CN43-1258/TP
ISSN 1007-130X

计算机工程与科学
COMPUTER ENGINEERING & SCIENCE

2012 年第 34 卷第 2 期
Vol. 34, No. 2, 2012

文章编号:1007-130X(2012)02-0172-04

一种基于知网的句子相似度计算方法*

A Method of Sentence Similarity Computing Based on Hownet

程传鹏,吴志刚

CHENG Chuan-peng, WU Zhi-gang

(中原工学院计算机学院,河南 郑州 450007)

(School of Computer Science, Zhongyuan Institute of Technology, Zhengzhou 450007, China)

摘　要:句子相似度是衡量文档相似度的基础,在自然语言处理领域中有着非常重要的作用。目前的句子相似度计算方法忽略了句子的结构对相似度的影响。本文在分析已有研究工作的基础上,提出了一种改进的句子相似度计算方法。依据知网对"实体概念"的描述,构造出义原的语义层次树,由各个义原在树中的相对位置,计算出义原之间的相似度。对三种义原加权求和得到词语之间的语义相似度。综合句子的表层相似度和句子的词语语义以及词语的相对位置关系,得到句子的整体相似度。实验表明,在同等的测试条件下,本文所提出的句子相似度计算方法在相似度比较上更符合人的直观感觉。

Abstract: Sentence similarity is the basis of document similarity, and sentence similarity computing plays an important role in the field of natural language processing. The current methods of sentence similarity computing neglect the influence of sentence structure. On the basis of the interrelated research, this paper proposes an improved method of similarity comparison. The semantic tree of sememe is constructed according to the description of entity conception in the Hownet, the semantic similarity of sememe is computed based on the relative positions in the sememe tree. Calculating of sentence similarity is based on surface similarity and semantic similarity. Under the same test conditions, the experiments show that the proposed method is much closer to the people's comprehension to the meanings of the sentences.

关键词:句子相似度;知网;表层相似度;语义偏移量

Key words: sentence similarity; hownet; surface similarity; semantic offset similarity

doi: 10.3969/j.issn.1007-130X.2012.02.031

中图分类号:TP391.1　　**文献标识码**:A

1　引言

句子相似度的比较作为中文信息处理研究领域中一个关键的问题,一直以来都是人们研究的热点和难点。句子相似度计算在自动问答、双语例句检索、文档文摘等领域都有很重要的应用价值。目前,句子相似度计算的方法主要有两种:一种是基于词语共现的统计方法,例如,北大计算语言所提出的一种句子相似度计算公式:$2c/(m+n)$(其中m、n分别表示两个句子的词数,c是两个句子中相同词的数目[1]。);另外一种是基于词汇的词法和语义信息的分析[2,3]。第一种方法简单、高效,但忽视了词汇的词法和语义信息,因此在计算句子整体相似度上不够准确;第二种方法虽然考虑到了词语的语义信息,但忽略了词语之间的相对位置信息。

* 收稿日期:2011-07-23;修订日期:2011-10-08
基金项目:河南省教育厅自然科学资助项目(2008B520046)
通讯地址:450007 河南省郑州市中原工学院计算机学院
Address: School of Computer Science, Zhongyuan Institute of Technology, Zhengzhou, Henan 450007, P. R. China

本文在已有研究工作的基础上,综合考虑了两种方法的优缺点,提出了一种新的句子相似度计算的方法。文章首先介绍了基于《知网》的词语相似度计算方法;接着介绍了句子相似度的计算方法;最后对该方法进行了实验和评价。

2 基于《知网》的词语相似度计算

句子的相似度主要取决于句中词语的相似度。本文采用了《知网》来计算词语的相似度。《知网》[4]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[5]。在《知网》中,所有的词语通过"概念"来描述,每一个词可以表达为几个概念,每一个"概念"由"义原"来描述,具体定义如表 1 所示。

《知网》中义原之间有 8 种关系,分别是"上下位"关系、"同义"关系、"反义"关系、"对义"关系、"属性-宿主"关系、"部件-整体"关系、"材料-成品"关系、"事件-角色"关系。所有的义原之间组成了一个复杂的网状结构。在这 8 种义原关系中,最重要的"上下位"关系,该关系可以用如图 1 所示的树状层次结构来表示。

