Journal of Computer Applications
计算机应用,2020,40(4):1069－1073
ISSN 1001-9081
CODEN JYIIDU
2020-04-10
http://www.joca.cn

文章编号:1001-9081(2020)04-1069-05
DOI:10.11772/j.issn.1001-9081.2019091540

基于共识和分类改善文档聚类的识别信息方法

王留洋*,俞扬信,陈伯伦,章　慧
(淮阴工学院 计算机与软件工程学院,江苏 淮安 223003)
(*通信作者电子邮箱 wangly@hyit.edu.cn)

摘　要:不同的聚类算法用于设计各自的策略,然而,每种技术在执行特定数据集时都有一定的局限性。选择恰当的识别信息方法(DIM)可确保文档聚类的进行。针对这些问题提出一种基于共识和分类的文档聚类(DCCC)的DIM。首先,选择识别信息最大化聚类(CDIM)作为数据集生成初始聚类的解决方法,并使用两种不同的CDIM方法生成两个初始聚集;其次,使用不同的参数方法对两初始聚集再进行初始化,通过簇标签信息间的关系建立共识,最大限度地提高文档的识别数总和;最后,选择识别文本权重分类(DTWC)作为文本分类器给共识分配新的簇标签,通过训练文本分类器更改基础分区,并根据预报标签信息生成最后的分区。采用8个网络数据集进行实验,选择BCubed的精度和召回率指标进行聚类验证。实验结果表明,所提出的共识分类方法的聚类结果优于对比方法的聚类结果。

关键词:共识聚类;文档聚类;识别信息;簇标签;文本分类器
中图分类号:TP391.3　　**文献标志码:**A

Discrimination information method based on consensus and classification for improving document clustering

WANG Liuyang*, YU Yangxin, CHEN Bolun, ZHANG Hui
(*Faculty of Computer & Software Engineering, Huaiyin Institute of Technology, Huai'an Jiangsu* 223003, *China*)

Abstract: Different clustering algorithms are used to design their own strategies. However, each technology has certain limitations when it executes a particular dataset. An adequate choice of Discrimination Information Method (DIM) can ensure the document clustering. To solve these problems, a DIM of Document Clustering based on Consensus and Classification (DCCC) was proposed. Firstly, Clustering by DIM (CDIM) was used to solve the generation of initial clustering for dataset, and two initial cluster sets were generated by two different CDIMs. Then, two initial cluster sets were initialized again by different parameter methods, and a consensus was established by using the relationship between the cluster label information, so as to maximize the sum of documents' discrimination number. Finally, Discrimination Text Weight Classification (DTWC) was chosen as text classifier to assign new cluster label to the consensus, the base partitions were altered by training the text classifier, and the final partition was obtained based on the predicted label information. Experiments on 8 network datasets for clustering verification by BCubed's precision and recall index were carried out. Experimental results show that the clustering results of the proposed consensus and classification method are superior to those of comparison methods.

Key words: consensus clustering; document clustering; discrimination information; cluster label; text classifier

0　引言

目前,大量数据通过不同的渠道源源不断地汇集到互联网数据库中,并有针对性地作为互联网的副产品进入社会。网络信息的巨大容量对搜索引擎的信息检索效率造成了巨大威胁。在搜索引擎返回数千个搜索结果中,只有极少部分与用户感兴趣的话题有关[1]。对无标记数据可进行无监督数据分析。数据聚类是一种无监督学习技术,聚类利用数据中的非结构化信息揭示数据间潜在的关系[2]。

个人资料库和公共论坛中的信息通常以文本形式出现,这样的数据无法用传统数据库工具进行抓取、管理和处理。文本数据一般属于未标记数据类别,可通过数据聚类技术进行分析。数据聚类技术在计算机中的应用已超过50年,现有的技术有共识聚类、多视图聚类、集合聚类和累积聚类等。

近年来,不同的聚类算法用于设计各自的策略,然而,每种技术在执行特定数据集时都有一定的局限性。因此,缺少一种通用的数据聚类方法,支持不同技术的混合聚类,即共识聚类。共识聚类可通过以下几步实现:首先,划分训练数据,并使用不同聚类方法对每个数据进行分区,以便获得基本结果;其次,使用相同的算法,进行不同的初始化或使用不同的参数,通过聚类技术间的关系建立共识;最后,使用不同的特

