UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Semantics

2021-1-F9101Q011

Aims

The main purpose of the course is to provide students with the knowledge and skills necessary to understand and solve problems of semantic interoperability in data science applications, with particular reference to problems of representation, reconciliation and integration of heterogeneous data.

The topics addressed in the course have a dual purpose: 1) to provide theoretical and practical tools to represent, organize, publish, query, reconcile, and explore information in real application scenarios (widely discussed during lectures and addressed during the exercises) using semantic technologies and 2) to acquire the necessary skills to understand new semantic interoperability problems and the necessary techniques to solve them adequately regardless of particular reference technologies.

Contents

The course presents computational methods to represent, harmonize and reconstruct the semantics of data used in data science applications, with a particular focus on:

- models and languages developed within the semantic web to support the integration of heterogeneous data (knowledge graph, data linking, ontologies, RDF, RDFS, OWL);
- techniques for the integration of data and vocabularies;
- techniques for extracting information from texts (outline);
- artificial intelligence models for data and knowledge exploration.

Detailed program

1. **Data Semantics:** the role of semantics in data analytics (big data, web sources, heterogeneous formats, information integration and semantic enrichment, data linking, knowledge graphs).
2. **Knowledge Graphs:** representation and interogation of data in the semantic web (RDF, SPARQL, semantic technologies and architectures, corporate knowledge graphs with graph databases). Excercise on querying RDF knowledge graphs with SPARQL.
3. **Semantics for Knowledge Graphs:** definition of shared vocabularies with ontologies and logic-based languages ??(from shared vocabularies to ontologies, taxonomies, lexical ontologies, axiomatic ontologies, automatic reasoning and semantics, RDFS, OWL, SWRL). Excercise on ontology modeling with RDFS and OWL.

4. **Semantic reconciliation I:** information integration and semantic reconciliation, instance-level and schema-level reconciliation, information extraction (named entity recognition, entity linking, relation extraction).
5. **Semantic reconciliation II:** reconciliation of ontologies and vocabularies (ontology matching, mapping, semantic similarity and matchers' combination, mapping selection). Exercise on taxonomy reconciliation.
6. **Semantic reconciliation III:** value and instance-level reconciliation (deduplication and record linkage, probabilistic reconciliation approaches, distance metrics and similarity measures, combination and learning of similarity measures, data fusion strategies, graph-based similarity measures). Exercise on reconciling data with the help of existing tools.
7. **Information Exploration:** semantic techniques for the information exploration (measures of relevance, semantic associations, active learning of relevant associations, semantic recommender systems).
8. **New approaches to data semantics:** data-driven semantics and frontier approaches (semantic profiling of knowledge graphs, distributional semantics, word embeddings and knowledge graph embeddings).

Prerequisites

Mathematics and computer science as taught in the compulsory courses of the first semester.

Teaching form

Lectures and exercise with students' personal computers. Moodle e-learning platform. Seminars about usage of semantics in real-world applications given by experts from the industry.

During the Covid-19 emergency period, lessons will take place in a mixed mode: partial attendance and asynchronous videotaped lessons. If this mode is not possible, the course will be held remotely with asynchronous videotaped lessons and synchronous events (QA sessions) in videoconference (as done for the 2019-2020 edition).

Teached in English

Textbook and teaching resource

ITA: Tommaso Di Noia, Roberto De Virgilio, Eugenio Di Sciascio, Francesco M. Donini. Semantic Web: tra ontologie e Open Data, Apogeo, 2013.

ENG: Grigoris Antoniou, Paul Groth, Frank van van Harmelen, *A Semantic Web Primer*, (Third Edition), MIT press, 2012.

Additional material such as presentations and articles is provided to cover novel topics that are not covered by the textbook.

Semester

Semester II

Assessment method

The final evaluation consists of the aggregation of the scores obtained in two independent assessments.

- The first assessment is based on an exam-tailored project or a survey, carried out individually or in groups, and aimed at bringing the student to have an in-depth knowledge and/or hands-on experience of a specific topic covered in the course or linked to topics covered in the course; the project and the survey are both discussed through an oral presentation supported by slides lasting about 20 minutes; it is possible, during the presentation, to include a short demo of the project; the survey consists of a bibliographic review on a topic, in which the student discusses and compares proposed solutions in the state of the art to a specific problem of interest for him. The evaluation is based on: significance of the project with respect to the topics covered in the course, methodological soundness (within the limits of what is reasonable to ask for an exam project); mastery of the in-depth topic demonstrated during the oral presentation.
- The second assessment is based on the verification of the knowledge acquired by the student about the topics addressed during the course in one of the following ways, freely chosen by the student:

1. oral exam taken in conjunction with the discussion of the first evaluation;
2. two ongoing tests consisting of exercises and open questions: one related to the topics covered in the first part of the course (knowledge graph, ontologies, RDF, RDFS, OWL), and one related to the topics covered in the second part of the course (data and vocabulary integration, information extraction, models for data and knowledge exploration).

In the Covid-19 emergency period, if it is not possible to carry out written exams in the presence, the ongoing tests will be replaced by assignments to be carried out individually and verified during the oral exam.

During the Covid-19 emergency period, oral exams will be online only.
They will be hosted using the WebEx platform and a public link will be published on the e-learning page to grant access to virtual spectators.

Office hours

Thursday 14.30-15.30