文章编号：2095-2163(2023)10-0045-08 中图分类号：TP391.1 文献标志码：A

基于文本融合特征的突发事件子话题聚类研究

芦子涵，郑中团
（上海工程技术大学 数理与统计学院，上海 201600）

摘 要：突发事件具有突发性、公共性、传播范围广等特点，研究同一突发事件中更细粒度的子话题聚类，对舆情管控部门实现精准化管控具有重要意义。针对以往话题聚类方法忽略了同一事件下更细粒度的子话题聚类，且为了更有效地表达微博文本的语义信息，提出一种基于 LDA 文档-主题分布与 Doc2Vec 句向量融合的文本特征表示方法与文本相似度计算方法，应用 Single-Pass 增量聚类算法实现同一突发事件下子话题聚类，并根据 *F*1 值与单一文本特征子话题聚类实验结果进行对比。结果表明，本文方法子话题聚类效果更佳，*F*1 值为 72.4%，表明该方法能够有效地表达文本特征，进而提高子话题聚类的准确度。

关键词：突发事件；子话题聚类；文本特征；LDA 主题模型；Doc2Vec 模型

Research on sub-topic clustering of emergencies based on text fusion features

LU Zihan, ZHENG Zhongtuan
(School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201600, China)

【Abstract】 Emergencies are characterized by suddenness, publicity and wide dissemination. It is of great significance for public opinion management and control departments to study finer-grained sub-topic clustering in the emergency. In view of the fact that previous topic clustering methods ignores more fine-grained sub-topic clustering under the same event, to more effectively express the semantic information of microblog text, a text feature representation method based on the fusion of LDA document topic distribution, Doc2Vec sentence vector and text similarity calculation method is proposed. The Single-Pass incremental clustering algorithm is applied to achieve sub-topic clustering under the same emergency. The experimental results of sub-topic clustering based on *F*1 value and single text feature are compared. The results show that the sub-topic clustering effect of this method is better with 72.4% *F*1 value, which shows that the proposed method can effectively express the text features to improve the accuracy of sub-topic clustering.

【Key words】 emergency; sub-topic clustering; text features; LDA topic model; Doc2Vec model

0 引 言

话题检测与追踪(Topic Detection and Tracking, TDT)是美国国防高级研究计划局(Defense Advanced Research Projects Agency, DARPA)于 1996 年开展的语言信息研究项目[1]，曾在评测会议上对话题等相关要素进行了定义，认为话题是由一个种子事件或活动，和全部与之直接关联的后续事件和活动构成[2]。而在国内，曾有学者定义子话题为话题内一组相关事件的集合，是话题内所有事件集合的一个子集[3]。近年来，突发事件时有发生。譬如 2022 年“3·20”东航航班坠机等事故灾难事件、2022 年 6 月河北唐山打人等社会安全事件、2021 年“7·20”河南特大暴雨等自然灾害事件与至今仍时有发生的 2020 年新冠肺炎疫情等公共卫生事件。与此同时，随着网民规模的扩大与社交平台的普及，像新浪微博这样传播范围广、普及率高的社交网络平台逐渐成为突发事件的曝光口。社会大众可自由地在网络平台上发表自身对突发事件的看法或评论，从而形成网络舆情。由于突发事件具有不确定性、危害性等特点[4]，通常会给社会大众带来负面的心理冲击。如若不能针对性地根据社会大众对于某一突发事件所关注的不同子话题来引导积极的舆论走向，并建立舆情治理机制，则会放大社会大众的

基金项目：全国统计科学研究项目(2020LY080)。

作者简介：芦子涵(1998-)，女，硕士研究生，主要研究方向：社会网络分析、文本挖掘；郑中团(1979-)，男，博士，副教授，主要研究方向：机器学习与数据挖掘、应用统计与综合评价、随机过程与复杂网络等。

通讯作者：郑中团 Email：zhongtuanzheng@163.com

收稿日期：2022-10-20

负面情绪,引起不必要的激进言论,甚至会对政府机构造成不良影响。现有研究大多基于事件这一粒度进行话题聚类,而忽略了同一事件下不同侧面的更细粒度子话题的研究。因此,如何有效地挖掘某一事件中的潜在子话题,逐渐成为了新兴研究热点,也对舆情管控相关部门实现舆情精准化管控具有重要现实意义。

本文针对以往话题聚类大多基于事件这一层次,而忽略了同一事件下更细粒度子话题的研究,且文本特征表示上缺乏上下文语义信息的缺陷,提出一种基于 LDA 文档-主题分布与 Doc2Vec 句向量融合的文本表示方法与文本相似度计算方法,最后通过 Single-Pass 增量聚类算法实现同一突发事件下子话题聚类。

