

基于改进 K-Means 算法的教学反思文本聚类研究

何聚厚^{1,2}, 范文静¹

(1. 陕西师范大学 计算机科学学院, 陕西 西安 710062;

2. 陕西师范大学 现代教学技术教育部重点实验室, 陕西 西安 710062)

摘要:对教学反思内容的准确评估是教师基于教学反思过程提升其专业能力的重要保障。基于改进的 K-Means 算法对相同主题的教学反思文本进行聚类,通过给定初始聚类中心 K 的取值范围使其可以在给定范围内自动增加,在聚类过程中加入相似度阈值以限定文本间相似度的取值范围,实现对教学反思文本的分类和对自我反思文本的定位。实验结果表明改进的 K-Means 算法在反思文本聚类的准确率和稳定性方面比传统算法有所提高,且能根据教学反思内容准确地进行自动分类。

关键词:K-Means 算法;文本聚类;教学反思;相似度;均值

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2013)11-0099-04

doi:10.3969/j.issn.1673-629X.2013.11.025

Research on Text Clustering of Teaching Reflection Based on Improved K-Means Algorithm

HE Ju-hou^{1,2}, FAN Wen-jing¹

(1. School of Computer Science, Shaanxi Normal University, Xi'an 710062, China;

2. Key Laboratory of Modern Teaching Technology of Ministry of Education, Shaanxi Normal University, Xi'an 710062, China)

Abstract: An accurate assessment of teaching reflection content is an important guarantee based on teachers' teaching reflection process to enhance their professional capabilities. Clustering the same theme of the teaching reflection text based on an improved K-Means algorithm, through given the initial cluster center K a value ranges, so that it can be automatically increased within the given range, during the clustering process, similarity threshold is introduced to limit the reflection texts' similarity ranges, realizing the teaching reflection text classification and the self-reflection text classification. The experiment result indicates the improved algorithm has a higher accuracy, better stability, and can accurately automatically classify according to the teaching reflection content.

Key words: K-Means algorithm; text clustering; teaching reflection; similarity; means

0 引言

教学反思指教师对教育教学实践的再认识、再思考,并以此来总结经验教训,进一步提高教育教学水平^[1-4]。随着网络与信息技术的快速发展,教学反思的形式也从传统的日记等形式转变为信息化平台。如 CLANDRA 等^[2]提出通过观看教学视频的方法基于网络平台帮助教师进行教学反思;Blog 由于其平台的广泛性与便利性,更是很多教师首选的教学反思形式^[3]。基于指定的反思主题,教师通过比较别人与自

己的教学反思文本,进而实现自我提升教育教学水平的目的^[1-4]。但通过网络搜索可以看出,教学反思文本数据广泛且分散,教师很难定位自己的反思文本与哪些教学反思内容进行比较。

文本聚类是一种典型的无指导机器学习方法,其目标是将文档集合分成若干聚类,要求同一聚类内容的相似度尽可能大,而不同聚类的相似度尽可能小^[5-8]。通过对反思文本的自动聚类,可以使教师更容易地查找到与自己反思内容相近的文本来进行比

收稿日期:2013-02-01

修回日期:2013-05-09

网络出版时间:2013-08-27

基金项目:中央高校基本科研业务费专项资金(GK201002028);国家985优势学科“教师教育创新平台”项目(GJ9850104)

作者简介:何聚厚(1972-),男,博士,副教授,通讯作者,研究方向为计算机网络安全、技术增强学习;范文静(1987-),女,硕士研究生,研究方向为智能信息处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130827.1432.011.html>

较。目前常用的文本聚类算法包括:以 K-Means^[9]和 K-Medoids^[10]为代表的划分法;以 AGNES^[10]和 DI-ANA^[10]为代表的层次聚类算法等。其中 K-Means 算法的时间复杂度与数据量之间呈线性关系,计算开销少,适用于大量文本集的聚类^[5-6]。

由于教学反思种类繁多、内容多样,而传统的 K-Means 算法对初始聚类中心的选取比较敏感,常常使聚类结果陷入局部最优解^[11]。为此文中通过以闭区间的形式限定初始聚类中心 K 的取值范围使 K 值可以在限定范围内自动增加以及聚类过程中文本间相似度计算的改进,从而更加准确地根据反思内容进行自动分类。

