融合近邻标题图的涉案新闻话题发现

卢天旭[1,2]，余正涛[1,2]，黄于欣[1,2+]

（1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500）

摘 要：针对涉案舆情领域同一案件下不同话题的新闻文档要素信息较为接近，已有的话题发现方法不能很好地进行表征和区分的问题，提出融合近邻标题图的涉案新闻话题发现方法。在话题发现的过程中引入标题的关联关系，构建近邻标题图，通过图卷积网络提取标题的全局特征，同时使用深度网络提取文档的局部特征，加入到标题的编码过程中去，更好地实现聚类。实验结果表明，联合标题和文档进行话题建模可以提升涉案新闻话题发现的准确性指标。

关键词：涉案新闻；话题发现；要素信息；聚类；近邻标题；图卷积；预训练模型

中图法分类号：TP391.1 **文献标识号**：A **文章编号**：1000-7024（2022）05-1249-09

doi：10.16208/j.issn1000-7024.2022.05.007

Topic discovery for case-related news combined with neighbor title graph

LU Tian-xu[1,2]，YU Zheng-tao[1,2]，HUANG Yu-xin[1,2+]

（1. Faculty of Information Engineering and Automation，Kunming University of Science and Technology，Kunming 650500，China；2. Yunnan Key Laboratory of Artificial Intelligence，Kunming University of Science and Technology，Kunming 650500，China）

Abstract：Aiming at the problem that the information of news documents of different topics in the same case in the field of public opinion involved is relatively close，and the existing topic discovery methods cannot be well characterized and differentiated，a method of topic discovery for case-related news combined with neighbor title graph was proposed. In the process of topic discovery，the association relationship of the title was introduced，the neighbor title graph was constructed，the global feature of the title was extracted through the graph convolution network，and the local feature of the document was extracted using the deep network，which was added to the encoding process of the title，so as to better achieve clustering. Experimental results show that the topic modeling with joint titles and documents can improve the accuracy index of case-related news topic discovery.

Key words：case-related news；topic discovery；element information；clustering；neighbor titles；graph convolution；pretrained model

0 引 言

涉案舆情由于其涉案的特殊性，通常具备敏感性和易爆发性，如何有效地进行涉案舆情监管是一个关键问题。而涉案话题包含了涉案舆情信息的准确凝练和大多数网民的关注点，及时发现涉案新闻的话题并疏导涉案舆情对于维护社会稳定而言至关重要。涉案新闻话题发现是指在司法案件相关的新闻信息中，针对同一案件把描述相同话题的新闻信息归到同一个话题簇中，可以转化为一个话题级的聚类任务。目前现有的话题发现模型主要是通过对文档进行表征和使用聚类算法计算文档相似度度量这两个问题上实现的。通过研究[1-3]发现这些方法在处理大规模涉案新

收稿日期：2022-01-01；**修订日期**：2022-03-24

基金项目：国家重点研发计划基金项目（2018YFC0830105、2018YFC0830100）；国家自然科学基金项目（61972186、61762056、61472168）；云南省重大科技专项计划基金项目（202002AD080001）；云南省高新技术产业专项基金项目（201606）；云南省基础研究专项面上基金项目（202001AT070046、202001AT070047）；云南省应用基础研究计划重点基金项目（2019FA023）

作者简介：卢天旭（1996-），男，北京人，硕士研究生，CCF 学生会员，研究方向为自然语言处理；余正涛（1970-），男，云南曲靖人，教授，博士生导师，CCF 高级会员，研究方向为自然语言处理、信息检索和机器翻译；+通讯作者：黄于欣（1983-），男，河南洛阳人，博士研究生，CCF 学生会员，研究方向为自然语言处理、文本摘要等。E-mail：huangyuxin2004@163.com

闻语料数据时依赖词频统计信息，表征质量不高，对于同一案件不同话题下的新闻文档，无法区分共现词较少但属于同一话题的情况，且使用的聚类方法对数据输入顺序敏感。此外，应用主题模型在话题检测发现、热点主题挖掘以及子话题关联等相关研究任务上也取得了一定的效果。但通过研究[4,5]发现，这些方法捕获的主题信息由于相似度过高而被归为同一个主题下，同样不能够很好地区分同一案件不同话题下的新闻文档，这些研究表明了话题发现任务很大程度上依赖于文档的表征能力。因此，认为提高涉案新闻文本表征的能力才能得到质量更好的涉案新闻话题簇，从而提高话题发现的准确性。

1 相关研究

近年来国内外学者针对涉案领域话题发现研究较少，在通用领域，目前话题发现方法集中于使用传统聚类模型、主题模型以及改进型的聚类模型等方法实现。

基于传统聚类模型的话题发现方法旨在利用基于划分、密度、增量等经典的聚类算法来计算文档样本之间的欧氏距离，根据相似度度量实现话题发现。Nur'aini 等[6]使用经典 K-means 聚类算法实现了 Twitter 社交媒体话题发现；Mustakim 等[7]使用基于密度的应用程序空间聚类 DBSCAN (density-based spatial clustering of applications with noise) 算法对 Twitter 文本数据进行聚类，挖掘社交媒体中用户近期感兴趣的热点话题；Zhang 等[8]提出了一种基于多视图文本语义和 Single-Pass 聚类算法的话题发现方法，在财经新闻数据集中，通过融合模型的特征可以实现从海量数据中获取对投资者有效的话题信息。

基于主题模型的话题发现方法通过 LDA (latent dirichlet allocation) 等常见的主题模型以及衍生模型，基于词袋模型考虑词条的共现，生成新闻文档的主题分布。Rortais 等[9]使用 LDA 主题模型快速检测媒体中特定的食品欺诈事件，通过探索大量文档，发现与欺诈事件相关的话题，并组织和总结识别其中包含的话题的文本文档；Kumar 等[10]提出了一种用于短文本流聚类的在线语义增强 Dirichlet 主题模型，将语义信息集成到一个新的图形模型中，并在每个输入的短文本中自动聚类，解决话题发现中短文本语义稀疏问题；Fan 等[11]提出了一种基于分层贝叶斯非参数框架在线新闻话题发现和跟踪方法，该方法允许在语料库中的不同新闻故事之间共享话题，应用于在线新闻数据流上取得了一定的效果。

