Journal of Computer Applications
计算机应用,2014,34(1):249-254
ISSN 1001-9081
CODEN JYIIDU
2014-01-10
http://www.joca.cn

文章编号:1001-9081(2014)01-0249-06
doi:10.11772/j.issn.1001-9081.2014.01.0249

基于 LDA 的改进 *K*-means 算法在文本聚类中的应用

王春龙[1*],张敬旭[2]
(1.华北电力大学 控制与计算机工程学院,北京 102206; 2.甘肃省电力公司,兰州 730030)
(*通信作者电子邮箱 wchlong0508@126.com)

摘 要:针对传统 *K*-means 算法初始聚类中心选择的随机性可能导致迭代次数增加、陷入局部最优和聚类结果不稳定现象的缺陷,提出一种基于隐含狄利克雷分布(LDA)主题概率模型的初始聚类中心选择算法。该算法选择蕴含在文本集中影响程度最大的前 m 个主题,并在这 m 个主题所在的维度上对文本集进行初步聚类,从而找到聚类中心,然后以这些聚类中心为初始聚类中心对文本集进行所有维度上的聚类,理论上保证了选择的初始聚类中心是基于概率可确定的。实验结果表明改进后算法聚类迭代次数明显减少,聚类结果更准确。

关键词:主题模型;*K*-means;聚类中心;文本聚类;隐含狄利克雷分布
中图分类号:TP301.6 **文献标志码**:A

Improved *K*-means algorithm based on latent Dirichlet allocation for text clustering

WANG Chunlong[1*], ZHANG Jingxu[2]
(1. *School of Control and Computer Engineering, North China Electric Power University, Beijing* 102206, *China*;
2. *Gansu Electric Power Corporation, Lanzhou Gansu* 730030, *China*)

Abstract: The traditional *K*-means algorithm has an increasing number of iterations, and often falls into local optimal solution and unstable clustering since the initial cluster centers are randomly selected. To solve these problems, an initial clustering centers selection algorithm based on Latent Dirichlet Allocation (LDA) model for the *K*-means algorithm was proposed. In this improved algorithm, the top-m most important topics in text corpora were first selected. Then, the text corpora was preliminarily clustered based on the m dimensions of topics. As a result, the m cluster centers could be got in the algorithm, which were used to further make clustering on all the dimensions of the text corpora. Theoretically, the center for each cluster can be determined based on the probability without randomly selecting them. The experiment demonstrates that the clustering results of the improved algorithm are more accurate with smaller number of iterations.

Key words: topic model; *K*-means; cluster center; text clustering; Latent Dirichlet Allocation (LDA)

0 引言

随着互联网的不断发展,网络上文本信息呈爆炸式增加,如何精准有效地发现、组织和利用海量文本背后的有用信息成为一个热门话题[1],文本聚类技术作为自然语言处理的预处理步骤[2],对文本进一步分析和处理产生了重要的影响。目前比较经典的文本聚类算法大致分为划分方法、层次方法、基于网格的方法、基于密度的方法以及基于模型的方法[3]。对于像大规模文本处理这样开销比较大的应用,划分方法相对来说具有较低的处理复杂度因而应用相对比较广泛。而基于划分的方法[4-7]有 *K*-means、*K*-prototypes、*K*-medoids 等,其中 *K*-means[8]是比较常用的算法。

K-means 的时间复杂度是线性的 $O(n*k*t)$,其中 k 为指定的类别数,n 为待聚类的样本数,t 为迭代次数的上限,算法简单且收敛速度快,所以更适应于处理大规模文本。然而也有着明显的不足[9]:该算法采用梯度法求解目标函数极值,如果初始聚类中心选择不好,结果很容易陷入局部最优,导致聚类结果不稳定。

在文本聚类方面,Dhillon 等[10]曾使用 *K*-means 算法,并利用余弦相似度来计算文本间的距离,而文本是非结构化或半结构化的,文本向量的维度高,具有稀疏性,不同簇之间相似度的差异性比较大,因此可能导致聚成一簇的文本之间的非相似性[11],传统 *K*-means 往往更容易陷入局部最值,导致较差的聚类结果,因此如何获得合适的初始聚类中心,在保证算法结果稳定性的同时保持其准确性,对提升算法的聚类性能显得尤为重要[4-5]。很多研究人员提出了一系列改进的 *K*-means 算法。文献[12]提出了一种改进的 *K*-means 算法,该算法在基于密度检测的基础上,在原来的算法中加入噪声数据特征检测的步骤,除去待聚类数据集中的噪声,提高了数据集的凝聚力,但是对于大规模文本数据集处理,该算法的复杂度是难以接受的。文献[13]算法不同于传统的算法对于每个类别只选择一个聚类中心,而是选择多个中心聚类点,此外该算法通过计算加权距离来分发数据点,并产生新的特征点集合,同时给出了相应的参数,然而这些参数没有可靠的理论依据。文献[14]提出了基于密度的初始聚类中心选择算法,即选择 k 个处于高密度区域的数据对象作为初始聚类中心,但是该方法中的密度参数难以确定,这一不足对其性能有很大的影响。Lai 等[15]提出了一种快速 *K*-means 聚类算法,该算法利用中心点去除一些不太适合作为候选中心点的样本点,该算法比基于 k-d 树的算法有很低的复杂度,但是依然受

收稿日期:2013-07-23;**修回日期**:2013-10-31。
基金项目:国家自然科学基金资助项目(61001197,61372182);国家电网公司科技项目(522722130292)。
作者简介:王春龙(1987-),男,河北保定人,硕士研究生,主要研究方向:信息检索、语义 Web; 张敬旭(1983-),男,山东莱芜人,硕士研究生,主要研究方向:信息系统。

