Automating Multidimensional Design from Ontologies

Oscar Romero

Alberto Abelló

Universitat Politècnica de Catalunya (UPC)

Outline

- Introduction & Motivation
- Our Proposal
- Method Foundations
- The Method
- Conclusions & Further Work

Introduction & Motivation (I)

- Many methodologies and approaches have been presented to design multidimensional Data Warehouses in the literature.
  - Most of them are completely carried out manually.
- In the last years, some efforts have tried to automate the design of multidimensional databases.
  - Mainly, these approaches start from a detailed analysis of the data sources to determine the multidimensional concepts in a reengineering process.

Introduction & Motivation (II)

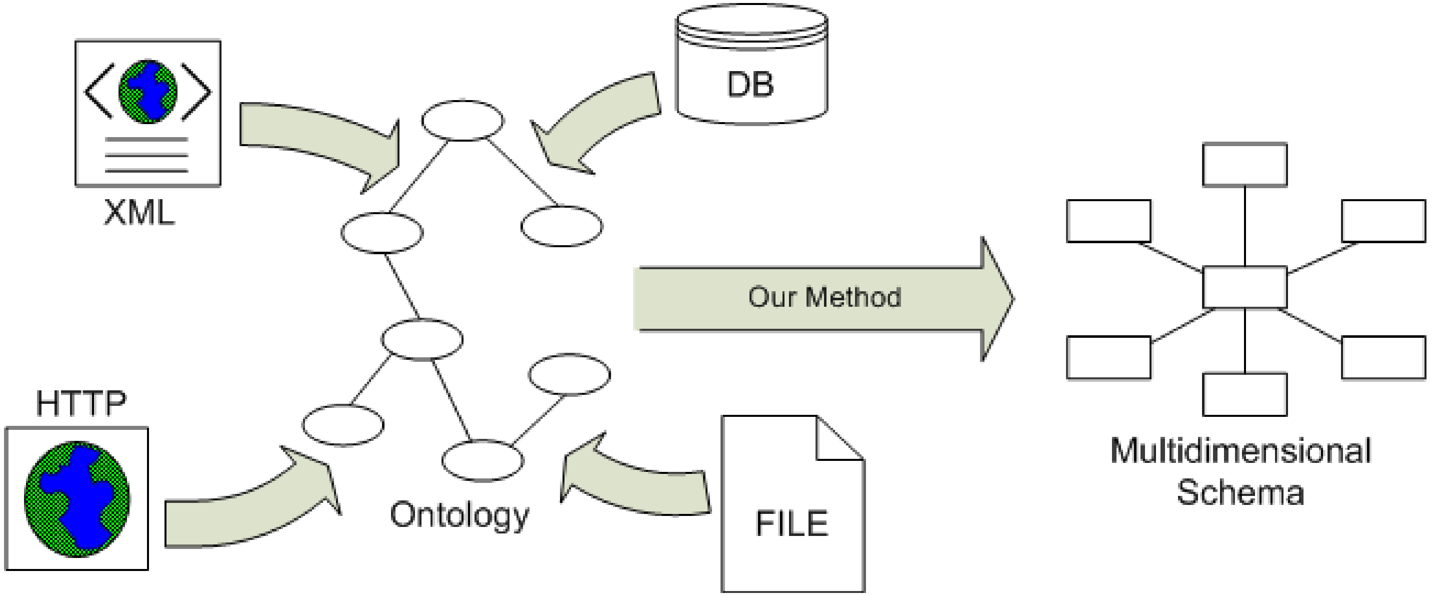
- Automated methods introduced in the literature share three main restrictions:
  - They exclusively work over relational sources.
  - The complexity of the source schemas must not be high.
  - They mainly work with a *table granularity*. Consequently, they require a certain degree of normalization to work properly.

Our proposal (I)

- We propose a semi-automatable method aimed to find the business multidimensional concepts from an ontology representing our business domain.
  - Our input ontology may represent different and potentially heterogeneous data sources.
- This method will point out our business multidimensional concepts contained in data sources of our domain having nothing in common but that they are all described by the same domain ontology.

