孙自梅, 黄承宁, 朱玉全. 基于并行聚类算法的高校舆情分析研究[J]. 智能计算机与应用,2026,16(8):180-185. DOI:10.20169/j.issn.2095-2163.24121303

基于并行聚类算法的高校舆情分析研究

孙自梅[1], 黄承宁[1], 朱玉全[2]
(1 南京工业大学 浦江学院, 南京 211222; 2 江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013)

摘 要: 本文针对高校网络舆情数据的聚类分析,提出了一种并行层次聚类算法。该算法通过分治策略和并行计算技术,实现了对大规模网络舆情数据的快速、精准聚类。采用对比实验,验证了本文提出的并行层次聚类算法的有效性。实验结果表明,该算法在准确率和召回率上优于传统的 K-means 和 FCM 算法,且具有较好的可扩展性,适用于大规模数据聚类分析。此外,该算法不仅提高了聚类的速度,而且保持了较高的聚类准确率和召回率,显示出了良好的性能和实用价值。因此,该算法为高校网络舆情监测和分析提供了一种新的有效工具,具有广泛的应用潜力和推广价值。

关键词: 高校网络舆情; 分治策略; 并行层次聚类; 准确率; 召回率
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2026)08-0180-06

Research on university online public opinion analysis based on parallel hierarchical clustering algorithm

SUN Zimei[1], HUANG Chengning[1], ZHU Yuquan[2]
(1 Pujiang Institute, Nanjing Technology University, Nanjing 211222, China;
2 School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China)

Abstract: This paper proposes a parallel hierarchical clustering algorithm for the cluster analysis of university online public opinion data. The algorithm achieves fast and accurate clustering of large-scale network public opinion data through the divide-and-conquer strategy and parallel computing technology. Contrastive experiments verified the effectiveness of the proposed parallel hierarchical clustering algorithm. The experimental results show that the algorithm outperforms traditional K-means and FCM algorithms in terms of accuracy and recall rate, and has good scalability, making it suitable for large-scale data clustering analysis. In addition, the algorithm not only improves the speed of clustering but also maintains high clustering accuracy and recall rate, demonstrating good performance and practical value. Therefore, the algorithm provides a new effective tool for the monitoring and analysis of university online public opinion, with broad application potential and popularization value.

Key words: university online public opinion; Divide-and-conquer strategy; Parallel hierarchical clustering; Accuracy rate; Recall rate

0 引 言

随着互联网的广泛渗透与社交媒体的飞速发展,网络舆情已成为影响社会、政治和文化等领域的关键因素。在高校里,网络舆情的波动直接关系到学生的思想教育和校园的稳定[1]。如何从海量且复杂的数据中快速提取有价值的信息并合理解读,是当前网络舆情分析的关键[2-3]。传统的舆情分析方法处理速度较慢,而且在大规模数据处理上存在一定的局限性,这对高校舆情的快速响应处置构成了挑战[4-5]。因此,本文拟提出一种高效率、高准确度的网络舆情分析技术,旨在提升校园舆情管理的质量和效率,为维护高校稳定和谐发展提供支撑。

1 相关技术分析

1.1 层次聚类算法

层次聚类是一种关键的数据分析方法,其通过逐步合并相似的数据点来构建一个分层的聚类结

基金项目: 国家自然科学基金(60572112);南京工业大学浦江学院自然课题(NJPJ2024-1-05)。
作者简介: 黄承宁(1985—),男,博士,副教授。主要研究方向:计算机应用,智能信息处理;朱玉全(1965—),男,博士,教授。主要研究方向:数据挖掘,机器学习。
通信作者: 孙自梅(1986—),女,博士,副教授。主要研究方向:计算机应用,智能数据分析。Email:736554824@qq.com。
收稿日期: 2024-12-13

构[6]。此算法主要分为凝聚型和分裂型两种策略。在凝聚型层次聚类中，每个数据点最初被视为一个独立的聚类，随后算法逐步将最相似的两个聚类合并，这个过程持续进行，直到所有数据点合并成一个聚类，或者达到预先设定的聚类数量。分裂型层次聚类以一个包含所有数据点的初始聚类开始，通过不断分裂，直至每个数据点单独成为一个聚类。这种技术适用于不同形状和大小的数据集，并能自动确定聚类数量，特别适用于在聚类结构未知的情况下进行数据探索[7]。

