2}, \cdots, \mathrm{sim}S_{\max_k}\} \tag{22}$$

其中 $\mathrm{sim}S_{\max}$ 为两段落中句子的最大相似度组合. 参照式(18)得到段落相似度计算公式：

$$\mathrm{sim}(t_1, t_2) = \frac{1}{k}\sum_{i=1}^{k}\mathrm{sim}S_{\max_i} \tag{23}$$

段落相似度计算过程由图 2 所示.

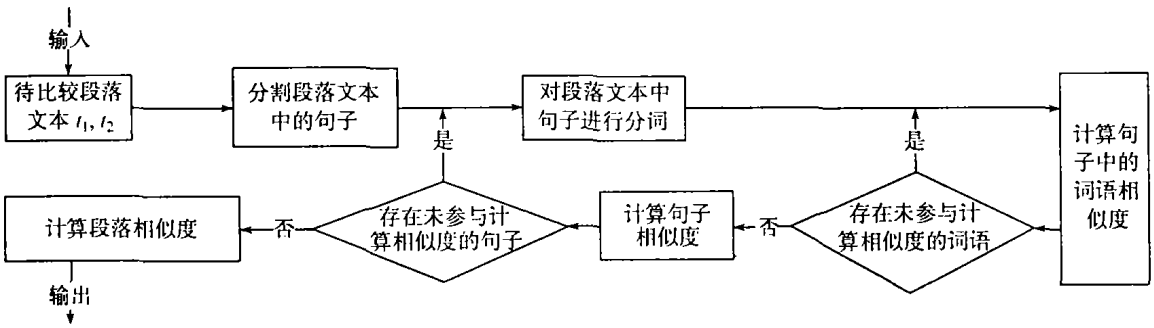

图 2　段落相似度计算过程

Fig. 2　Flow chart of paragraph similarity computation

3　实验及结果分析

利用上述方法，本文实现了基于知网的语义相似度计算程序模块. 根据本文研究中的词语、句子及段落 3 个层次，实验也从 3 个层次对结果进行分析.

在词语相似度层次，评价的方案最好是放在实际系统中(例如机器翻译系统)，以观察不同的计算方法对实际系统性能的影响. 但这需要一个完整的应用系统，在不具备这个条件的情况下，采用了人工判别的方法. 为此设置了一个对比实验，如表 2 所示. 其中方法 1 是仅使用知网语义表达式中第一独立义原来计算词语相似度；方法 2 是文献[7]中使用的词语语义相似度计算方法；方法 3 是文献[6]中使用的词语语义相似度计算方法；方法 4 是本文的相似度计算方法. 为了保证对比实验的可对比性，计算公式中的参数选取采用文献[6]中的参数值. 其中 $\alpha = 1.6, \beta_1 = 0.5, \beta_2 = 0.2, \beta_3 = 0.17, \beta_4 = 0.13, \gamma = 0.2, \delta = 0.2$.

表 2　词语相似度计算结果

Tab. 2　Results of word similarity computation experiment

词语 1	词语 2	方法 1	方法 2	方法 3	方法 4
男人	女人	1.000	0.668	0.833	0.673
男人	父亲	1.000	1.000	1.000	1.000
男人	母亲	1.000	0.668	0.833	0.673
男人	和尚	1.000	0.668	0.833	0.673
男人	经理	1.000	0.351	0.657	0.577
男人	高兴	0.016	0.024	0.013	0.007
男人	收音机	0.186	0.008	0.164	0.102
男人	鲤鱼	0.347	0.009	0.208	0.191
男人	苹果	0.285	0.004	0.166	0.157
男人	工作	0.186	0.035	0.164	0.102
男人	责任	0.016	0.005	0.010	0.115

由表2的实验结果可以看到，方法1的结果比较粗糙，只要是人相似度都为1，显然不够合理；方法2的结果比方法1细腻，能区分不同人之间的相似度，但有些不尽合理，如“男人”和“工作”之间的相似度大于“男人”和“鲤鱼”的相似度，不符合可替换性的要求，原因是其将相关度指标加入相似度的计算中；方法3以可替换性为相似度计算的主要性质，计算结果比较符合实际，但在某些词汇的计算中不尽准确，比如“男人”和“工作”之间的相似度与“男人”和“责任”之间的相似度差别较大，原因是其没有考虑义原深度等对相似度计算的影响；方法4加入了深度影响因子，计算结果更加准确，符合实际.

在句子相似度层次，设计了效果实验，利用的是文献[7]中自动问答系统的应用研究，将本文的句子相似度应用于一个QA系统中. 利用微软公司技术支持的FAQ数据，采取随机提问的方式进行测试，评判人判断查询答案是否正确. 在约50人参与评测的实验中，对查询出相似度最大的5个条目，95%以上的人认为找到合理的结果. 与文献[7]的效果(100名评测者对多数回答结果达到98%的满意率)相似.

在段落相似度层次，以自建的一个小的文本集为测试对象，该文本集共有包括机械、电子、航空、化工、计算机、物理、农业等7个领域的摘要共140篇，均采自相关领域核心期刊上公开发表的学术论文. 使用摘要作为测试对象是因为摘要文本的长度适中，一般在200字以内，句子在5句以内；此外，摘要的结构清晰，书写规范；同时，研究摘要的相似度计算对论文查抄及数据检索等领域的研究有重要的意义.

