面向短文本的神经网络聚类算法研究

孙昭颖 刘功申

(上海交通大学电子信息与电气工程学院 上海 200240)

摘 要 词汇个数少、描述信息弱的缺陷,导致短文本具有维度高、特征稀疏和噪声干扰等特点。现有的众多聚类算法在对大规模短文本进行聚类时,存在精度较低和效率低下的问题。针对该问题,提出一种基于深度学习卷积神经网络的短文本聚类算法。所提算法以大规模语料为基础,利用 word2vec 模型学习短文本中词语之间潜在的语义关联,用多维向量表示单个词语,进而将短文本也表示成多维的原始向量形式;结合深度学习卷积神经网络,对稀疏高维的原始向量进行特征提取,以此得到特征更为集中、有效的低维文本向量;最后,利用传统的聚类算法对短文本进行聚类。实验结果表明,所提聚类方法对文本向量的降维是可行、有效的,并且取得了 F 值达到 75% 以上的文本聚类效果。

关键词 短文本,文本聚类,深度学习,卷积神经网络,word2vec

中图法分类号 TP183 **文献标识码** A

Research on Neural Network Clustering Algorithm for Short Text

SUN Zhao-ying LIU Gong-shen

(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract Short text has a small number of vocabularies and weak description of information, resulting in the characteristics of high dimensionality, sparse features and noise interference. The existing clustering algorithms have low accuracy and efficiency for the large-scale short text. A short text clustering algorithm based on deep learning convolution neural network was proposed to solve this problem. The proposed clustering algorithm uses the word2vec model to learn the potential semantic association between words in the short text, and the multidimensional vector to represent the single word based on the large-scale corpus, and then the short text is also expressed as the multidimensional original vector form. Using convolution neural network, the feature vector is extracted from the original vector of sparse and high dimension to the low-dimensional text vector with more effective characteristics. Finally, the traditional clustering algorithm is used to cluster the short text. The proposed clustering method is feasible and effective for the reduction of text vector, and has achieved good short text clustering effect with F-measure of over 75%.

Keywords Short text, Document clustering, Deep learning, Convolution neural network, Word2vec

1 引言

互联网的高速发展使得人们获取信息的方式越来越依赖于网络,同时也催生出海量的诸如微博、书评影评、新闻标题、摘要等各种形式的短文本,而这些短文本也成为了广大网民实时了解与讨论新闻事件、展示自我以及人际交往的重要工具之一[1]。社交网络平台上充斥着大量与当前社会热点事件紧密相关的海量信息,这些与国内外政治、经济和文化紧密相关的海量社交网络数据具有重要的研究价值[1-3],其中包含的大量有价值的隐含信息对人们的日常生活产生了重要的影响。对海量的网络信息进行高效的聚类,可以有效地帮助政府、企业等相关决策人员更好、更快地了解社会热点事件的发展态势,这对于舆情疏导、危机公关、产品营销等都有着重要的应用意义。因此,人们迫切需要对网络上的海量短文本信息进行有效聚类,以便快速获取有用的信息。

与长文本相比,短文本词汇个数少且描述信息弱,具有稀疏性和不规范性等特点。基于向量空间模型(Vector Space Model, VSM)[4-8]的传统文本聚类方法应用网络中的海量短文本数据时,面临着复杂度高、特征稀疏度高、噪声干扰大等[9]挑战。文献[10-11]提出了基于语义进行短文本聚类的方法,而文献[12-17]则通过外部链接[12-15]或者知识库[15-17]来扩展词语的语义。基于扩充后的语义信息重构短文本的特征空间可以提高短文本之间相似性度量的准确性,但中文词语规模相对较大,相似词语纷繁多样,从而导致上述方法的效率较低。此外,网络短文本往往存在用语不规范的问题,因此,通过上述方法对词语进行语义扩展存在效率较低、准确性较差的问题。文献[18]使用人工免疫规划网络的方法实现了对短文本的聚类,但是该方法搜索空间较大,只能处理小规模短数据集,在大规模语料上的处理性能还达不到实际应用的要求。