1.2　并行计算技术

随着数据量的爆炸式增长，传统的串行计算方式已无法满足大数据处理的效率需求，效率低下成为了技术发展的主要障碍。并行计算技术通过同时使用多个计算资源（如多核处理器、多处理器或多计算机）来执行计算任务，显著提高了计算效率。特别是 MapReduce 编程模型的出现，其通过 Map（映射）和 Reduce（归纳）两个简单的函数，使得大规模数据集的并行处理变得简单而高效。这种模型不仅可扩展性强，而且具有良好的容错性，非常适合云计算平台[8]。

1.3　高校网络舆情分析

高校网络舆情分析的重要性不言而喻，其在洞察学生思想动态、优化教育管理、提升校园文化建设水平以及防范风险等方面发挥着至关重要的作用。首先，网络舆情分析是高校把握时代脉搏、紧跟学生思想动态的重要手段。在信息化、网络化日益发达的今天，互联网已成为学生表达意见、交流思想的主阵地[9]。通过对网络舆情的深入分析，高校能够及时了解学生对时政热点、教育政策、校园生活的态度和看法，这对于制定更加科学合理的教育管理策略具有指导性意义[10]。其次，聚类分析（如层次聚类和 K-means 聚类算法）的应用，极大地提升了高校网络舆情分析的效率与准确性。这些算法通过对海量文本数据的挖掘，将相似的观点和情绪进行归类，使得高校管理者能够迅速识别出校园内的热点话题和潜在风险[11]。这种高效的分析方式有助于高校在问题发酵前采取相应措施，维护校园稳定和谐。再者，情感分析作为网络舆情分析的核心环节，对于理解学生的真实情感具有重要意义[12]。通过对学生言论的情感倾向进行判断，高校可以更加精准地把握学生对各项政策、措施的情绪反应，从而有针对性地调整工作策略，提升学生满意度。

除聚类与情感分析外，LDA 模型在高校网络舆情分析中的应用，为揭示隐藏在文本背后的深层次主题提供了有力支持[13]。LDA 模型能够从大量非结构化的文本数据中提炼出关键主题，帮助研究人员发现学生关注的焦点问题，进而为政策制定和教学改革提供数据支撑。高校可通过训练 LDA 模型，从文本中挖掘出潜在的主题，并进一步探讨不同主题下的情感走向[14-16]。这不仅有助于高校管理者了解校园内的舆论趋势，而且为政策制定和舆情引导提供了数据支持。

2　并行层次聚类算法设计与实现

2.1　算法设计

2.1.1　问题定义

层次聚类算法，作为一种传统的聚类技术，与 K-means 或模糊 C 均值（FCM）等划分式聚类方法不同[16-17]，其依据数据点之间的相似度逐步构建聚类层次结构。该算法的核心优势在于其能够有效识别数据中的不规则形态和异质空间分布，尤其适用于解析复杂结构的数据，如文本聚类[18-19]。通过构建聚类树，层次聚类方法能够清晰地将数据集划分为多个主要类别，并在每个主要类别下进一步细分为子类别，这一特性极大地便利了网络舆情的监控与分析，为决策制定提供了有力支持。

2.1.2　并行策略的制定

研究设计的并行层次聚类算法分为 3 个关键步骤：数据分区、节点内舆情聚类处理以及聚类结果的汇总与融合。

（1）数据分区

为确保每个子问题都能在一个计算节点上单独处理，本研究采用分治策略，依据相似性指标对庞大的网络舆情数据集进行分割，将整体聚类问题分解为多个可独立解决的子问题。首先，从舆情数据集中随机选取 n 条记录（n 为计算节点的数量），记为 $\{S_i \mid i = 1,2,\cdots,n\}$，这些记录作为初始种子分配到 n 个计算节点上。接着，剩余的舆情数据根据与这些种子记录的相似度进行分配，相似度由余弦距离衡量，即舆情信息 X_i 将被分配到与之相似度最高的节点上，用 $P(X_i)$ 表示舆情信息 X_i 分配的节点，公式如下：

$$P(X_i) = \min_j \mathrm{Dist}(X_i, S_j),\ j = 1,2,\cdots,n \qquad (1)$$

其中，$\mathrm{Dist}(X_i, S_j)$ 表示舆情信息 X_i 与种子记录 X_j 之间的余弦距离。此过程持续进行，直到每个节点的舆情数据量达到 N/n（N 为舆情数据集的总长度）。

