计算机科学
COMPUTER SCIENCE

基于Doc2Vec增强特征的长文本主题聚类研究

陈洁

引用本文

陈洁. 基于Doc2Vec增强特征的长文本主题聚类研究[J]. 计算机科学, 2023, 50(6A): 220800192-6.

CHEN Jie. Study on Long Text Topic Clustering Based on Doc2Vec Enhanced Features[J]. Computer Science, 2023, 50(6A): 220800192-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800192

基于 Doc2Vec 增强特征的长文本主题聚类研究

陈　洁
中华女子学院数据科学与信息技术学院　北京 100101

摘　要　针对新闻长文本语义表征的难点,基于 Doc2Vec 文档嵌入和词向量加权方式构建增强的特征表示。利用 DV-sim 方法和 DV-tfidf 方法从文档首尾部分特定词性的内容中提取增强特征,再分别与 Doc2Vec 文档向量组合,形成新的全局表征。DV-sim 从语义角度,采用特征词与 Doc2Vec 向量的相似度获得词权重;DV-tfidf 从词频统计角度,采用词频-逆文档频率方式获得词权重,然后利用 HDBSCAN 算法在 THUCNews 和 Sogou 数据集上进行主题聚类。相比直接应用 Doc2Vec 向量,DV-sim 在两个数据集上的噪声数分别减少 60.82% 和 60.63%,准确率提高 12.14%和 20.58%,F1-Score 值提高 15.61%和 11.58%;DV-tfifd 在两个数据集上的噪声数分别减少 15.20%和 59.55%,准确率提高 10.85%和 17.93%,F1-Score 值提高 15.60%和 9.21%。实验结果表明,DV-sim 和 DV-tfidf 都可以提高主题聚类性能,且基于语义的增强特征比基于词频的效果更好,DV-sim 在优秀女性人物报道的主题聚类上也得到了有效应用。

关键词:主题聚类;文本表征;Doc2Vec;词向量;HDBSCAN
中图法分类号　TP391

Study on Long Text Topic Clustering Based on Doc2Vec Enhanced Features

CHEN Jie
School of Data Science and Information Technology,China Women's University,Beijing 100101,China

Abstract　Aimed at the difficulties of semantic representation of long news text,an enhanced document feature representation is constructed based on Doc2Vec embedding and word vector weighting. Enhanced features from the specific parts-of-speech contents on the head and tail of the document are extracted by the method of DV-sim or DV-tfidf. These features are then combined with doc2vec to form a new global representation. DV-sim uses the similarity between feature words and doc2vec vectors to obtain word weight from the semantic point of view,and DV-tfidf uses term frequency inverse document frequency to obtain word weight from the word frequency statistics point of view. Then the HDBSCAN algorithm is applied to cluster topics on the Thucnews and Sogou datasets. Compared with the Doc2Vec vector,the noise number on the two datasets reduces by 60.82% and 60.63%,the accuracy improves by 12.14% and 20.58%,and the F1-score increases by 15.61% and 11.58%,respectively,with DV-sim. The noise number on the two datasets reduces by 15.20% and 59.55%,the accuracy improves by 10.85% and 17.93%,and the F1-score increases by 15.60% and 9.21%,respectively,with DV-tfidf. Experiments show that both DV-sim and DV-tfidf can improve the performance of topic clustering,and the enhancement feature based on semantics is better than that based on word frequency. DV-sim has also been effectively applied in topic clustering of excellent female character reports.

Keywords　Topic clustering,Text representation,Doc2Vec,Word embedding,HDBSCAN

网络应用的飞速发展使得文本数量日益增多。文本主题聚类是一种无监督的机器学习方法,该方法通过提供一种有意义的分类,可以自动完成文本信息的有效组织、自动归类、话题发现等任务,帮助用户快速地从海量文本中获取特定主题下有用的信息。

文本相似度聚类是文本主题聚类的一种重要方法,是在文档集无类别标注的情况下,依据特定的标准将其划分为若干簇,每个簇代表一个主题,文本表征是文本相似度聚类的关键。对大型文本的有效分析是一个具有挑战性的问题,长文本因语义更加多样化、文本蕴含的主题不唯一且存在冗余和噪声等问题,增加了聚类难度。

传统的向量空间模型通过 One-Hot,TF-IDF 等方法提取文本特征,但不考虑词序,缺乏语义特征[1]。利用神经网络语言模型可以获得语义关联的文本表征[2],目前主要有基于词嵌入(Word Embedding)方式和基于预训练语言模型(PTM)方式。以 BERT[3]为代表的 PTM 采用深度双向 Transformer 结构,具有很强的特征提取功能,但是 BERT 要求输入序列的长度不超过 512 个字符,无法直接获得长文档的语义表征。Doc2Vec 模型[4]可应用于短语、句子、大型文档等任意长度的文本片段,在情感分析[5]、文本分类[6-7]、主题聚类[8-11]、个性化推荐[12]等 NLP 任务中得到了有效应用。

本文利用 Doc2Vec 模型和 HDBSCAN 聚类算法实现文本主题聚类。针对长文本语义表征的难点,构建基于 Doc2Vec 的增强文档向量,以提高主题区分度,提升聚类

基金项目:中华女子学院科研基金(ZKY200020228)
This work was supported by the Research Fund of China Women's College(ZKY200020228).
通信作者:陈洁(chenjie@cwu.edu.cn)

