文章编号:1006-2475(2022)03-0048-05

基于 RoBERTa-WWM 和 HDBSCAN 的文本聚类算法

刘　锟[1],曾　曦[1,2],邱梓珩[2],陈周国[1,2]

(1. 中国电子科技集团公司第三十研究所,四川 成都 610000; 2. 深圳市网联安瑞网络科技有限公司,广东 深圳 518000)

摘要:在大数据环境下,从海量的互联网数据中获取热点话题是研究当前互联网中民意民情的基础,其中文本聚类是得到热点话题最常用的方法之一,可以分为文本向量化表示和聚类 2 个步骤。然而在文本向量化表示任务中,传统的文本表示模型无法准确表示新闻、帖文等文本的上下文语境信息。在聚类任务中,最常使用的是 K-Means 算法和 DBSCAN 算法,但是它们对数据的聚类方式与实际中话题数据的分布不符,这使得现有的文本聚类算法在实际的互联网环境中应用效果很差。本文根据互联网中话题的数据分布情况,提出一种基于 RoBERTa-WWM 和 HDBSCAN 的文本聚类算法。首先利用预训练语言模型 RoBERTa-WWM 得到每一篇文本的文本向量,其次利用 t-SNE 算法对高维文本向量进行降维,最后利用基于层次的密度聚类算法的 HDBSCAN 算法对低维的文本向量进行聚类。实验结果表明提出的算法相较于现有的文本聚类算法,在含有噪声数据且分布不均衡的数据集上,聚类效果有很大的提升。

关键词:文本聚类; 预训练语言模型; 可视化降维; 密度聚类

中图分类号:TP391.1　　**文献标志码**:A　　**DOI**: 10.3969/j.issn.1006-2475.2022.03.009

Text Clustering Algorithm Based on RoBERTa-WWM and HDBSCAN

LIU kun[1], ZENG Xi[1,2], QIU Zi-heng[2], CHEN Zhou-guo[1,2]

(1. The 30th Research Institute of China Electronics Tecnnolpgy Group Corporation, Chengdu 610000, China;
2. Shenzhen CyberAray Network Technology Co., LTD., Shenzhen 518000, China)

Abstract: In the big data environment, obtaining hot topics from massive Internet data is the basis for studying public opinion and sentiments in the current Internet. Among them, text clustering is one of the most common methods to get hot topics, which can be divided into two steps: text vectorization representation and clustering. However, in the task of vectorized text representation, the traditional text representation model cannot accurately represent the contextual information of texts such as news and posts. In the clustering task, the K-Means algorithm and DBSCAN algorithm are most commonly used, but their clustering method is not consistent with the actual distribution of topic data, which makes the existing text clustering algorithms very poorly applied in the actual Internet environment. Therefore, this paper proposes a text clustering algorithm based on RoBERTa-WWM and HDBSCAN according to the data distribution of topics in the Internet. Firstly, the pre-trained language model RoBERTa-WWM is used to obtain the text vector of each text. Secondly, the t-SNE algorithm is used to reduce the dimension of the high-dimensional text vector. Finally, the HDBSCAN algorithm based on hierarchical density clustering algorithm is used to cluster the low-dimensional text vector. The experimental results show that compared with the existing text clustering algorithms, the proposed algorithm has a great improvement in the clustering effect on data sets that contain noisy data and are unevenly distributed.

Key words: text clustering; pre-training language model; visual dimensionality reduction; density clustering

0　引　言

随着互联网的普及,我国的网民规模不断扩大,各种信息呈指数级增长。如何从海量的稀疏数据中获取到有价值的信息和知识是摆在研究者面前的一道难题[1]。文本聚类作为自然语言处理领域里一种能够从海量的文本数据中快速发现热点话题的方法,近些年得到了学者们的广泛研究。

根据以往的研究,文本聚类分为文本向量化表示和聚类 2 个主要步骤[2],其中传统的用于文本向量化表示的模型有词袋模型、Word2Vec 模型、Doc2Vec 模型[3],用于聚类的算法有基于划分的聚类算法、基于层次的聚类算法、基于密度的聚类算法等[4]。现有的研究大多也是基于这 2 个步骤进行改进,例如李志强等[5]采用词袋模型和改进的 K-Means 算法来研究短文本聚类;毛郁欣等[6]基于 Word2Vec 模型和 K-Means 算法来做信息技术文档的聚类研究;吴德平等[7]基于 Word2Vec 词嵌入模型和 K-Means 聚类算法对安全事故的文本案例进行分类研究;阮光册等[8]基于 Doc2Vec 做期刊论文的热点选题识别;贾