基于改进型聚类模型的话题发现方法是在经典的聚类算法的基础上，融入其它模块以增强数据的表示，解决经典聚类算法自身的缺陷。Li 等[12]提出了一种基于时间窗口的改进的基于密度的 DBSCAN 算法，以实现更加准确的话题发现，并具有降低时间复杂度的辅助优势；Xiao 等[13]提出了一种基于图形分解的新型文档表示方法，将每个新闻文档分解为不同的语义单元，然后构建语义单元之间的关系以形成胶囊语义图，最后通过 Single-Pass 算法实现新闻文档的话题发现；Wu 等[14]基于 BTM (bayesian sparsetopic model) 和 GloVe (global vectors) 相似性线性融合的方法，将微博短文本分别使用 BTM 模型和 GloVe 词向量建模，计算两种不同的相似度，将两种相似度线性融合作为距离函数，实现 K-means 聚类，提高了微博短文本话题发现精度。

已有的话题发现方法在处理通用领域的任务时已经取得了不错的效果，但是在涉案领域的话题发现任务上效果表现较差，这是由于这些方法所使用相似度度量方法在计算高维数据时效率偏低，且不具备较强的涉案新闻表征能力。

近年来深度学习在大规模数据表征和处理方面表现突出，在聚类算法上融入深度学习强大的表征能力越来越受到重视。Xie 等提出模型的聚类不是从数据本身来聚类，而是学习到数据到隐空间的映射，然后设置了聚类优化目标来学习隐空间的聚类；Yang 等[15]未采用以往方法的先降维再进行聚类的模式，认为联合这两个过程可以得到更好的聚类效果，提出一种基于深度网络降维和 K-means 聚类的联合优化准则。这些方法在涉案新闻表征能力上都明显强于以往的话题发现方法，因此本文考虑将深度学习融入聚类算法，并应用到涉案新闻话题发现任务中以提高模型的准确性。本文在学习数据样本的表征中考虑到了数据样本之间关系的重要性，提出一种融合近邻标题图的涉案新闻话题发现方法，既考虑到新闻文档数据自身的特征，又学习到标题关系间的潜在相似性，通过深度网络和图卷积网络学习的表征融合，来提高涉案新闻话题发现聚类的准确性。

2 融合近邻标题图的涉案新闻话题发现方法

2.1 模型框架

针对已有的话题发现方法在涉案新闻话题发现任务中准确度不高，难以区分同一案件话题下新闻信息的问题，本文提出融合近邻标题图的涉案新闻话题发现模型，模型框架如图 1 所示。该模型主要分为 5 部分，分别为标题编码模块、近邻标题图的构建、文档特征提取模块、标题结构信息提取模块和指导模块。

2.2 标题编码模块

标题编码模块用于编码涉案新闻话题数据集中标题部分，通过 BERT (bidirectional encoder representation from transformers) 预训练模型[16]训练完成后能够获得标题的表征，以便接下来构建近邻标题图。BERT 模型是由多个 Transformer 模型[17]组合而成的，其训练方式分为两个任务：

其一是随机选择 15% 的词用于预测，其中 80% 采用

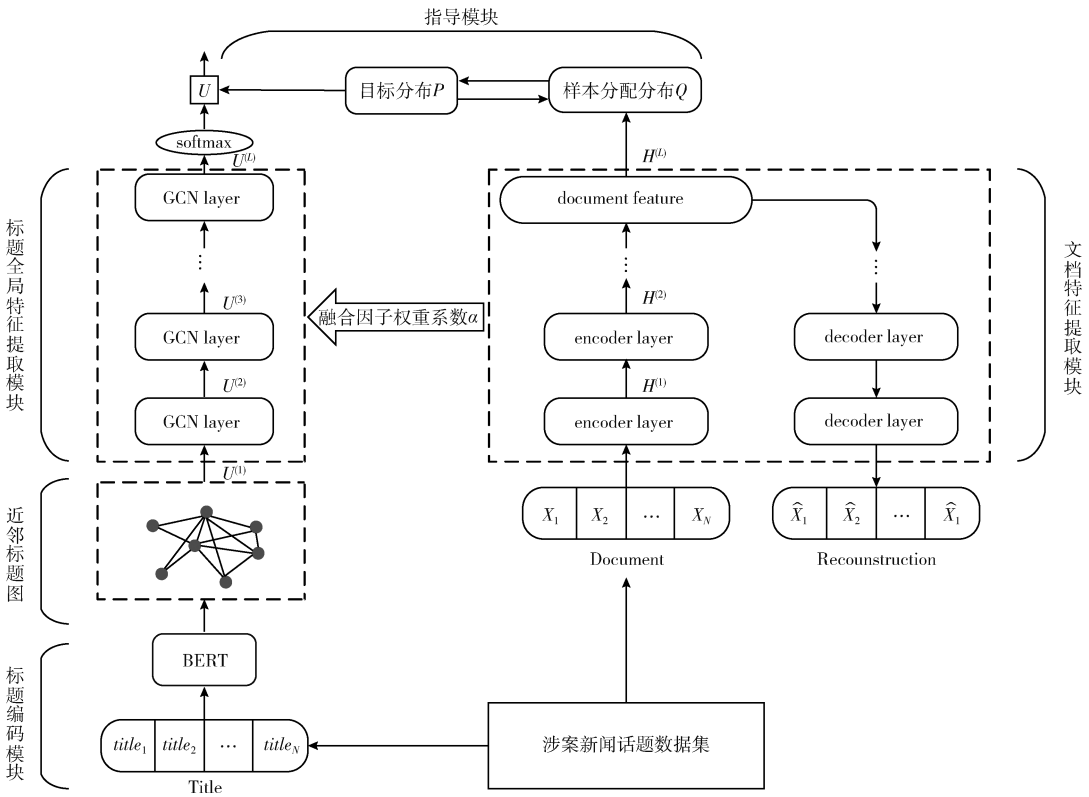

图 1　融合近邻标题图的涉案新闻话题发现模型

MASK 符号遮盖，10%用随机词替换，其余保持不变，这使得模型倾向于依赖上下文来预测词汇，具备一定的纠错能力；其二是预测两句话是否为连贯文本。因此 BERT 模型在结束训练后能够获得涉案新闻标题的单词表征和句子表征。Transformer 模型结构如图 2 所示。

