Journal of System Simulation

Volume 30 | Issue 10 Article 29

1-4-2019

Optimizing Initial Cluster Centroids by SVD in K-means Algorithm for Chinese Text Clustering

Yueming Dai
Engineering Research Center of Internet of Things Technology Applications Ministry of Education, Jiangnan University, Wuxi 214122, China;

Minghui Wang
Engineering Research Center of Internet of Things Technology Applications Ministry of Education, Jiangnan University, Wuxi 214122, China;

Zhang Ming
Engineering Research Center of Internet of Things Technology Applications Ministry of Education, Jiangnan University, Wuxi 214122, China;

Wang Yan
Engineering Research Center of Internet of Things Technology Applications Ministry of Education, Jiangnan University, Wuxi 214122, China;

Optimizing Initial Cluster Centroids by SVD in K-means Algorithm for Chinese Text Clustering

Abstract

Abstract: In process of clustering with traditional K-means algorithm, it is difficult to identify the value of the number of clusters K and its clustering results are influenced by initial centers. It has the weakness of sensitivity to noise and instability. Meanwhile, to solve the problems for the high dimensions, sparse spatial and latent semantic structure of the text data, an algorithm for Chinese text clustering was proposed. This new algorithm uses the physical significance of Singular Value Decomposition (SVD) to firstly classify the data rough, and then uses K-means for text clustering. *It applies SVD to decompose and keep semantic features, remove noise, make smoothing process of text data, meanwhile, it takes the advantage of physical significance of SVD to have rough set classification, and then regard classification results as initial centers of K-means.* Experiment results demonstrate that the F-Measure of cluster quality has been improved compared with other K-means algorithms.

Keywords

SVD, text clustering, K-means, initial center point

Recommended Citation

Dai Yueming, Wang Minghui, Zhang Ming, Wang Yan. Optimizing Initial Cluster Centroids by SVD in K-means Algorithm for Chinese Text Clustering[J]. Journal of System Simulation, 2018, 30(10): 3835-3842.

第 30 卷第 10 期 系统仿真学报© Vol. 30 No. 10
2018 年 10 月 Journal of System Simulation Oct., 2018

SVD 优化初始簇中心的 K-means 中文文本聚类算法

戴月明*，王明慧，张明，王艳
(江南大学 教育部物联网技术应用工程研究中心，江苏 无锡 214122)

摘要：为了改善传统 K-means 算法在聚类过程中，聚类数目 K 难以准确预设，聚类结果受初始中心影响，对噪声点敏感，不稳定等缺点，同时针对文本聚类中文本向量化后数据维数较高，空间分布稀疏，存在潜在语义结构等问题，提出了一种利用奇异值分解(Singular Value Decomposition, SVD)的物理意义进行粗糙分类，再结合 K-means 算法的中文文本聚类优化算法(SVD-Kmeans)。*新算法利用 SVD 分解的数学意义对文本数据进行了平滑处理，同时利用 SVD 分解的物理意义对文本数据进行粗糙分类，将分类的结果作为 K-means 算法的初始聚类中心点。*实验结果表明，相比其他 K-means 及其改进算法，SVD-Kmeans 算法的聚类质量 F-Measure 值有明显提升。

关键词：SVD; 文本聚类; K-means; *初始中心点*
中图分类号：TP317 文献标识码：A 文章编号：1004-731X (2018) 10-3835-08
DOI: 10.16182/j.issn1004731x.joss.201810029

Optimizing Initial Cluster Centroids by SVD in K-means Algorithm for Chinese Text Clustering

Dai Yueming, *Wang Minghui*, *Zhang Ming*, *Wang Yan*
(Engineering Research Center of Internet of Things Technology Applications Ministry of Education, Jiangnan University, Wuxi 214122, China)

Abstract: In process of clustering with traditional K-means algorithm, it is difficult to identify the value of the number of clusters K and its clustering results are influenced by initial centers. It has the weakness of sensitivity to noise and instability. Meanwhile, to solve the problems for the high dimensions, sparse spatial and latent semantic structure of the text data, an algorithm for Chinese text clustering was proposed. This new algorithm uses the physical significance of Singular Value Decomposition (SVD) to firstly classify the data rough, and then uses K-means for text clustering. *It applies SVD to decompose and keep semantic features, remove noise, make smoothing process of text data, meanwhile, it takes the advantage of physical significance of SVD to have rough set classification, and then regard classification results as initial centers of K-means.* Experiment results demonstrate that the F-Measure of cluster quality has been improved compared with other K-means algorithms.

