研究与开发

融入词汇共现的社交网络用户情感 Biterm 主题模型

顾秋阳[1,2]，吴宝[1,2]，琚春华[3]

（1. 浙江工业大学管理学院，浙江 杭州 310023;
2. 浙江工业大学中国中小企业研究院，浙江 杭州 310023;
3. 浙江工商大学管理工程与电子商务学院，浙江 杭州 310018）

摘　要：近年社交网络用户数量不断增加，基于文本的用户情感分析技术得到普遍关注和应用。但数据稀疏性、精度较低等问题往往会降低情感识别方法的精度和速度，提出了用户情感 Biterm 主题模型（US-BTM），从特定场所的文本中发现用户偏好及情感倾向，有效利用 Biterm 进行主题建模，并使用聚合策略形成伪文档，为整个文本集创建词汇配对以解决数据稀疏性和短文本等问题。通过词汇共现算法对主题进行研究，推断文本集级别信息的主题，并通过分析特定场景下的评论文本集中的词汇配对集及其相应主题的情感，达到准确预测用户对特定场景的兴趣、偏好和情感的目的。结果证明，所提方法能准确地捕捉用户的情感倾向，正确地揭示用户偏好，可广泛应用于社交网络的内容描述、推荐及社交网络用户兴趣描述、语义分析等多个领域。

关键词：词汇共现；社交网络；用户情感；Biterm 主题模型；聚合策略

中图分类号：TP18

文献标识码：A

doi: 10.11959/j.issn.1000−0801.2020302

Biterm topic model of social network users' sentiment by integrating word co-occurrence

GU Qiuyang[1,2], WU Bao[1,2], JU Chunhua[3]

1. School of Management, Zhejiang University of Technology, Hangzhou 310023, China
2. China Institute for Small and Medium Enterprises, Zhejiang University of Technology, Hangzhou 310023, China
3. School of Management Science & Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

Abstract: With the increasing number of social network users in recent years, text-based user sentiment analysis technology has been widely concerned and applied. However, data sparsity and low accuracy often reduce the accuracy and speed of emotion recognition methods. The user emotion Biterm topic model (US-BTM) was proposed which could find user preference and emotional tendency from the text of specific places, so as to effectively use Biterm for topic modeling. The strategy of user aggregation to form pseudo-documents was used, and word pairs were created for the whole corpus to solve the problems of data sparsity and short text. Then the topic was studied through the lexical co-occurrence model, so as to infer the topic with abundant corps-level information, and the purpose of accurately predicting the user's interest, preference and emotion to the specific scene was achieved by analyzing the lexical matching set in the comment corpus under the specific scene and the emotion of the corresponding topic. The

收稿日期：2020−03−17；修回日期：2020−11−10

experimental results show that the method proposed can accurately capture users' emotional tendency and correctly reveal users' preference, which can be widely used in social network content description, recommendation, social network user interest description, semantic analysis and other fields.

Key words: vocabulary co-occurrence, social network, user sentiment, Biterm topic model, aggregation strategy

1 引言

近年来社交网络用户量不断增加，其价值也受到了各方的广泛关注。据中国互联网信息中心统计：截至 2019 年 6 月，我国网民规模达 8.54 亿人，与 2018 年年底同比增长 2 598 万人，互联网普及率达 61.2%，比 2018 年年底同比提升 1.6%。截至 2018 年年底，新浪微博作为我国最大的社交网络应用，活跃用户超过 4.62 亿人，连续 3 年增长超过 7 000 万人；而其垂直领域数量则扩大至 60 个，月阅读量过百亿领域达 32 个。越来越多的用户在社交网络中进行信息、照片和语言等形式的交互。但社交网络中的信息交互往往以短文本的形式出现（即具有一定的嘈杂性和稀疏性），这使得情感分析和主题发现的难度大大提高[1]。

事实上，几乎所有现有市面上的应用都需要进行信息推荐、用户兴趣点判断、主题检测和语义分析。在进行上述判断的过程中，由于社交网络文本中的大多词汇只出现一次，且没有足够的上下文用于识别歧义词，社交网络中的短文本存在严重的数据稀疏性问题[2]。Devi 等[3]提出了融合文本聚合策略和潜伏狄利克雷分配（LDA）、融合 Unigrams 算法和稀疏主题模型以解决社交网络中普遍存在的数据稀疏性问题。在线社交网络中的信息一般是用户进行信息发布、评论和转发时产生的意见或感受。而社交网络的特征决定了用户会关注许多不同的主题，并使用不同方式表达自身情感。由于主题为一组相互关联的词汇，而关联效果又是通过词汇体现的，因此本文结合刘洺辛等[4]的做法，利用词汇共现模式，以更好地揭示主题。由于确定主题会涉及较广的范围，故概率主题模型被广泛应用于用户情感分析的研究中，为更好地模拟用户情感及其关注的主题，本研究通过考虑用户及其进行评论时所附带的情感，对其情感中所隐含的主题进行定位。

本文所提用户情感 Biterm 主题模型（user sentiment-Biterm model，US-BTM）结合用户及其情感取向，并利用 Biterm 主题模型（即词共现模式，一个常用且适合在短文本环境下进行文本分析的主题模型，用于挖掘短文本深层的语义关系）和词汇配对进行主题建模。该方法可以同时产生积极/消极主题词、用户积极/消极主题、动态主题分布。该方法融合经典 LDA 技术与文档聚合策略，为主题学习建立词汇共现模型，形成一种改进主题推理方法。同时，US-BTM 方法扩展了用户情感主题模型和作者主题模型，在作者用户与主题间增加了一个情感层，并通过 Biterm 对主题进行获取。US-BTM 法的理论框架旨在建立一个更加连贯的模型，能够清楚地捕捉用户喜欢或不喜欢的主题，即利用情绪导向信息寻找积极/消极的主题。该模型的一个重要特征是在无监督条件下，能够准确捕捉用户的情感趋势。

US-BTM 模型理论架构如图 1 所示。通过调查用户的意见，将有效的文本分析结合到主题提取中，而不是仅通过建模了解用户讨论形成文本的内容。该模型通过提取用户不同的情感趋势主题分析用户对不同主题的喜好和兴趣。该模型提取 Biterm，并通过情感计算极性。在完成情感极性分类后将结果反馈给 LDA 模型，并根据用户评论在正确的情绪标签下的特征提取积极/消极的主题。以此通过提取主题准确地分析用户情感。

