一种成对约束限制的半监督文本聚类算法

王纵虎[1,2] 刘 速[2]
(中国人民大学统计学院 北京 100872)[1] (中国石油规划总院计算机信息中心 北京 102206)[2]

摘 要 半监督聚类能利用少量标记数据来提高聚类算法性能,但大部分文本聚类算法无法直接应用成对约束等先验信息。针对文本数据高维稀疏的特点,提出了一种半监督文本聚类算法。将成对约束信息扩展后嵌入文档相似度矩阵,在此基础上根据已划分与未划分文档之间的统计信息逐步找出剩余未划分文本集合中密集的且与已划分聚类中心集合相似度较小的 K 个初始聚类中心集合,然后将剩余的相对较难区分的文档结合成对约束限制信息划分到 K 个初始聚类中心集合,最后通过融合成对约束违反惩罚的收敛准则函数对聚类结果进行进一步优化。算法在聚类过程中自动确定初始聚类中心集合,避免了 K 均值算法对初始聚类中心选择的敏感性。在几个中英文数据集上的实验结果表明,所提算法能有效地利用少量的成对约束先验信息提高聚类效果。

关键词 聚类,半监督,向量空间模型,成对约束,文本
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.033

Pairwise Constrained Semi-supervised Text Clustering Algorithm

WANG Zong-hu[1,2] LIU Su[2]
(School of Statistics, Renmin University of China, Beijing 100872, China)[1]
(Computer Information Center, Petrochina Planning and Engineering Institute, Beijing 102206, China)[2]

Abstract Semi-supervised clustering can use a small amount of tag data to improve the clustering performance, but most of the text clustering algorithms can not directly apply priori information such as pairwise constraints. As the characteristics of text data were high-dimensional and sparse, we proposed a semi-supervised document clustering algorithm. First, pairwise constraints were expanded and embedded in the document similarity matrix, then K density regions which have a small similarity with the already partitioned text collection were gradually searched in the remaining unpartitioned text collection as initial centroid. The remaining unpartitioned texts which are relatively difficult to distinguish were assigned to the K initial centroid according to the constraints. Finally, the clustering result was optimized by the convergence criterion function through integration of punish violations of pairwise constraints. In the clustering process, it can automatically determines the initial centroids to avoid the sensitivity to the initial centroids of K-means algorithm. Experimental results show that the proposed algorithm can effectively use a small amount of pairwise constraints to improve the clustering performance in Chinese and English text datasets.

Keywords Clustering, Semi-supervised, VSM, Pairwise constraints, Text

1 引言

文本聚类作为一种无监督的机器学习方法,由于不需要训练过程以及预先对文本类别进行手工标注,因此具有较高的灵活性和自动化处理能力,可以自动提取文本集合中的主题类别结构,辅助用户制定文档的分类体系或者改善现有分类体系,已经成为对文本信息进行有效组织、摘要和导航的重要手段。

在机器学习领域中,传统的学习方法有两种:监督学习和无监督学习。半监督学习(Semi-supervised Learning)是监督学习与无监督学习相结合的一种学习方法,它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。半监督学习对于减少标注代价、提高学习机器性能具有非常重大的实际意义,近年来逐渐成为机器学习领域的一个研究热点[1]。传统的聚类算法往往无法有效地利用先验监督信息,而半监督聚类算法则是研究如何在聚类算法中有效地利用这些先验信息来提高聚类性能,这些少量的先验监督信息可以是类别标记或样本对间的成对约束信息[2]。近年来,许多半监督聚类算法被提出并在很多实际领域(如生物信息处理、文本挖掘、图像处理)中获得了广泛应用[3]。目前国内对半监督聚类算法的研究在某些方面也有一些开创性工作。王玲等[4]将成对限制信息和空间一致性先验信息同时引入谱聚类中,提出了一种密度敏感的半监督谱聚类算法。尹学松等[5]提出了一种基于成对约束的判别型半监督聚类方法,利用监督信息集成数据降维和聚类。肖宇等[6]在近邻传播算法的基础上引入标记数据或成对约束信息对相似度矩阵

到稿日期:2016-01-20 返修日期:2016-03-26
王纵虎(1984—),男,博士后,工程师,主要研究方向为数据挖掘、大数据分析,E-mail:zonghuwang@petrochina.com.cn;刘 速(1984—),女,博士,工程师,主要研究方向为大数据分析,E-mail:sueliu@petrochina.com.cn。

