第 45 卷 第 1 期 计 算 机 工 程 2019 年 1 月
Vol. 45 No. 1 Computer Engineering January 2019

·人工智能及识别技术· 文章编号：1000-3428(2019)01-0210-07 文献标志码：A 中图分类号：TP391

基于双词语义扩展的 Biterm 主题模型

李思宇，谢 珺，邹雪君，续欣莹，冀小平

（太原理工大学 信息工程学院，山西 晋中 030600）

摘 要： 针对 Biterm 主题模型短文本文档的双词产生过程中词对之间缺乏语义联系的情况，提出一种融入词对语义扩展的 Biterm 主题模型。考虑双词的语义关系，引入词向量模型。通过训练词向量模型，判断词与词之间的语义距离，并根据语义距离对 Biterm 主题模型进行双词语义扩展。实验结果表明，与现有 Biterm 主题模型相比，该模型不仅具有较好的短文本主题分类效果，而且双词间的语义关联性能及主题词义聚类性能也得到明显提升。

关键词： Biterm 主题模型；双词；词向量；双词语义；吉布斯采样

中文引用格式： 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.

英文引用格式： LI Siyu，XIE Jun，ZOU Xuejun，et al. Biterm topic model based on semantic extension of double words[J]. Computer Engineering，2019，45(1)：210-216.

Biterm Topic Model Based on Semantic Extension of Double Words

LI Siyu，XIE Jun，ZOU Xuejun，XU Xinying，JI Xiaoping

（College of Information Engineering，Taiyuan University of Technology，Jinzhong，Shanxi 030600，China）

【Abstract】 Aiming at the lack of semantic connection between double words in Biterm Topic Model(BTM) short text documents，a BTM based on semantic extension of double words is proposed. Considering the semantic relationship between each word in double words，the word vector model is introduced. By training the word vector model，the semantic distance between word in double words is judged，and the BTM is extended according to the semantic distance. Experimental results show that，compared with the existing BTM，this model not only has better short text topic classification effect，but also improves the performance of semantic association and topic meaning clustering between double words.

【Key words】 Biterm Topic Model(BTM)；double words；word vector；double words semantic；Gibbs sampling

DOI： 10.19678/j.issn.1000-3428.0049745

0 概述

文献[1]指出主题模型主要是利用词项在文档级中的共现信息，组成具有语义相关性的主题集合，同时文档级的词空间变为主题空间，对文档进行了主题降维。主题模型如潜在语义分析(Latent Semantic Analysis，LSA)[2]、概率潜在语义索引(Probabilistic Latent Semantic Indexing，pLSA)[3]以及潜在狄利克雷分布(Latent Dirichlet Allocation，LDA)[4]等主题模型以及各种基于主题模型的扩展模型在文本分类领域已得到了广泛应用。但是文献[5]指出 LDA 模型对于短文本主题分类效果不佳，主要是由于文档太短不利于训练 LDA。近年来，文献[6-7]对 LDA 短文本处理进行改进，取得了较好的分类效果。

Biterm 主题模型(Biterm Topic Model，BTM)[8]通过提取文档中词共现的方式，提高主题模型对于短文本的分类效果，但仅考虑文档中双词共现，仍存在特征稀疏的问题。在 BTM 基础上，研究人员提出了很多改进方法。文献[9]引入奇异值分解方法，估计词与文档的相似性，将相似度高的词加入短文本中，用于缓解稀疏性，实现情感检测。文献[10]提取短文本中的卡方特征，然后利用 BTM 对短文本建模，有效降低了短文本的特征稀疏。文献[11]提出 TF-IWF 和 BTM 融合的短文本分类方法，缺少上下文的语义信息。文献[12]通过引入事件的概念，将高阶语义单元融入主题的生成过程中，取得了一定的成果。

在现有的 BTM 改进模型中，主要是通过统计方法来降低数据的稀疏性，缺少短文本相应的外部知

基金项目： 山西省回国留学人员科研项目(2015-045)。

作者简介： 李思宇(1992—)，男，硕士，主研方向为自然语言处理、文本主题模型；谢 珺，副教授、博士；邹雪君，硕士；续欣莹、冀小平，副教授、博士。

收稿日期： 2017-12-19 **修回日期：** 2018-01-24 **E-mail：** 593737285@qq.com

识,对语义信息的处理效果不佳。文献[13]指出深度学习在自然语言处理中取得了较好的成果。短文本外部知识在深度学习处理后能得到更好的语义信息。文献[14]将 BTM 与 word2vec 相结合,建立意图识别模型,使动词的主题聚类得到了更高的一致性,但该方法仅针对动词,忽视了其他词性对于主题的影响。本文将基于深度学习的词向量方法融入到主题模型的产生过程中,利用词向量的有序性对词袋模型进行相应的改进。在 BTM 的基础上,结合 word2vec[15] 词向量模型,提出一种双词扩展的 BTM 主题模型(WBW-BTM)。

