CMDC：一种差异互补的迭代式多维度文本聚类算法

黄瑞章[1,2]，白瑞娜[1]，陈艳平[1,2]，秦永彬[1,2]，程欣宇[1,3]，田有亮[1,2]

（1. 贵州大学计算机科学与技术学院，贵州 贵阳 550025；
2. 贵州省公共大数据重点实验室，贵州 贵阳 550025；
3. 贵州省智能人机交互工程技术研究中心，贵州 贵阳 550025）

摘 要：针对传统多维度文本聚类算法把文本表示与聚类过程分离，忽略了维度间的互补特性的问题，提出了一种差异互补的迭代式多维度文本聚类算法——CMDC，实现文本聚类与特征调整过程的统一优化。CMDC 算法挑选维度聚类间结果的互补文本，基于局部度量学习算法利用互补文本促进聚类的特征调优，以维度的度量一致性来解决多维度文本聚类的划分一致性。实验结果表明，CMDC 算法有效地提升了多维度聚类性能。

关键词：多维度文本聚类；互补文本；约束文本聚类；度量计算

中图分类号：TP301

文献标识码：A

doi: 10.11959/j.issn.1000−436x.2020152

CMDC: an iterative algorithm for complementary multi-view document clustering

HUANG Ruizhang[1,2], BAI Ruina[1], CHEN Yanping[1,2], QIN Yongbin[1,2],
CHENG Xinyu[1,3], TIAN Youliang[1,2]

1. College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
2. Guizhou Provincial Key Laboratory of Public Big Data, Guiyang 550025, China
3. Guizhou Intelligent Human-Computer Interaction Engineering Technology Research Center, Guiyang 550025, China

Abstract: In response to the problems traditional multi-view document clustering methods separate the multi-view document representation from the clustering process and ignore the complementary characteristics of multi-view document clustering, an iterative algorithm for complementary multi-view document clustering——CMDC was proposed, in which the multi-view document clustering process and the multi-view feature adjustment were conducted in a mutually unified manner. In CMDC algorithm, complementary text documents were selected from the clustering results to aid adjusting the contribution of view features via learning a local measurement metric of each document view. The complementary text document of the results among the dimensionality clusters was selected by CMDC, and used to promote the feature tuning of the clusters. The partition consistency of the multi-dimensional document clustering was solved by the measure consistency of the dimensions. Experimental results show that CMDC effectively improves multi-dimensional clustering performance.

Key words: multi-view document clustering, complementary text, constrained document clustering, metric calculation

收稿日期：2020−02−03；**修回日期**：2020−06−10

基金项目：国家自然科学基金资助项目（No.61462011，No.91746116）；国家自然科学基金联合基金资助项目（No.U1836205）；贵州省科学技术基金资助项目（No. [2020]1Z055）

Foundation Items: The National Natural Science Foundation of China (No.61462011, No.91746116), The Joint Funds of the National Natural Science Foundation of China (No.U1836205), The Key Projects of Science and Technology of Guizhou (No.[2020]1Z055)

1 引言

文本聚类，旨在按照文本的相似性自动挖掘文本的结构，是文本挖掘的重要任务，被众多应用所关注[1]。传统的文本聚类多从单一的文本内容维度出发，根据文本内容中的语义特征来挖掘文本数据集的结构关系。随着互联网和数据分析技术的发展，文本数据的表示逐渐从传统单一的内容维度向立体的多维度发展[2]。例如，互联网环境中的新闻文本数据，除表示为以词向量为代表的传统内容维度以外，还可表示为新闻文本的主题维度（如新闻用词所涵盖的主题），以及新闻文本在互联网传播过程中获得的传播行为维度（如新闻的转发用户、阅读用户、点赞用户等）；研究类论文的文本数据除表示为传统的内容维度以外，还可被描述为论文的研究行为维度，包含论文自身、引用论文和被引论文的作者等。这些多维度文本数据较传统的表示方式更为全面立体，如何有效利用文本的多维度数据来分析挖掘文本数据集的结构，为传统文本聚类问题带来了新的机遇和挑战。

多维度文本聚类可联合利用多个维度的信息改善单维度信息在文本聚类上的局限，为文本聚类带来了机遇。在实际的多维度文本聚类过程中，数据的多个维度特征对文本结构的发现有互补作用，传统文本内容维度中表现不佳的数据在其他维度可能获得更好的聚类结果。例如，在面向研究类论文的文本聚类问题中，同领域的研究论文涉及的具体研究细节不同，使论文中的内容和用词不尽相同，这导致相同领域的论文在传统文本内容表达维度中具有较大的差异，难以被划分到同一个聚类分组中。然而在论文的研究行为维度，这些论文普遍被同一批学者所关注，更倾向于被划分到同一类簇中。相似地，在新闻领域的文本聚类中，被相似人群关注的新闻一般具有相似的主题，但这些新闻文本聚类的内容表示往往因作者写作风格或新闻事件演变等原因存在差异，增加了新闻文本聚类的难度。

除了机遇，文本的多维度表示亦为文本聚类带来挑战。其中，多维度文本聚类的一个核心问题是如何从文本在多个维度的表示中获得一致的聚类结果。由于文本在不同维度上的表示具有差异性，使文本在维度上的距离测量不一致，导致各维度的聚类划分不一致。文本表示的差异性主要表现在以下 2 个方面：1)不同维度的文本表示中特征的含义不同，例如，新闻内容维度特征主要反映新闻的主题，新闻评论维度特征反映用户对新闻的态度，新闻行为维度特征反映新闻内容的传播受众群体；2)文本表示中的关键特征与噪声特征的分布不同，例如新闻的主题维度中的噪声信息相对较少，但新闻的内容维度中普遍包含大量的噪声信息，关键特征在距离测量中的贡献容易被噪声特征淹没，且各关键特征对距离度量的贡献各异。因此，如何有效利用多维度文本聚类的互补特性，设计合理的多维度文本聚类算法以弥补多维度聚类结果差异的问题，非常值得研究。