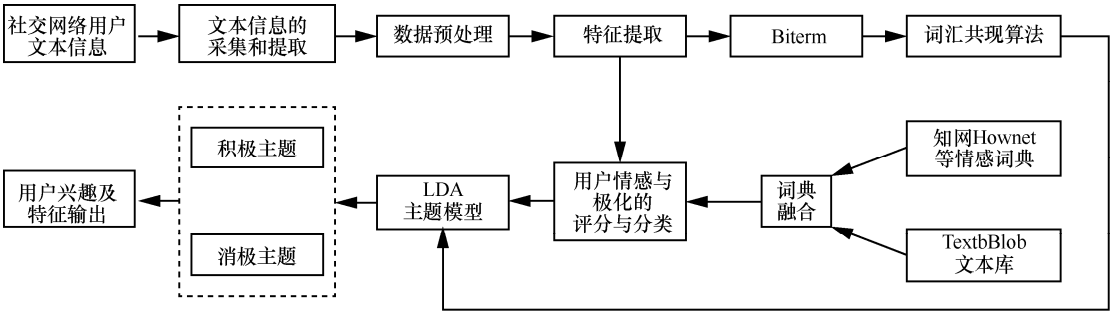

图 1　US-BTM 模型理论架构

2　文献综述

2.1　文献回顾

社交网络的用户评论是近年来通信领域出现的最新文本形式，即用户在网络上以图片、视频和语音等形式共享短文本。不同于博客、小说等形式，用户在社交网络上发布的信息多属于短文本，故传统的 LDA（linear discriminant analysis）模型不能直接应用于短文本的主题提取中，故从社交网络的文本中提取主题具有一定的难度，此问题主要反映在短文本分析中[5]。现今学术界对于社交网络中的用户评论的文本分析往往基于 LDA 模型，其假定文档服从主题分布，而主题服从词汇分布[6]。蔡永明等[7]指出：LDA 模型是一个概率生成模型，也是一个多层次的层次贝叶斯模型（Bayesian hierarchical modeling）。Li 等[8]提出了一种新的基于概率的短文本伪文档主题模型，该模型试图通过隐式聚合短文本克服数据稀疏性问题。此外，周孟等[9]构建了稀疏增强的 PTM（performance test model），该模型涉及应用先验，旨在消除伪文档和潜在主题间存在的非理想相关性问题。欧阳继红等[10]基于经典 LDA 模型提出了 3 个主题建模方案和一个扩展的作者主题模型（authortopic model，ATM）。实验结果表明，与其他基线模型相比，由同一用户聚合的信息取得的结果更为显著。

Li 等[11]提出了一种新的方法——Biterm 主题模型（BTM model），该模型在通过词汇共现方法直接研究文本主题。实证结果表明，与其他基准方法相比，BTM 模型的结果能挑选出文本中更有价值的主题；且与基准方法相比其性能更好、精度更高。为了解决短文本存在的数据稀疏性问题。Mei 等[12]提出了一种在整个语料库中捕获共现词的新方法。为了更有效地处理大规模数据集，孙锐等[13]引入了两种 BTM 在线算法。实验结果表明，该模型能够通过学习训练产生更高质量的主题，且文档主题正确的比例也高于基准。Priyan 等[14]提出了一种新的区分性 Biterm 主题模型（discriminative Biterm topic model），该模型通过区分主题词、一般词与文献特定词，去除了低指示性的 Biterm。实验结果表明，当上述方法应用于研究标题时，可以引发潜在主题；可以在实现 LDA 模型效果的基础上，将文本中语言区分为主题术语、文档特定术语和一般术语。为了对新闻标题的主题词进行归纳，Balan 等[15]引入了判别式构建了双词 Biterm 主题模型（d-BTM），Bicalho 等[16]提出了一种考虑用户个性化和聚合 Biterm 主题模型，以了解用户特定的主题分布。另外，Bicalho 等[16]还通过背景主题、背景词和主题词来研究用户偏好，构建 T-BTM 模型。实验结果显示，T-BTM 模型的性能超过了几种先进的基准主题词提取方法。Li 等[8]将 BTM 主题模型与 *K*-means 聚类算法相结合，以期对文本主题进行挖掘。实验结果显示这两者的结合在一定程度上解决了数据稀疏性问题，并提高了聚类精度。Nguyen 等[17]提出了一种基于词汇共存的 Dirichlet 过程，该过程能够自

动从短文本中进行挖掘主题词提取。实验结果表明，该模型在主题提取质量、精度等方面均优于其他基准方法。周孟等[9]在把词汇的情感极性的倍增过程分解为两个层次。第一层次为判断词汇为情感词汇或普通词汇；第二层次为如果发现该词是一个情感词汇，就自动对该词汇的极性进行定义。

Lin 等[18]对不同文本聚类的方法进行分析，旨在克服数据稀疏性问题。实验结果表明，通过散列标签实现推文聚合的新方法能够有效改进主题一致性的度量，同时自动散列标签也已达到进一步改善聚合效果的目的，从而改进 LDA 主题模型在信息聚类中的应用。朱宪莹等[19]提出了主题情感混合模型（topic sentiment model，TSM），并将网络信息的固有主题进行重新定义和分类。实证结果表明，该法对情感和主题的分类都非常有效。短文本的性质为解决数据稀疏性问题铺平了道路，也给主题建模带来了新的挑战。Kumar 等[20]通过使用线程池策略解决了短文本的碎片化本质，池策略根据地点和用户聚合用户评论。孙锐等[21]提出了基于位置的动态情绪主题模型，该模型对情绪、主题、时间和地理位置等细节信息进行综合建模，通过从社交网络文本中发现情感和主题，并研究情感与主题在自然灾害中处于不同地理位置所发生的变化，实验结果表明该模型表现良好。

Manogaran 等[22]在社交网络短文本环境中进行建模，提出了一种综合性的短文本主题模型（lifelong topic modeling，LTM），该模型提出了一种灵活的短文本聚类过程，并对用户感兴趣的内在变量进行了评估。王建成等[23]建立在多任务学习的框架下基于 BTM 模型提出了一种新的情感分析方法，同时推测一段对话的主题分布和每个句子的情感倾向。张佳明等[24]提出一种新的无监督微博情感倾向性分析方法，利用 BTM 模型挖掘文档中的隐含主题，通过已有情感词典分析隐含主题的情感分布，并实现情感倾向性分析。Rosen 等[25]提出联合情感主题模型，其将模型作为情感层附加在文档层与主题层之间。结果表明：在该模型中，文字与情感和主题相关，主题与情感标签相链接，情感标签与文档相链接。琚春华等[26]将 STDP 模型中的情感层分解为两个层次。首先确定该词为感情词还是主题词；其次，如果其应该为感情词，对其极性进行研究，且 STDP 模型需要前期知识区分情感词和主题词。

2.2 文献评述

通过上述对现有研究成果的梳理发现，社交网络中的用户情感分析方法已经受到了国内外学者的重视并得到丰富的研究。上述研究成果，既对本研究具有一定的借鉴意义，但也存在一些不足：（1）已有很多学者对文本主题词方法开展了一系列卓有成效的研究，但是多数集中在对简单文本特征提取进行研究（如阮光册等[27]），较少有融入用户情感要素对此过程进行探究。（2）现有文献大多集中于对 LDA 模型进行扩展和优化（如 Mukherjee 等[28]），但很少有使用 Biterm 主体模型进行研究的文献记录。（3）几乎所有文献都只在实验中对参数进行简单的设置（如张雄等[29]），很少有基于塌缩吉布斯抽样等方法推断参数的文献记录。（4）现有关于社交网络用户情感分析的文献主要进行定量仿真分析（如 Jo 等[30]），而当前还未有学者使用真实社交网络用户评论数据集相结合的角度进行分析。