1　相关工作

1.1　LDA 主题模型

LDA 主要是一个以主题-文档-主题词结构为主的 3 层生成模型,对于参数的确定,主要利用吉布斯采样进行估计。在采样完成后,找到每篇文档的主题分布以及每个主题的主题词。

1.2　BTM 主题模型

BTM 主题模型[8] 主要是针对短文本进行建模的主题模型,将整个语料库中的词共现作为特征,避免短文本特征稀疏的问题。BTM 模型中整个语料库是不同主题模型的混合组合,每个词对都是从一个独立的主题中产生。主题模型的表示方式有 2 种,一种是图模型,另一种是生成过程。BTM 图模型表示如图 1 所示,其符号及参数含义如表 1 所示。

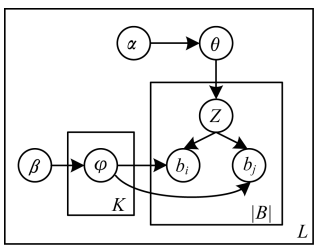

图 1　BTM 图模型表示

表 1　BTM 符号及参数含义

符号及参数	含义
α	狄利克雷分布,θ 的超参数
θ	语料库级的主题多项分布
β	狄利克雷分布,φ 的超参数
φ	某个主题下词的多项分布
Z	从主题多项分布 θ 中选出的主题
b_i, b_j	从主题词的多项分布 φ 中选出的词对
$\lvert B\rvert$	语料库中所有的词对
L	整个语料库

按照 BTM 的基本框架,BTM 模型的文档产生过程具体如下:

1)对每一个主题 Z,描述确定主题 Z 下的词分布 $\varphi \sim Dir(\beta)$。

2)对短文本语料库 L,描述一个语料库级别的主题分布 $\theta \sim Dir(\alpha)$。

3)对于词对 $|B|$ 中的每一个词对按照以下步骤产生,假设一个词对用 b 表示,则 $b=(b_i,b_j)$:

(1)从语料库级别的主题分布 θ 中抽取一个主题 Z,即 $Z \sim Muli(\theta)$。

(2)从被抽取到的主题 Z 中同时抽取 2 个词 b_i、b_j,服从基本假设,每一个词对都从一个独立主题中产生,即 $b_i, b_j \sim Muli(\varphi)$。

根据以上过程可以计算出词对 $b=(b_i,b_j)$ 的联合分布概率,概率如下:

$$P(b)=\sum_z P(z)P(b_i|z)P(b_j|z)=\sum_z \theta_z \varphi_{i|z}\varphi_{j|z} \tag{1}$$

对于整个语料库:

$$P(B)=\prod_{(i,j)}\sum_z \theta_z \varphi_{i|z}\varphi_{j|z} \tag{2}$$

对于整个语料库的主题分布参数 θ 和主题-词分布参数 φ 的估计如下:

$$\theta_z=\frac{n_z+\alpha}{|B|+K\alpha} \tag{3}$$

$$\varphi_{b|z}=\frac{n_{b|z}+\beta}{\sum_b n_{b|z}+M\beta} \tag{4}$$

其中,K 代表主题个数,M 为语料库中不同词语的个数,不考虑重复的词语,n_z 表示主题 Z 的个数,$n_{\omega|z}$ 表示主题词 Z 下 ω 的出现次数。

1.3　word2vec 模型

word2vec 模型[15] 主要是利用一个 3 层的神经网络模型,包括输入层、投影层和输出层,通过训练语料将词语转化为 K 维具有实数含义的向量。通过训练将词语特征映射为一个 K 维的向量,从而计算词语之间的语义相关性,可以将该结果应用于自然语言处理的很多方面,如聚类、同义词分析、情感分析等。

word2vec 包含 2 种训练模型:CBOW 和 Skip-gram。前者主要是通过上下文预测当前词语,后者主要是在已知当前词语的条件下,预测其上下文。两者的目标函数如下:

$$f(x)_{\text{CBOW}}=\sum_{\omega\in c}\lg\, p(\omega|Context(\omega)) \tag{5}$$

$$f(x)_{\text{Skip-gram}}=\sum_{\omega\in c}\lg\, p(Context(\omega)|\omega) \tag{6}$$