对于不同维度聚类结果不一致的问题，目前多维度文本聚类算法大多采用首先对各维度进行融合表示学习，在此基础上利用传统的单维度聚类实现文本聚类的整体划分，维度表示过程与聚类过程被分割成 2 个独立的步骤，无法利用多维度聚类的互补特点指导各维度特征的贡献。针对以上问题，本文构建一种差异互补的迭代式多维度文本聚类算法——CMDC（complementary multi-view document clustering）算法，使多维度文本聚类过程与文本维度特征的调整互相促进，利用多维度文本的互补特性弥补多维度文本聚类的划分的差异，实现聚类与特征调整过程的统一优化。本文需要解决 3 个问题，具体如下。

1) 如何从聚类划分中获取互补文本，即在维度类簇中聚类意见不一致的文本数据。由于各维度聚类类簇含义不同，不能简单地认为在各维度聚类结果中未被划分到同标签类簇的文本为互补文本。因此，如何挑选维度间的互补文本是本文需要解决的问题。

2) 如何利用互补文本促进聚类的特征调优。各维度聚类类簇关注的关键特征不同，需要有效利用互补文本改善各维度的特征在聚类过程中的贡献，使互补文本在文本的多个维度中呈现一致的聚类结果。

3) 如何使维度特征调优与聚类划分共同优化。区别于传统多维度文本聚类算法，本算法将设计聚类划分与维度特征调优的共同优化，利用维度间的互补文本帮助聚类划分与维度特征的调优互相迭代促进。

对互补文本的获取问题，CMDC 算法通过文本对的聚类结果一致性（即是否同属一个类簇）来判

断文本对在不同维度中的聚类意见，并设计了一个可信因子综合考虑当前及其他维度中文本对的聚类结果，评估文本对在当前聚类结果的可信程度。基于互补文本，CMDC 算法以维度的度量一致性来解决多维度文本聚类的划分一致性问题，通过度量学习调整维度特征对聚类的贡献，在此基础上本文提出了基于度量学习的约束文本聚类算法，为各维度的每个类簇设计了独立的度量矩阵，利用互补文本调节各类簇的度量计算方法，解决因文本差异性造成的维度和类簇间的度量差异。在基于度量学习的约束文本聚类中，设计了聚类与度量学习的共同优化目标函数，面向互补文本实现聚类结果与度量学习的共同调优。最终，令互补文本挑选及基于度量学习的约束文本聚类算法迭代进行，互相促进，提升各维度间聚类结果的一致性。本文采用 2 个真实的数据集进行验证，并与多个先进多维度文本聚类算法进行对比。从实验结果来看，CMDC 算法可有效地利用多维度数据的互补性改善多维度文本的差异性问题，聚类结果有明显提升，验证了算法的有效性。

2　相关工作

多维度聚类旨在通过对可用的多维度特征信息进行组合，以在不同维度之间搜索一致的聚类分配，将相似的主题分到同一类簇中[3]。多维度聚类问题提出[4]以来，相关算法受到了广泛关注，并运用于文本挖掘和信息检索等领域。目前，大多算法都是直接关注聚类目标，通过优化算法寻求最佳的聚类解决方案。与聚类算法类似，多维度聚类也分为不同维度的特征表示学习和聚类 2 个阶段。以是否将 2 个阶段融合为标准，现有算法被分为 2 类。最具有代表性的是利用典型相关分析（CCA, canonical correlation analysis）将多维度数据投影到低维空间融合[5]进行聚类。文献[6]针对 2 个维度的数据，基于协同训练思想提出了使用某一维度的拉普拉斯算子的特征向量对样本进行聚类，然后利用聚类结果来修正另一维度中的拉普拉斯算子，直到得到具有足够结构信息的特征向量，并将其作为下游聚类算法（k-means 或谱聚类等）的输入。该团队又从最小化数据不同维度的预测函数出发，将拉普拉斯图的特征向量矩阵作预测函数，提出了 2 种基于协同正则化的多维度谱聚类算法[7]。将数据的多维度信息作为子空间特征，为了使多个子空间获得一致的聚类结果，文献[8]通过强制最小化每对子空间系数矩阵来获得共享公共系数矩阵。近年来随着神经网络深入各个领域，基于深度学习框架的多维度聚类算法也不容忽视。文献[9]和文献[10]都采用基于深度学习的框架来学习不同维度间的特征表示，进行融合后再运用图聚类或子空间聚类等方法得到聚类结果。为了改善多维度聚类算法两阶段的断层，逐步出现了统一特征表示和聚类两阶段的多维度聚类算法。文献[11]将图片的每种类型的特征视为一个维度，提出了通过统一不同维度（即图像特征）来学习共享的图拉普拉斯矩阵的多模态光谱聚类（MMSC, multi-modal spectral clustering）算法，并直接求解聚类指标矩阵。文献[12]提出改进的低秩表示模型，可对维度特征空间中的局部数据流形结构进行建模，基于谱聚类实现多维度协议的共同优化。在多维度深度聚类的最新研究中，文献[13]改进单维度深度嵌入聚类（DEC, deep embedding for clustering analysis）模型[14]，利用文本聚类的结果来调整多维度融合参数。文献[15]运用多维度聚类解决对话意图来学习任务，提出了同时学习多维度特征表示和优化聚类的算法。目前，统一特征表示和聚类两阶段的多维度聚类算法中尚处于摸索阶段，聚类过程与特征表示过程虽然被同步优化，却忽视了多维度文本数据的差异性表示，未考虑利用具有争议的聚类文本改进聚类结果的不一致问题。

聚类算法（如 k-means 算法）依赖于底层距离函数，针对由多个维度表示组成的高维稀疏的文本数据，通常采用的距离函数或手动调整的度量方式显然是不适用的。文献[16]提出距离度量学习算法寻求在半监督或完全监督的设置中自动优化距离函数，其学习目标是优化反映当前问题领域特定概念的距离函数。文献[17]在文献[16]的基础上提出了基于无监督自适应度量学习算法，同时执行聚类和度量矩阵学习。文献[18]针对度量方式提出了一种非线性度量学习算法，通过学习非参数核矩阵来学习完全灵活的距离度量并用到聚类中。文献[19]也给出了运用到图像、分类等任务上的度量学习算法的实证评估，并指出使用依赖成对约束的度量算法可以产生与有监督算法相当的实验效果。然而，上述基于度量学习的算法都是面向单维度数据的，其约束对或者标签数据都来自于数据自身。在多维度文本数据的聚类上，文献[20]将文本数据中其他维度信息与文本维度聚类相结合，文献[21]则使用了基于