(2)节点内舆情聚类处理

数据被分配到各个计算节点后,每个节点会独立地对所分配的数据子集执行层次聚类算法。这一过程涉及计算数据点间的相似度,并根据相似度逐步构建聚类层次,直至所有数据点被合理地归入相应的聚类中。初始步骤是计算数据点之间的成对距离,形成距离矩阵 $\boldsymbol{D}$。该矩阵的具体计算方式如下:

$$D_{ij} = \mathrm{Dist}(X_i, X_j) \tag{2}$$

其中,D_{ij} 代表舆情数据点 i 和数据点 j 之间的距离,由 Dist(·,·)函数计算得出。

该距离矩阵是后续聚类过程的基础,记录了所有数据点之间的相似度或差异性。通过这种方式,能够确保在每个计算节点上高效地进行舆情数据的聚类运算,为最终合并成一个全局聚类结构奠定基础(如图 1 所示)。

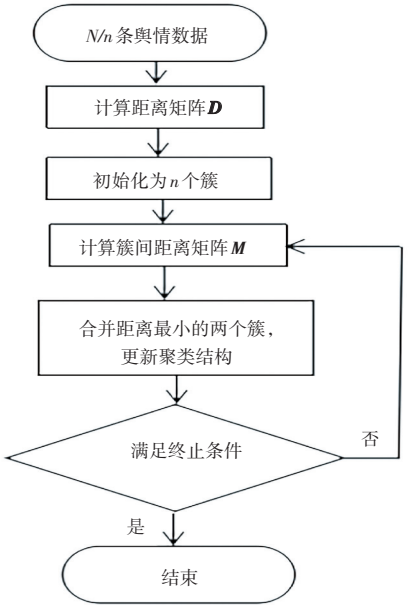


图 1 算法流程图

Fig. 1 Flowchart of the proposed algorithm

在计算簇间距离时,采用了平均距离法。具体而言,簇 C_i 与簇 C_j 之间的距离公式如下:

$$\mathrm{Dist}(C_i, C_j) = \frac{\sum_{k \in C_i} \sum_{l \in C_j} D_{kl}}{\mathrm{num}(C_i) \times \mathrm{num}(C_j)} \tag{3}$$

其中,$\mathrm{num}(C_i)$、$\mathrm{num}(C_j)$ 分别表示簇 C_i 和簇 C_j 的数据条数。

2.2 时间复杂度分析与参数 c 的选定

在每一个计算节点上,将 N/n 个初始簇合并成 c 个簇的过程中,总共需要进行 $N/n - c$ 次迭代。因此,每个计算节点的计算量公式如下:

$$T_1 \propto \frac{1}{2}\left[\frac{N}{n}\left(\frac{N}{n}-1\right) + \left(\frac{N}{n}-1\right)\left(\frac{N}{n}-2\right) + \cdots (c+1)c\right] = \frac{\left(\frac{m}{N}\right)^3 + 3\left(\frac{m}{N}\right)^2 + 2\frac{m}{N}}{3} - \frac{(c)^3 + 3(c)^2 + 2c}{3} \tag{4}$$

在聚类算法中,当需要将 nc 个簇合并为一个最终的簇时,则需要进行 $nc-1$ 次迭代,因为每次迭代合并两个簇,直到只剩下一个簇。每次合并操作的计算量取决于所使用的聚类算法和簇的特征维度。其计算量公式如下:

$$T_2 \propto \frac{1}{2}[nc(nc-1) + (nc-1)(nc-2) + \cdots + 2 \times 1] = \frac{(nc)^3 + 3(nc)^2 + 2nc}{3} \tag{5}$$

