基于融合矩阵的文本相似度计算实现检索结果聚类*

赵悦阳[1] 崔 雷[2]

([1] 中国医科大学附属盛京医院图书馆 沈阳 110004 [2] 中国医科大学医学健康管理学院 沈阳 110122)

〔摘要〕 **目的/意义** 弥补医学文本语义表示方面的不足，实现 PubMed 数据库检索结果聚类。**方法/过程** 采用 Jaccard 系数和 TF－IDF 构建融合矩阵方法，建立短语间、文档间、短语与文档内容间的相似性关系融合矩阵，训练聚类算法，将 PubMed 数据库检索结果集合分组，随后生成类别标签，描述每一类簇文档的含义。**结果/结论** 基于融合矩阵的聚类效果较好，提取出描述类别的高频词能很好地区分类别含义，对检索结果文本聚类任务有效。

〔关键词〕 文献检索；文本聚类；融合矩阵；文本相似度

〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2024.03.010

A Fusion Matrix－based Study on Text Clustering of Document Retrieval Results

ZHAO Yueyang[1], *CUI Lei*[2]

[1] *Library of Shengjing Hospital of China Medical University, Shenyang 110004, China*; [2] *School of Health Management, China Medical University, Shenyang 110122, China*

〔**Abstract**〕 **Purpose/Significance** To solve the deficiencies in the semantic representation of medical texts, and to realize the clustering of the retrieval results of the PubMed database. **Method/Process** The paper proposes a method to construct a fusion matrix by using the Jaccard coefficient and TF－IDF. Similarity relations between phrases, documents, and the contents of phrases and documents are combined to construct a fusion matrix, and several clustering algorithms are trained to group a collection of documents from the PubMed database. Category annotations are created to describe the meaning of each category of clustered documents. **Result/Conclusion** Experimental results show that the fusion matrix－based clustering is superior in grouping the document sets, and the extracted high－frequency words in the category descriptions distinguish the meanings of the categories well, so the fusion matrix design is effective for clustering descriptions of academic texts.

〔**Keywords**〕 document retrieval; text clustering; fusion matrix; text similarity

1 引言

文本聚类根据语义相似性将文本分组[1]，是网络索引、文本摘要、内容挖掘和信息检索等领域重要的研究手段[2－3]。对 PubMed 搜索结果进行聚类有助于科研人员快速理解相关子主题，并揭示生物医疗研究间的联系。此领域研究的关键是将医学文

〔修回日期〕 2023－12－29

〔作者简介〕 赵悦阳，副研究馆员，发表论文 13 篇。

〔基金项目〕 辽宁省社会科学规划基金资助项目（项目编号：L20BTQ003）。

本的高维稀疏数据转化为实质性的语义表示，进而提高聚类效果。

文本的语义表示对文本聚类的效果至关重要。在传统聚类中，文档通常采用向量空间模型（vector space model，VSM）[4]或词频－逆文档频率（term frequency－inverse document frequency，TF－IDF）[5]表示，但这导致特征空间较大且稀疏，影响了模型性能。与之相对，Jaccard 相似度[6]算法更适合处理高维稀疏数据，能够改进聚类效果[7]。基于“词－上下文”矩阵的语义相似性计算方法，利用语义关系动态构建并分析矩阵[8]，其中矩阵根据不同上下文（如文档、相邻词等）的关系而变化，导致矩阵稀疏度不同。通常，“词－文档”矩阵最为稀疏，对计算效果影响较大。

本研究通过 4 个核心步骤，改善医学文本检索结果的聚类效果。第 1 步：采用一种高效的模型来精确表述文档集合。通过整合融合技术，构建一种保留原邻近性信息及反映 3 种输入相似性矩阵中高阶关系能力的矩阵。第 2 步：为了简化医学文本的复杂性，实施降维处理。基于矩阵的方法将文本数据向量化，得到更易于处理的低维数据表示，以便后续分析和可视化。第 3 步：为了方便用户理解不同的文本集合，开发一种标签生成技术。利用词－文档共现矩阵，对文档中出现的关键词进行统计和排序，进而为每个文本类簇生成能够代表其内容的标签。第 4 步：为确保聚类结果的逻辑性和用户友好性，利用先进的聚类算法来确定聚类的最佳数量。这不仅提高了聚类的准确性，也使结果更容易被终端用户理解和使用。期望通过上述措施为医学信息检索提供一个更精准和用户友好的聚类框架，以便在海量的医学数据中迅速定位重要信息，促进医学研究和实践的效率。

