智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

一种三层加权文本聚类集成方法

李娜, 徐森, 徐秀芳, 许贺洋, 郭乃瑄, 刘轩绮, 周天

引用本文:
李娜, 徐森, 徐秀芳, 许贺洋, 郭乃瑄, 刘轩绮, 周天. 一种三层加权文本聚类集成方法[J]. 智能系统学报, 2024, 19(4): 807-816.
LI Na, XU Sen, XU Xiufang, et al. A three-level weighted approach for text clustering ensemble[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(4): 807-816.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202303029

DOI: 10.11992/tis.202303029
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20240313.1456.005

一种三层加权文本聚类集成方法

李娜[1,2], 徐森[1], 徐秀芳[1], 许贺洋[1], 郭乃瑄[1,2], 刘轩绮[1], 周天[3]
（1. 盐城工学院 信息工程学院, 江苏 盐城 224051; 2. 东南大学 计算机网络和信息集成教育部重点实验室, 江苏 南京 211189; 3. 哈尔滨工程大学 水声工程学院, 黑龙江 哈尔滨 150001）

摘　要: 为了提高聚类集成效果, 本文设计了一种对点、簇、划分进行加权的统一框架, 提出一种三层加权文本聚类集成方法。首先根据基聚类生成超图邻接矩阵, 然后依次对点、簇、划分进行加权获得加权邻接矩阵, 最后用层次凝聚聚类算法获得最终结果。在多个真实文本数据集上进行实验, 结果表明, 与未加权及其他层面加权相比, 三层加权方法可以获得更好的聚类效果, 三层加权相较于未加权的平均提升幅度为 12.02%; 与近年来的其他 8 种加权方法相比, 该方法在所有数据集上的平均排名位列第一, 验证了本文方法的有效性。

关键词: 文本聚类; 聚类集成; 加权聚类集成; 三层加权; 加权聚类; 多层加权; 聚类分析; 无监督学习
中图分类号: TP181;TP301　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)04−0807−10

中文引用格式: 李娜, 徐森, 徐秀芳, 等. 一种三层加权文本聚类集成方法 [J]. 智能系统学报, 2024, 19(4): 807–816.
英文引用格式: LI Na, XU Sen, XU Xiufang, et al. A three-level weighted approach for text clustering ensemble[J]. CAAI transactions on intelligent systems, 2024, 19(4): 807–816.

A three-level weighted approach for text clustering ensemble

LI Na[1,2], XU Sen[1], XU Xiufang[1], XU Heyang[1], GUO Naixuan[1,2], LIU Xuanqi[1], ZHOU Tian[3]
(1. School of Information Engineering, Yancheng Institute of Technology, Yancheng 224051, China; 2. Key Laboratory of Computer Network and Information Integration, Southeast University, Nanjing 211189, China; 3. School of Underwater Acoustic Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: To improve the clustering ensemble effect, this paper designs a unified framework for weighted points, clusters and partitions, and proposes a three-level weighted approach for text clustering ensemble. Firstly, the hypergraph adjacency matrix is generated according to the base clustering, and then the weighted adjacency matrix is obtained by successively weighting the points, clusters and partitions. Finally, the final result is obtained by the hierarchical condensation clustering algorithm. Experiments were carried out on multiple real text datasets. The results show that compared with the unweighted results and other level weighted results, this approach has better clustering effect. The average increase of three-layer weighted compared with that unweighted is 12.02%. Compared with the other 8 weighted methods in recent years, the average ranking of this algorithm is the first in all datasets, which verifies the effectiveness of the proposed method.

Keywords: text clustering; clustering ensemble; weighted clustering ensemble; three-level weighting; weighted clustering; multi-level weighting; cluster analysis; unsupervised learning

聚类分析是统计模式识别中非监督模式分类的一个重要分支, 其任务是把未标记的样本集按某种准则划分成若干子集/类/簇, 要求相似的样本尽可能地归于同一个簇, 而不相似的样本被归于不同的簇[1-5]。聚类被广泛应用于各个领域, 包括文本挖掘、分类、文档检索和图像分割等。

收稿日期: 2023−03−20.　**网络出版日期:** 2024−03−15.
基金项目: 国家自然科学基金项目(62076215); 江苏省高等学校自然科学研究面上项目(21KJD520006); 未来网络科研基金项目(FNSRFP-2021-YB-46); 盐城工学院研究生培养创新工程项目(SJCX21_XZ018); 教育部产学研合作协同育人计划项目(202102594034); 中央高校基本科研业务费专项(K93-9-2022-03); 江苏高校“青蓝工程”项目.
通信作者: 徐森. E-mail: xusen@ycit.cn.

给定一个数据集，不同的聚类算法即使是相同的算法在不同的初始化或参数下，也可能会产生不同的聚类结果，从而呈现数据的不同视角。为了融合多种聚类结果，文献 [6] 首次提出了聚类集成的概念，并提出了基于簇的相似度划分算法（cluster-based similarity partitioning algorithm, CSPA）、超图划分算法（hypergraph partitioning algorithm, HGPA）和元聚类算法（meta-clustering algorithm, MCLA）3 种聚类集成方法。聚类集成提出后，关于文本聚类集成的研究愈来愈多。例如，文献 [7] 提出了一种新的多视图文本聚类集成方法，该方法首先基于不同的文本表示模型生成不同的视图，然后对每个视图应用不同的聚类算法以获得不同的分区，最后对这些分区进行集成来获得最终的聚类结果。文献 [8] 用聚类集成方法来进行基于关键词的学术文本研究，验证了聚类集成方法相较于单一聚类方法的优越性。文献 [9] 用聚类集成算法来检测多作者文档中写作风格的变化。众多学者的研究表明聚类集成可以将多个基聚类结合在一起，降低单一聚类算法的局限性，从而获得更准确、更稳健的聚类结果[10-18]。

