计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):290−297 [doi: 10.15888/j.cnki.csa.008563]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于数据增强的地质文本主题模型①

张竞元[1], 刘　刚[1,2], 曾　粤[1], 周大双[1], 陈麒玉[1,2]

[1](中国地质大学 (武汉) 计算机学院, 武汉 430074)
[2](智能地学信息处理湖北省重点实验室, 武汉 430074)
通信作者: 刘　刚, E-mail: liugang@cug.edu.cn

摘　要: 直接利用主题模型对地质文本进行聚类时会出现主题准确性低、主题关键词连续性差等问题, 本文采取了相关改进方法. 首先在分词阶段采用基于词频统计的重复词串提取算法, 保留地质专业名词以准确提取文本主题, 同时减少冗余词串数量节约内存花销, 提升保留词的提取效率. 另外, 使用基于 TF-IDF 和词向量的文本数据增强算法, 对原始分词语料进行处理以强化文本主题特征. 之后该算法与主题模型相结合在处理后的语料上提取语料主题. 由于模型的先验信息得到增强, 故性能得以提高. 实验结果表明本文算法与 LDA 模型相结合的方法表现较好, 在相关指标及输出结果上均优于其他方法.

关键词: 地质文本; 主题模型; 数据增强; 词向量; TF-IDF

引用格式: 张竞元,刘刚,曾粤,周大双,陈麒玉.基于数据增强的地质文本主题模型.计算机系统应用,2022,31(7):290–297. http://www.c-s-a.org.cn/1003-3254/8563.html

Geological Text Topic Model Based on Data Augmentation

ZHANG Jing-Yuan[1], LIU Gang[1,2], ZENG Yue[1], ZHOU Da-Shuang[1], CHEN Qi-Yu[1,2]

[1](School of Computer Science, China University of Geosciences (Wuhan), Wuhan 430074, China)
[2](Hubei Key Laboratory of Intelligent Geo-Information Processing, Wuhan 430074, China)

Abstract: Problems such as low topic accuracy and poor continuity of topic keywords occur when geological texts are directly clustered by topic models. This study adopts relevant improvement methods. In the word segmentation stage, the repeated word string extraction algorithm based on word frequency statistics is adopted. Geological terms are retained to accurately extract text topics, and redundant word strings are reduced to save memory costs. In this way, the efficiency of retained word extraction is improved. In addition, a text data augmentation algorithm based on term frequency-inverse document frequency (TF-IDF) and word vector is used to process the original word segmentation corpus and thereby strengthen the text topic features. Then, the algorithm is combined with the topic model to extract the corpus topics on the processed corpus. The performance of the model is improved due to its enhanced prior information. The experimental results show that the method combining the proposed algorithm with the latent Dirichlet allocation (LDA) model performs well, superior to other methods in all the related indexes and output results.

Key words: geological text; topic model; data augmentation; word vector; term frequency-inverse document frequency (TF-IDF)

地质科学文献、地质勘查报告以及野外记录等地质类文本数据数量急剧增加, 人们如果使用以往的方法从海量的结构化与非结构化文本数据[1] 中发掘、获取信息意味着巨大的时间、精力的投入, 导致工作效率的低下. 地质文本数据相较于其他领域的数据, 在复杂度与专业程度上丝毫不低. 文本聚类[2] 作为一种无

① 基金项目: 国家自然科学基金联合重点项目 (U1711267); 水利部协作项目 (2019306340); 中国地质大学 (武汉) 国家级创新训练计划 (201810491232)
收稿时间: 2021-10-07; 修改时间: 2021-11-08; 采用时间: 2021-11-12; csa 在线出版时间: 2022-05-31

监督的机器学习方法, 其优势在于能够对文本数据进行较为有效地组织、摘要和导航. 而这些也正是地质文本数据处理所需要的. 由于地质学科的庞大复杂, 所以产生的文本数据也是种类繁多, 而面向主题的思想能够使我们有针对性地组织、管理和获取数据, 从而得到所需的信息, 是一个能对文档集进行整体分析的视角和工具[3].

为了完成文本聚类的同时挖掘文本数据的主题, 文本主题模型是常用的方法. 不过相较于普通文本, 专业领域的文本对文本主题控制会有进一步要求, 这对文本模型提取主题信息提出了挑战. 基于文本主题模型自身的优势, 目前有应用于地质大数据表示技术[4]、地质文本分类[5]、 地质实体识别[6] 等. 而在地质文本主题提取方面, 樊中奎[7] 使用信息提取技术对已进行粗分类的地质资料的具体内容进行按主题获取, 可以提高资料的利用效率; 王永志等[8] 研发了融合加权与词频两种方法的组合关键词提取算法, 该算法具有较高的地学关键词命中率, 能够反映文本的主题信息; 邱芹军[9] 使用基于本体与增强词向量的方法 (OEWE) 获取文本关键词从而提取主题信息; 陈喜文[10] 提出了基于地质资料特征的主题模型 GIC-LDA, 该方法基于时空权重, 同时联合摘要、目录等元信息进行联合建模, 从而提升模型的主题推荐效果. 但目前仍存在以下问题: (1) 需要大量的人力搜集较为齐全的外部词典等先验知识, 另外为提高关键词或专业名词命中率保留了大量冗余词, 故在存储效率和词语筛选效率上仍有待改进. (2) 较少关注对地质文本主题信息的挖掘分析, 这在主题关键词的连续性上有直接表现, 需要增强文本的地质主题特征, 减少杂词的干扰, 提高主题可描述性.

