计算机应用 2014, 34(1) 249-254 DOI: 10.11772/j.issn.1001-9081.2014.01.0249 ISSN: 1001-9081 CN: 51-1307/TP

人工智能

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

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

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

扩展功能

本文信息

Supporting info

PDF(932KB)

[HTML全文]

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

主题模型

K-means

聚类中心

文本聚类

隐含狄利克雷分布

本文作者相关文章

王春龙

张敬旭

PubMed

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.

Keywords: topic model K-means cluster center text clustering Latent Dirichlet Allocation (LDA)

收稿日期 2013-07-23 修回日期 2013-09-27 网络版发布日期 2014-02-14

DOI: 10.11772/j.issn.1001-9081.2014.01.0249

基金项目:

国家自然科学基金资助项目;国家电网公司科技项目

通讯作者: 王春龙

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。
作者Email: wchlong0508@126.com

参考文献:

本刊中的类似文章

1. 曹永春 蔡正琦 邵亚斌.基于K-means的改进人工蜂群聚类算法[J]. 计算机应用, 2014,34(1): 204-207
2. 江浩 陈兴蜀 杜敏.基于主题聚簇评价的论坛热点话题挖掘[J]. 计算机应用, 2013,33(11): 3071-3075
3. 史庆伟 李艳妮 郭朋亮.科技文献中作者研究兴趣动态发现[J]. 计算机应用, 2013,33(11): 3080-3083
4. 罗彪 闫维维 万亮.基于ANP和K-means聚类的客户价值分类模型及应用[J]. 计算机应用, 2013,33(10):

Article by Yu,C.L
Article by Zhang,J.X

2954-2959
5．洪留荣.无需设定阈值的图像边缘检测[J]. 计算机应用, 2013,33(08): 2330-2333
6．王丽娟 郝志峰 蔡瑞初 温雯.基于随机取样的选择性K-means聚类融合算法[J]. 计算机应用, 2013,33(07): 1969-1972
7．张利伟 苑津莎.基于智能互补策略的免疫算法[J]. 计算机应用, 2013,33(04): 953-956
8．冯汝伟 谢强 丁秋林.基于文本聚类与分布式Lucene的知识检索[J]. 计算机应用, 2013,33(01): 186-188
9．侯海霞 原民民 刘春霞.面向大文本数据集的间接谱聚类[J]. 计算机应用, 2012,32(12): 3274-3277
10．岑梓源 李彬 田联房.基于K-Means++聚类的体绘制高维传递函数设计方法[J]. 计算机应用, 2012,32(12): 3404-3407
11．王留正 何振峰.基于全局性分裂算子的进化K-means算法[J]. 计算机应用, 2012,32(11): 3005-3008
12．叶龙欢 王俊峰 高琳 袁军.复杂背景下的票据字符分割方法[J]. 计算机应用, 2012,32(11): 3198-3205
13．张妨妨 钱雪忠.改进的GK聚类算法[J]. 计算机应用, 2012,32(09): 2476-2479
14．郑丹 王潜平.K-means初始聚类中心的选择算法[J]. 计算机应用, 2012,32(08): 2186-2192
15．李劲 张华 吴浩雄 向军.基于特定领域的中文微博热点话题挖掘系统BTopicMiner[J]. 计算机应用, 2012,32(08): 2346-2349
16．张瑞丽 张继福.基于w-距离均值的模糊聚类算法[J]. 计算机应用, 2012,32(07): 1978-1982
17．杨玲 钟云飞 王彬.基于模糊规则的印刷图像专色分色研究[J]. 计算机应用, 2012,32(06): 1598-1600
18．张宜浩 金澎 孙锐.基于改进k-means算法的中文词义归纳[J]. 计算机应用, 2012,32(05): 1332-1334
19．昌燕 张仕斌.基于加权直觉模糊集合的聚类模型[J]. 计算机应用, 2012,32(04): 1070-1073
20．谢娟英 郭文娟 谢维信 高新波.基于样本空间分布密度的改进次胜者受罚竞争学习算法[J]. 计算机应用, 2012,32(03): 638-642
21．刘培奇 孙捷焓.基于LDA主题模型的标签传递算法[J]. 计算机应用, 2012,32(02): 403-410
22．原福永 张晓彩 罗思标.基于信息熵的精确属性赋权K-means聚类算法[J]. 计算机应用, 2011,31(06): 1675-1677
23．张燕平,张娟,何成刚,褚维翠,张利娜.基于佳点集与Leader方法的改进K-means聚类算法[J]. 计算机应用, 2011,31(05): 1359-1362
24．张玉芳 朱俊 熊忠阳.改进的概率潜在语义分析下的

文本聚类算法[J]. 计算机应用, 2011,31(03): 674-676
25. 张新伦 苏一丹 惠刚刚.核K-Means聚类在Folksonomy标签模糊和冗余中的应用[J]. 计算机应用, 2011,31(03): 680-682
26. 傅德胜 周辰.基于密度的改进K均值算法及实现[J]. 计算机应用, 2011,31(02): 432-434
27. 周世兵 徐振源 唐旭清.K-means算法最佳聚类数确