这些算法在将词语和文本转化为向量表示时,不管是基于词频还是语义,都存在着一定程度的主观选择和判断,而这

本文受国家自然科学基金项目(61472248,61431008)资助。

孙昭颖(1993-),女,硕士生,主要研究方向为机器学习、深度学习,E-mail:sunzy93@163.com(通信作者);刘功申 男,副教授,主要研究方向为内容安全、社交网络,E-mail:lgshen@sjtu.edu.cn。

种主观行为的介入很难完全保留词语和文本之间的相似性、差异性和相关性，从而丢失了一部分重要的有效信息。而在特征提取方面具有显著优势的深度神经网络目前主要应用于短文本分类领域[19-21]，在短文本聚类过程中使用深度神经网络的算法并不多见。

针对短文本聚类面临的特征高维、稀疏以及文本规模较大等问题，本文从降低短文本特征维数的角度入手，提出一种基于深度学习卷积神经网络的短文本聚类方法。这种方法以大规模语料库为基础，使用 word2vec 模型训练语料，学习词语之间的语义关系，并用词向量形式表示词语，进而将短文本转化为稀疏的原始向量形式；然后，利用深度学习的卷积神经网络对原始向量进行特征学习，以此构造维数较低的文本向量；最后，利用 k-means 方法实现对短文本的聚类，并通过实验表明了本文所提方法的有效性。

2　短文本神经聚类过程

2.1　短文本的向量化

2.1.1　短文本预处理

本文的中文文本的预处理过程包括分词、去除标点符号和去除停用词这 3 个步骤。分词过程一般指利用分词软件将连续的中文文本切割成若干个分离的词条，并将得到的词条作为后续提取文本特征的基础。由于自然语言中不断有新词涌现，已有的分词软件很难准确地识别出文本中所有的新词，同时为便于利用维基百科对词向量模型进行训练，本文在进行分词处理时对分词软件进行改进，将维基百科的概念词条加入词库。

对短文本进行预处理时，将标点符号去除，以便于将短文本向量化，让短文本的特征表示尽可能地只关注短文本的词汇构成和语义本身，而不会受到标点符号等相关性较低的其他因素的影响。

中文文本的处理过程通常会包含去除停用词，这是因为中文文本中存在着大量的高频但无具体含义的词语。用中文停用词表过滤这些噪音干扰，就能够只保留汉语句子中的所谓核心部分，可以在一定程度上降低特征维数，提高自然语言处理的效率和效果。

2.1.2　词语表示

将词语映射到一个新的空间中，并以多维的连续实数向量进行表示，该过程叫做“Word Represention”或“Word Embedding”。21 世纪以来，人们逐渐从原来的词向量稀疏表示法过渡到如今的低维空间中的密集表示。在解决实际问题时，使用稀疏表示法的过程中经常会遇到维数灾难、语义信息无法表示、无法揭示词语之间的潜在联系的问题。而采用低维空间表示法，不仅解决了维数灾难问题，而且还挖掘了词语之间潜在的关联属性，从而提高了向量语义上的准确度。

word2vec 通过神经网络机器学习算法来训练 N-gram 语言模型，并可以在训练过程中得到词语所对应的词向量这一附属产物[22]。

word2vec 中有两个重要的模型——CBOW 模型（Continuous Bag-of-Words Model）和 Skip-gram 模型（Skip-gram Model），其中，CBOW 利用词语的上下文来预测词语，而 Skip-gram 则利用词语来预测它的上下文。CBOW 模型如图 1 所示。

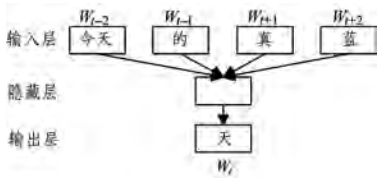

图 1　CBOW 模型示意图

如图 2 所示为 CBOW 模型的神经网络训练模型。其中，最下方表示的是待预测词语 w_t 周围的 $2n$（图中 $n=2$）个词语：$w_{t-n},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+n}$；$C(w_{t-n})$ 为词语 w_{t-n} 对应的词向量，利用此模型可以预测出词语 w_t，输出为 w_t 所有可能对应词语出现的概率，概率最高的词语即为 w_t 最有可能的选项。