收稿日期:2019-09-05;**修回日期:**2019-10-23;**录用日期:**2019-10-24。　**基金项目:**国家自然科学基金资助项目(61602202)。
作者简介:王留洋(1974—),男,江苏淮安人,副教授,硕士,主要研究方向:信息管理与信息系统、智能化信息处理、大数据挖掘;　俞扬信(1970—),男,江苏泰州人,教授,硕士,主要研究方向:信息管理与信息系统、智能化信息处理、知识组织;　陈伯伦(1986—),男,江苏淮安人,副教授,博士,主要研究方向:复杂网络的链路预测;　章慧(1970—),女,江苏南通人,教授,硕士,主要研究方向:信息管理与信息系统、智能化信息处理。

征空间给共识分配新标签，并在第一时间通过聚类更改基础分区，设计最终的聚类方案[3]。本文提出了一种使用监督学习方法集成文档聚类的共识构建方法，该共识构建方法是现有文档聚类的共识聚类技术的新补充，共识功能使用了一种基础聚类生成的标签训练分类器，且训练分类器通过预测标签生成最后的分区。

1 相关研究

数据聚类是开发未标记数据的底层结构的基本工具之一。文本数据种类繁多，数据聚类技术需不断发展，以适应未来的挑战。目前，大多数设计的数据聚类方法都具有已知的潜力，但在大多数情况下，没有一种通用的技术能够保持良好的性能。可靠的解决方案是将不同的聚类技术融合成一个单一策略。

将多种聚类技术结合起来进行最终数据划分的想法启发于分类器和信息融合中使用的类似策略启发。与此类似，K均值(K-means)方法先从n个数据对象任意选择k个对象作为初始聚类中心，然后计算每个所获新聚类的聚类中心，进行多次K均值聚类分区，最后直到每个聚类不再发生变化为止，以找出结果中的任何关联。然而，已有文献尚未意识到数据分布对K均值算法的影响，并未理解算法与数据具有很强的耦合关系。文献[4]提出了一种使用共同关联样本矩阵映射相关联的方法，通过在最终分区的关联矩阵上应用基于最小生成树(Minimum Spanning Tree, MST)的聚类，对任意形状的聚类进行扩展。Fred等[5]利用相似矩阵上的单链路和平均链路方法对EAC(Evidence Accumulation Clustering)的计算缺陷进行分类。和一般的集成聚类不同，EAC并不直接组合不同的划分，而是由这些不同的划分得到一个邻近度矩阵。之后便可在这个邻近度矩阵上运用层次聚类中的单连接算法得到最终的划分，即从多个分区中获得的公共信息，进行最终聚类。

聚类融合结合了多种数据集群技术，可以生成初始标签并形成最终统一的群聚解决方案。据了解，在早期的划分步骤中，在构造和积累知识时会丢失一些有价值的信息，如关联矩阵这样的中间表征特征缺少一些参与聚类技术的数据条目。文献[6]研究了丢失信息对融合聚类结果的影响，并提出了一种通过聚集间的相似性链接方法，通过揭示集群之间的相似性来猜测未知条目。最后，利用图分割技术得到最终的聚类结果。投影聚类集合结合了输入数据子空间聚类和集合聚类本身，利用集合聚类的知识对数据子空间进行优化。文献[7]提出了自适应集成聚类，并设计了一种输入数据子采样的自适应集成策略。子样本利用了以往的聚类结果，并强调在共识过程中存在不良历史的样本。聚类可以利用分类技术，通过比较基本聚类方法的比例结果来达成共识。

多视图聚类的灵感来自多视角学习的概念。多视图聚类通过对多维数据进行预处理和对处理后的数据进行聚类投影、分类解决聚类时所涉及的技术[8]。其中一个技术是描述无监督学习问题的共正则化，文献[9]提出了一个光谱聚类目标函数，该函数隐式地将来自多个数据视图的图结合起来，以获得更好的聚类。多视图聚类的另一个技术是使用标准化来解决共识的非负矩阵分解(Non-negative Matrix Factorization, NMF)。它从多个视图中学习到的集群结构的表示应该规范化，以达成共识。类似的想法将多流形正则化纳入NMF，从而保留了数据空间的局部几何结构。