具体如下，设涉案新闻话题数据集中标题 $Title$ 数量为 N，$Title=\{title_1,title_2,\cdots,title_N\}$，每条涉案新闻标题长度为 S，$E=\{e_1,e_2,\cdots,e_S\}$ 为每条标题中词向量的集合，将标题的词向量输入到 BERT 模型中进行编码，可以得到每条标题的向量表征。以编码一条标题为例，编码过程如图 3 所示。

BERT 模型要求每条标题输入的词元表征必须含有 3 种类型的嵌入，即词元嵌入 r_{wordi}、片段嵌入 r_A 和位置嵌入 r_i，每条标题的词元前都有一个［CLS］标记用来表示整个标题句子。将词向量集合 E 输入到 BERT 模型中，经过多层 Transformer 网络得到每个词元各自的表征。其中位于输出起始位置的［CLS］表征 T_i 即为整个标题句子的向量表征。将所有标题的词向量分别输入到 BERT 模型中编码，最终得到融合语义信息后的标题向量表征集合 T，$T=\{T_1,T_2,\cdots,T_N\}$。

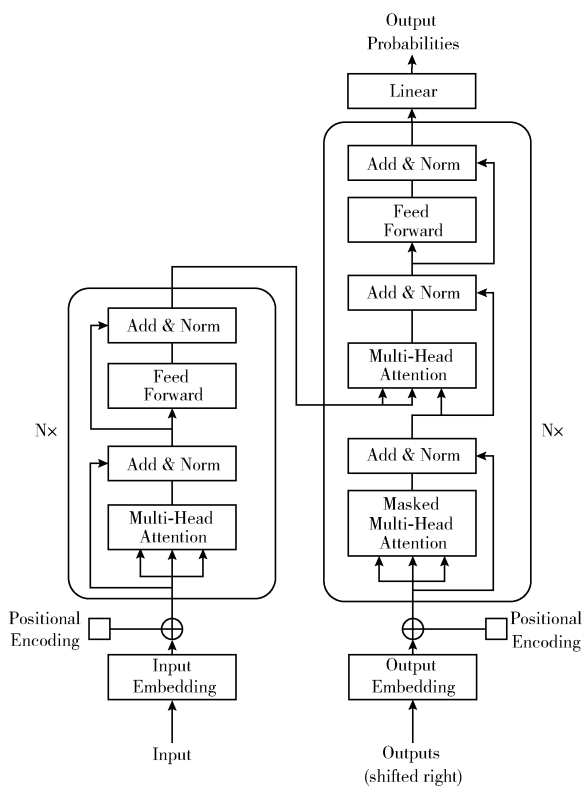

图 2　Transformer 模型结构

2.3　近邻标题图

近邻标题图构建模块采用 K 近邻算法构建近邻标题图来提取标题的全局特征。设标题数据 $T\in R^{N\times a}$，其中每行 T_i 代表第 i 个标题样本，N 是样本数，a 代表维度。对于

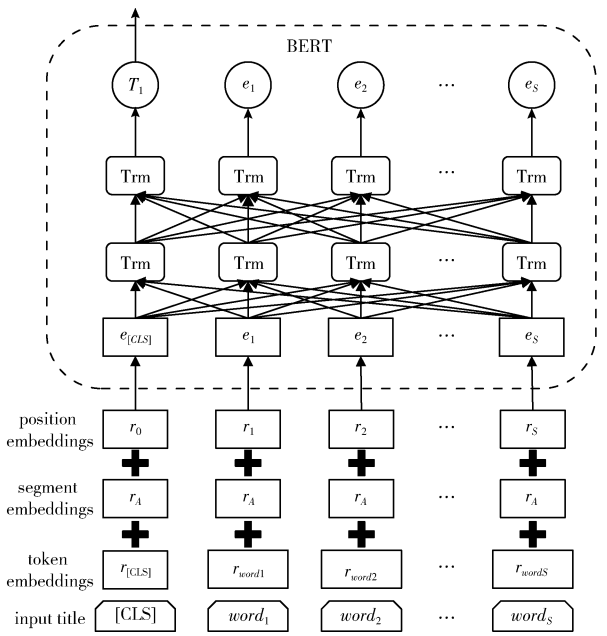

图 3 BERT 编码涉案新闻标题模型结构

每个标题样本，首先找到它的前 K 个相似度最高的邻居作为邻居节点，并通过边来连接，以构成近邻标题图。利用向量的点积运算来计算任意两个标题之间的相似度矩阵 S_{ij}，它是一个 $N\times N$ 维矩阵，如式（1）所示

$$S_{ij}=(t_j^{bert})^{\mathrm{T}}t_i^{bert} \tag{1}$$

对于任意两个标题节点 t_i 和 t_j，令 w_{ij} 为节点之间的权重。如果节点之间有边相连，则 $w_{ij}>0$，若没有边相连，则 $w_{ij}=0$。由于我们构建的近邻标题图是无向权重图，因此 $w_{ij}=w_{ji}$。图中任意节点的度为和它连接的所有边的权重之和，定义如式（2）所示

$$d_i=\sum_{j=1}^{n}w_{ij} \tag{2}$$

通过计算每个节点的度，得到一个只有主对角线有值的节点度矩阵 $D\in R^{N\times N}$，如式（3）所示

$$D=\begin{pmatrix} d_1 & \cdots & \cdots & \cdots \\ \cdots & d_2 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & d_n \end{pmatrix} \tag{3}$$

主对角线的值表示第 i 行第 i 个点的度数。计算所有节点之间的权重，得到 $N\times N$ 维的邻接矩阵 M，其第 i 行第 j 个元素就是权重 w_{ij}，$w_{ij}=s_{ij}$。

