DOI:10.11992/tis.201606007
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160808.0830.020.html

基于词加权 LDA 算法的无监督情感分类

郝洁,谢珺,苏婧琼,续欣莹,韩晓霞
(太原理工大学 信息工程学院,山西 晋中 030600)

摘 要:主题情感混合模型可以有效地提取语料的主题信息和情感倾向。本文针对现有主题/情感分析方法主题间区分度较低的问题提出了一种词加权 LDA 算法(weighted latent dirichlet allocation algorithm,WLDA),该算法可以实现无监督的主题提取和情感分析。通过计算语料中词汇与情感种子词的距离,在吉布斯采样中对不同词汇赋予不同权重,利用每个主题下的关键词判断主题的情感倾向,进而得到每篇文档的情感分布。这种方法增强了具有情感倾向的词汇在采样过程中的影响,从而改善了主题间的区分性。实验表明,与 JST(Joint Sentiment/Topic model)模型相比,WLDA 不仅在采样中迭代速度快,也能够更好地实现主题提取和情感分类。

关键词:情感分类;主题情感混合模型;主题模型;LDA;加权算法
中图分类号: TP391 **文献标志码**:A **文章编号**:1673-4785(2016)04-0539-07

中文引用格式:郝洁,谢珺,苏婧琼,等. 基于词加权 LDA 算法的无监督情感分类[J]. 智能系统学报, 2016, 11(4): 539-545.
英文引用格式:HAO Jie, XIE Jun, SU Jingqiong, et al. An unsupervised approach for sentiment classification based on weighted latent dirichlet allocation[J]. CAAI Transactions on Intelligent Systems, 2016, 11(4): 539-545.

An unsupervised approach for sentiment classification based on weighted latent dirichlet allocation

HAO Jie, XIE Jun, SU Jingqiong, XU Xinying, HAN Xiaoxia
(Information Engineering College, Taiyuan University of Technology, Jinzhong 030600, China)

Abstract: The topic and sentiment unification model can efficiently detect topics and emotions for a given corpus. Faced with the low discriminability of topics in sentiment/topic analysis methods, this paper proposes a novel method, the weighted latent dirichlet allocation algorithm (WLDA), which can acquire sentiments and topics without supervision. The model assigns weights to terms during Gibbs sampling by calculating the distance between seed words and terms, then counts the weights of key words to estimate the sentiment orientation of each topic and obtain the emotional distribution throughout documents. This method enhances the impact of words that convey emotional attitudes and obtains more discriminative topics as a consequence. The experiments show that WLDA, compared with the joint sentiment/topic model (JST), not only has a higher iteration sampling speed, but also gives better results for topic extraction and sentiment classification.

Keywords: sentiment classification; topic and sentiment unification model; topic model; LDA; weighting algorithm

互联网不仅是获取信息的重要途径,也是广大网民表达观点和看法的平台。随着博客、微博、微信等自媒体的流行,网络购物的盛行和网购评价体系的不断完善,对事件的观点、对物品的评价等具有情感倾向的文本飞速增长。这些信息对于政府部门的舆情监控、企业的经营决策和个人的购买决定都起着至关重要的作用。然而,这些评价信息数量巨大、变化迅速,仅依赖人工收集整理不仅成本高,也难以

收稿日期:2016-06-02. 网络出版日期:2016-08-08.
基金项目:山西省回国留学人员科研项目(2015-045,2013-033);山西省留学回国人员科技活动择优资助项目(2013);山西省自然科学基金项目(2014011018-2).
通信作者:谢珺. E-mail:xiejun@tyut.edu.cn.