本文将以多种大类主题的地质文本数据作为处理对象, 针对以上的问题, 在现有分词器的基础之上改进一般算法在保留地质专业词语的同时节省内存、时间花销以提升效率; 另外在对文本进行聚类时结合面向主题[11] 的思想, 利用主题模型提取准确度、连续性较高的主题描述词. 针对传统主题模型难以处理噪声词语和短文本的问题, 采用基于 TF-IDF 算法和词向量模型的数据增强算法, 增强文本的主题特征, 增强主题模型建模的先验信息, 提升模型效果. 最后展示文本集包含的主题信息及模型对比指标, 以此验证该方法的有效性及优势.

1 地质文本主题模型

为了实现基于数据增强的地质文本主题模型, 本方法包含以下步骤: (1) 将搜集到的地质文本逐一进行预处理. 预处理的过程分为两步: 第 1 步获取专业名词并将其作为保留词; 第 2 步利用获得的保留词进行二次分词. 经过预处理, 则得到经过分词处理的语料. (2) 计算词语的 TF-IDF 权重和词向量. (3) 利用 TF-IDF 权重和词向量模型使用数据增强算法处理分词语料. (4) 使用步骤 (3) 中的语料对整个语料库根据不同的主题个数建立模型, 通过主题关键词的描述选取主题个数合理的模型. 之后可以根据模型得到每篇文本的主题概率分布, 由主题概率分布确定每篇文本所属主题完成聚类. 具体处理流程如图 1 所示.

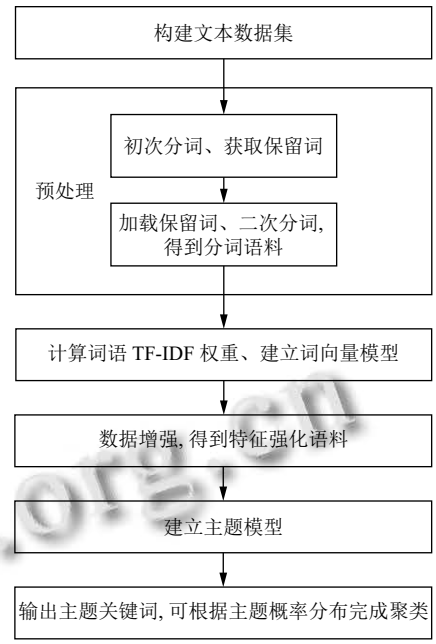

图 1 总体技术方案图

1.1 预处理

预处理阶段主要解决通用分词器对未登录词无法识别误将其切分的问题, 尽可能保留地质专业名词, 从而保证主题关键词的完整性, 增强主题可描述性. 如图 2, 该步骤细分为两个阶段: 在第一阶段, 使用通用分词器直接对地质语料进行分词, 得到首次切分语料, 使用的通用分词器为 jieba 分词器. 第二阶段, 在首次切分结果上使用重复词串提取算法, 该算法首先获取专业名词候选集, 之后添加约束条件对候选集进行过滤, 筛选出需要保留的地质专业名词, 以得到地质专业名词.

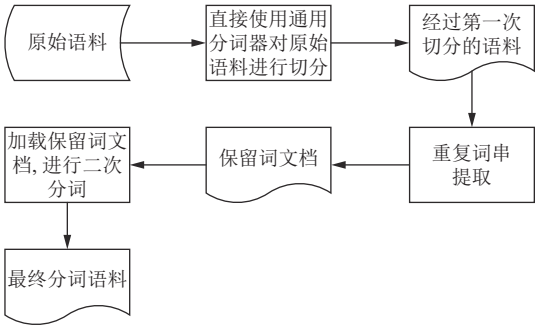

图 2 预处理流程图

为获取候选集本文采用了一种基于词频统计的重复词串提取算法, 通过统计各个切分部分在切分语料中出现的频率 (如表 1) 得到词频序列, 以词频序列中词频为 1 的词串为间隔, 对词频大于 1 的词串进行组合[12](如表 2). 此处使用的约束条件为词串组合频率和字符长度, 对于出现频率低于 2 的词串组合以及超过最大字符长度的词串组合直接过滤, 最终得到专业名词保留词文档. 由于《地质矿产术语分类代码》(GB 9649—1988) 中的地质专业术语长度大多数不超过 10, 所以字符长度阈值取该值. 该算法事先判断了高频词串的位置, 从而能直接对其进行组合, 避开了未重复部分, 大幅减少需要存储的垃圾或冗余词串, 提升处理效率. 虽然词频为 1 的词串也有可能组成专业名词, 但是由于其出现频率过低, 说明其与文本主题关联不大, 故可以忽略. 经过预处理的两个处理步骤后则得到保留词文档, 将该文档作为用户自定义词典, 再次使用通用分词器对原始语料进行分词处理, 得到最终分词语料. 由于加载了用户自定义词典, 通用分词器能识别一定的地质专业名词, 对其进行保留, 从而提升分词效果. 在两次分词过程中, 均有去停用词.

