ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 3,May 2017,pp. 629－635
DOI:10.16337/j.1004-9037.2017.03.023

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

一种基于 LDA 主题模型的评论文本情感分类方法

王　伟[1]　周咏梅[1,2]　阳爱民[1,2]　周剑峰[3]　林江豪[4]

(1.广东外语外贸大学思科信息学院,广州,510006;2.广东外语外贸大学语言工程与计算实验室,广州,510006;3.广东外语外贸大学图书馆,广州,510006;4.广东外语外贸大学财务处,广州,510420)

摘　要:针对互联网出现的评论文本情感分析,引入潜在狄利克雷分布(Latent Dirichlet allocation,LDA)模型,提出一种分类方法。该分类方法结合情感词典,依据指定的情感单元搭配模式,提取情感信息,包括情感词和上、下文。使用主题模型发掘情感信息中的关键特征,并融入到情感向量空间中。最后利用机器学习分类算法,实现中文评论文本的情感分类。实验结果表明,提出的方法有效降低了特征向量的维度,并且在文本情感分类上有很好的效果。

关键词:评论文本;情感单元;潜在主题;情感分析;机器学习

中图分类号:TP391　　**文献标志码**:A

Method of Sentiment Analysis for Comment Texts Based on LDA

Wang Wei[1], Zhou Yongmei[1,2], Yang Aimin[1,2], Zhou Jianfeng[3], Lin Jianghao[4]

(1. Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou, 510006, China;2. Laboratory for Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou, 510006, China;3. Library, Guangdong University of Foreign Studies, Guangzhou, 510006, China;4. Financial Department, Guangdong University of Foreign Studies, Guangzhou, 510420, China)

Abstract: A method of sentiment analysis for online comment texts is proposd based on the latent Dirichlet allocation (LDA) model. The method extracts the sentiment information containing sentiment words and context with the sentiment word dictionary according to the specified collocation patterns of sentiment unit. Use the LDA model to mine the key features of the sentiment information and then combine them into the sentiment vector space. The machine-learning algorithm is used to classify the sentiment polarity of Chinese comment texts. After experiment, the presented method is proved to be effective in reducing dimensionality and text sentiment classification.

Key words: comment text; sentiment unit; latent topic; sentiment analysis; machine learning

引　言

互联网的蓬勃发展方便了网民观点的表达与传播,导致出现了大量主观性的在线文本信息。这些

基金项目:国家社会科学基金(12BYY045)资助项目;教育部"新世纪"优秀人才支持计划(NCET-12-0939)资助项目;广东省教育厅科技创新(2013KJCX0067)资助项目;广州市社会科学规划(15Q16)资助项目;广东外语外贸大学研究生科研创新(14GWCXXM-36)资助项目;广东外语外贸大学校级(14Q3)资助项目;广东省普通高校青年创新人才类(299-X5122106)资助项目。

收稿日期:2015-06-19;**修订日期**:2015-07-31

在线文本的情感分析已经成为自然语言处理的一个研究热点。文本情感分析是指对包含用户表示的观点、喜好和情感等的主观性文本进行检测、分析以及挖掘[1]。对于一些群体性事件,分析网民情绪的变化过程实际上就是对网络舆情进行演化建模及趋势预测,为有关部门进行舆论引导提供决策依据[2]。除此之外,文本情感分析技术也被成功运用到产品营销、股价预测等领域,因此研究在线文本的情感倾向具有很重要的理论和实用价值。目前文本情感分析的研究成果主要可归结为基于语义分析和基于机器学习的两大类方法。基于语义分析的方法大多依靠已有的情感词典、语义规则等来判别情感极性。杨佳能等[3]提出基于 PageRank 算法判定情感词集的极性并计算其强度,进而构建新闻评论情感词典。唐浩浩等[4]提出一种基于词亲和度的算法识别微博词语语义倾向,以此构建出高质量的情感词典,从而提高微博文本情感分析的准确率。文献[5,6]也做了基于语义规则实现情感分类的相关研究。基于机器学习的方法主要是选取大量有意义的特征来实现分类。Pang 等[7]首次使用 3 种机器学习方法,对电影评论的"积极"和"消极"情感进行分类。文献[8]定义了 7 种词语搭配模型,以微博语料为基础,构建二元词语搭配词库。相关研究也探讨了利用深度学习(Deep learning)对文本情感进行分析。梁军等[9]利用递归神经网络来发现与任务相关的特征,算法性能接近当前采用许多手工标注特征的传统算法,节省了大量人工标注的工作量。

