结合语义相似度改进 LDA 的文本主题分析

赵林静

（中国民航飞行学院 计算机学院，四川 广汉 618307）

摘 要：为对评论文本进行准确的主题分类，提出一种结合 HowNet 语义相似度和隐含狄利克雷分配（LDA）模型的主题聚类方法。不同于传统 LDA 模型，该方法通过 HowNet 常识知识库计算输入单词与当前主题聚类中单词间的语义相似度，以此调整 LDA 模型中的超参数 β。为不同的单词分配不同的 β 值，以此监督聚类过程，在主题分析中实现从语法到语义的转变。实验结果表明，该方法能够有效提高主题聚类的准确性。

关键词：评论短文本；主题分析；HowNet 语义相似度；LDA 模型；超参数 β

中图法分类号：TP391 **文献标识号**：A **文章编号**：1000-7024（2019）12-3514-06

doi：10.16208/j.issn1000-7024.2019.12.025

Modified LDA model based on semantic similarity for topic analysis of text

ZHAO Lin-jing

(School of Computer, Civil Aviation Flight University of China, Guanghan 618307, China)

Abstract: To classify comment texts accurately, a topic clustering method based on HowNet semantic similarity and implicit Dirichlet distribution (LDA) model was proposed. Different from the traditional LDA model, the semantic similarity between the input words and the words in the current topic clustering was calculated through HowNet common sense knowledge base, and the hyper-parameter β in LDA model was adjusted. Different β values were assigned to different words to supervise the clustering process, thus realizing the transformation from grammar to semantics in thematic analysis. Experimental results show that the proposed method can effectively improve the accuracy of topic clustering.

Key words: comment short text; topic analysis; HowNet semantic similarity; LDA model; hyper-parametric β

0 引 言

随着电子商务的迅速发展，互联网上拥有消费者对产品和服务的大量评论信息，具有非常宝贵的研究价格。通过对这些评论进行情感分析，找出客户对商品的评论主题或感兴趣的内容，以此可以提高客户管理和推荐系统的能力[1,2]。但这些信息数量巨大，全部阅读这些评论十分困难. 另外，现有的大多数情感分析工具仅限于提取与整个文档相关的情感极性值，即识别正面或负面评价，而不是提取出评论的主题[3-5]。为了深层次的分析评论数据，需要一种有效的网络评论文本主题挖掘方法。

在文本情感分析中，主题模型（topic modeling）[6]是近些年研究的热点。主题模型对隐含主题进行建模来表示文本，能够显著减少特征数量，比传统 TF-IDF 表示方法有明显改善[7,8]。概率潜在语义分析（pLSA）[9]和隐含狄利克雷分配（LDA）[10]是目前主流的两种主题模型。这两种模型均在观测变量“文档”和“单词”之间引入一个潜在变量“主题”来分析文档的语义主题分布。在主题模型中，每个文档被表示为潜在主题上的随机混合，其中每个主题表示为单词上的分布。LDA 模型给出了文档-主题分布的 Dirichlet 概率生成过程，在每个文档中，根据由 Dirichlet 先验 α 控制的多项式分布选择潜在主题。然后，给定一个主题，根据由另一个 Dirichlet 先验 β 控制的另一个多项式分布来提取单词。LDA 模型以狄利克雷分布来描述主题与单词的生成，从而使模型参数数量不随样本增加而线性增长，解决了 pLSA 模型中的过拟合现象[11,12]。

为了进一步提高 LDA 在主题分类上的精度，学者对其

收稿日期：2019-02-18；**修订日期**：2019-04-08

基金项目：国家自然科学基金民航联合基金重点项目（U1233202/F01）

作者简介：赵林静（1982-），女，重庆人，硕士，副研究馆员，研究方向为计算机网络与信息处理。E-mail: cafuczx@163.com

进行了多种改进。例如，文献［13］提出基于带标签的 LDA 模型（L-LDA），通过加入标签因素解决强制分配隐性主题的问题，但是这种方法使 LDA 模型缺乏处理潜在类别标签的能力。文献［14］提出了一种弱监督分类器，称为 Classify-LDA，首先用 LDA 对训练数据生成一个主题模型，然后由领域专家，根据每个主题最可能的单词为其分配一个或多个类标签，形成一个新的主题模型。给文档分类时，根据文档中的主题比例和主题的类标签对文档进行分类，也取得了不错的分类效果。

