PhD Student: Luigi Libero Lucio Starace

Cycle: XXXV

Training and Research Activities Report

Academic year: 2020-21 - PhD Year: Second

Luigi Libero Lucio Storace

Tutor: Prof. Sergio Di Martino

Co-Tutor: Prof. Adriano Peron

Date: October 21, 2021

Training and Research Activities Report

PhD program in Information Technology and Electrical Engineering

PhD student: Luigi Libero Lucio Starace   **Cycle: XXXV**

_____

1. Information:

➢ **PhD student: Luigi Libero Lucio Starace**   **PhD Cycle: XXXV**

➢ **DR number: DR993893**

➢ **Date of birth: 25/02/1991**

➢ **Master Science degree: Computer Science**   **University: UNINA**

➢ **Scholarship type: Funded by NetCom Group S.p.A.**

➢ **Tutor: Prof. Sergio Di Martino**

➢ **Co-tutor: Prof. Adriano Peron**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
Robot Manipulation and Control	Seminar	2.5	0.5	17/11/2020	Proff. Paolo Dario and Cesare Stefanini	Y
Digital Project Management: practices, processes, techniques, tools, and scientific approach.	Seminar	1	0.2	18/11/2020	Part of the Picariello Lectures on Data Science	Y
Science, Reality and Credibility.	Seminar	1.5	0.3	24/11/2020	FuturoRemoto2020	N
#andràtuttobene: Images, Texts, Emojis and Geodata in a Sentiment Analysis Pipeline	Seminar	1.5	0.3	25/11/2020	Part of the Picariello Lectures on Data Science	Y
Patent Searching Best Practices with IEEE Xplore	Seminar	1	0.2	27/11/2020	Dr. Alessandra Scippa	Y
JISET: JavaScript IR-based Semantics Extraction Toolchain	Seminar	1	0.2	01/12/2020	Facebook Testing and Verification Symposium	N
Automatic Generation of Optimal Mutant Reduction Strategies	Seminar	1	0.2	01/12/2020	Facebook Testing and Verification Symposium	N

Testing Web Enabled Simulation at Scale using Metamorphic Testing	Seminar	1	0.2	02/12/2020	Facebook Testing and Verification Symposium	N
Metamorphic Testing of Android Graphics Drivers	Seminar	1	0.2	02/12/2020	Facebook Testing and Verification Symposium	N
Software robustness: A survey, a theory, and some prospects	Seminar	1	0.2	02/12/2020	Facebook Testing and Verification Symposium	N
Using Developers' Intuitions to Improve Property-Based Tests	Seminar	1	0.2	03/12/2020	Facebook Testing and Verification Symposium	N
CRISPR-cas9 screens and multi-omic data integration for identifying and prioritising anti-cancer therapeutic	Seminar	1.5	0.3	04/12/2020	Computational and Quantitative Biology Ph.D. Program (DIETI)	N
Exploiting Deep Learning and Probabilistic Modeling for Behaviour Analytics	Seminar	1	0.2	09/12/2020	Part of the Picariello Lectures on Data Science	Y
GDPR basics for computer scientists	Seminar	1.5	0.3	10/12/2020	Prof. Bonatti (DIETI)	Y
Data Driven Transformation in WINDTRE through Managers voice	Seminar	2	0.4	16/12/2020	Part of the Picariello Lectures on Data Science	Y
Computer Forensics' methods, practices, and tools	Ad-hoc Course	10	3	Credits acquired:18 /01/2021	ITEE/DIETI	Y

Training and Research Activities Report

PhD program in Information Technology and Electrical Engineering

PhD student: Luigi Libero Lucio Starace   **Cycle: XXXV**

From Photometric Redshifts to Improved Weather Forecasts: an interdisciplinary view on machine learning	Seminar	1	0.2	13/01/2020	Part of the Picariello Lectures on Data Science	Y
Cybercrime and e-evidence: the criminal justice response	Seminar	2	0.4	20/01/2021	Part of the Picariello Lectures on Data Science	Y
AI: Artificial Intelligence for notary's sector - a case study	Seminar	1	0.2	27/01/2021	Part of the Picariello Lectures on Data Science	Y
The era of Industry 4.0: new frontiers in business model innovation	Seminar	1	0.2	03/02/2021	Part of the Picariello Lectures on Data Science	Y
Machine learning: causality lost in translation	Seminar	1.5	0.3	10/02/2021	Part of the Picariello Lectures on Data Science	Y
Signature reversion and other computational strategies for identifying drug repositioning opportunities	Seminar	1.5	0.3	19/02/2021	Organized by the Computational and Quantitative Biology Ph.D. Program	N
Supporting machine learning with biological knowledge to extract insight from omics data	Seminar	1	0.2	26/02/2021	Organized by the Computational and Quantitative Biology Ph.D. Program	N
Data Science for Patient Records Analysis	Ad-hoc Course	10	2.5	10.02.2021 - 17.03.2021	ITEE/DIETI	Y

