

DOI:10.16652/j.issn.1004-373x.2020.23.015

基于情感主题特征词加权的微博评论聚类算法研究

吴迪^{1,2}, 杨瑞欣^{1,2}, 申超¹

(1. 河北工程大学 信息与电气工程学院, 河北 邯郸 056038;

2. 河北省安防信息感知与处理重点实验室, 河北 邯郸 056038)

摘要: 为了提高微博评论的情感分析聚类效果, 提出一种基于情感主题特征词加权的微博评论聚类算法。该算法首先通过定义情感主题词袋提取情感主题词; 然后利用语义相似度计算获得情感主题特征词, 进一步通过定义情感主题特征词重要度和分布度对其进行加权, 充分考虑情感主题特征词对主题的贡献度; 最后通过LDA聚类获得加权的情感主题特征词。实验结果表明, 该算法表现出较好的情感分析聚类效果, 对网络舆情分析有一定的指导作用。

关键词: 微博评论; 聚类算法; LDA聚类; 情感分析; 特征词加权; 主题词提取

中图分类号: TN911.1-34; TP391.1

文献标识码: A

文章编号: 1004-373X(2020)23-0067-05

Microblog comment clustering algorithm based on weighted sentiment topic feature words

WU Di^{1,2}, YANG Ruixin^{1,2}, SHEN Chao¹

(1. School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China;

2. Hebei Key Laboratory of Security & Protection Information Sensing and Processing, Handan 056038, China)

Abstract: In order to enhance clustering effect of sentiment analysis on microblog comments, a microblog comment clustering algorithm based on the weighted sentiment topic feature words is proposed in this paper. With the algorithm, the sentiment topic words are extracted by defining the sentiment topic word bag, and then the sentiment topic feature words are obtained by calculation of semantic similarity. The sentiment topic feature words are weighted by defining their importance and distribution. The contribution of sentiment topic feature words to the topic is considered fully in the algorithm. The weighted sentiment topic feature words have been obtained by latent Dirichlet allocation (LDA) clustering. The experimental results show that the algorithm has a good clustering effect of sentiment analysis, and has a certain guiding role for network public opinion analysis.

Keywords: microblog comment; clustering algorithm; LDA clustering; sentiment analysis; feature word weighting; topic word extraction

0 引言

随着网络的普及, 微博已成为获取社会资讯的重要渠道^[1]。用户微博上的评论包含主观情感的表达, 并携带大量用户信息和数据信息, 微博用户的情感表达也会对其他用户的情感甚至对政府的决策行为产生影响。因此, 亟需对微博评论进行有效的情感分析, 从而把握网络舆情走向^[2]。

目前, 主题模型已成为微博评论分析领域的热点。LDA (Latent Dirichlet Allocation, 隐含的狄利克雷分配) 主题模型是包括词项、主题和文档三层结构的文档生成

概率模型^[3], 可以识别大规模文档集或语料库中潜在的主题信息。传统的LDA主题模型只能对微博评论中隐含的主题进行建模, 仅考虑了微博评论表面语义信息之间的关系, 在情感分析方面能力欠佳。

文献[4]在LDA中加入情感层与微博用户关系参数, 提出基于LDA和微博用户关系的主题情感模型, 该模型具有良好的情感分类能力; 文献[5]在LDA主题模型的基础上提出了适用于短文本的情感分析模型, 该模型可增强情感词汇之间的共现频率, 提高了在情感极性分类上的效率; 文献[6]基于LDA主题模型, 对产业产品进行文本挖掘, 利用情感词典对提取主题词进行情感标

收稿日期: 2020-04-23

修回日期: 2020-05-14

基金项目: 国家自然科学基金资助项目(61802107); 国家重点研发计划(2018YFF0301004); 河北省高等学校科技攻关项目(ZD2018087, ZD2016017); 河北省专业学位研究生教学案例建设项目(KCJSZ2018083); 河北省自然科学基金青年基金项目(F2019402428)

