基于大模型的动态网络舆情热点信息特征聚类挖掘

赵伟康 [1]
ZHAO Weikang

摘 要 在对网络舆情数据处理时，精准挖掘热点信息特征并聚类能够助力政府和企业决策。但传统方法未将多样化的特征进行融合，导致特征的信息挖掘离散，挖掘结果可靠性差。为此，研究基于大模型的动态网络舆情热点信息特征聚类挖掘方法，从而提升舆情信息聚类挖掘精准性。通过大模型提取社交媒体平台的相关数据信息，并获得信息文本主题特征。针对从海量数据中提取出的多样化特征进行融合，将不同特征中的信息整合。运用无监督的主题聚类方法描述信息融合特征，从而挖掘文本深层次的语义内容。实验结果表明，进行数据特征聚类挖掘能够达到较高的聚敛度水平，数据聚类的融合性也表现良好；该方法的聚类正确率结果为 90% 以上，表现出良好的文本聚类效果。

关键词 大模型；网络舆情；热点信息；聚类；特征；挖掘

doi：10.3969/j.issn.1672-9528.2025.07.033

0 引言

在信息化时代的背景下，互联网已成为舆情传播的主要载体，网络舆情作为公众意见和情感的集中反映，对社会治理、公共决策等领域产生了深远影响。在海量的舆情数据中准确识别热点信息，挖掘其特征和规律，为舆情分析提供了新的技术手段。大模型因高效的特征提取能力，能够深度挖掘舆情数据中的隐含信息和潜在规律，为舆情热点信息的识别和聚类提供了更为精准和高效的解决方案。通过构建科学的分析框架，利用大模型技术对网络舆情数据进行多层次的分析，实现舆情热点的精准识别与特征提取。通过聚类算法对舆情信息进行特征分析，挖掘不同舆情热点之间的相似性与差异。

在研究过程中，文献 [1] 选择某个数据点作为初始聚类中心。计算每个聚类的中心点距离。在大数据场景下，对用户评论进行聚类分析。不同的初始聚类中心可能导致不同的聚类结果，算法的稳定性较差。文献 [2] 在文本编码层之后加入注意力层，使模型能够关注文本中的关键部分。注意力机制可以通过计算注意力权重，突出文本中的关键信息。虽然注意力机制能够处理长距离依赖关系，但对于非常长的文本序列，仍然存在一定的限制。因此现阶段，以动态网络舆情热点信息特征聚类挖掘为研究对象，结合大模型进行实际分析与讨论。

1. 郑州信息科技职业学院 河南郑州 450008
[基金项目] 2025 年河南省科技攻关项目（252102210123）

1 信息特征聚类挖掘

1.1 大模型提取主题特征

由于网络舆情数据规模庞大、内容繁杂，提取主题特征能够从海量信息中精准提炼出关键要点，确定每部分数据的核心主题。而大模型凭借海量文本处理能力、深度语义理解等，可高效应对网络舆情数据规模大、内容复杂、噪声干扰的问题。通过大模型提取社交媒体平台的相关数据信息，并获得信息文本主题，并结合情感词典分析出网络舆情中的情感倾向，提升理解公众对热点事件的态度和情绪。为此，大模型通过引入 D 先验分布，以概率分布的形式深入解析每个文本中的潜在主题，并依据这些主题对语料库进行有效分类 [3]。在模型中，设定主题 z 与单词 w，其分布遵循多项式分布。这一假设下，主题模型的变量联合分布可以表示为：

$$P(w,z,\theta)=\prod_{j=1}^{M}p(\theta;\alpha)\prod_{j=1}^{M}p(\vartheta;\beta) \tag{1}$$