1 相关研究

目前,在话题挖掘领域,多以基于概率主题模型的话题发现、基于文本特征表示的话题聚类两种为主要途径与方法。概率主题模型是对文本中隐含主题的一种非监督建模方法,其认为一篇文档中的每个词都是通过以一定概率选择某个主题,并从这个主题中以一定概率选择某个词的方式得到的。早期,为解决 TF-IDF 文本模型的缺陷,利用奇异值分解将高维共现矩阵映射到低维潜在语义空间的潜在语义分析模型(Latent Semantic Analysis,LSA)被提出。因其计算复杂度高且缺乏概率基础,Hofmann[5]在 1999 年将 LSA 的思想引入到概率模型中,提出概率潜在语义分析模型(Probabilistic Latent Semantic Analysis,PLSA)。2003 年,Blei 等[6]基于贝叶斯思想,认为文档-主题概率分布是服从狄利克雷概率分布的随机变量,提出了潜在狄利克雷模型(Latent Dirichlet Allocation,LDA)。在话题挖掘领域,LDA 主题模型也成为目前最为成熟的概率主题模型。由于概率主题模型以词袋模型为基础,通常忽略了单词与单词之间的语义信息,导致语义缺失、主题可解释性差等问题。基于此,赵林静等[7]通过 HowNet 常识知识库计算单词间的语义相似度,来调整 LDA 主题模型中的超参数β,提出 SS-LDA 模型以提高主题挖掘的精度。居亚亚等[8]为解决 LDA 主题模型语义连贯性较差等问题,在 LDA 框架下引入 GRU 模型加入单词—单词和文档—单词语义相似度来引导建模,提出了 SDS-TM 模型。闫盛枫[9]利用词嵌入技术进行语义向量编码,以此来合并同语义信息主题词并调整主题词分布及权重,增强了主题模型的语义表达性。也有学者通过优化 LDA 主题建模结果实现子话题的挖掘。如:周楠等[10]基于 PLSA 模型得到每个子话题下不同的词频分布,通过相似子话题合并、子话题更新优化主题建模结果,解决了传统方法的子话题区分度差等缺陷。夏丽华等[11]将概率主题模型融合词共现关系,提出 GPLSA 方法对原始子话题进行合并与更新,解决了描述同一产品的文档十分相似,难以保证子话题差异性的问题。

聚类是一种十分重要的非监督学习技术,其任务是按照某种标准或数据的内在性质及规律实现样本的聚类[12]。在话题挖掘领域,话题聚类基于文本的特征表示或文本间的相似度,将目标文档分为若干个簇,使得每个簇内文本间的相似度尽可能高,不同簇间文本的相似度尽可能低。因而,众多研究者基于文本特征表示或文本相似度进行话题发现。史剑虹等[13]利用隐主题模型挖掘微博内容中隐含主题—文档分布作为文本特征表示,并基于 K-means++聚类实现话题发现。颜端武等[14]针对微博文本高维稀疏与上下文语义缺失等问题,以 LDA 文档—主题分布特征和加权 Word2Vec 词向量特征构建文本融合特征,并通过 K-means 聚类实现主题聚类。肖巧翔等[15]提出一种基于 Word2Vec 扩充文本和 LDA 主题模型的 Web 服务聚类方法,将短文本主题建模转化为长文本主题建模,进而通过 K-means 算法更准确地实现了服务内容主题聚类。赵爱华等[16]针对子话题间文本相似度高的特点,引入主题特征词相关性分析,提出一种改进的文本相似度计算方法,并基于 Single-Pass 增量聚类实现新闻话题子话题挖掘。李湘东等[17]针对 LDA 建模结果较泛化的缺陷,将 LDA 建模结果主题—特征词分布作为文本较粗粒度的特征,将 TF-IDF 向量作为文本较细粒度的特征来融合表示文档,采用知网语义词典得到文本相似度,通过 Single-Pass 聚类实现国内各地时事新闻子话题划分。

综上,子话题挖掘多以 LDA 主题模型建模、LDA 主题模型建模结果优化、基于文本特征表示的话题聚类为主要方法。其中,对于评论短文本 LDA 主题模型具有文本向量高维稀疏、缺乏上下文语义信息等缺陷;改进的 LDA 主题模型以引入外部知识库来修改超参数β来引导建模,通用性低且计算复杂度高。基于文本特征表示的话题聚类多以事件为层次进行主题发现,而忽略了同一事件下更细粒度、更深层次的子话题聚类研究。基于此,本文提出一