网络的第一层为输入层：将 $2n$ 个词语向量的首尾拼接，组成 $2n\times m$ 维的长向量，其中 m 是初始规定的词向量维度。网络的第二层为隐藏层：与普通神经网络类似，偏置项可以随机初始化，激活函数选用 tanh。网络的第三层为输出层：使用 softmax 函数将输出值的概率归一化。

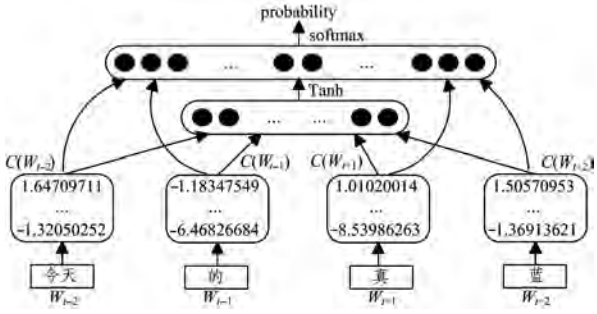

图 2　CBOW 神经网络训练模型

2.1.3　短文本表示

根据对实验语料的分析，获取语料中所有短文本的最大长度，然后把每个短文本填充到最大的短文本长度中，使得每个短文本都包含有相同的词语数量，且相同的长度有利于进行高效的批处理操作。根据 word2vec，短文本中的每一个词语都转化为相同形状的 $1\times k$ 向量，其中 k 为词向量的维数，将短文本中词语的词向量按照先后顺序进行排列，则得到一个与短文本长度和词向量长度相关的 $n\times k$ 向量，且所有短文本对应的向量形状是一致的。这一向量充分保留了短文本中的语义信息和位置信息，因此本文用这一向量来作为短文本的原始特征向量。

2.2　短文本神经网络的构建

如图 3 所示，由于本文用卷积神经网络来处理文本对象，而不是常见的图像对象，因此对传统的卷积神经网络进行了一定的调整[23]。

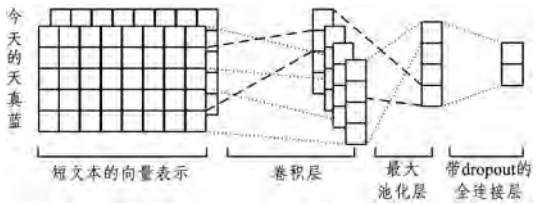

图 3　文本聚类卷积神经网络示意图

首先，本文所用卷积神经网络的输入是短文本对应的由词向量堆叠而成的词矩阵，而不是图像对应的像素矩阵。

在卷积层中，用于图像处理的卷积核通常是正方形的，如

5 * 5,因此该卷积核在整张图像上沿宽和高逐步移动来进行卷积操作。但是,在自然语言处理中,网络的输入是一个词矩阵,如 n 个词语,若每个词语用 128 维的词向量表示,则网络的输入就是 $n*200$ 的矩阵。为了保证词语作为语言处理的最小粒度,本文构建的卷积神经网络中的卷积核只能在高度上进行滑动,而其宽度和词向量的维度一致,即每次窗口滑动过的位置都是完整的单词,不会将几个单词的一部分词向量进行卷积。

由于卷积核与词向量的长度一致,一个卷积核对应一个短文本,卷积后得到的结果是一个向量,而其在池化后得到的就是一个标量。这一点与传统的图像卷积非常不同。

由于池化后得到的只是一个标量,本文会使用多个不同窗口大小的卷积核对输入的词矩阵进行卷积。一个卷积核得到的只是一个标量,将相同窗口大小卷积出来的若干个标量组合在一起,就可以得到这个窗口大小下的特征向量。最后,再将所有窗口大小下的特征向量组合成一个完整的特征向量,作为短文本最终的向量化表示以及聚类算法的输入。