注,有效地对产品评论进行情感分析。

上述文献在情感分析方面表现出较好的性能,但在兼顾情感与语义两方面的分析效果上不太理想。文献[7]提出一种情感主题联合模型(Joint Sentiment Topic Model for Sentiment Analysis, JST),该模型在LDA的基础上增加情感层,使之与文档、主题和词语相关联,具有一定的情感和语义分析能力;文献[8]提出了隐含情感模型(Latent Sentiment Model, LSM),将主题划分为三种带有情感的特殊主题,从而实现对文档的情感分析;文献[9]中提出了LTM模型,将情感先验知识以频繁项集的形式集成到主题模型中,提高了语义分析能力;文献[10]提出了平均加权的WLDA模型,引入平均加权值和支持向量机算法,有效提高主题区分度和语义分析方面的质量。

上述文献在情感和语义分析两方面都有所突破,但却忽略了主题内部词与词之间的相关度。本文充分考虑主题词的贡献度和词与词之间的相关度,建立更加完善的情感主题词袋,提出基于情感主题特征词加权的微博评论聚类算法MCCWSFW(Microblog Comment Clustering Based on the Weighted Sentiment Topic Feature Words),旨在挖掘出微博评论中的情感主题特征词集和隐含语义特征,以达到较好的情感分析聚类效果。

1 算法分析

对于情感分析,一般选取主观性文本,也就是分析用户用微博发布的对某事物或事件的评论。本文处理微博评论步骤主要包括:数据预处理、特征处理与词向量获取、词向量相似度计算、特征选择和特征加权,最后给出MCCWSFW算法的详细描述。

1.1 数据预处理

数据预处理主要包括三个步骤。首先,利用ACHE爬虫法爬取微博评论,并对微博评论数据进行清洗,消除词干、停止词,删除评论频率很低的词;然后采用jieba进行中文分词,并通过LDA主题模型训练语料库进行降维处理;最后,针对微博评论定义情感主题词袋,用来匹配情感主题词。定义的情感主题词袋是为后续情感分析工作做准备,其具体定义如下。

定义1:情感主题词袋 $F(ST)$ 。

$$F(ST) = c\left(\sum_i^k s(\text{adj})\right) \cup c\left(\sum_i^k s(\text{adv})\right) \cup c\left(\sum_i^j s(v)\right) \cup c\left(\sum_i^n s(\text{else})\right) \quad (1)$$

式中:adj代表形容词;adv代表副词;v代表动词;else代表其他词性; i, k, j, n 分别代表 $F(ST)$ 中的形容词、副词、动词以及其他词性的数量; $c\left(\sum_i^k s(\text{adj})\right)$,

$c\left(\sum_i^j s(v)\right)$, $c\left(\sum_i^n s(\text{else})\right)$ 分别代表形容词、副词、动词和其他词性词袋;“ \cup ”表示字符串的拼接。

1.2 特征处理与词向量获取

经过1.1节数据预处理过程,对其得到的情感主题词袋作进一步处理。处理过程主要包括情感主题词的提取和对其词向量的获取两部分。

1) 从情感主题词袋中提取情感主题词,此步骤由特征处理过程完成。特征处理主要用来降低向量空间维数,从而简化计算,提高文本处理的速度和效率。本文采用TF-IDF进行特征处理。

2) 对情感主题词进行词向量获取。首先获取语料并对语料进行预处理,进一步在数据预处理和特征处理的基础上,采用基于word2vec的词向量获取方式对处理过的微博评论做训练,最后采用Gibbs抽样法对模型求解。

1.3 词向量相似度计算

在完成TF-IDF特征处理和词向量获取后,微博评论之间的相似度问题转变成了向量之间的相似度问题。情感主题词是根据词与词之间的相似度检测的,本文采用文本处理中最常用的相似性度量方式,即余弦相似度。

利用 $\text{Sim}(\mathbf{a}, \mathbf{b})$ 表示计算向量 \mathbf{a} 和向量 \mathbf{b} 的相似度,则相似度公式描述如下:

$$\text{Sim}(\mathbf{a}, \mathbf{b}) = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \sqrt{x_2^2 + y_2^2}} \quad (2)$$

式中: $\mathbf{a}[x_1, y_1]$ 和 $\mathbf{b}[x_2, y_2]$ 代表两个不同的向量; x_1, y_1, x_2, y_2 分别表示两个向量的横、纵坐标。

