情报杂志
Journal of Intelligence
ISSN 1002-1965,CN 61-1167/G3

《情报杂志》网络首发论文

题目：基于 BBTM 改进算法的微博热点话题发现研究
作者：向卓元，吴玉，陈浩，张芙玮
网络首发日期：2021-11-12
引用格式：向卓元，吴玉，陈浩，张芙玮．基于 BBTM 改进算法的微博热点话题发现研究[J/OL]．情报杂志.
https://kns.cnki.net/kcms/detail/61.1167.G3.20211111.1400.002.html

网络首发时间：2021-11-12 16:41:09
网络首发地址：https://kns.cnki.net/kcms/detail/61.1167.G3.20211111.1400.002.html

基于 BBTM 改进算法的微博热点话题发现研究*

向卓元　吴　玉　陈　浩　张芙玮

（中南财经政法大学信息与安全工程学院　武汉　430073）

摘　要：[**研究目的**]针对主流话题发现模型存在数据稀疏、维度高等问题，提出了一种基于突发词对主题模型（bursty biterm topic model，BBTM）改进的微博热点话题发现方法（BiLSTM-HBBTM），以期在微博热点话题挖掘中获得更好的效果。[**研究方法**]首先，通过引入微博传播值、词项 H 指数和词对突发概率，从文档层面和词语层面进行特征选择，解决数据稀疏和高维度的问题。其次，通过双向长短期记忆（Bi-directional Long-Short Term Memory，BiLSTM）训练词语之间的关系，结合词语的逆文档频率作为词对的先验知识，考虑了词之间的关系，解决忽略词之间关系的问题。再次，利用基于密度的方法自适应选择 BBTM 的最优话题数目，解决了传统的主题模型需要人工指定话题数目的问题。最后，利用真实微博数据集在热点话题发现准确度、话题质量、一致性三个方面进行验证。[**研究结论**]实验表明 BiLSTM-HBBTM 在多种评价指标上都优于对比模型，实验结果验证了所提模型的有效性及可行性。

关键词：热点话题发现；主题模型；微博；短文本；BiLSTM；BBTM；Word2Vec

中图分类号：TP391.1

Research on Microblog Hot Topic Discovery Based on the Improved BBTM Algorithm*

Xiang Zhuo-yuan　Wu Yu　Chen Hao　Zhang Fuwei

（School of Information and Safty Engineering，Zhongnan University of Economics and Law，Wuhan　430073）

Abstract：[**Research purpose**]Aiming at the problems of sparse data and high dimension in mainstream topic discovery model，this paper proposes an improved microblog hot topic discovery method（BiLSTM-HBBTM）based on the bursty biterm topic model（BBTM），in order to get better performances in microblog hot topic mining.[**Research method**]First，microblog propagation value，H index of term and bursty probability of biterm are used to select characteristics. The characteristicsselection is carried out from the document level and the word level to solve the problem of data sparsity and high dimension. Second，Through the Bi-directional long-short term memory（BiLSTM）training，the relationship between words，combined with the inverse document frequency of words as the prior knowledge of biterms，the relationship between words is considered and solve the problem of ignoring the relationship between words. Third，a density based method is used to select optimal number of topics for the BBTM model，which solves the problem that the traditional topic model needs to manually specify the number of topics. Finally，the actual datasets are used to verify the accuracy of hot topic discovery，topic quality and consistency.[**Research conclusion**]The experiment shows that BiLSTM-HBBTM is better than the contrast model in a variety of evaluation indicators，and the experimental results have verified the effectiveness and feasibility of the model.

Key words：hot topic discovery；topic model；microblog；short texts；BiLSTM；BBTM；Word2Vec

---

基金项目：**国家自然科学基金面上项目“面向跨语言观点摘要的领域知识表示与融合模型研究”（编号：71974202）研究成果之一。**

作者简介：**向卓元**，男，1964 年生，硕士，教授，硕士生导师，网络信息智能处理和舆情分析；**吴　玉**，女，1998 年生，硕士，信息安全管理；**陈浩**，男，1995 年生，硕士，自然语言处理。

0 引 言

随着移动互联网的快速发展,微博、贴吧等社交媒体得以蓬勃发展,用户生成的社交媒体信息量呈现爆炸式的增长,同时互联网将产生海量的短文本信息。新浪微博是国内大型网络媒体之一,人们可以不受时间、空间限制,实现实时分享与传播互动。当某一个话题爆发时,微博对事件能够做出快速的反应,用户可以通过 PC 端、移动端等方式获取有关话题信息,或者参与信息交互(转发、评论、点赞等操作),在短时间内形成舆论焦点,从而使该话题形成一个热点话题。因此,微博具有短文本性、实时性和交互性的特点。如何从海量的短文本数据中高效、准确地挖掘热点话题是目前舆情分析中的一个研究热点问题。

1 相关工作概述

梳理已有文献可以发现传统主题模型[1-3]是为挖掘长文本主题设计的,当应用这些模型来处理短文本时,会面临数据稀疏、语义信息匮乏、向量维度高等问题,从而无法从短文本中有效的挖掘文本主题信息,失去了在长文本话题发现中所发挥的优势。

近些年来,为了解决短文本的数据稀疏问题,有一部分学者通过语料数据文档的词对共现信息来学习主题。Yan 等人在 2013 年提出词对主题模型(Biterm Topic Model,BTM)[4],通过构建词对解决短文本的稀疏性问题,实验表明该模型挖掘的话题质量不受文本长度的限制,在短文本上同样取得较好的效果;但 BTM 模型挖掘的主题可能属于普通话题,也可能属于热点话题,因此无法直接用于热点话题发现。王亚民等[5]利用 BTM 模型进行微博舆情热点发现,与改进 TF-IDF 算法进行特征提取及相似性度量,解决了传统短文本主题模型的高维度和稀疏性问题。李卫疆等[6]结合 BTM 话题模型和 K-means 聚类算法来检测微博话题,缓解了短文本数据稀疏的问题。这些主题模型及其改进方法虽然能解决短文本的稀疏问题,但是无法直接用于发现热点话题,需要一些启发式后处理等工作。Hoffman 等人提出了在线主题模型(Online for Latent Dirichlet Allocation,OnlineLDA)[7],但仍然存在需要手工标注话题数目等后处理问题。M. Gerlach 等人[8]提出的 hSBM 模型通过调整具有非参数先验的随机块模型(SBM),获得了一个更通用的主题建模框架,它能够自动检测主题的数量,并对单词和文档进行分层聚类。分析表明,在统计模型选择方面,SBM 方法比 LDA 方法能得到更好的主题模型。