进行调整，提出了一种基于近邻传播算法的半监督聚类方法。近年来，半监督学习也被用来提高文本聚类的质量。Huang等[7]提出了一种根据用户反馈主动选择有效的文档对来指导文本聚类的主动学习框架，通过估计词的共现概率，提出了一种增益指导的文本约束对主动选择方法。Wang等[8]认为类簇中心的选择对基于测度学习的半监督聚类质量的影响较大，因此可以利用对成对约束的满足程度来选择簇中心，通过将成对约束满足程度作为优化目标，通过迭代求解寻找最优结果。Shi Zhong等[9]比较了基于种子的、基于约束的和基于反馈的3种半监督聚类算法用于文本聚类时的性能，实现结果表明当标记信息足够多时，基于约束的半监督文本聚类算法性能最好；而当标记信息较少时，基于反馈的方法性能较好。文献[19]通过对约束集中所含信息量的衡量和对DBSCAN算法本身的分析，提出了一种启发式的主动学习算法，其能够选取含信息量大的成对约束集，从而能够更高效地辅助半监督文档聚类。

初始聚类中心选择不当将严重地影响K-Means类算法的有效性。Higgs[10]提出先用MaxMin算法对数据进行聚类，并将聚类结果的类中心作为K-Means的初始化值。这两种初始化方法本身就是一种聚类算法，因此其自身聚类结果的好坏将影响到K-Means算法。文献[11]借鉴了P. Pantel等[12]提出的CBC(Clustering By Committee)聚类算法，首先探测数据集中的相对密集区域，再利用这些密集区域生成初始类中心点。该方法能够很好地排除类边缘点和噪声点的影响，并且能够适应数据集中各个实际类别密度分布不平衡的情况。但此方法采用了两个固定的相似度阈值来判定一个数据区域能否成为一个密集区，算法对参数过于敏感。文献[13]结合谱图理论的思想，提出一种基于密度敏感的相似度量方法来初始样本值，有效确定K个代表性强的对象作为初始中心，加快了聚类过程。不过此算法中用到的密度参数和领域半径调节系数均需针对不同数据集事先给出合适的经验值，参数大小对于能否有效达到最优的聚类结果影响较大。这些算法虽然避免了传统K-Means算法中初始聚类中心的随机选取，但是还是存在着可能受噪声点或边缘点影响而使得算法陷入局部最优、需要事先人为设定一些阈值参数以及在高维数据聚类如文本聚类中可扩展性差等问题。

针对以上问题，提出了一种成对约束限制的半监督文本聚类算法。通过计算文本之间的相似度矩阵，将成对约束扩展并嵌入相似度矩阵，设计了一种新的初始聚类中心的选择方法，通过对已划分和未划分的文档集合的相似度矩阵信息进行统计，逐步寻找数据集中未划分归属类簇部分文档的相对密集区域的代表文档，若此密集区域代表文档与已划分类簇的覆盖度小于一定阈值，则将此代表文档及其所代表的密集区域中的文档集合作为一个初始化聚类中心。寻找到K个初始聚类中心后，根据文档与类簇的平均相似度信息和成对约束关联信息将剩余未划分的文档划分到最相似的类簇中，最后通过结合成对约束惩罚项的准则函数对聚类结果进行进一步优化。算法中的阈值均通过在聚类过程中对文档集合的嵌入成对约束的相似度矩阵信息进行动态统计时获得，避免了根据经验对不同数据集阈值设定的盲目性。该方法能够很好地排除类边缘点和噪声点的影响，能够适应数据集中各个实际类别大小及密度分布不平衡的情况，且在不同数据集上都能取得较好的结果，可伸缩性较强，能有效利用成对约束提高文档聚类效果。

2 成对约束监督信息嵌入

在许多聚类应用领域，以成对点限制形式出现的监督信息比样本类属信息更实用。因为对于用户而言，确定样本类别比较困难，而获得一些关于两个样本点是否属于一类的信息则相对容易。另外，基于限制的先验信息比类属信息更为一般化，可以从类别标记信息中获得等价的实例约束，反之则不然。Wagstaff等人最早引入两种类型的成对点约束[2](即must-link和cannot-link)来辅助聚类，must-link约束要求两个数据点必须在同一个类簇中，而cannot-link约束要求两个点不能在同一类簇中。Klein等人[14]在must-link和cannot-link成对约束限制的基础上，提出了融合二值传递关系方法的半监督聚类算法。该方法先使用最短路径算法添加扩展must-link限制，然后在改变了的测度空间中仅施加样本层面上的cannot-link限制，最后利用完全链接层次聚类算法将cannot-link限制进行扩散，从而充分利用成对约束信息指导聚类过程。这两种约束具有对称性：

$$\begin{aligned}&(x_i,x_j)\in must\text{-}link\Leftrightarrow(x_j,x_i)\in must\text{-}link\\&(x_i,x_j)\in cannot\text{-}link\Rightarrow(x_j,x_i)\in cannot\text{-}link\end{aligned}\tag{1}$$

以及具有限传递性：

$$\begin{aligned}&(x_i,x_j)\in must\text{-}link\ \&\ (x_j,x_k)\in must\text{-}link\\&\quad\Rightarrow(x_i,x_k)\in must\text{-}link\\&(x_i,x_j)\in must\text{-}link\ \&\ (x_j,x_k)\in cannot\text{-}link\\&\quad\Rightarrow(x_i,x_k)\in cannot\text{-}link\end{aligned}\tag{2}$$

