SparkER: Scaling Entity Resolution in Spark

Luca Gagliardelli  
University of Modena  
and Reggio Emilia, Italy  
luca.gagliardelli@unimore.it

Giovanni Simonini  
MIT CSAIL  
Cambridge, MA, USA  
giovanni@csail.mit.edu

Domenico Beneventano  
University of Modena  
and Reggio Emilia, Italy  
domenico.beneventano@unimore.it

Sonia Bergamaschi  
University of Modena  
and Reggio Emilia, Italy  
sonia.bergamaschi@unimore.it

Entity Resolution (ER)

Entity Resolution is the task of identifying different representations (profiles) that pertain to the same real-world entity.

Application areas:
Data Integration, Duplicate Detection, Fraud Detection, etc.

Hard to Scale:
all pairwise comparisons among profiles has complexity \(O(n^2)\)

SparkER

Implements our new broadcast meta-blocking method by using Apache Spark.
Provides both schema-agnostic and loosely schema-aware blocking techniques [4].
Covers the main ER tasks.

Scalable ER Techniques

Blocking and meta-blocking are used to reduce the number of comparisons. Parallel meta-blocking ER techniques [3] to manage Big Data were proposed: all of them are based on the repartition join (not efficient as generates a large materialized graph).

Broadcast meta-blocking

A novel optimized meta-blocking parallelization strategy that exploits broadcast join to avoid the graph materialization.

Experimental Results

Datasets characteristics

| Dataset | \(|P_1|\) | \(|P_2|\) | \(|D_E|\) |
|---------|--------|--------|--------|
| Citations | 1.8M | 2.5M | 0.6M |
| DBpedia  | 1.2M | 2.2M | 0.9M |
| Freebase | 4.2M | 3.7M | 1.5M |

Node characteristics

- CPU: 2 x Intel Xeon E5-2670v2 2.5 GHz (20 cores)
- RAM: 128 GB
- OS: Ubuntu 14.04

Broadcast meta-blocking always outperforms repartition meta-blocking: on 10 nodes it is 5 to 12 times faster.

Conclusions

Broadcast meta-blocking
- A novel meta-blocking parallelization strategy;
- Outperforms the state-of-the-art parallel implementations in term of execution time and resources consumption.

SparkER
- A highly scalable tool for Entity Resolution with Apache Spark;
- Implements both schema-agnostic and loosely schema-aware blocking techniques;
- Implements the main ER tasks.

References