到孤立点的影响。Chang 等[16]在 Jim 等[15]的基础上对算法做了进一步的改进。他们将文本分成多个独立的主题，并针对每个主题实施快速 K-means 算法。尽管这种方法能够降低算法复杂度，但是没有考虑到数据的分布模型，所以容易陷入局部最优，然而主题聚类方法比传统的方法要更有效[17]。

本文基于主题模型提出了一种改进的 K-means 算法，该算法采用了一种新的初始聚类中心点的选择方法，考虑到了文本-主题服从 Dirichlet 分布，主题-词服从多项式分布的分布特点，并使用了隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）模型对文本进行分析和处理。理论推导和实验结果表明本文算法选择的初始聚类中心是基于概率可确定的，迭代次数明显减少，聚类结果稳定且更准确。

1 文本主题空间模型及相似度定义

1.1 K-means 算法

K-means 算法是聚类分析中一种基本的划分方法。该算法首先对样本进行粗略聚类，然后根据某种修正原则对聚类结果进行不断优化，直到聚类结果比较合理为止。算法流程陈述如下：

输入　待聚类文本数 N，聚类数目 k，最大迭代次数 $Step_{max}$，迭代终止条件 ε。

输出　k 个聚类和迭代次数。

1）从 N 个待聚类的文本中任意选择 k 个文本作为初始聚类中心。

2）根据每个聚类文本的均值（中心文本或最靠近中心位置的文本），计算每个文本与这些中心文本的距离；并根据最小距离原则重新对相应样本进行划分。

3）重新计算每个（有变化）聚类文本的均值：计算标准测度函数 E，当迭代达到指定的迭代次数 $Step_{max}$ 或满足终止条件 $|E_{n+1}-E_n|\leqslant\varepsilon$，则算法终止；否则回到 2）。

其中：标准测度函数 $E=\sum_{i=1}^{k}\sum_{X\in C_i}|X-\overline{X}_i|^2$，$\overline{X}_i$ 为聚类 C_i 的中心文本。K-means 算法因其理论上可靠、算法简单、收敛速度快，能有效地处理大数据集而被广泛使用，研究表明[18]，对于文本聚类这样的应用使用 K-means 算法简单而有效；但传统的 K-means 算法对初始聚类中心敏感，使用不同的初始聚类中心聚类产生的聚类结果也不一样，容易陷入局部最优，此外它对噪声和离群点数据也是敏感的。因此本文提出了一种寻找初始聚类中心的方法，在提高抗噪声能力的同时使得初始聚类中心的分布尽可能体现数据的实际分布，从而找到比较具有代表性的文本作为初始聚类中心。

1.2 文本主题空间模型

本文需要使用 LDA 模型对文本进行建模并抽取主题，LDA 是一种非监督机器学习技术，是比较常用的概率主题模型，可用来识别大规模文本或语料库中隐含的主题信息，因此在文本相关领域中，如信息检索[19-20]、文本分类[21]等都得到了广泛的应用。

LDA 是一个三层贝叶斯概率模型，由词、主题和文本三层构成。该模型假设每个文本包含一定数量的隐含主题，而每个主题包含特定的词，文本和词汇间的关系通过隐含主题 Z_k 体现，其中文本到主题服从 Dirichlet 分布，主题到词服从多项式分布；隐含主题之间是相互独立的，这些主题被文本集中的所有文本所共享，而每个文本有一个特定的主题分布。将 LDA 模型引入聚类后如图 1 所示，可见主题总数 n 的确定对于最后的文本聚类有直接的影响。

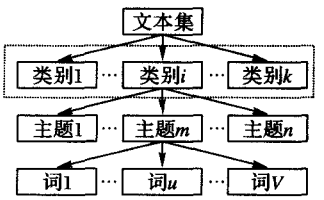

图 1　基于 LDA 的聚类模型

LDA 主题模型属于生成模型，对于包含 J 个文本、V 个不同词汇的文本集，LDA 定义了生成其中任意一个文本 $\boldsymbol{d}_j$ 的过程陈述如下：

1）对第 j 个文本 $\boldsymbol{d}_j$，从包含 N（反复实验后事先给定）个主题的主题多项式分布 $\boldsymbol{\theta}$ 中抽取一个主题 $\boldsymbol{Z}_k$；

2）从上述被抽到的主题 $\boldsymbol{Z}_k$ 所对应的单词多项式分布 $\boldsymbol{\varphi}$ 中抽取一个单词 $w_{j,n}$。

重复上述过程 N_j 次（N_j 是文本 $\boldsymbol{d}_j$ 中单词个数）直到遍历文本中的每个单词 $w_{j,n}$。

其中：对于涉及 N 个主题的 J 个文本，构成每个主题的词分布为一个从参数为 $\boldsymbol{\beta}$ 的 Dirichlet 先验分布中抽取的多项式分布。对于每个文本，其长度服从泊松分布，从一个参数为 $\boldsymbol{\alpha}$ 的 Dirichlet 先验分布中抽取一个多项式分布作为该文本中出现每个主题下词的概率分布；给定一个文本集合，根据 LDA 的生成文本过程，可以写出如式（1）所示的所有变量的联合分布。

$$p(\boldsymbol{w}_j,\boldsymbol{z}_j,\boldsymbol{\theta}_j,\boldsymbol{\phi}|\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{n=1}^{N_j}p(w_{j,n}|\boldsymbol{\varphi}_{z_{j,n}})p(z_{j,n}|\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j|\boldsymbol{\alpha})\cdot p(\boldsymbol{\phi}|\boldsymbol{\beta}) \quad (1)$$