1 反思文本聚类问题描述

在反思文本聚类中,基于文本的特征词权重,将 N 篇反思文本聚集成 K 个聚类集合,其目标是使各个聚类集合中反思文本的相似度最高而聚类间反思文本的相似度最低,从而便于对反思文本进行有效的自动分类。

定义1. 每篇反思文本 d 被表示为向量空间中的一个向量 \vec{d} 。反思文本集合 D 的特征空间就是出现在该反思文本集合中所有的词的集合,表示为:

$$T_D = \{t \mid \forall t \in d, \forall d \in D\} \quad (1)$$

其中, t 为反思文本 d 的特征词。

定义2. 反思文本 d 使用 TF-IDF (Term Frequency - Inverse Document Frequency)^[12-13] 来计算每个词的权重。TF-IDF 通过下式计算:

$$\text{tfidf}(d, t) = \text{tf}(d, t) \times \log \frac{N}{\text{df}(t)} \quad (2)$$

其中, $\text{tf}(d, t)$ 是词 t 在 d 中的词频; $\text{df}(t)$ 是 D 中包含词 t 的所有文本的数目; N 是反思文本集合 D 所包含的文件总数。

定义3. 根据(2)所计算出来的特征词权重,根据空间向量模型 (Vector Space Model)^[14] 每篇反思文本表示为:

$$\vec{d}_j = [\text{tf}_1 \log(N/\text{df}_1), \text{tf}_2 \log(N/\text{df}_2), \dots, \text{tf}_N \log(N/\text{df}_N)]^T \quad (3)$$

其中, $\text{tf}_i \log(N/\text{df}_i)$ 为第 j 篇反思文本中第 i 个词的权重; tf_i 为第 i 个词在第 j 中出现的频率; df_i 为反思文本集合 D 包含第 i 个词的反思文本数目。

定义4. 使用余弦距离度量反思文本之间的相似度,定义如下式:

$$\text{Sim}(\vec{d}_1, \vec{d}_2) = \frac{\vec{d}_1 \cdot \vec{d}_2}{\|\vec{d}_1\| \times \|\vec{d}_2\|} \quad (4)$$

若 $\text{Sim}(\vec{d}_1, \vec{d}_2)$ 越大,说明反思文本 d_1 和反思文

本 d_2 的相似度越高。为了减小不同长度的文本对于计算文本相似度的影响,每个文本向量都被归一化到单位长度。设:

$$\vec{d}_j = [\text{tf}_1 \log(N/\text{df}_1), \text{tf}_2 \log(N/\text{df}_2), \dots, \text{tf}_N \log(N/\text{df}_N)]^T$$

进行归一化:

$$\vec{d}_0 = \frac{\vec{d}}{\|\vec{d}\|} = \frac{[\text{tf}_1 \log(N/\text{df}_1), \dots, \text{tf}_N \log(N/\text{df}_N)]^T}{\sqrt{(\text{tf}_1 \log(N/\text{df}_1))^2 + \dots + (\text{tf}_N \log(N/\text{df}_N))^2}}$$

可以得出 $\|\vec{d}_j\| = 1$ 。

定义5. 对反思文本集合 $D = \{d_1, d_2, \dots, d_n\}$ 经过聚类算法得到一个反思文本聚类集合 $C = \{C_1, C_2, \dots, C_K\}$, 且满足 $\cup_{j=1}^K C_j = D$, 使得 $\forall d_i (d_i \in D), \exists C_j (C_j \in C), d_i \in C_j$ 且 $C_j \cap C_l = \emptyset, j \neq l$ 。对于其中的一个反思文本聚类 C_j , 其内所有反思文本向量之和为:

$$\vec{C}_j = \sum_{d_i \in C_j} \vec{d}_i$$

归一化后得到聚类 C_j 的中心向量:

$$\vec{Z}_j = \frac{\vec{C}_j}{\|\vec{C}_j\|} \quad (5)$$

定义6. 使用 F_1 -测量值 (F_1 -Measure)^[15] 来评价聚类的质量, F_1 -Measure 综合了文本聚类过程中的查全率和查准率的结果。定义如下所示:

$$\text{precision}(i, r) = \frac{N_{ir}}{N_r} \quad (6)$$

$$\text{recall}(i, r) = \frac{N_{ir}}{N_i} \quad (7)$$

$$F_1(i, r) = \frac{2 \times \text{recall}(i, r) \times \text{precision}(i, r)}{\text{recall}(i, r) + \text{precision}(i, r)} \quad (8)$$