共识聚类有助于确定最佳的聚类集。文献[10]提出了一种在分布式环境下实现幂迭代聚类的方法。利用某种相似性度量方法，将原始数据转换成一个可以视为图的亲和矩阵。通过顶点切割，把行归一化后的亲和矩阵切分成若干个小图，图的每一个划分子图对应一个类簇。通过从多个聚类算法形成共识矩阵来猜测聚类的数量，根据关联图的近似非耦合结构受益的有效性，确定最佳聚类集。

上下文中最新想法是将不同数据分区视图结果与来自不同集合技术的相似矩阵结合起来，为共识决策计算出最终的相似矩阵[11]。计算机矩阵的相似性度量方法有：基于聚类的相似性矩阵、相似性矩阵和成对相异性矩阵。相似矩阵中分配的权重用于聚集。共识聚类的基本原理是通过几种策略得出的：使用具有相似参数的不同聚类技术、使用具有不同聚类参数的单一策略或两者的组合。另一个想法是对训练数据进行分区，并使用不同聚类方法对数据进行分区，以获得基本结果。聚类的结果来自于早期的基础阶段。共识可以使用投票公式分配新的标签[12]。著名的共识策略有：层次凝聚-聚类共识、最远共识、基于聚类的共识、期望最大化共识、迭代投票共识和基于片段的聚类共识[13]。

近年来，共识聚类在文档聚类中得到了广泛的应用和有效的使用。Topchy于2003年使用QMI(Quadratic Mutual Information)作为效用函数，提出利用K均值算法来解决共识聚类的问题，即将共识聚类问题在QMI下转化成经典的K均值优化问题[14]，如何种效用函数的效果比较优越、如何处理样本不一致问题以及其收敛性等，都亟待解决。共识聚类相比单一聚类算法的优势体现在鲁棒性、新颖性、稳定性、并行性和扩展性等[15]。然而，共识聚类是一个具有挑战性的工作，其主要难点在于从不同聚类结果中求出一个共识划分，使得共识效用函数最大。学者从不同角度解释基础聚类器产生聚类结果的共性，从而找出共识聚类结果。

本文提出一种基于共识和分类的文档聚类(Document Clustering by Consensus and Classification，DCCC)策略，实现共识和分类文档自适应集成聚类。利用基于关键词DIM(Discrimination Information Method)的不同文档聚类算法的优势，揭示它们之间的潜在关系。从有关文献中可明显地看出，每个DIM都利用来自不同潜在客户的数据，因此会发现不同的聚类解决方案[16]。然而，簇重叠最有可能出现在每个解决方案中，并且有一些文档与其聚类解决方案相符。簇解决方案可以将这些文档保持在相似的簇中。这些带有簇标签的融合文档有助于训练共识分类器[17]。分类器使用从早期聚类解决方案中发现的知识来预测与初始聚集簇标签信息不一致的文档，从而确定最终的聚类解决方案。聚类利用分类技术是本文研究的主要焦点。

2 文档聚类识别信息方法

近年来，利用文档的K簇识别信息的文档聚类算法的高性能已得到有效的证明。这些聚类算法迭代地将文档投影到K维识别信息空间上，并将有最大值的簇分配给一个文档。在每次迭代中，关键词识别信息定义了识别信息空间，其中关键词识别信息是根据上一次迭代中生成的标记文档集进行估计的。该聚类方法作为优化问题被提出，目标函数可以使用各种可用的统计和信息理论度量[18]。人们经常使用以下方法

发布目标函数的结果:相对危险度(Relative Risk, RR)、识别信息方法(Method of Discrimination Information, MDI)、域相关性(Domain Relevance, DR)和域识别(Domain Consensus, DC)。

使用发布函数方法可能带来另一个风险,即在相同数据的不同解决方案之间如何选择最佳方案。虽然不同的聚类解决方案可选择不同的方法来计算最终簇数,但可计算的簇数之间存在显著的相似性。因此,对来自不同聚类方法的同一簇中的文档须有统一的簇标签。文本分类器使用这些同义文档的簇标签进行训练,随后该分类器再用于预测不同文档的簇标签,使用共识和分类改善文档聚类的工作流程如图1所示,图2是DCCC过程总结的可视化。

