PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Iliana Mineva Petrova

XXIX Cycle

Training and Research Activities Report – Third Year

Tutor: Piero Andrea Bonatti

1. Information

Iliana Mineva Petrova, MS in Computer Science (Information Systems) – Università degli Studi di Napoli Federico II.

XXIX Cycle - Information Technology and Electrical Engineering - Università degli Studi di Napoli Federico II.

Three-year fellowship grant by Università degli studi di Napoli Federico II.

Tutor: Prof. Dr. Piero Andrea Bonatti.

2. Study and Training activities

A. Courses

B. Seminars

"Reasoning over Large Amounts of Data via Abstraction and Refinement", Jr. Prof. Birte Glimm, 04 May 2016, 11.00 - 12.00, University of Ulm, Germany, 0.2 credits.

"Axiom Pinpointing", Peter Skocovsky, 25 May 2016, 10.15 - 11.15, University of Ulm, Germany, 0.2 credits.

"Automated Verbal Explanations for EL Subsumptions", Dr. Ing. Marvin Schiller, 8 June 2016, 10.15 - 11.15, University of Ulm, Germany, 0.2 credits.

"Language Is Unreliable - a common issue for AI and Buddhism", Zhangquan Zhou, 29 June 2016, 10.15 - 11.15, University of Ulm, Germany, 0.2 credits.

"Change the plan - how hard can that be", Gregor Behnke, 6 July 2016, 10.15 - 11.15, University of Ulm, Germany, 0.2 credits.

"Assessing the Expressivity of Planning Formalisms through the Comparison to Formal Languages", Daniel Höller, 13 July 2016, 10.15 - 11.15, University of Ulm, Germany, 0.2 credits.

"Rule based module extraction", by Markus Brenner, 20 July 2016, 10.15 - 11.15, University of Ulm, Germany, 0.2 credits.

C. External courses

"Algorithms for Knowledge Representation", held by Prof. Yevgeny Kazakov, University of Ulm, 6 CFU, exam passed on 28 July 2016

D. Summary of training activities for the third year

Student: Iliana Mineva Petrova ilianamineva.petrova@unina.it

Tutor: Piero Andrea Bonatti pieroandrea.bonatti@unina.it

Cycle XXIX

	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			3		6	8	**17**	13		3	4				**7**	6			6				**6**	**30**	30-70
Seminars	5				2 2/5	2 9/10	1 7/10	**7**	5	1 2/5				6 3/5		**8**	0		4/5	3/5				**1,4**	**16,4**	10-30
Research	35	10	10	7	7	2	3	**39**	42	9	7	5	10	4	10	**45**	54	10	9	4	10	10	10	**53**	**137**	80-140
	60	10	10	10	9,4	10,9	12,7	**63**	60	10,4	10	9	10	10,6	10	**60**	60	10	9,8	10,6	10	10	10	**60,4**	**183**	180

