ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 3,May 2017,pp. 579－587
DOI:10.16337/j.1004-9037.2017.03.017

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于词向量的跨领域中文情感词典构建方法

冯　超　梁　循　李亚平　周小平　李晓菲

(中国人民大学信息学院,北京,100872)

摘　要: 情感分析已经成为当今自然语言处理领域的热点问题。对于文本的自动化、半监督式的情感分析研究具有广泛的理论和实用价值。基于情感词典的情感倾向分析方法是文本情感分析的一种重要手段。然而,中文词汇在不同领域中的情感倾向不尽相同,一词多义现象明显。同时,不同领域中的情感词也具有专业性、领域性的特点。针对这些问题,本文提出一种基于词向量相似度的半监督情感极性判断算法(Sentiment orientation from word vector,SO-WV),并依据该算法设计出一种跨领域的中文情感词典构建方法。实验证明,本文所设计的情感词典构建方法能有效地对情感词情感倾向进行判断。算法不仅在不同领域的情感词典建立上具有良好的可移植性,同时还具有专业性、领域性的特点。

关键词: 情感分析;情感词典;词向量;跨领域

中图分类号: TP391.1　　**文献标志码:** A

Construction Method of Chinese Cross-Domain Sentiment Lexicon Based on Word Vector

Feng Chao, Liang Xun, Li Yaping, Zhou Xiaoping, Li Xiaofei

(School of Information, Renmin University of China, Beijing, 100872, China)

Abstract: Nowadays, sentiment analysis has become a hot research topic in the natural language processing field. The automated and semi-supervised way of text sentiment analysis makes a high value on practicing and theory studies. The sentiment orientation algorithm based on sentiment lexicon is an important approach in text sentiment analysis. Constructing a sentiment lexicon effectively is a basic task in the text sentiment analysis. However, Chinese words are very ambiguous in different domains. Meanwhile, different areas of sentiment words also have the characteristic of specialized. To solve these problems, we propose a semi-supervised sentiment orientation classification algorithm based on word vector similarity (SO-WV). Experiments show that, the algorithm can classify the sentiment orientation of words effectively. This algorithm has the versatility in different areas, and also offers professional and specialized characteristics.

Key words: sentiment analysis; sentiment lexicon; word vector; cross-domain

引　　言

伴随着通信技术的高速发展,网络世界变得越来越多姿多彩,人们也越来越乐于在网络世界表达自

基金项目: 日立企业合作(中国社交媒体的分析服务系统)资助项目;中国人民大学科学研究基金(10XNI029)资助项目;北京市自然科学基金(4132067)资助项目;国家自然科学基金(71271211)资助项目。
收稿日期: 2015-09-29;**修订日期:** 2015-10-29

己的态度和看法,发表对事物的评价。对于这些大量出现的评论信息,如何分析和处理它们成为当前信息学科发展的重点。文本情感分析正是针对处理这些文本信息的语意情感而发展起来的一个研究方向[1-3]。在各种文本情感分析的方法中,通过情感词典的方式来进行语义情感的判断是一种常用而高效的方式[4,5]。因而,情感词典的构建在情感分析的研究中具有基础性的作用。情感词典的构建主要是对词汇进行情感极性的判断,即将词汇构建划分成褒义、贬义或中性词词典的过程。对于情感词典的构建,当前的主要方法可分为两类:(1)有监督式的学习方法[6-8]。这种方式需要人工事先标注一定量的语料库,然后通过机器学习或者语义关联规则等方式判断词语的情感极性。这种方式的情感词典构建过程需要大量的人力标注语料,且可扩展性较差。(2)半监督或者无监督的方式[1,2,9-12]。其中半监督的情感词典构建方法是基于现有的种子情感词典,通过近义词的方式在已有的电子词典中寻找情感词,并扩充情感词典[11]。这种方法虽然有效,但是由于其涵盖的词汇有限,只能在词典已有的词汇中扩充,因而具有一定的局限性。另一种无监督或半监督算法首先人工标注出一个种子词典,再利用一定的度量方式来对词汇与种子词典中词汇的联系程度进行衡量,最后依据这种联系程度对词汇进行自动的情感极性分类[1,9,10]。中文词汇具有多义性、模糊性的现象,同一词汇在不同领域的意义可能不同[10],而不同领域的情感词也不尽相同,因而有必要针对不同的领域构建其各自的领域情感词典[13,14]。本文正是为了解决这一任务,提出了一种新的半监督式领域情感词典构建方法。词向量是一种对于文本词汇有效的符号化表示方法[15]。使用深度学习网络训练出的分布式词向量(以下简称词向量)进行自然语言处理已经被证明是行之有效的工具[15-19]。通过词向量表示词汇可以有效地度量词语间的相似度[16,17],这种相似度计算方法提供了一种新的情感分类思路[19]。基于此,本文提出了一种基于词向量的情感极性算法(Semantic orientation from word vector, SO-WV)计算词语的情感极性,并构建出情感词典。本文利用相关领域语料构建词向量模型,通过基于词向量的情感极向算法计算出情感词的情感极性和情感强度,并依据此构建情感词典。