2.3 短文本聚类

k-means 算法[24]最早由 MacQueen 提出。该算法具有简单且收敛速度快等特点,是数据挖掘及知识发现领域中的一种重要的聚类方法。

k-means 算法是一种基于划分的聚类方法。其基本思想为:对于给定的聚类数目 k,首先随机选择 k 个文本作为初始的类中心,然后将每个文本划进与其最相似的类别中,最后重新计算每个类别的中心。不断迭代以上过程,直到目标函数收敛。在文本聚类中采用的目标函数一般为:

$$S(C)=\sum_{i=1}^{k}s(C_i)=\sum_{i=1}^{k}\sum_{d\in C_i}d^{\mathrm{T}}H(C_i) \tag{1}$$

其中,$S(C_i)$为对应簇 C_i 内部聚类的相似度。

$$s(C_i)=\sum_{d\in C_i}d^{\mathrm{T}}H(C_i)=\|F(C_i)\| \tag{2}$$

式(1)的目标函数是基于余弦相似度的。如果对文本 d 及中心进行了归一化处理,则采用欧氏距离和余弦距离进行相似度度量的效果是等同的。

算法输入为:文本集合 D;输出为:簇的集合 C;定义 C 的复合向量 $F(C_i)=\sum_{d\in C_i}d$,集合 C 的中心 $H_i=H(C_i)=\frac{F(C_i)}{\|F(C_i)\|}$;记第 t 次迭代产生的簇的集合为 $C^{(t)}$,$t=0,1,2,\cdots$。下面给出传统的 k-means 聚类算法的流程:

1)随机选择类的初始中心向量 $H_1^{(0)},H_2^{(0)},\cdots,H_k^{(0)}$,初始迭代次数 $t=0$;

2)对每个文本向量 $d\in D$ 与每个中心向量 $H_j^{(t)}$ 进行比较,并根据相似程度将其分配到最相近的一类中,即:$j^*=\arg\max_j d^{\mathrm{T}}H_j^{(t)}$,$1\leqslant j\leqslant k$,$t$ 为迭代次数;

3)由步骤 2)得到新的簇分区集合 $C^{(t+1)}=nextKM(C^{(t)})$,并计算新的中心向量 $H_j^{(t+1)}$,$1\leqslant j\leqslant k$;

4)如果 $S(C^{(t+1)})-S(C^{(t)})>\varepsilon(\varepsilon>0$,为判断终止的阈值),则 $t=t+1$,转到步骤 2),否则终止算法,并输出聚类最终簇的集合 C^*。

针对 k-means 算法容易陷入局部最值的缺点,本文对 k-means 算法进行二次分区调整,调整完毕之后,重新进行 k-means 启发式的搜索过程,从而实现对 k-means 算法的改进;此外,由于 k-means 算法的聚类结果受初始类中心的影响,因此本文对同一输入进行多次 k-means 聚类并取均值,以降低初始类中心对聚类结果的影响。

3 实验分析

本文使用 Python2.7 中的 gensim 工具包(Google 的 word2vec 工具源码 python 封装)进行词向量的训练,选用大规模维基中文语料进行学习。另外,采用中科院的 ICTCLAS 中文分词系统对文本进行分词处理。

本文实验使用的语料库来自 2012 年 6 月—2014 年 7 月期间搜狐新闻主页上的新闻数据,选择具有代表性的 6 个类别共 484017 篇短文本的标题作为原始的实验数据,具体的类别构成如表 1 所列。