性能。通过 THUCNews 和 Sogou 两个公开的新闻数据集以及网站采集的优秀女性人物报道的长文本聚类结果，证明了增强文档向量在主题聚类上的有效性。

1 相关研究

1.1 文本表征

TF-IDF 可以评估某个单词在一个文档中的重要程度，Chen 等[13]将 TF-IDF 与其他特征构成混合向量模型，以弥补 TF-IDF 缺乏语义的不足，并应用于中文评论情感分析。Word2Vec[14]是一种词嵌入模型，可以获取词汇级别的语义，利用词嵌入的组合可获得更大文本的嵌入。Tang 等[15]采用 TF-IDF 和词嵌入的加权求和获得文本向量，以体现单个词对整篇文档的影响程度，并应用于搜狗新闻分类。

Doc2Vec 模型是在 Word2Vec 上发展起来的，其能够综合考虑文本对象及其结构，对文档的语义和语法进行建模，获得全局文本表征。PV-DM(Distributed Memory Model of Paragraph Vectors)是 Doc2Vec 的一种训练模型，其结构如图 1 所示。每个文档被视为一个特殊单词，与上下文中的词向量具有相同的维度，经过平均或拼接形成组合向量，并通过预测目标单词同时学习文档向量和词向量，从而在文档级别学习分布式向量表示。

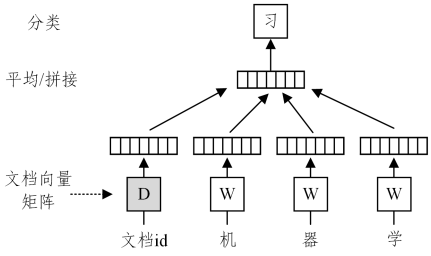


图 1 PV-DM 模型结构

Fig. 1 PV-DM model structure

LEE 等[5]利用 Doc2Vec 模型生成文档向量和情感极性向量，然后将两者拼接形成混合特征向量，用于影评数据的情感分类，结果表明使用 Doc2Vec 嵌入比 TF-IDF 具有更高的性能，而混合向量比仅使用文档向量或极性向量具有更高的精度。

Mandal 等[6]分别使用 IF-IDF、IF-IDF 与 Word2Vec 加权、Doc2Vec、BERT、LDA 等方法表征法律文档，用于相似性评估。由于 BERT 不能直接处理大型文本，因此他们将长文本分割为 500 字的重叠块，再通过平均池化获得全局特征表示。与专家评估结果相比，Doc2Vec 相似性得分最接近，BERT 嵌入不能很好地执行任务，可能是由长文本的截断造成的。

Adorno 等[7]使用 Doc2Vec 模型识别文章作者的写作风格，对内容涉及多个类别的文章采用丢弃的方法，避免类别重叠。Arif 等[10]利用 Doc2Vec 与球形聚类进行新闻主题建模，特别指出聚类时应考虑一篇新闻文章会涉及较为宽泛主题的问题，建议与分层聚类方法相结合。

现有研究表明融合的特征表示可以补充单一特征的不足，长文本需要排除文档中噪声的影响，更好地突出主要主题，但尚未有比较好的解决方法。

1.2 聚类算法

不同的文献采用了不同的文本相似度聚类算法，Jia 等[8]和 Ruan 等[9]采用改进的 K-means 聚类，Arif 等[10]采用球形聚类算法，Chang 等[11]采用基于密度的聚类算法。K-Means 算法需要事先确定聚类的主题数目，并且对离群点(也称为噪声或异常值)较为敏感；基于密度的聚类算法可以发现任意形状的簇，并且能够有效处理噪声点。因此，本文采用 HDBSCAN 算法[16-18](Hierarchical Density-Based Spatial Clustering of Applications with Noise)。

HDBSCAN 是 DBSCAN[19]算法与层次聚类算法的结合。DBSCAN 算法对邻域半径 ε 和邻域内最少点参数的设置很敏感。HDBSCAN 通过在不同的 ε 上执行 DBSCAN，并对结果进行集成，找到提供最佳稳定性的聚类，以此得到不同密度的簇类，对参数选择更具鲁棒性。而且，HDBSCAN 不需要预先确定聚类数目，可以提供一种聚类层次结构，较为客观地选择聚类数目。

2 基于 Doc2Vec 增强特征的主题聚类模型

2.1 基于 Doc2Vec 的增强特征表示

给定文档集 $D=\{d_1,d_2,d_3,\cdots,d_n\}$，利用 PV-DM 模型训练 Doc2Vec，可以得到文档嵌入 $\{\boldsymbol{v}(d_1),\boldsymbol{v}(d_2),\boldsymbol{v}(d_3),\cdots,\boldsymbol{v}(d_n)\}$ 和词嵌入 $\{\boldsymbol{v}(w_1),v(w_2),\boldsymbol{v}(w_3),\cdots,\boldsymbol{v}(w_N)\}$，其中 $\{w_1,w_2,w_3,\cdots,w_N\}$ 代表词汇表中的所有单词。两者来自相同的语料库，并且 $\boldsymbol{v}(d_i)$ 和 $\boldsymbol{v}(w_j)$ 具有相同的嵌入维度。

Doc2Vec 模型可以直接获得全局文本表征，即文档向量 $v(d)$，但长文档可能存在文本蕴含的主题不唯一、噪声较多等问题。为此，本文在 Doc2Vec 文档向量基础上，构建增强的文档特征表示，以提高主题区分度，降低噪声影响。