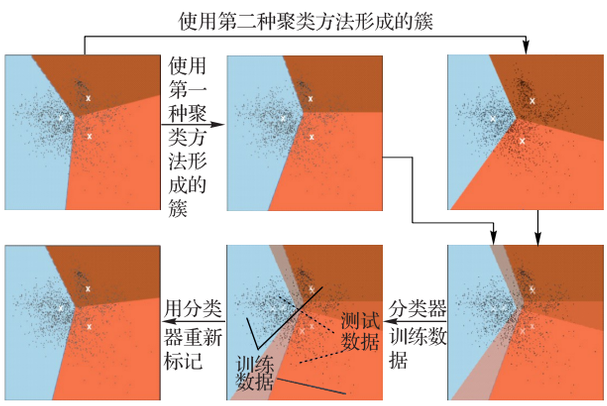

图1　本文方法的简单工作流程

Fig. 1　Simple flowchart of the method proposed in this paper

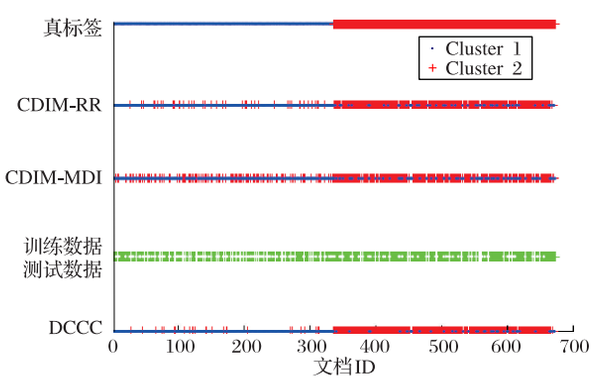

图2　DCCC过程总结的可视化

Fig. 2　Visualization of DCCC process summary

2.1　使用CDIM初始聚类

本文选择CDIM作为数据集生成初始聚类的解决方法。CDIM是一种迭代分区文档聚类方法,在从M维输入空间转换的K维识别信息空间中找到K组文档,其中M表示词汇表中不同关键词的总数。通过有效的文挡投影和分配可实现这一目标,最大限度地提高文档的识别数总和。CDIM-RR和CDIM-MDI是CDIM的两种变体,用来寻找初始聚集E_{RR}和E_{MDI}。第一类使用RR,第二类使用MDI、CDIM-RR和CDIM-MDI看作是CDIM识别项加权策略。

2.1.1　CDIM-RR

用CDIM-RR时,关键词x_j在簇C_k中的RR高于剩余簇$\overline{C}_k$的。簇C_k中的关键词x_j的识别信息(如w_{jk})和剩余簇$\overline{C}_k$中的关键词x_j的识别信息(如$\overline{w}_{jk}$)由式(1)~(2)给出。

$$w_{jk}=\begin{cases}\dfrac{p(x_j|C_k)}{p(x_j|\overline{C}_k)}, & p(x_j|C_k)-p(x_j|\overline{C}_k)>0\\ 0, & \text{其他}\end{cases}\tag{1}$$

$$\overline{w}_{jk}=\begin{cases}\dfrac{p(x_j|\overline{C}_k)}{p(x_j|C_k)}, & p(x_j|\overline{C}_k)-p(x_j|C_k)>0\\ 0, & \text{其他}\end{cases}\tag{2}$$

其中:$p(x_j|C_k)$是簇C_k中的关键词x_j的条件概率。关键词识别信息要么为0(无识别信息),要么大于1,较大值表示有较强的识别能力。

2.1.2　CDIM-MDI

用CDIM-MDI时,MDI用于计算关键词的识别信息,量化关键词间的语义相关性。类别1和类别2分别定义为:ifd_{I1}和ifd_{I2}。在文档聚类中,ifd_{I1}和ifd_{I2}分别对应于簇C_k和簇$\overline{C}_k$。方法定义如下:

$$ifd_{I1}(x_j)=p(x_j|C_k)\lg\frac{p(x_j|C_k)}{\lambda_1 p(x_j|C_k)+\lambda_2 p(x_j|\overline{C}_k)}\tag{3}$$

$$ifd_{I2}(x_j)=p(x_j|\overline{C}_k)\lg\frac{p(x_j|\overline{C}_k)}{\lambda_1 p(x_j|C_k)+\lambda_2 p(x_j|\overline{C}_k)}\tag{4}$$