Training and Research Activities Report

PhD student: Luigi Libero Lucio Starace   **Cycle: XXXV**

Statistical data analysis for science and engineering research	Ad-hoc Course	12	4	17.02.2021 – 04.03.2021	ITEE/DIETI	Y
Scientific programming and visualization with Python	Ad-hoc Course	10	2	08-10.03.2021	ITEE/DIETI	Y
Robo Ludens: A game design taxonomy for human-robot interaction	Seminar	1	0.2	05/03/2021	Professor Silvia Rossi, PRISCA Lab. - DIETI	Y
Dai Mainframe all'IoT: una retrospettiva sull'evoluzione delle architetture di calcolo	Seminar	1	0.2	08/03/2021	Professor Alessandro Cilardo, DIETI	Y
Big Data and Computational Linguistics	Seminar	2	0.4	10/03/2021	Part of the Picariello Lectures on Data Science	Y
The coming revolution of Data driven Discovery (a fourth Methodological Paradigm of Science)	Seminar	1.5	0.3	25/03/2021	Part of the Picariello Lectures on Data Science	Y
IEEE Authorship and Open Access Symposium: Best Practices to Get Published to Increase the Exposure and Impact of Your Research	Seminar	1.5	0.3	21/04/2021	Eszter Lukács, IEEE	Y
Artificial Intelligence and 5G combined with holographic technology: a new perspective for remote health monitoring	Seminar	2	0.4	27/04/2021	Prof. Antonia Tulino, DIETI	Y

Training and Research Activities Report

PhD program in Information Technology and Electrical Engineering

PhD student: Luigi Libero Lucio Starace **Cycle: XXXV**

_____

Distributional Semantics Methods: How Linguistic features can improve the semantic representation	Seminar	1	0.2	28/04/2021	Part of the Picariello Lectures on Data Science	N
Risk assessment in real life: experiences from the railway domain	Seminar	1.5	0.3	26/05/2021	Prof. Valeria Vittorini – DIETI	Y
'Sadas Engine', an innovative DBMS for the DATA WAREHOUSE, great performance in the VLDB environment	Seminar	1.5	0.3	23/06/2021	Part of the Picariello Lectures on Data Science	Y
L'esposizione ai campi elettromagnetici generati dal sistema 5G - Metodologie scalari e vettoriali di misura dell'esposizione e tecniche di estrapolazione	Seminar	4	0.8	16/07/2021	Prof. Nicola Pasquino (DIETI)	Y
SAE 2021 - Big4small, Data Science Methodology Transfer: Big to Small	Seminar	2	0.4	24/09/2021	Session of the SAE2021 conference, chaired by Prof. Giuseppe Longo	N
Analyzing and Supporting the Evolution of Data-Intensive Systems	Seminar	1	0.2	07/10/2021	Prof. Paolo Tonella, Software Institute, USI	N
Visualizing Discord Servers	Seminar	1	0.2	14/10/2021	Prof. Paolo Tonella, Software Institute, USI	N
Qiskit: state of the art and tools for Quantum Computers from IBM	Seminar	2	0.4	15/10/2021	Prof. Cacciapuoti (DIETI)	N

Quality Metrics and Oracles for Autonomous Vehicles Testing	Seminar	1	0.2	21/10/2021	Prof. Paolo Tonella, Software Institute, USI	N
Second Quantum Revolution: innovation trends and expected industrial impacts	Seminar	2	0.4	22/10/2021	Prof. Cacciapuoti (DIETI)	N

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	0	3.8	7	0	10.8
Bimonth 2	3	1.8	7	0	11.8
Bimonth 3	8.5	2	7	0.4	17.9
Bimonth 4	0	0.6	6	0.24	6.84
Bimonth 5	0	0.8	6	0	6.8
Bimonth 6	0	1.8	8	0.96	10.76
Total	11.5	10.8	41	1.6	64.9
Expected (2nd year)	10 - 20	5 - 10	30 - 45	0 - 1.6	45 – 76.6