种基于 LDA 文档-主题分布与 Doc2Vec 句向量融合的文本特征表示方法与文本相似度计算方法，通过 Single-Pass 增量聚类算法实现同一突发事件下子话题聚类。一方面，上述文本融合特征不仅通过 LDA 文档—主题分布提取了全局主题信息，同时也通过句向量的构建提取了局部上下文语义信息以补充 LDA 主题模型语义信息的缺乏。另一方面，不同于大多话题所基于的事件层次，针对同一事件下子话题相似度高、区分度低的问题，本文给出了一种同一事件下更细粒度、更深层次的子话题聚类方法。

2　预备知识

2.1　LDA 主题模型

主题模型是一种用来发现一系列文档中隐含主题的无监督统计模型，认为一篇文档中的每个词都是以一定概率而选择某个主题，并从该主题中以一定概率而选择某个词所生成的。如图 1 所示，LDA 主题模型是 2003 年被 Blei 等人[6]提出的文档—主题—单词的三层贝叶斯主题模型。该模型以词袋模型为基础，认为一篇文档是由词所组成的集合，而词与词之间没有语义联系与顺序。其能够将一篇文档表示为隐含主题的多项分布，即该文档属于每个主题的概率；将主题表示为词集上的多项分布，即该主题下各个词出现的概率。与其他概率主题模型不同的是，LDA 主题模型基于贝叶斯思想，认为文档—主题分布 θ_d 的先验分布为 Dirichlet 分布，即 $\theta_d = Dirichlet(\vec{\alpha})$。主题—词分布 β_k 的先验分布为 Dirichlet 分布，即 $\beta_k = Dirichlet(\vec{\eta})$。

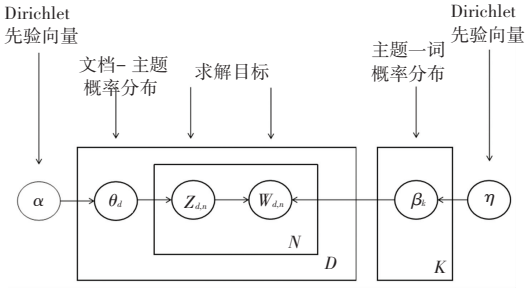

图 1　LDA 主题模型

Fig. 1　LDA topic model

在 LDA 主题模型中，通常使用 Gibbs 采样算法[18]来进行求解。α，η 作为已知的先验输入，目标是得到各个 $z_{d,n}$、$w_{d,n}$ 对应的整体文档—主题分布与主题—词分布。

2.2　Doc2Vec 模型

为表达整条文本评论或整篇文档的特征，常将由 Word2Vec 得到的词向量进行向量拼接，此方法导致信息损失较大，得到的新向量不能涵盖丰富语义信息内容[19]；或将由 Word2Vec 得到的词向量进行平均求和，但此方法未考虑到词与词之间的语序信息，一定程度上忽略了文本上下文语义信息。Mikolov 等人[20]在 Word2Vec 的基础上提出了 Doc2Vec 模型，以期构建文档的向量化表示。Word2Vec 模型本质上一个具有输入层、隐藏层、输出层的三层神经网络结构，其包含 CBOW(Continue Bag of Words)与 Skip-Gram 两种学习模型。CBOW 模型根据所输入的目标词上下文单词的 One-Hot 向量表示来输出对目标词的预测，而 Skip-Gram 则是输入当前词来预测上下文词。

与 Word2Vec 不同的是，Doc2Vec 模型在训练过程中增加了段落向量 Paragraph id，进而可以结合上下文词训练文本，从而得到句向量和文本向量[21]。在 Doc2Vec 模型中，段落向量与单词一样首先将被映射成一个句向量 Paragraph Vector，其次将段落向量与上下文词语所映射成的向量累加或拼接起来，作为输出层的输入。由于 Paragraph Vector 在同一个文档的每一次训练中是共享的，因此随着文档每次滑动窗口取上下文单词训练的过程中，Paragraph Vector 作为输入层向量的一部分每次都将被训练，向量所储存的段落信息将会越来越准确。Doc2Vec 模型同样包含 PV-DM(Distributed Memory)与 PV-DBOW(Distributed Bag of Words)两种学习模型。本文拟采用 PV-DM 模型，如图 2 所示。PV-DM 模型根据所输入目标词的上下文单词来预测目标词，而 PV-DBOW 则是输入当前词来预测上下文词。

