PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Enrico Caldarola

XXIX Cycle

Training and Research Activities Report – Second Year

Tutor: Prof. Antonio Picariello – co-Tutor: Prof. Antonio Rinaldi

UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

1. Information

I graduated in Computer Science Engineering at Politecnico di Bari in 2006. I have been working as computer programmer since 2006 and now I am attending the XXIX cycle of Phd course in Intofmation Technology and Electrical Engineering, at the Department of Electrical Engineering and Information Technologies (DIETI), University of Naples "Federico II", collaborating with the group headed by Prof. Antonio Picariello, my tutor, and Prof. Antonio Rinaldi, my co-tutor.

I am also research fellow at the Institute of Industrial Technologies and Automation, National Council of Researches in Bari in the EVA group (Enterprise Engineering and Virtual Applications).

2. Study and Training activities

During the second year of my PHD, I have attended different training activities. Namely, I attended the ad hoc modules: "Models, methods and software for Optimization" by Prof. A. Sforza and C. Sterle at University of Naples, and "Semantic web reasoners", by Prof. Bonatti. I also participated at SoDATA 2015 summer school on BI and Big Data Analytics, from June 10 to 12th, 2015 in the Island of Capri, Italy and at the 2nd International Winter School on Big Data, at University of Deusto, Bilbao, Spain, from 8th to the 12th of February, 2016. Some of the courses attended at the schools follow:

- Exploring the Mysteries of our Cosmos: the Big Deal between Big Data and Big Science;
- Decision Trees for Big Data Analytics;
- Multidimensional, Spatial, and Metric Data Structures for Applications in Spatial Big Data Algorithms that Aren't Machine Learning;
- Big Data: Comparison with Computational Models, Big Data Analytics for Social Networks Large Scale Graph Analytics and Mining;
- The Evolutionary Trends of BI and Analytics: A market Perspective;
- The Evolutionary Trends of BI and Analytics: A market Perspective;
- Social Network Analysis;
- Information Retrieval and User Profiling;
- Text Classification and Machine Learning: Real-life application scenarios;
- Big Data Integration;
- How the Analytics can improve business decisions;

In addition, I attended the following seminars:

- "The iCub project: An open platform for research in robotics & artificial intelligence", by dr. Giorgio Metta, from Istituto Italiano di Tecnologia;
- "Good Ontology Design Philosophical and Empirical Perspectives" held by dr. Ludger Jansen, Münster/Rostock University;
- "Partial Possibilistic regression path modelling", by Prof. Rosaria Romano from Università della Calabria, Cosenza;

- Conference on Substrate integrated waveguides and related technology – CSIWT 2015, Politecnico di Bari;
- "Geo-computation using open source software: combining research reproducibility and processing efficiency" by dr. G. Amatulli (Yale University, US), at CNR ISSIA, Bari;
- IEEE-IAS-Distinguisched Lecture on "V7f and I-f advanced of ac motor drives: recent progress and future trends", by Prof. Ion Boldea from Polytecnic University of Timisoara, Romania, at Politecnico di Bari;
- "Conferenza di dipartimento" at the National Research Council, University of Pisa, about Big Data and Big Data Analytics.

From 3rd of August to the 14th of August 2015, I took part at an Intensive English course (level B2, duration: 60 h) held from at F+U Academy of Languages in Heidelberg, Germany.

I've also attended the course: Big Data Analytics And Businness Intelligence, by Prof. A. Picariello at University of Naples Federico II.

Finally, I attended the online module: "Introduction to Big Data", by US San Diego University at coursera.org, and I get a certificate of accomplishment for this course.

