改进的 K-means 算法在网络舆情分析中的应用①

汤寒青[1,2]，王汉军[2]

[1](中国科学院 研究生院，北京 100049)
[2](中国科学院 沈阳计算技术研究所，沈阳 110171)

摘 要：结合网络舆情分析的应用需求背景，首先介绍了文本信息的处理，然后探讨了文本聚类中的 K-means 算法，针对其对初始聚类中心的依赖性的特点，对算法加以改进。基于文档标题能够代表文档内容的思想，改进算法采用稀疏特征向量表示文本标题，计算标题间的稀疏相似度，确定初始聚类中心。最后实验证明改进的 K-means 算法提高了聚类的准确度；与基于最大最小距离原则的初始中心选择算法比较，提高了执行效率，同时保证了聚类准确度。

关键词：网络舆情；K-means 算法；文本聚类；稀疏特征向量

Application of Improved K-Means Algorithm to Analysis of Online Public Opinions

TANG Han-Qing[1,2], WANG Han-Jun[2]

[1](Graduate University, Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110171, China)

Abstract: Combining background application requirement of online public opinion analysis, this paper firstly introduces the processing of text information, and then discusses the K-means algorithm of the text clustering, according to its characteristic that clustering results depend on the centers of initial clustering, and improves it. Based on the thought that text title can express its content, the improved algorithm uses sparse character vector to express text title, calculates the sparse similarity of them and ascertains the centers of initial clustering. The experiments show that the method improves the clustering accuracy. Compared with another algorithm based on the principle of maximum and minimum distance, the improved method heightens the efficiency and ensures the clustering accuracy.

Keywords: online public opinion; K-means clustering algorithm; text clustering; sparse character vector

1 引言

在当今网络信息技术快速发展的时代，网络舆情分析受到众多学者的广泛关注。网络舆情分析首先把收集到的网络信息转换为文本格式，然后对文本信息处理，包括文本分词，文本表示，文本聚类并发现热点问题，最后对热点问题进行趋势分析，并做相应的处理。

目前，人们已经提出了多种文本聚类算法[1,2]，其中基于划分的聚类方法和层次聚类方法应用最为广泛。K-means 算法[3]是由 Macqueen 提出的解决聚类分析问题的经典算法，Dhillon 将其应用到文本聚类领域[4]，并利用余弦相似度计算对象间的距离，该算法简单且收敛速度快，但也有着明显的缺点，因此本文针对 K-means 算法对初始聚类中心的依赖性，基于文本标题能够表达文本中心内容的思想，对算法加以改进。改进的算法具体采用稀疏特征向量表示文本标题，通过计算标题间的稀疏相似度，给出一种新的初始中心的选择算法，并对改进后 K-means 算法聚类效果进行了验证，实验结果表明文中改进的算法能够快速准确的发现的问题。

① 收稿时间:2010-07-07;收到修改稿时间:2010-08-04

2 文本信息处理

2.1 文本表示

文本表示主要有三种模型：布尔模型(Boolean Model)，概率模型和 VSM(Vector Space model)向量空间模型[5]。本文采用 VSM 表示文本内容，在向量空间模型中，每一个文本用 n 维的空间的一个点来表示，点由一组规范化的正交矢量组成。向量的维数由文本包含的特征词组成，而其中每个分量的值由该词在文本中的权重来表示。权重用来说明词对所在文本的重要程度。文本的向量空间表示可描述为：在有 n 个不同特征项的一组 d_1，d_2，…，d_n 的文本系统中，给定文本的传统特征向量表示：d_i=($\omega_1(d_i)$，$\omega_2(d_i)$，…，$\omega_j(d_i)$，…，$\omega_n(d_i)$)，由于 d_1，d_2，…，d_n 互不相同，可以把它们看作是 n 维欧氏空间 n 个坐标，把 d_i 看作是 n 维欧氏空间的向量。其中 $\omega_j(d_i)$ 表示第 j 个特征词在文档 d_i 中的权重。