为了解决 BTM 模型无法直接应用于短文本热点话题发现的问题,Yan 等人在 2015 年提出了突发词对主题模型(Bursty Biterm Topic Model,BBTM)[9],将词对突发概率作为模型的先验知识,可直接用于突发话题的发现。黄畅[10]改进 BBTM 模型,提出热点话题发现方法(Hot topic-Hot Biterm Topic Model,H-HBTM),用传播值来量化词对热值突发概率,设计了一种自适应学习话题数目的方法。林特[11]改进 BBTM 模型量化词对突发概率方法,提出了一种结合基于自动状态机的枚举突发词对和正态分布的方法来量化突发词对。

为了考虑词语间的语义信息,沈兰奔等人[12]结合注意力机制和 BiLSTM 用于检测中文事件。Yuan 等人[13]在 2016 年提出的词共现网络模型(WNTM)将文档中的词共现信息构建成词网络,提高了数据空间的语义密度。彭敏等人[14]提出了一个基于双向 LSTM 语义强化的概率主题模型,强化语义特征之间的关系。和志强等人[15]提出了基于双向 LSTM 的短文本分类算法,该算法能够有效解决短文本分类过程中语义缺乏的问题。

也有很多学者致力于将人工神经网络结合主题模型来研究短文本主题挖掘。Li 等人[16]提出了一种基于反馈递归神经网络的主题模型,将 LSTM 与主题模型结合,提升了模型挖掘文档集合主题的效率。石磊等人[17]利用 RNN 来学习词之间的关系作为先验知识加入到稀疏主题模型,结合主题模型发现社交网络突发话题。张翠[18]等人将 CNN 和 BiLSTM 获取的特征进行融合,能充分理解上下文信息,有效提取文本特征信息。Chitkara 等人[19]提出了一种具有自我注意力的层次模型,将深度学习技术应用于话题发现。

由上述内容可知主流主题模型存在未进行特征选择、没有考虑词语之间语义信息、未削弱高频中性词对主题的影响、需要人工指定话题数目等问题,针对这些问题设计一种基于密度的 BiLSTM-HBBTM 的最优话题数目选择方法,提出基于双向长短期记忆网络的热点突发词对主题模型(BiLSTM based on topic-hot Bursty Biterm Topic Model,BiLSTM-HBBTM)。

2 基本概念

2.1 微博传播值 微博传播值的计算如公式(1)所示。传播值越大,则该微博越有可能是热点微博。

$$spread_d = \gamma \cdot \max\{\ln(rep_d),0\} + \chi \cdot \max\{\ln(com_d),0\} + \mu \cdot \max\{\ln(att_d),0\} \quad (1)$$

其中,$spread_d$ 表示微博文档 d 的传播值,rep_d、com_d、att_d 分别表示微博文档 d 被转发数、被评论数、被点赞数。γ、χ、μ 分别表示微博文本被转发、评论和点赞对微博传播值的影响权重。当 $spread_d = 0$ 时,将该微博标记为噪声微博并将其删除。

2.2 **词项 H 指数** 受到 Hirsch[20] 提出的 H 指数的启发,本文提出词项 H 指数,将每篇微博文档被转发数作为该篇文档每个词语的被浏览次数,词项 H 指数的定义如下:假设有 N 条微博中包含词项 w_i,并且有 H 条微博的被转发频次大于或等于 H 次,那么该 H 值就是词项的 H 指数,用来确定该词项对微博语料库的重要性。

2.3 **词对先验知识** 为了解决 BBTM 模型没有考虑词语之间关系的问题,本文在 BBTM 模型的基础上融入了词对之间的关系作为共现词对分布的先验知识来强化词对主题的相关性。

基于 BiLSTM 的先验知识框架如图 1 所示。首先,BiLSTM-HBBTM 使用词嵌入算法表示文本向量,引入 BiLSTM 来学习词之间的关系。其次,为了过滤高频中性词对于主题质量的影响,BiLSTM-HBBTM 将改进词语的逆文档频率(IDF)作为先验知识的一部分。将 BiLSTM 的输出信息和 IDF 结果的加权值作为模型的先验知识。

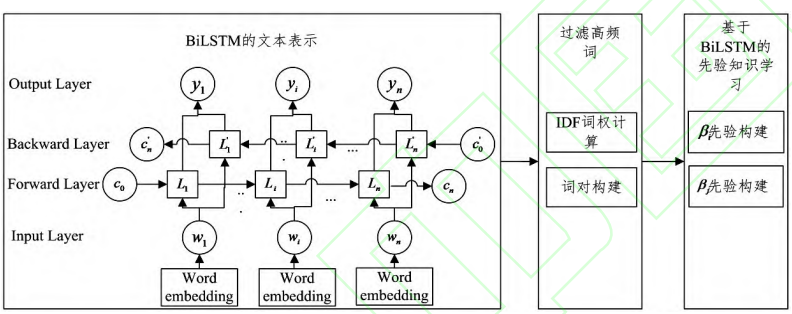

图 1 基于 BiLSTM 的先验知识框架

2.4 **词对突发概率** 假设时间片 t 内词对 b 在微博文本数据集中出现 $n_b^{(t)}$ 次,$n_{b,0}^{(t)}$ 表示出现在普通话题中的次数,$n_{b,1}^{(t)}$ 表示出现在热点话题中的次数,容易得到 $n_b^{(t)} = n_{b,0}^{(t)} + n_{b,1}^{(t)}$。

