A KDD PROCESS TO RETRIEVE AND AGGREGATE DATA FROM RELATIONAL DATABASES

Jérôme Dantan
LAMSAD Laboratory, Esitpa-APCA / CEDRIC Laboratory, CNAM
3 rue du Tronquet CS 40118 76134 Mont St Aignan Cedex, France / 292 rue Saint-Martin 75003 Paris, France

Yann Pollet
CEDRIC Laboratory, CNAM
292 rue Saint-Martin 75003 Paris, France

Salima Taibi
LAMSAD Laboratory, Esitpa-APCA
3 rue du Tronquet CS 40118 76134 Mont St Aignan Cedex, France

ABSTRACT

Relational databases are a standard for representing data models. SQL is the most widely used language for querying such databases. Consequently, in many research domains, scientists extract data from relational databases, compute them and do statistical treatments. But they have to deal with the complexity of relational databases models. In addition, it takes a long time for the scientists to manually retrieve and compute data. That's why we propose a system which automatically does. It contains the following layers: parameterizable extraction of data, automatic process of SQL queries, data aggregation, statistical parameters computation, writing the results to tables and final data processing by the scientist, thanks to a statistical analysis software. A use case on the research on a soil quality index from a large relational database will be presented.

KEYWORDS

Information extraction, Ontologies, Relational database, Semantics-based parametrization, Soil quality index, Structured Query Language.

1. INTRODUCTION

In many research domains, scientists have to deal with relational databases because of their advantages: easy insertion, easy modification, extensibility, research... When performing statistical treatments, researchers have to cope with the complexity of relational databases models. Indeed, relevant data are scattered in many different tables and fields and, consequently, executing SQL requests one by one is not an efficient way to retrieve relevant data, apply aggregation operators and put data in a format which is supported by statistical analysis software. In addition, it takes a long time for the scientists to manually retrieve and compute complex data every time the database is updated. That's why researchers need a generic architecture which automatically extracts and computes data from relational databases. In addition, the system must be easily customizable, thanks to user parametrization based on the ontology of the considered domain.

2. ARCHITECTURE OF THE SYSTEM

We have designed such an architecture that allows easily customizable extraction of data thanks to dynamic semantic-based mapping between SQL queries and OWL ontologies, that provide 'representation adequacy' for the user and 'inference efficiency' for computers (Christophe Cruz and Christophe Nicolle, 2011), *via* a graphical user interface. It also contains a java code repository dynamically updated by the user with

aggregation operators, statistical operators, indicators and so on, automatic generation and execution of SQL queries, computed data writing to tables and final data processing by the scientist, thanks to a statistical analysis software, as illustrated in figure 1.

