Actively Learning Blocking Scheme for Entity Resolution

Jingyu Shao and Qing Wang

Research School of Computer Science
The Australian National University
{jingyu.shao,qing.wang}@anu.edu.au

Entity resolution refers to the process of identifying records which represent the same real-world entity from one or more datasets. Blocking is an important part of entity resolution. It aims to improve the time efficiency of entity resolution by grouping potentially matched records into the same block. Both supervised and unsupervised approaches have been proposed in the past; Nonetheless, they still have some limitations: (1) Supervised blocking scheme learning approaches require a large number of labels, but it is an expensive task to obtain labels for entity resolution; (2) Existing unsupervised blocking scheme learning approaches, generate training sets based on the similarity of record pairs, instead of their true labels, thus the blocking quality can not be guaranteed.

In this lightning talk, we will present a blocking scheme learning approach based on active learning techniques. In our approach, we actively learn a blocking scheme using two strategies: (1) Active sampling strategy aims to tackle with the class imbalance problem for training sample selection; (2) Active branching aims to reduce the number of candidate blocking schemes instead of enumerating all blocking schemes. With a limited label budget, our approach can efficiently learn a blocking scheme which can generate high quality blocks.

We have conducted experiments over four real-world datasets *Cora*, *DBLP-Scholar*, *DBLP-ACM* and *North Carolina Voter Registration* with class imbalance ratio ranging from 1:49 to 1:31,440. We evaluate our proposed approach in terms of: *blocking quality* (including reduction ratio, pairs completeness, pairs quality and F-measure), *blocking efficiency* (including the number of record pairs generated by different approaches) and *labelling cost*. The experimental results show that our proposed approach outperforms several baseline approaches and yields high quality blocks within a specified error bound and a limited budget of labels.