文章编号：2095-2163(2023)06-0084-06 中图分类号：TP391.1 文献标志码：A

文本挖掘在新能源汽车领域中的应用

张 雨，黄润才
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：在新能源汽车领域中使用文本挖掘，可以回顾新能源汽车的发展历程、预测未来发展趋势及研究热点。本文从中国国家知识基础设施数据库（CNKI）和科学网（WOS）数据库中分别获取了 16 293 篇（2011～2020）和 10 328 篇（2012～2020）论文，并使用文本挖掘算法对这些论文进行研究，包括词嵌入、T-SNE 降维、小批量 K-Means 聚类等，得出国内外新能源汽车领域的研究热点、作者分布及其相互关系。最后，通过可视化分析，对新能源汽车领域未来的研究方向进行了展望。

关键词：文本挖掘；新能源汽车；小批量 K-Means 聚类；词嵌入；T-SNE 降维

Application of text mining in the field of new energy vehicles

ZHANG Yu, HUANG Runcai
（School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China）

【Abstract】 The application of text mining in the field of new energy vehicles can review the development of new energy vehicles and predict the development trend and research hotspots in the future. For this reason, this study obtains 16 293 papers (2011～2020) from CNKI (China National Knowledge Infrastructure) and 10 328 papers (2012～2020) from WOS (Web of Science) database, and uses text mining algorithms to study these papers, including word embedding, dimensionality reduction by using T-SNE (T-distributed Stochastic Neighbor Embedding), clustering of mini batch K-means and so on. Finally, the hot research trends, author distribution and interrelationship in the field of new energy vehicles at home and abroad are obtained. In the end, based on the visual analysis on the mining results, this study also predicts the future research direction in the field of new energy vehicles.

【Key words】 text mining; new energy vehicles; mini batch K-means; glove; T-SNE

0 引 言

文本挖掘是一种整合信息的工具，能够有效地提取文本中有用、创新、易懂和有价值的元素。用户可以自由访问科学研究、新闻资讯、商业信息、娱乐报道等各种类型的信息。这些信息构成了一个被广泛使用的异构性和开放性数据库，而在这个数据库中存放的是非结构化的文本数据。在人工智能的发展过程中，自然语言处理和计算机科学被整合到一起，从此网络挖掘和文本挖掘诞生了。

网页中包含很多类型的数据，如文本、链接和用户访问等，因此网络挖掘也有多种类型，例如文本挖掘、数据挖掘和图像挖掘。文本挖掘注重于把大量文本信息处理成可被人使用的信息。

在新能源汽车领域中，文本挖掘被用于分析中国新能源汽车产业政策，消费者评价新能源汽车时也使用了文本挖掘技术，在新能源汽车故障诊断中也会使用文本挖掘技术。

文本挖掘技术运用在新能源汽车领域的同时，也带来了挑战：

（1）数据来源多样化。新能源汽车领域数据发布主体主要有个人、企业、媒体、政府机构等，具体表现形式也有很多，如社交平台（推特、微博、论坛等）、研究论文、公司企业年报、季报，政府机构定期或不定期发布的各类信息等。

（2）数据信息展示。数据体量呈现了几何式增长，使用文本挖掘技术从海量数据中挖掘信息，还需要将数据信息展现出来。

本文采用文本挖掘技术，如中文中的分词、词向量、降维、聚类、数据可视化等方法对 CNKI 和 Web

作者简介：张 雨（1997-），女，硕士研究生，主要研究方向：人工智能、数据挖掘；黄润才（1966-），男，博士，副教授，主要研究方向：计算机网络与信息安全、智能计算、大数据等。
通讯作者：黄润才 Email：hrc@ sues.edu.cn

收稿日期：2022-06-12

Of Science 中的论文进行分析，展示新能源汽车领域的研究趋势及发展。

本文主要进行了以下几个方面的研究："数据与方法"部分描述了从中国知网（CNKI）获得的16 293篇文章的数据源，和从 Web Of Science 获取的10 328篇文章，并介绍了本文所使用的文本挖掘方法；"结果与讨论"部分通过主题河流图展现了论文中关键词随年份的演变，展现了新能源汽车领域的研究热点变化趋势；通过中国地图展示了研究新能源汽车的作者机构空间分布图，使用关系图来揭示 CNKI 中论文作者的关系；通过聚类算法和数据可视化揭示论文研究点的分布情况并给出相应的预测；讨论了本文所使用文本挖掘技术的局限性。本研究的潜在贡献体现在对新能源汽车领域的回顾和预测，有助于研究人员了解新能源汽车领域的研究趋势和研究热点。

