

文章编号:1001-9081(2013)08-2276-04

doi:10.11772/j.issn.1001-9081.2013.08.2276

改进的基于《知网》的词汇语义相似度计算

朱征宇^{1,2}, 孙俊华^{1,2*}

(1. 重庆大学 计算机学院, 重庆 400044; 2. 软件工程重庆市重点实验室, 重庆 400044)

(*通信作者电子邮箱 boyjunhuagirl@163.com)

摘要:针对当前基于《知网》的词汇语义相似度计算方法没有充分考虑知识库描述语言对概念描述的线性特征的情况,提出一种改进的词汇语义相似度计算方法。首先,充分考虑概念描述式中各义原之间的线性关系,提出一种位置相关的权重分配策略;然后,将所提出的策略结合二部图最大权匹配进行概念相似度计算。实验结果表明,采用改进方法得到的聚类结果F值较对比方法平均提高了5%,从而验证了改进方法的合理性和有效性。

关键词:知网;义原;概念;权重;语义相似度

中图分类号: TP391.1 **文献标志码:**A

Improved vocabulary semantic similarity calculation based on HowNet

ZHU Zhengyu^{1,2}, SUN Junhua^{1,2*}

(1. College of Computer Science, Chongqing University, Chongqing 400044, China;
2. Chongqing Key Laboratory of Software Engineering, Chongqing 400044, China)

Abstract: The present HowNet-based vocabulary semantic similarity calculation method fails to give due attention to the linear feature of conceptual description in knowledge database mark-up language. To resolve this shortcoming, an improved vocabulary semantic similarity calculation method was proposed. Firstly, fully considering the linear relationship between the sememes in the conceptual description formula, a position-related weight distribution strategy was proposed. Then concept similarity was calculated by combining the strategy above with bigraph maximum weight matching. The experimental results show that, compared with the contrast method, the F-measure of text clustering using improved method increases by 5% on average, thus verifying the rationality and validity of the improved method.

Key words: HowNet; sememe; concept; weight; semantic similarity

0 引言

词汇语义相似度计算在文本聚类^[1]、信息检索、机器翻译等领域有着广泛应用。当前词汇语义相似度计算方法大致可分为两类:一类利用大规模语料库进行统计,依据词汇上下文信息的概率分布进行计算;另一类基于某种世界知识来计算,通常是基于某个知识完备的语义词典中的层次结构关系进行计算,例如荀恩东等^[2]采用WordNet进行英语词语间的相似度计算,刘群等^[3]提出基于《知网》的词语相似度计算等。基于语料库的方法比较精准,但计算比较复杂并且结果容易受训练数据的噪声影响;而基于语义词典的方法简单有效,比较直观,但对词典依赖性较大,且易受人主观意识影响,当前词汇语义相似度计算大多采用该方法。

《知网》是一个以汉语和英语的词语所代表的概念(义项)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^[4]。在《知网》中,词汇对应于若干概念,而概念是以义原为基础通过知识库描述语言进行定义的,即概念的义项表达式,义原又通过多种关系进行描述,如上下位关系等,其具体含义可查阅相关文献[3-4]。目前大多数学者基于《知网》的词汇语义相似度计算思想是整体相似度可由部分相似度加权平均进行计算。其中比较有代表性的方法如刘群等^[3]首先提出的仅考虑义原之间距离因素的词汇语义相似度计算方法,李峰等^[5]在前者的基础上所提出的考虑义原深度因素计算方法,Dai等^[6]提

出的基于《知网》的中英文词间相似度算法,刘青磊等^[7]提出的基于信息论的计算方法,王小林等^[8]提出的变系数计算方法等。然而文献[9]指出知识库描述语言对概念的描述具有线性关系,但上述计算方法都没有充分考虑该线性关系,使得词汇相似度计算结果不够合理。

本文在深入研究和分析知识库描述语言的结构特征以及现有计算方法基础上,提出一种充分考虑知识库描述语言线性描述特征的词汇语义相似度计算方法,使得词汇间的相似度计算结果更为合理。特别指出,若非特殊说明本文所述的《知网》都是指《知网》2000版。

