计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(10):245−250 [doi: 10.15888/j.cnki.csa.006879]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于词向量的中文新词情感倾向性分析①

杨　政, 易绵竹

(信息工程大学 洛阳校区, 洛阳 471003)

摘　要: 为更具体表义社会新词的情感含义及其倾向性, 该文提出了一种基于词向量的新词情感倾向性分析方法. 在信息时代不断发展变化中, 由于语言应用场景不断发展变化以及扩展语义表达的丰富性, 网络上不断出现很多表达情感的新词, 但是这些新词的表达虽有丰富的含义但缺乏准确的定义, 因此对其情感倾向性分析具有一定困难. 该文在分析了新词发现方法和词向量训练工具 Word2Vec 的基础上, 研究了基于 Word2Vec 的情感词新词倾向性分析方法的可行性和架构设计, 并面向微博语料进行实验, 结果显示新词可以从与其相近的词中分析其情感倾向.

关键词: 词向量; 新词发现; 情感词; 倾向性分析; Word2Vec

引用格式: 杨政,易绵竹.基于词向量的中文新词情感倾向性分析.计算机系统应用,2019,28(10):245–250. http://www.c-s-a.org.cn/1003-3254/6879.html

Study on Emotional Tendency of New Words Based on Word Vector

YANG Zheng, YI Mian-Zhu

(Luoyang Campus, Information Engineering University, Luoyang 471003, China)

Abstract: In order to more specifically express the emotional meaning and tendency of social new words, this study proposes a new word sentiment orientation analysis method based on word vector. In the everchanging development of the information age, due to the continuous development of language application scenarios and the enrichment of extended semantic expressions, many new words expressing emotions appear on the Internet, while the expression of these new words has rich meaning but lacks accurate definition. Therefore, it is difficult to analyze its sentiment orientation. Based on the analysis of the new word discovery method and the word vector training tool Word2Vec, this study focuses on the feasibility and architecture design of the new word orientation analysis method based on Word2Vec, and conducts experiments for microblog corpus. The results show that new words can analyze their emotional tendencies from similar words.

Key words: word vector; new word discovery; emotional word; tendentiousness analysis; Word2Vec

由于信息科技日新月异的变革和网络数据洪流的冲击, 人们日常使用的语言在不断发生着变化, 有越来越少使用的旧词汇被淘汰, 也有更多的新词汇大量涌现. 每一种新词的出现和变化发展都蕴含了不同的情感意义. 因此, 对于情感词的新词发现以及新词的倾向性分析研究一直是情感分析中的热点问题.

但是, 新词的产生伴随着多种情绪的表达, 这使得一个新词可能包含多种情感倾向, 而其中每种倾向的重要度各不相同. 因此本文提出基于词向量的新词情感倾向性分析方法, 通过量化的方式分析新词可能含有的多种情感倾向. 同时, 通过训练词向量, 可以对具有相似情感倾向的情感词进行聚类分析, 寻找近义词.

1　相关工作

对新词的发现和对新词进行情感分类的研究, 国外目前已有大量的相关工作, 对于文本情感分析也取

① 收稿时间: 2018-11-02; 修改时间: 2018-11-22; 采用时间: 2018-12-05; csa 在线出版时间: 2019-10-15

得不少研究成果, 但是国内对于短文本, 例如情感短语或情感词的研究可分为两种, 一是基于情感词典和规则的短文本情感分析, 二是基于机器学习的短文本情感分析.

基于情感词典的情感分析有肖江[1]等采用相似度方法构建相关领域情感词典, 有 Jo[2]等基于“主题-句子”关系的情感分类方法, 在词上同时标记主题和情感两类标签, 有杨立月[3]等构建的微博情感词典, 包括开源情感词典、具有时代特征的网络情感词典以及具有明显情感倾向的语气情感词典. 基于机器学习的情感分类方法有 2013 年 Liu[4]等提出的将 Co-training 协同训练算法和 SVM 相结合进行的情感分析. 2016 年 Dey[5]等利用 Bayes 算法进行的情感分析.

