Query Driven Data Labeling with Experts: Why Pay Twice?

Eyal Dushkin **Shay Gershtein** **Tova Milo** **Slava Novgorodov***

Motivation

- Large scale systems involve ML methods (e.g. for search queries classification), that rely on data annotated by experts.
- Experts are an expensive resource in terms of monetary costs and latency, hence interaction with them is minimized.
- We present an efficient labeling algorithm that constructs a minimal set of classifiers (and minimizes the need of labeled data).

Model

- **Data & Queries:** The model consists of a relation R and a set of attributes, that is queried by a given set of queries.
- **Binary Classifier:** Missing values can be discovered by a binary classifier, based on labeled data generated by experts.
- **Graph Model:** A set of queries modeled as an undirected graph. The vertices are the query values and the edges connect values appearing in the same query.

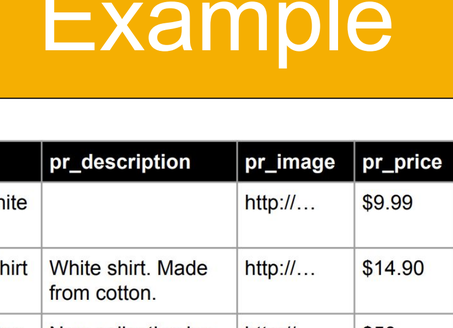
Example

Shirts

pr_id	pr_title	pr_description	pr_image	pr_price	color	material
P17892	Linen White Shirt		http://…	$9.99		
P42947	Cotton Shirt (White)	White shirt. Made from cotton.	http://…	$14.90		
P68203	Red Cotton Shirt (D&G)	New collection by D&G	http://…	$50		
P31415	Umbro Black Shirt	Perfect cotton sport shirt by Umbro	http://…	$39.99		
P86229	Linen Shirt	Material: Linen, Color: Blue	http://…	$25		

Queries:
SELECT * **FROM** `Shirts` **WHERE** `color` = ‘Red’ **AND** `material` = ‘Cotton’
SELECT * **FROM** `Shirts` **WHERE** `color` = ‘Black’ **AND** `material` = ‘Cotton’
SELECT * **FROM** `Shirts` **WHERE** `color` = ‘White’ **AND** `material` = ‘Polyester’
SELECT * **FROM** `Shirts` **WHERE** `color` = ‘Red’ **AND** `material` = ‘Linen’
SELECT * **FROM** `Shirts` **WHERE** `color` = ‘White’ **AND** `material` = ‘Linen’

Graph Representation:

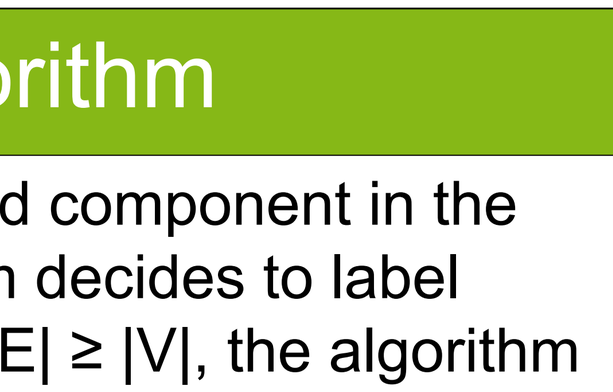


Algorithm

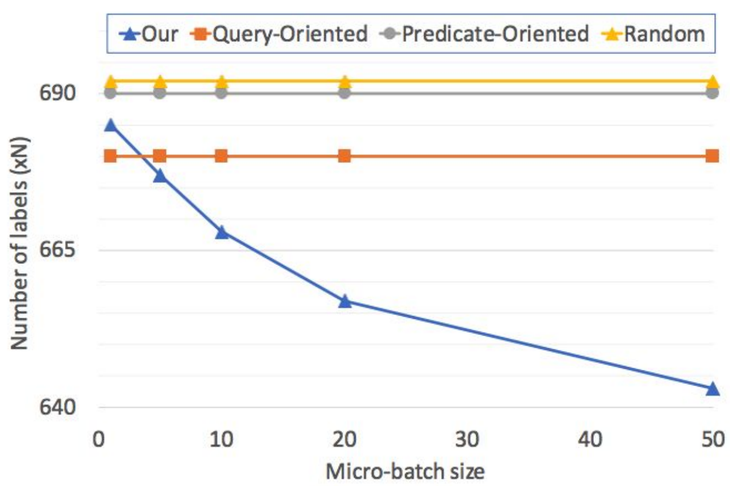
- For every connected component in the graph, the algorithm decides to label edges or nodes. If |E| ≥ |V|, the algorithm labels nodes and otherwise edges.
- Since the algorithm works per connected component, it labels edges if and only if |E| = |V| − 1 (it’s a tree).
- Different scenarios:

Best case ratio:

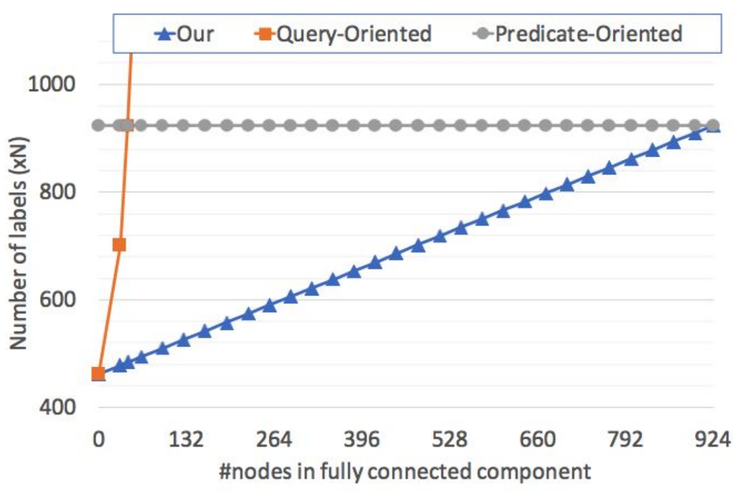
x2



(a) Offline results.



(b) Online results.



(c) Effect of queries dispersion.