其中,$f(x)$ 代表目标函数,ω 代表当前词语,$Context(\omega)$ 满足窗口距离 H 的当前词语 ω 的上下文。

对目标函数的构建有 2 种方式,一种是利用 Hierarchical Softmax 构造,另一种是采用随机负采样来估算。文献[16]提出一个简化版本,主要用来提高训练速度并改善所得词向量的质量,与 Hierarchical Softmax 相比,不使用相对复杂的 Huffman 树,而是利用相对简单的随机负采样,大幅提高了 word2vec 的训练性能。本文采用随机负采样对 word2vec 语料库进行训练。

2 WBW-BTM 主题模型

本文主要将基于深度学习的词向量方法融入到主题模型的产生过程中,利用词向量的有序性对词袋模型进行相应的改进。为了验证模型的有效性,在此对 LDA 及 BTM 主题模型进行相应的改进。

2.1 融入词对语义扩展的 W-LDA 主题模型

针对传统的 LDA 主题模型,主题词与文档之间缺乏相应的语义联系,在此采用词向量的方式将主题词与文档之间的语义关系量化,证明融入词向量模型对主题模型中主题词的产生有一定影响。

在 LDA 主题模型中,通过式(7)产生文档向量:

$$\boldsymbol{D}_m = \frac{1}{l}\sum_{i=1}^{l} f_i \times \boldsymbol{w}_i \tag{7}$$

其中,$\boldsymbol{D}_m$表示第 m 篇文档的向量,f_i 表示第 i 个词在文档 m 中出现的次数,$\boldsymbol{w}_i$表示第 i 个词的词向量,l 表示第 m 篇文档中词语的长度。

然后,利用余弦公式计算出每个主题词与文档之间的语义相关,进行语义相关词扩展。

2.2 针对 BTM 主题模型的改进 WBW-BTM 模型

对于传统的 BTM 主题模型来说,假设经常在一起出现的 2 个词语具有一定的相关性,这样可以降低短文本主题分类时的特征稀疏性,但只是利用了双词一起出现的频率进行双词的特征采样,并没有考虑到双词之间是否存在某种关系,比如天气非常好该句子,对于其预处理后分为天气、非常、好,结合后的双词分别为“天气-非常”“天气-好”“非常-好”“非常-天气”“好-天气”“非常-天气”。从以上6 个双词组合中可以看出,在“天气-非常”和“非常-天气”中,随机组合的 2 个词语,一是缺少语序的正常组合,二是缺少语义之间的相关联系。由上面的实例可以看出,传统的 BTM 模型不但减少了主题模型对于短文本进行主题建模的特征稀疏性,而且减弱了词语之间的语序状态以及语义相关性。

针对上述问题,本文在 BTM 主题模型的基础上,改进吉布斯采样中双词采样方式,使具有语义关系的双词共现,这样可以增强主题内部的语义关联,更好地对主题进行分类。

WBW-BTM 主题模型整体框架如图 2 所示。该模型主要利用 word2vec 计算的语义相似度,针对 BTM 双词抽取过程中缺少双词间的顺序及语义联系的问题,利用 word2vec 对 BTM 采样中双词语义进行扩展,使 BTM 主题模型中双词的特征具有语义以及上下文关联。

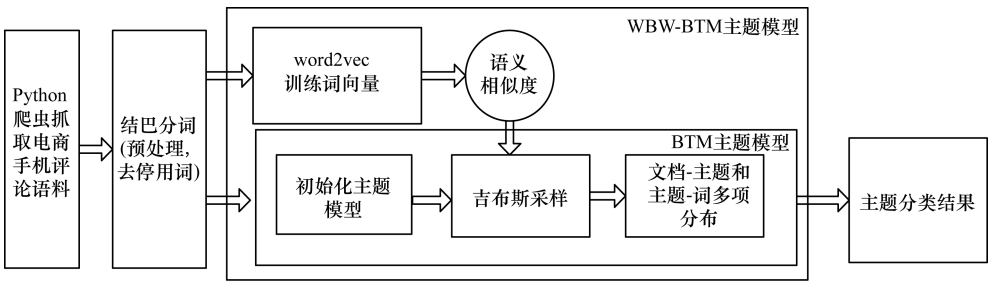

图 2 WBW-BTM 主题模型整体框架

本文利用 word2vec 得到词语之间的语义关联性,利用式(8)计算出词语b_i 和b_j 之间的语义空间距离。

$$\gamma = sim(b_i, b_j) = \frac{\boldsymbol{b}_i \cdot \boldsymbol{b}_j}{|b_i||b_j|} \tag{8}$$

