M.V. Hurtado, L. Garcia and J. Parets
HTS Computer Engineering, University of Granada, Granada, Spain

Semantic Views over Heterogeneous and Distributed Data Repositories: Integration of Information System Based on Ontologies.

Abstract: Nowadays, in the global information infrastructure, the great expansion of the communication networks allows users to work with a great quantity of heterogeneous and distributed data, which are stored in different repositories. In order to support co-ordinated access to this data, preserving the correct semantics of the data stored in different ways, an integrated view is needed. We propose, in a first step, obtaining an integrated schema (Semantic View) and an integrated terminology in order to describe the content of the repositories. Ontologies are suggested as an interesting semantic representation which can be used here. Next, in a second step, the integrated schema and integrated terminology must be enriched in order to allow us to manage the intrinsic evolutionary aspects which characterise the modelling process of global information systems.

I. Introduction

In recent years the development of the communication networks allow users to work with a great quantity of data repositories. These repositories present different structures, organisations, query languages and data semantics. A partial solution to the problem of lack of uniformity when dealing with the available repositories consists on defining new information retrieval techniques with a strategy that focuses on information content and semantics.

In the second section, we will describe some reasons that justify the use of Semantic Views over the repositories. In the third section, as a result of research carried out in the last decade, Ontologies are suggested as an interesting semantic representation. Following the approach of previous papers on software evolution (Parets and Torres, 1996), (Parets et al., 1999) and (Rodriguez et al., 2000) and on information integration (Hurtado and Parets, 2001), the fourth section of the paper presents the possible evolution mechanisms of the semantic views. Additionally, effective graph mechanisms for semantic evolution are provided.

2. Semantic Views: Implications and Needs

Semantic Views on specific domains minimise the problem of semantics associated with stored data and hide all the technical and organisational details associated with data. Furthermore, a logical schema of shared information is obtained which provides a logical description of the shared data, allowing application programs and databases to interoperate without having to share data structures. Indeed, in order to provide definitions for the vocabulary used to represent knowledge, a richer semantically integrated terminology is obtained too. In order to do this, the formalisms used for Semantic Views must allow:
• A common understanding of the structure of information among people or software agents.
• The reuse of domain knowledge.
• To make domain assumptions explicit.
• A separation of domain knowledge from the operational knowledge.
• An analysis of domain knowledge.

Typically, Semantic Views can be described by means of different formalisms (systems based on Description Logic (Devanbu, 1994)), terminological systems, ontologies and so on).

We advocate here for using ontologies, because they have moved beyond the domain of library science, philosophy and knowledge representation. Ontologies are used as central controlled vocabularies that are integrated into catalogues, databases, web publications, knowledge management applications, etc. and large ontologies are essential components in many online applications including retrieval (such as Yahoo and Lycos) and e-commerce (Such as Amazon and eBay) (McGuinness et al., 2000).

3. Using Ontologies

In the context of knowledge sharing, the term ontology is used as an explicit specification of a conceptualisation (Gruber1995). People typically have some notion of the meaning of the term, one of the simplest notions of an ontology may be a controlled vocabulary (i.e., a finite list of terms). Catalogues are an example of this category, they can provide an unambiguous interpretation of terms. Another potential ontology specification is a glossary (a list of terms and meaning) and also a thesaurus, because they provide additional semantic relationships among terms (such as synonym or hierarchical relationships).

Furthermore, because ontologies have to express more information about shared conceptualisation and also conceptual frameworks for modelling domain knowledge, we will require that an ontology holds the following properties in order to consider something as an ontology:

- Finite, controlled and extensible vocabulary
- Unambiguous interpretation of class and term relationships
- Hierarchical subclass relationships between classes
- Specification of arbitrary logical relationships between terms.

In this sense, we propose a specialised kind of ontologies based on the Object-Oriented Approach which satisfies the previous properties. In the O-O approach, classification is a core concept (Rumbaugh, 1991). The objects which share data structure and behaviour are grouped into classes, and the classes can be sub-classified by means of inheritance mechanisms (i.e. in figure 1, Author denotes a class and the class definition in natural language can be the following "An author is a person who writes things. An author must have created at least one document. In this ontology, an author is known by his or her name"). This implies a high degree of abstraction, which describes important properties and ignores irrelevant ones, and this is a conceptual process which is independent of the programming languages (i.e. the sentence gives necessary conditions on class membership, says that authors must also be persons or says that authors must have exactly one associated name. This means that the relation author.name maps every instance of
the class Author to the same name, with only one name per author). This process is usual in human knowledge acquisition and allows abstract concepts to be expressed.

The classification schema in O-O is always explicit, for instance, figure 1 shows a partial enriched view of the possible UML class diagram based on the Stanford-I ontology (http://www.ksl.stanford.edu) about bibliography data which can be considered as special ontology. In addition O-O classification implies a description of the static structure of the classes of the system and their relationships (associations, specialisation, generalisation, etc).