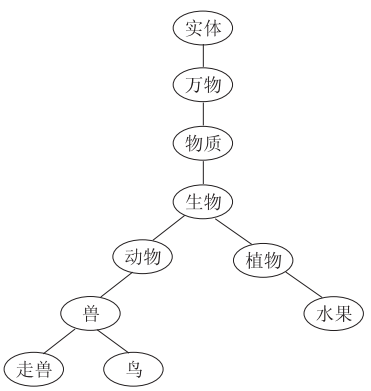

图 1 义原"上下位"关系树状图

为了便于后文的讨论,依据义原"上下位"关系树状图给出如下两个定义:

定义 1 绝对高度($Height(P_i)$)指的是节点到根节点的路径长度。

例如, $Height$ ("动物")=4。

定义 2 重合度指的是两个节点第一次到达同一个父节点所经过的最长路径长度,文中用 $Length(P_i,P_j)$ 表示。

例如, $Length$ ("走兽","水果")=3。

从图 1 的语义树形图中,可以得出如下结论:

对于重合度相同的节点对,处于语义树较高层的,其语义距离较大。例如,"动物"和"植物","走兽"和"鸟",这两对词语间的重合长度都是 1,但前一对词("动物"和"植物")绝对高度为 3,后一对词("走兽"和"鸟")绝对高度为 6。

Lin Dekang 认为任何两个事物的相似度取决于它们的共性(Commonality)和个性(Difference),并从信息理论的角度给出任意两个事物相似度的通用公式[7]:

$$Sim(x,y)=\frac{p(common(x,y))}{p(description(x,y))} \tag{1}$$

其中, $common(x,y)$ 描述了 x、y 共性所需要的信息量大小; $description(x,y)$ 描述了 x、y 所需的信息量大小。

在义原"上下位"树形图中,节点共性主要体现在两个节点的父节点,个性主要体现在节点之间重合度上,综合考虑节点的共性信息和个性信息,本文中给出如下的义原语义相似度计算公式:

$$Sim(p_i,p_j)=\frac{Height(pnode)}{Length(p_i,p_j)+Height(pnode)} \tag{2}$$

其中, $height(pnode)$ 表示义原 p_i、p_j 共同父节点的绝对高度。

有了义原相似度计算公式后,两个词语 W_1 和 W_2 的相似度为各个概念的相似度之最大值,计算方法参照文献[5],公式如下:

$$Sim(W_1,W_2)=\max_{i=1,\cdots,n,j=1,\cdots,m} Sim(S_{1i},S_{2j}) \tag{3}$$

其中, $S_{11},S_{12},\cdots,S_{1n}$ 为 W_1 的 n 个概念; $S_{21},S_{22},\cdots,S_{2m}$ 为 W_2 的 m 个概念。

两个概念语义表达式的整体相似度记为:

$$Sim(S_1,S_2)=\sum_{i=1}^{3}\beta_i Sim_i(S_1,S_2) \tag{4}$$

其中, $Sim_1(S_1,S_2)$ 为"上下位"义原关系相似度; $Sim_3(S_1,S_2)$ 为"符号"义原相似度; $Sim_2(S_1,S_2)$ 为其他独立义原相似度。

3 句子相似度计算

传统的方法只是简单地运用词语共现的方法

表 1 概念描述表

符号	NO	W_C	G_C	E_C	W_E	G_E	E_E	DEF
符号含义	概念编号	词语	词性	例子	对应的英文单词	英文词性	英文例子	概念定义
举例	017146	洗衣	V		wash clothes	V		{wash|洗涤:patient={clothing|衣物}}

来计算相似度,在计算句子相似度的时候并没有考虑句子中词语的语义距离。本文利用《知网》来计算词语之间的语义距离,有关计算如下。

3.1 表层相似度计算

表层相似度指的是两个句子形态上的相似程度,以两个句子中所含相同词或同义词的个数来衡量。设 P_1、P_2 为两个句子,则 P_1、P_2 的词形相似度为:

$$Sim_s(P_1,P_2)=\frac{2*\Gamma(\pi(P_1)\cap\pi(P_2))}{Len(P_1)+Len(P_2)} \quad (5)$$

其中,$\cap$ 表示集合的交运算;Γ 运算符表示求集合中的元素个数;$\pi(S_1)\cap\pi(S_2)$ 表示的是两个句子中所含有相同词或者同义词的集合;$Len(S_i)$ 表示句子的长度,即句子中含有的词语个数。表层相似度表明,两个句子中所含的相同词或者同义词越多,则表层相似度越大。