收稿日期:2021-08-27; **修回日期**:2021-10-09

基金项目:国家自然科学基金资助项目(61803352)

作者简介:刘锟(1996—),男,甘肃平凉人,硕士研究生,研究方向:自然语言处理,E-mail: 1368603690@qq.com; 曾曦(1969—),女,研究员级高级工程师,硕士,研究方向:网络安全与安全防御。

君霞等[9]基于 Doc2Vec 模型和卷积神经网络模型 CNN 对新闻文本数据做了聚类研究,相较于传统的模型,提高了准确率。但是传统的文本表示方法无法反映整篇文本的上下文语境信息,导致文本聚类的准确性较差。近些年,随着预训练语言模型 BERT[10]的产生并在各项 NLP 任务中都取得了很好的效果,一些研究者开始使用基于 BERT 的语言预训练模型来提高文本向量化表示的准确性[11]。例如曹凤仙[12]使用 BERT 模型和 K-Means 模型对市长公开电话的文本进行聚类研究,得到民众的诉求主题分布;朱良齐等[13]融合 BERT 和自编码网络来做文本表示,再利用 K-Means 算法做文本聚类,有效地提高了文本聚类的准确性。除此之外,由于基于划分聚类的 K-Means 算法需要预先指定类别数并且在含有噪声数据时聚类效果较差,而基于密度的聚类算法不需要预先指定簇数并且对噪声数据不敏感,近些年来得到了越来越广泛的使用,例如邹艳春[14]基于 DBSCAN 算法结合文本相似度做文本聚类,曹旭友等[15]基于 BERT + ATT 和 DBSCAN 来做专利文本分析,得到了很好的效果,蔡岳等[16]提出了一种基于簇关系树的改进 DBSCAN 算法对互联网中的文本做聚类研究,但是 DBSCAN 算法在数据的密度分布不均匀时聚类效果较差,而且在实际的应用中对输入参数异常敏感[17],而 HDBSCAN 算法相较于 DBSCAN 算法不需要复杂的调参过程,同时可以适应多密度的数据分布因此,本文针对现有文本聚类算法存在的问题,提出一种基于 RoBERTa-WWM + HDBSCAN 的文本聚类算法,从文本向量化表示和聚类 2 个方面来同时提升文本聚类的效果。

1　相关工作

1.1　RoBERTa-WWM 表示文本

RoBERTa(A Robustly Optimized BERT Pretraining Approach)[18]是在 BERT 模型的基础上,去除 NSP 任务,使用更大规模的数据集和更大的 batch-size 再次训练得到的预训练语言模型。虽然和 BERT 模型相比,RoBERTa 模型在多个 NLP 任务上的表现更好,但是原始的 RoBERTa 模型并不能很好地应用在中文语言环境下,因此哈工大的研究团队在不改变 RoBERTa 模型结构的前提下,根据中文的语言特点提出了 RoBERTa-WWM(Whole World Mask)模型,该模型在预训练时采用全词遮挡(WWM)的方式,极大地提升了 RoBERTa 模型在中文环境下的文本表示能力[19]。全词遮挡的具体工作原理如表 1 所示。

表 1　WWM 的工作原理示例

原始的句子	句子中需要遮挡的词	使用经典 MASK 方式后的输入	使用 WWM 方式后的输入
使用语言模型来预测下一个词出现的概率	使用语言**模型**来**预测**下一个词出现的**概率**	使用语言[MASK]型来[MASK]测下一个词出现的[MASK]率	使用语言[MASK]来[MASK]下一个词出现的[MASK]

1.2　t-SNE 降维

无论是传统的文本表示模型还是 RoBERTa 模型,最终得到的文本向量维数一般都会很高,而高维度的数据则会导致数据的可区分性变差,模型容易出现过拟合等问题。因此在进行聚类之前,为了避免高维数据带来的影响,同时节省存储和计算成本,一般需要对样本数据进行降维处理[20]。