2 相关研究

2.1 聚类算法

常用的文本聚类算法包括基于划分的、基于密度的和基于层次的。K－means[9]是一种广泛使用的划分聚类算法，其中一个聚类中心与该聚类的其他数据点之间的距离平方和被最小化，以获得给定数据集的最佳数据划分。MiniBatch K－means[10]是标准 K－means 的变体，用于处理大数据集。Agglomerative[11]是一种自下而上的层次聚类方法。BIRCH[12]也是一种常见的层次聚类。常用聚类算法的基本原理、主要应用、优势与不足等情况，见表 1。

表 1 常见的聚类算法比较

聚类技术	流行算法	基本原理	主要应用	优势	不足
基于层次的聚类技术	Agglomerative、BIRCH[12]	通过使用启发式“自顶向下”拆分或“自底向上”合并技术生成聚类树（或树状图）	主题发现、员工/客户分组、软件聚类、绩效评价、主题演化、智能问答、舆情分析	定义距离比较容易、自由；不用预先指定集群数目；发现层次间的关系；可以发现任意形状的簇	计算复杂度高，运行慢；对异常值敏感
基于划分的聚类技术	K－means[9]、Fuzzy、C－Means、Expectation－Maximization	将数据集划分为指定数目，通过对聚类中心的迭代重置，达到“聚类内部点足够近，类间点足够远”的目标效果，完成样本集的最优得分	主题划分、数据挖掘、模式识别、金融风控、主题发现	容易实施、灵活，运行快；可扩展性较好；对凸形簇效果好；容易解释	算法依赖于数据；依赖初始条件；得不到全局最优解；要预先指定集群数量
基于密度的聚类技术	DBSCAN[13]	基于“簇”和“噪声”的直观概念，根据样本的紧密程度，将密集区域当作一个一个的聚类簇	弹幕分析、热点话题分析、主题抽取、舆情分析、网络热点和媒体事件监测	不用预先指定集群数量；对噪声不敏感；擅长找到离群点	算法参数复杂，对结果影响大；高维数据聚类有困难

2.2 文本语义表示

2.2.1 表面文本相似度 文本语义表示的主要研究方向包括表面文本相似度和语义相似度[8]。表面文本相似度直接计算原始文本的字符串序列或字符组合的匹配程度或距离，衡量相似度，其研究历史较长，且原理简单、易于实现。其中，N－Gram、Jaccard 系数、Dice 系数、Overlap 系数用于计算字

符串序列或字符组合的匹配程度，比较样本集合之间的相似和差异。基于向量空间模型的方法则是通过空间距离上的相似度表达语义相似度。常用余弦距离、欧式距离将术语表示成向量。但是这种方法为了方便运算，只简单地将文本处理成向量，并不能含有语义信息。Jaccard 系数用于比较样本集之间的相似性和差异性，是计算机领域检验文本相似性的常用方法[14]。由于集合元素互不相似，用于文本相似度计算时不考虑词在文本中出现的频率。谢红[15]基于词频比改进了 Jaccard 相似度算法。医学文本的特征是不同主题之间的关键词基本不同，而同类主题的关键词相似度较大，决定了采用 Jaccard 系数判断医学文本间的相似度是不错的选择[16]。

2.2.2 语义相似度 基于矩阵分布的模型是文本语义表示的另一个方向，其针对词语发生的上下文关联构建矩阵。上下文关系可以是词所在的文档、词与邻近词或目标词的关系，使用的上下文不同，则构成的矩阵稀疏程度也不同，其中“词 - 文档”矩阵最稀疏，计算效果也最差。矩阵里的数值除了使用词与上下文的共现次数表示以外，许多研究还会使用 TF - IDF 或取对数，对元素值实现加权或平滑。最后使用奇异值分解或非负矩阵分解等技术将原始的“词 - 上下文”矩阵从高维稀疏向量压缩为低维稠密向量[17]。

综上，本文提出利用 Jaccard 系数和 TF - IDF 构建融合矩阵的方法，在保留原始邻近信息的同时，还保留 3 个输入相似性矩阵所隐含的高阶关系，以弥补医学文本语义表示方面的不足，在聚类分析 PubMed 检索结果时，提升聚类质量。