本研究基于以上研究综述，认为本文所提用户情感 Biterm 主题模型（US-BTM）应从特定文本出发，发现用户偏好及情感倾向，从而有效利用 Biterm 进行主题建模；应采用用户聚合策略形成伪文档，并为整个语料库创建词汇配对以解决数据稀疏性和短文本等问题；随后通过词汇共现模式对文本主题进行剖析，从而推断具有丰富语料库级信息的主题，并通过分析特定场景下的评论语料库中的词汇配对集及其相应主题的情感，

达到准确预测用户对特定场景的兴趣、偏好和情感的目的。以期为净化社交网络环境、分析用户情感提供参考；同时为有关部门控制用户情绪极性、维护社会稳定、保证我国网络安全提供借鉴。

3　模型构建

本文所提 US-BTM 方法对主要研究语料库层面的主题建模和情绪极性分类。模型通过基于用户的文本聚合方法生成伪文档，并直接对词对共现模式进行建模，从而解决了短文本与数据稀疏性等问题。本文所提半监督 US-BTM 法主要进行情感分类和主题建模。本研究所设计的 US-BTM 模型是一个概率生成模型（probabilistic generation model），通过在一定范围内的词汇共现模式进行建模，并以此对传统 LDA 模型的流程进行改进，以学习和理解用户短评文本中的情感与主题。根据本文所提 US-BTM 模型，整个语料库被转换为一个共生词汇语料库，从模型中对每个实例进行抽样，包括情感和主题模型，从而研究语料库中用户情感的主题分布。实验结果表明，与基准方法相比，本文所提 US-BTM 方法能够通过识别评论用户的极性准确发现文本主题和用户情感。

3.1　改进 Biterm 主题模型

对于许多现有社交网络应用开发者来说，在短文本中（微博、评论、转发等）发现主题是一项具有挑战性的任务，这是由于现有的经典主题挖掘模型并不能很好地处理短文本。为了克服这一难题，本文引入 Biterm 模型，以在整个语料库中创建词汇同现模式，已达到直接发现主题的目标，Biterm 为完全没有组织的词对集同时出现在短文本中的情况。故本文假设实验中的文本集由词汇组成，最终该模型总共生成 C_i^2 个词对。

Biterm 表示短文本中无序的词汇对。本文模型中的短文本是表示序列项上的一个短且确定的窗口。本文假设每个评论或微博都是一个单独的文本单元，在这种情况下，任何文档级别中任何两个不同的词汇都能构成 Biterm（例如一个包含 4 个不同词汇的文档将生成 6 个 Biterm）。具体如式（1）所示：

$$\begin{aligned}&\mathrm{BP}(w_1,w_2,w_3,\cdots,w_n)\Rightarrow\{(w_1,w_2),(w_1,w_3),\cdots,\\&(w_1,w_n),(w_2,w_3),\cdots,(w_2,w_n),\cdots,w_{n-1},w_n\}\end{aligned}\tag{1}$$

其中，$(-,-)$为无序的。对文本的 Biterm 提取完成后，只需对文本进行一次扫描，整个文本集就会变成一组 Biterm 集。

Biterm 主题模型（即 BTM）模拟了短文本中词汇共现模式的计算过程。但在为语料库级文本集生成词汇共现模式时，常常容易忽略用户独特性。因此本文认为 BTM 主题模型中应包含用户级的个性化信息，在完成基于用户的评论聚合后，可基于用户的 Biterm 聚合中学习用户特定的主题。本文将除了情感词汇以外的词汇区分为背景词和主题词，并形成一个简化的 Biterm 词集，以便更好地进行主题词学习。本文所提方法利用了用户评论的聚合，将背景主题合并到 BTM 主题模型中。

3.2　模型描述及构建

由于本文实验中的文本集是由一组用户、位置及文本所组成的，故本文假设对于任何文本而言，都有用户在位置处发布的文本或信息。另外，本文假设 S 与 T 为预定义的常数值，其中，S 为离散情绪标签的数量，T 为主题的总数量。由于社交网络中的评论、博文与转发等文本都较短，本文对这些评论进行单独理解意义不大；故本文引入线程池策略，以形成每个用户在特定地点的聚合伪文档。实验中的 A_v 是在位置 v 时进行书面评论的用户集；d_v 是位置文档，也就是在位置 v 时的所有文本的集合；N_d 为位置文档 d_v 中的词汇数量。本文所提模型中常用符号见表 1。

本文所提模型通过组合建模用户、情感、主题等要素，优化概括了传统 LDA 模型（如图 2 所示）；该模型描述了文本集中词对总集 B 的生成过程。根据本文所提方法，对每个文本中的用户情

表 1　社交网络用户情感 Biterm 主题模型参数设置

参数名称	参数符号	参数描述
位置集	V	位置 v 的集合 V
用户集	U	用户的集合
主题数量	Y	文本中所包含的主题总数
情感标签	S	用户情感标签
特定位置进行评论的用户集	A_v	在位置 v 中进行过评论的用户集合 A
特定位置的文本集	d_v	在位置 v 中的所有文本的集合 d
主题变量	z	主题中存在的变量 z
用户变量	u	用户变量 u
开关变量	c	开关变量 c
特定位置的文本集中的词汇数	N_d	特定位置 v 中的所有文本的集合 d 中的词汇数量
位置集的主题分布	θv	位置集 V 中的主体分布 θ
主题的词汇分布	$\varnothing_z$	主题 z 中的词汇分布 $\varnothing$
用户的情感分布	X^a	用户 u 的情感分布 X
主题的情感分布	θ^a	主题的情感分布 θ^a
位置中的情感分布	X^v	在位置 v 中的情感分布 X
狄利克雷先验概率	α,β,σ,η	狄利克雷先验概率参数
伯努利先验概率	γ	伯努利先验概率参数
词对集	w_i,w_j	词对集合
词对数量	M	词对数量

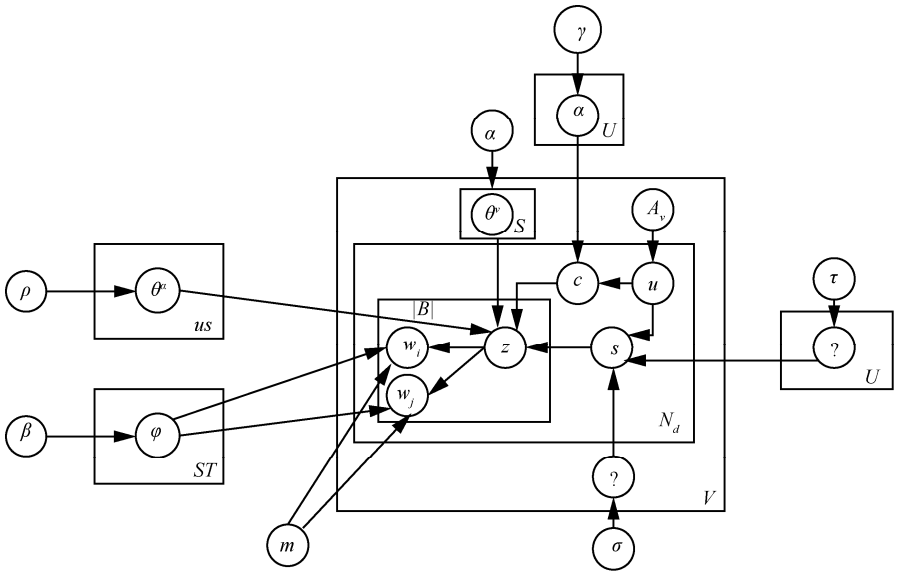

图 2　本文所提 US-BTM 模型的模型架构

感主题分布进行估计。

本文所描述模型的架构过程考虑用户偏好，并包含了一个已知的标签（即情感倾向）。为获得社交网络评论中词汇的情感极性 r，本文使用了知网 Hownet 情感词典来获取文本中词汇的极性值。故可推导出特定位置进行评论的用户集 A_v 中的地点文本集 d_v 中的词对集及情感极性值 (r,s)，其中，本文认为权威用户的情感会影响甚至决定预测的主题。而对于情感标签 S 下的每个主题 Z，$\varnothing_{s,z}$ 都表示主题情感狄利克雷分布，该分布由狄利克雷先验概率 β 得出。且每个用户的词对集 w_u 都服从 γ 的 Beta 分布。