特征词的权重有多种，例如这里我们采用的 TF-IDF[6]，TF-IDF 是一种应用广泛的特征项度量方法，这种方法给文本中每一个特征词一个权重，计算公式如下：

$$\omega_j(d_i)=\frac{tf_j(d_i)\times\log_2^{(n/n_j(d_i)+0.01)}}{\sqrt{\sum_{i=1}^{n}(tf_j(d_i)\times\log_2^{(n/n_j(d_i)+0.01)})^2}} \quad (1)$$

$\omega_j(d_i)$ 表示第 j 个特征词在文本 d_i 中的权重；$f_j(d_i)$ 表示第 j 个特征词在文本 d_i 中出现的频率，$n_j(d_i)$ 表示包含第 j 个特征词的文本个数，n 表示所有文本个数。

为了快速准确的确定 K-means 算法中初始聚类中心，对文本标题采用稀疏特征表示[7]。然后对数据压缩处理。 文本的稀疏特征表示方法：假设系统中 n 篇文档的标题共包含 m 个特征词，文档标题的形式化定义为如下格式：

$$t_i=(w_{1i},w_{2i},\cdots,w_{ji},\cdots,w_{mi})$$

$$w_{ji}=\begin{cases}1 & F_j(t_i)>0\\ 0 & F_j(t_i)=0\end{cases} i=1,2,...n;\ j=1,2,...m \quad (2)$$

其中 $F_j(t_i)$ 表示第 j 个特征词在标题 t_i 出现的频率。

2.2 文本特征向量降维

计算出每个特征词在文本中的权值以后，把 $\omega_j(d_i)$ 大于一定值的单词作为代表文本内容的特征词，其他权重小的单词不予考虑，以达到降低维数的目的，把所有文档中的特征词作为这组 n 篇文本的特征词。然后用所有的特征词表示每一篇文本，进行归一化处理。

稀疏特征向量降低维数原理：由经验可知，标题中出现的词语是文档内容的关键词汇，所以在这里无论特征词出现多少次，只要它出现就意味着对该文档非常关键。根据这个原理把在标题中没有出现的词语直接去掉降维。降低维数之后的结果为：

$$t_i=(w_1,w_2,...,w_j)$$

其中 $j\le m$，w_j 表示 t_i 中的第 j 个特征词，当 $w_{ji}>0$ 时有 $w_j=j$，为零时直接去掉降低维数。

2.3 文本相似度计算

本文采用余弦距离度量[8]表示文本之间的相似性，它定义两篇文档 d_i,d_j 的相似度如下：

$$simd(d_i,d_j)=\frac{\sum_{k=1}^{n}w_k(d_i)\times w_k(d_j)}{\sqrt{\sum_{k=1}^{n}w_k(d_i)^2}\times\sqrt{\sum_{k=1}^{n}w_k(d_j)^2}} \quad (3)$$

稀疏特征向量表示下的两标题的相似度，简称稀疏相似度，定义如下：$simt(t_i,t_j)=N_{t_i=t_j}$ (4)

$N_{t_i=t_j}$ 表示一组中的 t_i,t_j 两标题的相同特征项的个数。假设我们得知 t_i 中第 1，2，3 项出现，而 t_j 中第 1，3，4，6 项出现，用稀疏向量分别表示为：t_i={1，1，1，0，0，0}；t_j={1，0，1，1，0，1}。降低维数之后的标题为：t_i={1，2，3}，t_j={1，3，4，6}，可以看出两个标题当中只有第 1，3 特征项是相同的，标题中所有的特征项个数为 6 个。那么根据公式(4)得出：两标题的相似度为 $simt(t_i,t_j)$=2。

3 传统K-means算法

K-means 算法思想简单，就是把给定的 N 个文本分配到 k 类中，其中 k 是已知的。首先从输入文档中随机选择 k 个文本作为簇的初始聚类中心，然后根据簇内文本之间相似度大而不同簇间文本相似度小的原则，把剩余的文档指派到相应的类簇；一个聚类中心的所有样本的集合构成一个簇，把簇中所有文档的空间向量的平均值作为新的聚类中心，然后根据新的聚类中心并重新指派所有的样本点，直到聚类中心不再

改变为止。

K-means 算法可以用来处理大量的数据，而且效率高，复杂度低，所以可以用来处理大量的由 VSM 向量空间模型表示的文本聚类问题。但是很遗憾由于随机选择初始聚类中心使得聚类效果存在偶然性