其中:λ_1和λ_2分别是C_k和$\overline{C}_k$的已有概率。关键词x_j的识别具有以下不平等特征:

$$\psi_1=\lambda_1 ifd_{I1}-\lambda_2|ifd_{I2}|>0\tag{5}$$

$$\psi_2=\lambda_2 ifd_{I2}-\lambda_1|ifd_{I1}|>0\tag{6}$$

如果满足不等式(5),则关键词x_j支持ifd_{I1}超过ifd_{I2};当满足不等式(6),则关键词x_j支持ifd_{I2}超过ifd_{I1}。根据不等式(5)~(6),CDIM的w_{jk}和$\overline{w}_{jk}$的识别项权重如下:

$$w_{jk}=\begin{cases}\psi_1, & \psi_1>0\\ 0, & \text{其他}\end{cases}\tag{7}$$

$$\overline{w}_{jk}=\begin{cases}\psi_2, & \psi_2>0\\ 0, & \text{其他}\end{cases}\tag{8}$$

2.2　寻找共识

在分别使用CDIM-RR和CDIM-MDI获得两个簇集E_{RR}和E_{MDI}后,下一步就是将这两簇集结合起来,寻找相符的文档。由于无监督特性的原因,导致CDIM-RR和CDIM-MDI分配给文档的簇标签不同。为了解决这一问题,需将每个簇$C_i\in E_{\mathrm{RR}}$和$C_j\in E_{\mathrm{MDI}}$进行比较,将E_{RR}和E_{MDI}中最相似的两个聚类分配相同的簇标签,依次共产生K个簇标签,并使用Jaccard指数计算两簇集的相似度。Jaccard相似度为:C_i和C_j交集的大小与并集大小的比值,即

$$J(C_i,C_j)=\frac{|C_i\cap C_j|}{|C_i\cup C_j|};\ \forall C_i\in E_{\mathrm{RR}},C_j\in E_{\mathrm{MDI}}\tag{9}$$

相似度值越大说明两聚类集共识文档数越多。当C_i和C_j都为空时,$J(C_i,C_j)=1$。

将$E^{\mathrm{c}}_{\mathrm{RR}}$和$E^{\mathrm{c}}_{\mathrm{MDI}}$输出(上标c表示簇$E_{\mathrm{RR}}$和$E_{\mathrm{MDI}}$对应的簇标签),通过$E^{\mathrm{c}}_{\mathrm{RR}}$和$E^{\mathrm{c}}_{\mathrm{MDI}}$可找到具有初始聚类共识文档,即$E^{\mathrm{c}}_{\mathrm{RR}}$和$E^{\mathrm{c}}_{\mathrm{MDI}}$中具有相同簇标签的所有文档。

2.3　使用分类器进行改进

位于CDIM-RR和CDIM-MDI同一簇中的文档很可能正好聚集在一起,将这些文档与它们的簇标签信息一起生成文本分类器的训练集。

$$X_{\mathrm{train}}=\{x_i|E^{\mathrm{c}}_{\mathrm{RR}}(i)=E^{\mathrm{c}}_{\mathrm{MDI}}(i),\forall x_i\in X\}\tag{10}$$

其中:X_i是数据集X中第i个文档。

对于没有在CDIM-RR和CDIM-MDI同一簇中的剩余文档

则生成文本分类器的测试集。

$$X_{\text{test}} = X - X_{\text{train}} = \{ x_i | E^{\text{c}}_{\text{RR}}(i) \neq E^{\text{c}}_{\text{MDI}}(i), \forall x_i \in X \} \quad (11)$$

本文选择DTWC文本分类器。DTWC是一种基于识别式加权的线性识别方法，可用于文本分类，而且实践表明具有良好的分类结果[19]。尽管可以使用任何文本分类器，但除了其效率和良好结果外，选择DTWC的另一个原因是该方法的识别特性与初始聚类方法CDIM相匹配。E^{c}_f是簇标签的最终列表，融合了训练集中文档的共识簇标签$E^{\text{c}}_{\text{CDIM}}$和测试集中被DTWC预测的文档簇标签$E^{\text{p}}_{\text{DTWC}}$。算法1为本文算法的流程。

算法1 DCCC。

输出 X(关键词-文档数据集)，K(簇编号)。

Step 1 $E_{\text{RR}} \leftarrow$CDIM-RR(X,K)。

Step 2 $E_{\text{MDI}} \leftarrow$CDIM-MDI($X$,$K$)。

Step 3 $E^{\text{c}}_{\text{RR}}, E^{\text{c}}_{\text{MDI}} \leftarrow$解决($E_{\text{RR}}, E_{\text{MDI}}$)匹配问题。