1　数据与方法

1.1　数据获取

本文分析的所有数据均来自 CNKI 及 Web Of Science，搜索条件如下：

（1）在 CNKI 中以"new energy vehicle"为关键词获取相关硕博论文、以及发表在学术及行业期刊上的论文；

（2）在 Web Of Science 中同样以"new energy vehicle"为关键词，获取相关的会议或期刊论文。

1.2　数据构成

在 CNKI 中，获取的数据包含标题、作者、摘要、关键词、作者机构等信息；在 Web Of Science 中，获取的数据包含标题、作者、摘要等信息。由于部分论文存在缺失信息的情况，本文对所收集的数据进行了初步筛选，剔除了一部分不满足条件的数据，最终所获得 CNKI 论文16 293篇，Web Of Science 论文10 328篇。

1.3　中文分词

中文分词（Chinese Word Segmentation）就是将一句通顺的汉字序列根据特有规范分割为多个独立的词序列[1]。目前的分词方法可以归纳为 3 个类别：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法[2]。

基于统计的中文分词方法已然占据了主流位置，该方法是在已有大量被分词过的文本的基础上，使用统计机器学习模型来学习词语切分的规律（称为训练），以此实现对未知文本的切分。

在实际的应用中，基于统计的分词系统都需要使用分词词典来进行字符串匹配分词，同时使用统计方法识别一些新词，即将字符串频率统计和字符串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。在本文中使用的 jieba（结巴）分词方法。

1.4　词嵌入

本文使用 GloVe（Global Vectors for Word Representation）生成词向量。其是一个基于全局词频统计（Count-Based and Overall Statistics）的词表征（Word Representation）工具[3]。

GloVe 的构建过程：

（1）根据语料库构建一个共现矩阵，元素 Z_{ij} 表示在矩阵中任意单词 i 和其上下文单词 j 在规定范围内的上下文窗口中共同出现的次数；

（2）构建词向量（Word Vector）和共现矩阵之间的近似关系，其目标函数为式（1）：

$$J = \sum_{i,j=1}^{v} f(X_{ij}) \left(\boldsymbol{\omega}_j^{\mathrm{T}} \widetilde{\boldsymbol{\omega}}_j + b_i + b_j - \log X_{ij}\right)^2 \quad (1)$$

其中，$\boldsymbol{\omega}_i^{\mathrm{T}}$ 和 $\widetilde{\boldsymbol{\omega}}_j$ 是最终要求解的词向量，b_i 和 $\widetilde{b}_j$ 分别是两个词向量的偏置项。

这个损失函数的基本形式就是最简单的均方误差损失函数，只不过在此基础上加了一个分段权重函数 $f(X_{ij})$，式（2）：

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha} & \text{如果 } x < x_{\max} \\ 1 & \text{否则} \end{cases} \quad (2)$$

其中，x 为 X_{ij}，$x_{\max}$ 达到最大值时 x 的取值，当 x 小于 $x_{\max}$ 时为一个非递减函数，达到一定程度后取值不再增加。

从损失函数出发，只需要找到两个值，其中一个代表词向量，另外一个代表其真实标签，就可以借助平方误差损失函数让初始值与最终值越来越接近，最后得到词向量。

1.5　降维

T-SNE（T-Distributed Stochastic Neighbor Embedding）是用于降维的一种机器学习算法，由 Laurens van der Maaten 等在 2012 年提出[4]。T-SNE 是一种非线性降维算法，常用于高维数据降维到 2 维或者 3 维，以便进行可视化。该算法具有有效性，越相似的数据点，t 分布在低维空间中聚合更紧密；而对于不相似的数据点，t 分布在低维空间中的距离则需要远一点。

T-SNE 的梯度更新有两大优势：

(1)对于不同簇之间的点，可以利用短距离带来的大梯度使这些点互相疏远；

(2) 这种互相疏远不会变的无穷远(梯度中分母)，以避免不同簇的点过分疏远。

1.6 聚类

K-Means 算法是一种常用的聚类算法，但其算法本身存在一定的问题，如在大数据量下的计算时间过长等[5]。因此，一种基于 K-Means 的变种聚类算法 Mini Batch K-Means 应运而生。