2.4 文档特征提取模块

文档特征提取模块的作用是提取涉案新闻话题数据集中文档的局部特征，本文使用深度神经网络自编码器来学习有效的数据表征。自编码器是一种表示模型，利用输入数据作为参考，不利用标签监督，以用来提取特征和降维。自编码器将输入映射到特征空间，再映射回输入空间进行数据重构。设自编码器有 L 层，编码器学到的第 L 层的表征如式（4）所示

$$H^{(l)}=\sigma(W_{enc}{}^{(l)}H^{(l-1)}+b_{enc}{}^{(l)}) \tag{4}$$

其中，σ 为 relu 函数，$W_{enc}{}^{(l)}$ 为编码器中第 l 层的变换矩阵，$b_{enc}{}^{(l)}$ 为偏置。$H^{(0)}$ 表示为原始文档数据 X。

解码器部分将特征映射回输入空间，得到原始数据的重构 $\hat{X}$，如式（5）所示

$$H^{(l)}=\sigma(W_{dec}{}^{(l)}H^{(l-1)}+b_{dec}{}^{(l)}) \tag{5}$$

$W_{dec}{}^{(l)}$ 为解码器中第 l 层的变换矩阵，$b_{dec}{}^{(l)}$ 为偏置，重构数据 $\hat{X}=H^{(L)}$。

文档特征提取模块的损失函数如式（6）所示

$$\mathcal{L}_{dfe}=\frac{1}{2N}\sum_i\|x_i-\hat{x}_i\|_2^2 \tag{6}$$

通过最小化重构误差和梯度下降算法不断优化网络参数进行训练。

2.5 图卷积神经网络

图神经网络 GNN（graph neural network）是一类处理图结构信息的方法的统称，其中代表方法是图卷积神经网络。图卷积神经网络是一个对图数据进行特征提取的多层神经网络。传统的卷积神经网络可以处理有规则空间结构的数据，这些数据的结构可以用一维和二维的矩阵来表示。然而许多数据是不具备规则的空间结构的，传统的卷积神经网络就不能处理这些数据。在不规则空间结构的图数据中，每个节点有属于自己的特征信息，每个节点还具有结构信息且图的形状不规则，邻居节点也不固定。图卷积网络可以从这类数据中提取特征，得到图的嵌入表示，从而实现边预测、节点分类等任务。在模型计算过程中，图卷积网络结构如图 4 所示。

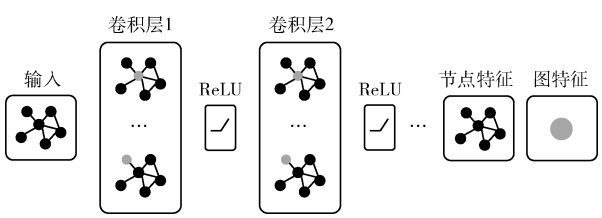

图 4 图卷积网络结构

网络首先对节点的特征进行抽取，将每个节点自身的属性信息变换后传送给邻居节点，每个节点收集邻居节点的特征，融合局部结构信息，聚集结构信息和属性信息后做非线性变换以增强网络的表达能力。图卷积网络处理图数据具有以下优势，首先网络中节点的表征与下游任务具有很好的适应性，节点表征与下游任务被统一到一个模型端到端训练，监督信号可以同时指导卷积层与分类层更新参数。其次图卷积网络可以同时学习节点的属性信息与结构信息，使它们协同影响节点的最终表征。

2.6　标题全局特征提取模块

2.4 节提到的文档特征提取模块能够从涉案新闻话题数据集的文档中提取有用的表征，但自编码器只提取到了文档局部特征，不能提取到样本之间的关联关系。2.3 节构建的近邻标题图蕴含了大量的标题全局结构信息，使用图卷积网络提取近邻标题图中的结构特征，并将自编码器提取到的文档局部特征集成到图卷积网络中，这样模型就可以同时提取到数据的两种不同特征。图卷积网络第 l 层提取的表征通过卷积运算得到，如式（7）所示

$$U^{(l)}=\sigma(\tilde{\mathcal{L}}_{sym}U^{(l-1)}W^{(l-1)}) \tag{7}$$

其中，$\tilde{\mathcal{L}}_{sym}=\tilde{D}^{-\frac{1}{2}}\tilde{M}\tilde{D}^{-\frac{1}{2}}$ 为归一化的拉普拉斯矩阵，$\tilde{M}=M+I$，$\tilde{D}_{ii}=\sum_j\tilde{M}_{ij}$。$I$ 为邻接矩阵 M 的单位对角阵，D 为节点度矩阵。$\tilde{\mathcal{L}}_{sym}$ 将图卷积网络学到的前一层的表征 $U^{(l-1)}$ 向下一层传播得到新的表征 $U^{(l)}$。

本文为了使图卷积网络学习到的涉案新闻话题数据特征同时具有标题的全局特征和文档的局部特征，将两种表征 $U^{(l-1)}$ 和 $H^{(l-1)}$ 通过融合因子结合在一起，得到一种更全面的数据表征，如式（8）所示

$$\tilde{U}^{(l-1)}=\alpha U^{(l-1)}+(1-\alpha)H^{(l-1)} \tag{8}$$

式中：α 是平衡两种表征的权重系数，通过融合因子逐层连接自编码器和图卷积网络可以将文档的局部特征有效融合到标题的全局特征中。融合两种表征后，将 $\tilde{U}^{(l-1)}$ 输入到图卷积网络中得到表征 $U^{(l)}$，如式（9）所示

$$U^{(l)}=\sigma(\tilde{\mathcal{L}}_{sym}\tilde{U}^{(l-1)}W^{(l-1)}) \tag{9}$$

以此类推得到图卷积网络最后一层输出的表征 $U^{(L)}$。网络的输出端连接了一个 softmax 多分类器，最终输出的结果如式（10）所示

$$U=\mathrm{softmax}(\tilde{\mathcal{L}}_{sym}U^{(L)}W^{(L)}) \tag{10}$$

模型得到的结果 U 是一个概率分布，其元素 u_{ij} 表示涉案新闻样本 i 属于簇中心 j 的概率。

2.7　指导模块

在上一节中已经将自编码器和图卷积网络学习到的表征通过融合因子结合了起来，并且得到了概率分布 U。但是自编码器的作用主要是用来学习文档的局部表征，是一种无监督的学习，而图卷积网络主要用来学习标题的关系特征，它们都不是直接用来做聚类任务的，需要在表征中引入聚类信号。因此本文使用指导模块将两个模块统一到一个框架中同时进行端到端的聚类优化训练。