在线评论文本存在大量新词、语法不规范等特点[8],使得中文评论文本情感分析存在困难与挑战。相关研究引入了近几年发展起来的主题模型。文献[10,11]利用潜在狄利克雷分布(Latent Dirichlet allocation,LDA)模型实现文本的聚类和分类。文献[12]提出基于主题的情感向量空间模型,它将文本的潜在主题特征融入到情感模型中,实验证明主题概率模型在情感分类任务上有良好的性能。本文结合上下文知识,提出一种基于 LDA 主题模型的中文评论文本情感分析方法。这种方法以语料库为基础,结合情感词典,依据指定的情感单元搭配模式,抽取出情感词和上下文知识,使用 LDA 模型挖掘文本中关键的情感特征,并利用支持向量机(Support vector machine, SVM)方法进行分类,实验表明了本文提出方法的有效性。

1 评论文本情感分类方法框架

本文提出的评论文本情感分类方法基本框架如图 1 所示。主要包括利用 LDA 主题模型训练情感单元和基于主题的情感向量空间建模。构建模型之前,先对评论文本进行预处理,主要是对语料进行分词、词性标注等,筛选出属于目标词性的词语。利用情感词典,依据提前定义的情感单元搭配模式,抽取能够表征评论文本情感的信息,即情感词和上下文。然后使用 LDA 主题模型,对选取出的情感信息进行训练,得到评论文本的关键情感特征。将得到的关键情感特征作为特征向量的特征项,构建基于主题的文本情感向量空间,利用支持向量机方法实现对评论文本的情感分类。其中,LDA 是一种 3 层贝叶斯概率模型,包含"文档-主题-词"3 层结构。2003 年 David M B 等[13]提出的最初模型只引入 1 个超参数 α 使每个文档的主题概率分布服从 Dirichlet 分布。随后,Griffiths 等[14]引入另一个超参数 β 使每个主题的词概率分布也服从 Dirichlet 分布。从而,LDA 模型发展为一个完整的产生式概率生成模型。LDA 是一种非监督机器学习方法,建模时做了词袋(Bag of words)假设,即只考虑词语出现的次数而不考虑词语的顺序。当有 X 篇文本,主题数为 K,词语数为 N 时,一篇文本中第 i 个词语的概率为

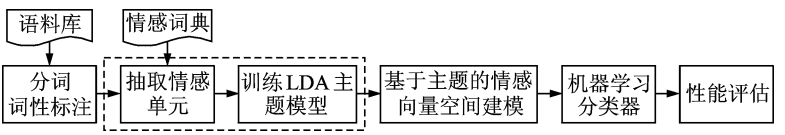


图 1 提出的评论文本情感分类方法基本框架

Fig. 1 Framework of sentiment analysis for comment texts

$$P(w_i)=\sum_{j=1}^{K}P(w_i\mid z_i=j)P(z_i=j) \tag{1}$$

式中:$P(z_i=j)$ 为取出的词语 w^* 属于主题 j 的概率;$P(w_i\mid z_i=j)$ 为词语 w^* 刚好为词语 w_i 的概率。对于文档 $\boldsymbol{d}$,$P(z_i=j)$ 和 $P(w_i\mid z_i=j)$ 分别为文档在主题上的多项分布和主题在单词上的多项分布,记为 $\theta_j^d=P(z=j)$ 和 $\varphi_w^j=P(w\mid z=j)$。多项分布服从 Dirichlet 分布,各层参数对应的依赖关系为[15] $w_i\mid z_i,\varphi^{z_i}\sim\text{Discrete}(\varphi^{z_i}),\varphi\sim\text{Dirichlet}(\beta),z_i\mid\theta^{d_i}\sim\text{Discrete}(\theta^{d_i}),\theta\sim\text{Dirichlet}(\alpha)$。