1.4 特征选择

将提取的情感主题词作为候选情感主题词,在词向量模型上将其转换成具有语义知识的词向量形式,再采用Pearson相关系数^[11]计算词向量,得到候选情感主题词之间的语义相似度值,以此对微博评论进行特征选择,最后将上述特征选择得到的主题词定义为情感主题特征词 S_i ,利用情感主题特征词对微博评论进行聚类。

情感主题特征词提取流程图如图1所示。

图1中,该流程通过匹配情感主题词袋,判断微博评论是否为情感主题词,“是”则进一步进行语义相似度计算,得到情感主题特征词集;“否”则继续匹配情感主题词袋,直到遍历整个微博评论集。

1.5 特征加权

由于在同主题内概率较高的词语间的相似度越高,则这个特征表达就越好。因此,本文充分考虑情感主题特征词对主题的贡献度,实现概率较高词语的相似度及

内部相似度计算,最终得到主题的权值计算结果。

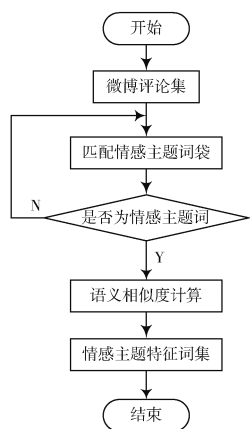


图1 情感主题特征词的提取流程图

本文定义了情感主题特征词重要度和情感主题特征词分布度两个参数,并分别对它们进行权重加权,再将情感主题特征词重要度和分布度进行线性相加。情感主题特征词重要度和分布度定义分别如下所示。

定义2:情感主题特征词重要度。假设微博评论 D 由 m 个情感主题特征词构成, $D=\{S_1, S_2, \dots, S_i, \dots, S_m\}$,则情感主题特征词 S_i 的重要度 $T(S_i)$ 的计算公式如下:

$$T(S_i) = \sum_{j=1}^m \text{Sim}(w_i, w_j) \quad (3)$$

式中 $\text{Sim}(w_i, w_j)$ 表示微博评论 D 中第 i 句和第 j 句之间的余弦相似度值。

情感主题特征词的重要度主要描述情感词对于主题的贡献作用。如果情感主题特征词中的词块与其他词块的相似度之和最大,则其很有可能就是微博评论的情感主题特征词或者与微博评论有较强的相关性。

与情感主题特征词相关的情感词的个数称为情感主题特征词分布度。情感主题特征词与越多的其他语句满足相似度阈值,则其覆盖的内容也就越多,因而成为情感主题特征词的可能性就越大。

定义3:情感主题特征词分布度。假设微博评论 D 的情感词数目为 m ,情感主题特征词 S_i 的分布度 $C(S_i)$ 计算公式如下:

$$C(S_i) = \frac{q(w_i)}{m} \quad (4)$$

式中 $q(w_i)$ 是微博评论 D 中与情感主题特征词 S_i 满足相似度阈值的评论个数。

根据前面对情感主题特征词的分析,情感主题特征词权值计算通过重要度和分布度线性相加得到。权值计算公式如下:

$$\text{Weight} = d \cdot T(S_i) + (1-d) \cdot C(S_i) \quad (5)$$

式中 d 表示阻尼系数,一般设置为0.85。

情感主题特征词加权算法流程如下:

算法1:情感主题特征词加权算法

输入:主题数为 n ,情感主题特征词集 T_i ,主题集 T

输出:情感主题特征词加权的主题集 T_i'

- 1) for 每个情感主题特征词集 $T_i \in T$;
- 2) 用式(5)计算情感主题特征词 T_n 的权重;
- 3) 得到情感主题特征词集 T_i' ;
- 4) end for;
- 5) output T_i' 。