根据 Yan 等人[9]对 BBTM 模型的分析,在时间片 t 上词对 b 的突发概率估计方法如式(2)所示:

$$\mu_b^{(t)} = \frac{n_{b,1}^{(t)}}{n_b^{(t)}} \tag{2}$$

2.5 **词对热值突发概率** 词对热值突发概率 $\gamma_{b,t}$ 可以表示为词对 b 在 t 时刻的热度值 $\varphi_{b,t}$ 相对于历史平均热度值 $\varphi_{b,h}$ 的增长率,$\varphi_{b,t}$ 和 $\varphi_{b,h}$ 的计算如式(3)和(4)所示。

$$\varphi_{b,h} = \frac{\sum_{j=t-s}^{t-1} \varphi_{b,j}}{s} \tag{3}$$

$$\varphi_{b,t} = \sum_{i=1}^{M_t} (1 + sprea\, d_{i,b}) \tag{4}$$

词对热值突发概率 $\gamma_{b,t}$ 如式(7):

$$\gamma_{b,t} = \frac{\varphi_{b,t} - \varphi_{b,h} - \delta}{\varphi_{b,t}} \tag{5}$$

其中,δ 用于过滤低频词对,s 表示相关时隙大小,M_t 表示 t 时隙内的微博数目,$sprea\, d_{i,b}$ 指词对 b 所在微博 i 的传播值。

3 BiLSTM-HBBTM 算法设计

3.1 **算法步骤** BiLSTM-HBBTM 算法步骤如图 2 所示,以下各节对主要部分进行详细阐述。

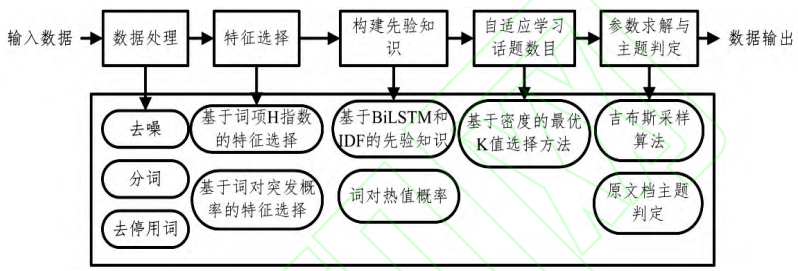

图 2 BiLSTM-HBBTM 算法流程图

3.2 **特征选择** 微博中的词可以分为热点词和非热点词。热点词是指与热点话题相关的词,在文本中出现的次数具有短期突增的特点,利用词项 H 指数和词的突发特性选择微博特征,选择词项 H 指数在前 80% 的词以及突发概率大于阈值 ω 的词作为微博的特征词。这样筛选出来的词能够更利于热点话题的发现,为后面建模减少了维度,降低了数据稀疏性和计算效率。特征选择算法如算法 1 所示。

算法 1 特征选择算法

输入:数据处理后的文本集 text,词突发概率阈值 ω,相关时间片段 s

输出:文本特征集 text_features

1. sorted($re\,p_w$,reverse = True) /按每个词的被浏览数降序排序/
2. calculate H_w /计算词 w 的 H 指数/
3. sorted(H_w ,reverse = True)
4. if H_w 排序在前 80%
5. 将词 w 加入 text_h
6. for w_i ,w_j in text_h
7. $b = (w_i, w_j)$ /获取词对/
8. calculate $\overline{n_b^{(t)}}$ by formula (2)
9. calculate $n_{b,1}^{(t)}$ by formula (3)
10. calculate $\mu_b^{(t)}$ by formula (4)
11. if $\mu_b^{(t)} > \omega$
12. 将词对 b 加入 text_features
13. end for

3.3 **词对热值概率化** BBTM 模型中的词对突发概率只考虑了词出现的频次,但是与热点话题相关的微博不只是表现为相关的微博数量变多,还表现为微

博的评论数、转发数和点赞数增多。热点词是热门微博文本的组成部分,同时具有突发性和传播性。因此,将词对热值突发概率代替词对突发概率作为 BBTM 模型的先验概率。词对热值概率化算法如算法 2 所示。

算法 2 词对热值概率化算法

输入:文本特征集 text_features,相关时间片段 s

输出:$(b,\gamma_{b,t})$

1. $b=(w_i,w_j)$ /读取词对/

2. calculate $\varphi_{b,h}$ by formula (5)

3. calculate $\varphi_{b,t}$ by formula (6)

4. calculate $\gamma_{b,t}$ by formula (7)

3.4 **自适应学习话题数目** 按照主题相似度最小时话题质量最佳的原则,在平均主题相似度最小时确定自适应学习主题数 K。根据文献[21]中的基于密度的自适应学习话题数目的选择方法,本文采用词嵌入 Word2Vec 算法的方式来表示话题向量。由于向量维度太高会淡化词之间的关系,维度太低又不能将词区分,因此将话题向量的维度设置在 300 维。改进的确定话题数目方法称为基于密度的 BiLSTM-HBBTM 最优话题数目 K 值确定方法。BiLSTM-HBBTM 算法中的词嵌入模型使用基于负方向采样的 Skip-gram 词向量来训练模型微博文本向量。基于密度的 BiLSTM-HBBTM 最优 K 值选择方法的基本过程如算法 3 所示。