由于节点间的计算并行完成,数据划分的时间可以忽略不计,因此总计算时间 T 为 T_1、T_2 之和,即:

$$T = T_1 + T_2 \tag{6}$$

聚类个数 c 的选择对计算时间的影响可通过公式(4)和(6)分析:T 为 c 的多项式函数,对 c 求导可得:

$$\frac{\mathrm{d}T}{\mathrm{d}c} = \frac{\mathrm{d}T_1}{\mathrm{d}c} + \frac{\mathrm{d}T_2}{\mathrm{d}c} = \frac{\alpha(3n^3c^2 + 6n^2c + 2n)}{3} - \frac{\beta(3c^2 + 6c + 2)}{3} = (\alpha n^3 - \beta)c^2 + 2(\alpha n^2 - \beta)c + \frac{2}{3}(\alpha n - \beta) \tag{7}$$

$\frac{\mathrm{d}T}{\mathrm{d}c}$ 为关于 c 的二次函数,开口向上,取最小值的点为:

$$c_{T\min} = -\frac{\alpha n^2 - \beta}{\alpha n^3 - \beta} \tag{8}$$

当 $c = 0$ 时,$\frac{\mathrm{d}T}{\mathrm{d}c}$ 的取值为:

$$T_0 = 1.5(\alpha n - \beta) \tag{9}$$

由于 $T_0 > 0$ 且 $c_{T\min} < 0$,因此当 c>0 时,有 $\frac{\mathrm{d}T}{\mathrm{d}c} > 0$。因此 T 不存在极值点,而是随 c 增大的函数。当 c 逐渐增大直至 $c = N/n$,意味着并行阶段的计算量不断减小,合并阶段的计算量不断增大,计算时间增大,聚类结果较好。当 c 逐渐减小直至 $c = 1$,意味着聚类操作大部分在并行阶段完成,总计算时间大幅缩减,聚类结果可能受到影响。

3　实验分析：高校网络舆情分析

为验证本文提出的并行层次聚类算法的有效性，本文进行了 2 组实验，一组旨在评估算法的性能，另一组用于与现有算法进行比较。算法的实现基于 Python 3.1 平台，分词处理采用开源工具 HanLP 1.7.2。

实验在 12 个节点组成的集群环境下进行，节点配置为：操作系统 Ubuntu 11.1、CPU 为 Intel i3-7100 3.9 GHz、内存 8 GB，硬盘容量 500 GB。实验数据使用自编的 Python 爬虫软件从高校的贴吧、论坛、微博等平台爬取近一年的 250 万条数据，记作 Zibian_data。通过对抓取的数据进行清洗，提炼与高校舆情紧密相关的信息。这一过程可通过关键词筛选、内容归类等技术来完成。只有那些涉及高校的帖子与评论会被筛选出来，而无关的资料则会被排除。考虑到聚类过程存在一定的随机性，每组实验均重复进行 10 次，并取平均值作为最终结果。

3.1　评价标准

本实验采用准确率（Accuracy，ACC）来衡量聚类算法的性能，计算公式如下：

$$\mathrm{ACC}=\frac{\sum_{i=1}^{N}\delta(r_i,\mathrm{map}(s_i))}{N} \tag{10}$$

其中，N 代表样本总数，函数在样本的预测标签 r_i 与其真实标签 s_i 相同时取值为 1，否则为 0。map（）函数用于实现最佳类别匹配，以确保统计的准确性，这一过程通常通过匈牙利算法（Hungarian Algorithm）来完成。

3.2　实验结果

本实验将探讨聚类参数 c 和节点数量 N 对实验效果的影响，并与其他聚类算法，如 K-means 和 FCM，进行对比实验，以评估算法的性能。

（1）聚类参数 c 对实验结果的影响

实验中逐步增加聚类参数 c 的值，从较小的数目到较大的数目，同时记录算法的运行时间和聚类准确率。实验结果分别展示在图 2 和图 3 中。通过这些图表，可以观察到聚类参数 c 的变化如何影响算法的性能和效率。

