基于文档分布式表达的新浪微博情感分类研究*

杨宇婷[1] 王名扬[1] 田宪允[2] 李鹏宇[2]

(1. 东北林业大学信息与计算机工程学院 哈尔滨 150040;
2. 哈尔滨工业大学经济与管理学院 哈尔滨 150001)

摘 要 [**目的/意义**]拥有庞大用户群体的新浪微博每天都产生海量的文本数据,对其进行情感分类有助于分析社会的舆论走向,为舆情监测提供帮助。其中,如何挖掘微博中的文本特征与情感信息是微博情感分类研究的关键。[**方法/过程**]将能有效考察上下文语境的基于文档分布式的特征表达方法引入到微博情感分类研究中,通过综合考虑上下文的语义、语序和情感信息,将微博文本转化为高维空间的特征向量,然后利用SVM分类器判断文本的情感极性。[**结果/结论**]实验表明,对微博文本进行文档分布式特征表达后,其分类准确率可达90.46%,优于其他特征表达方法。

关键词 微博 情感分类 文档分布式表达 Doc2vec

中图分类号 TP391 **文献标识码** A **文章编号** 1002-1965(2016)02-0151-06

引用格式 杨宇婷,王名扬,田宪允,等. 基于文档分布式表达的新浪微博情感分类研究[J]. 情报杂志,2016,35(2):151-156.

DOI 10.3969/j.issn.1002-1965.2016.02.027

Sina Microblog Sentiment Classification Based on Distributed Representation of Documents

Yang Yuting[1] Wang Mingyang[1] Tian Xianyun[2] Li Pengyu[2]

(1. College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040;
2. School of Management, Harbin Institute of Technology, Harbin 150001)

Abstract [**Purpose/Significance**] Sina Microblog produces a large amount of texts every day, the sentiment classification based on these data is meaningful in analyzing and monitoring public opinion. It's crucial for the microblogging sentiment classification research to mine the text characteristics and sentiment information. [**Method/Process**] A sentiment classification method based on distributed representation of documents is proposed, allowing a distributed representation of texts with the consideration of the contexts, the semantics, the word order and the syntactic structure of Chinese texts being introduced in the analysis, thus helps transfer the microblogging texts into vectors in higher space, then a SVM classification tool is used to facilitate judging the sentiment polarization of the texts. [**Result/Conclusion**] The result of classification accuracy of 90.46% shows the superiority of the method proposed to other methods.

Key words microblog sentiment classification distributed representation of documents Doc2vec

收稿日期:2015-12-11 修回日期:2016-01-07

基金项目:**中央高校基本科研业务费专项资金项目"基于社会网络特征提取的群体性突发事件预警方法研究"(编号:2572014DB05);国家自然科学基金项目"群体性突发事件预警的超网络方法研究"(编号:71473034)**。

作者简介:**杨宇婷**(ORCID:0000-0002-8799-6407),女,1990年生,硕士研究生,研究方向:社交网络挖掘;**王名扬**(ORCID:0000-0002-5022-6628),女,1980年生,博士,副教授,研究方向:数据挖掘、社交网络挖掘;**田宪允**(ORCID:0000-0002-3889-4287),男,1988年生,博士研究生,研究方向:数据挖掘、自然语言处理;**李鹏宇**(ORCID:0000-0001-5724-889X),男,1989年生,博士研究生,研究方向:数据挖掘、自然语言处理。

通讯作者:**王名扬**

0 引 言

作为中国最大的社交媒体平台,新浪微博在 2015 年 2 月已拥有超过 1 亿活跃用户。通过对含有主观感情色彩的中文微博进行情感分类研究,可分析社会舆论的走向与网民对于社会事件的情感倾向,实现政府及相关部门对网络异常和社会舆情的监制。另外,中文微博情感分类的结果还可以帮助商家了解消费者的需求,改进产品和营销策略,更好地服务用户[1-2]。