LDA 模型针对一个文本的生成过程为:(1)文本 d 的主题从主题分布中抽取得到,即从 Dirichlet(α) 抽样出 θ^d;(2)对于抽取出的主题 z_i,从 Dirichlet(β) 抽样出 φ^{z_i};(3)对于词语 w_i 和主题 z_i,抽样得到 $P(z_i\mid\theta)$ 和 $P(w_i\mid z_i,\varphi)$;(4)重复上述步骤直至遍历文本中每一个词语。具体模型如图 2 所示,各符号的含义如表 1 所示。本文引入 LDA 模型的生成思想对文本情感进行分析。一篇文本的生成过程基于某一类主要情感,有目的地选取能够表达相应情感或者潜在情感的关键词语,通过关键词语的组合和排列,得到主观性的情感文本。因此利用 LDA 模型发掘文本中基于主题的关键情感特征,并融入情感向量模型来实现文本的情感极性判别。

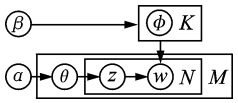


图 2 LDA 图模型表示形式

Fig. 2 Graph model of LDA

表 1 LDA 图模型各符号含义

Tab. 1 Meanings of parameters

符号	说明	符号	说明
α	文档-主题分布的超参数	w	词语
β	主题-词语分布的超参数	M	文档数
θ	文档-主题概率分布	N	词语数
φ	主题-词语概率分布	K	主题数
z	词语的主题分配		

2 情感信息抽取

仅仅依靠情感词难以完成实际的情感分类任务,因此将上下文知识融合到本文提出的模型中。利用语料库和情感词典,抽取指定词性搭配模式的情感词和上下文,构建三元搭配单元。

2.1 情感词典

本文研究包含情感词的文本情感极性,对不包含情感词的文本暂不考虑。一个较完整的情感词典对情感分析很重要。整合 HowNet 极性词典、台湾大学的 NTUSD 情感词典和大连理工大学信息检索研究室的情感本体库[16],去除重复词语,得到完整情感词集。利用各个词典的标注结果,对每一个情感词进行褒贬投票。对于投票倾向一致的情感词自动加入本文所用情感词典,否则采用人工标注方式并且多次校对。

2.2 提取情感单元

构造情感单元的目的是最大可能地获取文本中与情感有关的信息。本文提出的三元情感单元既包括与情感有直接关系的情感特征,也考虑了间接影响情感倾向的上下文。三元情感单元定义:$u=<e,w,f>$,其中 u 为情感单元;w 为情感词语;e,f 为上下文词语;w,e,f 三者满足以下两个条件:

(1) 词性搭配满足 8 种模式:$<\text{adj},\ \text{prep}_w,\ \text{adj}>$,$<\text{adj},\ \text{prep}_w,\ \text{n}>$,$<\text{adj},\ \text{prep}_w,\ \text{v}>$,$<\text{v},\ \text{prep}_w,\ \text{adj}>$,$<\text{v},\ \text{prep}_w,\ \text{n}>$,$<\text{n},\ \text{prep}_w,\ \text{adj}>$,$<\text{n},\ \text{prep}_w,\ \text{n}>$,$<\text{n},\ \text{prep}_w,\ \text{v}>$,其中 adj 为形容词,v 为动词,n 为名词,prep_w 为情感词 w 的词性并且 $\text{prep}_w\in\{\text{adj},\ \text{v},\ \text{n}\}$。

(2) 以一个句子为范围,在满足条件(1)的情况下,e、f 为距离 w 最近的上下文词语。此处的句子是指由标点符号分割而成的语言单位。在抽取之前,需要对语料先进行分词、词性标注以及删除停用词

