一种基于 LDA 主题模型的话题发现方法

郭蓝天，李扬，慕德俊，杨涛，李哲

（西北工业大学 自动化学院，陕西 西安 710072）

摘 要： 话题发现是提取热点话题并掌握其演化规律的关键技术之一。针对社交网络中海量短文本信息具有高维性导致主题模型难以处理以及主题分布不均导致主题不明确的问题，提出一种基于 LDA（latent dirichlet allocation）主题模型的 CBOW-LDA 主题建模方法，通过引入基于 CBOW（continuous bag-of-word）模型的词向量化方法对目标语料进行相似词的聚类，能够有效降低 LDA 模型输入文本的维度，并且使主题更明确。通过在真实数据集上计算分析，与现有基于词频权重的词向量化 LDA 方法相比，在相同主题词数情况下困惑度可降低约 3%。

关 键 词： 词向量；LDA 模型；话题发现；困惑度
中图分类号： TP391 **文献标志码：** A **文章编号：** 1000-2758（2016）04-0698-05

为了通过海量的社交网络数据及时的掌握热点话题和舆情的态势变化，需要对话题进行提取、追踪和预测。话题发现是解决该类问题的关键技术之一。LDA（latent dirichlet allocation，隐性狄利克雷分布）主题模型在新闻话题发现与检测方面获得了不错的效果，但由于社交网络文本（如微博客短文本）存在高维性及主题分布不均等问题，加之 LDA 自身的局限性，导致以概率化词汇抽取为基础的 LDA 主题模型在处理社交网络文本方面还存在模型难以降维处理和主题不明确的问题[1-4]。

CBOW 语言模型是 Mikolov 等[2]于 2013 年提出的一种基于类前馈神经网络的语言模型。它能利用文本词汇的上下文信息，通过模型训练将词转化为向量。通过向量空间上的相似度可以分析表示文本语义上的相似度。可作为词向量聚类方法用来寻找相似词汇，进而在有效表达语义信息的同时降低模型处理的维度[4]。

本文研究话题发现问题，通过对现有话题发现常用的 LDA 主题模型的局限性进行分析，提出一种基于 CBOW 语言模型的向量表示方法进行文本词相似性聚类，以聚类结果为基础利用 LDA 主题模型对文本进行隐含主题提取的话题发现方法。

1 相关工作

文献[4]提出一种将 LDA 与 VSM（vector space model，向量空间模型）结合的方法研究微博客话题发现。该方法基于 TF-IDF 的权重词向量，再将 2 种方法结果进行线性加权融合在一起，实现文本间相似度的计算。TF-IDF 向量方法仍然是对词频进行简单的概率统计，易受无用信息干扰。

为了减少代词和介词等无用文本信息对话题抽取模型的干扰，文献[5]提出在微博话题检测过程中，将中文词性标注后输入 LDA 主题模型进行话题抽取。该方法试图通过剔除大量无关词汇，使向量空间的维度降低。

利用 LDA 和基于神经网络语言模型的向量化方法进行文本的特征提取并对比分析。实验结果表明，LDA 直接应用在文本特征表示上的效果不理想，同时也面临着高维度的问题；基于神经网络语言模型的向量化方法应用于文本表示过程中能够带来一定的效果提升。

总结 LDA 模型的局限性主要表现在：

收稿日期： 2016-03-19
基金项目： 国家自然科学基金（61402373、61303224、61403311）与航空科学基金（20155553036、2013ZC53034）资助
作者简介： 郭蓝天（1987—），西北工业大学博士研究生，主要从事数据挖掘及机器学习等研究。

1）由于中文词义多样性,存在很多同义词、近义词、易混淆词等,致使基于概率化的单词抽取方法会存在文本的主题分散及主题混淆等问题。

2）社交网络文本数据量大,主题更新速度快以及训练语料数据维度特别高(通常上万维的向量),使得 LDA 主题模型规模很大,处理效率偏低。

2　基于 CBOW-LDA 的话题发现

2.1　话题发现的基本流程

话题发现的主要步骤为先根据聚类规则挖掘社会网络中的用户群组和抓取文本数据。进行数据预处理后,提取特征和模型表示,将杂乱的非结构化文本转化为结构化数据。然后利用聚类算法对文本主题词进行相似度的计算和聚类,从而找出群组的话题并进行分析,该过程如图 1 所示。