1 词汇语义相似度计算

当前基于《知网》的词汇语义相似度计算大致可以分为三个过程:义原相似度计算、概念相似度计算和词汇语义相似度计算。各具体的计算过程如下所述。

1.1 义原相似度计算

《知网》中义原间的相似度计算主要利用义原层次体系中义原之间的各种关系进行计算,例如刘群等^[3,5-6]利用义原之间的上下位关系进行计算等。本文选取当前计算方法中两种比较有代表性的义原相似度计算公式进行讨论。

刘群等^[3]提出的义原相似度计算公式如下:

$$\text{Sim}(S_1, S_2) = \frac{a}{\text{distance}(S_1, S_2) + a} \quad (1)$$

其中: S_1 和 S_2 表示两个义原; $\text{distance}(S_1, S_2)$ 表示 S_1, S_2 在义

收稿日期:2013-01-31;修回日期:2013-03-12。 基金项目:国家科技支撑计划项目(2011BAH25B04)。

作者简介:朱征宇(1959-),男,重庆人,教授,博士,CCF 高级会员,主要研究方向:Web 智能检索、智能交通、数据库; 孙俊华(1987-),男,河南驻马店人,硕士研究生,主要研究方向:数据挖掘、文本分析、自然语言处理。

原层次树中的路径长度,当 S_1, S_2 不在同一棵树中时取一个较大常数值; α 为可调节参数。

李峰等^[5]提出的义原相似度计算公式如下:

$$\text{Sim}(S_1, S_2) = \frac{\alpha \times \min(\text{depth}_{S_1}, \text{depth}_{S_2})}{\alpha \times \min(\text{depth}_{S_1}, \text{depth}_{S_2}) + \text{distance}(S_1, S_2)} \quad (2)$$

其中: S_1, S_2 表示两个义原; $\text{depth}_{S_1}, \text{depth}_{S_2}$ 分别为 S_1, S_2 所在层次树中的深度; $\text{distance}(S_1, S_2)$ 为义原在层次树中的路径长度,当 S_1, S_2 不在同一棵树中时取一个较大的常数值; α 为可调节参数。

分析以上两种公式可以看出,式(1)只考虑了义原层次体系中义原之间的距离因素对义原相似度的影响。例如,义原{“虫”,“鱼”}与{“物质”,“精神”}分别在义原层次树中的路径距离相等,则它们的相似度相等。但在人们直观理解上,显然前者之间的相似性应高于后者。所以只考虑义原间的距离因素往往计算得到的结果过于粗糙,不够合理。式(2)在式(1)的基础上充分考虑了义原在义原层次树中的深度因素对义原相似度的影响。同样以上述例子进行说明,前一组义原在义原层次树中的深度都为6,而后一组的深度都为2,则依据式(2)进行相似度计算,前者之间的相似度值大于后者,更加符合人们的主观理解,所以采用式(2)计算得到的结果较式(1)更为合理。

1.2 概念相似度计算

当前对于概念相似度计算也有许多方法,按照权重系数设置方式不同大致可以分为以下两类:

1)基于固定权重的概念相似度计算方法,比较有代表性的如刘群等^[3,5]提出的计算方法。其大致计算过程如下:首先对概念的义项表达式按描述式的形式不同划分为4类或2类;然后计算各相同类型描述式集合之间的相似度,记为 $\text{Sim}_i(C_1, C_2)$;最后赋予各集合间相似度值相应的权重 β_i ,加权求和得到概念相似度 $\text{Sim}(C_1, C_2)$ 。其具体的计算公式如下:

$$\text{Sim}(C_1, C_2) = \sum_{i=1}^n \beta_i \text{Sim}_i(C_1, C_2) \quad (3)$$

其中: n 的取值一般为2或4, $\beta_1, \beta_2, \dots, \beta_n$ 为可调参数且 $\beta_1 + \beta_2 + \dots + \beta_n = 1$,若 $n = 4$ 时,其权重参数关系须满足: $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$ 。参数 $\beta_1, \beta_2, \dots, \beta_n$ 的值确定后适用于计算任意概念间的相似度值。

2)基于变系数权重的概念相似度计算方法,比较有代表性的如王小林等^[8]提出的计算方法。其计算各部分集合间相似度的过程与基于固定权重计算方法的过程基本一致,但为各部分相似度值赋予权重系数大小时依赖于划分后各类型义原描述式集合中所包含的元素个数。其具体的计算公式如下:

$$\text{Sim}(C_1, C_2) = \sum_{i=1}^4 \beta_i \text{Sim}_i(C_1, C_2) \quad (4)$$

$$\beta_i = k_i / (m + n) \quad (5)$$

其中: k_i 为概念 C_1 和 C_2 的义项表达式划分后第 i 类义原描述式集合中的元素个数之和; m, n 分别为 C_1, C_2 的义项表达式中所包含的描述式个数;权重系数 β_i 通过式(5)进行计算, β_i 的取值与义项表达式中所包含的描述式数量相关。

