一种改进的 k–means 中文文本聚类算法

龚　静，李安民

（湖南环境生物职业技术学院 信息技术系，湖南 衡阳 421005）

摘　要：提出了 k-means 聚类算法中选取初始聚类中心及处理孤立点的新方法，改进了 k-means 算法对初始聚类中心和孤立点文本很敏感的不足之处，并将改进后的算法应用于中文文本聚类中。实验结果表明，改进的算法较原算法在准确率上有较大提高，并且具有更好的稳定性。

关键词：k-means 算法；文本聚类；中文文本；层次聚类

中图分类号：TP301　　**文献标识码**：A　　**文章编号**：1673–9833(2008)02–0052–03

Clustering Algorithm of One Improved *k*-Means Chinese Text

Gong Jing，Li Anming

（Department of Information Technology, Hunan Environment — Biological Polytechnic，Hengyang Hunan 421005，China）

Abstract This paper proposes a new way that selects initial cluster center and processes isolated points in the *k*-means clustering algorithm. And this method improves the deficiency that the *k*-means algorithm is very sensitive to the initial cluster center and the isolated point text. It applies the improved algorithm in Chinese text clustering. The experimental result indicates the improved algorithm has a higher accuracy compared with the original algorithm, and has a better stability.

Key words: *k*-means algorithm; text clustering; Chinese text; level clustering

0　引言

文本聚类基于“聚类假设”，相关文本之间的相似性比无关文本之间的相似性更大。文本聚类是一种无指导的文本分类，它把一个文本集分成若干称为簇（clustering）的子集，每个簇中的文本之间具有较大的相似性，而簇之间的文本具有较小的相似性。文本聚类可广泛应用于文本挖掘与信息检索的不同方面，在大规模文本集的组织与浏览、文本集层次归类的自动生成等方面都具有重要的应用价[1]。文本聚类中的文本表示模型通常采用向量空间模型[2]，文本聚类普遍采用的算法是基于划分的 k-means 算法。

k-means 算法是最早最典型的划分聚类算法，已经被广泛地应用于中文文本聚类。k-means 的聚类过程很简单，简述如下：首先随机地选择 k 个文本，每个文本初始地代表一个簇的平均值或者说中心。对剩余的每个文本，根据其与各个簇中心的距离，将它赋给离它最近的簇，然后重新计算每一个簇的中心，这个过程不断重复，直到准则函数收敛，常采用误差平方和准则函数作为聚类准则函数，误差平方和准则函数定义为：

$$J_e=\sum_{i=1}^{k}\sum_{p\in C_j}\left|p-m_i\right|^2,$$

其中 J_e 是所有文本对象平方误差的总和，p 是一个文本对象，m_j 是簇 C_j 的中心。此准则试图使生成的各个簇尽可能紧密，同时簇与簇之间尽可能分开。

从上面的算法过程中，不难看出 k 个初始聚类中心点的选取对聚类结果具有较大的影响，因为在该算法中是随机地选取任意 k 个文本作为初始聚类中心。存在所选的文本本应属于同一簇的可能，而依据 k-means 算法的思想，这些文本却被硬性地划分到不同的簇中去，必然会陷入局部极小，得到的解是局部最优解，而不是全局最优解。

针对这个问题，文献[3]取随机选取不同的初始值多次执行该算法，然后选取最好的结果作为初始聚类中心；文献[4]采用了全局优化方法中的模拟退火技术

收稿日期：2007–11–05

基金项目：湖南省教育厅基金资助项目 (07D036)

作者简介：龚　静 (1972–)，女，湖南岳阳人，湖南环境生物职业技术学院副教授，硕士，主要研究方向为自然语言处理.

以摆脱局部最优。

文章提出了采取m（m>1）次取样，对于每个样本分别用k-means算法进行聚类，得到$m\times k$个聚类中心，然后用凝聚的层次聚类算法Single−Link对$m\times k$个聚类中心进行聚类，产生k个聚类中心点作为k-means算法最终的初始聚类中心。另外，由于k-means算法对孤立点文本很敏感，造成聚类结果不稳定，因此，在用k-means算法进行聚类时，在进行第t轮聚类中心的计算时，采用簇中那些与第t-1轮聚类中心相似度较大的文本，计算它们的均值点（几何中心点）作为第t轮聚类的中心。将改进后的k-means算法应用于中文文本聚类，实验结果表明，改进算法较原算法在准确率上有较大提高，并具有较好的稳定性。