其中：$w_{j,n}$ 是文本 $\boldsymbol{d}_j$ 中的第 n 个词，$z_{j,n}$ 是词 $w_{j,n}$ 所属的主题，N_j 是文本 $\boldsymbol{d}_j$ 的单词总数。$\boldsymbol{\beta}$ 是每个主题下词的多项分布的 Dirichlet 先验参数，$\boldsymbol{\alpha}$ 是每个文本下主题的多项分布的 Dirichlet 先验参数。$\boldsymbol{\theta}_j$ 和 $\boldsymbol{\varphi}_k$ 这两个隐含变量分别表示文本 $\boldsymbol{d}_j$ 下的主题分布和第 k 个主题下词的分布，前者是 k 维（k 为主题个数）向量，后者是 v 维向量（v 为词典中词的个数）。

其中 $w_{j,n}$ 初始化为一个词 t 的概率为：

$$p(w_{j,n}=t|\boldsymbol{\theta}_j,\boldsymbol{\phi})=\sum_{k=1}^{N}p(w_{j,n}=t|\boldsymbol{\varphi}_k)p(z_{j,n}=k|\boldsymbol{\theta}_j) \quad (2)$$

即文本中出现主题 k 的概率与主题 k 下出现词 t 概率的乘积，然后遍历所有主题求和得到。于是文本集的似然函数为：

$$p(\boldsymbol{W}|\boldsymbol{\theta},\boldsymbol{\phi})=\prod_{j=1}^{J}p(\boldsymbol{w}_j|\boldsymbol{\theta}_j,\boldsymbol{\phi})=\prod_{j=1}^{J}\prod_{n=1}^{N_j}p(w_{j,n}|\boldsymbol{\theta}_j,\boldsymbol{\phi}) \quad (3)$$

最后通过 Gibbs 抽样法求得该模型中的两个参数："文本-主题"分布矩阵 $\boldsymbol{\theta}$ 和"主题-词"分布矩阵 $\boldsymbol{\varphi}$。

1.3 语义相似度

通过对文本进行主题模型建模，并采用 Gibbs 抽样法求解得到文本的主题概率向量，然后定义主题相关度函数用于计算两个文本之间的相似度。不同模型对应不同的相似度计算方法，在向量空间模型中文本用关键词的标准化 TF/IDF 值[22]来衡量，可以使用欧氏距离或余弦相似度计算。在 LDA 主题模型中文本用服从 Dirichlet 分布的主题概率向量来衡量。若仍使用余弦夹角来计算文本相似度就失去了主题模型的优势[23]，本文使用能够度量概率分布距离的相似度函数，即 JS（Jensen-Shannon）距离函数来定义主题概率向量 $\boldsymbol{p}$ =

$(p_1, p_2, \cdots, p_k)$ 到 $\boldsymbol{q} = (q_1, q_2, \cdots, q_k)$ 的距离,具体的距离定义为:

$$D_{js}(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{2}\left(\sum_{j=1}^{k} p_j \ln\frac{p_j}{q_j} + \sum_{j=1}^{k} q_j \ln\frac{q_j}{p_j}\right) \tag{4}$$

其中:$\boldsymbol{p}$、$\boldsymbol{q}$ 为主题概率分布。

2　本文算法

2.1　算法描述及解释

本文算法首先对语料库中的文本进行分词、去停用词、计算 TF/IDF 值、向量化、标准化等预处理操作,得到文本集的文本向量矩阵,并从文本集中提取词典(文本集中出现过的所有词的有序集合);然后引入 LDA 模型对文本向量进行建模,并使用 Gibbs 抽样法对建模后的文本向量矩阵进行求解,得到文本-主题矩阵和主题-词矩阵,从而抽取文本集中隐含的 n 个主题;最后从这 n 个主题中选取符合要求的 m 个主题,在这 m 个主题所在的维度上对文本集进行初步的 K-means聚类(预先设置为 k 类),从而得到 k 个聚类中心,进一步用这 k 个中心作为初始聚类中心在所有维度上对文本集进行聚类。其中本文的工作重点在于:1) 如何确定主题数目 n;2) 如何从 n 个主题中选择 m 个主题;3) 如何获取初始聚类中心。

2.1.1　主题数目的确定

由图 1 不难看出主题个数的选择会直接影响模型的好坏,从而直接影响聚类的精度,所以找到一个最优的主题个数很重要。文献[24]中证明主题间的平均相似度最小时的主题数目是最优的,相应的主题模型是最优的。本文使用最优主题数目选择算法[25]确定所用文本集的最优主题数目。首先定义主题 $\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j$ 间的相似度如下:

$$d_{ij} = \sum D_{js}(\boldsymbol{z}_i, \boldsymbol{z}_j) \tag{5}$$

设主题个数为 n,则主题间的平均相似度定义如下:

$$\bar{d} = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}}{n\cdot(n-1)} \tag{6}$$

而当 $\bar{d}$ 取到最小值时,即可得到最优主题数 n,详细过程参考文献[25]。然后通过对文本集进行 LDA 建模,并使用 Gibbs 抽样法对建模后的文本向量矩阵进行求解,即可抽取到文本集的 n 个主题。

2.1.2　最佳主题的选取

由于直接使用 LDA 模型提取出来的所有 n 主题用于聚类,会出现聚类结果不稳定现象[26],所以有必要从中选取 m 个有代表性的主题。本文分析了 n 个主题在文本中的分布情况,并定义了一个衡量指标 $TI_j(\boldsymbol{t}_i)$ 来计算主题 $\boldsymbol{t}_i$ 在文本 $\boldsymbol{d}_j = (\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n)$ 中的重要程度即包含的信息量大小,其中 $TI_j(\boldsymbol{t}_i)$ 值越大表示主题 $\boldsymbol{t}_i$ 在文本 $\boldsymbol{d}_j$ 中越重要,主题 $\boldsymbol{t}_i$ 包含的信息量越大,也即对聚类产生的作用越明显。