对于第 i 个样本和第 j 个簇，引用自由度为 1 的 *student-t* 分布作为核函数衡量自编码器的表征 h_i 和簇心 μ_i 之间的距离，如式（11）所示

$$q_{ij}=\frac{(1+\|h_i-\mu_j\|^2)^{-1}}{\sum_j(1+\|h_i-\mu_j\|^2)^{-1}} \tag{11}$$

其中，h_i 表示 $H^{(L)}$ 的第 i 行，μ_i 是经过 K-means 算法初始化后的簇心。我们将 q 视为文档样本 i 被分配到簇 j 的概率，Q 即为所有文档样本分配到簇的分布。

为了得到高置信度的分配来迭代聚类结果，提高聚类准确度，构造一个目标分布 P 来辅助模型训练，如式（12）所示

$$p_{ij}=\frac{q_{ij}^{\ 2}/\sum_i q_{ij}}{\sum_j(q_{ij}^{\ 2}/\sum_i q_{ij})} \tag{12}$$

在目标分布 P 中，每一个在文档样本分配分布 Q 中的聚类分配都被先平方再归一化处理，这样可以获得更高置信度的聚类分配，迫使簇内的样本更加接近簇心，簇与簇间的距离最大化，分配更加清晰。指导模块的损失函数之一为分布 Q 和目标分布 P 之间的 KL 散度损失，如式（13）所示

$$\mathcal{L}_{clustering}=\mathrm{KL}(P\|Q)=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{13}$$

通过最小化损失函数更新参数，目标分布 P 使自编码器学习到更接近簇心的样本文档聚类表征。

为了使标题全局特征提取模块和文档特征提取模块在训练迭代过程中趋于一致，需要将两个模块统一在同一目标分布中，因此也可以使用目标分布 P 指导图卷积网络输出的蕴含标题全局特征的样本分布 U。指导模块的损失函数之二为分布 U 和目标分布 P 之间的 KL 散度（Kullback-Leibler divergence）损失，如式（14）所示

$$\mathcal{L}_{tgfe}=\mathrm{KL}(P\|U)=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{u_{ij}} \tag{14}$$

通过指导模块的不同权重参数可以将两种不同表征的聚类分配统一在同一个损失函数中，模型的整体损失函数如式（15）所示

$$\mathcal{L}=\mathcal{L}_{dfe}+\beta\mathcal{L}_{clustering}+(1-\beta)\mathcal{L}_{tgfe} \tag{15}$$

β 为平衡损失函数一和损失函数二的权重参数。整个模型经过训练达到稳定后，可以将图卷积网络最终输出的聚类分布 U 作为涉案新闻话题发现的最终结果。

3　实验结果与分析

3.1　涉案新闻话题数据集

涉案新闻话题发现任务属于针对司法案件特定领域的任务，目前尚未有公开的涉案新闻话题数据集。因此本文在自行构建的涉案新闻话题数据集的基础上开展具体工作。

本文通过分析“百度新闻”、“新浪新闻”、“今日头条”等各大新闻网站和公众号平台近年来的涉案重点新闻，选取了“奔驰车主维权案”、“孙小果涉黑案”等十余个网民关注度较高的案件进行涉案新闻话题数据集的构建。使用爬虫技术根据新闻网站上的案件相关话题和案件关键词爬取有关的新闻数据，通过对爬取的新闻进行分析使每条涉

案新闻只属于一个案件话题，人工标注新闻与哪个案件话题相关，经过数据筛选和预处理，保存为 json 格式的文件。数据的筛选和预处理过程包括对新闻数据和案件话题相关性的人工校准，去除非案件话题相关的数据和重复的数据，去除特殊符号和链接等。最终得到每条清晰、准确的涉案新闻标题和文档，构建出涉案新闻话题数据集。数据集的具体信息见表 1。

表 1 实验数据集统计信息

案件数	12
涉案新闻标题和文档	17 889
同一案件话题数	5
同一案件最多新闻数	1326
同一案件最少新闻数	507
新闻文档平均长度	132

3.2 评价指标

对涉案新闻话题发现的结果进行评估，本文使用准确率（Accuracy，ACC）、标准化互信息（normalized mutual information，NMI）和调整兰德系数（adjusted rand index，ARI）作为模型的评价指标。

准确率（ACC）是衡量话题发现算法对话题簇划分准确程度的评价指标。具体计算如式（16）所示

$$\mathrm{ACC}=\frac{TP+TN}{TP+FP+FN+TN} \tag{16}$$

其中，TP，TN，FP，FN 为混淆矩阵中的每一项，TP 和 TN 分别表示模型与真实标签同时判定样本为正或负，即聚类准确的样本，反之 FP 和 FN 为聚类错误的样本。ACC 的取值在 0 到 1 之间，取值越大代表话题发现准确率越高。混淆矩阵见表 2。

表 2 样本混淆矩阵

真实簇	话题簇	
	N	P
P	FP	TP
N	TN	FN

标准化互信息（NMI）是衡量话题发现聚类结果与真实样本分布之间的熵，NMI 的取值在 0 到 1 之间，取值越大代表话题发现聚类效果好，如式（17）所示

$$\mathrm{NMI}(Y,C)=\frac{2\times I(Y;C)}{H(Y)+H(C)} \tag{17}$$

其中，Y 表示真实的样本分布，C 表示话题簇的分布，$I(Y;C)$ 表示 Y 分布与 C 分布之间的互信息，$H(Y)$ 与 $H(C)$ 表示信息熵。

调整兰德系数（ARI）是衡量话题簇分布和真实分布的重叠程度的评价指标。ARI 取值在 −1 到 1 之间，取值越大代表话题模型效果越好。其计算公式如式（18）所示

$$\mathrm{ARI}=\frac{\mathrm{RI}-\mathrm{E}(\mathrm{RI})}{\max(\mathrm{RI})-\mathrm{E}(\mathrm{RI})} \tag{18}$$