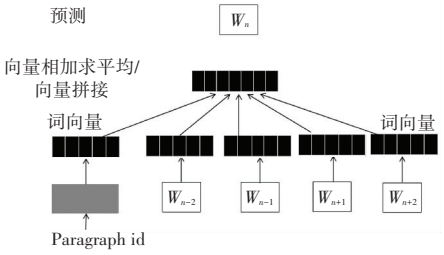

图 2　Doc2Vec 模型

Fig. 2　The model of Doc2vec

3　基于文本主题与语义融合特征的子话题聚类

3.1　思路与流程

本文针对同一突发事件下子话题具有相似度高而区分度低的特点，同时考虑到 LDA 主题模型以词袋模型为基础，其构建的单一主题特征常忽略文本

语义信息的问题,重点构建基于文本主题特征与文本语义特征的文本融合特征向量,并对上述两种不同特征的文本相似度进行线性结合,从而通过 Single-Pass 增量聚类实现突发事件下子话题聚类。首先,以新浪微博平台为数据来源,爬取突发事件评论文本构建语料库,并对数据进行清洗、分词、去停用词等预处理;其次,在全局主题层面通过 LDA 主题模型提取文档—主题分布以表达文本主题特征,在局部语义层面通过 Doc2Vec 模型提取文档句向量以表达文本语义特征,从而构建文本融合特征;然后将基于 KL 距离与余弦相似度线性结合计算融合特征相似度,以度量文本相似度;最后通过 Single-Pass 增量聚类实现子话题聚类。具体流程如图 3 所示。

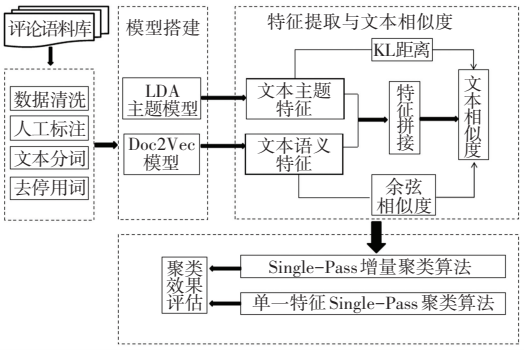

图 3 研究思路与流程

Fig. 3 Research process

3.2 文本融合特征的构建

假设预处理后的突发事件评论文本语料库 $D = \{d_1, d_2, \cdots, d_n\}$,其中 n 为语料库中评论文本的数目。首先,通过 LDA 主题模型提取文本主题特征。LDA 主题模型所提取的主题信息为 $T = \{t_1, t_2, \cdots, t_k\}$,$K$ 为主题个数,通常由人为自主设定,本文将采用困惑度这一指标来确定最优主题个数。本文采用 Gibbs 采样算法求解 LDA 主题模型,在初始时刻为每个单词随机地赋予主题,其次,对于每个文本 d 中的每个词,通过 Gibbs 采样公式获取其所对应的主题。Gibbs 采样公式如式(1)所示:

$$p(z_i = k \mid \vec{w}, \vec{z}_{-i}) = \frac{n_{d,-i}^{k} + \alpha_k}{\sum_{s=1}^{K} n_{d,-i}^{s} + \alpha_s} \frac{n_{k,-i}^{t} + \eta_t}{\sum_{f=1}^{V} n_{k,-i}^{f} + \eta_f} \quad (1)$$

其中,$n_d^{(k)}$ 表示在第 d 个文本中第 k 个主题词的个数,$n_k^{(v)}$ 表示第 k 个主题中第 v 个词的个数。

重复上述采样过程直至 Gibbs 采样收敛,即可得到所有词的采样主题。通过统计每个文本 d 对应词的主题计数,每个文本 d 可表示为 $\theta_d = \{(t_1, \theta_{t_1}), (t_2, \theta_{t_2}), \cdots, (t_k, \theta_{t_k})\}$ 的文档—主题分布,完成文本主题特征的提取。其次,通过 Doc2Vec 模型提取文本语义特征。本文采用 Doc2Vec 中的 PV-DM 模型,使用 Python 中 Gensim 库的 Doc2Vec 接口来训练语料库,从而得到语料库中每个文本 d 的句向量表示 $S_d = [s_{(d,1)}, s_{(d,2)}, \cdots, s_{(d,m)}]$.

由于基于词袋模型的 LDA 主题模型所提取的主题特征往往忽略了文本语义信息,而 Doc2Vec 模型所训练的文本句向量能够补充性地提取上下文语义信息,弥补 LDA 主题特征的这一缺陷。因此,本文将基于 LDA 主题模型与 Doc2Vec 模型所提取文本主题特征与文本语义特征进行横向拼接,构建文本融合特征矩阵 $\boldsymbol{ST}$。

$$\boldsymbol{ST} = \begin{bmatrix} \theta_{1,t_1} & \theta_{1,t} & \cdots & \theta_{1,t_k} & s_{(1,1)} & s_{(1,2)} & \cdots & s_{(1,m)} \\ \theta_{2,t_1} & \theta_{2,t_2} & \cdots & \theta_{2,t_k} & s_{(2,1)} & s_{(2,2)} & \cdots & s_{(2,m)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{n-1,t_1} & \theta_{n-1,t_2} & \cdots & \theta_{n-1,t_k} & s_{(n-1,1)} & s_{(n-1,2)} & \cdots & s_{(n-1,m)} \\ \theta_{n,t_1} & \theta_{n,t_2} & \cdots & \theta_{n,t_k} & s_{(n,1)} & s_{(n,2)} & \cdots & s_{(n,m)} \end{bmatrix}$$

