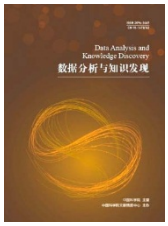

数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：　主题不平衡新闻文本数据集的主题识别方法研究
作者：　王红斌，王健雄，张亚飞，杨恒
网络首发日期：　2020-11-12
引用格式：　王红斌，王健雄，张亚飞，杨恒．主题不平衡新闻文本数据集的主题识别方法研究．数据分析与知识发现.
https://kns.cnki.net/kcms/detail/10.1478.g2.20201112.1113.010.html

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2020-11-12 14:01:34
网络首发地址：https://kns.cnki.net/kcms/detail/10.1478.g2.20201112.1113.010.html

主题不平衡新闻文本数据集的主题识别方法研究

王红斌[1,2]，王健雄[1,2]，张亚飞[1,2]，杨恒[3]
[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)
[3](云南唯恒基业科技有限公司 云南 昆明 650000)

摘要：
[目的] 传统LDA模型因新闻文本数据集中不同主题间文本数量不均衡导致文本主题识别不准确。**[方法]** 提出一种在主题不平衡新闻文本数据集上的主题识别方法，该方法基于传统LDA模型，结合独立性检测、方差检测和信息熵检测三种不同的特征检测方法来识别出文本的主题。**[结果]** 在10000篇新闻文本规模的数据集上实验验证，该方法相比传统的LDA主题识别方法在查全率上提高了0.2121、查准率上提高了0.0407和F1值提高了0.152。**[局限]** 由于新闻文本中新词较多，实验中使用的分词工具的分词准确率会降低，新闻文本主题识别的效果因对分词准确率的依赖而受影响。**[结论]** 实验证明，本研究所提的方法能在一定程度上解决了LDA对新闻文本数据集中不同主题间文本数量不均衡的主题识别问题。

关键词： 主题不平衡；新闻文本数据集；文本主题识别；潜在 Dirichlet 分配(LDA)
分类号： TP393，G250
DOI: 10.11925/infotech. 2020-0765.

Topic Recognition Research on Topic Imbalanced News Text Data Set

Wang Hongbin[1,2], Wang Jianxiong[1,2], Zhang Yafei[1,2], Yang Heng[3]
[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)
[3] (Yunnan Weiheng Jiye Technology Co., Ltd., Kunming 650000, China)

Abstract：
[Objective] The traditional LDA model is not accurate for text topic recognition，because of the number of different topic texts in news text dataset is not balanced. **[Methods]** This paper proposes a topic recognition method based on the traditional LDA model on unbalanced news text data sets, which combines three different feature detection methods: independence detection, variance detection and information entropy detection. **[Results]** Experiments are conducted on 10000 news texts, the proposed method improves recall by 0.2121, precision by 0.0407 and F1 value by 0.152, compared with the traditional LDA topic recognition method. **[Limitations]** Due to the large number of new words in news text, the segmentation accuracy of word segmentation tools used in the experiment will be reduced, and the effect of news text topic recognition is affected by the dependence on the accuracy of segmentation. **[Conclusions]** Experimental results show that the proposed method can solve the problem of LDA topic

recognition on unbalanced number of texts between different topics in news text dataset to a certain extent.
Keywords：Topic imbalanced; News text data set; Text topic recognition; Latent Dirichlet Allocation(LDA)

1 引言

在互联网和大数据日益发展的环境下，各类数据基于互联网平台被大量产生。新闻报道由于叙述详实规范，来源可靠，观点客观等特点，是互联网海量数据中一个重要的信息来源，新闻文本对于经济形势研究、国内国际政治研究、商业决策研究、社会文化研究、甚至科学技术发展方向研究等领域都有着十分重要的作用。从海量地新闻文本数据中分析筛选出有价值的信息需要耗费大量的资源，因此如何运用计算机技术自动归纳出新闻数据所包含的新闻主题，并通过一定方法将新闻数据所包含的信息清晰全面地呈现给用户，是一个重要的研究课题。

新闻文本主题识别已经是一个相对完备的过程，但对实际应用的新闻文本主题识别存在一些问题，对于一段时间内发生的新闻，热点新闻讨论热度较高，相应的新闻文本数量则相对的较多，而一些新闻热度相对较低的新闻则讨论程度较低，相应的新闻文本数量也相应的较少，但这些新闻文本所讨论的主题也是不可或缺的。采用传统的 LDA[1] 模型进行主题建模时，因新闻文本数据集中不同主题间文本数量不均衡导致文本主题识别不准确。基于 LDA 的主题识别模型是基于词频统计的，主题识别总受高频词影响，却忽略低频的词，对于某主题下文本数量多的主题词相对高频，从而过度训练产生噪声主题，对于某主题下文本数量少的主题词则相对低频，以致很难准确识别出这些少类新闻的文本主题。也就是说新闻文本主题识别存在文本数据不平衡问题。面对这样一个新闻文本数据集中的主题新闻文本数量不平衡的问题，准确识别出这些主题新闻文本数量少的新闻文本的主题变得很有必要。基于此，本文提出一种主题不平衡新闻文本数据集的主题识别方法来解决新闻文本数据集中不同主题的新闻文本之间存在的不平衡问题。

2 研究现状介绍

针对传统 LDA 模型在不平衡主题新闻文本数据的主题识别不准确问题，需要在现有的主题识别方法上面去做改进。现有的主题识别方法中，有基于主题模型和基于深度学习模型的方法。

目前各种文本挖掘模型被相继提出，包括文档表示模型(Term Frequency-Inverse Document Frequency, TF-IDF) [2]、潜在语义索引模型(Latent Semantic Index, LSI)[3]、概率潜在语义索引模型(probabilistic PLSI)[4] [5]和潜在狄利克雷分布(LDA)[1]。然而，基于神经网络的主题模型能更好的解决引入上下文的问题，近年也进行了关于神经主题模型的工作[6] [7] [8]。Adji B. Dieng 等 [9]考虑传统主题模型忽略了文本的上下文序列关系，在传统的主题模型基础上加入 RNN 提出 TopicRNN。Lau 等[10]结合主题模型和语言模型的好处，提出一个局部驱动的语言模型。现有的基于神经网络的解决不平衡问题的方法，都是针对有监督的分类算法，需要知道分类的标签，而 LDA 方法是无监督的学习方法，LDA 是一种处理非结构化文档集合的有效工具，是一种用于建模语料库的无监督生成概率

