Exploring Data Hierarchies to Discover Data on Different Domains

Giuseppe Ricupero

Supervisors: Prof. **Silvia Chiusano**, Prof. **Tania Cerquitelli**

Ph.D. in Computer and Control Engineering - XXXI cycle

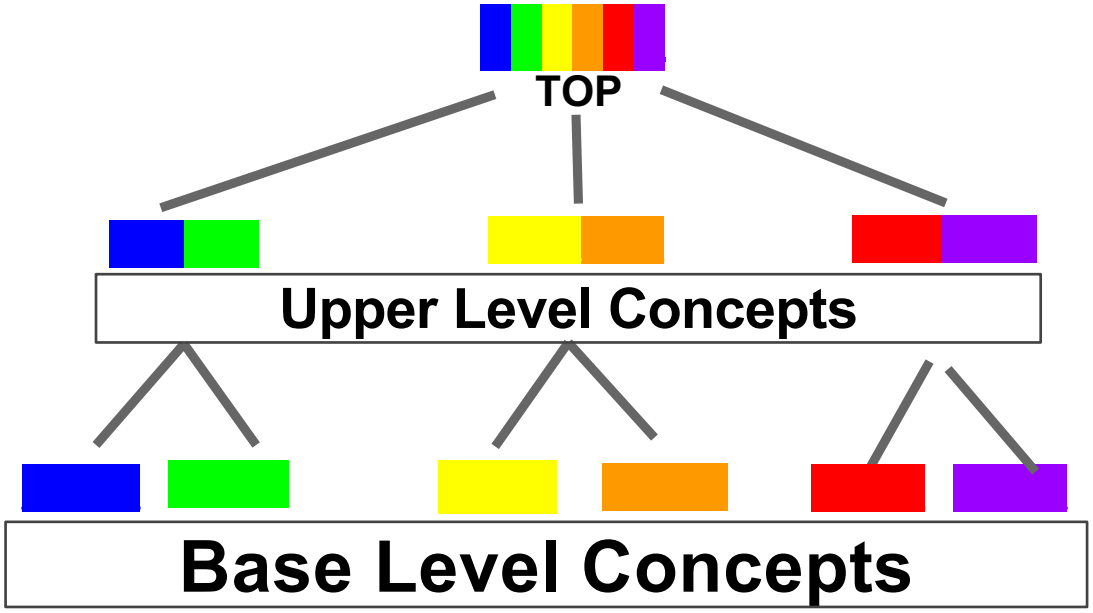
DBMG

POLITECNICO DI TORINO

Research Activity Context

- ICT generate huge amount of data
- **Data mining techniques** to extract unknown interesting patterns from these data: *cluster analysis*, *association rules analysis*, *classification*
- **Open Issue:** how to address the increasing *data cardinality*, *data dimensionality* and *variable data distribution* of these datasets to increase the **usability** of the information extracted

Research Activity Statement



The concept of **data generalization** applied to data mining techniques has been explored to extract information at **different levels of granularity** through **taxonomies** built on top of data.

Application Domains

DOMAIN	SUBJECT	TAXONOMY EXAMPLE
Urban	Air pollution	Concentration Criticality
Urban	Bike sharing	Temporal
Business	Retail Market	Product Groups

- Retail market
- Business directories integration

Theoretical Background

Frequent Itemset Mining

Def. (Itemset): a collection of *items* all belonging to distinct attributes

e.g., {Bread}, {Coke}, {Bread, Coke} are the possible itemsets in the example transaction **t4**

TID	Items
t1	Coke, Bread, Steak
t2	Water, Pasta, Steak
t3	Water, Bread
t4	Coke, Bread

FIMs	Support
{Bread}	3/4 (75%)
{Coke, Bread}	2/4 (50%)
{Water, Bread}	1/4 (25%)

$$\frac{f_{abs}(I)}{N}$$

Theoretical Background

Association Rules

Association Rule: $X \Rightarrow Y$
Where X,Y are itemsets and
$X \cap Y = \emptyset$

$$supp\,(X \Rightarrow Y) = \frac{f_{abs}(X \cup Y)}{N}$$

$$conf\,(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

$$lift\,(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)\,supp(Y)}$$

Example dataset

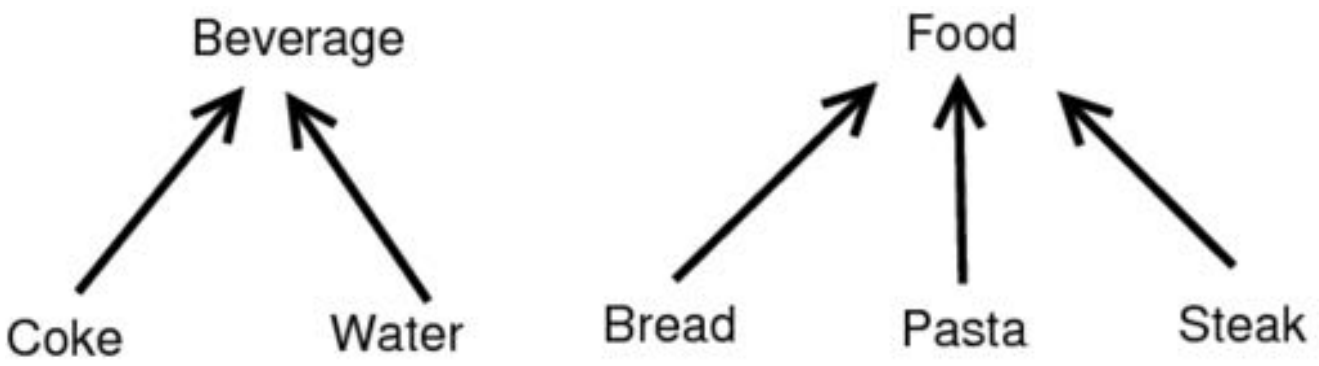
TID	Coke	Bread	Egg	Oat
t1	0	1	1	0
t2	1	1	0	1
t3	1	1	1	1

RULE (X => Y)	Support	Confidence	Lift
Coke ⇒ Oat	2/3 (66%)	1 (100%)	3/2 (> 1)
Bread ⇒ Coke	2/3 (66%)	2/3 (66%)	1 (= 1)
Coke,Bread ⇒ Egg	1/3 (33%)	1/2 (50%)	3/4 (< 1)

Theoretical Background

Generalized Items

Using a taxonomy the dataset is enriched describing the semantic is-a relationship among data items



Monitoring Air Pollution

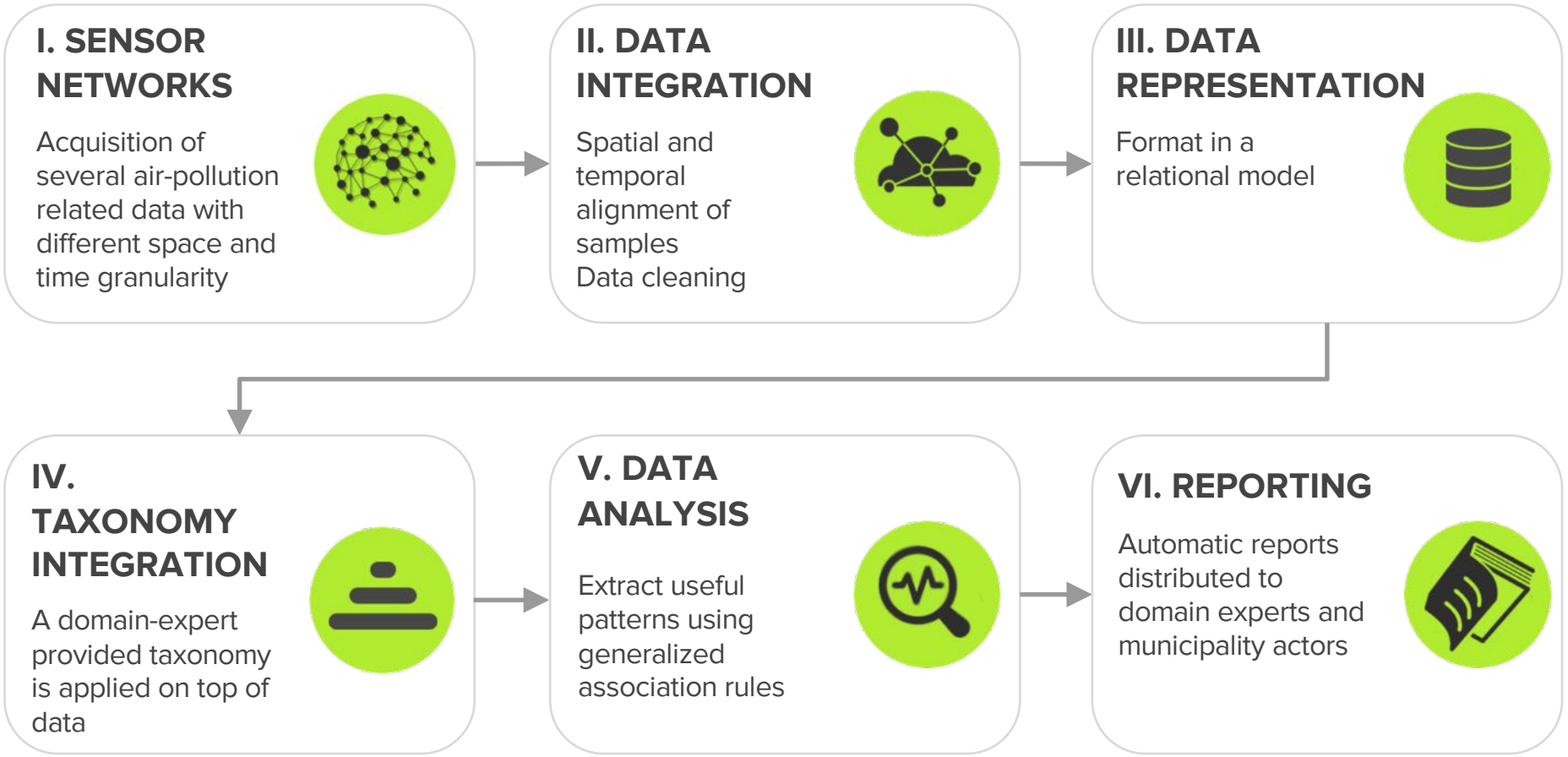
- Addressing the causes of air pollution entails discovering the **correlations** among heterogeneous data such as *pollutant concentrations*, *traffic flow measurements*, and *meteorological data*

