DOI：10.13718/j.cnki.xsxb.2018.09.007

基于主题加权 LDA 模型的情感分类方法①

王飞雪[1]，　李　芳[2]

1. 重庆人文科技学院 计算机工程学院，重庆 合川 401524；
2. 重庆大学 计算机学院，重庆 400044

摘要：针对 LDA(Latent Dirichlet Allocation)主题模型生成的大量 topic，很大部分 topic 内部词语相关度很低，可解释性差，对语言模型后的应用效果带来一定的影响. 针对这一问题，该文提出了一种基于主题加权 LDA 模型的情感分类方法，该模型实现不同主题中内部相关的词语特征加权计算，能够消除不同主题内具有相关度词语的相互影响. 实验结果表明，与传统 LDA 模型分类方法对比，该文提出的基于主题加权 LDA 模型的情感分类方法平均 $F1$ 值提高了 6.7%～8.1%，验证了该文提出的方法是有效的，提高了分类效果.

关　键　词：LDA 模型；特征加权；主题模型；情感分类

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1000－5471(2018)09－0038－07

随着“互联网＋”时代的快速发展，社交网络移动互联网时代已经渗透到了所有群体当中. 然而，广大互联网用户们在网上发表评论与意见也是朝着海量级的方向递增. 例如，微博评论[1]、豆瓣网电影评论、酒店评论、大众网点评评论、新闻评论等等. 近年来，如何挖掘文本蕴含的主题，准确地定位互联网用户们对相关主题的需求[2－3]，成为了学者们研究的热点.

LDA 模型是在 PLSA(概率潜在语义分析)模型的基础上进行了改进[4]，文献[5]利用 LDA 进行主题挖掘，分析用于对主题的偏好，文献[6]对 LDA 进行了参数改进，进行对主题的数字量化评分. 由于 LDA 模型有着很好的特征降维，被广泛使用在情感分类相关的领域当中，且分类效果比起其他方法，如基于心电信号的特征情感分类[7]效果更好. 但是，LDA 模型在进行主题分类时，有很大部分主题内的词语相关度很低，如果还是用来做相关度计算，会对语言模型之后的应用效果造成一定的影响.

鉴于 LDA 的这一问题，本文提出了一种基于主题加权 LDA 模型的情感分类方法. 该模型实现不同主题中内部相关的词语特征加权计算，能够消除不同主题内具有相关度词语的相互影响，并且用实验对改进前后的模型进行了对比，证明了改进的模型是有效的、可行的.

1　相关工作

LDA 模型是基于 PLSA 模型的改进模型，改进后在语义挖掘、文档的特征表示方面起到了显著效果，因此被广泛地应用于文本挖掘、信息检索、图像识别等方面的研究工作中. 构成部分主要分为文档，主题(topic)和词语 3 个部分. 由于 LDA 是一个生成模型，所以其成功步骤主要包括：

Step 1　　将主题分布 θ 从狄利克雷分布 α 中抽取；

① 收稿日期：2017－07－05
基金项目：国家自然科学基金项目(61662083).
作者简介：王飞雪(1974－)，女，硕士，讲师，主要从事计算机软件与应用技术研究.

Step 2　对 Step1 得到的主题分布 θ 进行采样得到主题 Z；

Step 3　将主题-词语分布 φ 从狄利克雷分布 β 中抽取出来；

Step 4　将词语 W 从主题-词语分布 φ 中选择出来.

其中，参数 α 和 β 会在抽样中不断更新至最优. LDA 模型示意图如图 1 所示.

模型联合概率的计算公式为

$$P(\theta, z, w, \varphi \mid \alpha, \beta) = \prod_{n=1}^{N} P(\theta \mid \alpha) P(z_n \mid \theta) P(\varphi \mid \beta) P(w_n \mid \theta) \tag{1}$$

其中 α 和 β 分别是文档和词语的狄利克雷超参数，属于文档集合级别的参数；z 和 w 是主题和词语级的参数，对图 1 模型中 4 个步骤循环 N 次，每一次采样得到一个词语，共得到 N 个词语.