方法，被广泛应用于文本分类[11]、信息检索[12]和主题识别等任务。使用上诉的方法只能造成以结果证结果而无法解决问题。

解决文本主题识别数据集不平衡问题，本质也是一种文本分类问题，需要从文本分类的不平衡问题上寻找解决办法，解决文本分类问题的方法主要有两种[13]：一种是数据层面，人为地对训练数据进行加工；另一种是算法层面，针对样本不平衡数据集设计新的分类算法。

数据层面的方法主要有两种，大类训练数据过采样、少类训练数据欠采样。训练数据重采样通过对少类样本进行重复采样，使得不平衡数据集变为平衡数据集，从而解决数据不平衡问题。最简单的重采样方法是直接复制少类数据使数据达到平衡，这样的方法容易造成过学习问题[14]。 改进的方法是有选择的复制或者生成新的样本，比如 SMOTE[15]， ADASYN[16]等。训练数据欠采样是通过丢弃一些多类样本数据，使得样不平衡数据集变为平衡数据集，最简单的方法是随机移除，但会导致信息丢失，其他方法有选择的移除一些多类样本，引入数据清除策略如 Tomek links 等[17]，也有研究者把重采样和欠采样综合使用，也取得了一定的效果。

很多研究者提出使用集成学习方法来解决不平衡数据集问题。王光等[18]发表了 CSVM 算法，该算法的核心思想是先对每个样本集聚类，得到多个子类，然后把子类样本集作为训练数据，该方法可以减少样本集的样本数量。使用代价敏感的方法，发表了对每类样本集给予不一样的惩罚系数[19]。李红莲[20]发表了 NN-SVM 算法，该算法的基本思想是在处理分类超平面周围的样本点时依据一些规则取舍使得分类的间隔增大，提高分类器的泛化能力。另外，居亚亚等[21]针对 LDA 主题模型中所有单词的重要性相同的问题，提出了一种基于动态权重的 LDA 算法，降低了高频词对建模的影响，提高了关键词的重要性。

综述所述，现有的解决数据集不平衡主题识别问题的方法，大多都是针对有监督的算法，需要知道分类的标签。另外使用上述的方法只能造成以结果证结果而无法解决问题。基于此，本文基于传统 LDA 主题模型提出一种主题不平衡新闻文本数据集的主题识别方法来解决新闻文本数据集中不同主题的文本之间存在的数据不平衡的主题识别问题。

3 方法设计

本节介绍了主题不平衡新闻文本数据集上的主题识别方法，该方法包括新闻语料库准备，数据整理，LDA 新闻文本主题识别，高质量主题筛选和文本最终主题获取五个主题识别框架模块以及评价方法。图 1 给出了本文提出的主题不平衡新闻文本数据的主题识别方法的总体框架图。

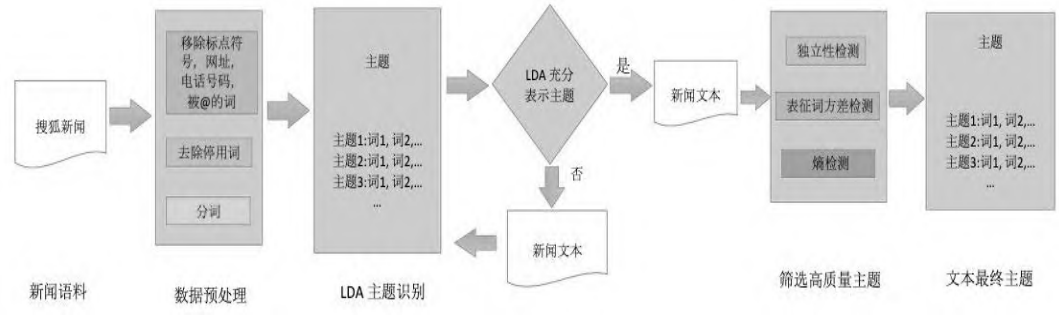

图 1 主题不平衡新闻文本数据的主题识别框架图

Fig.1 Topic recognition on topic imbalanced news text flow chart

3.1 模型构建思想

首先对新闻文本数据集进行预处理，包括文本分词，文本降噪处理。然后对预处理之后的数据集进行基于 LDA 的主题识别，利用基于密度峰值的聚类方法先识别出文本集的最优主题数目，并以此主题数目进行主题识别，得到文本-主题分布和主题-词分布。之后对于文本数据集中的每一篇文本，判断是否被识别出来的主题充分表示，得到被分成两类的文本，一类是被充分表示主题的文本，这些文本即是不平衡文本中的主题文本较多的文本，另一类是没有被充分表示主题的文本，这些文本即是不平衡文本集中的主题文本较少的文本。对于被充分表示主题的文本，由于有些不是主题的词的词频被提高，导致并不是识别出来的每一个主题都能作为一个有效的主题，我们对识别出来的主题进行判断，筛选出高质量的主题，去除噪声主题，以高质量主题作为这些被充分表示主题文本的主题；对于没有被充分表示主题的文本，重复上述的操作，这样每一次的循环都能识别出不平衡文本集中的大类文本的主题，最终在没有被充分表示主题的文本不再出现的时候结束循环。最后以多次识别出的高质量主题集合作为新闻文本数据集的文本主题。

3.2 新闻语料收集

新闻文本主题识别并没有公共的数据集，因此针对 2018 年全年的热门话题，并获取这些热门新闻话题的相关新闻文本内容，作为新闻文本主题识别的实验数据集。搜狐新闻是中国最大的移动媒体平台，众多平面、网络、电视、广播、意见领袖（自媒体）等各类媒体入驻，是权威的新闻文本获取平台，与微博、网页评论、知识问答等短文本相比，搜狐新闻数据，用户人群面向大众，文本内容包含内容更为丰富，主题性强，为文本主题识别实验提供了较好的条件。因此本研究中，选择使用 2018 年搜狐新闻的热门话题数据集来测试和评估所提出的方法。针对 2018 年全年的热门新闻话题，从头条分析中获得 2018 年的全年热门新闻话题，然后从搜狐新闻门户网页上搜索相关热门新闻话题，得到 2018 年全年的热门新闻话题的新闻文本，作为新闻文本主题识别数据集。

