PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Giovanni Cozzolino

XXXI Cycle

Training and Research Activities Report – First Year

Tutor: **Antonino Mazzeo**
co-Tutor: **Flora Amato**

1. Information

- Giovanni Cozzolino, Master's Degree in Computer Engineering, in 2013 from the University of Naples Federico II.
- XXXI Cycle – ITEE
- DIETI Grant
- Tutor: Antonino Mazzeo – co-Tutor: Flora Amato

2. Study and Training activities

Courses:

- *Advanced Computer Architecture and GPU Programming*, Alessandro Cilardo;
- *Big Data Analytics and Business Intelligence*, Antonio Picariello;
- *Intelligenza Artificiale*, Flora Amato;

Activities schedule

	Credits year 1								Credits year 2							
		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
Modules	20		6		6		6	18	12							0
Seminars	5							0	10							0
Research	35	6	5	8	6	4	6	35	45							0
	60	6	11	8	12	4	12	53	67	0	0	0	0	0	0	0

3. Research Activities

Title

Digital information integration and correlation through semantic techniques

Study

In the first phase of research activities, was performed a preliminary study on the State of Art of Integration of information from multiple disparate and heterogeneous sources. In particular, has been made a review of Semantic Web technologies proposed to address both representational and semantic heterogeneity in distributed and collaborative environments.

Together, during the "Intelligenza Artificiale" module, were studied in deep Natural Language Processing (NLP) techniques, in order to perform useful analysis through unstructured data written in natural languages humans use.

In a second phase, were investigated the challenges and the applications related to Big Data and Social network data analytics with the aim of collecting and analyzing information coming from social network stream to gain opinion from users or detect malicious behaviors.

Research description

Different sources of information generate every day huge amount of data. For example, let us consider social networks: here the number of active users is impressive; they process and publish information in different formats and data are heterogeneous in their topics and in the published media (text, video, images, audio, etc.).

Semantic Integration is an effort of bringing together different, often heterogeneous, sources of information interrelating them by leveraging the semantic information that is embedded inside these information sources. Its main benefit is that combined system usually contains more usable information than that is present in the individual sources themselves.

The heterogeneity of information and the huge scale of data makes the identification of interest information a challenging problem. The lack of integration and interoperability between information sources protract the analysis process and pauperize presentation results for non-technical parties.

During the research activities has been designed a methodology, based on Semantic Web technologies, to integrate, correlate and query different sources of data with the goal of more valuable data retrieval. The proposed approach can also improve, through semantics vocabularies, the automation of parts of the analysis with respect to data correlation.

The main potential advantages of such an approach regard:

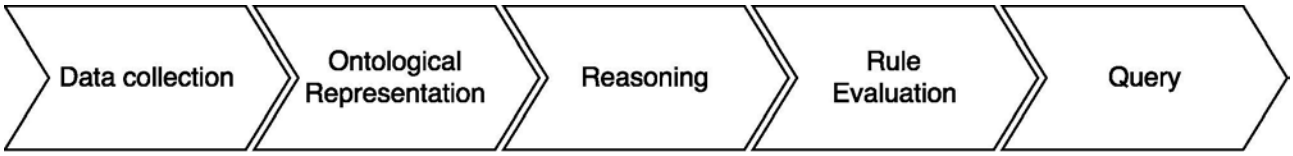
- Information Integration: the RDF data model promotes easy integration of heterogeneous data due to its schema independence and standardized statement representation of subject-predicate-object form.
- Classification and Inference: the RDF/OWL combination can infer class membership and typing information from data based on the ontological specified definitions. A semantic reasoner can use definitions and restrictions of the ontology, in order to infer dynamically class membership of the instances.
- Extensibility and Flexibility: RDF/OWL provide backward and forward compatibility, OWL provides the flexibility that different ontologies can be defined according to the

stakeholder's interests and scope, but still be able to integrate through ontology mapping processes.

- Search: Documents are enriched with semantic markup that is later used over or along traditional keywords for document indexing purposes. A query can be extended through a reasoning engine and generate more meaningful data.

Semantically-based approach can provide an automated framework to formalize and describe in a more expressive way a representation of existing domain knowledge.

The basic work-flow of the methodology is presented in Figure:



The Data Collection phase include all the data acquisition techniques. The goal of this phase is to generate the appropriate input for the subsequent phases and depends from the specific source of information.