模型实施的具体步骤如下所示。

步骤 1　对于情感标签 S 下的每个主题 Z 构建分布如式（2）所示：

$$\varnothing_{s,z} \sim \text{Dirichlet}\left(\beta_s, z\right) \tag{2}$$

步骤 2　对于每个用户 u，构建如式（3）、式（4）所示：

$$w_u \sim \text{Beta}\left(\gamma\right) \tag{3}$$

$$x_u^{(a)} \sim \text{Dirichlet}\left(\tau\right) \tag{4}$$

步骤 3　对于每个情感标签 S 构建主题分布如式（5）所示：

$$\theta_{u,s}^{(a)} \sim \text{Dirichlet}\left(\rho\right) \tag{5}$$

步骤 4　对于每个位置 v 中的文本 d 构建情感分布如式（6）所示：

$$x_{d_v}^{(v)} \sim \text{Dirichlet}\left(\alpha\right) \tag{6}$$

步骤 5　对于每个情感标签 S 构建主题分布如式（7）所示：

$$\theta_{d_v,s}^{(v)} \sim \text{Dirichlet}\left(\alpha\right) \tag{7}$$

步骤 6　对于每个位置 v 中的文本 d 中的每个词对 $b \in B$ 构建 $u \sim A_v$，并构建开关变量如式（8）所示：

$$c \sim \text{Bernoulli}\left(w_u\right) \tag{8}$$

步骤 7　如果 $c=0$，则分别构建情感分布与主题分布如式（9）、式（10）所示：

$$s_{d_v,i} \sim \text{Mult}\left(x_u^{(a)}\right) \tag{9}$$

$$z_{d_v,i} \sim \text{Mult}\left(\theta_{u,}^{(a)} s_{d_v,i}\right) \tag{10}$$

步骤 8　如果 $c=1$，则分别构建情感分布、主题分布与词对分布如式（11）、式（12）和式（13）所示：

$$s_{d_v,i} \sim \text{Mult}\left(x_{d_v}^{(v)}\right) \tag{11}$$

$$z_{d_v,i} \sim \text{Mult}\left(\theta_{d_{vi},}^{(v)} s_{d_v,i}\right) \tag{12}$$

$$w_t, w_j \sim \text{Mult}\left(\varnothing_{s_{d_{vi}}, z_{d_{vi}}}\right) \tag{13}$$

3.3　模型推导

在本文所提 US-BTM 模型中 $\alpha,\beta,\sigma,\eta,\gamma$ 为超参数，而社交网络用户发布文本中的潜在主题和情感取决于地点位置与用户偏好。变量 w 为伯努利分布的参数，其会影响地点位置与用户间的关系；而开关变量 c 是为了确定位置文本是由位置还是由用户决定。本文在参数推断方面，采用 Collapsed Gibbs Sampling 法找出未知隐藏变量 $\left\{\varnothing, w, \theta^a, \theta^v, X^a, X^v, \psi\right\}$。每个词的潜在变量的后验分布计算如式（14）所示。

$$P\left(U_{dvi}=U, C_{dvi}=0, S_{dvi}=S, Z_{dvi}=\frac{Z}{\theta}\right) \tag{14}$$

式（14）中满足 $\theta < U^{-dvi}, C^{-vi}, S^{-dvi}, Z^{-dvi}, w, t, A_v >$。其中，$-dvi$ 表示不包括当前单词，参数 α, x 的计算过程如式（15）和式（16）所示：

$$\alpha = \frac{n_{uc}^{-dvi}(0)+\gamma}{n_{uc}^{-dvi}+\gamma+\gamma'} \times \frac{n_{us}^{-dvi}+\tau_s}{\sum_S \left(n_{us} S^{-dvi}+\tau_s'\right)} \tag{15}$$

$$x = \frac{n_{uc}^{-dvi}+\rho z}{\sum_{z'}\left(n_{usz}^{-dvi}+\rho z'\right)} \times \frac{n_{szw}^{-dvi}+\beta\omega}{\sum_{\omega'}\left(n_{szw^1}^{-dvi}+\beta\omega'\right)} \tag{16}$$

当满足如式（17）条件时，可得 α 和 x 如式（18）和式（19）所示：

$$P\left(U_{dvi}=U,C_{dvi}=1,S_{dvi}=S,Z_{dvi}=\frac{Z}{\theta}\right) \quad (17)$$

$$\alpha=\frac{n_{uc}^{-dvi}(1)+\gamma}{n_{uc}^{-dvi}+\gamma+\gamma'}\times\frac{n_{vs}^{-dvi}+\sigma_s}{\sum_{s'}\left(n_{vs}S^{-dvi}+\sigma_s'\right)\sum_{s'}} \quad (18)$$

$$x=\frac{n_{vsz}^{-dvi}+\alpha z}{\sum_{z'}\left(n_{usz}^{-dvi}+\alpha z'\right)}\times\frac{n_{szw}^{-dvi}+\beta\omega}{\sum_{\omega'}\left(n_{szw^1}^{-dvi}+\beta\omega'\right)} \quad (19)$$

其中，$n_{uc}^{-dvi}(0)$ 和 $n_{uc}^{-dvi}(1)$ 是当 $c=0$ 和 $c=1$ 时的用户采样的计数结果。而 n_{us} 则表示用户 X^{α} 在情感分布中采样情感 S 的次数，n_{vs} 是从地点位置 X^{v} 的情感分布中采样的情感 S 的次数。n_{usz} 是从用户 u 和用户情感 s 的分布中采样主题 z 的次数。n_{vsz} 是从位置 v 分布和情感标签 s 中采样的主题 z 的次数，n_{szw} 是从情绪 s 和主题 z 分布中采样单词 w 的次数。在使用塌缩吉布斯抽样后，隐藏变量 $\{\varnothing,w,\theta^{a},\theta^{v},X^{a},\ X^{v},\psi\}$ 的计算如式(20)、式(21)所示：

$$\hat{\theta}_{vsz}=\frac{n_{vsz}+\alpha z}{\sum_{z'}\left(n'_{vsz}+\alpha z'\right)} \quad (20)$$

$$\hat{\theta}_{usz}=\frac{n_{usz}+\rho z}{\sum_{z'}\left(n'_{usz}+\rho z'\right)} \quad (21)$$

