汉语文本聚类及其算法设计

陈炯[1,2]，范卓华[2]，张虎[2]

（1．山西综合职业技术学院电子分院，山西 太原 030006；

2. 山西大学 计算机与信息技术学院 ,山西 太原 030006）

摘要:本文主要针对传统的聚类算法倾向于识别大小类似的球形聚类簇，且对离群数据较为敏感等问题，利用聚类簇代表点选取的方法，同时结合基于人进行聚类判断所遵循的基本原则，即聚类中对象间距离应小于聚类间距离，设计了一种有效的聚类算法，实验结果表明算法是有效的。

关键词: 聚类 代表点 聚类簇 聚类中心

中图分类号：TP391 **文献标识码：A**

引 言

随着互联网的出现，大量的文本信息如潮水般不断涌现，网络已经成为一个庞大而杂乱无章的桌面图书馆。对海量的文献人们迫切需要能够自动实现文本的分类处理，在节省时间的同时更好的定位查找自己需要的文献。有效的信息检索需要有良好的索引和文本内容概括，文本聚类便是解决这类问题的一种手段。

文本聚类就是将一个训练文献集分成若干称为聚类簇（cluster）的子集，每个聚类簇中的成员之间具有较大的相似性，而聚类簇之间的文本具有较小的相似性。文本分类一般是通过统计方法或知识工程方法来实现的。知识工程方法需要编制大量的推理规则，因此其开发费用相当昂贵。相比之下，统计方法由于其简单的机制，为大多数实用文本分类系统所采用。在基于统计的各种分类方法中,它们的共同点是从文本中提取词汇信息，并以特征向量的形式来表示文本。基于以向量来表示的文本，聚类算法有很多种，本文是通过在特征向量中选取代表点来完成聚类的。

1 聚类分析法

在日常生活、生产、科研、工作中，经常要对被研究的对象分类。研究和处理给定对象分类的数学方法称为聚类分析（Clustering Analysis)。

聚类算法是数据挖掘中常用的方法之一。通常可以分为层次式和非层次式两种。本文介绍的是层次式聚类方法。其优点是聚类的形成一般依赖于数据，而不是通过用户预定义的聚类数得到。它通过对初始数据构造一个聚类层次来完成聚类。初始，输入的每个数据点被看成一个单独的聚类簇，然后将成对的聚类簇一一合并，同时依据各聚类簇中各对象间的最大距离应小于各聚类簇之间的最小距离的原则，在包含 N 个对象的 m 维单位空间中，对象间的平均距离为$1/\sqrt[m]{N}$。按照“各聚类簇中对象间距离不应超过此标准，而各聚类簇之间距离不应低于此标准”规则，来结束聚类的。聚类簇合并的每一步，是合并距离最小的一对聚类簇。常用的聚类簇合并策略有：合并重心最为靠近的一对聚类簇；考察分属不同聚类簇的点之间的距离，并合并距离最小的一对数据点所对应的两个聚类簇；合并所有数据点间距离的平均值最小的两个聚类簇；考察分属不同聚类簇的点对之间的最大距离，合并该值最小的点对各自所在的聚类簇。如果待确定的各聚类簇内部数据点分布比较紧凑，且各聚类簇之间足够远离，这些策略都会得到较好的结果。然而，如果各聚类簇比较靠近（即使一些聚类簇之间是由离群数据连接的)，或者聚类簇的形状不是超球形的且聚类簇的大小差异较大，则采用不同的合并策略产生的结果有相当大的差别。如果待聚类的资料为长条形结构，可能会使长条形聚类簇被割裂开，而且会将割裂的属于不同聚类簇的子聚类簇合并成一个单独的聚类。从以上的分析可以看出，基于重心的方法和考虑所有点的方法都不适用于非球形、任意形状的聚类。基于重心的方法缺点在于它仅仅用一个点（即聚类簇的重心）来代表整个聚类簇。对于一个大的聚类簇或是一个任意形状的聚类簇，它的各个子聚类簇的重心可能会距离相当远，这样就会导致这个聚类簇被分割开。另一方面，考虑所有点的方法用一个聚类簇内的所有点来代表它，这样就会使得聚类算

