Managing Semantic Evolutions in Semi-Structured Data

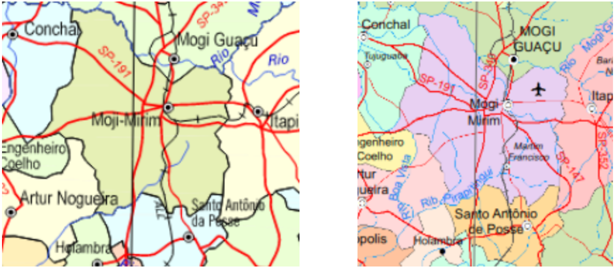
Pedro Ivo Siqueira Nepomuceno and Kelly Rosa Braghetto

DEXA - Database and Expert Systems Applications - 2023

Table of Contents

What is Semantic Evolution?

- Data semantic interpretation is not immutable. Meanings and grouping levels of long data series often change over time.
- Since 1992, for example, 86 Brazilian counties have changed names.



(a) Moji Mirim county in a 2009 IBGE map.

(b) Mogi Mirim county in a 2015 map.

Figure: Maps from the same region of São Paulo state.

Another Example

Date	GE Market Cap	GEHC Market Cap
2023-01-03	$92,854.94	N/A
2023-01-04	$76,705.30	$27,505.41

Table: Closing market cap of General Electric and General Electric HealthCare. In December 2022, GEHC was announced as a spin-off from GE common shares. Therefore, from the start of GEHC trading, part of the previous market cap of GE has been separated into GEHC company shares.

Query Impact

Although the query itself is not hard to rewrite, the results do not come ready. In the first example, they must be processed in order to rewrite “Moji Mirim” to “Mogi Mirim” if it is the same county in São Paulo for records before 2015. This way, proper aggregation can be executed.

$Q =$ {“County”:“Moji Mirim”}
    or ({“County”:“Mogi Mirim”} and {“Date”:{ $gte: “2015-01-01”}})

Objectives

- Create a storage model to annotate and keep track of semantic heterogeneity in semi-structured data
- Create algorithms that offer users a convenient way to query semantic heterogeneous data
- Evaluate performance results of developed models and algorithms

Table of Contents

1 Introduction
- What is Semantic Evolution?
- Query Impact

2 Theoretical Framework
- Semantic Evolution Operations
- Storage Model
- Querying the model

3 Concluding Remarks
- Performance Analysis
- Future Work

Semantic Evolution Operations

All semantic heterogeneity is generated by a **semantic evolution operation**, that happened in a **well-defined point in time**. Semantic evolution operations also have a **well-defined chronological order**.

Translation Operation

County	Year	Population
Moji Mirim	2014	91,027
Moji Mirim	2015	91,483
Mogi Mirim	2016	91,929
Mogi Mirim	2017	92,365

Table: Moji Mirim and Mogi Mirim population statistics **as IBGE published them.**

Translation Operation

County	Year	Population
~~Moji~~ Mogi Mirim	2014	91,027
~~Moji~~ Mogi Mirim	2015	91,483
Mogi Mirim	2016	91,929
Mogi Mirim	2017	92,365

Table: Mogi Mirim population statistics **after processing and standardization**.

Storage Model

A ***semantic version*** s is a tag assigned to a version of a document after a semantic evolution operation. Every semantic version has a number and a time t_s when it became active, that is, when the operation happened.

$$(d_1, s_1) \xrightarrow{T(args)} (d_1, s_2) \xrightarrow{G(args)} (d_1, s_3)$$

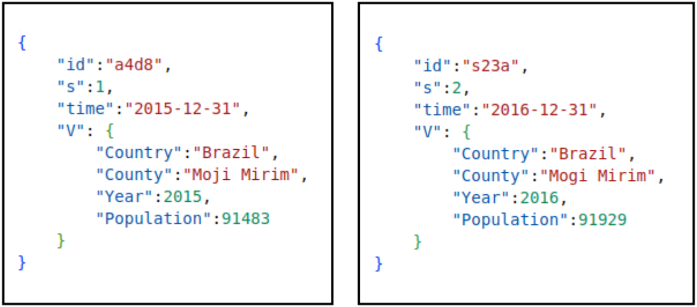
For the translation, it is also possible to revert the operation:

$$(d_1, s_1) \xleftarrow{T^{-1}(args)} (d_1, s_2)$$

C, a collection which has n documents and passes through m semantic evolution operations can be denoted by:

$$C = \{(d_1, s_1), (d_1, s_2), \ldots, (d_1, s_{m+1}), (d_2, s_1), \ldots, (d_n, s_{m+1})\}$$

