文章编号：2096-1472(2016)-06-13-04

改进的聚类分析算法在科研立项管理中的应用研究

姜 丹，张晓雯，周 丽
（海军大连舰艇学院基础部，辽宁 大连 116018）

摘 要：针对目前的科研项目管理信息系统仅对科研项目进行低水平管理，无法区分、甄别科研内容等问题，对k-means聚类分析技术进行改进，并进一步将该项技术应用于科研立项管理中，通过对科研立项申报书进行聚类分析，得出立项申请中相似的项目和创新的项目，为科研立项提供智能型的决策支持，避免了重复立项和重复研究，使得计算机应用技术更好地服务于科研项目管理。

关键词：聚类分析；k-means聚类算法；科研项目管理
中图分类号：TP391 **文献标识码**：A

The Research and Application of the Scientific Research Project Management Based on the Improve of Clustering Algorithm

JIANG Dan,ZHANG Xiaowen,ZHOU Li
(*Dept.of Basic Science,Dalian Naval Academy,Dalian* 116018,*China*)

Abstract: For current low-level management of scientific research project information and incapability to distinguish or identify the contents of research projects,the paper improves the k-means clustering analysis technique,and further applies this technique into research project initialization management.Through clustering analysis of research project initialization declaration documents,decision-makers can find out the repetitive studies and the innovative projects.It intelligently supports the decision-making in project initialization by avoiding repetitive projects and studies,and makes it possible for computer application technology to better serve scientific research project management.

Keywords:clustering analysis;k-means clustering algorithm;scientific research project management

1 引言(Introduction)

随着计算机应用技术的飞速发展，计算机信息系统已经渗透到人们生活、工作的各个方面，但是在科研管理中计算机信息系统的应用程度还仅仅停留在对科研项目进行查询、删除、维护等基本操作上。而实际应用中，随着科研项目数目的日益庞大，研究内容的日益繁复，如何对科研项目的内容进行深度分析，以避免在科研中普遍存在的重复立项和低水平重复研究等问题，是对计算机信息系统提出的更高要求。

聚类分析技术是数据挖掘中最常用的工具，可以对大量数据进行聚类，考察数据间的相似度或相异度。若将聚类分析技术应用于科研项目管理的计算机信息系统中，在科研立项环节对立项申请书进行聚类分析，找到众多申请项目中的相似性项目和创新性项目，避免重复立项和重复研究，为科研项目管理系统提供科学的、合理的立项决策支持，使得科研项目管理信息系统更加智能、功能更加强大，是一个亟待研究的课题。

2 聚类分析技术(Clustering)

2.1 聚类分析概述

聚类分析技术是数据挖掘领域最为常见的技术之一，用于发现数据库中未知的对象类，其核心是聚类[1]。所谓聚类即“物以类聚”，首先考察对象之间的相似度或相异度，然后将相似的对象划分在同一个组内，相异的对象划分在不同的组内，保证同一组内的数据对象尽可能的相似，不同组内的数据对象尽可能的相异，最终形成若干个类(或者簇)[2,3]。

聚类分析的定义如下：给定数据集合$V\{v_i|i=1,2,\cdots,n\}$，v_i为数据对象，根据数据对象v_i间的相似度或者相异度，将数据集合$V\{v_i|i=1,2,\cdots,n\}$分成k组$C_j(j=1,2,\cdots,k)$，并满足：

$$\begin{aligned} &C_j \mid j=\{1,2,...,k\} \\ &C_j \subseteq V \\ &C_i \cap C_j = \varnothing \\ &U_{i=1}^{k} C_i = V \end{aligned} \tag{1}$$

该过程称为聚类分析，$C_j(j=1,2,\cdots,k)$称为簇(类)[4,5]。

基金项目：中国人民解放军海军大连舰艇学院2015年学院科研发展基金资助项目(DJYKYKT2015-03).

www.rjgcz.com

2.2 *k*-means聚类分析算法

聚类分析的方法有层次聚类方法、划分聚类方法、基于密度的聚类方法、基于网格的聚类方法等。其中划分聚类中k−means算法具有算法思想简单、收敛速度快、可伸缩性好等优点，应用非常广泛。

k−means聚类算法的基本思想是：以数据对象之间的欧式距离作为相似度或者相异度来考察数据对象，距离越近的数据对象其相似性就越大，距离越远的数据对象其相异度越大，相应的簇是由离得近的数据对象组成。

