PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Vincenzo Schiano Di Cola

XXIX Cycle

Training and Research Activities Report – Second Year

Tutors: Nicola Mazzocca, Francesco Piccialli
co-Tutor (company): Paolo Benedusi

1. Information

This report is written by Vincenzo Schiano Di Cola. I received my master's degree in Mathematics from this same University. I belong to the XXIX Ph.D. Cycle of the ITEE at the Università di Napoli Federico II. I received a regional fellowship, the P.O.R. entitled: *Data Scientist for Predictive Analytics*; according to the "*Dottorati di ricerca con Caratterizzazione Industriale*" - DGR n. 156 del 21/03/2017 - DD n. 155 del 17/05/2018 - within the *POR Campania FSE 2014/2020 - Obiettivo Specifico 14 - Azione 10.4.5***.** The tutors who are supervising me are Prof. Nicola Mazzocca and Prof. Francesco Piccialli, and in this second year, I have been tutored by Ing. Paolo Benedusi, on behalf of the company DATABOOZ ITALIA S.r.l.

2. Study and Training activities

During this second year, I obtained a number of study and training credits summing up, roughly, to 26 ECTS. The credits are distributed among seminars and courses, both taken either online and within Naples. As advised by the ITEE guidelines, my training activity focused on the know-how, rather than knowledge, by the ad hoc teaching offered by the doctorate.
The courses are here listed, with their title, type, and number of credits:

- Virtualization technologies and their applications; Ad hoc module; 4
- Innovation management, entrepreneurship and intellectual property; Ad hoc course; 5
- 3rd Advanced Course on Data Science & Machine Learning (ACDL) 2020; Summer School; 7
- Oxford Machine Learning Summer School (OxML 2020); Summer School; 4

The seminars are listed, with their title, type, and number of credits:

- How to get published with the IEEE?; Seminar; 0,4
- IEEE: Access the eLearning Library; IEEE Xplore Webinar; 0,2
- Large Scale Training of Deep Neural Networks, Seminar; 0,5
- SAS Analytics; Seminar; 0,4
- Noninvasive Mapping of Electrical Properties using MRI; Seminar;0,3
- Enterprise Blockchain for Healthcare; IEEE online Seminar; 0,3
- Reasoning Web Summer School -RW 2020; Summer School; 3,2
- How to Publish Open Access with IEEE to Increase the Exposure and Impact of Your Research; IEEE Webinar; 0,3

3. Research activity

The POR project involves three main participants is enriching my research with their specific objectives and expertise, within the more general framework of Data Science. These three members of the POR fellowship are the DIETI at University of Naples "Federico II", the IT company DATABOOZ ITALIA S.r.l., and the Chung-Ang University in South Korea.

This second-year research continued the research started in the first year related to Artificial Intelligence techniques for IoT in the Cultural Heritage domain with a focus on Knowledge Graphs (KG) from a corporate point of view.

The research questions I tried to answer were:

1. Which tools are best for modeling KG and quickly build a knowledge base?
2. How to extract insight from KG, and use complex applications (entity resolution, link prediction, pattern discovery)?
3. How to correctly apply Machine Learning techniques to KG?
4. How to calculate all the possible distances of users paths in a museum?
5. How to best show statistical properties of users in a museum to the final stakeholders?

Coherently with the nature of Data Science, as the intersection of math, IT and domain knowledge, most of the collaborations have been pursued with the Department of Mathematics and Applications (DMA) "Renato Caccioppoli", and with CINI, the Consorzio Interuniversitario Nazionale per l'Informatica.

In the following, there will be a description of the research developed during this second year, being focused on a business aspect, thanks to DATABOOZ ITALIA S.r.l., with the study of tools for BI (Microsoft Power BI), and the involvement in the project *deliverables*.

I worked on Health care booking data, by using a legacy dataset developed in the 1980s to search insights with AI solutions. The data came from booking data of the Local Health Department of Naples, data of patients booking their health service with a prescription, referred by their practitioner. The main goals to tackle were to apply AI solutions on legacy data (i.e., gather insights with machine learning from that data) then generate a KG form tuple data, since the data was in the form of a legacy SQL-like database without any *Entity-relation schema*; and finally to real-time analytics and prediction. These goals were given because booking processes are widely used, the approach can be mimicked in different other examples and data. And since there were medical prescriptions, this data could be used to predict what kind of health service a patient might be needed, and also predict the load, and KPI of health service providers, by changing the point of view of the predicting subject.