- **AIM** - to support **city administration** in the decision-making process used to **control air pollution**

Related work

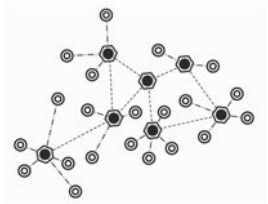
1. **Correlation** among air pollutants and between air pollutants and meteorological data
    - Maura Lodovici et al (PAH, Benzo(a)pyrene, Toxic equivalency)
    - Statheropoulos et al. (Principal component with meteo data)
2. **Classification** to predict air quality levels (Zheng et al.)
    - Predict air quality levels in unmonitored areas
    - Predict air quality levels in the future

GECKO Architecture



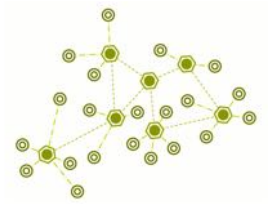
Sensor Networks

Weather



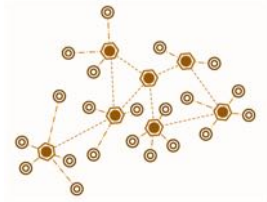
UV index
Humidity
Temperature
Precipitation
Wind Speed
Wind Direction
Atmospheric Pressure

Pollutants



Particulate Matter $10e^{-6}m$ (PM_{10})
Particulate Matter $2.5e^{-6}m$ ($PM_{2.5}$)
Ozone (O_3)
Nitrogen dioxide (NO_2)
Carbon Monoxide (CO)
Benzene (C_6H_6)

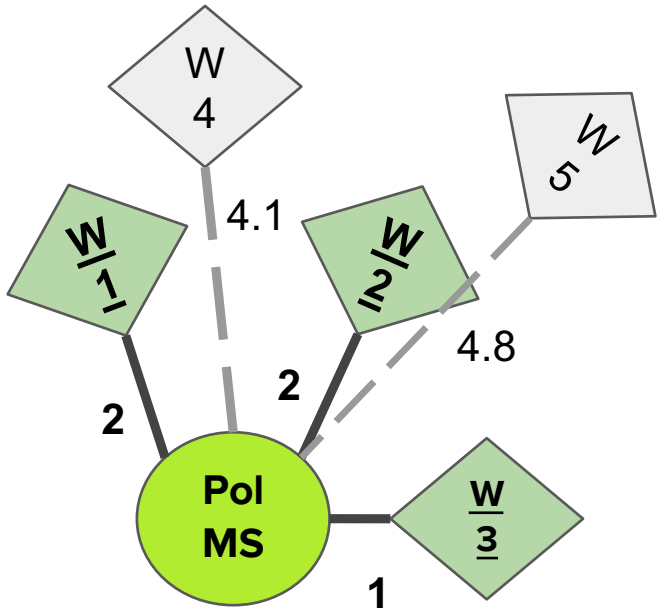
Traffic (by fuel type)



Petrol
Diesel
Electric
GPL / Metan
Hybrid (petrol/electric)

Data Integration

After being cleaned, starting from the **spatial coordinates** of the **PolMS** we can obtain the (Euclidian) distance of each weather station and calculate the weighted average mean.



WEATHER STATION (i)	DISTANCE (d)	VALUE (v)
W1	2	20
W2	2	26
W3	1	10

$$\overline{\chi}_W = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{v_i}{d_i}\right), n = 3$$

$$\Rightarrow \frac{1}{n}\left(\frac{v_1}{d_1} + \frac{v_2}{d_2} + \frac{v_3}{d_3}\right)$$

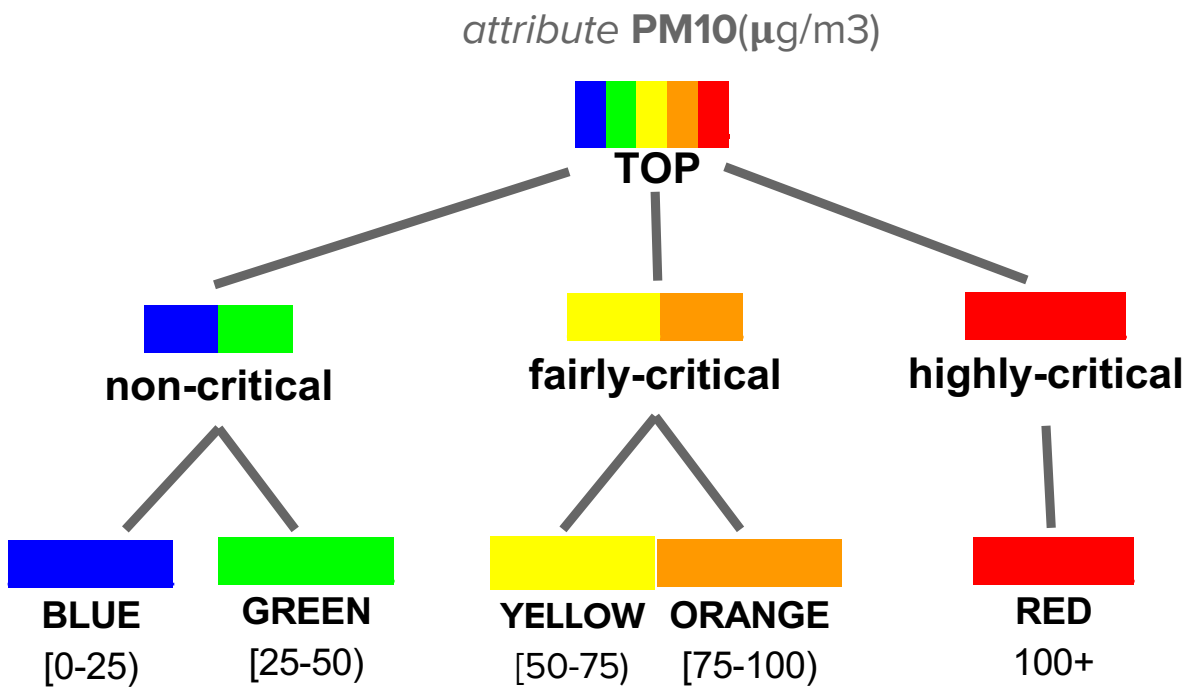
$$\Rightarrow \frac{1}{3}\left(\frac{20}{2} + \frac{26}{2} + \frac{10}{1}\right)$$

$$\Rightarrow \frac{1}{3}(33) = 11$$

Taxonomy Integration

To grasp correlations better and represent information more clearly, different levels of abstraction are assigned for each attribute by means of a domain-expert provided **taxonomy**