The Ontological Representation phase expects to parse the input data, with appropriate software or even hardware tools, and to represent them in a common and explicit format using the RDF data model and respective ontologies. In this phase can be considered, if existent, shared domain ontologies, or can be specified new ad-hoc ontologies. Parsing tools have to be flexible enough to operate with shared ontologies as well as custom made ontologies. The task of this step should be the transformation of the input data into a set of triplet constituting the RDF data model. The final output can be stored using the various ontology formats such as RDF/XML or RDF/OWL.

The Reasoning step involves the use of an OWL-based reasoner upon the previously generated ontological representation of data, in order to infer additional axioms based on the instances' relations. Different types of new axioms can be inferred, enriching the instances with information regarding the definition of their class, properties or subclasses. Thanks to property restrictions or subclass hierarchy, the reasoner can classify instances as members of specific class, so a resource can belong to multiple classes in parallel. Furthermore, a resource can dynamically be classified with higher precision compared to concepts defined in the original data. Moreover, the inference engine analyzes property relations in order to infer new property axioms that are attached to the individuals.

In the Rule Evaluation step, SWRL Rules are evaluated by a rule engine and the newly asserted axioms are inserted back to the ontology. SWRL Rules can assert new additional axioms that cannot be inferred due to the DL-safe expressivity limitations of OWL. This phase can be realized

also with the support of external Rule Engines, and the newly inferred axioms should be translated again back to the RDF data model and integrated in the existing ontology.

Finally, queries can be expressed in the SPARQL language and submitted to a SPARQL endpoint hosting the sets of asserted and inferred axioms. Both OWL-inference as well as SWRL evaluation relations in the form of predicates can be established between resources of the same or different datasets.

The methodology has been implemented in a system architecture. This architecture was deployed in different case study domains in order to test and validate the proposed method and evaluate the scalability and adaptability of the system in different scenarios.

Projects

- Big4H - Big Data for e-Health Applications
- Conferenza dei Rettori delle Università Italiane, Ministero della Giustizia, Dipartimento dell'organizzazione giudiziaria, del personale e dei servizi – Direzione Generale per i sistemi informativi automatizzati (DGSIA): "*La gestione del servizio di formazione qualificata, ricerca applicata e certificazione di professionalità a seguito della riorganizzazione del Ministero della Giustizia per il tramite dei sistemi ICT, su strumenti e funzionalità del Processo Civile e Penale telematico, nell'ambito della riduzione dei tempi della giustizia, su profili di sicurezza dei sistemi informativi in uso presso il Ministero della Giustizia e gli uffici Giudiziari*".

4. Products

Publications:

- Amato F., Cozzolino G., Mazzeo A., Romano S., "A semantic system for diagnoses suggestion and clinical record management" (2016) - Proceedings - IEEE 30th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2016, art. no. 7471186, pp. 133-138.
- Amato F., Cozzolino G., Mazzeo A., Romano S. "Malicious event detecting in twitter communities" (2016) - Smart Innovation, Systems and Technologies, Springer International Publishing, 55, pp. 63-72.
- Amato F., Cozzolino G., Di Martino S., Mazzeo A., Moscato V., Picariello A., Roman S., Sperlí G. "Opinions analysis in social networks for cultural heritage applications" (2016) - Smart Innovation, Systems and Technologies, 55, pp. 577-586.
- Amato F., Cozzolino G., Mazzeo A., Romano S. "An Architecture for processing of Heterogeneous Sources" (2016) - Advances on P2P, Parallel, Grid, Cloud and Internet Computing, Volume 1 of the series Lecture Notes on Data Engineering and Communications Technologies pp 679-688

- Amato F., Cozzolino G., Moscato F., Moscato V., Picariello A., "Modeling Approach for Specialist Domain" (2016) - Advances on P2P, Parallel, Grid, Cloud and Internet Computing - Volume 1 of the series Lecture Notes on Data Engineering and Communications Technologies pp 689-698
- Amato F., Cozzolino G., Mazzocca N., "Semantic Integration and Correlation of Digital Evidences in Forensic Investigations" (2016) - International Conference on P2P, Parallel, Grid, Cloud and Internet Computing pp 415-424

5. Tutorship

- Fondamenti di informatica
  - Type: Seminar
  - Subject: C++ programming
  - Hours: 4
- Sistemi informativi
  - Type: Seminar
  - Subject: BPMN and Bonita BPMS
  - Hours: 4