山东大学学报(理学版) 2008, 43(1) 60-64 DOI: ISSN: 1671-9352 CN: 37-1389/N



论文

# 一种用于文本聚类的改进k-means算法

索红光[1,2],王玉伟[2]

1. 北京理工大学计算机科学技术学院, 北京 100081；2. 中国石油大学计算机与通信工程学院, 山东 东营 257061

摘要：

k-means是目前常用的文本聚类算法,针对其最终搜索的局部极值与全局最优解偏差较大的缺点,采用一种基于局部搜索优化的思想来改进算法,并推导出目标函数的变化公式。根据目标函数值的改变对聚类结果作再次划分后,继续k-means迭代,拓展其搜索范围。理论分析和实验结果表明修改后的算法能有效地提高聚类的质量,且计算复杂度仍与数据集文本总数呈线性变化。

关键词： 文本聚类 k-means 向量空间模型 局部迭代

## An improved k-means algorithm for document clustering

SUO Hong-guang[1,2],WANG Yu-wei[2]

1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China;2. School of Computer & Communication Engineering,China University of Petroleum, Dongying 257061, Shandong, China

Abstract:

The k-means algorithm is a popular method for document clustering, but it often gets stuck at a local maximum far from the optimal solution. A procedure based on local search was used to improve this algorithm. The formula about object function change was also deduced, which can be used to again partition the clustering. This procedure makes appropriate iterations to enlarge the search space. Theory analysis and experimental results show that the improved algorithm efficiently improves k-means clustering and its computation is also linear in the size of document collection.

Keywords: document clustering k-means vector space model local iteration

收稿日期 1900-01-01 修回日期 1900-01-01 网络版发布日期 2006-10-24

DOI:

基金项目:

通讯作者: 索红光

作者简介:

本刊中的类似文章



扩展功能
本文信息
Supporting info
PDF(331KB)
[HTML全文](0KB)
参考文献[PDF]
参考文献
服务与反馈
把本文推荐给朋友
加入我的书架
加入引用管理器
引用本文
Email Alert
文章反馈
浏览反馈信息
本文关键词相关文章
文本聚类
k-means
向量空间模型
局部迭代
本文作者相关文章
索红光
王玉伟