通过主题特征词重要度和分布度两个参数对情感主题特征词进行加权,充分考虑情感主题特征词对主题的贡献度,从而优化情感分析能力。

1.6 MCCWSFW 算法

MCCWSFW 算法的图模型如图2所示。

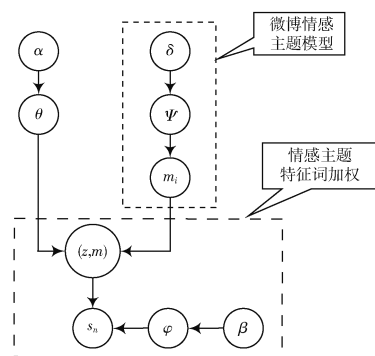


图2 MCCWSFW 算法图模型

图2中, θ, φ 和 Ψ 分别为超参数 α, β 和 δ 的概率分布, m_i 表示情感主题词袋, (z, m) 表示特征词加权过程, s_n 表示加权后情感主题特征词。

图模型主要描述微博情感主题建模和情感主题特征词加权过程,MCCWSFW 算法具体过程如图3所示。

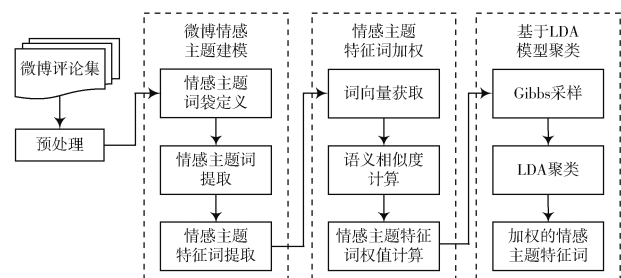


图3 MCCWSFW 算法框图

MCCWSFW 算法描述如下。

算法2:MCCWSFW 算法

输入:主题数为 n ,情感主题词袋 $F(ST)$,主题集 T

输出:情感主题特征词加权的主题集 T_i'

- 1) 利用情感主题词袋 $F(ST)$ 匹配情感主题词;
- 2) 利用语义相似度对情感主题词进行计算,提取

到情感主题特征词,构成情感主题特征词集;

3) 采用 word2vec 工具获取情感主题词集的词向量;

4) 得到词向量,使用情感主题特征词加权算法(算法1)进行加权计算;

5) 对主题特征词加权后的分布训练,利用 LDA 主题模型聚类,输出情感主题特征词加权的主题集 T_i' 。

MCCWSFW 算法挖掘语料库潜在的情感主题特征词,通过 LDA 构建微博情感主题模型。LDA 主题模型参数设置:主题数为 n , $\alpha=0.5$, $\beta=0.01$ 。采用 Gibbs 抽样法,将情感主题-词语分布从狄利克雷分布中抽取出来。

2 实验结果与分析

对 MCCWSFW 算法从情感主题特征词维度、准确率、综合指标值和时间效率进行测试,与 JST, LSM, LTM, WLDA 模型进行对比分析。实验首先确定 MCCWSFW 算法在 3 个数据集上最优情感主题特征词数,然后分别在自身最优情况下进行对比实验,并分析实验结果。

2.1 数据集

数据集从 2018 年 10 月的 52 个微博主题评论爬取,选取 3 个微博主题一周内的评论数据 6 557 条进行实验,其信息如表 1 所示。

表 1 实验数据基本信息

序号	主题	数量
1	国庆期间国内旅游乱象	1 819
2	北大第一医院医生被打事件	2 973
3	教师体罚学生被送派出所	1 765

用 Python 语言,在 CPU 为 Intel Core i5-7200U@2.50 GHz,内存 8 GB,操作系统为 Windows 7 环境下进行测试。

2.2 聚类评价指标

实验采用准确率(Accuracy)、召回率(Recall)和综合指标 F_1 值衡量聚类算法的有效性,计算公式如下:

$$Accuracy = \frac{\text{正确识别的微博评论情感个数}}{\text{识别出的微博评论总数}N} \quad (6)$$

$$Recall = \frac{\text{正确识别的微博评论情感个数}}{\text{测试集中微博评论情感的总数}n} \quad (7)$$

$$F_1 = \frac{2 \times Accuracy \times Recall}{Accuracy + Recall} \quad (8)$$

2.3 情感主题特征词维度测试

由于主题数对特征词提取有重要的影响,因此确定聚类最优时的主题数就很重要。下面通过设置不同情感主题特征词数,对微博评论语料进行 LDA 建模,最后采用综合指标 F_1 值衡量主题数最优值。