1　新的初始聚类中心选择方法

新的初始聚类中心选择方法的基本思想是假设已经知道文本集的分布情况，文章认为一个优良的初始聚类中心应该满足：

1)选择的初始中心各属于不同的簇，即任意两个初始中心不能属于同一簇；

2)选择的初始聚类中心应能够作为该簇代表，即应该尽量靠近簇中心。

要选出k个文本作为初始聚类中心，并且同时保证这k个文本恰好分别属于不同的簇，这种严格的约束很难通过随机抽样方式实现。于是想到：为了尽可能减小取样对初始聚类中心选取所产生的影响，采取m次取样，样本大小为n/m，其中，n为文本集中文本的个数，m的取值为每次抽取的样本大小应该能装入主存，并尽可能满足m次提取的样本之和等于原始文本集。对于每次提取的样本文本分别采用k-means算法进行聚类，分别产生一组具有k个聚类中心的文本簇；对于m次取样操作共生成$m\times k$个聚类中心，再用凝聚的层次聚类算法Single-link算法进行聚类，得到k个簇，取k个簇的均值为最终的k个初始聚类中心。

与k-means算法所采取的划分策略不同，在凝聚的层次聚类算法中不存在初始聚类中心的选择问题。最初它把每个文本均视为一个簇，文本就是该簇的中心，聚类的每一步，将最相似的两个簇合并为一个簇，直到所有的文本归为一个簇，或者只有k个簇为止。随着聚类的进行，相似的文本逐渐聚集成一簇，层次聚类能够自动生成不同层次上的聚类模型[5]。

结合凝聚的层次聚类算法和k-means算法的思想，提出了一个基于k-means的层次聚类算法来选取初始聚类中心，即使用k-means方法所产生的聚类中心来约束凝聚的层次聚类算法的凝聚空间。

选取初始聚类中心方法的总体描述如下：

1) 对文本集进行m次取样，并且划分文本集合为m个样本集$\{S_1,S_2,\cdots,S_m\}$；

2) 对每个样本集分别执行k-means算法，产生m组k个聚类中心；

3) 用凝聚的层次聚类算法（在此用Single−link算法）对$m\times k$个聚类中心进行再次聚类，直到只有k个簇为止，取每个簇的均值作为下一步k-means算法的初始聚类中心。

从上面的算法可知，提取的样本文本集比原始文本集小，因此，搜索初始聚类中心的过程量较少，迭代次数少，速度较快；同时也保证了最终聚类中心均属于不同的簇，具有足够的代表性。

2　改进的k−means算法

在用k-means算法进行文本聚类实验时，发现除了初始聚类中心的选择对聚类产生较大的影响外，还在实验结果的稳定性方面存在问题，即实验结果偶尔出现较大的偏差。通过分析表明，导致这种偏差产生的原因在于文本的分散性。由于存在少量文本远离高密度的文本密集区，但在进行k-means聚类计算时，是将聚类均值点作为新的聚类中心进行新一轮聚类计算，此时新的聚类中心将偏离真正的文本密集区。因此，为了使选择的初始聚类中心能够作为该簇代表，应该尽量靠近簇中心，在用k-means算法进行聚类时，在进行第t轮聚类中心的计算时，采用簇中那些与第$t-1$轮聚类中心相似度较大的文本，计算它们的均值点（几何中心点）作为第t轮聚类的中心。因此，综合新的初始中心选择方法，提出了改进的k-means算法。

改进的k-means算法描述如下：

1) 首先由基于k-means的层次聚类算法来选取初始k个聚类中心，其向量为$\{C_1,C_2,\cdots,C_k\}$；

2) 对文本集D中的每个文本d_i，用k-means算法依次计算文本d_i与各个簇C_i的相似度；

3) 如果文本d_i与某个簇C_i的相似度最大，则将d_i归入以C_i为簇中心的簇C_i，从而得到D的一个聚类$Clusters=\{C_1,C_2,\cdots,C_k\}$，同时将每个文本与所归入簇的相似度保存下来；

4) 对于$t-1$轮聚类所获得的簇C_i，找出簇中文本与该簇中心C_i的相似度最小的相似度，记为$MinOfSim$；

5) 选择簇C_i中与聚类中心C_i相似度大于$1-\beta*(1-MinOfSim)$的文本，其中β为0到1之间的常数，记该文本集合为C_i'；

6) 计算C_i'中文本的均值点，作为第t轮聚类的聚类中心；

7) 重复步骤2)、3)、4)、5)和6)若干次，直到聚类中心不再发生变化。

改进的k-means聚类算法的时间复杂度分析：

1) 在改进的k-means算法中，对提取的样本文本搜索

初始聚类中心的过程文本数较少，迭代次数很小，速度很快。对于文本集合非常大，提取样本以及搜索初始聚类中心的过程所耗费的时间在整个算法中可以忽略不计。

2) 在改进的 k-means 算法中，层次聚类算法的时间复杂度为 $0((mk)^2)$，相对于文本集来说，mk 的值很小，所耗费的时间在整个算法中可以忽略不计。