将 word2vec 的结果与 BTM 主题模型结合,在吉布斯采样的过程中对双词特征进行扩展。在吉布斯采样初始化时,利用 γ 与给定的语义距离值 C 之间的距离关系来确定词对扩展程度,若满足距离要求,则:

$$N_{bz} = N_{bz} + \gamma \times 10 \tag{9}$$

其中,N_{bz}表示在主题 Z 下采样词对 b 的次数。若不满足距离要求,则:

$$N_{bz} = N_{bz} + 1 \tag{10}$$

利用式(10)对 N_{bz}进行扩展。完成初始化过程后,在吉布斯采样过程中,依据 γ 与给定的语义距离值 C 之间的距离关系,每次采样过程中对主题的更新采用不同的方式,若满足距离要求,则:

$$P(z|z_{X\neg b}, B, \alpha, \beta, \gamma) \propto (n_z + \alpha) \cdot \frac{[n_{b_i|z} + \sum_{K=1}^{K}(\gamma \times 10) + \beta][n_{b_j|z} + \sum_{K=1}^{K}(\gamma \times 10) + \beta]}{[\sum_b n_{b|z} + 2\sum_{K=1}^{K}(\gamma \times 10) + M\beta]^2} \tag{11}$$

否则:

$$P(z|z_{X\neg b}, B, \alpha, \beta) \propto (n_z + \alpha) \cdot \frac{[n_{b_i|z} + \beta][n_{b_j|z} + \beta]}{[\sum_b n_{b|z} + M\beta]^2} \tag{12}$$

其中,$X\neg b$ 表示去除词对 b 之外的词对,M 表示语料中互不相同的词语。在完成吉布斯采样后,确定 n_z、$n_{\omega_i z}$和 $n_{\omega_j z}$。对语料中主题的多项分布参数 θ_z 和主题词下的多项分布参数 $\varphi_{b|z}$进行计算,确定文档-主题以及主题-词的概率分布。

$$\theta_z = \frac{n_z + (\sum_{K=1}^{K} \gamma \times 10) + \alpha}{|B| + (\sum_{l=1}^{l} \gamma \times 10) + K\alpha} \tag{13}$$

$$\varphi_{b|z} = \frac{n_{b|z} + (\sum_{K=1}^{K} \gamma \times 10) + \beta}{\sum_{b} n_{b|z} + (2\sum_{K=1}^{K} \gamma \times 10) + M\beta} \tag{14}$$

在式(13)中,l 代表满足距离条件的所有扩展词对的数量。对语料中主题的多项分布参数 θ_z 和主题词下的多项分布参数 $\varphi_{b|z}$ 进行计算,确定文档-主题以及主题-词的概率分布。

吉布斯采样主要采用马尔科夫模型。在采样的过程中,每次对符合语义距离的词进行扩展时,符合语义距离值扩展的数量多,不符合语义距离值扩展的数量少。所以在随机采样的过程中,容易造成采样的不平衡,导致最后更新的主题值矩阵的某些主题值为负值,即 $\boldsymbol{P}_j < 0$。在此将负值转化为正值,即 $\boldsymbol{P}_j = -\boldsymbol{P}_j$,主要是因为在随机采样过程中,如果经常采样到某个随机主题,那么说明在该主题下符合语义距离值的词对比较多,所以该主题值就越重要。这样一方面可以突出随机更新的主题值为负值的重要性,另一方面可以使吉布斯采样进入稳态。