整个语料 $N = \{w_n\}_{n=1}^{N}$ 的生成概率为

$$p(N \mid \alpha, \beta) = \prod_{n=1}^{N} p(w_n \mid \alpha, \beta) \tag{2}$$

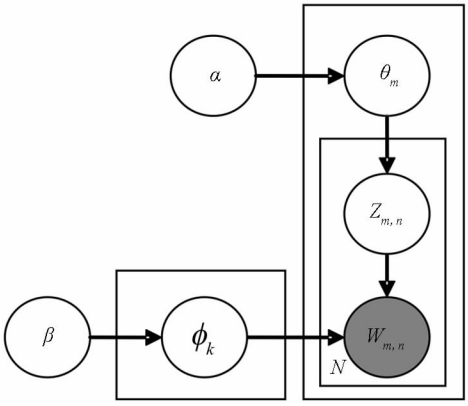

图 1　LDA 模型示意图

2　基于 LDA 模型的语料建模

2.1　挖掘 LDA 模型潜在的主题特征

在进行文档生成时，需要经过 2 个步骤的选取，分别是对主题的选择和对词语的选择，才能得到最终想要的文档生成结果. 对前者的选取是由文档-主题分布决定的，而对后者的选取是由主题-词语分布决定的. 对模型求解时，采用 Gibbs 抽样法[8].

由于主题数量对主题特征的提取有着至关重要的影响作用，因此确定一个最合适的主题数量显得非常重要. 下面通过分别设置不同的主题数量，分别对语料进行 LDA 建模，训练语料选择是豆瓣影评语料、网易新闻语料，最后采用分类 $F1$ 值来衡量主题数量最优值(图 2).

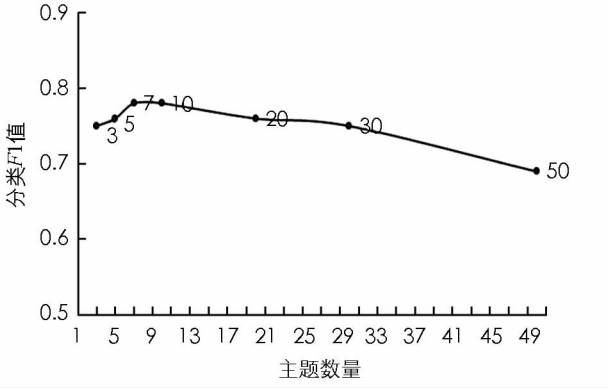

图 2　情感分类 $F1$ 值随主题数量变化图

从图 2 中可以很明显地得出，当数量在 7 之内，分类 $F1$ 值在不断地上升，在数量达到 7 时，达到最大值，7 以后开始逐渐地下降. 可以很好地说明，当维度过小时，不能充分表现出文本差异的特征来，从而影响实验效果；当维度过大时，会把本来表现明显的特征强行地分配到更多的特征上来，从而影响文本差异不明显且噪声增加，使得文本分类效果达不到想要的效果. 根据最佳数量对语料的 LDA 建模便可以得到文档-主题分布、主题-词语分布这 2 个分布.

2.2　主题特征维度的设置

根据 2.1 小节内容得到的主题－词语分布，可能会遇到一个主题下面的词语会同时出现正向评价词语，例如：“好看”、“喜欢”、“太棒”等；反向评价词语，例如：“糟糕”、“无趣”等；除了正向、反向评价词语，一些主题特征表现差的中性词语也有出现，例如：“没”、“想”等. 对于这些特征表现不明显的中性词，需要进行词语分布概率的最低阈值设置，从而获取到大于该阈值的词语，来对这些词语进行主题内容相关性研究. 阈值的取值影响每个主题下分布的若干词语的获取，最终影响算法的准确度，阈值经验最优值为

$\alpha=0.5$，$\beta=0.01$，对阈值 α 取不同的值做比对试验，找到最优阈值.

3 基于词语相关度的主题加权 LDA 方法

3.1 词向量的获取

在本文词向量的获取方式采用基于 word2vec 的词向量获取方式[9]. 进行词向量获取的第一步是获取语料. 对于情感分析，一般选取主观性文本，也就是大众在网上发布的对某一事物的看法和评价，可以使用前辈们准备好的语料也可以通过爬虫获取. 第二步需要对语料进行预处理，预处理也就是去掉标点、非中文字符、停用词、繁体字转化为简体字、分词等，其中分词采用 ICTCLAS 分词[10]，分类器使用 SVM 分类器[11]. 第三步使用 word2vec 模型，对第二步中处理过的语料做训练，训练之前的参数设置见表 1.