Research activity

My research activities aim at using, combining and eventually enhancing ontology integration and semantic matching techniques also evaluating their scalability in terms of Big Data. The ontology integration is a broad field of research, which is concerned with determining and overcoming mismatching between ontologies (or in general knowledge bases) in order to allow the reuse of such ontologies. Knowledge integration along with the paradigm shift involved by Big Data, as for theories and technologies in data management, offers a new perspective to deal with the huge pile of interconnected and (now even more) semantically related data over the Internet, fostering the concrete realization of the semantic web. In order to carry out the research topics above, I have spent much of my Phd first year doing a literature review of existing and well-known ontology matching techniques. I have collected and critically analysed several scientific papers describing linguistic or conceptual matching techniques, or more specifically ontology-oriented matching techniques, also detailing the similarity measures, which have been proposed throughout the past years from other researchers. I have also investigated existing frameworks dealing with the ontology integration problem from a methodological perspective. Parallel to researches on ontology matching techniques, I have carried out a literature review on Big Data, from the high-level characterization of the term "Big" throughout the Volume-Velocity-Variety model, to the existing technologies dealing with very large databases in terms of storage and computing. Furthermore, I have investigated the potential scenarios in which the Big Data paradigm shift matters, from data-intensive sciences to social networking, business intelligence and modern companies.

With the second year of my phd, I conducted researches about NoSQL and graph-oriented database management tool resulting in a survey of such technologies through a qualitative evaluation of their features (functional and non- functional). I conducted further developments and researches in my doctoral thesis about the integration of very large knowledge bases specifically focused on: collecting the state-of-the-art of matching techniques: string matching, linguistic/semantic matching and structural matching; exploiting the above techniques and metrics in order to find a general measure of similarity to relate entities coming from different knowledge bases each other. Moreover, I did experiments on matching two well-known large, generalistic knowledge bases, namely: DBpedia and Wordnet, as a benchmark or test suite to evaluate performances (in terms of precision and recall) of my matching algorithms and techniques. I learnt and used Neo4J as a new GraphDB to import, query and visualize very large databases. In particular, some experiments have been conducted in visualizing the WordNet lexical DB within Neo4J and Cytoscape. In addition, Neo4J has been used for a rapid prototyping of a graph structure to handle the similarity measure between entities coming from WordNet and DBpedia tested matching algorithms. In this regard, at the end of the second year I found a new way to visualize WordNet synsets in Neo4J and Cytoscape using a tag clouds-based approach.

I have also conducted a study of the main international journals related with my research topics, annotating indexing and impact factors for each of them, in order to individuate proper journals for the forthcoming articles I plan to write.

Products

During the second year, I submitted and published four conference papers coming from the research activities above described:

1) Caldarola E. and Rinaldi A., Big Data: A Survey - The New Paradigms, Methodologies and Tools. In Proceedings of 4th International Conference on Data Management Technologies and Applications (KomIS-2015), pages 362-370, ISBN: 978-989-758-103-8, DOI: 10.5220/0005580103620370.

Abstract: For several years we are living in the era of information. Since any human activity is carried out by means of information technologies and tends to be digitized, it produces a humongous stack of data that becomes more and more attractive to different stakeholders such as data scientists, entrepreneurs or just privates. All of them are interested in the possibility to gain a deep understanding about people and things, by accurately and wisely analyzing the gold mine of data they produce. The reason for such interest derives from the competitive advantage and the increase in revenues expected from this deep understanding. In order to help analysts in revealing the insights hidden behind data, new paradigms, methodologies and tools have emerged in the last years. There has been a great explosion of technological solutions that arises the need for a review of the current state of the art in the Big Data technologies scenario. Thus, after a characterization of the new paradigm under study, this work aims at surveying the most spread technologies under the Big Data umbrella, throughout a qualitative analysis of their characterizing features.

2) Enrico G. Caldarola, A. Picariello, A. Rinaldi, An Approach to Ontology Integration for Ontology Reuse in Knowledge Based Digital Ecosystems, in Proceeding MEDES '15 Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems Pages 1-8 ACM New York, NY, USA ©2015 , ISBN: 978-1-4503-3480-8, doi>10.1145/2857218.2857219

Abstract: In the last years, the large availability of information and knowledge models formalized by ontologies has demanded effective and efficient methodologies for reusing and integrating such models in global conceptualizations of a specific knowledge or application domain. The ability to effectively and efficiently perform knowledge reuse is a crucial factor in the development of ontologies, which are a potential solution to the problem of information standardization and a viaticum towards the realization of knowledge-based digital ecosystem. In this paper, an approach to ontology reuse based on heterogeneous matching techniques will be presented; in particular, we will show how the process of ontology building will be improved and simplified, by automating the selection and the reuse of existing data models to support the creation of digital ecosystems. The proposed approach has been applied to the food domain, specifically to food production.