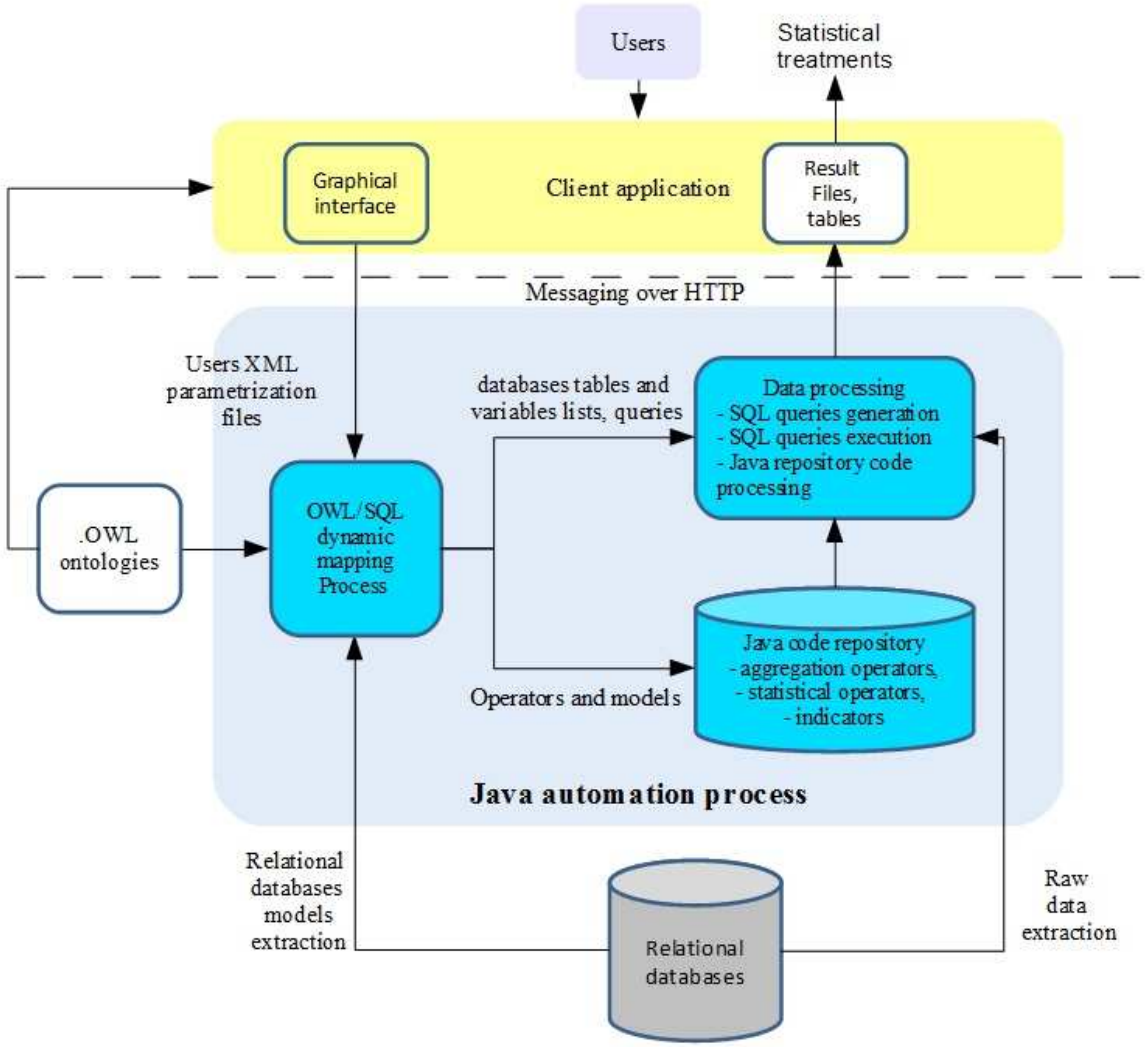

Figure 1. Software architecture

Complex data and indicators (e.g. median values or specific indicators; Joe Celko, 2010, Chris Date, 2003) are very complex to compute with SQL only. So, it was necessary to integate processes combining SQL requests with *ad hoc* Java code, that can run on any Java Virtual Machine regardless of the computer architecture. Furthermore, because of the flexibility due to ontologies and semantics-based parametrization prompted by the user, our system is independent from any relational database model.

3. USE CASE

For example, we have to calculate heavy metal rates to build a soil quality index from a very complex relational database. We obtain the comma-separated values file tables, as illustrated in table 1.

	Metal #1 (mg/m²)	Metal #2 (mg/m²)	Metal #3 (mg/m²)	Metal #4 (mg/m²)
Place #1 average	90.95	16.95	44.05	0.68
Place #1 standard deviation	71.10	8.10	27.9	0.48
Place #1 median	82.65	18.45	42.3	0.67
Place #2 average	57.90	22.90	25.67	37.80
Place #2 standard deviation	76.47	189.27	41.35	6880.00
Place #2 median	67.34	27.68	34.96	11.89
Place #3 average	52.33	139.87	25.71	2221.50
Place #3 standard deviation	51.84	28.67	36.29	930.00
Place #3 median	56.83	22.78	27.64	11.20

Table 1. Extract of a result table

4. CONCLUSION

The data extracted contain: statistical parameters such as average, median, standard deviation and soil quality indexes prompted by the user. These data will be useful to be then treated thanks to statistical software, for example to model aggregated soil quality indexes based on metal rates (Denis Baize, 1997).

Our approach relies on the combination of two ideas: (1) dynamic mapping between the real world (ontologies) and SQL queries and (2) dynamic mapping between Java repository codes and SQL queries, prompted by the user through a client application providing XML parametrization files.

REFERENCES

Denis Baize, 1997. *Teneurs totales en éléments métalliques dans les sols (France)*. INRA Editions, Nancy, France.

Howard Beck et al, 2010. Ontology-based simulation in agricultural systems modeling. *Agricultural Systems*, Vol. 103, pp 463-477.

Stefania Castellani et al, 2011. A knowledge-based system to support legal case construction. *Proceedings of Knowledge Engineering and Ontology Development conference.* Paris, France, pp. 15-27.

Joe Celko, 2010. *Joe Celko's SQL for Smarties, Fourth Edition: Advanced SQL Programming.* Morgan Kaufmann, San Francisco, USA.

Christophe Cruz and Christophe Nicolle, 2011. A graph-based tool for the translation of xml data to OWL-DL ontologies. *Proceedings of Knowledge Engineering and Ontology Development conference.* Paris, France, pp. 361-364.

Chris Date, 2003. *An Introduction to Database Systems (8th Edition)*. Addison Wesley, Boston, Massachusetts.

Quynh-Nhu Numi Tran, Graham Low, 2008. MOBMAS: A methodology for ontology-based multi-agent systems development. *Information and Software Technology,* Vol. 50, pp 697-722.

Naïma Souad Ougouti et al, 2011. Architecture of MEDPEER - A New P2P-based System for Integration of Heterogeneous Data Sources. *Proceedings of Knowledge Management and Information Sharing conference.* Paris, France, pp. 351-354.