满足时效性要求。因此文本情感分析受到了学术界与工业界越来越多的关注[1-2]。

情感分类是文本情感分析的重要组成部分。它是指根据文本所表达的含义和情感信息将文本划分为褒扬或贬义两种或几种类型,是对文本作倾向性、观点和态度的划分。目前,大多数情感分类方法都是监督模型或半监督模型,但标记好的语料常常难以获取,给情感分类造成困难。基于主题模型的情感分类,不仅具有无监督的优势,也具有较强的可移植性[3]。

Lin 等[4] 提出了 LSM 模型(latent sentiment model),该模型将情感作为主题的特例,认为文档中词汇的分布与情感有关,从而实现了文档的无监督情感分类,但无法识别出更细粒度的情感信息。Titov 等[5]提出的 MG-LDA 模型(multi-grain model)能够以较细的粒度提取主题,该算法是一个有监督学习模型,需要对样本类别进行人工标注。TAM(topic-aspect model)[6] 和 TSM(topic sentiment mixture)[7] 能够无监督地抽取文档的主题和情感信息。但这两种算法假定主题和情感的分布相互独立,忽略了二者的联系,也给解释主题和情感的关系造成困难。ASUM 模型(aspect and sentiment unification model)考虑了主题和情感的相关性,建立了“句子—主题—词”的 3 层模型,有效提取了情感和主题信息,但这种方法将每个句子视为一个文档,丢失了上下文信息[8]。JST 模型(joint sentiment/topic model)是一种可以无监督地提取文档主题和情感信息的 4 层贝叶斯网络,但该算法的复杂度较高,结果不够稳定[3]。欧阳继红等在 JST 模型的基础上,提出了多粒度的主题情感混合模型 MG-R-JST 和 MG-JST,该方法同时考虑到文档和局部两个粒度的情感主题分布,稳定性好,但面临复杂度较高的问题[9]。

本文在 LDA 模型的基础上,提出了应用于主题/情感分析的词加权 LDA 算法(weighted latent dirichlet allocation,WLDA),通过计算语料中词汇与情感种子词的距离,在吉布斯采样中对各词区分对待,利用每个主题下的关键词判断主题的情感倾向,进而得到每篇文档的情感分布。实验表明,WLDA 可提取细粒度情感,并且具有迭代速度快、分类精度高的优点。

1 LDA 模型

LDA(latent dirichlet allocation)[10] 是一种 3 层贝叶斯模型,它描述了文档、主题、词汇间的关系。LDA 模型自 2003 年提出以来,已经有了诸多的改进和变形算法,并在文本分类[11]、信息检索[12] 等领域得到了广泛应用。其图模型见图 1。

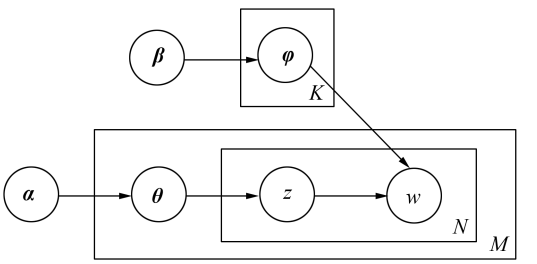


图 1 LDA 图模型[10]

Fig.1 Graphical model of LDA[10]

图 1 中,各个符号的含义见表 1。

表 1 LDA 符号含义对照表

Table 1 Symbols and its meanings in LDA

符号	含义
$\boldsymbol{\alpha}$	狄利克雷分布,$\boldsymbol{\theta}$ 的超参数
$\boldsymbol{\beta}$	狄利克雷分布,$\boldsymbol{\varphi}$ 的超参数
$\boldsymbol{\theta}$	“文档-主题”的多项式分布
$\boldsymbol{\varphi}$	“主题-词汇”的多项式分布
z	词的主题分配
w	词
K	主题数目
M	文档数目
N	一篇文档的词数

根据 LDA 模型,文档的产生过程见算法 1。

算法 1[10] LDA 文档产生过程。

输入 $\boldsymbol{\alpha}$、$\boldsymbol{\beta}$、K;