3. Research activity:

During the second year, I continued working with my research group on novel near-duplicate detection techniques to support automated model inference, and thus testing, of web applications. Indeed, state-based models of web applications are often used to support testing activities, and these models, in which each states represent a feature, are typically obtained by crawling. However, in practice, models obtained by crawling are often affected by the so-called near-duplicates, i.e. that correspond to web pages that are different, but nonetheless represent the same functionality. The example in Figure 1 depicts, Web page A and Web page B are different web pages, with different images and different text, but nonetheless they both represent the same "View book detail" functionality. The presence of near-duplicates in a model inferred by crawling has a negative impact on the precision of the model, since we are adding a number of "useless" states, and on the completeness of the model as well, since wasting time on exploring near-duplicate states could lead to not having enough time to explore other relevant parts of the application under test.
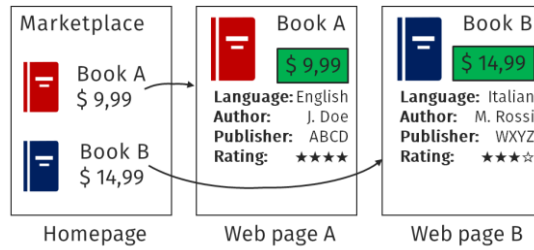
Figure 1 Example of near-duplicates

More in detail, during this year we defined a novel near-duplicate detection technique which leverages Tree Kernel functions, a class of kernel functions largely used in natural language processing, to measure the similarity between the tree-structured DOM representation of web pages. We implemented this technique into a modified version of the well-known Crawljax web crawler, and started working on the empirical evaluation of the proposed tree kernel-based technique by replicating a study that was recently presented at ICSE20 [1]. In that study, ten near-duplicate detection techniques are benchmarked on a model inference task, and the authors also provide a massive dataset of annotated web page pairs and several gold standard models for selected open-source web applications. Preliminary results, which have also been published at the *International Symposium on Empirical Software Engineering and Measurement* (ESEM'21) [2], show that the proposed tree kernel-based technique performs better than all the ten state-of-the-art techniques on a near-duplicate detection task. We are currently working on evaluating the quality of the models inferred using the proposed technique against the gold standard models provided by [1]. To this end, we are currently working on implementing a software component to compare a gold standard model, as provided in [1], with a crawl generated by our modified version of Crawljax.

3.1. References

[1] Yandrapally, Rahulkrishna, Andrea Stocco, and Ali Mesbah. "Near-duplicate detection in web app model inference." Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. 2020.

[2] Corazza, A., Di Martino, S., Peron, A., **Starace, L. L. L.** (2021). Web Application Testing: Using Tree Kernels to Detect Near-duplicate States in Automated Model Inference. Accepted for presentation at the *15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2021)*. DOI: https://doi.org/10.1145/3475716.3484187.

_____

4. Research products

4.1. Journal papers

Asprone, D., Di Martino, S., Festa, P., **Starace, L. L. L.** (2021). Vehicular crowd-sensing: a parametric routing algorithm to increase spatio-temporal road network coverage. *International Journal of Geographical Information Science* (IJGIS). Status: published. Venue indexed in Scopus/WoS. DOI: https://doi.org/10.1080/13658816.2021.1893737.

Di Martino, S., **Starace, L. L. L.** (2021). A Budget-aware Decentralized Incentivization Solution for Vehicular Crowd-Sensing. IEEE Internet of Things Journal (JIOT). Status: submitted, under revision. Venue indexed in Scopus/WoS.

4.2. Conference Papers

Corazza, A., Di Martino, S., Peron, A., **Starace, L. L. L.** (2021). Web Application Testing: Using Tree Kernels to Detect Near-duplicate States in Automated Model Inference. Accepted for presentation at the *15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2021)*. Status: published. Venue indexed in Scopus/WoS. DOI: https://doi.org/10.1145/3475716.3484187.

Benerecetti, M., Mogavero, F., Peron, A., & **Starace, L. L. L.** (2021). Expressing structural temporal properties of safety critical hierarchical systems. Presented at the *14th International Conference on the Quality of Information and Communications Technology* (QUATIC 2021). Status: published. Venue indexed in Scopus/WoS. DOI: https://doi.org/10.1007/978-3-030-85347-1_26.

Di Martino, S., **Starace, L. L. L.** (2021). Vehicular Crowd-Sensing on Complex Urban Road Networks: A Case Study in the City of Porto. Accepted for presentation at the *24th Euro Working Group on Transportation Meeting* (EWGT 2021). Status: to appear in the conference proceedings (Transportation Research Procedia). Venue indexed in Scopus/WoS.

4.3. Tools

We made available an open-source extension of the KNIME Analytics Platform, KNOT (KNime mObility Toolkit). The extension is designed to support scientists and practitioners working on knowledge discovery from mobility data, and has been acknowledged by Dr. Stefan Helfrich, manager of the KNIME Academic Alliance program. The official website for the extension is available at the address: https://luistar.github.io/knot/.