等非目标词性的词。抽取过程主要依赖于预先指定的词性搭配模式。抽取步骤为：

(1)利用整合得到的情感词典，匹配出文档 $\boldsymbol{d}$ 中出现的情感词 w_i。

(2)对于情感词 w_i，根据提出的 8 种词性搭配模式提取满足条件的上下文词语 e_i 和 f_i，组合得到情感单元 u_i。

(3)重复上述两个步骤，直至遍历文档 $\boldsymbol{d}$ 中所有词语。提取情感单元后的文档 d^* 表示为：$d^* = \{u_1, u_2, \cdots, u_m\}$，其中 m 为文档 $\boldsymbol{d}$ 中情感词数量。

3 融合主题的情感向量空间模型构建

3.1 情感特征

提取情感特征是文本情感分析的技术重点和难点之一，有效的特征项是正确分类的关键。类似 LDA 生成思想，本文认为一篇文本是基于某一类主要情感有目的地选取表达对应情感或者潜在情感的词语，组合之后得到的。本文得到情感特征的主要过程是抽取出上下文词汇，配合情感词，通过 LDA 模型选取出关键的情感特征，以此作为向量空间的特征项。本文用于分类的情感特征包括上下文词汇和情感词。每一篇文档可表示为

$$\boldsymbol{d} = [\mathrm{con}_1, \mathrm{con}_2, \cdots, \mathrm{con}_m, w_1, w_2, \cdots, w_n] \tag{2}$$

式中：$\boldsymbol{d}$ 为文档的向量表示；con_i 为上下文词语；m 为上下文词语数目；w_i 为情感词；n 为情感词数目。按照 2.2 节抽取得到的情感单元包含了全部上下文词汇和情感词，容易出现维数较大的问题，并不适合构造特征向量，需要结合 LDA 模型计算出关键特征项，实现降维的效果。本文所提取的关键特征项是指文档 d^* 所属最大概率主题中概率值较大的词语。LDA 模型训练后得到“文档-主题”概率矩阵 $\boldsymbol{D}_{\mathrm{T}}$ 和“主题-词语”概率矩阵 $\boldsymbol{T}_{\mathrm{W}}$。利用得到的矩阵抽取关键特征项，图 3 为步骤流程，具体步骤如下：

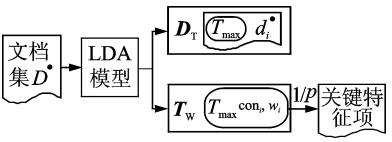


图 3 关键特征项抽取步聚

Fig. 3 Framework of extracting key features

(1)将已提取情感单元的文档集 D^* 作为 LDA 模型的输入。

(2)训练 LDA 模型得到“文档-主题”矩阵 $\boldsymbol{D}_{\mathrm{T}}$ 和“主题-词语”矩阵 $\boldsymbol{T}_{\mathrm{W}}$。

(3)针对文档 d_i^*，在矩阵 $\boldsymbol{D}_{\mathrm{T}}$ 中，确定其最大概率主题 $T_{\max}$。

(4)对应矩阵 $\boldsymbol{T}_{\mathrm{W}}$ 中的主题 $T_{\max}$，将词语按照模型训练后的概率值大小排序，然后以比例 $1/p$ 抽取得到关键特征项，降低特征项的维度，p 取正整数。

(5)重复步骤(3)，(4)直至遍历文档集 D^* 所有文档，然后整合全部关键特征项并去除重复项。

3.2 特征权重

向量空间的特征权重采用 tfidf 值。tfidf 值是一种普遍使用并且有效的权重计算方法。它强调某一个词在一篇文档中的重要性，表示为

$$\mathrm{tfidf} = \mathrm{TF} \times \mathrm{IDF} \tag{3}$$

式中：$\mathrm{TF} = h/g$，$\mathrm{IDF} = \log(1 + t/r)$，TF 为词频，$h$ 为词语 w 在文档 $\boldsymbol{d}$ 出现的次数，g 为文档 $\boldsymbol{d}$ 的词语数量，IDF 为逆向文件频率，t 为总文档数，r 为包含词语 w 的文档数量。