由图 2 可以看出，随着 c 值的增加，算法的执行时间急剧上升。当 c 达到最大值 N/n 时，算法本质上则退化为传统的层次聚类方法，难以处理类似网络舆情等大规模数据集的聚类任务，该现象与前面分析结果相吻合。图 3 则揭示了随着 c 值的提升，聚类准确率逐渐上升但其增长速率趋于平缓。当 c 超过 $N/10n$ 后，准确率的提升幅度变得非常小，几乎接近最优值。因此，在进行网络舆情聚类时，建议将 c 值设定为 $N/10n$ 或在其附近微调，这样不仅保证了较高的聚类准确率，也确保了高效率。

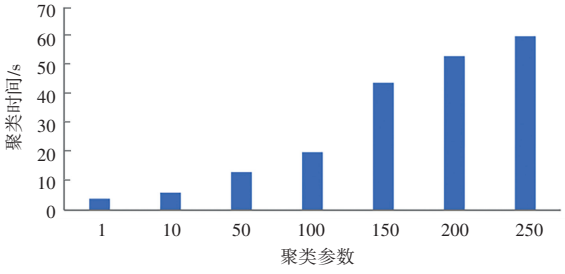


图 2　参数 c 对聚类时间的影响

Fig. 2　Impact of parameter c on clustering time

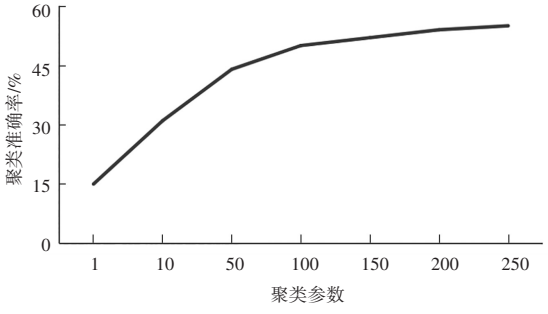


图 3　参数 c 对聚类准确率的影响

Fig. 3　Impact of parameter c on clustering accuracy

（2）节点数量对聚类时间的影响分析

为评估本文算法的可扩展性，本实验将调整服务器节点的数量。实验中，服务器节点数量 n 分别设置为 2、4、6、8、10 和 12，同时将聚类参数 c 固定为 $N/10n$。聚类时间的变化趋势和准确率结果分别如图 4 和图 5 所示。

从图 4 和图 5 可以看出，随着服务器节点数量的增加，聚类耗时逐渐降低，表明算法的并行处理能力较强，能够有效利用额外的计算资源。但聚类准确率基本保持不变，说明增加节点数量不会降低算法的准确性，算法具备良好的稳定性和可扩展性。

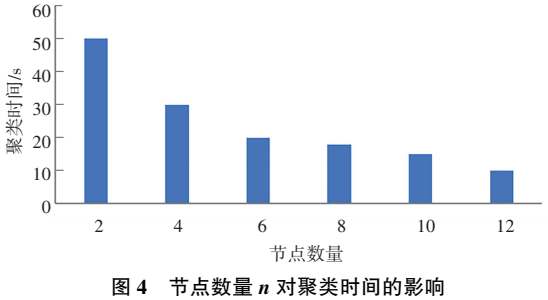


图 4　节点数量 n 对聚类时间的影响

Fig. 4　Impact of the number of nodes（n）on clustering time

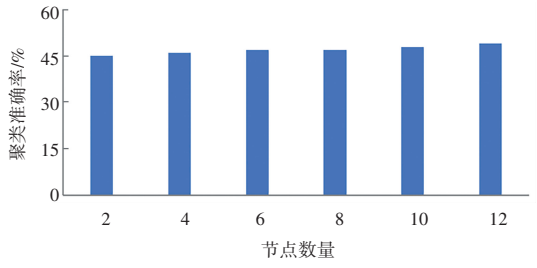


图 5 节点数量 n 对聚类准确率的影响

Fig. 5 Impact of the number of nodes (n) on clustering accuracy

(3)算法对比实验

为验证该算法的聚类效果,采用与经典算法 K-means 及 FCM 进行对比实验分析。本实验中该算法采用 12 个服务器节点的并行化处理,FCM 算法的模糊参数 m 取 2。对比实验结果见表 1,从中可以看出:在不同数据集中,该算法在聚类准确率上相较于 FCM 提高了约 10%,而 FCM 算法本身已比 K-means 准确率高出约 3%左右。同时,在聚类耗时上该算法略低于 K-means 与 FCM 算法。由此证实了该算法通过并行化加速了聚类速度,同时保持了较高的聚类准确率。