为了解决 SNE 算法难以优化以及降维到二维空间时容易出现拥挤现象(Crowding Problem)等问题,Maaten 等[21]提出了一种 t-SNE(t-Distributed Stochastic Neighbor Embedding)算法,首先在使用梯度下降法寻找最优解时,使用对称的损失函数,其次为了解决数据拥挤问题,在高维空间仍然保持数据的高斯分布,但是在低维空间使用 t 分布来构建数据分布。实验表明在有异常数据干扰时,t 分布比高斯分布对数据的低维拟合效果更好。

t-SNE 算法的流程如下:

输入:N 个 D 维向量 $\{X_1,\cdots,X_N\}$,设定困惑度 Perp,迭代次数 T,学习速率 η,动量 $\alpha(t)$

输出:二维或者三维向量 $\{Y_1,\cdots,Y_n\}$

1)计算高维空间中的条件概率 $P_{j|i}$,令:

$$P_{ij}=\frac{P_{j|i}+P_{i|j}}{2m} \tag{1}$$

2)使用正态分布 $N(0,10^{-4})$,随机初始化 $Y_{m\times k}$。

3)从 1 到 T 进行迭代并计算低维空间的条件概率 q_{ij}及损失函数 $C(y^i)$对 y^i 的梯度,同时更新 Y^t,其中 Y^t 的计算公式为:

$$Y^t=Y^{t-1}+\eta\frac{\partial C(Y)}{\partial Y}+\alpha(t)(Y^{t-1}-Y^{t-2}) \tag{2}$$

4)输出 Y,得到最终的低维数据分布。

1.3　HDBSCAN 聚类

尽管现在有几十种用于聚类的算法,然而每一种聚类算法都存在某种局限性,只能用来处理特定类型的数据。比如虽然 DBSCAN 算法在含有异常数据的数据集上相比于其他聚类算法有较好的效果,但是它只能对相同密度分布的数据进行聚类,其次在优化过程中的最小步长 Minpts 和邻域半径 Eps 这 2 个参数调整困难,这使得 DBSCAN 算法在应用时受到很大的限制。为了解决这一问题,Campello 等[22]在 DBSCAN 算法的基础上,引入层次聚类的思想构建了 HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise)算法。首先 HDBSCAN 算法定义了一种新的度量 2 个点之间距离的方法,这种度量方法可以更好地反映数据点的密度,计算公式为:

$$d_{\text{mr-k}}(A,B)=\max([d_{\text{corek}}(A),d_{\text{corek}}(B),d_{\text{AB}}]) \tag{3}$$

其中,$d_{\text{mr-k}}(A,B)$是指 A、B 之间的相互可达距离,d_{AB}指的是 A、B 之间的欧氏距离。

其次它使用最小生成树来构建点与点之间的层

次树模型,使得模型仅需给定簇所包含的最小样本数便能自动得到最优的聚类结果,避免了复杂的调参过程,极大地提升了模型准确性和适用范围。HDBSCAN 算法的具体步骤如下:

1)根据密度、稀疏度变换空间。
2)构建距离加权图的最小生成树。
3)构建关联点的簇层次结构。
4)根据最小簇的大小压缩簇的层次结构。
5)从压缩树中提取稳定的簇。

2 算法设计

2.1 算法流程

本文算法流程如图 1 所示。

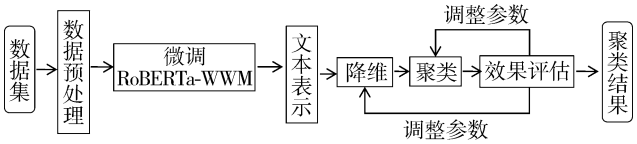

图 1 本文算法的流程

2.2 数据集说明

本文采用的是 THUCNews 新闻文本分类数据集的一个子集[23],本次实验一共选用了其中的 6 个类。为了更好地模拟现实网络环境中的数据分布不均匀的情况,在选取数据时,对每个类别选取不同数量的数据,同时从其他类别的数据集中抽取部分数据作为异常数据。其中正常的数据共 20945 条,干扰数据共 330 条。采用带有标签的数据集的目的是为了能够更好地对比不同算法的聚类效果。