实验分别在聚类标准准确率、召回率和 F_1 值下进行维度测试,确保在自身最优的情况下与其他模型进行对比。结果如图 4 所示。

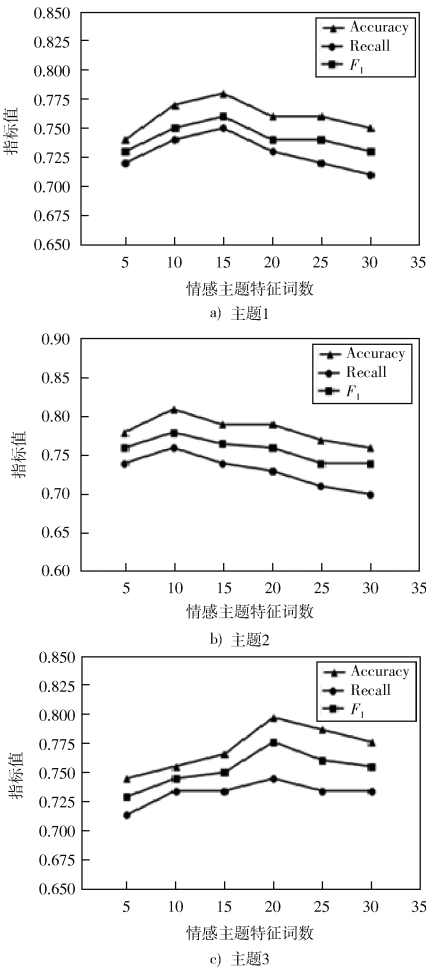


图 4 主题 1、主题 2 和主题 3 最优情感主题特征词数

如图 4 所示,在主题 1、主题 2 和主题 3 中,情感主题特征词数分别为 15,10,20 时, F_1 值达到最好效果。因此确定主题 1、主题 2 和主题 3 的最优情感主题特征词数分别为 15,10 和 20。

综上所述,3 个数据集情感主题特征词数在 10~20 时,均取得较好聚类效果。情感主题特征词代表主题的特征,当特征维数较小时,难以分辨主题特征;当特征维数较大,主题噪声会对实验结果产生干扰,从而降低聚类效果。

2.4 准确率测试

在主题 1、主题 2 和主题 3 的情感主题特征词数分别为 15,10,20 时,将 MCCWSFW 算法分别与 JST, LSM, LTM, WLDA 模型的情感分类准确率进行对比。对比结果如图 5 所示。

由图 5 可知, MCCWSFW 算法的准确率较 WLDA、LTM、LSM 和 JST 都有提升。在主题 2 上,该算法较

WLDA和LTM有明显的优势,正确率分别提高了8%和11%;在主题1和主题3上,该算法较其他模型的准确率均有提升。其中,较WLDA提高了5.33%,较LTM提高7%,较LSM提高8.6%,较JST提高10.7%。

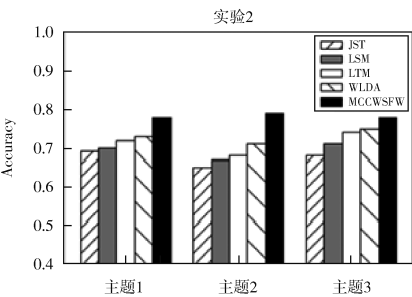


图5 三个主题下不同模型的准确率对比结果

2.5 指标值测试

实验在准确率、召回率和 F_1 值上,将MCCWSFW算法分别与JST,LSM,LTM和WLDA模型进行对比,结果如图6所示。

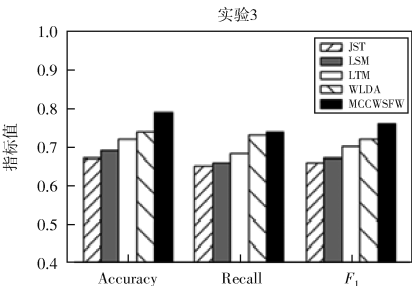


图6 不同算法的准确率、召回率、 F_1 值对比结果

由图6可知,MCCWSFW算法较其他模型的指标值均有所提升,具有一定的优越性。这是因为该算法在WLDA基础上,添加定义的情感主题词袋 $F(ST)$ 和加权的情感主题特征词集,从而可以在确保提取正确主题特征的前提下进行情感分析。