Mini Batch K-Means 既可以利用小批量的数据子集大幅度缩短计算时长，又可以优化目标函数。所谓的小批量是指每次训练算法时随机抽取数据子集进行训练，大大缩短了计算时长，与此同时还可以保持聚类的准确性，此算法的优势是减少了 K 均值的收敛时间。

该算法的迭代步骤有两步：

(1)首先从数据集中随机选取部分数据，分配给距离最近的聚簇中心点；

(2)通过计算平均值来更新聚簇的中心点值，并把数据分配给这个聚簇中心点值，迭代次数越多，聚簇中心点值变化越小，直到中心点趋于稳定或者达到迭代次数，才停止计算。

2 结果与讨论

2.1 发文量分析

将获取到的论文数量信息做可视化处理，得到的结果如图 1、图 2 和表 1 中所示。无论是在 WOS 数据库中还是在 CNKI 中，对新能源汽车领域的研究都呈现了一个上升的趋势。从表 2 的增速可以看出，最近五年与 2011～2015 年相比，分别增长了 105.65%和 137.54%，且都在 2019 年达到了各自的峰值，分别为1 570 条和3 496 条。在 WOS 数据库中，2012 年出现了一个猛增的趋势，接着就是逐年增长；而在 CNKI 中，2017 年以前都是逐步增长，而到了 2018 年反而下降，这说明在 2018 年中国国内汽车市场低迷，呈现出了负增长的趋势，新能源汽车领域受到了影响。

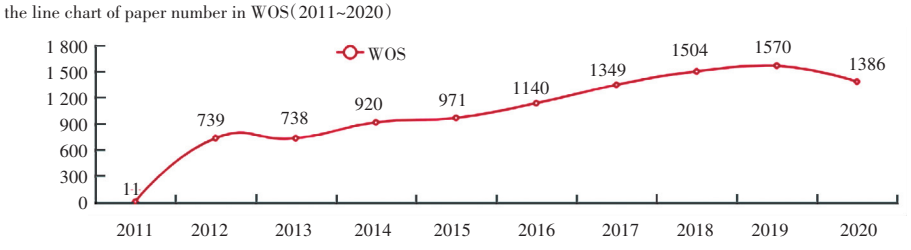

图 1 WOS 中论文数量折线图(2011～2020)

Fig. 1 Line chart of the number of papers in WOS (2011～2020)

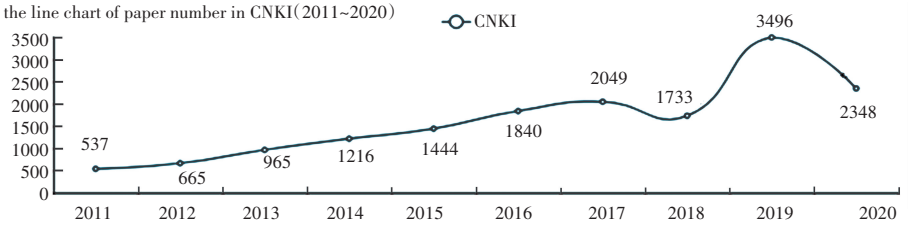

图 2 CNKI 中论文数量折线图(2011～2020)

Fig. 2 Line chart of the number of papers in CNKI (2011～2020)

2.2 研究热点变化趋

由于在 CNKI 中，2015～2020 年间新能源汽车领域的论文数量波动较大，本文选取了 2015～2020 年间 CNKI 论文，提取论文的关键词见表 2，利用中文分词的方法，根据词频进行分析，得到研究热点的变化趋势如图 3 所示。

从表 2 和图 3 可以发现，“新能源汽车”、“电动汽车”作为新能源汽车领域的主要特征词，在 2015～2020 年间的每一年都保持了一个极高的出现频率。而“发展战略”及“战略性新兴产业”则呈现了一个出现频率递减的趋势，这与中国推广新能源汽车政策有关，2015 年中国正处于新能源汽车发展的第二阶段。关键词 “锂离子电池”、“动力电池”、“永磁同步电机”的出现频率表现出了增长的趋势，年均增长率分别达到 25.55%、20.11%和 16.72%，说明在 CNKI 中与新能源汽车的动力电池有关的研究中，永磁同步电机及锂离子电池逐渐成为了研究热点。

