Giovanni Cozzolino

Tutor: Antonino Mazzeo – co-Tutor: Flora Amato

XXXI Cycle - II year presentation

Semantic Correlation of Digital information in Specialized Domain

CONTEXT

Heterogeneous information produced by systems and applications is growing up with the use of digital and computer-related technologies.

Problems:

- complex data analysis processes and poor presentation results
- Redundancy and Inconsistency of data
- Fragmented, outdated, inaccessible or indecipherable information

Integration of data improves the success rate of critical projects and key analytics initiatives

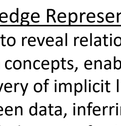
Semantic integration is the combination of technical and business processes used to combine data from different sources into meaningful and valuable information.

TRENDS

Knowledge Representation

Solution to reveal relationships between concepts, enabling:

- discovery of implicit links between data, inferring new knowledge out of existing facts.
- data integration, from various sources in various formats (structured and unstructured)
- data classification linked to other relevant datasets



Natural language processing

NLP break down the occurrence of terms in a sentence and store and create relationships in a graph database.

Various deep models have become the new state-of-the-art methods for NLP problems:

- NLP applications that employ reinforcement learning and unsupervised learning methods.

Text mining

Text mining structures text as important facts, key terms or persons. This makes information extraction easier and enables efficient indexing and improved search, or personalized recommendations to users.
Main trends:

- non-English text processing through resources (linguistic resources and annotated corpora)
- graph-based text representation

Internet of Things

Some of the key properties of IoT systems are:

- unstructured nature of data being generated
- speed and streaming nature of data
- fuzziness of data
- dynamics and heterogeneity of technical architectures.

The variety which needs to be controlled in such a systems is very high and therefore is as a natural application for semantic technologies

Data Analytics

Data Analytics models, together with semantic enrichment, are capable of:

- classifying and adapting content so that it can be easily reused
- anticipating user behavior
- apply predictive analytics to flag potential risks

GROUP ACTIVITIES

Public security protection: identification of fraudulent behavior through Social networks content analysis, Forensic Investigations, etc.

Justice Data Processing Improvement: Implicit correlation discovery, suggestion, search by content

Healthcare Improvement: Diagnoses suggestion and clinical record management

IDEA

Semantic Integration, 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.

Main benefit: combined system usually contains more usable information than that is present in the individual sources themselves

METHODOLOGY

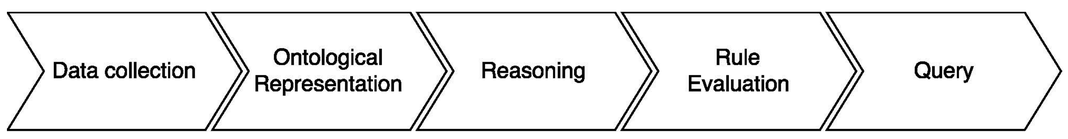
The **Data Collection** phase includes all the techniques for gathering information, depending from the particular source.

The **Ontological Representation** phase expects to parse the input data and to represent them in a common and explicit format (triplet) using the RDF data model. In this phase can be considered, if existent, shared domain ontologies, or can be specified new ad-hoc ontologies.

The **Reasoning** phase 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.

In the **Rule Evaluation** phase, SWRL Rules are evaluated by a rule engine and the newly asserted axioms are inserted back into the ontology: the newly inferred axioms should be translated again back to the RDF data model and integrated in the existing ontology.

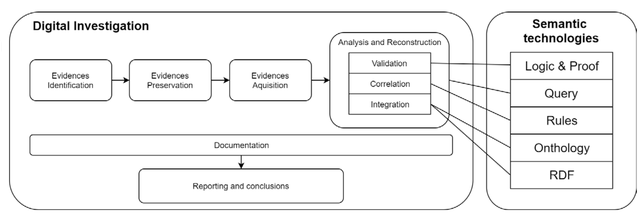
Queries can be expressed in the SPARQL language and submitted to a SPARQL endpoint hosting the sets of asserted and inferred axioms.



DEVELOPMENTS

Semantic technologies applied to Forensics Investigations

Objective: The goal of Digital Forensics is not only collection, acquisition and documentation of data stored on digital devices, but, above all, it is the interpretation of evidences. Correlation of information is a crucial phase in forensics analyses, because it is the only mean to allow for the contextualization of digital evidences, promoting them as clues.

