POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXX cycle

Supervisor

Prof. Silvia Chiusano

Prof. Tania Cerquitelli

Mining Heterogeneous Urban Data at Multiple Granularity Layers

PhD Candidate: *Antonio Attanasio*

1. Introduction

In the last few years, the capability to both generate and collect data of public interest within urban areas has increased at an unprecedented rate, to such an extent that data rapidly scale towards big urban data. The abundance of information collected through ad-hoc sensor networks in the city context provides a remarkable opportunity to tackle interesting urban challenges and to add intelligence in the urban environment. However, for each data source and type, *different spatial and temporal references* are potentially used. Moreover, the *high volume and heterogeneity of data* increases the complexity of the analysis and new suitable algorithms should be devised. When massive volumes of data are considered, alternative *efficient data storage and processing technologies* are required.

The research activity aims at the design and development of innovative data mining solutions and it focuses on *data analysis algorithms*, suitable to mine useful insights by *exploring large and heterogeneous data collections*, deployed on *cloud-based platforms* to *guarantee good performance* in the mining process [4]. The PhD thesis proposes novel data mining algorithms and it enhances the existing ones by fitting them to novel big data architectures. The adoption of high performance and cloud computing infrastructures is considered to leverage on parallel computations. Real datasets are used to assess the proposed approaches.

2. Methods and results

In the PhD research, data mining solutions have been studied and developed to address the following issues.

Exploratory analysis of heterogeneous urban data

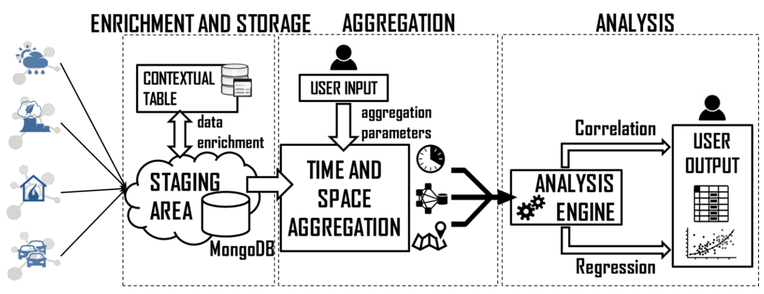

Figure 1. Architecture of the BI2CITY engine with space-time data aggregation framework.

A distributed business intelligence engine (BI2CITY) has been developed to efficiently support the integration and analysis of huge and various data collections. BI2CITY stores fine-grained data collected in the urban area. Then, it computes *the temporal and spatial data aggregation on the fly* to transpose the original data into the proper resolution as required by the analysis performed by the user. BI2CITY allows to *correlate different urban data* (weather, pollution, traffic, energy consumption) and to *forecast their expected values* for every spatio-temporal aggregation level. To efficiently deal with huge data collections, BI2CITY is based on *MongoDB* and the *MapReduce* paradigm. BI2CITY has been suitably applied to different use cases, where many types of data are used to provide insightful descriptive and predictive analyses. It has been validated on real data collected in a major Italian city [5].

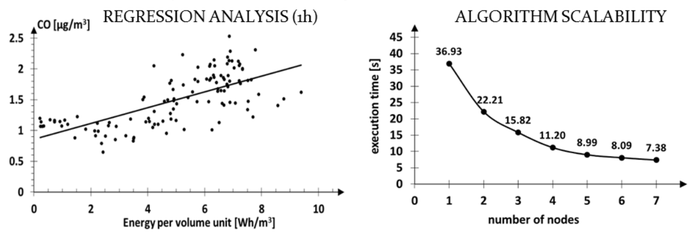

Figure 2. Regression analysis at 1 hour resolution and algorithm execution time vs number of nodes.

Buildings measured and asset energy rating

A distributed *Energy Signature Analysis* (ESA) system has been developed to efficiently compute the performance of a set of buildings (*measured rating*) in *near-real time* and to *increase energy awareness* [1] by informing users on their energy consumption. ESA collects energy data from smart meters deployed in thousands of buildings, indoor climate conditions from sensors, and other contextual data. *Enriched data* are modeled into a *document-oriented distributed data warehouse* including a MapReduce engine based on MongoDB over a cluster of 5 nodes. To characterize the performances of real buildings through energy signature, ESA performs a distributed computation, with the implemented algorithm scaling roughly linearly with the number of nodes [2].

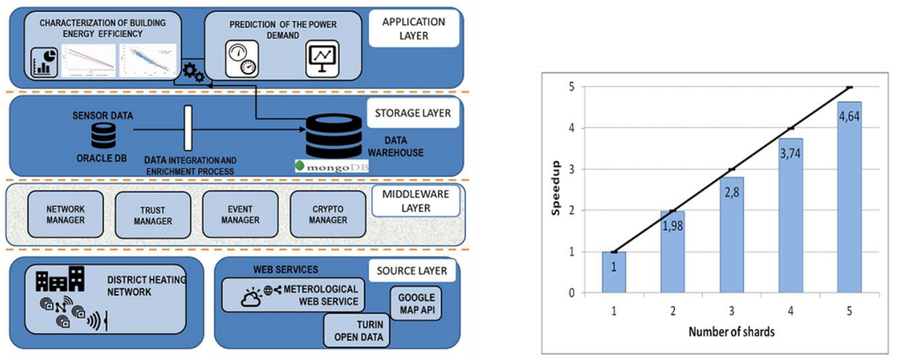

Figure 3. Architecture of the ESA system and algorithm speedup vs number of nodes.