国内外现有的文本情感分类技术主要有:a. 基于规则的方法。主要是指根据对数据集的观察制定出一些规则,而后用其对文本进行分类。Turey 等人利用文本中一些短语具有较强情感极性的特点提出了一种互信息的方法进行文本分类[3]。b. 基于机器学习的方法。主要是指利用机器学习方法进行建模,再利用模型将文本分类问题转换成常见的分类问题。文献[4]首次采用机器学习,将特征分类的方法用于篇章级的情感分类中,对中文文本的情感分类具有借鉴作用。Pang 等人将机器学习的方法与词袋模型相结合,应用于情感分析领域[5],由于传统的词袋模型忽略了文本的语义和语序,故分类效果没有达到期望的准确度。张成功等人提出基于极性词典的情感分析方法[6],但是该方法不仅忽略了词语在特定环境下的语义变化,且无法解决未登录词的问题。赵吉昌等在中文微博中利用表情符来进行情感分类[7],但未考虑不含表情符的微博文本。

2014 年,Mikolov 提出了一种较新颖的可将句子或段落直接转化为固定维度向量的文档分布式表达的方法“Doc2vec”[8]。由于该方法能很好地结合上下文语境、词语和段落中的语义和情感信息,减少忽略语序、词语歧义等问题对分类结果造成的影响,Doc2vec 已被运用到了 twitter 的情感分类研究中,并取得了良好的效果[9]。这为解决中文文本情感分类过程中同样存在的未考虑词序、词义及上下文逻辑等问题提供了解决思路。本文尝试将基于文档分布式表达的方法运用到中文文本情感分类研究中,探讨其对中文文本分类效果的影响。

在研究过程中,利用基于文档分布式表达的方法,以及常用的三种文本特征表达方法,分别实现微博文本的特征表达。然后利用分类器对表达好的文本进行情感极性分类,并通过分类准确率来探究基于文档分布式的特征表达方法对中文文本情感分类的效果和适用性。同时,本文通过分析对比采用不同规模训练数据集和设置不同维度得出的大量实验结果,找到该方法实现中文文本特征表达的最佳维度,以及训练数据集大小对结果产生的影响,为研究者们今后进一步探讨该方法的应用提供借鉴。

1 研究方法

本文研究框架与方法如图 1 所示。首先对从新浪微博中爬取的文本数据进行人工标注,分为“积极”与“消极”两类。在对数据集进行预处理后,利用基于文档的分布式表达,以及三种常用的文本特征表达方法(基于词频、基于表情符和基于情感词典的方法),将已标注的微博文本进行特征表达,形成等维度的输入向量,通过分类器 SVM 来对比四种方法的实验效果。

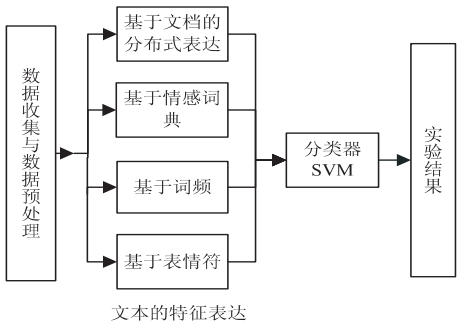

图 1 研究方法与框架

1.1 数据收集与预处理

1.1.1 数据收集 本文通过新浪微博开放平台 API 获取 2013 年 10 月 1 日至 2014 年 4 月 20 日间随机用户的随机微博文本,建立了两个分别包含 7 万条和 2 000 万条微博文本的数据集。然后,从 7 万条微博文本中标注出 2 500 条为“积极”和 2 500 条为“消极”的文本,用来测试分类器。另外,从 2 000 万条微博文本中分别随机抽取 8 000、4 万、20 万、100 万条微博文本构成 4 个训练数据集,用于训练 Doc2vec 的两种模型,对比用不同规模数据集训练模型对分类结果的影响。

1.1.2 数据预处理 相比其他文本,社会媒体中的微博文本具有短小且语法不规范等特点,包含许多无用的信息,所以需要对数据进行预处理,降低无用信息对实验的干扰。本文的数据预处理主要为降噪、分词处理。

a. 数据降噪:除去 URL 链接、用户昵称、地点信息,并将微博文本中所有的繁体中文转化为简体。

b. 分词处理:本文采用的分词工具为 mmseg4j,它是由谷歌发布,利用 Chih-Hao Tsai 的 MMSeg 算法实现的中文分词器,是目前分词效率与准确率较高的分词工具。