Keywords: SVD; text clustering; K-means; initial center point

引言

随着计算机的普及，互联网等技术的发展，世界逐步迈向“大数据”时代，人们接受的信息也呈现几何级数的增长[1]。这些数据大部分都以一种非结构化的形式存在，如表格、网页、文本等形式，如何从海量文本信息中获取有效的知识是当前数据挖掘研究的重点。数据挖掘建立在各种存储海量数据基础上，它致力于寻找隐藏在数据之中有用的模式[2]。聚类作为数据挖掘的重要分

收稿日期：2016-09-22 修回日期：2017-01-11；
基金项目：国家自然科学基金(61572238)，江苏省杰出青年基金(BK20160001)；
作者简介：戴月明(1964-)，男，江苏常熟，硕士，副教授，硕导，研究方向为人工智能和软件工程；王明慧(1992-)，女，黑龙江哈尔滨，硕士，研究方向数据挖掘和人工智能。

支，近年来逐渐引起广泛的重视。相对于其他数据挖掘方法，聚类是一种无监督学习，具有无须先验知识的优势，可以根据数据自然分布而获取有效知识[3]，它将数据集划分成若干簇，将相似的对象归到同一个簇，特性差异较大的对象分布到不同的簇[4]。

聚类算法有基于划分的、分层的、密度的、网格的、关联的、模型的等类型[5]。作为基于划分的聚类算法，K-means 算法具有算法简单、收敛速度快、能有效处理大数据集等多方面的优点[6]。但是 K-means 聚类算法对初始值敏感依赖，需要预先设定簇的个数 K，对噪声点敏感，容易陷入局部最优[7]。针对以上不足，许多学者做出了改进。在初始聚类中心确定方面，Shehroz 等[8]设计了一种近邻密度中心点选取方法，能避免传统 K-means 算法对聚类中心点的随机选取，在一定程度上加快算法收敛速度，但是该方法需要预设 K 值，并没有从根本上解决 K 值的设定对聚类结果造成的影响；牛琨等[9]提出了一种基于超三角的融合网格密度的选取办法，能避免预设 K 值的干扰，但是其引入的步长参数会对聚类中心的选取造成一定的影响；张建沛等[10]利用直方图法对数据样本空间进行最优划分从而确定初始聚类中心，无需预设 K 值且有较高的准确率，但是该算法在高维样本空间中会出现递归次数增多、超立方体几何运算过于复杂等问题，对于文本聚类中的高维数据样本并不适用。同时这些方法虽然有效的改善了 K-means 算法，但是它们大部分是从结构化数据集上对 K-means 算法进行改进和实验。对于非结构化数据，文本数据有着高维数，空间分布稀疏，含有大量噪声，存在潜在语义结构等问题，这些算法并没有考虑到这些特点。

在处理文本数据过程中，使用向量空间模型(Vector Space Model, VSM)将非结构化的文档表示成向量形式，使得计算机对文档进行数学处理成为可能。但是在向量空间模型中，对文档进行了正交假设，即认为词语之间的关系是相互独立的。这一假设在实际应用中是无法实现的，实际文本中出现的词条大部分对上下文具有相关性。针对这一问题，研究人员提出潜在语义索引模型，其核心即通过对“词项-文档”关联矩阵进行 SVD 分解，对文本数据去除噪声，转换到新的语义空间，从而保留词之间相关性，简化文档向量[11]。目前大部分研究只考虑了 SVD 分解的数学意义，将其应用于文本数据降维和去除噪声中，并没有考虑到 SVD 的物理意义，将其应用于粗糙分类中。

本文既应用 SVD 数学意义对文本数据进行降维和去除噪声操作，又应用其物理含义对文本进行粗糙分类，进而优化初始中心点，有效的解决了 K-means 算法因随机初始中心点带来敏感依赖性等缺点。文章设置奇异值下降速度阀值，自动确定降维维数与分类个数，解决了 K-means 算法需要人工设置 K 值的缺点。同时 SVD 对文本数据的处理，去除了噪声，对文本数据可能存在的稀疏矩阵进行了平滑处理，将非结构化的文本数据更好的转化为计算机能进行数学处理的结构化数据。

1 文本表示

1.1 文本预处理

不同于普通的数据聚类，文本聚类在聚类前需要对文本数据进行预处理，主要过程有：读取文本、文本分词、去除停用词、统计词频等。目前中文文本分词技术发展的比较成熟，国内较多的高校和科研机构对分词系统进行了研究，比如 ICTCLAS、盘古分词、CC-CEDICT、Jieba 等。本文的文本分词步骤是通过 Jieba 分词组件实现的，具体可以在 https://github.com/fxsjy/jieba 下载得到。Jieba 分词组件是目前 Python 编程领域中较好的一款中文分词开源组件，该组件可以实现中文分词，并对词性进行标注。

在对文本分词之后，文本中还存在大量语气助词、介词或者代词等，它们反复出现却没有实际意义，如“的”、“了”、“是”、“他们”等。这些词被称为停用词，它们的存在对文本的表示与聚类没有意

