Integration of data sources heterogeneous in warehouses data

Sara Djebrit[1]

March 4, 2025

Abstract

Data integration systems help combine information from different sources into a single format. This research looks at the challenge of merging these diverse data sources, which have different meanings and locations. The goal is to create integrated systems that convert various formats into one global format and effectively process queries to retrieve information. The report discusses an earlier study and the implementation of a mediator using a local view approach, focusing on optimizing queries and evaluating the accuracy of algorithms, achieving a general accuracy value of 70 %

Introduction

In recent times, computer science requires ways to access, process, and integrate data from different sources. Various integration systems have been developed, including TSIMMIS from Stanford, Picsel from Paris Sud University, and MOMIS from universities in Modena and Milan. Data integration aims to combine different formats into a single interface, but issues arise due to data heterogeneity, affecting integration and query manipulation. This work focuses on integrating heterogeneous sources in data warehouses and uses the LAV approach for mediation. It presents chapters on definitions, query evaluation, historical studies, and creating an integration system with matching algorithms and ETL processes, concluding with performance assessments.

Methodologies

Computer grid:Data reconciliation

Grille informative combines various data sources for administrative functions. Data integration involves matching similarities among sources without common information. Comparison of databases occurs in two steps: comparing common fields and analyzing the results for decision-making. Empirical methods measure field similarity directly. Advanced methods include string comparison techniques like LCS and Jaro-Winkler, while phonetic algorithms like Soundex are based on pronunciation.

BDBO(Database of Ontological Base)

The integration system combines various data sources into a single view, called a data warehouse. A complex issue is the automatic interpretation of the meaning of heterogeneous data, leading to conflicts. Ontologies play a crucial role in integration systems.

The BDBOs integration system defines operations based on ontologies extracted from data sources. It aims to create a shared ontology to facilitate combinations. The text presents an integration scenario based on ontology, outlines operations, and describes three integration scenarios linked to BDBO composition operators.

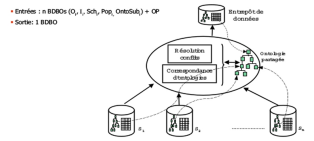

Figure 1: *general plan of BDBO*

VISS(Virtual Integration Support System

the VISS architecture improves data integration by optimizing existing approaches. VISS uses a mediator to analyze queries and identify matching data sources. It has a platform called metadata that accesses these sources and imports the requested information.

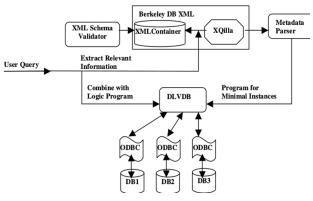

Figure 2: *VISS Architecture*

Construction of LAV approach

The LAV approach involves combining different data sources into a unified global schema. The goal is to

create a data integration system that offers a comprehensive view of all data sources. This section focuses on implementing the LAV approach, considering all components: the mediator and the wrapper, including algorithms for data matching and the ETL processes (Extract, Transform, Load). Finally, we analyze and assess the resulting global schema using performance calculations and a confusion matrix.

Wrapper

Introduction The wrapper acts as an intermediary component between data sources and the global schema.

Key Points

- It extracts data from the provided sources and stores it for querying within the global schema.
- Four platforms respresent 4 types of data (XML,HTML,JSON,SQL) are created, each representing the wrappers for the four proposed sources.
- These platforms perform the ETL process, specifically the extraction function, which retrieves attributes and values.
- The platforms utilize Hash Map types.
- The extraction function is implemented in four different ways based on the structure of the sources.

Approximation methods The Wrinkler-Jarro algorithm is one of the best matching algorithms. It measures the similarity between two sequences and gives higher percentages if they have the same meaning. We implement this algorithm in comparisons between attributes of different data sources, considering all types of conflicts.

Algorithm 4 Algorithme de Wrinkler-Jarro

$S_1 : chainesdecaracteres.$
$S_2 : chainesdecaracteres.$
$C_1 : chainedecaracteres;$
$C_2 : chainedecaracteres;$
$intt \leftarrow 0;$
$C_1 \leftarrow caracterecomunes(S_1, S_2)$
$C_2 \leftarrow caracterecomunes(S_2, S_1)$
$N \leftarrow lalongueurde C_1$

for $i \leftarrow 0\ N$ **do**
  if $C_1[i] \neq C_2[i]$ **then** $t \leftarrow t + 0.5$

$$réelR \leftarrow \frac{N}{S_1 \cdot longueur} + \frac{C_2 \cdot longueur()}{S_2 \cdot longueur} + \frac{\frac{C_1 \cdot longueur() - t}{C_2 \cdot longueur}}{3}$$

Figure 3

Results

The data experiments among the algorithms carried out give very important studies such as:

- the matching techniques: Wrinkler-Jaro algorithm with the performance of **70%** taking into account all types of data, as it gives the best values for similarity distances;
- the comparison with the data dictionary gives maximum performance according to the algo rithme of Wrinkler-jaro as it is **82.9%** and almost found most simulated values;
- the overall performance of our algorithm based on these two case studies We also study the execution time of our fusion algorithm among the effective machine to validate the results. We observe the time change in some numbers of data, such as by measuring tanque the fusion algorithm arrives the execution we have the following values

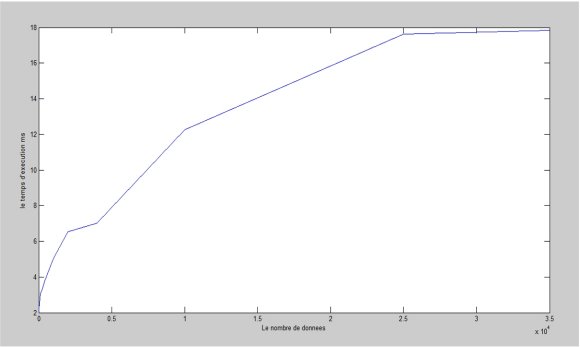

Figure 4: *Performance of Algorithm*

Conclusion

The project explores systems for integrating different data sources, focusing on creating a unified schema and efficient querying. It addresses issues like semantic and structural differences, discusses key concepts like wrappers and data mapping types, and includes query processing features. A mediation used the LAV approach with four data formats, applying algorithms for better data merging. Future work aims to improve structures for large data and enhance the VISS architecture.