AUTOMATIC INTEGRATION ISSUES OF TABULAR DATA FOR ON-LINE ANALYSIS PROCESSING

ANR-19-CE23-0005 BI4people (Business Intelligence for the people)

Yuzhao Yang, Jérôme Darmont,
Franck Ravat, Olivier Teste

Context

- Numerous tabular data are produced by companies and individuals (Excel, csv, pdf…)
- Automate the BI process for small organisations, companies and individuals
- In this paper, we propose:
  - *A typology of tabular data*
  - *A process of automatic multidimensionnal schema generation*

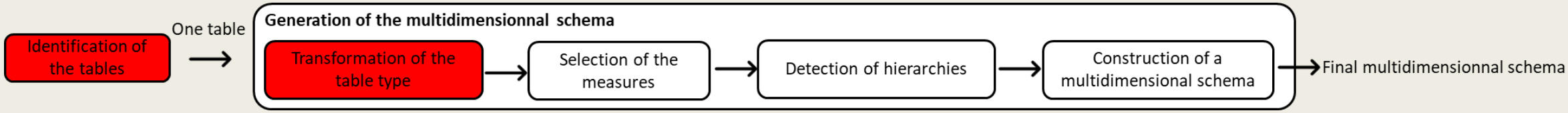
Related works

- Automatic multidimensional schema generation:
  - *Top-down: based on user requirements (Romero and Abelló,2009)*
  - *Bottom-up (Our orientation): based on data source (Relational database, XML…) (Phipps and Davis., 2002), (I.-Y.Song et al., 2007) (Golfarelli et al., 2001) (Usmanetal.,2011)*
  - *Hybrid (Romero and Abelló,2009)*
  - ➢ *No specific approach for tabular data:*
    - ➢ Difficult to retrieve keys, cardinalities, schemas, metadata.
    - ➢ Different types and structures of tabular data
    - ➢ Hard to undertand the semantics

- Typology of tabular data:
  - *Fonctionality of tables (Wang and Hu, 2002) (Crestan and Pantel, 2011)*
  - *Arrangement of the data(Crestan and Pantel, 2011) (Yoshida et al., 2001)*
  - *Dimensionality(Milosevic et al.,2016)*
  - *Characteristics of cells (Lautert et al., 2013)*
  - ➢ *None of these classifications is complete*

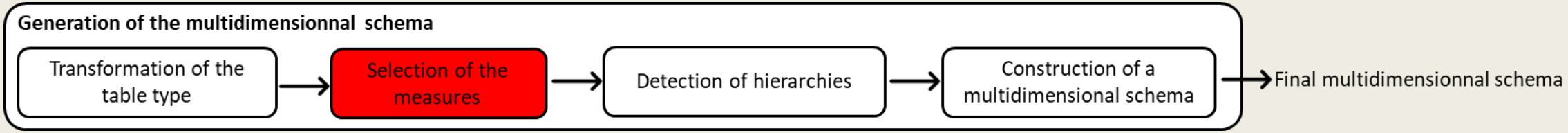
Augmented tabular data classification

Tabular data integration approach



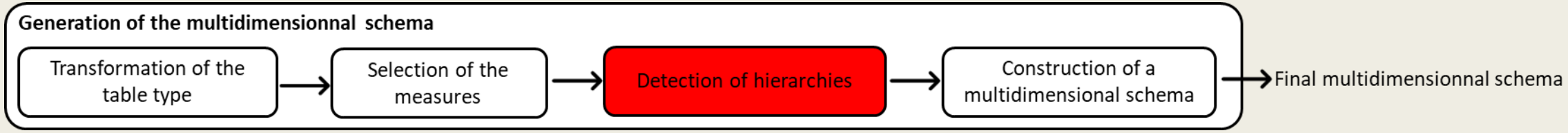
- Table identification:
  - *Table extraction (Zhang and Balog, 2020)*
  - *Detect tables from images (Schreiber et al., 2017); (Paliwal et al., 2019)*
- Table type transformation:
  - *Table structure:*
    - Multidimensional table: Transformed into one-dimensional table (Milosevic et al., 2016)
  - *Cell content*
    - Non-simple cells: Transformed into simple cells
  - *Header:*
    - Distributed and duplicated headers: Transformed into single-header table(Yoshida et al., 2001)
    - Hierarchical header: Extracted (Chen and Cafarella, 2013)
    - No attribute header: Column identification (Zhang and Balog, 2020)

