Information Integration

Projects

Marco Console

Master of Science in Engineering in Computer Science

Sapienza, University of Rome

Academic Year 2020/2021

Information Integration – A Quick Overview

Information Integration

- Information integration is the process of reconciling the information coming from multiple sources.

- The mismatch between sources can be (mainly) of two kinds
  - **Syntactic heterogeneity**: data are stored in different formats.
    - Different data types or data models.

  - **Semantic heterogeneity**: data talk about different aspects of the same reality.
    - Different information.

- How do we reconcile data?

Information Integration Systems -- Intuition

- At the formal level we make some simplifying assumptions.

- Data sources can be represented as a single relational database.
  1. No syntactic heterogeneity between sources
  2. A single set of predicates define the **source schema**

- The result of the integration takes the form of a relational database.
  1. No syntactic heterogeneity between sources
  2. A set of predicates define the target schema
  3. A set of constraints on the schema specifying further constraints

Information Integration Systems -- Syntax

- An information integration system is a triple $\langle G, M, S \rangle$
- G is the **global schema**
  - A logical theory over a relational alphabet A_G that describes how data resulting from the integration process should look like.
- S is the **source schema**
  - A logical theory over a relational alphabetA_S (disjoint from A_G) that describes data sources.
- M is the **mapping between S and G**
  - A set of pairs $\langle q_S, q_G \rangle$ where q_S is a query over A_S and q_G is a query over A_G .

Information Integration Systems -- Semantics

- Assume an information integration system $\mathrm{J} = \langle G, M, S \rangle$ and a relational database D that satisfies S.

- The semantics of J with respect to I are defined as follows:

$$sem^D(J) = \{C \mid C \vDash G \; and \; q_S^D \subseteq q_G^C \; for \; each \; \langle q_S^D, q_G^C \rangle \in M\}$$

- This semantics is based on **sound mappings**
  - Different semantics for mappings are possible.

Practical Information Integration

- The theoretical perspective of information integration hides several fundamental aspects of an information integration process.

1. Data sources usually are syntactic heterogenous
   - How do we practically reconcile the different data formats and data models?
   - What format should the result of the integration take?

2. How do we represent the result of the integration?
   - What do we want to do with the data? Query? Data management?
   - Data at the sources change? Are we interested in fresh data?

Projects

Projects -- Overview

- The project will require you to integrate data from one or more sources.
- **Starting point**: One or more data sources (databases, CSVs, excel sheets)
- **Goal**: Provide a reconciled view of the data.
  - Define a task e perform the integration of the data in such a way to allow the execution of the task.
- The project should (broadly) consist of three distinct phases.

Phase 1 – Data Gathering

- Define a domain of interest
  - Something that you like and won't mind spending some time on.

- Collect data about the domain of interest.
  - **Where** do I take the data?
    - Kaggle, USA data portal, European data portal...

- **Define a task that you want to perform concerning the domain.**
  - What kind of useful insight can I get from the data?

Phase 2 – Modelling

- Define a data integration system describing the domain of interest.
  - Describe how the source data should look like using the **source schema**.
  - Describe the result of the integration process using the **target schema.**
  - Reconcile source and target schema using the **mapping layer**.
  - Define queries on the target schema to perform the tasks you identified.

- For this step, you have to use the formal tools seen in the course.
  - First order logic.
  - More expressive query languages?

Phase 3 – Implementation

- Convert the output of the previous two phases into a software!
- You can use a free data integration tool (**preferable**)
  - Tallend, Pentaho, IBM Infosphere
- Or you can write your own code
  - Python, Java
- You need to be able to actually perform the tasks defined in Phase 1

Projects – Expected Outcome

- For the exam, you will be asked to report on your project.

- Prepare a 15-20 minutes presentation detailing the following.
  1. The scenario of interest, the data you collected, and the tasks you identified.
  2. The information integration system you defined
  3. How you implemented the output of the previous phases into a software

- Prepare a short demo of your software
  1. Show that you can actually perform the integration task you defined.

- Be prepared to answer instructor's questions on the topics of the course
  1. How your project connect with the concepts we see in the course

Projects – Practical Steps

1. Form a group of 1-2 students.
2. Choose data, a set of tasks and a tool for data integration (ETL).
3. **Contact the instructor for approval.**
4. If 3 succeeds, go ahead and prepare the material.
5. When ready, book an appointment with the instructor for the discussion.