定义 1　$TI_j(\boldsymbol{t}_i)$。主题 $\boldsymbol{t}_i$ 对于第 j 个文本 $\boldsymbol{d}_j = (\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n)$ 中的重要程度 $TI_j(\boldsymbol{t}_i)$ 定义如下:

$$TI_j(\boldsymbol{t}_i) = T_j(\boldsymbol{t}_i)\cdot I_j(\boldsymbol{t}_i) \tag{7}$$

其中 $T_j(\boldsymbol{t}_i)$ 指主题 $\boldsymbol{t}_i$ 对于文本 $\boldsymbol{d}_j$ 的条件概率,即:

$$T_j(\boldsymbol{t}_i) = P(\boldsymbol{t}_i \mid \boldsymbol{d}_j) = \frac{DT(j,i)}{\sum_{k=1}^{n} DT(j,k)} \tag{8}$$

其中 $\boldsymbol{DT}$ 为事先生成的"文本-主题"矩阵,而 $I_j(\boldsymbol{t}_i)$ 定义如下:

$$I_j(\boldsymbol{t}_i) = \ln\frac{1}{P(\boldsymbol{d}_j \mid \boldsymbol{t}_i)} = \ln\frac{\sum_{k=1}^{J} DT(k,i)}{DT(j,i)} \tag{9}$$

其中 J 为待聚类文本总数,结合式(7) ~ (9) 得到:

$$TI_j(\boldsymbol{t}_i) = P(\boldsymbol{t}_i \mid \boldsymbol{d}_j)\cdot\ln\frac{1}{P(\boldsymbol{d}_j \mid \boldsymbol{t}_i)} = \frac{DT(j,i)}{\sum_{k=1}^{n} DT(j,k)}\cdot\ln\frac{\sum_{k=1}^{J} DT(k,i)}{DT(j,i)} \tag{10}$$

定义 2　$TI(\boldsymbol{t}_i)$。主题 $\boldsymbol{t}_i$ 在文本集上的重要程度为 $TI(\boldsymbol{t}_i)$。

$$TI(\boldsymbol{t}_i) = \sum_{j=1}^{J} TI_j(\boldsymbol{t}_i) \tag{11}$$

其中 J 为文本总数。在 n 确定的情况下,选择前 m 个主题方法如下:首先将所有主题按照其 $TI(\boldsymbol{t}_i)$ 大小进行降序排序,得到如图 3 所示的主题重要程度分布图(本文使用搜狗提供的语料库,其主题总数为 124),然后选择前 m 个主题作为最佳的主题参与后续的聚类。定义 3 给出了如何确定 m 值的方法。

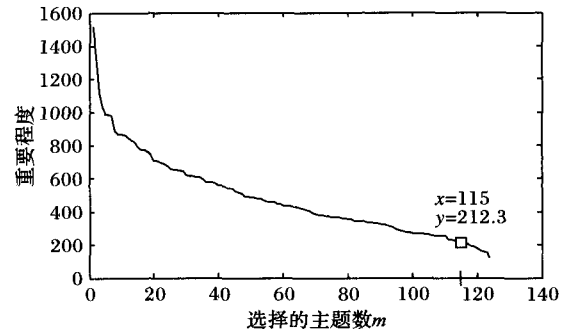

图 2　主题重要程度分布

定义 3　重要程度误差 δ。δ 表示选择的最佳 m 个主题重要程度之和与全部 n 个主题重要程度之和的误差,

$$1 - \frac{\sum_{i=1}^{m} TI(\boldsymbol{t}_i)}{\sum_{i=1}^{n} TI(\boldsymbol{t}_i)} = \frac{\sum_{i=m+1}^{n} TI(\boldsymbol{t}_i)}{\sum_{i=1}^{n} TI(\boldsymbol{t}_i)} = \delta \tag{12}$$

如图 2 所示,当 m 为 115 时,$\sum_{i=1}^{m} TI(\boldsymbol{t}_i)$ 可以理解为 m = 115 左侧部分曲线与坐标轴围成的面积大小,它在一定程度上反映了前 m 个主题包含信息量的大小;δ 可以理解为剩余 $n-m$ 个主题包含信息量大小与全部信息的比值。进一步,可以通过控制误差 δ 的大小来找到合适的 m 值。

2.1.3　基于 LDA 模型获取初始聚类中心

通过设置经验误差 δ 的值来确定最佳主题数目 m,并使用 2.1.2 节中介绍最佳主题的选取方法从 n 个主题中选择这 m 个主题,然后使用 K-means 算法在这 m 个主题所对应的维度上对文本集进行聚类,在计算文本间相似度时采用 JS 距离而不是欧氏距离或余弦相似度计算,并使用如下的目标函数:

$$H = \sum D_{js}(\boldsymbol{k}_i, \boldsymbol{d}_j^i) \tag{13}$$

即每个样本点 $\boldsymbol{d}_j^i$ 到其质心 $\boldsymbol{k}_i$ 的距离和,其中 $0 < i < k$(k 为聚类个数),这样可以得到 k 个聚类中心,这 k 个聚类中心就可以作为初始聚类中心在所有维度上对文本集进行 K-means聚类。通过该方法产生的初始聚类中心能够在很大程度上减少聚类的迭代次数,更重要的是该方法产生的初始聚类中心是基于概率可确定的,摆脱了选择时的任意性。