In my research, I analyzed different solutions and applications for KG like Neo4j, OrientDB, JanusGraph, and selected *Grakn*. I then worked on the data storing in a graph database, and then the graph data embedding phase. Since the data were not structured into an Entity-relation database and were in a tabular form; an in-depth study of the relationships between pairs of rows sharing the same values on one or more field has been performed. This heterogeneous kind of data presents much inconsistency, given the non-entity-relation structure. By observing that, when rows in a table share the same value on a fixed field, they might be linked to the same relation or entity. *Entity disambiguation* has been performed by counting the unique occurrences of couples of columns in the database in order to have an overall view of what pair of fields could be considered belonging or not to an entity or relation. This process has been refined with some domain experts, in order to obtain the final knowledge graph schema. We have worked in strict contact, with doctors, medical staff and the IT support system, to better gather insights from the available data. Grakn was useful to model a KG because of fo the multidimensionality of the data relations. I also added further meta relations with rules, that simplified the transformation of a SQL database to a KG.

The last step to is to turn the data into some kind of structured data that will be a viable input for our machine learning algorithms. Once we obtain this data, then we feed it into our unsupervised and supervised learning algorithms. What we will want in reality and what will make our life much easier is to avoid the feature engineering phase. So we want, basically, to automatically learn the features. This process is called

Embedding. The goal is to map each element of a graph in a low dimensional space. The challenge is to choose the number of dimensions, because given the size of this space, the algorithm will be able to squeeze there more or less information. In literature there exist many Knowledge Graph Embedding algorithms and selecting the most appropriate for a given dataset is an important challenge. One of the most simple is the TransE embedding. It's based on the idea that in a triple one entity is moved by the relation to another entity. So this algorithm optimizes the alignment of all the entities based on the relations. For this reason, it's a Translational embedding algorithm. Another kind of embedding algorithm is bilinear embedding. It's based on the decomposition of the tensor space. Some examples are CompleEx, DistMult, RESCAL. It is not always straightforward to obtain a high-quality model, given enough data to get useful and suitable insights. Each Knowledge Graph Embedding algorithm has different input parameters to be set up. In this perspective, we applied a random grid search to find the best choice for our data. The one that gave us the best results was ComplEx, but more research are still needed. Finally, to visualize the embedding space we projected all the entities to a 2Dimensional space with two different algorithms t-SNE and PCA.

In the following months, I worked in close contact with DATABOOZ, by being involved the research projects Cetra (http://www.cetra-project.it/) and Remian. We analysed the possibility to model sematic paths with Grakn, integrated with more company oriented tools, like Hortonworks HDP, Cloudera DataFlow (Ambari) and Superset. This research lead to generate a model in Grakn that could describe multiple kind of museum visitors, tracked with multiple devices. The data that we worked with came from two Neapolitan museums: MANN and Castel Nuovo. The National Archaeological Museum of Naples, which had installed IoT sensors so that we were able to track the Bluetooth devices of visitors. While the other Musume, the Castel Nuovo, has recorded the usage of an application installed on a tablet, that visitors could use in order to enhance their experience in the museum.

Since, Data Science processes, usually requires to organize, clean the data, and structure the data in a way that can be used by the corresponding algorithm, I investigated how is to remodel the data as a Knowledge Graph and use such modeling to apply graph techniques, like Node importance and Community detection. The two promising technological solutions so to create this framework, have been: Grakn, an Hyper relation database and AmpliGraph for KG embedding, based on Tensorflow. Still, there is no overall analysis about embedding that describes how well is the embedding suited for different kinds of data and the relations among those data and the Graph structure, and the kind of mathematical properties that the embedding preserves. Some tests of Machine Learning on embedded values might give a result that can be generalizable. Other possible research involve investigating further embedding algorithms and a mixture of embedding models (like ConvE with RotatE). The final model for the museum includes parameters for distances among Point Of Interests of a museum, calculated trough R, and statistical informations about users behaviours were presented with PowerBi.

