计算机科学
COMPUTER SCIENCE

# 基于集成学习的跨语言文本主题发现方法研究

李帅, 于娟, 巫邵诚

引用本文

李帅, 于娟, 巫邵诚. 基于集成学习的跨语言文本主题发现方法研究[J]. 计算机科学, 2024, 51(6A): 230300201-8.

LI Shuai, YU Juan, WU Shaocheng. Cross-lingual Text Topic Discovery Based on Ensemble Learning [J]. Computer Science, 2024, 51(6A): 230300201-8.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300201

# 基于集成学习的跨语言文本主题发现方法研究

李 帅 于 娟 巫邵诚
福州大学经济与管理学院 福州 350108
(lish1223@163.com)

**摘 要** 跨语言文本主题发现是跨语言文本挖掘领域的重要研究方向,对跨语言文本分析和组织各种文本数据具有较高的应用价值。基于Bagging和跨语言词嵌入改进LDA主题模型,提出跨语言文本主题发现方法BCL-LDA(Bagging,Cross-lingual word embedding with LDA),从多语言文本中挖掘关键信息。该方法首先将Bagging集成学习思想与LDA主题模型结合生成混合语言子主题集;然后利用跨语言词嵌入和K-means算法对混合子主题进行聚类分组;最后使用TF-IDF算法对主题词进行过滤排序。汉语-德语、汉语-法语主题发现实验表明,该方法在主题连贯性和多样性方面均表现优异,能够提取出语义更加相关且主题更加连贯多样的双语主题。

**关键词**:主题发现;跨语言;LDA;主题聚类;德语;法语
**中图分类号** TP391.1

## Cross-lingual Text Topic Discovery Based on Ensemble Learning

LI Shuai,YU Juan and WU Shaocheng
School of Economics and Management,Fuzhou University,Fuzhou 350108,China

**Abstract** Cross-lingual text topic discovery is an important research direction in the field of cross-lingual text mining,and it has high application value for cross-lingual text analysis and organization of various text data. Based on Bagging and cross-lingual word embedding to improve the LDA topic model,a cross-lingual text topic discovery method BCL-LDA(Bagging,cross-lingual word embedding with LDA) is proposed to mine key information from multilingual text. This method first combines the Bagging integrated learning idea with the LDA topic model to generate a mixed language subtopic set. Then it uses cross-lingual word embedding and K-means algorithm to cluster and group the mixed subtopics. Finally,the TF-IDF algorithm is used to filter and sort the subject words. The Chinese-German and Chinese-French topic discovery experiments show that this method performs well in terms of topic coherence and diversity,and can extract bilingual topics with more relevant semantics and more coherent and diverse topics.

**Keywords** Topic discovery,Cross-lingual,LDA,Topic clustering,German,French

## 1 引言

随着互联网的快速发展和全球一体化建设的不断深入,不同国家和地区之间的交流日益频繁,人们越来越关心国际时事并积极主动地参与到全球范围内的讨论中。由此产生的新闻报道和演讲评论层出不穷,互联网上各种语言的文本数据快速增长。面对海量的文本数据,如何从中准确地挖掘出有用的关键信息值得深入研究。同时,瞬息万变的国际局势使得单一语言文本数据所带来的信息量已无法满足国家政策制定和跨国组织发展决策的需要,跨语言比较不同国家和地区的人们对同一事件的不同态度对于国家和跨国组织越来越重要。主题发现能够有效提取文本中的关键信息,但目前已有的主题发现研究主要面向单语文本数据[1-2],而在跨语言文本主题发现方面的相关研究较少。

跨语言文本主题发现又称跨语言文本主题识别、跨语言文本主题抽取和跨语言文本主题挖掘等,属跨语言文本挖掘的范畴。跨语言文本主题发现源于对单语言文本的主题发现,旨在从两种及以上语言的文本集中发现和挖掘潜在的共同主题信息,用于对多种语言的混合文本进行主题发现。跨语言文本主题发现在跨语言信息资源的搜索、多语言新闻热点监测、跨语言文本分类和跨语言文本聚类等跨语言文本挖掘任务领域中都有着较高的应用价值,在挖掘海量的跨语言文本数据资源任务中发挥着巨大的作用。

现有的少量跨语言文本主题发现方法的研究大多是对单语主题模型的扩展,效果尚不够令人满意。为此,本文提出一种基于Bagging(Bootstrap Aggregating)[3]集成学习和跨语言词嵌入[4]的LDA(Latent Dirichlet Allocation)[5]跨语言主题发现方法BCL-LDA(Bagging,Cross-lingual Word Embedding with LDA),基于Bagging集成学习的思想训练多个LDA单语主题模型,使用跨语言词嵌入生成跨语言主题向量,并利用聚类算法实现主题分组。本文第2章分析相关研究的现状;第3章详细说明本文提出的跨语言主题发现方法BCL-LDA;

基金项目:国家自然科学基金(71771054,72171090)
This work was supported by the National Natural Science Foundation of China(71771054,72171090).
通讯作者:巫邵诚(2558861318@qq.com)

第 4 章为汉-德、汉-法跨语言主题发现实验与结果分析；最后总结全文。

## 2 相关研究

跨语言文本主题发现的关键在于如何跨越语言的障碍从两种及以上语言的文本中发现主题。相较于单语言主题发现，跨语言主题发现研究的难度更大，成果较少。已有相关研究主要可分为：文档链接法、词汇链接法和词嵌入法。

文档链接法将平行文档进行一对一链接，假设每对文档共享相同的主题分布，并对每个主题使用不同的主题词分布。例如，Mimno 等[6]提出的 PLTM 跨语言主题模型，利用单语 LDA 模型从平行文档对中提取出跨语言主题。Yu 等[7]假设双语文档语义相关，利用双语主题模型 BLDA 完成跨语言知识链接任务。Zosa 等[8]将单语动态主题模型 DTM(Dynamic Topic Models)与多语言主题模型 PLTM 结合起来，提出了多语言动态主题模型 ML-DTM，旨在捕获随时间演变的跨语言主题。该类方法对平行文档的质量要求较高，对低资源语言不够友好。

