SEMANTIC BASED APPROACH FOR ENTITY MATCHING ON NOISY SEMI STRUCTURED DATA

Nikhil Acharya

Supervisor- Diego Collarana

Entity Matching refers to matching data instances referring to same real world entities. The task is very relevant if the data come from rdf data sources and that is the input to our pipeline. For entity matching we use Deep Matcher as a parent algorithm. Deep matcher is designed to handle structured, textual or noisy data. However if there are semantic conflicts it performs poorly. The rdf datasets have many conflicts which we need to define for either of the datasets. Also we have labelled the pairs of entities through a semi automated method. Eventually we transform the data from semi structured to structured and design it such that it gives the best results for the matching task using the Deep Matcher Algorithm.

The design consists of a pipeline where we define 2 rdf datasets which can follow 2 different ontologies. The next stage involves resolving the conflicts which were defined beforehand. Important to note that the conflicts might be very specific to either of the datasets. Next we use semi automated labeller to label the pair of entities. Now we can merge the entities that are similar and also identify dissimilar entities. For the rest of the pairs we assume we are not sure of its similarity. Now we can merge 2 rdf datasets. The last stage of the pipeline involves data transformation to structured 2 dim datasets. We use BFS algorithm to parse the nodes of knowledge graph to fetch values. Once the data is ready we pass it to deep matcher and test for best results for Entity Matching by modifying its hyper parameters.