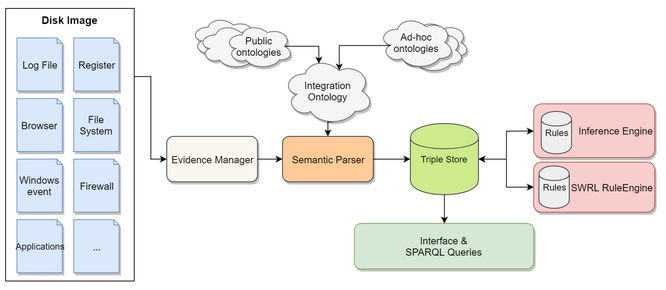


System Architecture: Our overall system consists of an ontology and five modules:

Evidence Manager: loads binary content of digital evidences, identifying the type of given source and verifying its integrity through hash values

Semantic Parser: generates an OWL representation of knowledge extracted from digital evidence;
instantiate the ontology, combined from public domain or custom ones.

Inference Engine: Inference engine performs automated reasoning, according to the OWL specifications, coming from domain ontologies or investigators knowledge.



SWRL Rule Engine: adopts SWRL rules in order to correlate different individuals or to establish relationships among individuals belonging to different ontologies but representing similar concepts.

SPARQL Queries: is responsible for accepting SPARQL queries from the user and evaluating them against a SPARQL query engine.

RESULTS

The configuration of the machine used to run the experiment and hosting the triple store (Stardog) has a 3.20 GHz Intel Core i5-6400 processor and 8 GB RAM.
we generated three disk image from a virtual machine Running Windows 7. On these machines we performed a set of user actions to simulate a malicious behavior caused by a malware.

TABLE 1
Volumes of processed data through the reconstruction and analysis process

Criterion	Dataset No1	Dataset No2	Dataset No3
Extracted entries	6798	4263	15983
Generated triples	8786	10475	32498
Deduced triples	18	21	31
Correlations	1231	535	14982

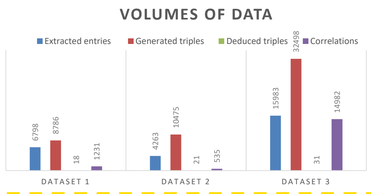
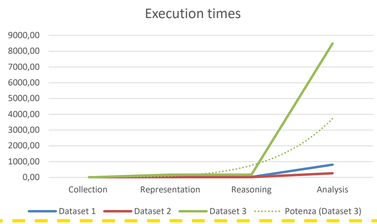


TABLE 2
Execution times (in seconds) for the different phases of the process

Steps	Dataset No1	Dataset No2	Dataset No3
Collection	1.07	1.43	1.67
Representation	24.37	18.3	167.45
Reasoning	0.29	0.247	0.314
Analysis	776.3	241.3	12314.7



giovanni.cozzolino@unina.it
antonino.mazzeo@unina.it
flora.amato@unina.it

EUROPEAN PROJECT

Founded European Project CREA (Conflict Resolution with Equitative Algorithms).
Justice Programme, Grant Agreement number: 766463 — CREA — JUST-AG-2016/JUST-AG-2016-05
Coordinated by University of Naples "Federico II".

COLLABORATION – NATIONAL PROJECT

Conferenza dei Rettori delle Università Italiane (CRUI) - Ministero della Giustizia – Direzione Generale per i Sistemi Informativi Automatizzati (DGSIA)
Applied research on tools, functionality and security protocols of Ministry of Justice information systems.

- Document management, encrypted full text indexing, semantic searches
- System architecture, virtualization, authentication and authorization protocol

FUTURE WORK

Methodology validation

In order to validate the proposed methodology (based on the semantic representation, integration and correlation of digital information) next year will be dedicated to the implementation of designed framework.

- heterogeneous data integration
- support for automation
- improved analytical capabilities
- expressive and flexible querying layer

Exploit proposed methodology to support mobile devices analysis:

- App data
- Phone functions (Call, sms, etc.)
- Sensors (GPS, Accelerometer, etc)