POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXXI cycle

Supervisor

Prof. Silvia Chiusano

Mining Data on Innovative Application Domains

PhD Candidate: *Giuseppe Ricupero*

1. Introduction

In the last few years, the use of Information and Communication Technologies has made available a huge amount of data in various complex application domains. For example, in the urban scenario, Internet of Things (IoT) systems generate and capture massive heterogeneous data collections describing human mobility, citizen's perception of provided services, and the overall urban environment as in terms of air quality and weather conditions.

These collections can be a valuable instrument to provide more convenient services and a better environments. However, their dimension and heterogeneousness limit the feasibility of an analysis based on the standard data mining techniques.

The research activity aims at creating ***novel data mining frameworks and patterns*** which leverage the concept of ***generalization*** to extract valuable information at different levels of granularity from the above-mentioned data with the aim of more efficiently and effectively mine useful insights.

2. Methods and results

Two fields of application have been considered as reference use cases: ***urban data*** coming from sensors, e.g. air pollution [1, 2], traffic [1], meteorological data [1], bike sharing services [4], and ***business data*** coming from online retail [3], e.g., Amazon, and web directory, e.g., Pagine Gialle.

Applications in Urban Context

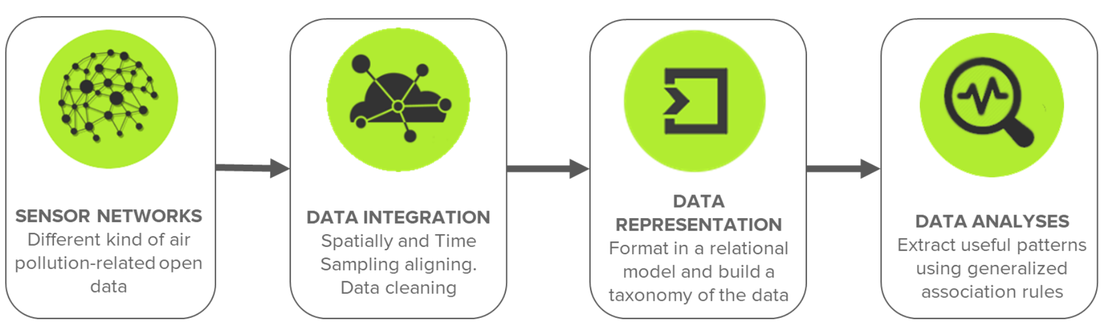

Figure 1. Architecture of GECkO framework.

The study [1] presents a data mining system, named ***GEneralized Correlation analyzer of pOllution data (GECkO)***, to extract interpretable correlations, at different abstraction levels, among a large variety of data related to air quality. Pollutant measurements are first integrated with traffic and meteorological data and enriched with an analyst-provided ***taxonomy***, which aggregates measurement values into the corresponding higher-level categories. Then, an established generalized association rule mining algorithm is applied to the prepared dataset. The extracted rules, namely the ***generalized association rules***, represent frequent co-occurrences between pollutant levels and environmental conditions at different abstraction levels (Fig. 2). The GECkO system (Fig. 1) was ***validated on real data*** collected in Milan (Italy).

In another study treating the same field [2], a data mining engine called ***AiR QUAlity patTern Analyzer (ARQUATA)*** has been designed to discover air quality patterns from air pollution-related data using a tailored ***frequent weighted itemsets*** approach.

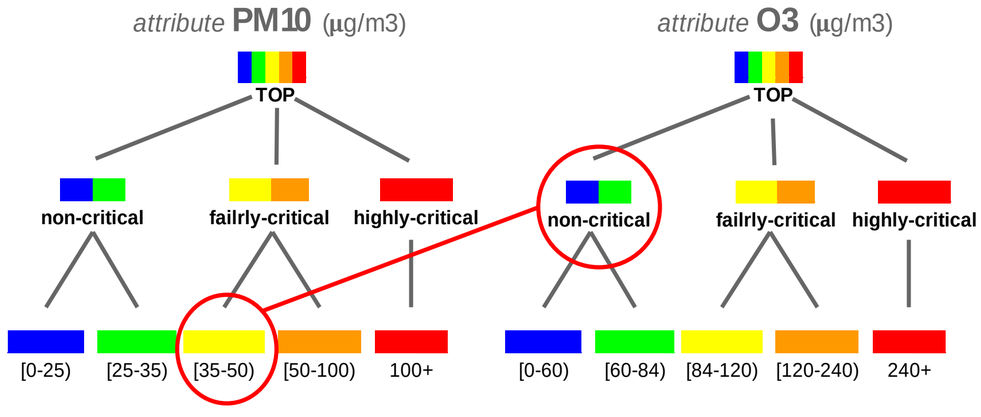

Figure 2. An example of cross-level generalized association rules.

In ***bike sharing systems***, bicycles are rented in a station and returned, usually after a short time, to any other station with free docks. To achieve a satisfactory user experience, all the stations in the system monitored through IoT devices must be neither overloaded nor empty. The goal of the study [4] was to analyze occupancy level data to discover situations of dock overload in stations belonging to the same spatial area and time interval.