表 1 词频序列 (示例)

分类	示例
原始语料	火山岩/花岗闪长岩方辉橄榄岩矿化作用花岗闪长岩华北克拉通玄武岩矿化作用方辉橄榄岩华北克拉通矿化作用矿井
首次切分语料	火山岩/花岗/闪长岩/方辉/橄榄岩/矿化/作用/花岗/闪长岩/华北/克拉通/玄武岩/矿化/作用/方辉/橄榄岩/华北/克拉通/矿化/作用/矿井
词频序列	1/2/2/2/2/3/3/2/2/2/2/1/3/3/2/2/2/2/3/3/1

基于词频统计的重复词串提取算法伪代码如算法 1.

算法 1. 基于词频统计的重复词串提取算法

输入: 首次切分语料 M
输出: 保留词串 K

初始化词频序列 N, 其值均为 0, 其长度等于 M 中词串数量;
for i in 0 to *length*(M)
 if $N[i]$>0
  i+=1;
 else 统计 $M[i]$ 在 M 中的出现次数 n, $N[i]$=n.
初始化索引序列 L, 其元素为词频为 1 的词串在 N 中的索引;
for i in 0 to *length*(N)
 if $N[i]$==1
  将 i 保存入 L;
if *length*(L)== 0
 对 M 进行词串组合并保存入 K;
else
 for i in 0 to *length*(L)
  if i!=*length*(L)–1 and $L[i+1]$–$L[i]$!=1
   对 $M[L[i]+1]$~ $M[L[i+1]]$ 进行词串组合并保存入 K;
for i in 0 to *length*(K)
 if $K[i]$ 在 K 中出现次数<2 and *length*($K[i]$)<10
  删除 $K[i]$;
return K.

表 2 词串组合 (示例)

分类	示例
原始语料	花岗闪长岩致密油储层
首次切分语料	花岗/闪长岩/致密/油/储层
第1趟组合	花岗闪长岩/闪长岩致密/致密油/油储层
第2趟组合	花岗闪长岩致密/闪长岩致密油/致密油储层
…	…

1.2 TF-IDF 算法、词向量模型与 LDA 主题模型

在最终分词语料基础上应该确定词串对其所在文本的重要程度, 本文采用 TF-IDF (term frequency-inverse document frequency) 权重作为度量标准. TF-IDF 算法[13]的主要原理是: 如果某个词语在一篇文本中出现的频率 TF 很高, 并且在其他文章中很少出现, 则认为该词或者短语具有很好的类别区分能力, 适合用来对文章进行分类. 其中 TF (词频) 的计算较为简单, 即对于任意一个词语其在文本中出现的次数与文本词语总数之比. 而 IDF (逆向文件频率) 的意义是, 对于某个词语, 得到出现该词语的文档数量, 然后使全部文本文档数目除以该文档数, 再求自然对数. 常用的 TF-IDF 公式如下:

$$\omega_{i,d} = TF \cdot IDF = \frac{n_{i,d}}{|d|} \cdot \log \frac{|D|}{n_{i,D}} \tag{1}$$

其中, $n_{i,d}$表示词条t_i在文档d中出现的次数, $|d|$表示全部样本文档的总数, $n_{i,D}$表示D中包含词条t_i的文档数. 根据该公式的性质, 文本数据集中包含某一词语的文本越多, 它区分文档类别的能力越低, 其权重越小; 在某

一文本中, 某一词语的出现频率越高, 说明区分文本类别的能力越高, 其权重就越大.

词向量技术能够将文本中单个词语转化为一个对应的高维空间向量, 通过该向量多维的属性来表征该词语. 词向量模型的编码表示主要有独热 (one-hot) 表示和分布式表示两种方式, 其中独热方式虽然简单但是由于只有该词对应的词典索引位置为 1 外其余全为 0, 造成数据稀疏; 此外如果数据量大时还会造成维度灾难. 而分布式表示方法能够将词语转化为一个对应的稠密向量, 当词语表示为该种方式的向量时, 则可以通过计算向量的距离来计算词语间的相似性.

LDA 模型[14] 目标在于分析文本的主题分布, 识别主题, 主要是用于文本主题分类[15] 由文本-词语矩阵生成文本-主题矩阵 (分布) 和主题-词语矩阵 (分布). LDA 模型是一个包含了词语、主题、文本 3 层的贝叶斯概率模型, 以主题层作为核心层, 包含多个相互独立的主题, 每个主题是词语层上的词语多项式分布, 每篇文本由多个主题随机混合而成, 是多个主题上的多项式分布[16]. 建立 LDA 模型其生成文档的过程如下.