法对离群数据极度敏感，而且数据点位置的微小变化都会导致算法结果有较大变动。此外大多数聚类学习算法，均需用户事先给定聚类个数 K，且聚类结果对 K 值大小都很敏感，不同 K 值的聚类学习结果往往大相径庭，但就目前的状况来说，如何选择合适的 K 值，本身就是一个难题。

2 算法设计及其实现

2.1 问题描述

给定任一训练文献集，根据某一聚类算法将其分为内容相近的几类。本文的聚类过程是建立在文献已经分词并进行了词频统计的基础上的。利用分词和词频统计结果产生的词长、词频以及其它有用数据来随机抽取特征词，根据抽取出的特征词形成每个文本的特征向量，分别求出任两个向量的欧氏距离，按具体算法对文本聚类。

2.2 算法思想

本文采用一种新的分层聚类方法，它介于仅基于重心点的算法和考虑所有点的算法之间，兼有两者的长处。首先在一个聚类中选取固定数量充分分散的点，这些分散的点捕捉到了整个聚类簇在形状和范围上的特征。然后把这些分散的点以一个介于 0 到 1 之间的因子向它们所在聚类簇的中心收缩，由此得到的点集即用来代表聚类簇，在分级运算的每一步中，哪一对聚类的代表点集距离最近，就合并这两个聚类簇。当计算两个聚类簇之间的距离时，从这两个聚类簇中各取一个代表点并计算它们的距离，所得距离的最小值即为这两个聚类簇之间的距离。采用这种方法，计算一个聚类簇与其它聚类簇之间的距离时只需用它的代表点集即可。

2.3 模块设计

系统主要由四个模块组成：分词模块、特征词抽取模块、向量距离计算模块、文本聚类模块，各模块的功能如下：

①分词模块：将文章分成词、词组或短语。

②特征词抽取模块：根据分词结果进行词频统计，抽取文本特征词。

③向量距离计算模块：对每个文本向量分别进行两两之间的欧氏距离计算。

④文本聚类模块：按各篇文本向量之间的距离，通过算法规则，对其进行归类合并，直到分类结果符合一定规则而不能再合并为止。

本文主要研究的是文本聚类模块，它的算法思想是基于中国科学技术大学计算机系陈恩红等人提出的一种利用代表点的有效聚类算法和中国科学技术大学自动化系朱明等人提出的一种聚类学习的新方法。

2.4 具体实现

2.4.1 基本概念

①训练文献集：待测的文本的集合。

②词频 f_k：第 k 个特征项在文本中出现的频率。

③反比文献率 idf_k：是第 k 个特征项在训练文献集中的反比文献频率

$$idf_k=[\log_2 n-\log_2 d_k]+1$$

其中，n 是训练文献集的文本总数；d_k 是训练文献集中含第 k 个特征项的文本数。

④文本特征向量 v：为完成文本的向量化表示，可利用中文分词技术将每个文本分解成若干词的序列，然后进行特征词抽取，抽取出的这些特征词就构成了与这个文本对应的多元组：

$$v=(w_1, w_2, \cdots, w_n) \qquad \text{其中，} w_k=f_k*idf_k$$

⑤距离 d（x，y）：根据欧氏距离公式，设文本的某一代表点 x 和 y 的特征向量分别为：

$$v_x=(W_{x1}, W_{x2}, \cdots, W_{xn}),\quad v_y=(W_{y1}, W_{y2}, \cdots, W_{yn})$$

$$\text{则：} d(x, y)=((W_{x1}-W_{y1})^2+(W_{x2}-{}_wW_{y2})^2+\cdots+(W_{xn}-W_{yn})^2)^{1/2}$$

