计算机科学
COMPUTER SCIENCE

基于深度学习的海洋热点新闻挖掘方法

覃娴萍, 丁昭旭, 仲国强, 王栋

引用本文

覃娴萍, 丁昭旭, 仲国强, 王栋. 基于深度学习的海洋热点新闻挖掘方法[J]. 计算机科学, 2024, 51(11A): 231200005-10.

QIN Xianping, DING Zhaoxu, ZHONG Guoqiang, WANG Dong. Deep Learning-based Method for Mining Ocean Hot Spot News [J]. Computer Science, 2024, 51(11A): 231200005-10.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231200005

基于深度学习的海洋热点新闻挖掘方法

覃娴萍[1]　丁昭旭[1]　仲国强[1]　王　栋[2]
1 中国海洋大学计算机科学与技术学院　山东 青岛 266404
2 中国海洋大学图书馆　山东 青岛 266404
(qinxianping_20@163.com)

摘　要　移动互联网的快速发展和现代移动客户端的普及推动了网络新闻行业、社交媒体和自媒体等的蓬勃发展,为用户提供了多元、丰富的海量信息。随着我国海洋强国战略的稳步推进,国民海洋意识的显著增强,有关海洋领域的多方面信息充斥着网络,相关媒体报道、公众舆论在网上大量涌现,热点事件频频发生。针对多来源、多属性的网络海洋信息,基于多源文本聚类和自动摘要技术,提出一种基于深度学习的海洋热点新闻自动挖掘系统,包括多源涉海数据自动采集、数据预处理、特征提取、文本聚类、自动摘要五大功能模块。具体而言,网络爬虫程序从多个数据源采集多样且分散的海洋数据,自动将数据结构化后存入数据库;根据文本特征的近似程度和文本间的关联关系实现聚类分析,聚类结果为后继摘要生成、主题发现提供数据支撑;基于预训练语言模型强大的上下文理解能力和丰富的语言表达能力,提出基于预训练语言模型的海洋新闻自动摘要生成方法。通过多组实验证明了所提方法在各个评估指标上的有效性,突显出其在多源异构网络海洋新闻挖掘方面的优势。该方法为处理分散的海洋资讯信息、生成可读性更强的内容摘要提供可行的解决方案,对提高海洋信息获取效率、监测公众舆论走向、推动海洋信息的应用与传播具有重要意义。

关键词:海洋新闻;文本聚类;自动摘要;深度学习;自然语言处理;预训练模型
中图分类号　TP391

Deep Learning-based Method for Mining Ocean Hot Spot News

QIN Xianping[1],DING Zhaoxu[1],ZHONG Guoqiang[1] and WANG Dong[2]
1 College of Computer Science and Technology,Ocean University of China,Qingdao,Shandong 266404,China
2 Library of Ocean University of China,Qingdao,Shandong 266404,China

Abstract　The rapid development of the mobile Internet and the popularity of modern mobile clients promote the vigorous development of the online news industry,social media and self-media,etc.,providing users with diverse and rich information. With the steady advancement of China's maritime power strategy and the significant enhancement of national maritime eawareness,the Internet is flooded with multifaceted information on the ocean field, with relevant media reports and public opinions proliferating online and hotspot events occurring frequently. Aiming at multi-source and multi-attribute network marine information,based on multi-source text clustering and automatic summarization technology,an automatic deep learning-based ocean hot news mining system is proposed,including five functional modules:automatic collection of multi-source ocean-related data,data preprocessing,feature extraction,text clustering,and automatic summarization. Specifically,the web crawler program collects diverse and scattered ocean data from multiple data sources,automatically structures the data and stores it in the database;clustering analysis is performed based on the similarity of text features and relationships between texts,which provides data support for subsequent summarization generation and topic discovery. Additionally,an automatic summary generation method for ocean news is proposed,leveraging the powerful contextual understanding and rich language expression abilities of the pre-trained language models. Multiple experiments demonstrate the effectiveness of the proposed method in each evaluation index,highlighting its superiority in mining news on multi-source heterogeneous networks. This method provides a feasible solution for processing scattered marine information and generating more readable content summaries,significantly contributing to the enhancement of marine information retrieval efficiency,monitoring public opinion trends,and promoting the application and dissemination of marine information.

Keywords　Ocean news,Text clustering,Automatic summarization,Deep learning,Natural language processing,Pre-trained model

基金项目:科技创新2030—"新一代人工智能"重大项目(2018AAA0100400);山东省自然科学基金(ZR2020MF131,ZR2021ZD19);青岛市科技计划项目(21-1-4-ny-19-nsh);中国海洋大学图书情报研究基金(202253006)
This work was supported by the Scientific and Technological Innovation 2030—Major Project for New Generation of AI(2018AAA0100400),Natural Science Foundation of Shandong Province,China(ZR2020MF131,ZR2021ZD19),Science and Technology Program of Qingdao(21-1-4-ny-19-nsh) and Fundamental Research Funds for the Central Universities(202253006).
通信作者:王栋(wangdong@ouc.edu.cn)

1 引言

21世纪是海洋的世纪，海洋在国家经济发展、国家安全领域、维护生态环境以及文化科教领域具有重要的战略意义和长远的发展价值。作为一个拥有丰富海洋资源和庞大海洋领域利益的国家，我国一直积极投入海洋事业建设，应对全球海洋治理问题、坚定维护国家利益，并推动全球海洋事务的可持续发展。近年来，移动互联网的快速发展和现代移动客户端的迅速普及带来了网络新闻行业、社交媒体和自媒体等的蓬勃发展，极大地拓宽了人们获取信息的渠道，为用户提供了形式多样、内容丰富的海量信息。有关海洋经济发展、海上安全形势、海洋生态环境、海洋科教文化等海洋领域的多方面信息充斥着网络，相关媒体报道、公众舆论在网上大量涌现，热点事件频频发生，海洋热点新闻正成为社会广泛关注的焦点。在此背景下，来源复杂、交叉重复、繁杂冗余的海洋信息构成了以文本为基础的极为庞大复杂的多属性异质信息网络，使得用户很难获取热点话题或兴趣话题，在海量冗余的信息中准确定位关键信息更是费时费力，因此对海洋新闻进行自动挖掘已成为亟待解决的挑战性问题。

热点新闻挖掘是信息检索、自然语言处理、机器学习和数据挖掘等领域交叉的研究方向，通过对大规模新闻文本数据的分析和处理，自动识别、归类、提取和总结其中的关键信息，帮助人们从海量的新闻数据中快速准确地发现和了解当前社会热点，为决策和行动提供重要参考。热点新闻挖掘的研究发展历程可以追溯到过去几十年，经历了从基于规则和关键词匹配的传统手工方法到利用机器学习、主题建模实现文本挖掘，再到社交媒体数据时期，应用深度学习技术捕捉文本复杂关系和上下文信息的演变。深度学习方法通过构建深层神经网络模型，利用强大的自动特征学习能力，能够从大规模数据中自动学习更抽象、更高级的特征表示和处理长文本间的复杂语义关系，已在热点新闻挖掘任务中取得显著的进展。

文本聚类分析[1-3]和自动摘要生成[4-7]是热点新闻挖掘任务的重要组成部分，具有关键地位并已成为其中的研究热点。文本聚类技术基于新闻文本的内容或主题相似性将其划分为不同的簇或类别，帮助发现热点话题和趋势，提供对新闻数据的结构化表示。有效的聚类分析是知识挖掘的重要前提条件。自动摘要技术通过自动化方法从长篇新闻文本中提取关键信息，生成简明扼要的概括，帮助用户快速浏览和获取关键信息，并提供一种有效的方式来摘要大量新闻报道。两者在热点新闻挖掘中协同作用，用户能够更好地理解和追踪新闻事件，提高了信息获取的效率和准确性。