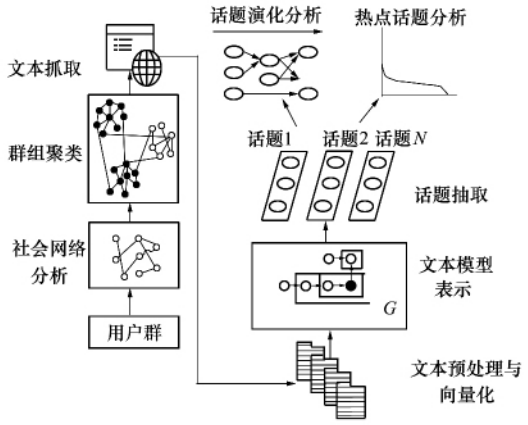

图 1　话题发现的示意图

2.2　CBOW-LDA 算法框架

在话题发现过程中,CBOW-LDA 算法的主要功能是将文本的向量化后进行文本模型表示。文本向量化是指将文本中的单词表示为多维向量的形式,然后输入 LDA 主题模型进行训练,得到文本的模型表示。算法核心思想是在 LDA 主题模型的文本表示基础上,利用词向量进行相似词聚类,这样做能使 LDA 模型的处理维度降低的同时改善主题分散混淆。

2.3　文本向量化

与传统的 one-hot representation 向量表示法不同,CBOW 属于 distributed representation 的词向量表示方式[7],该方法通过引入连续的分布式词表示方法,形成了不同于传统词袋模型连续词袋模型。

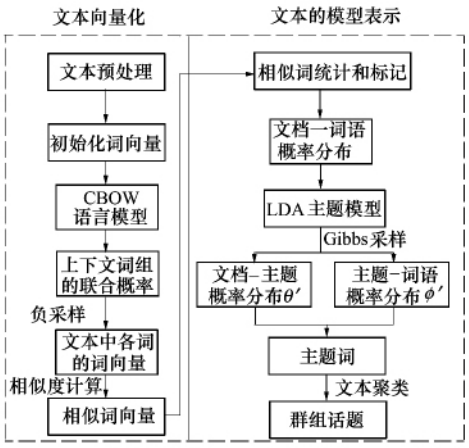

图 2　CBOW-LDA 算法框架

CBOW 模型的主要思想是根据语料中词的上下文信息生成其对应的词向量,并映射到高维空间中后,以词向量在高维空间中的相互关系来计算词与词之间的相似度。具体地,是将语料中的词通过左边的输入层映射到中间的投影层得到词典。

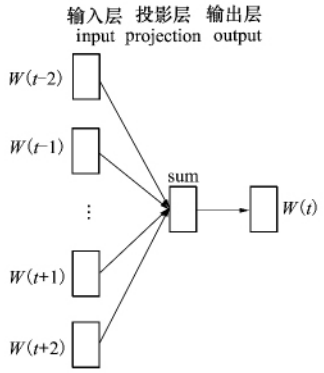

图 3　CBOW 语言模型

设有语料库 C 中的 t 个词汇 $C(t)$ 通过共享投影层,得到对应的唯一位置 $W(t)$,接着通过 $W(t)$ 的上下文信息来预测 $W(t)$。基本训练步骤如下

1）在输入层,通过窗口值 k 限定输入层中的上下文窗口大小,然后在读取窗口中的词 $C(t-k)$, $C(t-k+1)$,…, $C(t+k-1)$, $C(t+k)$,通过 hash 表得到投影层的相应位置 $W(t-k)$, $W(t-k+1)$, …, $W(t+k-1)$, $W(t+k)$,这样就可得到某个词 $W(t)$ 的上下文词汇 Context($W(t)$),因为 CBOW 的模型的目标是在已知当前词 $W(t)$ 的上下文 Context($W(t)$) 的情况下预测当前的词。

2）在中间的投影层,利用对 $W(t)$ 的上下文信

息 Context($W(t)$)进行累加操作。用公式可表达为

$$V(t) = \sum_{t-k}^{t+k} \text{Context}(W(t)) \tag{1}$$

3）从中间的投影层到右边的输出层，利用 $W(t)$ 的上下文 Context($W(t)$)建立条件概率表达式 $P(W(t) \mid \text{Context}(W(t)))$，用来表示生成 $W(t)$ 的向量值。