算法 3 基于密度的 BiLSTM-HBBTM 最优 K 值选择算法

输入:$(b,\gamma_{b,t},\max K_{it})$

输出:K

1. 随机初始化话题数目 K,$K\in(20,100)$。set flag=-1,simHis=1,topic=K,simBest=1。

2. while 话题数目 K 不再改变时 or 达到最大迭代次数

3. Calculate$Si\,m_{avg}$ by formula (12) and(13)

4. if$Si\,m_{avg}>simHis$ then

5. flag=-flag

6. else

7. flag=flag

8. if$simBest>Si\,m_{avg}$ then

9. $simBest=Si\,m_{avg}$, $simHis=Si\,m_{avg}$,topic= K

10. if$si\,m_{kd}<Si\,m_{avg}$ then

11. 统计每个话题的话题密度

12. 计算噪声话题数 C,即话题密度小于 K/3 的话题数。

13. update K, $K=K+\text{flag}\times C$

14. return K

3.5 **模型参数求解** BiLSTM-HBBTM 用 Gibbs 采样方法对参数进行求解,需要采样的参数变量有 z,θ 和 φ。词对的条件概率分布分别如式(6)和(7)所示。

$$P(e_i=0\mid e^{\neg i},z^{\neg i},B,\alpha,\beta,\mu)\propto(1-\mu_i)\frac{n_{0,w_{i,1}}^{\neg i}+\beta}{n_{0,\cdot}^{\neg i}+W\beta}\cdot\frac{n_{0,w_{i,2}}^{\neg i}+\beta}{1+n_{0,\cdot}^{\neg i}+W\beta} \tag{6}$$

$$P(e_i=1,z_i=k\mid e^{\neg i},z^{\neg i},B,\alpha,\beta,\mu)\propto\mu_{b_j}\cdot\frac{n_{\cdot}^{\neg i}+\alpha}{n_{\cdot}^{\neg i}+K\alpha}\cdot\frac{n_{k,w_{i,1}}^{\neg i}+\beta}{n_{k,\cdot}^{\neg i}+W\beta}\cdot\frac{n_{k,w_{i,2}}^{\neg i}+\beta}{1+n_{k,\cdot}^{\neg i}+W\beta} \tag{7}$$

式中,$e=\{e_i\}_{i=0}^{N_B}$,$Z=\{z_i\}_{i=0}^{N_B}$,$\mu=\{\mu_i\}_{i=0}^{N_B}$ 是词对分配给常态词分布的次数,$n_{0,\cdot}=\sum_{k=1}^{W}n_{0,w}$ 是分配给普通话题的词总数,n_k 是分配到热点话题的词对的总数,$n_{\cdot}=\sum_{k=1}^{K}n_k$ 表示分配到热点话题词的总数量,α,β 为先验知识,$n_{k,w}$ 表示词 w 分配给热点主题 k 的次数,$n_{k,\cdot}=\sum_{w=1}^{W}n_{\mathrm{k,w}}$ 是分配给热点主题 K 的词的总数,$\neg\, i$ 表示不包含词对 b。

经过足够次数的迭代后,将收集统计信息,并逐一更新每个单词对的主题类型表示变量和主题分配。这些统计信息可用于估计各种参数。在达到最大迭代次数之后,将学习到的参数的平均值用作参数估计值。最后,推导出微博热点话题分布和单词分布参数结果。如公式(8)和(9)所示。

$$\theta_k=\frac{n_k+\alpha}{n_{\cdot}+K\alpha} \tag{8}$$

$$\varphi_{k,w}=\frac{n_{k,w}+\beta}{n_{k,\cdot}+W\beta} \tag{9}$$

BiLSTM-HBBTM 的吉布斯采样算法具体描述如算法 4 所示。

算法 4 BiLSTM-HBBTM 吉布斯采样算法

输入:(K,α,β,B)

输出:(θ,φ)

1. 随机初始化 e,z

2. foriter=1 to N_{iter} do

3. for each $b_i=(w_{i,1},w_{i,2})\in B$ do

4. 从式(8)和(9)抽取 e_i,k

5. if $e_i=0$ then

6. update $n_{0,w_{i,1}},n_{0,w_{i,2}}$

7. else

8. update $n_k,n_{0,w_{i,1}},n_{0,w_{i,2}}$

9. end for

10. end for

11. returnθ and φ

4 实验分析

4.1 **实验环境** 本章实验均是在 Intel(R) Core(TM),i5-8250U CPU,1.60GHz 的主频,8GHe 内存,Windows 10 操作系统上进行的。应用软件是采用 3.7 版本的 Python 程序结合 2019.3.1 版本的 JetBrains Pycharm 进行实证分析。

4.2 **数据集** 本文选取对热点事件传播影响力较

大的官方微博,通过 Python 爬虫抓取从 2020 年 1 月 1 日至 2020 年 3 月 31 日的微博,共计 151240 条构成本实验原始数据集,用于发现疫情期间的微博热点话题。其中,每条微博数据包括发布时间、微博正文、点赞数、评论数以及转发数,有"广州一市场仍在偷买野生动物""钟南山称已有药物用于临床治疗"等热点话题。

4.3 参数设置

4.3.1 词对突发概率阈值 BiLSTM-HBBTM 在特征选择中的词对突发率阈值 ω 的取值大小会影响到最后话题的质量。ω 的取值范围为(0,1),如果取值太小不能有效的过滤非突发词,取值过大则容易丢失部分突发词,因此进行了阈值 ω 变化对话题质量影响的实验。实验结果如图 3 所示。实验结果纯度(Purity)指标的值越大表示话题质量越好,从图中所知,参数 ω 取 0.4 时纯度最高,说明此时生成的话题质量最好,因此实验中参数 ω 的取值为 0.4。

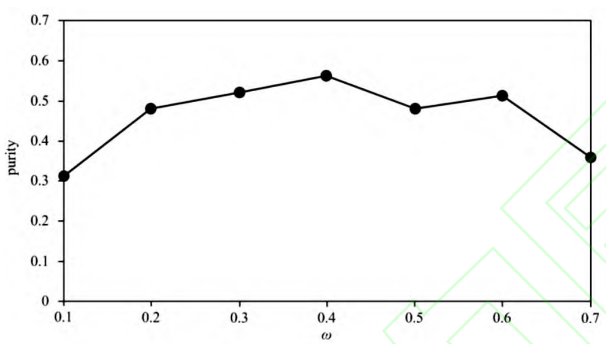

图 3 参数 ω 的实验结果

4.3.2 其他参数 根据参考文献[7],OnlineLDA 的参数 α,β 分别为 0.05,0.01;BTM、BBTM、BiLSTM-HBBTM 中的参数根据文献[4]和[8]的思想,设置为 $\alpha = 50/K$ 和 $\beta = 0.01$,热点主题的数量从 20 个到 100 个不等。Gibbs 采样过程的迭代次数都设置为 1000 次。根据文献[10],H-HBTM,BiLSTM-HBBTM 的其余参数取值为:$s = 4$, $\delta = 1$, $\gamma = 0.7$, $\chi = 0.2$, $\mu = 0.1$。