其中，RI 为兰德系数，E(RI) 为兰德系数的期望值，计算公式如式（19）所示

$$\mathrm{RI}=\frac{a+b}{a+b+c+d} \tag{19}$$

式中：a，b，c，d 为表 3 中的变量。兰德系数变量见表 3。

表 3 兰德系数变量

a	在真实集中为同一簇，在话题簇中也为同一簇的对象对数
b	在真实集中为同一簇，在话题簇中属于不同簇的对象对数
c	在真实集中为不同簇，在话题簇中属于同一簇的对象对数
d	在真实集中为不同簇，在话题簇中也为不同簇的对象对数

3.3 实验设置

在模型的参数设置方面，本文通过预先训练的 BERT 中文语料库来表征涉案新闻话题数据集中的标题，词表为 BERT 模型自带词表，BERT 模型包含 12 层 Transformer 网络，每层网络包含 12 个注意力头，模型参数为 110 M，隐藏层维数为 768；文档特征提取模块中自编码器的维数为“输入-768-768-2000-10”，标题全局特征提取模块中使用了 4 层图卷积神经网络来迭代近邻标题图的关系特征，近邻标题图中 K 的个数取值为 10，话题簇初始簇心由 K-means 算法经过 20 次初始化获得，融合因子中平衡系数 α 设置为 0.5；模型训练轮次为 200，学习率为 1e−3，优化器采用 Adam。

3.4 基线模型分析

为了验证融合近邻标题图联合标题和文档进行话题建模对提高涉案新闻话题发现任务聚类效果的有效性，本文选取 8 个模型作为基线模型，分别在涉案新闻话题数据集上进行实验，其基线模型为：经典 K-means 算法、LDA、AE+Kmeans、DeepLDA、DEC、DCN、IDEC 和 NMC。

（1）K-means[6]是一种经典的聚类算法，在给定数据和聚类数目 k 的基础上，根据某个距离函数将数据分入 k 个簇中。

（2）LDA 是一种经典的主题模型，可将每篇文档的主题以概率分布的形式给出，可根据主题分布进行聚类。

（3）AE+K-means 是一种同时利用自编码器的表征和数据重构并结合 K-means 算法的聚类模型。

（4）DeepLDA[18]是一种融合深度神经网络的主题模型，将文档的词袋表示输入深度神经网络中，将 LDA 的输出作为一个标签，对神经网络进行监督训练，使神经网络既能学习主题文档分布，又能学习主题词分布。

（5）DEC 利用深度网络进行数据降维，通过软分配构造数据样本的簇分布，构造辅助目标分布计算其与样本分

布的 KL 散度。

（6）DCN[15]联合优化降维和聚类任务，利用深度神经网络逼近任何非线性函数的能力的同时，保持降维和聚类共同优化的优势。

（7）IDEC[19]考虑到保留数据的结构，并利用聚类损失作为指导，操控特征空间分散数据点，即模型可以联合聚类并学习代表性特征。

（8）NMC[20]是一种神经主题模型，利用伽马分布的重参数化和泊松分布的高斯逼近，开发了神经变分推理算法来推断模型参数，在大规模数据和特征稀疏的短文本数据上具有优势。基线模型性能比较见表 4。

表 4　基线模型性能比较

基线模型	ACC	NMI	ARI
K-means	0.6134	0.6482	0.5074
LDA	0.6329	0.6681	0.5235
AE＋K-means	0.7172	0.7245	0.6015
DeepLDA	0.7361	0.7314	0.6346
DEC	0.7602	0.7426	0.6735
DCN	0.8012	0.7784	0.7044
IDEC	0.8266	0.8004	0.7485
NMC	0.8539	0.8346	0.7918
本文方法	**0.8972**	**0.8619**	**0.8311**

从表 4 的实验结果中能够看出，经典 K-means 算法在处理涉案新闻话题数据时效果最差，因为它使用原始数据，不能很好地进行表征，且易受孤立点的影响。LDA 主题模型应用于通用领域的话题发现任务可以取得不错的效果，但是由于涉案新闻数据的特殊性，LDA 依赖于统计特征，聚类结果经常出现同类不同案的现象，准确率仍然不高。AE＋K-means 方法通过自编码器对数据降维后，得到数据的表征，再利用 K-means 算法进行聚类，话题簇的准确性得到了较为明显的提高，说明构造准确有效的表征对提升聚类准确率非常重要。DeepLDA 方法通过深度网络加强表征，并将 LDA 作为监督信号后，模型的计算效率大幅提升，但是由于缺乏标题信息等外部知识和聚类监督信号对主题分布的帮助，模型的内聚性仍然不高。DEC 和 DCN 模型相比较以上基线模型取得了更好的效果，因为这两种模型都引入了损失函数或目标分布作为监督信号，可以同时学习数据表征和聚类分配，并优化聚类样本使其更加接近话题簇心。IDEC 模型相较于 DEC 和 DCN 模型效果又有了一定提升，因为模型引入了重构损失可以学习到数据中具有局部结构保护的代表性特征。NMC 模型是一个比较新型的神经主题模型，相较于其它基线模型，NMC 在准确性指标上具有优势，可以较好地模拟具有过度分散和层次依赖特征的随机变量，但受限于数据规模和涉案新闻的特点，通过统计分布学习文档局部特征仍然具有主题不一致问题。