1.2 文本的特征表达

经过预处理后的文本数据长短参差不齐,内容零散,需要通过特征表达的方式来对其进行规范化处理。因此,文本的特征表达是决定分类效果的关键。

2014 年,Mikolov 在基于词语分布式表达的方

法——“Word2vec[10]”的基础上加以改进,提出基于文档分布式表达的方法,又称为“Doc2vec[8]”。由于该方法能很好地结合上下文语境,挖掘语义、语法和情感信息,故本文将其引入进来,作为微博文本特征表达的方法,来探讨其在中文文本情感分类中的效果。

1.2.1　基于文档分布式表达的方法——Doc2vec

Doc2vec 是利用无监督的训练方法来获得任意长度文本向量,如语句、段落、文档,它能包含上下文中大量的语义和语序信息。Doc2vec 主要包含两种模型:DM(Distributed Memory Model)和 DBOW(Distributed Bag of Words),均以神经网络语言模型[11]为基础,去掉隐含层,利用上下文和段落特征来预测某词语出现的概率分布,段落向量与词向量[10]是训练过程的副产物。

DM 模型是利用段落向量和词向量相结合来预测下一个词的向量。假设给定前 n-1 个词 $w_{t-n+1}, w_{t-n+2}, \cdots, w_{t-1}$ 和其所在的段落 d,w_t 为要预测的词,结合统计语言模型[12]的知识,可知 DM 模型要使目标函数(1)最大化。

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c, j\neq 0}\log P(w_t \mid w_{t+j}; d) \tag{1}$$

其中,T 为文本中词语的数量,c 为滑动窗口大小。DM 模型结构如图 2 所示。

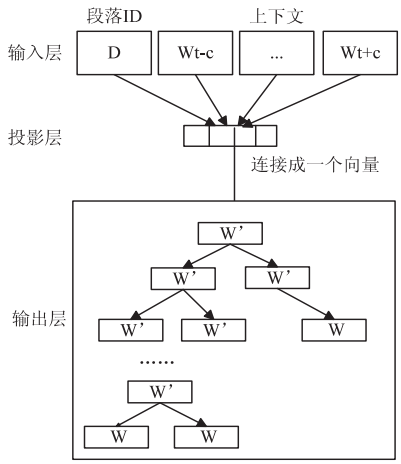

图 2　DM 模型结构

1)在输入层,将 $w_{t-n+1}, w_{t-n+2}, \cdots, w_{t-1}$ 分别映射成 $n-1$ 个词向量,所有词向量存储在词矩阵 C 中;将每一段文本映射成为一个唯一的段落向量,记为段落 ID,用来记录滑动窗口之外的段落信息,所有段落向量存储于段落矩阵 D 中。两种向量初始化均随机,向量维度需人工设置。由于该方法从未用于中文文本情感分类研究中,故没有经验值可借鉴,本文实验中设置了不同维度来进行对比分析。

2)在投影层,将词向量与段落向量相加或首尾连接(本文采用首尾连接,其在文献[8]的实验中取得了较高准确率),用于预测 W_t。

3)在输出层,用层次 Softmax[11]进行输出分类,如(2)式,求得 W_t 的概率分布。同时,通过图 2 中神经网络的反向传播[14]获得误差梯度,再运用随机梯度下降法,迭代更新模型中的参数以及段落向量与词向量。

$$P(w_t \mid w_{t-c}, \cdots, w_{t+c}; d) = \frac{e^{y_{(w_t,d)}}}{\sum_i e^{y_i}} \tag{2}$$

其中,输出层的节点对应 C 中的所有词语,而每一个 y_i 对应词 i 未标准化的对数概率:

$$y = b + Uh(w_{t-c}, \cdots, w_{t+c}, d; C, D) \tag{3}$$

其中,b 和 U 均为 Softmax 参数,h 表示将词向量与段落向量相连接。

为了加速训练过程,在输出层以词语在语料库中的词频作为权值构造的一棵哈夫曼 Huffman 二叉树,最大限度地忽略与待预测词特征无关的词语。叶子节点为词汇表中的所有词语,其对应的向量即为词向量。非叶子结点代表一类词语共有的特征。从根节点查找到该词的过程可被看成一个连续的二分类问题[13]。该词在给定上下文环境中出现的概率即为二分类概率的乘积。训练模型时,每一次窗口滑动,都先选取一段固定长度的文本,用上述方法获得待预测词向量的概率,再通过神经网络的反向传播与随机梯度下降法更新段落向量和词向量。