输出 文档。

对每个主题 $k \in [1,K]$,采样词分布 $\varphi_k \sim \mathrm{Dir}(\beta)$

对每篇文档 $m \in [1,M]$

采样一个主题分布 $\boldsymbol{\theta}_m \sim \mathrm{Dir}(\alpha)$

对文档 m 中的每个词 w

根据 θ_m 采样一个主题 $z \sim \mathrm{Mult}(\boldsymbol{\theta}_m)$

根据主题 z 采样一个词 $w \sim \mathrm{Mult}(\boldsymbol{\varphi}_z)$

其中,隐含变量 $\boldsymbol{\theta}$ 和 $\boldsymbol{\varphi}$ 可按式(1)和式(2)估计:

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_m^{(k)} + \alpha_k)} \tag{1}$$

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} (n_k^{(t)} + \beta_t)} \tag{2}$$

式中:$n_m^{(k)}$ 表示文本 m 中主题为 k 的词汇数目,$n_k^{(t)}$ 表示词 t 中主题为 k 的词汇数目。V 表示不计重复的词汇总数。

2　本文算法

LDA 模型假设每个词都是同等重要的。然而,无论是从信息论或是语言学来看,该假设都并不完美。文献[13]指出高频停用词对 LDA 模型的主题推理有很大影响。然而,对于文本情感分类任务,在去除通常的停用词后,仍有大量与领域相关但对情感分类作用较小的词,具有褒贬倾向的词汇淹没其中,而使得 LDA 模型主题间区分度较小,分类精度不高。以酒店评价语料为例,大量文档中都出现有“酒店”、“房间”、“前台”等词,这些词是情感分类时的广义“停用词”,若不加以处理,将随机散布在各个主题的关键词当中。

由于这些词与领域相关,无法通过构建统一的词表去除该类词汇,给主题的提取和情感倾向的划分造成困难。本文针对情感语料的词汇分布特点,根据每个词与情感种子词的点互信息(point mutual information,PMI)[14],赋予词汇不同权重,并将权值信息融入吉布斯采样过程,利用每个主题下的关键词判断主题的情感倾向,从而实现文档的情感分类。整个算法的步骤如图 2 所示。

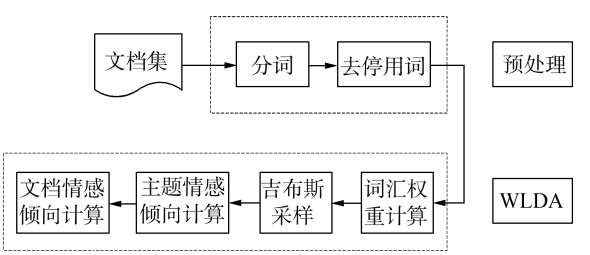


图 2　WLDA 算法步骤

Fig.2　Road map of WLDA algorithm

点互信息可根据两个离散随机变量的共现概率度量其相关性。对于两个变量 x 和 y,其点互信息:

$$\mathrm{PMI}(x,y)=\log\frac{p(x,y)}{p(x)\cdot p(y)} \tag{3}$$

显然,两个变量共现的概率越大,其 PMI 值越大。以此为理论基础,文献[15]根据某一词汇与正面情感种子词和负面情感种子词的 PMI 值度量该词的情感倾向。考虑到种子词在语料中的出现可能不均衡,本文对原公式稍加改动,根据语料中出现的正向和负向种子词个数添加归一化因子。对于词 w,其权重定义为

$$\mathrm{weight}(w)=\frac{1}{a}\sum_{i=1}^{a}\log\frac{p(w,\mathrm{pos}(i))}{p(w)\cdot p(\mathrm{pos}(i))}-\frac{1}{b}\sum_{j=1}^{b}\log\frac{p(w,\mathrm{neg}(j))}{p(w)\cdot p(\mathrm{neg}(j))} \tag{4}$$