表 1 数据集对比实验结果

Table 1 Comparative experimental results on public data sets

数据集	*K*-Means		FCM		本文算法	
	耗时/s	准确率/%	耗时/s	准确率/%	耗时/s	准确率/%
自编数据集	21.71	72.84	25.57	78.65	19.93	89.04

为评估本文算法对网络舆情数据的聚类效果,分别从 Zibian_data 数据集中选取了 50、100、150、200、250 万条舆情数据,共进行了 5 组实验。实验结果见表 2。

表 2 舆情聚类对比实验结果

Table 2 Comparative experimental results of public opinion clustering

实验组号	舆情数量(万条)	*K*-Means		FCM		层次聚类		本文算法	
		耗时/min	准确率/%	耗时/min	准确率/%	耗时/min	准确率/%	耗时/min	准确率/%
1	50	10.1	72.68	11.2	78.21	103	88.97	9.5	93.65
2	100	44.7	72.57	46.8	78.58	421.2	88.62	38.1	94.13
3	150	101.3	73.23	108.5	78.76	998.6	88.99	88.3	93.47
4	200	179.6	72.75	180.3	78.38	1 836.6	88.08	165.5	92.95
5	250	289.4	71.83	299.7	78.64	2 976.2	88.87	269.1	92.44

由表 2 可以看出,本文算法在聚类准确率和效率上均展现出显著优势。相较于 K-means 和 FCM 算法,该算法准确率提升了约 10%,并且在处理时间上更为高效。由此可以表明:本文算法对于海量舆情信息聚类是有效的。此外在聚类所需时间上,随着舆情信息的增长,尤其是对海量舆情数据的分析,本文算法的时间增长速率较慢,显著优于传统层次聚类算法。因此,本文算法更加适用于对更新速度快、数据规模大的高校舆情数据进行聚类分析。

为了全面评估算法识别不同类别舆情的能力,引入了召回率作为补充评估指标。召回率的计算公式为:特定类别中聚类正确的样本数量除以数据集中该类别的实际样本数量。与此同时,继续采用准确率作为评价指标。实验结果见表 3。

由此可以看出:本文所提算法在处理各类别舆情数据时,无论是在准确率还是召回率上,均展现出了相较于 K-Means 算法、FCM 算法以及层次聚类算法的优越性。这一结果凸显了本算法在精确识别热点话题方面的强大能力。

结合以上实验结果,本文提出的并行层次聚类算法在应对大规模网络舆情数据时,不仅显著提升了聚类效率,同时确保了高准确率和召回率。经过多轮的实验验证,该算法非常适合于云计算环境下的大规模数据聚类任务,特别是在高校舆情数据快速且精确分析方面,展现出巨大的应用前景和潜力。

表 3　舆情数据各特定类别的准确率和召回率

Table 3　Accuracy and recall rates for specific categories of public opinion data

主题	*K*-means		FCM		层次聚类		本文算法	
	准确率/%	召回率/%	准确率/%	召回率/%	准确率/%	召回率/%	准确率/%	召回率/%
毕业	76.5	78.7	75.3	76.7	88.9	88.0	91.1	90.6
考研	70.8	71.3	68.2	77.2	79.3	80.1	91.8	90.4
考证	71.5	70.1	72.6	76.2	78.7	79.2	91.9	90.7
出国	79.5	69.2	80.1	70.8	83.2	83.2	93.7	91.4
工作	75.0	74.1	76.8	75.5	80.1	79.2	91.5	89.9
抑郁	68.1	70.1	71.0	72.3	78.5	80.1	91.0	88.5
旅游	76.2	71.7	75.9	72.4	81.3	81.5	91.6	89.3
饮食	74.3	70.2	73.5	73.7	89.8	88.8	91.9	92.9
交友	75.5	73.4	76.2	74.8	84.4	83.2	91.8	92.6
运动	76.3	72.1	78.1	71.1	89.2	85.5	91.2	93.5