由于在输入层段落向量被看成一个词向量,与上下文词向量相连接,故每次预测下一个词时都利用了整个段落的语义和语法特征。段落向量只能被运用在同一篇文档中,而词向量可用于所有的文档中。若要预测新的段落向量,也借助梯度下降法,但在此过程中,模型中的参数和词向量不更新。具体训练过程请参见文献[8]。

Doc2vec 方法中的另一种模型——DBOW 是利用已知的段落向量 d 来预测同一窗口中的其他词语,即将下列目标函数最大化。

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c, j\neq 0}\log P(w_{t+j} \mid d) \tag{4}$$

DBOW 与 DM 训练方法基本一致,如图 3 所示,在仅给定段落向量的情况下预测段落中一组随机单词出现的概率,并用层次 Softmax 函数表示。同样通过随机梯度下降法和神经网络的反向传播不断更新模型中的参数以及段落向量与词向量。与 DM 不同的是,DBOW 的输入层仅为一个段落向量,而输出层为多个词向量的概率分布,且在训练过程只用存储 Softmax 参数,相比 DBOW 模型还需要存储词向量来说,节省了存储空间。

利用训练数据集训练好 DM 和 DBOW 模型后,通过已有的词向量和随机梯度下降法来预测本文中已标注的微博文本的段落向量,从而实现一段文本的特征表达。文本用同样的语料集同时训练这两种模型,将

两个模型中同一段落 ID 对应的向量在维度上进行首尾拼接,得到最终的段落向量。这样的方法在文献[8]中取得了较高的准确率。

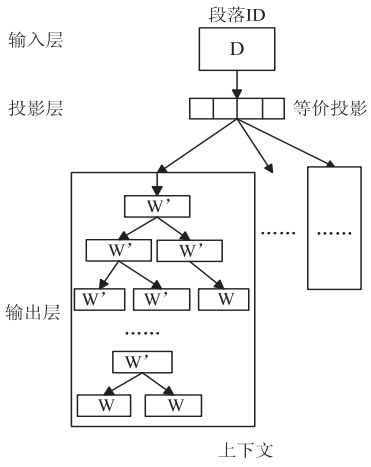

图 3 DBOW 模型结构

在中文文本的情感分类中,词语由于所处环境不同而会有不同的语义,而词语出现的顺序则体现了语言的逻辑性,一词多义或语序颠倒都会直接影响语句或段落的情感倾向。因此,在本文研究过程中,将 Doc2vec 引入到中文文本特征表达中,借鉴其能很好的考虑上下文语境的特点,提高中文文本情感分类中的应用效果。

为了更好地衡量基于文档分布式表达方法的分类效果,文中选取基于情感词典[15]、基于表情符[7]和基于词频[16]的三种常用的文本特征表达方法作为对比。

1.2.2 三种情感分类中常用的特征表达方法

在基于情感词典的特征表达方法中,本文采用的是《台湾大学情感词典(2006)》,该词典共包含 2 810 个积极词和 8 276 个消极词,经过分词后的微博文本对照该词典可确定文本中的词是否出现在词典中,从而被转化为固定维度的向量。

基于表情符的方法是将微博中的表情符号看作情感的标志。本文将新浪微博文本中由 234 个表情符号转化成的中文词语构成"表情符词典",根据表情符词典将每条微博表示为固定维度的向量。

基于词频的方法则选取在一类文档中出现次数最多的前 N 个词语构成词典。根据词语是否出现而在相应位置标记,将每一条微博文本转化成固定维度的向量。

对文本进行特征表达后,需要将表达好的文本向量输入分类器,实现文本的情感分类。本文采用常用的 SVM 分类器来探讨四种特征表达方法的分类效果。

1.3 **分类器** 支持向量机(SVM)的目的是构建距离点间隔最大超平面来区分高维空间中的点[17],其原理可以简化为找出一个超平面,将数据集$\{x_i, y_i\}$($i = 1,2,3,\cdots,n$, $x_i \in \mathrm{Rd}$, $y_i \in [-1,1]$)中的x_i归类。在 Rd 空间中的点必须满足以下条件。

$$\begin{cases} \min\limits_{w,b} \dfrac{1}{2} w^T w \\ y_i((w^T x_i) + b) \geqslant 1 \end{cases} \tag{5}$$