由于文本向量高维稀疏的特点，采用余弦距离计算两个文档间的相似度，构成数据集n个文档间的相似度矩阵S，用$n\times n$的矩阵表示：

$$\begin{bmatrix}0 & s(1,2) & s(1,3) & \cdots & s(1,n)\\ s(2,1) & 0 & s(2,3) & \cdots & s(2,n)\\ s(3,1) & s(3,2) & \cdots & \cdots & \vdots\\ \vdots & \vdots & \vdots & 0 & \vdots\\ s(n,1) & s(n,2) & \cdots & \cdots & 0\end{bmatrix}\tag{3}$$

相似值$s(i,j)$值越大，则两个样本越相似，如果$s(i,j)=1$，则样本i和j完全相同。在计算数据集相似度矩阵时，保存所有样本对之间的最大相似度值为$MaxValue$。

在获得的数据集样本间相似度矩阵的基础上，将已知的某些样本对之间的两种类型的成对点限制(即must-link和cannot-link)信息嵌入矩阵。实验中采用Random函数随机产生成对约束对，即将数据集中的样本编号，每次随机产生一对数字，如果编号和这对数字相同，则两个样本同属一类，加入must-link正约束集；否则加入cannot-link负约束集。

除了用最初输入的某些样本对之间的两种类型的成对点限制对相似度矩阵进行调整以外，还参考文献[15]中的距离矩阵调整方法对其他样本之间的相似度进行调整扩展。对样本间成对约束对称性和传递性进行调整扩展后，可以发现隐藏在最初输入的样本对之间成对约束先验监督信息背后其他样本间的成对约束情况。

在将成对约束监督信息嵌入数据集相似度矩阵后，会生成已知的某些样本对之间的两种类型的成对点限制中满足must-link约束和cannot-link约束的样本对以及由初始的must-link和cannot-link约束集根据式(1)和式(2)约束关系的对称性和传递性扩展得到新的must-link约束和cannot-link约束。由于采用余弦距离计算样本间相似度，如果两个

样本属于同一类，则为 must-link 约束，将其相似度值设置为数据集所有样本对之间的最大相似度 $MaxValue$；若两个样本不属于同一类，则为 cannot-link 约束，则将其相似度值设置为 0。

Floyd 算法可以计算多源目标间的最短路径，而这里只需求出两个样本之间的成对约束情况。将 Floyd 算法进行改造成求解扩展的成对约束的算法，如下：

步骤 1　根据最初加入的成对约束建立数据集成对约束标记矩阵 M，矩阵大小为样本的总数，定义如下：

$$m[i,j]=\begin{cases}1, & 当(x_i,x_j)\in must\text{-}link\\ -1, & 当(x_i,x_j)\in cannot\text{-}link\end{cases} \tag{4}$$

步骤 2　在成对约束标记矩阵的基础上利用改造过的 Floyd 算法求解成对约束扩展情况，算法如下。

输入：成对约束标记矩阵 M，样本总数 n

输出：经过扩展的成对约束标记矩阵 M_{new}

```
1. for k=1 to n
2.   for i=1 to n
3.     for j=1 to n
4.       if(M^{k-1}[i,k]==1 and M^{k-1}[k,j]==1)
5.         Then M^{k-1}[i,j]=1;
6.       if( (M^{k-1}[i,k] * M^{k-1}[k,j]==-1)
7.         Then M^{k-1}[i,j]=-1;
8.     end for
9.   end for
10. end for
```

步骤 3　根据扩展后的成对约束标记矩阵 M_{new} 修正数据集相似度矩阵 S 中样本对之间的相似度值，计算如下：

$$s[i,j]=\begin{cases}0, & 当\ m_{new}[i,j]=1\\ MaxValue, & 当\ m_{new}[i,j]=-1\end{cases} \tag{5}$$

实验表明本文提出的改造过的 Floyd 算法能有效地获取隐藏在最初获得的成对约束中经过传递而获得的更多的 must-link 和 cannot-link 成对约束先验信息。

3　基于聚类过程统计的文本划分聚类算法

文本由词构成，不同的词的集合反映了不同的类簇的主题，一个类簇的特征通过某些词的集合反映出来。在文本预处理及特征选择后获得的数据集特征词空间中，不同类簇中包含的文本的权重可能只在特征词空间中的某些维较大。当文本数量较多时，同一类簇中不同文本更只是占了反映该类簇主题特征的特征词子空间中的某些维。文本数据由于其高维性和稀疏性，一般在不同类簇之间文本的相似度往往较小，而同类之间文本数据之间相似度也可能只是类簇中某几个文本之间的较大。同类簇中两个文本可能只是通过某几个特征词的共现而聚集在一起，同类簇中不同文本通过特征词的交集的传递而产生联系。