算法　WBW-BTM 吉布斯采样算法

输入　主题数量 K,超参数 α、β,双词组合数量 $|B|$,迭代次数 N,语义距离阈值 C,主题值矩阵 $\boldsymbol{P}$

输出　主题分类结果

```
1. 利用 word2vec 计算语义距离
for B = 1 to |B|
  for b_i, b_j ∈ B do
      按式(8)计算语义相似度 γ
2. 初始化所有词对主题:
for B = 1 to |B|
   for b ∈ B do
     if γ > C
       式(9)
     else
       式(10)
3. Gibbs 吉布斯采样过程
  for iter = 1 to N do
   for b ∈ B do
     if r > c
       按式(11)计算出每次更新的主题
       for j in P:
        if P_j < 0:
          P_j = - P_j
       更新吉布斯采样的参数,n_z,n_ωz 和 n_ωz
     else
       按式(12)计算每次更新的主题
       for j in P:
         if P_j < 0:
           P_j = - P_j
       更新吉布斯采样的各参数,n_z,n_ωz 和 n_ωz
4. 按照式(13)和式(14)计算出 θ,φ
```

3　实验结果与分析

3.1　实验数据

实验训练的 word2vec 模型语料来自于网络爬虫,主要从各大电商网站进行手机评论的抓取,集中于手机评论的原因是为了测试主题模型时,主题可以有大概的范围,相比不同的评论语料,更具专一性。采集到的商品评论中共有原始单词 245 221 407 个,评论 4 904 600条,出现不同汉字 32 757 个。训练语料和测试语料用十折交叉验证法进行处理。经过预处理后的部分实验数据具体如下:

外观 不错 第一次 华为 手机 功能 一段时间 追加
宝贝 收到 棒棒 确认 收货 忘记 不好意思
送货 速度 超级 弟弟 喜欢 说 手感
手机 收到 喜欢 支持 国产机 杜绝 美国机
手机 确实 力 好用 支持 国产

3.2　word2vec 训练标准

word2vec 主要采用 negative-sampling 训练数据,词语维度为 200 维。为了适应短文本,窗口大小调整为 10,初始学习率为 0.025,去除语料中频率小于 5 的词语,语料库上的迭代次数为 20。

3.3　评价标准

现有主题模型的评估大多采用主题凝聚度(Topic Coherence,TC)和 KL(Kullback-Leibler)散度。两者的区别是主题凝聚度主要利用主题词之间的共现来反映主题的内聚程度,而 KL 散度主要是通过不同主题之间的区别来衡量主题的差异性。本文采用这 2 个指标从主题凝聚度和差异性两方面进行主题质量评估。

3.3.1　主题凝聚度

文献[17]提出一种针对主题模型检验主题参数的方法,主要是基于词共现的文档数目统计量,具体如下:

$$TC(t;B^{(t)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1} \lg \frac{N(b_m^{(t)}, b_l^{(t)}) + 1}{N(b_l^{(t)})} \tag{15}$$

其中,$N(b)$ 表示包含词对 b 的文档数目,$N(a,b)$ 表示存在词对 a 和 b 共现的文档数目,$B^{(t)} = (b_1^{(t)}, b_2^{(t)}, \cdots, b_M^{(t)})$ 表示主题 t 下概率最大的 M 个词对。TC 得分越高,主题凝聚度就越高。

3.3.2　JS 距离

传统的 KL 距离从信息熵的角度考虑评价相同空间中 2 个概率分布之间的差异大小。本文主要衡量程序运行后生成的不同主题-词分布之间的信息差异。KL 距离越大,获得的主题质量越高。KL 散度公式具体如下:

$$D_{\mathrm{KL}}(p \| q) = \sum_i p(i) \lg \frac{p(i)}{q(i)} \tag{16}$$

其中,p 和 q 分别表示不同主题下的主题-词分布,i 表示主题-词分布的数量。

由于传统 KL 距离的不对称性,不能完全表示

2 个主题-词分布之间的单向关系,因此采用 JS(Jensen-Shannon)距离先计算出每个主题-词分布与平均分布的 KL 距离,然后再求出 2 个不同的主题-词分布的 KL 距离。

$$m = (p + q)/2 \tag{17}$$

$$D_{JS} = 1/2 \times KL(p \| m) + 1/2 \times KL(q \| m) \tag{18}$$

其中,m 表示平均分布的 KL 距离,D_{JS}表示根据平均距离计算出的 KL 距离。

3.4 对照实验及结果分析

3.4.1 定性评估

定性评估是对主题模型的最终结果进行直观的判断,主要从词向量相关性和改进前后的主题模型性能 2 方面进行评估。

1)词向量相关性

本文将训练出的词向量进行定性测试,随机选取手机的多个属性作为中心词,然后依据中心词,通过词向量模型找到与属性词最相关的前 5 个属性,在此用余弦距离计算得出相关度,部分结果如表 2 所示。余弦距离越接近 1,说明语义相关性越高,从结果来看,训练出的词向量模型,符合实际期望。