本文方法与其它基准模型相比取得了更优的性能，与 NMC 基线模型相比，ACC 提升了 4.33%，NMI 提升了 2.73%，ARI 提升了 3.93%。这是因为基线方法在做涉案新闻话题发现任务时，通常只着重提取文档自身的局部特征，而同一涉案新闻不同话题下的新闻文档包含了许多相似案件要素信息，基线方法不能很好地区分。本文的模型利用图卷积网络提取了近邻标题间的关联关系，并将其与文档的局部特征融合起来以增强标题的表征，从而实现话题建模更好的效果。这也证明了通过融入近邻标题图，联合标题与文档进行话题建模是有效的。

3.5　消融实验分析

为了验证本文模型各个模块的有效性，将模型拆解为主模型去除文档特征模块和主模型去除标题全局特征模块两个子模型，3 个评价指标保持不变，最优结果用加粗表示。消融实验结果见表 5。

表 5　简化模型性能分析

消融模型和主模型	ACC	NMI	ARI
本文主模型 w/o 文档特征	0.8041	0.7864	0.7112
本文主模型 w/o 标题全局特征	0.7602	0.7426	0.6735
本文主模型	**0.8972**	**0.8619**	**0.8311**

从消融实验结果可以看出，去除模型中的标题特征部分，只利用文档局部特征和指导模块进行建模效果最差，ACC 下降了 13.7%，NMI 下降了 11.9%，ARI 下降了 15.7%。虽然文档中包含了大量的案件要素信息，但是同一案件下不同话题的新闻文档要素有很多相似之处，噪声数据多，容易出现同一案件下划分为同一话题簇的数据却本该属于不同话题，或属于同一类型的案件却不是同一案件的情况。只利用标题全局特征和指导模块建模，效果比仅用文档特征要好一些，ACC 下降了 9.3%，NMI 下降了 7.5%，ARI 下降了 11.9%。因为模型提取到了近邻标题间的结构关系，但是由于标题篇幅的限制，所涵盖案件话题信息的内容有限，容易出现标题的信息偏置。将标题特征与文档特征结合起来建模，即本文主模型，效果提升明显。在获取涉案新闻之间的关联关系的基础上，同时引入文档表征增强标题的表示避免偏置可以更好地实现涉案新闻话题发现，这也从侧面验证了本文模型的有效性。

3.6　不同融合因子权重系数实验分析

为了验证调整融合因子的权重系数，即式（8）中权重系数 α 是否对模型性能有提升，本文做了如下实验。取步长为 0.2 的多个 α 值分别做对比实验。实验结果如图 5 所示。

从实验结果中可以看出，当 α 取 0.5 时，本文模型达

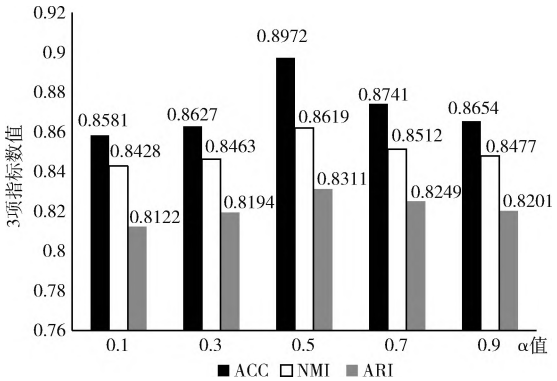

图 5 不同融合因子权重系数对模型的影响分析

到了最好的效果，而当 α 取值比 0.5 大或者比 0.5 小时，模型的性能都有所下降。因为 α 是融合因子的平衡权重系数，起到平衡标题全局特征和文档局部特征的作用。当 α 过大时，文档的局部特征权重就被削弱，模型只能学习到近邻标题图的关联关系，缺乏文档的内容信息，容易产生标题的信息偏置，图卷积网络容易产生过度平滑，同时模型失去了自编码器的重构损失，涉案新闻话题发现的准确性会降低；当 α 过小时，标题的全局特征权重被削弱，模型学习到的表征几乎全部来自文档自身，相似要素不能得到很好的区分，涉案新闻话题发现的准确性同样会降低。因此，将融合因子的权重系数 α 设置为 0.5 可以很好地融合两种特征。

3.7 随时间变化模型的准确率分析

为了验证时间指标对本文模型性能的影响，选取了 DEC、IDEC、NMC 这 3 个在基线对比实验中表现较好的模型和本文模型，在时间指标上进一步对比模型的准确率，如图 6 所示。

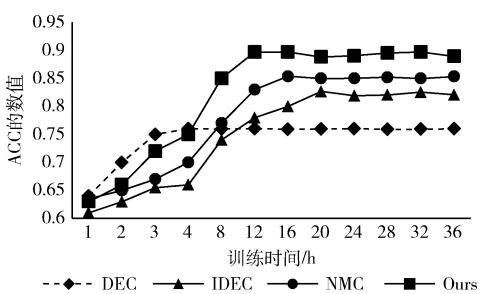

图 6 不同模型随训练时间增加准确率的变化分析

从训练模型的收敛时间上可以看出，DEC 模型收敛的时间最快，在模型训练 4 个小时左右即达到了该模型准确率的最优值，但是准确率最高仅有 0.7602，不能满足准确性的要求。而 NMC 和 IDEC 模型在准确性上要比 DEC 好很多，但受限于模型复杂程度的影响，需要训练 16 个小时以上才能达到收敛并达到最佳准确率，在实际应用中可操作性较差，不能及时发现涉案舆情话题。本文模型虽然没有 DEC 收敛速度快，但是相比另外两个对比模型，仅需一半的时间就可以达到收敛，且准确率可以达到 0.89 以上，在实际应用中非常适用于涉案舆情新闻早期传播的话题发现，对于有关部门开展舆情监管具有实际意义，也印证了本文方法的实用性。

3.8 实例分析

为了进一步验证本文方法模型的效果，通过实例分析对比了不同方法话题词的效果。以涉案话题“孙小果被判处死刑”为例，本文通过提取不同方法生成的话题簇中新闻文档的关键词，来直观地展示模型效果。实验结果见表 6。