(1)增强特征的表示

新闻的开头部分通常都是开宗明义，主体详细阐述，结尾强调意义与影响，或者先引入一段背景信息再进入正题，例如一则与健康相关的报道，开篇可能是一段描写熬夜看球赛的内容。因此，新闻的标题以及正文开头和结尾部分(HT-Segments)通常与新闻主题的相关性最大，可以提取 HT-Segments 中的内容特征增强 Doc2Vec 文档表征，以此获得增强的文档向量 $v_e(d)$。

本文利用词嵌入获取 HT-Segments 的特征表示，由于每个特征词对主题的贡献度是不同的，因此我们采用以下两种方式获取特征词的权重。

(1)基于语义相似度的词权重。利用 HT-Segments 中每个词与文档语义的相似度，选择与文档语义最相关的特征词，语义越相近，分值越高。本文采用 Scaled dot-product attention 方法获取相似度权值，第 d 篇文档中各特征词的相似度为 $sim(w,d)$：

$$sim(w,d)=\mathrm{softmax}\left(\frac{\boldsymbol{v}(d)\boldsymbol{v}(w)^{\mathrm{T}}}{\alpha}\right) \tag{1}$$

其中，$\boldsymbol{v}(d)$ 为文档向量；$\boldsymbol{v}(w)$ 为词向量；softmax 函数对分值进行归一化，得到(0,1)范围的相似度分布；α 是一个调节参数，较小的取值可以保持 softmax 函数的输入尽可能大，聚焦最重要的词语。

(2)基于词频的词权重。词频-逆文档频率(TF-IDF)反映了特征词表征文本的能力,TF-IDF 值越高,表示单词的重要性越强,即该单词在一篇文档中出现的频率高而在其他文档中较少出现。第 d 篇文档中的第 t 个特征词的 $tf\text{-}idf$ 值为 $tfidf(t,d)$:

$$tfidf(t,d)=tf(t,d)*idf(t)=\frac{n_{t,d}}{\sum_k n_{k,d}}*\log\frac{|D|}{df(t)+1} \tag{2}$$

其中,n 表示词频,k 为文档 d 的词汇数量,$|D|$ 为语料库的文档总数,$df(t)$ 为包含特征词 t 的文档数量。

(2)构建增强的文档特征向量

从文档开头(包含标题)和结尾部分分别截取 500 字符,将合并后的 1 000 字符作为 HT-Segments 的内容来源。由于名词和动词可以更有效地表达文档内容[20],因此只提取其中的名词、动词和动名词作为 HT-Segments 中的特征词。

将特征词权重 t(分别对应式(1)中的 sim 和式(2)中的 $tfidf$)与 Doc2Vec 词向量 $\boldsymbol{v}(w)$ 进行加权求和,得到 HT-Segments 的特征向量 $\boldsymbol{v}(s)$:

$$\boldsymbol{v}(s)=\sum_{j=1}^{M} t_{i,j}*\boldsymbol{v}(w_{i,j}) \tag{3}$$

其中,M 表示特征词的数量,词向量 $\boldsymbol{v}(w)$ 与文档向量 $\boldsymbol{v}(d)$ 来自相同的语义空间。

将文档向量 $\boldsymbol{v}(d)$ 与特征向量 $\boldsymbol{v}(s)$ 叠加,得到 $\boldsymbol{v}(d)$ 的增强语义表示 $\boldsymbol{v}_e(d)$,进行归一化处理后作为最终的文档特征向量 $\boldsymbol{v}_{\text{norm}}(d)$,第 i 个向量的第 j 个元素 $\boldsymbol{v}_{\text{norm}}(d'_{i,j})$ 表示为:

$$\boldsymbol{v}_e(d)=\boldsymbol{v}(d)+\boldsymbol{v}(s) \tag{4}$$

$$\boldsymbol{v}_{\text{norm}}(d'_{i,j})=\boldsymbol{v}_e(u_{i,j})/\|\boldsymbol{v}_e(u_i)\| \tag{5}$$

其中,$\|\boldsymbol{v}_e(u_i)\|$ 表示向量的模长。

2.2　主题聚类

HDBSCAN 是一种基于密度的聚类算法,容易受到维数灾难的影响。减少数据维度,可使密度更加明显,从而提高基于密度的聚类的性能。为此,采用 UMAP[21-22](Uniform Manifold Approxima_ tion and Projection)算法降维。UMAP 是一种非线性的流行降维技术,对高维数据很有效,并且能够尽可能多地保留全局数据结构。

在 HDBSCAN 聚类算法中,邻域内最小样本数的设置会影响聚类的主题数。取值较小时,聚类粒度也小,主题划分地更细,可以发现更多的主题,但也会产生较多相似主题,需要进一步合并;取值较大时,则只能得到较少的聚类,不利于发现更多有效的主题。将 HDBSCAND 的层次聚类结构可视化,可以更直观地选择合适的主题数目。

提取主题词采用基于簇的 TF-IDF 算法:

(1)将一个聚簇中的所有文档合并为一个大文档。

(2)对每个聚簇应用 TF-IDF 算法提取关键词作为主题描述,第 c 个聚簇中的第 t 个词汇的 $tfidf$ 值为 $tfidf(t,c)$:

$$tfidf(t,c)=\frac{n_{t,c}}{\sum_k n_{k,c}}*\log\frac{|W|}{df(t)+1} \tag{6}$$

其中,$|W|$ 为所有聚簇的平均单词数,$df(t)$ 为单词 t 在所有文档中出现的总次数。

2.3　聚类性能评估

在带标签的数据集上可以采用准确率(Acc)和 F1-score(F1)评估聚类性能。

准确率是聚类正确的样本数占总样本数的比例,计算公式为:

$$Acc=\frac{TP+TN}{TP+FP+TN+FN} \tag{7}$$

其中,TP 表示正确聚类的样本数,FP 表示错误聚类的样本数。

F1-score 是精确率 P 和召回率 R 的调和值,是一个综合指标,多类别性能评估时可以使用宏平均(MF1),类别不平衡时可以使用加权 F1-score(WF1),计算公式为:

$$P=\frac{TP}{TP+FP} \tag{8}$$

$$R=\frac{TP}{TP+FN} \tag{9}$$

$$F1=\frac{2*P*R}{P+R} \tag{10}$$

$$MF1=\frac{\sum_{i=1}^{V} F1_i}{V} \tag{11}$$

$$WF1=\sum_{i=1}^{V} w_i*F1_i \tag{12}$$

其中,V 是样本类别数,w_i 是数据集中第 i 类样本占总样本的比例,$F1_i$ 是第 i 类样本的 $F1$。

基于 Doc2Vec 增强特征的主题聚类流程如图 2 所示。

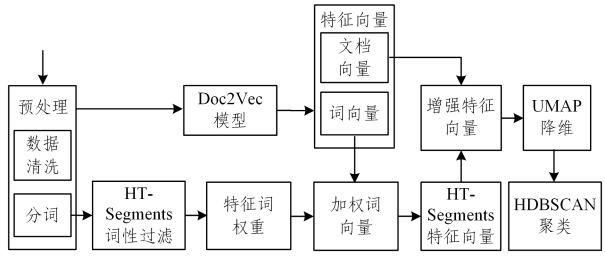


图 2　基于 Doc2Vec 增强特征的主题聚类流程

Fig. 2　Topic clustering process based on Doc2Vec enhanced features

3　实验与结果分析

3.1　实验设置

实验选择 THUCNews 和 Sogou 两个带标签的公开新闻数据集进行测试,然后应用于优秀女性人物报道的主题聚类。

(1)从 THUCNews 数据集的教育、科技、社会、体育、房产、时尚、游戏、财经、娱乐、家居这 10 个类别中随机选择 10 000 条数据,每条数据的长度在 300～10 000 之间,各类别样本数不相等,时尚和家居两个类别的样本数较少,占比分别为 7.78% 和 7.46%,其他类别的样本占比在 9.79%～10.88%之间。

(2)从 Sogou 数据集的科技、社会、健康、体育、军事、旅游、汽车、财经、房产、教育这 10 个类别中随机选择 10 000 条数据,每条数据的长度在 100～10 000 之间,各类别样本数量相等。

(3)从中华全国妇女联合会、中国妇女网和女性之声 3 个网站采集 2007 年－2021 年的优秀女性人物报道,通过数据清洗删除内容中包含的 HTML 标签和特殊符号,得到长度在 100～10 000 之间的 7 650 条无标签数据(Women)。

3 个数据集的基本信息如表 1 所列。

表 1 数据集的基本信息

Table 1 Basic information of data sets

dataset	average length	proportion of samples with length less than 1000	proportion of samples with length of 1000～3000	proportion of samples with length greater than 3000
THUCNews	1383.7938	46.80%	46.16%	7.04%
Sogou	1006.4038	64.73%	31.20%	4.07%
Women	1848.97268	12.54%	77.57%	9.90%

使用 jieba 工具进行分词，利用哈工大的停用词表去除常用的停用词(包括标点符号)，将处理后的结果送入 Doc2Vec 模型进行训练。

实验语言为 Python 3.7，调用 Gensim 库的 Doc2Vec 算法，选择 PV-DM 训练模型，滑动窗口大小设置为 7，嵌入维度为 192，经过训练，将文档和单词分别转换为 192 维的文档向量和词向量。

数据降维调用 umap 库提供的算法，嵌入的空间维度设置为 50；嵌入点之间的最小距离设置为 0，使样本点能够密集地打包在一起，有助于簇之间更清晰的分离。降维处理后，得到 50 维的新向量，再调用 hdbscan 库中的 HDBSCAN 算法进行聚类。在 HDBCAN 算法中，主要参数是设置邻域内最小样本数 *min_count*，其大小决定了聚类的粒度，在 THUCNews 和 Sogou 两个数据集上 *min_count* 设置为 100。

聚类的文档向量分别使用以下 3 种形式，维度均为 192。

(1)DV：Doc2Vec 模型获取的文档向量。

(2)DV-sim：Doc2Vec 向量与 sim 加权的增强特征向量叠加，构成新的文档向量。

(3)DV-tfidf：Doc2Vec 向量与 TF-IDF 加权的增强特征向量叠加，构成新的文档向量。

3.2 实验结果分析

(1)THUCNews 和 Sogou 数据集上的聚类

在 THUCNews 和 Sogou 数据集上分别使用上述 3 种文档向量进行聚类，并使用准确率(Acc)、F1 值的宏平均(MF1)或加权平均(WF1)进行性能评估。

在 THUCNews 上的聚类性能如表 2 所列。可以看出，相比 DV 方法，DV-sim 噪声数减少 60.82%，ACC 和 WF1 分别提高 12.14%和 15.61%；DV-TFIDF 噪声数减少 15.20%，ACC 和 WF1 分别提高 10.85%和 15.60%。DV-sim 增强文档向量的聚类性能更好，说明增强语义后，提高了文本类别的可识别度，增加了簇类数量，有更多的文本能够正确聚类。