式中:pos 为语料中包含的正面情感种子词集合,a 为正面情感种子词个数,neg 为语料中包含的负面情感种子词集合,b 为负面情感种子词个数。

受文献[16]启发,在得到词汇权重后,本文按照式(5)对每个词的主题进行吉布斯采样,式中,W 为词汇总数,$n_{mj}^{(k)}$ 表示文本 m 中,词 i 被分配给主题 k 的数目,$\neg i$ 表示采样过程中不计当前词影响:

$$\begin{aligned}p(z_i=k\mid z_{\neg i},w_{\neg i},\mathrm{weight})=&\\(\sum_{j=1}^{W}\mid\mathrm{weight}(j)\mid\cdot n_{mj,\neg i}^{(k)}+\alpha_k)\cdot&\\(\mid\mathrm{weight}(t)\mid\cdot n_{k,\neg i}^{(t)}+\beta_t)\cdot&\\(\sum_{k=1}^{K}(\sum_{j=1}^{W}\mid\mathrm{weight}(j)\mid\cdot n_{mj,\neg i}^{(k)}+\alpha_k))^{-1}\cdot&\\(\sum_{t=1}^{V}(\mid\mathrm{weight}(t)\mid\cdot n_{k,\neg i}^{(t)}+\beta_t))^{-1}&\end{aligned} \tag{5}$$

整个模型的“文档—主题”分布 $\boldsymbol{\theta}$ 和“主题—词汇”分布 $\boldsymbol{\varphi}$ 可分别按照式(6)和式(7)计算:

$$\theta_{m,k}=\frac{\sum_{j=1}^{W}\mid\mathrm{weight}(j)\mid\cdot n_{mj}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(\sum_{j=1}^{W}\mid\mathrm{weight}(j)\mid\cdot n_{mj}^{(k)}+\alpha_k)} \tag{6}$$

$$\varphi_{k,t}=\frac{\mid\mathrm{weight}(t)\mid\cdot n_{k}^{(t)}+\beta_t}{\sum_{t=1}^{V}(\mid\mathrm{weight}(t)\mid\cdot n_{k}^{(t)}+\beta_t)} \tag{7}$$

与 LDA 模型类似,此处选取每个主题下 φ 值最大的 S 个词作为该主题的关键词。定义主题 k 的情感倾向 $E(k)$:

$$E(k)=\sum_{i=1}^{S}\mathrm{weight}(i)\varphi_{ki} \tag{8}$$

“文档–情感”矩阵 $\boldsymbol{\pi}$ 表征了文档的情感分布,其规模为 $M\times 2$,由正面情感分布 $\boldsymbol{\pi}_{\mathrm{pos}}$ 和负面情感分布 $\boldsymbol{\pi}_{\mathrm{neg}}$ 组成。其定义见式(9)和式(10):

$$\boldsymbol{\pi}_{\mathrm{pos}}=\sum_{i=1}^{K}\theta_i,E(i)>0 \tag{9}$$

$$\boldsymbol{\pi}_{\mathrm{neg}}=\sum_{i=1}^{K}\theta_i,E(i)<0 \tag{10}$$

式中:θ_i 为每一篇文档分配给主题 i 的概率,$\boldsymbol{\pi}_{\mathrm{pos}}$ 和 $\boldsymbol{\pi}_{\mathrm{neg}}$ 分别是文档为正面或负面的概率值,刻画了每篇文档的情感分布情况。在后面的实验中,认为文档 d 的情感倾向:

$$\mathrm{E}(d)=\mathrm{argmax}(\boldsymbol{\pi}_d) \tag{11}$$

完整的 WLDA 算法如下:

算法 2　基于 WLDA 的情感分类算法。

输入　待分类文档,情感种子词;

输出　情感分类结果。

For $w\in W$

按式(4)计算 weight(w)

Repeat

For $m \in M$
For $n \in N$
按式(5)采样每个词的主题
Until 收敛 or 达到最大迭代次数
分别按照式(6)和式(7)计算 θ、φ
For $k \in K$
For $s \in S$
按式(8)计算主题 k 的情感倾向 $E(k)$
If $E(k) > 0$
$\pi_1 = \pi_1 + \theta_k$
If $E(k) < 0$
$\pi_2 = \pi_2 + \theta_k$
For $m \in M$
If $\pi_{m,1} > \pi_{m,2}$
文档情感为正面
Else
文档情感为负面