但是, 情感新词的发现不仅限于围绕情感词典发现的词语, 很多新词的产生都富含情感的标签. 例如“闺蜜”虽然是名词, 但是能表达一个人对另一个人蕴含的亲密情绪. 除此之外, 一个情感新词的情感也并不是单一的分类, 而是有多重情绪在其中的, 并且比重也不相同. 所以对情感新词的倾向应该是多分类的. 因此, 本文提出基于词向量的中文新词情感倾向性分析.

2 新词发现

自然语言处理领域一个重要攻克方向就是中文分词问题. 中文语言处理的过程不像西方语言一样在词与词之间有天然的分界线, 所以分词问题就可以在很大程度上影响接下来的很多步骤, 例如关系抽取、自动文摘等. 而现在大多的分词方法都是根据词库进行的, 那么未登录词的问题就显得更为重要[6,7]. 中文的书写没有首字母大写, 也没有专名号等, 因此计算机难以辨认人名地名等专有名词. 除专有名词外, 网络用语、品牌机构名、缩略语、简写词等词汇, 它们的出现和演变似乎完全无规律可寻. 随着语言处理的重要性不断提高, 中文分词领域的研究都在集中攻克这一难关. 自动发现新词成为了关键的环节.

新词挖掘的传统方法是, 首先对文本进行分词, 假定未能成功匹配的文本剩余片段是新词, 然后提取这些片段[8]. 但是, 分词结果的准确性是依赖于分词词库的完整性的, 如果分词词库中完全没有新词, 那么分词的结果就可能导致挖掘的“新词”难以成词.

那么新词挖掘需要另辟蹊径, 一种成熟的想法是, 不依赖于任何已建立好的词库, 而仅仅根据词本身含有的特征, 在较大规模的语料中将可能成词的文本片段全都提取出来, 不论该词是新词还是旧词. 然后, 将这个提取的词库中的词和已有词库进行比较, 就可以找出新词了[9–11].

2.1 凝合度

判断一个词是否可以成词的首要标准是该词的凝合度[12]. 例如在约 5900 万字的训练语料库中, 出现频率超过 150 的两个字段, “的设计”的出现频率比“设计感”要高, 但是在人们的认知中“设计感”才是一个词, 这就说明“设计”和“感”的结合程度更紧密, 但是“设计”和“的”的结合就没有达到人们认知中的紧密程度.

下面将通过计算证明“设计感”一词的内部凝固程度比“的设计”要高. 如果“设计”一词和“感”在文本中的出现是独立的, 并且是随机的, 那么计算这 3 个字被拼凑到一起的概率. 在该 5900 万字的语料库中, “设计”一词出现了 8491 次, 该词出现的概率约为 0.000 0143. “感”字在语料中出现了 59 448 次, 出现的概率约为 0.000 9321. 如果这两个字段之间随机且相互独立, 那么“设计感”一词出现的概率就应该是 0.000 0143×0.000 9321, 约为 1.33×10^{-7}. 但事实上, “设计感”一词在该语料中出现了 185 次, 出现概率约为 3.13×10^{-6}, 大约是预测值的 46 倍. 以此类推, 统计语料可发现“的”字的出现概率约为 0.0343, 因而“的”和“设计”随机组合到一起的概率应当为 0.0343×0.000 0143, 约为 4.904×10^{-6}, 出现的频次应当在 290 次, 这个数据与“的设计”在语料中出现的频次比较接近, 该字段出现的频次为 1816, 是预测值的 6.26 倍. 从以上的计算可以看出, “设计感”的结合更紧密, 该字段是一个有意义的搭配的可能性更大, 而“的设计”的出现则更有可能是“的”和“设计”这两个字段被随机拼接到一起的.

但是, 一个值得注意的问题是, 计算的过程中不存在先验知识. 换言之, “设计感”一词可能是由“设计”和“感”组合而成, 也可能是“设”和“计感”组合而成, 因此在计算的时候需要枚举一个字段的多种组合方法, 然后取概率最大的组合方式.

2.2 信息熵

一个词之所以成词除了其内部的凝合度之外, 还有一个标准就是该词外部的自由度[13]. 例如“辈子”的用法, 除了“一辈子”、“这辈子”、“上辈子”、“下辈子”等用法外, 在“辈子”前面添加字就没有很多选择了. “辈子”这个字段左边可以出现的字比较有限, 因此在

计算凝合度时, “辈子”并不单独成词, 真正成词的其实是“一辈子”、“这辈子”这样的整体.