在主题 z 中的情感标签 s 的词汇分布由式（22）和式（23）计算所得：

$$\hat{\varnothing}_{\alpha szw}=\frac{n_{szw}+\beta\omega'}{\sum_{\omega'}\left(n_{sz\omega'}+\beta\omega'\right)} \quad (22)$$

$$\hat{W}^{au}=\frac{n_{uc}(1)+\gamma}{n_{uc}+\gamma+\gamma^{1}} \quad (23)$$

在对本文所提 US-BTM 模型进行必要训练后，本文的目标是估计 $P(z/u,v)$，即词对 (w_i,w_j) 中的所有主题 z 被包含的概率。结合寇晓淮等[31]的做法，本文认为有必要对文本中的情感与主题进行近似估计，如式（24）所示：

$$S=\arg\max X^{a}X^{v} \quad (24)$$

在此基础上，可求解词汇 w_i 在文本 d 中出现主题 z 的概率如式（25）和式（26）所示：

$$P(z/w_i)=\frac{\sum_{B}P\left(\frac{z}{b}\right)XP\left(\frac{w_i}{b}\right)}{\sum_{B}p\left(\frac{w_i}{b}\right)} \quad (25)$$

$$P(z/b)=\frac{\sum_{l}P(s)P\left(\frac{z}{l}\right)P\left(\frac{w_i}{lz}\right)P\left(\frac{w_i}{lz}\right)}{\sum_{z}\sum_{s}P_d(S)P\left(\frac{z}{s}\right)P\left(\frac{w_i}{Sz}\right)P(\frac{w_i}{lz})} \quad (26)$$

将式（25）与式（26）化简计算可得如式（27）所示：

$$P(z/b)=\frac{\sum_{z}P_d(s)P\left(\frac{z}{s}\right)P\left(\frac{w_i}{sz}\right)P\left(\frac{w_i}{sz}\right)}{\sum_{z}\sum_{s}P_d(S)P\left(\frac{z}{s}\right)P\left(\frac{w_i}{sz}\right)P(\frac{w_i}{sz})} \quad (27)$$

同理可计算得到带有情感标签 S 的词汇 w_i 出现的概率如式（28）所示：

$$P(s/w_i)=\frac{\sum_{MM}P_d(s)P(s)P\left(\frac{w_i}{m}\right)}{\sum_{M}\left(P\left(\frac{w_i}{m}\right)\right)} \quad (28)$$

其中，M 表示文本集中的词对总数，而 $P\left(\frac{w_i}{m}\right)$ 为词汇 w_i 出现在文本 m 中的概率。

4 实验结果与分析

4.1 数据来源与预处理

为验证本文所提 US-BTM 模型的有效性，利用新浪微博数据集进行实验。利用 Python 工具从新浪微博 API 爬取真实用户数据集作为实验仿真的基础数据（内容见表 2），本文以微博热词“网红”为关键词，选取 20 个粉丝数超过 10 万的用户作为核心节点，并爬取其与其存在关注、转发和提及关系的用户发布的特征与文本信息，爬取时间为 2019 年 1 月 12 日—8 月 21 日，数据集共包含 38 916 个节点，289 326 条文本信息；并以 csv 格式保存在 MySQL 数据库中以便进行数据处理。本文结合张树森等[32]的方法，使用的改进神

表 2　新浪微博数据集特征

数据集名称	用户数量	用户平均评论数量	积极评论数量	消极评论数量	评论总数	平均用户评论长度	平均词汇（评论）
新浪微博	18 083 个	16.092 条	14 780 条	46 118 条	230 829 条	319.892 字符	14.893 个

经网络算法和随机森林算法对本文数据集进行僵尸和不活跃用户的剔除（保留存在 10 次评论以上的用户信息）；并使用刘啸剑等[33]的方法，利用 Python 中的 Jieba 分词工具及中国科学院开发的 NLPIR 汉语分词系统对去重后的文本进行分词，对数据进行清洗、分词、去除停用词等预处理后共得到 18 083 个用户的 230 829 条有效评论。并对用户进行情感打分，具体打分方法如第 4.2 节所述。

具体来说，为了减少低质量或不完整的评论对实验结果产生影响，本文通过以下步骤对评论进行预处理：（1）删除非中文字符、连接词；（2）删除重复评论；（3）过滤/剔除少于 5 个词的评论；（4）模型使用每个用户近 85%的数据以了解参数。

4.2　实验过程

为保证对本文所提 US-BTM 模型进行科学验证，本文结合 CHAN 等[34]的做法将主题数量 K 设置为 15，并将超参数 α 和 β 的初始值分别设置为 $50/K$ 和 0.01。用户数量、评论情感和平均评论字数等参数根据上文中提供的数据集进行实验。而模型 Perplexity 值被用于初始化 K 值。此外词汇分类是这项工作的基础，故本研究利用 Python 中的 TextbBlob 库（一个开源文本处理库，可用于执行自然语言处理，如词性标注、名词性成分提取、情感分析、文本翻译等）获取用户评论中不同词汇的极性值。另外，本文利用知网 Hownet 情感词典推导出词汇的积极、消极极性。由于模型 Perplexity 值表示模型的似然性，故模型参数的取值是基于模型的最小 Perplexity 值来选择的。

主题建模为一种无监督技术，其主要任务是从词汇的分布中提取主题，且所提取的主题可以任意以定性和定量的方式进行评估。定性的方法是观察主题模型提取出来的词；而定量方法是通过 Perplexity 值、主题一致性、内外部测量值、纯度等多种主题测量来评价。本文所提社交网络用户情感 Biterm 主题模型需要发现特定的情感话题，并根据观察提取出的基于词的情感主题，然后通过定性评价方法对发现的主题进行测量或检验，本文对每个实验均迭代运行 100 次仿真模拟并取平均数，以避免实验随机性带来的误差。

Perplexity 值（困惑度）常用于衡量给定模型与测试数据的拟合程度（即概率模型对样本的预测正确度），从而评估主题模型的预测能力[24]。模型 Perplexity 值越低，则表明预测能力越强；但当其值非常低时可能会导致给定本文的主题结构没有意义。因此需要使用数据集的知识有效评估 LDA 模型的主题数量。本文利用给定 K 值来估计 LDA 模型，然后给出主题的理论词汇分布，并与文档中的实际词汇分布进行比较。Perplexity 值的计算方法如式（29）所示：

$$\text{Perplexity}(W)=\sqrt[N]{\frac{1}{P\left(w_1,w_2,\cdots,w_N\right)}} \tag{29}$$

主题建模有利于系统化的总结大量短文本。主题模型是一种无监督的主题研究方法，即自动排列来自未标记文本的一组词汇。由于本文不能确保用于提取主题的 LDA 技术是否可行，故需要对其进行主题一致性（topic coherence）测量，从而区分积极主题和消极主题。进行主题一致性测度必须基于一致性框架，即一系列可以求和的组成部分，这种测量大致可分为内在度量和外部度量两类。内在度量是在一个有序词集中，分别利用先行词和后继词对一个词进行比较分析。为了避免计算零对数问题，本文在平滑排序中将对分