K-means 算法的聚类结果对于初始聚类中心有很大的依赖性，若选到了孤立点，聚类中心将发生偏移，聚类结果不可理解。

4 改进的k-means文本聚类算法

4.1 基于标题稀疏特征向量的初始聚类中心选择算法

在文本聚类算法中，选择初始点一般有经验选择、随机选择、最大最小原则[9]等方法。经验选择带有主观性，同时给用户增添了负担。随机选择可能选取“孤立点”、类边缘点，或者一个类中选取了两个以上的对象作为初始聚类中心，结果不理想。而最大最小原则依据待聚类对象的相似情况选择距离较远的对象作为聚类中心，第一个聚类中心仍为随机选取。针对此种情况，本文结合文本标题的稀疏特征表示方法提出一种简单而又高效的选择初始中心的聚类算法。

新的初始中心的选择方法基于以下假设：1)如果特征词在标题中出现，那么特征词是文本内容的关键词。2)如果两个文本标题相似属于同一类簇，那么这两个文本也属于同一类簇。

为了提高 k-means 算法聚类效果，结合文本标题的稀疏特征表示方法和 k-means 聚类算法，本文提出一种基于标题的稀疏特征向量初始聚类中心选择算法。与 k-means 算法随机选择初始中心方法不同，我们根据下列算法计算得出的。算法思想如下：首先用稀疏特征向量表示文本标题，计算任意两个文本标题间的稀疏相似度，找到最不相似的 k 个标题，但是与其稀疏相似度大于 λ 定值的文档个数大于 θ，然后把所有的标题分配到与其最相似的簇中，最后计算标题所对应的文本向量的平均值，使其作为簇的初始聚类中心。

选择初始中心算法步骤如下：

Step1: 对 N 篇文档标题的进行稀疏特征表示。T＝{t_1，t_2，…，t_n}，对文本向量空间表示 D={d_1，d_2，…，d_n}。

Step2: 计算任意两个标题间稀疏相似度 $simt(t_i,t_j)$，TS={s_{12}，s_{23}，…，s_{ij}}。

Step3: 从 TS 集合中选择 min(s_{ij})，分别计算与 t_i，t_j 相似度大于给定阈值 λ 的文档标题的个数，记为 α_i，α_j。

Step4: 当且仅当 α_i，α_j 都大于某一给定值 θ 时，把 t_i，t_j，放入集合 D_c 中；T=T－{t_i，t_j}；否则转步骤 step3。

Step5: 寻找集合 T 中标题与集合 D_c 中标题相似度最小的标题，min($simt(t_c,t_t)$)，其中 $t_c \in D_c$，$t_t \in$ T。即寻找与 D_c 中标题最不相似的标题，然后我们计算与 t_t 相似度大于给定阈值 λ 的标题个数 α_t。

Step6: 如果 $\alpha_t \geq \theta$ 时，$D_c = D_c$ +{t_t}，T=T－{t_t}，删除 TS 中 s_{ct}，k=k+1；否则删除 T 中 t_t，删除 TS 中与 t_t 有关的相似度,转步骤 step5。

Step7: 重复步骤 step5，step6，直到 k 为输入的数值为止。

Step8: 从 TS 中找到剩余标题与 D_c 中 k 个标题最大的稀疏相似度 max{$simt(t_i, t_c)$}，并把它分配到相应的簇中。

Step9: 分别计算 k 个簇中标题对应的文本向量空间的平均值，使其作为 k 个初始的聚类中心。D_c ={d_1，d_2，…，d_k}。

算法利用文本标题的稀疏相似度确定簇的初始聚类中心，大大减少了计算的数据量，缩短了算法的执行时间，避免选到孤立点、边缘点。

4.2 改进的 k-means 文本聚类算法

算法思想：采用改进的初始中心选择算法得到初始的聚类中心；按照传统的 k-means 算法重复迭代，计算每次分配完之后的新的簇中心，直到准则函数收敛为止。

对于给定的文本标题集 T＝{t_1，t_2，…，t_n}和文本集合 D={d_1，d_2，…，d_n}。

改进的 k-means 算法的步骤如下：

Step1: 按照改进的基于稀疏特征向量的初始聚类中心算法确定初始的聚类中心集合 D_c ={$C_1^{(0)}$，$C_2^{(0)}$，…，$C_k^{(0)}$}，其中 $C_j^{(0)}$ 表示第 j 类初始聚类中心。