表 1 实验数据集

类别	样本个数
汽车	138574
IT	199871
运动	44536
健康	23409
娱乐	50138
财经	27489

由于不同类别的样本个数相差较大,若直接用于实验,结果会存在较大的误差。因此,本文从每类样本中选取 20000 个具体样本用于实验分析。

本文共做了 4 组实验,文本数分别为 30000,60000,90000,120000,每次从各个类别中随机选取等量的文本数。

本文对海量短文本聚类效果的评价采用聚类中最常用的评价指标:F-measure。针对所有自然类 K_i 和聚类获得的簇 C_j,其计算方式如下:

$$p(K_i,C_j)=\frac{n_{ij}}{|C_j|} \tag{3}$$

$$r(K_i,C_j)=\frac{n_{ij}}{|K_i|} \tag{4}$$

$$F(K_i,C_j)=\frac{2p(K_i,C_j)\times r(K_i,C_j)}{p(K_i,C_j)+r(K_i,C_j)} \tag{5}$$

其中,n_{ij} 为自然类 K_i 中属于簇 C_j 的短文本个数,$|C_j|$ 为被划分到簇 C_j 中的短文本个数,$|K_i|$ 为自然类 K_i 中短文本的个数。

因此,对于一个聚类 C_j,其任意一个自然类 K_i 的 *F-measure* 值为其最大 $F(K_i,C_j)$ 值。整体聚类 C 的平均 *F-measure* 值的计算如下:

$$F(C)=\sum_{K_i\in K}\frac{|K_i|}{|D|}\max_{C_j\in C}F(K_i,C_j) \tag{6}$$

其中,$|D|$ 为所有测试文本的个数。

为了选取本文算法的最佳参数组合,选取不同的参数组合在同一数据集上做了 4 组实验,具体的实验结果如表 2 所列,其中 *win_size* 表示卷积神经网络中的卷积核窗口大小,*num* 表示每种窗口大小对应的卷积核的数目。

从表 2 中可以看出,不管短文本数量是多少,本文提出的神经聚类算法的聚类效果始终明显好于传统的 k-means 算法,神经聚类算法得到的 *F-measure* 均高于 0.7,而传统的 k-means 算法的 *F-measure* 则只略高于 0.5。此外,随着短文本数量的增加,神经聚类算法的聚类效果在整体上表现出提高的态势,而传统 k-means 算法的聚类效果却与短文本数量无太大关系,短文本数量的增加并不能有效改善其聚类效果。

表 2　神经聚类算法

文本数	*win_size*=3,4,5		*win_size*=5,5,5		*win_size*=4,5,6		传统 k-means 算法
	num=128	*num*=150	*num*=128	*num*=150	*num*=128	*num*=150	
30000	0.707	0.751	0.755	0.753	**0.756**	0.745	0.515
60000	0.737	0.743	0.742	**0.792**	0.753	0.757	0.515
90000	0.754	0.763	0.76	**0.776**	0.771	0.775	0.508
120000	0.739	0.749	**0.770**	0.757	0.763	0.760	0.518
平均值	0.734	0.751	0.757	**0.770**	0.761	0.76	0.514

观察神经聚类算法在不同参数下的聚类效果可以发现，卷积核数量的增加能够改善算法的聚类效果，但是由于资源的限制，卷积核的数量不可能无限增加，因此应该在充分利用现有资源的基础上尽可能找到其他参数的更优值。卷积核窗口的大小对聚类效果同样会产生影响，但是并未呈现单调的相关性，就实验结果来看，窗口大小为 5,5,5 的神经聚类算法在多数情况下都能表现出最佳的聚类效果。

结束语　本文针对短文本聚类过程中面临的特征高维与稀疏导致的聚类不准确、文本与词汇规模过大导致的时空复杂度过高等问题，提出了基于深度学习的短文本神经聚类算法。该算法首先利用词向量实现短文本的原始向量表示，然后利用卷积神经网络进行特征提取，从而实现短文本特征维数的减小和计算复杂度的降低，以及聚类效果的提高。利用 120000 万条搜狐新闻标题构成的短文本语料库验证了此算法的效果和鲁棒性。实验结果表明，与传统 k-means 聚类算法相比，该算法海量短文本的聚类效果有了明显提升，其能够更准确、有效地对海量短文本进行聚类。下一步将对此聚类算法进行改进，并借鉴分类算法中的反馈思想使聚类结果能够影响用于特征提取的卷积神经网络中的各项参数，从而实现整个聚类算法的自适应和自学习，有效提升短文本聚类的效果。

参 考 文 献

[1] 丁兆云，贾焰，周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展，2014，51(4)：691-706.