3 实验结果与分析

3.1 实验设置

语料 1 为中科院谭松波等收集整理的酒店评论语料,从中随机选取带有正向和负向情感倾向标注的评论各 500 篇;语料 2 为从互联网爬取的酒店评论 11 197 篇,包含正向文本 5 891 篇和负向文本 5 306篇。WLDA 和 JST 模型的正面和负面情感种子词来自知网的《中文情感分析用词语集》。实验前,首先对语料进行了分词、去停用词等预处理。

WLDA 参数取经验值 $\alpha = 50/K$, $\beta = 0.01$, $S = 100$。实验以 LSM 和 JST 两种经典算法作为对比,LSM 模型中,选取 $\alpha = 50/K$, $\beta = 0.01$;JST 模型参数设置与文献[6]保持一致。3 种算法的迭代次数均为 1 000 次。

3.2 加权方式对比

表 2 列举了部分词汇在 3 种加权方式下的权重值。

表 2 各加权方式下部分词汇权重对比

Table 2 Term weights in different weighting algorithms

词汇	PMI	IDF	二值化
舒适	3.31	0.63	2
实惠	2.92	0.55	2
很脏	2.16	0.71	0.5
破	3.16	0.70	0.5
服务员	0.10	0.43	0.5
酒店	0.36	0.19	0.5

方法 1 PMI 已在上文详述,方法 2IDF 权重计算方法来自文献[16],方法 3 的二值化见式(12):

$$\text{weight}(w) = \begin{cases} 2w, & \text{在已知情感种子词中} \\ 0.5, & \text{其他} \end{cases} \tag{12}$$

当一个词的权重大于 1 时,表明其作用在采样中将会被增强;小于 1 时,其重要性降低。若将全部权重置为 1,则为一般的吉布斯采样。

方法 1 和方法 3 均能将“舒适”、“实惠”等词赋以较大权重,将部分没有情感色彩的词如“服务员”、“酒店”等赋以较小权重,但对于未收录的情感词汇如“很脏”、“破”等,方法 3 表现不佳。方法 2 将提高出现次数较少的罕见词的权重,而同时降低高频情感词和高频非情感词的权重。综上,3 种方法中 PMI 加权最适用于本文,故以下实验中采用的均是 PMI 加权方式。

3.3 WLDA 和 LSM 模型对比

在主题模型中,通常以各个主题下的关键词来表征该主题的含义。表 3 为采用语料 1 时 WLDA 与 LSM 模型的关键词对比。

表 3 WLDA 和 LSM 关键词

Table 3 Keywords of WLDA and LSM

模型	正面	负面
WLDA	不错 方便 热情 免费 酒店 满意 总体 房间 舒服 干净 挺 下次 舒适 特色 周到 安静 推荐 很快 感动 交通	差 携程 不知道 不能 房间 根本 打电话 酒店 电话 太 前台 只能 告诉 不好 不要 洗澡 退房 失望 服务员 投诉
LSM	酒店 房间 不错 感觉 服务 入住 早餐 方便 免费 小 吃 设施 价格 干净 环境 大 餐厅 下次 服务员 晚上	酒店 房间 前台 入住 携程 服务员 服务 晚上 客人 发现 差 电话 退房 打电话 不能 房 不知道 点 这家 宾馆

在 WLDA 中,超过一半的关键词都具有明显的情感倾向,如“不错”、“方便”、“失望”等,使读者更容易区分主题的情感倾向;而在 LSM 模型中,正如上文所提到的,体现情感的词汇出现较少,而“酒店”、“房间”、“入住”等不能表达明确情感色彩的词散布在正面和负面两类情感的关键词中。