在聚类集成中，每个聚类结果称为一个聚类成员/基聚类/划分。低质量（甚至是病态）的基聚类会影响最终结果，从而降低聚类精度。为了避免低质量的聚类成员带来的不良影响，一些学者对基聚类进行评估和加权以提高共识性能[19-22]。对划分进行加权的方法通过分配不同的权重来控制每个聚类器的相对贡献，这意味着具有更好性能的聚类器可以分配更高的权重，而表现不佳的聚类器可以分配较低的权重。因此可以更好地利用每个聚类器的性能和相对优势，从而提高整体聚类性能。现实生活中，由于实际数据集的噪声和其固有的复杂性，同一个基聚类中的不同簇可能具有不同的质量。通过对簇进行加权，可以更好地利用每个簇之间的差异和相似性，从而减少噪声的影响，提高聚类结果的质量[23-31]。例如一些簇可能包含更多的核心样本，而另一些簇可能只包含少量的样本或者是噪声点。如果在未加权的聚类集成中，所有簇都被视为同等重要，那么这些差异可能被忽略，从而导致聚类结果的质量较低。对簇加权可以让包含更多核心样本的簇对聚类结果产生更大的贡献，而较小或者噪声簇的贡献则相应减少，从而提升整体的聚类性能。近期研究也表明点在不同的划分中会改变它的邻域，不同的点具有不同的关系稳定性，即点对底层数据结构的检测可能有不同的贡献[32-33]。对点进行加权可以更好地利用每个样本点之间的差异和相似性，让一些关键的样本点对聚类结果产生更大的贡献，而一些噪声点或者不太重要的样本点的贡献则相应减少，从而提升聚类的准确度。划分由一个或多个簇构成，簇由一个或多个点构成，只有同时考虑点、簇、划分三者的重要性，才能进一步提升聚类效果。

目前还缺少对不同研究对象（点、簇、划分）进行加权的统一框架，以进一步提升文本聚类集成的准确性。针对上述问题，本文提出一种三层加权文本聚类集成方法（three-level weighted approach for text clustering ensemble, TLWA），该方法针对文本数据集的特点设计权重，并通过超图邻接矩阵实现对点、簇、划分的三层加权。在多个文本数据集上进行了大量实验，与其他加权聚类集成方法相比，TLWA 获得了更加优越的聚类结果。

1　加权聚类集成相关工作

根据加权对象的不同可以将加权聚类集成的研究分为划分加权、簇加权及点加权 3 部分。

在划分加权方面，为了避免低质量基聚类的影响，学者们进行了一些研究，其中较为认可的一个思路是设计评价标准来评价基聚类的质量，并在集成过程中利用该评价指标对不同质量的基聚类进行加权以提高共识结果。其中，聚类成员的选择通过去除质量较差的基聚类，从而保留质量较高的聚类成员，是一种特殊的划分加权方法。文献 [19] 提出了一种改进的自适应聚类集成选择方法，兼顾了聚类成员的多样性与聚类整体的稳定性，在多个文本数据集上验证了其有效性。文献 [20] 使用标准化互信息（normalized mutual information, NMI）来衡量划分之间的相似度，并以此作为划分的权重，最后层次聚类进行集成。文献 [21] 通过利用信息熵计算类与类之间的相似性，并以此作为权重对基聚类加权。文献 [22] 提出了一种基于卷积神经网络的短文本聚类集成方法，实验基尼系数来度量基聚类的可靠性，并对其加权，最后使用层次聚类进行集成。

在簇加权方面，文献 [23] 提出了 3 种基于簇的加权聚类集成方法，通过对几个真实数据集（包括文本数据集）的实验验证了其有效性。文献 [24] 利用集合的链路网络模型估计出簇之间的相似性，并以此作为权重提出了 3 种新的基于链接的相似度评估算法。文献 [25] 通过计算每个簇在所有基划分下的不确定性构造出集成驱动聚类指标（ensemble-driven cluster index, ECI），并将此

作为权重对共协矩阵(co-association matrix, CA 矩阵)进行加权, 然后集成。文献 [26] 将点到簇中心的距离与簇内最大距离的比值作为簇的权重, 得到加权 CA 矩阵, 最后运用 K-means 得到最终结果。文献 [27] 通过评估簇与划分之间的集合匹配度来计算簇与划分之间的相似度, 并以相似度作为簇的权重, 随后根据多样性来选择基聚类, 该方法同时考虑了簇和划分的质量, 在包含文本在内的多个数据集上验证了其有效性。文献 [28] 通过信息论评估了簇的不可靠性, 提出了加权证据积累和加权图划分 2 种聚类方法。文献 [29] 通过熵和指数变换得到每个簇的可靠性, 并以此提出了 2 种簇权值计算方法, 最后对加权 CA 矩阵运用组平均(average link, AL)得到一致划分。文献 [30] 结合信息熵的概念和 Jaccard 系数提出一种衡量簇稳定性的评价标准, 并根据该指标对簇层面进行加权; 另一种是基于熵准则的文本聚类集成方法, 熵准则用于评估簇的不确定性, 根据簇的不确定性提出了 2 个指标, 进而选择高质量的基聚类进行集成。

在点加权方面, 文献 [31] 通过计算样本之间的距离来衡量样本之间的相似度, 进而评价一个类的可靠度。文献 [32] 通过计算样本的稳定性进而确定簇中心, 并将样本点分配到与其相似性最高的簇内, 最后用单链接(single link, SL)算法进行集成, 在包含文本数据集的多个数据集上验证了其有效性。文献 [33] 首先通过基聚类结果得到一个 CA 矩阵, 然后使用 CA 矩阵去描述每个样本的聚类困难程度, 并为其赋予相应的权重。

尽管人们已经从各个方面证明了加权的重要性, 但是大多数学者只是对聚类集成过程中的某个方面进行加权, 比如点加权或簇加权, 目前还缺少一种结合点、簇、划分 3 个层面的统一框架。

2　本文方法

设计一个三层加权文本聚类集成方法需要解决 2 个主要问题: 1)如何针对文本数据的特点, 设计点、簇、划分 3 个层面的权值, 2)如何构造一个三层加权框架来融合 3 个层面的权值。针对第 1 个问题, 本文根据文本数据的特点, 并结合多位学者的研究, 提出了针对点、簇及划分 3 个层面的权值设置方案, 详见 2.1 节。针对第 2 个问题, 本文通过超图邻接矩阵 $\boldsymbol{H}$ 来构造三层加权框架并进行权值的传递。对于超图邻接矩阵 $\boldsymbol{H}$ 而言, 其每一行代表一个点, 每一列代表一个簇, 具体权值逐层传递方案如 2.2 节所示。最后将加权后的 $\boldsymbol{H}$ 矩阵转化为加权 CA 矩阵, 对加权 CA 矩阵使用组平均法进行集成以得到最终的共识结果。下面依次介绍权值设置方案、权值逐层传递方案、加权 CA 矩阵的生成、算法流程及复杂度分析。