Step2: 对于文本集合中的数据 D ={d_1，d_2，…，d_n}中任一个数据 d_i，与集合 D_c 中 k 个中心点 $C_j^{(p)}$ 进行文本相似度计算，找到与其相似度最大的数据 Max ($simd(d_i, C_j^{(p)})$)，然后把 d_i 分配到

以 $C_j^{(p)}$ 为中心的簇中。以此类推把 D 中剩余的文档分配到相应的簇中。p 表示重复迭代次数。

Step3: 计算由步骤 step2 第 p 次迭代之后上的簇中心，D_c ={ $C_1^{(p+1)}$，$C_2^{(p+1)}$，…，$C_k^{(p+1)}$ }。

Step4: 重复步骤 step2，step3，直到准则函数 E 收敛为止。

Step5: 输出 k 个簇中心，输出由 N 个文本组成的 k 个簇。

改进后的算法簇的初始中心选择快速，并且执行效率很高；初始聚类中心比较有代表性，所以聚类结果更理想。

5 实验设计

实验步骤：

Step1: 把每一类文本都封装成一个文件夹。每一篇收集到文章转化为.txt 文本文档类型。

Step2: 采用中科院计算所分词词典对文本进行文本分词。

Step3: 用稀疏特征向量表示文本标题，对标题向量降维。

Step4: 计算所有文本出现的单词频率，计算每一单词在文本中的权重，对文本向量进行降维。

Step5: 聚类：准则函数：$E=\sum_{i=1}^{k}\sum_{d\in c_i}|d-c_i|^2$，其中 d 是 c_i 类中的文档，c_i 代表第 i 类的中心,即第 i 类中所有文本的平均值。由经验设定阈值 $\theta=\frac{N}{k}$，$\lambda=2$，考虑到标题中包含的单词较少，标题相似度阈值设置为 2，只要两标题中有两个以上的相同特征词，说明他们同属于一类。

6 数据分析

为了验证改进的 K-means 算法在网络舆情分析中发现热点问题的能力，本文采用 http://www.sogou.com/ 上下载的 sogou reduced 语料库，本文选取其中的 4 类共 4000 篇语料进行测试，每一类为 1000 篇，包括新闻类，军事类，体育类，教育类。

算法发现热点问题的能力高低可以由其对文本聚类的效果间接反映，所以验证其算法的有效性可以用文本聚类的指标进行衡量。衡量文本聚类结果指标通常采用算法执行时间和常用的 F_ measure[10]， F_measure 综合了信息检索领域中的查准率和查全率的概念。

表 1 是与传统的 K-means 算法 F 值的比较结果。显然算法执行时间要比 K-means 算法聚类时间消耗的时间长，所以这里仅仅比较 F 值。

表 1　与传统 K-means 算法 F 值的比较

文本	K-means	改进的 K-means
新闻类	72%	80%
军事类	67%	77%
体育类	69%	81%
教育类	70%	82%

由表 1 可以看出与传统的 K-means 算法比较，对于测试的 4 个文本集，改进后的 K-means 算法的 F_ measure 值平均高出 8%－12%，达到了 77%－82%，提高聚了类的准确性。 表 2 是与基于最大最小原则选择初始中心的 K-means 算法(简称基于 MMD 的 K-means 算法)所做的比较，表 2 显示了在执行时间 t 和 F_ measure 值上的比较结果。

表 2　与基于 MMD 的 K-means 算法的 F 值和 t 的比较

文本	改进的 k-means		基于 MMD 的 k-means	
	t(ms)	F	t(ms)	F
财经类	58800	80%	61275	82%
娱乐类	58724	77%	60986	80%
体育类	58243	81%	60378	84%
教育类	58573	82%	60864	85%

由表 2 显示结果可知，本文改进的 K-means 算法比基于 MMD 的 K-means 算法消耗的平均时间少约 2296(ms)，但是 F 值仅降低 2-3 个百分点。可以说明本文改进的 K-means 算法不仅降低了执行时间，同时保证了聚类结果的准确性。进而说明改进的 K-means 算法在网络舆情分析中能够快速准确的发现热点问题。