Our proposal (II)

- Depicting our proposal:



Our proposal (III)

- To our knowledge, this is the first method addressing this issue from ontologies, and in general, from non-relational sources.
- It opens new perspectives of work:
  - For instance, we can extend the DW and OLAP concepts to other areas like the Semantic Web, where ontologies play a key role to provide a common vocabulary.
  - One consequence would be that we would be able to integrate external data from the web into our DW to provide additional up-to-date information about our business domain.

Our proposal (IV)

- Main consequences we must bear in mind:
  - We cannot assume anymore that data sources are implemented over relational databases: we need to focus on the input ontology representing our data sources.
  - Ontologies are semantically richer than relational schemas metadata and therefore, our design process will be guided by knowledge contained in the input ontology.
    - However, in some cases, we may need to extract missing knowledge by means of data samples.

Method Foundations (I)

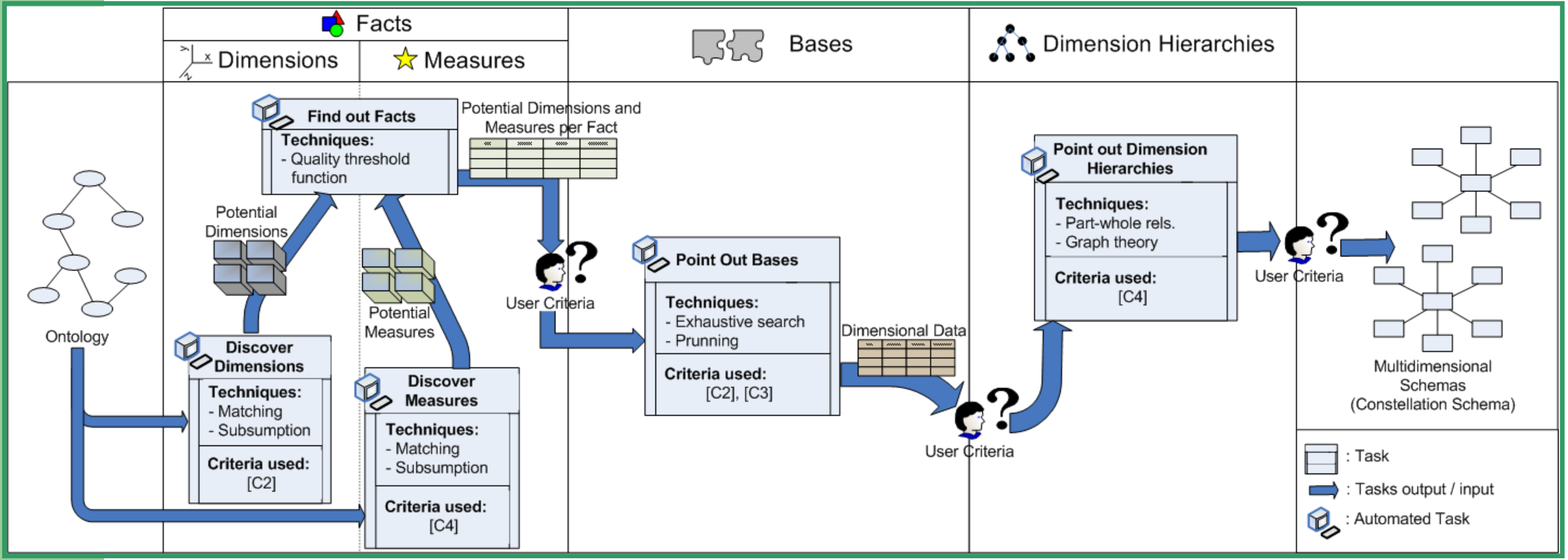
- Multidimensionality pays attention to two main aspects; placement of data in a multidimensional space and correct summarizability of data. Therefore, our method looks for meaningful conceptual schemas with orthogonal **Dimensions** fully functionally determining **Facts** and free of summarizability problems.

Method Foundations (II)

- [C1] **The Multidimensional Model** (**Fact** / **Dimension** dichotomy).
- [C2] **The Multidimensional Space Arrangement Constraint**: Each instance of data is identified (i.e., placed in the multidimensional space) by a point in each of its analysis **Dimensions**.
- [C3] **The Base Integrity Constraint**: We denote by **Base** a minimal set of **Levels** functionally determining a **Fact**.
  - Moreover, **Dimensions** giving rise to a **Base** must be *orthogonal.*
- [C4] **The Summarization Integrity Constraint**: Data summarization performed must be correct guaranteeing:
  - *Disjointness: The set of objects to be aggregated must be disjoint.*
  - *Completeness: The union of subsets must constitute the entire set.*
  - and *Compatibility* of the **Dimension**, kind of **Measure** being aggregated and the aggregation function.