*Sample taxonomy example on **PM10** Pollutant using **ARPA** color-based classification*



Data Analysis

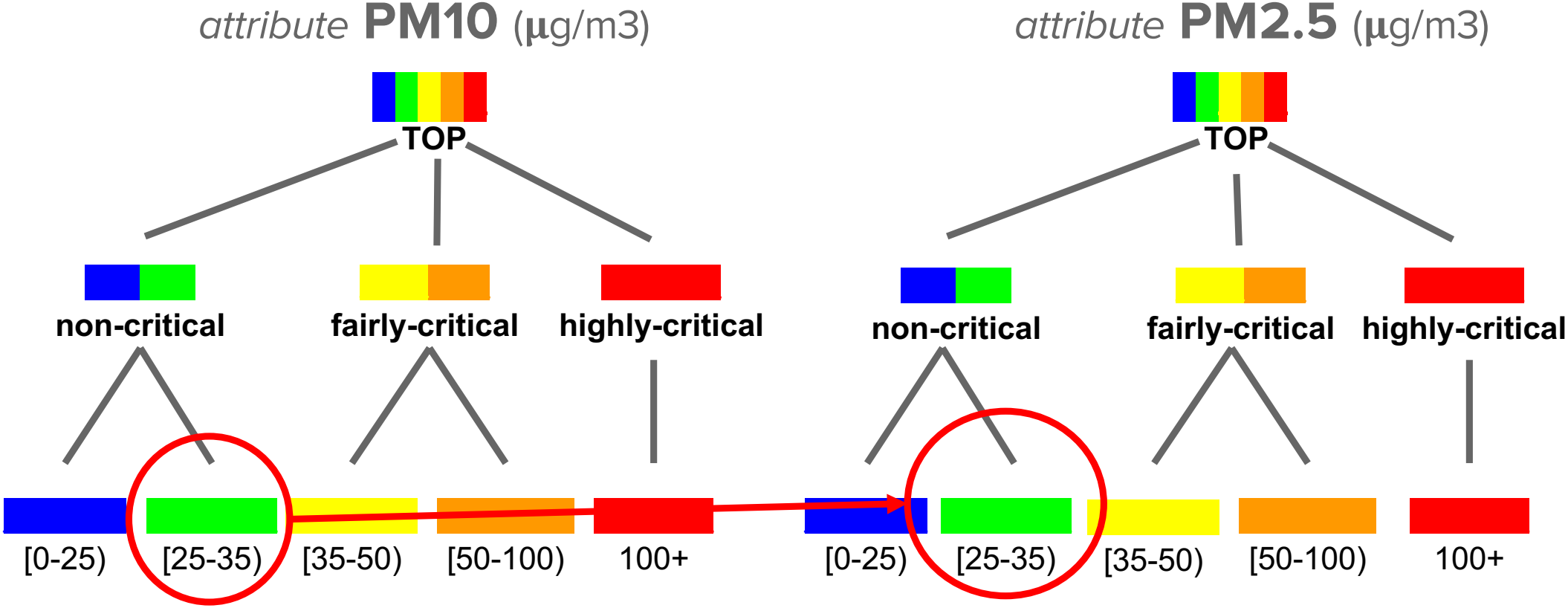
Generalized Association Rules

- **Generalized Itemset** is a set of *items* (i.e., (attribute,value)), and/or *generalized items* (obtained by means of a taxonomy)

- **Generalized Association Rules** is an implication **X -> Y,** where **X** and **Y** are disjoint generalized itemsets

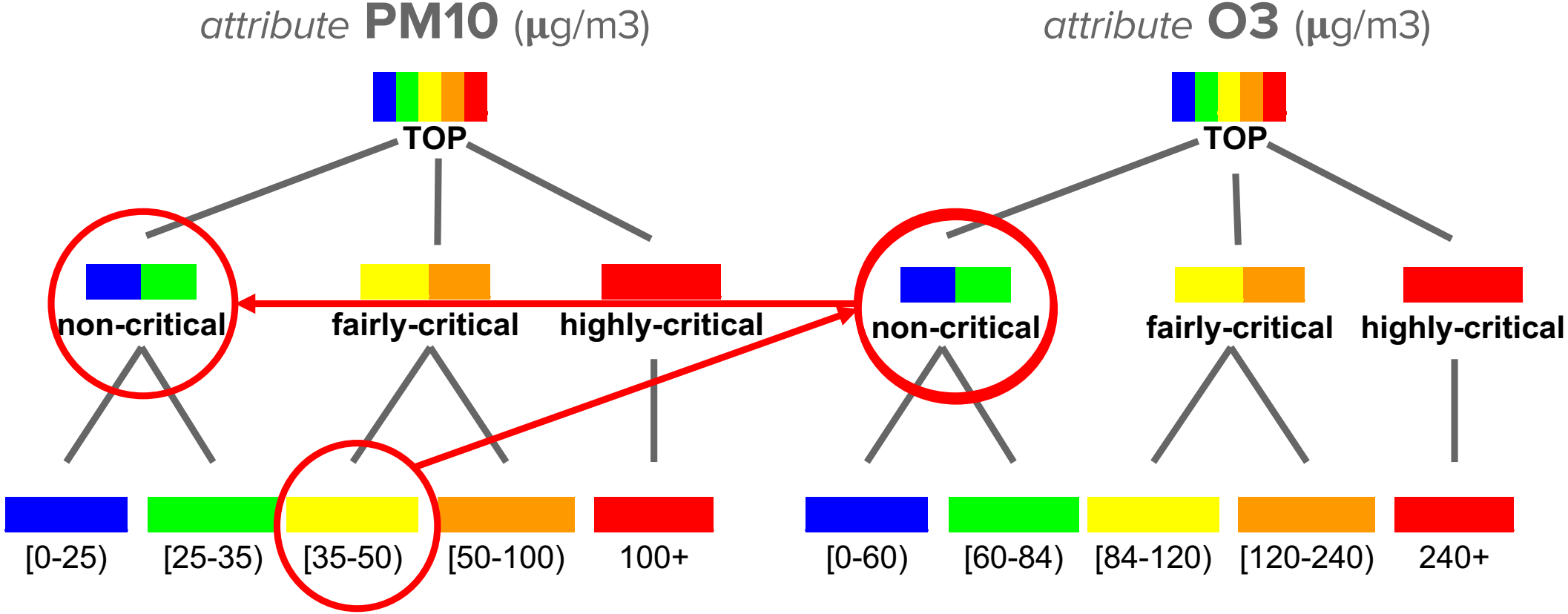
Data Analysis

Generalized Association Rules



Data Analysis

Generalized Association Rules



Knowledge Discovery

Class	Ante	Cons	Supp	Conf	Lift
PP*	O_3, *highly-critical*	NO_2, *non-critical*	5.9	51.6	1.5
PP*	O_3, *non-critical*	NO_2, *highly-critical*	11.3	24.1	1.7
PM*	Precipitations, drizzling PM_{10}, orange $PM_{2.5}$, red	Temp, very cold CO, fairly-high	1.1	40.0	20.1
PTE*	Date, *spring*	PM_{10}, green	5.6	55.6	1.4
PTE*	Date, *winter* O_3, *non-critical*	NO_2, *highly-critical*	5.9	26.9	1.9
PTR	N. Diesel, *high*	PM_{10}, green	14.7	34.4	0.9
PTR*	N. Diesel, *medium*	PM_{10}, fairly-high	9.7	70.0	1.8

GECKO Framework

Published

Cagliero L., Cerquitelli T., Chiusano S., Garza P. **Ricupero G.**, Xiao X.

Modeling Correlations Among Air Pollution-related Data through Generalized Association Rules, 2nd IEEE International Conference on Smart Computing (SMARTCOMP 2016), May 18-20 2016, St.Louis, Missouri.