2.1　权值设置方案

2.1.1　点层权值设置

2 个文本向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_m$ 的相似度可采用余弦函数求解, 余弦函数相较于其他距离函数而言更适合于文本数据[34-36], 余弦相似度的计算公式为

$$S(\boldsymbol{x}_i,\boldsymbol{x}_j)=\cos(\theta(\boldsymbol{x}_i,\boldsymbol{x}_j))=(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)\Big/(\|\boldsymbol{x}_i\|\cdot\|\boldsymbol{x}_j\|)=\boldsymbol{x}_i\boldsymbol{x}_j^{\mathrm{T}}$$

式中: $1\leqslant i\leqslant n$, $1\leqslant j\leqslant n$, n 为样本总数, $\boldsymbol{x}_j^{\mathrm{T}}$为 $\boldsymbol{x}_j$ 的转置矩阵。

显然, 文本 $\boldsymbol{x}_i$ 与其他文本的相似度越高, 其权值越大; 反之, 越小。因此, 本文首先设置文本 $\boldsymbol{x}_i$ 的权值 w_i 与 $\boldsymbol{x}_i$ 和其他文本的相似度之和成正比。然而, 对于类别不平衡的文本集, 在相似度相差不大的情况下, 基数大的簇中点的权值显然高于基数小的簇中点的权值。为了消除对簇大小的偏置, 本文进一步设置文本 $\boldsymbol{x}_i$ 的权值 w_i 与 $\boldsymbol{x}_i$ 在聚类集体中所属的所有簇的基数之和成反比。即点的权值为

$$w_i'=\left(\sum_{j=1}^{n}S(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)\Big/\sum_{m=1}^{M}\delta_{\mathrm{ism}}\left|C_s^m\right|$$

式中: C_s^m为第 m 个划分中的第 s 个簇, $1\leqslant m\leqslant M$, $1\leqslant s\leqslant k_m$; M 为划分的个数; k_m 为第 m 个划分中簇的个数; δ_{ism} 为点的隶属度函数; 若 $\boldsymbol{x}_i\in C_s{}^m$, 则 δ_{ism} 为 1, 否则为 0。

点层权值归一化得

$$w_i=w_i'\Big/\sum_{i=1}^{n}w_i' \tag{1}$$

2.1.2　簇层权值设置

由于簇由点构成, 本文设置簇的权值与其包含的所有点的权值之和成正比。即簇的权值为

$$u_s^{m'}=\sum_{i=1}^{n}\delta_{\mathrm{ism}}w_i$$

式中 δ_{ism} 为簇的隶属度函数, 若 $x_i\in C_s{}^m$, 则 δ_{ism} 为 1, 否则为 0。

簇层权值归一化得

$$u_s^m=u_s^{m'}\Big/\sum_{s=1}^{k_m}u_s^{m'} \tag{2}$$

2.1.3　划分层权值设置

NMI 值可以有效衡量划分之间的相似程度, 显然, 划分 $P^{(m)}$ 与其他划分的相似度越高, 其权值越大; 反之越小。因此, 本文设置划分 $P^{(m)}$ 的权值 v_m 与 $P^{(m)}$ 和其他划分的 NMI 值之和成正比。

考虑到划分由簇构成，本文进一步设置划分 $P^{(m)}$ 的权值 v_m 与 $P^{(m)}$ 包含的所有簇的权值之和成正比，即划分的权值为

$$v'_m = \sum_{q=1,q\neq m}^{M} \mathrm{NMI}(P^{(m)}, P^{(q)}) \times \sum_{s=1}^{k_m} u_s^m$$

划分层权值归一化得：

$$v_m = v'_m \Big/ \sum_{m=1}^{M} v'_m \tag{3}$$

2.2 权值逐层传递

2.2.1 对点层进行加权

对 H 的第 i 行（$1\leqslant i\leqslant n$）乘以点 x_i 的权值 w_i，得到点层加权矩阵：

$$\boldsymbol{H}_{pt} = \boldsymbol{W} \times \boldsymbol{H}$$

式中：$\boldsymbol{W}$=diag($w_1, w_2, \cdots, w_n$)，w_i 的计算如式 (1) 所示。

2.2.2 对簇层进行加权

对 $\boldsymbol{H}_{pt}$ 中第 m 个划分中的第 s 个簇 C_s^m 对应的列乘以其权值 u_s^m（$1\leqslant m\leqslant M$, $1\leqslant s\leqslant k_m$），得到簇层加权矩阵：

$$\boldsymbol{H}_{cr} = \boldsymbol{H}_{pt} \times \boldsymbol{U}$$

式中：$\boldsymbol{U} = \mathrm{diag}\left(u_1^1, u_2^1, \cdots, u_{k_1}^1, \cdots, u_1^m, u_2^m, \cdots, u_{k_m}^m, \cdots, u_1^M, u_2^M, \cdots, u_{k_M}^M\right)$，$u_s^m$ 的计算如式 (2) 所示。

2.2.3 对划分层进行加权

对 $\boldsymbol{H}_{cr}$ 中第 m 个划分 $P^{(m)}$ 对应的子矩阵乘以其权值 v_m，得到划分层加权矩阵：

$$\boldsymbol{H}_{pn} = \boldsymbol{H}_{cr} \times \boldsymbol{V}$$

式中：$\boldsymbol{V}$=diag($v_1, \cdots, v_1, v_m, \cdots, v_m, v_M, \cdots, v_M$)，$\boldsymbol{V}$ 中共有 k_1 个 v_1，k_m 个 v_m，k_M 个 v_M，v_m 的计算如式 (3) 所示。

2.3 生成加权 CA 矩阵

得到 $\boldsymbol{H}_{pn}$ 后，将其转化成加权共协（weighted co-association, WCA）矩阵：

$$\boldsymbol{W}_{CA} = \frac{\boldsymbol{H}_{pn} \times \boldsymbol{H}_{pn}^{\mathrm{T}}}{M} \tag{4}$$