1 领域情感词典构建步骤

情感词典是对文本进行情感分析的有效工具,在构建情感词典的过程中主要有两种方法:(1)有监督的学习方法,通过有标记的语料库利用机器学习的方式形成情感词典,这种方法需要大量的人工标注[6-8]。(2)无监督或半监督的方式,从未标记的语料中学习得到情感词典或者从已有的电子词典、知识库中进行扩展[1,2,9-10,12]。本文方法属于第(2)种方法,半监督的构建方法。本文所提出的中文领域情感词典的构建方法的基本架构如图 1 所示,主要包括 3 个步骤。

(1) 对原始语料进行处理,包括分词和词性标注等。在进行预处理后,本文所提出的方法一方面通过深度学习网络训练出词向量模型,同时另一方面通过词性分析和统计处理挑选出情感词。

(2) 在通过第 1 步获得的情感词的基础上,通过词频统计和人工标注的方式获得常出现并且情感强度较大的情感词作为种子词典。

(3) 利用 SO-WV 算法获得情感词的情感倾向和情感强度,进而得到情感词典。

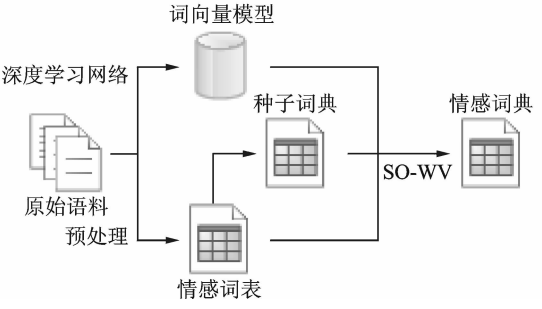

图 1 情感词典构建框架图

Fig. 1 Frame for construction of sentiment lexicon

2 基于词向量的中文情感词典构建方法

2.1 分布式词向量表示

将特征符号化表示是利用机器学习的方法解决问题的典型思路,在自然语言处理领域利用词向量的方式表示词语并进行分析与处理是这一思路的体现[15]。通过设置虚拟变量来表示文本词汇是较为

常用的词向量表示方式,如 One-hot Representation 表示方法,对于一个包含 n 个词汇的文本集合,利用 n 维向量表示一个词汇。显然这种方式在文本中词汇量比较大时其向量的维数将会非常高,产生维数灾难;并且这种方式的词汇向量表示方法损失了词汇之间的语义信息。为了解决以上问题,需要采用一种新的思路构造词向量。这种新的方法需要满足能够避免维数灾难、能够获取词汇与词汇间的语义信息,同时满足以下 3 个目标[15]:(1) 所有的词汇都可以用 m 维实空间中的特征向量表示。(2) 用联合概率函数表示词汇间共现的可能性。(3) 词汇的特征向量和概率函数的参数可以同时学习得到。通过分布式词向量表示方法可以解决上述问题。与 One-hot Representation 方法不同,分布式词向量表示方法利用一个较低维度的词向量来表示词汇,同时利用概率函数得到词汇与词汇间的语义联系[16]。本文所使用的词向量模型是利用深度递归神经网络中的 Skip-gram 模型训练出来的[15,16]。Skip-gram 模型是一个对称模型,在给定了一组训练样本 w_1, w_2,…, w_T 以后,模型的目标就是需要找到最大平均对数概率,目标函数为

$$F=\mathrm{Max}\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c,j\neq 0}\log p(w_{t+j}\mid w_t) \tag{1}$$