MCCWSFW算法聚类结果如表2所示,主题1、主题2和主题3加权的情感主题特征词数分别为自身最优维度下的15,10和20。

表2 聚类结果

主题	加权的情感主题特征词	情感主题特征词数
1	旅游、乱写、倡议、文明、素质、绿色出行、假特产、投诉、攀爬、收费、垃圾、强制、不当、消费、古迹	15
2	医闹、谴责暴力、呼吁理性、寒心、情理、剖宫产、指标、法律、身份、医患关系	10
3	警察、副所长、滥用职权、违法、指责、教师、体罚、教育尺度、监督、权利、公信力、责任、处罚、质疑、力度、父亲、不作为、公民、自觉、尊严	20

聚类结果表明,主题1关于游客不文明行为的评论居多,且评论中呼吁游客文明旅游、绿色出行。主题2

的评论质疑剖宫产指标的合理性,呼吁理性健康的医患关系。主题3的评论指责涉事副所长滥用职权,应注重对公民权利的保障,谨慎行使自己的职权,并认为教师的教育尺度越来越难以把握,维护老师尊严很有必要。由此可见,MCCWSFW算法对网络舆情有一定指导作用。

2.6 时间效率测试

对MCCWSFW算法效率进行测试。在主题1、主题2和主题3上对各模型进行20次平均执行时间对比实验,结果如表3所示。

表3 20次实验各类模型平均执行时间

数据集	JST	LSM	LTM	WLDA	MCCWSFW
主题1	2.087	1.683	1.982	1.925	2.033
主题2	2.782	2.571	2.638	2.536	3.190
主题3	1.961	1.965	1.864	1.865	1.912

由表3可知:在微博语料数据量小于2 000的主题1和主题3中,本文算法平均执行时间较其他模型相差不大;但在主题2中,可以看出,随着微博语料数据量的增加,MCCWSFW算法平均执行时间较其他模型明显增加,这是由于该算法利用情感主题词袋来匹配情感信息和获取主题情感特征,导致效率降低。

3 结 语

本文提出一种基于情感主题特征词加权的微博评论聚类算法,通过定义情感主题词袋,充分利用微博评论中包含的情感信息;在此基础上,通过加权情感主题特征词,充分利用微博评论之间的主题特征,考虑情感主题特征词对主题的贡献度,从而提高微博评论的情感分析聚类效果。

实验结果表明,MCCWSFW算法与JST、LSM、LTM和WLDA模型相比,在情感主题特征词维度、准确率、综合指标值和时间效率四个方面,表现出较好的情感分析聚类效果,能有效地应用到微博领域,提高网络舆情分析能力。下一步将针对提出算法时间效率不高的问题展开进一步研究。

注:本文通讯作者为申超。

参 考 文 献

[1] 代明竹,高嵩峰.聚类算法在高校学生微博的应用研究[J].现代电子技术,2019,42(7):177-180.

[2] LIU Zheng, LIU Chiyu, XIA Bin, et al. Multiple relational topic modeling for noisy short texts [J]. International journal of software engineering and knowledge engineering, 2018, 28(11/12): 1559-1574.

了预期水平。

5 结 语

本文从印刷体字符识别技术入手,提出了改进的Lenet-5Pro模型对字符进行识别,所做的工作包括:生成大量用于训练的印刷体汉字等字符的图片集以满足本文所改进的深度学习模型,同时,数据集可以扩展到其他文字和其他语言,扩展性强。对经典的深度学习网络进行了一定的改进,实验结果表明本文网络模型的有效性和准确性。但由于改进后的网络相对简单,针对较相似的字符识别容易发生错误,今后将尝试采用不同的网络来训练,以达到更好的鲁棒性。