1.3 词汇语义相似度计算

词汇语义相似度是个主观性较强的概念,因具体应用领域不同,其具体含义不尽相同。文献[3]在以《知网》为基础的词汇语义相似度计算研究中给出了形式化定义:词汇语义

相似度就是两个词汇在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。词汇一般可分为实词和虚词,虚词词汇之间的相似度计算比较简单,可参考文献[3]中的方法。实词词汇间的语义相似度计算大多采用以下方法。假定在已经计算得到词汇间所对应的任意概念间相似度的前提条件下,若词汇 W_1 对应有 n 个概念 $C_{11}, C_{12}, \dots, C_{1n}$,词汇 W_2 有 m 个概念 $C_{21}, C_{22}, \dots, C_{2m}$,则 W_1, W_2 之间的相似度如下:

$$\text{Sim}(W_1, W_2) = \max(\text{Sim}_{ij}(C_{1i}, C_{2j})) \quad (6)$$

其中: $i = 1, 2, \dots, n, j = 1, 2, \dots, m, \text{Sim}(W_1, W_2)$ 为词汇 W_1 与 W_2 之间的相似度值, $\text{Sim}_{ij}(C_{1i}, C_{2j})$ 为概念 C_{1i} 与 C_{2j} 的概念相似度值。

2 改进的词汇语义相似度计算

文献[9]指出知识库描述语言的描述方式具有线性顺序,它对义原的顺序是有规定的,如果破坏了这种顺序,就会导致意义上的错误。但当前的计算方法由于没有充分考虑义项表达式中义原描述式之间的顺序关系,使得概念的相似度计算不够合理,进而导致词汇间的语义相似度计算结果与人们的主观理解不一致。本文主要针对该缺点对现有的概念相似度计算方法进行改进,并结合1.1节和1.3节中所提出的方法进行词汇间的语义相似度计算。由于虚词所对应概念的定义方式比较简单,在此就不作赘述。接下来,本文就如何在概念相似度计算过程中充分考虑知识库描述语言对概念描述的线性关系进行讨论。

2.1 位置相关的义原描述式权重分配

分析1.2节中介绍的概念相似度计算方法可知,采用固定权重分配方案的方法需要在实际应用中设定权重系数,结果的合理性与权重系数有很大关系,但是往往很难找到一组适用于计算所有概念间相似度的权重系数,容易受主观因素影响。同时,权重系数固定使得各部分相似度值在合成得到整体概念相似度时所起的作用大小固定且划分后各集合中的描述式之间没有顺序关系,忽略了知识库描述语言的对其描述的线性关系,使得采用该策略的概念相似度计算具有其不合理性。与此相比较,王小林等^[8]采用与各描述式集合中元素数量相关的动态权重分配方案的方法,依据义项表达式不同,动态确定各部分在整体相似度中的作用大小;但是在确定权重系数时,完全没有考虑描述式之间的线性关系,所以采用该权重分配策略的计算方法也是不够合理的。下面举例说明上述方法的不合理性。

例如,概念“心脏”“CPU”“解码”“计算中心”在《知网》中的定义如下:

CPU:DEF = {part|部件, % computer|电脑, heart|心}

心脏:DEF = {part 部件, % animalHuman|动物, heart|心}

计算中心:DEF = {InstitutePlace|场所, #computer|电脑, #software|软件}

解码:DEF = {translate|翻译, #computer|电脑, #software|软件}

依据文献[3]的方法把概念的义项表达式划分后,分别按照文献[3]与文献[8]的方法进行概念的相似度计算,结果为表1中的 Sim_1 和 Sim_2 。

对比 Sim_1 与 Sim_2 的1,2行可以看出,显然 Sim_1 值过大。因为在计算它们之间的相似度值时,划分后的符号义原描述式集合相似度值按文献[3]的方法被赋予相对较低的权重系数,违背了其各自在概意义项表达式中的顺序才使得计算结