3.3 文本相似度计算

文本相似度的计算是子话题聚类的前提,本文将基于 KL 散度与余弦相似度计算文本主题概率分布相似度与句向量相似度,并将二者进行线性组合,从而得到本文所构建的融合特征相似度,即文本相似度,式(2):

$$sim(d_i, d_j) = sim_{\mathrm{LDA}}(d_i, d_j) + sim_{\mathrm{Doc2Vec}}(d_i, d_j) = KL(p_{\theta_{d_i}}, p_{\theta_{d_j}}) + cos\_distance(s_{d_i}, s_{d_j}) \quad (2)$$

其中,d_i 与 d_j 表示评论文本。

3.3.1 基于 KL 距离的文本主题特征相似度

KL 距离(Kullback-Leibler Divergence,KL)用来衡量相同事件空间里的两个概率分布的差异情况,又被称为相对熵。在本文中,评论文本 d_i 的文档—主题分布表示为 $p(t)$,评论文本 d_j 的文档—主题分布表示为 $q(t)$,$p(t)$ 与 $q(t)$ 的概率分布越相似,则两者之间的 KL 距离越小[16]。$p(t)$ 与 $q(t)$ 之间的 KL 距离如式(3)所示:

$$KL(p_{\theta_{d_i}} \| p_{\theta_{dj}}) = \sum_{t \in T} p(t) \cdot \log \frac{p(t)}{q(t)} \quad (3)$$

考虑到 KL 距离具有非对称性,交换 $p(t)$ 与 $q(t)$ 的位置后结果大不相同,参考文献[9]的做法,可采用公式(4)计算文档—主题概率分布之间的距离:

$$sim_{\mathrm{LDA}}(d_i,d_j)=KL(p_{\theta_{d_i}},q_{\theta_{d_j}})=\frac{1}{2}((KL(p_{\theta_{d_i}}\|q_{\theta_{d_j}}))+(KL(q_{\theta_{d_j}}\|p_{\theta_{d_i}}))) \quad (4)$$

3.3.2　基于余弦相似度的文本语义特征相似度

针对通过 Doc2Vec 模型训练所提取的表征文本语义特征的句向量，采用余弦相似度来计算文本语义特征相似度，如式(5)所示。

$$sim_{\mathrm{Doc2Vec}}(d_i,d_j)=\cos\_distance(s_{d_i},s_{d_j})=\frac{s_{d_i}\cdot s_{d_j}}{|\,|s_{d_i}|\times|s_{d_j}|\,|} \quad (5)$$

其中，S_{d_i}、S_{d_j} 为评论文本 d_i、d_j 的文本语义特征。

3.4　子话题聚类算法流程

本文采用 Single-Pass 增量聚类[22]实现子话题聚类，该算法是话题检测中一种常用算法，又称单通道法。在 Single-Pass 算法中，需要自主预设一个聚类阈值，对于所输入的评论文本，计算当前评论文本与已有话题聚类簇之间的相似度，若相似度大于预设的聚类阈值，则将该评论文本判为已有话题聚类簇；否则，将该评论文本作为簇核心创建新的话题簇。本文将所构建的文本融合特征与文本相似度计算嵌入 Single-Pass 聚类算法中，具体算法流程见表 1。

表 1　子话题聚类算法流程

Tab. 1　The process of sub-topic clustering algorithm

输入：文本语料库 $D=\{d_1,d_2,\cdots,d_n\}$
输出：各个子话题簇 $C_1,C_2,\cdots,C_m$
Step 1　训练 LDA 主题模型，提取文本主题特征文档—主题分布 θ_d
Step 2　训练 Doc2Vec 模型，提取文本语义特征句向量 $\boldsymbol{S}_d$
Step 3　构建文本融合特征矩阵 $\boldsymbol{ST}$
Step 4　顺序读入评论文本 d 的融合特征表示
Step 5　遍历计算评论文本 d 与每个子话题簇 C_i 簇核心之间的相似度 $sim(d,d_{c_i})$
Step 6　寻找与评论文本 d 相似度值最大的簇 C 令 $id=\mathrm{argmax}(sim(d,d_{c_i}))$ 获取簇 id
Step 7　若 $\max(sim(d,d_{c_i}))\geqslant\sigma$（聚类阈值），则将 d 加入子话题簇 C
Step 8　否则，则以当前文本为簇核心创建新簇
Step 9　转入 Step 4，直至所有评论文本特征循环结束