因此需要添加“信息熵”的概念, 信息熵反映了获得一个事件的结果后会带来多大的信息量. 如果一个事件某个结果的发生概率为 p, 当该结果出现时, 将被得到的信息量就被定义为$-\log(p)$. p 的值越小, 得到的信息量就越大.

邻接熵是 Huang 等[14]提出的判断一个字串是否成词的重要统计量. 邻接熵统计量利用信息熵来衡量候选新词 t 的左邻字符和右邻字符的不确定性. 不确定性越高, 表明候选新词 t 的前后字符串越混乱, 越不稳定, 所以其成词的可能性就越高.

例如在实验用语料中, “杯子”一词一共出现了 1080 次, “辈子”一词一共出现了 4030 次, 两者的右邻字集合的信息熵分别为 4.7374 和 6.1655, 数值上是接近的. 但“杯子”的左邻字则用例非常丰富. 例如“加杯子”、“拿杯子”、“用杯子”、“新杯子”、“旧杯子”、“收杯子”、“摔杯子”等几十种不同的用法. 计算得出“杯子”所有左邻字的信息熵为 4.9745. 但“辈子”的左邻字则相对少了很多, 在语料库中出现的 4030 个“辈子”中有 3240 个是“一辈子”, 414 个“这辈子”, 261 个“下辈子”, 78 个“上辈子”, 除此之外还有“n 辈子”、“两辈子”等 15 种比较罕见的用法. 所有左邻字的信息熵仅为 1.3679.

除了左邻字外, 一些文本片段虽然左邻字用法很多, 右邻字用例却非常贫乏, 例如“国庆”、“托儿”、“鹅卵”等, 这些词单独成词也是不符合常理的.

因此, 一个短语或词的自由运用程度可以定义为这个短语的左邻字信息熵和右邻字信息熵中的较小值.

3 词向量分析工具 Word2Vec

使用词向量来表示词的方法很早之前就出现了, 一般叫做 1-of-N representation,也有叫独热表示法等等, 但是这种方法使用的维度是整个词汇表的大小. 对于词表中每个词, 将该词对应位置上的 0 置为 1.例如一个有 5 个词的词表, 第二个词 answer 的向量表示就是 (0, 1, 0, 0, 0), 第五个词 hungry 的向量表示就是 (0, 0, 0, 0, 1). 因此可以看出, 一个词汇表的词汇量一般非常大, 所以这种词汇向量的表达方式稀疏程度非常大, 表达效率也不高[15].

解决这个问题的方法是 Dristributed representation, 该方法是通过训练, 将每个词表示为一个较短的向量, 向量的每个维度表达一个语义信息[16], 但是这个向量每个维度具体表达什么意义的可解释性不好. 在 Word2Vec 出现前, 一般使用神经网络训练词向量从而处理词. 一般分为 CBOW (Continuous Bag-of-Words 与 Skip-Gram 两种模型[17–19]. CBOW 模型的输入是文本中一个词对应的上下文词的词向量, 而输出是该词的词向量. 例如句子片段“…distributed representations which encode the relevant grammatical relations…”上下文大小为 6, 输出词是“encode”, 那么输入就应当是“encode”的前 3 个词和后 3 个词的词向量. 需要说明的是, 这 6 个词是没有先后顺序的, 使用的是词袋模型. 而 Skip-Gram 模型和 CBOW 模型相反的, 其输入是一个词的词向量, 而输出是该词上下文词语的词向量. 如上例中, Skip-Gram 模型的输入为“encode”的词向量, 而输出则是“encode”上下文各 3 个词, 一共 6 个词的词向量. 而 Word2Vec 则使用哈夫曼数的数据结构代替了神经网络模型, 同样也分为 CBOW 和 Skip-Gram 两种模型.

首先分析 CBOW 模型. 第一步需要定义词向量的维度大小为 M, 以及该字段的上下文大小 $2c$, 这样对于训练样本中的每一个词, 其前面的 c 个词和后面的 c 个词就作为 CBOW 模型的输入, 输出为所有词汇的词向量 w. 算法步骤如算法 1.