辅助数据约束的度量学习算法用于聚类，但这些算法在融入其他维度信息时也带入了文本噪声。

3 模型设计

3.1 符号与术语

本文使用的数据集中都是文本数据，为了方便数据及问题的描述，给出如下定义。

1) 特征 t^m 是维度 m 的基本组成。各维度 m 的特征库 $\boldsymbol{V}^m$ 不同，记为 $\boldsymbol{V}^m=\{t_1^m,t_2^m,\cdots,t_{|\boldsymbol{V}^m|}^m\}$，其中 t_i^m 是维度 m 的特征库 $\boldsymbol{V}^m$ 的第 i 个特征，i=1,2,⋯,$|\boldsymbol{V}^m|$。

2) 文本 $\boldsymbol{x}$ 包含 M 个维度，表示为 $\boldsymbol{x}=(\boldsymbol{x}^1,\boldsymbol{x}^2,\cdots,\boldsymbol{x}^M)$。每个维度 $\boldsymbol{x}^m$ 由特征库 $\boldsymbol{V}^m$ 特征权重组成 $\boldsymbol{x}^m=(w_1^m,w_2^m,\cdots,w_{|\boldsymbol{V}^m|}^m)$，其中 w_i^m 是第 i 个特征 t_i^m 的权重。

3) 数据集 D 包含 $|D|$ 个文本数据，记为 $D=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{|D|})$，数据集也根据维度表示为 $D=(D^1,D^2,\cdots,D^M)$。

4) 挑选的互补文本集为 $\{C^1,C^2,\cdots,C^M\}$。

5) 聚类结果表示为 $\{\theta^1,\theta^2,\cdots,\theta^M\}$。

3.2 CMDC 算法的整体设计

CMDC 算法利用多维度文本数据的互补性解决因多维度文本数据的差异性带来的聚类效果低下的问题。CMDC 算法通过识别各维度聚类结果中的互补文本数据来评估这些互补文本数据中的低质量聚类维度，应用于后续的文本聚类过程。在聚类过程中，为各个维度的每个类簇设计了一个度量矩阵，并自动地利用互补文本来调节，使互补文本在多个维度的度量具有相似的结果，以提升多维度文本聚类结果的一致性，最终实现多维度文本聚类整体效果的提升。CMDC 算法的具体过程如图 1 所示。

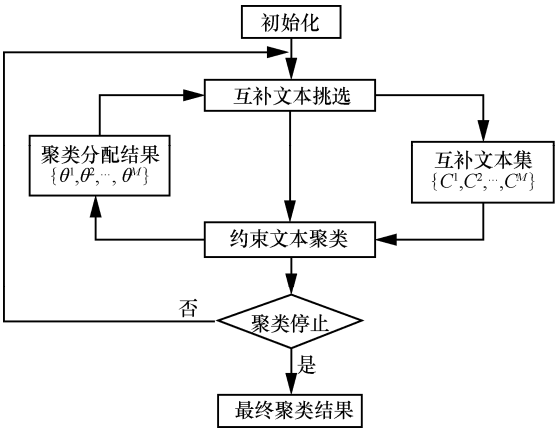


图 1 CMDC 算法过程

CMDC 算法包含 2 个关键组成部分，分别为互补文本挑选和约束文本聚类。互补文本挑选旨在自动学习每个维度 m 聚类结果中不一致的文本数据。本文以文本对 $(\boldsymbol{x}_i,\boldsymbol{x}_j)$ （$i,j=1,2,\cdots,|D|$）来评估聚类结果的一致性，若 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 在每个维度结果中都被分配到或都未被分配到同一类簇中，则认为 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的聚类意见一致，否则 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 在部分维度中属于同一类簇，在其他维度中被分配到不同的类簇，则($\boldsymbol{x}_i$, $\boldsymbol{x}_j$)为互补文本。互补文本挑选为每个文本维度自动学习互补文本集 C^m，其中包含在维度 m 中聚类质量可信度低的互补文本。互补文本集 C^m 将被加入后续的约束文本聚类过程中，对文本聚类进行约束，学习聚类过程中的合理距离度量。通过为各维度的每个类簇 k 学习不同的局部度量矩阵 $\boldsymbol{Q}_k^m=q_{k,i}^m(i=1,\cdots,|\boldsymbol{V}^m|)$，来调整各维度中各类簇中各特征的贡献权重，使关键特征在相似度测量中的贡献权重更高，并相应地降低噪声特征的影响，最终令互补文本集 C^m 中的文本对在约束文本聚类过程的度量一致，改善聚类结果。约束文本聚类为各维度学习新的聚类分配结果 θ_k^m 和局部度量矩阵 $\boldsymbol{Q}_k^m$，各聚类分配用于辅助下一轮互补文本挑选。在 CMDC 算法过程中，互补文本挑选与约束文本聚类互相促进，循环迭代直至聚类的结果收敛或互补文本的数量达到设置上限，CMDC 算法过程停止。聚类停止后，挑选互补文本最少的维度输出作为聚类的整体结果。

3.3 互补文本挑选

互补文本挑选重点考虑与维度间聚类结果不一致的文本对 $(\boldsymbol{x}_i,\boldsymbol{x}_j)$。通过设计可信因子 $\lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m}$ 来估算文本对 $(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 在维度 m 中的聚类可信度，该因子对 $(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 当前维度和其他维度的聚类被分配到同一类簇的概率差异进行对比。为减少计算量，首先选取在维度 m 不属于同一类簇中，但在其他维度中均属于同一类簇的文本来计算 $\lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m}$，如式(1)所示。

$$\lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m}=\frac{\max_k p((\boldsymbol{x}_i,\boldsymbol{x}_j)^m\in\theta_k^m)}{\sqrt[\widetilde{m}]{\prod_{\widetilde{m}}\max_k p((\boldsymbol{x}_i,\boldsymbol{x}_j)^{\widetilde{m}}\in\theta_k^{\widetilde{m}})}} \tag{1}$$

