Two-phase Parallel Learning to Identify Similar Strutures among Relational Databases

Debora Gomes dos Reis, Rommel Novaes Carvalho, Ricardo Silva Carvalho, and Marcelo Ladeira
Department of Computer Science at University of Brasilia (UnB) Brasilia DF Brazil 70910-900, and Ministry of Social Development of Brazil 70054-906

Motivation

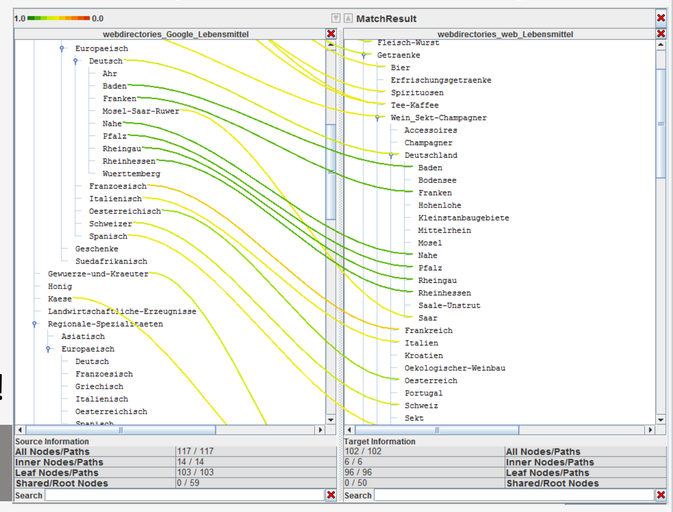
The Ministry of Social Development of Brazil maintains hundreds of **large databases.** Some of them are similar, which leads to:

- Wrong analysis
- The complexity of management
- The high cost of services
- The high cost of hardware

Identify similarities between large datasets are time-consuming and error-prone process.

We need more efficient techniques!

Example of Manual Matche's System



Previous Work

- Manual matches;
- Required all data dictionary to perform the matches;
- Depend on frequent human classification to continue perform the matches;
- Do not analyze large datasets.

Purpose

Apply **data mining** techniques to **classify similar schemas** based on its structure. **Parallel** and **Sequential** of:

- Generalized Linear Model (GLM)
- Random Forest (RF)
- Gradient Boost Machines (GBM)

Metrics:

- Precision
- Recall
- F-measure

Goals

- Automatic schema matches
- Less dependency on data dictionary
- Reduce overload of human work
- Analyze large datasets

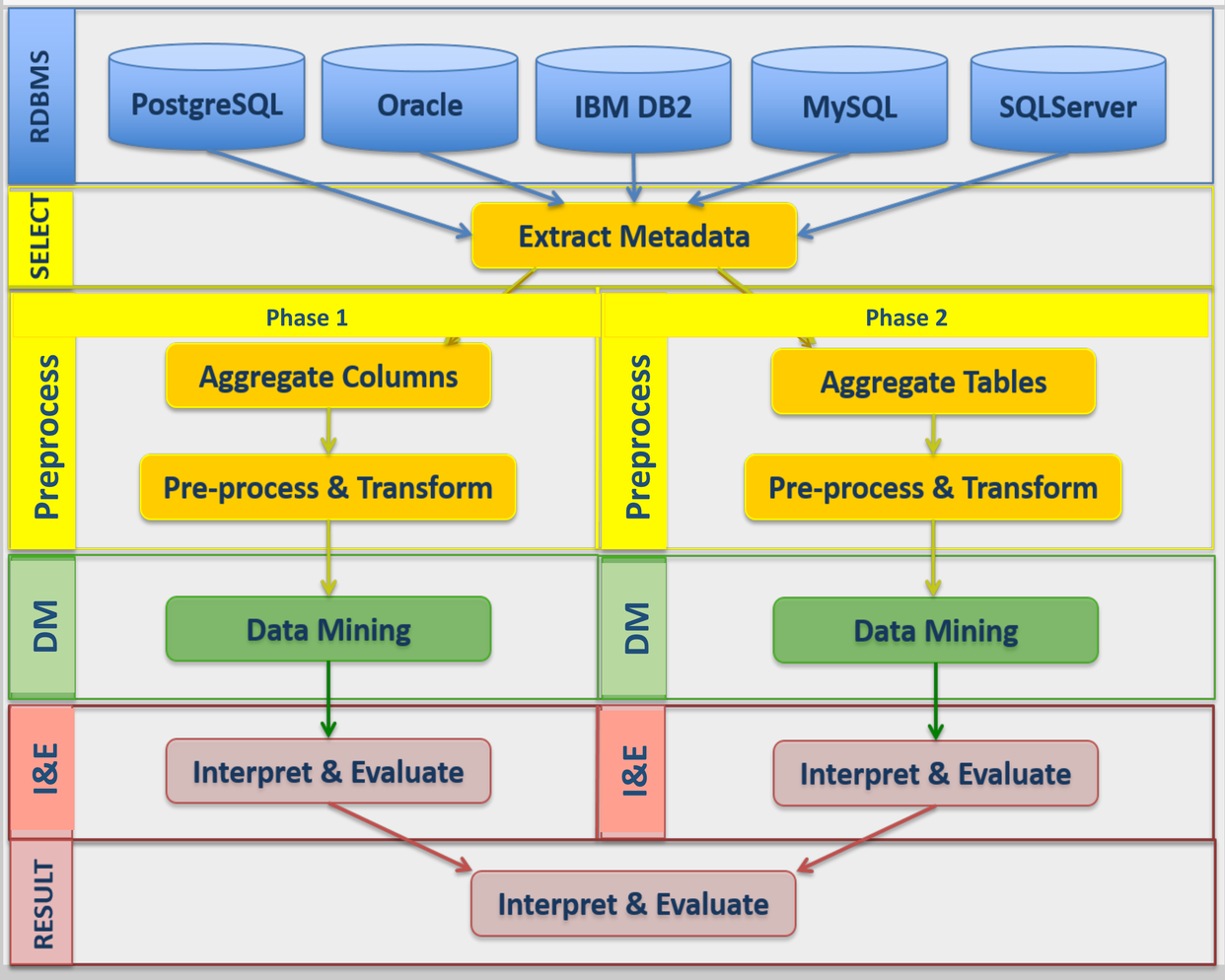
Which relational database have a similar structure?