算法 1. CBOW 模型算法

(1) 以训练语料为样本建立哈夫曼树;
(2) 随机初始化所有的模型参数 θ 和词向量;
(3) 进行梯度上升迭代过程, 对于训练集中的每一个样本 ($context(w)$, w) 做如下处理:
(3.1) e=0, 计算 $x_w = \frac{1}{2c}\sum_{i=1}^{2c} x_i$
(3.2) for j = 2 to l_w, 计算:
$f=\sigma\left(x_w^T\theta_{i-1}^w\right)$
$g=\left(1-d_i^w-f\right)$
$e=e+g\theta_{i-1}^w$
$\theta_{i-1}^w=\theta_{i-1}^w+gx_m$
(4) 对于 $context(w)$ 中的每一个词向量 X_i(共 $2c$ 个) 进行更新: $x_i=x_i+e$;
(5) 如果梯度收敛, 则结束梯度迭代, 否则返回第 (3) 步.

而对于 Skip-Gram 模型, 该模型的输入输出与 CBOW 模型相同, 训练算法如算法 2.

算法 2. Skip-Gram 模型算法

(1) 以训练语料为样本建立哈夫曼树;
(2) 随机初始化所有的模型参数 θ 和词向量;
(3) 进行梯度上升迭代过程, 对于训练集中的每一个样本 (w, $context(w)$) 做如下处理:

for i=1 to $2c$

e=0

for j = 2 to l_w, 计算:

$f=\sigma\left(x_i^T\theta_{i-1}^w\right)$

$g=\left(1-d_i^w-f\right)$

$e=e+g\theta_{i-1}^w$

$\theta_{i-1}^w=\theta_{i-1}^w+gx_i$

$x_i=x_i+e$

(4) 如果梯度收敛, 则结束梯度迭代, 否则返回第 (3) 步.

在 Word2Vec 中, 除了基于哈夫曼树的方法训练模型外, 还有基于负采样的方法[20]. 因为一个词如果过于生僻, 则哈夫曼树的查找层级就会比较多. 而采用负采样的方法时, 每次只是通过采样 n 个不同的中心词做负例, 就可以训练模型.

4 新词发现及情感倾向性分析方法

基于词向量的新词情感倾向性分析是首先利用上文中提及的凝合度、信息熵以及词频等计算量, 通过一定的阈值设定发现新词. 然后, 利用 Word2Vec 通过对训练语料的学习, 生成词表中所有词的词向量, 而后找出所有新词词表中与其相似度最高的几个词. 该新词发现及情感倾向分析方法的构建架构如图 1 所示.

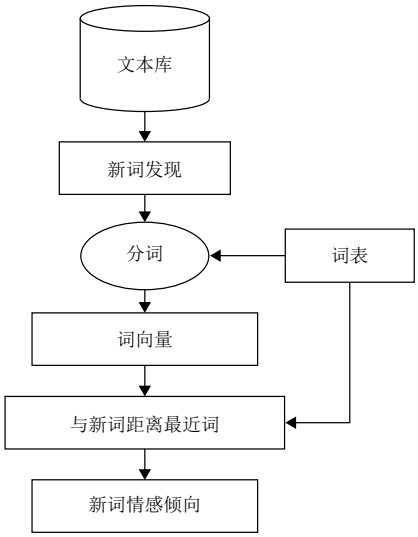

图 1 新词发现及情感倾向分析设计架构

该方法首先要将语料中所有字分成长度小于 5 的字段, 然后对字段进行计算凝合度以及信息熵. 计算完成后, 选择处于某阈值内的词作为新词. 将新词添加进分词词表中后, 对语料进行分词处理, 分词后训练所有词的词向量. 随后, 计算所有新词的中, 与每个词距离最近的前 n 个词. 最后进行情感倾向分析. 具体的构建过程如下:

步骤 1. 从待检索的文本库中通过对 xml 语言或 html 语言的分析, 将网页中的文本内容提取出来.

步骤 2. 对文本分成若干长度小于 5 的字段. 并计算每个字段的凝合度和信息熵. 根据实验得出的最佳阈值筛选得出新词表.

步骤 3. 对文本进行分词处理. 一般分词后要去除停用词, 但是词向量的学习要依据上下文, 停用词也会词其产生影响, 因此在这一步先不去除停用词.