表 1 词向量获取的参数设置

序号	参数设置	序号	参数设置
1	维数：200	6	是否使用 cbow 模型：是
2	训练窗口大小：8	7	线程数：20
3	采样阈值：le-4	8	迭代次数：15
4	是否使用 HS 方法：是	9	设置最低频率：5
5	是否使用 NEG 方法：否	10	学习速率：0.025

表 1 中参数 1 表示设置词向量的维度；参数 2 设置最大的词间窗口；参数 3 表示词共现的阈值；参数 4 表示是否采用 softmax 分层；参数 5 表示不使用 NEG 方法，使用 HS 方法；参数 6 表示使用 cbow 模型；参数 7 表示训练时线程数；参数 8 表示迭代次数；参数 9 表示设置最低频率；参数 10 表示学习率. 根据多次试验可以得到表 1 中参数，能够保证较好的训练效果.

训练后会得到词向量矩阵，大小为 $M\times N$ 维，M 表示词语的数量，N 表示词向量的维度. 词向量的表示形式是以一个词语开头，后面跟着这个词的词向量维度(表 2).

表 2 词向量的表示形式

好看	−0.006 495	−0.419 475	……	−0.006 500	−0.008 414
喜欢	−0.006 495	−0.419 475	……	−0.006 520	−0.006 847

3.2 词向量的词语相关度计算

基于词向量的词语相关度计算，采用向量余弦相似度(Cosine Similarity)进行计算，最后通过基于同义词词林的方法来与基于词向量的方法获取的词语相关度进行对比，验证基于词向量方法获取的词语相关度的效果. 通过余弦夹角来判断 2 个向量的相似程度大小，语义相关度是一个小数，范围是(−1，1)，相关度越接近于 1，说明夹角越小，相似度越高；相关度越接近于−1，说明夹角越大，相似度越低. 向量 a 和向量 b 的余弦相似度计算公式为

$$\mathrm{sim}(a,b)=\cos\theta=\frac{\vec{a}\cdot\vec{b}}{\|a\|\cdot\|b\|}=\frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\ \sqrt{\sum_i b_i^2}} \tag{3}$$

式中 θ 为向量夹角，a 和 b 是 2 个向量. 基于同义词词林的方法和基于词向量的方法实验对比如表 3.

表 3 词语相关度计算结果对比

对比词	同义词词林方法	词向量方法
感人	1.000	0.722
兴奋	0. 877	0.684
震撼	0.647	0.652
惊喜	0.100 0	0.509
欢喜	0.100 0	0.619
太棒	0.100 0	0.701

通过实验数据表明，基于词向量的方法在词语相关性表达上效果比基于同义词词林的方法更好. 基于

同义词词林的方法，语义依耐性太强，不能很好地体现出词语间的关联性.

3.3　词语相关度的主题加权

由于在同主题内概率较高的词语间的相关度越高，那么这个主题内部之间相关性则越强，也就是主题的特征表达越好. 为了达到这样的目的，需要加大表达能力强的主题的特征权值，减少表达能力差的主题的特征权值. 在满足这样的主题条件下，实现概率较高词语的相关度及内部相关度的计算，最终得到主题的权值计算结果. 主题内部相关度 L 与词语相关度的关系见式(4)，即词语两两相关度的总和求平均.

$$L=\frac{\mathrm{sim}(w_1,w_2)+\cdots+\mathrm{sim}(w_1,w_n)+\cdots\mathrm{sim}(w_2,w_3)+\cdots+\mathrm{sim}(w_2,w_n)+\cdots+\mathrm{sim}(w_{n-1},w_n)}{\frac{n(n-1)}{2}}=$$

$$\frac{2\sum_{i=1}^{n-1}\sum_{i+1}^{n}\mathrm{sim}(w_i,w_j)}{n(n-1)} \tag{4}$$

其中 w_n 表示词语，$\mathrm{sim}(w_i,w_j)$ 表示 2 个词语的相关度，由式(4) 中的主题内部相关度 L 可以计算得到权值 V 为

$$V=2^{k(L-\overline{L})} \tag{5}$$

式中 $\overline{L}$ 是 L 的均值，基准权重设置为 1，大于均值时，赋予权重 >1，小于均值赋予权重 <1. 参数 k 是调节参数，用于调整主题内 $\overline{L}$ 和 L 对权值 V 大小的影响. 图 3 中是 $k=1$ 时，主题权值 V 与 $\overline{L}$ 和 L 的函数关系.

