# 改进的基于知网词汇语义褒贬倾向性计算

杨昱昺,吴贤伟
YANG Yu-bing,WU Xian-wei

宁波大红鹰学院 电子信息学院,浙江 宁波 315175
Electronic Information Branch,Ningbo Dahongying College,Ningbo,Zhejiang 315175,China
E-mail:yybljf@163.com

**YANG Yu-bing,WU Xian-wei.Improved lexical semantic tendentiousness recognition computing.Computer Engineering and Applications,2009,45(21):91-93.**

**Abstract:** Lexical semantic tendentiousness recognition computing is the base of the sentence tendentiousness,and the sentence tendentiousness recognition is the text tendentiousness recognition and the chapter structure tendentiousness recognition foundation. Based on HowNet lexical semantic similarity computing,according to the current vocabulary appraise tendentious theory used by the method of computing similarity between words and benchmark words,from appraise benchmark words and computing formula, the paper puts forward an improved method.With experiment validation,in the same pair of benchmark words,accuracy rate has greatly been improved,arriving at 98.94%,there is some value in practical application.

**Key words:** semantic similarity;tendentiousness recognition;HowNet;appraise benchmark words

**摘 要:**词汇语义褒贬倾向性研究是句子褒贬倾向性识别的基础,而句子褒贬倾向性识别又是文本倾向性识别和篇章结构褒贬倾向性识别的基础。以《知网》的词汇语义相似度计算为基础,针对目前采用计算基准词对与词汇相似度的方法识别词汇褒贬倾向性理论,从褒贬基准词和计算公式入手,提出了改进办法。实验证明,在同样基准词对下,准确率得到了很大的提高,达到98.94%,具有实际应用价值。

**关键词:**语义相似度;倾向性识别;知网;褒贬基准词
**DOI:**10.3778/j.issn.1002-8331.2009.21.026 **文章编号:**1002-8331(2009)21-0091-03 **文献标识码:**A **中图分类号:**TP391

## 1 引言

词汇语义褒贬倾向性计算实际是计算某一词汇褒贬程度的度量值,为了便于处理,一般将度量值设定在[-1,+1]之间的实数。当度量值高于某阈值时,判别为褒义倾向;反之,则判为贬义倾向。这样,可以通过对句子中词汇的语义倾向值求平均的方式,获得句子的语义倾向,而句子又是构成篇章的基础,以此类推可获得篇章的语义倾向;另外,句子褒贬倾向性识别又是文本倾向性识别的基础,文本倾向性识别在信息过滤、自动文摘、文本分类等领域有广泛的应用前景。因此,对词汇的语义褒贬倾向性研究是此类研究中的关键工作。

自20世纪90年代,词汇倾向性的研究在国外得到了普遍的关注,并迅速发展起来。1997年,Hatzivassiloglou和McKeown通过对训练语料的学习进行形容词语义倾向判别,准确率达到82%[1]。2003年,Turney采用计算基准词对与词汇相似度的方法识别词汇倾向性,其准确率在包含形容词、副词、名词、动词的完整测试集上达到82.8%[2]。2002年,由刘群等人提出了基于《知网》[3]词汇语义相似度计算方法,成为目前中文词汇倾向性计算的主要依据[4]。

在朱嫣岚论文[5]词汇语义褒贬倾向性研究的基础上,指出了该算法中存在的一些不足之处,并对该算法进行一定的改进,通过实验证明该改进后的算法比原算法在准确率上有了较大的提高。

## 2 词汇语义褒贬倾向性计算

### 2.1 基于《知网》的语义相似度计算

知网(英文名称为HowNet)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。《知网》中两个主要的概念:“概念”与“义原”。“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念。“概念”是用一种“知识表示语言”来描述的,这种“知识表示语言”所用的“词汇”叫做“义原”。“义原”是用于描述一个“概念”的最小意义单位。

义原一方面作为描述概念的最基本单位,另一方面,义原之间又存在复杂的关系。在《知网》中,一共描述了义原之间的8种关系: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。可以看出,义原之间组成的是一个复杂的网状结构,而不是一个单纯的树状结构。不过,义原关系中最重要的还是上下位关