Tabular data integration approach



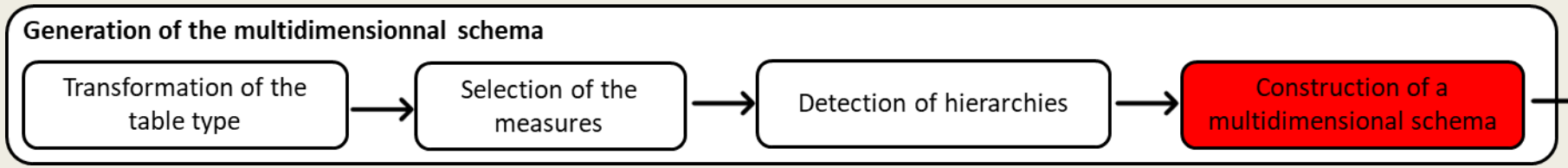
- (Alobaid et al., 2019): typology of numeric data:
  - *Nominal data*
  - *Ordinal data*
  - *Interval and ratios*
- Intervention of the user:
  - *Validate the proposed measures*
  - *Add specific measures*

Tabular data integration approach



- Automatic detection of all functional dependencies
- Get the minimal cover
  - *Get the elementary functional dependency*
  - *Remove all the transitivity and pseudo-transitivity dependencies*
- Approximate functional dependency (Liu et al., 2012)

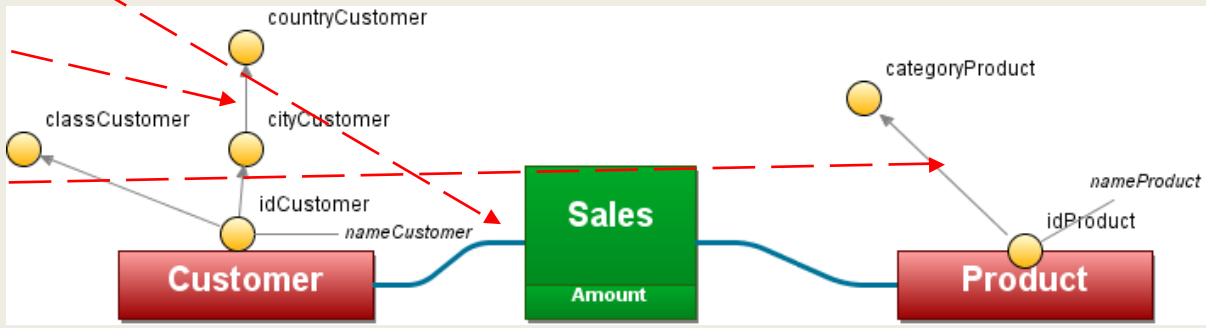
Tabular data integration approach



idCustomer → nameCustomer → cityCustomer → countryCustomer

classCustomer

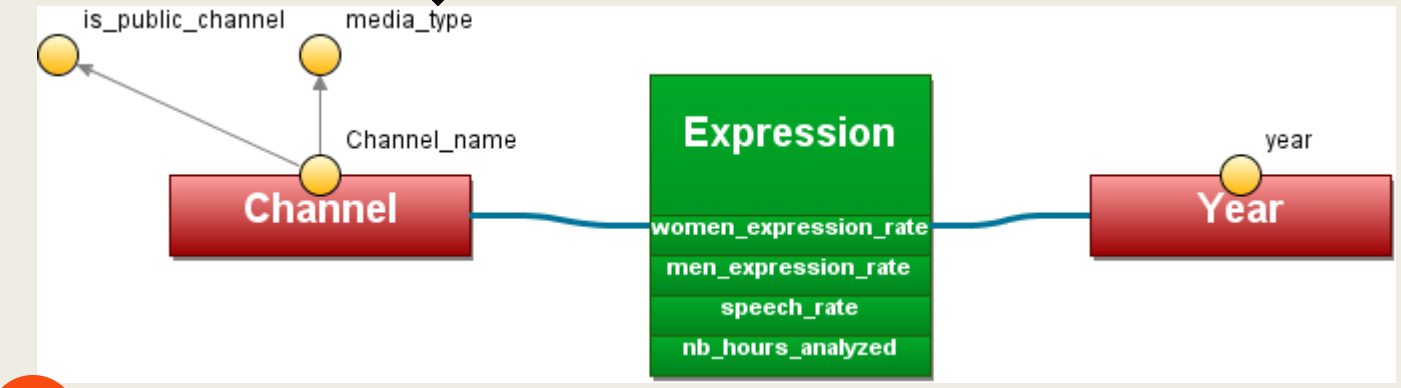
idProduct → nameProduct → categoryProduct