其中，θ_k^m 表示维度 m 聚类结果中分配的第 k 个类簇；$\widetilde{m}$ 表示除维度 m 以外的其他维度；$p((\boldsymbol{x}_i,\boldsymbol{x}_j)^m\in\theta_k^m)$ 表示在维度 m 中文本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 在聚

类结果中属于同一个类簇 θ_k 的概率，由于文本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 在聚类过程中彼此独立，$p((\boldsymbol{x}_i,\boldsymbol{x}_j)^m \in \theta_k^m)$ 与 $p(\boldsymbol{x}_i \in \theta_k^m)p(\boldsymbol{x}_j \in \theta_k^m)$ 可进行等价转化。

文本与所属的类簇越接近，文本的聚类结果越确定，因此 $p(\boldsymbol{x}^m \in \theta_k^m)$ 可按照式(2)来计算。

$$p(\boldsymbol{x}^m \in \theta_k^m) = \frac{s_{\boldsymbol{Q}_k^m}(\boldsymbol{x}^m, \boldsymbol{\mu}_k^m)}{\sum_{k'} s_{\boldsymbol{Q}_{k'}^m}(\boldsymbol{x}^m, \boldsymbol{\mu}_{k'}^m)} \tag{2}$$

其中，$\boldsymbol{\mu}_k^m$ 是类簇 θ_k^m 的质心；$s_{\boldsymbol{Q}_k^m}(\boldsymbol{x}^m, \boldsymbol{\mu}_k^m)$ 是文本 $\boldsymbol{x}$ 在维度 m 上距离类簇质心 $\boldsymbol{\mu}_k^m$ 的度量余弦相似，该度量余弦相似在传统余弦相似的基础上，利用度量矩阵调节相似度的计算结果；k' 与 k 含义相同，这里为了避免混淆，使用 k' 求和来做归一化。

$\lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m}$ 的值越小，文本对 $(\boldsymbol{x}_i,\boldsymbol{x}_j)^m$ 在不同维度间的聚类结果的差异越大，$(\boldsymbol{x}_i,\boldsymbol{x}_j)^m$ 在维度 m 的聚类结果中被归属于同一个类别的概率越低，在除 m 以外的其他维度被聚到同类簇的概率越高，则文本对在维度 m 的聚类可信度越低。因此，通过置信阈值选取 $\lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m}$ 值合理小的文本对。设置阈值 τ，并选取 $\lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m} < \tau$ 的那些文本对 C^m 加入后续的约束文本聚类中。

3.4　约束文本聚类模块

该模块由约束文本聚类算法构成。针对每一个维度 m，互补文本对抽取模块依赖前序的文本聚类结果，自动学习互补文本集合 C^m，互补文本 $(\boldsymbol{x}_i,\boldsymbol{x}_j)^m \in C^m$ 在单维度的聚类中结果较差。在约束文本聚类中，提出利用 C^m 改善维度 m 的聚类结果。在此过程中，需要计算文本与文本之间、文本与类簇质心之间的距离。由于余弦相似计算无法区分特征在距离计算中的贡献，本文在聚类过程中引入度量学习来进行调整。文本 $\boldsymbol{x}^m$ 与类簇质心 $\boldsymbol{\mu}_k^m$ 之间的度量余弦相似如式(3)所示。

$$s_{\boldsymbol{Q}_k^m}(\boldsymbol{x}^m, \boldsymbol{\mu}_k^m) = \frac{\sum_i^{|\boldsymbol{V}^m|} w_i^m q_{k,i}^m \boldsymbol{\mu}_{k,i}^m}{\| \boldsymbol{x}^m \|_{\boldsymbol{Q}_k^m} \| \boldsymbol{\mu}_k^m \|_{\boldsymbol{Q}_k^m}} \tag{3}$$

其中，$|\boldsymbol{V}^m|$ 表示 $\boldsymbol{Q}_k^m$ 中特征的个数，$\boldsymbol{\mu}_{k,i}^m$ 表示维度 m 的聚类结果中第 k 个类簇质心 $\boldsymbol{\mu}_k^m$ 中的第 i 个特征的权重，w_i^m 表示 $\boldsymbol{x}$ 在维度 m 中的第 i 个特征权重，$q_{k,i}^m$ 表示局部度量矩阵中的特征，$\| \boldsymbol{x}^m \|_{\boldsymbol{Q}_k^m}$ 表示 $\boldsymbol{x}^m$ 的加权 L_2 范数，其计算为方式为 $\| \boldsymbol{x}^m \|_{\boldsymbol{Q}_k^m} = \sqrt{\sum_i^{|\boldsymbol{V}^m|} (w_i^m)^2 q_{k,i}^m}$，相似地，$\| \boldsymbol{\mu}_k^m \|_{\boldsymbol{Q}^m}$ 表示 $\boldsymbol{\mu}_k^m$ 的加权 L_2 范数。

总体来说，CMDC 算法的目标函数如式(4)所示。

$$J^m = \alpha \Omega^m + (1-\alpha)\Phi^m \tag{4}$$

其中，Ω^m 表示当前维度 m 的聚类目标，评估当前聚类的总体结果质量；Φ^m 表示约束目标，评估当前维度的互补文本的符合情况。这 2 个部分以参数 α 进行线性连接。Ω^m 测量 D^m 中所有的文本数据到其分配类簇的距离，对所有的文本数据以及类簇的质心进行归一化处理后，Ω^m 的计算如式(5)所示。

$$\Omega^m = \sum_{\boldsymbol{x}^m \in D^m} (\| \boldsymbol{x}^m - \boldsymbol{\mu}_k^m \|_{\boldsymbol{Q}_k^m}^2 - \log\det(\boldsymbol{Q}_k^m)) = \sum_{\boldsymbol{x}^m \in D^m} (2 - 2s_{\boldsymbol{Q}_k^m}(\boldsymbol{x}^m, \boldsymbol{\mu}_k^m) - \log\det(\boldsymbol{Q}_k^m)) \tag{5}$$