4　实验与分析

本文将以新浪微博为数据来源，以“郑州地铁 7.20 事件”为突发事件评论语料库进行 3 组实验。第一组实验采用困惑度(Perplexity)评价指标，得出 1~10 个主题下的困惑度值，从而确定最优主题数；第二组实验采用 $F1$ 值寻找能够使 $F1$ 值达到最高的聚类阈值，从而确定最佳聚类阈值 σ；第三组实验生成 3 种评论文本特征向量，其中包括 LDA 文档—主题分布向量、Doc2Vec 句向量以及本文的融合特征向量，采用查准率(*Precision*)、召回率(*Recall*)与 $F1$ 值对比 3 种文本特征向量子话题聚类效果，以验证基于本文融合特征子话题聚类的有效性。

4.1　突发事件概述与数据预处理

2021 年 7 月 20 日，河南郑州发生罕见特大暴雨。当日晚 19 时左右，据郑州本地广播官方微博@MyRadio 发布的微博称，郑州地铁 5 号线雨水倒灌，车厢内积水已到达乘客胸部，数名乘客被困。随后该条微博被澎湃新闻官方微博@澎湃新闻转发，转发人次 5.2 万，评论人次 3.7 万，事件爆发。截至当日晚间 22 时左右，消防救援人员陆续疏散被困人员 500 余人。7 月 21 日上午，郑州地铁官方发布称此次事件导致 12 人遇难。随后，两名个人用户发布博文称有乘客邹某、沙某仍失联。26 日，乘客邹某、沙某确认遇难。27 日上午，郑州官方发布此次事件最终导致 14 人遇难，再次引起一波舆论高潮。2022 年 1 月 21 日，国务院调查组调查认定郑州地铁 5 号线亡人系责任事件，是造成重大人员伤亡与财产损失的突发事件。

本文以“郑州地铁 5 号线”、“多人被困”等为关键词，以 2021 年 7 月 20 日 19 时—2021 年 7 月 31 日 22 时为时间区间，每 2 小时为一个时间段，利用 Gooseeker 集搜客数据抓取器采集数据，共采集到 6 657条评论文本作为语料库。每条评论文本包含 5 个字段：用户 ID、发布时间、评论内容、点赞数与评论数。对语料库进行以下预处理操作：

(1)数据清洗。去除与话题不相关的评论文本，剔除特殊字符如表情、评论图片等；

(2)人工标注。结合郑州地铁 5 号线事件期间微博热搜内容，对评论文本进行话题标注，以便后续有效性验证；

(3)分词。采用 Python 中 Jieba 库对评论文本进行分词，同时加载分词词典以识别该事件特定词；

(4)去停用词。根据停用词表去除标点符号、语气助词等词语。

4.2 评估指标

本文采用查准率（*Precision*）、召回率（*Recall*）、*F*1 值来对比 3 种文本特征向量子话题聚类效果，其值越高，说明方法效果越好。

查准率（*Precision*）是指预测为属于子话题 C_i 的评论文本中，实际属于子话题 C_i 的评论文本比例；召回率（*Recall*）为实际属于子话题 C_i 的评论文本中，被预测为属于子话题 C_i 的评论文本比例。

$$F1_c = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

$$F1 = \frac{1}{n}\sum_{c=1}^{n} F1_c \tag{7}$$

其中，C 为子话题簇个数。

整体聚类效果采用 *F*1 对各个子话题的聚类效果求平均的方式来度量。

4.3 实验结果与分析

4.3.1 实验 1 确定最优话题个数

在 LDA 主题模型提取文本主题特征中，主题个数的选取能够直接影响到特征提取效果。若仅依赖人为设定，LDA 主题模型的性能将无法保证。因此，本实验采用困惑度（Perplexity）评价指标来确定最优主题个数。困惑度常被用来衡量概率分布或概率模型样本的优劣性[23]。在自然语言处理中，可用于 LDA 主题模型，确定最优主题个数，如式（8）所示：

$$Perplexity(V) = \exp\frac{\sum_{d=1}^{N}\log p(W_d)}{\sum_{d=1}^{N} M_d} \tag{8}$$

其中，V 表示语料库 D 中所有词的集合；N 表示语料库中评论文本的数量；W_d 表示评论文本 d 中的词；M_d 表示每个评论文本 d 中的词数；$p(W_d)$ 表示文本中词出现的概率。