本文提出了一个新的 LDA 框架，称为语义相似 LDA（semantic similarity LDA，SS-LDA）。将基于常识的语义相似度计算集成到 LDA 模型中的词分布计算中，用来动态调整超参数 β，以此来监督聚类过程，从而在主题分析中实现从语法到语义的转变。其主要创新点为：利用中文常识知识库 HowNet，通过信息量方法来计算词语间的语义相似度；将这种语义相似度计算嵌入到 LDA 模型中输入单词与主题词汇集相似性的计算中，并根据相似性来动态调整单词的超参数 β，以此提高主题聚类的速度和准确性。

1　传统 LDA 主题模型

LDA 模型是一种文档主题生成模型，其假设文档是主题的混合，每一篇文档表示为由一些主题所构成的一个概率分布。其将隐含主题看作是基于词特征的聚类，从而将表示文本的高维向量空间模型映射到低维潜在语义空间。模型建立如图 1 所示。

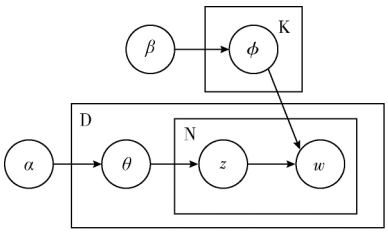

图 1　LDA 概率模型

设 V 是词汇集 $[w_1,\cdots,w_w]$ 的长度。对于每个文档 $d\in\{1,\cdots,D\}$，设定 S_d 是文档 d 中句子数量。对于每个句子 $s_d\in\{1,\cdots,S_d\}$，设定 $N_{(d,s)}$ 是句子 s_d 中单词的数量。假设每个句子都属于一个主题。θ 和 ϕ 分别为文档-主题概率分布和主题-单词概率分布。α 和 β 为分布的超参数，α 表示与 θ 分布相关的 Dirichlet 先验向量，在不损失一般性的情况下，可以对所有主题假定相同的 α 值[15]。β 表示 ϕ 分布相关的 Dirichlet 先验向量。

通常采用 Gibbs 采样算法[16]来求解 LDA，α 和 β 是已知的先验输入，求解目标是得到 z,w 对应的概率分布，即文档主题的分布和主题词的分布。

2　提出的语义相似 LDA 主题模型

2.1　方法描述

通过上述分析可知，对于词汇表中的每个单词，超参数 β 用于嵌入当前单词与每个主题之间的相似性。在传统 LDA 中，β 的值都为事先设定的一个固定值。为此，这种机制不能很好地反映出不同词语的语义信息。

本文提出了一种语义相似 LDA 模型（SS-LDA），将基于常识的语义相似度集成到 LDA 中主题-词分布 ϕ 计算中，以此来提高聚类准确性，其模型结构如图 2 所示。与传统 LDA 不同的是，本文根据常识知识计算当前词与当前主题集群中单词的相似度，为不同的单词赋予不同的 β 值，而不是一个固定值。其中，常识是指人类理解所依赖的背景知识。对于中文词汇，本文从知网中文词库（HowNet）的情感词典中提取这些知识。

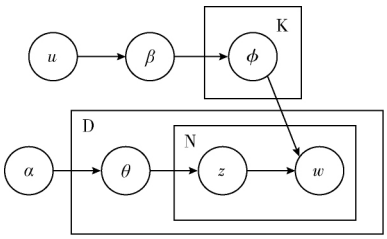

图 2　SS-LDA 模型框架

本文方法的主要思想是基于常识知识能够更好地匹配词之间的相似性，而词语共现频率仅在句法层面上表征相似性。语义相似性能够处理自然语言之间的概念相似性，从而能够有效地指导一致性词语-主题概率分布的创建，从而获得每个主题更可靠的主题词。

2.2　LDA 生成过程

用 μ 表示常识知识在计算 β 中的贡献。通过计算当前单词与当前主题词汇集中的前 50 个单词之间的平均相似度，在 Gibbs 采样算法的每次迭代中更新每个单词的 β 值。将阈值设置为 50 个单词是为了最小化聚类中的噪声，同时覆盖语料库中的几乎所有重要的主题词。