果不合理。

分析 Sim_1 与 Sim_2 的 3 行中的相似度值可以看出, 显然 Sim_2 的值过大, 由于依据文献[8]的方法赋予符号义原描述式集合之间的相似度值较大的权重, 才使得计算得到的概念相似度值不符合人们的主观理解。

表 1 概念相似度结果

概念 1	概念 2	Sim_1	Sim_2
心脏	CPU	0.897 368	0.670
案情	心脏	0.676 000	0.417
解码	计算中心	0.074 074	0.670

综上所述, 在概念相似度计算过程中, 若割裂了义项表达式中各描述式之间的线性关系, 则计算得到的结果是不合理的。

本文通过深入研究和分析知识库描述语言的结构特征, 认为知识库描述语言对概念的描述具有以下特点: 1) 义原描述式形式在位置关系上除义项表达式的首位置特定为基本义原描述式外, 其他位置以何种形式的描述式对概念进行描述与位置不相关。2) 义项表达式中的各义原描述式所能描述的概念含义抽象程度与其在表达式中的位置相关。即若义原描述式相对于表达式中的位置偏左, 描述式中的义原往往分布在义原层次体系的较高层, 其所代表的含义比较抽象, 更能代表概念的本质属性; 而位置偏右的描述式中, 义原一般位于义原层次体系中的较低层, 所能代表的含义比较具体, 能够描述不同概念间的细微差异。

基于以上描述特点, 本文提出一种位置相关的权重分配策略用于概念相似度计算。总体的分配思想为: 各类型集合间相似度权重系数依赖于集合中的各描述式在表达式中的位置。而描述式权重按如下原则进行分配: 表达式中位置偏左的描述式应该赋予较高的权重, 而位置偏右的义原描述式之间的相似度应该赋予较低权重。假设概念 C 的义项表达式中有 n 个义原描述式, 按照其在表达式中的顺序依次为 S_1, S_2, \dots, S_n , 则每个描述式的权重由以下公式计算所得:

$$weight(S_i) = (n + 1 - position(S_i))x \quad (7)$$

$$\sum_{i=1}^n weight(S_i) = 1 \quad (8)$$

其中: $weight(S_i)$ 为义原描述式 S_i 的权重, $position(S_i)$ 为 S_i 在表达式中的位置, i 的取值范围为 $[1, n]$ 。式(7)使得表达式中相邻描述式的权重差值为 x ; 式(8)保证概念相似度范为 $0 \sim 1$, 且完全相同的两个概念相似度为 $1^{[9]}$ 。

2.2 概念的语义相似度计算

在概念相似度计算介绍之前, 假定已经计算得到了义原之间的相似度。首先, 按照 2.1 节所述的权重分配策略为概念的义项表达式中各描述式分配权重, 并依照文献[3]提出的描述式类型划分方法把各描述式按形式不同划分为四个集合: 独立义原描述式集合、其他基本义原描述式集合、关系义原描述式集合和符号义原描述式集合。然后, 分别计算相同类型的描述式集合之间的相似度值。其原因在于一般只有相同类型的义原描述式集合间进行相似度计算才有意义^[3]。最后, 对得到的各部分相似度进行求和得到概念间的相似度值。接下来, 主要就集合间相似度计算方法进行讨论。

1) 独立义原描述式集合相似度: 概念的独立义原描述式有且仅有一个为基本义原, 故可先直接由义原相似度计算公式计算得到其基本义原间的相似度值, 然后乘以权重系数作

为该部分集合间的相似度值, 记为 $Sim_1(C_1, C_2)$ 。由于各义原描述式都有一个权重系数, 关于权重系数如何选取将在后续讨论中进行说明。

其他基本义原描述式集合相似度: 其他基本义原描述式集合中一般包含若干个基本义原描述式且形式为基本义原, 关于计算该集合之间的相似度值本文采用图论中的二部图最大权匹配算法。假设概念 C_1 和 C_2 所对应的其他基本义原描述式集合分别为 C_{11} 和 C_{21} , 其集合描述形式为:

$$C_{11} = \{(S_{1,1}, w_{11}), (S_{1,2}, w_{12}), \dots, (S_{1,m}, w_{1m})\}$$

$$C_{21} = \{(S_{2,1}, w_{21}), (S_{2,2}, w_{22}), \dots, (S_{2,n}, w_{2n})\}$$

其中: $S_{i,j}$ 为表达式中其他基本义原描述式; w_{ij} 为 $S_{i,j}$ 的权重; m, n 分别为概念 C_1, C_2 的义项表达式中所包含的其他基本义原描述式个数。