表 2 训练得到的词向量模型

中心词	相关词	
	词语	余弦距离
手机	机子	0.693 12
	东西	0.681 21
	宝贝	0.636 54
	机器	0.613 99
	外观	0.602 97
华为	魅族	0.549 80
	荣耀	0.370 02
	国产	0.323 89
	青春	0.319 23
	魅蓝	0.293 70
快递	物流	0.805 70
	发货	0.617 83
	配送	0.552 24
	送货	0.552 10
	递给	0.371 55
质量	品质	0.454 87
	不错	0.448 60
	东西	0.419 41
	京东	0.380 09
	手机	0.377 06

2)改进前后的主题模型性能对比

由于在主题模型中无法指定结果运行出的主题中心,因此选用不同算法结果中的不同主题进行对比,如表 3 所示。表 3 结果为在每个主题下有 20 个关键词,按照主题相关性,乱序抽取 10 个,可以看出使用词向量改进后的主题模型的关键词之间的语义相关性得到了明显增强。

表 3 主题模型语义相关性对比

主题模型	主题	关键词
LDA 模型	外观	照相、屏幕、外观、国产、漂亮、后盖、质量、京东、字体、小时
W-LDA 模型	手机	手机、机子、还好、东西、宝贝、音质、照相、机器、苏宁、服务
BTM 模型	电池	电池、手机、充电、感觉、耐用、发热、支持、信号、游戏、系统
WBW-BTM 模型	快递	速度、满意、发货、快递、好评、服务、配送、值得、好看、物流

3.4.2 定量评估

定量评估是对主题模型的最终结果进行量化判断,主要从词对采样数量、距离阈值、主题凝聚度和 JS 距离 4 个方面进行量化评估。

1)不同距离阈值 C 的词对采样数量对比

如果基于词扩展方式对 BTM 主题模型进行改进,那么在选取不同阈值 C 的条件下,扩展的词对数量是不一致的。从经验上来说,语义距离越相关的词语,词对的数量就越少。从图 3 可以看出,随着距离阈值的增大,不同阈值下的词对数量不断减少。同时,在不加入语义距离阈值 C 时,语料中词对的数量为 1 345 943,增加语义距离阈值 C 后,语料中词对的数量增加到 3 663 470,证明了基于语义距离阈值 C 的词扩展方式的有效性,结果如图 3 所示。

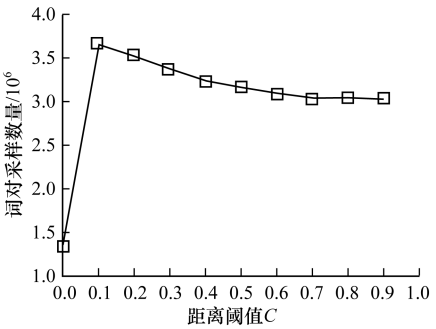

图 3 距离阈值 C 与词对采样数量的关系

2)距离阈值 C

基于词对扩展方式对 BTM 主题模型进行改进,主题的好坏不仅与训练过程有关,而且与语义阈值 C 也有一定的关系。不同的语义阈值 C 不仅可以扩充不同数量的词对,而且可以增强词对内部词语之间的语义相关性。因为本文主题模型比传统主题模型增加了一个语义参数 C,所以本文将通过实验找出最适合的阈值 C。实验中 BTM 主题模型及改进的 WBW-BTM 主题模型都采用相同的狄利克雷分布参数,$\alpha = 50/K$,$\beta = 0.01$,变量为语义阈值 C。从

图 4 可以看出,在主题数量为 5 时,不同的距离阈值 C 差别不大,随着主题数量的增加,不同距离阈值 C 和主题数目下的主题凝聚度越来越规律。可以看出,在语义距离阈值 C 为 0.4 时,主题凝聚度最高。在 BTM 主题模型以及 WBW-BTM 主题模型的对比中,语义距离阈值 C 的取值为 0.4。

