**Normalised Compression Distance for Clone Detection**

The Normalised Compression Distance (NCD) is a method that can be used to measure similarity between two given data objects, such data object could be files or search terms. The two selected data objects can be represented as data vectors, where these vectors are subjected to a compression algorithm. The compression algorithm computes a compressed length of each data vector and both data vectors concatenated. For example one might utilise Gzip as a compression method for the given data objects. The higher the compression value of the two concatenated data vectors relative to how each compressed data vector indicates that the more similar they are. The NCD between two data objects is capture more formally as:

Where is the compression method used. It was demonstrated that the NCD methods is generally applicable, parameter free, noise resistant and demonstrated theoretically to be optimal [1]. Furthermore NCD can be used with machine learning to induce object classification, where the classification can be achieved by utilising the computed NCD values. Such classification using NCD was shown to be effective by Stefan Axelsson [1]. Therefore it could be possible to apply this technique for clone detection in source code as the process of clone detection focuses on fragments within a source file that are similar.

Clone detection in source code heavily relies on a solid definition of what a clone is however, a single definition of what a clone is has yet to be universally agreed upon and thus clones are categorised with types. The clone types are as follows:

* Type 1 clones have identical code segments except for layout and comments.
* Type 2 clones are structurally identically with differences in identifies names, layout and comments.
* Type 3 clones are similar to type 2 but with variation in the addition, removal and modification of statements.
* Type 4 clones focus on the semantically equivalence of source files.

By using machine learning techniques it could be possible to have a training set that contains files that are labelled with the above clone types. For each one of the clone types it is then possible to calculate the NCD values and apply classification algorithms to partition the training set according to clone type. After the training phase, the new pairs of source file could then undergo similar procedures such that they could be classified as one the clone types or none. As this technique would only examine the information distance between the source file, it would be difficult to find at a granular i.e. the source level, where the duplication/clone actually occurs. However, at a higher level it would be recommended to apply such technique.

**References**

1. USING NORMALIZED COMPRESSION DISTANCE FOR CLASSIFYING FILE FRAGMENT, Stefan Axelsson, Blekinge Institute of Technology, 2010 International Conference on Availability, Reliability and Security.