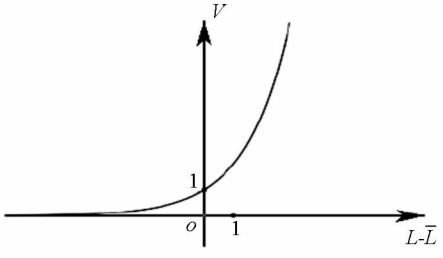

图 3　权值 V 与 $\overline{L}$ 和 L 的函数关系

3.4　基于主题加权 LDA 模型的情感分类算法

基于主题加权 LDA 模型情感分类算法过程如图 4 所示，算法的实现步骤如下：

第一步，挖掘语料库潜在的主题特征，构建 LDA 模型. 参数设置：主题数为 7，$\alpha=0.5$，$\beta=0.01$. 采用 Gibbs 抽样法，将主题－词语分布从狄利克雷分布中抽取出来. 最后，对阈值 α 取各种不同的值，统计分布概率高于 α 的所有主题下的词语进行实验对比，达到确保 α 最优.

第二步，对预处理后的语料库使用 word2vec 工具获取词向量，其维度取值为 200.

第三步，对于上述步骤得到维数为 200 的词向量，使用主题加权算法进行计算.

第四步，对主题加权后的分布进行训练，采用神经网络中知识向量机做分类.

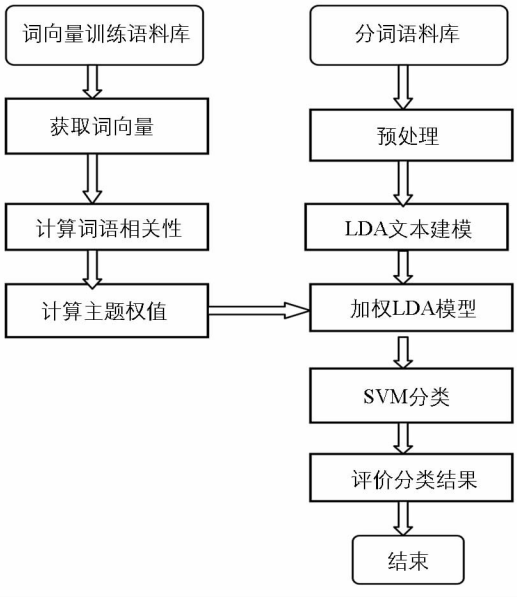

图 4　基于主题加权 LDA 模型分类算法流程图

4　实验结果与对比

4.1　实验语料

由于本实验是验证情感分类，为了验证本文提出的基于主题 LDA 模型的情感分类方法的性能，选择了主观性文本的评论语料来进行分析. 为了验证实验的正确性与精确性，本文选择了二种语料集.

第一种语料集是豆瓣网影评情感测试语料，来自豆瓣网对电影《ICE AGE3》的评论，评分标准均按照 5stars 评分在网页中标注，规模为 11 323 条评论. 豆瓣网影评原始页面，如图 5 所示.

第二种语料集是酒店评论语料，来自携程网，规模为 1 万篇.

4.2 性能指标

为了高可靠与稳定的模型，本文分类算法采取10次10折交叉验证方法，运用准确率 P(Precision)、召回率 R(Recall) 及 F(F-Measure) 值对实验结果进 行评价. 由于F-Measure是Precision和Recall加权调和平均，所以最后用 F 值作为分类效果度量. F(F-Measure) 值的计算公式为

$$F=\frac{(a^2+1)P\times R}{a^2(P+R)} \tag{6}$$

特殊地，当参数 $a=1$ 时，就是最常见的 $F1$，计算公式为

$$F1=\frac{2\times P\times R}{P+R} \tag{7}$$

可知 $F1$ 综合了 P 和 R 的结果，当 $F1$ 较高时，则能说明试验方法比较有效.

图5 豆瓣网影评原始页面

4.3 结果与分析

在不同主题之下，取不同的阈值 α，分布概率高于阈值的词语的平均数量如图6所示.

从图6可以看出，当 $\alpha=0.01$ 时，可从某一个主题中抽取得到11个词语，当 $\alpha=0.005$ 时，得到15个词语，比较适合；在 $\alpha>0.005$ 时，词语数量递增且每个主题下面的词语比较少，在 $\alpha=0.02$ 时，词语仅有7个，容易产生不确定性从而导致主题内部相关度产生一定的偏差；在 $\alpha=0.003$ 时，每个主题下抽取得到40个词语，此时内部相关度下降，而噪声上升，使得算法效果变差，所以在本实验中阈值取最佳值为 $\alpha=0.005$.

在使用内部相关度计算主题权值的时候对 k 进行调整，从而可以很好地起到减少他们两者之间差值的幅度变化. 图7给出了在第二种语料集中，分类 $F1$ 值随着 K 值变化的曲线.

