基于Laplacian图谱的短文本聚类算法

孟海宁[1,2],冯 锴[1],朱 磊[1],张贝贝[1],童新宇[1],黑新宏[1]
(1. 西安理工大学计算机科学与工程学院,陕西西安 710048;2. 陕西省网络计算与安全技术重点实验室,陕西西安710048)

摘 要: 提出基于词频处理的Laplacian图谱聚类算法,以解决短文本数据维数高、特征稀疏等问题.首先采用词频-逆文本频率指数TF-IDF(Term Frequency-Inverse Document Frequency)方法,将短文本数据集映射到文本向量空间得到词频权值矩阵;其次利用Laplacian矩阵的图谱聚类特性,对词频权值矩阵进行数据降维处理;然后依据Laplacian矩阵的特征值表示文本相似度的特点,选择前K个特征值对应的特征向量作为初始聚类中心,以减少聚类过程的迭代次数.在SSC、20 News Group及Microblog PCU数据集上进行相关实验,结果表明Laplacian图谱聚类算法比传统聚类算法,不仅具有更优的聚类结果与更快的收敛速度,而且受噪声点影响较小,有很好的鲁棒性.

关键词: Laplacian图谱;词频-逆文本频率指数;短文本聚类;向量空间模型;数据降维;特征权值
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2021)09-1716-08
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20201266

Short-Text Clustering Algorithm Based on Laplacian Graph

MENG Hai-ning[1,2],FENG Kai[1],ZHU Lei[1],ZHANG Bei-bei[1],TONG Xin-yu[1],HEI Xin-hong[1]
(1. *School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, Shaanxi* 710048, *China*;
2. *Shaanxi Key Lab Network Computer and Security Technology, Xi'an, Shaanxi* 710048, *China*)

Abstract: A Laplacian graph clustering algorithm based on word frequency processing is presented, to solve the problems of high feature dimension and sparse feature in short text. First, the term frequency-inverse document frequency (TF-IDF) method is used to map the short text dataset to the text vector space, to obtain the word frequency weight matrix. Secondly, the dimension of the word frequency weight matrix is reduced by using the graph clustering property of Laplacian matrix. Afterwards, according to the feature that the eigenvalues of Laplace matrix can represent the degree of text similarity, the eigenvectors corresponding to the first K eigenvalues are selected as the initial clustering center, thus reducing the number of iterations in the clustering process. We conduct extensive experiments on SSC, 20 News Group and Microblog PCU datasets. The results show that the Laplacian graph clustering algorithm not only has better clustering results and faster convergence speed compared with the traditional clustering algorithm, but also it is less affected by noises and has good robustness.

Key words: laplacian graph;term frequency-inverse document frequency;short-text clustering;vector space model;data dimensionality reduction;feature weight

1 引言

随着互联网技术的不断发展,互联网上的文本信息呈现急剧扩张的态势,如何精确快速地对文本信息进行分类和聚类是非常值得研究的问题.例如,识别垃圾短信的文本对信息进行分类,通常需要扫描数以千计的Web页面,其过程困难且耗时.该问题可通过文档聚类得以解决[1].聚类是将类似事物分成一类并将不同事物分成不同类别的过程,它是重要的数据分析手段.数据聚类方法根据数据的固有性质划分为不同类,使得同一类数据尽可能具有较高的相似性[2].

短文本是较常见的一种内容形式,手机短信、用户评论及微博话题等都属于短文本[3].对短文本进行聚类分析具有重要的应用价值,如对用户评论进行观点挖掘、对社交媒体进行话题检测以及舆情预警[4]等.由

收稿日期:2020-11-09;修回日期:2021-06-11;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61602375, No.61773313)

于短文本数据具有特征维数高、特征难提取、噪音数据多等特点,传统的聚类算法对短文本聚类精确度不高且速度较慢. 这主要有三方面的原因:一是对于非结构化或半结构化的短文本数据,文本向量维度高,聚类结果不够准确;二是初始聚类中心随机选择,可能导致算法的时间开销较大;三是短文本数据特征稀疏,聚类过程易受到噪声数据影响,算法鲁棒性差.

针对上述问题,很多学者提出了新的短文本聚类算法. 为优化初始聚类中心,Zhang等[5]利用高密度点的相对距离选取聚类中心点,避免将同类中两个高密度点同时选做初始聚类中心,从而提高聚类结果的正确率. 张雪松等[6]提出一种基于频繁词集的聚类算法,该算法不仅能降低文本维度,还可构建了文本间的关联关系,聚类效果较好. Ma等[7]通过计算支持度和置信度扩展集合,将信息增益引入TF-IDF,提高了短文本聚类算法的性能. 为解决传统文本聚类结果不精确的问题,Yang等[8]通过建立多任务光谱聚类模型,强化聚类标签和映射函数的学习过程,从而提高了聚类精确度. 唐俊等[9]提出一种基于拟Laplacian图谱的形状表示与聚类方法,针对形状数据集,将高维形状数据投影至低维空间进行聚类. Zeng等[10]将Laplacian图谱正则化引入子空间聚类中,用于无监督高光谱图像分类. 但上述算法都没有解决短文本数据聚类收敛速度慢的问题,也没有考虑短文本数据集中噪声数据对算法的鲁棒性影响.