步骤 4. 将分词后的文件利用 Word2Vec 进行训练, 并不断地调整参数以得到令人满意的结果. 训练结束得到所有词的词向量.

步骤 5. 找出新词表中与每个词距离最近的前 n 个词, 通过这些词的情感倾向分析新词的情感倾向.

该方法流程图如图 2 所示.

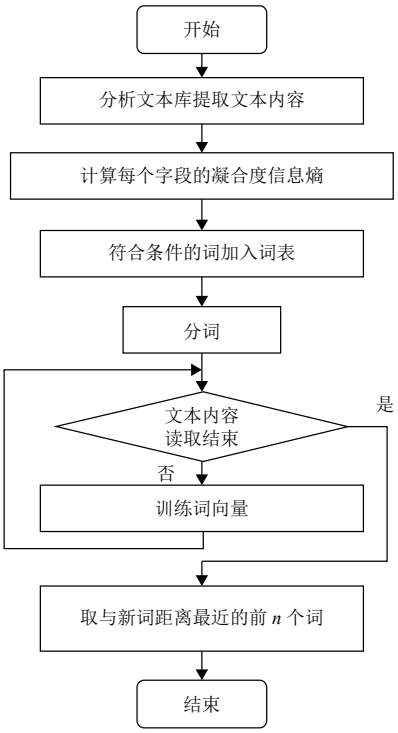

图 2 新词发现及情感倾向性分析方法流程图

5 实验及结果分析

为了检验提出的方法, 本文从新浪微博上抓取了 1200 万条数据, 主要分析内容为每条微博的文本内容作为语料库.

新词发现步骤中, 经过多次实验, 本文将新词获取

的凝合度阈值设定为 0.35, 信息熵设定为 0.5 至 1.5 之间. 添加的新词举例如表 1 所示.

表 1 新词发现举例

	频率	凝合度	信息熵
抓狂	647	0.808	0.79
屌丝	145	0.425	1.08
秒杀	109	0.332	0.91
闺蜜	98	0.685	0.88
坑爹	75	0.479	0.58

通过对模型训练时参数的调整, 除词向量维度外, 其余参数选定包括, 当前词与预测词在一个句子中的最大距离为 3, 使用 CBOW 算法. 词与词之间的距离计算公式为两向量的 Cosine 值:

$$similarity(A,B)=\cos(A,B)=\frac{A\cdot B}{\|A\|\cdot\|B\|} \quad (1)$$

程序输入为以上举例中的 5 个新词, 输出为训练模型中与每个词相似度最大的 4 个词, 当调整模型词向量空间维度为 100 时, 输出结果如表 2 所示. 每个新词对应 4 个距离最近的词, 对每个新词分别列出了这 4 个词以及他们之间的距离.

表 2 情感倾向分析结果

词距	词 1	词 2	词 3	词 4
抓狂	泪 0.7356	可怜 0.7038	衰 0.7027	晕 0.6737
屌丝	美女 0.5975	帅哥 0.5681	极品 0.5239	基佬 0.5068
秒杀	打折 0.5105	竞拍 0.4987	货 0.1702	倒 0.4611
闺蜜	朋友 0.6877	老友 0.6133	亲人 0.6098	姐妹 0.5939
坑爹	矛盾 0.6242	衰 0.6143	可怜 0.6070	泪流满面 0.5985

从表 2 的结果可以看出, 经过词向量训练的新词可以从与其相近的词中分析其情感倾向. 例如抓狂表达的情感与泪、可怜、衰等都是相近的. 同时, 通过两个词之间不同的距离也可以看出近似关系的远近.

6 结论与展望

为更具体表义社会新词的情感含义及其倾向性, 本文提出了一种基于词向量的新词情感倾向性分析方法. 本文在分析了新词发现方法和词向量训练工具 Word2Vec 的基础上, 研究了基于 Word2Vec 的新词情感倾向性分析方法的可行性和架构设计, 并面向微博语料进行实验. 从实验结果可以看出该方法具有较好的可行性和可以信服的结果, 但是在具体的新词情感倾向性分类上没有进行, 因此还有很多待完善的细节.