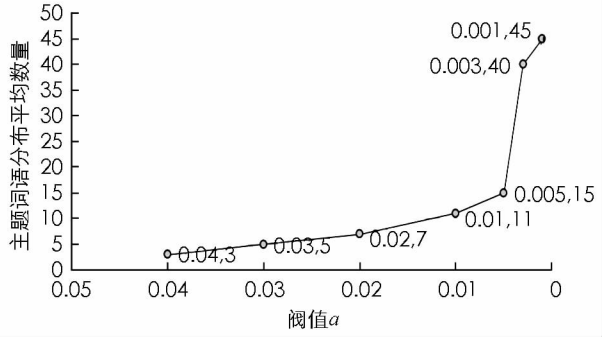

图6 主题的词语分布平均数量随阈值 α 变化曲线

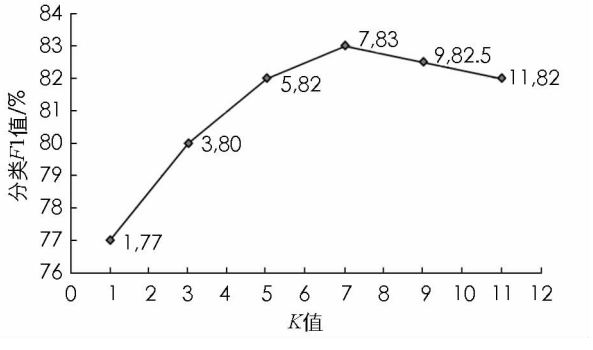

图7 $F1$ 值随着 K 值变化的曲线

通过曲线变化可以看出，在 k 值小于7时，随着 K 值递增趋势，分类 $F1$ 值比例也在逐渐呈递增趋势；在 K 值达到7时，$F1$ 值比例达到最佳；当 K 值大于7时，$F1$ 值比例开始逐渐地呈下降趋势. 所以 K 等于7为最优数值. 确定完最优参数 K 后，分别使用第一种语料库与第二种语料库对基于主题加权的LDA模型和没有对主题加权的LDA模型分别进行实验对比(表4).

表4 LDA主题加权模型与传统模型实验对比

	第一种语料			第二种语料		
	P	R	$F1$	P	R	$F1$
未加权模型	0.728	0.721	0.725	0.784	0.748	0.764
加权模型	0.776	0.774	0.774	0.841	0.809	0.827

经过在 2 个语料集上进行对比实验验证，发现在分类效果上基于主题加权 LDA 模型比传统 LDA 模型都有不同程度的提升，其中在第一种语料上 $F1$ 值提高了约 6.7%，对比效果如图 8；在第二种语料上 $F1$ 值提高了约 8.1%，对比效果如图 9. 那么，基于主题加权 LDA 模型在这 2 种语料上 $F1$ 值平均提高了 6.7%～8.1%，由此指标数据可以得出本文中 LDA 模型的主题加权方法能够有效地对情感进行分类，与传统 LDA 模型比较，能够得到更好的分类结果.