表 2 THUCNews 上的聚类性能

Table 2 Cluster performance on THUCNews dataset

文档向量	簇类数	主题数	噪声数	ACC	WF1
DV	12	9	513	0.7856	0.7698
DV-sim	14	10	201	0.8810	0.8900
DV-tfidf	15	10	435	0.8708	0.8899

图 3－图 5 分别为 3 种文档向量得到的聚类结果(－1 表示噪声)，根据每个簇类的主题词识别该簇的主题类别。在 DV 聚类结果中，1 号是关于“时尚的生活方式”，10 号是关于“时尚的穿搭设计”，两者都属于“时尚”主题。2 号则将游戏和科技内容聚为一类，合并同类主题后，DV 得到 9 个不同类别的主题，DV-sim 和 DV-tfidf 分别得到 10 个不同类别的主题。

主题	数量	主题词	类别
-1	513	期货 公司 企业 市场 交易 发展 设计 价格 活动 业务	
0	1076	比赛 球员 球队 火箭 对手 进攻 篮板 防守 表现 时间	体育
1	534	减肥 女性 食物 身体 运动 脂肪 方法 导语 按摩 饮食	时尚
2	1949	游戏 玩家 用户 手机 网游 公司 活动 网络 市场 产品	游戏
3	507	基金 投资 公司 股票 市场 收益 型基金 指数 投资者 银行	财经
4	473	市场 价格 上涨 期货 商品 合约 需求 经济 大豆 黄金	财经
5	1123	考生 学生 考试 学校 招生 录取 专业 高考 志愿 教育	教育
6	1016	男子 警方 民警 医院 发现 法院 发生 儿子 派出所 告诉	社会
7	960	观众 电影 导演 娱乐 音乐 演员 拍摄 节目 演唱会 角色	娱乐
8	196	房价 土地 市场 楼市 住房 开发商 政府 成交 地产 住宅	房产
9	825	项目 户型 别墅 热盘 样板间 区域 相册 均价 位于 生活	房产
10	198	搭配 时尚 黑色 设计 性感 外套 单品 造型 款式 组图	时尚
11	630	家具 品牌 企业 产品 家居 消费者 行业 装修 市场 设计	家居

图 3 THUCNews 使用 DV 文档向量的聚类结果

Fig. 3 Cluster result using DV vector on THUCNews dataset

主题	数量	主题词	类别
-1	201	网友 业主 网吧 银行 小区 收藏 公司 网络 物业 业委会	
0	1087	比赛 球员 球队 火箭 对手 篮板 进攻 防守 表现 联赛	体育
1	547	减肥 女性 食物 身体 运动 脂肪 方法 导语 按摩 胸部	时尚
2	767	项目 户型 别墅 热盘 样板间 相册 区域 位于 均价 生活	房产
3	407	采用 像素 英寸 笔记本 机身 功能 处理器 接口 性能 支持	科技
4	1140	考生 学生 考试 学校 招生 录取 专业 高考 志愿 教育	教育
5	475	基金 投资 公司 股票 市场 收益 型基金 指数 银行 净值	财经
6	534	期货 市场 价格 上涨 合约 商品 黄金 现货 需求 经济	财经
7	988	游戏 玩家 网游 活动 网络游戏 世界 技能 奖励 体验 系统	游戏
8	587	用户 公司 手机 业务 互联网 网络 运营商 网站 市场 电信	科技
9	966	观众 电影 导演 娱乐 音乐 演员 拍摄 演唱会 角色 节目	娱乐
10	1037	警方 男子 民警 医院 发现 法院 儿子 发生 派出所 调查	社会
11	258	房价 土地 开发商 市场 住房 楼市 地产 成交 政府 投资	房产
12	217	搭配 时尚 黑色 性感 造型 设计 外套 单品 款式 组图	时尚
13	789	家具 品牌 企业 家居 产品 消费者 行业 设计 装修 市场	家居

图 4 THUCNews 使用 DV-sim 文档向量的聚类结果

Fig. 4 Cluster result using DV-sim vector on THUCNews dataset

主题	数量	主题词	类别
-1	435	公司 企业 网友 银行 设计 保险公司 保障 生活 发展 建筑	
0	1082	比赛 球员 球队 火箭 对手 进攻 篮板 防守 表现 联赛	体育
1	1004	游戏 玩家 网游 活动 网络游戏 世界 技能 体验 奖励 系统	游戏
2	532	减肥 女性 食物 身体 运动 脂肪 方法 导语 按摩 瘦身	时尚
3	1166	考生 学生 考试 学校 招生 录取 专业 高考 志愿 教育	教育
4	521	期货 市场 价格 上涨 合约 商品 黄金 现货 需求 经济	财经
5	432	项目 户型 热盘 均价 位于 样板间 相册 为准 优惠 楼盘	房产
6	435	基金 投资 公司 股票 市场 型基金 收益 指数 净值 投资者	财经
7	597	用户 手机 网络 公司 互联网 业务 网站 运营商 电信 服务	科技
8	383	采用 笔记本 像素 英寸 机身 功能 处理器 接口 性能 支持	科技
9	1006	观众 电影 导演 娱乐 音乐 演员 拍摄 演唱会 节目 角色	娱乐
10	997	男子 警方 民警 医院 发现 法院 儿子 发生 派出所 司机	社会
11	230	房价 土地 住房 市场 楼市 开发商 成交 政府 地产 贷款	房产
12	318	项目 生活 区域 别墅 业主 户型 地产 社区 居住 样板间	房产
13	201	搭配 时尚 黑色 性感 设计 外套 造型 单品 款式 组图	时尚
14	661	家具 品牌 企业 家居 产品 消费者 行业 装修 市场 发展	家居