支持向量机在归纳和直推方法中都可以显著减少所需要的有类标的样本数,已广泛运用于文本和超文本的分类[18],且取得了良好的效果。

2 实验结果

2.1 **三种传统的文本特征分类表达结果** 本实验将标注好的数据利用常见的三种特征表达方法进行文本的特征表达,通过 SVM 分类器确定情感极性,用分类准确率衡量分类效果,并用十折交叉的方法验证分类的精确程度。

在基于词频的特征表达中,我们统计了被标注文本中所有词语的出现次数,选择出现次数最多的 1000 个词语作为特征维度(在前人的研究中取得了较好的实验效果),在其出现的位置进行标记,并将被标注的文本进行特征表达,形成输入向量。基于情感词典的特征表达方法则统计了被标注数据集中出现在《台湾大学情感词典(2006)》中的 1 363 个情感词作为特征词,用来将文本转化为输入向量。而基于表情符的特征表达方法则从被标注数据集中选取 234 个表情符作为特征符,将文本转化为输入向量。

将 5 000 条已标注好的微博文本通过上述三种特征表达方法转化为固定维度的特征向量,作为分类器的输入向量。本文使用台湾大学林智仁副教授开发的 LIBSVM 包[19]作为实现 SVM 分类的工具包,选择径向基函数[17]作为核函数,gamma 参数设为 0.01,惩罚因子 C 设为 0.7。运用十折交叉法,将不同的输入向量作为分类器中的训练集与测试集,利用 LIBSVM 提供的工具 svm-train.exe 对训练文本集训练出分类器,然后利用 svm-predict.exe 工具对测试文本集进行测试,得到分类结果后求得平均分类准确率,结果如图 4 所示。基于词频的特征表达方法取得了 84.54% 的准确率,基于情感词典和基于表情符的方法效果分别取

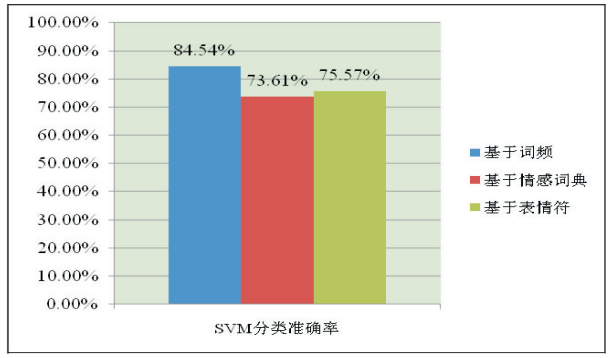

图 4 三种特征表达方法的分类结果

得了 73.61% 和 75.57% 的准确率。

2.2　基于文档分布式表达的中文文本分类结果

在基于文档分布式表达的特征表达方法中,实验机器配置为 Win7 系统,内存 16G,利用从 2 000 万条微博文本中随机抽取的 8000、4 万、20 万、100 万条微博文本(分别对应 DS1、DS2、DS3、DS4),来同时训练 Doc2vec 的 DM 和 DBOW 模型,目的是对比不同规模的数据集对文档分布式特征表达效果的影响。同时,为探究 Doc2vec 在中文微博情感分类实验中特征表达效果最佳的维度,本文设置了 50、100、200、300、400、500、600、700、800、900、1000 维度。然后将两种模型得到的同一段落的段落向量在维度上进行首尾拼接,变为 100、200、400、600、800、1000、1200、1400、1600、1800、2000 维,形成最终的段落向量[9]。最后将标注好的 5000 条微博文本分别利用不同数据集训练好的模型转化为不同维度的特征向量输入到 SVM 分类器中,测试基于文档分布式特征表达的方法对文本分类的准确率,结果如表 1。

表 1　基于文档分布式特征表达的新浪微博文本分类结果

数据集	100	200	400	600	800	1000	1200	1400	1600	1800	2000
DS1	86.80	86.44	86.42	86.44	86.71	86.25	86.23	86.60	86.62	86.50	86.08
DS2	86.82	86.49	86.40	86.47	86.42	86.43	86.33	85.43	86.41	86.40	86.47
DS3	88.52	87.41	85.87	85.92	85.45	83.36	83.48	81.24	81.47	80.12	80.49
DS4	90.46	87.76	87.47	86.48	84.21	71.49	68.79	66.12	66.56	66.32	66.47