3 研究框架与方法

使用 KeyBERT[18]抽取样本的重要关键词，将数据降维。构建融合矩阵，分别训练 K - means，MiniBatch K - means，Agglomerative 和 BIRCH 聚类算法，并以这 4 种算法作为基础训练集成矩阵。对 PubMed 检索结果文献集作聚类分组，生成每一类别的特征词语，随后评价聚类结果和类别标签。本实验通过 Python 3 编程实现。研究框架，见图 1。

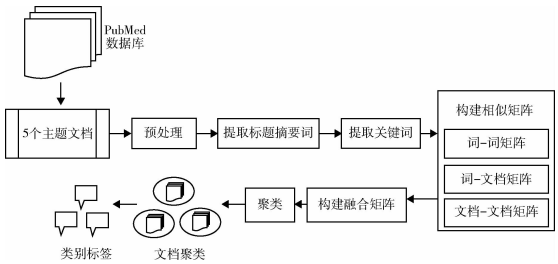

图 1 研究框架

3.1 数据获取

在 PubMed 数据库中以任意 5 个 MeSH 主题词进行检索。实验选取的主题词是：Aneurysm，Infected；Asthenopia；Glomerulonephritis，Membranoproliferative；Pelvic Floor Disorders；Restless Legs Syndrome。分别提取检索结果的标题和摘要，并将其混合，作为数据集。

3.2 抽取文档核心词，生成 top*N* 关键词

应用预训练的 KeyBERT 模型提取样本的 top*N* 关键词。KeyBERT 是非常基本但功能强大的关键词提取方法，使用 BERT 嵌入模型提取词/短语，使用简单语义相似度查找与文档本身最相似的关键词短语，可免除去停用词的步骤[18]。

3.3 构建相似性融合矩阵

3.3.1 构建词 - 词相似度矩阵 设关键词列表为 $W = \{w_1, \cdots, w_n\}$，文档集合为 $D = \{d_1, \cdots, d_m\}$。采用 Jaccard 算法计算词和词的相似性。词语 w 的 word - in - doc 列表 $L_d^w = \{L_{d_1}^{w_1}, \cdots, L_{d_j}^{w_i}, \cdots, L_{d_m}^{w_n}\}$，$L_{d_j}^{w_i} = id\,x_{d_j}$ 表示包含词 w_i 的文档 d_j 的索引值为 $id\,x_{d_j}$。Jaccard 公式为 $J(A,B) = \dfrac{|A \cap B|}{|A \cup B|}$，将 L_d^w 作为输入，计算两词之间的相似性分数。最终生成词 - 词相似性矩阵 $W_{n\times n}$，其中 $W_{ij} = J(L_d^{w_i}, L_d^{w_j})$。

3.3.2 构建词 - 文档相似性矩阵 词 - 文档相似性矩阵可以衡量词语和文档的相似性，采用 TF - IDF 算法计算词与文档的相似性。根据上一步得到的 L_d^w 构建词 - 文档 0 - 1 共现矩阵 $\text{occ\_}WD_{n\times m}$，其中 $\text{occ\_}WD_{i\times j} = 1$ 表示词语 w_i 存在文档 d_j 中。

$occ\_WD_{i\times j} = 1$ 的位置为：

$$TF - IDF(w_i, d_j) = tf(w_i, d_j) \times idf(w_i, d_j), \quad (1)$$

$$tf(w_i, d_j) = \frac{n_{wi}, d_j}{\sum_k n_k, d_j} \quad (2)$$

$$idf(w_i, d_j) = \log \frac{|D|}{1 + |D_{w_i}|} \quad (3)$$

其中，n_{wi}, d_j 指词 w_i 在文档 d_j 中出现的次数，$\sum_k n_{wk}, d_j$ 指文档 d_j 中所有关键词的数量，$|D|$ 指文档总数，$|D_{w_i}|$ 指包含词 w_i 的文档数量。

通过式（1）计算词语和文档的相似性分数，生成一个词 - 文档相似性矩阵 $tWD_{n\times m}$，其中当 $occ\_WD_{i\times j} = 1$ 时，$tWD_{i\times j} = TF - IDF(w_i, d_j)$；当 $occ\_WD_{i\times j} = 0$ 时，$tWD_{i\times j} = 0$。