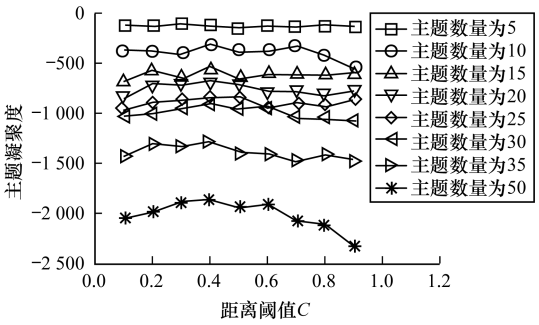

图 4　不同距离阈值 C 下的主题凝聚度对比

3)主题凝聚度

在对比实验中,参数设置为 $\alpha = 50/K$、$\beta = 0.01$、$C = 0.4$,在以上 3 个参数确定的情况下,通过评价标准主题凝聚度来进行验证。本文分别抽取的主题数量为 5、10、15、20、25、30、35、50。从图 5 可以看出,当主题数目为 5、10、15 时,WBW-BTM 主题模型与 BTM 主题模型的区别不大,随着主题数量的增加,TC 值区分越来越明显,聚类效果也越来越好。而对于 LDA 主题模型与 W-LDA 主题模型来说,在主题数目 10 ~ 25 之间,聚类性能逐渐下降。在加入词向量之后,LDA 和 W-LDA 主题模型的 TC 值都有明显上升,但是 WBW-BTM 主题模型的聚类效果要优于 W-LDA 主题模型。

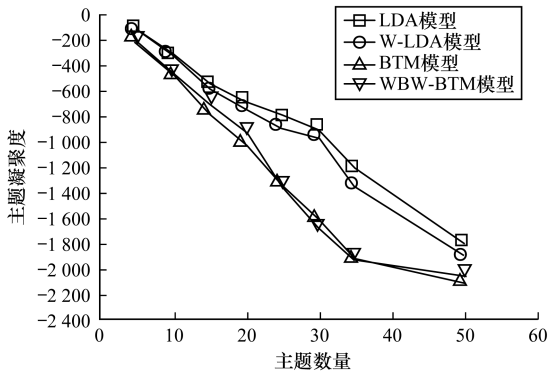

图 5　各主题模型的主题凝聚度对比

4)JS 距离

JS 距离主要用来衡量主题之间的差异,JS 距离值越大,聚类效果越好。从图 6 可以看出,BTM 主题模型与 WBW-BTM 主题模型的聚类效果整体要优于 LDA 主题模型和 W-LDA 主题模型。主要是因为前者采用词共现的方式,所以对短文本处理效果比较好。W-LDA 主题模型与 WBW-BTM 主题模型在不同主题数目上的聚类性能均有一定程度的提升。WBW-BTM 主题模型在主题数为 25 前,相对于 BTM 主题模型取得了不错的聚类效果,但是随着主题数目的增加聚类性能变差。其可能的原因是随着主题数目的增加,基于阈值扩展的双词数量越来越多,所以每个主题都有较好的内聚性,但是主题之间的差异反而变小,W-LDA也是同样的情况。

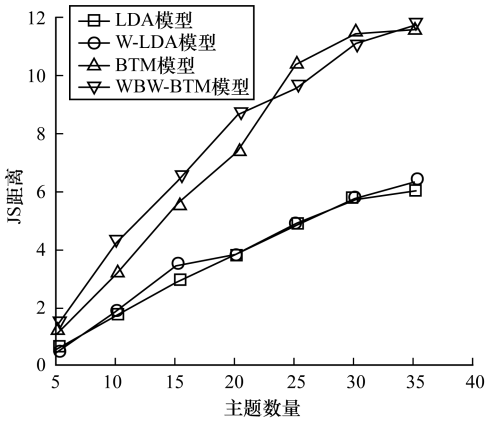

图 6　各主题模型的 JS 距离对比

4　结束语

本文结合词向量的方式,提出一种改进的 BTM 主题模型,对双词的构成融入语义关系,使得不同的语义关系取得的主题聚类效果不同。通过语义距离阈值实验,得到适合的双词语义距离阈值,并将其引入到 BTM 主题模型中,与传统 BTM 主题模型进行对比,具有较好的聚类性能。然而本文主要研究主题模型的产生过程,却忽视了不同主题下主题词的歧义问题,由于每个主题词可能对应不同的主题,因此下一步将利用主题模型对多义词进行词义消歧。

参考文献

[1] 徐戈,王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报,2011,34(8):1423-1436.