Φ^m 是约束目标，此目标计算互补文本集 C^m 的符合度。判断在互补文本集 C^m 中的文本对是否在聚类中被划分到一个类别中，若否，则对文本对进行惩罚。以文本对的 $\lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m}$ 结果计算惩罚的程度，具体计算方法如式(6)所示。

$$\Phi^m = \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j) \in C^m} (1 - \lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m}) \delta(S_{\boldsymbol{x}_i} \neq S_{\boldsymbol{x}_j}) \tag{6}$$

其中，$S_{\boldsymbol{x}}$ 表示文本 $\boldsymbol{x}$ 所属的类簇；δ 表示指示函数，$\delta(\text{true}) = 1$，$\delta(\text{false}) = 0$。

本文采用循环迭代机制来计算式(4)所示目标函数的最优解，如算法 1 所示。

算法 1　约束文本聚类算法

输入　数据集 D、文本维度 m

输出　D^m 对应的类簇质心 $\boldsymbol{\mu}_k^m$、度量矩阵 $\boldsymbol{Q}_k^m$

1) 初始化类簇起始点。

2) 给定聚类的类簇质心点，根据式(3)计算当前维度的文本到各类簇质心的相似度，选择相似度最高的类簇分配文本数据。

3) 给定聚类分配。

4) 更新各类簇的质心表示 $\boldsymbol{\mu}_k^m$。

5) 更新各类簇的度量矩阵 $\boldsymbol{Q}_k^m$。

6) 跳转到 2)重复直至收敛。

其中，类簇质心 $\boldsymbol{\mu}_k^m$ 根据被分配的所有文本进行更

新，计算方法如式(7)所示。

$$\boldsymbol{\mu}_k^m = \frac{1}{|\theta_k^m|}\sum_{\boldsymbol{x}^m \in \theta_k^m} \boldsymbol{x}^m \tag{7}$$

在度量矩阵的更新算法中，通过对目标函数 J^m 求偏导更新每个簇的度量矩阵 $\boldsymbol{Q}_k^m$。令 $\frac{\partial J^m}{\partial \boldsymbol{Q}_k^m}=0$，则有

$$\begin{aligned}\boldsymbol{Q}_k^m = &|D^m|\left[\sum_{\boldsymbol{x}^m \in D^m}(\boldsymbol{x}^m-\boldsymbol{\mu}_k^m)(\boldsymbol{x}^m-\boldsymbol{\mu}_k^m)^{\mathrm{T}} + \right.\\ &\frac{1-\alpha}{\alpha}\sqrt[\tilde{m}]{\prod_{\tilde{m}}\max_k p((\boldsymbol{x}_i,\boldsymbol{x}_j)^{\tilde{m}} \in \theta_k^{\tilde{m}})}\cdot\\ &\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in C^m}\frac{\sum_{k'} s_{\boldsymbol{Q}_{k'}^m}(\boldsymbol{x}^m,\boldsymbol{\mu}_{k'}^m)-s_{\boldsymbol{Q}_k^m}(\boldsymbol{x}^m,\boldsymbol{\mu}_k^m)}{\left(\sum_{k'} s_{\boldsymbol{Q}_{k'}^m}(\boldsymbol{x}^m,\boldsymbol{\mu}_{k'}^m)\right)^2}\cdot\\ &\left.(\boldsymbol{x}_i^m-\boldsymbol{x}_j^m)(\boldsymbol{x}_i^m-\boldsymbol{x}_j^m)^{\mathrm{T}}\delta(S_{\boldsymbol{x}_i}\neq S_{\boldsymbol{x}_j})\right]^{-1}\end{aligned} \tag{8}$$

4 实验

4.1 数据集及评估方法

实验使用 2 个真实数据集以验证 CMDC 算法的有效性。第一个真实数据集是英文论文数据集 AMiner。此数据集包含 3 个类簇，每个文本表达为 2 个维度，其中，以论文的摘要作为摘要维度，以论文的作者及参考文献的第一作者作为用户维度。本文爬取同一时期微博、百度和头条新闻等数据源 4 个重要新闻话题的热点新闻，构成一个多源热点新闻数据集（MHN, multi-source hot news），作为实验的第二个数据集。 MHN 共涉及 3 个维度，包含从新闻的正文中提取的正文维度、从新闻的标题中提取的关键内容作为标题维度，以及利用主题模型 LDA（latent Dirichlet allocation）提取的主题维度。从数据维度的构成上来看，AMiner 数据集的差异性大于 MHN 数据集。表 1 展示了数据集的详细信息。

表 1 数据集信息

数据集	样本个数/个	维度数量/维	类别数量/种
AMiner	1 500	2	3
MHN	2 605	3	4

本文使用归一化互信息指标（NMI, normalized mutual information）来评价实验的聚类效果，其计算式如式(9)所示。

$$\mathrm{NMI} = \frac{2I(R;S)}{H(R)+H(S)} \tag{9}$$

其中，$R=\{r_1,r_2,\cdots,r_k\}$ 表示算法聚类后的簇集合，$S=\{s_1,s_2,\cdots,s_j\}$ 表示标准的聚类标签；$I(R;S)=H(R)-H(R|S)$ 表示随机变量间的互信息，$H(R)$表示 R 的熵，$H(R|S)$ 表示给定 S 时 R 的条件熵。NMI 的取值范围为[0,1]，该值越大说明聚类效果越好。

4.2 实验参数设置

针对 AMiner 数据集，考虑用户维度覆盖学者和其所研究的领域（即摘要维度）具有一定的一致性，因此利用用户维度映射得出的表示也具有与摘要维度相同的意义，可以进行降维，达到更好的提炼特征的效果。本文结合深度学习特征的表现方式，将 AMiner 数据集提取的用户维度的特征映射到摘要维度，训练关于用户信息的嵌入（embedding）模型并以此来提取用户维度的特征表示。而对 MHN 数据集的标题维度，本文则选用了包含语义信息的 BERT（bidirectional encoder representation from transformer）模型[22]做文本表示，使输入增加语义信息。2 个数据集的其他维度都使用原始的词频向量表示。