SENSOR NETWORKS

Different kind of air pollution-related open data

DATA INTEGRATION AND CLEANING

Spatially and Time Sampling aligning. Data cleaning

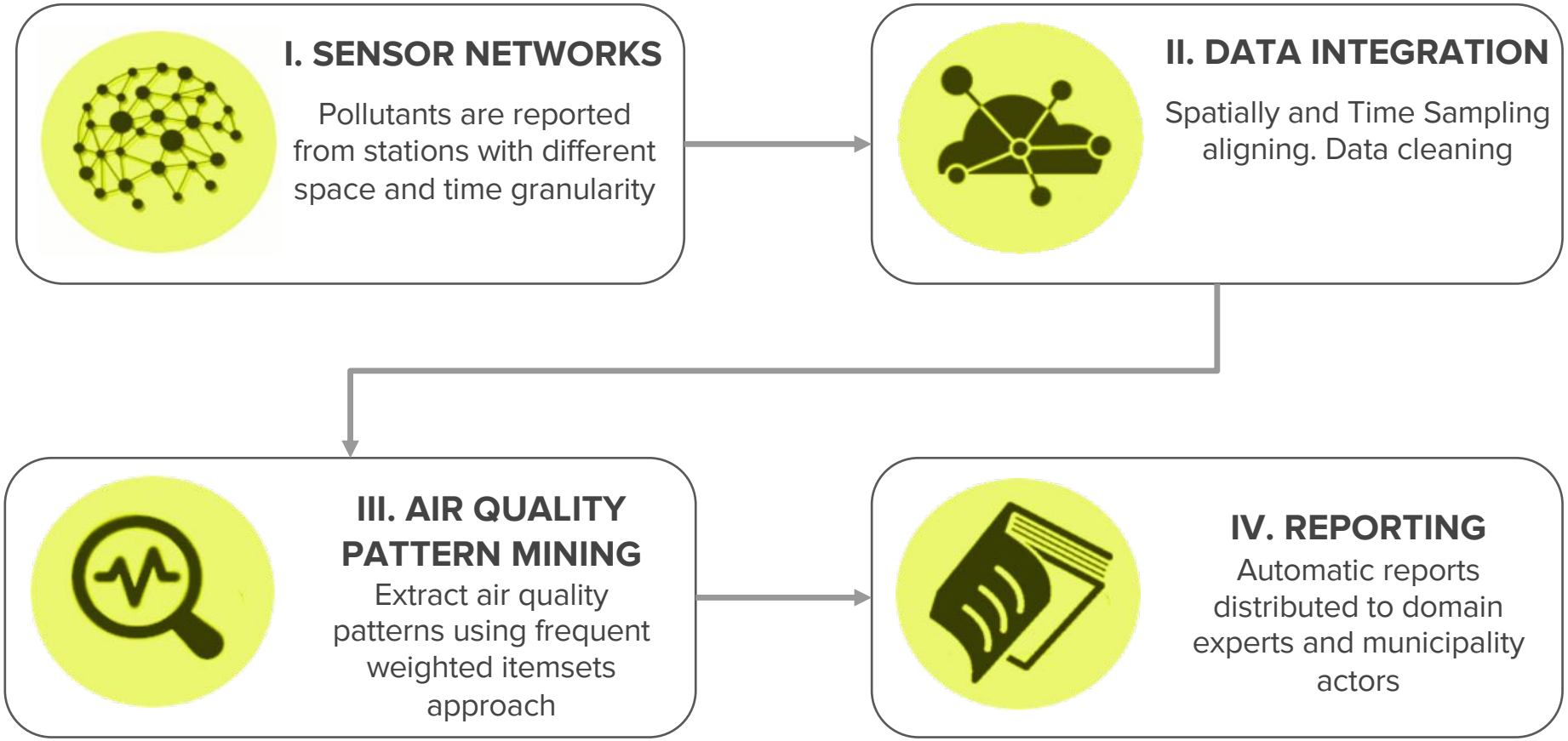
TAXONOMIES

Build a relational model, discretize and apply a taxonomy

DATA ANALYSIS AND REPORTING

Extract useful patterns using generalized association rules

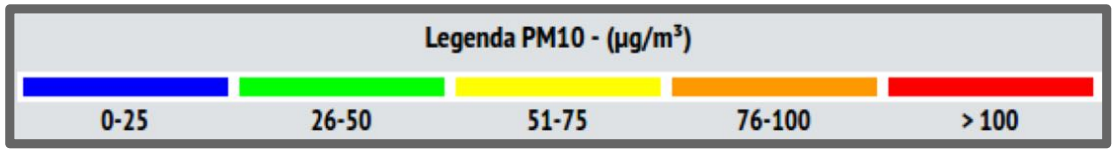
ARQUATA Architecture



Air Quality Pattern Mining

ARQUATA discovers combinations of pollutants whose concentration levels are all averagely critical in the considered time period

Time	P_1	P_2	...	P_n
t_1	10	23.1	...	46.4
...	...	...	...	...
t_m	12.2	65.3	...	28.2



A critical level (**CL**) for each pollutant is given by the domain expert (e.g., the critical level specified by law)

Air Quality Pattern Mining

A **weight (W)** is associated with each **pollutant** (P_j) at time t_i as the **average percentage variation** of its concentration with respect to the **critical level (CL)**.

$$W(P_j,t_i)=\frac{D(P_j,t_i)-CL(P_j)}{CL(P_j)}$$

A value called **Critical Gap (CG)** is associated with a combination of pollutants (**itemset I**). It is equal to the average weight calculated on the entire dataset.

$$CG(I)=\frac{1}{m}\sum_{i=1}^{m} min_{P_j\in I}(W(P_j,t_i))$$

Selected **itemsets** are those whose critical gap is above a **user-provided threshold (Th)**. For example, when Th=10%, itemset {$PM_{2.5}$, PM_{10}} is extracted if both pollutants have an average percentage variation above 10%.

Knowledge Discovery

Season	Itemset	Critical Gap (threshold 30%)
Autumn	{$PM_{2.5}$}	90.62
	{PM_{10}, $PM_{2.5}$}	77.64
	{NO_2, $PM_{2.5}$, PM_{10}}	34.12
Winter	{$PM_{2.5}$}	120.72
	{PM_{10}, $PM_{2.5}$}	93.49
	{NO_2, $PM_{2.5}$, PM_{10}}	45.92
Spring	{NO_2}	30.07
Summer	-	-

ARQUATA Framework

Published

Cagliero L., Cerquitelli T., Chiusano S., Garza P., Ricupero G.

Air Quality Patterns in Urban Environments, 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2016), September 12-16, Heidelberg, Germany.

SENSOR NETWORK

Pollutants are reported from stations with different space and time granularity

DATA INTEGRATION AND CLEANING

Spatially and Time Sampling aligning. Data cleaning

AIR QUALITY PATTERN MINING

Extract air quality patterns using frequent weighted itemsets approach

REPORTING

Automatic reports generated

Green Urban Mobility

Analyses of Bike Sharing Data

- Promoted by municipalities in order to **reduce pollutant emissions** and **traffic congestion**
- Bikes are **shortly rented** from stations scattered around the city
- Stations transmit their state through **IoT** devices
- To achieve a satisfactory user experience stations must not be **overloaded** nor **empty**

Bike Station Overload Analyzer (BELL)

- Novel **exploratory data-driven methodology** to analyze occupancy levels from geo-referenced bike stations
- It extract a new type of pattern, called ***Occupancy Monitoring Pattern*** (OMP)
- OMPs allow to detect groups of (potentially overlapped) **nearby stations** showing a **critical** or **alternate** usage patterns

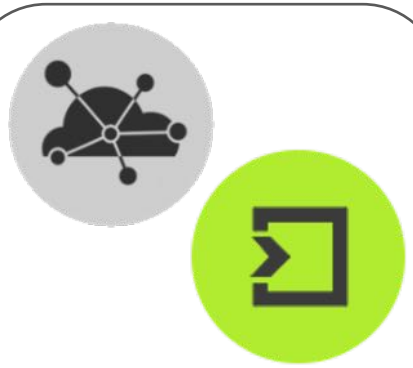
Related Work

1. **Grouping** stations based on their common usage profile
    - Clustering (Etienne, Latifa)
2. **Predicting** future station occupancy levels
    - Regression (Lozano et al.) or Classification (Hasan et al.)
3. **Bicycles repositioning** among the stations
    - Static (Liu et al.)
    - Dynamic (Nair et al.)
    - User-based (Fricker et al.)

BELL Architecture

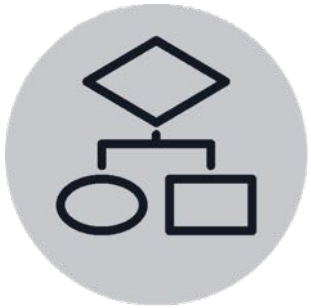
STATION
IoT SENSORS

Occupancy levels data are collected from the stations



DATA INTEGRATION

Data enriched with spatio-temporal info and saved into a relational dataset



PATTERN EXTRACTION OMP-MINER

A spatial constraint parameter is passed

KNOWLEDGE EXPLORATION

OMP ranking, pattern processing and visualization

Taxonomy Concept in BELL

To leverage the concept of taxonomy, the occupancy level data have been enriched with temporal information with a coarser granularity level.