4 实验及结果分析

4.1 实验数据和评测标准

实验数据来源于谭松波[17]搜集的关于酒店的中文情感评论语料。对于数据集中不包含情感词的文本暂不考虑。整理语料得到 10 000 条评论文本，其中包括 7 000 条正向文本，3 000 条负向文本。随机选取

3 000条语料作为实验语料,数据集信息如表 2 所示。数据预处理采用中科院 ICTCLAS 分词工具对实验语料进行分词、词性标注。实验中的机器学习分类器选用 SVM,工具选取台湾大学林智仁开发的 LibSVM。

表 2 实验数据

Tab. 2 Experimental data

数据集	负向数据集	正向数据集	总数
训练数据集	900	900	1 800
测试数据集	600	600	1 200

本文对不包含情感词的语料暂不考虑,并且认为包含情感词的文本具有单一情感极性,分类结果只有正向或负向。对于每一个文本都能进行分类的语料集,评判分类器性能的正确率(Precision)、召回率(Recall)和 F 相等。因此采用总体准确率作为本文方法的分类性能评价指标,公式为

$$O_{\text{accuracy}}=\frac{\sum_{c_i\in C}\text{Correct}(c_i)}{\sum_{c_i\in C}\text{Doc}(c_i)} \tag{4}$$

式中:O_{accuracy} 为总体准确率,$\text{Correct}(c_i)$ 是分类为 c_i 并且正确的文档数,$\text{Doc}(c_i)$ 是类别为 c_i 的文档总数。

4.2 实验结果分析

本文实验的情感类别分为正向情感和负向情感两类。利用本文方法与快速主成分分析法[18]分别提取出低维度空间下的情感特征,作情感极性判别实验对比分析。所用 LDA 模型参数设置如下:$\alpha=0.5$,$\beta=0.1$,主题数 K 选取不同的正整数进行实验对比分析,其中 α 和 β 为 LDA 模型的超参数。

(1)LDA 模型的参数实验。选取部分实验语料,对主题数 K 取不同的正整数进行实验,得到的实验结果如图 4 所示。明显可知总体准确率相对稳定,维持在 79%左右。总体准确率最高和最低的实验结果分别是 $K=1$ 和 $K=20$,两者相差只有 0.89%。实验语料针对同个话题下的评论语料,话题内容较集中,造成主题数 K 对本文方法的分类性能影响不大。

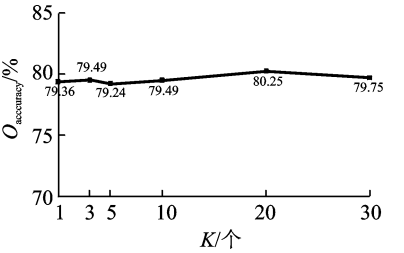


图 4 不同主题数下的实验结果

Fig. 4 Experimental results with different K

(2)情感分类实验。将 LDA 模型主题数 K 取 3,训练迭代次数为 100,对实验 1,2,3 和本文方法进行对比分析,实验结果如表 3 所示。实验 1 依据本文 2.2 节内容提取情感单元作为情感特征。实验 2 利用 LDA 模型提取文本中的特征。实验 3 则在实验 1 的基础上采用快速主成分分析法提取主要特征作为向量空间的特征项。本文方法利用 LDA 模型训练实验 1 得到的情感特征选取出概率值较大的主题词,并将其作为情感特征项。由表 3 可知,相较于实验 1~3,本文方法总体准确率有明显提高。实验 1 得到的负向准确率高达 90%,但是正向准确率较低,导致总体准确率只接近 70%,正负向分类性能明显不平衡。实验 1 和本文方法提取的正向情感词占总特征项词数的比例都接近 15%,但是实验 3 的正向准确率达到了 78.92%,说明特征项中情感词的比例并不是造成实验 1 正向准确率低的主要原因。主要原因是提取的情感词能否作为有效的特征项。实验 1 利用情感词典识别出情感词,但是部分正向情感词存在倾向性弱或者极性依附于语境的问题,例如“节省”,“随意”和“清淡”等词汇,因此该方法对情感词典的质量要求较高。本文方法则利用 LDA 模型自动训练出情感单元中的有效情感词作为情感向量空间的特征项,提高了分类的准确率,并且不过度依赖于情感词典。另外本文方法相对于实验 1,维数大幅度下降且取得了良好的分类效果,说明本文方法适用于大规模语料的分析任务。与本文方法不同,实验 2 不提取文本的情感单元,直接利用 LDA 模型训练文本的主题特征,得到的分类准确率低于本文方法的分类准确率,验证了对文本