CBOW 模型的优化目标函数取其对数似然函数

$$L = \sum_{w \in C} \ln P(W(t) \mid \text{Context}(W(t))) \tag{2}$$

直接使用梯度下降求解的运算复杂度非常高。通常使用负采样(negative sampling)的方法进行替换能简化求解计算[8]。根据负采样的原理，$\ln p(w_{t+j} \mid w_t)$ 模型中公式的表示为

$$\ln\sigma(\boldsymbol{v}'^{\mathrm{T}}_{w(t+j)} \cdot \boldsymbol{v}_{w(t)}) + \sum_{k=1}^{K} E_{w_k \sim P_v(w)} \ln(\sigma(-\boldsymbol{v}'^{\mathrm{T}}_{w_k} \cdot \boldsymbol{v}_{w(t)})) \tag{3}$$

$E_{w_k \sim P_v(w)}$ 表示 Huffman 树中上下文不出现某个词的期望值，$P_{v(w)}$ 表示整个语料中词频的分布，W_k 表示该词在 Huffman 树各层中非目标词组的节点向量和。

得到 Huffman 树中路径概率最大的词向量后，通过训练整个文本的词汇得到最终的词向量集。利用 cos 相似度计算词向量之间的相似度，记录相似词的词频和向量，输入 LDA 主题模型进行文本建模。

2.4 文本的模型表示

LDA 主题模型是包含文档－主题－词语的3层贝叶斯模型，其中主题是隐含层[9]。与传统主题模型输入不同，CBOW-LDA 算法中 LDA 主题模型输入的语料是经过相似性聚类的文档－词语的分布，使 LDA 主题模型处理维度降低及主题更明确。

LDA 主题模型采用概率产生模式，将文本表示为主题的混合分布 $p(z)$。LDA 的联合概率公式为

$$p(\theta, z \mid w, \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{4}$$

主题模型生成文本的过程如下：

1）对于主题 z，根据 Dirichlet 分布 Dir(β)得到该主题上的一个单词多项式分布向量 φ；

2）根据 Dirichlet 分布 Dir(α)得到该文档的一个主题分布概率向量 θ；

3）对于该文档 N 个单词中的每个单词 w_i($i \in [1, N]$)。从 θ 的多项式分布 Multi(θ)随机选择一个主题 z，即得到文档－主题的分布。从主题 z 的多项式条件概率分布 Multi(φ)选择一个单词作为 w_i，即得到主题－词语的分布。

利用 Gibbs 抽样简化求解 θ 和 φ 的值。由贝叶斯公式得到后验概率公式如下

$$p(z_i \mid z_{-i}, w, \alpha, \beta) = \frac{p(z_i, z_{-i}, w \mid \alpha, \beta)}{p(z_{-i}, w \mid \alpha, \beta)} \propto \frac{p(z, w \mid \alpha, \beta)}{p(z_{-i}, w \mid \alpha, \beta)} \tag{5}$$

引入 θ 和 φ，并积分可得

$$p(z_i = k \mid z_{-i}, w, \alpha, \beta) \propto \frac{n^k_{-i,m} + \alpha_k}{\sum_{k=1}^{K}(n^k_{-i,m} + \alpha_k)} \cdot \frac{n^w_{-i,k} + \beta_w}{\sum_{w=1}^{V}(n^w_{-i,k} + \beta_w)} \tag{6}$$

$n^k_{-i,m}$ 表示文档 m 中属于该主题 k 的包含词语个数；$n^w_{-i,k}$ 表示 w_i 属于主题 k 的次数；获得每个单词的主题标号 k 后，需要的参数计算公式可表示为

$$\theta_{m,k} = \frac{n^k_m + \alpha}{\sum_{k=1}^{K} n^k_m + k\alpha}, \varphi_{k,w} = \frac{n^w_k + \beta}{\sum_{w=1}^{V} n^w_k + W\beta} \tag{7}$$

式中，V 表示词语的数量；n^w_k 表示词项 w 在主题 k 中出现的次数；n^k_m 表述主题 k 在文档 m 中出现的次数。

对于同一文本而言，由于 CBOW-LDA 算法进行了词向量的相似性聚类，实质是优化了 LDA 主题模型输入的文档－词语分布，以致使求解得到 θ 和 φ 的结果产生了更新，得到了新的词项和主题项。