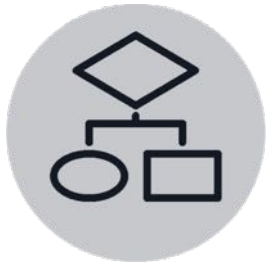
HOURLY TIME PERIOD

Determining the hour to rebalance the bikes among the stations

DAILY TIME PERIOD

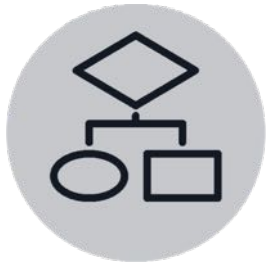
Spotting the most unbalanced days in terms of dock occupancy

Occupancy Monitoring Pattern (OMP)



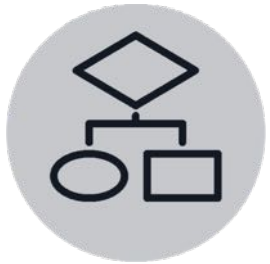
- OMP represents sets of stations showing a dock overload condition
- Based on the concept of **occupancy level** (*overloaded*, *normal)*
- Consider only groups of **nearby** stations
- **Critical situation:** occupancy level frequently overloaded for all stations at the same time
- **Intermittent situation:** group of stations in an overload condition in an alternate fashion

OMP-Miner



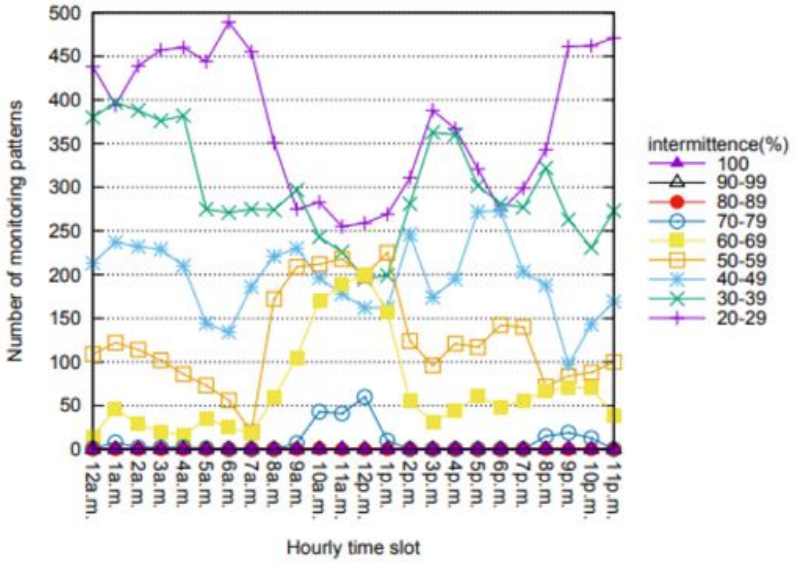
1. Creation of an in-memory FP-Tree from the transactional dataset
2. Mining all the homogeneously **critical** and **normal o-itemsets** using the spatial constraint **maxdist**
3. Generating the **OMPs** and their **criticality** and **intermittence levels**

OMPs Examples

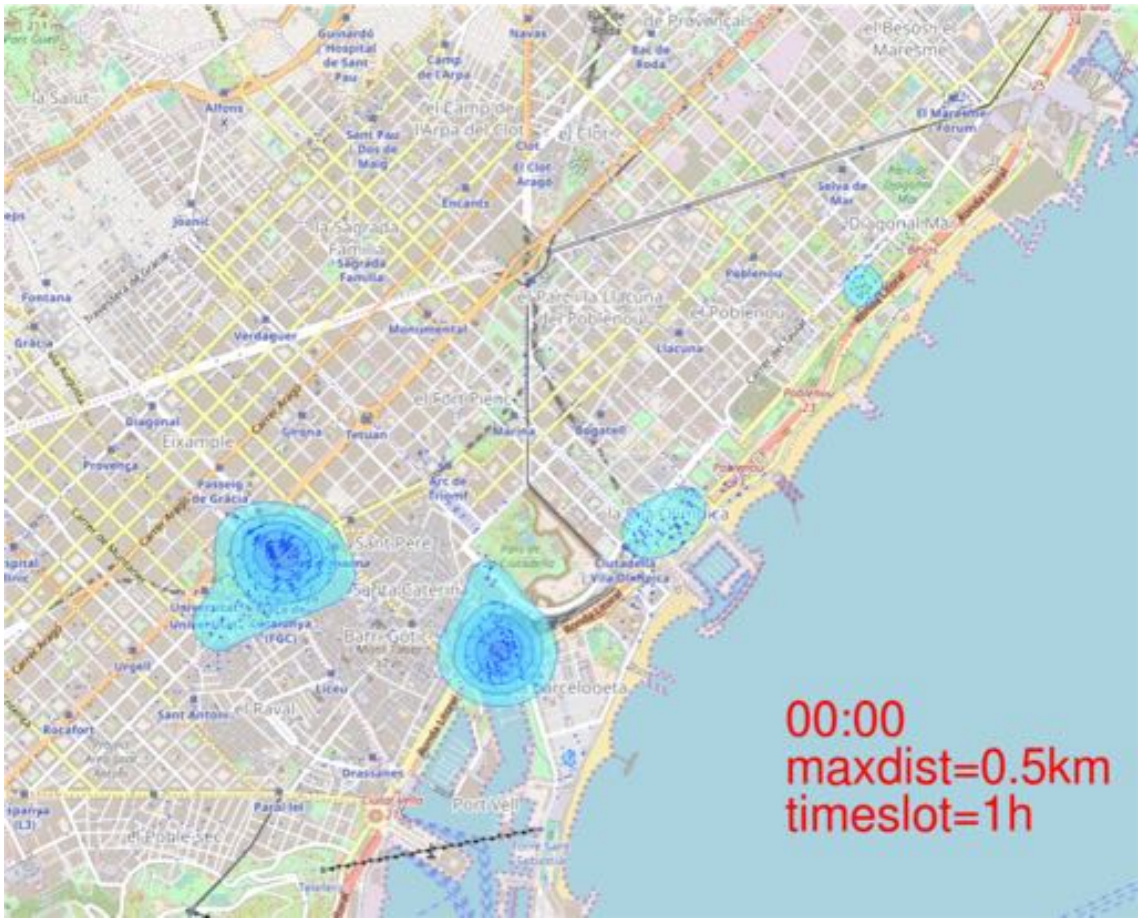


RID	Timestamp	Stations			Time period
		s_1	s_2	s_3	
1	t_1	Overloaded	Overloaded	Overloaded	TP_1
2	t_2	Overloaded	Normal	Overloaded	TP_1
3	t_3	Overloaded	Overloaded	Normal	TP_1
4	t_4	Overloaded	Normal	Normal	TP_1
5	t_5	Normal	Overloaded	Normal	TP_2
6	t_6	Normal	Overloaded	Normal	TP_2
7	t_7	Normal	Normal	Normal	TP_3