**基金项目:**浙江省教育厅科研项目(No.20071322)。
**作者简介:**杨昱昺(1969-),男,副教授,主要研究方向信息检索、中文信息处理;吴贤伟(1975-),男,系统分析师,主要研究方向图像检索。
**收稿日期:**2009-03-03 **修回日期:**2009-04-29

系。根据义原的上下位关系,所有的“基本义原”组成了一个义原层次体系。这个义原层次体系是一个树状结构,这也是进行语义相似度计算的基础。

在刘群论文中提出两个孤立词语之间的相似度计算最终归结到了两个概念之间的相似度计算。对于两个汉语词语 $W_1$ 和 $W_2$,如果 $W_1$ 有 $n$ 个义项(概念):$S_{11},S_{12},\cdots,S_{1n}$,$W_2$ 有 $m$ 个义项(概念):$S_{21},S_{22},\cdots,S_{2m}$,则 $W_1$ 和 $W_2$ 的相似度等于各个概念的相似度之最大值,即

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

而任一个义项可由四个部分组成:第一独立义原、其他独立义原、关系义原和符号义原,其中义原相似度的计算公式如下:

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \tag{2}$$

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。$\alpha$ 是一个可调节的参数。

这样两个义项(概念)语义表达式的整体相似度公式如下:

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

其中,$\beta_i(1\leqslant i\leqslant 4)$是可调节的参数,且有:$\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$。$Sim_1(S_1,S_2)$是第一独立义原描述式,$Sim_2(S_1,S_2)$是其他独立义原描述式,$Sim_3(S_1,S_2)$是关系义原描述式,$Sim_4(S_1,S_2)$是符号义原描述式。

## 2.2 词汇语义褒贬倾向性计算

在朱嫣岚论文中对某一词汇 $W$ 的语义褒贬倾向性计算指导思想是:先给定 $k$ 对基准词(其中 $k$ 个褒义词,$k$ 个贬义词),利用《知网》语义相似度计算公式,求出词汇 $W$ 与 $k$ 对基准词中每个词的语义相似度,并统计出 $k$ 个褒义词语义相似度的和 $S_1$、$k$ 个贬义词语义相似度的和 $S_2$,若 $S_1-S_2>0$ 则认为词汇 $W$ 更加接近褒义倾向,认定为褒义词,若 $S_1-S_2<0$ 则认为词汇 $W$ 更加接近贬义倾向,认定为贬义词。

词汇 $W$ 的语义褒贬倾向值计算公式如下:

$$Orientation(W)=\sum_{i=1}^{k}Similarity(key\text{-}p_i,W)-\sum_{j=1}^{k}Similarity(key\text{-}n_j,W) \tag{4}$$

其中 $k$ 表示 $k$ 对基准词,每对基准词包括一个褒义词和一个贬义词。褒义基准词表示为 $key\text{-}p$,贬义基准词表示为 $key\text{-}n$,$Similarity(key,W)$等于公式(3)中的 $Sim(key,W)$。

用于实验的 40 对基准词如表 1 所示。

**表 1 40 组褒贬基准词**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 褒义基准词: | | | | | | | | | |
| 健康 | 安全 | 天下第一 | 美丽 | 超级 | 保险 | 卫生 | 天使 | 英雄 | 精选 |
| 快乐 | 权威 | 稳定 | 优秀 | 高级 | 精英 | 最好 | 最佳 | 幸福 | 容易 |
| 高手 | 文明 | 积极 | 著名 | 漂亮 | 完美 | 简单 | 和平 | 开通 | 真实 |
| 先进 | 便宜 | 优质 | 欢乐 | 美好 | 良好 | 不错 | 出色 | 成熟 | 完善 |
| 贬义基准词: | | | | | | | | | |
| 不合作 | 黑客 | 疯狂 | 错误 | 事故 | 非法 | 失败 | 背后 | 麻烦 | 不良 |
| 病人 | 恶意 | 色情 | 暴力 | 黄色 | 浪费 | 落后 | 漏洞 | 有害 | 讨厌 |
| 自负 | 不安 | 魔鬼 | 花样 | 野蛮 | 陷阱 | 不当 | 腐败 | 无情 | 失误 |
| 淫秽 | 流氓 | 虚假 | 残酷 | 变态 | 脆弱 | 不合格 | 愚人 | 恶劣 | 恶魔 |

