基于抽样融合改进的大数据聚类方法

刘　岩 1，王存睿 2

(1 大连民族大学 网络与信息技术中心， 辽宁 大连 116600；

2 大连民族大学 计算机科学与工程技术学院， 辽宁 大连 116600）

摘　要：校园网络大数据集的有效挖掘以提高信息的使用价值，对校园网络优化有着极其深远的影响，为此，本文提出一种基于 leaders 算法的校园网络大数据聚类改进算法 leaders-k-means 算法，算法首先通过 leaders 算法对校园网大数据集进行初始聚类，并根据初始聚类中心对校园网络大数据进行多次随机抽样形成多个小样本集，然后利用初始聚类中心做为初始值对每个小样本集进行 k-means 聚类，既保证了 k-means 算法初始值设置的合理性，又使得算法在一个较小的样本集中聚类，提高效率，最后对聚类后的多样本集合并，利用自下而上的层次聚类方法重新聚类获得原始样本的聚类中心.算法融合了层次方法、划分方法以及密度方法的优势，通过对比实验验证，算法取得较好的聚类效果.

关键词：校园网络优化；大数据聚类；leaders 算法；多样本集聚类融合

An Improved Big Data Clustering Method Based on Sampling Fusion

LIU Yan 1，WANG Cun-rui 2

（1 Network and Information Technology Center，Dalian Minzu University,Dalian 116600, China；

2 School of Computer Science & Engineering,Dalian Minzu University,Dalian 116600, China)

Abstract：Effective mining of large data sets of campus network has been a very far-reaching impact on campus network optimization. So, in this paper, an improved large data clustering algorithm, named Leaders-k-means, was presented. In this method, the former Leaders algorithm is used to obtain initial cluster centers firstly and a number of small sample sets are formed on the basis of those centetrs by random sampling of the large data of the campus network, and then,the initial clustering center is used further as the initial value to carry out K-means clustering for each small sample set,which not only ensures the rationality of the initial value of K-means algorithm, but also makes the algorithm running in a small sample set improving the efficiency of the algorithm, and at last, these small sample sets which have been clustered by k-means method are combined into a larger sample set and the bottom-up hierarchical clustering method is used to obtain the final cluster centers of the original big data set. The proposed algorithm combines the advantages of hierarchical method, partition method and density method. The simulation results show further that the proposed algorithm has good clustering results.

Key words：campus network optimization; big data clustering; the leaders algorithm; multi-sample sets clustering

作者简介：

刘　岩　女，（1976-），硕士，工程师.研究方向为网络管理及计算机应用.E-mail:l_y_2016@sina.com.

王存睿　男，（1980-），硕士，讲师.研究方向为软件与理论。