式中 $\boldsymbol{H}_{pn}^{\mathrm{T}}$ 为 $\boldsymbol{H}_{pn}$ 的转置矩阵。

2.4 算法流程

本文设计的 TLWA 算法主要步骤如下：

算法 1 TLWA

输入 经过预处理的文本数据集 X={$x_1, x_2, \cdots, x_n$}，样本点的真实标签及真实类别数 k^*。

1）生成 M 个聚类；

for i = 1:M

运行基于余弦相似度的 K-means 算法，簇个数 $k\in[k^*, 2k^*]$

end

2）根据基聚类生成超图邻接矩阵 $\boldsymbol{H}$；

3）根据式 (1) 及 2.2.1 节对点加权；

4）根据式 (2) 及 2.2.2 节对簇加权；

5）根据式 (3) 及 2.2.3 节对划分加权；

6）根据式 (4) 将 $\boldsymbol{H}_{pn}$ 转化为 $\boldsymbol{W}_{\mathrm{CA}}$ 矩阵；

7）对 $\boldsymbol{W}_{\mathrm{CA}}$ 矩阵运行 AL 算法得到一致划分，簇的个数设置为 k^*；

8）输出一致划分，文本数据集 X 被分成 k^* 个类簇，即 C={$C_1, C_2, \cdots, C_{k*}$}。

2.5 复杂度分析

在上述算法流程中，步骤 1) 为运行基于余弦相似度的 K-means 算法 M 次，其时间复杂度为 $O(Mkdn)$，其中，k 为簇的个数，d 为样本的维度，n 为样本数。步骤 2) 生成超图邻接矩阵 $\boldsymbol{H}$ 的时间复杂度为 $O(Mkn)$。步骤 3)~5) 对各层加权的时间复杂度均为 $O(nC_K)$，其中 C_K 为 M 个划分中簇的总个数，即超图的边数。步骤 6) 将 $\boldsymbol{H}_{pn}$ 转化为 $\boldsymbol{W}_{\mathrm{CA}}$ 矩阵的时间复杂度为 $O(C_K\times n^2)$。步骤 7) 调用 AL 算法的时间复杂度为 $O(\log_2 n)$。即本文算法的步骤 1)~5) 均为线性阶，步骤 6) 为平方阶，步骤 7) 为对数阶。另外，本算法步骤 6) 为构建相似度矩阵，该方法能够获得更好的聚类结果，但复杂度较高，即该方法适用于中小规模的数据集。对于海量文本数据集，可直接在 $\boldsymbol{H}_{pn}$ 上运行基于矩阵低秩近似的方法（matrix low rank approximation-based algorithm, MLRAA）、基于深度低秩子空间集成（deep low-rank subspace ensemble, DLRSE）和 K-means 等算法，以进一步提高运行效率。

3 三层加权聚类集成方法实验探究

3.1 实验设置

实验采用 8 组公共文本测试集，具体描述如表 1 所示。Tr11、Tr12、Tr23、La12、Hitech、Reviews 和 Sports 由文本检索大会（http://trec.nist.gov）提供；数据集 K1b 来自于 WebACE project[37]，每个文本对应于 Yahoo!主题层次下的一个网页。

表 1 数据集介绍

Table 1 Dataset introduction

数据集	样本数	特征数	类别数
Tr11	414	6249	9
Tr12	313	5804	8
Tr23	204	5832	6
Reviews	4069	18483	5
La12	6279	31472	6
Hitech	2301	10080	6
K1b	2340	21839	6
Sports	8580	14870	7

因为文本类别标签已知, 本实验采用调兰德指数(adjusted rand index, ARI)和 F 值(F-measure)这 2 个评价指标进行评价。2 个评价指标均为值越大, 聚类质量越高; 反之, 越低。

实验分为 2 部分: 1)各层效果对比, 对每一层加权的效果进行详细对比, 验证三层加权聚类集成方法相较于其他层面加权的优越性; 2)与其他加权方法进行对比, 验证三层加权方法优于目前提出的其他加权方法。在本文的实验中, 所有文本数据均已经过 TF-IDF(term frequency-inverse document frequency)加权, 基聚类的生成方法为运行使用余弦相似度的 K-means 算法 100 次。其中关于 k 值的设定方面, 多数学者为了方便将 k 设置为 k^*(k^*为数据集的真实类别数), 为了使聚类成员更加多样化, 使其多方面反应数据内部结构, 本文将簇的个数 k 设置在 $[k^*, 2k^*]$ 的范围内[19,24]。为保证公平公正, 以下所有的比较均为在同样基聚类的基础上进行。

3.2　各层加权效果对比

表 2 给出了各层加权效果对比结果。表 2 中数值均为运行 10 次取平均值, 粗体标识代表三层加权后效果相较于未加权效果有所提高, 带有下划线的数据表示在各层加权效果的对比中最优, 提升幅度的计算方法为三层加权后评价指标提升的数值占未加权的评价指标的百分比。

表 2　各层加权效果对比

Table 2　Weighted effect comparison of each layer

数据集	评价指标	未加权	点加权	簇加权	划分加权	点、簇加权	点、划分加权	簇、划分加权	三层加权	提升幅度/%
Tr11	ARI	0.600±0.029	0.522±0.026	0.626±0.027	0.606±0.031	0.652±0.077	0.521±0.041	0.636±0.022	**0.683±0.084**	13.83
	F	0.762±0.015	0.700±0.012	0.768±0.025	0.770±0.015	0.761±0.044	0.704±0.026	0.773±0.016	**0.780±0.051**	2.36
Tr12	ARI	0.454±0.042	0.461±0.059	0.435±0.047	0.448±0.033	0.556±0.081	0.447±0.056	0.474±0.061	**0.578±0.064**	22.69
	F	0.704±0.019	0.705±0.035	0.686±0.028	0.696±0.016	0.760±0.034	0.702±0.033	0.710±0.028	**0.769±0.030**	9.39
Tr23	ARI	0.259±0.009	0.268±0.025	0.320±0.021	0.262±0.009	0.328±0.010	0.260±0.028	0.305±0.047	**0.318±0.020**	22.78
	F	0.528±0.010	0.540±0.010	0.566±0.014	0.527±0.010	0.563±0.009	0.530±0.010	0.549±0.041	**0.556±0.014**	5.50
Reviews	ARI	0.565±0.021	0.653±0.004	0.619±0.049	0.566±0.020	0.661±0.003	0.653±0.003	0.616±0.047	**0.661±0.003**	16.99
	F	0.728±0.014	0.767±0.002	0.751±0.022	0.729±0.013	0.769±0.001	0.767±0.002	0.749±0.022	**0.769±0.001**	5.63
La12	ARI	0.567±0.023	0.590±0.050	0.598±0.025	0.555±0.014	0.576±0.057	0.585±0.048	0.592±0.025	**0.579±0.061**	2.12
	F	0.728±0.027	0.765±0.041	0.763±0.030	0.720±0.019	0.764±0.034	0.762±0.041	0.761±0.030	**0.765±0.035**	5.08
Hitech	ARI	0.269±0.012	0.260±0.017	0.281±0.014	0.272±0.012	0.272±0.018	0.260±0.013	0.283±0.008	**0.270±0.015**	0.37
	F	0.517±0.017	0.511±0.024	0.530±0.017	0.518±0.013	0.525±0.023	0.519±0.022	0.531±0.016	**0.530±0.020**	2.51
K1b	ARI	0.555±0.091	0.580±0.098	0.700±0.079	0.564±0.111	0.726±0.017	0.532±0.098	0.673±0.104	**0.727±0.018**	30.99
	F	0.802±0.043	0.817±0.042	0.861±0.039	0.804±0.054	0.860±0.006	0.779±0.041	0.849±0.050	**0.861±0.006**	7.36
Sports	ARI	0.476±0.042	0.654±0.066	0.603±0.093	0.484±0.069	0.651±0.016	0.609±0.103	0.618±0.083	**0.655±0.017**	37.61
	F	0.722±0.025	0.792±0.042	0.786±0.048	0.733±0.040	0.772±0.002	0.762±0.064	0.797±0.036	**0.773±0.003**	7.06