Laplacian矩阵是图谱方法的一种表现形式,通过几何分析、代数计算,建立低维数据与高维数据的联系. 现有的Laplacian降维方法,通常针对图像和语音数据集[11],而非短文本数据集. 本文依据Laplacian矩阵是半正定对称阵的特性,对Laplacian矩阵求解特征值,将矩阵特征值按大小排列,即对短文本的多个特征按照与聚类主题关联程度进行排序,强关联特征在前,次要特征在后. 选择前K个特征值,去除统计特性较弱的次要特征,从而对数据集进行降维处理,避免发生维度爆炸问题.

本文提出一种基于Laplacian图谱的短文本聚类算法,主要工作包括以下三点:(1)通过对原始数据集的词频向量矩阵进行Laplacian矩阵化处理,将短文本数据在低维度下表示文本间的相似关系,提高聚类准确度;(2)依据Laplacian矩阵的图谱聚类特性,求解Laplacian矩阵的特征值,表示特征词在文本中的重要性. 选择前K个特征向量作为初始聚类中心,代替传统聚类算法中随机选取初始聚类中心的策略,减少了聚类过程迭代次数,提高聚类收敛速度;(3)在短文本数据集上进行大量实验,通过人工加入不同比例的噪声数据,验证了本文聚类算法具有较强的鲁棒性.

2　Laplacian矩阵

Laplacian矩阵[12]也称基尔霍夫矩阵,是一种表示带权无向图的矩阵. 若给定一个有n个顶点的图G,可定义Laplacian矩阵$\boldsymbol{L}$为:

$$\boldsymbol{L} = \boldsymbol{M} - \boldsymbol{A} \tag{1}$$

其中,$\boldsymbol{M}$为图的度矩阵(metric matrix),$\boldsymbol{A}$为图的邻接矩阵(adjacency matrix).

邻接矩阵为表示顶点之间相邻关系的矩阵,将邻接矩阵$\boldsymbol{A}$定义为:

$$\boldsymbol{A}_{pq} = \begin{cases} 0, & p\text{和}q\text{之间没有边连接} \\ 1, & p\text{和}q\text{之间有边连接} \end{cases} \tag{2}$$

度矩阵由每个顶点的度计算得出,将度矩阵$\boldsymbol{M}$定义为:

$$\boldsymbol{M}_{pq} = \begin{cases} 0, & p \neq q \\ \deg\left(v_p\right), & p = q \end{cases} \tag{3}$$

其中,$\deg(v_p)$为顶点p所有邻接边的特征权值之和.

Laplacian矩阵具有以下的性质:

(1)Laplacian矩阵是半正定对称矩阵;

(2)Laplacian矩阵所有特征值为非负实数;

(3)Laplacian矩阵最小特征值为0,任意行列和均为0.

3　Laplacian图谱聚类算法

3.1　算法描述

本文提出Laplacian图谱聚类算法,其流程如图1所示. 首先,采用自然语言处理工具包NLTK(Natural Language ToolKit),对数据集进行标记化分词、去除停用词、词性标注及词干提取. 其次,进行数据特征提取,第一步将预处理后的数据集映射到向量空间,转换为TF-IDF词频矩阵$\boldsymbol{Q}$;第二步计算词频矩阵$\boldsymbol{Q}$的Laplacian矩阵$\boldsymbol{L}$,矩阵中每个行向量表示一个样本数据的多维特征词. 然后求解$\boldsymbol{L}$的特征值,将特征值按照大小排序,即对特征词的特征权值排序. 之后选取前K个特征值对应的特征列向量,构成降维后的词频矩阵$\boldsymbol{R}$. 最后对Laplacian矩阵$\boldsymbol{R}$的特征值按照大小排序,即在样本数据中按照词频重要性(词语在样本中出现的频率高低)排序. 选取前K个特征值对应的特征行向量作为初始聚类中心点;然后对于剩余的数据对象,根据其与聚类中心的距离,分配到距离最近的类;之后重新计算每个类的平均值,更新聚类中心,不断迭代该过程,直到算法收敛则聚类结束.

3.2　文本向量化表示

根据Salton提出的向量空间模型(vector space model)[13],将预处理后的数据集映射到二维向量空间.