划分聚类算法希望选出的初始聚类中心具有以下性质：1）初始中心应该来自不同的类簇；2）初始中心应该尽量靠近类簇的中心；3）应该保证初始中心之间的相似度尽量小。本文基于上述性质和文本数据的特点提出了一种新的在聚类过程中选择初始聚类中心的方法。

在前面得到的嵌入成对约束限制的文本相似度矩阵的基础上，通过对已划分类簇文档集合和未划分类簇文档集合信息进行统计，逐步寻找未划分文档集合中的密集区域及该区域的代表文档，如果该密集区域与已划分类簇集合的覆盖度小于一定阈值，将该代表文档作为一个种子，将该密集区域中的满足阈值条件的文档一起作为一个初始聚类中心，直到获得预定数目的初始聚类中心，将剩余文档划分到最相似的初始聚类中心代表的类簇中，最后通过结合成对约束违反惩罚的准则函数对聚类结果进行优化。通过将成对约束先验信息嵌入文本相似度矩阵，使得已划分类簇集合和未划分类簇集合之间更容易区分。本文提出的文本聚类算法过程如图 1 所示。

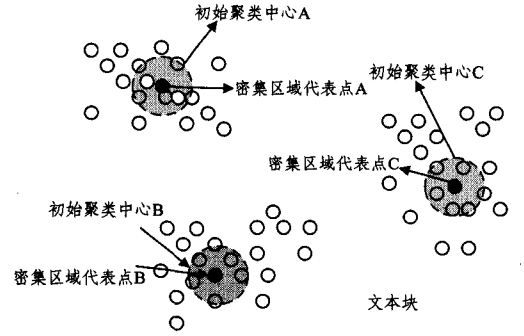

图 1　文本聚类算法聚类过程

如图 1 所示，数据集包含 3 个类簇，3 个虚线圆中标识的圆点代表密集区域代表点，虚线圆代表初始聚类中心。按照本文所提方法的思想，整个数据集数据划分的顺序为：密集区域代表点 A、初始聚类中心 A、密集区域代表点 B、初始聚类中心 B、密集区域代表点 C、初始聚类中心 C，形成了 3 个初始聚类中心后，最后将数据集剩余的其他类簇边缘文档添加到最相似的类簇中。算法中的密集区域代表点和初始聚类中心都是通过本文算法在聚类过程中自动识别生成。

算法在聚类过程中用到的针对已划分部分与未划分部分的文档集合统计信息定义如下。

定义 1　未划分类簇文档集合中文档 d_i 与未划分类簇文档集合中其他文档的平均相似度为 $RS(d_i)$，定义如下：

$$RS(d_i)=\frac{1}{n}\sum_{j=1}^{n}s(i,j) \tag{6}$$

定义 2　未划分类簇文档集合内部平均相似度为 RA，定义如下：

$$RA=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}s(i,j) \tag{7}$$

定义 3　未划分类簇文档 d_i 密集区域代表度为 $D(d_i)$，定义如下：

$$D(d_i)=\sum_{j=1}^{n}p(i,j) \tag{8}$$

其中，$p(i,j)=\begin{cases}s(i,j), & 当\ s(i,j)\geqslant RA\\ 0, & 当\ s(i,j)<RA\end{cases}$；$n$ 为尚未划分类簇的文档数目，计算 $D(d_i)$ 时，只有当文档间相似度大于未划分类簇文档总体平均相似度 RA 时才参加累加，从而排除那些与文档 d_i 相似度过小的一些噪声文档的影响。这一定义反映了一个文档在剩余未划分文档集合中对其所属的某个密集区域的代表能力，值越大，此文档越适合作为密集区域的代表。

定义 4　未划分类簇文档 d_i 与已划分初始聚类中心类簇的整体平均覆盖度定义如下：

$$AC(d_i)=\frac{1}{m}\left(\sum_{j=1}^{m}\sum_{d_j\in C_k}s(i,j)\right) \tag{9}$$

其中，m 为已划分的初始聚类中心类簇的数目。这一定义反映了文档 d_i 与所有已划分的初始聚类中心类簇的相似度，通

过对其进行统计，可以发现那些与已划分初始聚类中心类簇相似度都较大的文档，这类文档不适合作为初始聚类中心。

定义 5 未划分类簇文档 d_i 与某个已划分初始聚类中心类簇的单类簇覆盖度定义如下：

$$CC(d_i)=\max(\sum_{d_j\in C_k} s(i,j)) \tag{10}$$

其中，C_k 为已经划分出的初始聚类中心类簇，$k=1,\cdots,m$，m 为已划分的初始聚类中心类簇的数目。这一定义反映了文档 d_i 与某一个已划分初始聚类中心类簇的相似程度，与某个已划分初始聚类中心类簇相似度较大的文档同样不适合作为初始聚类中心。