通过表 2 可以得出以下结论:

1)总体来看, 对数据进行三层加权后, 所有数据集的 2 种评价指标值均有所上升, 8 个数据集在 2 种评价指标下的平均提升幅度为 12.02%。数据集 Tr12、K1b 及 Sports 提升效果较为显著, 其中数据集 K1b 与数据集 Sports 的 ARI 评价指标提升幅度最大, 分别为 30.99% 与 37.61%。可见, 本文提出的三层加权方法是提升聚类集成效果行之有效的方法。

2)从每一层的加权效果来看, 各层加权的效果不一, 单层加权效果如何与数据集本身有很大的关系, 很难找到一种加权方法适用于所有的数据集。个别数据集甚至会出现单层加权后效果变差的情况, 如 Tr11 及 Hitech 的点加权、Tr12 的簇加权、Tr12 及 La12 的划分加权。但是经过三层加权后的聚类效果在所有数据集上的表现均优于未加权的效果, 说明三层加权方法可以在一定程度上弥补单层加权方法的不足, 从而增强聚类结

果的稳定性，进而也在一定程度上提升了三层加权方法对于数据集的普适性。

3）从三层加权方法获得第 1 名的次数来看，点加权的提升效果位列第 1 的次数为 1 次，簇加权的提升效果位列第 1 的次数为 2 次，点、簇加权的提升效果位列第 1 的次数为 2 次，簇、划分加权的提升效果位列第 1 的次数为 3 次。相比之下，三层加权的提升效果位列第 1 的次数为 10 次，次数明显多于其他层面加权位列第 1 的次数。

综上，本文提出的三层加权方法相较于其他层面的加权方法而言，可以得到更好的聚类效果。并且能够在一定程度上弥补单层加权聚类方法的不足，从而增加聚类结果的稳定性，增强对数据集的普适性，即本文提出的三层加权聚类集成算法具有明显的优越性。

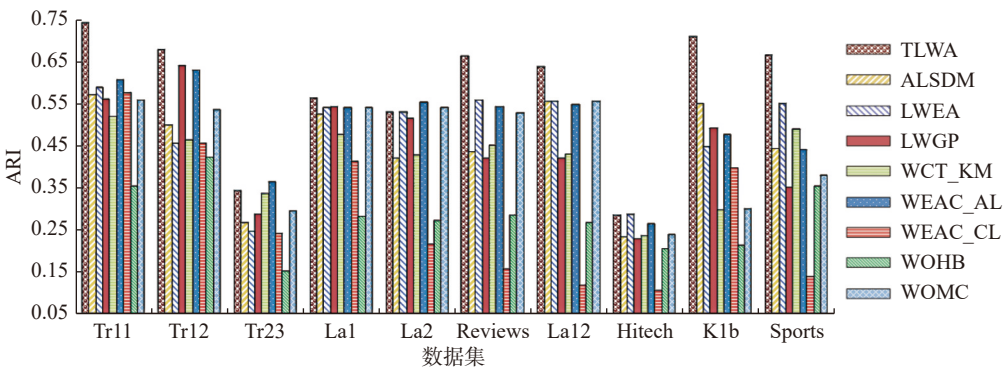
3.3　与其他加权方法进行对比

将本文方法 TLWA 与其他 8 种加权方法 ALSDM[29]、LWEA[25]、LWGP[25]、WCT_KM[24]、WEAC_AL[20]、WEAC_CL[20]、WOHB[33] 及 WOMC[33] 进行比较，结果如表 3 所示。表 3 中数据均为运行 10 次取平均值，粗体表示此结果在所有对比方法中排名第 1。另外，为了多方面展示本文结果，图 1 给出了 10 次运行结果中聚类效果较优的一次运行结果。

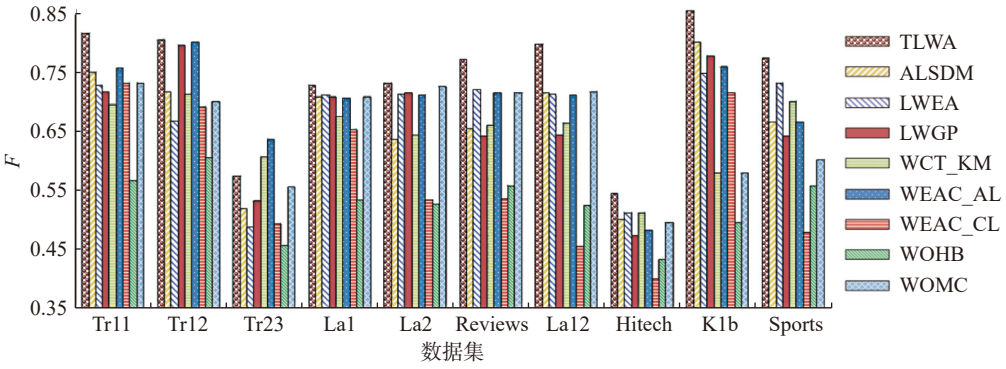
表 3　与其他加权方法对比结果 (平均值)

Table 3　Comparison results with other weighted methods (average)