3 实验及结果分析

3.1 评价指标

评价指标采用文本建模中常用的困惑度(perplexity)来度量，困惑度越小，主题词被选中的概率越大，表明语言模型吻合度越好。其定义如公式(8)所示

$$\text{Perplexity}(W) = \exp\left\{-\frac{\sum_{m=1}^{M} \ln(p(w_m))}{\sum_{m=1}^{M} N_m}\right\} \tag{8}$$

式中，W 为测试集，w_m 为测试集文档 m 中可观测到的单词，$p(w_m)$ 表示模型产生文本 w_m 的概率，N_m 为文档 m 的单词数。

3.2　语料获取

本文采集了新浪微博上有关 IT 互联网行业高管以及政府机关人员的微博语料,以及这个群组内的关注情况数据。数据集涉及约 6 000 名用户在 2015 年 3 月至 2015 年 4 月这 30 天内发表的约 43 万条微博。以这期间的某一天为例,抓取到 14 536 条微博,分析其包含词数多达 233 296。

3.3　实验步骤与参数设置

原始微博数据包含诸多无用信息,使用“Jieba”分词工具进行分词和过滤,得到的词典库包含 46 516个词;然后将其输入到 CBOW-LDA 模型使用 Word2vec 0.8(2015 年 7 月)开发向量程序进行向量化处理。

Word2vec 的参数设置如表 1 所示,其中 *Cbow* = 1 表示训练使用的是 CBOW 模型,*Hs* = 0 表示使用的是负采样简化求解计算。词向量聚类中,相似度的阈值设为 0.75。

表 1　Word2vec 参数设置

参数	含义	取值
Size	词向量维数	50
Window	上下文窗口大小	5
Sample	高频词亚采样阈值	0
Cbow	是否使用 *Cbow* 算法	1
Hs	是否使用层次 *Softmax*	0

CBOW-LDA 算法中的文本模型表示过程选用 *lda* 1.0.3 的 *Python* 工具集作为 LDA 的实现工具。该工具处理速度较快,适合分析大规模语料。已有文献大多数将模型参数中的 α 和 β 设置为:$\alpha = 50/K$,$\beta = 0.01$。K 为隐含主题词数,可根据文本规模和应用场景做相应调整。*Gibbs* 抽样迭代次数为 500 次。

所有参数设定好之后,程序便开始利用 Gibbs 抽样算法对模型求解。程序运行完成后可得到参数 θ 和 φ 的值,通过分析可得文本的主题词,进而总结语料的话题。

3.4　结果分析

在相同的参数设置和语料下,通过计算困惑度来度量模型的处理效果。对比方案参照文献[4]中基于 TF-IDF 的权重词向量 LDA 方法(本文简写为 TF-LDA),2 种方法困惑度随隐主题数目的变化情况如表 2 所示。

表 2　2 种方法困惑度比较

主题数	20	25	30	35	40	45	50	55	60
CBOW-LDA	151	138	125	116	112	106	100	96	92
TF-LDA	155	143	131	120	111	109	103	100	95

可以看出,随着主题数不断增加,二者困惑度都相应降低。将二者差值的百分数取平均值得出,在该 20 ~ 60 个主题数的范围内,CBOW-LDA 方法困惑度降低了约 3%。

在相同参数下,将主题数 K 设为 30,观察困惑度随迭代次数的变化情况如图 4 所示。文本实验主题模型求解的迭代次数为 500 次,为了便于展示,仅截取迭代 300 次的数据。