(1) 依照先验概率从语料集中选择一篇语料.

(2) 从超参数为 α 的狄利克雷分布中取样生成该篇文本语料的主题分布.

(3) 从主题的多项式分布中获取某一个主题.

(4) 从超参数为 β 的狄利克雷分布中取样生成该文本主题的词语分布.

(5) 从词语的多项式分布中获取词语.

其中, Dirichlet 的概率密度函数为:

$$f(x_1,x_2,\cdots,x_k;\alpha_1,\alpha_2,\cdots,\alpha_k)=\frac{1}{B(\alpha)}\prod_{i=1}^{k}x_i^{\alpha^i-1} \quad (2)$$

其中, $B(\alpha)=\dfrac{\prod_{i=1}^{k}\Gamma(\alpha^i)}{\Gamma\left(\sum_{i=1}^{k}\alpha^i\right)}, \sum_{i=1}^{k}x_i=1.$

多项分布概率密度函数为:

$$P(x_1,x_2,\cdots,x_k;p_1,p_2,\cdots,p_k)=\frac{n!}{\prod_{i=1}^{k}x_i!}\prod_{i=1}^{k}p_i^{x_i} \quad (3)$$

模型训练过程如下.

(1) 对每篇语料中的每一个词语赋予一个随机的编号.

(2) 再次扫描整个语料库, 使用 Gibbs sampling 方法对每个词语采样, 求出其归属的主题.

(3) 重复步骤 (2), 直至 Gibbs sampling 结果收敛.

(4) 统计整个语料库的主题-词语共现频率矩阵, 得到 LDA 主题模型.

利用主题模型可以得到数据集的主题概率分布, 通过主题概率分布可确定每篇文本的主题归属, 最终完成文本聚类. 所谓主题分布即每篇文本与每个主题相关的概率, 某一主题概率越高就越有可能归入该主题, 其分布形式如表 3 所示.

表 3 文本语料的主题分布 (示例)

语料编号	与主题1的相关概率	与主题2的相关概率	与主题3的相关概率	…
1	0.2612105	0.14390676	0.25281453	…
2	0.15169941	0.12245320 5	0.26336485	…
3	0.2504302	0.3192054	0.18799254	…

2 数据增强算法

一篇文本一般会围绕一个中心主题展开叙述, 为获取能描述文本主题的词语本文使用了 TF-IDF 算法寻找文本中权值较大的关键词, 这些词语往往与文本主题高度相关. 但是由于 TF-IDF 算法本身的局限性, 该算法无法体现词语间的关系, 为此采用词向量技术对每个词语生成对应的稠密向量, 从而能够计算词语间的相似性, 此处相似度选用余弦相似度. 首先, 通过 TF-IDF 算法得到每篇文本中一定数量的权值最大的关键词集合. 之后训练文本词向量模型, 利用关键词对应的词向量, 逐个计算每个关键词与其他关键词的相似度, 将相似性范围最广的关键词视为中心词. 此时可能会出现有多个中心词的情况, 那么则生成该篇文本的中心词集合. 之后, 计算各个中心词能覆盖到的关键词集合, 将所有中心词均无法覆盖的关键词剔除出关键词集合, 此时得到抽样集.

利用得到的抽样集开始对分词语料中的无关词语进行替换, 增强文本主题特征. 逐篇语料逐个词语的与对应的中心词集合中的中心词进行相似度计算判断是否需要进行抽样替换. 如果需要进行替换, 则需先判断当前文本词语对中心词的相似倾向程度, 从抽样集中抽取相应的中心词倾向程度的关键词对原词进行替换, 直至遍历结束. 在抽样阶段本文采用的是等概率抽样,

如果根据 TF-IDF 权重来分配抽样概率采取轮盘赌手法进行抽样, 算法效果则会严重依赖 TF-IDF 权重, 如果关键词采集出现偏差则会带来不利后果. 算法具体流程如图 3 所示.