词汇链接法通过构建不同语言词语之间的互译关系来链接双语主题，该方法主要分为两类。一类为借助已有的机器翻译工具将跨语言文本统一为单语言文本，从而避免语言不同给主题发现带来的影响。如，Leek 等[9]采用自制的术语翻译系统，结合双语词典将汉语文本翻译成英语文本，从而在同一语言下发现跨语言主题。Chen 等[10]在 LDA 模型的基础上提出了 ICE-LDA 模型，对中英文文档集分别进行主题建模，并利用百度翻译工具将不同语言主题映射到同一语言空间中计算相似度。该类方法应用较为简单，但机器翻译的性能很大程度上影响其效果。另一类通过构建双语词典来代替机器翻译工具。如，Jagarlamudi 等[11]基于双语词典提出的 JointLDA 多语言主题模型，可以从未对齐的不同语言的语料库中将相关主题合并为一个多语言主题。Zhang 等[12]使用基于双语词典定义的软约束正则化及其似然函数对概率潜在语义分析模型 PLSA(Probabilistic Latent Semantic Analysis)进行了扩展，提出的 PCLSA(Cross-Lingual Latent Semantic Analysis)模型可以有效地从多语言文本数据中提取跨语言潜在主题。Boyd-Graber 等[13]提出的 MUTO(Multilingual Topic Model)利用双语词典等双语信息从未对齐的文本中使用随机 EM(Expectation-Maximization)找到多语言语料库中的共享主题。Liu 等[14]以汉语-高棉语作为研究对象提出了双语主题模型 KCB-LDA，通过构建双语词典将汉语和高棉语映射到同一概念抽象层，然后将概念分组到相同的主题空间，实验显示该双语主题模型具有较好的预测能力。该类方法能有效地对不同语言的词语进行链接，但是双语词典的大小和质量对跨语言分析的影响很大，且无法避免一词多义的问题。

随着神经网络在文本挖掘领域研究和应用的不断深入，词嵌入得到了快速的发展，单语言和多语言词嵌入的研究成果也越来越丰富。在跨语言主题发现任务中，词嵌入的应用主要有两类，一类将两个单语空间向量映射到同一语义空间中进行比较。例如，Yang 等[15]提出了一种基于跨语言神经主题模型的汉越新闻话题发现方法，利用小规模的平行语料将汉越单语主题向量映射到同一语义空间中，然后使用 K-means 算法对双语主题表征进行聚类，从而发现新闻事件簇的话题。Chang 等[16]提出了一种基于后匹配的跨语言主题模型 PM-LDA。首先，分别构建单语言主题模型；其次，使用双语词典训练的翻译矩阵将汉英单语词嵌入映射到跨语言词空间中，将每个主题视为跨语言词空间中的向量；最后，利用聚类算法对主题向量进行聚类分组。Chang 等[17]利用基于映射的跨语言词嵌入来推断跨语言主题，提出了跨语言主题模型 Cb-CLTM。另一类结合基于大规模语料库预训练得到的多语言词嵌入进行跨语言主题建模。例如，Chen 等[18]基于 R(rectr)实现了跨语言主题的可重现提取，利用开源的跨语言词嵌入从英语、德语和法语新闻语料库中提取出了高质量的跨语言主题。Bianchi 等[19]提出的 ZeroShotTM 使用多语言词嵌入模型 M-BERT 来代替词袋表示，并通过结合 ProdLDA[20]的神经架构从段落表示中学习主题表示，实现了零镜头的跨语言主题建模。该类方法虽然需要大量的标注数据作为训练集，但可以综合考虑词语的上下文语义信息，有效解决一词多义问题。

为了充分利用跨语言词嵌入优秀的跨语言文本表示能力，并从文本聚类的角度研究跨语言文本主题发现，本文将集成学习的思想引入跨语言文本主题发现研究中。集成学习将多个基础机器学习算法按照一定的策略组合起来完成学习任务，从而获得一个效果更优的强学习器。目前集成学习的思想被广泛应用在文本挖掘研究中。Dai 等[21]基于 Spark 和 Stacking 提出了一种集成方法 S-FWS，用于舆情情感分析。Liang 等[22]通过集成多个模型的优势完成了命名实体识别，有效提高了模型的有效性和普适性。Feng 等[23]利用集成学习和回译的方法来解决维汉神经机器翻译由于语料匮乏导致的效果欠佳的问题，使用自助采样法进行重采样得到多个数据子集，并在此基础上训练多个子模型。

## 3 BCL-LDA 跨语言文本主题发现方法

BCL-LDA 是基于集成学习和跨语言词嵌入的跨语言文本主题发现方法。该方法首先构建单语言主题模型，然后运用词嵌入将主题编码为向量，最后使用聚类算法完成跨语言主题分组。本文的研究思路受到了文献[16]基于后匹配构建跨语言主题模型 PM-LDA 的启发。不同点在于，本文方法结合了 Bagging 集成学习的思想，从原始数据集中扩展出多个数据子集，并在此基础上构建多个单语言主题模型，使得模型可以从数据集中提取出更多的主题特征，生成更多子主题；并且，本文方法利用效果更好的跨语言词嵌入对主题进行向量表示，有效减少了 PM-LDA 在跨语言聚类分组时由于主题信息模糊而出现单语言主题聚集的现象。

BCL-LDA 跨语言主题发现方法的流程如图 1 所示。该方法主要可以分为 4 个模块。(1)文本预处理模块，根据语言结构的不同，为不同的语言选择不同的预处理工具；(2)子主题生成模块，BCL-LDA 基于 Bagging 集成学习的采样策略和单语 LDA 主题模型，生成语言 $l_1$ 和 $l_2$ 混合子主题集；(3)跨语言主题聚类模块，利用跨语言词嵌入和聚类算法将语言 $l_1$ 和 $l_2$ 混合子主题集中相似的主题聚为一类；(4)主题词过滤模块，通过对主题词进行过滤发现语言 $l_1$ 和 $l_2$ 双语主题，后文将按步骤详细说明该方法。

