# Evaluation of Semantic Metadata Pair Modelling Using Data Stream Clustering

Hiba Khalid, Esteban Zimanyi

Universite Libre de Bruxelles

**Abstract** One of the most expensive data tasks is cleaning noisy or messy data. Each dataset comes with specific and general pieces of information. This information could be both relevant or irrelevant to other datasets. To comprehend the connection between two disintegrated datasets; a middleware is required. Metadata presents such medium for connection, elaboration, examination and comprehension of relativity between two datasets. Metadata $\boldsymbol{MD_i}$ can be enriched to calculate the existence of connection $\boldsymbol{C_{(di,dk)}}$ between different disintegrated datasets. In order to do so, the very first task is to attain a generic metadata representation for domains. This representation narrows down the metadata search space $\boldsymbol{S_i}$. The metadata search space consists of attributes, tags, semantic content, annotations etc. to perform classification. The existing technologies limit the metadata bandwidth i.e. the operation set for matching purposes is restricted or limited. This research focuses on generating a mapper function called 'cognate' $\boldsymbol{CO_r}$ that can find mathematical relevance based on pairs of attributes between disintegrated datasets. Each pair is designed from one of the datasets under consideration using the existing metadata and available tags. After pairs $\boldsymbol{P(vd_i), (vd_k)}$ have been generated, samples are constructed using different combination of pairs. The similarity and relevance between two or more pairs is attained by using data stream clustering technique to generate large groups from smaller groups based on similarity index. The search space $\boldsymbol{S_i}$. Is divided using a domain divider function and smaller search spaces are created using relativity and tagging as main concept. For this research the initial datasets have been limited to textual information. Once all disjoint meta-collection(X1,…,Xn) have been generated the approximation algorithm calculates the centers of each meta-set. These centers serve the purpose of meta-pointers i.e. a collection of meta-domain representations. Each pointer can then join a cluster based on the content i.e. meta-content. All centers are then pooled through a domain channel and linear pointers are sent across meta-pointers. These linear pointers can then bind or leave a reference tag to the visited meta-node $Vmn_j$. Each visit is them recorded as a graph communication. The model designed facilitates the use of existing metadata to derive meta-operations and provide evidence of connection or disintegration between domains. It also facilitates the process of possible synonyms across cross-functional domains such as sports and food datasets still might share common attributes, people and details. This can be examined using meta-pointers and graph pools.

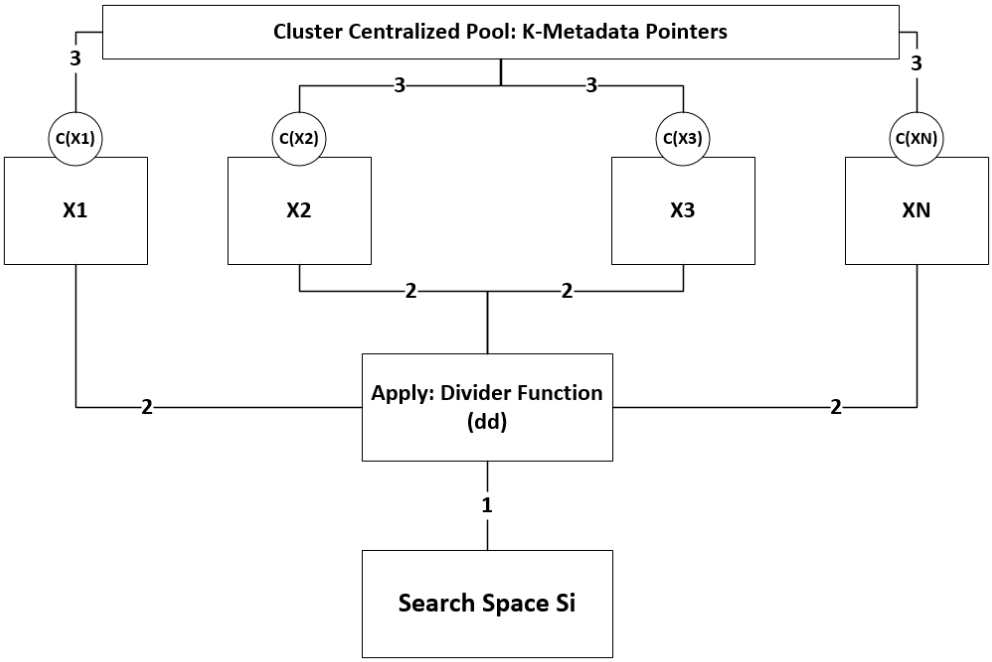


**Figure-I: Data Stream Approximation Clustering for Metadata Pointer Pairs**