算法的基本步骤包括：

(1)人为设定簇的个数k值。

(2)随机选取k个对象作为这k个类的初始聚类中心。

(3)计算其他对象到k个初始聚类中心的距离，然后按照就近原则分配对象。

(4)根据公式1重新计算每个类的质心，若给定簇$K_i=\{t_{i1},t_{i2},\cdots,t_{im}\}$，则簇的质心定义为：

$$M_i=\frac{1}{m}\sum_{j=1}^{m}t_{ij} \tag{2}$$

其中，m代表簇K_i中数据对象的个数，t_{ij}代表第j个对象到簇K_i的聚类中心的距离[6]。

(5)重复步骤(3)和步骤(4)，直至簇的质心不再变化或达到终止条件为止。

k−means算法思想简单，可伸缩性好，收敛速度快，适用于处理庞大的样本数据。但从k−means聚类算法存在着比较显著的缺点，其一，算法的第一步需人为设定簇的数目k，很显然k值很难在聚类前估计，对聚类结果影响也比较大；其二，算法随机选取k个初始聚类中心，一旦初始聚类中类中心选择不当，很难得到令人满意的聚类结果。

3 改进的*k*-means聚类分析算法(The improvement of clustering algorithm)

针对上述问题，引进网格和密度两个概念，提出一种改进的聚类分析算法——GBKM算法。

3.1 基本思想

首先对样本空间划分网格单元，划分方法为：设在第i维上数据空间取值范围为(l_i,h_i),$i=1,2,\cdots,n$，采用公式(3)将其划分为p个等长、不相交、左闭右开的区间。

$$\delta_i=(h_i-l_i)/p \tag{3}$$

数据空间被分割成p^n个不相交的、大小相等的网格单元。第i维上的第j个网单元可由公式(4)得出。

$$I_{ij}=\left[l_i+(j-1)\delta_i,l_i+j\delta_i\right],j=1,2,...,p \tag{4}$$

然后计算每个网格单元的密度和密度阀值，根据密度阀值区分高密度网格单元和低密度网格单元，密度阀值Minpt定义为：

$$Minpts=\left[\sum_{i=1}^{N}\mathrm{Denc}(C_i)^2\right]^{1/2}/N \tag{5}$$

其中，Denc(C_i)，$i=1,2,\cdots,n$为网格单元的密度的降序排列，如果Denc(C_i)与Denc(C_i+1)发生明显跳变，则$N=i$。

再后将相邻的高密度网格单元合并形成簇，称为“中间聚类”，将低密度网格单元中的数据对象标记为“自由数据”。

最后处理自由数据，计算每个簇的质心及自由数据到质心的距离，将自由数据分配到最近的簇中，重复此过程，直到聚类中心不再移动为止完成聚类[7]。

3.2 算法流程

算法的基本流程如图1所示。

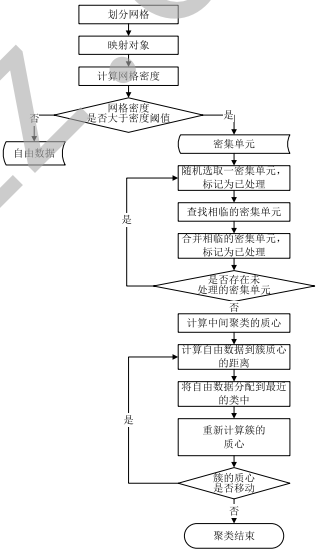

图1 GBKM算法流程

Fig.1 The process of the GBKM clustering

3.3 算法评价

改进的算法形成的初始聚类能够很好地捕获样本数据的原始分布情况，可以自动确定聚类过程所需要的k值及k个初始聚类中心，克服了k−means聚类算法人为确定k值，以及随机选择初始聚类中心这两大缺陷。

簇的纯度p_{ij}定义为簇C_i与第j类交集的大小，簇的纯度越高代表该算法的性能越好。使用Iris数据集进行多次试验，结果如表1所示。显而易见，改进的算法要明显优于传统的k−means聚类分析算法。由于改进的算法首先对高密度区域的数据进行聚类，再对低密度区域的数据进行聚类，聚类过程的迭代次数明显减少，时间上更加高效。表2为五次的实验结果，可以看出改进的算法在聚类所需时间明显小于传统的k−means聚类算法。

www.cjgczz.com

表1 簇的纯度结果比较

Tab.1 The comparison of the purity of cluster

	k−means算法	改进的算法
1	0.864	0.878
2	0.893	0.900
3	0.887	0.900
4	0.893	0.913
5	0.667	0.873