3.3.3 *构建文档 - 文档相似性矩阵* 采用 Jaccard 算法构建文档 - 文档相似性矩阵，与词 - 词相似性矩阵类似。计算文档 d 的 doc - has - word 列表 $L_w^d = \{L_{w_1}^{d_1}, \cdots, L_{w_i}^{d_j}, \cdots, L_{w_n}^{d_m}\}$，其中 $L_{w_i}^{d_j} = id\,x_{w_i}$，表示存在于文档 d_j 的词语 w_i 的索引值。将 L_w^d 作为 Jaccard 公式的输入，计算两文档间的相似性分数。最终生成一个文档 - 文档相似性矩阵 $D_{m\times m}$，其中 $D_{ij} = J(L_w^{d_i}, L_w^{d_j})$。

3.3.4 *构建融合矩阵* 根据前面得到的 $tWD_{n\times m}$ 和 $D_{m\times m}$ 矩阵，生成融合矩阵 $MD_{m\times m}$。定义：

$$MD_{ij} = simi(d_i, d_j) = D_{ij} + \frac{1}{n}\sum_{k=1}^{n} tWD_{ki} \times tWD_{kj} \quad (4)$$

其中，tWD_{ki} 表示第 k 个关键词与第 i 个文档的相似性分数，tWD_{kj} 表示第 k 个关键词与第 j 个文档的相似性分数，n 表示关键词数量。$\frac{1}{n}\sum_{k=1}^{n} tWD_{ki} \times tWD_{kj}$ 可理解为文档 i 和 j 之间的相似性增益分数。融合矩阵公式揭示了两个文档可以通过其共有的关键词与文档自身的相似性分数的求和平均提升相似性增益。求和后的平均是为了防止某些文档的相似性增益过大而进行的归一化处理。

3.4 训练集成聚类模型

根据 MD 相似矩阵分别训练 K - means，Mini-Batch K - means，Agglomerative 和 BIRCH 聚类算法，将结果作为特征，利用 K - means 训练。具体步骤如下。第 1 步：训练 4 种聚类后，每个聚类结果为一个相似性矩阵 Res_M。第 2 步：利用聚类算法 K - means 对若干个 Res_M 矩阵分别加权聚类，然后将这些聚类结果平均得到一致相似度矩阵 Con_M。第 3 步：基于 Con_M 矩阵，用 K - means 方法获得聚类集成结果。

3.5 聚类评价

使用聚类纯度（purity）、$F1$ 分数和调整兰德系数（adjusted rand index，ARI）评价聚类效果。聚类纯度将聚类的正确样本数除以样本总数，也称为聚类准确度，类似于分类任务中的准确率。

$$purity = (\Omega, C) = \frac{1}{N}\sum_k \max_j |w_k \cap c_j| \quad (5)$$

$F1$ 分数是精确率和召回率的调和平均值，可以准确地评价聚类算法的性能。其将聚类视为决策过程，当且仅当两个文档相似时，分组到同一聚类中。

$$精确率 = \frac{TP}{TP + FP} \quad (6)$$

$$召回率 = \frac{TP}{TP + FN} \quad (7)$$

TP 表示将两个相似的文档分组为一个簇（相同 - 相同），TN 表示将两个不同的文档放入不同的集群（不同 - 不同），FP 表示将两个不同的文档分组到同一个集群中（不同 - 相同），FN 表示将两个相似的文档分组到不同的集群中（相同 - 不同）。

$$F1 = \frac{2\ (精确率 \times 召回率)}{(精确率 + 召回率)} \quad (8)$$

调整兰德系数在 1985 年被提出，假设模型是随机分布的，每一类和类簇上的节点数目是固定的。

$$ARI = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)} \quad (9)$$

3.6 生成聚类特征词语

聚类以后输出聚类标签 $C = \{c_1, \cdots c_i, \cdots, c_m\}$，其中 $c_i \in \{0, \cdots, k\}$，k 表示聚类的簇数量。针对每个簇，利用词 - 文档共现矩阵 $occ\_WD_{n\times m}$ 对文档包含的关键词分别计数并排序，取前 n 个关键词，作为每个聚类簇生成的特征词语。从这 n 个词中，以高频前 5 位的词作为聚类标签词语。