实验中根据“郑州地铁 7.20 事件”期间新浪微博热搜词条，拟定 1～10 区间内的整数为实验主题数，得到困惑度变化如图 4 所示。

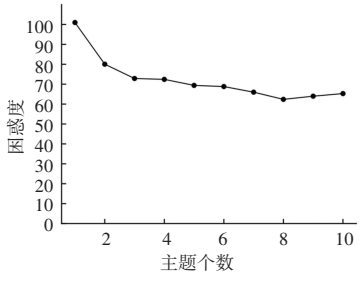

图 4 确定最优主题个数

Fig. 4 The determination of the optimal number of topics

通常情况下，困惑度随着主题数量的增加而呈现递减的规律。困惑度越小，意味着主题模型的生成能力越强[24]。通过图 4 可以看出，当 $T = 8$ 时 LDA 主题模型困惑度最小，因此本文将主题个数 T 设定为 8。

4.3.2 实验 2 确定最佳聚类阈值

实验中采用 4.2 节所描述的 *F*1 值来计算不同聚类阈值下聚类效果的优劣。经多次实验，当聚类阈值小于 0.3 时，所有评论文本被聚类为同一簇，聚类阈值过小。因此，本实验中拟定聚类阈值在 $\sigma \in (0.3,1)$ 这一区间内，分别进行 6 次实验，得到 *F*1 值变化如图 5 所示。可以看出，当聚类阈值 $\sigma = 0.52$ 时，聚类效果最好，此时的 *F*1 值为0.724，因此本文将确定聚类阈值 σ 为0.52。

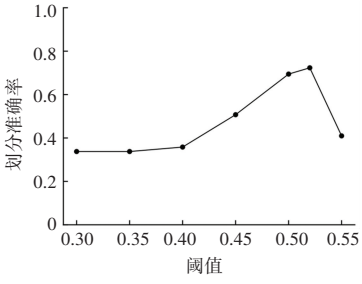

图 5 确定最佳聚类阈值

Fig. 5 The determination of threshold value in clustering

4.3.3 实验 3 对比实验与分析

为验证本文基于融合特征表示的子话题聚类方法的有效性，对于 LDA 主题模型所提取单一文本主题特征文档—主题分布、Doc2Vec 模型提取单一文本语义特征句向量、3.2 节所表述的文本融合特征分别进行 Single-Pass 子话题聚类实验，并采用精确率、召回率、*F*1 值来度量聚类效果的优劣。实验结果见表 2。

表 2 实验 3 结果对比

Tab. 2 The result of test 3

子话题聚类方法	精确率	召回率	*F*1 值
Doc2Vec 句向量+Single-Pass	0.722	0.677	0.673
LDA 文档-主题特征+Single-Pass	0.662	0.750	0.644
文本融合特征+Single-Pass	0.783	0.674	0.724

依据表 2 中数据分析可知：

（1）基于单一文本语义特征的子话题聚类的 *F*1 值为 67.3%。Doc2Vec 模型通过三层神经网络根据所输入的目标词来预测目标词的上下文单词，从而得到副产物句向量与词向量。一方面，相比将一条评论文本中每个词的词向量进行求和或加权平均求和来表示整条文本评论的方法，Doc2Vec 能够给出

整条文本评论的文档向量化表示，能够避免前者忽略单词在句子中的语序问题；另一方面，相比于 LDA 主题模型基于词袋模型，Doc2Vec 模型能够有效提取文本中的语序及上下文语义信息。但未考虑文本的全局信息，因而在 $F1$ 值位于另外两种特征子话题聚类之间。

（2）基于单一文本主题特征的子话题聚类的 $F1$ 值为 64.4%，相较于另外两种特征 $F1$ 值最低。LDA 主题模型将文本表示为维数为主题个数的多项分布，从而提取文本全局主题特征。LDA 主题模型所基于的词袋模型忽视了文本中单词的语序与语义表达，对于同一突发事件下相似度高、区分度差的评论文本而言，虽能够提取文本的主题特征，但仅用 LDA 主题特征来进行相似背景子话题聚类，则难以发挥 LDA 主题模型的优势与作用。

（3）基于融合特征的子话题聚类方法相较于单一特征聚类效果最佳，$F1$ 值达 72.4%。融合特征考虑到同一突发事件下子话题具有相似背景词而导致区分度差的特点，且 LDA 主题模型所提取主题特征基于词袋模型，缺乏语义信息，从文本主题层面与语义层面融合 LDA 文档—主题分布与 Doc2Vec 句向量，改善了单一特征进行子话题聚类的缺陷，能更加全面有效地表达文本特征，从而提高同一突发事件下子话题聚类效果。

