一种改进的 K-means 数字资源聚类算法

杨永涛[1],李 静[2]
(1. 燕山大学 信息化处,河北 秦皇岛 066004;
2. 燕山大学 信息科学与工程学院,河北 秦皇岛 066004)

摘 要:K-means 聚类算法在数据挖掘聚类分析方法中是一个基本的、使用最广泛的划分算法。为了对数字图书馆中大量的数字资源进行更加有效、快速的聚类,文中针对传统的 K-means 算法存在的问题,结合数字图书馆数字资源的特征,提出了一种改进的基于关键词特征向量的初始聚类中心选择算法,并在此基础上对传统的 K-means 聚类算法进行了改进,用于对数字资源进行聚类,并进行了算法的实验验证。通过对实验结果的分析证明,文中提出的算法降低了数字资源聚类的代价,提高了聚类的效率,从而验证了算法的可行性。

关键词:K-means 算法;数字资源;相似度;初始聚类中心
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)06-0107-03
doi:10.3969/j.issn.1673-629X.2014.06.027

An Improved K-means Clustering Algorithm for Digital Resources

YANG Yong-tao[1], LI Jing[2]
(1. Information Technology Office of Yashan University, Qinhuangdao 066004, China;
2. College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

Abstract: K-means clustering algorithm is a basic analysis method in data mining closeting analysis, which is also the most widely used partitioning algorithm. In this paper, in order to get more fast and effective clustering result from large number of digital resources in digital library, aiming at the problems of the traditional K-means algorithm, combining with the features of the digital resources, an improved selection algorithm based on the keyword feature vector for initial clustering center is proposed. On this basis, the traditional K-means clustering algorithm is improved for digital resources clustering and experiment verification. The analysis results show that the algorithm proposed reduces the digital resources clustering cost, improves the clustering efficiency, verifying the feasibility of the algorithm.

Key words: K-means clustering algorithm; digital resource; similarity; initial clustering center

0 引 言

K-means[1]算法属于聚类方法中一种典型的划分方法,但初始中心选择不当往往导致聚类效果出现偏差[2]。为了达到对资源项更好的聚类效果,针对经典的 K-means 算法的一些缺点,许多学者在原 K-means 的基础上提出了一些改进方法,主要集中在距离的计算、初始点簇中心的优化选择等方面[3]。

在文献[4]中,Nima Asgharbeygi 和 Arian Maleki 提到了使用测地距离[5-6](Geodesic Distance)代替经典的 K-means 算法中使用的欧几里得距离,改善传统 K-means 算法中的不足,如降低了对空间异常数据的敏感度[7]。

由于传统 K-means 算法中,初始点簇中心的选择是随机的,它是一种局部优化策略,容易陷入局部最优解[8]。好的初始中心的选择,能够极大地避免陷入局部最优,减少聚类结果的误差总和。因此,优化初始中心的选择成为重要的研究点。针对数字图书馆数字资源的种种特征,也就对数字资源的选择和利用提出了新的课题。

1 数字资源的表示

数字图书馆中的每一项数字资源都具备标题和关键词,文中以资源标题和关键词入手来表示每一项资源,其表示方法与文本处理中的文本表示方法非常相似。文本表示主要有三种模型:布尔模型(Boolean Model)、概率模型和向量空间模型(Vector Space Mod-

收稿日期:2013-08-09 **修回日期**:2013-11-16 **网络出版时间**:2014-02-24
基金项目:河北省自然科学基金面上项目(F2013203324)
作者简介:杨永涛(1981-),男,河北秦皇岛人,博士研究生,研究方向为数据挖掘、网络应用技术。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0930.065.html

el,VSM)[9],向量空间模型的方法可以较好地实现文本的抽象,该方法于 20 世纪 60 年代由 Gerard Salton 等人[9]提出。文中采用 VSM 的表示方法来表示数字资源,每一项数字资源表示为由正交特征向量组成的一组向量空间,数字资源中由关键词组成的每一个特征项对应向量空间的一维,不同的特征项对数字资源描述的重要程度不同,由向量空间对应维的数值表示,这样,数字资源 d 均可以表示为如公式(1)所示的规范化的特征向量。

$$d = \{(t_1, w_1), (t_2, w_2), \cdots, (t_n, w_n)\} \quad (1)$$

其中,t_i 代表资源的第 i 个特征项;w_i 代表 t_i 特征项在资源 d 中的权值。

1.1 特征项选择

数字资源的特征项用资源的关键词和资源标题所包含的基本语义单位(字、词、词组或短语等)来表示,关键词可以直接作为特征项,对于资源标题需要通过分词算法进行特征项提取。

目前分词算法主要有两类:一类是理解式分词,另一类是机械式分词。分词算法[10]主要有正向最大匹配法、逆向最大匹配法、最佳匹配法、逐词遍历法、最优路径选择法、最少分词法、特征词库法、邻接约束法、人工神经网络法、无词典分词法等。