4 实验结果与分析

数据集包括 5 种类别和 7 682 个文档。在聚类阶段选择聚类数目有很多标准[19]，要根据研究目标确定。本研究将聚类数目设定为 2 ~ 10。不同聚类数目前提下表现最好的降维方法和算法，见表 2，聚类数为 5 时，Agglomerative 算法的聚类表现最好。聚类数从 2 到 5 递增时，聚类效果越来越好，而从 6 到 10 则越来越差。聚类数目为 5 时不同聚类算法的聚类效果，见表 3，t - SNE 降维后 Agglomerative 的聚类效果最佳。

表 2 不同聚类数目设定下表现好的聚类算法

聚类算法	降维方法	聚类数（个）	聚类纯度	ARI	*F*1
K - means	t - SNE	2	0. 63	0. 476	0. 657
Agglomerative	t - SNE	3	0. 84	0. 792	0. 853
Agglomerative	t - SNE	4	0. 924	0. 939	0. 955
Agglomerative	t - SNE	5	0. 983	0. 972	0. 979
Agglomerative	t - SNE	6	0. 983	0. 818	0. 859
Agglomerative	t - SNE	7	0. 983	0. 675	0. 739
Agglomerative	t - SNE	8	0. 983	0. 602	0. 672
Agglomerative	t - SNE	9	0. 983	0. 542	0. 617
Agglomerative	t - SNE	10	0. 983	0. 479	0. 574

表 3 基于融合矩阵和共现矩阵的聚类算法结果（k=5）

矩阵构建	聚类算法	降维方法	聚类数（个）	聚类纯度	ARI	*F*1
融合矩阵	K - means	PCA	5	0. 889	0. 855	0. 891
		t - SNE	5	0. 908	0. 824	0. 867
	MiniBatch K - means	PCA	5	0. 900	0. 761	0. 818
		t - SNE	5	0. 917	0. 789	0. 840
	Agglomerative	PCA	5	0. 885	0. 866	0. 899
		t - SNE	5	0. 983	0. 972	0. 979
	Birch	PCA	5	0. 907	0. 881	0. 911
		t - SNE	5	0. 927	0. 834	0. 874
	Stack	t - SNE	5	0. 751	0. 548	0. 658
共现矩阵	K - means	PCA	5	0. 889	0. 855	0. 890
		t - SNE	5	0. 907	0. 823	0. 866
	MiniBatch K - means	PCA	5	0. 899	0. 761	0. 817
		t - SNE	5	0. 919	0. 794	0. 844
	Agglomerative	PCA	5	0. 917	0. 902	0. 927
		t - SNE	5	0. 926	0. 782	0. 834
	Birch	PCA	5	0. 903	0. 779	0. 832
		t - SNE	5	0. 926	0. 782	0. 834
	Stack	t - SNE	5	0. 790	0. 532	0. 638

比较基于融合矩阵与共现矩阵的结果，Agglomerative 算法基于 PCA 降维后共现矩阵效果好，但是 t - SNE 降维后共现矩阵效果远不如融合矩阵。BIRCH 算法同样如此。基于融合矩阵的集成算法（Stack）综合看略好于共现矩阵的结果。但是集成算法的结果并没有单独一种聚类算法有优势。说明融合矩阵的设计可以免去集成聚类的烦琐步骤，又能提高聚类效能。t - SNE 降维后的 Agglomerative 算法 5 个类别的颜色没有重叠或分布不均，类内的点足够近，类间的点足够远，证明文献集得到很好的分组，见图 2。

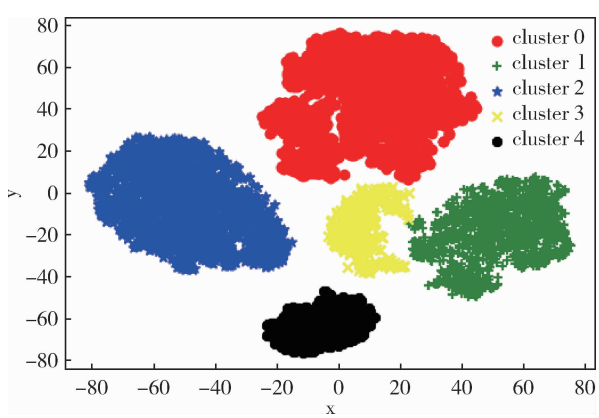

图 2 基于 t - SNE 降维后的 Agglomerative 算法聚类效果

将聚类后类簇数量与原始文档数比较，簇 1 和簇 2 的数量与原始文档接近；而簇 3、簇 4 和簇 0 基于融合矩阵的数量与原始文档数更接近，基于共现矩阵的簇 4 和簇 0 则相差很多，见表 4。