2.2 基于概率可确定性证明

本文的核心在于衡量主题重要程度指标 TI 的定义及基于 JS 距离的文本初始聚类中心选择方法的概率可确定性证明。文本集初始类别标识集合为 $\boldsymbol{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_i\}$，对应的类别中心集合为 $\boldsymbol{k}=\{\boldsymbol{k}_1,\boldsymbol{k}_2,\cdots,\boldsymbol{k}_i\}$，聚类后文本集的类别标识集合为 $\boldsymbol{C}'=\{\boldsymbol{C}'_1,\boldsymbol{C}'_2,\cdots,\boldsymbol{C}'_i\}$，对应的类别中心集合为 $\boldsymbol{k}'=\{\boldsymbol{k}'_1,\boldsymbol{k}'_2,\cdots,\boldsymbol{k}'_i\}$，这里 $\boldsymbol{C}_i$ 与 $\boldsymbol{C}'_i$ 一一对应，所以 $\boldsymbol{k}_i$ 与 $\boldsymbol{k}'_i$ 有相同的类别标识，即中心点 $\boldsymbol{k}_i$ 与 $\boldsymbol{k}'_i$ 属于同一类。

假设在 $\boldsymbol{k}_i$ 与 $\boldsymbol{k}'_i$ 属于同一类时，实验聚类中心点 $\{\boldsymbol{k}'_1,\boldsymbol{k}'_2,\cdots,\boldsymbol{k}'_i\}$ 与真实类别中心 $\{\boldsymbol{k}_1,\boldsymbol{k}_2,\cdots,\boldsymbol{k}_i\}$ 距离越小，算法聚类精度越高。基于该假设定义二者的距离如下：

$$D=\sum_{i=1}^{k}D_{js}(\boldsymbol{k}_i,\boldsymbol{k}'_i) \tag{14}$$

需要证明存在 $\{\boldsymbol{k}'_1,\boldsymbol{k}'_2,\cdots,\boldsymbol{k}'_i\}$ 使 D 取到最小值。其中式(14)中的类别中心点 $\boldsymbol{k}_i=(p_1^i,p_2^i,\cdots,p_m^i)$ 和 $\boldsymbol{k}'_i=(x_1^i,x_2^i,\cdots,x_m^i)$ 均为 m 维的主题向量：

$$D=\sum_{i=1}^{k}D_{js}(\boldsymbol{k}_i,\boldsymbol{k}'_i)=\sum_{i=1}^{k}\frac{1}{2}\left(\sum_{j=1}^{m}p_j^i\ln\frac{p_j^i}{q_j^i}+\sum_{j=1}^{m}q_j^i\ln\frac{q_j^i}{p_j^i}\right) \tag{15}$$

其中：$x_1^i,x_2^i,\cdots,x_m^i$ 是变化的，而真实类别中心 $\{\boldsymbol{k}_1,\boldsymbol{k}_2,\cdots,\boldsymbol{k}_i\}$ 是不变的，$\boldsymbol{k}_i=(p_1^i,p_2^i,\cdots,p_m^i)$ 也是不变的，于是

$$D'=\left(\sum_{i=1}^{k}\frac{1}{2}\left(\sum_{j=1}^{m}p_j^i\ln\frac{p_j^i}{x_j^i}+\sum_{j=1}^{m}x_j^i\ln\frac{x_j^i}{p_j^i}\right)\right)'=$$
$$\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(p_j^i\ln\frac{p_j^i}{x_j^i}+x_j^i\ln\frac{x_j^i}{p_j^i}\right)'=$$
$$\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(\frac{-p_j^i}{x_j^i}+\ln\frac{x_j^i}{p_j^i}+1\right)$$

记 $\varphi_j^i\triangleq\frac{-p_j^i}{x_j^i}+\ln\frac{x_j^i}{p_j^i}+1$

由于对文本进行的是主题概率模型建模，所以 $0<x_j^i<1$ 且 $0<p_j^i<1$。

当 $0<x_j^i<p_j^i<1$ 时，$\varphi_j^i\triangleq\left(1-\frac{p_j^i}{x_j^i}\right)+\ln\frac{x_j^i}{p_j^i}<0$，从而 $D'<0$ 时即 D 为关于 $x_1^i,x_2^i,\cdots,x_m^i$ 的减函数。

当 $0<p_j^i<x_j^i<1$ 时，$D'>0$ 即 D 为关于 $x_1^i,x_2^i,\cdots,x_m^i$ 的增函数。

所以 $D'=0$ 时存在唯一的聚类中心 $\{\boldsymbol{k}'_1,\boldsymbol{k}'_2,\cdots,\boldsymbol{k}'_i\}$ 使 D 取到最小值，从而证明根据重要程度 $TI_j(t_i)$ 选择的初始聚类中心是唯一的，即本文算法选择的初始聚类中心是基于概率可确定的，摆脱了任意性。

2.3 时间复杂度分析

本文提出了一种新的初始聚类中心选择方法，该方法用到了本文定义的指标量 $TI_j(t_i)$ 来计算主题 $\boldsymbol{t}_i$ 在文本 $\boldsymbol{d}_j=(t_1,t_2,\cdots,t_n)$ 中的重要程度，其中 $TI_j(t_i)$ 值越大表示主题 $\boldsymbol{t}_i$ 在文本 $\boldsymbol{d}_j$ 中越重要，也即对聚类产生的作用越明显，过程中需要事先计算并存储“文本-主题”矩阵 $\boldsymbol{DT}$ 每列之和 $\sum_{k=1}^{J}DT(k,i)$ $(1\leqslant i\leqslant n)$ 的时间复杂度为 $O(nJ)$，其中 J 为待聚类文本总数，n 为主题个数，需要计算每个主题 $\boldsymbol{t}_i$ 在文本集上的重要性 $TI_j(t_i)$，即 $TI(t_i)=\sum_{j=1}^{J}TI_j(t_i)$，则计算文本-主题重要程度矩阵时间复杂度为 $O(nJ)$，找到 $TI_j(t_i)$ 值前 m 个最大的主题时间复杂度为 $O(\lg n)$，利用这 m 个主题进行初步 K-means 聚类时间复杂度为 $O(ktJ)$，其中，k 为指定的类别数，t 为迭代次数的上限，所以总的时间复杂度为：