表 4 展示了 WLDA 和 LSM 模型对文档的情感分类精度。在关键词部分,虽然 LSM 中涉及的具有情感倾向的词汇较少,仍可辨别两类关键词的正负

情感倾向。但具体到刻画各个文档的情感,其精度远低于 WLDA,可见这类广义停用词对模型性能的影响。

表 4　WLDA 和 LSM 模型情感分类精度

Table 4　Sentiment classification accuracy of WLDA and LSM　%

模型	正面	负面	总
WLDA	86.8	92.6	89.7
LSM	80.4	70.0	75.2

此处以 LSM 为对比,说明了词汇加权对吉布斯采样结果的影响,但由于 LSM 模型只能将文档划分为正面、负面两类或正面、负面、中性三类,无法提取更细粒度的主题和情感信息,后文的实验均采用 WLDA 与 JST 两个模型的对比。

3.4　WLDA 和 JST 模型的情感分类精度对比

图 3 为 WLDA 和 JST 模型选取不同主题数目时,在语料 1 和语料 2 下的情感分类精度。

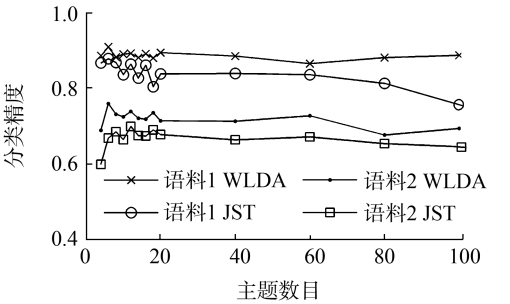


图 3　WLDA 和 JST 模型分类精度对比

Fig.3　Sentiment classification accuracy of WLDA and JST

对于语料 1 和语料 2,WLDA 不仅在情感分类上均有良好表现,受主题数目选取的影响也比 JST 模型更小。

3.5　WLDA 和 JST 模型的关键词对比

在语料 1 中,当 $K=6$ 时,两种算法的分类精度达到最高。表 5 列举了 $K=6$ 时,WLDA 和 JST 模型得到的关键词,并归纳了关键词的主要内容。

表 5　WLDA 和 JST 关键词

Table 5　Keywords of WLDA and JST

编号	情感倾向	主题归纳	WLDA	主题归纳	JST
1	褒义	房间舒适	不错 免费 满意 舒适 周到 舒服 特色 享受 总体 温馨 房间 酒店 宽敞 推荐 安静	房间舒适	酒店 房间 房 不错 入住 携程 大床 服务 感觉 免费 豪华 价格 行政 大堂 设施
2	褒义	服务热情	热情 感动 酒店 不错 帮 小姐 打电话 服务员 安排 工作人员 员工 花园 下次 很快 感谢	房间 服务 餐饮	酒店 服务 入住 服务员 行李 房间 客人 大堂 帮 吃 免费 热情 朋友 安排 早餐
3	褒义	交通方便	不错 方便 总体 酒店 干净 香港 房间 交通 满意 步行太 安静 位置 齐全 免费	房间 餐饮	酒店 不错 房间 感觉 方便 入住 小 干净 环境 吃 早餐 价格 设施 晚上 服务
4	贬义	卫生设施差	洗澡 差 水 根本 太 毛巾 地毯 只能 床单 门 最差 如家 不如 酒店 卫生间	房间设施差	房间 酒店 早餐 服务员 晚上 空调 感觉 不好 差 卫生间 宾馆 装修 不能 太 不知道
5	贬义	投诉交涉	携程 电话 告诉 不能 前台 打电话 不知道 酒店 投诉 退房 收 经理 根本 结帐	投诉交涉	酒店 前台 房间 入住 携程 服务员 电话 客人 服务 退房 打电话 告诉 发现 经理
6	贬义	房间设施差	差 空调 房间 不好 太 失望 不知道 吵 实在 不能 根本 很差 声音 只好 打电话	房间 服务 交通	酒店 房间 携程 服务 价格 感觉 前台 朋友 小 这家 入住 机场 四星 补充 出租车