DOI: 10.16182/j.issn1004731x.joss.201810029

义，因此要将这些词从文档中去除。本文中将一些常用的停用词归纳在停用词词表中，然后将文本数据中出现在停用词词表中的词语去除掉。

1.2 文本向量化表示

Gerard Salton 等人提出向量空间模型 VSM，并在自然语言处理中得到了广泛的应用，该模型以文本特征项作为维度属性，并用统计模型对文本进行表示[12]。向量空间模型中每一维由文档特征项和其对应的权重组成，对应一个文档 d_j 可以表示为 $d_j=(t_{j1},w_{j1};t_{j2},w_{j2};\cdots\cdots;t_{jn},w_{jn})$，其中 n 表示特征项格式，t 表示提取的特征词，w 表示该特征词在该文档中对应的特征权重。VSM 模型构造图如图 1 所示。

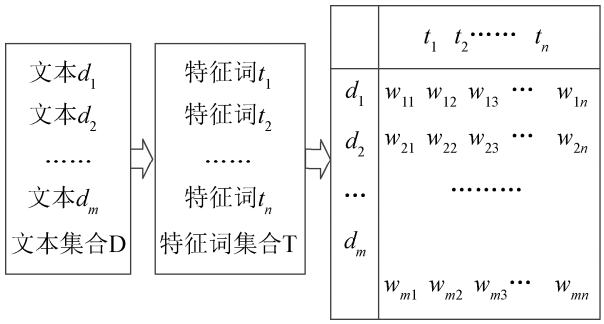

图 1 VSM 模型向量构造图
Fig. 1 Construction of VSM Model Vector

特征词的权重通过 TF-IDF (term frequency inverse document frequency, TF-IDF)表示[13]，TF-IDF 由词频(term frequency, TF)和逆文本词频(inverse document frequency, IDF)组成，可以有效的过滤掉常见的词语，保留重要的词语。其中词频 $tf(t_i,d_m)$，表示词 t_i 在文档 d_m 中出现的频率次数。逆文本词频 $IDF=\log\frac{N}{n_i}$，其中 N 表示文本集中所有文档的数量，n_i 表示文本集中所有含有特征词 t_i 的文档数量。TF-IDF 公式表示为(1)：

$$w(t_i,d_m)=tf(t_i,d_m)*\log\frac{N}{n_i} \tag{1}$$

1.3 文本相似度度量

将文本数据转化为向量后，可以通过计算向量的距离表示文本之间的相似性，本文使用文本特征向量夹角的余弦值来计算文本相似度[14]。用 $sim(d_i,d_j)$ 表示两个文本向量之间的相似度，其表示如公式(2)：

$$sim(d_i,d_j)=\cos\theta=\frac{d_i\cdot d_j}{|d_i|\times|d_j|} \tag{2}$$

两个文本向量夹角的余弦取值范围为 $[0,1]$，当两个文本特征向量正交时，$sim(d_i,d_j)$ 取值为 0，此时两个文本完全不相似；当两个向量接近平行时，$sim(d_i,d_j)$ 取值接近于 1，此时两个文本接近相同。

1.4 聚类效果评估

聚类算法在执行后会产生一个聚类结果，如何评价聚类算法的有效性就要看聚类结果，本文将基于人工分类的结果与聚类结果进行对比，并使用 F-Measure 值作为评价指标[15]。F-Measure 值是 Van Rijsbergen 在 1979 年提出的[16]，该聚类评价指标综合考虑了召回率(recall)和准确率(precision)。语料 X 上存在一个人工判定的类结构 $P=\{P_1,P_2,\cdots,P_s\}$，通过聚类算法可以得到一个待评价的聚类结果 $C=\{C_1,C_2,\cdots,C_m\}$，其中 P_i 、C_j 均为聚类簇，m 不一定等于 s。

对于任意人工主题 P_j 和聚类簇 C_i，准确率 $P(P_j,C_i)$ 为公式(3)所示，召回率 $R(P_j,C_i)$ 为公式(4)所示，F-Measure 值 $F(P_j,C_i)$ 为公式(5)所示：

$$P(P_j,C_i)=\frac{|P_j\cap C_i|}{|C_i|} \tag{3}$$

$$R(P_j,C_i)=\frac{|P_j\cap C_i|}{|P_j|} \tag{4}$$

$$F(P_j,C_i)=\frac{2P(P_j,C_i)*R(P_j,C_i)}{P(P_j,C_i)+R(P_j,C_i)} \tag{5}$$

对于每个人工主题 P_j，从中挑选最优指标值 $F(P_j)$ 及其对应的簇，以该最优的指标值来判定 P_j 的质量，如公式(6)所示：

$$F(P_j)=\max_{i=1,2\ldots m}\{F(P_j,C_i)\} \tag{6}$$

最终的 F 值为：

$$F=\frac{\sum_{j=1}^{s}(|P_j|*F(P_j))}{\sum_{j=1}^{s}|P_j|} \tag{7}$$

2 SVD 分解

2.1 SVD 数学意义

SVD 分解是线性代数中正规矩阵酉对角化的推广，能适用于任意一个矩阵[17]。对于任意矩阵 A，其 SVD 数学定义如下：

$$A=U\Sigma V^T \tag{8}$$

在式(8)中， A 是一个 $m\times n$ 的矩阵， U 一个 $m\times m$ 阶酉矩阵，Σ 是半正定 $m\times n$ 阶对角矩阵，V^T 为 V 的共轭转置，是 $n\times n$ 阶酉矩阵。Σ 矩阵为对角矩阵，如(9)所示，其对角线上的元素为 A 的奇异值，并按照从大到小的顺序排列，其中 Σ 可以由 A 唯一确定。