如果单词 w_i 与主题 z_k 的相关性较大，表明这个词是该主题的重要相关词，则对应的 β 赋予较高的值；如果它是一般词，则 β 值较低。这是因为重要相关词能够引导 LDA 聚类过程，通过修改概率分布以获得更可靠的主题。β 先验可以用如下结构的矩阵来表示

$$\begin{bmatrix} & z_1 & z_2 & \cdots & z_K \\ w_1 & \beta_{11} & \beta_{12} & \cdots & \beta_{1K} \\ \cdots & & & & \\ w_i & \beta_{i1} & \beta_{i2} & \cdots & \beta_{iK} \\ \cdots & & & & \\ w_N & \beta_{N1} & \beta_{N2} & \cdots & \beta_{NK} \end{bmatrix} \tag{1}$$

其中，z 表示主题，K 表示主题的数目，w 表示单词，N 表示单词的数目。

LDA 生成过程如下：

$\forall k \in \{1,\cdots,K\}$：

$\forall w_i \in [w_1,\cdots,w_W]$：

$\varphi_{k,i} \sim Dir(\beta_i)$ 主题在词上的分布

$\forall d \in \{1,\cdots,D\}$：

$\theta_d \sim Dir(\alpha)$ 文档在主题上的分布

$\forall s_d \in d$：

$z_{(d,s)} \sim Multi(\theta_d)$ 构建一个主题

$\forall w_i \in s_d$：

$w_i \sim Multi(\varphi_{z(d,s)})$ 构建一个词

发出 w_i

采用 Gibbs 抽样来估计 $P(z/w;\alpha,\beta)$ 的后验分布。由于 $P(z/w;\alpha,\beta) = \dfrac{P(z/w;\alpha,\beta)}{\sum_z P(z/w;\alpha,\beta)}$，Gibbs 抽样方程可由具有主题 z 的单词 w 的联合似然导出。

给定模型的总概率

$$P(z,w,\theta,\phi;\alpha,\beta) = P(\phi;\beta)P(\theta;\alpha)P(z/\theta)P(w/\phi) \quad (2)$$

在 θ 和 ϕ 上积分可以得到 w 和 z 的联合似然概率

$$P(z,w;\alpha,\beta) = \int P(z/\theta)P(\theta;\alpha)d\theta \cdot \int P(w/\phi)P(\phi;\beta)d\phi \quad (3)$$

由于所有项都是具有 Dirichlet 先验的多项式，并且 Dirichlet 分布与多项式分布共轭，因此可以导出

$$P(z,w;\alpha,\beta) = \prod_d \frac{B(n_{d,(.)}^{sent}+\alpha)}{B(\alpha)} \cdot \prod_k \frac{B(n_{(.),k}^{U}+\beta)}{B(\beta)} \quad (4)$$

其中，$n_{d,k}^{sent}$ 表示文档 d 中分配主题 z_k 的句子的数量，$n_{d,k}^{U}$ 表示文档 d 中分配主题 z_k 的单词的数量。另外，$B(\alpha) = \dfrac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$ 和 $B(\beta) = \dfrac{\prod_r \Gamma(\beta_r)}{\Gamma(\sum_r \beta_r)}$（对词汇集 V 中任意单词 r）。

假定，在马尔可夫链的每个步骤处的后验分布可以表示如下

$$P(z_{(d,s,i)} = k/z_{-(d,s,i)},w;\alpha,\beta) = \frac{P(z,w;\alpha,\beta)}{P(z_{-(d,s,i)},w;\alpha,\beta)} = \prod_d \frac{B(n_{d,(.)}^{sent}+\alpha)}{B(n_{d,(.)}^{sent,-(d,s,i)}+\alpha)} \cdot \prod_k \frac{B(n_{(.),k}^{U}+\beta)}{B(n_{(.),k}^{U,-(d,s,i)}+\beta)} \quad (5)$$

其中，i 是当前单词编号，s 是当前句子，d 是当前文档，(d,s,i) 表示省略文档 d 中句子 s 的第 i 个单词。那么，可以得到

$$P(z_{(d,s,i)} = k/z_{-(d,s,i)},w;\alpha,\beta) \propto (n_{d,k}^{sent,-(d,s,i)}+\alpha_k) \cdot \frac{n_{v,k}^{U,-(d,s,i)}+\beta_v}{\sum_{r=1}^{V} n_{r,k}^{U,-(d,s,i)}+\beta_r} \quad (6)$$