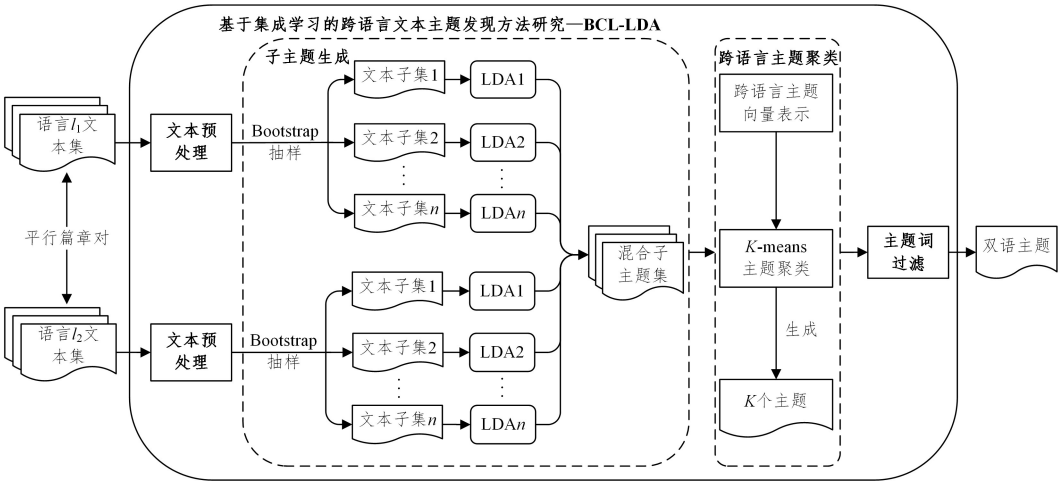


图 1　基于集成学习的跨语言文本主题发现方法 BCL-LDA 的流程图

Fig. 1　Flow chart of cross-lingual topic discovery based on ensemble learning－BCL-LDA

### 3.1　文本预处理

文本预处理模块针对不同语言的结构属性选择不同的方法进行预处理，主要步骤包括文本清洗、分词或词形还原、词性标注、停用词及停用词性的删除。对于汉语文本，使用 NLPIR[24] 工具进行分词和词性标注；对于以德语、法语为代表的印欧语系语言，选用 TreeTagger[25] 工具完成词语的词性标注和词形还原工作[26]。最后，根据停用词表和停用词性表过滤文本中的停用词，得到预处理后的文本。

### 3.2　子主题生成

子主题生成模块的输入是经过文本预处理的语言 $l_1$ 和 $l_2$ 平行文本，输出是语言 $l_1$ 和 $l_2$ 混合子主题集。该模块分为数据子集构建和基于 LDA 的单语主题建模两部分。其中，数据子集构建是基于 Bagging 集成学习的采样策略，利用 Bootstrap[27] 算法从预处理后的文档数据集中进行有放回地随机重复抽样，生成 $N$ 个大小相近但内容存在差异的数据子集。基于 LDA 的单语主题建模是在 $N$ 份数据子集的基础上分别构建基于 LDA 的单语主题模型，每个 LDA 模型生成一定数量的单语子主题，将不同语言的子主题组合生成混合子主题集。

LDA[5] 主题模型是一种从大量文本数据中进行无监督学习的机器学习方法，用于发现文档集中隐含的主题信息和关键信息。模型是基于“文档-主题-词”三级分层的贝叶斯概率模型，其示意图如图 2 所示。其中，$k$ 表示文本主题设定个数，$M$ 和 $N$ 分别表示文档中文档数量和主题下的词语数量；$\alpha$ 和 $\beta$ 分别为给定的文档-主题和主题-词的 Dirichlet 先验分布；$\theta$ 和 $\Phi(k)$ 分别为文档-主题分布和主题-词分布；$z$ 表示文档中词语的主题；$w$ 为文档中词语。

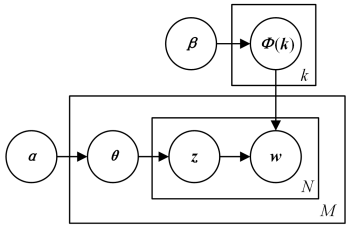


图 2　LDA 主题模型示意图

Fig. 2　Schematic diagram of LDA topic model

通过 Bootstrap 自助采样得到的 $N$ 份数据子集之间近似服从同一分布，但内容略微存在差异。数据子集包含原始数据集中部分重复文档，对原始数据集中的隐含子主题进行了强化，使得生成的子主题信息更加清晰和多样化。

### 3.3　跨语言主题聚类

跨语言主题聚类模块利用跨语言文本聚类的方法完成混合语言子主题集的主题聚类分组，从而获得多语言数据集的关键主题信息。该模块分为跨语言主题向量表示和 K-means 主题聚类两个子模块。

#### 3.3.1　跨语言主题向量表示

跨语言主题向量表示的输入是语言 $l_1$ 和 $l_2$ 的混合子主题集，输出由主题向量组成的嵌入矩阵。该模块通过使用跨语言词嵌入将不同语言的子主题在同一个向量空间中进行表示。首先，计算子主题中每个主题词的词向量；其次，对这些词向量进行平均，从而得到表示子主题信息的主题向量。本文选择 2019 年由 Google 提供的跨语言词嵌入工具 USEM (Universal-Sentence-Encoder-Multilingual)[28] 对语言 $l_1$ 和 $l_2$ 子主题进行向量表示。

例如，语言 $l_1$ 和 $l_2$ 混合子主题集中共有 $2n$ 个子主题，每个子主题都由 $m$ 个主题词组成。本文使用 USEM 对每个子主题的主题词进行向量化表示，其中语言 $l_1$ 子主题的主题词向量化为 $\{\overrightarrow{w_1^{l_1}},\overrightarrow{w_2^{l_1}},\cdots,\overrightarrow{w_m^{l_1}}\}$，平均后得到语言 $l_1$ 子主题向量 $\overrightarrow{t^{l_1}}$；语言 $l_2$ 子主题的主题词向量化为 $\{\overrightarrow{w_1^{l_2}},\overrightarrow{w_2^{l_2}},\cdots,\overrightarrow{w_m^{l_2}}\}$，平均后得到语言 $l_2$ 子主题向量 $\overrightarrow{t^{l_2}}$。最终，将语言 $l_1$ 和 $l_2$ 的混合子主题集表示为主题向量嵌入矩阵 $\{\overrightarrow{t_1^{l_1}},\overrightarrow{t_2^{l_1}},\cdots,\overrightarrow{t_n^{l_1}},\overrightarrow{t_1^{l_2}},\overrightarrow{t_2^{l_2}},\cdots,\overrightarrow{t_n^{l_2}}\}$。

#### 3.3.2　K-means 主题聚类

通过 USEM 对两种语言的子主题进行跨语言主题向量化表示后，本文选择使用 K-means 聚类算法对混合子主题进行聚类分组，输入为主题向量嵌入矩阵，输出为 $k$ 个主题，聚类类别常数 $k$ 与 LDA 建模主题数设置保持一致。选择 K-means 算法进行聚类的原因在于可以通过设置聚类数 $k$ 与 LDA 单语主题建模数保持一致，从而保证较为稳定的主题输出。主题向量之间的相似度计算采用余弦距离：

$$M_{ij}=\frac{\vec{t}_i\cdot\vec{t}_j}{\|\vec{t}_i\|\cdot\|\vec{t}_j\|} \tag{1}$$