根据上述定义，本文提出的基于过程统计的半监督文本划分聚类算法如下。

步骤 1　根据式(6)－式(8)计算所有未划分类簇整体和每个文档的平均相似度及密集代表度，将其加入候选种子列表。

步骤 2　若已划分类簇集合不为空，则根据式(9)、式(10)计算未划分类簇文档与已划分初始聚类中心类簇的整体平均类簇覆盖度和单类簇覆盖度。

步骤 3　对步骤 1 中的候选种子列表进行筛选，删除覆盖度大于整体平均类簇覆盖度和平均单类簇覆盖度的文档，对余下的文档按密集代表度进行降序排序。

步骤 4　返回候选种子列表的第一个文档作为下一个初始聚类中心种子，初始化初始聚类中心类簇，加入初始聚类中心类簇集合。

步骤 5　计算所有未划分类簇文档与步骤 4 中得到新的初始聚类中心类簇的相似度。根据相似度对所有文档进行降序排序，选择相似度最大的文档，若此文档与当前初始聚类中心类簇的相似度大于此文档的平均相似度 $RS(d_i)$，则将此文档加入当前初始聚类中心类簇，重复步骤 5，否则进入步骤 6。

步骤 6　若当前初始聚类中心类簇中的文档数目不再增加，则进入步骤 1，否则进入步骤 7。

步骤 7　若初始聚类中心类簇集合中类簇数目到达 K 个，则按照式(10)计算剩余未划分类簇的文档与 K 个初始聚类中心类簇的相似度，将其划入相似度值最大的类簇。

步骤 8　得到 K 个类簇后根据聚类结果计算准则函数的值。

步骤 9　计算每个文档与 K 个类簇的相似度情况，若文档的最大相似类簇不是包含此文档的类簇，则将此文档从原类簇移动到新的相似度最大的类簇。

步骤 10　根据聚类结果计算准则函数的值，若值不再发生变化则停止计算，返回 K 个类簇，否则进入步骤 9 继续进行迭代优化。

通过计算文档的密集代表度，能较好地搜索数据集中能较好代表密集区域中心的文档作为初始聚类中心，减小了选择边缘点和噪声点作为初始聚类中心的可能，通过计算覆盖度保证了这些作为候选聚类中心的不同的密集区域间的相似度较小。算法动态地计算平均文档相似度、平均覆盖度和单类簇覆盖度，以其作为候选类簇聚类中心种子的阈值，较好地把握了数据集未划分类簇的文档集合的整体信息。本文算法数据集中类内相似度较大、类内包含文档较多的类簇会被优先选作初始聚类中心。

步骤 7 中划分剩余文档、计算文档与类簇相似度时，使用的相似度公式如下：

$$Similarity(d_i,C_j)=\begin{cases}\dfrac{\sum_{d_k\in C_j} s(i,k)}{|C_j|}, & \text{当}\sum_{d_k\in C_j} m(i,k)=0\\ \sum_{d_k\in C_j} m(i,k), & \text{当}\sum_{d_k\in C_j} m(i,k)\neq 0\end{cases} \tag{11}$$

其中，$s(i,k)$ 为文档 d_i 和类簇 C_j 中文档 d_k 的余弦相似度，$m(i,k)$ 为扩展后的成对约束标记矩阵 M_{new} 中文档 d_i 和类簇 C_j 中文档 d_k 按式(4)取值的成对约束情况。算法在计算剩余文档与已划分类簇相似度时，同时统计文档与类簇的平均相似度信息和成对约束关联信息，这些值可以直接通过前面计算的文档相似度矩阵和扩展后的成对约束标记矩阵中的对应项累加得到。如式(11)所示，若一个文档与某一类簇中存在 must-link 和 cannot-link 情况时，则式(11)返回成对约束关联信息的统计值作为相似度值，否则返回文档与类簇的平均相似度信息。步骤 7 中的剩余未划分类簇文档往往是较难区别的类别，通过引入成对约束信息后，可以通过式(11)直接根据其与已划分类簇中文档的成对约束关联信息的统计来情况确定其类别归属，从而有效地提高聚类的准确率。

本文采用的聚类收敛准则函数定义如下。

由于采用余弦距离计算文档间相似度，算法通过最大化式(13)来寻找最优聚类结果。

$$f(X)=\delta\cdot\sum_{r=1}^{m}\sum_{d_i\in C_r}(\frac{1}{n}\sum_{d_j\in C_r} s(i,j)) \tag{12}$$

其中，δ 为划分的类簇中文档间的成对约束违反情况，在算法中当无成对约束信息时值为 1，当添加成对约束时定义如下：

$$\delta=(C/A+M/B)/2 \tag{13}$$

其中，A 为嵌入到相似度矩阵中的所有 cannot-link 成对约束的总数，B 为嵌入到相似度矩阵中的所有 must-link 成对约束的总数，C 为聚类结果中各个划分的类簇中遵守 cannot-link 成对约束的样本对的总和，M 为各个划分的类簇中遵守 must-link 成对约束的样本对的总和。在适应度函数中添加 δ，当加入成对约束监督信息后可以统计聚类结果中违反成对约束先验信息的情况来指导算法收敛，δ 根据聚类生成类簇中成对约束的分布正确与否的情况来指导聚类。采用余弦距离计算文档间相似度，采用的适应度函数同时考察了各类簇样本与类簇中心的平均距离和违反成对约束的情况。