Raw Collection

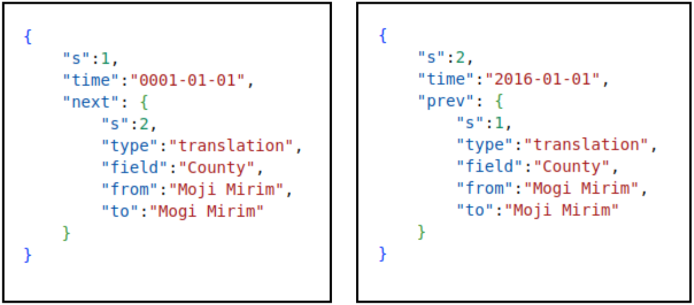


```
{
    "id":"a4d8",
    "s":1,
    "time":"2015-12-31",
    "V": {
        "Country":"Brazil",
        "County":"Moji Mirim",
        "Year":2015,
        "Population":91483
    }
}
```

```
{
    "id":"s23a",
    "s":2,
    "time":"2016-12-31",
    "V": {
        "Country":"Brazil",
        "County":"Mogi Mirim",
        "Year":2016,
        "Population":91929
    }
}
```

Figure: Raw version of records is kept intact.

Versions Collection

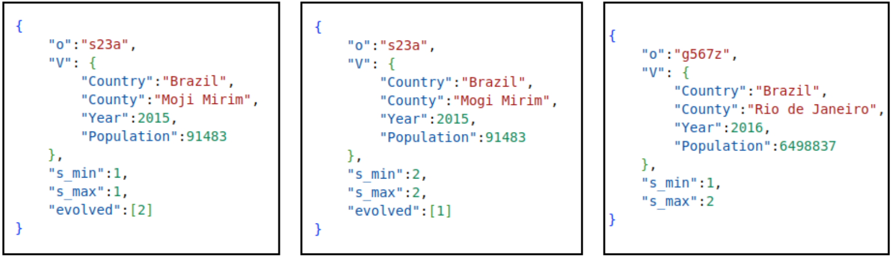


```
{
    "s":1,
    "time":"0001-01-01",
    "next": {
        "s":2,
        "type":"translation",
        "field":"County",
        "from":"Moji Mirim",
        "to":"Mogi Mirim"
    }
}
```

```
{
    "s":2,
    "time":"2016-01-01",
    "prev": {
        "s":1,
        "type":"translation",
        "field":"County",
        "from":"Mogi Mirim",
        "to":"Moji Mirim"
    }
}
```

Figure: The versions collection stores all semantic evolution operations. The model resembles a double-linked list.