$$\Sigma=\begin{bmatrix}\sigma_1 & 0 & \cdots & 0 & \cdots & 0\\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots\\ 0 & 0 & \cdots & \sigma_r & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0 & \cdots & 0\end{bmatrix} \tag{9}$$

其中， $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_r\geqslant 0$。

SVD 分解有如下定理[18]：

定理 1 矩阵 A 的非零奇异值的个数等于 $Rank(A)$。

定理 2 设矩阵 A 的 SVD 分解式由式(8)给出，且 $r=Rank(A)\leqslant p=\min(m,n)$，那么 $A_k=\sum_{i=1}^{k}u_i\sigma_i v_i=U_k\Sigma_k V_k{}^T$ 表示 A 的 k 阶矩阵。其中 $U_k=(u_1,u_2,\cdots,u_k),V_k=(v_1,v_2,\cdots,v_k),\Sigma_k=diag(\sigma_1,\sigma_2,\cdots,\sigma_k)$， $\min\limits_{Rank(B)<k}(m,n)\|A-B\|_F=\|A-A_k\|_F=\sqrt{\sum_{i=k+1}^{P}\sigma_i^2}$， $\min\limits_{Rank(B)<k}(m,n)\|A-B\|_2=\|A-A_k\|_2=\sigma_{k+1}$。

该定理说明，在 Frobenius 范数下，A_k 是和 A 最接近的 k 阶矩阵。文本矩阵在 SVD 分解后，可以剔除较小的、不重要的奇异值，从而实现矩阵的降维。

定理 3 假设 $A,B\in R^{m*n}$， A 和 B 的奇异值分别为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$ 和 $\tau_1\geqslant\tau_2\geqslant\cdots\geqslant\tau_p$，其中 $p=\min(m,n)$，则有 $|\lambda_i-\tau_i|\leqslant\|A-B\|_2$。

定理 3 表明当矩阵 A 有微小的扰动时，扰动前后矩阵奇异值的变化不会大于扰动矩阵的-2 范数[17]。这说明对于文本特征矩阵中存在的近义词、同义词等噪声干扰情况时，文本特征向量在 SVD 分解后不会出现较大的变化。这一性质显示了奇异值的稳定性。

定理 4 若对于任意一个矩阵 A 的奇异值为 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_r\geqslant 0$，矩阵 $aA(a\neq 0)$ 的奇异值为 $\sigma_1^*\geqslant\sigma_2^*\geqslant\cdots\geqslant\sigma_r^*\geqslant 0$，则有：

$$|a|(\sigma_1,\sigma_2,\cdots,\sigma_r)=(\sigma_1^*,\sigma_2^*,\cdots,\sigma_r^*)$$

定理 4 说明了奇异值的比例不变性，这表示将文本矩阵进行归一化处理不会从本质上改变奇异值的相对大小[17]。

2.2 SVD 物理意义

在文本向量化表示之后，由全部文档集生成一个词_文本关联矩阵 A，一般该矩阵的维数很大，包含了许多冗余信息。将该文本矩阵按式(2-1)进行奇异值分解之后，会生成 3 个矩阵，即 U 、Σ 和 V^T。这 3 个矩阵有着非常清晰的物理含义[19]。

矩阵 U 的行向量对应矩阵 A 的词向量，是对词进行分类的结果。SVD 分解将词_文档向量映射到一个子空间中，在该子空间中，词_文档的语义关系被保留。矩阵 U 的每一行表示一个特征词 t_i，每一列表示一个语义相近的词类 C_j。对于矩阵 U 的每一行 i，该行的每个非零元素 $(u_{i1},u_{i2},\cdots,u_{im})$ 表示特征词 t_i 在各个语义类 C_j 中的权重，数值越大越相关。

矩阵 V^T 对应矩阵 A 的文档向量，它表示对文本分类的结果，它的每一列对应一个文本 d_i，每一行对应一个主题 C_j^*。对于矩阵 V^T 的每一列，该列的每个元素 $(v_{i1},v_{i2},\cdots,v_{in})^T$ 表示文本 d_i 在不同主题 C_j^* 中的相关性，数值越大越相关。矩阵 Σ 则表示词的类 C_j 和文章的类 C_j^* 之间的相关性。3 个矩阵的物理意义较为明确，关联矩阵在 SVD 分解后，就可以同时完成近义词分类和文章分类。