图 5 THUCNews 使用 DV-tfidf 文档向量的聚类结果

Fig. 5 Cluster result using DV-tfidf vector on THUCNews dataset

3 种方法在各类别上的 WF1 值如表 3 所列，DV-sim 和 DV-tfidf 两种特征增强结果对不同类别文档的影响程度不同。

表 3 DV，DV-sim 和 DV-tfidf 在各类别上的 WF1(THUCNews)

Table 3 WF1 of DV, DV-sim and DV-tfidf on each class of THUCNews dataset

	教育	科技	社会	体育	房产	时尚	游戏	财经	娱乐	家居
DV	0.896	—	0.815	0.982	0.889	0.886	0.688	0.88	0.877	0.827
DV-sim	0.903	0.847	0.795	0.978	0.911	0.890	0.916	0.923	0.868	0.853
DV-tfidf	0.902	0.861	0.807	0.978	0.895	0.892	0.932	0.909	0.871	0.830

在 Sogou 上的聚类性能如表 4 所列。

表 4　Sogou 上的聚类性能

Table 4　Cluster performance on Sogou dataset

文档向量	簇类数目	主题数目	噪声数量	ACC	MF1
DV	11	10	2586	0.6503	0.7372
DV-sim	12	10	1018	*0.7841*	*0.8226*
DV-tfidf	12	10	1046	0.7669	0.8051

在 Sogou 数据集上,DV-sim 方法得到的聚类结果如图 6 所示,其中 8 号和 9 号分别与身体健康和生理心理健康相关。相比 DV 方法,DV-sim 的噪声数减少 60.63%,ACC 和 WF1 分别提高 20.58% 和 11.58%;DV-tfidf 的噪声数减少 59.55%,ACC 和 WF1 分别提高 17.93% 和 9.21%,DV-sim 增强向量的聚类性能更好。

主题	数量	主题词	类别
-1	1018	孩子 学生 教师 教育 学校 学习 社会 家长 发展 老师	
0	1005	比赛 球队 球员 队员 联赛 对手 主场 俱乐部 冠军 决赛	体育
1	1018	导弹 军事 作战 部队 飞机 演习 海军 系统 武器 训练	军事
2	913	市场 成交 楼市 房价 调控 住房 政策 上涨 土地 项目	房产
3	1117	旅游 游客 旅行社 图库 酒店 航班 旅客 文化 活动 景区	旅游
4	843	公司 股东 股权 股份 行情 证券 市场 投资 有限公司 机构	财经
5	788	车型 汽车 配置 报价 图片 品牌 销量 市场 经销商 发动机	汽车
6	860	用户 手机 公司 市场 业务 产品 服务 资费 软件 企业	科技
7	904	民警 警方 发生 现场 男子 犯罪 司机 发现 人员 事故	社会
8	777	治疗 医院 患者 药品 疾病 手术 病人 药物 医生 女性	健康
9	158	男人 女人 喜欢 女性 心理 生活 职业 事情 老板 感觉	健康
10	150	考试 复习 考生 知识 阅读 时间 高考 学生 考研 作文	教育
11	449	考生 招生 专业 志愿 录取 学校 考试 高校 学生 大学	教育

图 6　Sogous 使用 DV-sim 文档向量的聚类结果

Fig. 6　Cluster result using DV-sim vector on Sogou dataset

3 种方法在各类别上的 WF1 值如表 5 所列。

表 5　DV,DV-sim 和 DV-tfidf 在各类别上的 WF1(Sogou)

Table 5　WF1 of DV,DV-sim and DV-tfidf on each class of Sogou dataset

	科技	社会	健康	体育	军事	旅游	汽车	财经	房产	教育
DV	0.614	0.560	0.668	0.949	0.924	0.743	0.746	0.651	0.719	0.798
DV-sim	0.794	0.682	0.794	0.976	0.945	0.830	0.845	0.803	0.865	0.692
DV-tfidf	0.795	0.629	0.658	0.976	0.950	0.860	0.78	0.806	0.806	0.791

(2)Women 数据集上的主题聚类

由于不同时期妇联工作重点不同,语料涉及的主题是不均衡的,因此将 HDBSCAN 算法的 min_count 参数设置为 50,即每个簇类最少有 50 个样本。另外,对照中华全国妇女联合会网站上介绍的妇联工作重点,可以得到语料涉及的主题包括妇联工作、农村工作、社区工作、教育工作、医护工作、农业生产、企业经营、科研创新、法律维权、公安执法、抢险救灾、一线工人(包括环卫工人、技术工人等)、服务工作(包括交通、银行、税务等)、文艺工作、爱心助人(包括敬老孝亲等)、体育竞赛、部队训练等方面。

使用 DV 向量聚类,噪声数为 1906,有 23 个簇类;使用 DV-sim 增强的文档向量聚类,结果如图 7 所示,噪声数减少 43.07%,增加了 11 号和 17 号主题,簇类划分更细。图 8 所示为层次聚类图,可以看出各个簇类之间的距离关系,其中 0 号和 1 号、9 号和 10 号、21 号和 22 号距离都比较近,内容的相关性很高。根据各个簇类的主题词和层次聚类图可以更直观地选择合适的主题数目。