5　结束语

本文提出的基于文本融合特征的子话题聚类方法，结合 LDA 主题模型提取的文本主题特征与 Doc2Vec 模型提取的文本语义特征构建一种文本融合特征，并通过 Single-Pass 增量聚类实现子话题聚类。研究中使用本文方法，以新浪微博为数据来源平台，对“郑州地铁 7.20 事件”这一突发事件评论文本进行实验分析。在对比实验中，采用 $F1$ 值与两种单一特征子话题聚类进行聚类效果评估。实验结果表明，融合特征能更加全面地表达文本特征，改善了单一特征进行子话题聚类缺乏上下文语义信息及忽略语序的问题，有效地提高了突发事件中子话题聚类的准确率。

受各方面因素所限，本文还存在一定的局限与不足。在突发事件网络舆论中，网民往往带有浓烈的正向或负向的情感色彩。因此，在文本的特征表达中，如何提取评论文本的情感特征并将其进行融合处理，从而更有效地进行子话题挖掘，在后续的研究中仍有待进一步深入和突破。

参考文献

[1] NALLAPATI R, FENG A, FU C, et al. Event threading within news topics [C] // Proceedings of the 13th ACM Conference on Information and Knowledge Management.New York: ACM, 2004: 446-453.

[2] JAMES A. Topic Detection and Tracking: Event - Based Information Organization [M]. Norwell: Kluwer Academic Publisher, 2002: 1-16.

[3] 李军，李涓子. 新闻专题内子话题划分 [C] // 第 4 届全国信息检索与内容安全学术会议，中国中文信息学会，2009：442-451.

[4] 李纲，李阳. 关于突发事件情报失察的若干探讨 [J]. 情报理论与实践，2015，38(7)：1-6.

[5] HOFMANN T. Probabilistic latent semantic indexing [C]// Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999：50-57.

[6] BLEI D，NG A，JORDAN M. Latent Dirichlet allocation [J]. Journal of Machine Learning Research，2003，3：993-1022.

[7] 赵林静. 结合语义相似度改进 LDA 的文本主题分析 [J]. 计算机工程与设计，2019，40(12)：3514-3519.

[8] 居亚亚，杨璐，严建峰. 基于语义分布相似度的主题模型 [J]. 计算机应用研究，2019，36(12)：3553-3557.

[9] 闫盛枫. 融合词向量语义增强和 DTM 模型的公共政策文本时序建模与演化分析——以“大数据领域”为例 [J]. 情报科学，2021，39(9)：146-154.

[10] 周楠，杜攀，靳小龙，等. 面向舆情事件的子话题标签生成模型 ET-TAG [J]. 计算机学报，2018，41(7)：1490-1503.

[11] 夏丽华，韩冬梅. 面向社交媒体评论的子话题挖掘研究 [J]. 情报杂志，2020，39(4)：110-116.

[12] 理姗姗，杨文忠，王婷，等. 基于网络社交媒体的子话题检测技术综述 [J]. 计算机应用，2020，40(6)：1565-1573.

[13] 史剑虹，陈兴蜀，王文贤. 基于隐主题分析的中文微博话题发现 [J]. 计算机应用研究，2014，31(3)：700-704.

[14] 颜端武，梅喜瑞，杨雄飞，等. 基于主题模型和词向量融合的微博文本主题聚类研究[J]. 现代情报，2021，41(10)：67-74.

[15] 肖巧翔，曹步清，张祥平，等. 基于 Word2Vec 和 LDA 主题模型的 Web 服务聚类方法 [J]. 中南大学学报(自然科学版)，2018，49(12)：2979-2985.

[16] 赵爱华，刘培玉，郑燕. 基于 LDA 的新闻话题子话题划分方法 [J]. 小型微型计算机系统，2013，34(4)：732-737.

[17] 李湘东，巴志超，黄莉. 基于 LDA 模型和 HowNet 的多粒度子话题划分方法 [J]. 计算机应用研究，2015，32(6)：1625-1629.

[18] CHEUNG S H，BANSAL S. A new Gibbs sampling based algorithm for Bayesian model updating with incomplete complex model data [J]. Mechanical System and Signal Processing，2017，92(1)：156-172.

[19] 宁宁. 基于 LDA 和句向量的文本分类研究[D]. 天津：天津理工大学，2021.

[20] LE Q，MIKOLOV T. Distributed representations of sentences and documents [C]//International conference on machine learning. PMLR，2014：1188-1196.

[21] 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]. 计算机工程与应用，2022，58(16)：123-128.

[22] PAPKA R，ALLAN J. On-line new event detection using single pass clustering[J]. University of Massachusetts，Amherst，1998，10：290941-290954.

（下转第 55 页）