Two-phase Parallel Learning

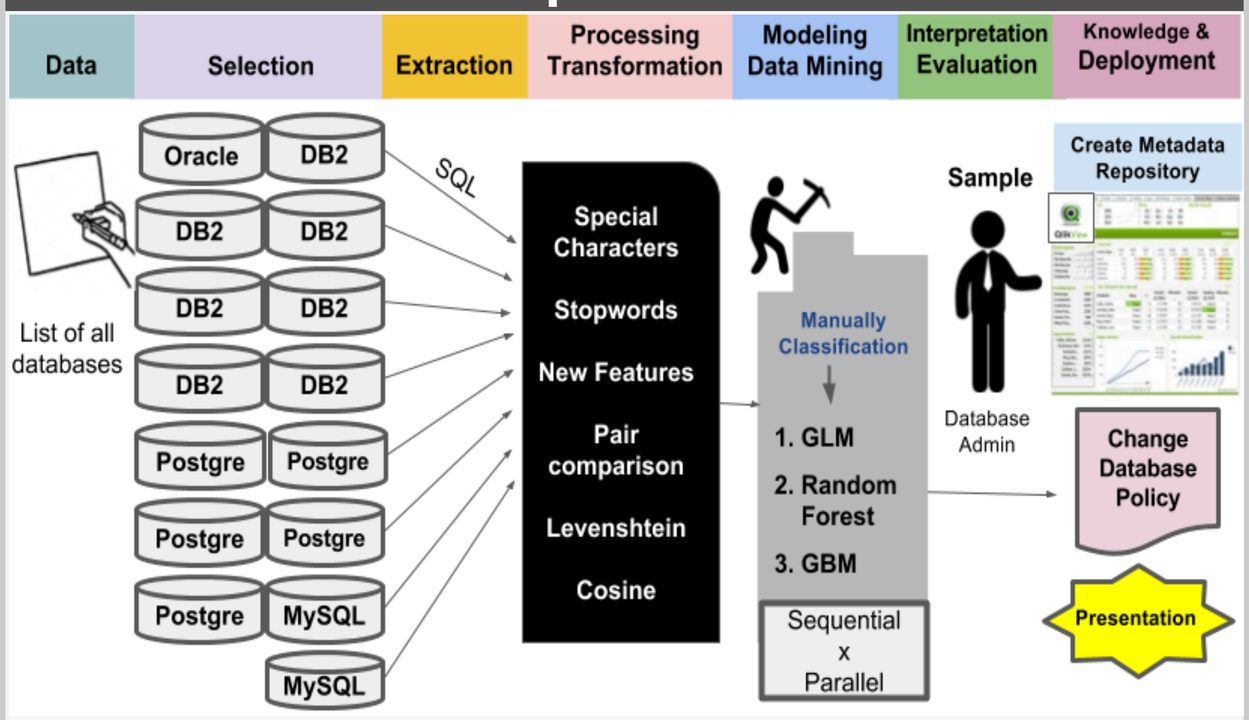
A new approach to identify similar structures of relational databases fast. The extraction of metadata is the same for any RDBMSs, which brings **flexibility**. Each phase performs the steps of **KDD** as methodology, where:

- **Phase 1:** identify similar columns
- **Phase 2:** identify similar tables

The Data Mining (DM) steps brings **parallelism.**



Experiment



Example of metadata extraction:

servidor	schema	banco	tabelas	tamanho_GB	qtd_linhas	qtd_colunas
supghm01	adesan	adesan	rltipoperfilmenu, adesaomunicipal, anexoadesaomunic...	0.0005874634	6228	115
supghm02	adesan	adesan	rltipoperfilmenu, adesaomunicipal, anexoadesaomunic...	0.0005111694	6069	115

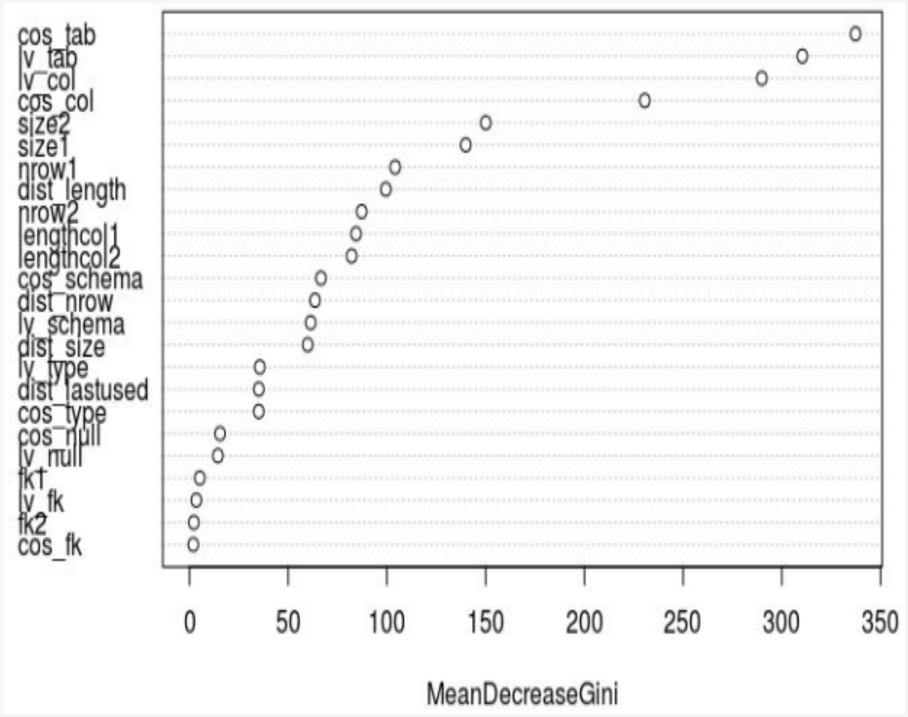
Paired comparison from 13 large schemas:
Sums the total of 78 schemas comparisons *(calculate by 13*13-13/2 = 78)*

Groups:

- 60% training
- 20% validation
- 20% test

Hardware & Software

64-bit, 8 cores
32GB RAM
Ubuntu Server
R 3.3.1, Rstudio
1.0.143 H2O
3.10.4.6



Importance of Variables

Classes are unbalanced.

- Undersampling with 10-folds cross-validation repeating 3 times.
- Balanced classes, validated using test dataset.

Results:

- Parallel processing had 1.0 of Precision, Recall, and F-measure.
- Duration was a decisive factor in choosing the best algorithm.
- The parallel execution of GBM took 3 mins and was at least 10 times faster than the sequential processing, which took 40 mins.

Conclusion

- ✓ Created the two-phase parallel learning approach to schema matching.
- ✓ Validate it by an experiment that classified similar datasets structures, using GLM, RF, and GBM in parallel and sequential processing mode.
- ✓ The GBM in parallel mode was the faster and better than others.
- ✓ The final result shows **35% of similar structures.**

Future Work

- Apply these techniques to all Ministry's datasets
- Compare the results with first-order approach

More Info