2.3 数据预处理

对于基于 RoBERTa-WWM 模型的算法,首先将每个文本处理成连续的句子序列,然后去除其中的特殊符号和网址等,其次,去除新闻文本中类似于欢迎关注、快讯等对文本的正确语义造成干扰的无效文本,最后由于大部分文本的长度小于 RoBERTa-WWM 模型的最大输入序列长度 512,并且新闻文本的主要信息都集中在文本的开头部分,因此对文本长度大于 512 的文本做头截断(head-only)处理。

2.4 文本向量表示

2.4.1 微调 RoBERTa-WWM 模型

在聚类时,要求相似度高的文本在表示成文本向量时也要具有较高的相似度,其中原始的 RoBERTa-WWM 模型在寻找相似度最高的句子对时,需要将所有句子两两传入模型进行计算,这导致在文本集数量较大时,算法的计算开销很大($O = C_N^2$),其中 O 为算法的计算成本,N 为句子的数量。因此为了使模型能够在下游的聚类中任务中有较好的效果,需要微调 RoBERTa-WWM 模型。在微调时,参考 Reimers 等[24]提出的 SBERT 结构,在 RoBERTa-WWM 模型的输出端加入 3 层网络结构,构建 RoBERTa-WWM 微调模型,具体结构如图 2 所示。

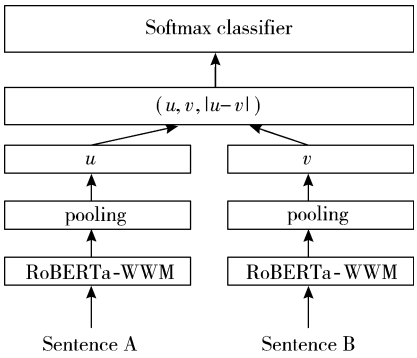

图 2 RoBERTa-WWM 的微调结构

微调的具体流程是:首先将标注好的相似句子对和不相似句子对数据集输入到 RoBERTa-WWM 模型,其中相似句子对和不相似句子对数据集从原有的 THUCNews 新闻文本分类数据集中获得,挑选同一类新闻中语义相似的 2 个句子作为相似句子对,随机选择不同类别新闻中的 2 个句子作为不相似句子对。得到每个句子中每个词的词向量 $\boldsymbol{w}_x[a_1, a_2, \cdots, a_{768}]$,然后通过 pooling 层采用 MEAN 的方式得到固定大小的句子嵌入向量 $\boldsymbol{u}[u_1, u_2, \cdots, u_{768}]$ 和 $\boldsymbol{v}[v_1, v_2, \cdots, v_{768}]$,并计算 2 个句子嵌入向量的差向量 $|\boldsymbol{u}-\boldsymbol{v}| = \boldsymbol{c}[c_1, c_2, \cdots, c_{768}]$ 以及它和 $\boldsymbol{u}, \boldsymbol{v}$ 这 2 个向量的拼接向量 $\boldsymbol{s}[u_1, u_2, \cdots, u_{768}, v_1, \cdots, v_{768}, c_1, \cdots, c_{768}]$,最后乘以训练得到的权重 $\boldsymbol{\omega}_{3n\times 2}$,得到最终的目标分类函数 softmax。在具体计算过程中,使用句子向量的余弦相似度的均方差函数作为损失函数,通过优化损失函数来更新模型参数,达到微调的目的。

2.4.2 计算文本向量

将预处理后的文本数据输入微调后的 RoBERTa-WWM 模型中,得到每个文本的文本向量 $\boldsymbol{T}_i[t_1, t_2, \cdots, t_{768}]$,$i$ 表示第 i 个文本。最终得到的文本向量如表 2 所示。

表 2 文本向量表示示例

文本	类别	前 10 维向量(共 768 维)
古巴雪茄制作师何塞·卡斯特拉尔制作的长 81.8 米的雪茄烟创吉尼斯世界纪录	娱乐	0.5172357 0.11064557 −0.21470352 −0.197296 −0.04500669 −0.6972283 −0.0084492 −0.01120344 0.3232498 0.28340638
最终作《圣域 2:冰与血》游戏宣传片奇幻 RPG《圣域 2》的资料片《冰与血》即将于月底在德国发售	游戏	−0.06114106 0.4234198 0.4794094 −0.0636165 0.25681195 −0.34394372 0.097863 0.12085832 0.0994001 0.43068302