所提方法将成对约束监督信息扩展后嵌入数据集相似度矩阵中，使得在寻找 K 个初始聚类中心集合时，已划分子簇与未划分类簇文档间的相似信息更加清晰。在获得 K 个初始聚类中心集合后，同时统计文档与已划分类簇的平均相似度信息和成对约束关联信息，从而有效地根据文档与某个类簇的成对约束关系判定文档的类别归属。最后通过最大化融合各类簇样本与类簇中心的平均距离和违反成对约束情况的乘积作为收敛准则函数进行迭代计算，使得算法收敛于最优结果。

4　实验结果

实验系统采用 C# 语言实现，在 CPU 为 3.2GHz、2GB 内存环境下测试。

本文对比实验采用明尼苏达大学 Karypis 教授提供的 CLUTO 聚类工具集[16]中的聚类算法。CLUTO 可以用于高维数据如文本数据的聚类分析，提供了多种聚类方法，并可以对聚类结果按多种标准进行评价。将实验数据转化为 CLU-

TO 定义的输入格式并选择了其中性能较好的二分划分聚类(rb)、凝聚层次聚类(agglo)、图划分聚类(graph) 3 种方法与本文的方法进行了对比,CLUTO 的参数都采用默认设置。

实验中采用的中文语料库为搜狗实验室提供的搜狗中文语料库。英文数据集为著名的 20Newsgroups 和 reuters-21578。本文采用自己编写的数据抽取程序从上述数据集中抽取了部分数据生成实验数据集,在用程序抽取测试集时,分别循环地从每个类抽取若干次直到抽取到预定数量的文档,使得数据集中同一类文档不会按顺序聚集在一起,以检验算法对输入记录次序的敏感性。其中实验数据集 SG1 和 SG2 从搜狗语料库抽取,NG1 和 NG2 从 20Newsgroups 抽取,RS1 和 RS2 从 reuters-21578 抽取。实验中采用中科院 ICTCLAS 中文分词系统和从网址(http://tartarus.org/~martin/PorterStemmer/)下载的英文分词算法分别对中文和英文文本数据集进行分词预处理,然后进行停用词过滤、特征词选择、向量表示等预处理,特征词选择时去除了那些只在一篇文档中出现的词从而进行降维,各个数据集的详细信息如表 1 和表 2 所列。

表 1 数据集信息

数据集	类数	文档数	未降维词数	降维后词数	特征词数	最小类文档数	最大类文档数
SG1	5	500	262448	18708	8798	30	187
SG2	7	845	425812	123468	12057	156	51
NG1	5	500	109269	10738	5462	55	144
NG2	9	1204	261512	17702	9293	66	184
RS1	5	500	86995	42716	3383	27	192
RS2	8	846	108572	55315	3824	51	163

表 2 成对约束扩展情况

数据集	约束类型	100	200	300	400	500	600	700	800	900	1000	1500
SG1	ML	24	54	80	95	135	150	191	220	269	257	374
	CL	76	146	220	305	365	450	509	580	631	743	1126
	EML	28	256	144	176	471	448	613	1033	5194	5831	14367
	ECL	98	75	479	828	2201	2651	4472	8953	28062	27578	59607
NG1	ML	26	70	78	108	148	170	201	236	259	283	426
	CL	74	130	222	292	352	430	499	564	641	727	1074
	EML	28	113	125	203	437	1493	2094	7995	9625	7922	21454
	ECL	93	227	406	885	1576	6180	8945	26636	33853	33032	64214
RS1	ML	26	55	86	113	125	167	201	225	248	272	413
	CL	74	145	214	287	375	433	499	575	652	728	1087
	EML	30	85	144	224	301	824	1359	2988	6123	6995	20119
	ECL	91	229	477	990	1416	5135	9453	15879	28018	30489	67776

从表 1 可以看出,本文实验采用的 6 个数据集包含的类簇数目和类簇内文档数目都各不相同,且类簇间包含的文档数目相差较大,这些都给文档聚类带来了一定的挑战。

图 2 为本文算法在 6 个文本数据集上未添加成对约束信息时与 CLUTO 聚类器中 3 种方法的聚类纯度值的比较。从图中可以看出本文在各个数据集上都取得了最优效果,特别是在 20Newsgroups 数据集上,聚类效果提升明显,说明本文提出的基于聚类过程统计的文本划分聚类算法的思想是正确的,能较好地处理不同类簇数目与类内文档数目的文本数据集。