实验时首先按照不同领域内容确定不同的查询式，共7个. 查询式是一个与摘要文本规模相似的文本，使用简单的陈述句表达各领域中所研究的主要内容. 其结果的评价方法借鉴信息检索的评价方法[12]，其主要的评价指标包括召回率(R)、查准率(P)等. 召回率是实际识别出的正确结果与数据库中总的正确结果的百分比；查准率是返回结果中正确结果的百分比. 实验结果如表3所示.

可以看到，与传统的向量空间模型段落相似度计算方法相比，本文的基于知网语义理解的段落相似度计算在召回率及查准率上有明显的提高.

表3 段落相似度计算结果

Tab.3 Result of paragraph similarity computation experiment

文献类别	R/%		P/%	
	向量空间模型	语义理解	向量空间模型	语义理解
机械	36	63	50	70
电子	20	42	38	57
航空	35	47	40	52
化工	35	52	48	62
计算机	40	63	60	66
物理	26	31	55	42
农业	21	42	30	44

4 结 语

本文通过使用知网知识模型实现了文本相似度的计算，参考刘群等的词语相似度研究[6]，改进了词语相似度的计算公式，进而将基于知网语义理解相似度计算推广到句子及段落相似度计算，从实验效果来看具有较强的可操作性和可应用性. 其特点表现在以下几方面：

(1) 使用知网进行语义理解，知网的知识结构符合汉语语言习惯，意思理解准确.

(2) 在词语的相似度计算中、充分考虑了知网义原树状结构及知网知识的网状结构的特点，并加以利用，计算较为全面可靠.

(3) 跳过句法分析的难点，通过对实词集合的相似度计算，句子相似度的计算更为有效.

(4) 将目前的基于知网语义理解相似度计算推广到句子及段落范围的文本相似度计算，使相似度计算更具有实用价值.

目前，本算法还存在一些问题有待于进一步研究. 首先是计算效率，相似度计算中一个重要的应用领域就是信息检索，但如果没有很高的效率，信息检索将无法实际应用，这个问题在于算法的数据结构需要进一步优化，对数据库的也需要加以改进，如使用内存数据库等方法；其次是样本库的问题，在实际应用中，需要有较大规模的样本库，至少能够覆盖某一领域，这需要加强样本库的建设.

致谢：本文的词语相似度计算部分参考了中国科学院计算技术研究所刘群的基于知网的词汇语义相似度计算软件包；中文分词程序采用了中国科

学院计算技术研究所软件研究室的汉语词法分析系统 ICTCLAS. 以上程序及文档均来自中文自然语言处理开放平台 http://www.nlp.org.cn.

参考文献：

[1] WILLETT P. Recent trends in hierarchical document clustering: a critical review [J]. **Inf Process and Manage**, 1988, **24**(5):577-597.

[2] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval [J]. **Inf Process and Manage**, 1988, **24**(5):513-523.

[3] CALLAN J P. Passage-level evidence in document retrieval [A]. **Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval** [C]. Dublin: [s n], 1994. 302-310.

[4] AGIRRE E, RIGAU G. A proposal for word sense disambiguation using conceptual distance [A]. **International Conference on Recent Advances in Natural Language Processing** [C]. Velingrad: [s n], 1995. 258-264.

[5] 车万翔，刘　挺，秦　兵，等. 面向双语句对检索的汉语句子相似度计算[A]. 全国第七届计算语言学联合学术会议[C]. 北京：清华大学出版社，2003. 81-88.

[6] 刘　群，李素建. 基于《知网》的词汇语义相似度计算[A]. 第三届汉语词汇语义学研讨会论文集[C]. 台北：[s n], 2002. 59-76.

[7] 李素建. 基于语义计算的语句相关度研究[J]. 计算机工程与应用，2002(7)：75-78.

[8] 董振东，董　强. 知网[EB/OL]. http://www.keenage.com., 2003-07-12.

[9] ZHANG Hua-ping, Yu Hong-kui, Xiong De-yi, *et al*. HHMM-based Chinese lexical analyzer ICTCLAS [A]. **41st Annual Meeting of the Association for Computational Linguistics** [C]. Sapporo: [s n], 2003. 184-187.

[10] 陆善采. 实用汉语语义学[M]. 上海：学林出版社，1993.

[11] 胡壮麟. 系统功能语法概论[M]. 长沙：湖南教育出版社，1989.

[12] 俞士汶，段慧明，田剪秋. 机械文摘自动评测的原理及实现[A]. 吴泉源. 智能计算机接口与应用进展——第三届中国计算机智能接口与智能应用学术会议论文集[C]. 北京：电子工业出版社，1998. 230-233.

Similarity algorithm of text based on semantic understanding

JIN Bo[1,2], SHI Yan-jun[1,2], TENG Hong-fei*[1]

(1. School of Mech. Eng., Dalian Univ. of Technol., Dalian 116024, China;
2. Dept. of Comput. Sci. and Eng., Dalian Univ. of Technol., Dalian 116024, China)

Abstract: Text similarity counting has been widely used in several fields, for example, the field of copy detection and the field of information retrieval, etc.. With the study of text similarity computing and semantic understanding, the textural similarity counting can be expanded to paragraph similarity counting, then the paragraph similarity counting can be expanded to article similarity counting. A new set of textural (including words, sentences and paragraphs) similarity algorithm is given. This algorithm can count out the similarity rate of two texts. Compared with other methods of similarity computing, the algorithm can raise the recall rate.

Key words: HowNet; semantic; text similarity; copy detection; information retrieval