朱嫣岚论文中通过实验选用词频最高的一部分词作为测试集 3,而基准词根据词频选取前 1 对、4 对、5 对、10 对、15 对、20 对、30 对、40 对褒贬词进行测试,随着基准词对的变化,准确率也从 22%变化到 87%左右。最终得出的结论是:基于《知网》的语义倾向判别,只需利用《知网》的本地资源和少量的基准词,比较容易实现且不受外界条件(如网络环境)的干扰。从实验结果来看,基准词的增加对判别的准确性提高有明显作用,但即使是极少量基准词,在常用词集中同样可以达到 80%以上的准确率。最后也提到了两方面存在的不足:(1)基准词的选取不够科学全面;(2)算法比较直观,不够科学。

## 2.3 改进的词汇语义褒贬倾向性计算

针对朱嫣岚论文中算法存在的问题,将从基准词的选取和算法的改进两方面着手,最后通过实验证明在同样基准词对下,准确率得到了很大的提高,达到 98.94%。

### 2.3.1 基准词的选取

表 1 中基准词的选取原则是按照 Google 搜索返回 Hits 数,即它们在 Web 上的词频前 40 组褒贬词得到。但使用频率高不等于词汇的覆盖面广,这样就造成了基准词中许多词汇在《知网》中的语义是相同的。如褒义词中的“天下第一”、“优秀”、“漂亮”、“优质”、“良好”、“出色”、“完善”在《知网》中的义项都是“aValue|属性值,GoodBad|好坏,good|好,desired|良”、“高级”、“最好”、“最佳”在《知网》中的义项都是“aValue|属性值,rank|等级,HighRank|高等,desired|良”,而贬义词中的“不良”、“落后”、“有害”、“恶劣”在《知网》中的义项都是“aValue|属性值,GoodBad|好坏,bad|坏,undesired|莠”,“疯狂”、“野蛮”、“无情”、“残酷”在《知网》中的义项都是“aValue|属性值,behavior|举止,fierce|暴,undesired|莠”,等。由于表 1 基准词中有不少词汇的语义是相同的,在一定程度上影响了基准词词汇的覆盖面,结果也影响了需判别词汇语义褒贬倾向性的准确率。

基准词选取的原则是基于朱嫣岚论文基准词选取原则,将表 1 中语义重复的词替换成新的有较高 Hits 数的褒义词或贬义词,最后得到新的 40 组褒贬基准词(见表 2),这 40 组褒贬基准词的特点是在保留较高使用频率外,排除了语义相同的情况,提高了词汇的覆盖面。

**表 2 改进后 40 组褒贬基准词**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 褒义基准词: | | | | | | | | | |
| 健康 | 友善 | 美丽 | 保险 | 卫生 | 天使 | 精选 | 权威 | 优秀 | 精英 |
| 欢喜 | 幸福 | 容易 | 文明 | 积极 | 著名 | 完美 | 简单 | 和平 | 开通 |
| 真实 | 先进 | 便宜 | 不错 | 成熟 | 诚信 | 乖巧 | 勤俭 | 坚定 | 精神 |
| 茂盛 | 安静 | 成绩 | 雄心 | 奖牌 | 完整 | 新 | 亮点 | 捷报 | 利润 |
| 贬义基准词: | | | | | | | | | |
| 不合作 | 黑客 | 疯狂 | 错误 | 事故 | 非法 | 失败 | 背后 | 麻烦 | 不良 |
| 病人 | 恶意 | 色情 | 暴力 | 黄色 | 浪费 | 漏洞 | 讨厌 | 自负 | 不安 |
| 花样 | 陷阱 | 敌对 | 失误 | 流氓 | 虚假 | 变态 | 脆弱 | 不合格 | 愚 |
| 谣言 | 淫秽 | 嘈杂 | 残 | 恶势力 | 缺失 | 脏 | 陈旧 | 丑陋 | 毒 |