面向多来源、多属性的网络海洋信息，本文聚焦于多源文本聚类和自动摘要技术，提出了一种基于深度学习的海洋热点新闻自动挖掘系统，包括多源涉海数据自动采集、数据预处理、特征提取、文本聚类、自动摘要五大功能模块。网络爬虫程序从多个数据源采集多种类型的海洋数据，经过文本预处理、特征提取和网络构建，根据文本特征的近似程度和文本间关联关系进行聚类分析，聚类结果为后继摘要生成、主题发现提供数据支撑。通过对多源文本聚类和摘要生成方法进行充分研究和实验验证，实现了多源异质网络海洋新闻的深层次挖掘，对提供可靠新闻报道、反映热点事件、监测舆论走向具有重要意义。本文的主要贡献如下：

1)开发了一套基于深度学习的海洋热点新闻挖掘系统，包括数据自动采集、数据预处理、特征提取、文本聚类、自动摘要五大功能模块。该方法能够实现多源网络涉海数据的自动采集，并自动将数据结构化后存入数据库，通过对预处理后的海洋数据进行文本聚类和摘要生成，实现了多源异质网络海洋新闻的深层次挖掘。

2)有效的聚类分析是实现热点发现和知识挖掘的重要基础。面对主题分散、类型不同、结构各异的海洋数据，通过TF-IDF方法提取特征实现文本向量化，并使用潜在语义分析进行降维处理，捕捉数据中的隐藏语义结构，最终应用改进的K均值聚类算法实现文本聚类分析。通过实验分析在6个评估指标上证明了该方法的有效性。

3)对聚类后不同簇中的海洋数据进行摘要生成有利于进一步简化信息理解、提供全局视角，摆脱传统繁杂的人工编撰。基于预训练语言模型强大的上下文理解能力和丰富的语言表达能力，提出了基于预训练语言模型的海洋新闻自动摘要生成方法，通过对mT5和DistilBART预训练模型进行微调，实现聚类关键词的提取，并应用百川智能最新推出的Baichuan2系列大模型对最终聚类结果进行摘要生成，采用人工专家直接对自动摘要结果进行评分。该方法提升了摘要生成的质量，同时易于迁移使用。

2 相关工作

2.1 文本聚类

文本聚类[8-10]作为一种重要的文本挖掘技术，旨在将相似的文本样本聚集在一起形成簇群，以实现对文本数据的有效组织和概括。通过文本聚类，可以发现文本数据中的隐藏模式、主题结构以及相关性，从而帮助人们理解和分析大规模文本数据集。在热点新闻挖掘领域，文本聚类被广泛应用于主题建模[11]、热点检测和新闻推荐等任务，因此有效的聚类分析是实现深度数据挖掘的重要基础条件。在文本聚类研究中首先要解决的是文本表示问题，网络文档自身包含多种文本属性，有效的特征提取算法是构建文本表示模型的关键。文本聚类主要流程如图1所示。

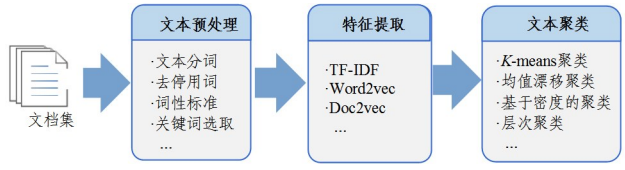

图1 文本聚类基本流程图

Fig. 1 Basic flow chart of text clustering

文本预处理是聚类分析的第一步，通过对原始数据进行去噪、分词、去除停用词、词干化或词形还原等操作，能够提高数据质量和一致性，减少噪声对聚类结果的干扰，同时将文本数据转化为适合特征提取的形式；特征提取将文本数据进一步转化为可用于聚类的特征向量表示，如典型的词袋模型、TF-IDF权重、Word Embeddings嵌入等，不同的特征提取方法可以捕捉文本数据中的不同信息，以便于后续的相似度计算和聚类算法的应用；聚类算法是实现聚类的核心部分，如K-means聚类[12-14]、基于图的聚类算法[15-16]和层次聚类[17-18]

等，根据不同的相似性度量将样本分配到合适的簇群中，形成有意义的聚类结果，其中选择合适的聚类算法和确定聚类数目是保证聚类结果质量的关键因素。*K*-means 聚类算法由于灵活、高效、易于实现，仍然是最流行和最直接的聚类算法，在许多聚类应用领域具有广泛的适用性。Ikotun 等[19]对 *K*-means 聚类算法及其变体进行了全面的回顾，提出了一种新的聚类分类方法。Huang 等[20]提出了一种新的鲁棒深度模型来分层执行 *K*-means，从而探索数据的分层语义，同一类的数据样本被有效地逐层收集，提供了清晰的聚类结构。Shrifan 等[21]提出了一种基于 Tukey 规则的 *K*-means 算法，使用修改的 Tukey 规则自适应地删除数据集中的异常值，并结合新的距离度量，将每个数据点分配到最近的聚类中，减轻在质心测量过程中异常值的影响，以获得更好的聚类精度。He 等[22]提出一种融合简单数据增强方法的深度聚类模型，采用 SBERT 对短文本进行嵌入表示，利用无监督 SimCSE 方法联合深度聚类 *K*-means 算法对文本嵌入模型进行微调，改善短文本的嵌入表示使其适于聚类。Li 等[23]揭示了特征矩阵的行和列本质上对应于实例和聚类表示，提出在实例级和聚类级进行双对比学习（Twin Contrast Learning，TCL）来实现在线聚类。近年来，随着深度学习的快速发展，基于深度学习的文本聚类方法也取得了显著进展[24-26]。研究者们研究各种神经网络架构、特征表示和聚类算法的组合，以获得更好的文本聚类性能。Cai 等[27]提出了一种新的同时学习特征表示和标签分配的联合优化聚类框架，通过在特征学习中引入收缩表示，在聚类层中利用焦点损失，该框架的两个组成部分可以相互促进学习聚类的表示。Dang 等[3]提出了一种新颖的深度聚类框架，称为基于两级最近邻匹配（NNM），分别从局部和全局将样本与其最近邻进行匹配，进一步提升聚类性能。Ronen 等[28]引入了一种能推断聚类数量的深度聚类方法，该方法不需要预先知道 *K* 的值，而是在学习过程中进行推断。通过分裂/合并框架、适应变化 *K* 的动态架构和新颖的损失，该方法优于现有的非参数方法。

2.2 自动摘要

自动摘要[29-32]作为一项重要的自然语言处理任务，旨在从大量文本中提取关键信息，通过压缩和概括原始文本，自动生成简洁、准确的文本摘要。它提供了对文本主题的准确理解，同时保留了文本内容的核心信息，帮助用户从大量文本中快速获取关键信息，从而节省时间和精力。在热点新闻挖掘中，自动摘要对于快速理解新闻内容、追踪热点问题、准确把握事件的要点和进展至关重要，从而实现对热点新闻的有效挖掘和跟踪。发展至今，自动摘要技术经历了日新月异的变化，其基本流程如图 2 所示。