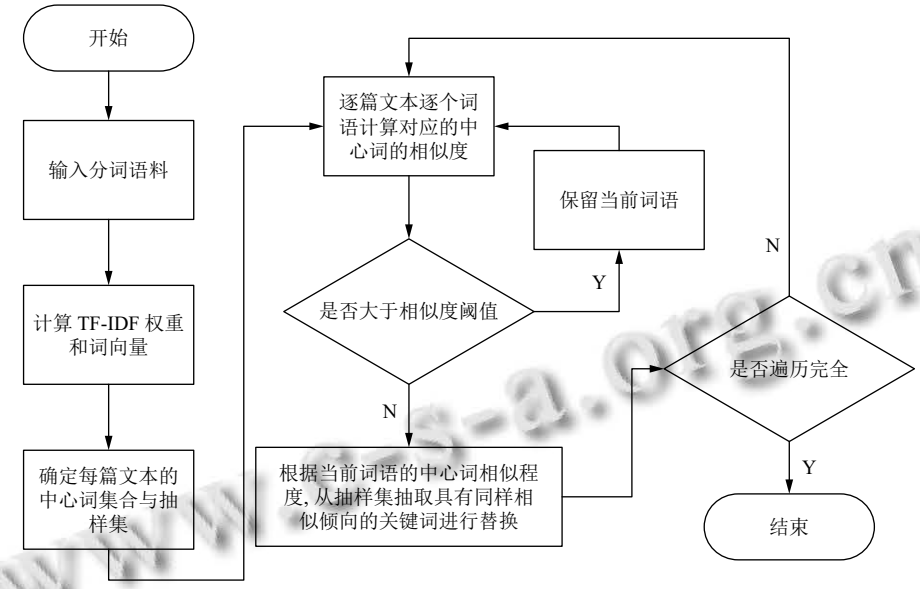

图 3 数据增强算法流程图

3 实验分析

3.1 实验数据集

本文从知网等文献资源网站收集整理 10 006 篇地质文献摘要, 共选取矿物、岩石、地质工程技术、地球化学、地质灾害、地质构造等十余种主题, 构建地质文本数据集. 每一篇文献摘要生成一个文本文档, 文献标题作为该文本文档的文件名. 之后将所有文本文档放入一个文件夹, 则原始语料库建立完成.

3.2 实验及结果讨论

文本数据集制作完成后首先需要对数据集进行预分词处理. 以语料《三维旋转水射流与水力压裂联作增透技术研究》为例, 该篇文本包含较多少见且冗长的专业术语. 表 4 为使用通用分词器处理后的语料与经本文重复词串提取算法处理后的语料对比结果, 通过准确率 (Precision)、召回率 (Recall) 和 F 值 (F-measure) 对结果进行评价. 由表可知, 通用分词器在地质专业名词充斥的情况下分词效果很差, 而本文算法一定程度保留了专业名词, 故分词效果显著. 本文在一般重复词串提取算法的基础上进行改进, 通过统计词频, 直接对高频部分进行组合, 忽略词频为 1 的切分部分, 减少词串组合时需要存储的候选词串数量, 节约存储空间; 另外, 由于候选集词串数量减少, 加快了词串的筛选过滤. 两者的对比结果见表 5, 经改进的算法得到的候选集词串数量仅为一般算法的 1.24%, 一般算法得到的词串候选集最终真正得到保留的只有原来的 0.31%, 经改进后达到原来的 22.45%.

表 4 分词评价结果 (%)

分类	Precision	Recall	F-measure
通用分词器处理语料	57.9	62.02	59.89
经本文算法处理后的语料	88.89	80.62	84.55

表 5 两种算法获取的词串数量对比

结果	一般重复词串提取算法	本文改进算法
候选词串数量	44 253	548
保留词串数量	136	123

经过分词处理, 接下来要对处理后的语料计算其 TF-IDF 权重和词向量. 其中词向量模型采用了 Word2vec (CBOW 模型) 和 Glove 两种常用方法. 之后, 利用 TF-IDF 权重和词向量文件进行数据增强处理得到特征强化语料, 在该语料上建立主题模型. 本文除 LDA 模型外还使用 BTM 模型进行了实验对比, 实验参数 α=50/K, β=0.01, Gibbs sampling 最大迭代数为 800, 其中 K 为主题个数. 数据增强算法中的相似度阈值根据具体数据集以及词向量模型来确定, 选择标准是既能保证排除无关词又能保证抽样集有相对充足的样本, 本文实验使用的相似度阈值范围是 0.11–0.13. 在实验过程中发

现, 经数据增强处理后, 两种模型训练的时间得到了减少, LDA 模型表现更为显著, 如图 4、图 5 所示.

为定量地衡量模型的优劣程度, 本文采用了主题间距离和模型困惑度 (perplexity) 两种常用指标对 LDA 模型进行评估. 其中主题距离采用 JS (Jensen-Shannon) 散度, 相较于 KL 散度, 它解决了计算结果非对称的问题. 其计算公式如下:

$$JS\left(P_1\|P_2\right)=\frac{1}{2}KL\left(P_1\|\frac{P_1+P_2}{2}\right)+\frac{1}{2}KL\left(P_2\|\frac{P_1+P_2}{2}\right) \tag{4}$$

其中, KL为 KL 散度. 模型困惑度是评价 LDA 模型最常用的方法之一, 其值越小表示模型的泛化性能越优. 其计算公式如下:

$$P=-\frac{1}{N}\sum_{d=1}^{N}\frac{1}{N_d}\log p\left(d\right) \tag{5}$$

其中, N为文本数量, N_d为文本d中包含的词语数量.