表 1　在 2011~2015 年的论文数量

Tab. 1　Number of papers in 2011~2015

年份	WOS	CNKI
2011	11	537
2012	739	665
2013	738	965
2014	920	1 216
2015	971	1 444
2016	1 140	1 840
2017	1 349	2 049
2018	1 504	1 733
2019	1 570	3 496
2020	1 386	2 348
总数(2011~2015)	3 379	4 827
总数(2016~2020)	6 949	11 466
增长率	1.056 5	1.375 4

表 2　特征词在论文中被提及的频率(2011~2015)

Tab. 2　The frequency of feature words mentioned in papers (2011~2015)

特征词	2015	2016	2017	2018	2019	2020	年均增长率/%
新能源汽车	180	219	300	165	507	401	17.37
电动汽车	156	189	216	187	267	177	2.56
纯电动汽车	44	59	61	48	92	82	13.26
锂电子电池	25	54	47	57	116	78	25.55
战略性新兴产业	84	70	66	28	60	25	-21.52
新能源	40	51	58	30	81	54	6.19
动力电池	22	39	44	25	99	55	20.11
永磁同步电机	18	25	31	37	61	39	16.72
控制策略	29	37	32	34	44	22	-5.38
发展战略	24	19	20	34	38	26	1.61

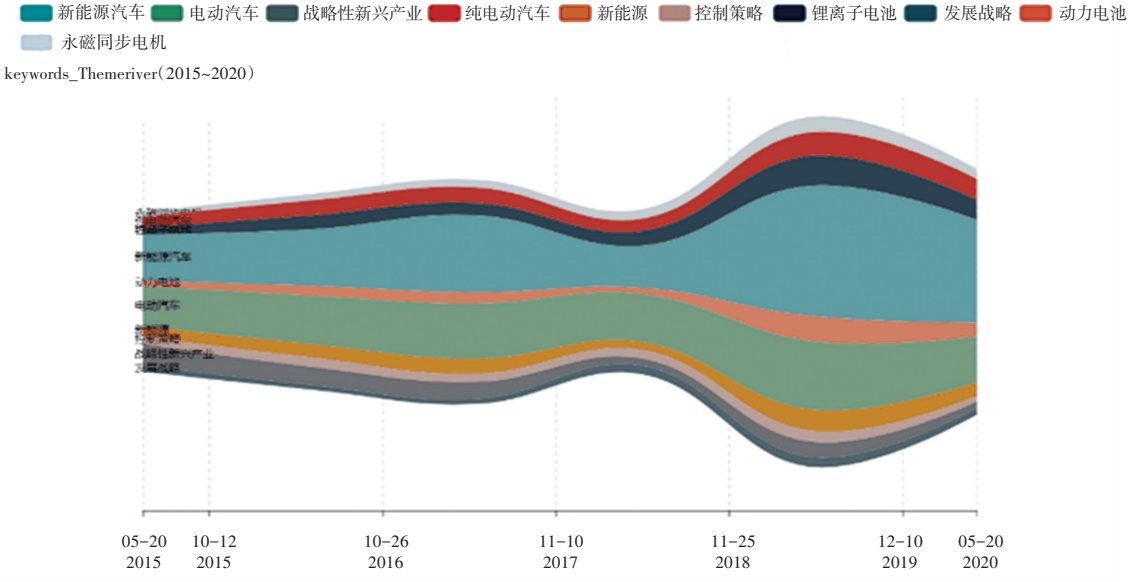

图 3　CNKI 中关键词主题河流图(2015~2020)

Fig. 3　River diagram of keyword theme in CNKI (2015~2020)

2.3　作者空间分布及关系

将 CNKI 中的文章分为期刊论文和硕博论文。对于期刊论文,筛选出在新能源汽车领域文章数量前十的期刊,并以玫瑰图的形式展现出来。本文使用的英语处理工具(Kadriu 2013),在中文摘要中使用 jieba 分词,实验结果如图 4 所示,可以直观的看出相关的期刊都是与汽车相关的,其次便是与电源有关,这与大多新能源汽车是以电池作为动力源有关。对于硕博论文,则以作者所在的单位进行研究,用同样的实验方法根据其所在省份得到中国对新能源汽车领域研究的空间分布,显示各个省份对新能源汽车都有研究,其中北京、上海、重庆、天津研究人数较多。