Year	Lecture/Activity	Type	Credits	Certification	Notes
1	First EATCS Young Researchers School on Automata, Logic and Games	International Summer School	3	x	
1	Fractional Programming for Energy Efficiency in Wireless Networks: Part 1, Theory	Seminar	2/5	x	
1	Fractional Programming for Energy Efficiency in Wireless Networks: Part 2, Theory	Seminar	2/5	x	
1	Nano-carbon based components and materials for high frequency electronic	Seminar	4/5	x	
1	Developmental Robotics: From Babies to Robots	Seminar	2/5	x	
1	Quantum teleportation	Seminar	2/5	x	
1	EuroProgettazione	Ad hoc module	3	x	
1	IBM-Bluemix Corso tecnico sulla nuova piattaforma di sviluppo in Cloud (PaaS)	Ad hoc module	3	x	
1	Reliability and Availability Modeling in Practice	Seminar	3/5	x	
1	Capacity Planning for Infrastracture-as-a-Service Cloud	Seminar	2/5	x	
1	Heterogeneities in temporal networks emerging from adaptive social interactions	Seminar	1/5	x	
1	Methods and tools for smart device integration and simulation	Seminar	2/5	x	
1	UML Profiles for the specification of non functional properties of software systems	Seminar	1/2	x	
1	Site Reliability Engineering at Google	Seminar	2/5	x	
1	Verification of Mobile Agents in Partially Known Environments	Seminar	2/5	x	
1	Machine Learning e applicazioni mod A: Machine learning applicato all'information retrieval	MS Module	6	x	
1	Three core issues for the Internet: things, security and economics	Ad hoc Module	2	x	
1	Applications for software development: types, interactions and continuousintegration	Seminar	2/5	x	
1	Risk management meets model checking: fault tree analysis and model-based testing via games	Seminar	2/5	x	
1	Workshop "Efficient service distribution in next generation cloud networks"	Seminar	9/10	x	
2	Answering queries over inconsistent databases	Seminar	2/5	x	
2	Colloquium on Robotics Six Keynote Talks by International Experts	Seminar	1	x	
2	Designing and writing scientific manuscripts for publication in English language scholarly journals, and related topics	Ad hoc Module	3		
2	11th Summer School on Ontology Engineering and the Semantic Web (SSSW'15)	International Summer School	2	x	
2	The 11th Reasoning Web Summer School (RW 2015)	International Summer School	2	x	
2	Memory technologies for Android based systems	Seminar	2/5	x	
2	Real-Time Embedded Control Systems	Seminar	1 1/5	x	
2	Test and Diagnosis of Integrated Circuits	Seminar	2 2/5	x	
2	Hardware Security and Trust	Seminar	2 2/5	x	
2	Armi autonome: problemi etici e decisioni politiche	Seminar	1/5	x	
3	Reasoning over Large Amounts of Data via Abstraction and Refinement	Seminar	1/5		
3	Axiom Pinpointing	Seminar	1/5		
3	Automated Verbal Explanations for EL Subsumptions	Seminar	1/5		
3	Language Is Unreliable - a common issue for AI and Buddhism	Seminar	1/5		
3	Change the plan - how hard can that be	Seminar	1/5		
3	Assessing the Expressivity of Planning Formalisms through the Comparison to Formal Languages	Seminar	1/5		
3	Rule based module extraction	Seminar	1/5		
3	Algorithms for Knowledge Representation	External module	6	x	

3. Research activity

Design, implementation and optimization of non-standard reasoning services

Nowadays, within the field of knowledge representation there is the increasingly growing need to extend the ontology languages and reasoning engines with non-standard characteristics. Here we focus on two concrete application contexts - Nonmonotonic Description Logics (DLs) and Confidentiality Models for Ontologies. Several authors advocated nonmonotonic inference as useful means to address real-world problems in important fields such as biomedical domains and access control policies. On the other hand, OWL and Linked Open Data have been increasingly used to encode sensitive knowledge on individuals (e.g. health), companies and public organizations. As Semantic Web techniques make it possible to extract implicit information, any access control method that does not deal with inference fails to ensure privacy.

In both cases the available classical tools are just not sufficient to offer the required non-standard reasoning services. A considerable additional effort reveals mandatory in order to provide a practical and scalable support of these kind of applications. As a consequence my research activity is currently focused on the design, implementation and optimization of the above mentioned frameworks. We adopt a suitable, easy to grasp, representation of the new constructs introduced by the specific context (non monotonic axioms in the first and background metarules in the second case). Due to the very large size of knowledge bases (KBs) that both applications need to deal with, general purpose optimization techniques such as incremental reasoning and *suitably adapted* module extraction proves to improve performance drastically. Effective conjunctive query answering techniques are as well required for the efficient implementation of the second framework. When it comes to a standard reasoning tasks these applications rely on publicly available classical reasoners.

- **Nonmonotonic Description Logics (DLs)**

  As far as the first context is concerned, my working group contribute to the practical support of nonmonotonic inferences in DLs by introducing a new semantics (cf. [1]) expressly designed to model priority-based overriding. This formalism makes it possible to design knowledge bases by describing prototypical instances whose general properties can be refined later, by adding suitable exceptions. The new family of nonmonotonic description logics is called DL^N. Automated reasoning in the new logic is carried out by means of polynomial reduction to classical DLs.