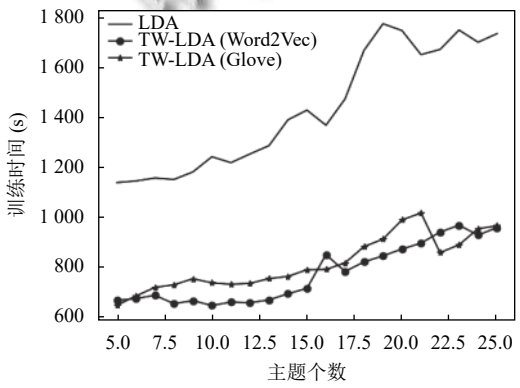

图 4 LDA 模型训练时间折线图

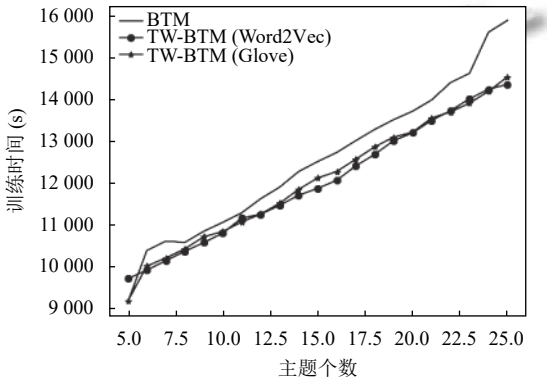

图 5 BTM 模型训练时间折线图

结果见图 6 和图 7. 可以看出, 经本文方法处理过后建立的 LDA 模型 (TW-LDA) 在主题距离和模型困惑度两项指标上均优于传统 LDA 模型, 即经数据强化后模型泛化性能和主题独立性均有提高, 体现了本文方法的优越性. 其中, 使用 Glove 方法训练得到的词向量最终得出的 LDA 模型的困惑度低于 Word2Vec 方法, 这是由于相较于 Word2Vec, Glove 引用了词共现矩阵, 同时考虑了词语的局部和整体信息; 而 Word2Vec 只关注窗口内的局部信息, 故生成的词向量准确率相对较低. BTM 模型的实验结果见图 8、图 9 所示. 因为 BTM 模型没有对文档的生成过程进行建模, 所以无法使用困惑度指标进行评估[17]. 故选择 *H-score* 对其进行评价. *H-score* 在文本聚类的结果上同时考虑类内和类间因素进行考量, 以评价文本主题模型, 其值越小则代表模型输出结果越优. 其计算方法如下:

$$H\text{-}score=\frac{Intra\_Dis}{Inter\_Dis} \tag{6}$$

其中, $Intra_Dis$为类内文本的平均距离, $Inter_Dis$为类间文本的平均距离.

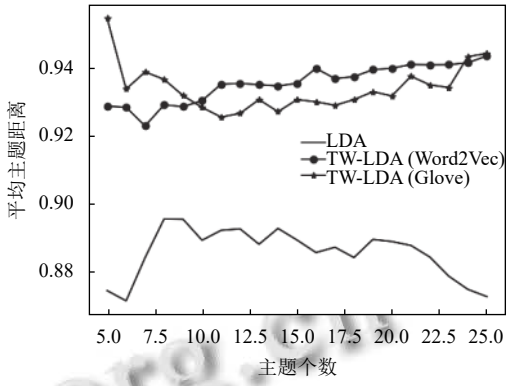

图 6 LDA 模型平均主题距离折线图

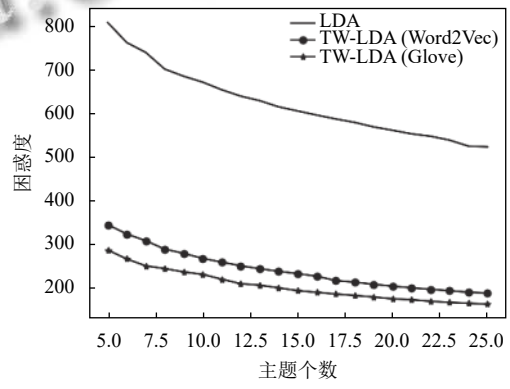

图 7 LDA 模型困惑度折线图

两者计算公式如下:

$$Intra\_Dis(C)=\frac{1}{K}\sum_{k=1}^{K}\left[\sum_{d_i,d_j\in C_k,i\neq j}\frac{2JS\left(d_i\|d_j\right)}{|C_k||C_k-1|}\right] \tag{7}$$

$$Inter\_Dis(C)=\frac{2}{K(K-1)}\sum_{C_k,C_{k'}\in C,k\neq k'}\left[\sum_{d_i\in C_k}\sum_{d_j\in C_{k'}}\frac{JS\left(d_i\|d_j\right)}{|C_k||C_{k'}|}\right] \tag{8}$$

其中, C为文本聚类的类簇集合, C_k为C中第k个类簇, K为主题个数, d_i为第i个文本.