其中，$\vec{t}_i$ 和 $\vec{t}_j$ 表示主题向量集合中第 $i$ 个和第 $j$ 个主题向量，$\|\vec{t}\|$ 表示向量的模运算。

通过使用 K-means 聚类算法，将混合主题向量集合分配给每一个聚类中心，形成 $k$ 个主题，每个主题由两种语言的词语组成。

### 3.4 主题词过滤

通过对子主题集进行聚类，将 $2n$ 个子主题划分为 $k$ 个主题。本文参考单语言主题模型 BERTopic[29] 的思想，将每一个聚类主题簇作为一个文档，采用 TF-IDF(Term Frequency-Inverse Document Frequency)[30] 算法对每个聚类簇的主题词进行重要性排序，通过选取前 Top-$n$ 个主题词对主题信息进行表示。

TF-IDF 通过计算词语频率 $TF$ 和逆文档频率 $IDF$ 来计算词语的重要性，如式(2)—式(4)所示。

$$TF_i=\frac{N_i}{N} \tag{2}$$

$$IDF_i=\log\left(\frac{Y}{Y_i+1}\right) \tag{3}$$

$$TF\text{-}IDF_i=TF_i*IDF_i \tag{4}$$

其中，$TF_i$ 表示词频，$N_i$ 表示单词 $i$ 在该文档中出现的次数，$N$ 表示该文档中的总词数。$IDF_i$ 表示逆文档频率，$Y$ 表示语料库文档总数，$Y_i$ 表示语料库中包含词语 $i$ 的文档数。

## 4 实验分析

为了说明跨语言主题发现方法的有效性，采用实验将本文提出的 BCL-LDA 方法与经典代表方法 PLTM[6]，JointLDA[11] 和 PM-LDA[16] 进行对比分析。其中，PLTM 属于文档链接法，是基于平行文档链接的代表，假设同一对双语文档共享同一主题分布。JointLDA 是词汇链接法的代表，该方法以双语词典作为跨语言链接的纽带来生成跨语言主题分布。PM-LDA 是词嵌入法的代表，该方法首先使用训练的翻译模型将两个单语词向量空间映射到跨语言词向量空间中，然后通过 LDA 分别得到单语主题，将主题向量化后，利用 DBSCAN 聚类算法进行跨语言主题分组。

### 4.1 数据集

由于当前缺乏检验跨语言主题发现方法优劣的标准语料库，本文选取多语言公开数据集 TED2020(TED-Talk 2020)[31] 中的汉语-德语(汉-德)、汉语-法语(汉-法)平行语料作为本文的实验数据集。TED2020 是一个从 TED 网站上爬取的经过翻译后的多语言公开语料库，提供包括汉语、德语和法语在内的 108 种语言的平行文档。其中，汉-德平行语料包含平行句子 293 124 对，按空行划分得到有效平行文档 2 650 对，共计 21.6 MB；汉-法平行语料包含平行句子 396 360 对，按空行划分得到有效平行文档 3 710 对，共计 29 MB。

### 4.2 评价指标

采用常见的文本主题发现评价指标[17,32-33]，本文主要从主题的连贯性和多样性两个方面对跨语言主题质量进行评估。

主题连贯性是指同一个主题下的主题词之间要尽可能地保持连贯一致。为此，本文选用由 Hao 等[34] 提出的一种多语言主题模型的评估指标——跨语言标准化逐点互信息 CNPMI(Cross-lingual Normalized Pointwise Mutual Information)。CNPMI 是对标准化逐点互信息 NPMI(Normalized Pointwise Mutual Information)[35] 的多语言扩展，用于衡量同一主题中不同语言单词之间的接近程度。NPMI 主要用于评估单语主题模型中主题词分布的连贯性，与人类的判断密切相关。NPMI 的计算式如式(5)所示。

$$NPMI(w_i,w_j)=\frac{\log\left(\frac{P(w_i,w_j)}{P(w_i)P(w_j)}\right)}{-\log(P(w_i,w_j))} \tag{5}$$

其中，$P(w)$ 表示语料库中单词 $w$ 出现的概率，$P(w_i,w_j)$ 表示单词 $w_i$ 和 $w_j$ 在语料库中的共现概率。

Hao 等[34] 基于大量可比较的维基百科文档作为计算 CNPMI 的参照语料库，来估计 $P(w)$ 和 $P(w_i,w_j)$。CNPMI 主要是通过平均同一主题下每个双语词对之间的 NPMI 值得到的，同一种语言词对之间不计算 NPMI 值。CNPMI 的计算式如式(6)所示。

$$CNPMI(l_1,l_2,k)=\frac{\sum_{i,j}^{C}NPMI(w_{i,l_1},w_{j,l_2})}{C^2} \tag{6}$$

其中，$l_1$ 和 $l_2$ 表示两种不同的语言，$k$ 表示第 $k$ 个主题，$C^2$ 表示第 $k$ 个主题下有 $C^2$ 个双语词对。

主题多样性是指主题之间是有区别的，通过计算主题中贡献度最高的前 $N$ 个主题词为主题所独有的比例来衡量主题之间的区别性[36]，比例越高，主题多样性越高。本文使用逆平均 Jaccard 相似度[17](Inverse Average Jaccard Similarity, inverse-AJS)进行评估。inverse-AJS 是将平均 Jaccard 相似度计算扩展到跨语言主题发现的评估中来探索主题之间的重合度，其定义如式(7)所示。

$$inverse\text{-}AJS(T)=1-\frac{\sum_{l\in L}\sum_{t_1,t_2\in T}\frac{Top(t_1,l)\cap Top(t_2,l)}{Top(t_1,l)\cup Top(t_2,l)}}{|L|\times|T|\times(|T|-1)/2} \tag{7}$$

其中，$L$ 表示语言集，$T$ 表示主题集，$|L|$ 表示语言种类数，$|T|$ 表示主题个数，$Top(t,l)$ 表示主题 $t$ 中属于语言 $l$ 的前 $N$ 个主题词。inverse-AJS 度量范围为[0,1]，0 表示冗余主题，1 表示更多样性主题。

### 4.3 结果分析

为保证实验的科学性，对各模型的共享参数采用相同的设置。其中文档-主题分布 $\alpha=1/|T|$，主题-词分布 $\beta=0.01$，$|T|$ 表示主题个数，每个模型进行 1 000 次迭代采样。