2.3 SVD 粗糙分类

关联矩阵 A 在 SVD 分解后，根据阀值等参数，忽略较小的奇异值，提取前 k 个奇异值，并保留对应的 U_k 和 V_k^T，此时 U_k 和 V_k^T 中元素可能会出现负数。文章[20]指出这里的负数不能表示文档的负相关性。因为在矩阵 A 中没有区分同义词和反义词的表示形式，对文档中出现的同义词和反义词的处理是相同的，所以 SVD 分解中出现的负数与负相关性无关。根据 V_k^T 矩阵的物理意义，在每个文本列 $(v_{i1},v_{i2},\cdots,v_{ik})^T$ 中，只需找到最大的权值对应的类，即可完成对该文章的分类 C_j^*，此时文本集最多分为 k 类，分类结果会略显粗糙。

对于保留的奇异值个数 k，在矩阵 Σ 中奇异值 σ_i 由大到小排列，一般前 10%的奇异值就包含了近 90%的信息[21]。奇异值的下降速度是保留奇异值个数的一个重要因素，下降速度越快，表明该奇异值变化越大，所携带的信息含量也越大。本文设置奇异值下降速度阀值 θ，当奇异值下降速度小于阀值 θ 时，选取此时的 k 值，保留前 k 个奇异值。速度阀值 θ 不能过大，否则 k 值会过小，会使有用的信息丢失，导致准确率下降；速度阀值 θ 也不能过小，否则 k 值过大，会引入较多噪音。速度阀值 θ 可以根据语料文本数据的特点，通过实验获得。

3 SVD-Kmeans 算法

3.1 算法及流程

SVD-Kmeans 算法利用 SVD 分解的物理意义对文本数据进行粗糙分类，将分类的结果作为 K-means 算法的初始聚类中心点，其基本流程可概括为如下。

算法 1：

步骤 1：文本预处理，对文本进行分词和去停用词操作；

步骤 2：计算文本中每个词的权重 TF-IDF，对每个文本中的词按照 TF-IDF 从大到小排列，取前 20 个词作为特征词构造特征集。利用特征集和文本集构造 VSM 模型，得到文本关联矩阵 A，并将其进行归一化处理；

步骤 3：对关联矩阵 A 进行 SVD 分解，得到 U、Σ 和 V^T 矩阵；

步骤 4：计算 Σ 矩阵中的奇异值下降速度，并与速度阀值 θ 进行比较，当下降速度小于阀值 θ 时，得到 k 值。

步骤 5：对矩阵 V_k^T 中每一列按照最大值对应的类别进行分类，获得每个文本的粗糙分类结果 $C_k=\{C_1,C_2,\cdots,C_k\}$；

步骤 6：计算每个类中的均值向量；

步骤 7：利用公式(1-2)计算各个文档与每个类均值向量的相似度，并记录与均值向量最为相似的类；

步骤 8：将各个文档归到最为相似的类中；

步骤 9：如果没有文档归到新的类中或者超过最大迭代次数，则停止计算，否则转到步骤 6。

3.2 算法分析

SVD-Kmeans 算法首先利用 SVD 进行分解，保留了文本数据中的语义关系，并实现了文本数据降维、去除同义词噪声干扰等功能，完成了对高维、稀疏的文本数据的平滑处理；其次利用 SVD 分解的物理意义，对 V_k^T 矩阵中每一列文本进行分类，得到粗糙分类结果 $C_k=\{C_1,C_2,\cdots,C_k\}$；并将该分类结果作为 K-means 算法的初始聚类中心，最后利用 K-means 算法得到聚类结果。

不同于其他算法在分类中需要构造分类器和有监督的训练实验，新算法步骤更为简单，无需监督检验，在文本数据 SVD 分解的同时完成了数据的去噪声、降维以及分类等一系列步骤，得到粗糙分类结果，该分类结果考虑到了语义之间的关系，满足文本数据的特征。同时考虑到 K-means 算法对初始值的敏感依赖性，避免了因随机初始化带来的不稳定性等缺点。新算法设置奇异值下降速度阀值，自动获取类别数与降维数，避免了 K-means 算法依靠人工经验确定 K 值的缺陷。新算法同时也继承了 K-means 算法简单、收敛速度快等优点。

4 仿真实验

4.1 奇异值阀值选取实验

为了确定奇异值下降速度阀值θ，进行了奇异值下降速度阀值实验。实验环境在 Python 3.5 中进行，CPU 为 Intel®Core™i5-4200U 4.9GHz，内存为 4 GB。本文实验的语料库采用搜狗实验室发布的搜狗文本分类语料库库精简版，具体可以在 http://download.labs.sogou.com 上下载得到。该语料库来源于 Sohu 新闻网站保存的大量经过编辑手工整理与分类的新闻语料与对应的分类信息，共有 9 个分类，每个分类有 1 990 个文档。在本文的实验中，选用了互联网、体育、健康、军事和教育 5 个类别进行实验，并选取不同的文档数进行实验。在文本预处理过程中，选用 Jieba 分词组件对文本实现分词操作。使用 numpy 组件中的 linalg 包的 svd 函数对关联文本矩阵实现 SVD 分解操作，并得到奇异值矩阵 Σ。计算奇异值的下降速度，并绘制折线图。如图 2~6 所示，分别是文本数为 100、200、300、400、500 的实验数据，在 SVD 分解后奇异值下降速度曲线。