首先依据集合中的元素构造一个完备二部图, 假设把集合 C_{11}, C_{21} 中的描述式分别作为二部图中的 X 部和 Y 部各顶点, 分别计算 X 部与 Y 部各顶点间的相似度值作为顶点间边的权值, 可直接由义原相似度公式进行计算。若集合 C_{11} 与 C_{21} 中的元素个数不相等, 则应当在元素较少的一侧 (X 部或者 Y 部) 虚拟若干空节点以满足完备二部图的基本条件: $|X| = |Y|$ 。假设 $|C_{11}| > |C_{21}|$ 时, 则所构成的二部图如图 1 所示。

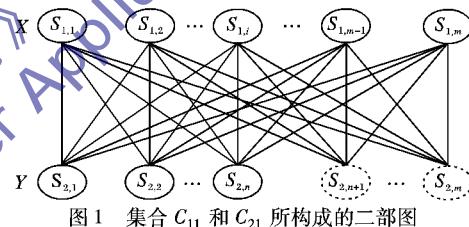


图 1 集合 C_{11} 和 C_{21} 所构成的二部图

其中节点 $S_{2,n+1}, \dots, S_{2,m}$ 为 Y 部中虚拟的若干空节点 (如图 1 中的虚线节点所示), 且图中任意非空节点与空节点的相似度为一个固定的较小值。

然后, 利用 Kuhn-Munkres 算法^[10] 得到二部图的最大权匹配 $M = \{e_{1x}, e_{2y}, \dots, e_{ij}, \dots\}$, 其中 e_{ij} 为 X 部顶点 $S_{1,i}$ 与 Y 部顶点 $S_{2,j}$ 的匹配边, 该边权重为 $S_{1,i}$ 与 $S_{2,j}$ 的相似度值。

最后, 将匹配 M 与顶点的权重加权得到集合之间的相似度值。其计算公式如下:

$$Sim(C_{11}, C_{21}) = (w_1 * W_{1x} + w_2 * W_{2x} + \dots + w_t * W_{tx}) / t \quad (9)$$

其中: w_1 为集合 C_{11} 中 $S_{1,1}$ 的权重或 C_{21} 中 $S_{2,1}$ 的权重, t 为集合 C_{11} 或 C_{21} 中的元素个数, 关于权重的选择以及 t 值的确定随后将对该问题进行讨论; $Sim(C_{11}, C_{21})$ 为概念 C_1 与 C_2 的其他基本义原描述式集合之间的相似度值, 记为 $Sim_2(C_1, C_2)$ 。

2) 关系义原描述式集合相似度: 关系义原描述式集合相似度与其他基本义原描述式集合相似度计算过程类似, 其不同点在于关系义原描述式之间的相似度计算, 关系义原描述式由关系义原与基本义原组成, 计算两个关系义原描述式间的相似度必须在它们的关系义原相同时才有意义^[3], 即当关系义原相同时直接计算描述式的基本义原之间的相似度值作为关系义原描述式之间的相似度值; 否则, 其相似度值设定为一个较小固定常数。

3) 符号义原描述式集合相似度: 符号义原描述式集合相似度与基本义原描述式集合相似度计算过程类似, 不同点在于符号义原描述式间的相似度计算。符号义原描述式由关系符号和基本义原构成, 在计算描述式之间的相似度值时, 只有

在关系符号一致的情况下计算才有意义^[3],即在关系符号一致时直接计算描述式中基本义原之间的相似度值作为它们之间的相似度值;否则,其相似度值设定为一个较小固定常数。

关于上述计算过程的权重选择,本文采取下述策略:假设概念 C_1 和 C_2 的义项表达式分别包含的义原描述式个数为 m 和 n ,且 $m > n$,则计算过程中选择概念的义原描述式个数较多一侧的各顶点权重,即在计算独立义原描述式集合相似度时,选择 C_1 中的描述式权重进行计算;而在对其他集合进行计算时, w_1, w_2, \dots, w_t 的权重选择 C_{11} 中各描述式的权重,且 t 取值为 m 。之所以采取这样的策略是因为这样可以避免计算得到的概念间相似度值过于相似且能保证得到的值介于 0 到 1 之间。下面以示例说明该策略的合理性。