表 4 Agglomerative 算法聚类后类簇数量与原始文档数比较

文档名	簇 ID	基于共现矩阵	基于融合矩阵
Aneurysm_ Infected. txt	2	2 275/2 308	2 275/2 308
Asthenopia. txt	3	681/470	583/470
Glomerulonephritis_ Membranoproliferative. txt	1	1 601/1 638	1 601/1 638
Pelvic_Floor_Disorders. txt	4	1 561/697	681/697
Restless_Legs_Syndrome. txt	0	1 564/2 569	2 542/2 569
总数		7 682/7 682	7 682/7 682

本文开展了预实验，分别使用 Cosine、Dice 和 Jaccard 相似性构建融合矩阵，结果显示，使用 Jaccard 训练 Agglomerative 算法的结果最好。2016 年 Mu T 等[20]构建 CEDL 框架，将词 - 词、词 - 文档和文档 - 文档矩阵共同嵌入到框架内，同时完成文档聚类和标签生成。这种构建嵌入框架的方式是通过计算余弦相似度。本文构建的融合矩阵是通过 Jaccard 计算文本相似度，对于有重复词语的文本，如果应用余弦计算相似度会变化，而应用 Jaccard 计算的重复词的相似度则不变[21]。这是 Jaccard 的优势，而且本实验也证实了其较共现矩阵得到的聚类效果好。

分别归纳每一类别中融合矩阵生成的频次最高的 30 个词语。通过 Agglomerative 算法生成的特征词只出现在自己类里，其他 3 类中没有出现，充分满足聚类标签表现类别含义的要求。而其他 3 种聚类得到的特征词，都存在高频词在其他类中出现的情况，比如 aneurysm，aortic。

对于每个簇，按出现频次由高到低，选择前 5 个短语作为标签词语，并与下载文献时选取的 MeSH 主题词比较，见表 5。Asthenopia 是视疲劳，标签词语是 visual，symptom，work，symptoms 和 ocular，组合起来是眼睛工作带来的症状，与视疲劳接近；Glomerulonephritis_Membranoproliferative 是肾小球肾炎膜增生，标签词语是 nephritis，life，proliferative，deposit 和 renal，说的是肾脏某些物质增殖产生沉淀导致肾炎对生命有影响，也有增生肾炎的意思。其他 3 类标签词语基本与主题词含义相同。

表 5 标签词语与 MeSH 主题词比较

聚类数目	标签词语（5 个）	MeSH 主题词（1 个）
2	aneurysm；case；aortic；report；aorta	Aneurysm_Infected
3	visual；symptom；work；symptoms；ocular	Asthenopia
1	nephritis；life；proliferative；deposit；renal	Glomerulonephritis_Membranoproliferative
4	floor；pelvic；pelvic floor；women；disorder	Pelvic_Floor_Disorders
0	syndrome；legs；restless leg；restless；restless legs	Restless_Legs_Syndrome

综上，基于融合矩阵的聚类效果整体比基于文档共现矩阵的好，文献集分组良好；基于融合矩阵训练集成聚类时，聚类效能反而下降，说明融合矩阵的设计可免去集成聚类的烦琐步骤，用简单的 Agglomerative 算法就能提高聚类效能；同时提取的描述类别的高频词也能很好地区别类别含义，所以融合矩阵的设计对于学术文本聚类描述的任务是有效的。

5 结语

本文研究通过构建融合矩阵来提高文本聚类效果的方法，主要探讨 4 个问题：如何构建有效的融合矩阵；基于融合矩阵的聚类效果是否有所提升；提出基于融合矩阵集成聚类的方法，并探讨其聚类效果；生成聚类标签，描述聚类结果。通过利用融