Step 4 指标(训练集)$\leftarrow$查找($E^{\text{c}}_{\text{RR}}=E^{\text{c}}_{\text{MDI}}$)。

Step 5 $X_{\text{train}} \leftarrow X$(指标(训练集))。

Step 6 $E^{\text{c}}_{\text{CDIM}} \leftarrow E^{\text{c}}_{\text{RR}}$(指标(训练集))。

Step 7 指标(测试集)$\leftarrow$查找($E^{\text{c}}_{\text{RR}} \neq E^{\text{c}}_{\text{MDI}}$)。

Step 8 $X_{\text{test}} \leftarrow X$(指标(测试集))。

Step 9 $E^{\text{p}}_{\text{DTWC}} \leftarrow$DTWC($X$(测试集)，$X$(训练集))。

Step 10 $E^{\text{c}}_f \leftarrow E^{\text{c}}_{\text{CDIM}} \cup E^{\text{p}}_{\text{DTWC}}$。

Step 11 返回E^{c}_f。

3 实验与结果分析

3.1 数据集

在8个网络数据集上评估了本文的聚类方法DCCC。表1给出了这些网络数据集的关键特征。数据集1、3到8进行预处理，同时对数据集2进行了停用词的删除和封堵。

表1 数据集及其特征

Tab. 1 Datasets and their characteristics

编号	数据集	文档数N	关键词数M	种类数K
1	pu	672	19 868	2
2	movie	1 200	38 408	2
3	citeseer	3 312	38 408	6
4	hitech	2 301	13 170	6
5	tr31	927	10 128	7
6	cora	2 708	1 433	7
7	re0	1 504	2 886	13
8	wap	1 560	8 460	20

数据集pu是从Internet Content Filtering Group的网站获得，包含标记为垃圾邮件或非垃圾邮件的特定用户收到的电子邮件；数据集movie来自Internet电影数据库(Internet Movie DataBase，IMDB)的电影评论，每个电影文档都有正面或负面评论；hitech、tr31、re0和wap数据集来自明尼苏达大学的Karypis实验室；数据集hitech来源于圣何塞水星报，这些文章作为TREC(Text Retrieval Conference)系列的一部分发布；数据集tr31来自TREC-6，此数据集中的查询类别与其最相关的文档对应；数据集re0来自Reuters-21578文本分类测试集分发版1.0；数据集wap是从WebACE项目中获得的，每个文档对应于Yahoo！主题层结构中列出的网页；数据集cora是一个指标矩阵，表示文档中某个关键词的存在或不存在；citeseer是出现在网站上的文章的集合，这些文章的视图与关键词-文档矩阵相对应。

3.2 聚类验证方法

选择BCubed的精度和召回率指标进行聚类验证。每个文档的精度(BCubed Precision，BP)和召回率(BCubed Recall，BR)都可计算，如图3所示。根据BP和BR的值可计算(Bcubed F-measure，BF)，公式为$BF = 2 \times \dfrac{BP \times BR}{BP + BR}$，$BF \in (0, 1)$，值越大表示聚类效果越好。

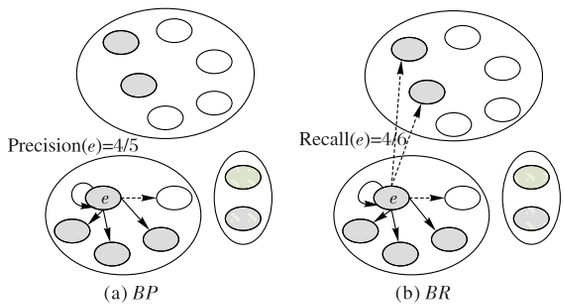

图3 计算一个文档的BP和BR示例

Fig. 3 Example of calculating BP and BR for a Document

3.3 实验结果与分析

通过CDIM进行聚类可获得高质量结果，但需在方法RR和MDI之间进行选择。

尽管CDIM-RR和CDIM-MDI的性能不同没有统计学意义，但从结果中可以看到一个简单的模式，如图4所示。对于较小的种类数，RR较强；而对于较大的种类数，MDI较强。这种模式促进了共识聚类方法的融合发展。