The Method (I)
At first sight

- Overview of the method presented:



The Method (II)
Pointing out Facts

- We consider a concept to be a potential subject of analysis if it is related to as many potential **Dimensions** and **Measures** as possible.
  - Discover potential **Dimensions** of analysis and potential **Measures**.

The Method (III)
Pointing out Dimensions

- According to [C2], a concept is a potential **Dimension** of analysis if it is related to a **Fact** by a one-to-many relationship; that is, every instance of data is related to one, and just one, of its instances.
- We have formalized this in a DL multidimensional pattern.
  - We are looking for concepts (*D*) such that every instance of a given **Fact** (*F*) is related, directly or by composition through a set of properties (*r*), to, at least and at most, one of its instances.

The Method (III)
Pointing out Measures

- Typically, **Measures** are numeric attributes allowing data aggregation. In our method, we consider any numeric datatype to be a **Measure** of a given **Fact** *F* if, according to [C4], it preserves a correct data aggregation from *F*.
- We use a similar multidimensional pattern to the one presented to identify **Dimensions**.
  - Essentially, it is the same idea. Thus, this algorithm raises the same computational complexity.

The Method (IV)
Looking for potential Bases

- This step points out potential **Bases** for each identified **Fact** among those concepts labelled as its potential **Dimensions** of analysis.
- According to [C3], we need to find a set of **Dimensions** identifying the **Fact** univocally. Moreover, **Bases** must contain orthogonal **Dimensions**, and a set of potential **Dimensions** will be considered a feasible **Base** if they are able to identify all the instances of a Fact:

$$\prod |D_i| \geq |F|$$

- To do so, we need to work with multiplicities provided by the ontology or with data samples.

The Method (V)
Looking for potential Bases

- We need to combine **Dimensions** to point out those combinations identifying the **Fact**.
  - Naïve solution: All potential combinations… **Wrong**! Not only expensive but also inappropriate.
- Three heuristics to prune the search space:
  - If *B* results to be a **Base**, then, all Bases containing *B* are overlooked (since they are not minimal).
  - Two concepts (*a* and *b*) that have been proved to not be **Bases** of a given **Fact,** are generated (i.e. {*a, b*}) as a combination to be proved as a **Base** if *a* and *b* are orthogonal.
  - A given combination may be a potential **Base** of a given **Fact** if its *intermediate* **Bases** have been proved to be **Bases**.

The Method (VI)
Giving rise to Dimension hierarchies

- This step shapes **Dimension** hierarchies in order to allow summarizability of data.
- We look for to-one relationships (also known as "*Roll-up*" relationships) giving rise to hierarchies allowing a correct data aggregation.
- Starting from each concept identified as a **Dimension**, a directed graph following all to-one relationships paths is depicted.
- At this moment, in general, we cannot differentiate the role played by each graph node (i.e., concepts); either as a **Level** or as a **Descriptor**.
  - It is a design decision to spot each concept as an attribute of an existing **Level** or as a new **Level** within a **Dimension** hierarchy; giving rise to star or snowflake schemas.

Conclusions and Further Work (I)

- We have presented a semi-automated method to point out multidimensional concepts from an ontology representing our business domain.
- In our approach, we use ontologies as well as reasoning tools provided by ontology languages to look for multidimensional patterns.
- Up to now, traditional approaches were typically carried out manually and were restricted to work over relational sources. Conversely, our approach is able to integrate information from heterogeneous data sources describing their domain through ontologies.
  - One of the most promising areas where to apply our method is the Semantic Web, giving rise to new possibilities like extracting and integrating external data into our DW.

Conclusions and Further Work (II)

- We have justified the method feasibility by:
  - Carrying out a complete simulation of a real case study to validate our algorithms.
  - We have also presented an in depth theoretical study of the algorithms complexity.
- As further work:
  - We aim to consider **Bases** pointed out along the method to extract more flexible **Dimension** hierarchies,
  - Find out better and more accurate multidimensional patterns that may be implemented through logic reasoners.

Thanks for your attention

Questions?