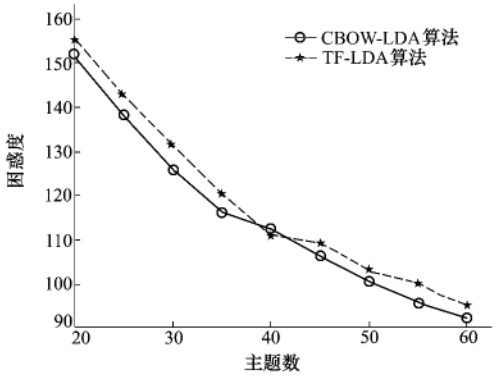

图 4　困惑度随主题数目的变化情况

从结果可以看出本文的 CBOW-LDA 方法虽然采用比 TF-LDA 更为复杂的向量化方法,但是收敛速度并没有随之减慢,表现出较好的响应能力。

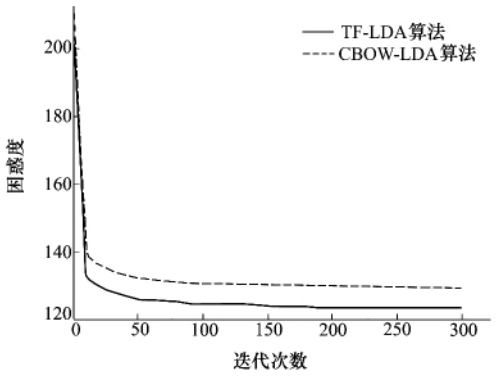

图 5　困惑度随迭代次数的变化情况

4　结　论

文本针对社交网络中短文本信息的特点,提出

将文本深度表示模型的词向量化方法与LDA主题模型结合进行话题发现的方法。通过对LDA模型的输入进行相似词的聚类,使得话题抽取模糊度更低,话题含义的表达更加明确。今后的研究工作将进一步深度研究和优化模型,加强话题发现的效果。

参考文献:

[1] Cheng Xueqi, Yan Xiaohui, Lan Yanyan, et al. BTM: Topic Modeling Over Short Texts [J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(12): 2928-2941

[2] Mikolow Tomas, Yih Wentau Scott, Zweiq Geoffery. Linguistic Requlrities in Contrmcous Space Word Representations [C] // Proceedings of the 12nd Conference of the North Anerican Chapter of the Association for Computational Linguistics, Atlanta, USA: NAACL, 2013

[3] Dermouche M, Velcin J, Khouas L, et al. A Joint Model for Topic-Sentiment Evolution Over Time [C] // Proceedings of 14th IEEE International Conference on Data Mining. Shenzhen, China, 2014

[4] Huang Bo, Yang Yan, Mahmood Amjad, et al. Microblog Topic Detection Based on LDA Model and Single-Pass Clustering [C] // Proceedings of 7th International Conference on Rough Sets and Current Trends in Computing. Chengdu, China, 2012

[5] Darling M William, Song Fei. Probabilistic Topic and Syntax Modeling with Part-of-Speech LDA [J]. ArXiv:1303.2826, 2013

[6] Bai Xue, Chen Fu, Zhan Shaobin. A New Clustering Model Based on Word2vec Mining on Sina Weibo Users´ Tags [J]. International Journal of Grid Distribution Computing, 2014, 7(3): 41-48

[7] Zhou Xinjie, Wan Xiaojun, Xiao Jianguo. Repre-Sntation Learning for Aspect Category Detection in Online Reviews [C]. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas, USA, 2015

[8] Mikolov Tomas, Sutskever Hya. Distributed Representutions of Words and Phrases and Their Compositionality [C] // Proceedings of the Ilth Newral Information Processing Systems Conference Lake Tahoe, USA: NIPS, 2013

[9] Cao Ziqiang, Li Sujian, Liu Yang, et al. A Novel Neural Topic Model and Its Supervised Extension [C] // Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas, USA, 2015

A LDA Model Based Topic Detection Method

Guo Lantian, Li Yang, Mu Dejun, Yang Tao, Li Zhe

(School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Topic Detection is one of the most important techniques in hot topic extraction and evolution tracking. Due to the high dimensionality problem which hinders processing efficiency and topics mal-distribution problem which makes topics unclear, it is difficult to detect topics from a large number of short texts in social network. To address these challenges, we proposed a new LDA (Latent Dirichlet Allocation) model based topic detection method called CBOW-LDA topic modeling method. It utilizes a CBOW(Continuous Bag-of-Word) method to cluster the words, which generate word vectors and clustering by vectors similarity. This method decreases the dimensions of LDA output, and makes topic more clearly. Through the analysis of topic perplexity in the real-world dataset, it is obvious that topics detected by our method has a lower perplexity, comparing with word frequency weighing based vectors. In a condition of same number of topic words, perplexity is reduced by about 3%.

Keywords: word vectors; LDA model; topic detection; perplexity