4.4 评价指标

4.4.1 主题相关性评估

(1) 平均话题相似度

根据文献[21]的思想,主题间平均话题相似度最小时模型发现的各个话题相关程度最低,表明此时模型达到最优。两个文本向量 k 和 d 的相似度 $Sim_{k,d}$ 用余弦距离与 IDF 结合计算,能够削弱高频中性词对主题的影响,文本的相似性计算方法见公式(10)。

$$Sim_{k,d} = \frac{\sum_{i=1}^{n} vec(k_i) \times IDF(k_i) \times vec(d_i) \times IDF(d_i)}{\sqrt{(vec(k_i) \times IDF(k_i))^2} \times \sqrt{(vec(d_i) \times IDF(d_i))^2}} \tag{10}$$

式中,k_i 表示 k 向量对应 i 维上的值;d_i 表示 d 向量对应 i 维上的值。

用两两话题向量间相似度的平均值来表示平均话题相似度,其计算公式如式(11):

$$Sim_{avg} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Sim_{i,j}}{\frac{n(n-1)}{2}} \tag{11}$$

式中,$Sim_{i,j}$ 是第 i 个话题与第 j 个话题之间的相似度,n 表示为话题向量维度。

(2) 点互信息

受到信息论中互信息的启发,本文主题一致性评估采用点互信息(Pointwise Mutual Information,PMI)指标,点互信息的值越高,说明词语的相关性越大,越能解释同一个主题。本文 PMI 计算公式如下式(12)所示。

$$\varphi_{z,w_i} PMI(z) = \frac{2}{N(N-1)} \sum_{1 \leqslant i \leqslant j \leqslant N} \log_2 \frac{p(w_i, w_j)}{p(w_i) p(w_j)} \tag{12}$$

其中,$w_1, w_2, \cdots, w_N$ 为主题 z 前 N 个可能的词,$p(w_i, w_j)$ 是词对 $< w_i, w_j >$ 共现在同一滑动窗口的联合概率分布,$p(w_i)$ 表示为词 w_i 出现在滑动窗口内的边缘概率分布。

4.4.2 话题质量评估

(1) 平均准确度

将在不同热点话题数目 K 下的平均准确度(P@K)作为发现热点话题准确度的评价指标。将算法生成的话题随机混合在一起,邀请 7 个志愿者根据给出的话题信息对生成的热点话题进行人工标注。一个话题被认定是热点话题的标准是:当有超过 50% 志愿者都将该话题标注为热点话题,则该话题被认定是一个热点话题。平均准确度计算如式(13)。

$$P@K = \frac{K_v}{K} \tag{13}$$

其中,K 为算法生成热点话题数目,K_v 为人工标注的热点话题数目。

(2) 熵值和纯度

话题质量评估采用熵值(entropy)[22]和纯度(purity)[23]来度量。整个聚类划分的熵值和纯度的计算如式(14)和(15)。

$$entropy = \sum_{i=1}^{K} \frac{m_i}{m} \left(- \sum_{j=1}^{L} \frac{m_{ij}}{m_i} \log_2 \frac{m_{ij}}{m_i} \right) \tag{14}$$

$$purity = \sum_{i=1}^{K} \frac{m_i}{m} \left[\max \left(\frac{m_{ij}}{m_i} \right) \right] \tag{15}$$

式中,m_i 表示为在聚类 i 中所有成员的个数,m_{ij} 表示为聚类 i 中的成员属于类 j 的个数,m 是整个聚类划分所涉及到的成员个数,K 是聚类的数目,L 是类的个数。

4.5 实验过程和结果分析

为了证明算法的有效性,选取了当前三个业界主流模型 OnlineLDA、BTM 和 BBTM 以及 BBTM 模型的改进算法 H-HBTM 作为基准模型,以平均话题相似度、平均准确度熵值、纯度、点互信息作为指标,在自适应学习话题数目、发现准确度、发现质量和主题一致性这四个角度上,对改进的模型与基准模型进行比较。

4.5.1 自适应学习话题数目实验　针对传统主题模型存在需要人工确定话题数目的问题,BiLSTM-HBBTM 算法中采用了基于密度的 BiLSTM-HBBTM 最优 K 值选择方法(KBiLSTM-HBBTM),用于确定话题数目 K。当主题之间平均余弦距离最小时,话题质量最佳,模型最优。为了证明该方法能够自适应学习话题数目,将其与原方法 KLDA 进行比较,并在运行时间角度同时与 hSBM[8] 比较。

在不同话题数目下,KBiLSTM-HBBTM 与 KLDA 方法平均话题相似度的变化情况如图 2 所示。当话题数目 K=67 时,两种方法的平均话题相似度都最低。由于基于密度的最优 K 值选择方法是根据 LDA 的模型结构提出,将最优 K 值选择与模型参数估计统一在一个框架里,因而基于密度的最优 K 值选择方法会更适合 LDA 模型,KLDA 的平均话题相似度也低于 KBiLSTM-HBBTM。从图 4 的实验结果可知,基于密度的 BiLSTM-HBBTM 的自适应学习话题数目的方法,能够在较好地确定最优话题数目 K。

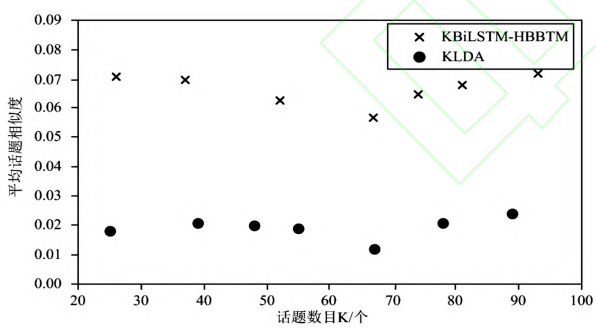

图 4　KBiLSTM-HBBTM 和 KLDA 选择 K 的性能表现

为了证明 BiLSTM-HBBTM 模型使用的词嵌入算法能够改善 LDA 模型的高维度问题,提高运算效率,与 KLDA、hSBM 分别在 10 组不同话题数目下比较完整运行一次算法所用的时间,实验结果如图 5 所示。