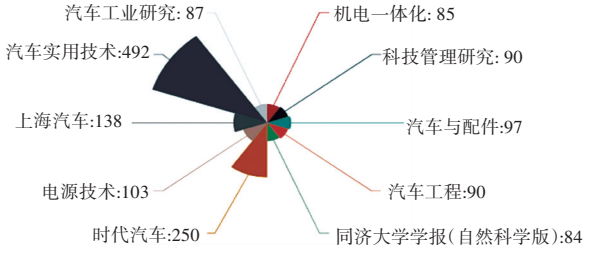

图 4　CNKI 新能源汽车领域的文章数量 top10 期刊

Fig. 4　The top-10 journals in CNKI with most articles in the field of new energy vehicles

为了得到 WOS 数据库和 CNKI 中新能源汽车领域论文作者的关系图,本文对所收集的数据进行预处理,选取了 2019 年两大数据库的作者信息,分别得到了各自的作者关系图如图 5 和图 6 所示,可以得到在 CNKI 中由于论文多数来自于硕博论文,作者关系相比 WOS 数据库中更为简单。

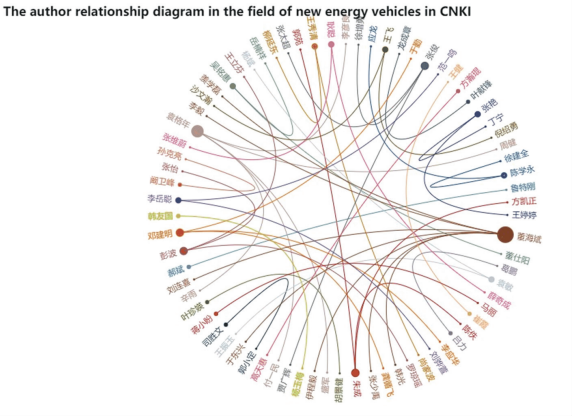

图 5 CNKI 新能源领域作者关系图

Fig. 5 Diagram of author relationship in CNKI in the field of new energy

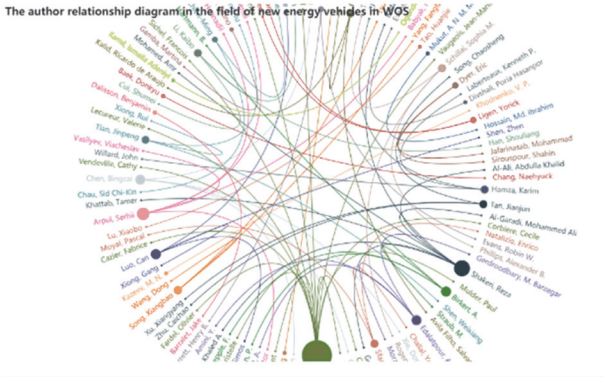

图 6 WOS 新能源领域作者关系图

Fig. 6 Diagram of author relationship in WOS in the field of new energy

2.4 论文研究热点聚类分析

本文选取了 CNKI 中获取的数据进行研究,对数据中的摘要部分进行提取,使用 jieba 算法进行中文分词,通过分词和去除停用词得到处理后的摘要数据,使用 glove 训练获得词向量。训练后得到的词向量为 200 维,词向量维度过高会导致维度爆炸,因此采用了 T-SNE 算法进行降维,将数据降维至 2 维,实验结果如图 7 所示,可以看出,这些词分为 4 类。并使用小批量 K 均值聚类得到如图 8 所示的更详细的信息。