将本文提出的跨语言文本主题发现方法 BCL-LDA 设置 50 次和 100 次 Bootstrap 自助抽样进行效果对比，分别记为 BCL-LDA50 和 BCL-LDA100。对于 Bootstrap 抽样数据训练的 LDA 子主题集，选取每个子主题前 5 个词进行跨语言主题聚类。

#### 4.3.1 主题连贯性分析

跨语言文本主题发现方法在 TED2020 的汉-德、汉-法数据集上的 CNPMI 值比较结果如图 3 所示。图中，“ZH-DE”表示 TED2020 的汉-德数据集；“ZH-FR”表示 TED2020 的汉-法数据集；“Topics: $n$”表示主题个数 $|T|$ 为 $n$；“Top Words”表示每个主题的前 Top-$n$ 个词语，后文保持相同设置。

由图 3 可知：

(1)PM-LDA 和 PLTM 在汉-德和汉-法两个数据集上的主题连贯性较低。PM-LDA 是基于 LDA 单语建模后使用词嵌入进行主题的后匹配，使用 DBSCAN 进行聚类分组，主题的单一语言聚集问题降低了 PM-LDA 的主题连贯性。PLTM 是基于文档链接法，假设每对文档共享相同主题分布，数据集的对齐质量对其结果有较大影响。

(2) JointLDA 的主题连贯性表现优于 PM-LDA 和 PLTM，在汉-法语料集上优于 BCL-LDA50，低于 BCL-LDA100。JointLDA 是通过构建双语词典链接两种不同语言的主题，能较好地发现文本的主题信息，并通过双语词典建立

语言之间的联系。但双语词典的规模大小、质量以及分词方法的不同都会很大程度地影响到JointLDA主题发现的效果;随着更多新词的出现,容易出现较多的未登录词,且不能解决一词多义的现象。

(3)本文提出的跨语言主题发现方法BCL-LDA在TED2020汉-德和汉-法数据集上的主题连贯性表现最优。BCL-LDA既考虑了词语之间的概率统计信息,又考虑了子主题之间的语义信息,并且不需要额外构建双语词典。通过预训练的跨语言词嵌入建立不同语言之间的联系,在低资源语言主题发现任务中也具有较好的适用性。

(4)相较于其他模型,BCL-LDA显著提高了CNPMI值,且表现稳定。当主题个数$|T|=10,20,30,40,50$时,其都有较为平稳的表现。本文将自主抽样的次数分别设置了50和100两个数值,其中迭代100次的模型表现效果优于50次。考虑到设置更多自主抽样次数会使训练模型的时间大幅增加,本文选择将自主抽样的次数设置为100次。

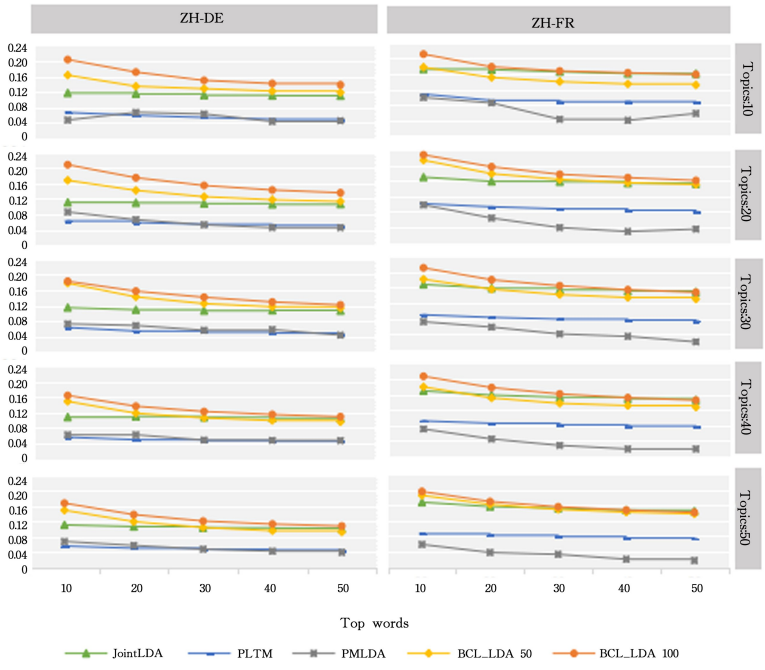


图3　各模型方法在TED2020汉-德、汉-法数据集上的CNPMI值对比

Fig. 3　Comparison of CNPMI values of each model method on TED2020 Chinese-German and Chinese-French datasets

### 4.3.2　子主题词个数的影响分析

为了探究跨语言主题聚类时子主题词的数量对跨语言文本聚类及最终跨语言主题发现效果的影响,本文选择效果更好的自助抽样100次的模型BCL-LDA100进行实验。通过选取不同数量的子主题词进行分析和比较,选择的子主题词分别为5,10,15和20。

在汉-德和汉-法两个数据集上的CNPMI值实验结果如所示。

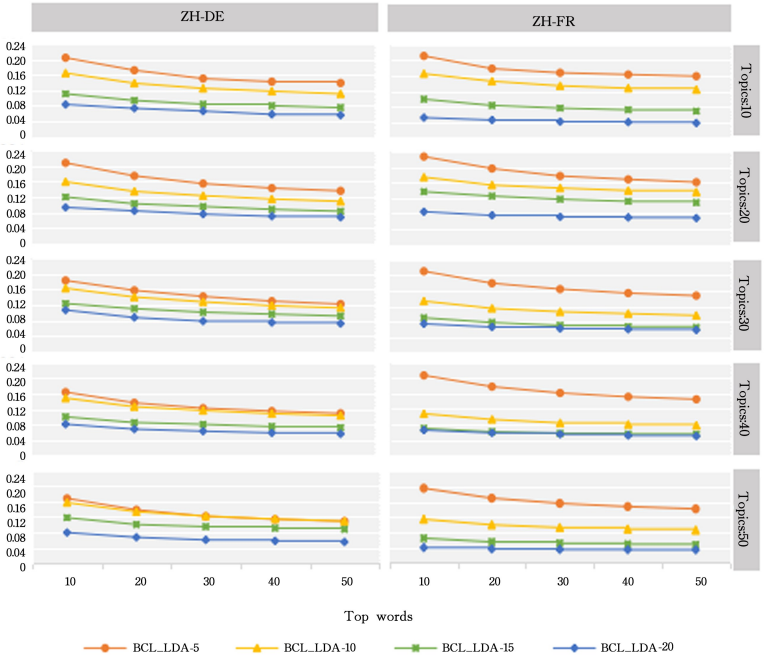


图4　BCL-LDA在两个数据集上选择不同数量子主题词实验的CNPMI值对比图

Fig. 4　Comparison of CNPMI values in BCL-LDA with different numbers of sub-theme words selected on two datasets