从图 5~6 可看出奇异值下降速度会随着 k 值的增大呈波动下降趋势，其中速度在下降到 0.1 左右后下降趋势会逐渐趋于平稳，并且随着文本数的增多，这一趋势越加明显。因此奇异值下降速度阀值θ设置为 0.1 较为合理，且 SVD 分解对于文本数较大，维数较高的数据进行处理时会有较多的优势。

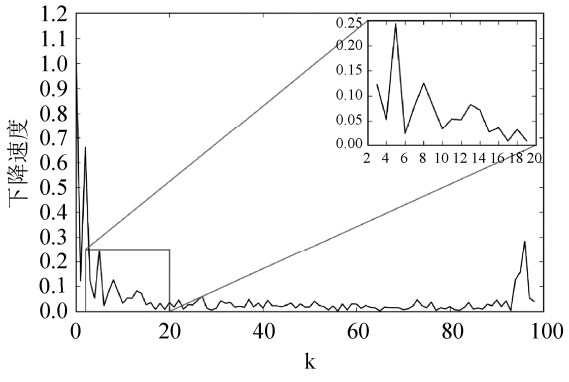

图 2 文本数 100 的奇异值下降速度曲线
Fig. 2 Singular Value Descent Speed of 100 Texts

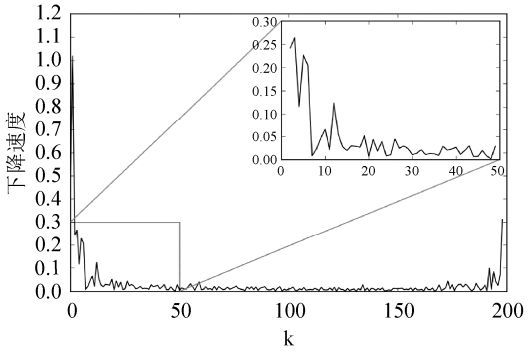

图 3 文本数 200 的奇异值下降速度曲线
Fig. 3 Singular Value Descent Speed of 200 Texts

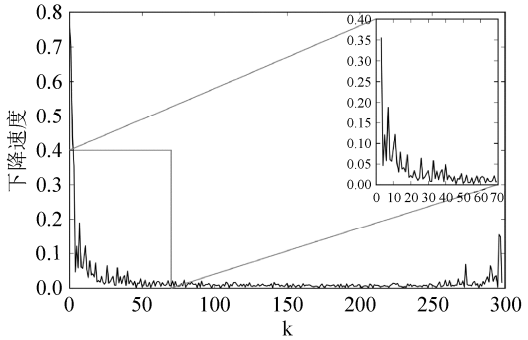

图 4 文本数 300 的奇异值下降速度曲线
Fig. 4 Singular Value Descent Speed of 300 Texts

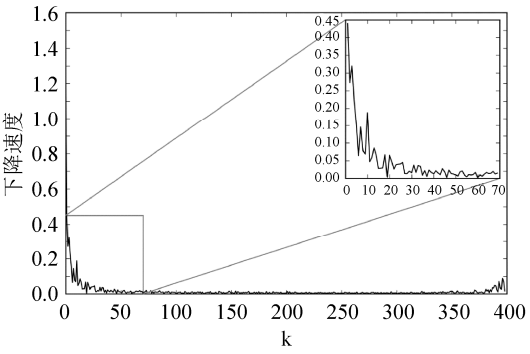

图 5 文本数 400 的奇异值下降速度曲线
Fig. 5 Singular Value Descent Speed of 400 Texts

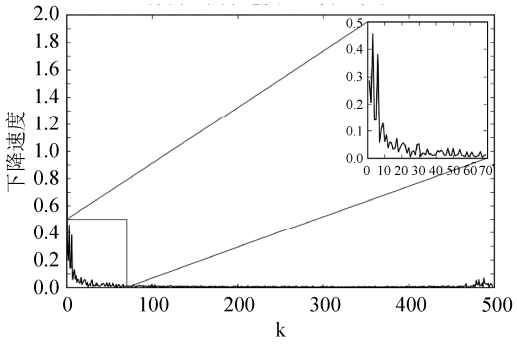

图 6 文本数 500 的奇异值下降速度曲线
Fig. 6 Singular Value Descent Speed of 500 Texts

DOI: 10.16182/j.issn1004731x.joss.201810029

4.2 文本聚类实验

在确定了奇异值下降速度阀值θ之后，需要对提出的新算法 SVD-Kmeans 的聚类效果进行实验和评估。实验环境与上一节实验环境相同，奇异值下降速度阀值$\theta=0.1$，最大迭代次数 1000，聚类效果评价指标采用 1.4 节介绍的 F-Measure 值。