7 结束语

结合网络舆情分析的应用需求背景，本文介绍了

(下转第 196 页)

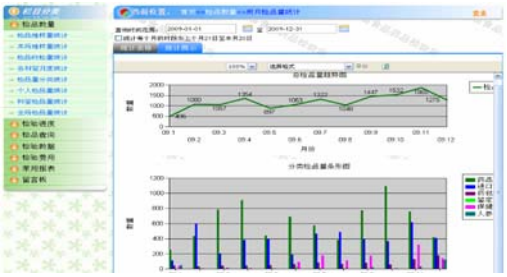

图 3 网站截图

4 结束语

本文介绍了 Reporting Services 在检品查询系统中的整合方案及实施关键技术，成功解决了原药品检验管理系统数据统计功能不足的缺点，系统具有查询简便、易于扩充等优点。由于 SQL Server Reporting Services 和 SQL Server 数据库结合紧密并且具有良好的扩展性，以及对企业商业智能(BI) 的支持，今后的主要工作是运用商业智能工具对海量检验数据进行智能化处理，为业务人员和领导决策提供支持和帮助。

参考文献

1 陈锋,郑晓琼.全国省级药品检验所信息化现状调研.中国药事,2008,22(1):34－35.

2 康苏明.Lotus Notes 数据库与关系型数据库的数据转换.山西大同大学学报(自然科学版),2008,24(5):56－58.

3 Mike Gunderloy, Jorden JL, Tschanz DW.曲丽君,李军田,毛选等译.SQL Server 2005 从入门到精通.北京:电子工业出版社,2006.748－803.

4 Mundy J.闫雷鸣,冯飞,译.数据仓库工具箱——面向 SQL Server 2005 和 Microsoft 商业智能工具集.北京:清华大学出版社,2007.295－297.

5 Larson B.赵志恒,武海锋译.Microsoft SQL Server 2005 商业智能实现.北京:清华大学出版社,2008.468－470.

6 韩敏,尤枫,赵恒永.基于 SQL2005 的企业报表系统的研究与实现.电脑知识与技术,2008,12:410－412.

7 Turley P, et al.谢文亮译.SQL Server 2005 报表服务高级编程.北京:清华大学出版社,2007.432－475.

(上接第 168 页)

文本信息的处理，重点研究了基于 K-means 算法的文本聚类方法，针对其对初始聚类中心的依赖性以及其它初始中心选择算法的效率较低的情况，给出了一种初始中心选择算法。实验表明改进算法能够快速准确发现网络舆情信息中的热点问题，为进一步实现话题追踪功能奠定了基础。

参考文献

1 Likas A, Vlassis N, Verbeek J. The global k-means clustering algorithm. Pattern Recognition, 2003,36(2):451.

2 李凡,林爱武,陈国社.一种基于 VSM 的文本分类系统的设计与实现.华中科技大学学报:自然科学版,2005,33(3):53.

3 MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. of the 5th Berkeley Symp. on Mathematics Statistic Problem, 1967: 281－297.

4 Dhillon IS, Modha DS.Concept decompositions for large sparse text data using clustering. Machine Learning, 2001, 42(1):143－175.

5 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of ACM, 1975,18(5): 613－620.

6 Bun KK. Topic Extraction from News Archive Using TF*PDT Algorithm. Proceedings of the 3rd International Conference on Web Information Systems Engineering. 2002.

7 赵亚琴,邹红艳.基于信息粒度的文本聚类算法.计算机工程与设计,2009,30(22):51－72.

8 Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques Proceeding of the 6th ACM-SIGKDD International Conference on Text Mining, Boston,MA,USA: ACM Press, 2000:103－122.

9 张睿.基于 K-means 算法的中文文本聚类算法的研究与实现[硕士学位论文].西安:西北大学,2009.29－30.

10 Steinbach M, Karypis G, Kumara V. A Comparison of Document Clustering Techniques. KDD-2000 Workshop on Text Mining, Boston MA, August 20-23, 2000: 109－110.