⑥聚类簇的中心点：两个聚类簇 C_1，C_2 合并成新的聚类簇 C 时，C 的中心可以由 C_1 和 C_2 的中心算出。设 C_1 和 C_2 中的数据点数分别为 n_1 和 n_2，且每个点的重量一样，则：

C．mean=（n_1* C_1.mean+n_2*C_2..mean）/（n_1+n_2）

其中 C．mean 用来存储 C 中点集的中心点

⑦代表点的选取：把聚类簇的数据空间根据代表点的个数以聚类簇的中心为中心分为数目相等的区域，在每个区域中选择一个离中心的最远的点作为该区域内的代表点。

⑧对象间的平均距离：$\Phi = 1/\sqrt[m]{N}$。其中 N 是包含的对象的总数；m 指的是数据维数。

2.4.2 具体实现

①算法描述

a、 读入数据。读入每篇文章的特征词提取结果及其词频。

b、 构造文章 i 的特征向量 v_i=（W_{i1}，W_{i2}，…，W_{in}）

W_k（i，j）=f_k（i，j）*idf_k（i，j）

f_k（i，j）表示第 i 篇文章第 j 个特征词的词频

idf_k（i，j）表示文献中第 j 个特征词的反比文献频率

c、利用代表点开始聚类

②输出聚类结果。

3 实验及结果分析

选用包含 100 篇文本的文献集进行测试，结果如下：

测试文本	分的类数	含一篇以上文章（类数）		包含文本最多的数	含一篇文章（类数）	
100	21	16	76.2%	16	5	23.8%

对实验结果分析发现：

①实验结果与代表点的个数和收缩因子的值有很大关系，选取不同数目的代表点和收缩因子取不同的值对边缘点的影响都会比较大；因收缩因子等于 1 时本算法就是基于重心的算法，而收缩因子等于 0 代表点取所有点时本算法就是基于所有代表点的算法。经过测试代表点选 10，而收缩因子选取 0.15 时效果较好。

②因为测试数据选取有限，所以没能进行更广泛的测试，通过观察分类结果，发现其中有很多是一篇文章一类，而一类最多包含的文章数是 16 篇，可见每一类所包含文本数不太平衡，说明在许多方面还有待改进。

4 结束语

传统的聚类算法要么只对球形聚类簇效果比较好，要么只对相似形状的聚类簇效果较好，要么对离群数据太敏感，当涉及到聚类簇的形状不规则，各个聚类簇的形状差异较大时，或离群数据较多的情况下，这些算法往往难以得出满意的结果。本文的聚类算法不但可以识别形状各异的聚类簇，而且几乎不受离群数据的影响。

参考文献

1 陈恩红，王上飞，宁岩，王煦法. 一种利用代表点的有效聚类算法设计与实现[J]. 模式识别与人工智能，2001.12：417～421

2 李聪，张勇，高智. 一种新的聚类算法[J]. 模式识别与人工智能，1999，6：205~208

3 Leouski A V, Croft W B.An Evaluation of Techniques for Clustering Search Results. Technical Report IR-76,University of Massachusetts,1996

4 朱明，王俊普，一种聚类学习的新方法[J]。模式识别与人工智能，2000.9：262～265。

Chinese Text Clustering and Algorithm Designing

Chen Jiong[1], Fan Zhuohua[2], Zhang Hu[2]

(1.Electronic Branch of Technological Institute of Comprehensive Profession of Shanxi ,Taiyuan

Shanxi 030006 ,China ;2.School of Computer & Information Technology Shanxi University,Taiyuan Shanxi 030006 ,China;)

Abstract: Based on the method of selecting the representative points of clustering clusters and the principle that the distance between objects in clusters must be shorter than that between clusters when judging clusters, this paper focuses on problem that the algorithm of traditional clustering is sensitive to the independent datas and inclined to recognize the spheral clustering clusters which are similar in size, and designs an effective clustering algorithm. The result indicates that this algorithm is effective when handling complicated datas.

Keywords: clustering representative point clustering cluster clustering center

第一作者 陈炯 男 34 岁