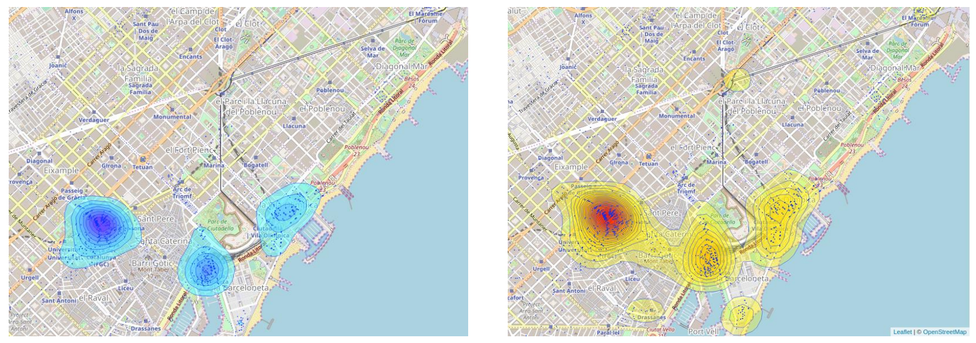

Figure 3. Heat maps representing intermittence (left) and criticality (right) situations in Barcelona at the hourly time slot [11am-12am), maxdist=0.5km, and time slot size=1h

A methodology that relies on mining a new pattern type, called ***Occupancy Monitoring Pattern (OMPs)***, is introduced to characterize situations of dock overload in multiple stations. OMPs model two complementary situations (Fig. 3): (i) ***intermittent situations*** in which a set of stations are overloaded in an alternate fashion and (ii) ***critical situations*** in which the docks of a set of stations are frequently overloaded at the same time. The real open data of two bike-sharing services (i.e., *Bicing* in Barcelona and *Citi Bike* in New York) were used to validate the approach. The results can improve the service exploitation for the end user by suggesting alternative stations in the neighborhood and support system managers in rebalancing bikes between stations.

Applications in Business Context

The contribution [3] includes the development of a novel pattern, named ***Generalized High-Utility Itemset (GHUI)***, which allows to discover recurrent combinations of items characterized averagely by a high profit from transactional data sets. Due to a taxonomy built on top of the analyzed data, the patterns are extracted at different abstraction levels. E.g., not only {Coke, Bread} but also {Beverage, Food} (Fig. 4). A novel algorithm, called ML-HUI Miner, has been defined to mine GHUIs in a single-phase mining session.

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

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

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

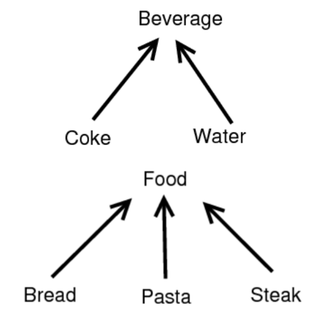

Figure 4. An example GHUI (left) and its taxonomy (right) in retail domain.

Another study defines a ***classification model*** to support the integration of business activities among different ***web directories*** (e.g., Pagine Gialle, Google Maps, Facebook pages) characterized by ***taxonomies of different granularity levels***. For instance, a generic restaurant category, which lacks any specification, cannot be mapped to a finer level of granularity (Fig. 5), such as a Chinese or Italian restaurant. The goal is to provide the most cohesive set of categories from different taxonomies corresponding to one and the same business activity.

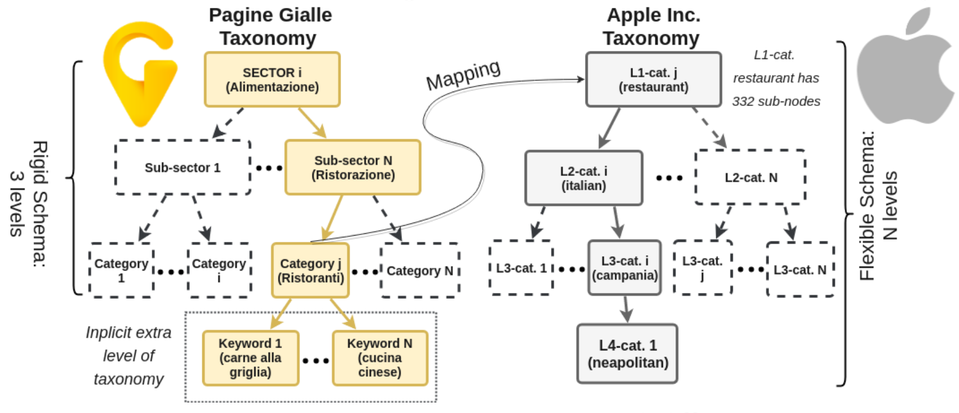

Figure 5. Comparison among two taxonomies with different granularity levels

3. Conclusion

The proposed approaches proved to be effective to get useful knowledge from heterogeneous data in complex urban and business application domains. The results confirm that the use of the generalization concept allows to discover patterns and relationships at a higher abstraction level otherwise unavailable.

4. References

[1] Cagliero L., Cerquitelli T., Chiusano S., Garza P.**,** **Ricupero G.**, Xiao X. (2016) Modeling correlations among air pollution-related data through generalized association rules. SSC 2016 (SmartComp 2016), St. Louis (USA), pp. 1-6

[2] Cagliero L., Cerquitelli T., Chiusano S., Garza P., **Ricupero G.** (2016) Discovering air quality patterns in urban environments. UbiComp 2016, Heidelberg (Germany), pp. 25-28

[3] Cagliero L., Chiusano S., Garza P., **Ricupero G.** (2017) Discovering High-Utility Itemsets at Multiple Abstraction Levels. DaS 2017 (ADBIS 2017), Nicosia (Cyprus), pp. 224-234

[4] Cagliero L., Cerquitelli T., Chiusano S., Garza P., **Ricupero G.**, Baralis, E. (2018) Characterizing situations of dock overload in bicycle sharing stations. Submitted to Applied Sciences.