其中，v 是第 i 个单词在词汇集 V 中的位置。

然后修改 LDA Gibbs 采样算法，用来添加常识，具体步骤如下所示：

设置：文档个数为 D，每个文档 $d \in \{1,\cdots,D\}$ 有 S_d 个句子，迭代数为 T，文档 d 中句子 s 中词的数量为 $N_{(d,s)}$，主题数为 K，w_i 是句子中第 i 个词，$s_{i,k}$ 是句子第 i 个词与主题 z_k 对应的聚类之间的相似性度量。

随机初始化 z；$\forall s_d \in \{1,\cdots,S_d\}$，其中 $d \in \{1,\cdots,D\}$；

for $t = 1 \rightarrow T$

for $i = 1 \rightarrow N_{d,s}$

$n_{d,s,z} = n_{d,s,z} - 1$，$n_{i,z} = n_{i,z} - 1$；

for $k = 1 \rightarrow K$

计算相似度 $s_{i,k,t}$；

if $(s_{i,k,t} > s_{i,k,t-1} + 0.1)$，then $\beta_v = \beta_v + 0.001$；

if $(\beta_v > 1)$，then $\beta_v = \beta_v - 0.001$；

Else if $(s_{i,k,t} < s_{i,k,t-1} - 0.1)$，then $\beta_v = \beta_v - 0.001$；

if $(\beta_v < 0)$，then $\beta_v = \beta_v + 0.001$；

$$P(z = k/w) = (n_{d,k}^{sent} + \alpha_k) \frac{n_{v,k}^{U} + \beta_v}{\sum_{r=1}^{V} n_{r,k}^{U} + \beta_r} \frac{n_{v,k,l}^{S} + \gamma_l}{\sum_{m=1}^{V_l} n_{m,k,l}^{S} + \gamma_m};$$

从 $P(z/w)$ 采样得到主题；

$n_{d,s,z} = n_{d,s,z} + 1$；$n_{i,z} = n_{i,z} + 1$；

在初始化变量之后，算法会在一定数量的迭代次数内运行。在每个循环中，为句子中的每个单词分配主题。在建立分布之前，必须删除当前分配，并减少计数。然后计算相似性度量 s，并相应地更新 β。

改变 β 的值意味着修改单词属于某个主题的概率。在对单词 w_i 的每次迭代中，计算 w_i 与某个主题聚类中前 50 个单词之间的平均相似度。相似性度量的计算将在下一节中进行描述。如果第 i 个单词的相似度值与前一次迭代相比减少了一定值，那么相应的 β 也必须减少。实验中，我们设置增加或减少 β 的步长为 0.001。只有当单词的相似性分数与前一次迭代相比增加或减少至少 0.1 时，我们才改变 β。这个值也是根据多次实验经验来设定的。最后，计算后验概率，然后对分布进行采样以获得主题，并且增加相关计数。

2.3 基于 HowNet 的单词语义相似度计算

通过上文可知，本文是通过计算单词与某个主题聚类中前 50 个单词之间的平均相似度来调整对应的 β 值。为了准确计算词语间的语义相似性，采用了 HowNet 知识库和通过信息量表征相似性的方法。

HowNet 是一个汉语常识知识库，可以表示概念之间和概念属性之间的联系。与英文 WordNet 知识库不同，HowNet 采用分层系统，其中概念（又称为义项）是对词汇语义的一种描述，每个词可以表达为几个概念，概念由

一系列义原描述。

在传统相似度计算中，义原的语义相似度计算公式

$$S(p_1,p_2)=\frac{k}{d+k} \tag{7}$$

其中，p_1 和 p_2 表示两个义原，d 表示 p_1 和 p_2 在层次体系中的路径距离，d 越大则相似度越小。相似度的取值范围为 [0,1]。k 是一个可调节的参数，通常默认设置为 20。

目前国内计算汉语词语相似度均是基于语义距离的方法。在 WordNet 中，可以通过信息量计算来计算相似度。其基本思想是基于拥有很多下位关系节点的概念比叶子节点概念所含的信息量少这一原理。为了提高中文词汇相似度计算的准确性，本文在 HowNet 词典的层次体系树中也使用信息量来计算词汇的语义相似度，取代传统的义原路径距离。