同时,在进行特征词提取的时候,需要注意停用词的过滤,停用词是一系列没有实际意义的词[11],它们对于资源的表示没有任何贡献。如英语中的“an,a,of,the”,汉语中的“一种、的、基于”等。

1.2 数字资源的相似度计算

要想合理地对一个数据集中的对象聚类,必须描述数据集中对象之间的亲疏远近程度[12]。同样,为了度量数字资源间的接近或相似程度,必须定义划分类别的计量指标,相似系数是其中最常用的计量指标。

在传统向量空间模型中,每一个特征词组成的向量表示数据集中的一条记录,文中研究的数字资源中的每一项资源的关键词对资源的描述程度是有区别的,因此每个资源的特征项根据其对资源描述的重要程度赋予一定的权值,然后通过向量元素间夹角的余弦来计算它们之间的相似度,如公式(2)所示。

$$\mathrm{sim}(d_i, d_j) = \frac{\sum_{k=1}^{m} (w_{ki} w_{kj})}{\sqrt{\sum_{k=1}^{m} w_{ki}^2 \sum_{k=1}^{m} w_{kj}^2}} \quad (2)$$

式中,w_{ki} 表示第 i 项数字资源的第 k 维属性的权值。

2 K-means 算法

K-means 算法的基本思想是:给定一个包含 n 个数据对象的数据库,以及要生成簇的数目 k,先随机选取 k 个对象作为簇的初始中心,然后计算剩余各个对象到每一个簇中心的距离,把它们划分到距离它最近的簇中心所在的簇,计算各个簇的平均值作为新的簇中心,重复以上过程,直至所有 k 个簇的中心点都不再发生变化。

K-means 算法的过程如下所示[13]:

输入:簇的数目 K 和包含 n 个对象的数据库;

输出:平均误差和最小的 K 个簇。

K-means(k, D, $\{c_1, c_2, \cdots, c_i\}$)

begin

(1)任意选择 K 个对象作为初始聚类中心;

(2)for 数据库中的每一个对象;

(3){将其划分到对应的簇中;}

(4)for 每一个簇

(5){计算各个簇的对象平均值,重新将所有对象划分给最类似的簇;

(6)Repeat;

(7)until 所有簇中心不再发生变化}

end

K-means 方法是解决聚类问题的一种简单、快速的经典算法。它的缺陷在于生成硬性划分的聚类,即每个数据点只唯一地分配给一个聚类。由于事先不知道实际的聚类情况,因此这可能是一种严重的局限[14]。同时,K-means 算法很容易陷入局部极小值从而无法获取全局最优解,在大矢量空间搜索中性能下降。

3 改进的 K-means 算法

为了提高数字图书馆的电子资源的聚类效率,文中提出了一种改进的初始聚类中心选择算法。与经典 K-means 算法随机选择初始中心的方法不同,该方法的思想是:首先用特征向量表示书籍资源的关键词,然后计算出相异度矩阵,根据相异度矩阵找到最不相似的 $m(m>k)$ 个资源,但是与其相似的资源个数大于给定值 θ,然后将所有的资源分配到与其相似的簇中,最后计算每个簇的平均值,作为簇的新的聚类中心,即初始聚类中心。

选择初始聚类中心算法如下所示:

输入:关键词特征向量集 $D = \{d_1, d_2, \cdots, d_n\}$;

输出:m 个聚类中心 $D_c = \{c_1, c_2, \cdots, c_m\}$。

SL(D, D_c)

begin

(1)计算资源的相异度矩阵;

(2)选择相似度最小的两个资源 $t_1, t_2 = \min\{\mathrm{sim}(d_i, d_j)\}$,分别计算与它们相似的资源个数 a_1, a_2;

(3)if a_1, a_2 均大于 θ;

(4){将 t_1, t_2 放入集合 D_c 中;

(5) $D=D-\{t_1,t_2\}$;
(6)}
(7) else
(8){删除 t_1,t_2;
(9) goto(2);}
(10) for $i=1$ to $m-2$
(11){寻找 D 中与 D_c 相似度最小的资源 t,计算与它相似的资源个数 a_t;
(12) if $a_t \geqslant \theta$
(13){ $D_c = D_c + \{t\}$;
(14) $D = D - \{t\}$;}
(15)}
(16) return D_c
end

一旦确定了初始聚类中心,按照经典 K-means 算法重复迭代,计算每次分配完之后的新的簇中心,直至收敛为止,产生 m 个聚类中心,然后合并聚类中心距离最近的簇,直到聚类数减少到 k。

改进的 K-means 算法 SK 如下所示:

输入:簇的数目 K、数据库 I 以及初始聚类中心 O_c ;
输出:k 个聚类数据对象集合 C_k。
SK ($k,I,O_c,\{c_1,c_2,\cdots,c_k\}$)
begin
(1) 调用 SL(D, D_c)选定 m 个初始的聚类中心;
(2) for 数据库中的每一个对象
(3) 计算资源项 i_i 和聚类中心 o_i 的相似度 $\mathrm{sim}(i_i,o_i)$;
(4) if $\mathrm{sim}(i_i,o_k) = \max\{\mathrm{sim}(i_i,o_1),\mathrm{sim}(i_i,o_2),\cdots,\mathrm{sim}(i_i,o_i)\}$;
(5) $c_k = c_k \cup i_i$;}
(6) for 每一个簇
(7){计算每个簇中对象的平均值,将所有对象重新赋给类似的簇;
(8) Repeat;
(9) until $\{c_1,c_2,\cdots,c_m\}$ 不再发生变化;}
(10) 分别合并聚类中心距离最近的簇;
(11) until 聚类数减少到 k,产生 $\{c_1,c_2,\cdots,c_k\}$;
(12) return($\{c_1,c_2,\cdots,c_k\}$)
end

4　算法分析

为了验证改进的 K-means 算法的聚类效果,在数字图书馆的资源中选取了 18 000 本电子图书,提取每本电子书的关键词和标题的有效分词作为特征项进行聚类。

分三次进行计算,每次 6 000 本,包括经济、计算机学、材料学、文学、管理、历史六大类,每类 1 000 项。

通过衡量算法三次执行的平均时间 T 和平均准确率 A 来验证算法的有效性,如表 1 所示。

由表 1 可以看出,在此次实验测试中,改进后的 K-means 算法的准确率比传统的 K-means 算法提高了 7 到 15 个百分点;同时消耗的平均时间均有所减少,可以说明改进的 K-means 算法不仅降低了聚类所需的执行时间,同时提高了资源聚类的准确率,进而说明了改进的 K-means 算法能够快速准确地对数字图书馆的数字资源进行聚类。

表 1　文中提出算法与传统 K-means 算法的比较

资源类别	K-means 算法		文中提出算法	
	T/ms	A/%	T/ms	A/%
经济类	61 275	73	58 600	80
计算机学类	73 809	63	60 947	78
材料学类	60 378	71	58 203	81
文学类	58 830	75	58 023	79
管理类	60 123	69	59 524	77
历史类	60 850	72	60 590	83

5　结束语

文中通过对传统 K-means 聚类算法和数字图书馆数字资源特征的研究,分析了传统 K-means 聚类算法的局限性,设计了一种基于关键词特征向量的初始聚类中心选择算法,并以此为基础对传统的 K-means 聚类算法进行了改进,用于对数字图书馆中大量的数字资源进行快速、有效地聚类,最后,通过实验分析验证了算法的有效性。

参考文献:

[1] MacQueen J. Some methods for classification and analysis of multi-variate observations[C]//Proc of the 5th Berkeley symposium on mathematical statistics and probability. Berkeley, USA: Univ of Calif Press, 1967: 281-297.
[2] 周爱武,于亚飞. K-Means 聚类算法的研究[J]. 计算机技术与发展,2011,21(2):62-65.
[3] 张　靖,段　富. 优化初始聚类中心的改进 K-means 算法[J]. 计算机工程与设计,2013,34(5):1691-1694.
[4] Asgharbeygi N, Maleki A. Geodesic K-means clustering[C]//Proc of 19th international conference on pattern recognition. Tampa, FL: IEEE, 2008: 1-4.
[5] Lanthier M, Maheshwari A, Sack J R. Approximating weighted shortest paths on polyhedral surfaces[C]//Proc of symposium on computational geometry. [s. l.]: [s. n.], 1999: 274-283.
[6] Mitchell J S B, Mount D M, Papadimitriou C H. The discrete geodeic problem[J]. SIAM Journal on Computing, 1987, 16(4): 647-668.
[7] 黄　韬,刘胜辉,谭艳娜. 基于 k-means 聚类算法的研究[J]. 计算机技术与发展,2011,21(7):54-57.
[8] 殷君伟,陈建明,薛百里,等. 一种基于排序划分的聚类初

(下转第 113 页)

练数据的分类性能。

图 4 是环境音数据对应各方法分类性能的比较。总体来看,EPS 和 SDS 都比单一分类器性能好,而且曲线比较平滑和稳定(除了 SDS 在标记样本比例增大到 60% ~80% 外)。在标记数据占总训练样本数量 10% ~50% 时,主动学习能得到比较理想和稳定的分类状态,性能的提高比较明显。因为在学习的过程中,未标记数据可能会被误标记,以致在扩大的训练标记样本中引入了噪音数据,在下一次学习训练迭代时,包含了错误标记的训练数据,引入了训练误差。尤其 SDS 算法,虽然训练的时间减少了,但是当已标记样本比例增加时,性能反而有所降低。