3.3 预处理

对要进行文本识别的文本进行文本数据整理，文本数据整理包括中文分词，数据去噪等内容。

对于新闻文本数据集，用结巴分词工具进行分词处理。

数据去噪主要是去除不包含任何主题意义和干扰主题区分度的词语，具体包含以下内容：

(1) 停用词

停用词通常不包含主题意义，如很，非常等副词。

(2) 数字，地名

数字，地名等对于主题的区分会产生很大的干扰，一般去除。

(3) 单个汉字

单个汉字通常不包含有用的主题信息。

(4) 白名单词典

一些有意义的词语会被当作噪声词，比如 LDA 等。

(5) 没有任何主题意义的词语

邮箱，电话号码，不包含汉字的词语不包含任何主题意义。

3.4 主题识别

按照基于困惑度方法[22-23]判断出文本集的最佳主题数目去识别出对应数目的主题，并以主题下高概率的词语作为表征词展现文本集的主题，事实上，并非所有表征词所构成的主题都能表示一个主题，这些主题仅仅是根据数据集中的词语共现关系等而自动获取的，且识别出的最优主题数目并不能使主题识别的查准率和查全率达到百分百，存在一定误差，主题数量不足可能导致 LDA 模型过于粗糙而无法准确区分主题，另一方面，过多的主题可能导致模型过于复杂，使一个主题被过度细分或者一个主题被分作多个主题，使主题的解释变得困难。表 1 和表 2 分别列出了对收集到的 2018 年全年的热门新闻经过 LDA 主题识别得到的其中一个高质量主题实例和一个非高质量主题实例。

如表 1 中的高质量主题实例，该主题的前几个高概率主题词代表的都是“长生生物”的相关表征词，且没有与长生生物主题不相关的主题表征词，则该主题则可以作为一个高质量主题。

如表 2 中低质量的主题示例所示，诸如“服务”之类的通用术语通常无法描述该主题，但是诸如“服务”之类的词汇在该主题中频率最高，而且概率分布不均匀。 仅单词“服务”的概率为 0.03175。其次，该主题是噪音主题，因为他的常用词有两个不同的主题，即“长生生物疫苗假”和“滴滴乘客遇害”。

表 1 高质量主题实例

Table 1 High-quality theme examples

山东省	疫苗	长生	合格	疾控中心	长春	…
0.00914	0.00714	0.00304	0.00208	0.00108	0.00108	…

表 2 非高质量主题实例

Table 2 Examples of non-high-quality topics

服务	滴滴	疫苗	长春	顺风车	失信	…
0.03175	0.00139	0.00138	0.00133	0.00132	0.00077	…

综合分析，构成高质量主题的表征词应具有以下特征：

a．独立性：一个高质量主题的表征词仅作为该主题的表征词，不在其他高质量主题的高频率表征词中出现，具有一定的独立性。

b．不均衡性：高质量主题的表征词是不均衡分布的，少量的表征词以较高的频率出现，大量的表征词以较低的频率出现且对主题表征没有意义。

基于高质量主题和非高质量主题的不同特征，我们提出了三种不同的特征检测来识别出文本的高质量主题，针对高质量主题表征词的独立性进行独立性检测，针对高质量主题表征词的不均衡性进行方差检测和信息熵检测。

（1）主题表征词的独立性检测

对于一个高质量主题，它的表征词通常具有独立性，即一个高质量主题的表征词只出现在该主题的高频词下，而不出现在其他主题的高频词下，比如“长生生物疫苗造假事件”这个主题，像“疫苗”，“疾控中心”， “长春”这些高频词，他很大概率只以高概率出现在“长生生物疫苗造假事件”这个主题下，而

以极小的概率出现在其他主题中，比如“滴滴：8 月 27 日起全国下线顺风车业务”这个主题中这些词语则应该以很小的概率出现。

为计算一个主题的独立性，我们计算该主题的表征词与其余所有主题的表征词中共同出现的词的相似度，并将该主题所有共现词表征词的相似度相加，作为该主题的独立性结果，我们需要计算所有主题的独立性，即计算所有主题表征词与其他主题表征词相似度的相加结果。

主题独立性结果太大，说明该主题与其他主题相似度太高，该主题的表征词由一些通用词构成，比如“服务”，“中心”等，通用词对不同主题区分有很大的干扰，需要把这些词判断为噪声。因此，主题独立性结果太大，主题对应的表征词大多由通用词构成，该主题不能作为一个高质量的主题。

主题独立性结果太小，说明该主题的表征词由一些随机杂乱的词语构成，也是没有任何主题意义的，比如其中一个主题“交通”，“交互”，“交代”，“主体”，该主题完全由杂乱无章的词语构成表征词，该主题也需要被判断为噪声主题，不能作为一个高质量的主题。

由此分析可见，主题的独立性需要保持在一定合理的区间内，太大主题表征词由通用词构成，太小主题表征词则由随机杂论的词构成。

定义 i 主题与 j 主题表征词中共现词的相似度为 $sim(i, j)$，主题 i 中表征词 w 出现的概率为 P_{iw}，语料词语总数为 N，相似度计算公式为（1）所示[24]：

$$sim(i, j) = \sum_{w=1}^{N} \mathrm{P}_{iw} * P_{jw} \quad (1)$$

主题数目为 K，则该文本语料的独立性 dep 计算如（2）所示[24]：

$$dep = \sum_{i=1, j\neq i}^{N} sim(i, j) \quad (2)$$

（2）主题表征词的方差检测

高质量主题的表征词是不均衡分布的，少量的表征词以较高的频率出现，大量的表征词以较低的频率出现且对主题表征没有意义。方差可以用来衡量一组数据的波动程度。

一个主题表征词方差太大，说明该主题的波动程度较大，则该主题分布失衡，大部分的表征词由少部分的通用词构成并作为高频词，对于一个主题来说没有区分，不能作为一个高质量的主题。

一个主题的表征词方差太小，说明该主题的波动程度太小，则该主题分布过于均衡，大部分的表征词由随机杂乱的词语构成，也把主题表征词方差太小的主题作为一个噪声主题。

由此分析可见，一个高质量主题的表征词方差需要保持在一定合理的区间内，太大则分布过于不均衡，主题表征词由通用词构成，太小则分布过于均衡，主题表征词由随机杂乱的词语组成。

定义一个主题的表征词概率期望为 $\overline{X}$，该主题中表征词 w 出现的概率为 P_w，语料词语总数为 N，主题的表征词概率期望计算公式为（3）所示[24]：

$$\overline{X} = \sum_{w=1}^{N} \frac{P_w}{N} \quad (3)$$

一个主题的表征词方差计算公式为（4）所示[24]：

$$variance = \frac{1}{N}\left[\sum_{w=1}^{N}\left(P_w - \overline{X}\right)^2\right] \quad (4)$$

（3） 主题表征词的熵检测

信息熵是信息论中用于度量信息量的一个概念。一个系统越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。所以，信息熵也可以说是系统有序化程度的一个度量。