式中:F 为最大平均对数概率;T 为样本中词汇总数;w_t 为词汇在模型中的分布式表示向量;c 为模型在训练过程中所使用的窗口,表示以 w_t 为中心词汇,其前后的 c 个词汇都用于模型的训练。显然 c 越大则表示用于训练的样本更多,其精度也就越大,但同时所需要的训练时间也越多,算法复杂度也越高[16]。利用深度学习网络训练出来的词向量有很好的性质。可以直接利用向量的相似度表示词汇间的相似关系[16-18]。

2.2　利用词向量计算相似度

相关研究表明[1,2]语义相似的词汇更有可能拥有相同的情感极性。词向量相似度可以很好地度量词汇间的语义相似度,因而可以通过词向量相似度的方式来对词汇的情感极性进行计算和分类[19]。本文利用标准化后的词向量相似度衡量词汇语义的相似度,其计算步骤如下:

(1)计算相似度。对于两个词的词向量 $\boldsymbol{v}_1$ 与 $\boldsymbol{v}_2$,其相似度利用余弦相似度来度量,计算过程为

$$S(\boldsymbol{v}_1,\boldsymbol{v}_2)=\frac{\boldsymbol{v}_1\cdot\boldsymbol{v}_2}{\|\boldsymbol{v}_1\|\cdot\|\boldsymbol{v}_2\|} \tag{2}$$

式中:$\boldsymbol{v}_1\cdot\boldsymbol{v}_2$ 为 $\boldsymbol{v}_1$ 与 $\boldsymbol{v}_2$ 的内积,$\|\boldsymbol{v}\|$ 为词向量 $\boldsymbol{v}$ 的模。

(2)标准化。将相似度标准化到[0, 1],便于后续计算,标准化过程为

$$\mathrm{NS}(\boldsymbol{v}_1,\boldsymbol{v}_2)=\frac{S(\boldsymbol{v}_1,\boldsymbol{v}_2)+1}{2} \tag{3}$$

式中:NS 的范围为[0, 1]。

2.3　基于词向量的 SO-WV 算法

利用点互信息(Pointwise mutual information,PMI)的方式进行半监督式情感词典构造算法是一种常用方法[20]。Turney 提出了基于点互信息的情感分类算法(Semantic orientation from pointwise mutual information,SO-PMI)算法[1],实现了半监督的情感分析模型,该算法利用搜索引擎来判断文本的情感倾向。Wang 和 Araki 将该方法引入日文领域,结果显示该方法在不同语言情感分析上具有通用性特征[9]。但基于信息检索的算法对搜索引擎的依赖度过高,具有一定的局限性。本文利用词向量相似度计算方法,提出了基于词向量的 SO-WV 算法。首先在情感词表选取 $2m$ 个具有情感极性的词作为种子词典,其中正向情感词 m 个,记为集合 POS,负向情感词 m 个,记为集合 NEG。对于词汇 word,其与正负向种子情感词集相似度计算分别为

$$\mathrm{NS}(\mathrm{word},\mathrm{POS})=\sum_{i=1}^{m}\mathrm{NS}(\mathrm{word},\mathrm{POS}_i) \tag{4}$$

$$\mathrm{NS}(\mathrm{word},\mathrm{NEG})=\sum_{j=1}^{m}\mathrm{NS}(\mathrm{word},\mathrm{NEG}_j) \tag{5}$$