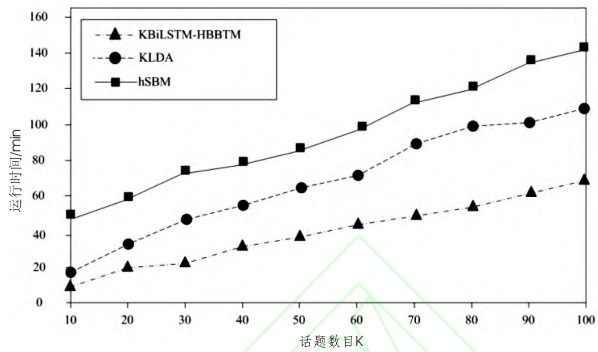

图 5　不同话题数目 K 下的运行时间

从图 5 看出,KBiLSTM-HBBTM、KLDA 与 hSBM 三种方法的运行时间,都随着话题数目的增加而增长。在相同数量的话题数目下,KBiLSTM-HBBTM 的运行时间都少于 KLDA 和 hSBM 模型的运行时间,主要原因是 KLDA 使用向量空间表示话题分布,向量的维数与微博文本中特征词的数量相同,会出现高维度的问题。hSBM 算法构造无向图,数据量越大,运行时间呈几何增长,因此当数据集很大时,所耗费的时间会非常久,运算效率低。而在 KBiLSTM-HBBTM 方法中,采用 Word2Vec 词嵌入的方式来表示话题向量,将话题向量维数设置在 300,极大地降低了话题向量维度,缩短了在不同话题数目下每轮迭代过程中计算话题相似度所消耗的时间。由此可知,与 KLDA、hSBM 方法相比,KBiLSTM-HBBTM 方法能够改善维度过高的问题,缩短计算时间。

4.5.2　特征选择结合 BBTM 与 BBTM、H-HBTM 发现热点话题对比　本文引入了基于传播值和词项 H 指数的特征选择方法,从文档层面和词语层面进行特征选择。为了验证本文提出的特征选择方法的有效性,对比传播值与词项 H 指数结合 BBTM 建模(HBBTM)与 H-HBTM 模型、BBTM 模型,分别得出每类热点主题下的词分布以及词语之间的 PMI。提取“各地医生驰援湖北”、“火神山医院施工现场”两个话题中出现概率最大前 5 个词,实验结果如表 1 所示。

表 1　模型改进前后挖掘主题比较

主题		Topic1--“各地医生驰援湖北”	Topic2--“火神山医院施工现场”
BBTM	话题词	驰援、湖北、奉命、危重症、行者	火神山、医院、搭建、安装、观看
	PMI	1.21	1.18
H-HBTM	话题词	驰援、湖北、医院、抗击、集结	火神山、医院、建设、监工、设施
	PMI	1.38	1.27
HBBTM	话题词	各地、医生、驰援、湖北、奔赴	火神山、医院、施工、建设、进度
	PMI	1.45	1.32

由表 1 可以看出,通过 BBTM 模型得出的关键词从语义上来看有些和主题无关,例如“危重症”和“观看”;通过 H-HBTM 算法提取的话题词能较好地描述主题,但也存在少量干扰词。前两种方法计算出的

PMI 均比 HBBTM 低,说明 BBTM 和 H-HBTM 模型发现的热点话题中词语之间的相关性比 HBBTM 低,HBBTM 能够更好地发现热点话题。实验结果表明,通过微博传播值和词项热度结合 BBTM 建模,每个话题得出的词语与主题高度相关,能与主题相吻合,实验结果优于 BBTM 和 H-HBTM 方法。这是因为传播值综合考虑了微博的被转发数、点赞数、评论数对微博文本的影响;而词项 H 指数考虑了词项的热度。因此,使用结合传播值和词项 H 指数的特征选择法建模得出的词语能够覆盖整个话题的表述。

4.5.3 BiLSTM-HBBTM 与 BBTM、H-HBTM 发现热点话题对比 针对传统主题模型存在忽略词之间关系的问题,本文引入了 BiLSTM 来双向学习词语之间的关系。为了验证 BiLSTM 学习的词之间的关系对挖掘热点话题的的有效性,将 BiLSTM-HBBTM 建模与 H-HBTM 模型、BBTM 模型对比,得出每类热点主题下的词分布以及词之间的 PMI。对比两个话题中出现概率最大的前 5 个词,实验结果如表 2 所示。

表 2 模型改进前后所得的话题词及 PMI

主题		Topic1--"广东发现 10 起家庭聚集性疫情"	Topic2--"李文亮仍在抢救"
BBTM	话题词	确诊、疫情、广东、通报、接触	抢救、生命、抗击、医院、问题
	PMI	1.12	1.07
H-HBTM	话题词	聚集、家庭、传染、新增、场所	李文亮、抢救、重症、病情、无效
	PMI	1.29	1.22
BiLSTM-HBBTM	话题词	广东、家庭、聚集、疫情、病例	李文亮、病危、抢救、医生、感染
	PMI	1.38	1.26

由表 2 可以看出,BiLSTM-HBBTM 模型得出的每个话题的词语与热点主题语义相近,而通过 H-HBTM 和 BBTM 建模得出的每个主题的词语中,有一部分词语与主题语义无关或者语义相差较远。从 PMI 得分也能看出 BiLSTM-HBBTM 输出的词语之间关联程度更高,能够更好地描述主题。这是因为 BBTM 是以概率的方法来计算词的突发概率,并将其作为模型的先验知识,只从统计的角度考虑词语的热度;而 BiLSTM 考虑了词语之间的语义关系,并且利用逆文档频率削弱了高频中性词的影响,因此,引入 BiLSTM 能够更加准确地提取各个热点话题下的关键词,更有利于热点话题的发现。