实验将 SVD-Kmeans 算法与经典的随机初始聚类中心的 K-means 和 K-means++算法，以及 WLV-Kmeans 算法[22]进行对比实验，其中 WLV-Kmeans 算法利用加权局部方差对初始聚类中心点的选取进行了改进。实验选取不同的文档数进行聚类，将聚类结果进行评估并记录 F-Measure 值。其中 K-means 和 K-means++选用 sklearn 组件实现，sklearn 是一个 Python 的科学计算库，用于机器学习，可以实现多种聚类和分类算法。K-means、K-means++和 WLV-Kmeans 算法的文本预处理和构造关联向量与 SVD-Kmeans 算法相同，它们的 K 值均为语料库分类数 5，迭代次数 1000，进行 10 次实验，保留最佳聚类结果。选取文本数为 100、200、300、400 和 500 的语料进行实验，实验结果记录在表 1 中。

表 1 文本聚类 F-measure 评估结果(百分比)
Tab. 1 Assessment Results of F-measure for Text Clustering(Percentage)

算法\文本数	100	200	300	400	500
SVD-Kmeans	58.82	76.03	60.75	84	78.51
K-means	37.12	33.04	36,08	33.09	50.07
K-means++	37.34	34.13	39.7	57.06	43.3
WLV-Kmeans	48.57	48.59	56.17	59.05	68.78

从表 1 中可以看出，在相同文本数的情况下，SVD-Kmeans 算法相较于传统的 K-means、K-means++算法与优化初始聚类中心点的 WLV-Kmeans 算法，其聚类评价指标 F-Measure 均为最大值，这说明所提算法聚类质量较好，在文本聚类中有较好的聚类结果，显示出了算法的有效性和优越性。同时，随着样本文本数的增加，聚类效果也逐渐加强，在文本数为 400 时，聚类效果最好，之后虽有所下降，但评价指标仍然比其他算法好。这说明所提 SVD-Kmeans 算法对于高维数据的聚类效果较好，可以有效的解决高维数据下的文本聚类问题。

5 结论

SVD-Kmeans 算法利用 SVD 分解的物理意义进行粗糙分类，再结合 K-means 算法进行聚类，有效的解决了 K-means 算法对初始值的敏感依赖性，同时保留了文本数据中语义结构的特点，对文本数据进行去噪降维的平滑处理。实验结果表明，SVD-Kmeans 算法的聚类质量有明显的提升。

本文实验中 SVD 分解使用 numpy 组件中的 linalg 包实现，但是对于更高维的数据计算量会非常大，这一点还有待改进。在今后的实验中，可以考虑使用 MapReduce 或 OpenMP 等实现 SVD 的并行化计算。

参考文献:

[1] 翟东海, 余江, 高飞, 等. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究 [J]. 计算机应用研究, 2014, 31(3): 713-715.
Zhai Donghai, Yu Jiang, Gao Fei. K-means text clustering algorithm based on initial cluster centers selection according to maximum distance [J]. Application Research of Computers, 2014, 31(3): 713-715.

[2] Sholom M Weiss, Nitin Indurkhya, Tong Zhang. Fundamentals of predictive text mining [M]. Xi'an, China: Xi'an Jiaotong University Press, 2012: 97-103.

[3] 彭京, 杨冬青, 唐世渭, 等. 一种基于语义内积空间模型的文本聚类算法 [J]. 计算机学报, 2007, 30(8): 1354-1363.
Peng Jing, Yang Dongqing, Tang Shiwei. A Novel Text Clustering Algorithm Based on Inner Product Space Model of Semantic [J]. Chnese Journal of Computers, 2007, 30(8): 1354-1363.

[4] 邓海, 覃华, 孙欣. 一种优化初始中心的 K-means 聚类算法 [J]. 计算机技术与发展, 2013, 23(11): 42-45.
Deng Hai, Tan Hua, Sun Xin. A K-means Clustering Algorithm of Meliorated Initial Center [J]. Computer Technology and Development, 2013, 23(11): 42-45.

[5] Wong K C. A short survey on data clustering algorithms

[C]// International Conference on Soft Computing and Machine Intelligence. USA: IEEE, 2015: 64-68.

[6] 熊忠阳, 陈若田, 张玉芳. 一种有效的 K-means 聚类中心初始化方法 [J]. 计算机应用研究, 2011, 28(11): 4188-4190.
Xiong Zhongyang, Chen Ruotian, Zhang Yufang. Effective method for cluster centers's initialization in K-means clustering [J]. Application Research of Computers, 2011, 28(11): 4188-4190.

