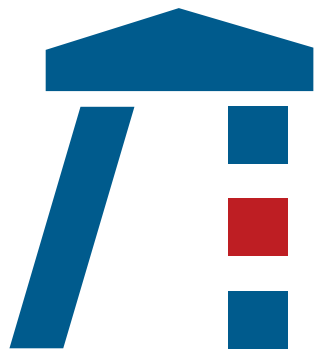
Learning and improving data partitioning for distributed stream joins

Constantin Roudsarabi



Department of Computer Science
Technical University of Kaiserslautern

Problem Overview

- Problem: Joining Json documents arriving on data streams over multiple machines
- Challenge: Similar documents have to be sent to the same machine for joining
- Json documents are not labeled -> unsupervised learning required

General Approach

- Need to find a good representation of the similarity of the Json documents
- Use a clustering algorithm to classify them as different groups of documents
- partition the different clusters over multiple machines using Reinforcement Learning

Clustering Approach

- Count the number of co-occurrences of different attribute-value pairs within documents
- Metric: The higher the number of co-occurrences in all documents the smaller the distance between attribute-value pairs
- cluster the attribute-value pairs with k-means and the distance metric (clusters are the equivalent to association groups in the base approach)
- a JSON document is assigned to all clusters where it has matching attribute-value pairs within