如果一个主题的文本表征词 X 被看作一个离散型随机变量，其取值空间为 R，其概率分布为 $P(w) = P(X = w), w \in R$，那么， X 的熵的定义为公式（5）所示[25]：

$$H(X) = -\sum_{w \in R} p(w)\log_2 P(w) \quad (5)$$

一个随机变量的熵越大，它的不确定性越大，系统越是无序，熵越小，不确定性越小，系统越是有序。使用信息熵来度量主题表征词的不确定性，则熵越大，主题表征词系统趋于稳定，熵越小，主题表征词系统越是无序。

一个高质量的主题的表征词是少部分的主题表征词占据很大的概率，而大部分的主题表征词都以很低的概率分布。因此，主题表征词系统是趋于稳定的，信息熵应该越低越好。但是若只选择少量的核心表征词来计算信息熵，由于核心词概率较大，核心词之间的概率分布不均衡，因此，序列的信息熵越大越好。我们为了方便计算则只选取前 200 个表征词计算，则信息熵越大越好。

3.5 识别被 LDA 主题模型充分表示主题的文本

被 LDA 主题模型充分表示主题的文本，即该篇文本能被识别出来的主题的表征词所表示。一篇文本被 LDA 主题模型充分表示主题，则该篇文本的主题不能是一个噪声主题，主题概率要高于一定的阈值，且该篇文本的词语也应大概率出现在主题的高概率表征词中。

新闻文本 LDA 主题识别之后，每篇文本得到一组文本-主题概率分布，概率和为 1，但识别出的主题并不是全部具有主题意义，只有挑选出的高质量主题才能作为文本的主题，概率和就不一定为 1。针对新闻文本的特点，一篇新闻文本一般只具有一个单一主题，我们选择出的高质量主题中，选择概率最高的主题，作为一篇文本的唯一主题。下面给出一个简单实例，100 篇文本的 5 个主题，通过 LDA 主题识别之后，以一篇文本为例，得到表 3 所示的 5 个主题。

表 3 判断一篇文本的最终主题

Table 3 Determine the final theme of a text

	概率	表征词
主题 1(长生生物造假)	0.3329	长生、账户、上市公司、关税、疫苗
主题 2(北京房租上涨)	0.0293	租赁、房租、租金、北京、上涨
主题 3(噪声主题)	0.2364	项目、副董事长、产业园、证券、质押
主题 4(美国加征关税)	0.0138	美国、关税、特朗普、公安分局、钢铁

主题 5(噪声主题)	0.0085	洗衣机、追加、千亿美元、电子、美国政府

由表 3 可以看出，经过 LDA 主题识别之后识别出 5 个主题，其中主题 3 和主题 5 是噪声主题，高质量主题是主题 1，主题 2 和主题 4，而在这一篇文本中，这些高质量主题的概率是 0.3329，0.0293，0.0138，就以高质量主题中概率最高的主题 1 作为该篇文本的最终主题。

一篇新闻文本只有一到两个主题，只有其中一到两个主题概率很高，其他主题的概率是很低的，文本主题概率也是不均匀分布的，而且一篇文本主题概率高于一定的阈值，说明这篇文本被一个高质量主题所表示，且其他主题的概率很低，不出现在这篇文本中，这样也能把噪声主题是概率最高的主题的情况给排除。

识别出一篇文本的主题，且得到了一篇文本中主题的概率，如果该主题概率大于一定的阈值，则该篇文本是被充分表示主题的文本，低于阈值则该篇文本的主题没有被识别出来。

主题模型是根据词语的共现关系识别出文本的主题，在一个被 LDA 主题模型充分表示主题的文本中，该篇文本的词语也应大概率出现在主题的高概率表征词中。对于一篇文本，我们计算这篇文本的词语在所有高质量主题表征词中出现的概率的加和，并以文本主题概率作为权重，得到这篇文本词语和主题表征词的共现度，以此来计算判断文本的词语是否大概率出现在主题的高概率表征词中，高质量主题数为T，主题概率为P_t，文本有n个词语，文本词语的主题词概率为P_w，本文提出的文本词语和主题表征词的共现度Co为公式(6)所示:

$$Co = \sum_{t=1}^{T} (P_t \sum_{w=1}^{n} P_w) \tag{6}$$

识别出一篇文本的主题，得到文本词语和主题表征词的共现度，如果该共现度大于一定的阈值，则该篇文本是被充分表示主题的文本，低于阈值则该篇文本的主题没有被识别出来。

对于被 LDA 主题模型充分表示主题的文本，就以识别出来的文本高质量主题作为这些文本的主题。

对于没有被 LDA 主题模型充分表示主题的文本，即是相对被识别出主题的文本的少类文本，对于这些文本，我们再进行一次 LDA 主题识别，识别出最佳主题数目，然后再用 LDA 进行主题识别，筛选出高质量主题，并作为这些没有被 LDA 主题模型充分表示主题的文本的主题，如果第二次主题识别之后还有没有被 LDA 主题模型充分表示主题的文本，继续对此次没有被 LDA 主题模型充分表示主题的文本重复操作，直到不再出现没有被 LDA 主题模型充分表示主题的文本。

最终我们得到多次文本主题识别之后的高质量主题表示，我们把多次主题识别之后的高质量主题表示合起来，作为最终的文本集的主题。对于主题不平衡新闻文本的主题识别方法的算法伪代码如下所示：

算法 主题不平衡新闻文本的主题识别

输入：新闻文本数据集

输出：Topic

1. for 全部新闻文本 do
2. 从语料库读取新闻文本
3. 对新闻文本分词
4. 对分词后的文本进行删除标点符号，停用词等预处理
5. 通过基于困惑度方法获得新闻文本中最佳主题数 N
6. N topic, 每篇文本 T_j 的文本-主题分 D_j ← 对所有新闻文本 LDA 主题识别
7. for 每个主题 N_i in N topic
8. $H(N_i)$ ← 计算主题表征词的信息熵
9. if ($H(N_i) > 1.1$) then
10. add N_i to 高质量主题集合 C_1
11. end for
12. for 每篇新闻文本 T_j in 新闻文本数据集
13. $Co(T_i)$ ← 计算文本单词和主题表征词的共现度
14. if ($D_j > 0.1$ and $Co(T_i) < 0.5$) then
15. 添加 T_j 到 LDA 主题模型未充分表示的文本集合 S_1 中
16. if ($S_1 \neq \varnothing$) then
17. C_j , S_j ← 重复步骤 6-15，将步骤 6 中的所有新闻文本替换为集合 S
18. end $S_1 \neq \varnothing$