例 1 概念“男亲属”和“男人”在知网中的定义如下:

男亲属:DEF = {human|人, male|男, family|家}

男人:DEF = {human|人, male|男}

若在计算集合相似度时,选择概念的义原描述式数较少一侧,即上例中选择概念“男人”的各描述式权重进行计算,则不能体现出两个概念在描述式“family|家”上的细微差异性,使得相似度值过大,所以本文采用上述策略进行各集合间相似度计算。

最后,基于整体相似度可由部分相似度加权合成计算的思想,概念 C_1 和 C_2 之间的相似度值如下:

$$\text{Sim}(C_1, C_2) = \sum_{i=1}^4 \text{Sim}_i(C_1, C_2) \quad (10)$$

其中: $\text{Sim}(C_1, C_2)$ 为概念 C_1 与 C_2 之间的相似度值, $\text{Sim}_i(C_1, C_2)$ 为各部分集合间的相似度值。在得到词语间的任意概念间相似度值后,可根据式(6)计算得到词汇间的语义相似度。

3 实验结果与分析

为了验证上述方法的有效性,本文从两个方面对其进行对比实验验证。一方面,采用本文与文献[8]的方法进行词汇间语义相似度计算并将实验结果进行对比分析;另外,将本文和文献[8]的方法应用于文本聚类,并对聚类结果进行对比分析,从而间接地验证方法有效性。

3.1 词汇语义相似度实验

本文选取了若干组具有代表性的词汇进行词汇语义相似度计算的对比实验验证。在计算过程中,由于文献[8]及本文主要就概念相似度计算方法作出改进,所以本文将这两种概念相似度计算方法分别结合 1.1 节和 1.3 节中所介绍的已有方法进行对比实验分析。表 2 中的 Sim_1 和 Sim_2 为文献[8]结合文献[3]和文献[5]的义原相似度计算方法得到的结果;表 2 中的 Sim_3 和 Sim_4 为本文方法分别结合相同的两种义原相似度计算方法得到的结果。其中,实验中的词汇语义相似度计算方法均采用 1.3 节中所介绍的方法。

对比分析 $\text{Sim}_1, \text{Sim}_3$ 两列,从 1 ~ 6 行可以看出这两列结果都相对比较符合人们的主观理解,但是对比分析这两列中 7 ~ 12 行的结果可以看出,显然 Sim_3 比 Sim_1 更合理。比如“厂部”和“厂规”是关于“厂”的两个不同概念,而 Sim_1 的结果显然不是很合理。同样对比分析 Sim_4 与 Sim_2 两列中 7 ~ 12 行的结果,显然 Sim_4 比 Sim_2 更合理。其原因在于采用文献[8]的方法进行概念相似度计算,由于没有充分考虑知识库描述语言对概念描述的线性特征才使得计算得到的词汇语义相似度值不够合理;而本文方法充分考虑了知识库描述语言对概念描述的线性特征,能够更为准确地计算出概念相似度,使最终得到的词汇语义相似度更能合理反映词汇间的相似性。所以采用本

文方法能够更为合理地计算出词汇间的语义相似度值。

表 2 词汇语义相似度结果

$Word_1$	$Word_2$	Sim_1	Sim_2	Sim_3	Sim_4
男人	经理	0.416	0.384	0.476	0.465
男人	和尚	0.815	0.778	0.815	0.778
男人	父亲	1.000	1.000	1.000	1.000
心脏	案情	0.417	0.385	0.529	0.507
心脏	CPU	0.670	0.670	0.700	0.700
CPU	案情	0.417	0.385	0.529	0.507
厂	厂长	0.380	0.388	0.346	0.357
厂部	厂规	0.571	0.568	0.427	0.422
返防	换岗	0.387	0.397	0.275	0.291
接防	换岗	0.722	0.733	0.583	0.600
返防	接防	0.399	0.458	0.293	0.382
军籍	军官	0.230	0.230	0.106	0.106

3.2 聚类实验

由于词汇间的语义相似度计算结果合理性评价往往采用人工方法进行判别,容易受人的主观因素影响。所以本文为了更为客观地验证本文方法的有效性,将文献[8]与本文的概念相似度计算方法应用于词汇语义相似度计算,并以此为基础进行基于语义的文本聚类对比实验。实验语料采用 CNLP Platform 中一个中文文本语料库^[11]的子集,共 300 篇文档。分别从语料库中选取不同主题中的文档进行 3 组聚类分析实验,其中第一组为环境(11 篇)、经济(10 篇)、环境(13 篇)、艺术(12 篇),第二组为教育、交通、环境和艺术各 20 篇,第三组为环境、艺术、教育和医药各 20 篇。具体实验过程如下:

1) 对每组实验文档进行分词、去停用词等一系列预处理后建立每个文本的特征向量。

2) 采用文献[12]提出的文本相似度计算方法计算任意文档之间的相似度值。在该过程中,分别采用了文献[8]和本文的方法计算文本间的词汇语义相似度值。

3) 采用 K 中心点算法 (Partitioning Around Mediod, PAM)^[13] 对文本特征向量进行聚类。

3 组实验分别采用不同的概念相似度计算方法进行聚类,每组实验结果分别采用准确率 (Precision)、召回率 (Recall) 和 F 值 (F-measure)^[14] 的均值进行评价分析,结果如表 3 所示。

表 3 文本聚类实验结果

实验	算法	Precision/%	Recall/%	F-measure
1	文献[8]方法	67.56	64.19	0.6461
	本文方法	70.27	66.31	0.6761
2	文献[8]方法	77.00	75.00	0.7451
	本文方法	79.82	78.75	0.7917
3	文献[8]方法	78.71	75.00	0.7509
	本文方法	83.99	82.50	0.8227

由表 3 中的每组对比实验可见,采用本文方法比文献[8]的方法得到的聚类结果 F 值均有不同程度提高。究其原因,由于本文相比文献[8]的方法充分考虑了知识库描述语言的顺序性特征,使得采用本文的概念相似度计算方法计算得到的词汇语义相似度更具合理性,所以计算得到的文本相似度能够更准确地反映文本间的语义相似性,从而能够有效地改善聚类的质量。综上所述,采用本文提出的方法能够更为合理地计算出词汇间的语义相似度。

(下转第 2288 页)

参考文献:

- [1] 王志文, 郭戈. 移动机器人导航技术现状与展望[J]. 机器人, 2003, 25(5): 470–474.
- [2] FLOREANO D, MONDADA F. Evolutionary neuro-controller for autonomous mobile robots [J]. Neural Networks, 1998, 11(7/8): 1461–1478.
- [3] YEN J, PFLUGER N. A fuzzy logic based extension to Payton and Rosenblatt's command fusion method for mobile robot navigation [J]. IEEE Transactions on Systems, Man and Cybernetics, 1995, 25(6): 971–978.
- [4] KERMICHE S, SAIDI M L, ABBASSI H A. Gradient descent adjusting Takagi-Sugeno controller for a navigation of robot manipulator [J]. Journal of Engineering and Applied Science, 2006, 1(1): 24–29.
- [5] JOO ER M, CHANG D. Obstacle avoidance of a mobile robot using hybrid learning approach [J]. IEEE Transactions on Industrial Electronics, 2005, 52(3): 898–905.
- [6] JOO ER M, ZHOU Y. Automatic generation of fuzzy inference systems via unsupervised learning [J]. Neural Networks, 2008, 21(10): 1556–1566.
- [7] BOUBERTAKH H, TADJINE M, GLORENNEC P-Y. A new mobile robot navigation method using fuzzy logic and a modified Q-learning algorithm [J]. Journal of Intelligent & Fuzzy Systems, 2010, 21(1/2): 113–119.
- [8] SUTTON R S, BARTO A G. Reinforcement learning [M]. London: MIT Press, 1998: 1–12.
- [9] SU S F, Hsieh S H. Embedding fuzzy mechanisms and knowledge in box-type reinforcement learning controllers [J]. IEEE Transactions on Systems, Man and Cybernetics: Part B, 2002, 32(5): 645–653.
- [10] ZEYBEK Z. Role of adaptive heuristic criticism in cascade temperature control of an industrial tubular furnace [J]. Applied Thermal Engineering, 2006, 26(2/3): 152–160.
- [11] MUCIENTES M, ALCALA-FDEZ J, ALCALA R, et al. A case study for learning behaviors in mobile robotics by evolutionary fuzzy system [J]. Expert Systems with Application, 2010, 37(2): 1471–1493.
- [12] DESOUKY S F, SCHWARTZ H M. Self-learning fuzzy logic controllers for pursuit-evasion differential games [J]. Robotics and Autonomous Systems, 2011, 59(1): 22–33.
- [13] KNUDSON M, TUMER K. Adaptive navigation for autonomous robots [J]. Robotics and Autonomous Systems, 2011, 59(6): 410–420.
- [14] TOURETZKY D S, SAKSIDA L M. Operant conditioning in Skinerbots [J]. Adaptive Behavior, 1997, 5(3/4): 219–247.
- [15] GUTNISKY D A, ZANUTTO B S. Learning obstacle avoidance with an operant behavior model [J]. Artificial Life, 2004, 10(1): 65–81.
- [16] SAKSIDA L M, RAYMOND S M, TOURETZKY D S. Shaping robot behavior using principles from instrumental conditioning [J]. Robotics and Autonomous Systems, 1998, 22(3/4): 231–249.
- [17] GAUDIANO P, CHANG C. Adaptive obstacle avoidance with a neural network for operant conditioning: Experiments with real robots [C]// CIRA 97: Proceedings of 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Piscataway: IEEE, 1997: 13–18.
- [18] ITOH K, MIWA H, MATSUMOTO M, et al. Behavior model of humanoid robots based on operant conditioning [C]// Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robots. Piscataway: IEEE, 2005: 220–225.
- [19] 蔡建羨, 阮晓钢. OCPA 仿生自主学习系统及在机器人姿态平衡控制上的应用[J]. 模式识别与人工智能, 2011, 24(1): 138–146.
- [20] 段勇, 崔宝侠, 徐心如. 进化强化学习及其在机器人路径跟踪中的应用[J]. 控制与决策, 2009, 24(4): 532–536.