$$O(nJ)+O(nJ)+O(\lg n)+O(ktJ)=O(J)$$

因此本文提出的初始聚类中心确定方法的时间复杂度为 $O(J)$，即线性时间复杂度。

3 实验及结果分析

3.1 数据集简介及结果评估指标

本文通过实验，考察了使用新的初始聚类中心选择算法后的 K-means 聚类性能。文中使用搜狗实验室提供的文本分类语料库[27]，其中包含 IT、财经、健康、教育、军事、旅游、汽车、体育、文化、招聘 10 个类别，每个类别有 8 000 个文本，实验随机抽取每个类别中的 6 000 个文本进行分析，并使用 IK 分词器、Lucene 工具对文本进行分词、去停用词、字典提取等预处理操作，同时采用信息抽取领域的评测方法，使用查准率和 F 值来评价本文算法的性能，F 度量值是信息检索中一种组合查全率和查准率指标的平衡指标，F 度量值越大，聚类效果越好。

查全率：

$$recall(r,s)=\frac{n(r,s)}{ns} \tag{16}$$

查准率：

$$precision(r,s)=\frac{n(r,s)}{nr} \tag{17}$$

其中：r 为聚类结果中的某个类别，s 为真实类别，$n(r,s)$ 是聚类 r 中包含类别 s 中文本的个数，nr 是聚类类别 r 中文本的个数，ns 是预定类别 s 中文本的个数，则聚类 r 和类别 s 之间的 F 值为

$$F(r,s)=\frac{2\,recall(r,s)\cdot precision(r,s)}{recall(r,s)+precision(r,s)} \tag{18}$$

总体评价函数为：

$$F=\sum_{i}\frac{n_i}{n}\max\{F(i,j)\} \tag{19}$$

其中：j 为文本集的真实类别，其取值为区间[1,10]内的整数；n 是所有测试文本的个数，而 n_i 表示预定类别 i 中的文本个数。

3.2 实验结果分析

由于 α、β 存在经验值，这里设置 $\alpha=50/K$、$\beta=0.01$，主题数目设置合适与否，直接影响模型的精度，从而会影响聚类结果的准确性。经计算得到语料库的最佳主题数为 124，并用 LDA 模型对其建模，最后抽取这些主题，主题-词分布表如表 1 所示。实验分为两部分：1）分析使用所有主题进行聚类会产生不稳定现象；2）验证本文算法聚类精度的提高。

表 1 主题-词分布表(部分)

主题 1	主题 2	…	主题 124
美元(0.030)	中国(0.224)	…	问题(0.011)
市场(0.021)	美国(0.029)	…	一个(0.009)
公司(0.020)	印度(0.024)	…	方面(0.009)
全球(0.016)	认为(0.012)	…	发展(0.008)
企业(0.010)	欧洲(0.009)	…	需要(0.005)
中国(0.010)	国家(0.009)	…	因此(0.005)
表示(0.009)	发展(0.008)	…	分析(0.004)

注：“美元(0.030)”表示“美元”这个词在主题 1 中出现的概率。

为说明直接使用所有主题(即 δ 值为 0 时)进行聚类会产

生不稳定现象,实验通过设置 δ 值分别为 0.35,0.25,0.15,0.05,0 来选取前 m 个 $TI(t_i)$ 最大的主题,其中 m 值分别对应为40,65,90,115,124,如图 3 所示。不难发现不同 m 值得到的初始聚类中心用于 K-means 算法将导致不同的迭代次数和不同的 F 值。

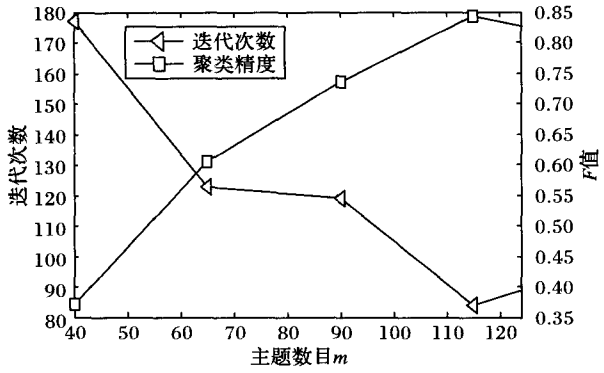

图 3　不同 m 值对 K-means 算法的影响对比

聚类结果稳定时的迭代次数随着 m 的增大会不断减小,而 F 值会不断提高;当 m 超过某个值时(这里实验结果在 115 附近)迭代次数会逐渐增大,而 F 值会出现降低的趋势。通过分析发现开始时随着 m 的不断增大,被选取的主题包含的信息量越来越多,δ 值减小,产生的初始聚类中心的质量也越来越高,即越来越接近真实的聚类结果中心,有利于 K-means 算法的快速收敛和 F 值的提高;但当 m 超过某个值时,选择的主题中包含的"噪声"信息也开始变多,从而会导致聚类结果产生不稳定的现象,可见主题选择的重要性。

本文算法从所有主题中选取最佳的主题用于产生合适的初始聚类中心,进一步提高最后的 F 值,如表 2 所示。实验中选取 δ 值为 0.05,即 m 为 115 时产生的初始聚类中心进行 K-means算法聚类,并对 K-means 算法和本文算法从查准率和 F 值两个方面做了对比,分析发现本文算法在这两个方面都有不同程度的提高,验证了初始聚类中心选择方法的有效性。实验数据表明当 $m=115$ 时,原始 K-means 算法达到稳定的聚类结果平均需要迭代 80 次,而本文算法达到稳定的聚类结果仅平均需要迭代 53 次。通过对上述实验结果的分析发现主题个数会影响初始聚类中心的选择,从而影响 K-means 算法迭代次数和 F 值。