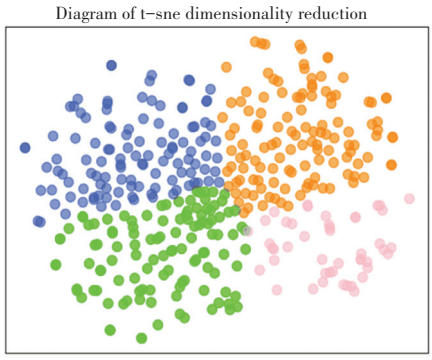

图 7 T-SNE 降维图

Fig. 7 T-SNE dimension reduction result

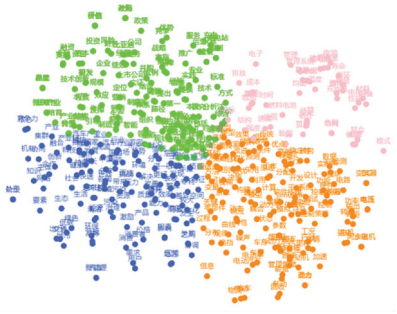

图 8 CNKI 论文研究点分布图

Fig. 8 Distribution map of research topics in CNKI

图 8 中显示为绿色的这一大类主要分布的词有“制造业”、“生产”、“营销”、“产业结构”、“资本”、“建设”等,说明在新能源汽车领域中,学者们很注重新能源汽车从制造到生产、销售整个产业的结构建设方面的研究;显示为蓝色的这一类别中,“汽车行业”、“环境污染”、“压力”、“质量”、“低碳”、“消费者”、“新能源”、“价格”等词作为主要关键词,反映了新能源汽车由于使用新能源能够缓解环境污染问题,使消费者能够低碳出行,同时价格也影响着新能源汽车行业;显示为粉色的这一类别中,分布的词数较少,主要有“燃料电池”、“成本”、“电化学”、“材料”、“电网负荷”等词,这一类别代表了对新能源汽车领域的燃料及所需成本的研究;显示为橙色的类别里,可以看到“新能源汽车”、“simulink”、“开发”、“设计”、“方案”、“电动汽车”、“车身”、“控制策略”、“发动机”等关键词,本文认为这象征着对新能源汽车进行开发设计时通常包含了车身、发动机等汽车的主要部件的研究,同时新能源汽车研究较多的是电动汽车,在设计方案时使用了 simulink 等仿真软件确定最优的策略。

最后,由于新能源汽车领域关于燃料的研究较少,所以结合图 2 所示的主题河流图,本文做出以下预测:

（1）在未来关于新能源汽车领域的研究中，可以着重关注于纯电动汽车及各种电池及永磁同步机的研究；

（2）由于各种电池如锂离子电池等所需要的电化学反应不同，所需材料的成本也不同，可以针对不同的燃料电池所需花费的成本及大规模投放后对电网所产生的负荷影响进行研究。

2.5　文本挖掘技术的不足

通过本文所介绍的文本挖掘算法及数据可视化方法，虽然已经得到了 2011～2020 十年间 WOS 数据库和 CNKI 中的数据信息，但仍有不足：

（1）T-SNE 倾向于保存局部特征，没有唯一最优解，而且在 T-SNE 中距离本身没有意义，都是概率分布问题；

（2）Mini Batch K-Means 为了减少数据规模，随机从整体选取出一小部分数据代替整体，虽然算法收敛速度大大加快，但是代价是聚类的精确度相比标准算法会有一些降低。

3　结束语

本文使用文本挖掘算法，如 jieba 分词、glove 词向量、T-SNE 降维、Mini Batch K-Means 聚类算法，研究了 WOS 数据库和 CNKI 中以新能源汽车为主题的论文，通过主题河流图、作者关系图、作者空间分布图、研究点分布图直观的展示了文本挖掘所得到的数据信息。研究结果表明在国内外对于新能源汽车领域都在持续关注，尤其是电动汽车；预测了研究人员对于新能源汽车应当在燃料电池、成本及电网负荷方面进行关注。

参考文献

[1] HUANG C, ZHAO H. Chinese word segmentation: A decade review [J]. Journal of Chinese Information Processing, 2007, 21(3): 8-20.

[2] ZHAI F, HE F, ZUO W. Chinese word segmentation based on dictionary and statistics [J]. Minimicro Systems - Shenyang -, 2006, 27(9): 1766.

[3] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[4] VAN DER MAATEN L, HINTON G. User's Guide for t-SNE Software[J]. Structure, 2008.

[5] SARMA T H, VISWANATH P, REDDY B E. Single pass kernel k-means clustering method[J]. Sadhana, 2013, 38(3): 407-419.

（上接第 83 页）

[3] AQEL D, AL-ZUBI S, MUGHAID A, et al. Extreme learning machine for plant diseases classification: a sustainable approach for smart agriculture[J]. Cluster Computing, 2022: 1-14.

[4] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [C]// Advances in Neural Information Processing Systems, 2015: 28.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[6] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[7] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS--improving object detection with one line of code[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5561-5569.

[8] ZHANG X, ZENG H, GUO S, et al. Efficient long-range attention network for image super-resolution [C]//Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVII. Cham: Springer Nature Switzerland, 2022: 649-667.

[9] PANG J, CHEN K, SHI J, et al. Libra r-cnn: Towards balanced learning for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 821-830.

[10] LIU Y, SHAO Z, HOFFMANN N. Global attention mechanism: Retain information to enhance channel-spatial interactions [J]. arXiv preprint arXiv:2112.05561, 2021.

[11] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS--improving object detection with one line of code[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5561-5569.