  The first experimental scalability tests on a semi-naive implementation (cf. [1], relying only on the optimization techniques of the underlying classical reasoner), called NMReasoner,

revealed promising. The result graphs showed that nonmonotonic constructs cause an approximately linear increase in reasoning time.
More recently we studied optimization techniques ([3, 4]) based on module extraction and optimistic computation techniques for reasoning in DL^N :

I. Module Extraction Techniques

Module extraction is an optimization technique for querying large ontologies. In practice many of the axioms in a large KB are expected to be irrelevant to a given query. Module extractors can be used to focus reasoning on relevant axioms only. More precisely we have introduced an iterated module extraction procedure for DL^N reasoning. It is worth noting that the approach is not trivial (module extractors are unsound for most nonmonotonic logics, including default logics and circumscription) and requires an articulated correctness proof. Furthermore, a module extraction algorithm for ABoxes has been proposed and proved to be correct under the assumption of consistency of the knowledge base (this hypothesis, in practice, is compatible with some of the main intended uses of module extraction). The commonly employed bottom-module extractors are not very effective in the presence of non empty ABoxes. This phenomenon is even amplified in DL^N, where reasoning requires repeated incremental classifications of the knowledge base. The new module extraction algorithm is proved to discards significantly more axioms in the presence of non empty ABoxes. Finally, we developed a parallel module extraction procedure to further improve the performance. It is designed to take advantage of the multiprocessor architectures by evaluating the relevance of axioms in parallel.

II. Optimistic computation (OPT)

The optimistic computation algorithm for query answering that has been introduced is expected to exploit incremental reasoners at their best. Incremental reasoning is crucial as DLN's reasoning method iterates consistency tests on a set of KBs with large intersections. While the assertion of new axioms is processed very efficiently, the computational cost of axiom deletion is generally not negligible. The *optimistic reasoning method is thus expected to reduce the number of deletions*.

III. One-step retraction with bookkeeping

The price for the significant reduction of the number of retractions that the optimistic computation algorithm pays is given by the inherently greater size[1] of the manipulated knowledge base. Better result can be obtained using an inference engine that allows for efficient incremental retraction of a single axiom. For the purpose an ad-hoc one-step bookkeeping strategy have been designed and integrated into the ELK reasoner. In particular, using a form of bookkeeping to trace the consequences that no longer hold as a result of a retraction of an axiom saves the cost of performing extra reasoning while

[1] In one to one correspondencewith the number of skipped retractions.

maintaining the memory comsumption overhead due to the extra information saved by the algorithm reasonably low.

Currently no "*real*" KBs encoded in a nonmonotonic DLs exist, as standard DL technology does not support nonmonotonic reasoning. Our scalability tests have been carried out on synthetic test cases automatically generated in a principled way. The test suites have been obtained by modifying large biomedical ontologies (*GALEN, Gene Ontology, Fly Anatomy*) and proved to be nontrivial w.r.t. a number of structural parameters.

All optimizations have been integrated in NMReasoner and validated experimentally on large synthetic KBs with more than 30K axioms. The experimental results show that the iterated module-based method make DLN reasoning at least one order of magnitude faster (and up to ~780 times faster in some cases); the OPT speedup factor is about two. The conditional module extractor for nonempty ABoxes is very effective when the ABox assertions are loosely interconnected, with speedups up to 75%.

To briefly summarise, all optimizations introduced are sound and complete. To the best of our knowledge, for the first time response times compatible with real-time reasoning are obtained with nonmonotonic KBs of this size (more than 20K concept names and over 30K inclusions).

❖ Confidentiality model for ontologies

The main contribution to the area of knowledge base confidentiality of my work group consists in the introduction of a knowledge base confidentiality model, based on a fully generic formalization of the user's background knowledge, and the definition of a method for computing secure knowledge views that generalizes some previous approaches. In order to compute secure views in practice, we adopt a safe, generic method for approximating background knowledge, together with a specific rule-based language for expressing metaknowledge.

We further provide a tractable implementation of the new framework (cf. [2]), a secure view generator called SOVGen. The generator exploits module extractors on the background knowledge bases (such as SNOMED) in order to make reasoning focus on relevant knowledge only. Experimental analysis show that the extracted modules result on average two or three orders of magnitude smaller which drastically improves performance.
Evaluation of metarules requires techniques for effective query evaluation. We provide three different implementations based on *Apache Jena-ARQ*, *OWL-BGP* and ad hoc module, called Metarule Evaluation Engine (MEE), designed to take advantage of the specific nature of Horn metarules. MEE extensively exploit incremental reasoning, short-circuit evaluation and memoization techniques.

Synthetic test cases have been specifically designed to simulate the employment of the confidentiality model in a e-health scenario as part of the SmartHealt 2.0 Project. Each test case represents the encoding of sensitive data in a CDA[2]-compliant electronic health record. According to the CDA standards a disease is represented by a ICD-9CM code, pharmaceutical products and procedures by SNOMED CT codes, while diagnostic tests and laboratory data by LOINC codes.

