# Mining information for intelligent transformation

An information-mining laboratory in southern China is attracting top researchers whose **INNOVATIVE MODELS AND THEORIES** lead to disruptive technologies.

**Efficiently and securely** extracting useful information from massive data requires new methods and technologies. These are explored at Guangxi Key Laboratory of Multi-Source Information Mining and Security, in the School of Computer Science and Technology (SCST) of Guangxi Normal University.

The laboratory director, Shichao Zhang has led a team to dig deep into 'non-deterministic polynomial time problems,' which are fundamental computability theories. They have developed matrix computation to represent uncertain reasoning, which can be transformed into computational formula. They have also proposed strategies to deal with temporal reasoning, using matrices to represent time interval relations. This matrix approach is also used in k-Nearest Neighbour (k-NN) classification, an algorithm often used for classifying data points separated into several groups, to obtain the k value and the nearest neighbouring points from test data.

k-NN is an easily applicable algorithm used in data mining. However, this classification has traditionally been a trial-and-error process. Zhang's team presented a new approach called shell neighbour imputation (SNI), which fills in missing values in a given dataset by using only its left and right neighbours, outperforming k-NN in classification accuracy. They also proposed two more k-NN strategies, K*Tree classification and one-step computation, which reduce costs, improve classification performance, and allow for use on big data.

A useful divide-and-conquer approach to dealing with big data is mining by blocks.

Zhang has proposed methods for multi-source mining for local pattern analysis of heterogeneous data to transform the massive data problem into a weighted fusion algorithm, mining strategies for effective maintenance of dynamic data that measure bias for predictive models, and extracting infrequent patterns from non-repetitive data for association rule learning.

Targeting high storage costs, and inefficient knowledge discovery from high-dimensional big data, Zhang's colleague, Xiaofeng Zhu, has proposed a dimensionality reduction method based on sparse modelling. This has improved feature selection in computer-aided diagnostic systems based on neuroimaging.

Zhu's group also developed a linear hashing strategy for high-dimensional big data retrieval. This links to their research on patterns and mechanisms for use in populating missing data in sets.

Zhenjun Tang focuses on image hashing, which can be used for processing large-scale image data, for example, finding duplicates, to support efficient image retrieval. His group proposed robust feature extraction methods based on ring partition to solve feature representation of rotated images. Their discovery of the image features of dimensionality reduction paves the way for similarity calculations. Their feature compression with invariant vector distance enables efficient coding and compact representation for image classification.

"These results demonstrate the rapid development of this laboratory," says Xianxian Li, dean of SCST. With many innovative research teams and platforms, along with researchers selected for national-level talent programmes, the laboratory publishes more than 100 academic papers every year, and produces a series of patents, software copyrights, and technologies for applications. "We welcome more talented minds to join us, driving research innovation together," he says. ■

www.ci.gxnu.edu.cn



The laboratory hosts many academic events, including this conference on information retrieval.



New data extraction methods are leading to diverse applications from medical diagnosis to image processing.

MF3D/PORTUGAL/GETTY