Topic=$\{S_1, S_2, S_3, \cdots, S_j\}$

4 实验及结果分析

在本节中，我们判断得到选择高质量主题的最优参数区间，并描述在新闻文本数据集上的主题不平衡主题识别过程和结果。

4.1 实验数据和数据预处理

实验数据来源于搜狐门户网站的新闻文本数据约 10000 条。选择使用 2018 年搜狐新闻的热门话题数据集来测试和评估所提出的方法。针对 2018 年全年的热门新闻话题，从头条分析中获得 2018 年的 145 个热门新闻话题，然后从搜狐新闻门户网页上搜索相关热门新闻话题，共获得 145 个话题总共 10000 多篇新闻文本，文本包括新闻的全文。部分所得到的主题及主题文本数如表 4 所示：

表 4 部分语料主题以及主题文章数目

Table 4 Some corpus topics and the number of topic articles

主题	数目	房租价格上涨	24
莫焕晶被执行死刑	38	电影《大轰炸》宣布取消公映	19
“鸿茅药酒事件”引关注	25	赵丽颖冯绍峰宣布结婚	53
美股大跌，市值蒸发超 8 万亿	144	中美贸易关税	32
美国年度最强飓风”迈克尔“登陆佛罗里达州	6	数学家阿提亚宣称自己证明了黎曼猜想	6

对于整个过程，获得的 145 个话题总共 10000 多篇新闻文本，文本篇幅较大

且话题较多，难以展示出整个过程，就选取其中少量文本和少量话题的不平衡文本展示整个过程，既能验证方法的有效性，方便展示，又能判断在不同数据集文本的不平衡文本下的适用性。所得到的主题及主题文本数如表 5 所示，得到 6 个主题的 116 篇不平衡新闻文本。

将新闻文本数据集进行中文分词和数据去噪处理，使用结巴分词工具进行中文分词，然后进行相关的数据去噪处理，去掉不包含主题词的词条和没有任何主题意义的词语。

表 5　少量语料主题以及主题文章数

Table 5 A small number of corpus topics and topic articles

主题	数目
莫焕晶被执行死刑	19
中美贸易关税	19
房租价格上涨	17
滴滴顺风车	23
云南通海地震	3
长生生物疫苗造假事件	35

4.2　判断高质量主题选择的最优参数及方法

对于选择高质量主题的独立性选择，方差选择，信息熵选择三种方法[24]，都需要判断最优参数取值空间才能选择出高质量主题，我们通过在数据集上的实验判断出三种方法选择高质量主题所必需的参数，并选择其中效果最好的方法作为最终选择高质量主题的方法。

对于独立性选择高质量主题的方法，对 145 个主题近 10000 篇文本 LDA 主题识别，表征词数采用前 200 个词，然后人工标记高质量主题，并计算不同独立性区间内的查准率，查全率，F1 值，判断最佳独立性取值空间。

实验结果如表 6 所示，由于实验数据较多，仅选择 4 个阈值区间作为代表作为展示。

表 6　主题独立性验证

Table 6 Subject independence test

独立性阈值区间	查准率	查全率	F1 值
（1×10^{-5}，1×10^{-1}）	0.9241	0.5172	0.6632
（1×10^{-4}，1×10^{-1}）	0.9241	0.5241	0.6689
（1×10^{-4}，1×10^{-2}）	0.8966	0.5448	0.6781
（1×10^{-3}，1×10^{-2}）	0.8276	0.5379	0.6520

由表 6 结果所示，在阈值区间在（1×10^{-4}，1×10^{-2}）内时，所筛选出的高质量主题 F1 值最高，效果最好。

对于方差选择高质量主题的方法，同样对 145 个主题近 10000 篇文本 LDA 主题识别，表征词数采用前 200 个词，然后人工标记高质量主题，并计算不同方

差区间内的查准率，查全率，F1 值，判断最佳方差取值空间。

实验结果如表 7 所示，由于实验数据较多，仅选择 4 个阈值区间作为代表作为展示。

表 7　主题方差验证

Table 7 Subject variance test

方差阈值区间	查准率	查全率	F1 值
（0.01，1）	0.9793	0.4542	0.6206
（0.03，0.5）	0.9517	0.4828	0.6406
（0.05，0.5）	0.8966	0.5172	0.6560
（0.08，0.5）	0.8620	0.5103	0.6410

由表 7 结果所示，在阈值区间在（0.05，0.5）内时，所筛选出的高质量主题 F1 值最高，效果最好。

对于信息熵选择高质量主题的方法，同样对 145 个主题近 10000 篇文本 LDA 主题识别，表征词数采用前 200 个词，然后人工标记高质量主题，并计算不同信息熵下的查准率，查全率，F1 值，判断最佳方差取值空间。

实验结果如表 8 所示，由于实验数据较多，仅选择 4 个信息熵值作为代表作为展示。

表 8　主题信息熵验证

Table 8 Subject information entropy test

信息熵取值	查准率	查全率	F1 值
0.8	0.9103	0.6690	0.7712
1.0	0.8690	0.7586	0.8101
1.1	0.8276	0.8138	0.8184
1.2	0.7793	0.8413	0.8091

由表 8 结果所示，在信息熵大于 1.1 时，所筛选出的高质量主题 F1 值最高，效果最好。

高质量主题选择的独立性选择，方差选择，信息熵选择三种方法[25]，熵检测方法选择高质量主题的方法最优，效果最好，F1 值是 0.8184，独立性检测方法其次，F1 值是 0.6781，方差检测方法效果最差，F1 值是 0.6560，因此选择高质量主题选择最优的熵检测方法来进行接下来的实验，信息熵大于 1.1 时，所筛选出的主题则是高质量主题。

4.3　LDA 最优主题数目识别

在用 LDA 主题模型进行大规模的主题挖掘时，需要给予一个最优的主题数目参数，基于本篇论文重点在解决主题识别的过程，最优主题数目选择通过基于困惑度的最优主题数目识别方法识别。

我们对一定数量的主题进行迭代，当困惑度在一定范围内变慢时，此时的主

题数可以最好地表示数据集中的最佳主题数。表 9 和图 2 显示了 116 篇新闻文本的实验结果：

表 9 基于困惑度的最优主题数目识别

Table 9 Identification of optimal number of topics based on confusion

n	perplexity
1	-11.080522603744297
8	-11.828040290223517
15	-11.686841810925171
23	-11.983907486919497
31	-11.949689567973851
38	-12.04940281527981