_____

5. Conferences and seminars attended

I attended the following conferences/events (in chronological order).

- *32nd IFIP International Conference on Testing Software and Systems* (ICTSS 2020). Naples, Italy (virtual event). 8.12.2020 to 10.12.2020. Attended and presented a paper titled "*Inspecting Code Churns to Prioritize Test Cases*".
- *ACM SIGSOFT International Symposium on Software Testing and Analysis* (ISSTA 2021) *Doctoral Symposium*. Aarhus, Denmark (virtual event). 13.07.2021. Attended and presented a short paper titled "*Detecting Near-duplicate States in Web Application Model Inference: a Tree Kernel-based Approach*".
- *24th Euro Working Group on Transportation Meeting* (EWGT 2021). Aveiro, Portugal (virtual event). 8.09.2021 to 10.09.2021. Presented a paper titled "*Vehicular Crowd-Sensing on Complex Urban Road Networks: A Case Study in the City of Porto*".
- *14th International Conference on the Quality of Information and Communications Technology* (QUATIC 2021). Faro, Portugal (virtual event). 8.09.2021 to 10.09.2021. Presented a paper titled "*Expressing structural temporal properties of safety critical hierarchical systems*".
- *15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (ESEM 2021). Bari, Italy (virtual event). 11.10.2021 to 15.10.2021. Presented a paper titled "*Web Application Testing: Using Tree Kernels to Detect Near-duplicate States in Automated Model Inference*".

6. Periods abroad and/or in international research institutions

I have been spending a study and research period abroad at the Università della Svizzera Italiana (USI) in Lugano, Switzerland, from 1.10.2021 to 30.10.2021. In particular, I am working at TAU Laboratory, in the USI Software Institute, under the supervision of Prof. Paolo Tonella. During this period, I am attending weekly meetings and seminars with the other members of the research group led by Prof. Tonella, and I am working with Dr. Andrea Stocco on the implementation of a software component to assess the quality of a web application model inferred by crawling against reference models available in the literature. This is a key step to evaluate the effectiveness of the near-duplicate detection techniques I am investigating. Moreover, I am also investigating the feasibility of using Deep Learning to learn a mapping from web pages to numerical vectors, which could then be used to measure the similarity between web pages in a more accurate way. In the next year, I plan to continue these activities and to spend two additional months in Lugano, from 01.11.2021 to 31.12.2021, working with the same institution and supervisor.

Overall, in my second year, I spent 1 month abroad.

7. Tutorship

I performed tutorship and teaching activities, supervised by Prof. Sergio Di Martino, in courses both within the B.Sc. in Computer Science program ("Object Orientation" and "Ingegneria del Software"), and within the M.Sc. in Computer Science program ("Software Project Management and Evolution") at the Università degli Studi di Napoli Federico II.

More in detail, within the Software Project Management and Evolution course, I prepared and presented five seminar lectures on the following topics:

- Advanced cloud-based and serverless architectures.
- Unit Testing with Junit 5, Hamcrest.
- Unit Testing in isolation (Inversion of Control, Mockito).
- Informal Introduction to Formal Methods for Software Engineers (with NuSMV demonstration).
- Mutation Testing and Metamorphic Testing.

Moreover, I also prepared and presented a seminar lecture on UML State Machines for the Software Engineering course, a seminar lecture with a practical demonstration on the git versioning tool and on Apache Maven, and assisted Prof. Sergio Di Martino in managing project assignments for the Software Engineering and the Object Orientation course. I also tutored students from the Software Project Management and Evolution course working on their final assignment.

During this second year, I also tutored 7 B.Sc. students in Computer Science in their internal curricular internship and theses activities, and 3 M.Sc. students in Computer Science working on their master's thesis projects.

8. Plan for year three

In year three, I plan to continue working on the design, development and validation of the proposed Tree Kernel-based near-duplicate detection technique. Moreover, I also plan to investigate the feasibility of using a deep learning-based approach to learn a mapping from web pages to numerical vectors, which could then be used to implement a novel near-duplicate detection technique. I plan to validate this additional technique by using the same experimental framework I am using for the Tree Kernel-based one.

I plan to work on these activities also during the two months research period abroad, at the Università della Svizzera Italiana in Lugano, Switzerland, supervised by Prof. Paolo Tonella and Dr. Andrea Stocco. In particular, I plan on staying in Lugano from November 1, 2021, to December 31, 2021.

As for my Ph.D. thesis, my plan is to generalize the proposed near-duplicate detection solutions and propose a general near-duplicate detection framework to support both the definition of novel near-duplicate detection techniques (possibly by combining multiple existing approaches), and the empirical assessment of their effectiveness.