数据集	评价指标	TLWA	ALSDM	LWEA	LWGP	WCT_KM	WEAC_AL	WEAC_CL	WOHB	WOMC
Tr11	ARI	**0.683±0.084**	0.600±0.029	0.584±0.017	0.574±0.028	0.500±0.050	0.619±0.025	0.583±0.061	0.408±0.048	0.507±0.036
	F	**0.780±0.051**	0.762±0.015	0.720±0.011	0.720±0.019	0.664±0.038	0.762±0.021	0.733±0.029	0.607±0.047	0.680±0.033
Tr12	ARI	**0.577±0.064**	0.454±0.042	0.496±0.042	0.572±0.045	0.391±0.060	0.493±0.092	0.451±0.066	0.460±0.027	0.466±0.029
	F	**0.769±0.030**	0.704±0.019	0.697±0.031	0.762±0.021	0.650±0.059	0.728±0.047	0.696±0.042	0.679±0.035	0.684±0.015
Tr23	ARI	0.318±0.019	0.259±0.009	**0.331±0.039**	0.272±0.017	0.228±0.063	0.320±0.059	0.269±0.041	0.130±0.030	0.261±0.036
	F	0.556±0.014	0.527±0.010	0.578±0.039	0.517±0.024	0.513±0.051	**0.584±0.068**	0.526±0.044	0.437±0.026	0.554±0.015
Reviews	ARI	**0.661±0.003**	0.459±0.051	0.565±0.019	0.502±0.064	0.462±0.108	0.565±0.021	0.318±0.094	0.239±0.057	0.514±0.012
	F	**0.769±0.001**	0.679±0.031	0.725±0.011	0.691±0.038	0.703±0.055	0.729±0.013	0.605±0.055	0.538±0.054	0.714±0.009
La12	ARI	**0.579±0.061**	0.550±0.014	0.557±0.005	0.486±0.056	0.487±0.063	0.559±0.012	0.145±0.059	0.251±0.019	0.553±0.006
	F	**0.765±0.035**	0.717±0.007	0.718±0.005	0.689±0.031	0.683±0.049	0.721±0.017	0.479±0.033	0.504±0.022	0.719±0.004
Hitech	ARI	0.270±0.015	0.255±0.022	**0.290±0.009**	0.240±0.006	0.238±0.034	0.273±0.006	0.097±0.022	0.196±0.019	0.262±0.018
	F	**0.530±0.020**	0.517±0.022	0.526±0.010	0.497±0.011	0.500±0.025	0.519±0.014	0.418±0.023	0.444±0.023	0.512±0.017
K1b	ARI	**0.727±0.018**	0.491±0.048	0.475±0.032	0.496±0.005	0.358±0.050	0.570±0.108	0.347±0.075	0.234±0.009	0.313±0.029
	F	**0.861±0.006**	0.756±0.041	0.762±0.019	0.780±0.003	0.647±0.045	0.807±0.051	0.668±0.056	0.517±0.015	0.612±0.041
Sports	ARI	**0.655±0.016**	0.390±0.049	0.639±0.083	0.445±0.065	0.320±0.107	0.458±0.066	0.211±0.080	0.328±0.032	0.406±0.034
	F	0.773±0.003	0.640±0.030	**0.792±0.051**	0.707±0.043	0.591±0.070	0.708±0.041	0.548±0.067	0.561±0.026	0.621±0.036



(a) ARI 值



(b) F 值

图 1　与其他加权算法对比结果 (最优值)

Fig. 1　Comparison results with other weighting algorithms (optimal value)

由表 3 可以看出, 除 Tr23 的 ARI 评价指标和 F 评价指标、Hitech 的 ARI 评价指标及 Sports 的 F 评价指标外, 本文提出的 TLWA 算法均处于第 1 名, 处于第 1 名的比例为 12/16。由图 1 可以看出, 本文提出的 TLWA 方法在数据集 Tr11、Reviews、La12、K1b 及数据集 Sports 上的聚类效果明显优于其他加权聚类集成方法。为了使各方法排名情况更加直观, ARI 及 F 评价指标下各加权方法 10 次运行结果的平均值排名表如表 4 及表 5 所示, 平均序值为各方法在 8 个数据集下排名的均值。

表 4　ARI 评价指标下各加权方法排名表

Table 4　Ranking of weighting methods under ARI evaluation index

数据集	TLWA	ALSDM	LWEA	LWGP	WCT_KM	WEAC_AL	WEAC_CL	WOHB	WOMC
Tr11	1	3	4	6	8	2	5	9	7
Tr12	1	7	3	2	9	4	8	6	5
Tr23	3	7	1	4	8	2	5	9	6
Reviews	1	7	2.5	5	6	2.5	8	9	4
La12	1	5	3	7	6	2	9	8	4
Hitech	3	5	1	6	7	2	9	8	4
K1b	1	4	5	3	6	2	7	9	8
Sports	1	6	2	4	8	3	9	7	5
平均序值	1.500	5.500	2.688	4.625	7.250	2.438	7.500	8.125	5.375

表 5　F 评价指标下各加权方法排名表

Table 5　Ranking of weighting methods under F evaluation index

数据集	TLWA	ALSDM	LWEA	LWGP	WCT_KM	WEAC_AL	WEAC_CL	WOHB	WOMC
Tr11	1	2.5	5.5	5.5	8	2.5	4	9	7
Tr12	1	4	5	2	9	3	6	8	7
Tr23	3	5	2	7	8	1	6	9	4
Reviews	1	7	3	6	5	2	8	9	4
La12	1	5	4	6	7	2	9	8	3
Hitech	1	4	2	7	6	3	9	8	5
K1b	1	5	4	3	7	2	6	9	8
Sports	2	5	1	4	7	3	9	8	6
平均序值	1.375	4.688	3.313	5.063	7.125	2.313	7.125	8.500	5.500

接下来, 本文通过 Friedman 检验及 Nemenyi 检验来判断本方法与其他方法是否具有显著性差异。

下面首先使用 ARI 评价指标的排名来进行算法的 Friedman 检验, 来判断这些方法的性能是否都相同。

$$\chi_F^2=\frac{12\times8}{9\times10}\left(1.500^2+5.500^2+2.688^2+4.625^2+7.250^2+2.438^2+7.500^2+8.125^2+5.375^2-\frac{9\times10^2}{4}\right)=48.830$$

$$F_F=\frac{7\times48.830}{8\times8-48.830}=22.532$$

F_F 服从自由度为 9−1=8 和 (9−1)×(8−1)=56 的 F 分布，给定 α=0.1，查表 F(8, 56) 为 2.109，小于 F_F，因此拒绝“所有算法性能相同”这个假设。

