基于 MapReduce 的 CTK 加权聚类改进算法

王万良，胡禹
(浙江工业大学 计算机科学与技术学院，浙江 杭州 311023)

摘要： 本文提出了一种基于 MapReduce 的分布式聚类改进算法，该算法将聚类分两阶段进行，首先改进了 Canopy 算法，引入梯度变化来确定初始中心点以及最佳簇数，解决了传统算法对初始值的依赖性以及对聚类个数的不确定性.设置了区域半径并动态改变，避免了聚类过程中的局部最优，并采用了信息熵加权，解决了相似度计算的特征权重的问题.最终结合 MapReduce 分布式计算模型，设计了算法的并行策略与方案.试验结果表明该算法在准确性、加速比、扩展性上具有良好的性能。

关键词： 大数据;聚类算法;Canopy 算法;MapReduce;

Improved CTK Weighting Clustering Algorithm Based on MapReduce

WANG Wan-liang，HU Yu
(College of Computer Science & Technology， Zhejiang University of Technology，
Hangzhou 311023， China)

Abstract： This paper introduces an improved algorithm of Distributed Clustering Based on MapReduce， the process of clustering will be divided into two stages， firstly， introduce Canopy algorithm， find out the suitable K of clustering algorithm by the change of Gradient value. That reduce the number of iterations and avoid the uncertainty of initial center point results in. Then dynamically change the radius of the region to solve the problem of similarity of high-dimensional data sets and solve the problem of characteristic weight of similarity calculation with the weighting of information entropy. Finally， the parallel strategy and scheme of the algorithm are designed according to the MapReduce distributed computing model. Experimental results show that the proposed algorithm has good performance in accuracy， speedup and scalability.

Key words： Big data， Cluster， Canopy algorithm， MapReduce

作者简介：
王万良男，（1957-），博士，教授.研究方向为智能调度等.
胡禹（通讯作者）男，（1992-），硕士研究生.研究方向为聚类算法.E-mail:huyuchina@163.com.