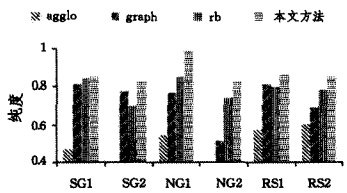

图 2 聚类效果比较(纯度)

表 2 为 SG1,NG1 和 RS1 在添加不同数目的初始 Must-link(ML)和 Cannot-link(CL)时,经过本文提出的改造过的 Floyd 算法进行扩展后得到的 Must-link(EML)和 Cannot-link(ECL)数目的情况。从表 2 中可以看出,经过本文提出的改造过的 Floyd 算法进行扩展后,可以利用的成对约束数目显著增加,从而能充分地挖掘隐藏在初始加入的成对约束对背后数据集中文档间的成对约束关联情况。

图 3—图 5 为在添加不同数目(100~1500)成对约束对条件下,本文算法和二分划分聚类(rb)在不同数据集上分别运行 20 次获得的平均 F-Measure 值的变化曲线图,每次运行都随机获得一定数目的成对约束对从而避免选择不同成对约束对对聚类结果改进贡献的差异。

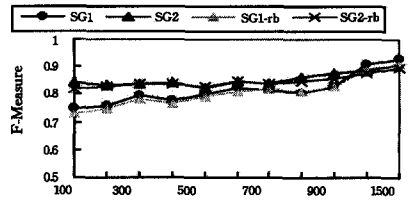

图 3 SG 不同约束对数目下聚类结果的 F-Measure 值变化曲线

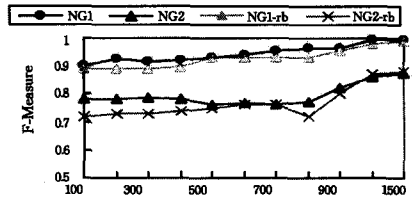

图 4 NG 不同约束对数目下聚类结果的 F-Measure 值变化曲线

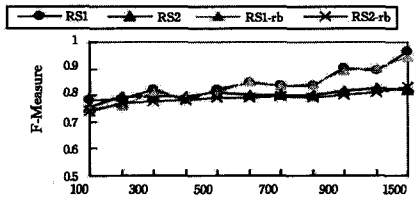

图 5 RS 不同约束对数目下聚类结果的 F-Measure 值变化曲线

从图中可以看出,本文方法在添加少量成对约束后能有效提高各个数据集的聚类结果的 F-Measure 值。在 SG1,SG2,NG1 和 RS1 上,当成对约束数目达到 1500 对以上时,聚类结果的 F-Measure 值都能达到 0.9 以上,且聚类结果随着成对约束数目的增加不断改进。当数据集包含的类簇数目较

多时，成对约束的增加对聚类效果的改进效果增幅，SG2 包含 7 类，RS2 包含 8 类，其 F-Measure 曲线增幅较包含较少类簇的数据集明显减缓。由于添加不同的成对约束对集合对聚类结果的贡献可能有所差别，因此图 3－图 5 中曲线并不是随着成对约束数目的增加而不断上升的，而是有所波动，但当成对约束达到一定数目后，波动变小。二分划分聚类（rb）F-Measure 值也随着成对约束的增加而提升，但本文算法整体提升效果更加明显。

结束语 本文提出了一种基于聚类过程统计的半监督文本聚类算法，通过在数据集相似度矩阵中嵌入经过扩展的成对约束监督信息，同时通过分析文本数据的特点和聚类过程中已划分类簇文档与未划分类簇文档间的关系，从而更好地识别出初始聚类中心，获得 K 个较能代表数据集组成的初始聚类中心子簇集合后，将剩余的相对较难区分的文档划分到最相似的子簇时，直接根据其与已划分类簇中文档的成对约束关联信息的统计情况确定其类别归属，从而充分利用成对约束信息指导聚类过程，最后对最大化融合各类簇样本与类簇中心的平均距离和与违反成对约束情况的收敛准则函数进行迭代计算，使得算法收敛于最优结果。实验结果显示本文方法能有效地根据文本数据的特点利用半监督信息提高文本聚类效果。本文添加的成对约束对都是随机选择的，但是通常相似度较小，但实际上属于同一类的样本对之间的约束信息对聚类的贡献更大，近年来基于主动学习的半监督聚类方法[17-20]被提了出来，下一步工作准备研究如何主动地获得成对约束信息，如获取采用本文所提算法在获得初始聚类中心集合后剩下的较难区分的未划分类簇文本与已划分类簇文本的成对约束情况，从而更有效地利用成对约束信息指导聚类。

参 考 文 献

[1] Wagstaff K, Cardie C. Clustering with Instance-Level Constraints [C]// Proc of the17th International Conference on Machine Learning, San Francisco, USA. 2000:1103-1110

[2] Li Xue-mei, Wang Li-hong, Song Yi-bin. A Hybrid Constrained Semi-Supervised Clustering Algorithm[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(4):452-456(in Chinese)
李雪梅，王立宏，宋宜斌. 一种混合约束的半监督聚类算法[J]. 模式识别与人工智能，2011，24(4)：452-456