接下来，在两两比较中使用 Nemenyi 检验，9 种算法在 $q_{0.1}$ 处的临界值为 2.855，对应的 CD 为$2.855\times\sqrt{\frac{9\times10}{6\times8}}=3.909$，即算法 TLWA 与算法 ALSDM、WCT_KM、WEAC_CL、WOHB 及算法 WOMC 有显著性差异，与算法 LWEA、LWGP 及算法 WEAC_AL 有差异。同理，计算 F 评价指标的 Friedman 检验，来判断这些算法的性能是否都相同。

$$\chi_F^2=\frac{12\times8}{9\times10}\left(1.375^2+4.688^2+3.313^2+5.063^2+7.125^2+2.313^2+7.125^2+8.500^2+5.500^2-\frac{9\times10^2}{4}\right)=47.850$$

$$F_F=\frac{7\times47.850}{8\times8-47.850}=20.740$$

服从自由度为 8 和 56 的 F 分布，给定 α=0.1，查表 F(8,56) 为 2.109，小于 F_F，因此拒绝“所有算法性能相同”这个假设。

接下来，在两两比较中使用 Nemenyi 检验，9 种算法在 $q_{0.1}$ 处的临界值为 2.855，对应的 CD 为 3.909，即本文提出的算法 TLWA 与算法 WCT_KM、WEAC_CL、WOHB 及算法 WOMC 有显著性差异，与算法 ALSDM、LWEA、LWGP 及算法 WEAC_AL 有差异。

综上所述，无论是平均值比较还是在最优值比较，本文提出的 TLWA 方法总能获得较为优异的结果。并且表 4 与表 5 也表明对于不同的加权方法而言，其在不同的评价指标下的排名也不一样，但是本文提出的 TLWA 方法在 2 种评价指标下的平均序值均为第 1 名，由此可见三层加权聚类集成方法能获得更好的共识结果。

4　结束语

本文提出了一种三层加权文本聚类集成方法 TLWA，该方法针对文本数据集的特点设计点、簇、划分三层的权值，并通过超图邻接矩阵实现三层加权。基于多个数据集上的实验表明：

1）基于三层加权后的聚类效果优于未加权及其他层面加权的聚类效果；

2）与其他方法进行比较，本文提出的 TLWA 方法较为突出。

综上，本文提出的三层加权文本聚类集成方法是提升聚类性能的行之有效的方法。

参考文献：

[1] 李洁, 高新波, 焦李成. 基于特征加权的模糊聚类新算法 [J]. 电子学报, 2006, 34(1): 89–92.
LI Jie, GAO Xinbo, JIAO Licheng. A new feature weighted fuzzy clustering algorithm[J]. Acta electronica sinica, 2006, 34(1): 89–92.

[2] JIA Caiyan, CARSON M B, WANG Xiaoyang, et al. Concept decompositions for short text clustering by identifying word communities[J]. Pattern recognition, 2018, 76(4): 691–703.

[3] XIE Junyuan, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]//The 33rd International Conference on Machine Learning. New York: W&CP, 2016: 478−487.

[4] 冯冰, 李绍滋. 中医脉诊信号的无监督聚类分析研究 [J]. 智能系统学报, 2018, 13(4): 564–570.
FENG Bing, LI Shaozi. Unsupervised clustering analysis of human-pulse signal in traditional Chinese medicine[J]. CAAI transactions on intelligent systems, 2018, 13(4): 564–570.

[5] 张智, 毕晓君. 基于风格转换的无监督聚类行人重识别 [J]. 智能系统学报, 2021, 16(1): 48–56.
ZHANG Zhi, BI Xiaojun. Clustering approach based on style transfer for unsupervised person re-identification[J]. CAAI transactions on intelligent systems, 2021, 16(1): 48–56.

[6] STREHL A, GHOSH J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions[J]. Journal of machine learning research, 2002, 3(3): 583–617.

[7] FRAJ M, BEN HAJKACEM M A, ESSOUSSI N. Ensemble method for multi-view text clustering[C]//International Conference on Computational Collective Intelligence. Hendaye: Springer, 2019: 219−231.

[8] 张颖怡, 章成志, 陈果. 基于关键词的学术文本聚类集成研究 [J]. 情报学报, 2019, 38(8): 860–871.
ZHANG Yingyi, ZHANG Chengzhi, CHEN Guo. Research on clustering integration of academic texts based on keywords[J]. Journal of the China society for scientific and technical information, 2019, 38(8): 860–871.

[9] AL-SHAMASI S, MENAI M. Ensemble-based clustering for writing style change detection in multi-authored textual documents[C]//Proceedings of the Working Notes

of CLEF 2022. Bologna: CEUR Workshop Proc, 2022: 2357−2374.

[10] 张美琴, 白亮, 王俊斌. 基于加权聚类集成的标签传播算法 [J]. 智能系统学报, 2018, 13(6): 994–998.
ZHANG Meiqin, BAI Liang, WANG Junbin. Label propagation algorithm based on weighted clustering ensemble[J]. CAAI transactions on intelligent systems, 2018, 13(6): 994–998.

[11] 廖彬, 黄静莱, 王鑫, 等. SCEA: 一种适应高维海量数据的并行聚类集成算法 [J]. 电子学报, 2021, 49(6): 1077–1087.
LIAO Bin, HUANG Jinlai, WANG Xin, et al. SCEA: a parallel clustering ensemble algorithm for high-dimensional massive data[J]. Acta electronica sinica, 2021, 49(6): 1077–1087.

[12] ZHANG Mimi. Weighted clustering ensemble: a review[J]. Pattern recognition, 2022, 124: 108428.

[13] SHEN Qiaoyun, QIU Yican. A novel text ensemble clustering based on weighted entropy filtering model[J]. Journal of physics: conference series, 2021, 2024(1): 012045.

[14] NAJAFI F, PARVIN H, MIRZAIE K, et al. Dependability‐based cluster weighting in clustering ensemble[J]. Statistical analysis and data mining: the ASA data science journal, 2020, 13(2): 151–164.

[15] JI Xia, LIU Shuaishuai, ZHAO Peng, et al. Clustering ensemble based on sample's certainty[J]. Cognitive computation, 2021, 13(3): 1034–1046.