对于 3.2 节中 CMDC 算法的停止条件，本文设置互补文本集数量上限为 12 000；对于 3.3 节中互补文本挑选模块中的参数，设置置信阈值 $\tau\in(0,1)$。为了更完善地捕获多维度数据的互补性，通过对实验涉及的 2 个数据集进行统计分析，将 $\{\lambda_{(\boldsymbol{x}_i,\boldsymbol{x}_j)^m};(\boldsymbol{x}_i,\boldsymbol{x}_j)^m\in C^m, i\neq j\}$ 的第三、四分位数设置为当前维度阈值 τ（针对不同的数据集特性，该算法需要根据自身任务及经验设置合适的参数，为获得足够数量的互补文本，阈值 τ 可进行放大）。在本文的实验中，AMiner 数据集的摘要维度和用户维度均设置为 0.08，MHN 的标题维度、正文维度以及主题维度设置为 0.52。

为使当前维度的聚类目标和约束目标同时发挥作用，实验中，对于式(4)所示的目标函数，本文设置 α=0.5 。

4.3 对比实验及结果分析

CMDC 算法的本质是采用质量互补文本挑选模块及基于度量学习的约束文本聚类模块迭代进行的，利用多维度文本数据的维度互补性来弥补文本数据在单个维度聚类过程中质量的不足，最终提升各维度间聚类结果的一致性。本文

通过 2 个数据集共记 5 个维度进行分析，从单维度和多维度 2 个方面对 CMDC 算法进行对比；除此之外，还对 CMDC 算法的互补性和一致性进行了探究。

4.3.1 单维度聚类实验

为验证质量互补文本挑选和基于度量学习的约束文本聚类的作用，本文选取了几种对比算法，具体如下。

1) k-means 算法。该算法是传统无监督聚类，作为单维度聚类性能比较的基线方法；

2) MTCUBC[21]（multi-dimensional text clustering with user behavior characteristics）算法。该算法是基于辅助维度的约束信息单向进行基于度量学习的约束文本聚类算法。

实验设置互补文本对的数目为 12 000 条，3 种算法的单维度聚类性能如表 2 所示，CMDC 算法在所有的维度上都高于 k-means。相比于 MTCUBC 算法，在除 MHN 数据集的正文维度之外的其他 4 个维度上，CMDC 算法分别有 0.035、0.117、0.062 和 0.135 的提升。同时也分析了不同数据集的提升差异，从 k-means 算法的聚类效果可以看出，AMiner 数据集中 2 个维度的聚类效果差距较小，而 MHN 数据集中不同维度间的差异性虽然较低，但是各维度的性能相差高达 0.3。因而在 CMDC 算法过程不同维度相互迭代、相互促进的过程中，AMiner 数据集的 2 个单维度性能都得到了提高；而在 MHN 数据集的 3 个维度中，标题维度和主题维度的性能都有了较大的提升，而这种提升在整个 CMDC 算法中是以正文维度性能小提升（基于基线方法）为代价的，因此在正文维度，CMDC 算法的性能略低于 MTCUBC 算法。也正是出于对度量差异的极度不平衡性（而非来源差异性）的考虑，本文在数据预处理阶段添加了语义嵌入来降低这种不平衡。在单维度聚类的对比实验中，CMDC 算法在 4 个维度上取得了最好的效果。从 CMDC 算法和 MTCUBC 算法的差异而言，充分说明本文设计的互补文本集学习策略在聚类过程中是有效的。

4.3.2 多维度聚类实验

针对多维度数据差异性和互补性 2 个特点，除了基线方法外，本文选取了多种算法进行对比。

1) Mv+k-means。对多维度信息进行无差别拼接组合后进行 k-means 聚类，作为多维度聚类基线方法。

2) P-MLRSSC 和 C-MLRSSC。MLRSSC（multi-view low-rank sparse subspace clustering）[23]系列算法通过构造亲和力矩阵，学习多维度之间共享的联合子空间表示来改善多维度文本的差异性问题，本文选取了适用于当前数据集的 2 个算法：P-MLRSSC（pairwise MLRSSC）、C-MLRSSC（centroid MLRSSC）用于对比，参数设置请见文献[23]。

3）MSC_IAS （multi-view subspace clustering with intactness-aware similarity）。为充分利用多维度数据的互补性，Wang 等[24]提出通过集成编码的补充信息来学习完整空间，记为 MSC_IAS。实验设置参数如下。AMiner 数据集：k=30，d=600。MHN 数据集：k=6，d=1 500。参数释义请见文献[24]。

同样设置互补文本对的个数为 1 2000 条来验证 CMDC 算法在多维度的效果。由表 3 可以看出，CMDC 算法在 AMiner 数据集的提升效果最为明显，其原因是论文在摘要维度中多关注论文解决的实际问题和使用方法，用词差异大，使用的用户维度由相关文献的第一作者组成，其共同关注度更高，AMiner 数据集的互补性更好。相比于 AMiner 数据集，MHN 数据集的 3 个表示维度的关联度更高，互补性更弱，因此多维度性能提升效果不显著。

表 3　　多维度聚类性能 NMI

算法	AMiner	MHN
Mv+k-means	0.782	0.760
P-MLRSSC	0.826	0.748
C-MLRSSC	0.833	0.817
MSC_IAS	0.769	**0.854**
CMDC	**0.850**	0.837

表 2　　单维度聚类性能 NMI

算法	AMiner 摘要维度	AMiner 用户维度	MHN 标题维度	MHN 正文维度	MHN 主题维度
k-means	0.679	0.710	0.525	0.825	0.709
MTCUBC	0.783	0.733	0.606	**0.851**	0.621
CMDC	**0.818**	**0.850**	**0.668**	0.837	**0.756**