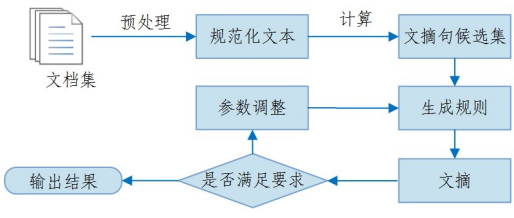

图 2　自动摘要基本流程图

Fig. 2　Basic flow chart of automatic summarization

预处理阶段对原始文本进行清洗和规范化，为后续的特征提取和分析建模提供干净的数据；文本表示将文本转化为计算机能够处理和比较的数值形式；特征提取是自动摘要的关键步骤之一，基于特征表示，分析词句的语义特征或综合考虑句子的词频、语义和句法特征，区分重要信息和次要信息。自动摘要根据文摘生成的方式分为概括式和抽取式两种。概括式主要依托自然语言理解技术，在理解语义的基础上补充一些原文中没有出现过的词组和句子，但该技术距实际应用还有一定距离。目前研究主要集中于抽取式，该方式通过摘录原文中的词和句子，按照定义的生成规则对每个词和句子进行评分和排序，考虑句子的显著性和新颖性两大指标，排序靠前的句子作为文摘句[33-34]。摘录词句的方案也从传统的基于规则和统计的方法发展到侧重于理解文本语法和语义结构的语言模型，再到通过构建句子间的关系图来识别重要句子的图模型方法，以及基于特征工程和标注数据的机器学习方法，再到如今使用编码器-解码器架构和预训练语言模型对数据进行表示学习和语义建模的端到端深度学习技术。此外，最近的研究中还将自动摘要生成视为序列决策问题，通过与环境的交互来优化生成的摘要，所生成摘要的内容日趋丰富全面。Mao 等[35]提出一种基于关键词异构图的生成式摘要模型，从原始文本中提取关键词，将其与句子共同作为输入构建异构图，进而学习关键词和句子之间的依赖关系。Srivastava 等[36]提出将聚类与主题建模的无监督摘要提取方法，减少了主题偏差。Khurana 等[37]提出了一种用于无监督提取文档摘要的信息理论方法，使用香农熵来捕获句子的信息量，使用非负矩阵分解（NMF）来揭示潜在空间中计算术语、主题和句子熵的概率分布，可用于实时的文档摘要。Jing 等[38]提出一种新颖的多重图卷积网络（Multi-GCN）来联合建模句子和单词之间不同类型的关系，并在此基础上提出一种用于提取文本摘要的多重图摘要（Multi-GraS）模型。Joshi 等[34]提出一种基于主题建模和词嵌入的新方法 DeepSumm，通过主题和语言向量编码的组合来捕获文档的结构和语义特征。基于概率主题分布和词嵌入两种不同的递归神经网络对句子进行编码，使用句子的局部和全局语义结构来计算句子的显著性，同时引入一种新颖性计算度量 SNS，以生成无冗余和多样化的文档摘要。Kouris 等[39]提出一个基于深度学习技术和语义数据转换的概括式文本摘要新框架，将深度编码器-解码器架构与基于语义的内容方法相结合，以产生泛化形式的概括摘要，并将其转换为人类可读的形式。

3　海洋热点新闻挖掘

本文聚焦于多源文本聚类和自动摘要技术，提出了一种基于深度学习的海洋热点新闻自动挖掘方法，包括数据自动采集、数据预处理、特征提取、文本聚类、自动摘要五大功能模块。该方法能够实现多源网络涉海数据的自动采集，自动将数据结构化后存入数据库；通过对预处理后的海洋数据进行文本聚类和摘要生成，可以从互联网上挖掘出分散各处的海洋资讯信息，并将同样内容的资讯合并优化，处理为可读性更强的内容摘要，从而减少数据冗余，获取丰富信息，实现对多源异质网络海洋新闻的深层次挖掘。本文提出的海洋热点新闻挖掘总体方案流程图如图 3 所示。下面对特征提取、文本聚类和自动摘要模块进行详细介绍。

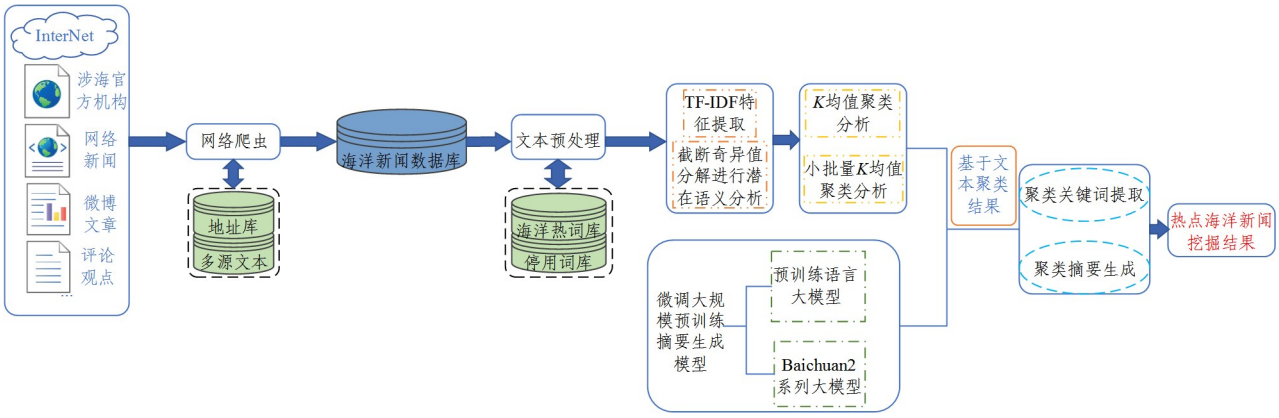

图 3 海洋热点新闻挖掘整体方法流程图

Fig. 3 Overall flow chart of ocean hot spot news mining

3.1 海洋新闻文本聚类

3.1.1 通过 TF-IDF 方法对文本进行特征提取

文本特征提取是海洋新闻聚类分析的关键步骤，它将文本数据转化为机器可理解和处理的数值形式，与聚类算法共同决定新闻文本最终的聚类效果。为了平衡效率和有效性，降低特征维度，加快文本处理速度，首先通过 TF-IDF（Term Frequency-Inverse Document Frequency）方法对预处理后的新闻文本进行特征提取。该方法量化词频和文档频率，根据词汇在当前新闻中的重要性对其进行加权，对在当前新闻文本中出现次数多且在其他新闻文本中出现次数少的单词赋予较高权重，从而提取出具有区分度和代表性的关键词特征，允许度量文本之间的相似性。

其中词频（Term Frequency，TF）用于衡量词语在当前文档中的重要性，即词语t_i在当前新闻d_j中出现的频率。

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

逆文档频率（Inverse Document Frequency，IDF）用于衡量词语在整个数据集中的普遍重要性。计算方法为总新闻数目除以包含该词语的新闻文本数目，再取一个底为 10 的对数。

$$idf_i=\log_{10}\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{2}$$

二者的乘积 $tf_{i,j}\times idf_i$ 作为词汇在文档中的 TF-IDF 值，用于衡量词语t_i对于新闻d_j的重要性。TF-IDF 值越大，表明该词汇在当前文本中频繁出现，但在整个语料库中较少出现，因此具备更好的代表性和区分度。通过 TF-IDF 特征提取，可得到用于表示新闻文本的特征向量。具体来说，一篇海洋新闻文本表示为一个特征向量，向量的维度等于该新闻文本中对应的词汇数，每个维度上的值即为该词在文本中的 TF-IDF 值，表示该词在文本中的相对重要性。在海洋新闻文本聚类中，得到的 TF-IDF 特征向量有助于识别关键词、发现文本间的相似性和差异性，从而比较和聚合不同新闻文章，以识别相关主题和趋势。

