基于 R-树索引的高维相似重复记录检测改进算法

宋国兴 1，2，3 ，周　喜 1，3，马　博 1，3，赵　凡 1，3
（1 中国科学院 新疆理化技术研究所，新疆 乌鲁木齐 830011；2 中国科学院大学，
北京 100049；3 新疆民族语音语言信息处理实验室， 新疆 乌鲁木齐 830011）

摘　要：经典的相似重复记录检测算法 SNM 算法随着记录维度的增加，投影过程不仅会导致数据丢失，算法的误差率也会明显增大.针对 SNM 算法的不足，提出 DRR 算法，利用 R-树构建索引保留记录的高维空间特性，通过聚类减少记录在叶子节点中的比较次数提高效率，同时改进度量记录相似性的距离算法，避免高维数据稀疏性的影响.最后，通过真实数据在不同维度上分别与 SNM 算法进行对比，验证了算法的有效性.

关键词：SNM 算法；R-树索引；高维空间特性；改进距离算法；数据稀疏性

Research on High Dimensional Similarity Duplicate Record Detection Algorithm Based on R- tree Index

SONG Guo-xing 1，2，3 ，ZHOU Xi 1，3，MA Bo 1，3，ZHAO Fan 1，3
（1 Xinjiang Institute of Physical and Chemical Technology， Chinese Academy of Sciences，Urumqi 830011，China；2 University of the Chinese Academy of Sciences，Beijing 100049, China；3 Xinjiang Laboratory of Minority Speech and Language Information Processing，Urumqi 830011, China）

Abstract：The classic similar duplicate record detection algorithm SNM, With the increase of the recording dimension， the process of projecting can not only lead to the loss of data， but also the error rate of the algorithm will increase obviously. Aiming at the deficiency of SNM algorithm，using R- tree to construct index maintains the high dimension space characteristic of records. By clustering， the times of records comparing was reduced， so that the efficiency was improved. In order to avoid the influence of high dimensional data scarcity， an improved distance algorithm for measuring record similarity is proposed. Finally， the validity of the algorithm is verified by comparing the real data with the SNM algorithm in different dimensions.

Key words：SNM algorithm; R- tree index; high dimensional space characteristics; improved distance algorithm;data scarcity

作者简介：
宋国兴　男，（1989-），硕士研究生.研究方向为大数据分析、数据挖掘.E-mail:sgx805560893@163.com.
周　喜　男，（1978-），博士，研究员.研究方向为物联网应用技术、大数据分析.
马　博　男，（1984-），博士，副研究员.研究方向为数据分析与知识发现、机器学习.
赵　凡　男，（1980-），博士研究生，副研究员.研究方向为信息安全、大数据分析.