表2 聚类时间的比较

Tab.2 The comparison of the time of clustering

	k−means算法	改进的算法
1	119062.2063	119061.1902
2	119062.2089	119062.1935
3	119062.2138	119061.1835
4	119062.2144	119061.1805
5	119062.2145	119061.1882

综上所述，理论分析及实验证明了改进的算法优于传统的k−means算法。

4 改进聚类分析算法在科研立项管理中的应用(The application of the improve clustering algorithm in the project management)

4.1 基本思想

将改进的聚类分析算法应用与科研立项管理的基本是：首先将科研立项申报书作为输入，然后采用改进的*k*−means聚类算法进行聚类，然后对聚类结果进行分析，将分析结果作为输出，输出包括两方面的内容：(1)创新性项目有哪些；(2)相似的项目有哪些。

4.2 基本流程

由于聚类的对象——科研立项申报书属于中文文本，对中文文本的聚类与传统的聚类有很大的区别，主要在于文本是一种非结构化的数据，不能够直接进行聚类，必须经过一系列的预处理生成文本集合之后才可以聚类，聚类结束后对簇分析时也需要进行一系列的操作，所以对科研立项申报书进行聚类的基本流程如图2所示。

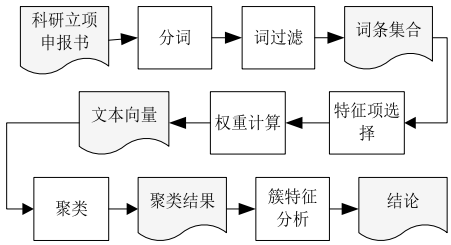

图2 对科研立项书聚类分析的基本流程

Fig.2 The process of the project declaration clustering analysis that

4.3 关键技术

(1)文本预处理

首先是分词。所谓分词就是把句子按照词的含义划分。基于字符串匹配的分词技术是目前最常用的分词技术，也称为机械分词，首先按照一定的策略将待分析的字符串与词库中的词条进行匹配，直到找到一个词匹配成功，否则重新匹配。

然后是词性标记。文本当中有些词性比如助词、叹词等对文本表示不起决定性作用，反思却对聚类结果产生影响，所以应该对分词形成的词进行词性标记，在聚类前将其过滤掉，比较成熟的有中科院开发的ICTCLAS系统。

接着是停用词过滤。首先使用停用词表，然后依次判断文本中的词是否停用，尽量减少文本的词量，以减轻聚类算法的负担，得到更优的聚类结果。

(2)文本的表示模型

通过上面的处理，得到的特征词语仍然数量庞大，很可能导致数据的维数过高，以至于无法聚类，所以需要进一步对文本进行特征选择。主要的特征选择方法有文档频率、单词权、单词熵、信息增益、互信息等，本文采取文档频率方法，该方法基于统计的思想，不但要考虑词在文本中出现的频率，还要考虑这个词在文本集合中出现的频率，文本集中有多少个文本包含了这个词即文档频率越高，这个词的对于所在文本的标识能力就越弱，就越不重要，权重就应该越低。特征项的权值采用经典的TF−IDF方法来计算，特征T_i在文本D_j中的权重值的计算公式如下：

$$TF-IDF(T_i,D_j)=TF(T_i,D_j)*\log\frac{|D|}{|DF(T_i)|} \tag{6}$$

其中，T_i代表某个特征项，D_j表示其所在的文本，$TF(T_i,D_j)$表示特征项T_i在文本D_j中出现的频率，称为词频，$|D|$表示集合中文本的总数目，$|DF(T_i)|$为包含特征项T_i的文本的数目，即文档频率。特征项Ti在文本D_j中出现的频率(词频)越大，则该项对于文本D_j的内容越有代表性，并且在其他文本中出现的频率(文档频率)越大，对文本D_j的内容越没有代表性。

文本作为非结构化的数据必须转化成计算机能够识别的数学模型，文本表示的数学模型有布尔模型、概率检索模型、向量空间模型等多种方法，本文采用向量空间模型，Vector Space Model，简称VSM模型[8]。