2.5 降维

选择 t-SNE 算法的初始化方式为 pca 降维模式,

设置困惑度和学习率,将文本向量 $\boldsymbol{T}_i$ 输入 t-SNE 算法进行降维,得到降维后的二维向量 $\boldsymbol{t}_i[x,y]$,并对降维后的数据进行可视化。通过观察降维后的数据分布,调整模型的困惑度和学习率,直到达到最优的效果。降维后最终的数据分布如图 3 所示。

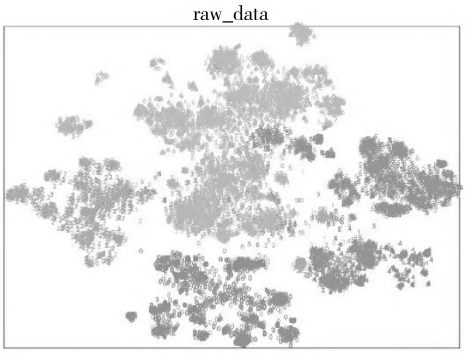

图 3　降维后的数据分布

从图 3 可以看出,数据可以分为 6 个比较大的簇,与实际的数据分布相同。

2.6　效果评估

评估聚类效果时,常用的有轮廓系数、互信息(MI)指数以及 Fowlkes-Mallows(FM)指数等,其中轮廓系数主要是衡量聚类结果中簇内数据点的相似程度和簇之间的区别程度,它一般适用于没有标签的数据,而后两者适用于有标签的情况。互信息指数主要是衡量原始的数据分布和聚类结果分布的相似程度,取值范围为[0,1],值越大,说明聚类效果越好。FM 指数用来综合衡量准确率和召回率。本文选择 FM 指数和 MI 指数作为作为聚类效果的衡量标准。FM 的计算公式为:

$$\mathrm{FM}=\frac{\mathrm{TP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})}} \tag{4}$$

其中,TP 为真实标签和预测标签中属于同一簇的点的数量,FP 为在真实标签中属于同一个簇而在预测中不是的点的个数,FN 为在预测中属于同一个簇而在实际中不是的点的个数,MI 的计算公式为:

$$\mathrm{MI}=\sum\sum p_{i,j}\log\left(\frac{p_{i,j}}{p_i\times p_j}\right) \tag{5}$$

其中:

$$p_{i,j}=\frac{m_{ij}}{N} \tag{6}$$

其中,p_i 为归属于 i 类的数据个数占数据总量的比例,p_j 同理。m_{ij}表示第 1 个序列中的 i 与第 2 个序列中 j 的交集的个数,N 为序列的长度。

2.7　聚类

使用 t-SNE 算法降维后的二维向量作为 HDBSCAN 算法的输入,设置参数的变化范围得到聚类结果,并计算 FM 指数和 MI 指数,根据它们的变化曲线选择最优的参数的结果作为最终的聚类结果,同时对该结果可视化,FM 指数和 MI 指数的变化情况和最终的聚类结果如图 4 ~ 图 6 所示。