Another relevant research topic concerns the energy performance rating of buildings according to their physical and structural features (*asset rating*). A methodology has been defined here to estimate the *thermal energy demand* of new buildings and to extract the principal rules that contribute to determine it.

Spatio-temporal characterization of people's interests from social networks

Data from social networks can provide decision makers with useful information to better understand people's interests and reactions [3]. The *Tweets Characterization Methodology* (TCharM) framework has been developed to explore collections of Twitter posts along three dimensions - *text content, posting time and place* – and to support context-aware topic trend analysis. TCharM is based on *cluster analysis*, to identify cohesive groups of tweets, and on *association rule analysis*, to describe clusters with significant patterns. To scale up to larger datasets, TCharM runs on Apache Spark and is deployed on a cluster of 8 nodes.

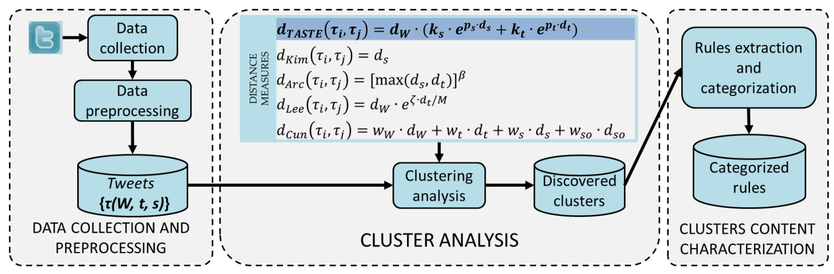

Figure 4. Architecture of TCharM.

TCharM integrates a novel distance measure for clustering (TASTE) which combines all three tweet features in one step. Moreover, TCharM integrates other distance measures proposed in literature which combine the three features, or a subset of them. The analytical comparison demonstrates that TASTE extracts clusters with a cohesion better balanced between the three tweet dimensions [6].

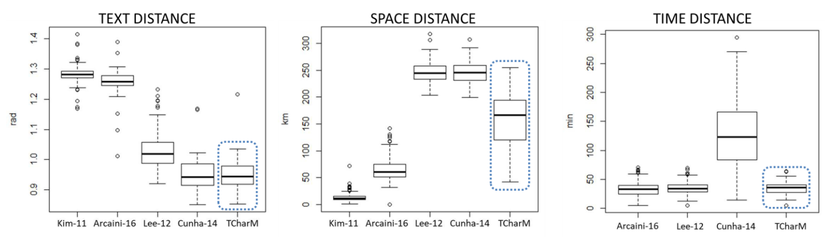

Figure 5. Average intra-cluster distances for the analyzed distance measures.

3. Conclusion

The proposed data mining approaches proved to be effective solutions to get useful knowledge from heterogeneous data in complex urban application domains. The described results confirm the importance of analyzing data with suitable granularity levels, in order to extract patterns and relationships among variables that are significant for the purposes of the analysis. Moreover, the implemented algorithms exhibited a good scalability with big datasets as well.

4. References

[1] Acquaviva, A.; Apiletti, D.; Attanasio, A.; Baralis, E.; Boni Castagnetti, F.; Cerquitelli, T.; Chiusano, S.; Macii, E.; Martellacci, D.; Patti, E. (2015), Enhancing energy awareness through the analysis of thermal energy consumption. In: The 4th workshop on Energy Data Management, Brussels, March 27th 2015. pp. 64-71

[2] Acquaviva, A.; Apiletti, D; Attanasio, A; Baralis, E; Bottaccioli, L; Boni Castagnetti, F.; Cerquitelli, T.; Chiusano, S.; Macii, E.; Martellacci, D.; Patti, E. (2015), Energy Signature Analysis: Knowledge at Your Fingertips. In: IEEE International Congress on Big Data 2015, New York, USA, 27 June 2015 - 2 July 2015. pp. 543-550

[3] Attanasio, A.; Jallet, L.; Lotito, A.; Osella, M.; Ruà, F.; (2015), Fast and Effective Decision Support for Crisis Management by the Analysis of People's Reactions Collected from Twitter. In: 2nd International Workshop on Big Data Applications and Principles (BigDap 2015), Poitiers, September 8-11, 2015. pp. 229-234

[4] Tosatto, A.; Ruiu, P.; Attanasio, A. (2015), Container-Based Orchestration in Cloud: State of the Art and Challenges. In: 2015 Ninth International Conference on Complex, Intelligent, and Software Intensive Systems. pp. 70-75

[5] Attanasio, A.; Cerquitelli, T.; Chiusano, S. (2016), Supporting the analysis of urban data through NOSQL technologies. In: The 7th International Conference on Information, Intelligence, Systems and Applications, Chalkidiki, Greece, 13-15 July, 2016. pp. 1-6

[6] Xin, Xiao; Antonio, Attanasio; Silvia, Chiusano; Tania, Cerquitelli (2017), Twitter data laid almost bare: An insightful exploratory analyser. In: EXPERT SYSTEMS WITH APPLICATIONS, vol. 90, pp. 501-517.