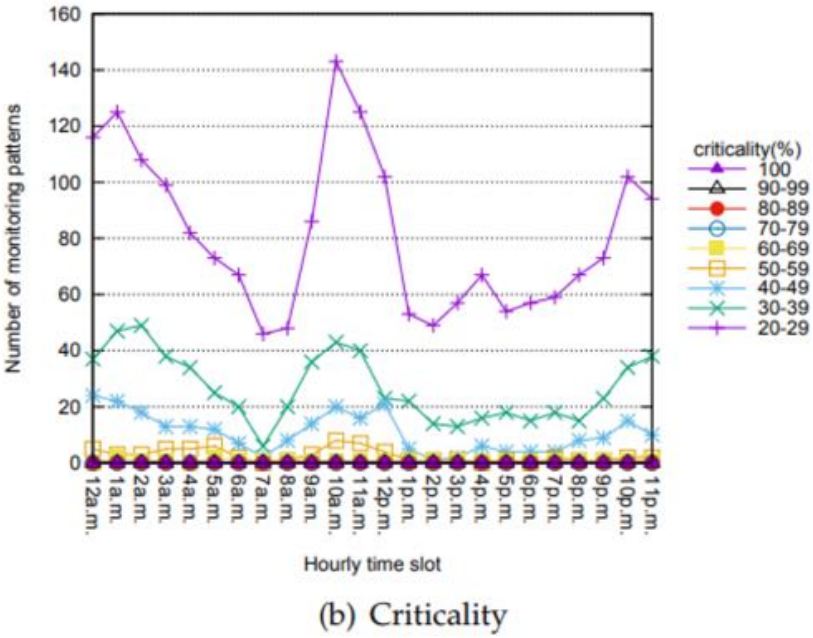
Intermittence situation visualization



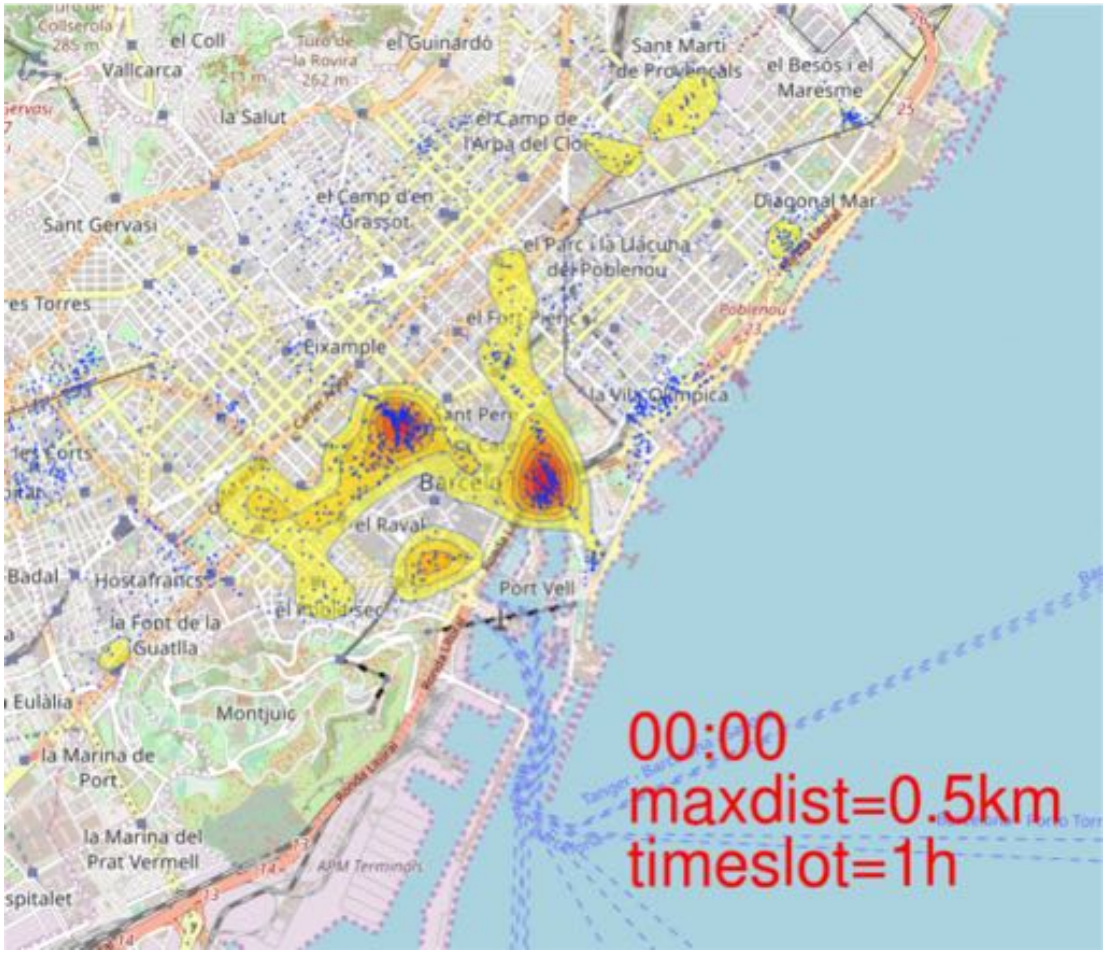
(a) Intermittence



Criticality situation visualization



(b) Criticality



BELL Framework

Published to: Applied Sciences

Cagliero L., Cerquitelli T., Chiusano S., Garza P., Ricupero G., Baralis E.

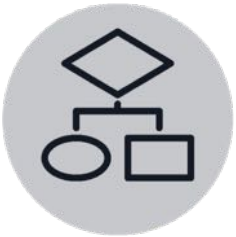
Characterizing situations of dock overload in bicycle sharing stations

IoT SENSORS

Occupancy levels data are collected from the stations

DATA INTEGRATION

Data enriched with spatio-temporal info, saved into a relational dataset



PATTERN EXTRACTION OMP-MINER

A spatial constraint parameter is passed

KNOWLEDGE EXPLORATION

OMP ranking, pattern processing and visualization

Business Domain

- A **novel pattern** will be presented on the **retail market: GHUI**

- A methodology **to align the taxonomies** of business activities web directories: **TACOMA**

GHUI Mining

A novel pattern, called **Generalized High-utility Itemset (GHUI),** that combines two data mining patterns:

- High-utility Itemsets (HUI)
- Generalized itemsets

Theoretical Background

High Utility Itemsets

Utility of {Coke, Steak} = (2x5)+(1x10) = 20

TID	Items and internal utility
t1	(Coke, 2), (Bread, 2), (Steak, 1)
t2	(Water, 3), (Pasta, 2), (Steak, 1)
t3	(Water, 2), (Bread, 2)
t4	(Coke, 1), (Bread, 2)

Item	External Utility
Water	1
Coke	5
Bread	1
Pasta	2
Steak	10

Generalized High-utility Itemsets

- Preserve the value of the High-utility Itemsets
- Enrich with the expressive power and compact representation of generalization
- Define rules to deal with the utility calculation and the threshold applied at multiple levels

Generalized High-utility Itemsets

TID	Items and internal utility
t1	(Coke, 2), (Bread, 2), (Steak, 1)
gt1	(Beverage, 2), (Food, 2)
t2	(Water, 3), (Coke, 3)
gt2	Beverage

Itemset	Utility
{Coke, Bread, Steak}	22
{Beverage, Food}	22
{Water, Coke}	18
{Beverage}	28

Beverage

Item	Coke	Bread	Steak	Water
eu(i)	5	1	10	1

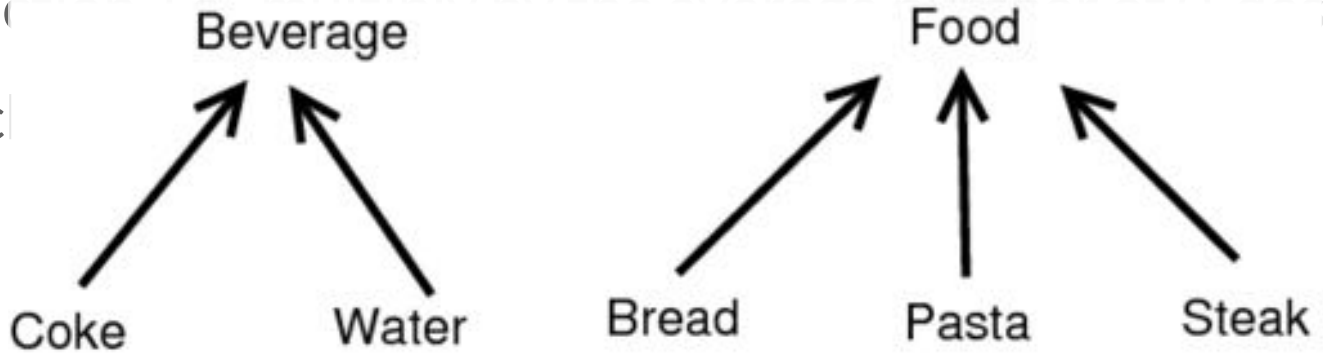
Generalized High-utility Itemsets

Calculate the level 1 minutil

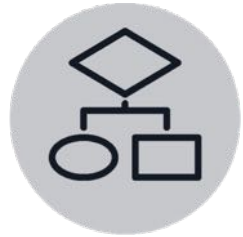
- Higher abstraction levels easier satisfy minutil constraint
- To overcome this problem, t **minutil** increase at each taxonomy level according to a monotonically increasing user-specified function i.e. multiply by the average number of 0 descendants at each

level 1 avg descendants 2.5

level	minutil
0	12
1	12 * 2.5 = 30



The ML-HUI Miner Algorithm



1. Taxonomy-driven ***HUI Mining***
2. *Single-phase extraction of generalized and non-generalized HUIs*
3. *Prevent the generation of uninteresting combinations of items avoiding cross levels combinations*

