基于改进 K-means 算法的海量数据分析技术研究

李　欢，刘　锋，朱二周

(安徽大学 计算机科学与技术学院，安徽 合肥 230601)

摘　要： 针对海量数据难处理的难题，利用 Hadoop 平台下的 Map-Reduce 模型，实施了一种改进的对海量数据进行并行处理的 K-means 算法.为了解决传统的 K-means 算法对初始聚类中心和聚类数敏感的问题，改进算法首先对海量数据进行多次采样，找出采样数据的聚类个数；其次，利用密度法找出采样数据的聚类中心；最后，将各个样本中心点归并得到原始数据的全局初始聚类中心点.通过在 Hadoop 集群上部署的实验结果表明，改进后的算法相比较于传统的算法具有高效、准确、可扩展以及良好的加速比等特性.

关键词： Map-Reduce；K-means；并行挖掘

Research of an Impoved K-means Algorithm for Aanalyzing Mass Data

LI Huan，LIU Feng，ZHU Er-zhou

(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract： Aiming at solving the problem of mass data, this paper proposes an improved K-means algorithm for processing massive data by making use of the Map-Reduce model on the Hadoop platform. In order to solve the problem that faced by traditional K-means algorithm, such as it is sensitive to initial clustering center and clustering number, the improved algorithm firstly finds out the clustering number from sampling data by implementing multiple sampling of massive data; Secondly, with the help of density method the clustering center of data sampling is founded. Finally, the global initial clustering centers of original data are obtained by merging the central points of each sample. The results of the experiments deployed on the Hadoop cluster have shown that the improved algorithm is more efficient, accurate, scalable and has better acceleration ratio than the traditional algorithms.

Key words： Map-Reduce；K-means；parallel mining

作者简介：

李　欢　男，(1992-)，硕士研究生.研究方向为大数据与云计算.E-mail:lh1219508753@qq.com.

刘　锋　男，(1962-)，博士，教授，硕士生导师.研究方向为并行计算与云计算.

朱二周　男，(1981-)，博士，副教授，硕士生导师.研究方向为虚拟化与程序分析.

〖FL)〗