$D(T_1,T_2,\cdots,T_n)$是文本D的特征项集合，T_k要求互异又不要求先后顺序，W_k是特征项T_k的权值，表示该特征项在文本D中的重要程度，文本D可以表示成向量$D(T_1,W_1;T_2,W_2;\cdots;T_n,W_n)$，称为文本的特征向量，基本上可以完全代表文本的特性。

www.rjgcz.com

(3)文本的相似度计算

经过上述步骤的处理，生成文本向量集后可以采取聚类算法进行聚类，聚类过程需要衡量文本之间的相似度或者相异度。文本的相似度对于文本聚类有着直接且重要的影响，常用的度量方法有：欧氏距离、内积距离、向量余弦距离，综合考虑。本文采用向量内积方法，内积值越大，相似度就越大。

文本D_1和D_2之间的向量之间的内积公式为：

$$\mathrm{Sim}(D_1, D_2) = \sum_{k=1}^{n} W_{1k} * W_{2k} \tag{7}$$

(4)簇的特征词提取

对于聚类后形成的文本簇，对其进行特征词提取，取得一些能够代表该簇的特征项，用于标识这个文本簇。簇的特征词提取方法为在采用文本向量空间模型的基础上，运用权重计算公式得出文本簇的向量矩阵，再进一步对权重进行排序，输出权重最高的五个特征词来代表该簇。

4.4 系统模型设计

科研立项管理信息系统主要分成文本预处理模块、文本表示模块、文本聚类模块、簇特征分析模块、立项决策输出模块等五大模块，如图3所示。文本预处理模块主要是将科研立项书进行分词、词性过滤、生成词条集合。文本表示模块主要对词条集合进行特征项选择，特征项权值计算，构建文本向量集合。文本聚类模块采用改进的聚类分析算法对前面生成的文本向量集合进行聚类，生成聚类结果。簇特征分析模块主要是对生成的簇进行特征提取及分析，并通过立项决策支持输出模块将创新性项目和相似性项目进行输出。

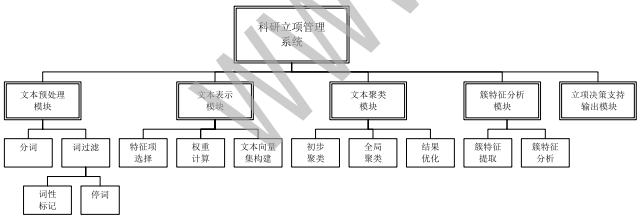

图3 科研立项管理系统的模型设计

Fig.3 The design of the scientific research project management

5 结论(Conclusion)

本文将传统的k-means聚类算法进行改进，并基于改进的聚类算法进一步研究中文文本聚类方法，在此基础上设计了科研立项管理信息系统。该系统使用改进的聚类算法对大量的科研立项申报书进行文本聚类，并对聚类结果进一步分析，从中发现相似的或创新性的项目，为科研立项提供了更加科学的、合理的决策支持，避免了科研中出现重复立项和低水平重复研究等问题。本文的研究成果不仅是对聚类分析技术的改进，也是对计算机应用技术的创新应用，不仅具有较高的学术价值，还具有较高的经济价值和实用价值。

参考文献(References)

[1] Hua Yuan,et al.From Trajectories to Path Network:An Endpoints-Based GPS Trajectory Partition and Clustering Framework(C).The 15th International Conference on Web-Age Information Management (WAIM'2014),Macau,China,June 16-18,2014:740-743.

[2] Huaping Zhang,Ruiqi Zhang,Yanping Zhao.Big Data Modeling and Analysis of Microblog Ecosystem[J].International Journal of Automation and Computing,2014,11(2):119-127.

[3] 罗可,李莲,周博翔.一种蜜蜂交配优化聚类算法[J].电子学报,2014(12):145-149.

[4] Jiawei Han,Micheline Kamber.Data Mining:Concepts and Techniques(Second Edition)[M].San Francisco:Morgan Kaufmann Publisher,2006:383-385.

[5] Anil K.Jain,Richard C.Dubes.Algorithms for Clustering Data[M].New Jersey:Prentice Hall,2006:402-403.

[6] 姜丹,周丽,唐红杰.聚类分析技术在教学指导中的应用研究[J].湖北:软件导刊,2014,13(10):135-138.

[7] Sushmita Mitra,Haider Banka.Collaborative Rough Clustering [J].Lecture Notes in Computer Science,2005,376(1):768-773.

[8] 郑韫旸.基于K-平均算法的文本聚类系统研究与实现[D].武汉理工大学,2009:10-12.

作者简介：

姜　丹(1982-)，女，硕士，讲师.研究领域：数据挖掘.

张晓雯(1978-)，女，硕士，讲师.研究领域：数据库系统.

周　丽(1981-)，女，硕士，实验师.研究领域：物理实验教学.

www.rjgcz.com