式中：α、β 为控制参数；θ 为控制主体的分布；ϑ为控制单词的分布参数；z 为单词对应的主题；w 为右单词分布与主题形成的一系列单词。

这些单词的有序组合构成了一篇热点信息文档，重复此过程即可生成语料库中的大量热点信息文档 [4]。随后，通过将这些生成的热点信息文档与原始热点信息文档进行分析，可以找出隐式分布中的最佳参数排布，从而最大化模型的概率解释力。在大模型的应用中，设置怀疑度用来衡量语言模型对数据集解释能力好坏。通过计算模型在数据集上产生的概率的倒数，来反映模型的整体泛化能力。怀疑度的计算公式为：

$$p(w)=\sum_{d}\prod_{j=1}^{M}\sum_{j=1}^{T}p(w\mid z=j)p(d) \tag{2}$$

式中：d 为语料库中的测试集；M 为热点信息文档的数量；$p(w)$ 为单词序列在模型下产生的概率。

进一步地，$p(w)$ 可以通过对热点信息文档中每个单词的主题分配和主题下的单词分布进行求和来得到。通过计算单个热点信息文档的概率，获得主题和单词的联合概率分布[5]。同时，设置 PMI 值来判断语料库中目标词和基准词相互关系。设目标词集合为 c，其中每个目标词与基准词 s 的 PMI 值计算式为：

$$\mathrm{PMI}(s,c)=\log_2\frac{P(s,c)}{P(s)P(c)} \tag{3}$$

式中：$P(s,c)$ 为两个单词同时出现的概率。

PMI 值的大小反映了两个词之间的关联强度，正值表示两词正相关，负值则表示负相关。使用预处理后的文本数据训练大模型，得到每个文档的主题分布和每个主题下的单词分布。根据大模型训练结果，识别出语料库中的主题，并为每个主题分配一个标签或描述。对于每个主题，提取出概率最高的几个单词作为该主题的特征词，这些特征词能够反映主题的核心内容。在实际应用中，大模型通过捕捉到随时间变化的热点词语及其关系，从而更准确地反映网络舆情热点的动态变化情况。通过大模型提取的主题特征，可以对热点事件的发展趋势进行预测，为决策提供有力支持。

1.2 网络舆情热点信息特征融合

1.1 节提取的特征包括文本的情感倾向、关键词的频率等，从不同角度反映了网络舆情的内在规律和外在表现。为了更全面地理解和把握网络舆情的动态变化，需要将这些多样化特征进行融合[6]。将文本情感倾向与关键词频率等多样化特征融合，能把分散在不同特征中的信息整合起来，形成更全面的舆情画像。为此，采用快速特征收敛法，该方法通过迭代优化的方式，逐步调整各特征在融合过程中的权重和贡献，使得融合结果能够更快地收敛到最优解。建立网络多属性融合的大模型，将这些属性特征有机地融合在一起，形成网络舆情表征。通过灰度化的处理方式将这些灰度特征信息进行重新组合和编码。通过灰度特征信息重组，可以得到网络多属性聚类的更新参数解。其公式为：

$$\lambda=E+Y\cdot\eta+f(t) \tag{4}$$

式中：E、Y 为特征分布相关的参数；η 为调整因子；$f(t)$ 为与特征有关的函数。

在灰度特征重组的过程中，引入差分进化能够在保持各属性原有信息的基础上，挖掘出它们之间的隐含联系[7]。差分进化对属性差异的迭代计算，逐步逼近最优的特征组合方式，从而得到差分进化的约束相关性因子。这个因子反映了在灰度特征重组过程中，各属性之间相互制约、相互影响的关系。构建网络多属性大数据分类的联合特征解，反映网络数据的内在特性。为了得到这个联合特征解，设计了一个计算公式为：

$$\mu=j\cdot X\cdot E_s(j) \tag{5}$$

式中：$E_s(j)$ 为特征分布主题中的分布集扰动。根据差分进化的约束相关性因子和各个属性特征向量的信息，通过计算得到联合特征解[8]。根据融合解对这些属性进行有机融合。以 P 为网络概率密度，根据融合参数得到其融合结果为：

$$T(t)=\int U_{ij}(t)\mathrm{d}t+j \tag{6}$$

式中：$U_{ij}(t)$ 为动态特征分布信息熵。

根据融合参数进行有机融合，充分挖掘网络多属性数据价值，得到更精确的特征融合结果。

在网络舆情热点信息特征融合中，采用快速特征收敛法、建立网络多属性融合大模型，结合灰度化处理和差分进化等手段[9]，能够有效整合文本情感倾向、关键词频率等多样化特征，挖掘特征间隐含联系，得到更精确的融合结果，充分发挥多属性数据价值。这不仅能全面呈现网络舆情态势，还为后续的聚类挖掘文本语义提供了坚实的数据基础。