[2] DEERWESTER S,DUMAIS S T,FURNAS G W,et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science,1990,41(6):391-407.

[3] HOFMANN T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and development in Information Retrieval. New York,USA:ACM Press,1999:50-57.

[4] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003,3:993-1022.

[5] HONG L,DAVISON B. Empirical study of topic modelingin Twitter[C]//Proceedings of the 1st Workshop on Social Media Analytics. New York,USA:ACM Press,2010:80-88.

[6] SRIDHAR V K R. Unsupervised topic modeling for short texts using distributed representations of words[C]//Proceedings of the Workshop on Vector Space Modeling for Natural Language Processing. New York,USA:ACM Press,2015:192-200.
[7] 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.
[8] 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.
[9] PANG Jianhui, LI Xiangsheng, XIE Haoran, et al. SBTM: topic modeling over short texts [C]// Proceedings of Database Systems for Advanced Applications. Berlin,Germany:Springer,2016:43-56.
[10] 李振兴,王松. 基于卡方特征和 BTM 融合的短文本分类方法[J]. 兰州交通大学学报,2016,35(1):36-41.
[11] 郑诚,吴文岫,代宁. 融合 BTM 主题特征的短文本分类方法[J]. 计算机工程与应用,2016,52(13):95-100.
[12] 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.
[13] 奚雪峰,周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报,2016,42(10):1445-1465.
[14] LU Tingting,HOU Shifeng,CHEN Zhenxiang,et al. An intention-topic model based on verbs clustering and short texts topic mining [C]//Proceedings of 2015 IEEE International Conference on Computer and Information Technology;Ubiquitous Computing and Communications;Dependable,Autonomic and Secure Computing;Pervasive Intelligence and Computing. Washington D. C.,USA:IEEE Press,2015:837-842.
[15] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space [EB/OL]. [2017-12-07]. http://cn. arxiv. org/abs/1301. 3781.
[16] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems,2013,26:3111-3119.
[17] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing semantic coherence in topic models[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2011: 262-272.

编辑 陆燕菲

(上接第 209 页)

[5] AZADJALAL M M,MORADI P,ABDOLLAHPOURI A,et al. A trust-aware recommendation method based on Pareto dominance and confidence concepts[J]. Knowledge-based Systems,2017,116:130-143.
[6] SHAMBOUR Q,HOURANI M,FRAIHAT S. An item-based multi-criteria collaborative filtering algorithm for personalized recommender systems[J]. International Journal of Advanced Computer Science and Applications,2016,7(8):274-279.
[7] ALHAMID M F,RAWASHDEH M,HOSSAIN M A,et al. Towards context-aware media recommendation based on social tagging [J]. Journal of Intelligent Information Systems,2016,46(3):499-516.
[8] 余凯,贾磊,陈雨强,等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展,2013,50(9):1799-1804.
[9] SCHMIDHUBER J. Deep learning in neural networks:an overview[J]. Neural Networks,2015,61:85-117.
[10] SZE V,CHEN Y H,YANG T J,et al. Efficient processing of deep neural networks:a tutorial and survey[J]. Proceedings of the IEEE,2017,105(12):2295-2329.
[11] 田垚,蔡猛,何亮,等. 基于深度神经网络和 Bottleneck 特征的说话人识别系统[J]. 清华大学学报(自然科学版),2016,56(11):1143-1148.
[12] DAHL G E,YU D,DENG L,et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J]. IEEE Transactions on Audio Speech and Language Processing,2012,20(1):30-42.
[13] DAN C,MEIER U,SCHMIDHUBER J. Multi-column deep neural networks for image classification [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2012:3642-3649.
[14] 邹冲,蔡敦波,赵娜,等. 基于 SVM-LeNet 模型融合的行人检测算法[J]. 计算机工程,2017,43(5):169-173.
[15] SATO T,TAKANO Y,MIYASHIRO R,et al. Feature subset selection for logistic regression via mixed integer optimization [J]. Computational Optimization and Applications,2016,64(3):865-880.
[16] PHAISANGITTISAGUL E. An analysis of the regularization between L2 and dropout in single hidden layer neural network [C]//Proceedings of the 7th International Conference on Intelligent Systems,Modelling and Simulation. Washington D.C.,USA:IEEE Press,2016:174-179.

编辑 赵 辉