式中：POS_i 与 NEG_j 为第 i 个正向种子情感词与第 j 个负向种子情感词，式(4)表示词汇 word 与种子正向情感词集相似度，式(5)为词汇 word 与种子负向情感词集相似度。然后利用 word 的正负向相似度之差来计算它的情感倾向值 SO。词汇 word 的 S_O 值表示为

$$SO(word) = NS(word, POS) - NS(word, NEG) \tag{6}$$

当满足 SO(word)>0 表示 word 的情感极向为正向；若 SO(word)<0 则表示 word 的情感极向为负向；若 SO(word)=0 则表示 word 为中性词。

2.4 情感词典构建

本文利用 2.3 节的基于词向量的 SO-WV 算法对所找出来的初始情感词进行情感极性和情感强度的判断，得到情感词情感极性和情感强度表。对于得到的情感词表中存在情感极性不强的词，在构建极性情感词典的时候需要设置正向阈值 T_p 与负向阈值 T_n，筛除中性词。对于词汇 word，如其满足

$$SO(word) > T_p \quad 或 \quad SO(word) < T_n \tag{7}$$

则将其加入情感词典，否则认为其情感强度太低而不能加入情感词典。阈值的选取可由具体任务而定。

3 基于词向量的中文情感词典构建方法实验及分析

本文提出了一种中文领域情感词典的构造方法，针对不同的领域可以构建出适合该领域的情感词典。

3.1 实验前期工作

为了验证方法的有效性，本文收集了 4 个不同领域的中文语料数据，分别是汽车、金融、电脑和化妆品。原始语料的数据来源、数量等详细信息如表 1 所示。

表 1 领域原始语料统计信息

Tab. 1 Domain text data statistics

领域	数据来源	语料数量	情感词数量
汽车	太平洋汽车网，汽车之家网，易车网	180 930	4 561
金融	新浪财经网，中国证券网	115 109	3 906
电脑	太平洋电脑网，天极网，中关村在线，亚马逊	39 004	3 055
化妆品	乐蜂网，聚美优品网	461 402	3 701

在进行情感分析之前，首先对原始语料进行分词处理，本文所使用的中文分词处理工具是中国科学院计算机研究所的分词处理软件 ICTCLAS。在对语料信息进分词后，本文采用 Google 的开源词向量工具 Word2VEC 对各领域原始语料分别进行词向量的训练。在种子词典的构建过程中，本文首先提取出不同领域原始语料中的情感词，人工挑选出种子词典。各领域种子词典都分别包含 40 个正向情感词和 40 个负向情感。为了验证本文提出的算法能有效地对情感词的情感倾向进行分类，本文利用综合正确率、正向正确率和负向正确率 3 个指标来衡量算法的有效性。其中综合正确率表示所有词汇中算法判断准确的词汇比例。正向正确率表示正向词汇中算法判断正确的比例。负向正确率表示负向词汇中算法判断正确的比例。计算步骤如下：(1)首先对情感词汇进行人工情感标注，分出正向词汇与负向词汇。(2)利用本文提出的算法对情感词汇进行情感倾向计算，自动化地求出词汇情感倾向。(3)将算法计算的结果与人工标注的结果进行比较，评价算法的有效性。