可以看到,WLDA 得到的关键词多为单一方面评价,一致性较强,易于人的理解。而在 JST 模型中,部分主题由多个方面的评价组成,如主题 2,在 15 个关键词中,同时涉及到房间、服务、餐饮三方面

内容;主题 6 同时涉及房间、服务、交通三方面内容。除此之外,WLDA 的关键词中涵盖的情感词汇更丰富,主题的情感倾向也更加突出。与 JST 模型相比,WLDA 得到的各个主题的关键词语义和情感都更加明晰。

3.6 WLDA 和 JST 模型的主题 KL 距离对比

上文通过关键词的列举直观展示了 WLDA 的性能,本部分将借助主题与背景主题的平均 KL 距离定量描述主题的区分性。其核心思想是一个合理的主题总倾向于在部分文档集中出现,主题在所有文档中出现的概率越平均,说明该主题越可能为垃圾/非重要主题[17]。极端情况,当某个主题在所有文档中出现的概率都相同,该主题对文档的区分能力为零。主题与背景主题的平均 KL 距离 KL_b 定义如下:

$$KL\_b = \frac{\sum_{k=1}^{K} D_{KL}(\theta_k \,||\, \theta\_b_k)}{K} \tag{11}$$

式中:θ_b_k 为 θ_k 的背景主题,其规模与 θ_k 相同,为 $M \times 1$, $\forall i \in M$, $\theta\_b_{ki} = \sum_{i=1}^{M} \theta_{ik}/M$。

表 6 WLDA 和 JST 模型中主题与背景主题的平均 KL 距离

Table 6 Kullback-Leibler divergence of WLDA and JST

主题数目	WLDA	JST
4	201.7	75.2
6	170.7	85.9
8	149.8	76.5
10	131.4	72.0
12	119.2	62.7
14	105.9	58.5
16	93.9	52.3
18	89.8	49.8
20	80.5	46.7
40	46.5	26.6
60	31.1	17.8
80	22.5	12.3
100	16.9	10.1

表 6 展示了 WLDA 和 JST 模型主题与背景主题的平均 KL 距离,其值越大,说明主题与背景主题的距离越远,主题的可区分性越强。可以看到,在各个主题数目下,WLDA 的主题区分能力均优于 JST 模型。

3.7 WLDA 和 JST 模型的时间消耗对比

以语料 1 为例,图 4 对比了 $K = 6$ 时 WLDA 和 JST 模型不同迭代次数所需的时间。

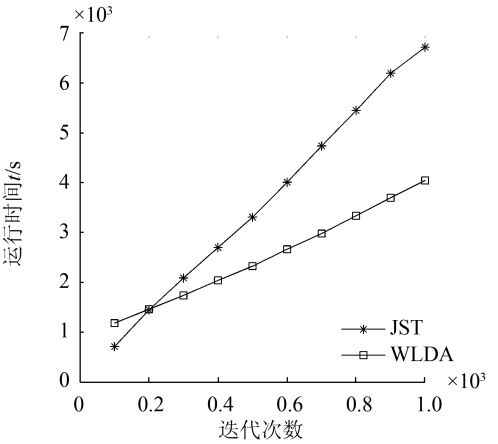


图 4 WLDA 和 JST 模型运行时间对比

Fig.4 Time consumption comparison of WLDA and JST

由于本文算法需要首先计算词汇权重,故吉布斯采样前的处理时间比 JST 模型长,但单次迭代速度比 JST 更快。当吉布斯采样的次数较小时,JST 模型消耗时间更短,然而,随着采样次数的增加,WLDA 的时间优势愈发明显。另外,对于同一语料库,取不同 K 值或其他参数发生改变时无需重复计算词汇权重,故在多次试验中,其平均运行时间将比图 3 所展示的更短。

4 结束语

本文提出了一种用于情感分类的词加权 LDA 算法,通过度量词汇与情感种子词的点互信息,在吉布斯采样中为不同词汇赋予不同权重,并利用每个主题下的关键词判断主题的情感倾向,从而实现文档的情感分类。实验表明,WLDA 不仅具有无监督、可提取细粒度情感的优点,而且分类精度较高,在采样中迭代速度较快。由于 WLDA 采用的是“词袋”模型,忽略了词与词之间的联系,可能会出现局部情感判断错误,因此,如何将词序信息融入 WLDA 是下一步的工作重点。

参考文献:

[1] AGARWAL B, PORIA S, MITTAL N, et al. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach[J]. Cognitive computation, 2015, 7(4): 487-499.