由图 4 可知，在两个数据集上，随着子主题词个数的增加，本文方法的主题连贯性效果越来越差；当子主题词个数为 5 时表现最好，这表明前 5 个词可以较好地表示子主题信息。由于本文用子主题中所有主题词的词向量平均后表示主题向量，所以随着子主题词个数的增加，主题向量的主题信息被稀释，不能有效地表现出子主题的鲜明性；如果选择更少的子主题词个数，则会导致主题不够明确。因此，本文选择前 5 个子主题词来表示子主题信息。

#### 4.3.3 主题多样性分析

表 1 列出了主题个数 $|T|=10,20,30,40,50$ 时，各模型在两个数据集上的主题多样性表现。本文选择前 10 个主题词进行实验，可以观察到每个模型都有较高的 inverse-AJS 值，且差距不大。其中在汉-德和汉-法语料集上，当 $|T|=10$，20 时，BCL-LDA100 具有最高值；随着主题个数增加，PLTM 模型占据优势；JointLDA 模型和 PM-LDA 模型的主题多样性最低，BCL-LDA50 和 BCL-LDA100 模型的主题多样性始终高于 JointLDA 和 PM-LDA 模型。

表 1 各模型方法在两个数据集上的主题多样性实验对比

Table 1 Comparison of topic diversity of each model on two datasets

| 数据集 | 模型 | 主题数 | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 | 50 |
| ZH-DE | PLTM | 0.970 | 0.977 | 0.989 | **0.993** | **0.998** |
| | JointLDA | 0.903 | 0.939 | 0.963 | 0.971 | 0.972 |
| | PM-LDA | 0.948 | 0.972 | 0.969 | 0.980 | 0.980 |
| | BCL-LDA50 | 0.961 | 0.981 | **0.990** | 0.992 | 0.994 |
| | BCL-LDA100 | **0.977** | **0.987** | 0.989 | 0.992 | 0.990 |
| ZH-FR | PLTM | 0.968 | 0.973 | **0.987** | **0.992** | **0.995** |
| | JointLDA | 0.774 | 0.858 | 0.884 | 0.910 | 0.917 |
| | PM-LDA | 0.811 | 0.958 | 0.958 | 0.956 | 0.948 |
| | BCL-LDA50 | 0.967 | 0.971 | 0.970 | 0.973 | 0.974 |
| | BCL-LDA100 | **0.972** | **0.978** | 0.977 | 0.980 | 0.978 |

基于上述观察可知，本文方法在不同语言对的数据集上都保持着较好的竞争性，随着主题数的增加，PLTM 模型的主题多样性表现更好；由于 PM-LDA 在主题聚类时出现了单语言主题聚集的现象，且未对聚类后的主题词进行过滤，导致 PM-LDA 模型在主题多样性方面表现最差；同时，JointLDA 模型是通过双语词典实现跨语言建模，无法避免一词多义问题，且词典质量也会对主题建模结果产生影响，导致 JointLDA 模型的主题多样性表现也较差。

此外，PLTM 和 JointLDA 在主题建模时都会忽略低频词，更多地关注高频词，而本文方法通过基于集成学习的思想抽样训练多个 LDA 模型，降低了模型对高频词的依赖；同时，本文通过结合跨语言文本聚类方法，综合考虑了文本的概率统计信息和语义信息；TD-IDF 算法更是降低了主题的冗余度，所以本文方法有较为优异的表现。由于本文结合了文本聚类方法对主题进行分组聚类，因此本文方法在主题多样性方面还有提升空间。

#### 4.3.4 消融实验

跨语言文本主题发现方法 BCL-LDA 旨在利用 Bagging 集成学习和文本聚类的思想对两种语言的文本进行主题发现，并通过 TF-IDF 算法进行主题词过滤排序。为验证本文基于 Bagging 集成学习的思想和 TF-IDF 算法对主题连贯性和多样性提升的有效性，本文设计了一组消融实验。模型 1 不考虑集成学习的思想，只利用一对 LDA 主题模型分别对不同语言的文本进行主题建模，然后进行聚类分组；模型 2 不添加 TF-IDF 主题词过滤层，直接按照主题词的个数进行排序。其中，主题连贯性值为当主题数 $|T|=10,20,30,40,50$ 时，分别计算 10，20，30，40 和 50 个主题词的 CNPMI 值的平均值；主题多样性为当主题数 $|T|=10,20,30,40,50$ 时，前 10 个主题词的 inverse-AJS 值。实验结果如表 2 和表 3 所列。

表 2 主题连贯性实验对比

Table 2 Comparison of topic coherence experiment

| 数据集 | 模型 | 主题数 | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 | 50 |
| ZH-DE | 模型 1 | 0.066 | 0.079 | 0.079 | 0.052 | 0.045 |
| | 模型 2 | 0.153 | 0.159 | 0.141 | 0.126 | 0.127 |
| | 本文方法 | **0.161** | **0.163** | **0.147** | **0.129** | **0.130** |
| ZH-FR | 模型 1 | 0.078 | 0.058 | 0.059 | 0.038 | 0.045 |
| | 模型 2 | 0.172 | 0.179 | 0.166 | 0.163 | 0.159 |
| | 本文方法 | **0.176** | **0.186** | **0.172** | **0.168** | **0.163** |

表 3 主题多样性实验对比

Table 3 Comparison of topic diversity experiment

| 数据集 | 模型 | 主题数 | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 | 50 |
| ZH-DE | 模型 1 | 0.973 | 0.964 | 0.961 | 0.958 | 0.956 |
| | 模型 2 | 0.957 | 0.964 | 0.972 | 0.978 | 0.981 |
| | 本文方法 | **0.977** | **0.987** | **0.989** | **0.992** | **0.990** |
| ZH-FR | 模型 1 | 0.958 | 0.957 | 0.958 | 0.956 | 0.959 |
| | 模型 2 | 0.952 | 0.964 | 0.965 | 0.968 | 0.967 |
| | 本文方法 | **0.972** | **0.978** | **0.977** | **0.980** | **0.980** |