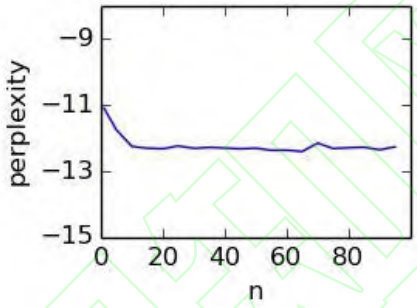

图 2 基于困惑度的最优主题数目识别

Fig.2 Identification of optimal number of topics based on confusion

可以知道主题数目在 8 时变化明显，并趋于稳定保持稳定，因此获得的最佳主题数为 8。

4.4 基于 LDA 的主题识别

识别出新闻文本数据集的最优主题数目，并以此作为 LDA 主题识别的参数进行 LDA 主题识别，最终得到的 8 个主题如表 10 所示：

表 10 主题识别结果

Table 10 Subject recognition results

主题 1	('美国', 0.005509399)	('关税', 0.004643482)	('特朗普', 0.0037529143)	('钢铁', 0.0029704035)
主题 2	('滴滴', 0.0063447123)	('莫焕晶', 0.004508633)	('死刑', 0.0028377306)	('顺风', 0.0026869758)
主题 3	('长春', 0.002711079)	('长生', 0.002532903)	('疫苗', 0.0018195861)	('公安分局', 0.0014926636)
主题 4	('疫苗', 0.0038461)	('长生', 0.002201729)	('地震', 0.001993109)	('账户', 0.0015869305)
主题 5	('租赁', 0.004220571)	('房租', 0.0038221274)	('租金', 0.0031279188)	('上涨', 0.002675991)
主题 6	('办法', 0.0002167317	('判决', 0.0002167317)	('发表声明', 0.0002167317)	('发生冲突', 0.0002167317)
主题 7	(莫焕晶', 0.006745752)	('死刑', 0.004233984)	('放火', 0.0034928808)	('保姆', 0.0024844697)
主题 8	('疫苗', 0.004375345)	('我省', 0.0018545929)	('长生', 0.001846968)	('接种', 0.0018227079),

4.5 筛选高质量主题识别出文本主题

识别出文本数据集的 6 个主题，需要筛选高质量主题识别出文本主题，选择高质量主题的方法中，效果最好的方法。

是熵检测，对于识别出来的每个主题，计算主题表征词的信息熵，如果这个表征词熵大于 1.1 时，则判断它是一个高质量主题，否则就是一个噪声主题，计算得到的 8 个主题的表征词信息熵，结果如表 11 所示：

表 11 主题信息熵

Table 11 Topic information entropy

主题	主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7	主题 8
信息熵	1.1326447 7147999	0.4799639 87322595	1.3008808 5328203	0.7741111 77981054	1.473561 18221953	0.6180361 66375225	1.2935727 6526454	1.5427371 3681934

8 个主题的信息熵中，主题 1，3，5，7，8 的信息熵高于 1.1，判断这五个主题是高质量主题，而主题 2，4，6 则判断为噪声主题。

4.6 识别被 LDA 主题模型充分表示主题的文本

对于每一篇文本，接下来判断是否被 LDA 主题模型充分表示主题。识别出一篇文本的主题，且得到了一篇文本中主题的概率，如果该主题概率大于 0.1，则该篇文本是被充分表示主题的文本，低于阈值则该篇文本的主题没有被识别出来。得到文本词语和主题表征词的共现度Co，如果该共现度大于 0.5，则该篇文本是被充分表示主题的文本，低于阈值则该篇文本的主题没有被识别出来。最终得到被 LDA 主题模型充分表示主题的文本。表 12 是所有文本的最终主题概率和共现度计算结果，由于文本较多，只展示了其中典型的例子：

表 12 识别文本是否被 LDA 充分表示

Table 12 Recognize if text is fully represented by LDA

	文本 1	文本 2	文本 3	文本 4	…
最终主题概率	主题 1 (0.3283)	主题 3 (0.3641)	主题 3 (0.2923)	主题 5 (0.0731)	…
共现度	0.8439	0.2136	0.7855	0.7781	…

表 12 是一部分文本主题识别之后，计算主题概率和共现度的结果，文本 1 最终主题概率大于 0.1 且共现度大于 0.5，则文本 1 是被主题模型充分表示主题的文本，文本 2 最终主题大于 0.1 但是共现度小于 0.5，文本 2 是没被主题充分表示的文本，文本 3 也是被主题模型充分表示主题的文本，主题 4 最种主题概率低于 0.1，是没被主题充分表示的文本，则最终文本 1，3 是被主题充分表示的文本，文本 2，4 是没被主题充分表示的文本，对于文本 1，3 则用筛选出来的高质量主题表示，文本 2，4 则进一步处理。

在 8 个主题的 116 篇文本中，有 33 篇文本是没有被 LDA 主题模型充分表示主题的文本，83 篇文本是被 LDA 主题模型充分表示主题的文本。

4.7 对没有被 LDA 主题模型充分表示主题的文本处理

对于没有被 LDA 主题模型充分表示主题的 33 篇文本，即是相对被识别出主题的文本的少类文本，对于这些文本，我们再进行一次 LDA 主题识别，识别出

最佳主题数目，然后再用 LDA 进行主题识别，筛选出高质量主题，并作为这些没有被 LDA 主题模型充分表示主题的文本的主题，如果第二次主题识别之后还有没有被 LDA 主题模型充分表示主题的文本，仔次重复上诉操作，直到不再出现没被主题充分表示的文本。最终得到多次文本主题识别之后的主题表示，我们把多次主题识别之后的主题表示合起来，作为最终的文本集的主题。

第二次对 33 篇没有被 LDA 主题模型充分表示主题的文本进行 LDA 主题识别度，识别出最优主题数目是 6，并以 6 为主题数目进行主题识别结果如表 13 所示：

表 13 第二次主题识别结果

Table 13 Second subject recognition result

	主题表征词				
主题 1	万元	公告	证券	账户	项目
	0.0019965528	0.0019172332	0.0019158046	0.0018948785	0.0018698205

再用熵检测的方法筛选高质量主题识，计算得到的 5 个主题的表征词信息熵，结果如表 14 所示：

表 14 第二次主题信息熵

Table 14 Second topic information entropy

主题	主题 1	主题 2	主题 3	主题 4	主题 5
信息熵	1.6139630952428	0.7160501545873	1.3008808532820	0.5072421856513	1.1301493928524

5 个主题的信息熵中，主题 1，3，5 的信息熵高于 1.1，判断这三个主题是高质量主题，而主题 2，4 则判断为噪声主题。

接下来判断每一篇文本是否被 LDA 主题模型充分表示主题，在 6 个主题的 33 篇文本中，有 17 篇文本是没有被 LDA 主题模型充分表示主题的文本，16 篇文本是被 LDA 主题模型充分表示主题的文本。