4.5.4 BiLSTM-HBBTM 与对比算法在话题发现准确度上的比较与分析 为了评估本文方法与基准模型发现热点话题的准确性,计算在不同的热点话题数目 K 下对应的平均准确度(P@K),作为各方法发现热点话题准确度的评价指标。实验结果如表 3 所示。

表 3 不同话题数目下的准确度

方法	P@30	P@60	P@90
OnlineLDA	0.233	0.250	0.344
BTM	0.567	0.483	0.456
BBTM	0.633	0.733	0.756
H-HBTM	0.767	0.817	0.811
BiLSTM-HBBTM	0.833	0.867	0.844

由表 3 可知:BiLSTM-HBBTM 方法的平均准确度都是大于 0.8,明显优于其他方法。这说明 BiLSTM-HBBTM 结合传播值和词项 H 指数进行特征选择,利用 BiLSTM 学习词之间的关系,并且将词对热值突发概率代替词对突发概率作为 BBTM 模型的先验概率,过滤掉一些非热点词,提高了热点话题发现的准确度。

4.5.5 BiLSTM-HBBTM 与对比算法在话题发现质量上的比较与分析 为了评价热点话题发现的质量,选择纯度和熵作为评价指标,纯度越大、熵值越小表示性能越好。话题数目设置为 $K \in [40,65]$。各个方法在不同话题数目下的热点话题聚类结果如图 6 和图 7 所示。

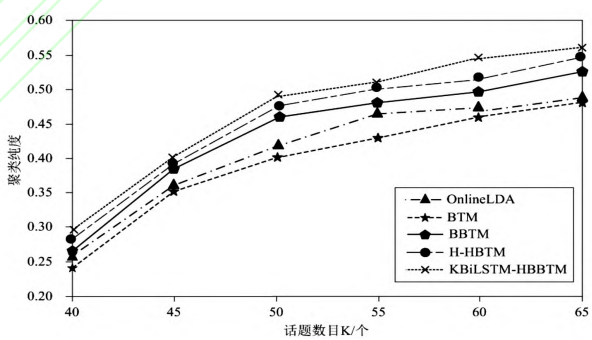

图 6 不同话题数目下的聚类纯度

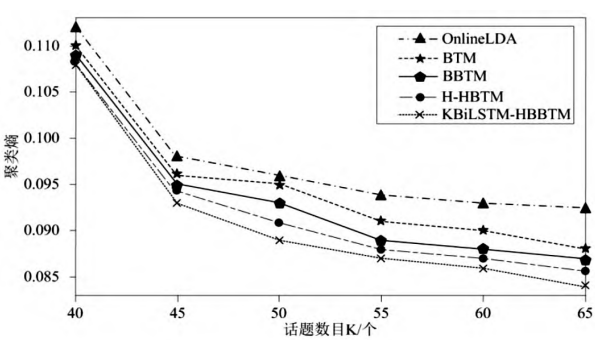

图 7 不同话题数目下的聚类熵

由图 6 和图 7 的实验结果可以看出,相比其他对比算法,本文提出的 BiLSTM-HBBTM 方法在纯度和熵指标的实验结果更好。BBTM、H-HBTM 的实验效果较好,但稍微差于本文所提方法,这是因为 BiLSTM-HBBTM 利用微博传播值和词项 H 指数选择微博文本

和特征词,并且将词对热值突发概率作为模型的先验概率,更好地表征词对热度;考虑了词语关系,过滤掉高频中性词,能够更准确地发现热点问题。

4.5.6 BiLSTM-HBBTM 与对比算法在主题一致性上的比较与分析 本文选用点互信息(PMI)指标来度量 BiLSTM-HBBTM 方法的主题一致性,当 PMI 越高时,表明该主题的主题一致性更强。BiLSTM-HBBTM 与对比算法在不同热点话题数量下的热点话题的主题一致性结果如图 8 所示。

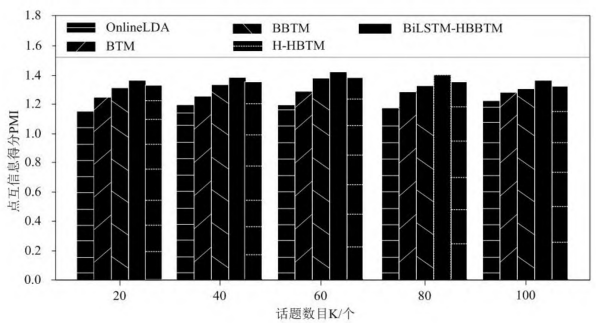

图 8 不同话题数目下的主题一致性

由图 8 可知,相比于其他对比方法,BiLSTM-HBBTM 方法的主题一致性实验效果更优,说明 BiLSTM-HBBTM 方法发现的热点话题里,各个词语之间的一致性更高。

为了定性分析主题一致性,此次实验随机挑选了一个出现频次较高且是热点话题的话题标签。抽取话题"双黄连可抑制新型冠状病毒"的实验结果,分别列出概率最高的前 10 个词语,如表 4 所示。

表 4 话题"双黄连可抑制新型冠状病毒"发现概率最高的前 10 个词语及 PMI

算法	前 10 个词	PMI
OnlineLDA	售罄、恐慌、证实、抢购、人用、公布、蔗糖、联合、发现、价格	0.92
BTM	口服液、药物、中成药、抑制、热度、成分、生产、证实、早期、购买	1.07
BBTM	双黄连、口服液、抑制、拔高、药物、生产、疗效、不适、抢购、热门	1.21
H-HBTM	双黄连、新型、药物、冠状病毒、抑制、上海、口服液、发现、临床、病情	1.28
BiLSTM- HBBTM	双黄连、抑制、冠状病毒、中成药、口服液、服用、脱销、控制、预防、好处	1.36