由表 1 可知,通过分类器后,在训练数据集为 100 万、维度为 100 维时取得了最高准确率 90.46%。在同一数据集上,最高准确率均在维度为 100 维时取得,说明两种模型均设为 50 维时特征表达效果最佳。且在 100 维时,训练模型的数据集越大,分类准确率越高。同时,我们发现,对同一个数据集,随着文本特征表达维度的增高,分类准确率均表现出一定的下降趋势,尤其是当训练数据集增大为 20 万(对应 DS3)、100 万(对应 DS4)时,用同一数据集训练模型后对应的分类准确率随着维度的增加呈明显的下降趋势。分析原因如下:维度越高,虽能更好的表达对应数据集的文本特征,但由此也容易带来对数据集文本特征的过拟合问题,使得在过高的维度下训练模型后对未知文本进行分类时产生了干扰。

在实验过程中,我们还发现训练 Doc2vec 中的两种模型是整个实验中耗时较长的部分。其时间复杂度与训练数据集和维度有着密切的关系,数据集越大,维度越高,则训练模型时间越长。这是因为模型中的需要调整的参数个数为:N * P+M * Q,它与语料大小及维度大小呈正相关,其中 N 为数据集中的段落数,M 为词的个数,P 为段落向量的维度,Q 为词向量的维度。当数据集增大至 100 万,维度为 2000 维时,实验在前文描述的实验配置上耗时达到最长,为 1.5 小时。在今后的实验中,为减少实验时长,可考虑搭建集群来用大规模语料训练模型。

表 2　四种文本特征表达方法分类结果准确率

特征工程	准确率(%)
基于词频的方法	84.54
基于情感词典的方法	73.61
基于表情符的方法	75.57
基于文档分布式表达的方法	90.46

由表 2 可知,相比于其他常用的特征表达方法而言,基于文档分布式表达的方法取得了最好的分类效果。其最高准确率较基于词频的方法上升 5.92%,较基于情感词典的方法上升 16.85%,较基于表情符号的方法上升 14.89%。由于其较基于词频的方法的优势不如较其他两种明显,故对比分类实验中取得最高准确率时的标准差,以分辨两种方法的稳定性。

表 3　基于词频与基于文档分布式表达的文本分类结果对比

文本特征表达方法	最高准确率/%	标准差
基于词频的方法	84.54	0.05
基于文档分布式表达的方法	90.46	0.04

由表 3 可知,中文微博文本通过基于文档分布式表达的方法进行特征表达后,输入分类器 SVM 所取得最高准确率较基于词频的方法高 5.92%,且同时标准差低 0.01,这说明使用该方法的稳定性要更强,即优于基于词频的方法。

3　结　语

针对中文文本情感分类研究中未全面考察文本上下文的语义、语序和情感信息的问题,本文将能很好地考虑上下文语境的基于文档分布式表达的方法引入进来,探讨其在中文文本情感分类研究中的适用性。

通过研究,发现相对于传统的基于词频、基于情感词典和基于表情符的文本特征表达方法,基于文档分布式表达的方法能更加有效地实现对中文文本的情感分类。同时,通过大量的实验分析,本文也得到了关于文档分布式的表达方法用于中文文本情感分类的经验:针对不同规模的数据集,DM 和 DBOW 模型均设为 50 维,也即两个模型段落向量首尾拼接 100 维的特征表达维度下,取得最佳的文本分类效果。且在该维度下,训练数据集越大,对应的分类效果越好。而这一

结论可作为经验值用于今后的研究中。

参 考 文 献

[1] 李光敏,许新山,张 磊. 微博中产品意见挖掘研究[J]. 情报杂志,2014,33(4):135.

[2] Pang Bo, Lillian Lee. Opinion Mining and Sentiment analysis [C]//Foundations and Trends in Information Retrieval, 2(1-2):1-135.

[3] Turney P D. Thumbs up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews[C]// Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002, 417-24

[4] 赵妍妍,秦 兵,刘 挺. 文本情感分析[J]. 软件学报,2010,21(8):1834-1848.

[5] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[6] 张成功,刘培玉,朱振芳,等. 一种基于极性词典的情感分析方法[J]. 山东大学学报,2012,47(3):47-50.

[7] Zhao J, Dong L, Wu J and Xu K. Moodlens: an Emoticon-based Sentiment Analysis System for Chinese Tweets[C]// Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, 1528-31.

[8] Quoc L E, Mikolov T. Distributed Representations of Sentences and Documents[EB/OL]. [2014-05-22]. http://arxiv.org/abs/1405.4053v2.