首先，计算义原 p 的信息量

$$ic_{hn}(p)=1-\frac{\log(hypo(p)+1)}{\log(\max_{hn})} \tag{8}$$

函数 $hypo(p)$ 表示义原子节点的数量，$\max_{hn}$ 表示义原的总数量。那么，两个义原之间的语义相似度为

$$S(p_1,p_2)=\max_{p\in(p_1,p_2)} ic_{hn}(p) \tag{9}$$

然后，通过义原相似度来计算概念相似度。由于词语一般有多个概念（多义词），每个概念又由多个义原描述。假设概念 n_1 包含 n 个义原 $N_1=\{p_{11},p_{12},\cdots,p_{1n}\}$，概念 n_2 包含 m 个义原 $N_2=\{p_{21},p_{22},\cdots,p_{2m}\}$，那么概念 n_1 和 n_2 间的相似度为

$$S_{hn}(n_1,n_2)=S_L(N_1,N_2)\frac{\min(C_1,C_2)}{\sqrt{C_1C_2}} \tag{10}$$

式中：$S_L(N_1,N_2)$ 是两个概念集合的相似度，为集合中所有概念对的相似度的算术平均。C_1 和 C_2 分别表示概念 n_1 和概念 n_2 的记录数目，用于修正 $S_L(N_1,N_2)$ 的误差。

最后，通过获得的两个词语的概念相似度来计算词语相似度。假设词语 w_1 包含 k 个概念 $\{n_{11},n_{12},\cdots,n_{1k}\}$，词语 w_2 包含 p 个概念 $\{n_{21},n_{22},\cdots,n_{2p}\}$，那么词语 w_1 和词语 w_2 间的语义相似度为

$$S_{hn}(w_1,w_2)=\frac{\sum_{i=1}^{k}\sum_{j=1}^{p}S(n_{1i},n_{2j})}{k\cdot p} \tag{11}$$

上式在计算词语相似度时考虑了一词多义现象，为此本文采用了各概念的平均值，使计算结果更接近 HowNet 对于词语的客观描述。

3 实验及评估

3.1 实验设置

为了评估提出的方法，使用了 Semeval 2014 数据集[17]，该基准数据集是从酒店评论网站 TripAdvisor. com 上收集的。本文从中提取了 5793 条酒店评论短文本，并设置以下几个评论主题类别：价格，位置，服务，房间设施，安全，环境。

将本文 SS-LDA 模型与现有的几种现有技术进行比较，分别为标准 LDA 模型和文献 [14] 提出的 Classify-LDA 模型。对于所有模型，在 3000 次 Gibbs 采样和 200 次迭代后，取一个样本作为后验推理。设 K 为主题数，数值为 5，设定 α 为 $50/K$。对于 LDA 模型和 Classify-LDA 模型，设置 β 为 0.1。对于本文方法，每个单词存在不同的 β，初始化为 0.1。

另外，在对所有评论文本进行主题分类之前，通过自然语言处理工具包（NLTK）对句子进行分词，并删除停止词和标点符号。然后，根据非情感词列表从句子中清除嘈杂的非情感词。在具有 Intel 酷睿 I5 5250@2.7GHz CPU，8 GB 内存和 Windows 7 64 位 PC 器上，通过实验工具 Gibbs LDA++用于 LDA 模型的训练和推断。

3.2 主题词聚类分析

表 1 显示了几种方法通过训练生成的其中 3 个主题的主要特征词汇，其中粗体标注的为不相关的词汇。可以看出，LDA 和 Classify-LDA 产生的特征词中都存在一些与主题相关性不大的错误词汇。与其它方法相比，本文 SS-LDA 产生了各主题最相关的特性词。

表 1 生成的一些主题词汇示例

主题	LDA	Classify-LDA	SS-LDA
价格	价格、支付、厨房、便宜、折扣、不吸烟、花费、人民币、整洁	价格、昂贵、花费、房间、便宜、价格、质量、位置、钱	价格、花费、便宜、值得、钱、支付、人民币、昂贵、实惠
位置	位置、汽车、价格、出租车、员工、区域、停车场、地点、酒店	位置、站点、停车场、餐厅、酒店、服务、区域、街道、地点	位置、区域、站点、酒店、机场、距离、汽车、停车场、方便
服务	服务、食品、酒吧、毛巾、早餐、床、大、经理	服务、员工、招待、食物、干净、经理、维护、休息室	服务、员工、招待、早餐、经理、帮助、酒吧、接待、游泳池