对情感词的新词发现和倾向性分析是为了更好理解用户通过文本表达的情感, 也是为中文分词的未登录词挖掘提供了一种探索的方法. 虽然该方向的研究仍存在诸多困难, 但是在不断深入创新的过程中必会取得令人满意的效果.

参考文献

1 肖江, 丁星, 何荣杰. 基于领域情感词典的中文微博情感分析. 电子设计工程, 2015, 23(12): 18–21. [doi: 10.3969/j.issn.1674-6236.2015.12.006]

2 Jo Y, Oh AH. Aspect and sentiment unification model for online review analysis. ACM International Conference on Web Search and Data Mining. Hong Kong, China. 2011. 815–824.

3 杨立月, 王移芝. 微博情感分析的情感词典构造及分析方法研究. 计算机技术与发展, 2019, (2): 1–6. http://kns.cnki.net/kcms/detail/61.1450.tp.20181115.1046.008.html.

4 Liu SH, Li FX, Li FT, *et al*. Adaptive co-training SVM for sentiment classification on tweets. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, CA, USA. 2013. 2079–2088.

5 Dey L, Chakraborty S, Biswas A, *et al*. Sentiment analysis of review datasets using naive Bayes and K-NN classifier. arXiv: 1610.09982, 2016.

6 郭胜国, 邢丹丹. 基于词向量的句子相似度计算及其应用研究. 现代电子技术, 2016, 39(13): 99–102, 107.

7 唐明, 朱磊, 邹显春. 基于 Word2Vec 的一种文档向量表示. 计算机科学, 2016, 43(6): 214–217, 269. [doi: 10.11896/j.issn.1002-137X.2016.06.043]

8 冯冲, 石戈, 郭宇航, 等. 基于词向量语义分类的微博实体链接方法. 自动化学报, 2016, 42(6): 915–922.

9 杜丽萍, 李晓戈, 于根, 等. 基于互信息改进算法的新词发现对中文分词系统改进. 北京大学学报 (自然科学版), 2016, 52(1): 35–40.

10 张佳明, 席耀一, 王波, 等. 基于词向量的微博事件追踪方法. 计算机工程与应用, 2016, 52(17): 73–78, 117. [doi: 10.3778/j.issn.1002-8331.1412-0144]

11 王馨, 王煜, 王亮. 基于新词发现的网络新闻热点排名. 图书情报工作, 2015, 59(6): 68–74.

12 张剑, 屈丹, 李真. 基于词向量特征的循环神经网络语言模型. 模式识别与人工智能, 2015, 28(4): 299–305.

13 李文坤, 张仰森, 陈若愚. 基于词内部结合度和边界自由度

的新词发现. 计算机应用研究, 2015, 32(8): 2302–2304, 2342. [doi: 10.3969/j.issn.1001-3695.2015.08.015]

14 周练. Word2Vec 的工作原理及应用探究. 科技情报开发与经济, 2015, 25(2): 145–148. [doi: 10.3969/j.issn.1005-6033.2015.02.061]

15 霍帅, 张敏, 刘奕群, 等. 基于微博内容的新词发现方法. 模式识别与人工智能, 2014, 27(2): 141–145. [doi: 10.3969/j.issn.1003-6059.2014.02.007]

16 陈飞, 刘奕群, 魏超, 等. 基于条件随机场方法的开放领域新词发现. 软件学报, 2013, 24(5): 1051–1060. [doi: 10.3724/SP.J.1001.2013.04254]

17 Huang JH, Powers D. Chinese word segmentation based on contextual entropy. Proceedings of the 17th Asian Pacific Conference on Language. Sentosa, Singapore. 2003. 152–158.

18 Ghosh M, Sanyal G. Performance assessment of multiple classifiers based on ensemble feature selection scheme for sentiment analysis. Applied Computational Intelligence and Soft Computing, 2018, 2018: 8909357.

19 Mondal A, Cambria E, Das D, *et al*. Relation extraction of medical concepts using categorization and sentiment analysis. Cognitive Computation, 2018, 10(4): 670–685. [doi: 10.1007/s12559-018-9567-8]

20 Zhu YJ, Yan EJ, Wang F. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of Word2Vec. BMC Medical Informatics and Decision Making, 2017, 17: 95. [doi: 10.1186/s12911-017-0498-1]