Processed Collection



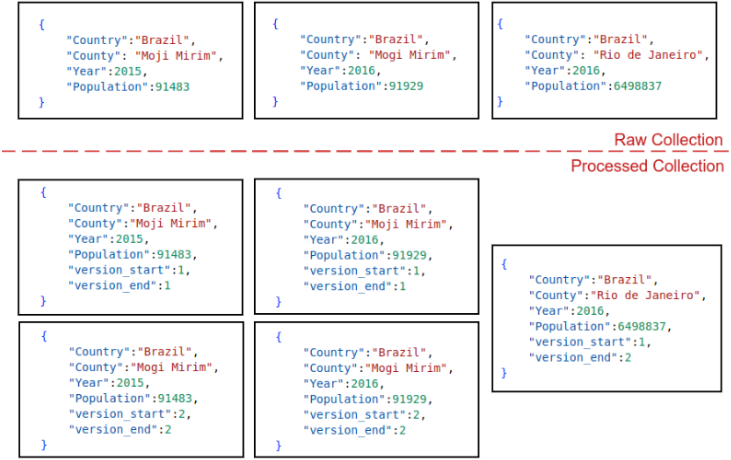
```
{
    "o":"s23a",
    "V": {
        "Country":"Brazil",
        "County":"Moji Mirim",
        "Year":2015,
        "Population":91483
    },
    "s_min":1,
    "s_max":1,
    "evolved":[2]
}
```

```
{
    "o":"s23a",
    "V": {
        "Country":"Brazil",
        "County":"Mogi Mirim",
        "Year":2015,
        "Population":91483
    },
    "s_min":2,
    "s_max":2,
    "evolved":[1]
}
```

```
{
    "o":"g567z",
    "V": {
        "Country":"Brazil",
        "County":"Rio de Janeiro",
        "Year":2016,
        "Population":6498837
    },
    "s_min":1,
    "s_max":2
}
```

Figure: Processed collection. All documents from the raw collection are made available in all semantic versions. Queries can be executed directly in any version desired.

Storage Model



```
{
    "Country":"Brazil",
    "County": "Moji Mirim",
    "Year":2015,
    "Population":91483
}
```

```
{
    "Country":"Brazil",
    "County": "Mogi Mirim",
    "Year":2016,
    "Population":91929
}
```

```
{
    "Country":"Brazil",
    "County": "Rio de Janeiro",
    "Year":2016,
    "Population":6498837
}
```

```
{
    "Country":"Brazil",
    "County":"Moji Mirim",
    "Year":2015,
    "Population":91483,
    "version_start":1,
    "version_end":1
}
```

```
{
    "Country":"Brazil",
    "County":"Moji Mirim",
    "Year":2016,
    "Population":91929,
    "version_start":1,
    "version_end":1
}
```

```
{
    "Country":"Brazil",
    "County":"Mogi Mirim",
    "Year":2015,
    "Population":91483,
    "version_start":2,
    "version_end":2
}
```

```
{
    "Country":"Brazil",
    "County":"Mogi Mirim",
    "Year":2016,
    "Population":91929,
    "version_start":2,
    "version_end":2
}
```

```
{
    "Country":"Brazil",
    "County":"Rio de Janeiro",
    "Year":2016,
    "Population":6498837,
    "version_start":1,
    "version_end":2
}
```

Figure: Copies of documents affected by semantic evolution operations. Unaffected documents only have their version interval extended.

Query Algorithm

- Documents are already processed when inserted into the processed collection. Therefore, no transformation in the results is needed.

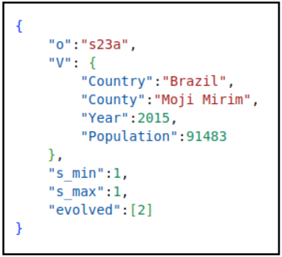
- It is necessary, however, to take into consideration possible evolution in terms queried.

Query Algorithm

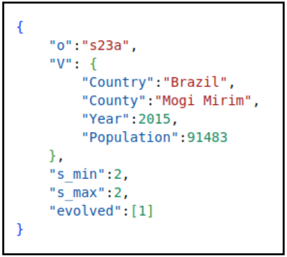
Take the following query as an example: *{"County":"Moji Mirim"}*
When querying in any semantic version, it must be taken into consideration:

- Counties that **are called** "Moji Mirim" in the queried version
- Counties that **were once called** "Moji Mirim" and use another name in the queried version
- Counties that **will be named** "Moji Mirim" in a later semantic version