Previous characteristics increase maintainability in ontologies based on the object model because:

- Explicit classification facilitates the introduction of new classes and the re-structuring of previous ones.
- Inheritance allows code reusability (the code of the behaviour of a class is reused by its subclasses), code sharing (different users and systems can share the same classes), interface consistency (inheritance guarantees that the inherited behaviour is the same for the subclasses and that the objects of the subclasses interact in a very similar way) and rapid prototyping (the classes developed in previous systems can be reused and refined).
- Information hiding allows the code of the behaviour of a class to be changed without changes in the uses of the class.

4. Evolutionary aspects in the integrated schema

The use of semantic views based on Ontologies allows, on the one hand, the flexible encapsulation of data in repositories and, on the other hand, the sharing of information between different users or systems. Many disciplines develop standardised ontologies that domain experts can use to share and annotate
information in their fields. An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. We have proposed that ontologies be described using an object-oriented approach and therefore using graphical representation and graph restriction.

This paper shows the application of our experience in graph evolution to semantic views, specially the idea of defining semantic restrictions over the relationships represented in the graph. Finally, once again our approach provides mechanisms to propagate changes, and it focuses on the evolutionary nature of Global Information Systems.

There is an additional and important fact that cannot be forget, the intrinsic evolutionary aspects, which characterise the modelling process of global information systems. That is to say changes in the underlying repositories or in the Semantic View may produce a propagation of changes in the rest of the global information system.

A previous paper (Hurtado&Parets2001) shows that certain advantages in the specification of the evolution of semantic views can be obtained when advanced object-oriented techniques, such as dynamic classification, multiple inheritance, etc. is used. Figure 2 shows that the class Proceedings can maintain inheritance relationships with Document and Conference, this fact it's not allowed in classical formalisation of ontologies (Guarino, 1998)).

These techniques include the notions of identity, classification, polymorphism and inheritance, which provide interesting ways of organising the objects and their activity and improve the evolutionary nature of semantic views.

Moreover, in order to provide concrete mechanisms of change, we propose that the previous ontology (figure 1) can be represented as heterogeneous graph where the nodes are classes (ellipses in the graph) and arcs are different types of relationships between them (associations, specialization, generalization, etc.) or relationships defined by users (write, Read-in etc) (see figure 2).

Additionally, this kind of graph has not only special semantics adopted from O-O relationships but also some implicit and explicit restrictions for these relationships (for instance, the relationship KindOf always verifies some restrictions such as a-cyclic restriction, anti-symmetric property, and it is incompatible with PartOf and with Is-a). The relationships defined by the user and their restrictions can also be explicitly defined.

We are now researching on practical ways of specifying these graphs. In previous papers (García and Parets,2000) and (García-Cabrera et al.,2001) we presented a set of basic operations on the graph (such as Create_node, Del_node, Create-arc, Del_arc, Nodes-connected-to, Connection-by, etc.) and a set of basics restrictions on the graph (tree, acyclic_graph, weak_connected, etc.), restrictions on the nodes (i.e. each node must have a label, unique name assumption, etc.) and restrictions on the arcs (have a label, anti-reflexive, anti-symmetrical, transitive, a-cyclic, incompatible with, incompatible transmission, and so on).

These operations allow that the structure of the graph to be changed, and the set of restrictions on the graph, nodes and arcs helps in propagating the changes because, the preconditions of each operation must be checked before a change is made. This fact implies that propagation of changes can be represented and automated.
For instance, figure 3 shows how the semantic view evolves as a result of introducing new classes and specialisation relations among them, this fact causes, for example, a change propagation in the mapping rules. Figure 3 shows the new class Best-seller which is a Trade-Book with more than 1000 copies sold and the prohibition of adding a new KindOf relation between Book and MiscellaneousPublication as result of the acyclical restriction. More complex changes, as the introduction of intermediate new classes or the deletion of classes could be introduced.

In order to make changes in the integrated information, a graphical approach has important advantages. From a formal point of view, to check the preconditions of operations and restrictions on the graph is easier for propagation changes. In addition abstract data types or classes can be used for implementation. Therefore, from a user point of view, this approach is easier to understand because it encodes the shared information by means of a simple graph with the labelled concepts, relationships and restrictions.

5. Conclusion

The use of semantic views based on Ontologies allows, on the one hand, the flexible encapsulation of data in repositories and, on the other hand, the sharing of information between different users or systems. Many disciplines develop standardised ontologies that domain experts can use to share and annotate information in their fields. An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. We have proposed that ontologies be described using an object-oriented approach and therefore using graphical representation and graph restriction.

This paper shows the application of our experience in graph evolution to semantic views, specially the idea of defining semantic restrictions over the
relationships represented in the graph. Finally, once again our approach provides mechanisms to propagate changes, and it focuses on the evolutionary nature of Global Information Systems.

![Diagram of relationships](image)

**Figure 3 Changes in the graph**

**Notes**
1. This research is partially supported by a R+D project of the Spanish CICYT (TIC2000-1673-C06-04).

**References**