通过对大量语料句子的观察,我们发现句子中的名词和动词更能够体现句子的中心思想,需要对这类词赋以较大的权重,因此对公式(5)进行修正为:

$$Sim_s(P_1,P_2)=\frac{2*(\lambda_1\Gamma_1(\pi(P_1)\cap\pi(P_2))+\lambda_2\Gamma_2(\pi(P_1)\cap\pi(P_2)))}{Len(P_i)+Len(P_j)} \quad (6)$$

其中,$\Gamma_1(\pi(P_1)\cap\pi(P_2))$ 指的是句子中所含有相同或者相近的名词、动词个数;$\Gamma_2(\pi(P_1)\cap\pi(P_2))$ 指的是含有其它词的个数;λ_1、λ_2 为常数,且 $\lambda_1+\lambda_2=1$,$\lambda_1>\lambda_2>0$。

3.2 语义偏移量相似度计算

语义偏移量相似度综合考虑了句中词语的语义相似性,以及词语在句子中的相对位置,它反映两个句子中的词语在语义以及位置关系上的相似程度。设 P_1、P_2 为两个句子,则两个句子的语义词序相似度计算公式如下:

$$Sim_p(P_1,P_2)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}sim(W_i,W_j)\times(1-|pos(W_i)-pos(W_j)|)}{len(P_1)\times len(P_2)} \quad (7)$$

其中,n、m 分别为两个句子中词语的数量;$sim(W_i,W_j)$ 为两个词语的相似度;$pos(W_i)$ 为词语 W_i 在句子中的相对位置,$pos(W_i)=\frac{i}{len(P_i)}$。

3.3 句子相似度的计算

根据以上分析,句子相似度取决于表层相似度和语义词序相似度,综合考虑二者对句子相似度的影响,本文给出如下句子相似度计算公式:

$$Sim(P_1,P_2)=\alpha_1\times Sim_s(P_1,P_2)+\alpha_2\times Sim_p(P_1,P_2) \quad (8)$$

其中,$Sim_s(P_1,P_2)$ 为 P_1、P_2 的表层相似度;$Sim_p(P_1,P_2)$ 为语义偏移量相似度;α_1、α_2 为常数,且满足 $\alpha_1+\alpha_2=1$。句子相似度反映了两个句子之间的相似程度,取值在 0 和 1 之间。如果两个句子完全没有任何关系,则相似度计算结果为 0;如果两个句子完全一样,则相似度为 1。

句子相似度的具体计算步骤如下:

(1)对句子 P_1、P_2 进行分词,并去掉无意义的停止词。

(2)依照同义词词典查找两个句子中同义词,计算两个句子中同义词和相同词的个数。

(3)按照公式(6)计算两个句子的表层相似度。

(4)依据公式(3)获得两个句中词语的语义相似度后,按照公式(7)计算句子的语义偏移量相似度。

(5)综合句子的表层相似度和语义偏移量相似度,按照公式(8)计算两个句子的相似度。

本文中所提到的方法,克服了传统基于词语共现方法的不足,能够从语义方面更深层次地挖掘两个句子的相似度。

4 实验及评价

本文的测试语料选择的是文献[1]中所提到的句子库,该语料是由清华大学周强博士提供的,语料中所有的句子都已经经过切分和词性的标注[1],格式如下:

30 [zj-XX [dj-ZW 梦雅/nP [vp-AD 呆/v 了/u]]。/。]

31 [zj-XX [fj-LG [vp-ZZ 从此/d [vp-ZZ 不/dN [vp-ZZ 再/d [vp-PO 有/v /n]]]],/, [vp-ZZ 老/d 失眠/v]]。/。]

32 [zj-XX [dj-ZZ 后来/t ,/,[dj-ZW 他们/rN [vp-AD 离婚/v 了/u]]]。/。]

在原有语料的基础上,人工添加 3 个句子,分别是句子 2、句子 3、句子 4,并且认为这 3 个句子与句子 1 的相似度相近。表 2 中的第一种方法指的是文章前言中所提到的词语共现的方法,具体公式是北大计算语言所提出的公式;第二种方法是采用文献[1]中所提到的方法。