的情感单元进行提取能够有效地优化情感分类的效果。实验 3 利用 FastPCA 方法对特征向量进行主成分分析,实现了同样程度的降维效果。从表 3 可知,本文方法相对实验 3 分类总体准确率高,说明本文方法在降维方面表现更出色,可以有效地提取出评论文本的关键特征项。

表 3 实验结果

Tab. 3 Experimental results

实验	特征提取方法	特征维数	负向准确率/%	正向准确率/%	总体准确率/%
1	情感单元	5 273	87.50	51.83	69.67
2	LDA	600	67.17	76.67	71.92
3	情感单元+FastPCA	600	48.33	82.50	65.42
本文方法	情感单元+LDA	600	77.83	80.00	78.92

5 结束语

本文将 LDA 模型引入到文本情感分析的研究中。基于 LDA 模型的生成思想,认为一篇评论文本是基于某一类主要情感有目的地选取词语,表达相应的情感或者潜在情感。因此本文通过构建一个较完整的情感词典,以中文评论语料库为基础,依据指定的情感单元搭配模式,匹配出情感词和上下文词汇,构造情感单元。通过 LDA 模型训练文本的情感单元,计算得到"文档-主题"矩阵和"主题-词语"矩阵,以此抽取出能够有效表征情感的关键特征项,并将其融入到情感模型中。最后利用机器学习的方法,对中文评论文本的情感进行分类,实验取得了很好的分类效果。同时实验证明相比于一般的降维方法,本文提出的方法更有优势。该方法能够结合主题模型挖掘词语之间潜在的语义关联,对文本进行有效的降维。本文研究还有很多可以改进的空间,在情感单元的构造过程中只考虑了上下文词汇,对更复杂的句子语境缺乏深入讨论。下一步会考虑利用依存句法的知识,挖掘句子中潜在语境和情感信息,并应用到文本情感分类中。

参考文献:

[1] 魏韡,向阳,陈千. 中文文本情感分析综述[J]. 计算机应用,2011,31(12):3321-3323.
Wei Wei, Xiang Yang, Chen Qian. Survey on Chinese text sentiment analysis[J]. Journal of Computer Applications, 2011, 31(12):3321-3323.

[2] 周耀明,李弼程. 一种自适应网络舆情演化建模方法[J]. 数据采集与处理,2013,28(1):69-76.
Zhou Yaoming, Li Bicheng. Adaptive evolution modeling method of internet public opinion[J]. Journal of Data Acquisition and Processing, 2013, 28(1):69-76.

[3] 杨佳能,阳爱民,周咏梅. 基于语义分析的中文微博情感分类方法[J]. 山东大学学报:理学版,2014,49(11):14-21,30.
Yang Jianeng, Yang Aimin, Zhou Yongmei. Sentiment classification method of Chinese micro-blog based on semantic analysis [J]. Journal of Shandong University: Natural Science , 2014, 49(11): 14-21, 30.

[4] 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 terms from Chinese micro-blogs based on word affinity measure[J]. Journal of Data Acquisition and Processing, 2015, 30(1): 137-147.

[5] 张晶,朱波,梁琳琳,等. 基于情绪因子的中文微博情绪识别与分类[J]. 北京大学学报:自然科学版,2014,50(1):79-84.
Zhang Jing, Zhu Bo, Liang Linlin, et al. Recognition and classification of emotions in the Chinese microblog based on emotional factor[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1):79-84.