对于没有被 LDA 主题模型充分表示主题的 17 篇文本，第三次对 16 篇没有被 LDA 主题模型充分表示主题的文本进行 LDA 主题识别度，识别出最优主题数目是 1，并以 1 为主题数目进行主题识别结果如表 15 所示：

表 15 第三次主题识别结果

Table 15 Third subject recognition result

主题 1	('疫苗', 0.004152856)	('账户', 0.0026989763)	('项目', 0.0025991187)	('失信', 0.0024023636)
主题 2	('截图', 0.0028408777)	('证券', 0.002368685)	('微博', 0.0022601166)	('上市公司', 0.0022336715)
主题 3	('疫苗', 0.004152856)	('账户', 0.0026989763	('项目', 0.0025991187	('失信', 0.0024023636)
主题 4	('钟元', 0.0028253077)	('补助', 0.0025135675)	('监委', 0.0023336564)	长春', 0.0021948302)
主题 5	('地震', 0.0059272447)	('通海县', 0.004201092)	('云南省', 0.0023276533)	('发生', 0.002031991)

再用熵检测的方法筛选高质量主题识，计算得到的这 1 个主题的表征词信息

熵是 1.3342942058423092。对于每一篇文本，判断每一篇文本是否被 LDA 主题模型充分表示主题，在 17 篇文本中，已经没有文本是没有被 LDA 主题模型充分表示主题的文本。

最终，我们把三次主题识别得到的高质量主题做一个集合，作为最终的文本集的主题。得到的主题集合如表 16 所示：

表 16 最终主题

Table 16 Final theme

主题 1	('美国', 0.005509399)	('关税', 0.004643482)	('特朗普', 0.0037529143)	('钢铁', 0.0029704035)
主题 2	('长春', 0.002711079)	('长生', 0.002532903)	('疫苗', 0.0018195861)	('公安分局', 0.0014926636)
主题 3	租赁', 0.004220571)	('房租', 0.0038221274)	('租金', 0.0031279188)	('上涨', 0.002675991)
主题 4	('莫焕晶', 0.006745752)	('死刑', 0.004233984)	('放火', 0.0034928808)	('保姆', 0.0024844697)
主题 5	('疫苗', 0.004375345)	('我省', 0.0018545929)	('长生', 0.001846968)	('接种', 0.0018227079),
主题 6	('疫苗', 0.004152856)	('账户', 0.0026989763)	('项目', 0.0025991187)	('失信', 0.0024023636)
主题 7	('疫苗', 0.004152856)	('账户', 0.0026989763	('项目', 0.0025991187	('失信', 0.0024023636)
主题 8	('地震', 0.0059272447)	('通海县', 0.004201092)	('云南省', 0.0023276533)	('发生', 0.002031991)
主题 9	('万元', 0.0019965528)	(公告', 0.0019172332)	('证券', 0.0019158046)	('账户', 0.0018948785)

我们把得到的结果与不解决不平衡问题的主题模型识别方法进行比较，并计算两种方法的 LDA 主题抽取的查准率 P ，查全率 R 与 F1 度量，评估方法的准确性指标。根据文本内容知识[26]得出本文的评价指标公式如式（7）所示：

$$R = \frac{N_2}{N_4}, P = \frac{N_1}{N_3}, F_1 = \frac{2PR}{P + R} \quad (7)$$

其中，N_4 为最终得到的主题数目，N_2 为主题抽取的有效主题的数目；N_1 为有效主题中正确抽取的主题数目，所谓正确抽取的主题指 LDA 所抽取的主题包含在专家评判的领域研究主题之中；N_3 为专家评判的领域主题数目。对两种方法进行比较，解决了不平衡问题的方法得到 9 个主题，有 8 个主题包含在人工判断的主题中，且人工判断的 6 个主题中全部被识别出来；没解决不平衡问题的方主题识别方法通过基于密度峰值的方法得到 8 个主题，有 5 个主题包含在人工判断的主题中，且人工判断的 6 个主题中有 3 个主题被识别出来，比较结果如表 17 所示。

表 17 不同方法的 LDA 主题抽取效果比较

Table 17 Comparison of LDA Topic Extraction Effects by Different Methods

	查准率 P	查全率 R	F_1
LDA	0.6250	0.5000	0.5556
解决不平衡问题的 LDA	0.8889	1．00	0.9412

如表 17 所示，用本文的方法解决不平衡问题后，LDA 模型在主题识别上的查全率，查准率和 F1 值得到大幅提升，对主题识别有很大的改善。

4.8 评估模型的适用性

为了验证本文所提方法在不同数据集大小上的适用性，对于获得的 145 个话题总共 10000 多篇新闻文本，将其分成不同大小批次。共得到 100 篇，500 篇，2000 篇，5000 篇的数据，对每个不同大小的数据，通过解决不平衡问题的方法得最终主题，并与没解决不平衡问题的方法进行比较，最终的比较结果如表 18 所示。

表 18 不同方法的 LDA 主题抽取效果比较

Table 18 Comparison of LDA Topic Extraction Effects by Different Methods

文本数量	方法	查准率 P	查全率 R	F_1
100 篇	LDA	0.8621	0.4552	0.5958
	解决不平衡问题的 LDA	0.9517	0.5379	0.6873
1000 篇	LDA	0.7000	0.5333	0.6054
	解决不平衡问题的 LDA	0.7941	0.7333	0.7637
5000 篇	LDA	0.8800	0.7619	0.8167
	解决不平衡问题的 LDA	0.9136	0.8572	0.8845
10000 篇	LDA	0.8291	0.5455	0.6578
	解决不平衡问题的 LDA	0.8698	0.7576	0.8098

如表 18 结果所示，在不同数据集上，解决不平衡问题的主题识别方法都比不解决平衡方法的查全率，查准率和 F1 值都有一定程度的提升，因此证明模型在不同大小的数据集上都适用，具有适用性。

5 结论

本文针对传统 LDA 模型难以解决不同主题之间预料不平衡的主题识别问题，因此提出一种主题不平衡新闻文本数据集的主题识别方法用于主题识别。该方法判断出文本是否被 LDA 主题模型所充分表示主题，对于充分表示主题的文本，识别高质量主题，去除噪声，对于没被充分展示主题的文本，重复进行 LDA 主题识别操作，直到没有没被充分展示主题的文本，并以多次识别的主题作为最终文本数据集主题，解决了传统 LDA 模型对各文本主题下文本不平衡的主题识别

问题，提高了传统 LDA 主题识别的查全率，查准率和 F_1 值。

参考文献：

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research Archive, 2003, 3:993-1022.