3.2 实验结果及分析

对于汽车、金融、电脑和化妆品 4 个领域的语料，在利用本文 2.1 节提出的方法做出预处理以后，分

别得到了各领域的情感词、词向量模型和种子词典。利用本文提出的基于词向量的 SO-WV 算法,对各领域情感词做出情感极向计算。

3.2.1　训练方式比较

词向量模型的优劣将会对算法的正确率产生一定的影响。在词向量维度和训练窗口大小上,本文分别采用维度为 100、窗口为 5,维度为 200、窗口为 10,维度为 200、窗口为 10 的 3 种不同方式,并在各自领域得到 3 个不同的模型。本文将对 3 个模型的有效性进行评价,找到一种最为有效的训练方式。对于每一个领域,本文随机挑选出 100 条评论作为测试集 I,经过相关预处理后,找到评论中的情感词。汽车、金融、电脑和化妆品 4 个领域测试集 I 中的情感词数量分别为 592,655,1 117,726。对于各领域测试集 I 中的情感词,分别利用基于 3 个不同训练方式的词向量模型的SO-WV算法进行情感极性计算,判断其准确性,并进行比较。3 种训练方式在各个领域中的正确率指标如表 2 所示。总体来说,维度为 200,窗口为 10 的词向量模型的综合正确率最高;维度 200,窗口 5 的模型优于维度 100,窗口 5 的模型。原因如下:(1) 维度越高表示的信息越多,越不容易出现失真现象,有利于算法的相似度计算。(2) 窗口越大所用到的原始语料信息越多,损失的语义信息越少,结果越精确。(3) 过高的维度易出现过拟合,并且增加了计算复杂度。本文利用维度为 200,窗口为 10 的模型作为构造情感词典的词向量模型。

表 2　不同训练方式下各领域正确率

Tab. 2　Accuracy of each domain under deferent training methods　%

模型	汽车	金融	电脑	化妆品
维度 100,窗口 5	86.0	76.9	77.7	84.2
维度 200,窗口 5	84.5	79.4	79.7	84.4
维度 200,窗口 10	87.2	80.3	81.7	85.9

3.2.2　基于词向量的情感词典构造方法

(1) 有效性与通用性

本节通过实验,验证 4 个领域情感词典的有效性、通用性、专业性与领域性。对于实验构造出来的 4 个领域情感词典,分析利用正态核密度函数对各领域情感词情感值概率密度做出估计(图 2),发现实验构造出的情感词典中情感词情感值分布总体呈现正态分布特征,符合自然语言规律,表明算法在对情感词强度的判断上有效性较高。对于算法构造出的情感词典,表 3 统计了它们的综合、正向及负向正确率,如表 3 所示。算法的各项正确率都达到了 79%以上,有较高的精度,说明算法的有效性较高。

表 3　领域情感词典评价

Tab. 3　Evaluation of each domain sentiment lexicon　%

正确率	汽车	金融	电脑	化妆品
综合正确率	84.7	88.9	82.9	84.5
正向正确率	96.0	95.7	93.4	94.2
负向正确率	81.9	86.9	79.5	78.3

实验通过人工校验的方式对情感词的综合正确率随词量变化进行统计(图 3)。实验表明在各领域情感词中,情感值排名前 20%的词汇正确率达到了 90%以上,前 70%的情感词正确率达到了 80%以上,表明本文所提出的算法可以有效地判断情感词的情感极性。同时各领域间的正确率差别不大,表明算法具有跨领域通用性的特点。