[6] 赵文清,侯小可,沙海虹. 语义规则在微博热点话题情感分析中的应用[J]. 智能系统学报,2014,9(1):121-125.
Zhao Wenqing, Hou Xiaoke, Sha Haihong. Application of semantic rules to sentiment analysis of microblog hot topics[J]. CAAI Transactions on Intelligent Systems, 2014, 9(1):121-125.

[7] Pang B, Lee L, Vaithyanathan S. Thumbs up: Sentiment classification using machine learning techniques[C]// Conference on Empirical Methods in Natural Language Processing. [S. l.]: Association for Computational Linguistics, 2002:79--86. .

[8] 周剑峰,阳爱民,周咏梅,等. 基于二元搭配词的微博情感特征选择[J]. 计算机工程,2014,40(6):162-165.

Zhou Jianfeng, Yang Aimin, Zhou Yongmei, et al. Micro-blog sentiment feature selection based on bigram collocation[J]. Computer Engineering, 2014, 40(6): 162-165.

[9] 梁军,柴玉梅,原慧斌,等. 基于深度学习的微博情感分析[J]. 中文信息学报,2014,28(5):155-161.
Liang Jun, Chai Yumei, Yuan Huibin, et al. Deep learning for Chinese micro-blog sentiment analysis[J]. Journal of Chinese Information, 2014, 28(5): 155-161

[10] 王鹏,高铖,陈晓美. 基于 LDA 模型的文本聚类研究[J]. 情报科学,2015,33(1):63-68.
Wang Peng, Gao Cheng, Chen Xiaomei. Research on LDA model based on text clustering[J]. Information Science, 2015, 33(1): 63-68

[11] 李湘东,廖香鹏,黄莉. LDA 模型下书目信息分类系统的研究与实现[J]. 现代图书情报技术,2014,30(5):18-25.
Li Xiangdong, Liao Xiangpeng, Huang Li. Research and implementation of bibliographic information classification system in LDA model[J]. New Technology of Library and Information Service, 2014, 30(5): 18-25.

[12] 王磊,苗夺谦,张志飞,等. 基于主题的文本句情感分析[J]. 计算机科学,2014,41(3):32-35.
Wang Lei, Miao Duoqian, Zhang Zhifei, et al. Emotion analysis on text sentences based on topics[J]. Computer Science, 2014, 41(3): 32-35.

[13] David M B. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 1-6.

[14] Griffiths T L. Finding scientific topics[J]. Proceedings of the National Academy of Sciences, 2004, 101: 5228-5235.

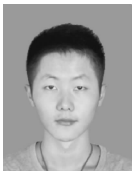
[15] Dowling J E, Wald G. The biological function of vitamin A acid[J]. Proceeding of the National Academy of Sciences of the United States of America, 1960, 46(5): 587.

[16] 徐琳宏,林鸿飞,潘宇,等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185.
Xu Linhong, Lin Hongfei, Pan Yu, et al. Construction the affective lexicon ontology[J]. Journal of The China Society For Scientific and Technical Information, 2008, 27(2): 180-185.

[17] 罗毅, 李利, 谭松波,等. 基于中文微博语料的情感倾向性分析[J]. 山东大学学报(理学版), 2014,49(11):1-7.
Luo Yi, Li Li, Tan Songbo, et al. Sentiment analysis on Chinese Micro-blog corpus[J]. Journal of Shandong University Natural Science, 2014, 49(11): 1-7.

[18] Sharma A, Paliwal K K. Fast principal component analysis using fixed-point algorithm[J]. Pattern Recognition Letters, 2007, 28(10): 1151-1155.

作者简介:



王伟(1991-),男,硕士研究生,研究方向:文本情感分析、机器学习和自然语言处理,E-mail:20131010007@gdufs.edu.cn。

周咏梅(1971-),女,教授,研究方向:自然语言处理、文本情感分析和机器学习。

阳爱民(1970-),男,教授,研究方向:自然语言处理、文本情感分析和机器学习。

周剑峰(1986-),男,硕士研究生,研究方向:自然语言处理、文本情感分析和机器学习。

林江豪(1985-),男,硕士研究生,研究方向:自然语言处理、文本情感分析和机器学习。