由表 4 的实验结果可知,BiLSTM-HBBTM 中词语之间的 PMI 最大,说明各个词语间语义相关性最强,与话题的一致性也更强。BBTM、H-HBTM 发现的话题关键词的相关性也较大,但也存在与话题不相关或相差较远的词语。OnlineLDA 中的 PMI 最低,词以日常通用词语为主,与话题相关的词语比较少,因此,在所有对比方法中,OnlineLDA 得出的结果与主题相关性最低。BTM 的实验结果虽然略优于 OnlineLDA 方法,但也包含了较多的日常通用词,比如说"生产"和"早期",表明 BTM 模型挖掘的主题有可能是普通话题,不是热点话题。

4.5.7 总结 从前文的分析可以得到以下各个模型功能模块的对比结果,详见表 5。

表 5 模型功能模块对比

功能模块	BiLSTM-HBBTM	Online-LDA	BTM	BBTM	H-HBTM
数据稀疏	√	√	√	√	√
传播特性	√	×	×	×	√
高维度	×	√	×	×	×
自适应学习话题数	√	×	×	×	√
词对突发概率	√	×	×	√	√
词对热值概率化	√	×	×	×	√
削弱高频词影响	√	×	×	×	×
词之间关系	√	×	×	×	×
启发式后处理	×	√	√	×	×

从表 5 可以看出,传统主题模型仍然存在一定的缺陷。本文提出的基于双向长短期记忆网络的热点突发词对主题模型(BiLSTM-HBBTM)在话题发现准确度、话题质量、话题一致性方面都取得了较好的实验效果。这是因为 BiLSTM-HBBTM 结合微博的传播性与词项热度进行了文档和词项的特征选择,将词之间的关系和词对热值概率作为词对的先验知识,同时削弱高频中性词对话题的影响,采用基于密度的自适应学习话题数目方法,能够从嘈杂的微博文本中挖掘出高质量的热点话题。

5 结 语

本文提出了一种基于基于突发词对主题模型改进的微博热点话题发现方法(BiLSTM-HBBTM),用来发现微博中的热点话题。BiLSTM-HBBTM 先引入微博传播值、词项 H 指数和词对突发概率,从文档和词语两个层面进行特征选择,再通过 BiLSTM 训练词语之间的关系,计算词对热值突发概率,为 BBTM 模型提供了更加准确的先验知识,最后使用基于密度的方法自适应选择话题数目,解决了传统的主题模型需要人工指定话题数目的问题。然而,本文数据集只选取微博的文本进行建模,但微博数据中还包含有图片、视频、音频、表情包等相关能反映话题的信息,未来或许可以考虑结合多方面的数据信息建模来更精确的挖掘热点话题。

参 考 文 献

[1] Deerwester S C, Dumais S T, Furnas G W, et al. Computer information retrieval using latent semantic structure: US, CA19890596524 [P]. 1989-04-12.

[2] Hofmann T. Probabilistic latent semantic analysis[C]. Fifteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc. 1999:391-407.

[3] Blei D M,Ng A Y,Jordan M I. Latentdirichlet allocation[J]. J Machine Learning Research Archive,2003,3:993-1022.

[4] Yan X H,Guo J F,Lan Y Y,et al. A biterm topic model for short texts[C]//Proceedings of the 22nd International World Wide Web Conference,Rio de Janeiro,May 13- 17,2013. New York:ACM,2013:1445-1456.

[5] 王亚民,胡 悦. 基于 BTM 的微博舆情热点发现[J]. 情报杂志,2016,35(11):119-124.

[6] 李卫疆,王真真,余正涛. 基于 BTM 和 K-means 的微博话题检测[J]. 计算机科学,2017,44(2):257-261.

[7] Hoffman M D ,Blei D M ,Bach F R . Onlinelearning for latent dirichlet allocation[C]//Advances in Neural Information Processing Systems 23:24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010,Vancouver,British Columbia,Canada. Curran Associates Inc. 2010.

[8] Gerlach M,Peixoto T P,Altmann E G. A network approach to topic models[J]. Science advances,2018,4(7):eaaq1360.

[9] Yan X H,Guo J F,Lan Y Y,et al. A probabilistic model for bursty topic discovery in microblogs[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, Jan 25 - 30, 2015. Menlo Park:AAAI,2015:353-359.

[10] 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.

[11] 林特. 短文本流突发性话题发现:BBTM 改进算法[J]. 电脑知识与技术,2017,13(1):248-250.

[12] 沈兰奔,武志昊,纪宇泽,林友芳,万怀宇. 结合注意力机制与双向 LSTM 的中文事件检测方法[J]. 中文信息学报,2019,33(9):79-87.

[13] Zuo Y,Zhao J,Xu K . Word network topic model:A simple but general solution for short and imbalanced texts[J]. Knowledge & Information Systems,2016,48(2):379-398.

[14] 彭 敏,杨绍雄,朱佳晖. 基于双向 LSTM 语义强化的主题建模[J]. 中文信息学报,2018,32(4):40-49.

[15] 和志强,杨 建,罗长玲. 基于 BiLSTM 神经网络的特征融合短文本分类算法[J]. 智能计算机与应用,2019,9(2):21-27.

[16] Li L S,Gan S J,Yin X D. Feedback recurrent neural network-based embedded vector and its application in topic model[J]. EURASIP Journal on Embedded Systems,2017,2017(1):5.

[17] 石 磊,杜军平,梁美玉. 基于 RNN 和主题模型的社交网络突发话题发现[J]. 通信学报,2018,39(4):189-198.

[18] 张 翠,周茂杰. 一种基于 CNN 与双向 LSTM 融合的文本情感分类方法[J]. 计算机时代,2019(12):38-41.

[19] Chitkara P,Modi A,Avvaru P,et al. Topicspotting using hierarchical networks with Self Attention[C]// Proceedings of the 2019 Conference of the North. 2019.

[20] Hirsch J E. An index to quantify an individual's scientific research output[J]. Proceedings of the National Academy of Sciences of the United States of America,2005,102(46):16569-16572.

[21] 曹 娟,张勇东,李锦涛,等. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报,2008,31(10):1780-1787.

[22] Gockay, Erhan, Principe, et al. Information theoretic clustering.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2002.

[23] Murphy K P. Machine Learning:A Probabilistic Perspective[M]. MIT Press,2012.