1.3 网络舆情热点信息文本语义聚类挖掘

1.2 节完成了网络舆情热点信息特征融合，为了有效地利用融合后的信息特征，选择海量舆情信息数据，结合来源运用无监督的主题聚类方法进行文本读取和处理，从而挖掘文本深层次的语义内容。运用无监督的主题聚类方法，从数据的标准化处理、聚类划分，到语义挖掘和降维处理，各个环节紧密配合，有效解决了聚敛度低的问题，为网络舆情热点信息的分析和挖掘提供了有力支持。将原始的文本数据转化为标准化源数据。在这个过程中，需要完成文本向量化。设定待聚类的文档集为 $U=\{b_1,b_2,\cdots,b_n\}$，其中 b 表示待聚类的文档。为进行聚类，首先随机选取某个文档 b_1，创建首个聚类中心 ε。设定 O 为文档集合 U 的几何中心。几何中心 O 可以通过计算所有文档向量的平均值来得到。其公式为：

$$O=\frac{1}{n}\sum_{i=1}^{n}b_i \tag{7}$$

式中：n 为文档向量的数量。

后计算几何中心 O 与聚类中心 ε 的距离。这个距离可以通过欧氏距离方法来计算。其公式为：

$$D=\sqrt{\sum_{j=1}^{d}(O-\varepsilon)^2} \tag{8}$$

式中：D 为聚类中心坐标向量与几何中心坐标向量之间的欧式距离。将这个距离与预先设定的阈值 K 进行比较。如果 $D<K$，将文档集合 U 中的文档划分到 ε 对应的聚类中。如果 $D>K$，形成一个新的聚类，并将 ε 更新为当前考虑的文档 U。直到待聚类文档无法再划分停止，从而得到了若干个聚类，每个聚类都包含了一组相似的文档[10]。聚类完成后，舆情文本可自动标注，其中可选取聚类内部相似度最高的文档作为挖掘对象。

对挖掘对象进行文本语义挖掘。深层语义间存在关联性，这种关联可通过大规模文本统计分析来得到。在此过程中，可统计词语在文档中的共现频次，若某些词语频繁共现，则表明其之间存在强关联。为构建新的语义空间，需对挖掘对象进行预处理，并设定降维矩阵，其中降维矩阵的元素能够反映词语之间的联系。通过计算这个降维矩阵，将挖掘对象的高维语义空间映射到一个低维的语义空间中，同时保留原始语义信息的大部分内容。其计算公式为：

$$Z=\frac{1}{m}X^{T}W \tag{9}$$

式中：m 为样本数量；X 为原始数据；W 为降维矩阵。经过上述操作，通过聚类挖掘来获得原始语义信息，结合挖掘不同的语义关系，形成更为完整的舆情热点信息集合。这样，用户就可以更加准确地获取舆情信息，降低舆情分析结果的偏差。

2 实验测试与分析

2.1 实验环境

为验证基于大模型的网络舆情热点信息特征聚类挖掘方法的有效性，通过对比实验评估其聚类性能和准确性。收集近期网络上的舆情热点信息，包括新闻、社交媒体帖子、论坛讨论等。对收集到的数据进行预处理，构建实验数据集，确保数据集包含多个不同的舆情热点主题。搭建实验环境，配置 8 GB 的内存。安装并配置好所需的大模型软件框架 PyTorch。确定聚类效果的评估指标，设定特征维度为 200 维，使用大模型对预处理后的数据进行特征提取。根据设定的评估指标对聚类结果进行评估，记录实验结果。对比不同方法的聚类效果，分析基于大模型的聚类方法有效性。

