Contribution ID: **237** Type: **Sectional**

Blocking strategies to accelerate record matching for Big Data integration

Thursday, 3 October 2019 12:15 (15 minutes)

Record matching represents a key step in Big Data analysis, especially important to leverage dis-parate large data sources. Methods of probabilistic record linkage provide a good framework to estimate and interpret partial record matches. However, they require combining string distances for the compared records. That is, direct use of probabilistic record linkage requires processing the Cartesian product of record sets.
A "blocking"step is often used where candidate record pairs are required to match exactly on a categorical column, greatly limiting the number of record comparisons and computational cost. However, this method requires a level of data quality and agreement between sources on the cat-egorical column. We propose a more flexible approach for situations where no good blocking col-umn can be chosen.

The key idea is to use approximate nearest neighbor search as the blocking filter. One possible method is to vectorize one string column with TF or TF/IDF into term frequency vectors, then use Location Sensitive Hashing to quickly search for approximate nearest neighbors in this vector space. Apache Spark libraries were used to show the effectiveness of this approach for linking open company registration datasets.

Primary author: Mr KADOCHNIKOV, Ivan (JINR, PRUE)

Co-author: VLADIMIR, Papoyan (JINR)

Presenter: Mr KADOCHNIKOV, Ivan (JINR, PRUE)

Session Classification: Machine Learning Algorithms and Big Data Analytics

Track Classification: Machine Learning Algorithms and Big Data Analytics