表 2　本文算法与原始 K-means 算法查准率、F 值对比

类别	查准率		F 值	
	原始 K-means 算法	本文算法	原始 K-means 算法	本文算法
IT	0.693	0.872	0.818	0.929
财经	0.847	0.858	0.570	0.875
健康	0.692	0.921	0.707	0.835
教育	0.775	0.847	0.865	0.872
军事	0.833	0.862	0.567	0.907
旅游	0.850	0.871	0.813	0.829
汽车	0.837	0.861	0.809	0.824
体育	0.861	0.913	0.580	0.766
文化	0.750	0.855	0.732	0.847
招聘	0.832	0.918	0.812	0.892

本文算法引入了衡量主题重要性的指标量 TI,该量度综合考虑到局部信息和整体信息,通过合理地选择参与聚类的主题,进一步降低了噪声的影响,降低对孤立点的敏感性,进而找到符合文本语义分布的初始聚类中心,促使 K-means 算法更快地收敛,在降低了算法复杂度的同时得到了更高精度的聚类结果。

4　结语

本文算法摆脱了初始聚类中心选择的任意性,减少了算法的迭代次数,选择出的初始聚类中心是基于概率可确定的,符合文本分布特点且更具代表性。该方法对大规模文本处理有着重大的意义,更重要的是提高了最终的聚类精度,使聚类结果更接近真实文本类别。LDA 模型对大规模文本集中的隐含主题抽取的准确度,直接影响后期的聚类精度,所以找到精度更高的文本-主题概率模型是进一步研究的重点。

参考文献:

[1] LIKAS A, VLASSIS N J, VERBEEK J. The global *K*-means clustering algorithm[J]. Pattern Recognition, 2003, 36(2): 451 - 461.

[2] HATZIVASSILOGLOU V, KLAVANS J L, HOLCOMBE M L, *et al*. SIMFINDER: a flexible clustering tool for summarization [C]// Proceedings of NAACL Workshop on Automatic Summarization, Association for Computational Linguistics. Pittsburgh: [s. n.], 2001: 4 - 14.

[3] SUN J G, LIU J, ZHAO L Y. Clustering algorithms research[J]. Journal of Software, 2008, 19(1): 48 - 61. (孙吉贵, 刘杰, 赵连宇. 聚类算法综述[J]. 软件学报, 2008, (19)1: 48 - 61.)

[4] HAMERLY G, ELKAN C. Learning the *k* in *K*-means[EB/OL]. [2012-10-10]. http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2003_AA36.pdf.

[5] YI S. Global optimization for neural network training[J]. IEEE Computer, 2006, 29(3): 45 - 54.

[6] ZHANG Y F, MAO J L, XIONG Z. An improved *K*-means algorithm[J]. Computer Applications, 2003, 23(8): 31 - 33. (张玉芳, 毛嘉丽, 熊忠阳. 一种改进的 *K*-meana 算法[J]. 计算机应用, 2003, 23(8): 31 - 33.)

[7] ORDONEZ C, OMIECINSKI E. Efficient disk-based *K*-means clustering for relational databases[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(8): 909 - 921.

[8] MACQUEEN J B. Some methods for classification and analysis of multivariate observations [C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967, 1: 281 - 297.

[9] XING X S, PAN J, JIAO L C. A novel *K*-means clustering based on the immune programming algorithm[J]. Chinese Journal of Computers, 2003, 26(5): 605 - 610. (行小帅, 潘进, 焦李成. 基于免疫规划的 *K*-means 聚类算法[J]. 计算机学报, 2003, 26(5): 605 - 610.)

[10] DHILLON I S, MODHA D S. Concept decompositions for large sparse text data using clustering[J]. Machine Learning, 2001, 42 (1/2): 143 - 175.

[11] ZHANG M, WANG D L, YU G. A text clustering method based on auto-selected threshold[J]. Journal of Computer Research and Development, 2004, 41(10): 1748 - 1753. (张猛, 王大玲, 于戈. 一种基于自动阈值发现的文本聚类方法[J]. 计算机研究与发展, 2004, 41(10): 1748 - 1753.)

[12] WANG J, SU X. An improved *K*-means clustering algorithm[C]// Proceedings of the 3rd International Conference on Communication Software and Networks. Washington, DC: IEEE Computer Society, 2011: 44 - 46.

[13] WANG Z, LIU G Q, GUO J C. An improved *K*-means algorithm

based on multiple feature points[C]// Proceedings of International Workshop on Intelligent Systems and Applications. Washington, DC: IEEE Computer Society, 2009: 1-5.

[14] YUAN F, ZHOU Z, SONG X. *K*-means clustering algorithm with meliorated initial centers[J]. Computer Engineering, 2007, 32(3):65-66. (袁方, 周志勇, 宋鑫. 初始聚类中心优化的*k*-means算法[J]. 计算机工程, 2007, 33(3):65-66.)

[15] LAI J Z C, HUANG T J, LIAW Y C. A fast *K*-means clustering algorithm using cluster center displacement[J]. Pattern Recognition, 2009, 42(11):2551-2556.

[16] CHANG C T, LAI J Z C, JENG M D. A fuzzy *K*-means clustering algorithm using cluster center displacement[J]. Journal of Information Science and Engineering, 2011, 27(3):995-1009.

[17] ZHANG M W, LIU Y, ZHANG B, *et al.* Concept-based data clustering model[J]. Journal of Software, 2009, 20(9): 2387-2396. (张明卫, 刘莹, 张斌, 等. 一种基于概念的数据聚类模型[J]. 软件学报, 2009, 20(9):2387-2396.)

[18] LARSEN B, AONE C. Fast and effective text mining using linear-time document clustering[C]// Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 1999: 16-22.