综合单维度聚类性能，由表 3 可以看出 CMDC 算法在不同特点的 AMiner 和 MHN 数据集上性能都较为稳定，而 MSC_IAS 在差异性较大的数据集（AMiner 数据集）的性能低于基线方法，甚至可能导致丢失单维度数据的有效特征；面向文本数据低秩和稀疏等特点，MLRSSC 系列算法也很好地改善了多维度间的度量差异，但相对于使用互补文本学习和约束文本聚类迭代进行的 CMDC 算法，聚类性能还有约 0.02 的差距。由此证明，CMDC 算法利用多维度文本数据的互补性，有效地解决了因多维度文本数据差异性带来的聚类效果低下的问题。多维度文本数据聚类的维度来源差异性越大、互补性越好，CMDC 算法的聚类结果越好。

4.3.3 多维度文本的互补效果实验

本节实验中，采用逐步提升互补文本对数量的方式来验证其对多维度聚类结果的影响。实验结果如图 2 所示，从整体趋势看，2 个数据集在迭代的过程中都有很好的表现。

图 2(a)所示，在 AMiner 数据集中，用户维度嵌入的特征信息对摘要维度聚类效果的影响表现很稳定，明显上升后开始进入收敛阶段，充分说明在互补性较强的数据集上 CMDC 算法效果显著。反观 MHN 数据集（如图 2(b)所示），正文维度初始聚类 NMI 达到 0.86，加入挑选互补文本集初期有较为明显的下降后进入收敛，产生此结果的原因是 MHN 数据集的维度之间有较强的关联性，使正文维度在整体聚类过程中有最好的结果，而从其他 2 个维度提取的互补文本数据对正文维度的互补性不强，甚至可能提取了含有噪声的文本用于互补，导致性能降低。相应地，标题维度和主题维度中包含的特征较少，效果不佳，而从正文维度中可提取更多的互补文本，呈现较明显的提升趋势。

4.3.4 多维度文本聚类一致性提升效果实验

CMDC 算法使用互补文本对挑选和约束文本聚类模块改善了不同维度之间的差异性问题，使置信度较高的样本在不同维度中得到一致的聚类结果。其中，互补文本旨在捕获不同维度的互补性，通过度量学习调整不同维度、不同类簇的测量方式，从而使式(4)所示的目标函数最小化。以数据集 AMiner 为例，多维度聚类一致性的趋势如图 3 所示。

图 3 展示的是 AMiner 数据集摘要维度使用 3.3 节自动挑选的 900 对互补文本进行约束聚类的情况，其中 694 对与实际样本类别一致，206 对为噪声样本。在约束文本聚类的过程中，一致的约束互补文本对最高攀升到了 568 对，并平均维持在 476 对，说明通过 CMDC 算法使来自用户维度的聚类信息运用到了摘要维度，通过约束文本聚类模块实现了聚类信息的共享；同时统计了在这些一致样本对中与实际样本类别相同的数量，可以看出与一致样本对数量趋势相同，并且差值保持在 90 对左右，这些样本是会影响聚类的噪声样本对。聚类过程中一致样本对数量的震荡也是聚类性能趋势的体现。

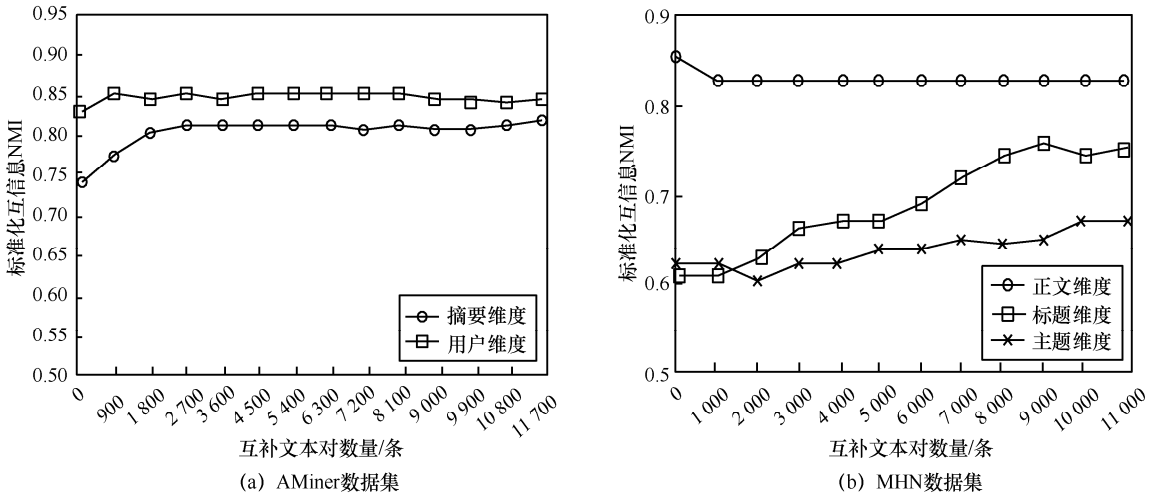


图 2 CMDC 算法在 2 种数据集的 NMI 性能

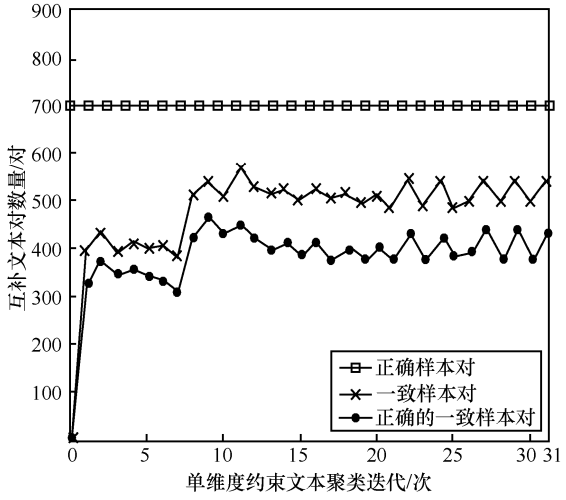


图 3　多维度聚类一致性趋势曲线

总体而言，CMDC 算法自动挑选各维度间的测量不一致的样本作为互补文本，利用基于度量学习的约束文本聚类模块，通过递增互补文本促进聚类，提高不同维度间聚类性能的一致性。CMDC 算法过程可以实现改进不同类别形状达到更好的聚类效果。

5　结束语

本文提出的 CMDC 算法是一种多维度文本聚类算法，算法中的互补文本挑选模块和约束聚类模块通过相互促进的迭代模式形成整体；有效地利用数据的互补性改善了多维度文本的差异性问题，实现聚类结果与度量学习的共同调优。CMDC 算法是基于度量学习在多维度文本聚类算法的改进，其思路亦可以应用于其他算法中，具有很好的通用效果。

本文算法还有需要进一步改进的地方，未来除了学习互补文本做约束外，将探索不同类簇之间潜在的聚类相关性语义[25]，以及在选择约束文本聚类的过程中，解决由低基线维度导致的互补文本集噪声问题。

参考文献：

[1] ALLAHYARI M, POURIYEH S, ASSEFI M, et al. Text summarization techniques: a brief survey[J]. International Journal of Advanced Computer Science and Applications, 2017, 8(10):397-405.