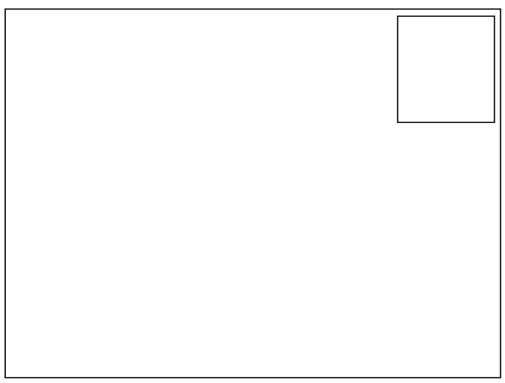

图 4　数据集各方法的性能

4　结束语

在训练标记数据较少,存在大量未标记数据的情况下,主动学习的分类效果明显优于传统的监督分类,因而提供了一条解决分类性能的有效途径。文中采用了 EPS 和 SDS 主动学习对实验中环境音数据进行分类取得了较好的预测效果。

如何提高选择样本的多样性,研究采样策略仍然是主动学习的研究热点。一方面将集成学习引入到主动学习中以保证选择样本的多样性;另一方面,采样策略算法应结合实际研究数据对象环境音的音频结构和提取的音频特征,更有效地选择具有高信息量的样本,减少计算的复杂度,提高分类的准确度和分类器的泛化能力。

参考文献:

[1] 白　亮,老松杨,陈剑赟,等. 基于支持向量机的音频分类与分割[J]. 计算机科学,2005,32(4):87-90.

[2] 张小梅,杨鼎才. 基于支持向量机模型的环境音分类研究[J]. 电子测量技术,2008,31(9):121-123.

[3] 余清清,李　应,李　勇. 基于 SVM 模型的自然环境声音的分类[J]. 计算机与数字工程,2010,38(7):1-5.

[4] 郑怡文. 典型的音频分类算法[J]. 计算机与现代化,2007(8):59-63.

[5] Zhou Zhihua, Wang Yu. Machine learning and application[M]. Beijing:Tsinghua University Press,2007.

[6] Seeger M. Learning with labeled and unlabeled data[R]. Edinburgh,UK:University of Edinburgh,2001.

[7] Zhu X. Semi-supervised learning literature survey[R]. Wisconsin:University of Wisconsin at Madison,WI,2008.

[8] 龙　军,殷建平,祝　恩,等. 主动学习研究综述[J]. 计算机研究与发展,2008,45(Sup):300-304.

[9] 刘　康,钱　旭,王自强. 主动学习算法综述[J]. 计算机工程与应用,2012,48(34):1-4.

[10] 吴伟宁,刘　扬,郭茂祖,等. 基于采样策略的主动学习算法研究进展[J]. 计算机研究与发展,2012,49(6):1162-1173.

[11] Seuong H S, Opper M, Sompolinski H. Query by committee[C]//Proceedings of the 5th annual workshop on computational learning theory. New York,NY,USA:ACM,1992:287-294.

[12] Campbell C, Cristianini N, Smola A J. Query learning with large margin classifiers[C]//Proc of ICML. San Francisco, CA,USA:Morgan Kaufmann Publishers Inc,2000:111-118.

[13] Roy N, McCallum A. Toward optimal active learning through sampling estimation of error reduction[C]//Proc of ICML. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2001:441-448.

[14] Witten I H, Frank E, Hall M A. Data mining: practical machine learning tools and techniques[M]. 3rd ed. San Francisco,CA,USA:Morgan Kaufmann Publishers Inc,2011.

(上接第 109 页)

始化方法[J]. 微电子学与计算机,2013,30(6):80-83.

[9] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of ACM,1975,18(11):613-620.

[10] 林春实,方　燕,全吉成. 汉语文献自动分词与标引技术发展浅析[J]. 情报学报,1997(S1):45-49.

[11] 化柏林. 知识抽取中的停用词处理技术[J]. 现代图书情报技术,2007(8):48-51.

[12] 李　健. 聚类分析及其在文本挖掘中的应用[D]. 西安:西安电子科技大学,2005.

[13] 李永森,杨善林,马溪骏,等. 空间聚类算法中的 K 值优化问题研究[J]. 系统仿真学报,2006,18(3):573-576.

[14] Shekhar S, Chawla S. 空间数据库[M]. 谢昆青,译. 北京:机械工业出版社,2004.

一种改进的K-means数字资源聚类算法

作者： 杨永涛， 李静， YANG Yong-tao， LI Jing
作者单位： 杨永涛,YANG Yong-tao(燕山大学 信息化处,河北 秦皇岛,066004)， 李静,LI Jing(燕山大学 信息科学与工程学院,河北 秦皇岛,066004)
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(6)