2.2 结果分析

在网络舆情热点信息采集与分析中，实验设定了样本数量为 800 个，数据分类的类别为 10，相似度融合系数为 0.30，迭代次数为 120。基于这些参数设定，通过 K-means 聚类算法对样本数据进行处理，并绘制轮廓分析图，包括轮廓图和聚类数据可视化图，以评估聚类效果并直观呈现数据分布特征，得到了网络舆情热点数据的统计特征量分布，具体如图 1 所示。

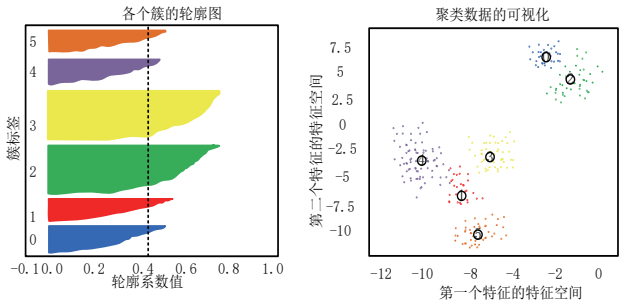

图 1 特征量分布结果

通过对特征量分布进行分析后发现，采用本文提出的方法进行数据特征聚类挖掘，能够达到较高的聚敛度水平。同时，数据聚类的融合性也表现良好，说明本文方法在处理网络多属性大数据时，能够有效地将具有相似特征的数据点聚集在一起，为后续的数据分析和挖掘提供了有力的支持。

为验证本文基于大模型的动态网络舆情热点信息特征聚类挖掘方法的卓越性能，本实验选取了 10 个具有广泛代表性的舆情热点类别进行深入研究。分别为 A：社会民生；B：科技前沿；C：教育领域；D：环境保护；E：文化娱乐；F：体育赛事；G：经济发展；H：交通出行；I：食品安全；J：自然灾害。通过对这些不同类别舆情数据的聚类分析，对比实际类别与模型聚类结果，以此全面评估本文方法的聚类准确度，进而验证其有效挖掘网络舆情热点信息的能力。混淆矩阵结果如图 2 所示。

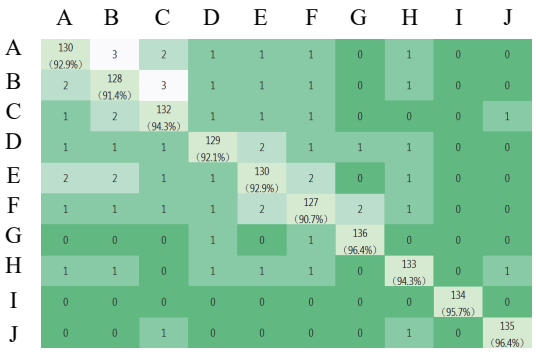

图 2 混淆矩阵结果

如图 2 所示，本文所提基于大模型的动态网络舆情热点信息特征聚类挖掘方法在各舆情热点类别上均展现出较高的聚类准确度，聚类正确率均在 90% 以上。这是因为对舆情文本自动标注并挖掘深层语义关联，利用降维矩阵在保留关键语义信息的同时降低数据维度，使得在处理专业性较强的舆情时，也能实现精准聚类。这些方法相互配合，共同提升了整体的聚类效果，有效挖掘了网络舆情热点信息。本文提出的基于大模型的动态网络舆情热点信息特征聚类挖掘方法具有显著的优势。

3 结语

在本次基于大模型的网络舆情热点信息特征聚类挖掘研究中，通过深入分析海量网络数据，结合先进的大模型技术，成功实现了对舆情热点信息的精准识别与特征提取。在处理多维度的网络舆情数据时表现出了良好能力。通过对不同时间维度、地域维度以及用户行为维度的综合分析，不仅能够迅速捕捉到舆情的动态变化，还能够深入挖掘舆情背后的社会心理和文化背景。通过聚类分析，能够更好地理解舆情的传播规律，预测其发展趋势，进而为相关部门提供科学决策支持。以期在全球化和多元化的背景下，实现对网络舆情的全面监测与深度挖掘。