[9] H Liang, R Fothergill, T Baldwin. RoseMerry: A Baseline Message-level Sentiment Classification System[C]// Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 551-555, Denver, Colorado, June 4-5, 2015.

[10] Mikolov Tomas, Chen Kai, Greg Corrado, et al. Efficient Estimation of Word Representations in Vector Space[EB/OL]. [2013-09-07]. http://arxiv.org/abs/1301.3781v3.

[11] Bengio Y, Ducharme R, Vincent P. A neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3(7):1137-1155.

[12] P R Clarkson, A J Robinson. Language Model Adaptation using Mixtures and and Exponentially Decaying Cache. In Proceedings IEEE ICASSP, 1997.

[13] 周 练. Word2vec 的工作原理及应用探究[J]. 情报科技开发与经济,2015(25):2.

[14] 孙玲芳,周加波,林伟健,等. 基于 BP 神经网络和遗传算法的网络舆情危机预警研究[J]. 情报杂志,2014,33(11):18.

[15] Yuan B, Liu Y, Li H, et al. Sentiment Classification in Chinese Microblogs: Lexicon-based and Learning-based Approaches. International Proceedings of Economics Development & Research, 2013; 68.

[16] Yang S M. Relative Term-frequency Based Feature Selection for Text Categorization[C]//Machine Learning and Cybernet, 2002-ieeexplore.ieee.org.

[17] Thorsten Joachims. Making large scale SVM learning practical. Research Reports of the unit no[J]. VIII (AI) Computer Science Department of the University of Dortmund ISSN 0943-4135.

[18] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features[M]. Lecture Notes in Computer Science 137-142.

[19] Chih Chung Chang, Chih-Jen Lin. LIBSVM: a Library for Support Vector Machines[EB/OL]. [2015-12-29]. shttp://www.csie.ntu.edu.tw/~cjlin/libsvm.

(责编:贺小利)

(上接第 109 页)

Guilford Press, 1994.

[32] STERN N. Age and Achievement in Mathematics: A Case-study in the Sociology of Science[J]. Social Studies of Science, 1978, 8(1):127-140.

[33] Simonton D K. Creative Productivity and Age: A Mathematical Model Based on a Two-step Cognitive Process[J]. Developmental Review, 1984, 4(1):77-111.

[34] Simonton D K. Creative Productivity: A Predictive and Explanatory Model of Career Trajectories and Landmarks[J]. Psychological Review, 1997, 104(1):66.

[35] Kuhn T S. The Structure of Scientific Revolutions[M]. University of Chicago Press, 1962.

[36] Cole S. Age and Scientific Performance[J]. American Journal of Sociology, 1979:958-977.

[37] Dennis W. Age and Productivity Among Scientists[J]. Science as a Carreer Choice: Theoretical and Empirical Studies, 1973:469.

[38] Wray K B. Is Science Really a Young Man's Game[J]. Social Studies of Science, 2003, 33(1):137-149.

[39] Wray K B. An Examination of the Contributions of Young Scientists in New Fields[J]. Scientometrics, 2004, 61(1):117-128.

[40] Kyvik S, OLSEN T B. Does the Aging of Tenured Academic Staff Affect the Research Performance of Universities[J]. Scientometrics, 2008, 76(3):439-455.

[41] Merton R K. The Sociology of Science: Theoretical and Empirical Investigations [M]. University of Chicago Press, 1973.

[42] 危怀安,钟书华. 国家科技奖励获奖人员的年龄结构分析[J]. 科技进步与对策,2008,25(1):180-182.

[43] Beard G M. Legal Responsibility in Old Age[J]. Russells, New York, 1874:5-42.

[44] Kyvik S. Age and Scientific Productivity. Differences Between Fields of Learning[J]. Higher Education, 1990, 19(1):37-55.

[45] Price R L, Thompson P H, Dalton G W. A Longitudinal Study of Technological Obsolescence[J]. IEEE Engineering Management Review, 1978, 3(6):45-51.

[46] Stephan P, Levin S. Age and the Nobel Prize Revisited[J]. Scientometrics, 1993, 28(3):387-399.

[47] Levin S G, Stephan P E, Walker M B. Planck's Principle Revisited: a Note[J]. Social Studies of Science, 1995:275-283.

(责编:刘影梅)