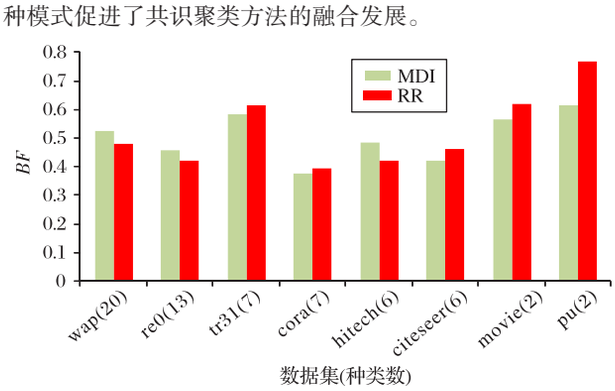

图4 不同数据集的RR和MDI的比较

Fig. 4 Comparison of RR and MDI of different datasets

表2是本文提出的DCCC方法与CDIM-RR、CDIM-MDI和HierLink方法进行的比较。

表2 DCCC方法与CDIM-RR、CDIM-MDI和HierLink方法的比较

Tab. 2 Comparison of DCCC with CDIM-RR, CDIM-MDI and HierLink methods

数据集	CDIM-RR	CDIM-MDI	HierLink	DCCC	Imp. DCCC/%
wap	0.481	0.522	0.456	**0.537**	7.03
cora	0.374	0.348	0.355	**0.382**	6.10
pu	0.773	0.618	0.618	**0.849**	21.54
citeseer	0.449	0.414	0.459	**0.453**	5.23
hitech	0.440	0.491	0.454	**0.497**	7.04
tr31	0.638	0.608	0.618	**0.650**	4.52
re0	0.421	0.447	0.402	**0.455**	5.03
movie	0.615	0.587	0.587	**0.632**	5.41

表2中最后一列显示的是DCCC方法比单个聚类方法的平均值提高的百分比。在HierLink方法中，首先，通过基于相同的簇成员计算所有对象的成对相似度矩阵来实现基础的共识聚类。然后，利用“区”链接进行分层聚类，最终达到聚类的目的。利用CDIM-RR和CDIM-MDI的两种初始聚类方法计算

相似矩阵。与 CDIM-RR、CDIM-MDI 和 HierLink 相比,DCCC 的结果都有显著改善。

表3是DCCC方法与其他共识聚类方法[8]的比较,表中显示的是精度值,本文提出的DCCC方法明显优于其他方法。

表3　DCCC与其他共识聚类方法的精度对比

Tab. 3　Comparison accuracy of DCCC and other consensus clustering methods

数据集	CSPA	WHC	HICC	HEC	MCE	DCCC
cora	0.211	0.306	0.351	0.389	0.193	**0.501**
citeseer	0.243	0.313	0.268	0.182	0.546	**0.641**

4　结语

本文提出了一种文档聚类的共识构建方法。在不同的数据聚类工具生成的簇解决方法中,DCCC使用分类工具进行共识度量。本文选择CDIM寻找初始簇:CDIM使用识别信息最大化进行文档聚类。起始阶段,使用不同聚类的解决方法寻找共识文档,即哪些文档应该属于最相似的簇。同时,识别文本分类器DTWC接受共识文档的培训。而后,DTWC文本分类器预报与初始聚集簇标签信息不一致的文档。为了获得不同的初始视图,本文采用了RR和MDI两种识别方法。RR和MDI具有不同的优势,在DCCC中进行融合可达到提升性能的效果。利用8种标准网络数据集验证了本文方法的有效性。

本文方法是一种通用的共识方法,可应用于文档聚类之外的其他领域,并可测试不同的初始聚类方法和不同的分类方法。目前,只有两种方法用于初始聚类。在未来,本文的目标是测试其他识别方法的融合、聚类和分类。

参考文献 (References)

[1] YU Y, WANG L, ZHU Q. Intelligent fuzzy information retrieval based on ontology knowledge-base [J]. International Journal of Internet Protocol Technology, 2018, 11(3): 180-191.

[2] 刘钰峰,李仁发. 基于查询—文档异构信息网络的半监督学习[J]. 通信学报, 2014, 35(8): 40-47. (LIU Y F, LI R F. Semi-supervised learning by constructing query-document heterogeneous information network [J]. Journal on Communications, 2014, 35(8): 40-47.)

[3] 刘兆军. XML文档数据集聚类问题研究[D]. 长春:吉林大学, 2015: 31-33. (LIU Z J. Study on clustering for XML document collection[D]. Changchun: Jilin University, 2015: 31-33.)