3.1.2 潜在语义分析

经过 TF-IDF 方法得到的特征向量可能仍然非常高维，包含大量词汇。而在新闻数据中，往往存在语义上相关的词汇，但这些相关性可能难以通过传统的 TF-IDF 方法来捕捉。为了更好地理解文本中的潜在语义结构并进一步降低数据维度，在 TF-IDF 特征提取的基础上，采用截断奇异值分解（Truncated Singular Value Decomposition，LSA）来进行潜在语义分析。该方法通过将文本-词汇矩阵进行 SVD 分解，降低特征维度，提取文本中的潜在语义信息。通过选择合适的截断维度，可以构建文本的语义空间表示，用于揭示文本数据之间的潜在语义关联、主题结构和文本相似性。使用 3.1.1 小节中得到的 TF-IDF 特征向量构建文本-词汇矩阵 $\boldsymbol{A}$（Document-Term Matrix），每行代表一条新闻文本，每列代表一个词汇，矩阵中的元素表示对应词在对应文本中的 TF-IDF 值。通过对矩阵 $\boldsymbol{A}$ 进行奇异值分解（Singular Value Decomposition，SVD），将其分解为 3 个矩阵的乘积，如式(3)所示：

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{3}$$

其中，矩阵 $\boldsymbol{U}$ 代表文档-主题矩阵，每行对应一个新闻文本，每列对应一个潜在语义主题，每个元素代表该文本在对应语义维度上的重要性或权重；矩阵 $\boldsymbol{V}^{\mathrm{T}}$ 代表词汇-主题矩阵，每行对应一个词汇，每列对应一个潜在语义主题，每个元素代表该词在对应语义维度上的重要性或权重；$\boldsymbol{\Sigma}$ 矩阵是一个对角矩阵，其对角线上的元素是奇异值，奇异值按降序排列，代表文档-词汇矩阵在对应语义维度上的重要性。

在 SVD 中，奇异值代表文本在语义空间上的重要性，奇异向量代表文本在语义空间上的表示。为了降低维度并保留最重要的语义信息，在奇异值分解的过程中只保留前 k 个最大的奇异值和对应的奇异向量部分，若矩阵 $\boldsymbol{A}$ 的秩为 r，则 $0<k<r$。具体形式如式(4)所示：

$$\boldsymbol{A}\approx\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}} \tag{4}$$

得到降维后的文档-主题矩阵$\boldsymbol{U}_k$和词汇-主题矩阵$\boldsymbol{V}_k^{\mathrm{T}}$，分别由 $\boldsymbol{U}$ 的前 k 列、$\boldsymbol{V}$ 的前 k 列和 $\boldsymbol{\Sigma}$ 的前 k 列组成，$\boldsymbol{\Sigma}_k$是 k 阶对角矩阵。

通过上述潜在语义分析，可以将海洋新闻数据从高维的 TF-IDF 特征向量转化为更富有语义信息的低维表示，能够揭示新闻数据的潜在结构和语义关系，从而提高了文本聚类的效果和解释性。

3.1.3 K 均值（K-Means）和小批量 K 均值（Mini-batch K-Means）聚类算法

综上，可以完成海洋新闻数据的特征提取和潜在语义分析，将文本数据转化为具有语义信息的低维表示。在此基础上，应用改进的 K 均值和小批量 K 均值聚类算法，实现对

海洋新闻的聚类。K 均值聚类由于灵活高效且可扩展,因此一直是最流行和最直接的无监督学习算法[19],它根据内部相似性对数据进行重新组织,将数据集划分为 k 个不同的簇,基于距离度量,通过迭代优化来找到最优的簇中心点,使得每个样本点与其所属簇的中心点之间的距离最小化。具体来说,算法首先初始化 k 个聚类中心 $a=a_1,a_2,\cdots,a_k$,并根据样本与聚类中心的距离将每个样本 x_i 分到距离最小的聚类中心对应的类,之后对每个类别 a_j 通过式(5)计算其聚类中心,迭代上述步骤,直到满足收敛条件。

$$a_j=\frac{1}{|\mathscr{C}_i|}\sum_{x\in\mathscr{C}_i}x \tag{5}$$

小批量 K 均值聚类是 K 均值算法的改进版本,旨在提高算法的效率和可扩展性。每次迭代时采用随机抽样的方式,只处理一小批量的数据点而不是整个数据集,缩小了每次传入网络的参数规模,加快算法的收敛速度。通过构建聚类模型进行迭代训练,完成对多源海洋新闻的聚类研究,将同样内容的海洋新闻资讯合并优化,识别具有相似主题或特性的海洋新闻文章。

3.2 海洋新闻自动摘要

近年来,预训练语言模型(Pretrained Language Models,PLM)在自然语言处理领域取得了巨大进展。其基本思想是先在大规模无监督语料库上预训练模型,然后通过微调来适应特定的下游监督任务。这种预训练-微调框架已成为自然语言处理领域的主流技术范式。PLM 能够从大规模预训练语料库中学习丰富的语言知识,并将其编码到模型参数中,通过特殊设计的目标来学习语言的通用表示和上下文信息,这使得 PLM 在各种下游任务中都表现出色。这种模型复用不仅提升了性能,还提高了研究效率,降低了复杂性。PLM 在众多 NLP 任务中的成功促使我们将其应用于海洋新闻自动摘要生成任务中。

在大规模语料库上预训练的 PLM 能够准确理解自然语言,并以人类语言流畅地表达,这是完成自动摘要任务的关键能力。在实现海洋新闻文本聚类后,对聚类的结果进行进一步分析,基于预训练语言模型实现关键词提取和自动摘要。

3.2.1 mT5 预训练模型

mT5(Multilingual T5)是基于 Transformer 架构的多语言预训练模型,是 Google Research 团队针对多语言数据集提出的 T5(Text-to-Text Transfer Transformer)模型变种,在 C4 数据集的多语言变种 mC4 上进行预训练。T5 的核心思想是将所有自然语言处理任务转化为文本到文本的任务,而 mT5 扩展了 T5,使其能够支持多种语言,其优势在于它的多语言能力和通用性。mT5 通过掩码语言建模(Masked Language Modeling)和文本翻译任务进行预训练。在掩码语言建模中,模型需要根据上下文预测被掩码的单词或短语。而在文本翻译任务中,模型需要将一种语言的文本翻译成另一种语言。mC4 数据集包括通过 Common Crawl 数据集爬取得到的 101 种语言的自然文本,总词语量达到了 250 000。为了使不同语言数据的训练程度较为均衡,mT5 对每个语言 L 使用了 $p(L)\propto|L|^{\alpha}$ 的采样率,其中 $|L|$ 是语言 L 的数据量,α 是控制采样次数的超参数。通过在大规模语料库上进行预训练,mT5 学习到丰富的语言知识和上下文表示能力,从而为下游任务提供强大的语言理解和生成能力。

3.2.2 DistilBART 预训练语言模型

BART 模型是 Facebook 开发的用于序列到序列生成的预训练语言模型,融合了BERT 模型的双向编码器和 GPT 模型的自回归解码器结构。它通过自编码器学习输入文本的表示,并通过自回归模型生成目标文本。其预训练任务包括掩码语言建模和文本生成,在摘要生成和机器翻译任务上表现优异。DistilBART 是通过“缩小和微调”方法对 BART 模型进行知识蒸馏得到的,保留了 BART 的核心思想,同时大幅减小了模型规模,使其在资源受限的环境中依然能够提供高质量的预测结果。

3.2.3 Baichuan2 系列生成式语言大模型

Baichuan2 系列模型是由百川智能推出的预训练语言大模型,于 2023 年 11 月正式发布。在 2.6 万亿高质量多语言数据训练的基础之上,通过融合意图理解、信息检索以及强化学习技术,结合有监督微调与人类意图对齐,设计多模态知识注入方法以及多模态多阶段的训练策略,使得 Baichuan2 系列模型不仅具有流畅的多轮对话能力以及良好的生成与创作能力,而且在数学、代码、逻辑推理、语义理解等多任务上实现了显著提升。凭借其强大的语言理解和高效的多样化文本生成能力,根据实际任务自适应调整,能够在无人工干预的情况下,自动完成新闻摘要任务,提高获取信息的效率和准确性。