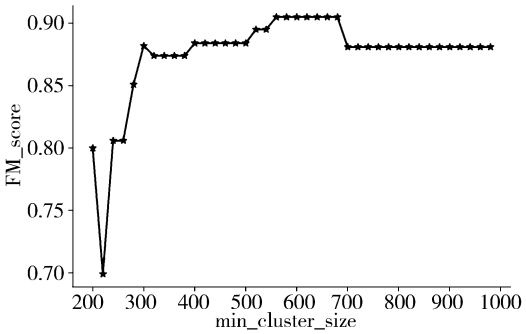

图 4　FM 指数随 min_cluster_size 的变化情况

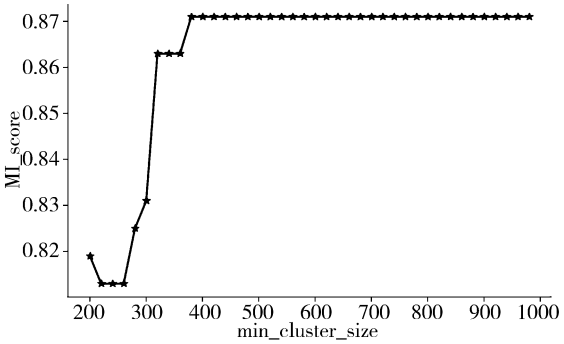

图 5　MI 指数随 min_cluster_size 的变化情况

从图 4 和图 5 可以看出,类别最小样本数在 540 ~680 时,聚类效果最好。

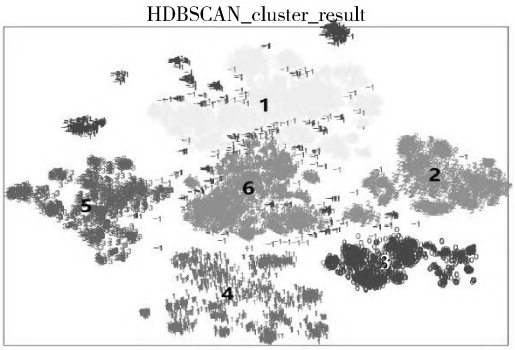

图 6　聚类后的每个类别的数据分布情况

从图 6 可以看出,聚类后的结果分布和原始分布大致相同,除了异常点外,总共可以有 6 个类别,每种类别中的文本如表 3 所示。

表 3　每个类别的文本向量

类别	标签	每个类别前 2 个文本
-1	异常	1. 万圣节饰品“喜乐哀怒”怒:鲜血淋漓骷髅头 2. 另外一个角度再来一张另外一个角度再来一张
0	星座	1. 有缘无分的星座组合编者按:大提琴奏不出你的心声,长笛吹不尽我的哀伤向左,向右,只是一个转身的距离,却是今生的再见 2. 12 星座谁最容易被欺骗编者按:梦也该结束了,她告诉自己。毕竟自己比任何人都清楚他的多情

续表 3

类别	标签	每个类别前 2 个文本
1	体育	1. 火箭击败步行者您如何评价米勒及全队表现？新浪体育讯北京时间 11 月 13 日，火箭在客场以 102-99 力取步行者，避免遭到二连败。火箭结束对步行者的四连败 2. 火箭常规训练犀利哥这是发烧了？新浪体育讯北京时间 3 月 19 日，火箭队在丰田中心进行了常规训练，训练结束后教练和部分球员接受了媒体的采访
2	游戏	1. 游戏十年评选最受期待网络游戏奖，成吉思汗、永恒之塔、万王之王、剑侠情缘网络版 2.《GT 赛车 5》最新官方宣传游戏视频本作是 SCE 旗下王牌竞速游戏《GT 赛车》系列在 PS3 上的第一款正统续作，在《GT 赛车 5 序章》的基础上进一步升级进化，带来令人惊异的真实赛车模拟游戏体验
3	时政	1. 澳大利亚政府警告称恐怖分子可能袭击沙特石油设施 2. 国务院：将建立杂交玉米和杂交水稻种子储备体系
4	股票	1. 大成行业轮动股票基金聘杨建勋任基金经理大成基金周五公告，大成行业轮动股票基金聘任杨建勋担任基金经理 2. 东方稳健回报债券基金发行东方基金旗下首只债券产品，稳健回报基金于昨日起在建行、工行、中行、招行、邮储行及东北证券等代销机构全国认购
5	教育	1. 09 年考研数学二大纲变化比较综述 2. 2004—2010 年管理类考研分数线走势图

3 实验结果分析

对不同的文本表示模型以及聚类模型进行实验，取每种模型最好的聚类效果作为最终结果，最终的结果如表 4 所示。

表 4 不同模型的聚类效果

模型(降维均采用 t-SNE)	FM 指数	MI 指数
词袋模型 + K-Means	0.70	0.64
Word2Vec + K-Means	0.77	0.69
RoBERTa-WWM + K-Means	0.88	0.79
词袋模型 + HDBSCAN	0.79	0.73
Word2Vec + HDBSCAN	0.81	0.76
RoBERTa-WWM + HDBSCAN	0.89	0.84
RoBERTa-WWM(微调) + HDBSCAN	0.91	0.87