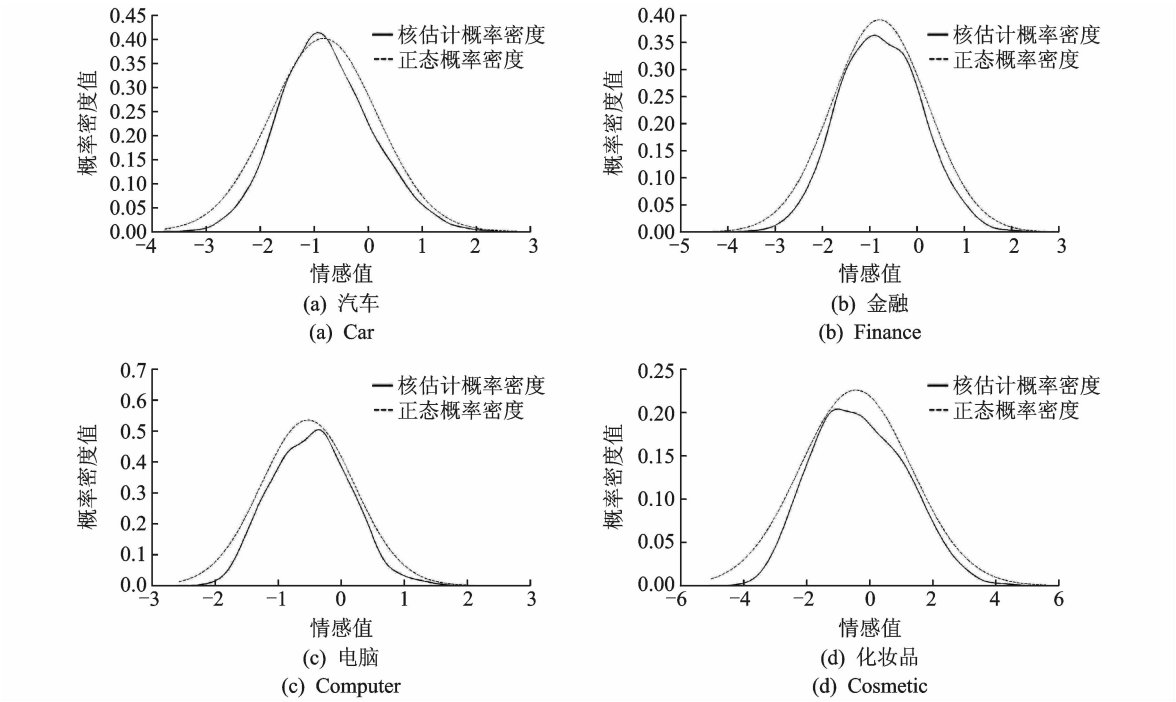

图 2 利用正态核函数对情感值进行概率密度估计

Fig. 2 Probability density function estimation of each domain sentiment value using Gaussian kernel function

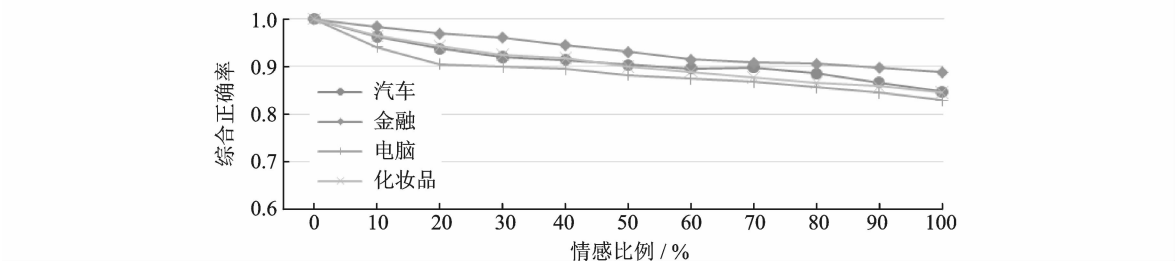

图 3 情感词典正确率统计

Fig. 3 Sentiment lexicon accuracy statistics

(2) 专业性与领域性

算法构造出的各领域的情感词典也具有很强的领域性与专业性。由于各情感词可能在各领域情感倾向不同,故针对每个领域的情感词典,实验分别选出情感值强度 Top 100,Top 200 和 Top 300 的情感词集,统计各个词集中在多个领域共同出现的情感词数量(如图 4 所示)。在 4 个领域情感强度 Top 100 的情感词中,没有同时都出现在 4 个领域的词;在 Top 200 的情感词中,4 个领域都出现的情感词比例为 1.5%;在 Top 300 的词汇中,4 个领域都出现的词比例为 2.7%,同时超过 3 个领域中出现的词的比例也都小于 20%。对于各领域构建出的情感词典,本文还分别列出了各领域情感值排名前 3 的正向及负向情感词,如表 4 所示。从表 4 可以看出,各领域情感词具有较强的领域性特点。

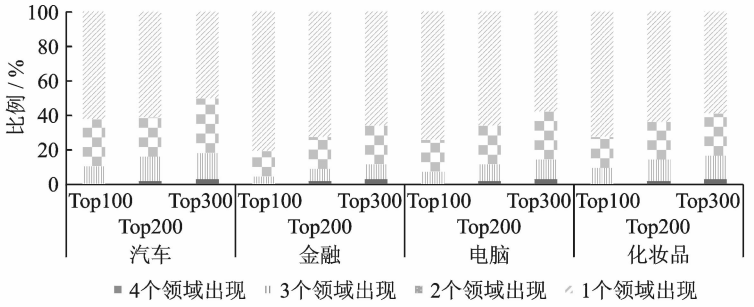

图 4　Top 100 中在多个领域出现的情感词统计

Fig. 4　Number of Top 100 sentiment words appeared in different domain sentiment lexicons