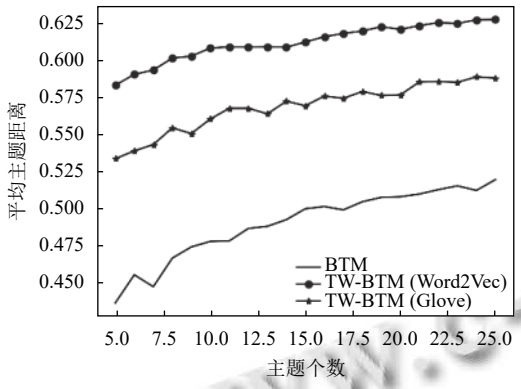

图 8 BTM 模型平均主题距离折线图

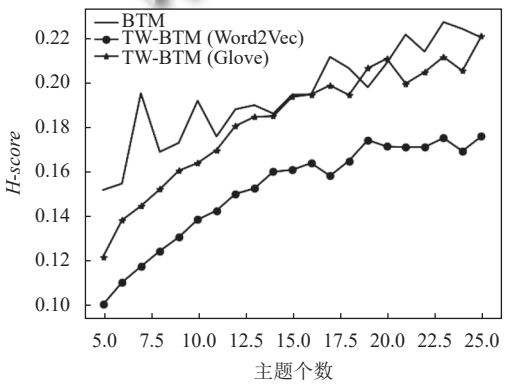

图 9 BTM 模型 *H-score* 折线图

由平均主题距离指标可知, BTM 模型与 Glove 结合的方法 (TW-BTM(Glove)) 效果并没有与 Word2Vec 结合 (TW-BTM(Word2Vec)) 的效果好, 该方法没有发挥出 Glove 全局词向量的优势. 同时, BTM 模型的主题独立性较差, 3 种模型中最高平均主题距离仍低于传统 LDA 模型. 另外, *H-score* 指标同样是 TW-BTM (Word2Vec) 有最好的表现. 综上, 诚然 BTM 模型能很好地缓解短文本建模稀疏的问题, 使用词对建模能够挖掘词语间一定的隐藏关系, 有助于提取文本主题, 但如此便削弱了词向量技术带来的提升, Glove 词向量因利用词共现矩阵采集词对用于训练故情况尤甚. 而 BTM 模型对规模较大的主题有较差细分能力的劣势便体现出来, 即主题间独立性差. 另外, 由于使用词对建模, BTM 模型的训练时间以及模型收敛迭代次数对比 LDA 模型没有优势.

下面对主题模型进行定性分析. 在传统 LDA、BTM 模型中, 主题关键词均出现了重复以及杂词, 这导致主题之间独立性较差, 这也表现了直接使用主题模型对大类主题难以进行适当的细分. 如表 6 所示, LDA 模型中主题 9 和主题 12 的最高主题关键词均是“储层”, 主题间产生了重叠, 并且主题 9 杂糅了储层和矿物浮选两个主题, 而主题 14 关键词连续性差难以对主题进行描述, BTM 模型也出现了上述类似情况. 经数据增强算法 (Word2Vec) 处理后的模型, 有了较大改观, 见表 7. 但 LDA 模型主题 1 的主题关键词的连续性很弱, BTM 模型主题关键词虽然连续性较高, 但仍有主题词重复的问题.

表 6 传统 LDA、BTM 模型主题关键词表 (局部)

LDA		BTM	
主题序号	主题关键词	主题序号	主题关键词
9	储层, 孔隙, 浮选, 捕收剂, 储层物性, …	5	储层, 勘探, 评价, …
12	储层, 层序, 发育, 沉积相, 裂缝, …	8	储层, 评价, 地层, 模型, …
14	隧道, 盾构, 砂卵石地层, 预报, 钻孔雷达, 喹啉酸, …	9	储层, 孔隙, 成岩作用, …

表 7 TW-LDA (Word2Vec)、TW-BTM (Word2Vec) 模型主题关键词表 (局部)

TW-LDA (Word2Vec)		TW-BTM (Word2Vec)	
主题序号	主题关键词	主题序号	主题关键词
1	构造煤, 卡林型金矿床, 火山岩储层, 裂缝, 钻孔灌注桩, …	2	储层, 发育, 孔隙, 成岩作用, …
	—	7	储层, 裂缝, 预测, 评价, …

采用 Glove 词向量技术后, BTM 模型每个主题相关概率最高的关键词已经没有重复出现, 但主题间仍有交叉. 由表 8 可知, 主题 2 与主题 7 虽然都与矿物有关, 但主题 2 为成矿主题, 主题 7 为找矿主题, 而主题 7 最高概率词语仍为“成矿”; 另外主题 15 应为隧道、岩溶作用以及溶洞类主题, 但与主题 5 产生重叠. 另外, 由于使用 Glove 发掘了隧道、岩溶主题, 但是由于 BTM 模型本身的特性并没有剥离与主题 5 的交叉部分; 然而使用 Word2Vec 的模型并没有该主题, 说明数据增强的效果相对 Glove 较明显, 故在前面指标评估

上逊于 TW-BTM (Word2Vec). LDA 模型的主题关键词的主题描述性最强, 内部没有杂词, 且主题之间没有重叠, 如表 9 所示. 通过对比各个主题个数的模型的主题描述词, 发现主题个数为 16 时, 主题关键词能够更好地对主题进行解释, 故选取 16 为最优主题个数. 由实验得数据集包含的主题有: 地质灾害、矿物浮选、油藏开采、岩土工程、城市地下空间、地质构造、花岗岩与岩浆、岩土力学、沉积矿物、土壤、储层、矿床及成矿、地质遗迹、化石、地质数据建模、火山岩及其储层.