(上接第 2279 页)

4 结语

本文在充分考虑知识库描述语言线性特征前提下, 提出了一种有效的义原描述式权重分配方案, 并结合二部图的最大权匹配算法以及现有方法进行词汇的语义相似度计算。实验结果表明, 采用本文方法计算得到的词汇语义相似度能够更合理地体现词语间语义上的差异性, 更加符合人们的主观理解。接下来, 将深入研究《知网》对词汇的描述特点, 从而更进一步改善词汇语义相似度计算的合理性。

参考文献:

- [1] ZHU Z Y, DONG S J, YU C L, et al. A text hybrid clustering algorithm based on HowNet semantics [C]// ICAMCS 2011: 2011 International Conference on Advanced Materials and Computer Science. Zurich: Trans Tech Publications Ltd, 2011: 474–476.
- [2] 荀恩东, 颜伟. 基于语义网计算英语词语相似度[J]. 情报学报, 2006, 25(1): 43–48.
- [3] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[C]// 第三届汉语词汇语义学研讨会论文集. 台北: [出版者不详], 2002: 59–76.
- [4] 董强, 董振东. 知网简介[EB/OL].[2013-01-29]. <http://www.keenage.com/>.
- [5] 李峰, 李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报, 2007, 21(3): 99–105.
- [6] DAI L L, LIU B, XIA Y N, et al. Measuring semantic similarity between words using HowNet [C]// ICCSIT'08: 2008 International Conference on Computer Science and Information Technology. Washington, DC: IEEE Computer Society, 2008: 601–605.
- [7] 刘青磊, 顾晓峰. 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.
- [8] 王小林, 王义. 改进的基于知网的词语相似度算法[J]. 计算机应用, 2011, 31(11): 3075–3077.
- [9] 郝长伶, 董强. 知网知识库描述语言[C]// 全国第七届计算语言学联合学术会议论文集. 北京: 清华大学出版社, 2003: 371–377.
- [10] 龚劬. 图论与网络最优化算法[M]. 重庆: 重庆大学出版社, 2009: 86–95.
- [11] 李荣陆. 中文本分类语料 2003 [DB/OL].[2013-01-29]. <http://www.nlpirc.org/download/tc-corpus-answer.rar>.
- [12] 余刚, 裴仰军, 朱征宇, 等. 基于词汇语义计算的文本相似度研究[J]. 计算机工程与设计, 2006, 27(2): 241–244.
- [13] HAN J W, KAMBER M. 数据挖掘: 概念与技术[M]. 范明, 译. 2 版. 北京: 机械工业出版社, 2007: 263–266.
- [14] LARSEN B, AONE C. Fast and effective text mining using linear-time document clustering [C]// Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1999: 16–22.