Experiments

DATASETS
Four from UCI:
Chess, Connect, Mushroom, **Retail**

RETAIL
Use **real taxonomy** and real **prices**

DATASET TRANSACTIONS
Min: Chess (3196)
Max: Retail (67557)

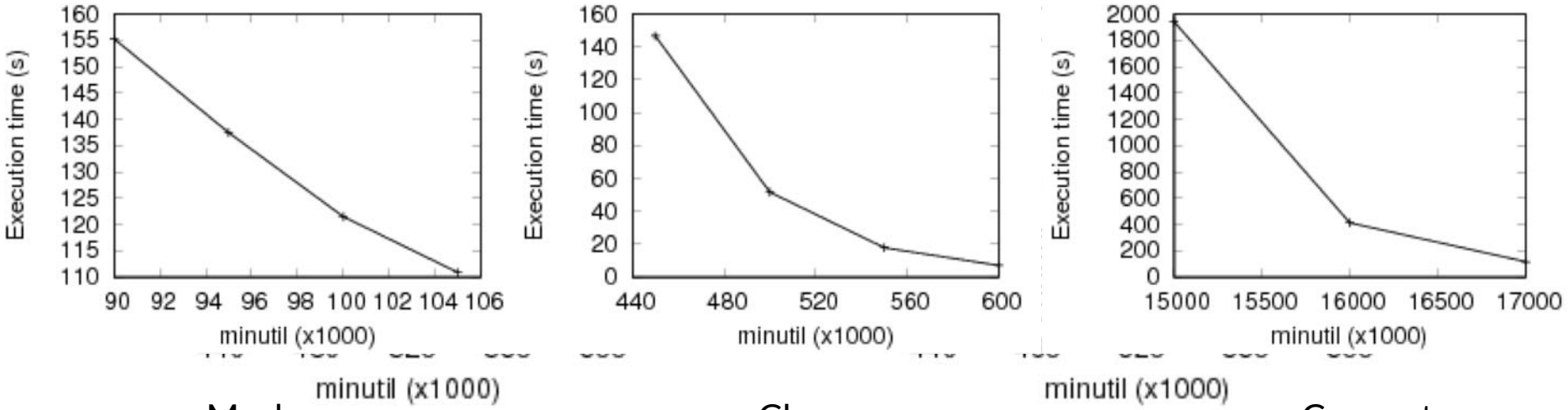
HW USED
Intel Xeon 2.67 GHz quadcore 32GB RAM (Ubuntu 12.04)

SYNTHETIC TAXONOMIES
Internal utility range: **1-5**
External utilities range: **1-1000**

DATASET ITEMS
Min: Chess (75)
Max: Retail (2741)

Experiments: performance

- Both patterns are **inversely proportional** to *minutil*
- *GHUIs are 2 orders of magnitude lower than HUI*



Mushroom Chess Connect

Experiments: regular HUIM comparison

- In respect to the standard FHM **execution time are linearly increased** with the number of generalized transactions generated (e.g., the time doubles with a level-2 taxonomy)
- Due to the pyramid structure of a taxonomy, and to the reduced number of per-level combinations, **higher order taxonomies does not increase execution time significantly**

Experiments: Example Patterns

lenght	GHUI
1	{Kitchen}
1	{Toy}
2	{Musical-Instrument, Kitchen}
2	{Musical-Instrument, Toy}
2	{Musical-Instrument, Home}
2	{Musical-Instrument, Office}

Experiments: example patterns

type	Itemsets	Utility
GHUI	{Musical-Instrument}	40k
HUI	{Red-Harmonica}	25k
HUI	{Blue-Harmonica}	10k

GHUIM Pattern

Published

Cagliero L., Cerquitelli T., Chiusano S., Garza P., Ricupero G.

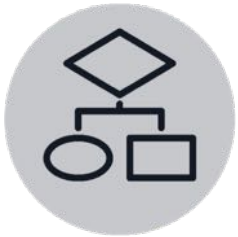
Discovering High-Utility Itemsets at Multiple Abstraction Levels, DaS 2017 (ADBIS 2017), Nicosia (Cyprus), pp. 224-234

RETAIL DATA

The main experiment has been performed on real data with real prices

TAXONOMIES

A taxonomy has been reconstructed to feed the algorithm



ML-HUI ALGORITHM

A single-phase algorithm to extract HUIs and GHUIs

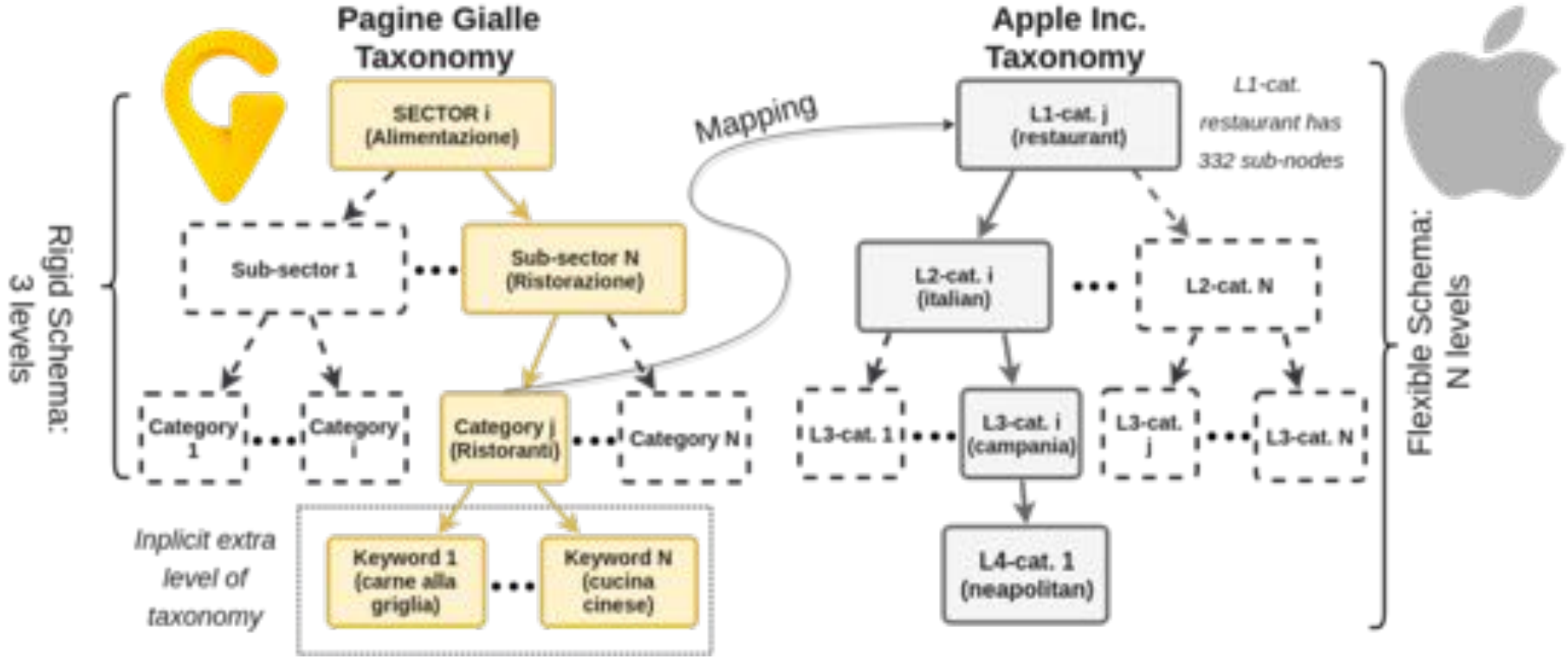
GENERALIZED ASSOCIATION RULES

Due to generalization compact rules are provided

The mapping problem

- **Integration** of business activities among **different web directories**
- The taxonomies used by external entities are often **significantly different** from the source taxonomy
- The usual but limited approach is to create a **static mapping** which can’t handle different granularity levels of taxonomies
- **Real industry case**