假设短文本集合为$D=\{d_j|j=1,2,3,\cdots,n\}$,$n$为文本

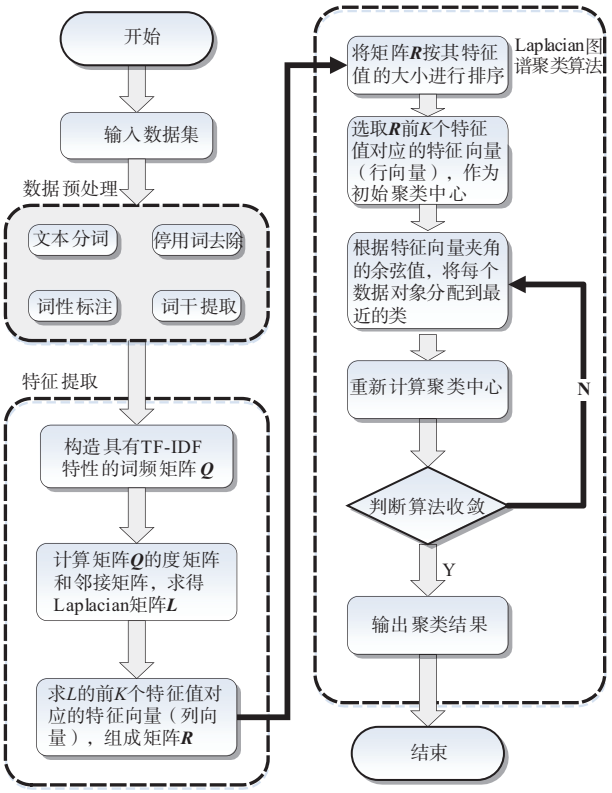


图1 Laplacian图谱聚类算法流程

总数. 数据集D中文本特征词构成的集合为$C=\{t_i|i=1,2,3,\cdots,m\}$，其中$m$为文本特征词总数. 向量空间模型中每一维由特征词和其对应的特征权值组成，对应地将数据集D中第j个文本表示为$d_j=\{(t_1,w_{j1}),(t_2,w_{j2}),\cdots,(t_i,w_{ji}),\cdots,(t_m,w_{jm})\}$，其中$w_{ji}$表示该特征词$t_i$在文本$d_j$中对应的特征权值. 本文采用TF-IDF方法表示特征权值，TF-IDF由词频TF和逆文本频率指数IDF组成. TF按式(1)计算：

$$\mathrm{TF}(t_i,d_j)=\frac{N(t_i,d_j)}{N(d_j)} \tag{4}$$

其中，$N(t_i,d_j)$表示文本d_j中特征词t_i出现的次数，$N(d_j)$表示文本d_j中的词条总数.

IDF按式(5)计算：

$$\mathrm{IDF}(t_i)=\log\left(\frac{n}{M(t_i)+1}\right) \tag{5}$$

其中，n是短文本数，$M(t_i)$是包含特征词t_i的短文本数，分母加1是为了避免特征词t_i未在任何文本中出现进而导致分母为零的情况.

根据TF及IDF值，特征词t_i的TF-IDF特征权值按式(6)计算：

$$\text{TF-IDF}(t_i,d_j)=\mathrm{TF}(t_i,d_j)\times\mathrm{IDF}(t_i) \tag{6}$$

然后，将TF-IDF特征权值代入数据集D，将其转化为词频矩阵$\boldsymbol{Q}_{n\times m}$. 计算词频矩阵$\boldsymbol{Q}$的邻接矩阵$\boldsymbol{A}$和度矩阵$\boldsymbol{M}$，得到Laplacian矩阵$\boldsymbol{L}$. 最后在聚类过程中计算Laplacian矩阵$\boldsymbol{L}$的特征值，将其从小到大排列，其对应的特征向量也按特征值递增排列，取前K个特征向量进行聚类，得到聚类结果.

3.3 Laplacian矩阵的图谱聚类特性

从图论角度看，聚类问题就是图的分割问题. 假定图$G=(V,E)$，顶点集$V=\{v_1,v_2,\cdots,v_n\}$表示各个样本，带特征权值的边可表示样本之间的相似度. 聚类等价于寻找一种优化算法，将图G分割成若干子图(对应聚类中的簇或类)，使子图间的特征权值或相似度较小，而子图内的特征权值或相似度较大.

对于任意实数列向量$\boldsymbol{f}\in\mathbf{R}^n$，$\boldsymbol{f}'$是$\boldsymbol{f}$的转置，求解Laplacian矩阵$\boldsymbol{L}=\boldsymbol{M}-\boldsymbol{A}$，则有式(7)成立.

$$\boldsymbol{f}'\boldsymbol{L}\boldsymbol{f}=\boldsymbol{f}'\boldsymbol{M}\boldsymbol{f}-\boldsymbol{f}'\boldsymbol{A}\boldsymbol{f} \tag{7}$$

采用RatioCut算法[14]，通过求割边的特征权值之和的最小值，求解图的分割问题. 若假定子图集合$B=\{G_1,G_2,\cdots,G_n\}$，则RatioCut的目标函数为：

$$\min_{G_i\subset V}\mathrm{RatioCut}(G_i,\bar{G}_i) \tag{8}$$

其中G_i为图G的第i个子图，$\bar{G}_i$为G_i的补图.

同时，定义向量$\boldsymbol{f}$的第p个样本f_p，如式(9)所示.

$$f_p=\begin{cases}\sqrt{|\bar{G}_i|/|G_i|}, & v_p\in G_i\\ -\sqrt{|G_i|/|\bar{G}_i|}, & v_p\in\bar{G}_i\end{cases},\quad p\in\{1,\cdots,n\} \tag{9}$$

然后将式(2)、(3)与(9)代入式(7)，得：

$$\begin{aligned}\boldsymbol{f}'\boldsymbol{L}\boldsymbol{f}&=\sum_{p=1}^{n}\deg(v_p)f_p^2-\sum_{p,q=1}^{n}f_pf_qU_{pq}\\&=\frac{1}{2}\sum_{p,q=1}^{n}Upq(f_p-f_q)^2\end{aligned} \tag{10}$$

其中U_{pq}为顶点p到顶点q的权值，$q\in\{1,\cdots,n\}$.