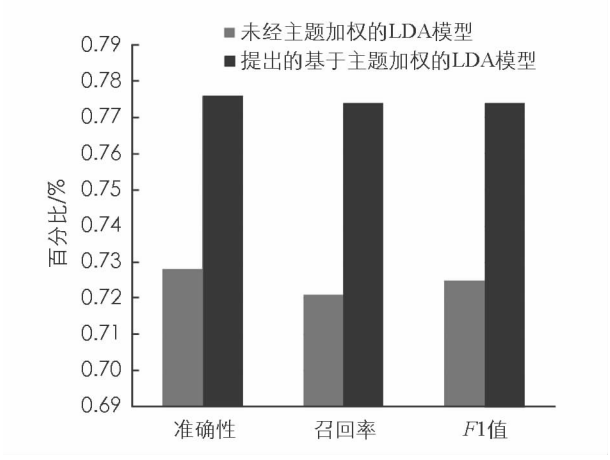

图 8　LDA 加权模型与传统模型在第一语料集上的对比

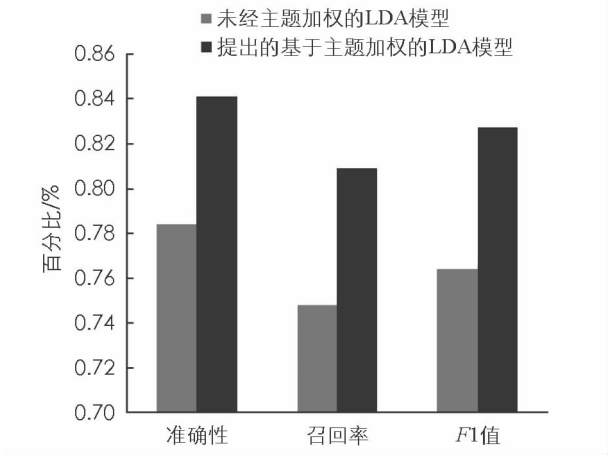

图 9　LDA 加权模型与传统模型在第二语料集上的对比

同时，从表 4 数据中还可以得出，在第一种语料集中分类效果比第二种语料集的分类效果稍差，其中有可能是不同语料的预处理方式有误差，不同语料下设置的最优参数会不一样，第一种语料中的评论语言没有经过人为的后期整理，内容不规范，且有一定的噪声. 这 3 个原因导致第一种语料集的分类效果比不上第二种语料集分类效果.

5　结　　论

针对 LDA(Latent Dirichlet Allocation)主题模型在 topic 上主题内部词语相关度低下，导致不能明确表达出该 topic 的主要语义这一问题，本文提出一种基于主题加权的 LDA 模型方法，并实现对情感进行分类. 该模型实现了不同主题中内部相关的词语特征加权计算，能够消除不同主题内具有相关度词语的相互影响. 最后通过电影短评和酒店评论 2 组语料库的分类实验，得出与传统 LDA 模型相比，本文提出的主题加权 LDA 模型的 $F1$ 值提升了 6.7%～8.1%，从而证明了本文模型对于情感分类的可行性与有效性.

参考文献：

[1] 尹书华. 基于复杂网络的微博用户关系网络特性研究 [J]. 西南师范大学学报(自然科学版)，2011，36(6)：57－61.

[2] 孙平安，谭秋月. 基于多属性决策理论的文本信息挖掘技术研究 [J]. 西南师范大学学报(自然科学版)，2016，41(11)：155－159.

[3] 李红波，孟欣赏，吴　渝，等. Web 访问挖掘中的匿名用户识别算法研究 [J]. 西南师范大学学报(自然科学版)，2015，40(9)：78－84.

[4] SOCHER R，PENNINGTON J，HUANG E H，et al. Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh：Association for Computational Linguistics，2011.

[5] TAI K S，SOCHER R，MANNING C D. Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing：ACL，2015.

[6] LIU Y, LI S, ZHANG X, et al. Implicit Discourse Relation Classification via Multi-Task Neural Networks [C]//Proceedings of the Thirtieth Conference on the Association for the Advance of Artificial Intelligence. Phoenix: AAAL, 2016.
[7] 程 静, 刘光远. 基于情感心电信号的去趋势波动分析研究 [J]. 西南大学学报(自然科学版), 2016, 38(2): 169－175.
[8] 刘真臻, 徐东平. 微博个性化标签图形化 RTM 模型 Gibbs 采样推荐 [J]. 微电子学与计算机, 2017, 34(12): 138－144.
[9] 张志昌, 周慧霞, 姚东任, 等. 基于词向量的中文词汇蕴涵关系识别 [J]. 计算机工程, 2016, 42(2): 169－174.
[10] 李湘东, 高 凡, 丁 丛. LDA 模型下不同分词方法对文本分类性能的影响研究 [J]. 计算机应用研究, 2017, 34(1): 62－66.
[11] 王 见, 陈 义, 邓 帅. 基于改进 SVM 分类器的动作识别方法 [J]. 重庆大学学报(自然科学版), 2016, 37(1): 12－17.

Emotion Classification Method Based on Topic Weighted LDA Model

WANG Fei-xue[1], LI Fang[2]

1. *School of Computer Engineering, Chongqing Institute of Humanities and Technology, Hechuan Chongqing 401524, China;*
2. *School of Computer Science, Chongqing University, Chongqing 400044, China*

Abstract: For the large number of topics generated by the LDA (Latent Dirichlet Allocation) theme model, the relevance of the internal words is very low, poor interpretation, and the effect of the language model is affected. In order to solve this problem, an emotion classification method based on topic weighted LDA model has been proposed in this paper, which can realize the weighting calculation of words in different themes, and can eliminate the influence of words with relevance in different themes. The experimental results show that compared with the traditional LDA model classification method, the average F1 value of the emotion classification method based on the topic weighted LDA model is improved by 6.7%－8.1%, which proves that our proposed method is effective and improved classification effect.

Key words: LDA model; weighted feature; topic model; emotion classification

责任编辑 夏 娟