表 6 实例分析

对比方法	话题词
K-means	孙小果 公开 再审 杜少平 恶势力 死刑 判决
LDA	孙小果 云南省高院 依法 判决 认为 决定 死刑 新晃
AE+KM	孙小果 终审 判决 故意伤害 寻衅滋事 黑社会 杜少平
DeepLDA	孙小果 云南省高院 寻衅滋事 湖南省高院 判决 依法 死刑
DEC	孙小果 再审 死刑 调查 有关部门 依法 判决
DCN	孙小果 社会关切 死刑 服刑 督办 法网 宣判
IDEC	孙小果 死刑 领导黑社会 扫黑 涉案人员 最高法院 督办
NMC	孙小果 专项斗争 云南高院 公开 死刑 重点 案件
本文方法	**孙小果 云南省高院 故意伤害 宣判 死刑 判决 剥夺**

从话题词的质量上可以看出，传统的聚类方法和主题模型方法以及它们的改进型方法的话题词中混入了同类型案件话题词，提取出了与“孙小果被判处死刑”话题同类型的“操场埋尸案杜少平被判处死刑”的话题词，说明使用原始数据以及依赖统计特征不能区分涉案新闻的要素信息，导致同类不同案的情况发生。而使用融入深度学习表征的聚类方法的话题词虽然描述的是同一案件，但是掺杂了同一案件下不同话题的词语，比如“孙小果被判处死刑”的话题词掺杂了“孙小果案挂牌督办”的话题词，这是因为此类方法在话题发现的过程中只重视文档自身的表征，没有考虑文档之间的关联，也没有融入外部信息指导。本文方法的话题词全部来自同一话题，话题发现准确率较高，充分说明引入标题的关联关系以及聚类指导模块，适用于涉案新闻话题发现任务。可以取得较好的效果，也验证了本文方法的有效性。

4 结束语

本文针对涉案新闻话题发现任务，提出一种融合近邻

标题图，联合标题和文档的表征进行话题建模的方法。解决了同一案件下话题新闻要素信息较为接近，表征不理想的问题，并提升了话题发现的准确性指标。基于涉案新闻话题数据集的实验结果表明，本文方法不仅可以得到质量更高的话题簇，而且在模型训练的时间指标上也有优势。

在未来的工作中，将探索如何从话题簇中得到准确的话题表示，并考虑话题关键信息的摘要抽取，以及长文本的处理工作，来进一步提高话题模型的性能。

参考文献：

[1] SUN Hongguang, GAO Xing, SUN Tieli, et al. Network news discovery based on improved Single-Pass algorithm [J]. Journal of Jilin University (Science Edition), 2018, 56 (1): 114-118 (in Chinese). [孙红光，高星，孙铁利，等. 基于改进 Single-Pass 算法的网络新闻话题发现 [J]. 吉林大学学报(理学版)，2018，56 (1)：114-118.]

[2] Zhou Qian, Shi Lei, Xu Liancheng, et al. An improved single-pass topic detection method [C] //10th International Conference on Measuring Technology and Mechatronics Automation, 2018: 317-320.

[3] Zhu Zhiliang, Liang Jie, Li Deyang, et al. Hot topic detection based on a refined TF-IDF algorithm [J]. IEEE Access, 2019, 7: 26996-27007.

[4] Mutanga MB, Abayomi A. Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach [J]. African Journal of Science, Technology, Innovation and Development, 2020, 14 (1): 163-172.

[5] Tarifa A, Hedhili A, Chaari WL. A filtering process to enhance topic detection and labelling [J]. Procedia Computer Science, 2020, 176: 695-705.

[6] Aini KN, Najahaty I, Hidayati L, et al. Combination of singular value decomposition and K-means clustering methods for topic detection on twitter [C]//International Conference on Advanced Computer Science and Information Systems, 2015: 123-128.

[7] Mustakim, Indah RNG, Novita R, et al. DBSCAN algorithm: Twitter text clustering of trend topic pilkada pekanbaru [C] //Journal of Physics: Conference Series, 2019: 1363 (1): 012001.

[8] Zhang Qiang, Du Junping, Kou Feifei. Financial topic detection algorithm based on multi-feature fusion [C] //Proceedings of Chinese Intelligent Systems Conference, 2020: 11-19.

[9] Rortais A, Barrucci F, Ercolano V, et al. A topic model approach to identify and track emerging risks from beeswax adulteration in the media [J]. Food Control, 2021, 119: 107435.

[10] Kumar J, Shao Junming, Uddin S, et al. An online semantic-enhanced Dirichlet model for short text stream clustering [C] //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 766-776.

[11] Fan Wentao, Guo Zhiyan, Bouguila N, et al. Clustering-based online news topic detection and tracking through hierarchical bayesian nonparametric models [C] //Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Association for Computing Machinery, 2021: 2126-2130.

[12] Li Chuanzhen, Liu Minqiao, Cai Juanjuan, et al. Topic detection and tracking based on windowed DBSCAN and parallel KNN [J]. IEEE Access, 2021, 9: 3858-3870.

[13] Xiao Kejing, Qian Zhaopeng, Qin Biao. A graphical decomposition and similarity measurement approach for topic detection from online news [J]. Information Sciences, 2021, 570: 262-277.

[14] Wu Di, Zhang Mengtian, Shen Chao, et al. BTM and GloVe similarity linear fusion-based short text clustering algorithm for microblog hot topic discovery [J]. IEEE Access, 2020, 8: 32215-32225.

[15] Yang Bo, Fu Xiao, Sidiropoulos ND, et al. Towards K-means-friendly spaces: Simultaneous deep learning and clustering [C] //International Conference on Machine Learning, 2017: 3861-3870.

[16] Devlin J, Chang M, Lee K. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv: 1810.04805, 2018.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[18] Bhat M, Kundroo M, Tarray TA, et al. Deep LDA: A new way to topic model [J]. Journal of Information and Optimization Sciences, 2019, 41 (3): 823-834.

[19] Guo Xifeng, Gao Long, Liu Xinwang, et al. Improved deep embedded clustering with local structure preservation [C] //International Joint Conferences on Artificial Intelligence Organization, 2017: 1753-1759.

[20] Wu Jiemin, Rao Yanghui, Zhang Zusheng, et al. Neural mixed counting models for dispersed topic discovery [C] //In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6159-6169.