3.2.4 海洋新闻自动摘要生成模型

在对分散且多样的海洋新闻进行有效聚类分析后,为了对聚类后的新闻内容进行更直观的认识,快速定位并提取关键信息,在此基础上通过特定数据集对预训练语言模型进行微调,来实现新闻文本主题关键词的提取和摘要的自动生成。通过微调后的 mT5 和 DistilBART 模型来提取聚类中新闻文本的关键词。其中 mT5 选用了两个版本的模型,分别在 XL-Sum 数据集和中文简体 CrossSum 数据集上进行了预训练;DistilBART 也选用了两个版本的模型,分别在 CNN Daily-Mail 数据集和 XSum 数据集上进行了预训练。通过对话调优和对齐后的 Baichuan2-13B-chat 模型来对聚类后的海洋新闻进行摘要生成,在无人工干预的情况下,自动完成新闻摘要的任务。

微调使用的数据集为 NLPCC 2017 单文档摘要数据集,包括 5 万条中文“文章-摘要”数据对。通过在微调数据集上对模型进行微调训练,使用微调后的模型对每个簇群的代表性海洋新闻报道进行关键词抽取和自动摘要可生成。提取出的关键词可以作为新闻聚类的类别标签,方便后期对新闻进行整理;生成的摘要可帮助用户快速获取新闻关键信息,节省时间精力,实现对复杂多源海洋新闻的知识挖掘。

4 实验结果及分析

4.1 海洋新闻数据集

海洋新闻数据集的制作是进行该项研究工作的基础和前提。针对数十家海洋领域政府机构网站、微博、百度新闻等开发了专门的网络爬虫程序,进行自动数据采集,通过关键词分析记录了多种关系信息,如转载、评论等,并提取出多种属性信息,如网站、点击量、发布时间等。最终本文实验采用的数据集是在各大新闻网站上爬取到的海洋新闻,总数为 65 640

条，涵盖了海洋科学、海洋保护、海洋工程等最新的资讯信息。数据集的制作过程主要分为多源网络文本采集和数据预处理操作。

4.1.1 多源网络文本采集

通过开发的专门网络爬虫程序进行海洋新闻数据的自动采集，根据保存在地址库中的新闻网站、社交网络等来源地址信息，进行自动地周期性网络文本采集。这些网络地址来源多样，包括新闻网站、社交网络、微博、微信和百度资讯等，保证了数据的全面性和多样化。在采集过程中，对网页内容中的垃圾信息、广告等“脏数据”进行了清理，只保留必要的标题、正文、作者、来源网站、发表时间等信息。程序自动将数据结构化后存入数据库中，每周自动收集数据近千条，目前已经收集涉海新闻、评论、博文等数万条。

4.1.2 数据预处理

由于新闻源很多，采集到的新闻质量参差不齐，会出现一些包含海洋的关键词，但与海洋并不相关的新闻，如“海洋馆”或包含海洋两个字的人名等。同时由于很多新闻转载的情况，数据中会出现大量重复和冗余。因此，在采集的基础上进行了一系列的预处理操作来清洗和规范文本数据，以确保后续分析和建模工作的准确性和可靠性。首先去除文本中的噪声、HTML 标签、特殊字符等不必要的内容，以净化数据，并对重复的新闻内容进行去除；同时通过分词工具对文本进行分割，引入停用词库和热词库对分词结果进行优化，加强对实时事件、专业术语和重点对象的抽取能力。通过文本预处理，为后续的特征提取和实验分析提供了高质量的数据。

4.2 海洋新闻聚类

4.2.1 参数设置与评判标准

对于海洋新闻聚类，分别使用改进的 K-Means 聚类模型和 Mini-Batch K-Means 聚类模型。算法对应的具体参数设置如表 1 所列。

表 1 聚类模型的具体参数设置

Table 1 Specific parameter settings of clustering model

参数名	Mini-Batch K-means	K-Means
n_init	5	5
max_iter	300	300
batch_size	500	—
init_size	1 500	—

其中，*n_init* 代表初始化质心运行算法的次数；*max_iter* 代表迭代运算的最大次数；*batch_size* 是 *Mini-Batch K-Means* 算法的采样批次大小；*init_size* 是候选质心的样本数据集大小。

为了证明聚类算法对海洋新闻文本聚类的有效性，同时考虑了以下 6 个指标：

1）同质性（Homogeneity）：每个簇类包含单个类的成员的比例。值越接近 1，表示每个簇内的样本类别越一致；值越小，表示每个簇内的样本类别越不一致。

2）完整性（Completeness）：给定类的所有成员都分配给同一个簇类的比例。值越接近 1，表示每个簇内的样本类别越完整；值越小，表示每个簇内的样本类别越不完整。

3）V-Measure：同质性和完整性的调和平均数，同时考虑了簇内的一致性和簇间的分离程度。值越接近 1，表示聚类效果越好。

4）轮廓系数（Silhouette Coefficient）：衡量一个样本在所属簇中的紧密程度和与其他簇的分离程度，适用于实际类别信息未知的情况。一个高轮廓系数分数表示聚类结果具有较好的紧密性和分离性。对于单个样本，设 a 是它与聚类中其他样本的平均距离，b 是与它距离最近不同样本的平均距离。轮廓系数定义如式（6）所示：

$$S=\frac{b-a}{\max(a,b)} \tag{6}$$

5）Calinski－Harabaz 指数：对簇间距离和簇内距离综合考虑的指标，度量了聚类结果的紧密性和分离性之间的平衡。一个高 Calinski-Harabaz 指数表示聚类结果具有较好的紧密性和分离性。其具体计算如式（7）所示：

$$CH=\frac{\left[\frac{\sum_{k=1}^{K} n_k \| c_k-c \|^2}{K-1}\right]}{\left[\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k} \| d_i-c_k \|^2}{N-K}\right]} \tag{7}$$

6）Davies-Bouldin 分数：衡量簇间的相似度和簇内的紧凑度。一个低的 Davies-Bouldin 分数表示聚类结果簇间距离越大，簇内距离越小，聚类效果越好。该分数的计算首先需要得出 S_i 和 $M_{i,j}$，分别表示类内数据到质心的聚类和聚类之间的距离。其中 X_j 表示簇类 i 中第 j 个数据点，A_i 是簇类 i 的质心，T_i 是簇类 i 中数据的个数，具体计算如式（8）和式（9）所示：

$$S_i=\left(\frac{1}{T_i}\sum_{j=1}^{T_i}|X_j-A_i|\right)^{\frac{1}{p}} \tag{8}$$

$$M_{i,j}=\| A_i-A_j \| \tag{9}$$

最终 Davies-Bouldin 分数由以上两个数值计算而得，如式（10）所示：

$$DB=\frac{1}{n}\sum_{i=1}^{N}\max_{j\neq i}\frac{S_i+S_j}{M_{i,j}} \tag{10}$$

需要指出的是，前 3 个指标更关注聚类结果的一致性和数据分布的匹配程度；后 3 个指标则更关注聚类结果的紧密性、分离度和簇内数据的一致性。因此，根据实际情况，需要综合考虑通过交叉验证并进行权衡。

4.2.2 肘部方法确定聚类数

在海洋新闻聚类过程中，首先通过肘部方法（Elbow Method）来确定最合适的聚类个数。由于 K-Means 的成本函数是各个类的畸变程度之和，而平均畸变程度会随着 K 的增大先减小后变大，因此可以通过可视化观察畸变程度与聚类数 K 之间的关系来确定最合适的聚类个数，而不是凭直觉或采用试错的方式进行选择。肘部方法提供了一种相对客观的方式来确定聚类个数，帮助提高聚类算法的效果和可解释性。通过实验绘制了畸变程度与聚类数量 K 的关系图，并找到图像中的“肘部”点，结果如图 4 所示。