通过消融实验对比结果可知，在汉-德和汉-法两个数据集上，本文方法和模型 2 通过基于 Bagging 集成学习的思想，相较于模型 1，无论是主题连贯性还是主题多样性都有较大提升，同时能够生成语义更加相关的主题。此外，通过 TF-IDF 算法对主题词过滤，在汉-德和汉-法数据集上，本文方法相较于模型 2 主题连贯性提升不高；但在汉-德数据集上，主题多样性平均增加了 1.65%，在汉-法数据集上，主题多样性平均增加了 1.24%。

#### 4.3.5 双语主题词提取

为进一步说明本文方法的有效性，对 TED2020 汉-德、汉-法数据集均进行了双语主题词提取实验，提取的主题示例如表 4 和表 5 所列。

表 4 TED2020 汉-德主题示例

Table 4 Examples of TED2020 Chinese-German topics

| 主题 | 汉-德主题词 |
|---|---|
| 主题 1 | 电脑 技术 信息 数据 网络 公司 系统 产品<br>Computer(电脑) Technologie(技术) Internet(互联网) Seite(页面)<br>System(系统) Information(信息) Netzwerk(网络) Google(谷歌) |
| 主题 2 | 设计 艺术 项目 建筑 空间 作品 照片 城市<br>Gebäude(建筑) Stadt(城市) Ort(地方) Bauen(建筑物)<br>Design(设计) Wasser(水) Projekt(项目) Haus(房子) |
| 主题 3 | 疾病 病人 治疗 医生 细胞 健康 医疗 大脑<br>Krankheit(疾病) Zelle(细胞) Körper(身体) Medikament(药物)<br>Patient(病人) Organ(器官) Gehirn(脑) Arzt(医生) |
| 主题 4 | 气候 地球 能量 变化 火星 黑洞 全球 星系<br>Planet(行星) Wasser(水) Ozean(大海) Tier(动物) Raum(空间)<br>Universum(宇宙) Energie(能量) Licht(光) Stern(恒星) |
| 主题 5 | 学校 感觉 女性 教育 男人 故事 父母 家庭<br>Mann(男人) Mädchen(女孩) Mutter(母亲) Fühlen(情感)<br>Schule(学校) Familie(家庭) Kind(孩子) Eltern(父母) |
| 主题 6 | 艺术 作品 感觉 故事 喜欢 声音 经历 语言<br>Spielen(游戏) Fühlen(感觉) schreiben(写作) Musik(音乐)<br>Lernen(学习) Kunst(艺术) Film(电影) Sprache(语言) |

表 4 是 TED2020 汉-德数据集的跨语言主题示例，从表

中可以看出抽取的主题主要为科技、城市建设、医疗、空间探索、教育和人文艺术 6 个方面。每个主题都生成了较为直观、连贯的主题词，且同一主题下不同语言的主题词都由语义相同或相近的主题词构成，可以较好地表示主题信息。

表 5 是关于 TED2020 汉-法语料的双语主题示例，从表中可以发现抽取的主题主要是关于科技、生态环境、国际政治、医疗卫生、大脑神经研究和宇宙探索 6 个方面。由此可以看出，TED2020 汉-法语料的演讲主题大多与人类的高科技、生存、起源、国际局势和医疗卫生等相关。通过跨语言主题发现方法可以快速掌握其中的关键信息，为研究人员节省大量的时间，也为相关研究提供有力的工具支持。

表 5　TED2020 汉-法主题示例

Table 5　Examples of TED2020 Chinese-French topic

| 主题 | 汉-德主题词 |
|---|---|
| 主题 1 | 机器 电脑 技术 机器人 计算机 人类 设计 科技<br>machine(机器) technologie(技术) donnée(数据) système(系统)<br>ordinateur(计算机) construire(构建) humain(人类) robot(机器人) |
| 主题 2 | 动物 生物 海洋 地球 物种 人类 植物 保护<br>espèce(物种) animal(动物) dinosaurs(恐龙) forêt(森林)<br>changement(变化) climat(气候) evolution(进化) humain(人类) |
| 主题 3 | 国家 美国 政府 政治 社会 经济 中国 战争<br>gouvernement(政府) politique(政治) Chine(中国) Afrique(非洲)<br>guerre(战争) économie(经济) mondial(全球) américain(美国) |
| 主题 4 | 细胞 感染 疾病 基因 病毒 细菌 生物 癌症<br>cellule(细胞) gene(基因) bactérie(细菌) adn(DNA)<br>cancer(癌症) virus(病毒) maladie(疾病) molecule(分子) |
| 主题 5 | 神经 大脑 身体 记忆 细胞 意识 神经元 区域<br>neurone(神经) cerveau(大脑) corps(身体) cellule(细胞)<br>signal(信号) capacité(能力) humain(人类) étude(研究) |
| 主题 6 | 地球 行星 宇宙 理论 火星 物质 能量 星系<br>planète(行星) Terre(地球) univers(宇宙) lumière(光)<br>espace(空间) théorie(理论) énergie(能源) particule(粒子) |

综上所述，在跨语言文本主题发现过程中，通过利用集成学习的思想将多个 LDA 模型集成在一起，并结合跨语言文本聚类方法，能够有效地从文档集中发现双语主题信息。本文提出的跨语言主题发现方法思想容易理解，实现原理简单，在主题连贯性方面能获得比传统模型更好的表现，在主题多样性方面也保持竞争力。本文方法得到的双语主题语义相关度高，可以较好地表现主题信息。

**结束语**　面对全球范围内飞速增长的多语言文本信息以及当前跨语言主题发现方法研究较少的现状，本文提出了一种基于集成学习和跨语言词嵌入的跨语言文本主题发现方法 BCL-LDA。BCL-LDA 通过集成学习的思想，结合了跨语言词嵌入与跨语言主题发现方法，对单语主题模型 LDA 进行跨语言扩展，并结合跨语言词嵌入跨越语言的障碍，使用聚类算法完成主题分组。BCL-LDA 充分利用跨语言词嵌入优秀的跨语言文本表示能力，使得主题内不同语言主题词之间具有较高的语义相似度。在 TED2020 汉-德和汉-法平行语料上的实验表明，本文方法的 CNPMI 值显著高于经典跨语言主题发现方法 JointLDA，PLTM 和 PM-LDA，主题多样性也保持竞争性。

另一方面，由于 BCL-LDA 是基于单语 LDA 主题模型进行跨语言子主题建模，因此具有单语 LDA 主题模型的一些弊端。BCL-LDA 对于小数据集不够友好，且对于大规模语料集存在建模时间长的缺陷。若想进一步提升本文的跨语言主题发现方法的表现效果，需要保证 LDA 主题模型在基于 Bootstrap 自助采样得到的单语语料库上能够产生更加鲜明的子主题。后续研究将尝试在保证主题质量的前提下，尽可能降低主题建模的计算成本。此外，还需要研究如何更好地结合跨语言词嵌入与主题模型，在跨语言主题模型中充分利用其优异的跨语言文本表示能力，并选择效果更好的跨语言词嵌入模型进行跨语言表示。

## 参 考 文 献


[1] HARANDIZADEH B, PRINISKI H, MORSTA-TTER F. Keyword Assisted Embedded Topic Model[C]// Proceedings of the 15th ACM International Conference on Web Search and Data Mining. 2022: 372-380.