[16] WU Junjie, LIU Hongfu, XIONG Hui, et al. K-means-based consensus clustering: a unified view[J]. IEEE transactions on knowledge and data engineering, 2015, 27(1): 155–169.

[17] TAO Zhiqiang, LIU Hongfu, FU Yun. Simultaneous clustering and ensemble[C]//The 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017: 1546−1552.

[18] ZHONG Caiming, HU Lianyu, YUE Xiaodong, et al. Ensemble clustering based on evidence extracted from the co-association matrix[J]. Pattern recognition, 2019, 92(8): 93–106.

[19] 徐森, 皋军, 花小朋, 等. 一种改进的自适应聚类集成选择方法 [J]. 自动化学报, 2018, 44(11): 2103–2112.
XU Sen, GAO Jun, HUA Xiaopeng, et al. An improved adaptive cluster ensemble selection approach[J]. ACTA automatica sinica, 2018, 44(11): 2103–2112.

[20] HUANG Dong, LAI Jianhuang, WANG Changdong. Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis[J]. Neurocomputing, 2015, 170: 240–250.

[21] BAI Liang, LIANG Jiye, DU Hangyuan, et al. An information-theoretical framework for cluster ensemble[J]. IEEE transactions on knowledge and data engineering, 2019, 31(8): 1464–1477.

[22] WAN Haowen, NING Bo, TAO Xiaoyu, et al. Artificial intelligence in China[M]. Singapore: Springer, 2020: 622-628.

[23] DOMENICONI C, AL-RAZGAN M. Weighted cluster ensembles: methods and analysis[J]. ACM transactions on knowledge discovery from data, 2009, 2(4): 1–40.

[24] IAM-ON N, BOONGOEN T, GARRETT S, et al. A link-based approach to the cluster ensemble problem[J]. IEEE transactions on pattern analysis and machine Intelligence, 2011, 33(12): 2396–2409.

[25] HUANG Dong, WANG Changdong, LAI Jianhuang. Locally weighted ensemble clustering[J]. IEEE transactions on cybernetics, 2018, 48(5): 1460–1473.

[26] VO C T N, NGUYEN P H. A weighted object-cluster association-based ensemble method for clustering undergraduate students[C]//Asian Conference on Intelligent Information and Database Systems. Cham: Springer, 2018: 587−598.

[27] LI Feijiang, QIAN Yuhua, WANG Jieting, et al. Cluster's quality evaluation and selective clustering ensemble[J]. ACM transactions on knowledge discovery from data, 2018, 12(5): 1–27.

[28] RASHIDI F, NEJATIAN S, PARVIN H, et al. Diversity based cluster weighting in cluster ensemble: an information theory approach[J]. Artificial intelligence review, 2019, 52: 1341–1368.

[29] BANERJEE A, PUJARI A K, RANI PANIGRAHI C, et al. A new method for weighted ensemble clustering and coupled ensemble selection[J]. Connection Science, 2021, 33(3): 623–644.

[30] 邵长龙, 孙统风, 丁世飞. 基于信息熵加权的集成聚类算法 [J]. 南京大学学报(自然科学版), 2021, 57(2): 189–196.
SHAO Changlong, SUN Tongfeng, DING Shifei. Ensemble clustering based on information entropy weighted[J]. Journal of Nanjing University (natural science edition), 2021, 57(2): 189–196.

[31] ZHONG Caiming, YUE Xiaodong, ZHANG Zehua, et al. A clustering ensemble: two-level-refined co-association matrix with path-based transformation[J]. Pattern recognition, 2015, 48(8): 2699–2709.

[32] LI Feijiang, QIAN Yuhua, WANG Jieting, et al. Clustering ensemble based on sample's stability[J]. Artificial intelligence, 2019, 273: 37–55.

[33] REN Yazhou, DOMENICONI C, ZHANG Guoji, et al. Weighted-object ensemble clustering: methods and analysis[J]. Knowledge and information systems, 2017, 51(2): 661–689.

[34] 武永亮, 赵书良, 李长镜, 等. 基于 TF-IDF 和余弦相似度的文本分类方法 [J]. 中文信息学报, 2017, 31(5): 138–145.
WU Yongliang, ZHAO Shuliang, LI Changjing, et al. Text classification method based on TF-IDF and cosine similarity[J]. Journal of Chinese information processing, 2017, 31(5): 138–145.

[35] THENMOZHI D, KANNAN K, ARAVINDAN C. A text similarity approach for precedence retrieval from legal documents[C]//FIRE (Working Notes). Bangalore: CEUR Workshop Proceedings, 2017: 90−91.

[36] 刘梦迪, 梁循. 基于偏旁部首知识表示学习的汉字字形相似度计算方法 [J]. 中文信息学报, 2021, 35(12): 47–59.
LIU Mengdi, LIANG Xun. A method of Chinese character glyph similarity calculation[J]. Journal of Chinese information processing, 2021, 35(12): 47–59.

[37] HAN E-H, BOLEY D, GINI M, et al. WebACE: a web agent for document categorization and exploration[C]// The 2nd International Conference on Autonomous Agents. Minneapolis: ACM, 1998: 408-415.

作者简介：

李娜，女，硕士研究生，主要研究方向为文本挖掘、机器学习和模式识别。E-mail：lina980104@163.com。

徐森，教授，博士，主要研究方向为机器学习、模式识别和文本挖掘。主持完成国家自然科学基金青年基金项目、江苏省教育厅国际科技合作聘请外国专家重点项目、江苏省高校自然科学面上项目各 1 项，主持江苏省政策引导类计划（产学研合作）–前瞻性联合研究项目 1 项，作为主要成员参与完成国家自然科学基金 5 项，省部级项目 5 项。发表学术论文 40 余篇，申请中国发明专利 20 余项，获得授权 8 项。国家自然科学基金通讯评审专家库成员，江苏省人工智能学会机器学习专委会常务委员，江苏省计算机学会大数据专家委员会委员，盐城市计算机学会理事，盐城市人工智能学会监事长，美国计算机协会会员，中国计算机学会会员，江苏省计算机学会会员。E-mail：xusen@ycit.cn。

徐秀芳，高级实验师，主要研究方向为数据挖掘和智能信息处理。以第一发明人申请国家专利 4 项，取得省级以上科研成果 3 项，市级科研成果 2 项，先后主持或参与完成 8 项省市级纵横向科研项目。主编或参与编写教科书 4 部。E-mail：xxf@ycit.cn。