为了进一步评估聚类模型对主题特征词的聚类准确性，使用 Rand-index 和熵（Entropy）作为度量，将各方法获得的主题特征词与一个标准特征词集合进行比较。设定 LDA 模型生成数据 $D_{LDA}=d_1,d_2,\cdots,d_n$，$n$ 是主题的数量，d_i 表示第 i 个主题的前 k 个特征单词集合。本文从 Semeval 2014 数据集中通过人工根据常识获得一个标准的主题特征词集合 $D_{gold}=g_1,g_2,\cdots,g_n$，对于每个主题，找出前 50 个特征单词。

（1）Rand-index：用于衡量聚类算法输出与标准数据集之间的相似性。Rand-index 的值越大，主题聚类获得的结果越好。对于一个特征词，设定 x 表示该词在 D_{LDA} 和 D_{gold} 中属于同一主题的次数，y 表示不在同一主题中出现的次数。Rand-index 计算如下

$$RandIndex = 2 * \frac{(x+y)}{n*(n-1)} \tag{12}$$

（2）Entropy：用于测量在聚类算法生成的主题词集合 d_i 中，包含标准集合 g_i 中对应词汇的比例。较小的熵值表示更好的准确性。Entropy 计算如下

$$Entropy(d_i) = -\sum_{j=1}^{n} Pr_i(g_j)\log_2 Pr_i(g_j)$$

$$Entropy(d_{LDA}) = -\sum_{i=1}^{n} \frac{|d_i|}{|d_{LDA}|} Entropy(d_i) \tag{13}$$

表 2 显示了各种方法的主题词聚类性能结果，其中分别计算了聚类中前 15 和前 30 个单词的准确性。可以注意到，LDA 和 Classify-LDA 在 Rand-index 方面表现较为相似。SS-LDA 在 Rand-index 方面优于 LDA 和 Classify-LDA。对于 Entropy 指标，Classify-LDA 优于 LDA，而 SS-LDA 再次获得最佳性能。这是因为当主题的个数较多时，在 Classify-LDA 算法中的人工标注阶段，可能会出现某个词与多个主题都较为对应的情况，就可能会出现标注错误的后果，这对实验的结果会造成一定的影响。而本文 SS-LDA 使用常识知识充当 LDA 模型的监督，基于单词之间的相似性将两个相关或相似单词分配到同一个主题中。此外，所有算法的性能随着考虑的单词数量的增加而降低。

表 2 各种算法的主题词挖掘性能比较

	单词数	LDA	Classify-LDA	SS-LDA
Rand-index	top 15	0.763	0.794	0.891
	top 30	0.742	0.776	0.863
Entropy	top 15	1.434	1.202	0.917
	top 30	1.758	1.456	0.951

3.3 文本主题分类性能

为了测量评论文本的主题分类性能，使用了两个度量标准，分别为准确性和 F-度量。其中设定 TP 表示正确分类的阳性样本，FP 表示错误分类的阳性样本，TN 表示正确分类的阴性样本，FN 表示错误分类的阴性样本。

（1）准确性度量（Accuracy），用于评估分类器正确分类的性能，表示如下

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

（2）F-度量（F-measure），用于评估分类器的整体性能。其中，F-度量由精确性（Precision）和召回率（Recall）计算而来，分别表示如下

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

$$Recall = \frac{TN}{TN+FN} \tag{16}$$

$$F\text{-}Measure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{17}$$

基于上述各种算法挖掘的主题词，在 Semeval 2014 数据集上进行评论文本的主题分类。表 3 显示了实验结果。结果表明 SS-LDA 在准确性和 F-度量方面均优其它方法。由于 LDA 和 Classify-LDA 方法中提取的主题词存在噪声，这会大大影响其召回率，所以获得的 F-度量性能不佳。

表 3 评论文本主题分类结果

分类器	准确性/%	精度/%	召回率/%	F-度量
SS-LDA	0.92	0.87	0.89	0.88
Classify-LDA	0.87	0.84	0.80	0.83
LDA	0.83	0.81	0.75	0.77

4 结束语

大多数情感分析工具仅限于提取与整个文档相关的极性值，而不是区分为单个主题。此外，这些方法主要依赖于明确表达情绪状态的文本部分的统计属性，因此无法捕捉隐含表达的观点和情绪。本文提出了一个框架，称为语义相似 LDA 模型，它在 LDA 算法中计算单词分布时集成了常识推理，从而在主题情感分析中实现了从语法到语义的转换，SS-LDA 通过利用与单词相关的语义，超越了仅仅依靠统计的方法，因此大大提高了聚类。

在今后工作中，将考虑融入情感分析部分，即实现一条评论文本的主题提取和情感分析的双重功能。

参考文献：

[1] LI Hanyu，QIAN Li，ZHOU Pengfei. Sentiment analysis and mining of product reviews [J]. Information Science，2017，35 (1)：51-55 (in Chinese). [李涵昱，钱力，周鹏飞. 面向商品评论文本的情感分析与挖掘 [J]. 情报科学，2017，35 (1)：51-55.]