Finally, at the and of the second year, two months were spent in analysing literature and writing a perspective on Deep Learning used in the COVID-19 pandemic.

As for further studies. There is still research to do within Cultural Heritage domani like: spatially model the nodes of the museum, identification of groups in the museum, and cluster of kind of days. In the contex of KG, future research can be carried trough: Node classification, Community detection and Link prediction. A challenging topics will be, also, to investigate Explainable AI (XAI), Generative Adversarial Networks (GAN) and Reinforcement Learning (RL). Any one of these topics will be investigated trough different applications based on the kind of data that will be available during the third year of research.

4. Products

This second year of research lead to three publications:

1. “Unsupervised learning on multimedia data: a Cultural Heritage case study” on the *Multimedia Tools and Applications*. Published: 14 March 2020. DOI: 10.1007/s11042-020-08781-1
2. “Path prediction in IoT systems through Markov Chain algorithm” on *Future Generation Computer Systems*. Available online: 6 April 2020. DOI: 10.1016/j.future.2020.03.053
3. “A Deep Learning approach for Path Prediction in a Location-based IoT system” on *Pervasive and Mobile Computing*. Available online: 17 June 2020. DOI: 10.1016/j.pmcj.2020.101210

Others are in preparation and one paper is under review at Elsevier’s Journal *Information Fusion*, which is “A Knowledge Graph approach for the Insights Extraction of e-health Data”. Further papers will, eventually, be on the most recent research developments.

5. Conferences and Seminars

I presented my work at a conference and at a summer school.

I presented my work with GRAKN, “Insights Extraction Through a Knowledge Graph of Medical Booking Data” in February at *Grakn Cosmos: The Universe of Orderly Systems*, together with my colleague G.Casolla.

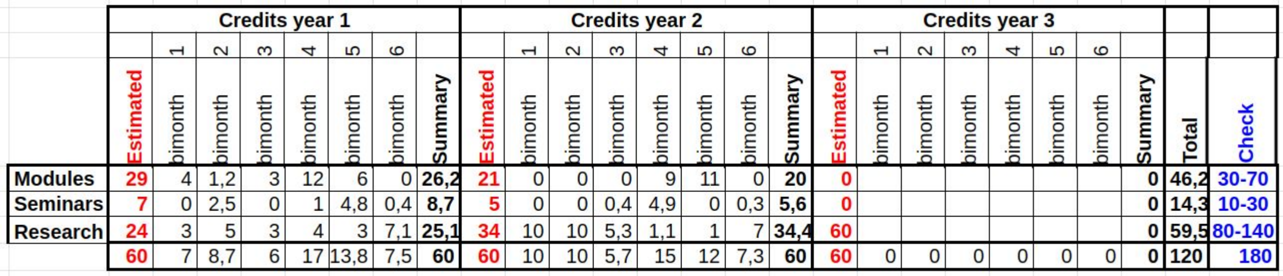
My most recent results were presented in July at the 3rd Advanced Course on Data Science &Machine Learning – ACDL 2020, with the title of: “Knowledge Graph modeling for IoT data in Cultural Heritage framework”

6. Activity abroad

For this second year, there has been no activity aboard since, within the POR scholarship, this was the year devoted to the indultral collaborations. Next year, will be spent aboard in best condition that the COVID-19 pandemic will allow.

7. Tutorship

No mentoring has been done, since my POR fellowship is on a tight schedule, and each hour is reported and accounted for the Campania Region.



	Credits year 1								Credits year 2								Credits year 3									
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6			
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Total	Check
Modules	29	4	1,2	3	12	6	0	26,2	21	0	0	0	9	11	0	20	0							0	46,2	30-70
Seminars	7	0	2,5	0	1	4,8	0,4	8,7	5	0	0	0,4	4,9	0	0,3	5,6	0							0	14,3	10-30
Research	24	3	5	3	4	3	7,1	25,1	34	10	10	5,3	1,1	1	7	34,4	60							0	59,5	80-140
	60	7	8,7	6	17	13,8	7,5	60	60	10	10	5,7	15	12	7,3	60	60	0	0	0	0	0	0	0	120	180