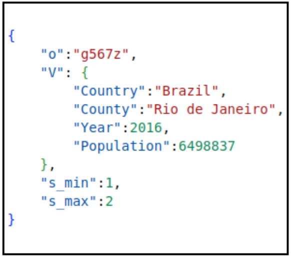
Query Algorithm



```
{
    "o":"s23a",
    "V": {
        "Country":"Brazil",
        "County":"Moji Mirim",
        "Year":2015,
        "Population":91483
    },
    "s_min":1,
    "s_max":1,
    "evolved":[2]
}
```



```
{
    "o":"s23a",
    "V": {
        "Country":"Brazil",
        "County":"Mogi Mirim",
        "Year":2015,
        "Population":91483
    },
    "s_min":2,
    "s_max":2,
    "evolved":[1]
}
```



```
{
    "o":"g567z",
    "V": {
        "Country":"Brazil",
        "County":"Rio de Janeiro",
        "Year":2016,
        "Population":6498837
    },
    "s_min":1,
    "s_max":2
}
```

Original query: *{“County”:“Moji Mirim”}*

Re-written query:

Q ={“County”:“Moji Mirim”}
or ({“county”:“Mogi Mirim”} and {“evolved__contains”:“1”})

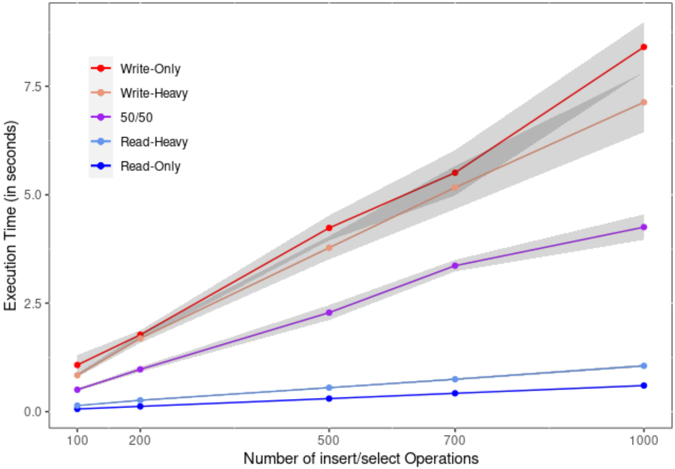
Table of Contents

1 Introduction
- What is Semantic Evolution?
- Query Impact

2 Theoretical Framework
- Semantic Evolution Operations
- Storage Model
- Querying the model

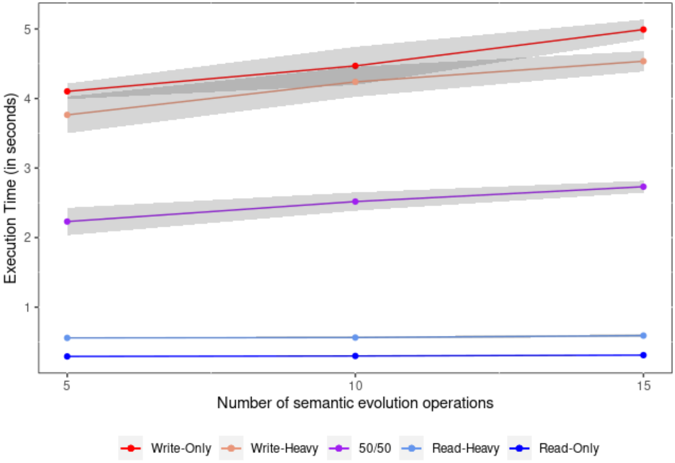
3 Concluding Remarks
- Performance Analysis
- Future Work

Impact of Number of Operations



Variation of the execution time with the number of executed operations for different workload scenarios. 500,000 documents in each execution. (95% CI)

Impact of Number of Evolutions



Variation of the execution time with the number of semantic evolution operations for different workload scenarios. (95% CI)

Future Work

- Development of a query rewriting algorithm to compare performances;
- Creation of a benchmark to compare different strategies, models, and algorithms that deal with semantic heterogeneity;
- Application of the models in two real use cases: counties name changes and corporate events in the stock market. Include cases where the conditions for the evolution involve more than one field;
- Improvement of performance tests: increase sample size, execute deeper statistical tests, and test using different hardware and network scenarios;
- Analyze possible impact regarding distributed environments.

Questions?

pedro.siqueira@ime.usp.br

This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, FAPESP proc. 14/50937-1, and FAPESP proc. 15/24485-9. The research is also funded by grant #2023/00779-0, São Paulo Research Foundation (FAPESP).