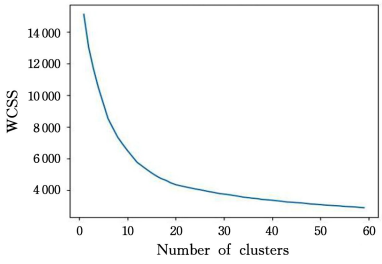

图 4 肘部方法

Fig. 4 Elbow method

在此基础上，进一步进行验证，以更准确地确定聚类个数 K，在不同 K 值下对新闻数据的聚类结果如表 2 所列。

表 2　聚类数量分析

Table 2　Cluster number analysis

K 值	轮廓系数	CH 指数	DB 系数
11	0.259	3163.779	1.308
12	0.267	3339.493	1.264
13	0.264	3237.465	1.252
14	**0.278**	**3348.787**	**1.249**
15	0.263	3173.027	1.267
16	0.270	3077.920	1.290

根据图 4 和实验分析，畸变程度和在聚类数目为 14 时出现一个形似手肘的拐点，该点表示了一种平衡：增加更多的簇类不再显著降低畸变程度，而只会增加模型的复杂性；该点在保持足够紧密度的同时避免了过度分割数据，因此将 14 作为较为理想的聚类数目。

4.2.3　聚类结果分析

在使用截断奇异值分解进行潜在语义分析阶段，通过保留最重要的奇异值和对应的奇异向量来减小数据的维度，因此选择适当的截断奇异值个数至关重要，需要在数据压缩、计算复杂度、模型性能和可解释性之间找到一个平衡点。在此过程中，对截断奇异值个数的选择进行研究，通过改进的 K-Means 方法进行聚类，得到的聚类结果在 3 个指标上的表现如表 3 所列。

表 3　截断奇异值数量分析

Table 3　Truncated singular value number analysis

截断奇异值个数	同质性	完整性	V-Measure
8	0.304	1.000	0.467
12	**0.310**	**1.000**	**0.473**
16	0.307	1.000	0.471
32	0.303	1.000	0.465
64	0.302	1.000	0.464

由表 3 可知，截断奇异值个数的选择会对最终聚类效果产生直接影响，选择最佳的截断奇异值个数能获得最优的潜在语义分析结果。因此在后续实验中，将截断奇异值个数默认设置为 12。

最终，使用改进的 K-Means 聚类模型和 Mini-Batch K-Means 聚类模型实现对海洋新闻数据的聚类分析。表 4 列出了在前面介绍的多项指标上算法的具体表现，实验中的参数设置前面已详细给出。

表 4　聚类算法性能比较

Table 4　Performance comparison of clustering algorithms

评估指标	Mini-Batch Kmeans	K-Means
同质性	0.308	0.310
完整性	1.000	1.000
V-Measure	0.471	0.473
轮廓系数	0.272	0.278
Calinski-Harabaz 指数	3234.926	3348.787
Davies-Bouldin 分数	1.254	1.249
耗时/s	14.770	22.277

从表 4 可以看出，两种聚类算法在各项指标上都能实现较好的聚类效果，能够根据提取的文本特征对海洋新闻进行有效的聚类。同时 Mini-Batch K-Means 聚类方法耗时较短，但在评判指标上逊色于 K-Means 算法。因此在实际应用中，需要根据数据集规模、资源成本、聚类性能确定最适合的算法以实现权衡。在实验中采用改进的 K-Means 算法作为聚类模型。

通过对聚类结果的可视化展示，海洋新闻最终聚类结果的图像化如图 5 所示。

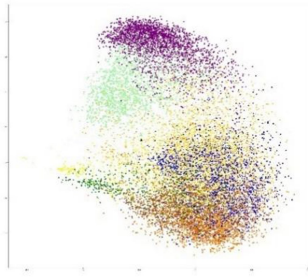

图 5　聚类结果可视化

Fig. 5　Visualization of clustering results

4.3　海洋新闻自动摘要

4.3.1　参数设置与评判标准

针对预训练语言模型的微调，分别对 4 种模型在 NLPCC 2017 数据集上进行了微调训练。微调的具体超参数设置如表 5 所列。

表 5　微调参数设置

Table 5　Fine-tuning parameter settings

参数名称	参数值设置
Num_epochs	20
Batch_size	16
Max_input_length	512
Max_target_length	128

表 5 中，*num_epochs* 代表学习轮数；*batch_size* 代表每个设备上的样本批量大小；*max_input_length* 代表输入文本的最大长度；*max_target_length* 代表输出文本的最大长度。

实验使用测试指标 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)来评估自动摘要生成的质量，包括 ROUGE-1，ROUGE-2 和 ROUGE-L 3 个主要子指标，用于衡量生成摘要与参考摘要之间的相似程度。

ROUGE-1 衡量生成摘要与参考摘要之间重叠单个词(Unigram)的数量，计算生成摘要和参考摘要间共享的一元词(每个词视为一个单一的元素)的比例；ROUGE-2 是一种二元语法级别的重叠匹配度量，考虑生成摘要与参考摘要之间重叠连续两个词(Bigram)的数量，相比 ROUGE-1，它考虑了相邻词语的匹配情况，提供了更多上下文信息；ROUGE-L 通过计算生成摘要和参考摘要间的最长公共子序列(Longest Common Subsequence，LCS)的长度来评估它们之间的相似性，它不仅考虑单词顺序，并且允许生成摘要中的单词与参考摘要中的单词不完全匹配，因此对于长文本的评估更合适。

4.3.2　微调模型结果

按照统一的参数设置分别对以上 4 种模型进行微调训练，并在 NLPCC 2017 测试集上对微调保存后的模型进行测试。不同模型在以上 3 个指标上的准确率、召回率和 F1-Score 分别如表 6－表 8 所列。

表 6 微调模型 ROUGE-1 性能比较

Table 6 Performance comparison of fine-tuning model ROUGE-1

模型名称	ROUGE-1 准确率	ROUGE-1 召回率	ROUGE-1 F1-score
distilbart-cnn-12-3	0.62	0.26	0.36
distilbart-xsum-12-3	0.58	0.38	**0.45**
mT5_multilingual_XLSum	0.55	0.23	0.32
mT5_m2o_chinese_simplified_crossSum	0.45	0.35	0.38

表 7 微调模型 ROUGE-2 性能比较

Table 7 Performance comparison of fine-tuning model ROUGE-2

模型名称	ROUGE-2 准确率	ROUGE-2 召回率	ROUGE-2 F1-score
distilbart-cnn-12-3	0.42	0.16	0.23
distilbart-xsum-12-3	0.37	0.23	**0.27**
mT5_multilingual_XLSum	0.29	0.11	0.15
mT5_m2o_chinese_simplified_crossSum	0.25	0.18	0.21

表 8 微调模型 ROUGE-L 性能比较

Table 8 Performance comparison of fine-tuning model ROUGE-L

模型名称	ROUGE-L 准确率	ROUGE-L 召回率	ROUGE-L FI-score
distilbart-cnn-12-3	0.56	0.20	0.29
distilbart-xsum-12-3	0.51	0.30	**0.37**
mT5_multilingual_XLSum	0.45	0.17	0.25
mT5_m2o_chinese_simplified_crossSum	0.43	0.30	0.35

