Bootstrapped Authorship Attribution in Compression Space

Ramon de Graaf
Leiden Institute of Advanced Computer Science

Cor Veenman
Digital Technology and Biometrics Department

PAN Authorship Attribution Problem

- Multi-class statistical pattern recognition problem
  - Proper feature representation

- Dataset properties
  - Very few training document samples
  - Low number of authors
  - Large documents

- Performance measure
  - Average precision, recall, and F1 score over all authors

Approach

- Low dimensional feature representation
  - Compression Distances to Prototypes (CDP)
    - > Compression distance measure (CDM)
    - > Compressor: Prediction by Partial Matching (PPM)

- Prototypes required to compute distance to
  - Draw one from each training document without replacement

- To learn a statistical model, more samples required
  - Bootstrapping from the large training document