计算机应用研究
Application Research Of Computers

基于内码序值聚类的相似重复记录检测方法*

Approach for detecting approximately duplicate records based on cluster of inner code's sequence value

中文关键词：相似重复记录 内码序值 聚类 等级法

英文关键词：approximately duplicate records inner code' s sequence value cluster rank method

基金项目：国家火炬计划资助项目(2004EB33006[0]);江苏省高校自然科学指导性计划资助项目(05JKD520050)

作者	单位
鲁均云， 李星毅， 施化吉， 马素琴	（江苏大学 计算机科学与通信工程学院， 江苏 镇江 212013）

中文摘要：

检测和消除相似重复记录是数据清理和提高数据质量要解决的关键问题之一，针对相似重复记录问题，提出了基于内码序值聚类的相似重复记录检测方法。该方法先选择关键字段或字段某些位，根据字符的内码序值，利用聚类思想将大数据集聚集成多个小数据集；然后，通过等级法计算各字段的权值，并将其应用在相似重复记录的检测算法中；最后，在各个小数据集中检测和消除相似重复记录。为避免关键字选择不当而造成记录漏查问题，采用多趟检测方法进行多次检测。通过实验表明，该方法具有较好的检测精度和时间效率，能很好地应用到中英文字符集，通用性很强，

英文摘要：

Detecting and eliminating approximately duplicated records is one of main problems needed to solve for data cleaning and improving data quality. As to such problem, this paper presented an approach for detecting approximately duplicate records based on cluster of inner code' s sequence value. The proposed method firstly chose the key field or some bits of it, and according to the inner code' s sequence value of character, clustered large datasets into many small datasets by cluster thought. Then in term of rank-based weights method, endowed each attribute with certain weight using in detecting approximately duplicate records. Finally, detected approximately duplicated records and eliminated in each small dataset. To avoid missing some records caused by choosing improper key field, the multiple-detecting method could be adopted. Experimental results show the proposed method has good detection precision and time efficiency, can be applied to English and Chinese character set, and therefore is an effective approach to solve approximately duplicate records for massive data.