表 4　各领域排名情感词实例

Tab. 4　Representative sentiment words in each domain

领域	汽车	金融	电脑	化妆品
正向	优越	有力	清晰	滋润
	灵动	坚实	舒适	细腻
	温驯	扎实	安全	润
负向	心惊胆颤	困惑	莫名	受不了
	气人	惊恐	无奈	凄惨
	心虚	焦虑	不行	为难

结果表明,利用本文算法可以有针对性地对每个领域构建出专业性的情感词典。上述实验证明本文算法所构造出的情感词典不仅有良好的跨领域有效性,同时还具有良好的领域性和专业性特征,可以有效地找出各领域专业情感词。

3.2.3　不同情感词典构建方法的比较

为了检验本文算法与其他半监督情感词典构造算法在有效性方面的差异,本节用本文算法与基于信息检索的 SO-PMI 算法[1]和基于标签传播算法(Label propagation algorithm,LP)[10]分别构建情感词典,并进行了比较研究。实验在每个领域中分别随机选取 100 条评论作为测试集 II,各领域情感词数量如表 5 所示。对于测试集 II 中的情感词,分别利用两种算法进行情感极性的计算,得到情感值。表 6 给出了算法在 4 个领域情感词典各评价指标下的表现。从表 6 可以看出,SO-WV 算法在 4 个领域的有效性均优于 SO-PMI 算法和 LP 算法。而同时 SO-PMI 算法严重依赖于搜索网络状况与引擎,具有很强的缺憾性,而本文提出的 SO-WV 算法克服了这一问题。上述实验表明,本文提出的基于词向量的情感极性判断算法 SO-WV 有效,具有通用性、专业性和领域性特点,适合跨领域情感词典的构建。

表 5　测试集 II 中各领域情感词数量

Tab. 5　Number of sentiment words in test set II

情感词数量	汽车	金融	电脑	化妆品
总情感词数量	401	334	646	580
正向情感词数量	232	218	404	470
负向情感词数量	169	116	242	110

表 6 算法计算结果统计

Tab. 6 Statistics results of different algorithms %

正确率	算法	汽车	金融	电脑	化妆品
综合正确率	SO-WV	77.1	73.1	70.3	86.7
	SO-PMI	60.3	50.6	59.3	81.2
	LP	63.5	55.4	69.7	83.3
正向正确率	SO-WV	61.2	62.8	57.9	88.9
	SO-PMI	87.1	34.4	82.9	99.8
	LP	67.6	64.4	70.9	85.7
负向正确率	SO-WV	98.8	92.2	90.9	85.5
	SO-PMI	23.7	81.0	19.8	1.8
	LP	57.9	38.5	67.7	73.0

4 结束语

如何在不同领域语言环境中自动地判断词汇的情感倾向是当前自然语言处理的重点。中文词汇具有模糊性、多义性的特点，而不同领域的情感词往往也不相同，这些问题都是跨领域情感词典构造过程中的难点。针对这些问题，本文设计了一种基于词向量的中文情感词情感倾向计算方法 SO-WV，并在其基础上构造出一种跨领域中文情感词典构建方法。实验证明，本文提出的情感词典构建方法可以有效地针对不同领域构建出不同的领域情感词典，具有较高的精确度，拥有良好的可移植性，并且具有专业性和领域性。本文方法可以用于推荐系统设计、社区发现[21]等具体应用场景中，具有很强的实践意义。在研究过程中发现，当情感词的情感值较低时，算法的有效性会在一定程度上下降，这是因为有高频非情感词与中性词汇的干扰。下一步的工作会继续优化算法，对词向量模型本身进行改进，消除中性词汇和非情感词汇的干扰，提高算法的精度。希望将本算法应用到中文情感分析系统中，设计出一种新的情感分析方法提高文本情感分析的效率和效果。

参考文献：

[1] Turney P D. Thumbs up or thumbs down semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, Pennsylvania, USA:[s. n.],2002:417-424.