其中, N_{ir} 是聚类 r 中包含类别 i 的反思文本个数; N_r 是聚类 r 中实包含反思文本的总数; N_i 是预定义类别 i 应有的反思文本个数。 F_1 -测量值越高,说明聚类集合中的各个文本相似度高,聚类的效果越好。

定义7. 根据(4)计算反思文本向量 \vec{d} 与各个聚类中心向量 \vec{Z}_j 的相似度,即:

$$V_j(d) = \text{Sim}(\vec{d}, \vec{Z}_j) = \frac{\vec{d} \cdot \vec{Z}_j}{\|\vec{d}\| \times \|\vec{Z}_j\|} \quad (9)$$

对于所有的 $\vec{Z}_j (j = 1, 2, \dots, K)$, $V_j(d)$ 的值越大表示反思文本 \vec{d} 与聚类中心向量 \vec{Z}_j 越接近,若对于 $\forall \vec{Z}_j (j = 1, 2, \dots, K), \exists \vec{Z}_m (1 \leq m \leq K)$, 使 $\vec{Z}_m \geq \vec{Z}_j$, 则反思文本 d 与聚类 m 中的反思文本最相似,这样教师只需参照聚类 m 中的反思内容对自己的反思内容进行分析,进而有效实现其教学能力的目的。

2 K-Means 算法对于教学反思文本聚类过程的描述

2.1 算法流程图及描述

如图 1 所示,教学反思文本算法的关键步骤有:文本预处理、文本间相似度的计算、聚类中心的更新、选择适合的聚类进行归并。算法初期首先选择 k 个文本作为初始聚类中心,计算每个文本与这些聚类中各个反思文本的相似度, S_{\min} 记录这些相似度中的最小值, S_{\max} 记录这些相似度中的最大值。计算出 S_{\min} 和 S_{\max} 的平均值 S_{avg} 。如果 S_{avg} 大于 S ,则当前对象是孤立点,作为新的聚类中心,同时 k 加 1,如果 k 大于 K ,选择最相近的两个聚类进行合并。如果聚类中心不再发生变化,则聚类结束,否则继续进行聚类。

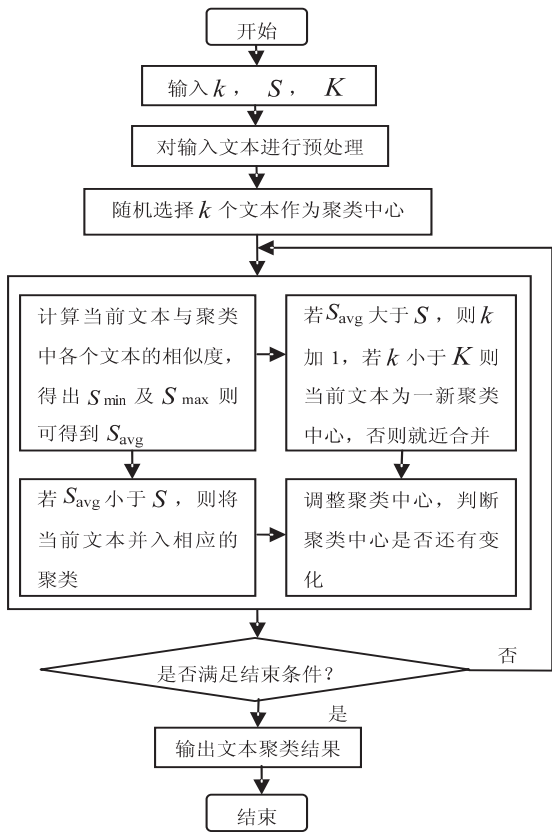


图 1 教学反思文本聚类算法流程图

2.2 算法步骤

Step1: 初始化参数。

初始聚类中心个数 k , 对象与聚类中心的最小相似度阈值 S , 聚类中心个数最大值 K 。判断 k 值的大小, k 的值小于 K 程序才能正常进行, 若 k 大于 K , 则输入不合法程序异常结束。

Step2: 文本预处理。

使用中科院的中文分词系统 ICTALAS 对待聚类教学反思文本进行中文分词, 之后与教学反思词库进行对照, 提取特征词。根据定义 2 计算所得到的特征词的权重 $\text{tfidf}(d, t)$, 各个特征词的权重就是定义 3