Taxonomies with different granularity levels



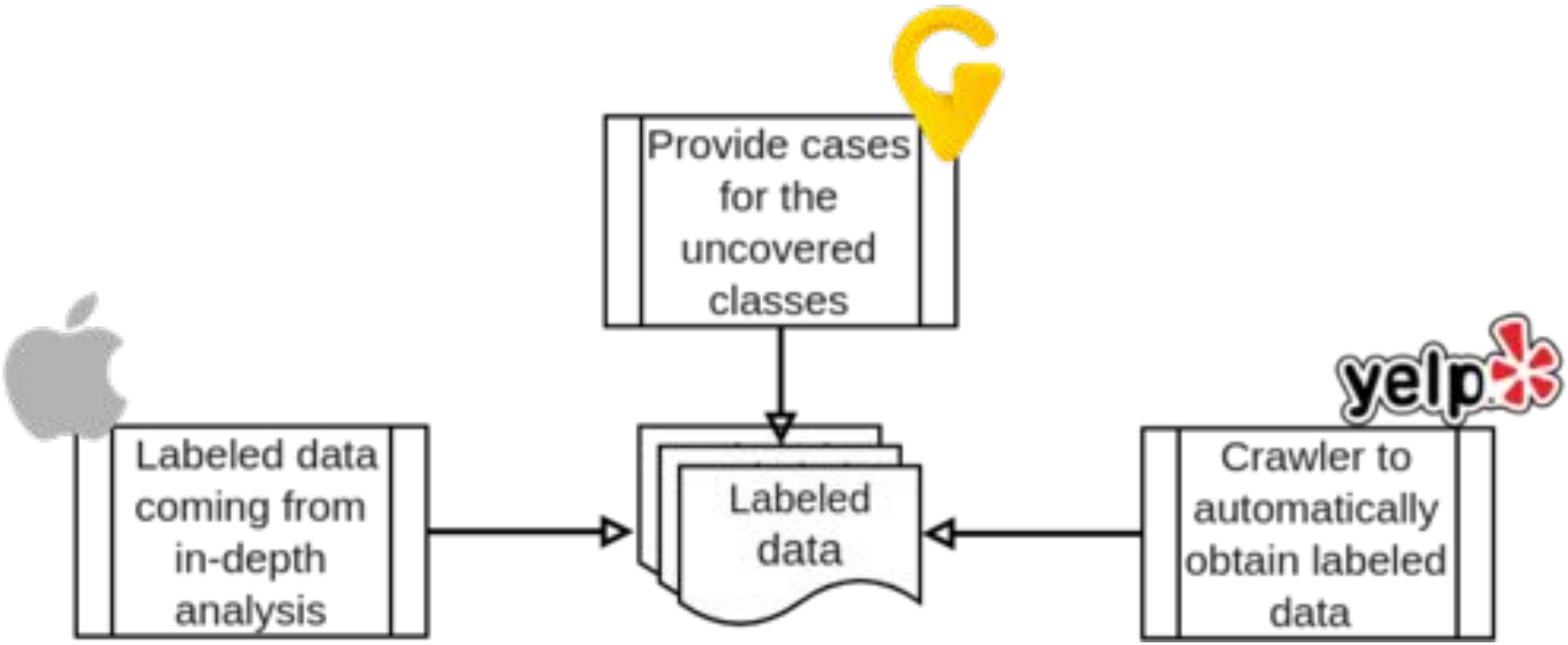
Related work

- Ontology mapping (Thor at al.)
  - metadata-based
  - **instance-based** (similarity metrics)
  - mixed forms
- Ontology Matching Website
  - Boost to minimize training set (Kejriwal et al.)
- Ontology Alignment Evaluation Initiative (OAEI)
  - Multi-lingual (Destro et al)

TACOMA
Key points

- Reformulate as a **classification** problem
- **Mixed form: instance** information and target **taxonomy metadata** are used to train the classifier
- **Generalize the target category** when the prediction confidence of the classifier is below the threshold

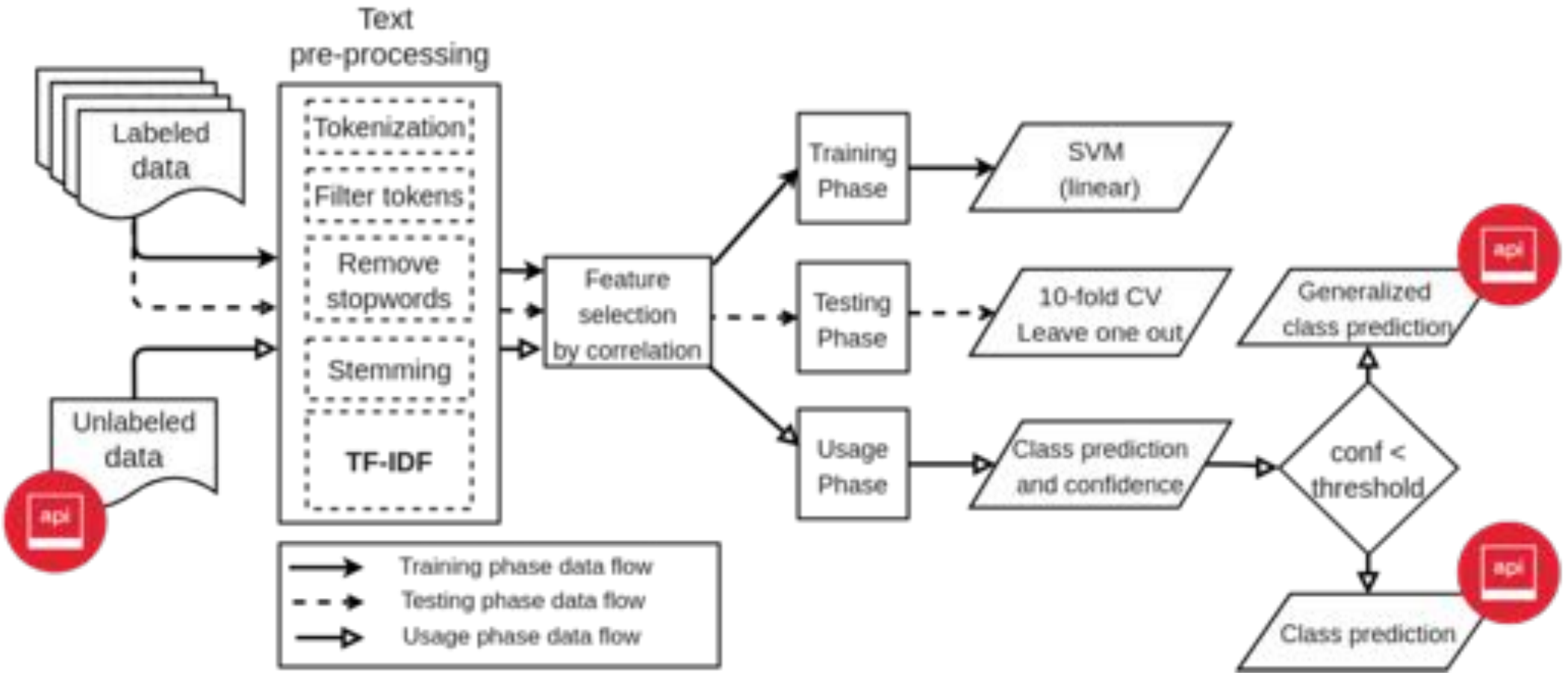
Dataset building process



Labeled record example

Attribute	Value
Label	restaurants.creperies
Name	Crispus
Description	Located in the town of Alberobello ... it is a place where you can stay until late at night for a tasty snack.
Activities	"dinner aperitif"
Specialties	"main dishes" / "traditional cuisine" / "homemade desserts"
Services	-
Products	"sandwiches" / "pizza" / "ice cream"
Brands	-

TACOMA Architecture



Experiments: datasets

Dataset	Level	Number of classes	Samples per class	Entries
restaurants	1	25	17	425
restaurants.italian	2	15	17	255
food	1	32	17	544
shopping.fashion	1	15	17	255

TACOMA performance

Class	Level	Precision	Recall
restaurants.seafood	1	93,75%	88,24%
restaurants.asianfusion	1	86,67%	76,47%
restaurants.italian	1	56,67%	94,44%
restaurants.italian.lumbard	2	61,54%	94,12%
restaurants.italian.napoletana	2	93,33%	82,35%
restaurants.italian.sardinian	2	89,47%	100%

Summary

The concept of data generalization applied to data mining techniques has been explored to extract information at different levels of granularity through taxonomies built on top of data

- Future work
  - **BELL**
    - applied to free floating bike sharing systems
  - **TACOMA**
    - usage of techniques to decrease the construction of the dataset
    - Improve the overall accuracy

Thank you for your attention