然后，将式(9)代入式(10)，可得：

$$\begin{aligned}\boldsymbol{f}'\boldsymbol{L}\boldsymbol{f}&=\frac{1}{2}\sum_{p,q=1}^{N}U_{pq}(fp-f_q)^2\\&=\frac{1}{2}\sum_{p\in G_i,q\in\bar{G}_i}U_{pq}\left(\sqrt{\frac{|\bar{G}_i|}{|G_i|}}+\sqrt{\frac{|G_i|}{|\bar{G}_i|}}\right)^2\\&\quad+\frac{1}{2}\sum_{p\in\bar{G}_i,q\in G_i}U_{pq}\left(-\sqrt{\frac{|\bar{G}_i|}{|G_i|}}-\sqrt{\frac{|G_i|}{|\bar{G}_i|}}\right)^2\\&=\mathrm{cut}(G_i,\bar{G}_i)\left(\sqrt{\frac{|\bar{G}_i|}{|G_i|}}+\sqrt{\frac{|G_i|}{|\bar{G}_i|}}+2\right)\\&=|V|\times\mathrm{RatioCut}(G_i,\bar{G}_i)\end{aligned} \tag{11}$$

其中$|V|$是常量. 可以看出，最小化$\boldsymbol{f}'\boldsymbol{L}\boldsymbol{f}$等价于最小化

RatioCut目标函数,即对短文本Laplacian矩阵L进行聚类,等价于图G的子图分割.

求解$f'Lf$的最小化问题,令λ是L的特征值,f是特征值λ对应的特征向量,则有$Lf=\lambda f$. 然后对$Lf=\lambda f$等式两边,同时左乘f',可得:

$$f'Lf=\lambda f'f \tag{12}$$

其中,f是列向量,则$f'f$值是一个实数,因此最小化$f'Lf$等价于最小化λ,只需求解L的最小特征值及其对应的特征向量.

根据Laplacian矩阵的图谱聚类特性,可以将高维的短文本数据,表示成低维数据. 通过求解Laplacian矩阵的特征值及其对应的特征向量,可得出Laplacian图谱聚类的目标. 即互有关系的顶点,在降维后的空间中尽可能靠拢;相互无关的顶点,在降维后的空间中尽可能远离.

3.4　基于词频处理的Laplacian图谱聚类算法

本文提出Laplacian图谱聚类算法. 首先利用Laplacian算子对短文本数据集进行降维处理. 具体地,计算词频矩阵Q的度矩阵M和邻接矩阵A,得到Laplacian矩阵L. 然后将矩阵L中前K个特征值对应的特征向量,组成词频权值矩阵R,作为降维后的短文本数据集再进行聚类. 算法的具体过程描述,如算法1所示。

算法1应用Laplacian矩阵对短文本数据集的词频矩阵进行特征提取,实现了短文本数据降维. 利用Laplacian矩阵求特征值的方法,决定特征词在文本的重要程度,以此为依据确定了初始聚类中心点. 最后进行K-means聚类. 本文方法不仅限于K-means算法,也适用于其他传统的聚类方法,如K-近邻、C-means等.

算法1　基于词频处理的Laplacian图谱聚类算法

输入:测试数据集词频矩阵Q

输出:聚类算法正确率

步骤:

1. 根据词频矩阵Q计算测试数据集的邻接矩阵A;
2. 根据式(3),计算度矩阵M;
3. 根据式(1),计算Laplacian矩阵L;
4. 采用肘方法[15]计算最优的文本聚类数K;
5. 提取Laplacian矩阵L的前K个特征值,按照从小到大排序,得到对应的K个特征列向量,即将短文本数据集降至K维,组成矩阵R;
6. 使用文本特征向量夹角的余弦值来计算文本相似度;
7. 将矩阵R作为数据集输入K-means聚类算法;
8. 选取R的前K个特征向量,作为初始聚类中心;
9. 计算剩余样本点到聚类中心的距离,并将其分配到距离最近的类内;
10. 若样本所属的类改变,则更新该类的聚类中心点;
11. 返回K个类;
12. 根据误差平方和准则函数,检验算法是否收敛.

3.5　基于Laplacian矩阵的初始聚类中心选择

传统聚类算法一般随机选取初始聚类中心,并且在算法迭代中更新聚类中心,聚类结果和算法性能很大程度上取决K个初始聚类中心,且收敛速度较慢. 为提高聚类算法性能,本文提出的Laplacian图谱聚类算法,初始聚类中心依据Laplacian矩阵的图谱特性决定. 具体地,依据Laplacian矩阵特征分解后特征值非负,对特征值按照大小排序,得到特征词在文本中的重要性排序. 前K个特征值对应的特征向量,即表示最有可能成为聚类中心的K个词语,选取其作为初始聚类中心,然后进行聚类,可以减少算法迭代次数.