由表 6－表 8 可以看出，通过在目标任务数据集上进行微调，预训练语言模型可以根据具体任务的特点和特征进行调整，从而提升在下游任务上的性能，在 3 个指标上均取得了优秀的测试结果。微调模型使得在大规模数据上学习到的丰富语言知识和表示能够迁移到海洋新闻的自动挖掘任务中，从而实现对聚类后的海洋新闻进行主题挖掘和摘要生成。

distilbart-xsum-12-3 模型在中文文本摘要生成任务上的性能表现最优，因此默认采用该模型来进行后续海洋新闻的信息挖掘。

4.3.3 聚类关键词提取

通过海洋新闻文本聚类实验，成功将新闻数据集中的样本划分为 14 个不同的簇群。在此基础上，综合考虑基于语义相似度和距离度量的方法，选择出每个簇群中最能反映簇群主题的中心句子。在该过程中，尝试了两种方式进行中心句子的挑选。对于每个簇群，通过计算该簇中所有样本与聚类中心之间的语义相似度和距离，综合选择与聚类中心最相似且最近的句子作为其中心句子。另外，采用关键词提取的方式，选择与簇类关键词具有最高重叠率的句子作为簇群表征。在此基础上，使用 distilbart-xsum-12-3 模型对簇群的中心句子进行主题挖掘。表 9 列出了使用该模型对海洋新闻进行聚类分析的最终结果。

表 9 海洋新闻聚类结果主题挖掘

Table 9 Topic mining of ocean news clustering results

簇类序号	新闻比例/%	簇群主题关键词
1	1403(0.062)	监测体系、海洋预报、海域、风险评估、观测、减灾、应急、海平面
2	1762(0.078)	北极海域、石油、天然气、能源供应、政策、法律、气候变化
3	1839(0.081)	海洋经济、海洋产业、涉海规划、示范区、高质量发展、强国战略
4	1351(0.060)	水产种业、渔业、海南、自由贸易、自贸港、服务、产业
5	1185(0.052)	联合国、命运共同体、蓝色家园、海洋公约、蓝色伙伴、国际合作、可持续、治理
6	1483(0.066)	极地、科考、考察队、破冰船、深海、冰川、海洋作业、航次
7	1910(0.084)	海洋生态、修复、填海、湿地、生态系统、环境保护、统筹治理、海岸带
8	862(0.038)	科普、海洋教育、主题活动、宣传、文化知识、国民海洋意识
9	2178(0.096)	科研、创新、研究、海洋科技、实验、科技成果、技术、研发、评审
10	1288(0.057)	海洋垃圾、微塑料、藻类、污染、危害、海洋生物、深海环境、全球变暖
11	736(0.033)	养殖、牧场、增殖、渔业资源、海洋生态、产业、发展
12	1599(0.071)	自然资源部、宗旨意识、政治建设、海洋强国、干部队伍、强化、服务
13	2448(0.108)	南海、美国、海军、俄罗斯、潜艇、演习、护卫舰、战略部署、主权、威胁、和平
14	2600(0.115)	船舶、港口、集装箱、海洋经济区、海运产业、海事、航线运输、一带一路

由簇群主题关键词挖掘实验结果可以看出，新闻簇类的主题关键词均很好体现了海洋领域相关的高热度新闻内容，涵盖了海洋经济、海洋生态环境、海上安全形势、海洋文化教育以及海洋科技等多方面多属性的社会热点内容，并且与人工挑选的新闻热点相一致。同时，可以看到不同聚类的主题关键词之间有明显的差异，而聚类内部的主题关键词具有很强的相关性，这也进一步证明了该方法聚类的有效性。

4.3.4 聚类自动摘要生成

在以上实验的基础上，利用微调后的生成式模型以及 Baichuan2-13B-chat 大模型，通过本地部署，对簇类中心句子进行自动摘要生成，表 10 清楚列出了使用该方法进行海洋新闻自动摘要的最终结果。

表 10 海洋新闻聚类自动摘要结果

Table 10 Automatic summary results of ocean news clustering

簇类编号	生成摘要信息
1	摘要：全国沿海海平面变化影响调查评估工作会召开，强调科学把握海面上升影响，建立评估机制，推进风险评估试点
2	摘要：美国内政部建议未来五年所有海上石油和天然气钻井拍卖限定在墨西哥湾进行，终止阿拉斯加北极水域的租赁销售，以平衡能源供应和气候变化承诺
3	摘要：广东省启动海洋产业项目，以“双合作区”为背景，研究六大海洋产业，构建政策与科普教育体系，促进高质量海洋经济发展

（续表）

簇类编号	生成摘要信息
4	摘要：中国国际进口博览会聚焦海南自由贸易试验区和自由贸易港建设，国家领导人表示支持深化改革，聚焦建设国家水产南繁硅谷，齐心协力助推海南渔业高质量发展，各方对海南发展前景表示期待，企业积极参与合作
5	摘要：联合国海洋大会发布《蓝色伙伴关系原则》，旨在通过全球蓝色伙伴合作促进海洋可持续发展，包括生态保护、气候变化适应、污染防治、可持续资源利用、蓝色经济发展、科技创新引领等原则，致力于建设共享、透明、公正的蓝色家园
6	摘要：中国第13次北极科学考察圆满结束，取得显著成果，深化了对全球气候变化、北极生态环境和空间环境的认知，促进了北极航道的利用，推动了特种装备研发和国际科技合作
7	摘要：2023海洋保护大会在荣成召开，聚焦海洋生态文明、污染治理和生态系统保护，旨在通过对话促进全民参与海洋生态建设
8	摘要：厦门海洋讲师团成立于“6.8全国海洋宣传日”，通过科普讲座和实地体验，向中小学生普及海洋知识，激发青少年对海洋科学的热爱和理解，播撒蓝色梦想
9	摘要：中国海洋工程咨询协会评选2021年度海洋工程科学技术奖45项，包括两项特等奖、12项一等奖、31项二等奖，涵盖海洋工程多个领域。评审历时10年，吸引了国内外顶级涉海科学家参会，成为集聚海洋高端创新资源的有力依托
10	摘要：微塑料与微藻群落相互作用，促进有害微藻传播，加剧赤潮风险，威胁海洋生态系统和食物网。UNEP预计到2050年全球海洋中微塑料重量将超过鱼类，加强对微塑料附生微藻的研究对了解海洋生态系统至关重要
11	摘要：中国农业农村部推动全国海洋牧场建设，到2025年计划创建178个国家级海洋牧场示范区。海洋牧场采用科技先导，通过人工鱼礁、智能平台等手段促进海洋经济发展，政策支持和技术创新推动海洋牧场成为综合发展的新引擎，解决海洋生态和渔业问题
12	摘要：战略所党委组织会议，以“加强党的政治建设，推动海洋强国建设”为主题，强调加强党的政治建设和海洋强国建设的重要性，提升研究和思维能力，推动党建与科研相融合
13	摘要：美国在南海举行大规模军演，突显紧张关系。维护地区和平与稳定需双方通过对话与合作解决争端，避免激化矛盾，促进共赢
14	摘要：中国港口在“一带一路”合作中发挥重要作用，港口活力指数显著增长，基础设施改善，全球运输能力提升，第三方市场合作推动港口深化，守护海洋专属经济区

由最终自动摘要生成的结果可以看出，针对每个簇类的中心句子，其生成的摘要能够准确地捕捉到原始文本的关键信息和主要内容，并提供了一个简洁且完整的概述。此外，能够直观看出，生成的摘要具有良好的语言流畅性，能完全满足读者的阅读要求，最终在准确性、流畅性、简洁性等方面取得平衡，提供了有价值且易于理解的文本摘要，实现了对海洋热点新闻的挖掘任务。

结束语 本文基于文本聚类技术和自动摘要方法，成功开发了一套基于深度学习的海洋热点新闻挖掘方案，包括数据采集、预处理、特征提取、文本聚类和自动摘要五大功能模块。该方案针对海洋新闻的特点，设计一系列特定的数据处理和模型微调策略，使得系统能够更准确地抓取和理解海洋新闻内容；实现了多源网络涉海数据的自动采集、结构化存储，并通过文本聚类和摘要生成实现了深层次挖掘多源异质网络海洋新闻的目标。

同时，仍然存在一些问题需要进一步研究和改进。在多源文本聚类方面，可以探索更多特征提取方法和聚类算法，以提高准确性和效率；在自动摘要生成方面，可以优化预训练语言模型的微调策略，提高生成摘要的一致性和流畅性。在未来的研究中，我们将进一步探索多模态信息的融合和情报学的前沿技术，将图像、视频等多种形式的海洋数据纳入考虑范围，以获取更全面的海洋信息。此外，也将进一步探索大语言模型在处理海洋新闻方面的巨大潜力，开发专门的海洋大模型，为提供可靠新闻报道、反映热点事件和监测舆论走向作出更大贡献。

参 考 文 献

[1] LIU Z C，LIN G S，GOH W L. Bottom-up scene text detection with Markov clustering networks[J]. International Journal of Computer Vision，2020，128：1786-1809.