向量空间模型中向量的分量, 则教学反思文本就可以表示为 $\vec{d}_i = [(\text{tf}_1 \log(n/\text{df}_1), \dots, \text{tf}_n \log(n/\text{df}_n))]^T$ 。

Step3: 教学反思文本聚类。

依次计算每个对象与这些聚类中各个对象的相似度, S_{\min} 记录这些相似度中的最小值, S_{\max} 记录这些相似度中的最大值。计算出 S_{\min} 和 S_{\max} 的平均值 S_{avg} 。如果 S_{avg} 大于 S , 则当前对象是孤立点, 作为新的聚类中心, 同时 k 加 1, 如果 k 大于 K , 选择最相近的两个聚类进行合并。

伪代码如下:

```
for 每篇文本
{for 每个聚类
```

```
    { 计算当前文本  $d_i$  与该聚类中各个文本  $d_j$  的相似度, 得到最小相似度  $S_{\min}$  和最大相似度  $S_{\max}$  并计算平均值  $S_{\text{avg}}$ ;
```

```
    if (  $S_{\text{avg}} \leq S$  )
```

```
        将当前文本归入该聚类;
```

```
        调整聚类中心;
```

```
    else
```

```
         $k++$ ;
```

```
        if (  $k \leq K$  )
```

```
            该文本为一新聚类中心;
```

```
        else
```

```
            计算各个聚类中心的相似度以及该文本和各个聚类中心的相似度。若是两聚类最相似, 则进行聚类合并, 该文本为一新聚类中心, 若是该文本和某一聚类中心相似, 则将当前文本归入该聚类;
```

```
            调整聚类中心;
```

```
    }
```

```
}
```

Step4: 文本聚类效果评估。

根据定义 4 中的式(6)、(7)计算出 $F_1(i, r)$, 根据 $F_1(i, r)$ 值的大小判断聚类的优劣, 一般, $F_1(i, r)$ 值越大, 聚类内的相似度越高, 聚类效果越好。

Step5: 教学反思文本自动归类。

根据定义 7 中的式(9)计算当前反思文本与各个聚类中心的相似度, 将当前反思文本加入计算所得最大 $V_j(d)$ 值所对应的聚类, 完成对当前反思文本的自动归类。

3 实验及结果分析

3.1 参数设置

实验采用 VC6.0 实现该算法。实验数据选用从网上收集到的数据结构教学反思文本 1 150 篇, 对 $N = 1\ 000$ 篇反思文本进行聚类, 聚类后对剩余的 $n = 150$

篇反思文本进行自动归类。初始聚类中心个数 k 为 3, 相似度阈值 S 为 0.12, 聚类中心个数最大值 K 为 5。对通过改进的 K-Means 算法进行反思文本的聚类后, 聚类中心数据及其包含文本数如表 1 所示。

表 1 聚类分布 ($K=5$)

聚类	单链表	排序	队列	二叉树	栈
篇数	180	230	160	250	180

3.2 文本聚类算法性能分析

为验证文本聚类算法的准确性, 文中选取 3.1 中的两组数据分别用经典 K-Means 算法和改进的 K-Means 算法进行聚类, 对结果进行比较。

当 $N=1\,000$ 时, 经典 K-Means 算法和改进的 K-Means 算法分别进行文本聚类后的查全率和查准率对比如表 2、表 3 所示。

表 2 查全率结果对比 %

	K-Means	改进的 K-Means
单链表	81.3	89.6
排序	73.4	84.5
队列	86.8	93.2
二叉树	70.2	81.4
栈	69.3	80.6

表 3 查准率结果对比 %

	K-Means	改进的 K-Means
单链表	76.8	83.2
排序	71.6	79.8
队列	79.5	83.6
二叉树	72.3	80.5
栈	77.5	79.3

