Querying Heterogeneous Spatial Databases: Combining an Ontology with Similarity Functions

Mariella Gutiérrez[1] and Andrea Rodríguez[2]

[1] School of Engineering, Universidad Católica de la Santísima Concepción
Caupolicán 490, Concepción, Chile
marielag@ucsc.cl
[2] Department of Computer Science, Universidad de Concepción
Edmundo Larenas 215, Concepción, Chile
andrea@udec.cl

Abstract. This paper uses a knowledge-based approach to querying heterogeneous spatial databases based on an ontology and conceptual and attribute similarities. The ontology, which may be independent of the databases, expands and filters a user query. Then, queries are translated into a formal specification of entity classes, which are compared against definitions in databases. This process is carried out by determining the conceptual similarity between entities in a user ontology and by comparing these entities in the ontology with entities in the conceptual models of databases. In addition, the specification of a query is done not only by identifying entity classes but also by considering constraints based on attribute values. The paper describes the system architecture and presents a case study with data from a forestry information system.

1 Introduction

This paper presents a system architecture for accessing information across heterogeneous spatial databases based on a user ontology and similarity functions. The focus of the paper is at the semantic level, where the ontological definitions of geographic features are independent of their geometric representations.

Studies that use an ontology for data integration require that databases subscribe to a common ontology, which is similar to subscribing to a shared schema at the schematic level. This common ontology is obtained by a single ontology or by the integration of multiple and independent ontologies [2, 17–19, 25, 29]. This work, in contrast, relaxes this strategy of using a common ontology, since it does not force databases either to subscribe to a common ontology or to have a complete semantic description of their information content. The approach of this work is to use semantic similarity measures to associate dynamically entities from different conceptualizations while maintening these conceptualizations independent [12].

This work follows and extends ideas from [12–14] that define similarity functions between ontologies and between ontologies and databases. Unlike these

S. Wang et al. (Eds.): ER Workshops 2004, LNCS 3289, pp. 160–171, 2004.

previous works, in this paper we define a mechanism that retrieves data from heterogeneous databases based on the identification not only of entity classes, but also of instances that are similar to a user request. This work assumes that each database has a conceptual schema. The use of the logical schema was explored in [14], but this approach has strong limitations respect to the description of the information content of a database. Conceptual schemas and the user ontology are expressed in OWL, a standard language for the definition of Ontologies in the Semantic Web [3, 15].

The organization of this paper is as follows. Section 2 reviews related work about querying heterogeneous databases. Section 3 describes the system architecture followed by Section 4 that addresses the description of the user ontology and conceptual schemas of databases. Section 5 adapts similarity functions of previous works [12–14] to evaluate similarity within the user ontology and between the ontology and conceptual schemas. A case study in the area of a forestry information system illustrates the access to databases in Section 6. Conclusions and future work are presented in Section 7.

2 Related Work on Querying Heterogeneous Data Repositories

Many studies have treated the problem of accessing independent databases as a problem of solving heterogeneities among these databases. Focusing on semantic heterogeneities, studies have proposed the use of ontologies to specify queries and describe the content information of databases [2, 8, 18, 19]. In current ontology-based information systems, semantic matching has meant the agreement on the vocabulary used by different agents. This implies sharing the same conceptualization or agreeing to adopt a common conceptualization, which is usually the intersection of the original conceptualizations [10, 11]. Consequently, the general approach to handling semantic heterogeneity has been to map the local terms in a database onto a shared or common ontology. Most of these approaches use the terms interrelationships to determine semantic similarity between concepts [4–6]. Other approaches are measures based on graph matches and probabilistic measures that predict the probability that an instance of a concept in a differentiated ontology will satisfy a request [30].

In environments with multiple and independent information systems, however, each system may have its own conceptualization and, therefore, its own intended model or ontology. Nonetheless, if existing ontologies are well defined, their integration may reduce the cost of building a global ontology from scratch [2, 16]. Ontology integration is a complex problem, because concepts can overlap or definitions of concepts may be inconsistent across ontologies [27]. Some systematic approaches to handling ontology integration are composition algebras [21], lexical interrelations [2, 18, 19], mappings with mediator agents [22], inheritance from top-level ontologies [8], and semantic correspondence that relies on a common vocabulary for defining concepts across different concepts [25, 29]. All of these approaches are manual or semi-automatic, requiring some input from domain experts.