4　实验及结果分析

4.1　数据集

表1　测试数据集参数表

编号	数据集	样本数	类别数	维度
1	SSC	5574	2	19
2	20 News Group	3960	5	159
3	Microblog PCU	3900	3	20

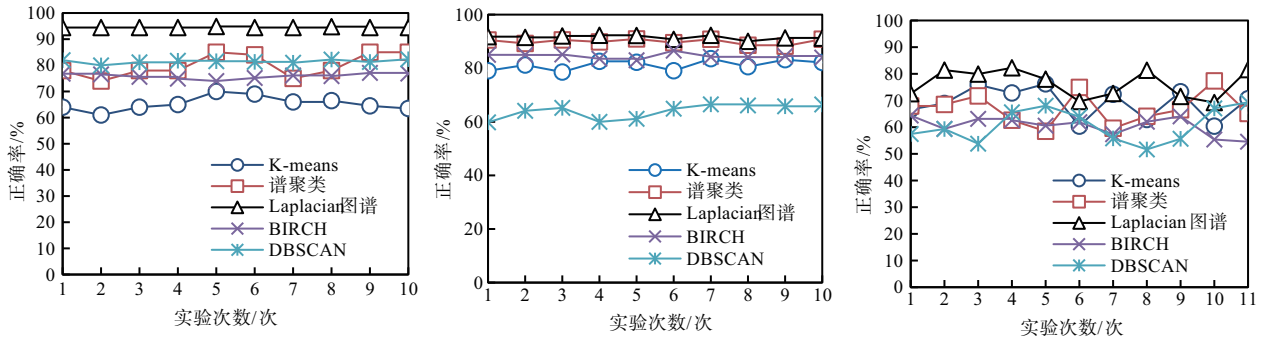
本文英文测试数据集采用UCI的SSC(SMS Spam Collection)数据集和20 News Group数据集,中文数据集采用Microblog PCU数据集. SSC是带有SMS数据标签的短文本数据集,共计5574份数据,包括4827条SMS正常邮件和747条垃圾邮件. 20 News Group数据集收集大约20000个新闻组文档,分20个不同主题的新闻组集合. Microblog PCU是新浪微博中文短文本数据集,用于研究机器学习方法和社会关系,其中包括3900条数据. 表1给出每个数据集的样本数、类别数及维度.

4.2　聚类结果

本文采用正确率(accuracy)作为聚类结果的度量标准. 聚类正确率越高,聚类算法的性能越好. 实验中通过统计被正确聚类的文本数占所有短文本总数的比例来计算正确率,并采用十折交叉验证法(10-fold cross-validation)对聚类结果的正确率进行验证. 具体地,将数据集分成10份,其中9份作为训练集,1份作为测试集. 每次实验得出聚类结果正确率,用10次正确率平均值评估算法性能.

将本文Laplacian图谱算法与K-means[16]、谱聚类[17]、BIRCH[18]及DBSCAN[19]算法,在SSC、20 News Group和Microblog PCU数据集上进行聚类. 实验中最大迭代次数设置为20,K值采取肘方法确定,K为各数据集的类别数. 其中BIRCH和DBSCAN算法无需预先选择K值. 图2给出了五种算法的聚类正确率对比.

从图2可以看出,本文Laplacian图谱聚类算法与传统K-means、谱聚类、BIRCH以及DBSCAN算法相比,聚类结果正确率最高. 本文Laplacian图谱聚类算法在



(a)SCC 数据集 (b)20 News Group 数据集 (c)Microb log PCU 数据集

图 2 聚类正确率对比

SSC 数据集上较 K-means，正确率提高 29. 17%，较谱聚类提高 14. 52%，较 BIRCH 提高 18. 99%，较 DBSCAN 提高 12. 24%；在 20 News Group 数据集上，本文算法较 K-means 正确率提高了 11. 25%，较谱聚类提高 1. 55%，较 BIRCH 提高 6. 56%，较 DBSCAN 提高 31. 75%；在 Microblog PCU 数据集上，本文算法较 K-means 正确率提高了 7. 24%，较谱聚类提高 9. 39%，较 BIRCH 提高 15. 9%，较 DBSCAN 提高 15. 69%.

4. 3 算法鲁棒性分析

聚类算法的鲁棒性指算法受到数据集中的干扰数据、噪声数据及离群点数据的影响程度. 本文通过对 SSC 数据集人工加入噪音比例依次为 5%、10%、15%、20%、25%、30% 的噪音数据，得到 6 个含不同比例噪音的数据集，在每个数据集上评估本文提出的 Laplacian 图谱聚类算法的正确率，从而检测本文算法的鲁棒性，实验结果如图 3 所示. 从正确率的对比来看，本文提出的 Laplacian 图谱聚类算法受噪音数据影响较小. K-means 算法和谱聚类算法受到噪音影响相对较大，当加入的噪音比例超过 20% 时，两个算法的正确率均出现明显下降；BIRCH 和 DBSCAN 两种算法在噪音比例超过 20% 时，也出现了正确率下降的现象；而本文算法在 0~30% 噪音比例的数据集上聚类结果正确率基本保持在 90% 以上，说明本文提出的算法具有较强的鲁棒性.