[2] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval,2008,2(1/2):1-135.

[3] Godbole N, Srinivasaiah M, Skiena S. Large-scale sentiment analysis for news and blogs[J]. ICWSM,2007,7:21.

[4] Mishne G. Experiments with mood classification in blog posts[C]// Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access. Citeseer: ACM,2005:19.

[5] Tsai A C R, Wu C E, Tsai R T H, et al. Building a concept-level sentiment dictionary based on common sense knowledge [J]. IEEE Intelligent Systems,2013,28(2):22-30.

[6] Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources[C]// EMNLP. Barcelona: ACL,2004:412-418.

[7] Wiebe J. Learning subjective adjectives from corpora[C]// Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. [S. l.]: AAAI Press,2000:735-740.

[8] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics,2004:271.

[9] Wang G, Araki K. Modifying SO-PMI for Japanese weblog opinion mining by using a balancing factor and detecting neutral

expressions[C]// Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2007: 189-192.

[10] 李寿山, 李逸薇, 黄居仁, 等. 基于双语信息和标签传播算法的中文情感词典构建方法[J]. 中文信息学报, 2013, 27(6): 75-81.
Li Shoushan, Li Yiwei, Huang Juren, et al. Construction of Chinese sentiment lexicon using bilingual information and label propagation algorithm[J]. Journal of Chinese Information Processing, 2013, 27(6): 75-81.

[11] 唐浩浩, 王波, 周杰, 等. 基于词亲和度的微博词语语义倾向识别算法[J]. 数据采集与处理, 2015, 30(1): 137-147.
Tang Haohao, Wang Bo, Zhou Jie, et al. Semantic orientation identification for terms from chinese micro-blogs based on word affinity measure[J]. Journal of Data Acquisition and Processing, 2015, 30(1): 137-147.

[12] Hu M, Liu B. Mining and summarizing customer reviews[C]// Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l.]: ACM, 2004: 168-177.

[13] Pan S J, Ni X, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment[C]// Proceedings of the 19th International Conference on World Wide Web. [S. l.]: ACM, 2010: 751-760.

[14] He Y, Lin C, Alani H. Automatically extracting polarity-bearing topics for cross-domain sentiment classification[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2011: 123-131.

[15] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The Journal of Machine Learning Research, 2003: 1137-1155.

[16] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[17] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems. Lake Tahoe: NIPS, 2013: 3111-3119.

[18] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]// 11th Annual Conference of the International Speech Communication Association. Makuhari, Chiba, Japan: [s. n.], 2010: 1045-1048.

[19] Xue B, Fu C, Shaobin Z. A study on sentiment computing and classification of sina Weibo with word2vec[C]// Big Data (BigData Congress), 2014 IEEE International Congress on. [S. l.]: IEEE, 2014: 358-363.

[20] Turney P. Mining the web for synonyms: PMI-IR versus LSA on TOEFL[C]// Proceedings of the Twelth European Conference on Machine Learning (ECML-2001). Freiburg, Germany: [s. n.], 2001: 491-502.

[21] 张海燕, 梁循, 周小平. 针对有向图的局部扩展的重叠社区发现算法[J]. 数据采集与处理, 2015, 30(3): 683-693.
Zhang Haiyan, Liang Xun, Zhou Xiaoping. Overlapping community detection from local extension in directed graph[J]. Journal of Data Acquisition and Processing, 2015, 30(3): 683-693.

作者简介:

冯超(1992-),男,硕士研究生,研究方向:数据挖掘、自然语言处理,E-mail: lukefeng@outlook.com。

梁循(1965-),男,教授,研究方向:数据挖掘、社会网络分析和网络金融。

李亚平(1989-),女,硕士研究生,研究方向:数据挖掘、自然语言处理。

周小平(1985-),男,讲师,研究方向:数据挖掘、社会网络分析。

李晓菲(1990-),女,硕士研究生,研究方向:数据挖掘、社会网络分析。