### 2.3.2 算法的改进

在朱嫣岚论文中,词汇语义褒贬倾向性计算方法是根据所要判断词汇 $W$ 与预先设定的褒贬基准词对中的每一个词进行语义相似度计算,累加词汇 $W$ 和所有褒义基准词的语义相似度 $S_1$,累加词汇 $W$ 和所有贬义基准词的语义相似度 $S_2$,最后判断 $S_1$ 和 $S_2$ 的大小,若 $S_1>S_2$,则认为词汇 $W$ 更具有褒义倾向性,否则若 $S_1<S_2$,则词汇 $W$ 更具有贬义倾向性。该算法采用了统计方法,即利用词汇 $W$ 与褒贬义基准词集合的相似度和进行比较,最后得到词汇 $W$ 更倾向于哪一边。但实验发现,对贬义词处理的准确率较高,而褒义词的准确率偏低,出现这种结果的原因在于词汇语义褒贬倾向性计算值比实际偏小,这估计

与褒贬基准词对的选择有关。但要选择计算值与实际完全一致的褒贬基准词对是很困难的,通过加入语义相似度最大值并进行适当的调节,且能获得这种整体的平衡。具体思想方法如下:

假设词汇 $W$ 是褒义的,则一般该词的 $S_1$ 应该大于 $S_2$,而该词与褒义词集合中语义相似度的最大值 $M_1$ 一般也应该大于该词与贬义词集合中语义相似度最大值 $M_2$。而且实验发现,如果直接采用 $M_1$ 与 $M_2$ 来代替 $S_1$ 与 $S_2$ 之间的比较,准确率也较高。另外,还发现该方法对褒义词处理的准确率较高,而贬义词的准确率较低,即词汇语义褒贬倾向性计算值比实际偏大。能否通过$(S_1+M_1)-(S_2+M_2)$代替原 $S_1-S_2$,获得一种平衡,从而提高词汇语义褒贬倾向性判别的准确率。

但实验又发现,单纯地将原算法 $S_1-S_2$ 改成$(S_1+M_1)-(S_2+M_2)$准确率并没有得到很大改观,这是因为一般 $S$ 比 $M$ 要大很多,如果不适当调低 $S$ 的值,$M$ 所起的作用将不会很大。经过综合考虑,最后将公式(4)改成公式(5),改进后的算法如下:

$$Orientation(W)=(\frac{1}{\alpha}\sum_{i=1}^{k}Similarity(key-p_i,W)+\frac{1}{\beta}\max_{i=1..k}Similarity(key-p_i,W))-(\frac{1}{\alpha}\sum_{j=1}^{k}Similarity(key-n_j,W)+\frac{1}{\beta}\max_{j=1..k}Similarity(key-n_j,W)) \quad (5)$$

其中 $k$ 表示 $k$ 对基准词,每对基准词包括一个褒义词和一个贬义词。褒义基准词表示为 $key-p$,贬义基准词表示为 $key-n$,$Similarity(key,W)$等于公式(3)中的 $Sim(key,W)$。$\alpha,\beta$ 是可调节参数,根据给定的基准词对,可通过对 $\alpha,\beta$ 的调节提高算法的准确率。

## 3 实验及结果分析

### 3.1 褒贬基准词之间的比较

采用相同的算法,都是使用原算法公式(4),但分别采用表 1 和表 2 中的褒贬基准词。

在实验中(包括后面的实验),默认使用 0 为阈值,即倾向值大于 0 则判断为褒义,小于 0 则判断为贬义。语义倾向判别准确率=判别正确的词数/测试集总词数,以此来衡量算法效果。

测试集使用了《知网》2000 中文词表中标注“良”(褒义),“莠”(贬义)属性的词汇。排除了既有“良”又有“莠”的词,因为这些词汇在不同语境下,或为褒义,或为贬义,并不能简单地将其判断为褒义词或贬义词。例如:词语“好看”,在描述事物时,可作褒义,如“这花真好看”,而在“要你好看”这样的语句中,显然带有强烈的贬义。这样共选用 5 930 个词。其中褒义词 2 884 个,贬义词 3 046 个。