由表 3 所示的结果, 根据定义 4 中的式(7)可以计算出各个聚类的 F_1 -测量值, 计算结果如图 2 所示。

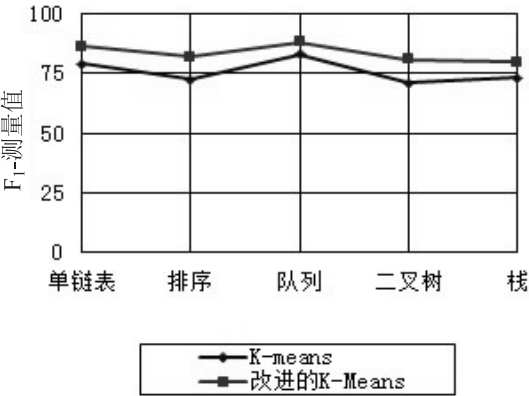


图 2 F_1 -测量值对比

3.3 反思文本自动归类

对 $N=1\,000$ 篇反思文本进行聚类后, 再对 $n=150$ 篇反思文本进行自动归类, 分别计算其中各篇反思文

本与各个聚类中心向量的相似度 $V_j(d)$ 的值, 将反思文本加入计算所得最大 $V_j(d)$ 值所对应的聚类, 之后对全部 1 150 篇反思文本重新聚类, 记录其中属于 n 的反思文本聚类结果, 对比两次聚类结果, 如图 3 所示。

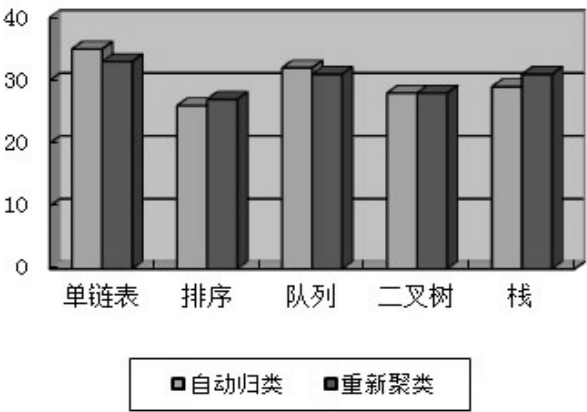


图 3 反思文本自动归类

由图 3 可以看出当所要进行自动归类的文本数远小于聚类中文本总数时, 通过计算反思文本与各个聚类中心相似度来进行归类, 既高效又准确, 且对聚类中心不会产生太大的影响。

4 结束语

文中基于种类多样的教学反思文本提出了一种改进的 K-Means 聚类方法。通过对初始聚类中心取值的优化以及聚类过程中文本间相似度计算的改进, 避免聚类结果过早陷入局部最优解, 从而提高了教学反思文本聚类质量, 教师可以通过聚类查找到与自己反思内容相近的反思文本, 经过比较, 从而更好地进行个人反思的评估以此进一步提升教育教学水平。试验中选用了数据结构教学反思文本进行测试, 实验数据表明该改进算法相对传统算法聚类结果更加准确并且能够高效地对反思文本进行自动归类。此外, 加入特定的教学反思词库, 使文本聚类在教育教学反思方面有了更好的应用。

参考文献:

[1] 申继亮, 刘加霞. 论教师的教学反思[J]. 华东师范大学学报(教育科学版), 2004, 22(3): 44-49.

[2] Calandra B, Dias L B, Lee J, et al. Using video editing to cultivate novice teachers' practice[J]. Journal of research on technology in education, 2009, 42(1): 73-94.

[3] 钟志贤, 曹东云. 基于信息技术的反思学习[J]. 远程教育杂志, 2004(4): 7-10.

[4] Kong S C, Shroff R H, Hung H K V. A web enabled video system for self reflection by student teachers using a guiding

由表中可以看出改进后,反正切函数精度大大地提高了,且反正切的值只与 x,y 的比值有关,而与向量的长度无关。而改进之前,反正切的值是与向量的长度有关的,向量长度越小,误差值越大,当 x,y 值小于1时,甚至发生严重的错误;当 x,y 为小数时,由于浮点转整形的截尾会产生误差,所以反正切函数值误差也会很大,而改进后这种误差会被大大地减弱。由于算法的迭代次数有限,所以对于极小角度,得到的反正切值误差较大,如果需提高小角度的精度,可以增加迭代次数,这样造成的结果是占用的资源变多,并且运行的周期变长,在实际应用中需要在精度和速度两者之间进行折中。

4 结束语

文中提出了 CORDIC 算法的改进方法,给出了浮点反正切函数实现的硬件结构图,并在 Quartus ii 9.0 环境下,在 Altera 公司的 FPGA Cyclone 系列 EP2C35F484C6 芯片上进行验证,其精度得到了大大的提高,运行速度快,硬件资源占用少。文中提出的改进方法已经在某型分布式导航系统和光电吊舱的实现过程中得到应用,该算法也适用于其他基于 FGPA 的系统实现中。

参考文献:

[1] Heredia G, Ollero A. Virtual sensor for failure detection, identification and recovery in the transition phase of a morphing aircraft[J]. Sensors, 2010, 10(3): 2188-2201.