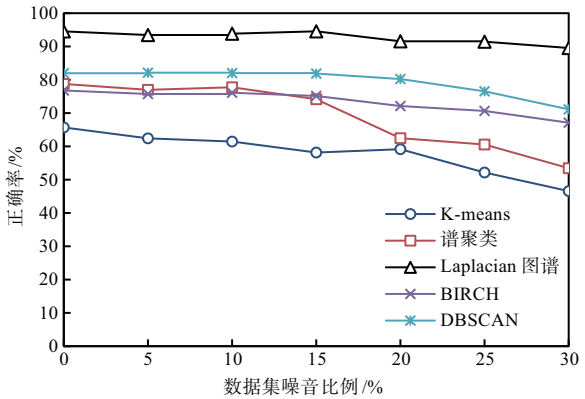
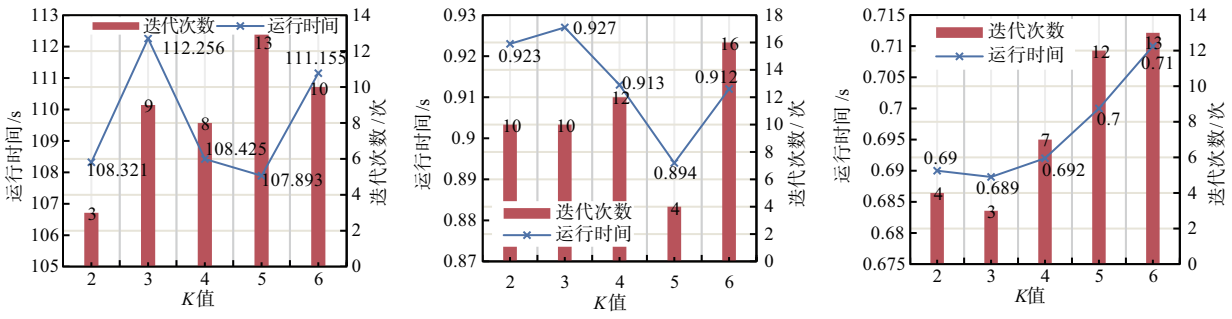


图 3 鲁棒性测试实验结果

4. 4 *K* 值对聚类算法的影响

为探究 *K* 的取值对本文所提出 Laplacian 图谱聚类算法的影响，我们采用肘方法，依据迭代次数和运行时间两个标准指标，来选取 *K* 值. 针对 SSC、20 News Group 以及 Microblog PCU 三个数据集，图 4 给出了 *K* 值对聚类效果影响的实验结果.

从图 4 可知，在 SSC 数据集上，*K* 值选择 2 时，算法的迭代次数和运行时间均为最少；在 20 News Group 数据集上，*K* 值为 5 时，算法的迭代次数和运行时间均为最少；在 Microblog PCU 数据集上，*K* 值为 3 时，算法在迭代次数和运行时间上均为最少. 此外，由表 1 可知，SSC



(a)SCC 数据集 (b)20 News Group 数据集 (c)Microb log PCU 数据集

图 4 *K* 值对 Laplacian 图谱聚类算法的影响

类别数为2,20 News Group类别数为5,Microblog PCU的类别数为3,这与上述肘方法确定的K值一致.

4.5 算法性能评估

聚类性能评估包括聚类质量和聚类收敛速度两个方面,其中,聚类质量采用查准率(precision)、查全率(recall)和F-score值三个评价准则.

查准率表示被正确分类的样本数与参与分类的样本总数之比,如式(13)所示:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

其中,TP(True Positives)为正确地划分为正例的个数;FP(False Positives)为错误地划分为正例的个数.

查全率表示被正确分类的样本数与应当被正确分类的样本数之比,如式(14)所示:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

其中,FN(False Negatives)为错误地划分为负例的个数.

F-score是查准率和查全率的调和平均,如式(15)所示.

$$F_b = \frac{\left[(1 + b^2) \times \text{precision} \times \text{recall}\right]}{\left(b^2 \times \text{precision} + \text{recall}\right)} \tag{15}$$

其中,b为调节查准率和查全率的权值系数,本文选取b=1,采用F_1作为性能评价标准.

首先,在聚类质量对比方面,表2给出了本文算法与传统的K-means、谱聚类、BIRCH及DBSCAN算法,在多次实验中查准率、查全率以及F-score值的对比结果.从平均值可以看出,本文算法在三个评价指标上均最优,K-means算法效果最差,BIRCH、DBSCAN和谱聚类算法的聚类效果居中.这是因为本文提出的Laplacian图谱聚类算法,利用Laplacian矩阵的图谱聚类特性,对数据集的词频权值矩阵进行了降维处理,又可表示文本之间的相似关系,对高维短文本数据的聚类效果较好.