Scalability experimental evaluations carried out on the synthetic test cases show that the use of *module extraction techniques improves the computation time of two-three orders of magnitude* at an acceptable cost of about 30 sec of overhead given the amount of background knowledge (consider that SNOMED-CT describes about 300K concepts). The experimental results obtained by using MEE (resp. OWL-BGP) to evaluate metarules *show MEE is 1-2 orders of magnitude faster* (for Jena-ARQ all the test cases exceeded 1 hour time-out).

Summing up, the experimental analysis confirm that module extraction techniques and a suitable, ad-hoc metarule evaluation engine – which intensively exploit incremental reasoning - largely outperform general conjunctive query evaluation engines.

Future work will include identification of different parallelization strategies, based on suitable reorderings of the operations performed during the translation of DLN and further possible optimizations for DLN , such as caching the translations used for previous queries. We further plan to extend SOVGen for general metarules. Last but not least, an extension of the synthetic test suites is currently in progress as part of the preparation of a journal paper describing the advanced optimisations for DLN.

References

[1] P.A. Bonatti, M. Faella, I. M. Petrova and L. Sauro: A new semantics for overriding in description logics. Artificial Intelligence Journal, 222:1–48, 2015

[2] “Optimized Construction of Secure Knowledge-Base Views”, P. A. Bonatti, I. M. Petrova and L. Sauro, Proceedings of the 28th International Workshop on Description Logics. CEUR Workshop Proceedings 1350, CEUR-WS.org

[3] “Optimizing the computation of overriding”, P. A. Bonatti, I. M. Petrova, L. Sauro: . In M. Arenas et all. (eds.): Proceedings of the 14th International Semantic Web Conference (ISWC-14). LNCS 9366, 356-372. Springer 2015

[2] Clinical Document Architecture (CDA) is an international standard for electronic health information exchange, based on the Health Level 7 Reference Information Model.

[4] "Optimizing the computation of overriding", P. A. Bonatti, I. M. Petrova, L. Sauro, CoRR abs/1507.04630 (2015)

Cooperations

Currently I am working in collaboration with with Prof. Dr. Piero A. Bonatti and Dr. Luigi Sauro, members of the Department of Electrical Engineering and Information Technologies. During my stay abroad I was supervised by Dr. Yevgeny Kazakov, member of the Institute of Artificial Intelligence of the University of Ulm, Germany.

4. Products

A. Conference papers
B. International journal papers
C. Technical reports
D. Papers in preparation

"Improving Module Extraction for Nonmonotonic and Classical DL", P. A. Bonatti, I. M. Petrova, L. Sauro, to be submitted to Journal of Automated Reasoning.

5. Conferences and Seminars

A. Conferences

I participated in the 3rd Italian Conference on Computational Linguistics (CLiC-it 2016), held in Naples, Italy from December 5th to December 6th, 2016.

B. Workshop Presentations made

"Evaluation of NLP and Speech Tools for Italian (EVALITA 2016)", 7 December 2016, Naples, Italy

C. Presentations

On March 17th, 2016 I held a talk entiteled "New semantics for overriding in Description Logics: The story so far" as a visiting student at the Institute of Artificial Intelligence, University of Ulm, Germany.

6. Activity abroad

During my third year, I spent a period abroad - from March to August 2016 - as a visiting student at the Institute of Artificial Intelligence, University of Ulm, Germany under the supervision of Dr. Yevgeny Kazakov. The research activity there concerned the study of incremental reasoning techniques and the design and implementation of one-step retraction into ELK[3], an OWL reasoner developed by Dr. Kazakov and his group. During my stay, I also attended the course "Algorithms for Knowledge Representation" and several seminars related to my research interests.

7. Tutorship

- I am teaching assistant for Web Technologies, course of the the Master's degree in Computer science, Department of Electrical Engineering and Information Technologies, Università degli Studi di Napoli Federico II.
- On 23 September 2016 I have been awarded with "Assegno di attività di tutorato" within the "Scuola Politecnica e delle Scienze di Base" of University of Naples Federico II. In this occasion I have been involved also in the course Programmazione II: Laboratorio, course of the Bachelor's degree in Computer science, Department of Electrical Engineering and Information Technologies, Università degli Studi di Napoli Federico II.

[3] https://github.com/liveontologies/elk-reasoner/