表 8 TW-BTM (Glove) 模型主题关键词表 (局部)

主题编号	主题关键词
2	矿床, 成矿, 成矿流体, 流体, 矿石, …
5	岩石, 试验, 应力, 损伤, 破坏, …
7	成矿, 矿床, 异常, 找矿, 构造, 成矿预测, …
15	隧道, 工程, 试验, 性能, 土体, …

表 9 TW-LDA (Glove) 模型主题关键词表

主题编号	主题关键词	主题编号	主题关键词
0	地质灾害, 滑坡, 地质环境, 崩塌, 灾害	8	沉积物, 磁化率, 磁性矿物, 重矿物, 气候
1	浮选, 表面, 捕收剂, 回收率, 矿物	9	土壤, gt, cd, 重金属, 浓度
2	油田, 井筒, 注水, 裂缝, 钻井液, 油藏, 伤害, 剩余油	10	储层, 成岩作用, 孔隙, 发育, 砂岩
3	隧道, 岩溶, 探测, 活动断裂, 溶洞	11	矿床, 成矿, 成矿流体, 矿体, 矽卡岩
4	城市, 地下空间, 异常, 发展, 管理, 城市地下空间	12	保护, 地质公园, 旅游, 地质遗迹, 景观
5	构造, 断裂, 断层, 盆地, 构造变形	13	化石, 沉积, 烃源岩, 盆地, 层序, 生物
6	花岗岩, 岩体, ma, 锆石, 熔融	14	反演, 模型, 计算, 光谱, 数据
7	试验, 岩石, 应力, 钻孔, 损伤	15	火山岩, 火成岩, 岩相, 喷发, 火山岩储层

4 结论与展望

在本文所建立的地质文本聚类流程基础之上, 有以下总结:

(1) 本文采用基于统计词频序列的重复词串提取算法, 避开低频词语, 减少冗余词串的产生节省存储空间. 通过实验结果可以看出能够有效保留专业词语. 但是本文算法是运行在通用分词器的分词结果之上的, 如何提升和保障第一次分词的准确度, 是进一步需要研究的问题.

(2) TW-LDA 算法虽然使用了词向量技术提取语义信息, 但是对于地质专业名词效果仍欠佳, 在设置相似度阈值时难以确定, 而地质专业名词对于地质文本的主题又至关重要, 这也是需待解决的问题.

参考文献

1 廖建新. 大数据技术的应用现状与展望. 电信科学, 2015, 31(7): 1–12.

2 曹晓. 文本聚类研究综述. 情报探索, 2016, (1): 131–134. [doi: 10.3969/j.issn.1005-8095.2016.01.030]

3 黄钊炜. 面向主题的文本挖掘研究与应用 [硕士学位论文]. 武汉: 华中科技大学, 2018.

4 马凯. 地质大数据表示与关联关键技术研究 [博士学位论文]. 武汉: 中国地质大学, 2018.

5 杜晓敏, 潘晓. 基于 BERT 深度学习模型的地质资料目录自动分类研究. 中国矿业, 2021, 30(S2): 143–148.

6 张雪英, 叶鹏, 王曙, 等. 基于深度信念网络的地质实体识别方法. 岩石学报, 2018, 34(2): 343–351.

7 樊中奎. 地质资料全文聚类分析及信息提取的研究 [硕士学位论文]. 北京: 中国地质大学 (北京), 2014.

8 王永志, 金樑, 朱月琴, 等. 基于大数据技术的地学文档关键词提取算法研发. 地球物理学进展, 2018, 33(3): 1274–1281. [doi: 10.6038/pg2018CC0124]

9 邱芹军. 基于地质报告文本的时空及主题提取关键技术研究 [博士学位论文]. 武汉: 中国地质大学, 2020.

10 陈喜文. 地质资料管理关键技术研究及主题模型构建 [硕士学位论文]. 北京: 中国地质大学 (北京), 2016.

11 谢昊, 江红. 一种面向微博主题挖掘的改进 LDA 模型. 华东师范大学学报 (自然科学版), 2013, (6): 93–101.

12 王宏, 朱学立, 曾涛, 等. 一种基于统计的地质专业词语识别方法. 软件导刊, 2020, 19(4): 211–218.

13 林永民, 吕震宇, 赵爽, 等. 文本特征加权方法 TF·IDF 的分析与改进. 计算机工程与设计, 2008, 29(11): 2923–2925, 2929.

14 Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022.

15 李清. 基于机器学习的文本摘要技术的研究与实现 [硕士学位论文]. 成都: 电子科技大学, 2020.

16 陈晓美. 网络评论观点知识发现研究 [博士学位论文]. 长春: 吉林大学, 2014.

17 Yan XH, Guo JF, Lan YY, *et al*. A biterm topic model for short texts. Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro: ACM, 2013. 1445–1456.

(校对责编: 牛欣悦)