[4] BOONGOEN T, IAM-ON N. Cluster ensembles: a survey of approaches with recent extensions and applications [J]. Computer Science Review, 2018, 28: 1-25.

[5] FRED A L N, JAIN A K. Combining multiple clustering using evidence accumulation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850.

[6] IAM-ON N, BOONGOEN T, GARRETT S, et al. A link-based cluster ensemble approach for categorical data clustering[J]. IEEE Transactions on Knowledge Data Engineering, 2012, 24(3): 413-425.

[7] MINAEI-BIDGOLI B, PARVIN H, ALINEJAD-ROKNY H, et al. Effects of resampling method and adaptation on clustering ensemble efficacy[J]. Artificial Intelligence Review, 2014, 41(1): 27-48.

[8] HUSSAIN S F, MUSHTAQ M, HALIM Z. Multi-view document clustering via ensemble method[J]. Journal of Intelligent Information Systems, 2014, 43(1): 81-99.

[9] 李玉,甄畅,石雪,等. 基于熵加权*K*-means全局信息聚类的高光谱图像分类[J]. 中国图象图形学报, 2019, 24(4): 630-638. (LI Y, ZHEN C, SHI X, et al. Hyper spectral image classification algorithm based on entropy weighted *K*-means with global information [J]. Journal of Image and Graphics, 2019, 24(4): 630-638.)

[10] 赵军,徐晓燕. 基于GraphX的分布式幂迭代聚类[J]. 计算机应用, 2016, 36(10): 2710-2714. (ZHAO J, XU X Y. Distributed power iteration clustering based on GraphX[J]. Journal of Computer Applications, 2016, 36(10): 2710-2714.)

[11] YU Y, LIU Z. Document topic mining algorithm without parameters clustering based on dynamic threshold[J]. Journal of Computational Information Systems, 2013, 9(5): 1965-1972.

[12] CHANG C H, DAI B R. A fragment-based iterative consensus clustering algorithm with a robust similarity[J]. Knowledge Information System, 2014, 41(3): 591-609.

[13] 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 S, GAO J, HUA X P, et al. An improved adaptive cluster ensemble selection approach [J]. Acta Automatica Sinica, 2018, 44(11): 2103-2112.)

[14] 陈黎飞,姜青山,王声瑞. 基于层次划分的最佳聚类数确定方法[J]. 软件学报, 2008, 19(1): 62-72. (CHEN L F, JIANG Q S, WANG S R. A hierarchical method for determining the number of clusters[J]. Journal of Software, 2008, 19(1): 62-72.)

[15] GAN H, SANG N, HUANG R, et al. Using clustering analysis to improve semi-supervised classification [J]. Neurocomputing, 2013, 101: 290-298.

[16] HASSAN M T, KARIM A, KIM J B, et al. CDIM: document clustering by discrimination information maximization[J]. Information Science, 2015, 316: 87-106.

[17] 赵孝礼,赵荣珍. 全局与局部判别信息融合的转子故障数据集降维方法研究[J]. 自动化学报, 2017, 43(4): 560-567. (ZHAO X L, ZHAO R Z. A method of dimension reduction of rotor faults data set based on fusion of global and local discriminant information [J]. Acta Automatica Sinica, 2017, 43(4): 560-567.)

[18] 胡凌超,于洪. 一种基于投票的三支决策聚类集成方法[J]. 小型微型计算机系统, 2016, 37(8): 1741-1745. (HU L C, YU H. Voting cluster ensemble approach based on three-way decisions [J]. Journal of Chinese Computer Systems, 2016, 37(8): 1741-1745.)

[19] 魏霖静,练智超,王联国,等. 基于词条与语意差异度量的文档聚类算法[J]. 计算机科学, 2016, 43(12): 229-233, 259. (WEI L J, LIAN Z C, WANG L G, et al. Term and semantic difference metric based document clustering algorithm[J]. Computer Science, 2016, 43(12): 229-233, 259.)

This work is partially supported by the National Natural Science Foundation of China (61602202).

WANG Liuyang, born in 1974, M. S., associate professor. His research interests include information management and information system, intelligent information processing, big data mining.

YU Yangxin, born in 1970, M. S., professor. His research interests include information management and information system, intelligent information processing, knowledge organization.

CHEN Bolun, born in 1986, Ph. D., associate professor. His research interests include link prediction for complex networks.

ZHANG Hui, born in 1970, M. S., professor. Her research interests include information management and information system, intelligent information processing.