数函数设置为条件对数概率计算。而外部度量是对于一个无序的词集将不同的词汇进行配对。外部度量使用点间互信息，而主题一致性评分为主题词$w_1,w_2,\cdots,w_n$对分的总和。而LDA模型的构建应采用不同的主题数量（即K值），并使K值具有最高的一致性评分。具体计算式如式（30）所示。

$$C\left(z,S^z\right)=\sum_{n=2}^{N}\sum_{l=1}^{n-1}\log\frac{D_2\left(W_n^z,W_l^z\right)+1}{D_1\left(W_l^z\right)} \quad (30)$$

其中，前N个z中的单词集合可表示如式（31）所示，$D_1\left(W_l^z\right)$是单词的文档W频率，而$D_2\left(W_n^z,W_l^z\right)$表示单词$W_n^z,W_l^z$的共现文档频率。更高的一致性分数表示更好的可解释性，且语义上更连贯。本文使用模型准确率（Accuracy值）对本文所提模型与其他基准模型进行比较，具体如式（31）所示。

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (31)$$

模型而情感识别通常由不同模型的情感准确性（sentiment accuracy）值来衡量，结合BOYD-GRABER等[35]所提方法可知其通常根据用户标签进行计算，也可以通过对情感标签进行配对t检验来衡量，通常p值小于0.05为显著。

4.3 实验结果

本文使用3种不同方法（基于所发现的主题；基于特定情感的主题发现；基于情感识别）对本文所提US-BTM模型的性能进行了分析。并对发现的主题进行了定性和定量评价。由US-BTM模型结合情感极性与共词现象发现的部分情感主题词列表见表3。

为了使结果更便于理解，本研究为情感词的主题进行了标签化，使得词汇在主题内更加连贯。在评论中通常会出现一些干扰词，这些干扰词应该在数据的预处理步骤中进行适当去除，同时还需要删除一些更常见的细粒度（fine-grained）分析主题。由表3所示结果可知US-BTM模型能够清楚地捕捉到积极和消极的主题。

接下来本研究将每个用户评论文本按照积极、消极和中性的标签进行分类，并测量该模型在积极和消极标签下的整体性能。Perplexity值是一个计算给定模型值的函数，具体来说是为模型计算不同的K值。LDA模型[36]、JST模型[30]、ASUM模型[30]、BTM-LDA模型[37]、W-LDA模型[38]、WBW-BTM模型[39]和本文所提US-BTM模型中的Perplexity值与主题数量关系如图3所示，由Perplexity值越低模型预测能力越强的特征可见，主题数量越多模型预测能力越弱。相对于其他模

表3 由US-BTM模型发现的积极、消极词列表（部分）

种类	积极词				消极词			
主题	感情	轻松	社会	服务	感情	经济	行为	否定
	多样	关系	工作	民生	无聊	问题	无所谓	不支持
	就好	很好	强烈	放假	代价	野心	关闭	恶心
	关心	调休	中国	服务	涨价	穷	坏	无聊
	友好	经历	中心	专业	无趣	少	等待	不坚持
情感词	享受	放松	舒适	接受	粗心	卑鄙	渣男	诽谤
	愉悦	吃饭	非常好	爱情	嫉妒	冒充	懦夫	挑拨
	主要的	魄力	磅礴	理想	屠杀	虚假	骗子	偏心
	楷模	精致	培养	谦虚	浪费	奸诈	后遗症	狠心
	神圣	胜利	甜美	淳朴	勾引	哄骗	糊涂	恐吓

型而言，US-BTM 模型的 Perplexity 值相对较小，故可知其预测能力相对最强。

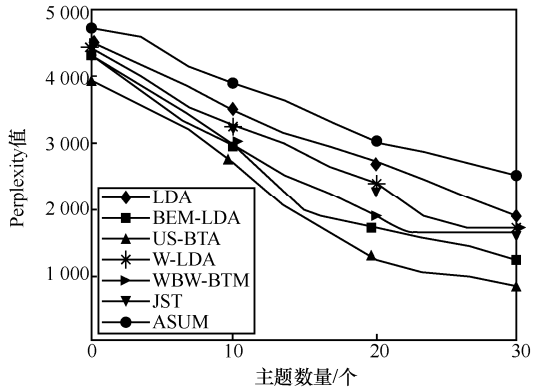

图 3　不同模型下的主题数量与 Perplexity 值关系

积极主题数量与主题一致性评分间的关系如图 4 所示，图 4 中的 C、O 和 P 分别代表在实验用户集中随机挑选的 3 个用户的评论的示意图。可见，随着主题数量 K 的增加，一致性得分达到最高值，随后反转下降。

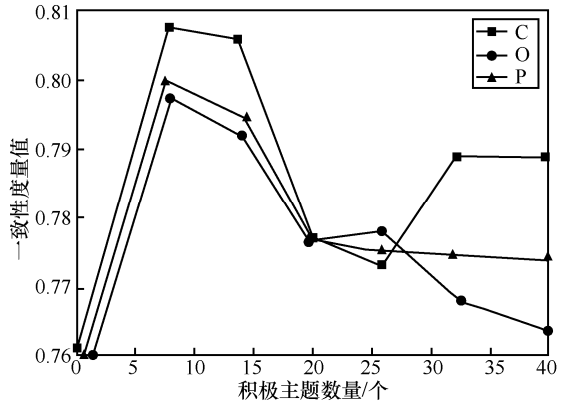

图 4　主题一致性与积极主题数量的关系

LDA 模型、BTM-LDA 模型、JST 模型、ASUM 模型、W-LDA 模型、WBW-BTM 模型和 US-BTM 模型下的主题一致性测度值（coherence measure）[40] 如图 5 所示。本文所提 US-BTM 模型的一致性比例最高，而 BTM-LDA 模型略低于 US-BTM 模型，其他模型一致性表现较差。且与 BTM-LDA 模型相比，本文所提 US-BTM 模型能确保词对中的两个词属于同一种主题和情感，这种约束条件更适用于利用评论对场景或产品做出评价。

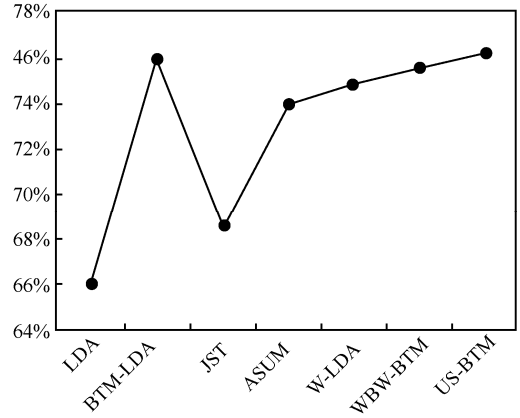

图 5　不同模型中的一致性比例测度结果

本文所提 US-BTM 模型以及 LDA 模型、BTM-LDA 模型、JST 模型、ASUM 模型、W-LDA 模型和 WBW-BTM 模型的 Accuracy 值（准确度）与主题数量关系如图 6 所示。实验结果显示，与其他模型相比，US-BTM 的准确性最高。总体来说当主题数量增加时，模型 Accuracy 值往往呈先升后降态势，只有本文所提 US-BTM 模型能够一直保持高水平的 Accuracy 值。