表2　算法聚类质量对比

算法	平均值								
	SSC数据集			20 News Group数据集			Microblog PCU数据集		
	查准率	查全率	F-score	查准率	查全率	F-score	查准率	查全率	F-score
K-means	64.80%	87.05%	74.27%	79.41%	85.41%	82.29%	70.00%	88.87%	78.32%
谱聚类	78.60%	95.93%	86.38%	88.25%	91.03%	89.61%	85.00%	97.34%	90.75%
Laplacian图谱	**94.52%**	**98.27%**	**96.36%**	**90.59%**	**93.34%**	**91.94%**	**94.84%**	**98.98%**	**96.87%**
BIRCH	76.78%	96.05%	85.34%	85.05%	97.15%	90.70%	84.01%	97.13%	90.09%
DBSCAN	81.96%	95.38%	88.16%	59.75%	88.94%	71.48%	78.00%	95.83%	86.00%

其次,在聚类收敛速度方面,图5给出了五种算法在SSC数据集上的聚类过程中迭代次数对比情况.实验可得本文算法在SSC数据集上相较K-means收敛速度提高60.95%,相较谱聚类提高38.81%,相较BIRCH提高51.19%,相较DBSCAN提高62.04%.这表明本文算法在收敛速度上具有明显的优势.此外,我们又对比了本文算法中未采用Laplacian矩阵确定初始聚类中心的情况,得出本文算法在SSC数据集上收敛速度提高了51.76%.这是因为本文提出的Laplacian图谱聚类算法,依照词语在文本中的重要性排序,针对SSC选取了前2个特征值对应的特征向量作为初始聚类中心,聚类后对应SSC中的正常邮件和垃圾邮件两类,可以减少聚类的迭代次数,提高算法的收敛速度.该结果也适用于其他2个数据集.

综上,本文算法根据Laplacian矩阵在图谱聚类上的特性,对向量空间模型的文本进行Laplacian矩阵转化,解决传统算法对短文本聚类中数据维数过高的问题.Laplacian矩阵表示特征词权值矩阵,将特征词按其

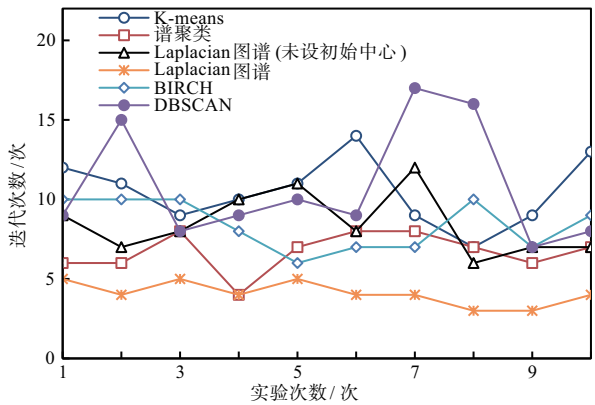


图5　SSC数据集上算法迭代次数对比

在短文本中重要度进行排序,可计算短文本间的文本相似度.因此,大大提高了聚类质量.本文算法以图论为基础进行聚类分析,所以对聚类的噪声信息不敏感.另外,聚类中初始聚类中心点不是随机分配,而是利用Laplacian矩阵的性质加以确定,因此聚类过程中减少了算法迭代次数,提高了聚类收敛速度.

5 结论与展望

研究如何提高短文本聚类质量和效率具有重要的应用价值. 对于短文本数据面临的问题,如特征稀疏、维数高、噪音干扰等,传统的聚类算法不适应于短文本处理. 本文提出一种基于Laplacian图谱的短文本聚类算法. 采用Laplacian矩阵对聚类数据进行降维处理,通过对原始数据集进行Laplacian矩阵化,将原始数据集在低维下表示,能有效处理冗余数据和噪声数据,提高了聚类质量. 另外,相较于经典聚类算法随机选取初始聚类中心的方法,本文算法通过对原始数据集词频矩阵的Laplacian映射,对词语在文本中的重要性排序,能够表示文本间的相似关系,确定初始聚类中心,从而更能提高聚类收敛速度. 再者,通过实验加入噪声数据进行聚类,验证了本文算法聚类过程能够较好抵抗噪声干扰,具有较强的鲁棒性.

今后工作将更多地采用智能计算方法对聚类算法进行优化,进一步提高算法的精度及算法的执行效率.

参考文献

[1] Habib S T, Zahid A. An analysis of map reduce efficiency in document clustering using parallel K-means algorithm [J]. Future Computing & Informatics Journal, 2018, 3(2): 200 – 209.

[2] Deng H, Qin H, Sun X, et al. A K-means clustering algorithm of meliorated initial center[J]. Computer Technology and Development, 2013, 11:42 – 45.

[3] 贺超波, 汤庸, 张琼, 等. 基于增量式鲁棒非负矩阵分解的短文本在线聚类[J]. 电子学报, 2019, 47(5): 1086 – 1093.
He Chao-bo, Tang Yong, Zhang Qiong, et al. Short text online clustering based on incremental robust nonnegative matrix factorization [J]. Acta Electronica Sinica, 2019, 47 (5): 1086 – 1093.(in Chinese)

[4] Yang K, Miao R. Research on improvement of text processing and clustering algorithms in public opinion early warning system[A]. Proceedings of the 5th International Conference on Systems and Informatics[C]. NY, USA: IEEE, 2018.333 – 337.