[2] CAMBRIA E. Affective computing and sentiment analysis [J]. IEEE intelligent systems, 2016, 31(2): 102-107.

[3] LIN Chenghua, HE Yulan. Joint sentiment/topic model for sentiment analysis[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China: ACM, 2009: 375-384.

[4] LIN Chenghua, HE Yulan, EVERSON R. A comparative study of Bayesian models for unsupervised sentiment detection [C]//Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Stroudsburg, PA, USA: ACM, 2011: 144-152.

[5] TITOV I, MCDONALD R. A joint model of text and aspect ratings for sentiment summarization[C]//Proceedings of Annual Meeting of the Computational Linguistics. Columbus, USA: Association for Computational Linguistics, 2008: 308-316.

[6] PAUL M, GIRJU R. A two-dimensional topic-aspect model for discovering multi-faceted topics[C]//Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta, USA: AAAI, 2010: 545-550.

[7] MEI Qiaozhu, LING Xu, WONDRA M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs [C]//Proceedings of the 16th International Conference on World Wide Web. North Carolina, USA: ACM, 2010: 171-180.

[8] JO Y, OH A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. Hong Kong, China: ACM, 2011: 815-824.

[9] 欧阳继红, 刘燕辉, 李熙铭, 等. 基于 LDA 的多粒度主题情感混合模型[J]. 电子学报, 2015, 43(9): 1875-1880.
OUYANG Jihong, LIU Yanhui, LI Ximing, et al. Multi-grain sentiment/topic model based on LDA[J]. Acta electronica sinica, 2015, 43(9): 1875-1880.

[10] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. The journal of machine learning research, 2003, 3: 993-1022.

[11] RUBIN T N, CHAMBERS A, SMYTH P, et al. Statistical topic models for multi-label document classification [J]. Machine learning, 2012, 88(1/2): 157-208.

[12] ANDRZEJEWSKI D, BUTTLER D. Latent topic feedback for information retrieval[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA: ACM, 2011: 600-608.

[13] WALLACH H M. Topic modeling: beyond bag-of-words [C]//Proceedings of the 23rd International Conference on Machine Learning. New York, USA: ACM, 2006: 977-984.

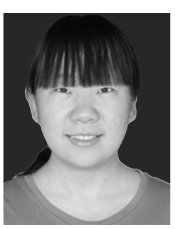
[14] CHURCH K W, HANKS P. Word association norms, mutual information, and lexicography[J]. Computational linguistics, 1990, 16(1): 22-29.

[15] TURNEY P D, LITTMAN M L. Measuring praise and criticism: inference of semantic orientation from association [J]. ACM transactions on information systems, 2003, 21(4): 315-346.

[16] 张小平. 主题模型及其在中医临床诊疗中的应用研究[D]. 北京: 北京交通大学, 2011: 57-58.
ZHANG Xiaoping. Study on topic model and its application to TCM clinical diagnosis and treatment[D]. Beijing: Beijing Jiaotong University, 2011: 57-58.

[17] ALSUMAIT L, BARBARá D, GENTLE J, et al. Topic significance ranking of LDA generative models[C]//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Bled, Slovenia: ACM, 2009: 67-82.

作者简介:



郝洁,女,1992 年生,硕士研究生,主要研究方向为自然语言处理、粗糙集。

谢珺,女,1979 年生,副教授,主要研究方向为粒计算、粗糙集、数据挖掘、智能信息处理。

苏婧琼,女,1991 年生,硕士研究生,主要研究方向为自然语言处理、粒计算。