[2] Volder J E. The cordic trigonometric computing technique[J]. IRE Trans on Electronic Computers, 1959, EC-8(3): 330-334.

[3] Vachhani L, Sridharan K, Meher P K. Efficient CORDIC algorithms and architectures for low area and high throughput implementation[J]. IEEE Transactions on Circuits and Systems-II: Express Briefs, 2009, 56(1): 61-65.

[4] 徐光辉, 程东旭, 黄如. 基于 FPGA 的嵌入式开发与应用[M]. 北京: 电子工业出版社, 2006.

[5] 张建斌, 梁芳, 刘乃安. 一种改进型 CORDIC 算法的 FPGA 实现[J]. 微电子学与计算机, 2010, 27(11): 181-184.

[6] 李全, 李晓环, 陈石平. 基于 CORDIC 算法高精度浮点超越函数的 FPGA 实现[J]. 电子技术应用, 2009(5): 166-170.

[7] 段文伟, 于龙洋, 李署坚. 一种改进的 CORDIC 算法及其 FPGA 实现[J]. 微电子学与计算机, 2012, 29(2): 95-98.

[8] 李美俊, 李光明. 基于嵌入式的 CORDIC 算法的改进及实现[J]. 微电子学与计算机, 2012, 29(2): 142-145.

[9] 骆艳卜, 张会生, 张斌, 等. 一种 CORDIC 算法的 FPGA 实现[J]. 计算机仿真, 2009, 26(9): 305-307.

[10] 张天瑜. 基于旋转模式的改进型 CORDIC 算法研究[J]. 微电子学与计算机, 2010, 27(3): 93-97.

[11] 辛艳, 李环. 改进型 CORDIC 算法及 FPGA 的实现[J]. 沈阳理工大学学报, 2010, 29(5): 34-37.

[12] 张俊涛, 王红仓. 基于 FPGA 的 CORDIC 算法通用 IP 核设计[J]. 微计算机信息, 2008, 24(7-3): 238-240.

+++++

(上接第 102 页)

framework[J]. Australasian journal of educational technology, 2009, 25(4): 544-558.

[5] 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.

[6] Han Jiawei, Kamber M. Data mining: concepts and techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2000.

[7] Grabmeier J, Rudolph A. Techniques of cluster algorithms in data mining[J]. Data mining and knowledge discovery, 2002, 6(4): 303-360.

[8] Jain A K, Murty M N, Flynn P J. Data clustering: a review[J]. ACM computing surveys, 1999, 31(3): 264-323.

[9] MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proc of the 5th symposium on mathematical statistics and probability. Berkeley: [s. n.], 1967: 281-297.

[10] Kaufman J, Rousseeuw P J. Finding groups in data: an intro-

duction to cluster analysis[M]. New York: John Wiley & Sons, 1990.

[11] 马帅, 王腾蛟, 唐世渭, 等. 一种基于参考点和密度的快速聚类算法[J]. 软件学报, 2003, 14(6): 1089-1095.

[12] Gospodnetic O, Hatcher E. Lucene in action2[M]. Stamford: Manning Publications Co, 2010.

[13] Salton G, Clement T Y. On the construction of effective vocabularies for information retrieval[EB/OL]. 2011-02-04. <http://dl.acm.org/citation.cfm?id=951766>.

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

[15] 刘远超, 王晓龙, 徐志明, 等. 文本聚类综述[J]. 中文信息学报, 2006, 20(3): 55-62.

基于改进K-Means算法的教学反思文本聚类研究

作者:

何聚厚, 范文静, HE Ju-hou, FAN Wen-jing

作者单位:

何聚厚, HE Ju-hou(陕西师范大学 计算机科学学院, 陕西 西安 710062; 陕西师范大学 现代教学技术教育部重点实验室, 陕西 西安 710062), 范文静, FAN Wen-jing(陕西师范大学 计算机科学学院, 陕西 西安, 710062)

刊名:

计算机技术与发展

ISTIC

英文刊名:

Computer Technology and Development

年, 卷(期):

2013(11)

本文链接: http://d.wanfangdata.com.cn/Periodical_wjz201311026.aspx