[5] Zhang X, Qiang S, Gao H, et al. A density-based method for selection of the initial clustering centers of K-means algorithm[A]. Proceedings of the 2nd Advanced Information Technology, Electronic and Automation Control Conference[C]. NY,USA: IEEE, 2017.2565 – 2568.

[6] 张雪松, 贾彩燕. 一种基于频繁词集表示的新文本聚类方法[J]. 计算机研究与发展, 2018, 55(1):102 – 112.
Zhang Xue-song, Jia Cai-yan. A new documents clustering method based on frequent itemsets[J]. Journal of Computer Research and Development, 2018, 55(1): 102 – 112. (in Chinese)

[7] Ma H, Lei D, Zeng X, et al. Short text feature extension based on improved frequent term sets[A]. Proceedings of Intelligent Information Processing[C]. Berlin, Germany: Springer Cham, 2016.169 – 178.

[8] Yang Y, Ma Z, Yang Y, et al. Multitask spectral clustering by exploring intertask correlation[J]. IEEE Transactions on Cybernetics, 2015, 45(5):1085 – 1090.

[9] 唐俊, 梁亮, 梁栋, 等. 基于拟Laplace谱的形状表示与聚类[J]. 华东理工大学学报, 2011, 37(6):749 – 753.
Tang Jun, Liang Liang, Liang Dong, et al. Shape representation and clustering based on quasi-Laplace spectrum[J]. Journal of East China University of Science and Technology, 2011, 37(6):749 – 753.(in Chinese)

[10] Zeng M, Cai Y, Liu X, et al. Spectral-spatial clustering of hyperspectral image based on Laplacian regularized deep subspace clustering[A]. Proceedings of IEEE International Geoscience and Remote Sensing Symposium[C]. NY, USA: IEEE, 2019. 2694 – 2697.

[11] Lei X, Zheng L, Liu Z, et al. Laplacian eigenmaps for automatic story segmentation of broadcast news[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1):276 – 289.

[12] Pirani M, Sundaram S. On the smallest eigenvalue of grounded Laplacian matrices[J]. IEEE Transactions on Automatic Control, 2016, 6(2):509 – 514.

[13] Liu X, Xiong H, Shen N. A hybrid model of VSM and LDA for text clustering[A]. Proceedings of the 2nd IEEE International Conference on Computational Intelligence and Applications[C]. NY,USA: IEEE, 2017.230 – 233.

[14] Li J, Nie F, Li X. Directly solving the original Ratiocut problem for effective data clustering[A]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing[C]. NY,USA: IEEE, 2018.2306 – 2310.

[15] Marutho D, Handaka S H, Wijaya E, et al. The determination of cluster number at K-means using elbow method and purity evaluation on headline news[A]. Proceedings of International Seminar on Application for Technology of Information and Communication[C]. NY,USA: IEEE, 2018.533 – 538.

[16] Xu T, Chiang H, Liu G, et al. Hierarchical K-means method for clustering large-scale advanced metering infrastructure data [J]. IEEE Transactions on Power Delivery, 2017, 32(2):609 – 616.

[17] Sapkota N, Alsadoon A, Prasad P W C, et al. Data summarization using clustering and classification: spectral clustering combined with K-means using NFPH[A]. Proceedings of International Conference on Machine Learning, Big Data, Cloud and Parallel Computing[C]. NY, USA: IEEE, 2019.146 – 151.

[18] Fontanini A D, Abreu J. A data-driven BIRCH clustering method for extracting typical load profiles for big data [A]. Proceedings of IEEE Power & Energy Society General Meeting [C]. NY,USA: IEEE, 2018. 1 – 5.

[19] Deng D. DBSCAN clustering algorithm based on density [A]. Proceedings of 7th International Forum on Electrical Engineering and Automation [C]. NY,USA: IEEE, 2020. 949 – 953.

作者简介

孟海宁　女,1979年生于内蒙古乌海.现为西安理工大学计算机科学与工程学院副教授、硕士生导师,主要研究方向为数据挖掘算法和可靠性建模.
E-mail:hnmeng@xaut.edu.cn

张贝贝　男,1978年生于陕西延安.现为西安理工大学计算机科学与工程学院讲师,主要研究方向为数据挖掘算法和大数据处理技术.
E-mail:bbzhang115@hotmail.com

冯　锴　男,1997年生于内蒙古锡林郭勒盟.现为西安理工大学计算机科学与工程学院硕士研究生.主要研究方向数据挖掘算法.
E-mail:wang389331557@163.com

童新宇　男,1996年生于陕西西安.现为西安理工大学计算机科学与工程学院硕士研究生.主要研究方向数据挖掘算法.
E-mail: tongxinyu@stu.xaut.edu.cn

朱　磊　男,1983年生于陕西咸阳.现为西安理工大学计算机科学与工程学院讲师,主要研究方向为数据挖掘算法和自然语言处理.
E-mail:leizhu@xaut.edu.cn

黑新宏　男,1976年生于陕西延安.现为西安理工大学计算机科学与工程学院教授、博士生导师,主要研究方向为机器学习和安全性评估.
E-mail: heixinhong@xaut.edu.cn