从表 4 可以看出，首先在使用相同的聚类模型的情况下，基于 RoBERTa 模型做文本表示的算法，它的 FM 指数和 MI 指数比基于词袋模型和 Word2Vec 模型做文本表示的算法平均高出 8 个百分点以上。其次，在使用相同的文本表示模型时，基于 HDBSCAN 模型的算法比基于 K-Means 模型的算法具有更好的聚类效果。除此之外，经过微调的 RoBERTa-WWM 模型，相比于原始的 RoBERTa-WWM 模型，聚类效果有所提升。

4 结束语

本文针对现有的文本聚类算法在处理现实网络环境中的数据时存在的问题，提出了一种基于 RoBERTa-WWM + HDBSCAN 的文本聚类算法，经过实验验证，首先该算法相比于传统的文本聚类算法，聚类效果有了很大的提升，也更适合现实网络环境下的数据，其次本文使用的 t-SNE 降维算法，除了可以用来降维，还可以对降维后的数据进行可视化，这使得在后续的聚类工作中，可以更好地确定最终的类别数，从而缩小了算法优化过程中参数的调整范围。然而本文仍存在可以改进的点，首先由于本文使用的有标注的微调数据较少，因此对 RoBERTa-WWM 模型微调后，对算法整体的聚类效果提升有限，如果进一步扩充数据集，聚类效果的提升会更明显，其次，本文只研究了 THUNews 数据集，在后续的工作中可以针对其他领域的数据进行研究以进一步提升该算法的泛化能力。

参考文献：

[1] 张长利. 面向特定领域的互联网舆情分析技术研究[D]. 长春：吉林大学, 2011.

[2] 许强. 基于 Spark 的话题检测与跟踪技术研究[D]. 成都：电子科技大学, 2018.

[3] 冀宇轩. 文本向量化表示方法的总结与分析[J]. 电子世界, 2018(22)：10-12.

[4] 陈新泉，周灵晶，刘耀中. 聚类算法研究综述[J]. 集成技术, 2017，6(3)：41-49.

[5] 李志强，王俊丰，贾晓霞. 基于 K-Means 算法改进的短文本聚类研究与实现[J]. 信息技术, 2019，43(12)：76-80.

[6] 毛郁欣，邱智学. 基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究[J]. 中国信息技术教育, 2020(8)：99-101.

[7] 吴德平，华钢. 基于 Word2Vec 词嵌入和聚类模型的安全生产事故文本案例分类[J]. 计算机系统应用, 2021，30(1)：141-145.

[8] 阮光册，夏磊. 基于 Doc2Vec 的期刊论文热点选题识别[J]. 情报理论与实践, 2019，42(4)：107-111.

[9] 贾君霞，王会真，任凯，等. 基于句向量和卷积神经网络的文本聚类研究[J/OL]. 计算机工程与应用：1-6[2021-10-08]. http://kns.cnki.net/kcms/detail/11.2127.TP.20210622.0840.002.html.

[10] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. Computation and Language, arXiv preprint arXiv：1810.04805, 2018.

[11] 李舟军，范宇，吴贤杰. 面向自然语言处理的预训练技术研究综述[J]. 计算机科学, 2020，47(3)：162-173.

[12] 曹凤仙. 基于 K-Means 的市长公开电话文本聚类[D]. 长春：东北师范大学, 2021.

[13] 朱良奇，黄勃，黄季涛，等. 融合 BERT 和自编码网络的短文本聚类研究[J/OL]. 计算机工程与应用：1-10[2021-08-24]. http://kns.cnki.net/kcms/detail/11.2127.TP.20210527.0850.002.html.

[14] 邹艳春. 基于 DBSCAN 算法的文本聚类研究[J]. 软件导刊, 2016，15(8)：36-38.

(下转第 63 页)