[2] Ravi K，Ravi V. A survey on opinion mining and sentiment analysis [J]. Knowledge-Based Systems，2015，89 (3)：14-46.

[3] Tajinder Singh，Madhu Kumari. Role of text pre-processing in twitter sentiment analysis [J]. Procedia Computer Science，2016，89 (3)：549-554.

[4] ZHANG Lin，QIAN Guanqun，FAN Weiguo，et al. Sentiment analysis based on light reviews [J]. Journal of Software，2014，25 (12)：2790-2807 (in Chinese). [张林，钱冠群，樊卫国，等. 轻型评论的情感分析研究 [J]. 软件学报，2014，25 (12)：2790-2807.]

[5] Wankhede R，Thakare A N. Design approach for accuracy in movies reviews using sentiment analysis [C] //IEEE International Conference of Electronics，Communication and Aerospace Technology，2017：212-216.

[6] Pavlinek M，Podgorelec V. Text classification method based on self-training and LDA topic models [J]. Expert Systems with Applications，2017，80 (1)：83-93.

[7] Callaghan D，Greene D，Carthy J. An analysis of the coherence of descriptors in topic modeling [J]. Expert Systems with Applications An International Journal，2015，42 (13)：5645-5657.

[8] CHEN Ting，LIU Jianxun，CAO Buqing，et al. Web services clustering based on Biterm topic model [J]. Computer Engineering and Science，2018，40 (10)：1737-1745 (in Chinese). [**陈婷，刘建勋，曹步清，等. 基于 BTM 主题模型的 Web 服务聚类方法研究** [J]. **计算机工程与科学**，2018，40 (10)：1737-1745.]

[9] Fernández Beltrán R，Pla F. Incremental probabilistic latent semantic analysis for video retrieval [J]. Image & Vision Computing，2015，38 (2)：1-12.

[10] Anandkumar A，Foster D P，Hsu D，et al. A spectral algorithm for latent dirichlet allocation [J]. Algorithmica，2015，72 (1)：193-214.

[11] YANG Mengmeng，HUANG Hao，CHENG Luhong，et al. Short text classification based on LDA topic model [J]. Computer Engineering and Design，2016，37 (12)：3371-3377 (in Chinese). [**杨萌萌，黄浩，程露红，等. 基于 LDA 主题模型的短文本分类** [J]. **计算机工程与设计**，2016，37 (12)：3371-3377.]

[12] PENG Yun，WAN Changxuan，JIANG Tengjiao，et al. Extracting product aspects and user opinions based on semantic constrained LDA model [J]. Journal of Software，2017，28 (3)：676-693 (in Chinese). [**彭云，万常选，江腾蛟，等. 基于语义约束 LDA 的商品特征和情感词提取** [J]. **软件学报**，2017，28 (3)：676-693.]

[13] LI Wenbo，SUN Le，HUANG Ruihong，et al. Text classification based on Labeled-LDA model [J]. Chinese Journal of Computers，2008，31 (4)：620-627 (in Chinese). [**李文波，孙乐，黄瑞红，等. 基于 Labeled-LDA 模型的文本分类新算法** [J]. **计算机学报**，2008，31 (4)：620-627.]

[14] Hingmire S，Chakraborti S. Topic labeled text classification：A weakly supervised approach [C] //37th International ACM SIGIR Conference on Research & Development in Information Retrieval. New York：ACM，2014：385-394.

[15] Al-Salemi B，Aziz M J A，Noah S A. LDA-AdaBoost. MH：Accelerated AdaBoost. MH based on latent Dirichlet allocation for text categorization [J]. Journal of Information Science，2015，41 (1)：27-40.

[16] Cheung S H，Bansal S. A new Gibbs sampling based algorithm for Bayesian model updating with incomplete complex modal data [J]. Mechanical Systems & Signal Processing，2017，92 (1)：156-172.

[17] Wang H，Yue L，Zhai C. Latent aspect rating analysis on review text data：A rating regression approach [C] //ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. Washington：ACM，2010：123-126.