词汇语义相似度计算使用基于《知网》语义相似度的方法,下同。

实验结果如表 3 所示。

**表 3 利用相同算法不同褒贬基准词集合的语义褒贬倾向性准确率** (%)

| 褒贬基准词来源 | 褒义词准确率 | 贬义词准确率 | 平均准确率 |
|---|---|---|---|
| 表 1 | 80.17 | 93.89 | 87.03 |
| 表 2 | 86.51 | 99.93 | 93.22 |

从表 3 实验结果可以比较明显地看出,用相同的算法,当采用改进后的褒贬基准词后,平均准确率提高了 6.19%。

### 3.2 算法之间的比较

在算法之间的比较实验中,包含两个子实验,第一个是通过相同基准词,不同算法之间的比较,第二个是通过不同算法,不同褒贬基准词集合之间的比较。

子实验 1,采用原算法公式(4)和改进算法公式(5)两种不同的算法,基准词集合采用表 2 中的 40 组褒贬基准词,测试集同前。公式(5)中 $\alpha$ 取 12,$\beta$ 取 1。

实验结果如表 4 所示。

**表 4 利用不同算法相同褒贬基准词集合的语义褒贬倾向性准确率** (%)

| 算法 | 褒义词准确率 | 贬义词准确率 | 平均准确率 |
|---|---|---|---|
| 原算法公式(4) | 86.51 | 99.93 | 93.22 |
| 改进算法公式(5) | 98.93 | 98.95 | 98.94 |

子实验 2,采用原算法公式(4)和改进算法公式(5)两种不同的算法,而基准词集合采用表 2 中根据顺序选取前 1 对、4 对、5 对、10 对、15 对、20 对、30 对、40 对褒贬词进行测试。测试集同前。公式(5)中 $\alpha$ 取 12,$\beta$ 取 1。

实验结果如图 1 所示。

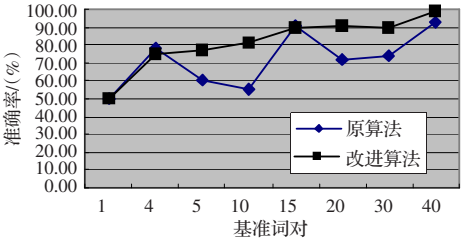


**图 1 基准词数量对实验效果的影响**

从表 4 实验 1 结果可以比较明显地发现,用相同的褒贬基准词,但当采用改进算法处理后,平均准确率提高了 5.72%,达到 98.94%。另外,从图 1 实验 2 结果发现,随着基准词数量的增加,词汇语义褒贬倾向性准确率都得到相应的改善,且改善情况改进算法明显比原算法要好。

### 3.3 改变测试集

前面几个实验测试集是采用《知网》中已经明确的褒贬词汇,那么对于那些未明确标识的褒贬词,在使用原算法和改进算法进行语义褒贬倾向性判别时的效果如何。下面是针对这个问题进行的实验。

测试集是一组由 10 个褒义词和 10 个贬义词组成的数据集,且这些词没有在《知网》中被标识为“良”或“莠”,褒贬基准词同表 2,分别使用原算法和改进算法进行语义褒贬倾向性计算,最后实验结果如下:

**表 5 10 组词的语义褒贬倾向性计算结果**

| 10 组褒贬词: | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 算法 | 拜寿 | 出生 | 安康 | 红光满面 | 喜欢 | 喜剧 | 放心 | 开心 | 侠客 | 保护 |
| 原 | –0.17 | –0.37 | 0.35 | 0.35 | 0.29 | 0.13 | –0.19 | 0.33 | 1.10 | 0.96 |
| 改进 | 0.01 | 0.01 | 0.79 | 0.79 | 0.74 | 0.04 | 0.14 | 0.74 | 0.22 | 0.08 |
| 10 组贬义词: | | | | | | | | | | |
| 算法 | 濒临灭绝 | 覆没 | 抱恙 | 受伤 | 厌恶 | 悲剧 | 担心 | 烦恼 | 小人 | 破坏 |
| 原 | –0.23 | –0.48 | –0.90 | –0.66 | –1.13 | –0.99 | –0.50 | –0.50 | –0.10 | –0.89 |
| 改进 | –0.01 | –0.08 | –0.48 | –0.31 | –0.88 | –0.18 | –0.20 | –0.20 | –0.41 | –1.62 |

(下转 108 页)