[2] FAN J C. Large-scale subspace clustering via k-factorization [C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021：342-352.

[3] DANG Z Y，DENG C，YANG X，et al. Nearest Neighbor Matching for Deep Clustering[C]// Proceedings of the IEEE/CVF Conference on Compute Vision and Pattern Recognition. 2021：13693-13702.

[4] HARTL P，KRUSCHWITZ U. Applying Automatic Text Summarization for Fake News Detection[J]. arXiv：2204. 01841，2022.

[5] LI H R，ZhU J N，ZhANG J J，et al. Keywords-guided abstractive sentence summarization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：8196-8203.

[6] ABDI A，HASAN S，SHAMMUDDIN S M，et al. A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion[J]. Knowledge-Based Systems，2021，213：106658.

[7] ALAMI N，MEKNASSI M，EN-NAHNAHI N，et al. Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling[J]. Expert Systems with Applications，2021，172：114652.

[8] STEFANOVITCH N，JACQUET G，LONGUEVILLE B D. Graph and Embedding based Approach for Text Clustering：Topic Detection in a Large Multilingual Public Consultation [C]// Companion Proceedings of the ACM Web Conference 2023. 2023：694-700.

[9] MCCONVILLE R，SANTOS-RODRIGUEZ R，PIECHOCKI R J，et al. N2D：(not too) deep clustering via clustering the local

manifold of an autoencoded embedding[C]// 2020 25th international conference on pattern recognition (ICPR). IEEE, 2021: 5145-5152.
[10] WANG D X, LI T R, DENG P, et al. A Generalized Deep Learning Algorithm based on NMF for Multi-view Clustering[J]. IEEE Transactions on Big Data, 2022, 9(1): 328-340.
[11] GEORGE L, SUMATHY P. An integrated clustering and BERT framework for improved topic modeling[J]. International Journal of Information Technology, 2023, 15(4): 2187-2195.
[12] OLUKANMI P, NELWAMONDO F, MARWALA T, et al. Automatic detection of outliers and the number of clusters in k-means clustering via Chebyshev-type inequalities[J]. Neural Computing and Applications, 2022, 34(8): 5939-5958.
[13] SAHA J, MUKHERJEE J. CNAK: Cluster number assisted K-means[J]. Pattern Recognition, 2021, 110: 107625.
[14] ZHAO X W, NIE F P, WANG R, et al. Improving projected fuzzy K-means clustering via robust learning[J]. Neurocomputing, 2022, 101, 491: 34-43.
[15] UNGER H, KUBEK M, HLOCH M, et al. A survey on innovative graph-based clustering algorithms[J]. The Autonomous Web, 2022, 101: 95-110.
[16] WANG C, PAN S R, CELINA P Y, et al. Deep neighbor-aware embedding for node clustering in attributed graphs[J]. Pattern Recognition, 2022, 122: 108230.
[17] RAN X C, XI Y, LU Y, et al. Comprehensive survey on hierarchical clustering algorithms and the recent developments[J]. Artificial Intelligence Review, 2023, 56(8): 8219-8264.
[18] DOGAN A, BIRANT D. K-centroid link: a novel hierarchical clusteringlinkage method[J]. Applied Intelligence, 2022, 52(5): 5537-5560.
[19] IKOTUN A M, EZUGWU A E, ABUALIGAH L, et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data[J]. Information Sciences, 2023, 622: 178-210.
[20] HUANG S D, KANGZ, XU Z, et al. Robust deep k-means: An effective and simple method for data clustering[J]. Pattern Recognition, 2021, 117: 107996.
[21] SHRIFAN N H M M, AKBAR M F, ISA N A M. An adaptive outlier removal aided k-means clustering algorithm[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(8): 6365-6376.
[22] HE W H, WU C J, ZHOU S J, et al. Study on Short Text Clustering with Unsupervised SimCSE[J]. Computer Science, 2023, 50(11): 71-76.
[23] LI Y F, YANG M, X PENG D Z, et al. Twin contrastive learning for online clustering[J]. International Journal of Computer Vision, 2022, 130(9): 2205-2221.
[24] LIU Y, TU W X, ZHOU S H, et al. Deep graph clustering via dual correlation reduction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(7): 7603-7611.
[25] CAI J Y, FAN J C, GUO W Z, et al. Efficient deep embedded subspace clustering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1-10.
[26] SUBAKTI A, MURFI H, HARIADI N. The performance of BERT as data representation of text clustering[J]. Journal of big Data, 2022, 9(1): 1-21.
[27] CAI J Y, WANG S P, XU C, Y et al. Unsupervised deep clustering via contractive feature representation and focal loss[J]. Pattern Recognition, 2022, 123: 108386.
[28] RONEN M, FINDER S E, FREIFELD O. DeepDPM: Deep clustering with an unknown number of clusters[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 9861-9870.
[29] EL-KASSAS W S, SALAMA C R, RAFEA A A, et al. Automatic text summarization: A comprehensive survey[J]. Expert systems with applications, 2021, 165: 113679.
[30] CAI X Y, SHI K L, JIANG Y H, et al. HITS-based attentional neural model for abstractive summarization[J]. Knowledge-Based Systems, 2021, 222: 106996.
[31] LIU Y X, LIU P F, RADEV D, et al. BRIO: Bringing order to abstractive summarization[J]. arXiv: 2203. 16804, 2022.
[32] JIN H Q, WANG T M, WAN X J. SemSUM: Semantic dependency guided neural abstractive summarization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 8026-8033.
[33] JOSHI A, FIDALGO E, ALEGRE E, et al. RankSum—An unsupervised extractive text summarization based on rank fusion [J]. Expert Systems with Applications, 2022, 200: 116846.
[34] JOSHI A, FIDALGO E, ALEGRE E, et al. DeepSumm: Exploiting topic models and sequence to sequence networks for extractive text summarization[J]. Expert Systems with Applications, 2023, 211: 118442.
[35] MAO X J, WEI Y, YANG Y R, et al. KHGAS: Keywords Guided Heterogeneous Graph for Abstractive Summarization [J]. Computer Science, 2024, 51(7): 278-286.
[36] SRIVASTAVA R, SINGH P, RANA K P S, et al. A topic modeled unsupervised approach to single document extractive text summarization [J]. Knowledge-Based Systems, 2022, 246: 108636.
[37] KHURANA A, BHATNAGAR V. Investigating entropy for extractive document summarization[J]. Expert Systems with Applications, 2022, 187: 115820.
[38] JING B Y, YOU Z Y, YANG T, et al. Multiplex graph neural network for extractive text summarization[J]. arXiv: 2108. 12870, 2021.
[39] KOURIS P, ALEXANDRIDIS G, STAFYLOPATIS A. Abstractive text summarization based on deep learning and semantic content generalization[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5082-5092.

QIN Xianping, born in 2000, postgraduate. Her main research interests include neural architecture search and natural language processing.

WANG Dong, born in 1979, Ph.D, senior engineer. His main research interests include machine vision, embedded system, software programming, and IoT design.