[2] YANG X，GHOTING A，RUAN Y，et al. A framework for summarizing and analyzing twitter feeds[C]// 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2012：370-378.

[3] ZHANG X，ZHU S，LIANG W. Detecting spam and promoting campaigns in the twitter social network[C]// 2012 IEEE 12th International Conference on Data Mining (ICDM). IEEE，2012：1194-1199.

[4] LIN D. An information-theoretic definition of similarity[C]// ICML. 1998：296-304.

[5] SCHÜTZE H，SILVERSTEIN C. Projections for efficient document clustering[C]// International ACM Sigir Conference on Research & Development in Information Retrieval. ACM，1997：74-81.

[6] RAMAGE D，HEYMANN P，MANNING C D，et al. Clustering the tagged Web[C]// Second ACM International Conference on Web Search and Data Mining. ACM，2009：54-63.

[7] FREEMAN R，YIN H. Self-organising maps for hierarchical tree view document clustering using contextual information[C]// International Conference on Intelligent Data Engineering and Automated Learning. Springer Berlin Heidelberg，2002：123-128.

[8] 索红光，王玉伟. 一种用于文本聚类的改进 k-means 算法[J]. 山东大学学报(理学版)，2006，43(1)：60-64.

[9] 刘金岭. 基于主题的中文短信文本分类研究[J]. 计算机工程，2010，36(4)：30-32.

[10] 杨震，王来涛，赖英旭. 基于改进语义距离的网络评论聚类研究[J]. 软件学报，2014，25(12)：2777-2789.

[11] 张群，王红军，王伦文. 一种结合上下文语义的短文本聚类算法[J]. 计算机科学，2016，43(s2)：443-446.

[12] SAHAMI M，HEILMAN T D. A web-based kernel function for measuring the similarity of short text snippets[C]// International Conference on World Wide Web. ACM，2006：377-386.

[13] BOLLEGALA D，MATSUO Y，ISHIZUKA M. Measuring semantic similarity between words using web search engines[C]// WWW 2007. 2007：757-766.

[14] BANERJEE S，RAMANATHAN K，GUPTA A. Clustering short texts using wikipedia[C]// 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM，2007：787-788.

[15] HU X，SUN N，ZHANG C，et al. Exploiting internal and external semantics for the clustering of short texts using world knowledge[C]// 18th ACM Conference on Information and Knowledge Management. ACM，2009：919-928.

[16] TIAN Y，LI H，CAI Q，et al. Measuring the similarity of short texts by word similarity and tree kernels[C]// 2010 IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT). IEEE，2010：363-366.

[17] CHEN X，ZHANG Y，CAO L，et al. An Improved Feature Selection Method for Chinese Short Texts Clustering Based on HowNet[M]// Computer Engineering and Networking. Springer International Publishing，2014：635-642.

[18] 行小帅，潘进，焦李成. 基于免疫规划的 K—means 聚类算法[J]. 计算机学报，2003，26(5)：605-610.

[19] 卢玲，杨武，杨有俊，等. 结合语义扩展和卷积神经网络的中文短文本分类方法[J]. 计算机应用，2017(12)：3498-3503.

[20] 张绮琦，张树群，雷兆宜. 基于改进的卷积神经网络的中文情感分类[J]. 计算机工程与应用，2017，53(22)：111-115.

[21] 郭东亮，刘小明，郑秋生. 基于卷积神经网络的互联网短文本分类方法[J]. 计算机与现代化，2017(4)：78-81.

[22] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems. 2013：3111-3119.

[23] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv：1408.5882，2014.

[24] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1966：281-297.