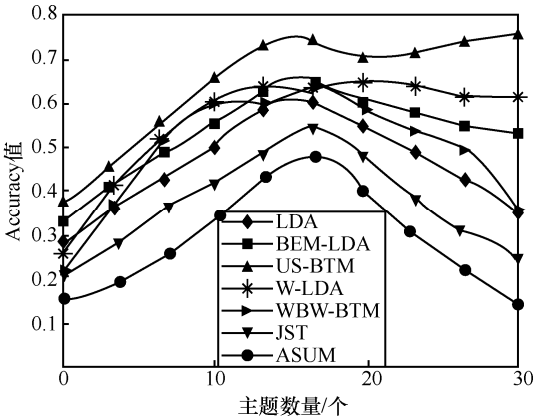

图 6　不同模型准确度与主题数量关系

而图 6 和表 4 分别报告了当主题数量 K=15 时，模型中主题数量对准确度的影响和不同模型的情感准确度。由于此计算需要使用知网 Hownet 情感词典和 TextBlob 文本处理库，故在对用户情绪标签进行配对和 t 检验时，规定 p 值要小于 0.05。由表 4 结果可知，相对其他模型，US-BTM 模型的情感准确度最高，其次为 ASUM 模型。结合参

表 4　不同模型的情感准确性计算结果

模型名称	LDA	BTM-LDA	JST	ASUM	W-LDA	WBW-BTM	US-BTM
情感准确性	59.43%	64.03%	61.94%	63.68%	67.21%	72.08%	86.48%

表 5　不同迭代次数下的情感准确性模型比较结果

模型名称	LDA	BTM-LDA	JST	ASUM	W-LDA	WBW-BTM	US-BTM
500 次迭代的情感准确性	46.79%	53.49%	54.78%	58.28%	62.38%	67.39%	81.20%
1 000 次迭代的情感准确性	50.36%	55.83%	58.42%	61.29%	65.42%	69.01%	84.38%

考文献[41]可知，JST 模型中的每个词汇都存在不同的主题，故准确性较差；而 ASUM 模型中认为一句话中的词汇都属于同一主题，故相对准确度较高。但由于社交网络用户的文本往往存在稀疏性问题，而 JST 和 ASUM 模型不适合短文本（没有足够的句子样本来获得参数）。本文所提 US-BTM 模型解决了上述难题，且允许句子中的词汇存在不同主题，故从这一角度也可以看出本文所提 US-BTM 模型优于其他模型。

为保证实验结果不受迭代次数的影响，本研究针对迭代次数（500 次和 1 000 次）进行了敏感性分析，并给出了相关结果。实验表明当增加迭代次数时，本文所提 US-BTM 模型的困惑率、准确率、主题一致性和情感准确度等指标值的排名基本不发生变化，仍优于其他主题词模型，故可知本文所提模型并不会受迭代次数的影响。本研究只罗列了不同迭代次数下的模型情感准确性计算结果见表 5，其他结果限于篇幅不在此进行展示。

5　结束语

本文提出的社交网络用户情感 Biterm 主题模型（US-BTM 模型）是一种将用户与他们的情感取向性结合起来的新方法，该方法融入聚合策略和情感极性技术，利用 Biterm 和词对共现方法同时生成积极主题词、消极主题词、用户积极话题、用户消极话题和基于位置的话题分布。本文通过使用真实社交网络用户文本数据集进行实验，结果表明本文所提模型相对其他基准模型而言具有更高的效率和准确度，却所提取主题更加连贯，更具有代表性。本文认为本文所提方法采用精确的情感极性技术，能准确地捕捉用户的情感倾向，正确地揭示用户偏好，可广泛应用于社交网络的内容描述、推荐及社交网络用户兴趣描述、语义分析等多个领域。例如本文所提模型可应用于新闻推荐系统，可利用新闻社交网络用户情感 Biterm 主题模型结合用户情感能够更准确地捕捉用户情感倾向和情感偏好，达到提升社交网络中用户新闻推荐精度的目的。

参考文献：

[1] 熊蜀峰, 姬东鸿. 面向产品评论分析的短文本情感主题模型[J]. 自动化学报, 2016, 42(8): 1227-1237.
XIONG S F, JI D H. A short text sentiment-topic model for product review analysis[J]. Acta Automatica Sinica, 2016, 42(08): 1227-1237.

[2] ZUO Y, WU J, ZHANG H, et al. Topic modeling of short texts: a pseudo-document view[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 2105-2114.

[3] DEVI G U, PRIVAN M K, GOKULNATH C. Wireless camera network with enhanced SIFT algorithm for human tracking mechanism[J]. International Journal of Internet Technology and Secured Transactions, 2018, 8(2): 185-194.

[4] 刘洺辛, 陈晶, 王麒媛. 基于改进特征选择方法的文本情感分类研究[J]. 电信科学, 2018, 34(10): 85-95.
LIU M X, CHEN J, WANG Q Y. Research on text sentiment classification based on improved feature selection method[J]. Telecommunications Science, 2018, 34(10): 85-95.

[5] CHENG X, YAN X, LAN Y, et al. Btm: topic modeling over short texts[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2928-2941.

[6] VIJAYAKUMAR V, PRIVAN M K, USHADEVI G, et al. E-health cloud security using timing enabled proxy re-encryption[J]. Mobile Networks and Applications, 2019, 24(3): 1034-1045.

[7] 蔡永明, 长青. 共词网络 LDA 模型的中文短文本主题分析[J]. 情报学报, 2018, 37(3): 305-317.
CAI Y M, CHANG Q. Chinese short text topic analysis by latent dirichlet allocation model with co-word network analysis[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 305-317.

[8] LI W, FENG Y, LI D, et al. Micro-blog topic detection method based on BTM topic model and *K*-means clustering algorithm[J]. Automatic Control and Computer Sciences, 2016, 50(4): 271-277.

[9] 周孟, 朱福喜. 基于情感标签的极性分类[J]. 电子学报, 2017, 45(4): 1018-1024.
ZHOU M, ZHU F X. Polarity classification based on sentiment tags[J]. Acta Electronica Sinica, 2017, 45(4): 1018-1024.

[10] 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 J H, LIU Y H, LI X M, et al. Multi-grain sentiment topic , model based on LDA[J]. Acta Electronica Sinica, 2015, 43(9): 1875-1880.

[11] LI C, ZHANG J, SUN J T, et al. Sentiment topic model with decomposed prior[C]//Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. [S.n.: s.l.], 2013: 767-775.

[12] MEI Q, LING X, WONDRA M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs[C]//Proceedings of the 16th international conference on World Wide Web. New York: ACM Press, 2007: 171-180.

[13] 孙锐, 郭晟, 姬东鸿. 融入事件知识的主题表示方法[J]. 计算机学报, 2017, 40(4): 791-804.
SUN R, GUO S, JI D H. Topic representation integrated with event knowledge[J]. Chinese Journal of Computers, 2017, 40(4): 791-804.

[14] PRIVAN M K, DEVI G U. A survey on internet of vehicles: applications, technologies, challenges and opportunities[J]. International Journal of Advanced Intelligence Paradigms, 2019, 12(1-2): 98-119.