主题	数量	主题词	类别
-1	1085	孩子 生活 妈妈 女儿 儿子 儿童 母亲 老人 爱心 家政	
0	1092	种植 养殖 技术 农民 发展 农业 蔬菜 养猪 创业 基地	农业生产
1	716	村民 群众 书记 全村 农村 发展 妇女 建设 村官 村干部	农村工作
2	72	妇女 妇女儿童 维权 活动 组织 主席 家庭 建设 培训 群众	妇联工作
3	113	地震 受灾 群众 灾区 灾情 防汛 救人 灾民 转移 救灾	抢险救灾
4	367	案件 法律 当事人 法官 法院 调解 律师 审判 办案 法庭	法律维权
5	98	清扫 垃圾 环卫工人 路段 打捞 路面 职工 环卫处 公路 扫帚	一线工人
6	116	旅客 乘客 服务 公交 铁路 车厢 公交车 列车 老人 客运	服务工作
7	225	社区 居民 小区 老人 服务 群众 党员 辖区 志愿者 街道	社区工作
8	239	民警 群众 派出所 执法 警察 嫌疑人 公安 社区 案件 女子	公安执法
9	360	病人 患者 医院 护理 医生 护士 治疗 手术 科室 医疗	医务工作
10	223	疫情 防控 医院 患者 肺炎 口罩 隔离 新冠 社区 抗疫	医务工作
11	53	客户 业务 服务 用户 公司 员工 支行 营销 营业部 银行	服务工作
12	117	纳税人 服务 税收 税务 审计 地税 业务 办税 系统 分局	服务工作
13	110	比赛 训练 教练 运动员 冠军 金牌 女子 选手 成绩 参加	体育竞赛
14	162	部队 女兵 战士 军人 训练 丈夫 军嫂 官兵 民兵 飞行	部队训练
15	256	公司 技术 车间 设计 工程 生产 员工 操作 项目 企业	一线工人
16	56	残疾人 轮椅 盲人 父母 朋友 按摩 生命 生活 母亲 身体	自强自立
17	55	妈妈 孩子 爱心 资助 女童 儿童 救助 母亲 社会 生活	爱心助人
18	479	学生 老师 学校 孩子 教师 教学 教育 小学 家长 学习	教育工作
19	101	研究 科学家 病毒 科研 领域 实验室 国家 科学 国际 基因	科研工作
20	875	企业 公司 创业 员工 产品 发展 市场 经营 生产 有限公司	企业经营
21	77	老人 院长 老年公寓 照顾 护理 院民 护理员 生活 服务 困女	爱心助人
22	245	婆婆 老人 丈夫 照顾 公公 家庭 儿子 媳妇 生活 公婆	爱心助人
23	191	制绣 作品 艺术 手工 剪纸 技艺 创作 编织 传统 制作	手工技艺
24	167	演出 文化 舞蹈 艺术 表演 音乐 演员 节目 观众 文艺	文艺工作

图 7　Women 使用 DV-sim 文档向量的聚类结果

Fig. 7　Cluster result using DV-sim vector on Women dataset

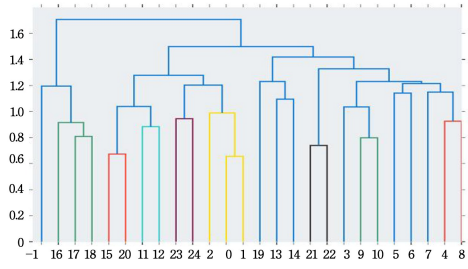


图 8　Women 使用 DV-sim 文档向量的层次聚类图

Fig. 8　Hierarchical cluster diagram on Women dataset

从长度 1000～3000 的语料中选择 10 个样本,DV 和 DV-sim 的聚类结果如表 6 所列。

表 6　DV 和 DV-sim 在 Women 数据集部分样本上的聚类结果

Table 6　Cluster results of DV and DV-sim on some samples of Women dataset

url	title	DV	DV-sim
http://www.women.org.cn/art/2009/2/17/art_24_139836.html	许惠芬:“看现场”的女探长	-1	公安执法
http://www.women.org.cn/art/2017/7/13/art_24_151491.html	杜丽群坚守抗艾一线为患者带来“绝地阳光”	-1	医务工作
http://www.women.org.cn/art/2009/7/16/art_24_143315.html	钱虹:地产“律政佳人”	-1	法律维权
http://www.women.org.cn/art/2009/6/1/art_24_140132.html	应景芳:“轮椅姑娘”书写绚丽人生	-1	自强自立
http://www.cnwomen.com.cn/2021/01/27/99219765.html	清除网上政务服务“僵尸”只是及格线	公安执法	服务工作
http://www.women.org.cn/art/2007/12/17/art_24_141245.html	魏翠:青春献环卫 我心永不悔	服务工作	一线工人
http://www.women.org.cn/art/2008/12/3/art_24_143702.html	程叶兰:缤纷银杏梦	爱心助人	农业生产
http://www.women.org.cn/art/2007/12/14/art_24_141278.html	冯菊荷:助百名贫困学子尽“爱心妈妈”职责	教育工作	爱心助人
http://www.women.org.cn/art/2011/4/14/art_24_142919.html	祝菠:“三味”女所长	社区工作	公安执法
http://www.women.org.cn/art/2014/5/27/art_24_145643.html	武汉女特警:本领过硬 性格可爱	部队训练	公安执法