Implementation

	A	B	C	D	E	F	G
1	media_type	channel_name	is_public_channel	year	women_expression_rate	speech_rate	nb_hours_analyzed
2	radio	Chérie FM	False	2002	47.10994424	15.73869436	718
3	radio	Chérie FM	False	2003	46.03444471	16.25025819	1617
4	radio	Chérie FM	False	2004	48.38360747	15.03544979	1644
5	radio	Chérie FM	False	2005	45.45162675	16.06377772	1624
6	radio	Chérie FM	False	2006	47.81930726	15.6028421	1604
7	radio	Chérie FM	False	2007	43.94302035	15.67245605	1571
8	radio	Chérie FM	False	2008	51.39922002	17.53800278	1663
9	radio	Chérie FM	False	2009	49.95436887	15.93191099	1677
10	radio	Chérie FM	False	2010	48.06286688	15.94817147	1585
11	radio	Chérie FM	False	2011	47.83215331	17.98112396	1601
12	radio	Chérie FM	False	2012	51.89108386	18.37799131	1636
13	radio	Chérie FM	False	2013	54.11972167	20.89188492	1658
14	radio	Chérie FM	False	2014	50.30261056	18.14301169	1699
15	radio	Chérie FM	False	2015	51.22796482	22.28877115	1616
16	radio	Chérie FM	False	2016	51.17050513	21.62808475	1599
17	radio	Chérie FM	False	2017	53.08764883	19.57975861	1829
18	radio	Chérie FM	False	2018	52.14109768	18.50494497	2753

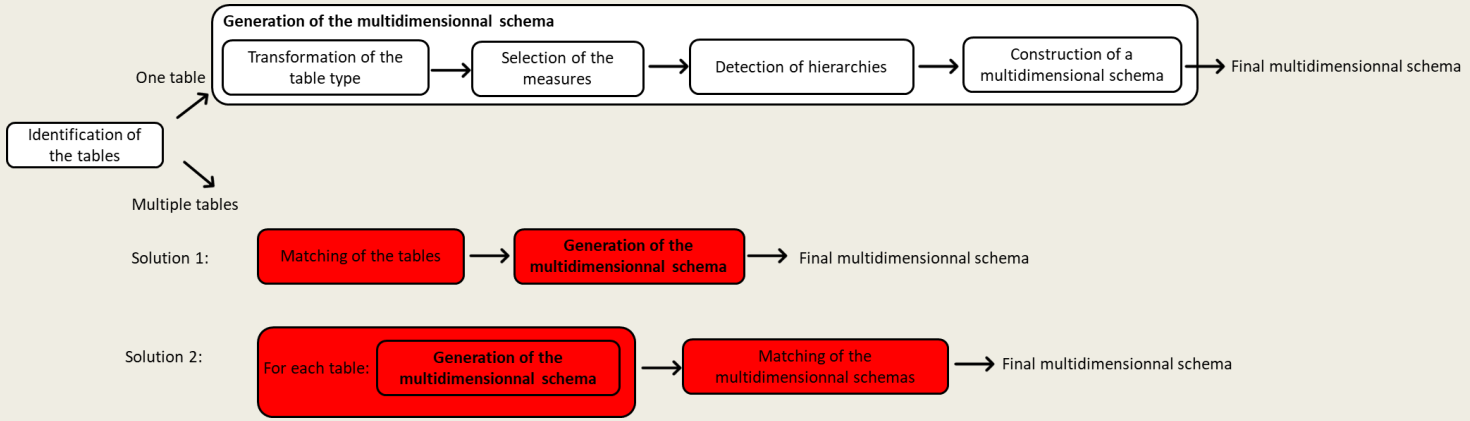
```
Hierarchies :
['channel_name', 'media_type']
['channel_name', 'is_public_channel']
['year']

Mesure :
['women_expression_rate', 'speech_rate', 'nb_hours_analyzed']
```



- Csv file from data.gouv.fr
- Link:https://www.data.gouv.fr/fr/datasets/temps-de-parole-des-hommes-et-des-femmes-a-la-television-et-a-la-radio/
- 7 Column, 701 rows
- A file recording speaking time of men and women corresponding to more than a million hours of programs broadcast from 1995 to February 28, 2019.

Conclusion and future research

- In this paper:
  - *Typology of tabular data*
  - *Process of automatic generation of multidimensional schema*
- In the future:
  - *Study the 2 solutions for multi-sources*



  - *Full implementation*
  - *Process of automatic population of the schema from data*

Thank you!