Entity Resolution (ER): 50+ year old Artificial Intelligence problem

Problem: How do we automatically resolve pairs of entities that occur in different syntactic and structural contexts, but refer to the same underlying entity?

Resolving such entity pairs in large, cross-domain knowledge bases has emerged as a Big Data problem.

Target Application: Data Integration

The Emerald architecture

API exposing Entity Name System

In order to meet ER needs...

...we must populate an Entity Name System (ENS) over knowledge bases describing Linked Data.

Research Question and Thesis Statement

Can current state-of-the-art ER solutions address needs of Big Data integration?

Thesis statement

ER for Big Data integration requires resolving four challenges simultaneously:

- Automation
- Scalability
- Heterogeneity
- High Quality

Heterogeneity Challenge 1: Single-class schema heterogeneity

Heterogeneity Challenge 2: Cross-domain heterogeneity

Scalability

A feasible system needs to work with distributed sources, scales near-linearly and is preferably implementable in the cloud. Proposed parallel framework: MapReduce

Current Status of Work

Automation: Generate a training set automatically using inexpensive heuristics, and make learning algorithms robust to potential noise

Single-class schema heterogeneity: Perform property matching between datasets using the automatically generated training set

Future Work: This system will be implemented in MapReduce and be extended to accommodate multiple classes and domains (cross-domain heterogeneity).

Experimental Results

Automation: Noisy duplicates generation

Quality: Impact of single iteration on performance

References


Acknowledgement: We would like to thank the National Science Foundation (NSF) for supporting the research and this trip.

Updated August 11, 2015