结束语　文本特征表示是影响主题聚类效果的关键因素,利用 Doc2Vec 模型可以直接获取长文档的全局语义表征,为减少长文档中噪声的影响,本文利用文档开始和结尾部分特定词性的内容构建增强的文档特征表示,DV-sim 基于

语义相似度获得增强的文档特征，DV-tfidf 基于词频获得增强的文档特征，THUCNews 和 Sogou 是两个带标签数据集，数据质量不完全相同，但聚类结果都表明 DV-sim 和 DV-tfidf 两种方法均优于 DV 方法，且 DV-sim 的聚类性能更好，可以提升主题区分度，降低噪声影响，提高聚类准确性和 F1-Score。将 DV-sim 方法应用于优秀女性人物报道也获得了很好的主题聚类结果。

不同数据集上的聚类结果表明采用增强的文档特征可以提高主题聚类效果，基于语义的特征和基于词频的特征在不同数据集上对不同类别文档的影响程度不同，后续将进一步研究如何将两种方法融合在一起以获得更好的聚类性能。

参 考 文 献

[1] ZHAO J S, SONG M X, GAO X, et al. Research on Text Representation in Natural Language Processing [J]. Journal of Software, 2022, 33(1): 102-128.

[2] XIONG H X, YANG M T, LI Y Y. A Survey of Information Organization and Retrieval Based on Deep Learning[J]. Information Science, 2020, 38(3): 3-10.

[3] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/ OL]. https://arxiv.org/pdf/1810.04805.pdf.

[4] LE Q, MIKOLOV T. Distributed Representations of Sentences and Documents[C]// Proceedings of the 31st International Conference on Machine Learning. PMLR, 2014, 32(2): 1188-1196.

[5] LEE S, JIN X, KIM W. Sentiment classification for unlabeled dataset using Doc2Vec with JST[C]// Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart Connected World. ACM New York, NY, USA, 2016.

[6] MANDAL A, GHOSH K, GHOSH S, et al. Unsupervised approaches for measuring textual similarity between legal court case reports. Artificial Intelligence and Law[J]. Artificial Intelligence and Law, 2021, 29(3): 417-451.

[7] ADORNO H G, DURAN J, SIDOROV G, et al. Document embeddings learned on various types of n-grams for cross-topic authorship attribution[J]. Computing. 2018, 100(7): 741-756.

[8] JIA X T, WANG M Y, CAO Y. Automatic Abstracting of Chinese Document with Doc2Vec and Improved Clustering Algorithm[J]. Data Analysis and Knowledge Discovery, 2018, 2(2): 86-95.

[9] RUAN G C, XIA L. Hot Topic Detection in Journal Papers Based on Doc2Vec[J]. Information Studies: Theory & Application, 2019, 42(4): 107-111, 106.

[10] ARIF B, REZA R, NOVYANTARA P H, et al. Unsupervised News Topic Modelling with Doc2Vec and Spherical Clustering [J]. Procedia Computer Science, 2021, 179: 40-46.

[11] CHANG W B, XU Z Z, ZHOU S H, et al. Research on detection methods based on Doc2vec abnormal comments[J]. Future Generation Computer Systems, 2018, 86: 656-662.

[12] AMIRI M Z, SHOBI A. A Link Prediction Strategy for Personalized Tweet Recommendation through Doc2Vec Approach[C]// Proceedings of 17th International Conference on IT Applications and Management, Babolsar, Iran. Korean Database Society(KDBS), 2017: 72-82.

[13] CHEN X, ZHU X D, GAO G K, et al. Sentiment Analysis of Chinese Comments Based on Hybrid Vector Model[J]. Computer Engineering, 2020, 46(1): 309-314.

[14] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv: 1301.3781, 2013.

[15] TANG M, ZHU L, ZOU X C. Document Vector Representation Based on Word2Vec[J]. Computer Science, 2016, 43(6): 214-217, 269.

[16] CAMPELLO R, MOULAVI D, SANDER J. Density Based Clustering Based on Hierarchical Density Estimates[C]// Advances in Knowledge Discovery and Data Mining(PAKDD 2013). Gold Coast, Australia. Springer, 2013: 160-172.

[17] MELVIN R L, XIAO J J, GODWIN R, et al. Visualizing correlated motion with HDBSCAN clustering[J]. Protein Science, 2018, 27(1): 62-75.

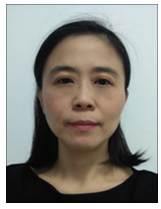
[18] TAHVILI S, HATVANI L, FELDERER M, et al. Automated Functional Dependency Detection Between Test Cases Using Doc2Vec and Clustering [C]// Proceedings of 2019 IEEE International Conference on Artificial Intelligence Testing(AITest). Newark, CA, USA. IEEE, 2019.

[19] ESTER M, KRIEGEL H P, SANDER J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C] // Proceedings of International Conference on Knowledge Discovery and Data Mining. AAAI, 1996: 226-231.

[20] WU H, AI S H, KA H R J, et al. Method of computing Chinese sentence similarity based on part-of-speech feature[J]. Computer Engineering and Design, 2020, 41(1): 150-155.

[21] MCINNES L, HEALY J, MELVILLE J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [J]. arXiv: 1802.03426, 2018.

[22] ASYAKY M S, MANDALA R. Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP[C] // Proceedings of 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications(ICAICTA). Bandung, Indonesia. IEEE, 2021.



CHEN Jie, born in 1969, postgraduate, associate professor. Her main research interests include information processing and text mining.