基于计算机通信的电子信息工程数据共享技术研究

陈 钊[1]
CHEN Zhao

摘 要 针对基于计算机通信的电子信息工程数据共享技术，文章提出了一种基于区块链的数据共享方案。通过构建包含数据提供节点、区块链服务器、数据需求节点和云服务平台的共享框架，实现数据的高效共享。并采用区块链技术、哈希时间锁定合约和闪电网络链下交易机制，对数据共享流程进行设计。仿真分析表明，所提方案在吞吐量和交易延迟方面具有较好优势。在低提现率下，闪电网络的处理能力远超传统比特币网络，有效提升了电子信息工程数据共享的效率。

关键词 区块链；计算机通信；电子信息工程；数据共享

doi：10.3969/j.issn.1672-9528.2025.07.034

0 引言

随着我国电子信息产业的快速发展，数据共享已成为推动产业创新、提高企业竞争力的关键因素。然而，在数据共享过程中，数据泄露、隐私侵犯、信任缺失等问题日益凸显，严重制约了电子信息工程领域的发展。区块链技术，作为一种去中心化、安全可靠、透明度高的新型技术，近年来在金融、物联网、供应链等领域取得了显著成果。区块链技术的核心优势在于其去中心化的特点，通过加密算法和网络共识机制，实现了数据的安全存储和高效传输。在此基础上，本文将区块链技术引入电子信息工程数据共享领域，旨在解决现有技术中存在的诸多问题。

1 区块链技术

区块链技术是一种数据存储方式，是将数据块按照时间顺序链接，形成一条不断扩展的链。每个区块由区块头和区块主体组成。其中，区块头包含版本号、区块高度、区块哈希值、时间戳、默克尔（Merkle）根以及一个与共识机制相关的随机数；区块主体则记录具体的交易信息。为确保数据的完整性和提高交易验证的效率，区块链利用哈希树进行数据校验[1]。

哈希树是一种基于哈希算法构建的二叉树结构，该结构

1. 安徽电信规划设计有限责任公司 安徽合肥 230000

参考文献：

[1] 王红林，李忠伟．大数据场景下用户评论聚类文本挖掘算法 [J]. 计算机仿真，2024,41(3):352-358.

[2] 孟凡会，王玉亮，汪卫霞．基于注意力机制的在线用户痛点信息挖掘 [J]. 情报理论与实践，2023,46(10):192-199.

[3] 周艳秋，高宏伟，何婷，等．电子监控部分遮挡目标单模态自监督信息挖掘技术 [J]. 现代电子技术，2024,47(10):47-51.

[4] 王宇琪，周庆山，赵菲菲．面向信息弱势群体的电子公共服务网络评论观点挖掘与诉求主题分析 [J]. 情报资料工作，2023, 44(4): 77-84.

[5] 邢春玉，张莉，冯卿松．基于可视化技术的审计信息挖掘及分析研究 [J]. 财会通讯，2023(13):17-23.

[6] 李红艳，徐寅森，张子栋．蜂窝移动网络大数据聚类异常挖掘方法仿真 [J]. 计算机仿真，2024,41(2):406-409.

[7] 王延，周凯，沈守枫．基于熵权法的教务大数据的挖掘和聚类分析 [J]. 浙江工业大学学报，2023,51(1):84-87.

[8] 蒋希文，王丽珍，Vanha TRAN. 基于模糊密度峰值聚类的区域同位模式并行挖掘算法 [J]. 中国科学：信息科学，2023, 53(7): 1281-1298.

[9] 周燕，肖莉．基于改进关联聚类算法的网络异常数据挖掘 [J]. 计算机工程与设计，2023,44(1):108-115.

[10] 康耀龙，冯丽露，张景安，等．基于谱聚类的不确定数据集中快速离群点挖掘算法 [J]. 吉林大学学报（工学版），2023, 53(4): 1181-1186.

【作者简介】

赵伟康（1990—），男，河南郑州人，博士，讲师、高级工程师，研究方向：大数据、人工智能。

（收稿日期：2025-03-21 修回日期：2025-07-08）