[15] BALAN E V, PRIVAN M K, NATH C G, et al. Efficient energy scheme for wireless sensor network application[C]//Proceedings of 2014 IEEE International Conference on Computational Intelligence and Computing Research. Piscataway: IEEE Press, 2014: 1-5.

[16] BICALHO P, PITA M, PEDROSA G, et al. A general framework to expand short text for topic modeling[J]. Information Sciences, 2017(393): 66-81.

[17] NGUYEN T S, LAUW H W, TSAPARAS P. Review synthesis for micro-review summarization[C]//Proceedings of the eighth ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2015: 169-178.

[18] LIN C, HE Y. Joint sentiment/topic model for sentiment analysis[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2009: 375-384.

[19] 朱宪莹, 刘箴, 金炜, 等. 基于特征融合的层次结构微博情感分类[J]. 电信科学, 2016, 32(7): 106-114.
ZHU X Y, LIU Z, JIN W, et al. Hierarchical micro-blog sentiment classification based on feature fusion[J]. Telecommunications Science, 2016, 32(7): 106-114

[20] KUMAR P M, DEVI U, MANOGARAN G, et al. Ant colony optimization algorithm with Internet of vehicles for intelligent traffic control system[J]. Computer Networks, 2018(144): 154-162.

[21] LIU S M, CHEN J H. A multi-label classification-based approach for sentiment classification [J]. Expert Systems with Applications, 2015, 42(3): 1083-1093.

[22] MANOGARAN G, SHAKEEL P M, HASSANEIN A S, et al. Machine learning approach-based gamma distribution for brain tumor detection and data sample imbalance analysis[J]. IEEE Access, 2018, 7(1): 12-19.

[23] 王建成, 徐扬, 刘启元, 等. 基于神经主题模型的对话情感分析[J]. 中文信息学报, 2020, 34(1): 106-112.
WANG J C, XU Y, LIU Q Y, et al. Dialog sentiment analysis with neural topic model[J]. Journal of Chinese Information Processing, 2020, 34(1): 106-112.

[24] 张佳明, 王波, 唐浩浩, 等. 基于 Biterm 主题模型的无监督微博情感倾向性分析[J]. 计算机工程, 2015, 41(7): 219-223, 229.
ZHANG J M, WANG B, TANG H H, et al. Unsupervised sentiment orientation analysis on micro-blog based on biterm topic model[J]. Computer Engineering, 2015, 41(7): 219-223, 229.

[25] ROSEN-ZVI M, GRIFFITHS T, STEVVERS M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Barcelona: AUAI Press, 2004: 487-494.

[26] 琚春华, 鲍福光, 戴俊彦. 一种融入公众情感投入分析的微博话题发现与细分方法[J]. 电信科学, 2016, 32(7): 97-105.
JU C H, BAO F G, DAI J Y. Discovery and segmentation method in micro-blog topics based on public emotional engagement analysis[J]. Telecommunications Science, 2016, 32(7): 97-105.

[27] 阮光册, 夏磊. 基于词共现关系的检索结果知识关联研究[J]. 情报学报, 2017, 36(12): 1247-1254.
RUAN G C, XIA L. Knowledge connection of retrieval results based on co-word analysis[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(12): 1247-1254.

[28] MUKHERJEE S, BASU G, JOSHI S. Joint author sentiment topic model[C]//Proceedings of the 2014 SIAM International

Conference on Data Mining. [S.n.:s.l.], 2014: 370-378.

[29] 张雄, 陈福才, 黄瑞阳. 基于双词主题模型的半监督实体消歧方法研究[J]. 电子学报, 2018, 46(3): 607-613.
ZHANG X, CHEN F C, HUANG R Y. Semi-supervised entity disambiguation method research based on biterm topic model[J]. Acta Electronica Sinica, 2018, 46(3): 607-613.

[30] 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. New York: ACM Press, 2011: 815-824.

[31] 寇晓淮, 程华. 基于主题模型的垃圾邮件过滤系统的设计与实现[J]. 电信科学, 2017, 33(11): 73-82.
KOU X H, CHENG H. Design and implementation of spam filtering system based on topic model[J]. Telecommunications Science, 2017, 33(11): 73-82.

[32] 张树森, 梁循, 弭宝瞳, 等. 基于内容的社交网络用户身份识别方法[J]. 计算机学报, 2019, 42(8): 1739-1754.
ZHANG S S, LIANG X, MI B T, et al. Content-based social network user identification methods[J]. Chinese Journal of Computers, 2019, 42(8): 1739-1754.

[33] 刘啸剑, 谢飞, 吴信东. 基于图和 LDA 主题模型的关键词抽取算法[J]. 情报学报, 2016, 35(6): 664-672.
LIU X J, XIE F, WU X D. Graph based keyphrase extraction using LDA topic model[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(6): 664-672.

[34] CHAN P P K, YANG C, YEUNG D S, et al. Spam filtering for short messages in adversarial environment[J]. Neurocomputing, 2015, 155(C): 167-176.

[35] BOYD-GRABER J, BLEI D. Syntactic topic models[J]. Advances in Neural Information Processing Systems, 2010(2): 185-192.

[36] BLEID M, ANDREWY N G, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993–1022,

[37] XIA Y, TAN G, HUSSIAN A, et al. Discriminative biterm topic model for headline-based social news clustering[C]// Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference. Piscataway: IEEE Press, 2008, 3(1): 32-45.

[38] 李思宇, 谢珺, 邹雪君, 等. 基于双词语义扩展的 Biterm 主题模型[J]. 计算机工程, 2019, 45(1): 210-216.
LI S Y, XIE J, ZOU X J, et al. Biterm topic model based on semantic extension of double words[J]. Computer Engineering, 2019, 45(1): 210-216.

[39] 黄畅, 郭文忠, 郭昆. 面向微博热点话题发现的改进 BBTM 模型研究[J]. 计算机科学与探索, 2019, 13(7): 1102-1113.
HUANG C, GUO W Z, GUO K. Research on improved BBTM model for microblog hot topic discovery[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(7): 1102-1113.

[40] 郑承利, 姚银红. 基于高阶一致风险测度的组合优化[J]. 系统管理学报, 2017, 26(5): 857-868.
ZHENG C L, YAO Y H. Portfolio optimization based on higher moment coherent risk measure[J]. Journal of Systems & Management, 2017, 26(5): 857-868.

[41] 崔雪莲, 那日萨, 刘晓君. 基于主题相似性的在线评论情感分析[J]. 系统管理学报, 2018, 27(5): 821-827.
CUI X L, NA R S, LIU X J. Sentiment analysis of online reviews based on topic similarity[J]. Journal of Systems & Management, 2018, 27(5): 821-827.

[作者简介]

顾秋阳（1995− ），男，浙江工业大学博士生，主要研究方向为智能信息处理、数据挖掘、电子商务与物流优化等。

吴宝（1979− ），男，博士，浙江工业大学教授、博士生导师，主要研究方向为社会网络、企业社会责任与高质量发展等。

琚春华（1962− ），男，博士，浙江工商大学教授、博士生导师，主要研究方向为智能信息处理、数据挖掘、电子商务与物流优化等。