参 考 文 献

- [1] 刘冬民. 基于深度学习的印刷体汉字识别[D]. 广州: 广州大学, 2018.
- [2] 孙巍巍. 基于深度学习的手写汉字识别技术研究[D]. 哈尔滨: 哈尔滨理工大学, 2017.
- [3] 汪一文. 深度卷积神经网络在OCR问题中的应用研究[D]. 成都: 电子科技大学, 2018.
- [4] RAWLS S, CAO H, KUMAR S, et al. Combining convolutional neural networks and LSTMs for segmentation-free OCR [C]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan: IEEE, 2017: 155-160.
- [5] KAKANI B V, GANDHI D, JANI S. Improved OCR based automatic vehicle number plate recognition using features trained neural network [C]// 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). Delhi, India: IEEE, 2017: 1-6.
- [6] BARTZ C, YANG H J, MEINEL C. STN-OCR: a single neural network for text detection and text recognition [EB/OL]. [2017-07-27]. <https://deeplearn.org/arciv/11984/stn-ocr>.
- [7] JIANG Y X, DONG H W, EI SADDIK A. Baidu Meizu deep learning competition: arithmetic operation recognition using end-to-end learning OCR technologies [J]. IEEE access, 2018, 6: 60128-60136.
- [8] KANG G H, KO J H, KWON Y J, et al. A study on improvement of Korean OCR accuracy using deep learning [C]// Proceedings of the Korean Institute of Information and Communication Sciences Conference - The Korea Institute of Information and Communication Engineering. [S.l.: s.n.], 2018: 693-695.
- [9] HE Yang, YUAN Jingling, LI Lin. Enhancing RNN based OCR by transductive transfer learning from text to images [C]// Thirty-second AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, USA: AAAI Press, 2018: 8083-8084.
- [10] WICK C, REUL C, PUPPE F. Improving OCR accuracy on early printed books using deep convolutional networks [EB/OL]. [2018-02-27]. <https://www.researchgate.net/publication/323444203>.
- [3] ZHANG Y, EICK C F. Tracking events in twitter by combining an LDA-based approach and a density-contour clustering approach [J]. International journal of semantic computing, 2019, 13(1): 87-110.
- [4] 黄发良, 于戈, 张继连, 等. 基于社交关系的微博主题情感挖掘 [J]. 软件学报, 2017, 28(3): 694-707.
- [5] 沈冀, 马志强, 李图雅, 等. 面向短文本情感分析的词扩充LDA模型[J]. 山东大学学报(工学版), 2018, 48(3): 120-126.
- [6] TAGO K, JIN Q. Influence analysis of emotional behaviors and user relationships based on twitter data [J]. Tsinghua science and technology, 2018, 23(1): 104-113.
- [7] YANG Qinjuan, RAO Yanghui, XIE Haoran, et al. Segment-level joint topic-sentiment model for online review analysis [J]. IEEE intelligent systems, 2019, 34(1): 43-50.
- [8] CHEN Zhenpeng, SHEN Sheng, HU Ziniu, et al. Emoji-powered representation learning for cross-lingual sentiment classification [C]// The World Wide Web Conference. San Francisco, CA, USA: ACM, 2019: 251-262.
- [9] CHEN Zhiyuan, LIU Bing. Topic modeling using topics from many domains, lifelong learning and big data [C]// International Conference on Machine Learning (ICML). Beijing: [s.n.], 2014: 2168-2177.
- [10] 郭晓慧. 基于LDA主题模型的文本语料情感分类改进方法 [J]. 延边大学学报(自然科学版), 2018(3): 266-273.
- [11] MU Yashuang, LIU Xiaodong, WANG Lidong. A Pearson's correlation coefficient based decision tree and its parallel implementation [J]. Information sciences, 2017, 435: 40-58.

(上接第71页)

作者简介: 吴迪(1984—), 女, 河北邯郸人, 博士, 副教授, 主要研究方向为数据挖掘、自然语言处理、软件安全分析。

杨瑞欣(1993—), 女, 河北邢台人, 硕士研究生, 主要研究方向为数据挖掘、自然语言处理。

申超(1985—), 男, 河北邯郸人, 博士, 讲师, 主要研究方向为人工智能、大数据、云计算。