[3] Li Kun-lun, Cao Zheng, Cao Li-ping, et al. Some Developments on Semi-Supervised Clustering[J]. Pattern Recognition and Artificial Intelligence, 2009, 22(5):735-742(in Chinese)
李昆仑，曹铮，曹丽萍，等. 半监督聚类的若干新进展[J]. 模式识别与人工智能，2009，22(5)：735-742

[4] Wang Ling, Bo Lie-feng, Jiao Li-cheng. Density-Sensitive Semi-Supervised Spectral Clustering[J]. Journal of Software, 2007, 18(10):2412-2422(in Chinese)
王玲，薄列峰，焦李成. 密度敏感的半监督谱聚类[J]. 软件学报，2007，18(10)：2412-2422

[5] Yin Xue-song, Hu En-liang, Chen Song-can. Discriminative Semi-Supervised Clustering Analysis with Pairwise Constraints[J]. Journal of Software, 2008, 19(11):2791-2802(in Chinese)
尹学松，胡恩良，陈松灿. 基于成对约束的判别型半监督聚类分析[J]. 软件学报，2008，19(11)：2791-2802

[6] Xiao Yu, Yu Jian. Semi-Supervised Clustering Based on Affinity Propagation Algorithm[J]. Journal of Software, 2008, 19 (11): 2803-2813(in Chinese)
肖宇，于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报，2008，19(11)：2803-2813

[7] Huang R Z, Lamb W. An active framework for semi-supervised document clustering with languages modeling [J]. Data & Knowledge Engineering, 2008, 68(1):1136-1155

[8] Wang J LWu S Y, Vu H Q, et al. Text document clustering with metric learning[C]// Proc. of the 33rd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval. 2010:783-784

[9] Shi Zhong. Semi-supervised model-based document clustering: A comparative study[J]. Machine Learing, 2006, 65(3):3-29

[10] Higgs R E, Bemis K G, Watson I A, et al. Experi-mental designs for selecting molecules from large chemical databases[J]. Journal of Chemical Information and Computer Sciences, 1997, 37(5):861-870

[11] Qin Yu, Jing Ji-wu, Xiang Ji, et al. An improved K-means algorithm based on optimizing initial points[J]. Journal of the Graduate School of the Chinese Academy of Science, 2007, 24(6): 771-777(in Chinese)
秦钰，荆继武，向继，等. 基于优化初始类中心点的 K-means 改进算法[J]. 中国科学院研究生院学报，2007，24(6)：771-777

[12] Pantel P, Lin D. Document clustering with committees[C]// Proceedings of the 25 th International Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM Press, 2002:199-206

[13] Wang Zhong, Liu Gui-quan, Chen En-hong. A K-means Algorithm Based on Optimized Initial Center Points[J]. Pattern Recognition and Artificial Intelligence, 2009, 22(2): 299-304(in Chinese)
汪中，刘贵全，陈恩红. 一种优化初始中心点的 K-means 算法[J]. 模式识别与人工智能，2009，22(2)：299-304

[14] Klein D, Kamvar S D, Manning C. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering [C] // Proceedings of the Nineteenth International Conference on Machine Learning. 2002: 307-314

[15] Bilenko M, Basu S, Mooney R J. Integrating Constraints and Metric Learning in Semi-Supervised Clustering[C]// Proc of the 21st International Conference on Machine Learning. 2004:81-88

[16] Karypis G. CLUTO: a clustering toolkit [OL]. http://glaros.dtc. umn. edu/gkhome/cluto/cluto

[17] Wang Na, Li Xia. Active Semi-Supervised Spectral Clustering Based on Pairwise Constraints [J]. Acta Electronica Sinica, 2010, 38(1):172-176(in Chinese)
王娜，李霞. 基于监督信息特性的主动半监督谱聚类算法[J]. 电子学报，2010，38(1)：172-176

[18] Su Ying-bin, Du Xue-hui, et al. Sensitive Information Inference Method Based on Semi-supervised Document Clustering [J]. Computer Science, 2015, 42(10):132-137(in Chinese)
苏赢彬，杜学绘，等. 基于半监督聚类的文档敏感信息推导方法[J]. 计算机科学，2015，42(10)：132-137

[19] Zhao Wei-zhong, Ma Hui-fang, Li Zhi-qing, et al. Efficiently active learning for Semi-Supervised document clustering[J]. Journal of Software, 2012, 23(6):1486-1499(in Chinese)
赵卫中，马慧芳，李志清，等. 一种结合主动学习的半监督文档聚类算法[J]. 软件学报，2012，23(6)：1486-1499

[20] Xiong S, Azimi J, Fem X Z. Active learing of constraints for semi-supervised clustering[J]. IEEE Transactions on know-ledge and Data Engineering, 2014, 26(1):43-45