[2] SALTON G, MCGILL M J. Introduction to Modern Information Retrieval [M]. New York:McGraw-Hill,1983:239-240.

[3] DEERWESTER S. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science & Technology, 1990, 41(6):391-407.

[4] HOFMANNT. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: IEEE Press,1999:50-57.

[5] T. Hofmann. Probabilistic Latent Semantic Indexing[C]//Proc. 22nd Ann. Int'l ACM Conf. Research and Development in Information Retrieval (SIGIR '99), Aug. 1999: 50-57.

[6] Li Wan, Leo Zhu, and Rob Fergus. A hybrid neural network-latent topic model[C]//In Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12). La Palma, Canary Islands, 2012:1287-1294.

[7] Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model[C]//In Advances in Neural Information Processing Systems, 2012:2708-2716.

[8] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Replicated softmax: an undirected topic model[C]//In Advances in Neural Information Processing Systems 21 (NIPS-09). Vancouver, Canada, 2009:1607-1614.

[9] A. B. Dieng, C. Wang, J. Gao, and J. W. Paisley. TopicRNN: A recurrent neural network with long-range semantic dependency[C]//CoRR, abs/1611.01702, 2016.

[10] Lau J H, Baldwin T, Cohn T. Topically Driven Neural Language Model[J]. arXiv preprint arXiv:1704.08012, 2017.

[11] Li Ximing, Ouyang Jihong, Zhou Xiaotang. Labelset topic model for multilabel document classification[J]. Journal of Intelligent Information Systems,2016,46(1):83-97.

[12] Wu Mengsung. Modeling query-document dependencies with topic language models for information retrieval[J]. Information Sciences, 2015, 312:1-12.

[13] 刘定祥, 乔少杰, 张永清,等. 不平衡分类的数据采样方法综述[J]. 重庆理工大学学报(自然科学), 2019, 33(7):102-112.

(Liu Dingxiang，Qiao Shaojie，Zhang Yongqing，et al. A Survey on Data Sampling Methods in Imbalance Classification[J]. Journal of Chongqing University of Technology(Natural Science), 2019, 33(7):102 - 112.)

[14] 骆凯敏. 文本分类中不平衡数据的处理[D]. 广东：中山大学硕士学位论文，2005.

(Luo Kaimin. Imbalanced Data Processing in Text Categorization[D]. guangdong : Sun Yat-sen University, 2005.)

[15] CHAWLA N V，BOWYER K W，HALL L O，et a1. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[16] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C] //Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008: 1322-1328.

[17] TomekI. Two modifications of CNN[J]. IEEE Transactions on Systems. Man and Cybernetics, 1976，11(6): 769-772.

[18] 王光，邱云飞，史庆伟. 一种用于中文主题分类的 CSVM 算法[J]. 计算机工程, 2012, 38(8):131-133.

(Wang Guang, Qiu Yunfei, Shi Qingwei. CSVM Algorithm for Chinese Theme Classification[J].Computer Engineering, 2012, 38(8): 131-133.)

[19] 吴雨茜, 王俊丽, 杨丽,等. 代价敏感深度学习方法研究综述[J]. 计算机科学, 2019, 46(5):8-19.
(Wu yuqian, Wang Junli, Yang li, et al. Survey cost-sensitive deep learning methods[J]. Computer Science, 2019, 46(5): 8-19.)
[20] 李红莲, 王春花, 袁保宗. 一种改进的支持向量机 NN-SVM[J]. 计算机学报, 2003, 26(8):1015-1020.
(Li Honglian, Wang Chunhua, Ruan Baozong. An imporved SVM :NN-SVM[J].Chinese Journal of Computers. 2003, 26(8):1015-1020.)
[21] 居亚亚, 杨璐, 严建峰. 基于动态权重的 LDA 算法[J]. 计算机科学, 2019, 46(8):260-265.
(Ju Yaya, Yang Lu, Yan Jianfeng. LDA Algorithm Based on dynamic weight[J]. Computer Science, 2019, 46(8): 260-265.)
[22] 廖列法, 勒孚刚, 朱亚兰. LDA 模型在专利文本分类中的应用[J]. 现代情报, 2017,37(3):35-39.
(Liao Liefa, Le Fugang, Zhu Yalan. The Application of LDA Model in Patent Text Classification[J]. Journal of Modern Information, 2017, 37(3): 35-39.)
[23] 刘江华. 一种基于kmeans聚类算法和LDA主题模型的文本检索方法及有效性验证[J]. 情报科学, 2017, 35(2):16-21.
(Liu Jianghua. A Text Retrieval Method and Validation Based on kmeans Clustering Algorithm and LDA Topic Model[J]. Information Science, 2017, 35(2): 16-21.)
[24] 郭剑飞. 基于 LDA 多模型中文短文本主题分类体现构建与分类[D].黑龙江：哈尔滨工业大学，2014.
(Guo Jianfei. Classification for Chinese Short Text Based on Multi LDA Models [D]. Heilongjiang: Harbin Institute of Technology, 2014.)
[25] 东北大学. 基于优质主题扩展的微博文本分类方法及系统与流程：CN201811064231.3[P]. 2019-02-15.
(Northeastern University. Microblog Text Classification Method, System and Process Based on High Quality Topic Extension: CN201811064231.3[P]. 2019-02-15.)
[26]https://blog.csdn.net/watkinsong/article/details/9836167?utm_medium=distribute.pc_relevant.none-task-blog-OPENSEARCH-2.channel_param&depth_1-utm_source=distribute.pc_relevant.none-task-blog-OPENSEARCH-2.channel_param

(通讯作者: 张亚飞，ORCID：0000-0003-2347-5642，E-mail: zyfeimail@163.com。)

基金项目：本文系国家自然科学基金项目“基于时序要素的跨文本热点新闻话题事件信息融合关键技术研究”（项目编号：61966020），国家自然科学基金项目“汉越双语新闻事件关联分析及摘要方法研究”（项目编号：61762056）和云南省重大科技专项项目“云南省生物医药信息化平台提升建设”（项目编号：2018ZF019）的研究成果之一。

作者贡献声明：

王红斌负责论文的思想的提出和实现；
王健雄参与论文思想的具体实现及论文撰写；
张亚飞和杨恒负责论文的修改完善。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

王健雄. THUCNews 数据集（http://thuctc.thunlp.org/）和自己爬取的搜狐新闻网页上热门新闻话题数据，数据格式为 txt，由作者自己存储，E-mail：Jianxiong_wong@icloud.com