4　结束语

本文提出云计算环境下的并行层次聚类算法，旨在提升高校网络舆情的分析效率及质量。采用分治策略，本算法大幅增强了处理大规模数据集的能力，不仅加快了聚类速度，而且在保持高准确性和召回率方面表现优异。实验结果证实，在处理海量数据时，本算法有效减少了计算时间和资源消耗，其准确率和召回率均超过了传统的 K-means 和 FCM 算法。因此，本研究提出的并行层次聚类算法为高校网络舆情监测和分析提供了一种新的有效工具，且具有很高的应用价值和推广前景。

参考文献

[1] 姜良杰，王星淇. 自媒体时代高校网络舆情事件应对研究[J]. 华北电力大学学报(社会科学版)，2020，21 (2)：134-140.

[2] 吴素华. 思想政治教育视角下高校网络舆情引导研究[J]. 教育理论与实践，2021，41 (3)：127-132.

[3] 周春梅，冯林，张华辉. 网络舆情对新冠疫情下青少年情感态度的分析[J]. 计算机仿真，2023，40 (1)：553-558.

[4] 赵好好. 基于数据挖掘的高校网络舆情热点话题分析及引导策略研究[J]. 信息系统工程，2023 (2)：117-119.

[5] 程名，刘虎，郑涵予，等. 舆情大数据在 PMI 数据分析预测中的应用研究[J]. 中国国情国力，2023(2)：44-50.

[6] 李文全，毛伊敏，彭新东. 基于犹豫模糊集的凝聚式层次聚类算法[J]. 计算机应用，2023，43(12)：3755-3763.

[7] 王鹏宇，王国宇，苏天赟. 一种基于局部特征的层次聚类算法[J]. 中国海洋大学学报(自然科学版)，2019，49(S2)：176-184.

[8] 陈雪刚. 基于云计算的网络舆情藤蔓式串并行聚类方法研究[J]. 情报杂志，2017，36(2)：126-127.

[9] Kaufman L，Rousseeuw P J. Finding groups in data：An Introduction to Cluster Analysis[M]. New York：Wiley，2008：253-279.

[10] 韩磊，夏明亮，施展. 基于大数据技术的高校舆情分析模型研究[J]. 智能计算机与应用，2024，14(11)：194-199.

[11] 陈涛，栾禹鑫，谭英杰，等. 基于爬虫技术的校园网络舆情分析和监测系统[J]. 网络安全技术与应用，2018(12)：54-55.

[12] 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.

[13] 曾莉，杨添宝，周慧. 基于 LDA 与注意力机制 BiLSTM 的微博舆情分析模型[J]. 南京理工大学学报，2022，46(6)：742-748.

[14] 钟金宏，黎梦萍，宣占祥. 基于 LDA 模型的高校疫情防控舆情分析[J]. 信息技术与信息化，2022(12)：72-75.

[15] Zhu Yalin，Li Xiangwei，Wang Juan. Analysis and research of weibo public opinion based on text[J]. Journal of Physics Conference Series，2021，1769(1)：012018.

[16] 唐欣. 优化的 K-means 聚类算法在客户细分中的应用研究[J]. 智能计算机与应用，2023，13(9)：194-197.

[17] 陈洪涛. 基于 K-means 算法的学生行为数据与学习成绩关联性分析[J]. 中国科技信息，2024(23)：86-88.

[18] 高群，薛超凯. 考虑局部密度的电力通信大数据并行聚类算法[J]. 计算机仿真，2024，41 (10)：59-62.

[19] 马文博，巫朝霞. 基于差分隐私保护的二分 k 均值聚类算法研究[J]. 智能计算机与应用，2023，13(2)：155-160.