3) Caldarola, E., Picariello, A. and Rinaldi, A., Big Graph-based Data Visualization Experiences - The WordNet Case Study. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR, pages 104-115 ISBN: 978-989-758-158-8, Copyright 2015 by SCITEPRESS – Science and Technology Publications

Abstract: In the Big Data era, the visualization of large data sets is becoming an increasingly relevant task due to the great impact that data have from a human perspective. Since visualization is the closer phase to the users within the data life cycle's phases, there is no doubt that an effective, efficient and impressive representation of the analyzed data may result as important as the analytic process itself. This paper presents an experience for importing, querying and visualizing graph database and in particular, we describe as a case study theWordNet database using Neo4J and Cytoscape. We will describe each step in this study focusing on the used strategies for overcoming the different problems mainly due to the intricate nature of the case study. Finally, an attempt to define some criteria to simplify the large-scale visualization of WordNet will be made, providing some examples and considerations which have arisen.

4) Caldarola, E., Picariello, A. and Rinaldi, A., WordNet exploration and visualization in Neo4J.A Tag Cloud-based approach. Accepted as full paper, at the Eighth International Conference on Information, Process, and Knowledge Management, eKNOW 2016, Venice, Italy.

Abstract: In the Big Data era, the visualization of large data sets is becoming an increasingly relevant task due to the great impact that data have from a human perspective. Since visualization is the closer phase to the users within the data life cycles phases, there is no doubt that an effective, efficient and impressive representation of the analyzed data may result as important as the analytic process itself. Starting from previous experiences in importing, querying and visualizing WordNet database within Neo4J and Cytoscape, this work aims at improving the WordNet Graph visualization by exploiting the features and concepts behind tag clouds. The ultimate goal of this work is, on the one hand, to facilitate the comprehension of Wordnet itself and, on the other hand, to investigate techniques and approaches to get more insights from the visual representation and analytics of large graph databases.

Conferences and Seminars

Conferences:

4th International Conference on Data Management Technologies and Applications (KomIS-2015), Colmar, France;

MEDES '15 Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems, San Paulo, Brasil;

7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Lisbon, Portugal.

Seminars:

- “The iCub project: An open platform for research in robotics & artificial intelligence”, by dr. Giorgio Metta, from Istituto Italiano di Tecnologia;
- “Good Ontology Design Philosophical and Empirical Perspectives” held by dr. Ludger Jansen, Münster/Rostock University;
- “Partial Possibilistic regression path modelling”, by Prof. Rosaria Romano from Università della Calabria, Cosenza;
- Conference on Substrate integrated waveguides and related technology – CSIWT 2015, Politecnico di Bari;
- “Geo-computation using open source software: combining research reproducibility and processing efficiency” by dr. G. Amatulli (Yale University, US), at CNR ISSIA, Bari;
- IEEE-IAS-Distinguisched Lecture on “V7f and I-f advanced of ac motor drives: recent progress and future trends”, by Prof. Ion Boldea from Polytecnic University of Timisoara, Romania, at Politecnico di Bari;
- “Conferenza di dipartimento” at the National Research Council, University of Pisa, about Big Data and Big Data Analytics.

Tutorship

This year I prepared the teaching material for a tutor activity to be carried out in the next semester for the Information Retrieval Systems degree course at University of Naples.

CS summary

The following table reports my credits for the second year of my phd activities.

	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	20		2	3	2	2	3	**12**	15	2	4	2	4	4	1	**17**	21	0	5	0	0	0	0	**5**	**34**	**30-70**
Seminars	8	1	1	1	1	1	1	**6**	6	1	1	1	2	1	1	**7**	12	2	0	0	0	0	0	**2**	**15**	**10-30**
Research	32	9	7	6	7	7	6	**42**	39	7	5	7	4	5	8	**36**	30	10	10	10	10	10	10	**60**	**138**	**80-140**
	60	10	10	10	10	10	10	**60**	60	10	10	10	10	10	10	**60**	63	12	15	10	10	10	10	**67**	**187**	**180**