[2] WANG D, XU Y, LI M, et al. Knowledge-aware Bayesian deep topic model[C]// Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). 2022.

[3] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140.

[4] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(Mar): 1137-1155.

[5] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(Jan): 993-1022.

[6] MIMNO D, WALLACH H, NARADOWSKY J, et al. Polylingual topic models[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 880-889.

[7] YU Y Y, CHAO W H, HE Y Y, et al. Cross-language Knowledge Linking Based on Bilingual Topic Model and Bilingual Embedding[J]. Computer Science, 2019, 46(1): 238-244.

[8] ZOSA E, GRANROTH-WILDING M. Multilingual dynamic topic model[C]// Proceedings of the International Conference on Recent Advances in Natural Language Processing. 2019: 1388-1396.

[9] LEEK T, JIN H, SISTA S, et al. The BBN crosslingual topic detection and tracking system[C]// Working Notes of the Third Topic Detection and Tracking Workshop. 2000: 894-901.

[10] CHEN X S, LUO L, WANG H Z, et al. Analysis and Research on Cross Language Topic Discovery in Chinese and English[J]. Advanced En-gineering Sciences, 2017, 49(2): 100-106.

[11] JAGARLAMUDI J, DAUMÉ H. Extracting multilingual topics from unaligned comparable corpora[C]// European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2010: 444-456.

[12] ZHANG D, MEI Q, ZHAI C X. Cross-lingual latent topic extraction[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 1128-1137.

[13] BOYD-GRABER J, BLEI D. Multilingual topic models for unaligned text[C]// Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. 2009: 75-82.

[14] LIU X, YAN X, XU G, et al. Khmer-Chinese bilingual LDA topic model based on dictionary [J]. International Journal of Computing Science and Mathematics, 2019, 10(6): 557-565.

[15] YANG W Y, YU Z T, GAO S X, et al. Chinese-Vietnamese news topic discovery method based on crosslingual neural topic model[J]. Journal of Computer Applications, 2021, 41(10):

2879-2884.
[16] CHANG C H, HWANG S Y, XUI T H. Incorporating word embedding into cross-lingual topic modeling[C]// 2018 IEEE International Congress on Big Data(BigData Congress). IEEE, 2018: 17-24.
[17] CHANG C H, HWANG S Y. A word embedding-based approach to cross-lingual topic modeling[J]. Knowledge and Information Systems, 2021, 63(6): 1529-1555.
[18] CHAN C H, ZENG J, WESSLER H, et al. Reproducible extraction of cross-lingual topics(rectr)[J]. Communication Methods and Measures, 2020, 14(4): 285-305.
[19] BIANCHI F, TERRAGNI S, HOVY D, et al. Cross-lingual contextualized topic models with zero-shot learning[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 1676-1683.
[20] SRIVASTAVA A, SUTTON C. Autoencoding variational inference for topic models[C]// Proceedings of the 5th International Conference on Learning Representations(ICLR 2017). 2017.
[21] DAI H L, ZHONG G J, YOU Z M, et al. Public Opinion Sentiment Big Data Analysis Ensemble Method Based on Spark[J]. Computer Science, 2021, 48(9): 118-124.
[22] LIANG B T, NI Y F. Chinese Named Entity Recognition Based on Integrated Learning[J]. Journal of Nanjing Normal University(Natural Science Edition), 2022, 45(3): 123-131.
[23] FENG X, YANG Y T, DONG R, et al. Uygur-Chinese Neural Machine Translation Method Based on Back-translation and Ensemble Learning[J]. Journal of Lanzhou University of Technology, 2022, 48(5): 99-106.
[24] Big Data Search and Control Laboratory. NL-PIRICTCLAS Chinese Word Segmentation System [EB/OL]. [2023-03-16]. www. nlpir. org/.
[25] HELMUT S. TreeTagger [EB/OL]. [2023-02-09]. https://www. cis. lmu. de/~schmid/tools/TreeTagger/.
[26] JIAN Z W, YU J. German Text Clustering Based on Feature Word Pairing[J]. Information Research, 2022, 299(9): 86-93.
[27] EFRON B, TIBSHIRANI R J. An introduction to the bootstrap [M]. Boca Raton, Florida: CRC Press, 1994.
[28] YANG Y, CER D, AHMAD A, et al. Multilin-gual universal sentence encoder for semantic retrieval[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020: 87-94.
[29] GROOTENDORST M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure[J]. arXiv: 2203. 05794, 2022.
[30] SALTON G, BUCKLEY C. Term-Weighting Approaches in Automatic Text Retrieval[J]. Information Processing and Management, 1988, 24(5): 513-23.
[31] REIMERS N, GUREVYCH I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 4512-4525.
[32] WU X, LI C, ZHU Y, et al. Learning multil-ingual topics with neural variational inference[C]// Natural Language Processing and Chinese Computing: 9th CCF International Conference. 2020: 840-851.
[33] HAO S, PAUL M. Learning multilingual topics from incomparable corpora[C]// Proceedings of the 27th International Conference on Computational Linguistics. 2018: 2595-2609.
[34] HAO S, BOYD-GRABER J, PAUL M J. Lessons from the Bible on modern topics: Low-resource multilingual topic model evaluation[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1(Long Papers). 2018: 1090-1100.
[35] BOUMA G. Normalized(pointwise) mutual information in collocation extraction[C]// Proceedings of the Biennial GSCL Conference. 2009: 31-40.
[36] BISCHOF J, AIROLDI E M. Summarizing topical content with word frequency and exclusivity[C]// Proceedings of the 29th International Conference on Machine Learning(ICML-12). 2012: 201-208.

**LI Shuai**, born in 1997, postgraduate. His main research interests include data mining and decision support system.

**WU Shaocheng**, born in 1997, doctoral candidate. His main research interests include text mining and data science.