cal dimming for high-dynamic-range liquid crystal displays [J]. Optics Express, 2017,25(3):1973-1984.
[5] MA J, REN X, YUREVICH T V. A novel fast iterative parallel thinning algorithm[C]// Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing. 2020:1-5.
[6] 袁良友,周航,韩丹,等. 引入平滑迭代的骨架提取改进算法[J]. 计算机工程与应用, 2020,56(24):194-199.
[7] CHAO X, XIAO X, LUO Y, et al. New skeleton extraction method based on distance transform[J]. Chinese Journal of Scientific Instrument, 2012,33(12):2851-2856.
[8] TELEA A, VAN WIJK J J. An augmented fast marching method for computing skeletons and centerlines[C]// Proceedings of the Symposium on Data Visualisation. 2002: 251-259.
[9] GUO Y H, SENGUR A. A novel 3D skeleton algorithm based on neutrosophic cost function[J]. Applied Soft Computing, 2015,36:210-217.
[10] KOTSUR D, TERESHCHENKO V. Optimization heuristics for computing the Voronoi skeleton[C]// International Conference on Computational Science. 2019:96-111.
[11] SUÁREZ A J F, HUBERT E. Scaffolding skeletons using spherical Voronoi diagrams[J]. Electronic Notes in Discrete Mathematics, 2017,62:45-50.
[12] HASSOUNA M S, FARAG A A. Robust centerline extraction framework using level sets[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005:458-465.
[13] SZWEDOWSKI T D, FIALKOV J, PAKDEL A, et al. An optimized process flow for rapid segmentation of cortical bones of the craniofacial skeleton using the level-set method [J]. Dentomaxillofacial Radiology, 2013,42(4):1-6.
[14] YANG Z H, GUO F F, DONG P. Robust skeleton extraction of gray images based on level set approach[J]. Journal of Multimedia, 2013,8(1):24-31.
[15] SETHIAN J A, VLADIMIRSKY A. Fast methods for the Eikonal and related Hamilton-Jacobi equations on unstructured meshes[J]. Proceedings of the National Academy of Sciences, 2000,97(11):5699-5703.
[16] MIREBEAU J M, PORTEGIES J. Hamiltonian fast marching: A numerical solver for anisotropic and non-holonomic Eikonal PDEs[J]. Image Processing on Line, 2019,9:47-93.
[17] CRANE K, WEISCHEDEL C, WARDETZKY M. Geodesics in heat: A new approach to computing distance based on heat flow[J]. ACM Transactions on Graphics (TOG), 2013,32(5):1-11.
[18] YANG J M, STERN F. A highly scalable massively parallel fast marching method for the Eikonal equation[J]. Journal of Computational Physics, 2017,332:333-362.
[19] CRANE K, WEISCHEDEL C, WARDETZKY M. The heat method for distance computation[J]. Communications of the ACM, 2017,60(11):90-99.
[20] SHARP N, CRANE K. A laplacian for nonmanifold triangle meshes[J]. Computer Graphics Forum, 2020,39(5): 69-80.
[21] YANG F, COHEN L D. Geodesic distance and curves through isotropic and anisotropic heat equations on images and surfaces[J]. Journal of Mathematical Imaging and Vision, 2016, 55(2):210-228.
[22] ROUCHDY Y, COHEN L D. Geodesic voting for the automatic extraction of tree structures. Methods and applications[J]. Computer Vision and Image Understanding, 2013,117(10):1453-1467.
[23] BOUIX S, MARTIN-FERNANDEZ M, UNGAR L, et al. On evaluating brain tissue classifiers without a ground truth [J]. Neuroimage, 2007,36(4):1207-1224.

(上接第 52 页)

[15] 曹旭友,周志平,王利,等. 基于 BERT + ATT 和 DBSCAN 的长三角专利匹配算法[J]. 信息技术, 2020,44(3):1-5.
[16] 蔡岳,袁津生. 基于改进 DBSCAN 算法的文本聚类[J]. 计算机工程, 2011,37(12):50-52.
[17] 王纵虎. 聚类分析优化关键技术研究[D]. 西安:西安电子科技大学, 2012.
[18] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized bert pretraining approach[J]. Computation and Language, arXiv preprint arXiv:1907.11692, 2019.
[19] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for chinese BERT[J]. Computation and Language, arXiv preprint arXiv:1906.08101, 2019.
[20] HUANG X, WU L, YE Y S. A review on dimensionality reduction techniques[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(10): 1950017.1-1950017.25.
[21] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008,9(11):2579-2605.
[22] CAMPELLO R J G B, MOULAVI D, ZIMEK A, et al. Hierarchical density estimates for data clustering, visualization, and outlier detection[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2015,10(1):1-51.
[23] 孙茂松,李景阳,郭志芃,等. THUCTC: 一个高效的中文文本分类工具包[EB/OL]. [2018-10-20]. http://thuctc.thunlp.org/.
[24] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using siamese BERT-networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.