[2] QIAN M, ZHAI C. Unsupervised feature selection for multi-view clustering on text-image Web news data[C]//Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2014: 1963-1966.

[3] YANG Y, WANG H. Multi-view clustering: a survey[J]. Big Data Mining and Analytics, 2018, 1(2): 83-107.

[4] BICKEL S, SCHEFFER T. Multi-view clustering[C]//Industrial Conference on Data Mining. Piscataway: IEEE Press,2004: 19-26.

[5] CHAUDHURI K, KAKADE S M, LIVESCU K, et al. Multi-view clustering via canonical correlation analysis[C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 129-136.

[6] KUMAR A, DAUMÉ H. A co-training approach for multi-view spectral clustering[C]//Proceedings of the 28th International Conference on Machine Learning.Washington:IMLS, 2011: 393-400.

[7] KUMAR A, RAI P, DAUME H. Co-regularized multi-view spectral clustering[C]//Advances in Neural Information Processing Systems. [S.n.:s.l], 2011: 1413-1421.

[8] YIN Q, WU S, HE R, et al. Multi-view clustering via pairwise sparse subspace representation[J]. Neurocomputing, 2015(156): 12-21.

[9] TIAN F, GAO B, CUI Q, et al. Learning deep representations for graph clustering[C]//28th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2014: 1293-1299.

[10] PENG X, XIAO S, FENG J, et al. Deep subspace clustering with sparsity prior[C]//International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 1925-1931.

[11] CAI X, NIE F, HUANG H, et al. Heterogeneous image feature integration via multi-modal spectral clustering[C]//Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2011: 1977-1984.

[12] WANG Y, WU L. Beyond low-rank representations: orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering[J]. Neural Networks, 2018(103): 1-8.

[13] XIE Y, LIN B, QU Y, et al. Joint deep multi-view learning for image clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2020 (99): 1.

[14] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]//International Conference on Machine Learning. New York: IMLS, 2016: 478-487.

[15] PERKINS H, YANG Y. Dialog Intent induction with deep multi-view clustering[J]. arXiv Preprint, arXiv: 1908. 11487, 2019.

[16] XING E P, JORDAN M I, RUSSELL S J, et al. Distance metric learning with application to clustering with side-information[C]// Advances in Neural Information Processing Systems. [S.n.:s.l.], 2003: 521-528.

[17] YE J, ZHAO Z, LIU H. Adaptive distance metric learning for cluster-

ing[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2007: 1-7.

[18] BAGHSHAH M S, SHOURAKI S B. Kernel-based metric learning for semi-supervised clustering[J]. Neurocomputing, 2010, 73(7-9): 1352-1361.

[19] MOUTAFIS P, LENG M, KAKADIARIS I A. An overview and empirical comparison of distance metric learning methods[J]. IEEE Transactions on Cybernetics, 2016, 47(3): 612-625.

[20] HYUN Y, KIM N, CHO Y. A multi-dimensional issue clustering from the perspective consumers' interests and R&D[J]. Journal of the Korea Society of IT Services, 2015, 14(1): 237-249.

[21] 黎万英, 黄瑞章, 丁志远, 等. 基于用户行为特征的多维度文本聚类[J]. 计算机应用, 2018, 38(11): 3127-3131.
LI W Y, HUANG R Z, DING Z Y, et al. Multi-dimensional text clustering with user behavior characteristics[J]. Journal of Computer Applications, 2018, 38(11): 3127-3131.

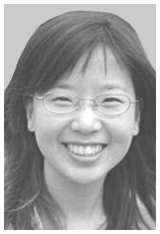
[22] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint, arXiv: 1810. 04805, 2018.

[23] BRBIĆ M, KOPRIVA I. Multi-view low-rank sparse subspace clustering[J]. Pattern Recognition, 2018(73): 247-258.

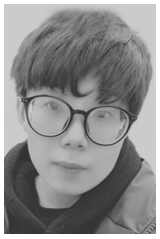
[24] WANG X, LEI Z, GUO X, et al. Multi-view subspace clustering with intactness-aware similarity[J]. Pattern Recognition, 2018(88): 50-63.

[25] RASIWASIA N, MAHAJAN D, MAHADEVAN V, et al. Cluster canonical correlation analysis[J]. Aistats, 2014: 823-831.

[作者简介]



黄瑞章（1979− ），女，天津人，博士，贵州大学副教授、硕士生导师，主要研究方向为数据挖掘、文本挖掘、机器学习和信息检索。



白瑞娜（1994− ），女，山西兴县人，贵州大学硕士生，主要研究方向为文本挖掘、机器学习。

陈艳平（1980− ），男，贵州长顺人，博士，贵州大学副教授、硕士生导师，主要研究方向为人工智能、自然语言处理。

秦永彬（1980− ），男，山东招远人，博士，贵州大学教授、博士生导师，主要研究方向为智慧计算与智能计算、大数据管理与应用。

程欣宇（1978− ），男，贵州绥阳人，贵州大学副教授，主要研究方向为机器学习和计算机视觉。

田有亮（1982− ），男，贵州盘县人，博士，贵州大学教授，主要研究方向为算法博弈论、密码学与安全协议、大数据安全与隐私保护、电子货币与区块链技术。