合矩阵文本聚类效果显著提升，不仅在文档分类方面表现出色，也能有效提取有代表性的词汇来描述每个类别的特征，证实该方法在处理学术文本聚类任务方面的有效性。

本文仍存在不足之处。一是数据集仅使用 PubMed 数据库，应将融合矩阵的构建扩展到更多类型的文本资料库，验证聚类效果。二是提取的聚类描述关键词的语义粒度较粗，只能反映查询结果内主题间的关系。应根据细粒度文本表示概念[22]，开展知识单元细粒度层面的分析，例如论文中的研究范畴、方法、数据、指标、指标值等信息提取，为医学数据的智能分析提供重要的方法支撑。

未来可通过本文的研究方法将 PubMed 的检索结果聚类，通过提供关键词简明扼要地描述检索结果，反映特定查询的内容分布。在此基础上，通过细粒度分析，从文本中识别研究范畴、研究方法、实验数据和评价指标及取值等知识，推出面向不同知识层次的定制服务，提高医学图书馆的服务质量。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 薛菁菁，秦永彬，黄瑞章，等．SSVAE：一种补充语义信息的深度变分文本聚类模型［J］．数据分析与知识发现，2022，6（6）：71－83.

2 冯小东，惠康欣．基于异构图神经网络的社交媒体文本主题聚类［J］．数据分析与知识发现，2022，6（10）：1－14.

3 刘婷，张娴，许海云，等．面向技术路径识别的文本挖掘方法应用研究述评［J］．情报理论与实践，2020，43（7）：179－185.

4 牛奉高，张亚宇．基于共现潜在语义向量空间模型的语义核构建［J］．情报学报，2017，36（8）：834－842.

5 肖悦珺，李红莲，张乐，等．特征融合的中文专利文本分类方法研究［J］．数据分析与知识发现，2022，6（4）：49－59.

6 高洪臻．基于杰卡德相似系数的 OPAC 用户检索行为研究［J］．图书馆研究与工作，2022（6）：74－79.

7 张晓琳，付英姿，褚培肖．杰卡德相似系数在推荐系统中的应用［J］．计算机技术与发展，2015，25（4）：158－161，165.

8 赖辉源，王春柳，杨永辉，等．文本相似度计算方法研究综述［J］．情报科学，2019，37（3）：158－168.

9 孙海霞，李军莲，吴英杰．基于 K－means 的机构归一化研究［J］．医学信息学杂志，2013，34（7）：41－44，71.

10 SCULLEY D. Web－scale k－means clustering［C］. North Carolina：International Conference on World Wide Web，2010.

11 NIELSEN F. Introduction to HPC with MPI for data science［M］. Switzerland：Springer International Publishing，2016.

12 杨秀璋，夏换，于小民，等．基于特征词典构建和 BIRCH 算法的中文百科文本聚类研究［J］．计算机时代，2019（11）：23－27，31.

13 陈氢，冯进杰．融合地理标签数据的个性化信息服务应用研究［J］．现代情报，2019，39（10）：24－31.

14 于鹏．逻辑公式间的 Jaccard 距离及其应用［J］．计算机科学与探索，2020，14（11）：1975－1980.

15 谢红．基于词频比的改进 Jaccard 系数文本相似度计算［J］．内江科技，2021，42（8）：27.

16 王安瑾．一种基于 MinHash 的改进新闻文本聚类算法［J］．计算机技术与发展，2019，29（2）：39－42.

17 来斯惟．基于神经网络的词和文档语义向量表示方法研究［D］．北京：中国科学院大学，2016.

18 Github. KeyBERT［EB/OL］.［2023－05－19］. https：//maartengr.github.io/KeyBERT/.

19 NIASI K，SIDHESHWARI P. Self－tuned descriptive document clustering using a predictive network［J］. IEEE transactions on knowledge and data engineering，2018，30（10）：1929－1942.

20 MU T，GOULERMAS J Y，KORKONTZELOS I，et al. Descriptive document clustering via discriminant learning in a co－embedded space of multilevel similarities［J］. Journal of the American society for information science and technology，2016，67（1）：106－133.

21 LEYDESDORFF L. On the normalization and visualization of author co－citation data：Salton's cosine versus the Jaccard index［J］. Journal of the American society for information science and technology，2008，59（1）：77－85.

22 余丽，钱力，付常雷，等．基于深度学习的文本中细粒度知识元抽取方法研究［J］．数据分析与知识发现，2019，3（1）：38－45.