[19] ANDRZEJEWSKI D, BUTTLER D. Latent topic feedback for information retrieval[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 600-608.

[20] WANG X, McCALLUM A, WEI X. Topical N-grams: phrase and topic discovery, with an application to information retrieval[C]// Proceedings of the Seventh IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2007: 697-702.

[21] RUBIN T N, CHAMBERS A, SMYTH P. Statistical topic models for multi-label document classification[J]. Machine Learning, 2012, 88(1/2):57-208.

[22] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5):513-523.

[23] JIA X P, LIU H Z H. Latent document similarity model[J]. Computer Engineering, 2009, 35(15):32-34. (贾西平, 刘海珠. 一种潜在文档相似模型[J]. 计算机工程, 2009, 35(15):32-34.)

[24] CAO J, ZHANG Y D, LI J T, *et al.* A method of adaptively selecting best LDA model based on density[J]. Chinese Journal of Computers, 2008, 31(10):1-8. (曹娟, 张勇东, 李锦涛. 一种基于密度的自适应最优模型选择方法[J]. 计算机学报, 2008, 31(10):1-8.)

[25] WANG L D, WEI B G, YUAN J. Document clustering based on probabilistic topic model[J]. Acta Electronica Sinica, 2012, 40(11):2346-2350. (王李冬, 魏宝刚, 袁杰. 基于概率主题模型的文档聚类[J]. 电子学报, 2012, 40(11):2346-2350.)

[26] ZHANG M X, WANG S G, WANG Z Q. A feature selection algorithm based on LDA for texts clustering[J]. Computer Development & Applications, 2012, 25(1):1-5. (张梦笑, 王素格, 王智强. 基于LDA特征选择的文本聚类[J]. 电脑开发与应用, 2012, 25(1):1-5.)

[27] Web Page Collection(SogouT)[EB/OL]. [2013-05-01]. http://www.sogou.com/labs/dl/t.html. (搜狗互联网语料库[EB/OL]. [2013-05-01]. http://www.sogou.com/labs/dl/t.html.)

(上接第238页)

[2] WANG D Q, DING F. Input-output data filtering based recursive least squares parameter estimation for CARARMA systems[J]. Digital Signal Processing, 2010, 20(4): 991-999.

[3] XIAO Y S, YUE N. Parameter estimation for nonlinear dynamical adjustment models[J]. Mathematical and Computer Modelling, 2011, 54(5/6): 1561-1568.

[4] XIE L, YANG H Z, DING F. Recursive least squares parameter estimation for non-uniformly sampled systems based on the data filtering[J]. Mathematical and Computer Modelling, 2011, 54(1/2): 315-324.

[5] DING F, CHEN T, QIU L. Bias compensation based recursive least-squares identification algorithm for MISO systems[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2006, 53(5): 349-353.

[6] DING Feng. System identification new theory and methods[M]. Beijing: Science Press, 2013. (丁锋. 系统辨识新论[M]. 北京: 科学出版社, 2013.)

[7] SHEN C, SUN Y X, HUANG L P, *et al.* Improved fuzzy auto-regressive model for connection rate prediction[J]. Journal of Computer Applications, 2013, 33(5): 1222-1229. (申晨, 孙永雄, 黄丽平等. 改进模糊自回归模型在预测网络接通率中的应用[J]. 计算机应用, 2013, 33(5):1222-1229.)

[8] LIU S R, ZHU W T, YANG F, *et al.* Multi-feature fusion based particle filter algorithm for object tracking[J]. Information and Control, 2012, 41(6):752-759. (刘士荣, 朱伟涛, 杨帆, 等. 基于多特征融合的粒子滤波目标跟踪法[J]. 信息与控制, 2012, 41(6):752-759.)

[9] LI M Y, BAI P, WANG X H, *et al.* A precise synchronization method based on iterative least square algorithm[J]. Journal of Electronics & Information Technology, 2013, 35(4):832-837. (李明阳, 柏鹏, 王徐华, 等. 基于一种迭代最小二乘法的精确同步方法[J]. 电子与信息学报, 2013, 35(4):832-837.)

[10] WANG H B. Research of information fusion technologies and existing problems in the Internet of things[J]. Application Research of Computers, 2013, 30(8): 2252-2255. (王洪波. 物联网信息融合技术及存在的问题研究[J]. 计算机应用研究, 2013, 30(8):2252-2255.)

[11] CHEN P, QIAN H, ZHU M L. Weighted minimum mean square Kalman filter[J]. Computer Science, 2009, 36(11): 230-231. (陈鹏, 钱徽, 朱淼良. 基于加权最小二乘的卡尔曼滤波算法[J]. 计算机科学, 2009, 36(11): 230-231.)

[12] XIANG W, CHEN Z H. New identification method of nonlinear systems based on Hammerstein models[J]. Control Theory & Applications, 2007, 24(1):143-147. (向微, 陈宗海. 基于Hammerstein模型描述的非线性系统辨识新方法[J]. 控制理论与应用, 2007, 24(1):143-147.)

[13] WANG Y, LIAO Z, PENG C. *et al.* Subspace identification of distributed order systems in time-domain[J]. Control and Decision, 2013, 28(1): 67-72. (王永, 廖增, 彭程, 等. 分布阶次系统时域子空间辨识[J]. 控制与决策, 2013, 28(1): 67-72.)

[14] LIU F, ZHAO F J, DENG Y K, *et al.* A new high resolution DBS imaging algorithm based least squares linear fitting[J]. Journal of Electronics & Information Technology, 2011, 33(4): 787-837. (刘凡, 赵凤军, 邓云凯, 等. 基于一种最小二乘直线拟合的高分辨率DBS成像算法[J]. 电子与信息学报, 2011, 33(4):787-837.)