在上述理论研究的基础上,采用 VB.net 开发

了一个句子相似度计算系统,程序运行界面如图 2 所示。

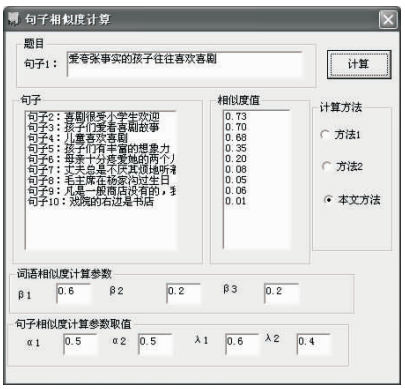

图 2 句子相似度计算图

实验中,$\sum_{i=1}^{3}\beta_i Sim_i(S_1,S_2)$ 中 $\beta_1=0.6$,$\beta_2=0.2$,$\beta_3=0.2$。$Sim_s(P_1,P_2)$ 中的 λ_1、λ_2 取了 3 组值:(1) $\lambda_1=0.6$,$\lambda_2=0.4$;(2) $\lambda_1=0.5$,$\lambda_2=0.5$;(3) $\lambda_1=0.4$,$\lambda_2=0.6$。$Sim_s(P_1,P_2)$ 公式中的 α_1、α_2 也分别取了三组值:(1) $\alpha_1=0.6$,$\alpha_2=0.4$;(2) $\alpha_1=0.7$,$\alpha_2=0.3$;(3) $\alpha_1=0.5$,$\alpha_2=0.5$。对 9 组参数取值所得到的结果取其平均值。按照方法一、方法二以及本文的方法,将 10 个句子分别与第一个句子进行相似度计算,计算结果如表 2 所示。

表 2 实验结果

实验编号	句子实例	相似度值		
		第一种方法	第二种方法	本文方法
1	爱夸张事实的孩子往往喜欢喜剧	1.00	1.00	1.00
2	喜剧很受小学生欢迎	0.18	0.66	0.73
3	孩子们爱看喜剧故事	0.18	0.53	0.70
4	儿童喜欢喜剧	0.20	0.55	0.68
5	孩子们有丰富的想象力	0.18	0.15	0.35
6	母亲十分疼爱她的两个儿子	0.00	0.10	0.20
7	丈夫总是不厌其烦地听着	0.00	0.05	0.08
8	毛主席在杨家沟过生日	0.00	0.05	0.05
9	凡是一般商店没有的,我们这里都有	0.00	0.05	0.06
10	戏院的右边是书店	0.00	0.02	0.01

第一种方法测试的结果中,句子 2、句子 3、句子 4 与句子 1 的相似度值都要小于 0.2,明显与实际情况不符。按照本文的方法测试,句子 2、句子 3、句子4与句子1的相似度值都要大于其它两种方法测试的值。可以看出,本文中所提出的相似度计算方法更符合实际情况。因此,本文所提出的句子相似度计算方法具备一定的实用性。

5 结束语

句子相似度的计算在自然语言处理领域中有着非常重要的意义。在分析已有算法的优点和缺点的基础上,本文提出了一种改进的句子相似度计算方法。首先依据《知网》来计算词语之间的语义相似度,然后从表层相似度和语义词序相似度来计算句子的相似度。实验结果表明,在同等的测试条件下,本文所提出的句子相似度计算方法测试结果更符合实际情况,因此本文中所提到的句子相似度计算方法有一定的应用价值。此外,本文在计算词义相似度时,并没有考虑到《知网》中未收录的词,这将在一定程度上影响词语相似度计算的准确性,在下一步的工作中,将对未登录词的语义相似性做进一步的研究。

参考文献:

[1] 王荣波,池哲儒,常宝宝, 等. 基于词串粒度及权值的汉语句子相似度衡量[J]. 计算机工程, 2005, 31(13):142-144.

[2] 吕学强,任飞亮,黄志丹, 等. 句子相似模型和最相似句子查找算法[J]. 东北大学学报(自然科学版), 2003, 24(6):531-534.

[3] 张民, 李生, 赵铁军, 等. 一种汉语句子间相似度的度量算法及其实现[C]//计算语言学进展与应用, 1995.

[4] 董振东,董强. 知网[DB/OL]. [2011-06-23]. http://www.keenage.com.

[5] 刘群,李素建. 基于《知网》的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会,2002.

[6] 梅家驹,竺一鸣, 高蕴琦, 等. 同义词词林[M]. 上海:上海辞书出版社,1993.

[7] Lin Dekang. An Information-Theoretic Definition of Similarity Semantic distance in WorldNet[C]//Proc of the Fifteenth International Conference on Machine Learning, 1998.

程传鹏(1977 -),男,河南信阳人,硕士,讲师,研究方向为自然语言处理。**E-mail**:Cheng8444@sina.com

CHENG Chuan-peng, born in 1977, MS, lecturer, his research interest includes natural language processing.