3) 对获得的初始聚类中心进行 k-means 聚类所耗费的时间为 $0(nkt)$，n 为文本集的大小，k 为聚类数目，t 为迭代次数。

因此，改进的 k-means 算法所需总的时间为 $0(nkt)$，与原 k-means 算法相比，耗费时间增加不多。

3 实验结果

为了检验改进算法的有效性，对原始算法和改进算法进行了对比实验。实验采用 VC++ 实现，在 Celeron(R) 2.0G，512M 内存的计算机上进行，实验用的数据取自人民网(www.people.com.cn)和新华网(www.xinhuanet.com)。

实验结果如表 1 所示。表 1 中各符号定义如下：M_i 为类别 i 的文本总数；N_j 为聚类 j 的文本总数；$M(n_{ij})$ 为类别 i 达到最大 F-measure 值时聚类 j 中包含类别 i 的文本总数；$M(F(i,j))$ 为类别 i 和不同聚类 j 的 F-measure 值中最大的值。

表 1 一组聚类对比实验结果

Tab. 1 A set of comparison clustering results

类 别	普通的 k-means 算法				改进后的 k-means 算法		
	M_i	N_j	$M(n_{ij})$	$M(F(i,j))$	N_j	$M(n_{ij})$	$M(F(i,j))$
政 治	63	60	38	0.618	61	54	0.871
艺 术	85	79	63	0.768	82	71	0.850
计 算 机	34	36	27	0.773	35	29	0.842
环 境	45	46	33	0.706	45	39	0.867
教 育	78	80	52	0.658	80	62	0.765
军 事	54	56	42	0.764	55	48	0.881
体 育	39	42	22	0.543	41	30	0.745
医 药	82	81	60	0.745	81	71	0.871
平均值			F=0.697			F=0.837	

表 1 数据表明，采用改进后的聚类算法使得聚类结果的准确性有了一些提高。为了验证改进后聚类算法结果的稳定性，采用了多组数据分别利用两种算法进行对比实验，获得 30 组实验数据，实验结果中 F-measure 值的分布情况如表 2 所示。

表 2 聚类对比实验统计结果

Tab. 2 Comparison experimental clustering results

F-measure 区间	F-measure 典型值	原 k-means 算法 F-measure 值落入此区间的实验次数	改进的 k-means 算法 F-measure 值落入此区间的实验次数
[0.45,0.55]	0.50	3	0
[0.55,0.65]	0.60	7	0
[0.65,0.75]	0.70	11	7
[0.75,0.85]	0.80	9	16
[0.85,0.95]	0.90	0	7
[0.95,1.00]	1.0	0	0

从表 2 可以看出，采用普通 k-means 算法得到的聚类结果稳定性不好，F-measure 值比较分散；而采用改进后的聚类算法得到的聚类结果稳定性较好，F-measure 值比较集中，F-measure 的平均值更高。

实验表明，采用改进后的聚类算法准确性和稳定性都有较大的提高。采用普通 k-means 算法，聚类结果的 F 值分散在 0.60 ~ 0.75 之间；而采用改进的算法，其值稳定在 0.75 ~ 0.85 之间。

4 结语

文章针对 k-means 算法在中文文本聚类过程中对初始点的选择及孤立点敏感的问题提出了改进的算法，实验表明了该改进算法的有效性。关于文本聚类，将进行新的中文文档向量模型改进聚类质量的研究，以便真正将中文文本聚类推向实用化。

参考文献：

[1] Steinbach M，Karypis G，Kumar V. A Comparison of Document Clustering Techniques[R]. Technical Report，Dept. of Computer and Information Science，Linkoping，1995：143-150.

[2] Fasulo D. An Analysis of Recent Work in Clustering Algorithms[R]. Technical Report UW-CSE-01-03-02，University of Washington，1999：176-186.

[3] Duda Ro，Hart PE. Pattern Classification and Scene Analysis [M]. New York：John Wiley and Sons，1973：143-146.

[4] Selim SZ，Alsultan K. A Simulated Annealing Algorithm for the Clustering Problem[J]. Pattern Recognition，191，24(10)：1003-1008.

[5] 朱红灿，孟志青. 一种基于 SOM 和层次凝聚的中文文本聚类方法[J]. 湘潭大学学报：自然科学版，2005，27(3)：36-38.

（责任编辑：罗立宇）