[7] 龚静, 李安民. 一种改进的 k-means 中文文本聚类算法 [J]. 湖南工业大学学报, 2008, 22(2): 52-54.
Gong Jing, Li Anmin. Clustering Algorithm of One Improved K-means Chinese Text [J]. Journal of Hunan University of Technology, 2008, 22(2): 52-54.

[8] Shehroz S Khan, Amir Ahmad. Cluster center initialization algorithm for K-Means clustering [J]. Pattern Recognition Letters (S0167-8655), 2004, 25(11): 1293-1302.

[9] 牛棍, 张舒博, 陈俊亮. 融合网格密度的聚类中心初始化方案 [J]. 北京邮电大学学报, 2005, 30(2): 7-10.
Niu Kun, Zhang Shubo, Chen Junliang. A Cell Density Enabled Schema for Initializing Cluster Centers [J]. Journal of Beijing University of Posts and Telecommunications, 2005, 30(2): 7-10.

[10] 张健沛, 杨悦, 杨静, 等. 基于最优划分的 K-Means 初始聚类中心选取算法 [J]. 系统仿真学报, 2009, 21(9): 2586-2590.
Zhang Jianpei, Yang Yue, Yang Jing. Algorithm for Initialization of K-means Clustering Center Based on Optimized-Division [J]. Journal of System Simulation (S1004-731X), 2009, 21(9): 2586-2590.

[11] 何亮亮. SVD 在文本分类中的应用 [D]. 广州: 华南理工大学, 2012.
He Liangliang. Application of the SVD in text classification [D]. Guangzhou, China: South China University of Technology, 2012.

[12] 吴夙慧, 成颖, 郑彦宁, 等. 文本聚类中文本表示和相似度计算研究综述 [J]. 情报科学, 2012, 22(4): 22-25.
Wu Suhui, Cheng Ying, Zhen Yanyu. A Review of Text Representation and Similarity Calculation in Text Clustering [J]. Information Science, 2012, 22(4): 22-25.

[13] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报, 2011, 34(5): 856-864.
Huang Chenghui, Yin Jian, Hou Fang. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method [J]. Chinese Journal of Computers, 2011, 34(5): 856-864.

[14] 林少波, 杨丹, 徐玲. 基于类别相关的新文本特征提取方法 [J]. 计算机应用研究, 2012, 29(5): 1680-1683.
Lin Shaobo, Yang Dan, Xu Ling. New Approach to Feature Selection for Text Categorization Using Class Correlation [J]. Application Research of Computers, 2012, 29(5): 1680-1683.

[15] 周昭涛. 文本聚类分析效果评价及文本表示研究 [D]. 北京: 中国科学院研究生院(计算技术研究所), 2005.
Zhou Zhaotao.Quality Evaluation of Text Clustering Results and Investigation on Text Representation [D]. Beijing, China: Graduate University of Chinese Academy of Science (Computer Software and Theory), 2005.

[16] K Van Rijsbergen. Information retrieval [M]. London, UK: Butterworths Press, 1979: 267 -301.

[17] Gu M, Demmel J, Dhillon I. LAPACK Working Note 88: Efficient Computation of the Singular Value Decomposition with Applications to Least Squares Problems [M]. USA: University of Tennessee, 1994: 68-70.

[18] 廖安平, 刘建州. 矩阵论 [M]. 湖南: 湖南大学出版社, 2005: 57-58.
Liao Anping,Liu Jianzhou.Matrix Theory [M]. Hunan, China: Hunan University Press, 2005: 57-58.

[19] 吴军. 数学之美 [M]. 北京: 人民邮电出版社, 2014: 136-141.
Wu Jun. The beauty of Mathematics [M]. Beijing, China: People Post Press, 2014: 136-141.

[20] 王怡, 盖杰, 武港山, 等. 基于潜在语义分析的中文文本层次分类技术 [J]. 计算机应用研究, 2004, 21(8): 151-154.
Wang Yi, Gai Jie, Wu Gangshan. Technology of Chinese Documents Multi-hierarchy Categorization Based on Latent Semantic Analysis [J]. Application Research of Computers, 2004, 21(8): 151-154.

[21] Golub G, Kahan W. Calculating the singular values and pseudo-inverse of matrix [J]. Siam Journal on Numerical Analysis (S1095-7170), 1965, 2(2): 205-224.

[22] 蔡宇浩, 梁永全, 樊建聪, 等. 加权局部方差优化初始簇中心的 K-means 算法 [J]. 计算机科学与探索, 2016, 10(5): 732-741.
Cai Yuhao, Liang Yongquan, Fan Jiancong. Optimizing Initial Cluster Centroids by Weighted Local Variance in K-means Algorithm [J]. Journal of Frontiers of Computer Science and Technology, 2016, 10(5): 732-741.

DOI: 10.16182/j.issn1004731x.joss.201810029